Overview of the relational analysis approach in data-mining and multi-criteria decision making

Julien Ah-Pine, Jean-François Marcotorchino

authoryear,round,citesep=;,aysep=,,yysep=;

0

Overview of the Relational Analysis approach in Data-Mining and Multi-criteria Decision Making

Julien Ah-Pine
Xerox Research Centre Europe
France

Jean-François Marcotorchino
Thales Communications
France

1. General scope

In this chapter we introduce a general framework called the Relational Analysis approach and its related contributions and applications in the fields of data analysis, data mining and multi-criteria decision making. This approach was initiated by J.F. Marcotorchino and P. Michaud at the end of the 70's and has generated many research activities. However, the aspects of this framework that we would like to focus on are of a theoretical kind. Indeed, we are aimed at recalling the background and the basics of this framework, the unifying results and the modeling contributions that it has allowed to achieve. Besides, the main tasks that we are interested in are the ranking aggregation problem, the clustering problem and the block seriation problem. Those problems are combinatorial ones and the computational considerations of such tasks in the context of the RA methodology will not be covered. However, among the list of references that we give thoughout this chapter, there are numerous articles that the interested reader could consult to this end.

In order to introduce the Relational Analysis approach (denoted "RA" in the rest of the document), let us first introduce several problems that one could encounter in the data analysis field. To this end, let us consider a data table concerning a set of N objects $\mathcal{O} = \{O^1,\ldots,O^i,\ldots,O^N\}$ described by a set of M variables $\mathbb{V} = \{V^1,\ldots,V^k,\ldots,V^M\}$. These data can be represented using a $(N \times M)$ feature matrix denoted T given by the following eqs. (1) and (2); $\forall i = 1,\ldots,N; k = 1,\ldots,M$:

$$T_{ik} = \quad V_i^k = \text{Numerical value assigned to object } O^i \text{ according to } V^k \tag{1}$$

$$T = \begin{array}{c} \\ O^1 \\ O^2 \\ \vdots \\ O^i \\ \vdots \\ O^N \end{array} \begin{array}{c} \begin{array}{cccccc} V^1 & V^2 & \ldots, & V^k & \ldots & V^M \end{array} \\ \begin{pmatrix} V_1^1 & V_1^2 & \ldots & V_1^k & \ldots & V_1^M \\ V_2^1 & V_2^2 & \ldots & V_2^k & \ldots & V_2^M \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ V_i^1 & V_i^2 & \ldots & V_i^k & \ldots & V_i^M \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ V_N^1 & V_N^2 & \ldots & V_N^k & \ldots & V_N^M \end{pmatrix} \end{array} \tag{2}$$

Depending on the nature of the features of T, we can list the different following problems tackled in the data analysis field and related domains:

- The variables in $\mathbb{V}$ can correspond to M criteria that give M different rankings on the objects $\mathbb{O}$. In that case, the objects could be N different alternatives. Then, one can be interested in finding a consensual ranking that "sums up" these M different rankings in order to determine the best (the most consensual) alternative[1]. This task is known as the ranking aggregation problem and it can be encountered in other domains, than data analysis, such as social choice theory, multi-criteria decision making or multi-agent systems.
- The variables in $\mathbb{V}$ can be categorical[2] variables. In that case, one can be motivated by discovering patterns among the set of objects $\mathbb{O}$. In other words, one would want to find a partition such that objects belonging to the same cluster have high similarities and objects belonging to different clusters have low similarities. This problem, known as the clustering problem, is studied in statistical data analysis, and data-mining fields and it has many applications such as customer relationship management, text mining or web data clustering for instance.
- Finally, one could also be interested in finding a bi-partition which simultaneously decompose both sets $\mathbb{O}$ and $\mathbb{V}$. In that context, a bicluster is the association of a cluster of objects and a cluster of variables. This task is known as the biclustering or block seriation problem and it is studied notably in gene-mining or in group technology problems for example.

The previously mentioned problems are often modeled and solved by means of different approaches. One of the main advantages of the method presented here, is that it allows to synthesize in a unique formal way, all those different data analysis tasks, as particular cases of a general model. In the RA framework, those different problems can be formalized as binary relations aggregation problems. Departing from the feature matrix T, the different aforementioned tasks, can be seen as finding a consensual binary relation that aggregates and summarizes a set of individual binary relations (the variables) of T.

We briefly give in what follows some illustrative examples:

- When the variables $V^k; k = 1, \ldots, M$, are numerical criteria, they induce M different rankings $\mathcal{R}^k; k = 1, \ldots, M$, on the set of alternatives $\mathbb{O}$. These rankings are order relations (partial or total orders) and the solution we look for is a consensual relation $\mathcal{R}$ (a total order for example) that fits "as good as possible", the M individual rankings.
- When the variables $V^k; k = 1, \ldots, M$, are categorical features, the latter induce M different partitions $\mathcal{R}^k; k = 1, \ldots, M$, of the set of objects. Clustering those objects can thus be seen as looking for a consensual partition (or an equivalence relation) $\mathcal{R}$, that sums up the M individual partitions.
- When the feature matrix T consists of $0/1$ values such as indicator tables that encode categorical variables, the biclustering problem amounts to determine a bi-partition of objects and categories. This problem can also be interpreted as the search for a consensual relation $\mathcal{R}$. This kind of relation is called a "block seriation" relation.

[1] The alternative that is ranked first in the consensual ranking.

[2] And more generally, numerical variables.

