RAIRO. RECHERCHE OPÉRATIONNELLE

H. MESSATFA

Maximal association for the sum of squares of a contingency table

RAIRO. Recherche opérationnelle, tome 24, n° 1 (1990), p. 29-47

<http://www.numdam.org/item?id=RO_1990__24_1_29_0>

Recherche opérationnelle/Operations Research
(vol. 24, n° 1, 1990, p. 29 à 47)

MAXIMAL ASSOCIATION FOR THE SUM OF SQUARES OF A CONTINGENCY TABLE (*)

by H. MESSATFA (1)

Abstract. – *In this paper, we show how to approximate, in the quickest and the most realistic possible way, the maximum of the sum of squares of a contingency table* ($\sum_{u,v} n_{uv}^2$), *with fixed margins. The trivail case, where the margins are not fixed, corresponds to the matrix structure known as "Complete Association". For practical problems, no methods exist which guarantee an exact optimal solution. Bounds dues to mathematics inequality are proposed. We don't talk about the combinatory complexity of the problem; let us quote with regards to Hubert and Arabie* [3]*"... Constructing an exact bound, conditional on the fixed row and column totals of the given contingency table, is a very difficult problem of combinatorial optimization..." Lerman* [5] *proposed a recursive algorithm, which determine step by step an optimal solution, based on the notion of "points extremaux". Unfortunately the computing time increases exponentially. We shall propose two finer bounds that those proposed in the literature. The distribution* n_{uv} *corresponding to these bounds is not often reached. Then, we shall propose a very fast heuristic procedure, based on classical assignment techniques, to find such optimal distribution.*

Keywords : Measure of association; contingency table; maximal association.

Résumé. – *Dans cet article nous montrons comment approximer de la façon la plus rapide et la plus réaliste possible le Maximum de la somme des carrés d'une contingence* ($\sum_{u,v} n_{uv}^2$), *à marges fixées. Le cas trivial où les marges ne sont pas fixées correspond à la structure du tableau dite : « Association Complète ». Pour les problèmes pratiques, il n'existe pas de méthodes qui garantissent une solution optimale exacte. Des bornes dues à des inégalités mathématiques ont été employées. On ne va pas parler de la complexité combinatoire du problème, citons en effet à cet égard Hubert et Arabie* [3] *« . . . Constructing an exact bound, conditionnal on the fixed row and column totals of the given contingency table, is a very difficult problem of combinatorial optimization... » Lerman a proposé un algorithme récursif qui détermine la solution pas à pas en s'appuyant sur la notion de « points extrémaux ». Malheureusement le temps de calcul croît exponentiellement. On proposera deux bornes plus fines que celles présentées dans la littérature (elles serrent de plus près le critère des associations positives). La distribution* n_{uv} *correspondante n'est pas souvent atteinte. On proposera une méthode heuristique très rapide, basée sur des problèmes d'affectation classiques, pour déterminer une telle distribution.*

Mots clés : Mesure d'association; table de contingence; association maximale.

(*) Received January 1989.
(1) University Paris-VI, I.B.M. Scientific Center, 3,5, Place Vendome, 75001 Paris.

Recherche opérationnelle/Operations Research, 0399-0559/90/01 29 19/$ 3.90

1.1 INTRODUCTION

The normalization problem of criteria of association between two partitions (qualitative variables) has been considered in various studies related to classification (Morey and Agresti [1], Hubert and Arabie [3], Lerman [5], Marcotorchino [6], H. Messatfa [8]). The principle of normalization is implemented using the general formula:

$$I_N = \frac{I - \tau}{I_{\max} - \tau}$$

where:

- I is a measure of association between two qualitative variables;
- τ represents a structure for which $I_N = 0$;
- $I_{\max}$ represents the maximum value of the measure I.

From a previous study [8] it has been shown that if a measure I can be expressed as a linear function of $A = \sum_{u,\, v} n_{uv}^2$ then it can be stated that

$$\frac{I - \tau_I}{I_{\max} - \tau_I} = \frac{A - \tau_A}{A_{\max} - \tau_A}$$

where $A_{\max}$ is the maximum of A for a particular structure of the contingency table.

The above shows that the discriminant element between the normalized criteria lies in the selection of $A_{\max}$, which in most cases cannot be computed and will therefore be approximated by a value dependent of the table.

