Statistical, geometrical and logical independences between categorical variables

Julien Ah-Pine, Jean-François Marcotorchino

Statistical, geometrical and logical independences between categorical variables

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. Classical association criteria, used for measuring statistical independence between categorical variables, are initially defined using contingency tables. There is another way for representing categorical variables : Relational Analysis which uses binary pairwise comparison matrices formalism. There exists correspondance formulas that enable to get from one representation to the other. By using these formulas, and these two representations, we can have a better understanding of the main differences between some famous association criteria. In fact, several types of independence, namely statistical, geometrical and logical, appear using one representation or the other. The aim of this paper is to present in a unified framework, these different kinds of independence and their relationships by studying the expression of the following association criteria in the two different representations : Belson, Lerman, χ^2 of Tchuprow, Jordan, Rand and Janson and Vegelius. This paper is based upon previous results obtained in [Marcotorchino, 1984], [Messatfa, 1989], [Marcotorchino and El Ayoubi, 1991], [Najah Idrissi, 2000].
Keywords: Relational Analysis, Association criteria, Independences, Nominal categorical variables.

Relational Analysis (RA) is concerned with the analysis of binary relations and their applications in different mathematical fields [Marcotorchino, 2006]. This approach represents the binary relations as pairwise comparison matrices, and it is basically related to different tools 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.

Here, we are particularly interested in the applications of RA in the analysis of association criteria. Previous work has been done in this area and different association criteria have been partially unified under the concept of geometrical independence using RA [Marcotorchino, 1984], [Najah Idrissi, 2000]. Our aim is to recall these results and to go further by adding other association criteria and by underlying another independence concept, called indetermination which is based on a logical approach [Marcotorchino, 1984].

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)

1 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 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), Lerman (L), χ^2 of Tchuprow (T), Jordan (J), Rand (R), and Janson-Vegelius (JV). These criteria are initially defined using the contingency table. We recall their definitions. We precise that the Rand and the Lerman criterion we mention, is a modified version according to [Marcotorchino, 1984] and [Najah Idrissi, 2000].

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

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

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

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

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

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

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

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

RA is another way of representing nominal categorical variables. This representation uses pairwise comparison 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 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, the studied binary relations are equivalence relations (or partitions) and we have for V^k :

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

Following Kendall [Kendall, 1970], Marcotorchino has developped correspondance formulas between contingency and relational representations. Using these correspondance formulas we can express the mentionned association criteria using the relational coding [Marcotorchino, 1984].

We give in Table 1 the classical correspondance formulas [Marcotorchino, 1984].

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

Table 1. Correspondance formulas between contingency representation and relational representation

In [Najah Idrissi, 2000], it is given the symmetric relational expression for Rand, Janson-Vegelius, Lerman, and Tchuprow criteria. We extend these results by giving the symmetric expression of Belson and Jordan criteria.

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

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

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

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

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

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

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 Rand criterion :

$$2R(C^k, C^l) - 1 = \frac{\sum_{i,i'} (C^k_{ii'} - \frac{1}{2})(C^l_{ii'} - \frac{1}{2})}{\sqrt{\sum_{i,i'} (C^k_{ii'} - \frac{1}{2})^2 \sum_{i,i'} (C^l_{ii'} - \frac{1}{2})^2}}$$

2 From statistical independence to geometrical independence

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

$$V^k \perp_S V^l \Leftrightarrow \frac{\mathbf{n}^{kl}_{uv}}{N} = \frac{\mathbf{n}^{kl}_{u.}}{N} \frac{\mathbf{n}^{kl}_{.v}}{N} \quad \forall (D^k_u, D^l_v) : D^k_u \in V^k, D^l_v \in V^l \qquad (1)$$

According to this principle, we can measure the association (or the relationship) between two nominal categorical variables by measuring their de-

viation from the statistical independence situation. In the contingency representation, many of the presented association criteria are based upon this approach. Indeed, we clearly see, that the Belson, the Lerman, the Tchuprow and the Jordan criteria are null if V^k and V^l are statistically independent.

In the relational representation, we can interpret each relational matrix as a binary tensor in a tensor space of canonical basis $\{e_i \otimes {}^t e_{i'}; i, i' = 1, \ldots, N\}$ where e_i is the canonical vector and $\otimes$ the Kronecker product.