The RA approach uses algebraic concepts related to binary relations in order to turn the previously mentioned problems into binary relations aggregation problems. Basically, if we denote by $\mathcal{R}^k$ the binary relation associated to a variable V^k, then all the aforementioned problems could be seen as a particular instance of the following problem:

$$\max_{\mathcal{R}} \text{Aggreg}(\mathcal{R}^1, \ldots, \mathcal{R}^k, \ldots, \mathcal{R}^M; \mathcal{R}) \tag{3}$$

where Aggreg is an aggregation procedure.

One of the main characteristics of the RA methodology is to use pairwise comparisons matrices similar as adjacency matrices in order to represent binary relations. We will see that the representation of binary relations through that coding has many properties. The second principle of the RA approach consists in using a criterion called "the Condorcet's criterion" as a global measure of consensus. This criterion has its origins in mathematical social sciences and was first studied in its literal form by A. de Condorcet in 1785. This criterion is nothing but a voting criterion, which was applied first to the ranking aggregation problem. It was shown that this criterion satisfies many axioms in the context of social choice theory. The Condorcet's criterion was, then, extended to the partitions aggregation problem. Therefore it can be interpreted as a partitioning criterion as well. The RA approach notably contributed to show that the ranking aggregation problem and the clustering one were particular cases of a unique model, as we will show in the next paragraphs.

The rest of this chapter is organized as follows. In section 2, we basically recall some key properties of the RA approach: the individual relational matrices, that represent individual binary relations; the collective relational matrix, that is aimed at aggregating the individual binary relations in a simple yet efficient way; the relational properties of binary relations that can be expressed as linear inequalities or equalities using relational matrices; and the general expression of the Condorcet's criterion.
In section 3, 4, 5, we respectively detail the applications of the RA approach in ranking aggregation problems, clustering problems, and block seriation problems. Particularly, in the RA framework, all these problems are modeled using the same formalism, based on $0/1$ integer linear programming.
Recently, "Correlation Clustering" (CC) problems were proposed in (6). This setting tackles the clustering problem from a graphic point of view and has many relationships with different concepts underlined in the RA approach. Indeed, the partitioning criterion used in the CC approach is very similar to the Condorcet criterion. Moreover, the linear program used by this approach for modeling the clustering problem (see for example (15), (22), (18)), is the same as in the RA method. Accordingly, we also introduce in section 6, other results of interest obtained in the clustering field by using the RA method in order to strengthen the graphical and the linear programming point of views for addressing clustering problems.

2. Introduction to the Relational Analysis approach

We first recall some previous contributions concerning the analysis of relational data, that is to say, data which have particular structures such as binary relations: order relations, equivalence relations or graph relations in general. There has been a growing interest for such kinds of data since the end of the 70's.

Concerning order relations, Condorcet's work opened up the mathematical field of decision making in the social sciences (14). In France, A. de Condorcet's work got a particular interest in the 80's, rediscovered and updated by J.F. Marcotorchino and P. Michaud (36), (40), B. Monjardet, J.P. Barthélémy and B. Leclerc (8), (26)... In the USA, we can also mention the Nobel Prize's laureate, K.J. Arrow who has contributed to the social choice theory (5) and also the book of Kemeny and Snell (23) in mathematical sciences as well.

Concerning equivalence relations (categorical data), since this type of data has been mostly studied by statisticians and data analysts, it is an other set of contributors which has to be quoted. We can firstly mention S. Régnier (46), I.C. Lerman (27) and J.F. Marcotorchino and P. Michaud (37), on the french side. Apart from french researchers, we can particularly mention H.T. Zahn's work (52) as well as B. Mirkin's work, see (42) for example.

Obviously, this is not an exhaustive reference list of scientists who have contributed to this area. We have only mentioned some main papers that are closely related to the approach proposed by the RA framework.

The RA methods, presented here, are mainly the approaches, studied and developed by J.F. Marcotorchino and P. Michaud and colleagues. Their work is essentially based upon the study of relational data from the graph theory, the statistical and the integer linear programming standpoints (36), (41).

First of all, let us recall basic definitions about binary relations.

A binary relation $\mathcal{R}$ on two sets of objects $\mathbb{O}$ (the domain) and $\mathbb{D}$ (the codomain[3]), is a triple $(\mathbb{O},\mathbb{D},G(\mathcal{R}))$, where $G(\mathcal{R})$ called the graph of the relation $\mathcal{R}$, is a subset of the Cartesian product $\mathbb{O}\times\mathbb{D}$. If we have $(O^i,D^j)\in G(\mathcal{R})$, then we say that object O^i is in relation with object D^j for the relation $\mathcal{R}$. This will be denoted by $O^i\mathcal{R}D^j$.

We can also associate to $\mathcal{R}$, its complement which is a binary relation denoted by $\overline{\mathcal{R}}$ and which is the subset of the cartesian product $\mathbb{O}\times\mathbb{D}$ such that $(O^i,D^j)\notin G(\mathcal{R})$.

When $\mathbb{D}=\mathbb{O}$, we talk about binary relations on a single set $\mathbb{O}$. This particular kind of binary relations is of interest and it will be referred as $(\mathbb{O},G(\mathcal{R}))$.
There exist different properties that a binary relation $(\mathbb{O},G(\mathcal{R}))$ can satisfy. Among all those relational properties, the most useful ones are given in Table 1.
Those properties allow us to characterize the type of a binary relation $(\mathbb{O},G(\mathcal{R}))$. We have the following definitions:

- A preorder is a binary relation that is reflexive and transitive.
- A strict total order is a binary relation that is irreflexive, asymmetric, transitive and total.
- An equivalence relation is a binary relation that is reflexive, symmetric, and transitive.