$A_{\max}$ has been approximated by many different ways. An exhaustive survey of approximation of $A_{\max}$ is found in [8]. Lerman [5] show that the symmetric boundary

$$B_3 = \min \left(\sum_u n_{u.}^2 , \sum_v n_{.v}^2 \right)$$

proposed by Hubert and Arabie [3] is in fact less than all the bounds based on application of Cauchy-Schartz inequality.

In this paper, we shall analytically define two boundaries of A which may be proved to be better than those mentioned above. The first bound (noted $\hat{B}$) due the inequality of Hoffman-Wielandt and the second (noted $\tilde{B}$) is due to the linear writing of criteria A.

In a second part of this study, we will determine an optimal distribution n_{uv} of a contingency table with fixed margins ([2]) such as: the positive association $\sum_{u, v} n_{uv}^2$ is maximal. For most practical problems, no methods exist which guarantee an optimum solution. In [5] an algorithmic bound is introduced, but the computational time required increases exponentially. It is computationally feasible for very small problem. We derive an algorithmic bound noted B'. It will be compared to the boundary determined by Lerman [5].

1.2. NOTATIONS

The study starts from the following configuration:

- C represents a variable with p modalities (partition with p classes);
- Y being a variable with q modalities (partition with q classes);
- n_{uv} denote the number of objects that are common to class u of C and to class v of Y.

The information on class overlap between the two partitions can be written in the form of a contingency table as in table I.

	1	v	q
1			
u		n_{uv}	$n_{u.}$
p			
		$n_{.v}$	N

Table T_1

n_{uv}: number of objects having modality u of C and modality v of Y;

$n_{u.}$: number of objects having modality u of C;

([2]) The trivial case where margins are not fixed corresponds to the matrix structure known as "Complete Association".

$n_{.v}$: number of objects having modality v of Y

$$n_{u.}=\sum_{v} n_{uv}, \qquad n_{.v}=\sum_{u} n_{uv}$$

Partitions C and Y can be represented by using paired comparison matrices of size $n\times n$ with value in $\{0, 1\}$ and their general term c_{ij} and y_{ij} are defined by:

$$\begin{cases} c_{ij}=1, & \text{if } i \text{ and } j \text{ are in the same class of } C \\ c_{ij}=0, & \text{otherwise} \end{cases}$$

$$\begin{cases} y_{ij}=1, & \text{if } i \text{ and } j \text{ are in the same class of } Y \\ y_{ij}=0, & \text{otherwise.} \end{cases}$$

Classical contingency formulas are linked to paired comparisons formulas by the means of the one to one correspondence relations (Marcotorchino [6]).

$$\begin{gathered} \sum_{u,v} n_{uv}^2=\sum_{i,j} c_{ij}y_{ij}; \qquad \sum_{v} n_{u.}^2=\sum_{i,j} c_{ij}; \qquad \sum_{u} n_{.v}^2=\sum_{i,j} y_{ij} \\ c_{i.}=\sum_{j} c_{ij}, \qquad c_{.j}=\sum_{i} c_{ij}, \qquad y_{i.}=\sum_{j} y_{ij}, \qquad y_{.j}=\sum_{i} y_{ij} \end{gathered} \tag{1}$$

1.3. DETERMINING THE RELATIONAL BOUND

1.3.1. Analytic bound

Let $\|C\|=\sqrt{\sum_{i,j} c_{ij}^2}$ denote the Frobenius norm of a matrix $C=(c_{ij})$. An analytic bound can be provided by the Hoffman-Wielandt inequality. If A and B are real $n\times n$ symmetric matrices with eigenvalues $a_1 \geqq \ldots \geqq a_n$, $b_1 \geqq \ldots \geqq b_n$ respectively then

$$\|A-B\|^2 \geqq \sum_{i=1}^{n} (a_i-b_i)^2 \tag{2}$$

The eigenvalues of each partition C are given by the number of elements in each class because the rows (columns) of C corresponding to elements in the same class are identical. Each partition has exactly p (the number of class) distinct rows (columns). For example, if C is a partition with 3 classes with 4, 3, 2 elements in each class respectively, then the associated matrix

can be written as followed:

$$C = \begin{matrix} 1&1&1&1&0&0&0&0&0 \\ 1&1&1&1&0&0&0&0&0 \\ 1&1&1&1&0&0&0&0&0 \\ 1&1&1&1&0&0&0&0&0 \\ 0&0&0&0&1&1&1&0&0 \\ 0&0&0&0&1&1&1&0&0 \\ 0&0&0&0&1&1&1&0&0 \\ 0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&1&1 \end{matrix}$$