Consequently, we can interpret the association criteria expressed in their relational representation, from a geometrical point of view.

More precisely, let $\langle C^k, C^l \rangle_F$ be the Frobenius scalar product derived from the Frobenius norm of a matrix. We have :

$$\langle C^k, C^l \rangle_F = \sum_{i=1}^{N} \sum_{i'=1}^{N} C^k_{ii'} C^l_{ii'} = \text{Trace}({}^t C^k \cdot C^l)$$

We can show that, using the relational coding, one can express the main different association criteria that we have recalled, as particular cases of a general Bravais-Pearson like correlation coefficient between relational matrices [Marcotorchino, 1984], [Marcotorchino and El Ayoubi, 1991], [Najah Idrissi, 2000] :

$$\begin{aligned} \Delta(C^k, C^l, f, \mu^k, \mu^l) &= \frac{\sum_{i,i'} \big(f(C^k_{ii'}) - \mu^k\big)\big(f(C^l_{ii'}) - \mu^l\big)}{\sqrt{\sum_{i,i'} \big(f(C^k_{ii'}) - \mu^k\big)^2 \sum_{i,i'} \big(f(C^l_{ii'}) - \mu^l\big)^2}} \\ &= \frac{\langle f(C^k) - \mu^k U_N, f(C^l) - \mu^l U_N \rangle_F}{\sqrt{\langle f(C^k) - \mu^k U_N, f(C^k) - \mu^k U_N \rangle_F \langle f(C^l) - \mu^l U_N, f(C^l) - \mu^l U_N \rangle_F}} \end{aligned} \tag{2}$$

According to (2), we can see that the differences between the 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, the different values of the parameters (f, μ^k, μ^l), which define a particular coefficient. These different coefficients are related to the mentionned association criteria.

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

$$\begin{aligned} V^k \perp_G V^l &\Leftrightarrow \Delta(C^k, C^l, f, \mu^k, \mu^l) = 0 \\ &\Leftrightarrow \sum_{i,i'} \big(f(C^k_{ii'}) - \mu^k\big)\big(f(C^l_{ii'}) - \mu^l\big) = 0 \end{aligned} \tag{3}$$

The RA approach has enabled to get a better understanding of the main differences between the presented association criteria. We have seen that there exists a particular relationship between contingency representation /

$f(C_{ii'})$	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	The Belson criteria is the numerator of the defined coefficient
$f(C_{ii'}) = C_{ii'}$	$C^k_{..}/N^2$	$C^l_{..}/N^2$	Lerman	$L(C^k, C^l)$
$f(C_{ii'}) = C_{ii'}/C_{i.}$	$1/N$	$1/N$	Tchuprow	$T(C^k, C^l)$
$f(C_{ii'}) = C_{ii'}$	$C^k_{i.}/N$	$C^l_{i.}/N$	Jordan	The Jordan criteria is the numerator of the defined coefficient $\times 1/N$
$f(C_{ii'}) = C_{ii'}$	$1/2$	$1/2$	Rand	$2R(C^k, C^l) - 1$
$f(C_{ii'}) = C_{ii'}$	$1/p_k$	$1/p_l$	Janson-Vegelius	$JV(C^k, C^l)$

Table 2. Correspondance between correlation coefficient and association criteria

statistical independence and relational representation / geometrical independence.

We give in the following, two other results that strengthen this duality between the contingency and the relational representations. Firstly, we show the link between the modified Lerman criterion and the tetrachoric correlation coefficient. Secondly, we establish a specific statistical / geometrical independence duality between the Belson criterion and the Janson-Vegelius (its numerator) criterion.

In the relational representation, 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 3.

	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 3. Agreements and disagreements between relational matrices

Considering Table 3, we can express the statistical independence concept by using another measure called the odds-ratio (OR). Then, we also say that two nominal categorical variables are statistically independent if we have the following relation :

$$\begin{aligned} V^k \perp_S V^l &\Leftrightarrow OR(C^k, C^l) = 11_{kl}00_{kl}/10_{kl}01_{kl} = 1 \\ &\Leftrightarrow 11_{kl}00_{kl} - 10_{kl}01_{kl} = 0 \end{aligned} \tag{4}$$