[3]For a real continuous quantitative variable, $\mathbb{D}$ could equal $\mathbb{R}$ for example; for a categorical variable, $\mathbb{D}$ could be the set of categories of this variable.

Relational property	Logical definition
Reflexivity	$O^i \mathcal{R} O^i \quad \forall O^i \in \mathbb{O}$
Irreflexivity	$O^i \overline{\mathcal{R}}^k O^i \quad \forall O^i \in \mathbb{O}$
Symmetry	$O^i \mathcal{R} O^{i'} \Rightarrow O^{i'} \mathcal{R} O^i \quad \forall (O^i, O^{i'}) \in \mathbb{O}^2$
Asymmetry	$O^i \mathcal{R} O^{i'} \Rightarrow O^{i'} \overline{\mathcal{R}}^k O^i \quad \forall (O^i, O^{i'}) \in \mathbb{O}^2 : O^i \neq O^{i'}$
Transitivity	$O^i \mathcal{R} O^{i'} \wedge O^{i'} \mathcal{R} O^{i''} \Rightarrow O^i \mathcal{R} O^{i''} \quad \forall (O^i, O^{i'}, O^{i''}) \in \mathbb{O}^3$
Totality	$O^i \mathcal{R} O^{i'} \vee O^{i'} \mathcal{R} O^i \quad \forall (O^i, O^{i'}) \in \mathbb{O}^2 : O^i \neq O^{i'}$

Table 1. Relational properties for $(\mathbb{O}, G(\mathcal{R}))$

After recalling briefly, basic concepts inherent to binary relations, we now present how the RA approach copes with such data structures.

The first principle of the RA methodology amounts to represent binary relations as pairwise comparisons matrices, called "relational matrices", which are made of $0/1$ values.
Let $(\mathbb{O}, G(\mathcal{R}))$ be a binary relation on a single set $\mathbb{O}$, where $\#\mathbb{O} = N$. Then in the RA method, we represent this binary relation by its $(N \times N)$ relational matrix[4] X where, $\forall i, i' = 1, \ldots, N$:

$$X_{ii'} = \begin{cases} 1 & \text{if } O^i \mathcal{R} O^{i'} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Using the RA formalism, we can easily define the relational matrix $\overline{X}$, associated to the complement of the relation $(\mathbb{O}, G(\mathcal{R}))$. Indeed, we have, $\forall i, i' = 1, \ldots, N$:

$$\overline{X}_{ii'} = 1 - X_{ii'} \tag{5}$$

In the block seriation problems, we are no longer faced with relations on the same set, as previously done, but we will consider, in that case, binary relations on two different sets $(\mathbb{O}, \mathbb{D}, G(\mathcal{R}))$.

[4] In graph theory those matrices are adjacency matrices but in the case of particular binary relations, these matrices have special properties as we will see later.

For clarity reasons, we will use other notations for that type of binary relations. Suppose that, $\#\mathbb{O} = N$ and $\#\mathbb{D} = P$; then the relational matrix that represents the binary relation is the $(N \times P)$ binary matrix Z, where, $\forall i = 1,\ldots,N; j = 1,\ldots,P$:

$$Z_{ij} = \begin{cases} 1 & \text{if } O^i \mathcal{R} D^j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The representation of binary relations by using pairwise comparison matrices allows to express the relational properties as linear equations. This is a strong property of the RA formalism. We give in Table 2, the linear equations related to the relational properties already presented in Table 1.

Relational properties	Linear equations using the relational matrix
Reflexivity	$X_{ii} = 1 \quad \forall i = 1,\ldots,N$
Irreflexivity	$X_{ii} = 0 \quad \forall i = 1,\ldots,N$
Symmetry	$X_{ii'} = X_{i'i} \quad \forall i,i' = 1,\ldots,N$
Asymmetry	$X_{ii'} + X_{i'i} \leq 1 \quad \forall i,i' = 1,\ldots,N : i \neq i'$
Transitivity	$X_{ii'} + X_{i'i''} - X_{ii''} \leq 1 \quad \forall i,i',i'' = 1,\ldots,N$
Totality	$X_{ii'} + X_{i'i} \geq 1 \quad \forall i,i' = 1,\ldots,N : i \neq i'$

Table 2. Relational properties as linear equations in the RA formalism for $(\mathbb{O}, G(\mathcal{R}))$ binary relations

The second principle of the RA approach is the use of Condorcet's criterion as an association and consensus criterion. This criterion is based upon a similarity or association measure between objects, and a dissimilarity or non association measure between the same items as well.

Let us call respectively, A and $\overline{A}$, the matrices associated to relation[5] and to non relation[6] between pairs of objects. Then the Condorcet's criterion applied to binary relations on a single set is given as follows:

$$\text{Condorcet}(A, \overline{A}, X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(A_{ii'} X_{ii'} + \overline{A}_{ii'} \overline{X}_{ii'} \right) \tag{7}$$

For the case of binary relations on two different sets[7], we have:

$$\text{Condorcet}(A, \overline{A}, Z) = \sum_{i=1}^{N} \sum_{j=1}^{P} \left(A_{ij} Z_{ij} + \overline{A}_{ij} \overline{Z}_{ij} \right) \tag{8}$$

As we can see, the Condorcet's criterion is a quite "logical" criterion: it measures the total agreements ("positive" and "negative" agreements) between two relations. Likewise, the greater the values of A_{ij} (resp. $\overline{A}_{ij}$), the more (resp. less) likely objects O^i and D^j should be in relation from a consensus standpoint.