We are now in a position to apply the Hoffman-Wielandt inequality. If C and Y are two partition matrices, the inequality states that

$$\| C - Y \|^2 \geqq \sum_{u=1}^{p} (n_{u.} - n_{.u})^2 + \sum_{u=p+1}^{q} n_{.u}^2 \qquad \text{with} \quad p < q$$

According to the previous notations we have

$$\| C \|^2 = \sum_{u=1}^{p} n_{u.}^2, \qquad \| Y \|^2 = \sum_{u=1}^{q} n_{.u}^2, \qquad \sum_{i,j} c_{ij} y_{ij} = \sum_{u,v} n_{uv}^2$$

We now obtain

$$\| C - Y \|^2 = \| C \|^2 + \| Y \|^2 - 2 \sum_{i,j} c_{ij} y_{ij} \geqq \sum_{u=1}^{p} n_{u.}^2 + \sum_{u=1}^{q} n_{.u}^2 - 2 \sum_{u=1}^{p} n_{u.} n_{.u}$$

Then

$$\sum_{i,j} c_{ij} y_{ij} = \sum_{u,v} n_{uv}^2 \leqslant \sum_{u=1}^{p} n_{u.} n_{.u}$$

Thus, the bound $\hat{B}$ can be derived as:

$$\hat{B} = \min \left(\sum_{u} n_{u.}^2, \sum_{v} n_{.v}^2, \sum_{u=1}^{p} n_{u.} n_{.u} \right) \tag{3}$$

In this boundary, we have introduce the third parameter, which, in some occasion gives a better solution than

$$B_3 = \min\left(\sum_u n_{u.}^2, \sum_v n_{.v}^2\right)$$

1.3.2. Determining the relational bound

We now show that it is more convenient to use the relational expression ([3]) of $A = \sum_{u, v} n_{uv}^2$, to find a bound lower than the previous one. As a result, it will be referred as relational bound. The following lemmas will be used to define this bound.

1.3.2.1. *Lemma* 1

Using the previous notations we have:

$$\sum_{u, v} n_{uv}^2 = \sum_{i, j} c_{ij} y_{ij} = \sum_{i, j} \min(c_{ij}, y_{ij}).$$

1.3.2.2. *Proof*

As $c_{ij} \in \{0, 1\}$ and $y_{ij} \in \{0, 1\}$ then:

$$c_{ij} y_{ij} = 1 \quad \text{if} \quad \min(c_{ij} y_{ij}) = 1$$
$$c_{ij} y_{ij} = 0 \quad \text{if} \quad \min(c_{ij} y_{ij}) = 0$$

this completes the proof of the lemma.

1.3.2.3. *Lemma* 2

For every contingency table we have

$$\sum_{u, v} n_{uv}^2 \leqq \sum_i \min(c_{i.}, y_{i.})$$

where

$$c_{i.} = \sum_j c_{ij} \quad \text{and} \quad y_{i.} = \sum_j c_{ij}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, n$$

([3]) The relational expression of A is defined as follow: $A = \sum_{i, j} c_{ij} y_{ij}$

1.3.2.4. *Proof*

From lemma 1 and the property of the Min operator, we have

$$\sum_{u,\,v} n_{uv}^2 = \sum_{i,\,j} \min(c_{ij},\, y_{ij}) \leqq \sum_{i} \min\left(\sum_{j} c_{ij},\, \sum_{j} y_{ij}\right)$$

as $c_{i.} = \sum_{j} c_{ij}$ and $y_{i.} = \sum_{j} c_{ij}$,

$$\sum_{u,\,v} n_{uv}^2 \leqq \sum_{i} \min(c_{i.},\, y_{i.})$$

This proves our claim.

1.3.2.5. *Corollary*

Let $I = \{1, \ldots, n\}$ and let the permutation σ be a bijection on I. From the previous lemmas,

$$B = \sum_{i \in I} \min(c_{i.},\, y_{\sigma(i).})$$

is a boundary of A.

In other hand σ correspond to a value $[n_{uv};\ 1 \leqslant u \leqslant p,\ 1 \leqslant v \leqslant q]$ of a contingency table. Then B is obtained by formula (4).