Indeed, using the correspondance formulas we can show that :

$$\begin{aligned} 11_{kl}00_{kl} - 10_{kl}01_{kl} &= \sum_{u,v} (\mathbf{n}^{kl}_{uv})^2 - \frac{\sum_u (\mathbf{n}^{kl}_{u.})^2 \sum_v (\mathbf{n}^{kl}_{.v})^2}{N^2} \\ &= \sum_{i,i'} \left(C^k_{ii'} - \sum_{i,i'} \frac{C^k_{ii'}}{N^2} \right) \left(C^l_{ii'} - \sum_{i,i'} \frac{C^l_{ii'}}{N^2} \right) \end{aligned}$$

Actually, we have the following identity :

$$\begin{aligned} L(C^k, C^l) &= \text{Tetrachoric}(C^k, C^l) \\ &= \frac{11_{kl}00_{kl} - 10_{kl}01_{kl}}{\sqrt{(11_{kl}+10_{kl})(01_{kl}+00_{kl})(11_{kl}+01_{kl})(10_{kl}+00_{kl})}} \end{aligned}$$

We consider now the Belson criterion and the Janson-Vegelius criterion's numerator. We establish a particular relationship between these two criteria : on the one hand the Belson criterion, in its contingency representation, is based on the statistical independence and in the relational presentation, it is based on a geometrical independence associated to the Torgerson[1] transformation; on the other hand, the Janson-Vegelius criterion's numerator, in its

[1] $C^k_{ii'} - \frac{C^k_{i.}+C^k_{.i'}}{N} + \frac{C^k_{..}}{N^2} = \langle O^i_k - G_k, O^{i'}_k - G_k \rangle$. $\{O^i_k; i = 1, \ldots, N\}$ are $(p_k \times 1)$ binary vectors where $[O^i_k]_u = K^k_{iu}; u = 1, \ldots, p_k$ and $G_k = \sum_{i=1}^N O^i_k$

contingency representation, is based on the geometrical independence associated to the Torgerson transformation, and in the relational representation it is based on the statistical independence upon equiprobability assumption.

We represent this relationship in Table 4.

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

Table 4. Dual relationship between Belson and Janson-Vegelius criteria due to the contingency / relational presentation duality

Firstly, considering the Janson-Vegelius criterion's relational representation, we can interpret the term $(C_{ii'}^{k}-\frac{1}{p_k})$, as a deviation from statistical independence 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 assume moreover, that the different classes $D_u^k; u=1,\ldots,p_k;$ are equiprobable. This involves 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 :

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

Secondly, considering the Janson-Vegelius criterion's contingency presentation, we can interpret the term $(\mathbf{n}_{uv}^{kl}-\left[\frac{\mathbf{n}_{u.}^{kl}}{p_l}+\frac{\mathbf{n}_{.v}^{kl}}{p_k}-\frac{\mathbf{n}_{..}^{kl}}{p_kp_l}\right])$, 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

$\mathbf{n}^{kl}_{uv} = \langle D^k_u, D^l_v \rangle$ and the following relation :

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

where $G^k = \frac{1}{p_k} \sum_{D^k_u \in V^k} D^k_u$ and $G^l = \frac{1}{p_l} \sum_{D^l_v \in V^l} D^l_v$.

3 A logical independence approach called "indetermination"

Among the previous studied association criteria, the one which is related to the Rand criterion and given by the paramaters ($f = Id, \mu^k = 1/2, \mu^l = 1/2$), is a special case. Although it has through its relational coding, an interpretation based on the geometrical independence, we can also interpret it using a logical approach which is called indetermination [Marcotorchino, 1984]. We say that two nominal categorical variables are indetermined or logically independent if we have the following relation :

$$\begin{aligned} V^k \perp_L V^l &\Leftrightarrow 11_{kl} + 00_{kl} - 10_{kl} - 01_{kl} = 0 \\ &\Leftrightarrow \textstyle\sum_{i,i'} C^k_{ii'} C^l_{ii'} + \sum_{i,i'} \overline{C}^k_{ii'} \overline{C}^l_{ii'} - \sum_{i,i'} C^k_{ii'} \overline{C}^l_{ii'} - \sum_{i,i'} \overline{C}^k_{ii'} C^l_{ii'} = 0 \\ &\Leftrightarrow \textstyle\sum_{i,i'} \left(C^k_{ii'} - \overline{C}^k_{ii'} \right) \left(C^l_{ii'} - \overline{C}^l_{ii'} \right) = 0 \\ &\Leftrightarrow 4 \textstyle\sum_{i,i'} \left(C^k_{ii'} - 1/2 \right) \left(C^l_{ii'} - 1/2 \right) = 0 \end{aligned} \tag{5}$$