3. The ranking aggregation problem

The ranking aggregation problem consists in looking for a consensual ranking (or order relation) on a set of objects (or alternatives) that summarizes a set of individual rankings (or a set of several criteria). This problem was firstly mathematically addressed by A. de Condorcet (14) in the context of voting theory and decision making. Historically, it is the first background of the RA approach and the first consistent foundation of such a theoretical framework.

The first aspect of this task consists in aggregating individual rankings in a natural manner. Indeed, suppose that we have items who are described by two real continuous quantitative variables such as their height (in centimeters) and their weight (in kilograms). These two real continuous quantitative variables induce a ranking among the items (the smallest to the tallest for example). How can we compute a consensual ranking that could efficiently summarize the rankings given by the height and the weight ?

In statistics, suppose we want to measure a central trend for the variable "height", it is then possible to compute the mean of this variable for instance. While computing the mean, we have to sum up the heights over all the items and we divide the obtained value by N. Here, the addition is possible because we aggregate "centimeters with centimeters". On the contrary, it is not obvious to aggregate for each item, his height and his weight. Indeed, it is a non sense to add centimeters with kilograms. Thus, how could we proceed to aggregate both variables ?

In order to answer this question, the RA suggests to compute the relational matrices associated to the individual rankings induced by the real continuous quantitative variables. Let us

[5] For two objects O^i and $O^{i'}$, this measure gives the "strength" of the relation $O^i \mathcal{R} O^{i'}$.

[6] For two objects O^i and $O^{i'}$, this measure gives the "strength" of the relation $O^i \overline{\mathcal{R}} O^{i'}$.

[7] In the case of two different sets, A_{ij} gives the "strength" of the relation $O^i \mathcal{R} D^j$ and $\overline{A}_{ij}$ gives the "strength" of the relation $O^i \overline{\mathcal{R}} D^j$.

suppose that we have M real continuous quantitative variables denoted by $V^k; k = 1, \ldots, M$. Let V_i^k be the value assigned to item O^i with respect to variable V^k. Then, for each variable, we can associate its following relational matrix:

$$C_{ii'}^k = \begin{cases} 1 & \text{if } V_i^k \leq V_{i'}^k \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

For example we build up a relational matrix as follows:

$$\begin{array}{|c|c|} \hline & V^k \\ \hline O^1 & 0.5 \\ O^2 & 0.2 \\ O^3 & 0.6 \\ O^4 & 0.9 \\ \hline \end{array} \quad \rightarrow \quad \begin{array}{c} \\ O^1 \\ O^2 \\ O^3 \\ O^4 \end{array} \begin{array}{c} \begin{array}{cccc} O^1 & O^2 & O^3 & O^4 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{10}$$

Considering the M individual relational matrices, the relation aggregation procedure becomes possible: it is just given by the sum over all the individual relational matrices. By doing so, we define the collective relational matrix also called the "collective Condorcet's matrix", denoted $\mathbf{C}$. The general term of the latter matrix is given by:

$$\mathbf{C}_{ii'} = \sum_{k=1}^{M} C_{ii'}^k = \begin{cases} \text{Nb of variables for which } O^i \\ \text{has a lower rank than } O^{i'} \end{cases} \tag{11}$$

Similarly, we can also define the collective relational matrix $\overline{\mathbf{C}}$ related to the aggregation of the M individual relational matrices $\overline{C}^k$ where:

$$\overline{\mathbf{C}}_{ii'} = \sum_{k=1}^{M} \overline{C}_{ii'}^k = \begin{cases} \text{Nb of variables for which } O^i \\ \text{has not a lower rank than } O^{i'} \end{cases} \tag{12}$$

This aggregation process that we have just introduced, gives a partial answer to the initial problem represented by eq. (3). It gives a subsequent process for aggregating order relations in a natural manner. But now that we have aggregated the different rankings, how can we determine a consensual ranking $\mathcal{R}$ that agrees as much as possible with the M individual rankings ?

In the RA methodology, the consensus ranking is the one that maximizes the Condorcet's criterion where $A_{ii'} = \mathbf{C}_{ii'}$ and $\overline{A}_{ii'} = \overline{\mathbf{C}}_{ii'}$:

$$\text{Condorcet}(A, \overline{A}, X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\mathbf{C}_{ii'} X_{ii'} + \overline{\mathbf{C}}_{ii'} \overline{X}_{ii'} \right) \tag{13}$$

Replacing $\overline{X}_{ii'}$ with $1 - X_{ii'}$ and selecting only the part of the criterion which is dependent on X, we then have the following Condorcet's criterion:

$$\text{Condorcet}(\mathbf{C}, \overline{\mathbf{C}}, X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\mathbf{C}_{ii'} - \overline{\mathbf{C}}_{ii'} \right) X_{ii'} \tag{14}$$

This means that O^i should have more chances to have a lower rank than $O^{i'}$ in the consensual ranking[8], if the number of variables that give a lower rank to O^i than to $O^{i'}$ is greater or equal than the number of variables that give a higher rank to O^i than to $O^{i'}$.