1.3.2.6. *Proposition*

Let σ be a given permutation then we may obtain

$$B = \sum_{i \in I} \min(c_{i.},\, y_{\sigma(i).}) = \frac{\sum_{u} n_{u.}^2 + \sum_{v} n_{.v}^2}{2} - \frac{1}{2} \sum_{u,\,v} n_{uv} \,|\, n_{u.} - n_{.v} \,| \qquad (4)$$

1.3.2.7. *Proof*

To prove this proposition we need to prove only the following equality:

$$\sum_{i \in I} \min(c_{i.},\, y_{\sigma(i).}) = \sum_{u,\,v} n_{uv} \min(n_{u.},\, n_{.v})$$

Knowing that $\forall i \in I,\ |I| = N$, i belongs to one and only one subset I_{uv} of objects having the u modality of C and the v modality of Y. $|I_{uv}| = n_{uv}$ and $\bigcup I_{uv} = I$, therefore

$$\sum_{i \in I} \min(c_{i.},\, y_{\sigma(i).}) = \sum_{u,\,v} \sum_{i \in I_{uv}} \min(c_{i.},\, y_{\sigma(i).})$$

and since, by definition, for every $i \in I_{uv}$ we have: $n_{u.} = c_{i.}$ and $n_{.v} = y_{i.}$. Then

$$\sum_{i \in I_{uv}} \min(c_{i.}, y_{\sigma(i).}) = n_{uv} \min(n_{u.}, n_{.v}).$$

Since

$$\min(a, b) = \frac{a+b}{2} - \frac{1}{2}|a-b| \tag{5}$$

we may obtain

$$\sum_{i \in I} \min(c_{i.}, y_{\sigma(i).}) = \sum_{u, v} n_{uv} \min(n_{u.}, n_{.v}) = \sum_{uv} n_{uv} \frac{(n_{u.} + n_{.v})}{2} - \frac{1}{2} \sum_{u, v} n_{uv} |n_{u.} - n_{.v}$$

or

$$B = \frac{\sum n_{u.}^2 + \sum_v n_{.v}^2}{2} - \frac{1}{2} \sum_{u, v} n_{uv} |n_{u.} - n_{.v}|$$

1.3.2.8. *Proposition*

For every contingency table we have

$$\sum_{u, v} n_{uv}^2 \leqslant \frac{(\sum_i c_{i.} + \sum_i y_{i.})}{2} - \frac{1}{2} \min_{\sigma} \sum_{i \in I} |c_{i.} - y_{\sigma(i).}|$$

1.3.2.9. *Proof*

We have shown that

$$\sum_{u, v} n_{uv}^2 \leqslant \sum_{i \in I} \min(c_{i.}, y_{\sigma(i).})$$

From equality (5) we obtain

$$\sum_{i \in I} \min(c_{i.}, y_{\sigma(i).}) = \frac{(\sum_{i \in I} c_{i.} + \sum_{i \in I} y_{i.})}{2} - \frac{1}{2} \sum_{i \in I} |c_{i.} - y_{\sigma(i).}|$$

This equality is obtained for all σ and

$$\sum_{u, v} n_{uv}^2 \leqslant \sum_{i \in I} \min(c_{i.} \cdot y_{\sigma * (i).})$$

where σ^* is the permutation which realize the maximum of the quantity

$$\sum_{i \in I} \min (c_{i.}, y_{\sigma(i).})$$

we get

$$\sum_{u, v} n_{uv}^2 \leqslant \max_{\sigma} \left(\frac{\sum_i c_{i.} + \sum_i y_{i.}}{2} - \frac{1}{2} \sum_{i \in I} | c_{i.} - y_{\sigma(i).} | \right)$$

Because the margins are fixed, the quantities $\sum_i c_{i.}$ and $\sum_i y_{i.}$ are constant, we finally obtain

$$\sum_{u, v} n_{uv}^2 \leqslant \sum_{i \in I} \min (c_{i.}, y_{\sigma(i).}) = \frac{(\sum_{i \in I} c_{i.} + \sum_{i \in I} y_{i.})}{2} - \frac{1}{2} \min_{\sigma} \sum_{i \in I} | c_{i.} - y_{\sigma(i).} |$$

which proves the proposition.

We are now ready to define the relational bound.