This corresponds to the situation where, for two nominal categorical variables, the number of agreements given by $11_{kl} + 00_{kl}$ is the same as the number of disagreements given by $10_{kl} + 01_{kl}$. The statistical independence defined by (4) is a multiplicative model associated to the number of agreements and disagreements. On the contrary, the indetermination or logical independence defined by (5), is an additive model which is a different way of measuring the association between two nominal categorical variables.

The indetermination concept can also be extended by giving weightings to agreements and disagreements :

$$V^k \perp_L V^l \Leftrightarrow \mu^k_1 \mu^l_1 11_{kl} + \mu^k_0 \mu^k_0 00_{kl} - \mu^k_1 \mu^l_0 10_{kl} - \mu^k_0 \mu^l_1 01_{kl} = 0 \tag{6}$$

We will consider the following general formula which gives a normalized coefficient that measures the deviation from the indetermination's situation between two categorical variables. This coefficient denoted Λ is null in the case of indetermination :

$$\Lambda(C^k, C^l, \mu^k_1, \mu^k_0, \mu^l_1, \mu^l_0) = \frac{\sum_{i,i'} \left(\mu^k_1 C^k_{ii'} - \mu^k_0 \overline{C}^k_{ii'} \right) \left(\mu^l_1 C^l_{ii'} - \mu^l_0 \overline{C}^l_{ii'} \right)}{\sqrt{\sum_{i,i'} \left(\mu^k_1 C^k_{ii'} - \mu^k_0 \overline{C}^k_{ii'} \right)^2 \sum_{i,i'} \left(\mu^l_1 C^l_{ii'} - \mu^l_0 \overline{C}^l_{ii'} \right)^2}} \tag{7}$$

Finally, we give the relation below which shows the formal link between geometrical independence and indetermination in the relational representation :

$$\Lambda(C^k, C^l, \mu_1^k, \mu_0^k, \mu_1^l, \mu_0^l) = \Delta\left(C^k, C^l, Id, \frac{\mu_0^k}{\mu_1^k + \mu_0^k}, \frac{\mu_0^l}{\mu_1^l + \mu_0^l}\right) \qquad (8)$$

For example, when $\mu_1^k = \mu_0^k = \mu_1^l = \mu_0^l = \mu$ then we have $\Lambda(C^k, C^l, \mu, \mu, \mu, \mu) = \Delta(C^k, C^l, Id, 1/2, 1/2) = 2R(C^k, C^l) - 1$. Furthermore, when $\mu_1^k = (p_k - 1), \mu_0^k = 1, \mu_1^l = (p_l - 1), \mu_0^l = 1$, then we have $\Lambda(C^k, C^l, (p_k - 1), 1, (p_l - 1), 1) = \Delta(C^k, C^l, Id, 1/p_k, 1/p_l) = JV(C^k, C^l)$.

Finally, we have seen that using one representation or the other, we can have different concepts of independence for measuring the relationship between two categorical variables. It is important to have a good understanding of the differences between association criteria that are proposed in the litterature. RA has enabled to contribute to this issue by showing the relationship between contingency representation / statistical independence and relational representation / geometrical independence. But RA enables to go further by pointing out another kind of independence concept : indetermination between two categorical variables. This approach, that is inherent to the Rand[2] criterion, has a logical definition and can be interpreted as a voting rule. Furthermore, the RA method has enabled to define partitioning criteria using association criteria. This was defined by Marcotorchino, as the maximal association model for the clustering problem [Marcotorchino, 1986a], [Marcotorchino, 1986b].

References

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

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

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

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

[Janson and Vegelius, 1992]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.

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

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

[2] the Rand criterion is highly related to the Condorcet criterion which is originally used as a voting criterion [Marcotorchino, 1984]

[Marcotorchino and El Ayoubi, 1991]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.

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

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

[Marcotorchino, 1984]J.F. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingences partie I,II,III. *Etudes IBM F069 - F071 - F081*, 1984.

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

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

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

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

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

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

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

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