If we suppose, moreover, that there is no missing rank value among the individual rankings, then we have: $\overline{\mathbf{C}}_{ii'} = M - \mathbf{C}_{ii'}$. If we take into account this expression in eq. (14), then we obtain the following simplified Condorcet's criterion:

$$\text{Condorcet}(\mathbf{C}, X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\mathbf{C}_{ii'} - \frac{M}{2} \right) X_{ii'} \tag{15}$$

In the particular case where we do not have any missing value, the consensual ranking will more likely give to O^i a rank lower than to $O^{i'}$[9], provided that the number of variables or criteria which ranked O^i before $O^{i'}$ is greater or equal to the simple majority.

If we simply apply the previous rule we will observe a paradoxical situation most of the time. Indeed, aggregating order relations using the simple majority decision rule does not guarantee to obtain an order relation as a solution[10]. This is the famous "Condorcet's paradox", which states that if a majority of voters prefer "i" to "j" and a majority of voters prefer "j" to "k", it could happen that a majority of voters prefer "k" to "i", thus violating the transitivity condition. Consequently the real problem, we want to solve is now based upon the following key question: how can we determine the strict total order relation that maximizes the Condorcet's criterion ? This problem is unfortunately not so simple, since it is an NP-hard problem (51). Without any algorithmic process, it would need a complete enumeration of all the $N!$ possible solutions to get the final result (by the way, just for N=70, notice that $70! \approx 1.2^{100}$). Thanks to the RA approach we can solve the ranking aggregation problem, through an $0/1$ integer linear programming approach. Furthermore, it is also possible to use a simpler model, based upon binarity relaxation scheme by considering $0 \leq X_{ii'} \leq 1$. In that case, we could use continuous linear programming technique and the "dual of dual process" described in (36) for instance.

According to Table 1, the use of pairwise comparisons matrices allows to turn the relational properties, characterizing an order relation, into linear constraints. Furthermore, once we are given A and $\overline{A}$, we can see that the Condorcet's criterion is a linear objective function with respect to X, the solution we are looking for. As a result, we can model the ranking aggregation problem by maximizing a linear consensus criterion subject to linear constraints. Hence, we can get the exact optimal solution using an integer linear programming solver as mentioned beforehand and as described in (36), (41), (39):

$$\begin{array}{lll} & \max_X \text{Condorcet}(A, \overline{A}, X) & \\ & \text{st} & \\ X_{ii'} \in \{0,1\} & & \text{(binarity)} \\ X_{ii'} + X_{i'i} \leq 1 & \forall i, i' = 1, \ldots, N : i \neq i' & \text{(asymmetry)} \\ X_{ii'} + X_{i'i} \geq 1 & \forall i, i' = 1, \ldots, N : i \neq i' & \text{(totality)} \\ X_{ii'} + X_{i'i''} - X_{ii''} \leq 1 & \forall i, i', i'' = 1, \ldots, N & \text{(transitivity)} \end{array} \tag{16}$$

[8] ie $X_{ii'} = 1$.
[9] ie $X_{ii'} = 1$.
[10] That is to say: irreflexive, asymmetric, total and transitive.

We have presented the general model to solve the ranking aggregation problem using the RA methodology. In eq. (16), X must respect the linear constraints of a strict and total order, but other types of order relations could also be modeled in a similar way. The interested reader could find in (41), (2), other relational properties and their associated linear equations (in terms of the RA formalism).
Other works, related to the ranking aggregation problem, can be found in (4), (21), (47) or (7) for example. In those papers, the ranking aggregation problem, which is also referred as the "median linear ordering problem" or shortly the "linear ordering problem"; is treated from a combinatorial optimization viewpoint. For a study of the complexity of problems like relations aggregation, see (51) for instance.

Among the different contributions in ranking aggregation problems, for which RA approach was used as a basic concept, special attention must be paid to the results obtained by S. Ghashghaie in (17). In this work, it is shown that statistical association criteria for comparing rankings such as Goodman and Kruskal, Somers, Kendall, Deuchler and Kim; differ from the Condorcet's criterion, just by slight changes. We can also mention the following reference too (19), where the author provides a theoretical and axiomatic comparison of Condorcet's criterion against other aggregation criteria.

4. The clustering problem

The RA methodology is still valid when we want to consider other relations and aggregation problems than the ranking aggregation task.

From an algebraic point of view, we can observe that the only difference between a linear order and an equivalence relation mainly consists in replacing the asymmetry property with the symmetry one. From this observation, J.F. Marcotorchino and P. Michaud extended the $0/1$ integer linear programming that optimally solves the rank aggregation problem to the similarities aggregation problem (41), (37). Hence, we get the second main application of the RA methods in data analysis: modeling the clustering of categorical data problem as a linear program.

Suppose that we have at our disposal $(N \times N)$ matrices A and $\overline{A}$ of pairwise similarities and dissimilarities between pairs of objects that we want to cluster. Then we can use the Condorcet's criterion as a clustering function similarly as for the ranking aggregation:

$$\begin{aligned} \text{Condorcet}(A, \overline{A}, X) &= \sum_{i=1}^{N}\sum_{i'=1}^{N} \left(A_{ii'} X_{ii'} + \overline{A}_{ii'} \overline{X}_{ii'} \right) \\ &= \sum_{i=1}^{N}\sum_{i'=1}^{N} \left(A_{ii'} - \overline{A}_{ii'} \right) X_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N} \overline{A}_{ii'} \end{aligned} \tag{17}$$

If we consider the part of eq. (17) which is only a function of X, we can notice that maximizing the Condorcet's criterion in the clustering task, consists in putting[11] objects O^i and $O^{i'}$ in the same cluster[12] if their measure of similarity $A_{ii'}$ is higher than their measure of dissimilarity

[11] In condition to satisfy the relational properties of an order relation see eq. (18).
[12] ie $X_{ii'} = 1$.