1.3.2.10. *Definition*

Let σ be a permutation then the relational bound $\tilde{B}$ is defined as follow:

$$\tilde{B} = \frac{(\sum_{i \in I} c_{i.} + \sum_{i \in I} y_{i.})}{2} - \frac{1}{2} \min_{\sigma} \sum_{i \in I} | c_{i.} - y_{\sigma(i).} | \tag{6}$$

1.3.2.11. *Remarks*

In contingential notations, $\tilde{B}$ can be written as:

$$\tilde{B} = \frac{1}{2} (\sum_u n_{u.}^2 + \sum_v n_{.v}^2) - \frac{1}{2} \min \sum_{u, v} n_{uv} | n_{u.} - n_{.v} | \tag{7}$$

From equalities (6) and (7) the computed value of $\tilde{B}$ depends on minimizing the quantities

$$\min_{\sigma} \sum_{i \in I} | c_{i.} - y_{\sigma(i).} |; \qquad \min \sum_{u, v} n_{uv} | n_{u.} - n_{.v} |$$

1.3.3. Computing the relational boundary

To compute boundary $\tilde{B}$ of A, when margins are fixed, we can proceed solving one of the 2 following equivalent problem:

Problem 1:

$$\begin{aligned} &\min \sum_{u,\,v} n_{uv} \,|\, n_{u.} - n_{.v} \\ &n_{u.} = \sum_{v} n_{uv} \\ &n_{.v} = \sum_{u} n_{uv} \\ &n_{uv} \geqslant 0 \end{aligned} \tag{8}$$

This is a classical transportation problem where the variables are n_{uv} and the costs are given by $C_{uv} = |\, n_{u.} - n_{.v} \,|$.

Problem 2:

$$\begin{aligned} &\min \sum_{ij} |\, c_{i.} - y_{i.} \,|\, x_{ij} \\ &\sum_{i} x_{ij} = 1, \qquad i = 1, \ldots, n \\ &\sum_{j} x_{ij} = 1, \qquad j = 1, \ldots, n \end{aligned} \tag{9}$$

The first constraint, implies that each component of C is assigned to one and only one component of Y. The second implying that each component of Y has one and only one component of vector C assigned to it. The unknown indicator x_{ij} is either 1 or 0 depending on whether the component of the two vectors are assigned. The solving process associated to this problem is not time consuming, because we deal with a classical linear assignment problem. The appropriate set of dichotomous indicator functions can be obtained by an application of the methods for solving integer linear programming.

It is interesting to examine that $\tilde{B}$ can be obtain analytically. The following proposition generalizes this observation.

1.3.3.1. *Proposition*

Consider that the margins of the contingency table are fixed. Let

$$(n_{1.}, n_{2.}, \ldots, n_{p.}) \qquad \text{and} \qquad (n_{.1}, n_{.2}, \ldots, n_{.q})$$

those margins respectively. The elements $c_{i.}$ and $y_{i.}$ are respectively components of C and Y defined as:

$$C=(n_{1.}, \ldots, n_{2.}, \ldots, n_{p.}) \quad \text{with} \quad n_{1.} > n_{2.} > \ldots > n_{p.}$$

$$Y=(n_{.1}, \ldots, n_{.2}, \ldots, n_{.q}) \quad \text{with} \quad n_{.1} > n_{.2} > \ldots > n_{.q}$$

In C, $n_{i.}$ is repeated $n_{i.}$ times, for $i=1, \ldots, p$. In Y, $n_{.j}$ is repeated $n_{.j}$ times, for $j=1, \ldots, q$.

- For the variable C, the objects i are ranked with respect to an uniform decreasing order of $c_{i.}$, for $i=1, \ldots, n$.
- For the variable Y, σ^1 is the complete order on the group of the objects i with respect to an uniform decreasing order of $y_{.i}$, for $i=1, \ldots, n$.

Thus

$$\tilde{B}=\frac{1}{2}\left(\sum_{i,j} c_{ij}+\sum_{i,j} y_{ij}\right)-\frac{1}{2}H$$

where

$$H=\sum_{i \in I} |c_{i.}-y_{\sigma^1(i).}| = \min_{\sigma} \sum_{i \in I} |c_{i.}-y_{\sigma(i).}|$$

1.3.3.2. *Proof*

We shall show that if the components $y_{i.}$ are ranked according to an order $\sigma \neq \sigma^1$, then

$$\sum_{i \in I} |c_{i.}-y_{\sigma^1(i).}| \leqslant \sum_{i \in I} |c_{i.}-y_{\sigma(i).}|$$

According to orders σ^1, C and Y can be written as followed:

$$C=(c_1, c_2, \ldots, c_p, \ldots, c_n) \quad \text{and} \quad Y=(y_1, y_2, \ldots, y_m, \ldots, y_n)$$

Suppose that the components y_1 and y_m are exchanged, referred to a permutation σ of I.