$\overline{A}_{ii'}$.

Given A and $\overline{A}$, then looking for a partition which is represented by a relational matrix X and which is aimed at maximizing the Condorcet's criterion, can be obtained by means of 0/1 integer linear programming (41), (37), (32), (31).

$$\begin{array}{l} \max_X \text{Condorcet}(A, \overline{A}, X) \\ \text{st} \\ \begin{array}{lll} X_{ii'} \in \{0,1\} & & \text{(binarity)} \\ X_{ii} = 1 & \forall i = 1,\ldots,N & \text{(reflexivity)} \\ X_{ii'} - X_{i'i} = 1 & \forall i, i' = 1,\ldots,N : i \neq i' & \text{(symmetry)} \\ X_{ii'} + X_{i'i''} - X_{ii''} \leq 1 & \forall i, i', i'' = 1,\ldots,N & \text{(transitivity)} \end{array} \end{array} \tag{18}$$

This model is particularly adapted for clustering objects, described by categorical variables[13]. Hence, the same aggregation method introduced beforehand for dealing with orders relation can be applied here as well.

Suppose that we have M categorical variables denoted by $V^k; k = 1,\ldots,M$, and let denote by V_i^k the class of V^k assigned to object O^i. Then each variable induce an equivalence relation on the objects. As a result, we can associate to each V^k a relational matrix C^k:

$$C_{ii'}^k = \begin{cases} 1 & \text{if } V_i^k = V_{i'}^k \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

For instance, we can get the following relational matrix:

	V^k
O^1	blue
O^2	brown
O^3	brown
O^4	blue

$\rightarrow$

	O^1	O^2	O^3	O^4
O^1	1	0	0	1
O^2	0	1	1	0
O^3	0	1	1	0
O^4	1	0	0	1

(20)

Just by considering the M individual relational matrices, we can, as for order relations, aggregate equivalence relations by summing up the individual relational matrices. We then define the collective relational matrix which general term is given by:

$$\mathbf{C}_{ii'} = \sum_{k=1}^{M} C_{ii'}^k = \begin{cases} \text{Nb of variables for which } O^i \text{ and } O^{i'} \\ \text{are in the same cluster} \end{cases} \tag{21}$$

We can also define the collective relational matrix $\overline{\mathbf{C}}$ related to the aggregation of the individual relational matrices $\overline{C}^k$ where:

$$\overline{\mathbf{C}}_{ii'} = \sum_{k=1}^{M} \overline{C}_{ii'}^k = \begin{cases} \text{Nb of variables for which } O^i \text{ and } O^{i'} \\ \text{are not in the same cluster} \end{cases} \tag{22}$$

[13] However, in section 6, we will consider the case where objects are described by real continuous quantitative variables.

Similarly to the previous section, if we take $A = \mathbf{C}$ and $\overline{A} = \overline{\mathbf{C}}$ and if we replace $\overline{X}_{ii'}$ with $1 - X_{ii'}$ then we first obtain:

$$\text{Condorcet}(\mathbf{C}, \overline{\mathbf{C}}, X) \quad = \quad \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\mathbf{C}_{ii'} - \overline{\mathbf{C}}_{ii'} \right) X_{ii'} \tag{23}$$

Secondly, if we suppose that there is no missing value then we have, $\overline{\mathbf{C}}_{ii'} = M - \mathbf{C}_{ii'}$, and the following simplified Condorcet's criterion:

$$\text{Condorcet}(\mathbf{C}, X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\mathbf{C}_{ii'} - \frac{M}{2} \right) X_{ii'} \tag{24}$$

Maximizing the Condorcet's criterion in order to cluster categorical data, amounts to highly consider to put O^i and $O^{i'}$ in the same cluster[14] of the consensus partition[15], if the number of variables considering that O^i and $O^{i'}$ are in the same cluster is higher than the number of variables considering that O^i and $O^{i'}$ are not in the same cluster. Moreover, if there is no missing value then it is equivalent to say that O^i and $O^{i'}$ are more likely in the same cluster of the consensual partition, if the number of variables indicating that O^i and $O^{i'}$ are in the same cluster is greater or equal to the simple majority $\frac{M}{2}$ of the total number of variables.

Here, it is worth mentioning that the integer linear program given in eq. (18) does not require as an "a priori" hypothesis, the knowledge of the expected number of clusters of the partition we are looking for. This is quite an attractive and interesting property of the RA approach in the clustering context: the number of clusters obtained solving eq. (18) is an optimal inherent value according to Condorcet's criterion.

There are other problems related to the clustering task that have been studied in the context of the RA framework. We quote here some references[16]. In (13), the RA approach is used for studying binary relations over triples of objects. This work led to the definition of association and partitioning criteria for heterogeneous data. In (10), an application in computational linguistics is proposed and particularly for the automatic building of synonyms dictionaries. We can also mention other theoretical contributions from (8), (45), or (50) for example.

More recently, the "Correlation Clustering" (CC) method has been proposed by G. Bansal and al in (6). The similarity matrix, considered here, is built up as follows: we put 1 if objects O^i and $O^{i'}$ are considered as similar and -1 otherwise. In terms of the notations presented in this chapter, this corresponds to the particular case where $A_{ii'} = 1$ if O^i and $O^{i'}$ are similar and $\overline{A}_{ii'} = 1$ if they are not. In (15), the linear program used for approximating the clustering problem is equivalent to eq. (18) except that the unknown relational matrix is $\overline{X}$ with general term $\overline{X}_{ii'} = 1 - X_{ii'}$. This latter representation leads to a distance relation which is irreflexive, symmetric and which satisfies the "triangle inequality" which is the exact dual of the transitivity property[17]. In a recent work, L. Labiod (25) has studied the possible connections between the

[14] In condition to satisfy the relational properties of an equivalence relation see eq. (18).

[15] ie $X_{ii'} = 1$.

[16] But we will show in section 6 other main results obtained by using the RA formalism.

[17] If X satisfies the transitivity inequality given in eq. (18), then it is easy to see that $\overline{X}$ satisfies the triangle inequality: $\overline{X}_{ii''} \leq \overline{X}_{ii'} + \overline{X}_{i'i''}$.

CC methods, the N-Cuts methods and other clustering functions and the RA approach. Accordingly, his conclusions corroborate the fact that the Condorcet's criterion is a central and focal concept.

5. The block seriation problem

Let us consider the case where we have as an input, a $(N \times P)$ 0/1 indicator table $\mathbf{K}$. Then in its original form, the problem of seriation consists in finding two permutations, the first one τ, corresponding to a permutation of the rows of $\mathbf{K}$, and the other one σ, corresponding to a permutation of the columns of $\mathbf{K}$; such that a dense structure "appears" along the diagonal of the permuted K'. A simple example is given below:

$$\mathbf{K} = \begin{array}{c} \\ O^1 \\ O^2 \\ O^3 \\ O^4 \\ O^5 \\ O^6 \\ O^7 \end{array} \begin{array}{c} \begin{array}{ccccccccc} D^1 & D^2 & D^3 & D^4 & D^5 & D^6 & D^7 & D^8 & D^9 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$\downarrow (\tau, \sigma) \tag{25}$$

$$\mathbf{K}' = \begin{array}{c} \\ O^1 \\ O^3 \\ O^2 \\ O^4 \\ O^5 \\ O^6 \\ O^7 \end{array} \begin{array}{c} \begin{array}{ccccccccc} D^1 & D^4 & D^7 & D^2 & D^8 & D^5 & D^3 & D^6 & D^9 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

In a more general perspective, let suppose that we have two $(N \times P)$ matrices, A and $\overline{A}$, such that A_{ij} gives the "strength" of the relation $O^i \mathcal{R} D^j$ and $\overline{A}_{ij}$ gives the "strength" of the relation $O^i \overline{\mathcal{R}} D^j$, $\forall i = 1, \ldots, N; j = 1, \ldots, P$. For example, considering the previous example, we can take $A_{ij} = \mathbf{K}_{ij}$ and $\overline{A}_{ij} = 1 - \mathbf{K}_{ij}$.

Let us moreover denote by $P = P^1 \cup \ldots \cup P^k \cup \ldots P^\kappa$ and $Q = Q^1 \cup \ldots \cup Q^k \cup \ldots Q^\kappa$, two partitions with regards to the set of objects $\mathbb{O}$ and the set of descriptors $\mathbb{D}$. These two partitions have the same number of clusters κ. Then, the block seriation problem can be reshaped under the maximization of the following criterion:

$$F(\kappa, P, Q) = \sum_{k=1}^{\kappa} \left(\sum_{\substack{O^i \in P_k \\ D^j \in Q_k}} A_{ij} + \sum_{\substack{O^i \in P_k \\ D^j \notin Q_k}} \overline{A}_{ij} \right) \tag{26}$$

We can see that the solution $\kappa = 3$; $P^1 = \{O^1, O^3\}, P^2 = \{O^2, O^4, O^5\}, P^3 = \{O^6, O^7\}$ and $Q^1 = \{D^1, D^4, D^7\}, Q^2 = \{D^2, D^8, D^5\}, Q^3 = \{D^3, Dv, O^9\}$; is the triple that maximizes the criterion considering the example given in eq. (25).

We can therefore define the two following assignment matrices:

$$P_{ik} = \begin{cases} 1 & \text{if } O^i \text{ belongs to } P^k \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

$$Q_{jk} = \begin{cases} 1 & \text{if } D^j \text{ belongs to } Q^k \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Using those assignment matrices, the problem can be re-stated as follows:

$$\begin{array}{l} \max_{\kappa,P,Q} F(\kappa,P,Q) = \sum_{k=1}^{\kappa} \left(\sum_{i=1}^{N} \sum_{j=1}^{P} A_{ij} P_{ik} Q_{jk} + \sum_{i=1}^{N} \sum_{j=1}^{P} \overline{A}_{ij} P_{ik} (1 - Q_{jk}) \right) \\ \qquad\qquad \text{st} \\ \begin{array}{ll} P_{ik} \in \{0,1\} & \forall i = 1,\ldots,N; k = 1,\ldots,\kappa \\ Q_{jk} \in \{0,1\} & \forall i = 1,\ldots,N; k = 1,\ldots,\kappa \\ \sum_{k=1}^{\kappa} P_{ik} = 1 & \forall i = 1,\ldots,N \\ \sum_{k=1}^{\kappa} Q_{jk} = 1 & \forall j = 1,\ldots,P \end{array} \end{array} \tag{29}$$

There are $N + P$ linear constraints but the criterion that we have to maximize is quadratic according to $P_{ik}Q_{jk}$ thus we cannot use integer linear programming solvers.