According to σ we get

$$Y=(y_m, y_2, \ldots, y_1, \ldots, y_n)$$

If $c_1 \geqq y_1$ and y_m is assigned to c_k then

$$\Delta H = \sum_{i \in I} |c_{i.} - y_{\sigma^1(i).}| - \sum_{i \in I} |c_{i.} - y_{\sigma(i).}|$$

$$= |c_1 - y_1| + |c_k - y_m| - |c_1 - y_m| - |c_k - y_1|$$

As $c_1 \geqslant y_1$, $c_1 \geqslant y_m$, ΔH can be written

$$\Delta H = |c_1 - y_1| + |c_k - y_m| - |c_1 - y_m| - |c_k - y_1|$$

$$\Delta H = y_m - y_1 + |c_k - y_m| - |c_k - y_1|$$

Let us consider ΔH in both cases:

First case $c_k \geqslant y_m$

$$\Delta H = c_k - y_1 - |c_k - y_1| \leqslant 0.$$

As $X - |X| \leqslant 0$, then $\Delta H \leqslant 0$

Second case $c_k \leqslant y_m$

$$\Delta H = 2y_m - 2y_1 \leqslant 0$$

As $y_m - y_1 \leqslant 0$, then $\Delta H \leqslant 0$.

We have in both cases

$$\sum_{i \in I} |c_{i.} - y_{\sigma^1(i).}| \leqslant \sum_{i \in I} |c_{i.} - y_{\sigma(i).}|$$

We conclude that every $\sigma \neq \sigma^1$ does not improve H.

The same proof is also applied to the case $c_1 \geqslant y_1$.

According to the proposition, we can see that $\tilde{B}$ is easily computed through decreasing order of $c_{i.}$ and of $y_{i.}$. Vectors C and Y are ranked according to the decreasing order of their components c_i and y_i. We compute term by term their difference in absolute value.

1.3.4. Example of the computation of H

Let us consider the following fixed margins:

$$(10, 7); (5, 5, 4, 3)$$

So, C is decomposed in 10 values of $c_{1.} = 10$ and 7 values of $c_{2.} = 7$

Y is decomposed in 5 values of $y_{.1}=5$, 5 values of $y_2=5$, 4 values of $y_{.3}=4$ and 3 values of $y_{.4}=3$.

$$C=(10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 7, 7, 7, 7, 7, 7, 7).$$

$$Y=(5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 3, 3).$$

$$H=10(10-5)+4(7-4)+3(7-3)=74.$$

1.3.5. The relational bound compared to the others boundaries

1.3.5.1. *Proposition*

The boundary $\tilde{B}$ is lower than the smallest of the proposed boundaries. We have

$$\tilde{B} \leqslant B_3 = \min\left(\sum_u n_{u.}^2, \sum_v n_{.v}^2\right)$$

1.3.5.2. *Proof*

From formula (5) and by using the "Min" operator, we get

$$\tilde{B}=\sum_i \min\left(\sum_j c_{ij}, \sum_j y_{ij}\right) \leqslant \min\left(\sum_{ij} c_{ij}, \sum_{ij} y_{ij}\right)=\min\left(\sum_u n_{u.}^2, \sum_v n_{.v}^2\right)$$

1.3.5.3. *Example*

Let the fixed margins defined as follows:

$$(12, 11, 5, 2, 2, 1, 1);\ (10, 9, 9, 6)$$

The bounds are computed and we get

$$B_3=298; \qquad \hat{B}=276; \qquad \tilde{B}=252$$

1.4. DETERMINING THE ALGORITHMIC BOUND B'

In many cases the contingency table associated to the relational bound is not reachable. That mean, it does not exist a distribution n_{uv} such that:

$$\tilde{B}=\sum_{u,\,v} n_{uv}^2$$

Now, we are interested to find a bound of A and its associated contingency table. We will define an heuristic strategy. The determined relational boundary $\tilde{B}$ has required the allocation of y_i to c_i in such a way that $\sum_{i \in I} \min (c_{i.}, y_{\sigma(i).})$ is minimal.

Let's consider A_k a group of y_k allocated to c_k; let the function F_k be defined by:

$$F_k = (c_k + y_k - 2 | A_k |) | A_k |$$

where $| A_k |$ is the cardinal of the set A_k; F_k is the cost of allocating y_k to c_k.

We begin with an example that illustrate the basic concept of the procedure. Consider the following 3×3 contingency table:

0	6	6	12
8	0	0	8
1	0	0	1
9	6	6	21

Table T_2

We have $\tilde{B} = 148$ and $A = 137$.

In terms of assignment the contingency table can be also described by the following representation

12	12	12	12	12	12		12	12	12	12	12	12		8	8	8	8	8	8	8	8		1	
6	6	6	6	6	6		6	6	6	6	6	6		9	9	9	9	9	9	9	9		9	

For this representation we have four A_i corresponding to the following costs:

$$F_1 = (c_1 + y_1 - 2 | A_1 |) | A_1 |$$

$$F_1 = (12 + 6 - 2 \times 6)\, 6; \qquad F_1 = 36, \qquad F_2 = 36, \qquad F_3 = 8, \qquad F_4 = 8$$

The total costs is computed as:

$$F = F_1 + F_2 + F_3 + F_4 = 88$$

For the same distribution of margins, let us consider the following contingency table:

9	3	0	12
0	3	5	8
0	0	1	1
9	6	6	21

Table T_3

We have $\tilde{B}=148$ and $A=125$.

T_3 can be also described by the following representation

12	12	12	12	12	12	12	12	12		12	12	12		8	8	8		8	8	8	8	8		1	
9	9	9	9	9	9	9	9	9		6	6	6		6	6	6		6	6	6	6	6		6	

For this representation we have five $A_i\ i=1\ldots5$ corresponding to the following costs:

$$F_1=(c_1+y_1-2|A_1|)|A_1|, \qquad |F_1|=(12+9-2\times9)9$$

$$F_1=27, \qquad F_2=36, \qquad F_3=24, \qquad F_4=20, \qquad F_5=5$$

The total costs is computed as:

$$F=F_1+F_2+F_3+F_4+F_5=112$$

It follows from these results that if F is minimum then $A=\sum_{u,\,v} n_{uv}^2$ is maximal.

1.4.1.1. *Proposition*

For every contingency table, we have the following formulas

$$A=\frac{1}{2}\left(\sum_{i,\,j} c_{ij}+\sum_{i,\,j} y_{ij}\right)-\frac{1}{2}F$$

1.4.1.2. *Proof*

For any $|A_k|$ there exists a couple (u, v) such as $n_{uv}=|A_k|$. Therefore

$$\sum_k F_k=\sum_k (c_k+y_k)\times|A_k|-2\sum_k|A_k|^2=(\sum_v n_{u.}^2+\sum_u n_{.v}^2)-2\sum_{u,v} n_{uv}^2$$

and finally

$$\frac{\sum_{i,j} c_{ij}+\sum_{i,j} y_{ij}}{2}-\frac{F}{2}=\frac{\sum_v n_{u.}^2+\sum_u n_{.v}^2}{2}-\frac{\sum_v n_{u.}^2+\sum_u n_{.v}^2}{2}+\sum_{u,v} n_{uv}^2=A$$

1.4.1.3. *Definition*

According to the previous proposition we define the algorithmic bound by:

$$B'=\frac{\sum_{i,j} c_{ij}+\sum_{i,j} y_{ij}}{2}-\frac{1}{2}\min F$$

1.4.1.4. *Proposition*

Boundary B' is more precise than $\tilde{B}$, that is $A \leqslant B' \leqslant \tilde{B}$

1.4.2. Proof

We have:

$$|A_i| \leqslant \min(c_i, y_i)$$

and thus

$$c_i+y_i-2|A_i| \geqslant |c_{i.}-y_{i.}|$$

As $|A_i| \geqq 0$, we have:

$$(c_i+y_i-2|A_i|)|A_i| \geqslant |c_{i.}-y_{i.}|$$

Finaly we obtain

$$\max(\sum_i (c_i+y_i-2|A_i|))|A_i| \geqslant \sum_i |c_{i.}-y_{i.}|$$

This proves our claim.

Note that if for every k, we have:

$$|A_k| = \min(c_k, y_k)$$

then

$$B' = \tilde{B}$$

In order to determine B', it is required to minimize the function F (quadratic problem). For this purpose an heuristic procedure is used minimizing every allocation (F_k). We summarize here an algorithm that has been proved theoretically to yield the maximal distribution of a contingency table. It is based upon assignment and transportation techniques.

1.4.3. The heuristic

The heuristic operates in the following manner:

STEP 0

We start with an initial empty contingency table T^* with given margins.

STEP 1

We built a table T_1 such as $n_{uv} = \min(n_{u.}, n_{.v})$ ($n_{u.}$ and $n_{.v}$ are the fixed margins). $u = 1, \ldots, p$; $v = 1, \ldots, q$.

STEP 2

Determine an element of T_1 which minimizes the cost F. Assign this element in the solution T^*. We built a table T_2 such as $n_{uv} = \min(n_{u.}, n_{.v})$ ($n_{u.}$ and $n_{.v}$ are the fixed margins). $u = 1, \ldots, p-1$; $v = 1, \ldots, q$. For every step k, the size of table T_k is reduced.

Step 1 and step 2 are repeated until T_k will be empty.

1.4.3.1. *Remark*

The affectation principle used in the heuristic is derived from the notion of "déchargement" defined by Lerman [5]. In Lerman context an element is affected (déchargé) if he corresponds to an extremal point, while in our procedure an element is affected if the cost is minimum.

To illustrate the heuristic let us consider an example.

1.4.3.2. *Example*

We consider the following marginal distributions of a contingency table

$$(403, 120, 117, 111, 104, 87, 35, 14)$$

$$(377, 252, 220, 139, 3)$$

The example has been treated with our heuristic and the Lerman algorithm [5].

For this example, we compute:

$\tilde{B}$ (the relational boundary), B_3 (the minimum between squared sums of margins), A_H (the value of A computed by the heuristic) and A_L (the algorithmic bound computed by Lerman).

The solution given by the two methods are.

377	1	22	0	3	403
0	120	0	0	0	120
0	117	0	0	0	117
0	0	111	0	0	111
0	0	0	104	0	104
0	0	87	0	0	87
0	0	0	35	0	35
0	14	0	0	0	14
377	252	220	139	3	

Table given by our heuristic.

377	1	22	0	3	403
0	120	0	0	0	120
0	117	0	0	0	117
0	0	111	0	0	111
0	0	0	104	0	104
0	0	87	0	0	87
0	0	0	35	0	35
0	14	0	0	0	14
377	252	220	139	3	

Table given by Lerman (33706 tables computed).

$$\tilde{B} = 208\,864, \qquad B_3 = 222\,625, \qquad A_H = A_L = 202\,839$$

In this example the two methods give the same result but the one given by the algorithm of Lerman requires computing 33706 tables.

1.5. CONCLUSION

The proposed algorithm has the virtues of being relatively simple, mathematically tractable and very fast. The algorithm we have described requires, some calculations for its implementation. First, we determine the value of

the relational bound; in general this portion of the algorithm requires a small number of operations. The computation of the costs is also easier to solve. We cannot say that the solution obtained by the heuristic will be optimal as we don't know it. However, we expect the one found by the procedure, is very close to a global optimum. This is due to the comparison of the optimum value with the relational bound. There exists other improvement possibilities of the final solution, they will be developed in the forthcoming articles. Thus, the author believes that the procedure may used for many other applications. Additional research could be both interesting and enlightening.

ACKNOWLEDGMENTS

The author is grateful to Dr F. Marcotorchino for his interest and helpful advises during the development of this study. We would like also to thank an anonymous referee for several useful suggestions.

REFERENCES

1. AGRESTI A. and L. MOREY, *An Adjustment of the Rand Statistic for Chance Agreement*, Educational and Psychological Measurement, Vol. 44, 1984, pp. 33-37.
2. ANTONIO MANGO, *Sulla costruzione de la indici relativi di contingenza basati sulla contingenzi ipotesi per una ricerca teorica*, Ricerche Di statistica Universita di Napoli, 1983.
3. ARABIE P. and L. HUBERT, *Comparing Partitions*, Fourth European meeting of the classification Societies, Cambridge, July 1985.
4. FAWLKES E. B. and C. L. MALLOWS, *A Method for Comparing two Hierarchical Clusterings*, J.A.S.A., Vol. 78, 1983, pp. 553-584.
5. LERMAN I. C. and P. PETER, *Structure maximale pour la somme des carrés d'une contingence aux marges fixées : une solution algorithmique programmée*, RAIRO-Rech. Op., Vol. 22, No. .
2, 1988, pp. 83-136.
6. MARCOTORCHINO F., *Utilisation des comparaisons par paires en statistiques des contingences*, Étude du Centre Scientifique IBM-France, N° F-071, 1984.
7. MARCOTORCHINO F., *Maximal Association Theory as a Tool of Research*, in W. GAUL and M. SCHADER (editors), Proceedings of the 9th Annual meeting of the classification society (F.R.G.), North-Holland, 1986.
8. MESSATFA H., *Unification de certains critères d'association par linéarisation et normalisation*, Étude du Centre Scientifique IBM-France, N° F-114, 1987.
9. POPPING R., *Traces of Agreement: On the Dot-Product as a Coefficient of Agreement*, Quality and Quantity, 17, 1983.