But, considering the criterion given in eq. (29), one can distribute the sum over k into the brackets and introduce the following variable (33), (35):

$$\begin{array}{rcl} Z_{ij} & = & \sum_{k=1}^{\kappa} P_{ik} Q_{jk} \\ & = & \begin{cases} 1 & \text{if } O^i \text{ and } D^j \text{ belong to the same block} \\ 0 & \text{otherwise} \end{cases} \end{array} \tag{30}$$

For example, according to this variable, the optimal solution corresponding to the example given previously is the following one (to be clear we give this solution according to permutations τ and σ):

$$Z = \begin{array}{c} \\ O^1 \\ O^3 \\ O^2 \\ O^4 \\ O^5 \\ O^6 \\ O^7 \end{array} \begin{array}{c} \begin{array}{ccccccccc} D^1 & D^4 & D^7 & D^2 & D^8 & D^5 & D^3 & D^6 & D^9 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array} \tag{31}$$

For instance we can see that the first block constituted of $(P^1;Q^1)$ is given by $(\{O^1,O^3\};\{D^1,D^4,D^7\})$ and all terms of Z corresponding to the Cartesian product $\{O^1,O^3\} \times \{D^1,D^4,D^7\}$ are assigned 1.

The $(N \times P)$ binary matrix Z is interpreted as a relational matrix associated to a binary relation on two sets $\mathbb{O}$ and $\mathbb{D}$. The important facts are that this approach firstly allows us to have a criterion that is independent of κ and which is linear according to Z, and secondly, it is possible to express the relational properties of this particular relation using linear equations.

These linear equations were given in (33):

$$\begin{array}{ll}
\sum_{j=1}^{P} Z_{ij} \geq 1 & \forall i = 1,\ldots,N \\
\sum_{i=1}^{N} Z_{ij} \geq 1 & \forall j = 1,\ldots,P \\
\left\{ \begin{array}{l} Z_{ij} + Z_{i'j} + Z_{ij'} - Z_{i'j'} \leq 2 \\ Z_{ij} + Z_{i'j} + Z_{i'j'} - Z_{ij'} \leq 2 \\ Z_{ij} + Z_{ij'} + Z_{i'j'} - Z_{i'j} \leq 2 \\ Z_{i'j} + Z_{ij'} + Z_{i'j'} - Z_{ij} \leq 2 \end{array} \right. & \forall i = 1,\ldots,N
\end{array} \tag{32}$$

These four latter constraints are called "impossible triads": "at the crossing of two rows and two columns of Z, one cannot get a "1" value three times". More precisely, let us suppose that for the optimal solution, $Z_{ij} = 1$, that is to say, O^i and D^j are in the same block. If furthermore, $Z_{ij'} = 1$ and $Z_{i'j'} = 1$ then we must also have $Z_{i'j} = 1$. In other words, if O^i is in the same block as D^j and $D^{j'}$, and if $O^{i'}$ is in the same block as D^j then $O^{i'}$ should be in the same block as $D^{j'}$.

Using the binary relation formalism, we can see that we can also obtain the relational matrix Z by means of integer linear programming solver:

$$\begin{array}{lll}
\multicolumn{3}{c}{\max_Z \text{Condorcet}(A, \overline{A}, Z)} \\
\multicolumn{3}{c}{\text{st}} \\
Z_{ij} \in \{0,1\} & & \text{(binarity)} \\
\left\{ \begin{array}{l} \sum_{j=1}^{P} Z_{ij} \geq 1 \\ \sum_{i=1}^{N} Z_{ij} \geq 1 \end{array} \right. & \begin{array}{l} \forall i = 1,\ldots,N \\ \forall j = 1,\ldots,P \end{array} & \text{(assignment)} \\
\left\{ \begin{array}{l} Z_{ij} + Z_{i'j} + Z_{ij'} - Z_{i'j'} \leq 2 \\ Z_{ij} + Z_{i'j} + Z_{i'j'} - Z_{ij'} \leq 2 \\ Z_{ij} + Z_{ij'} + Z_{i'j'} - Z_{i'j} \leq 2 \\ Z_{i'j} + Z_{ij'} + Z_{i'j'} - Z_{ij} \leq 2 \end{array} \right. & \forall i, i' = 1,\ldots,N \\ & \forall j, j' = 1,\ldots,P & \text{(impossible triads)}
\end{array} \tag{33}$$

Let us recall that the Condorcet's criterion under its general form, is given as follows:

$$\text{Condorcet}(A, \overline{A}, Z) = \sum_{i=1}^{N} \sum_{j=1}^{P} \left(A_{ij} Z_{ij} + \overline{A}_{ij} \overline{Z}_{ij} \right) \tag{34}$$

If $A = \mathbf{K}$, $\mathbf{K}$ being an indicator table, commonly used in categorical data analysis, then the block seriation model given in eq. (33) gives rise to a biclustering method for this type of table. In this particular case, we have $\overline{A}_{ij} = 1 - \mathbf{K}_{ij}$ and by replacing $\overline{Z}_{ij}$ with $1 - Z_{ij}$, we obtain the following simplified Condorcet's criterion:

$$\text{Condorcet}(\mathbf{K}, Z) = \sum_{i=1}^{N} \sum_{j=1}^{P} \left(\mathbf{K}_{ij} - \frac{1}{2} \right) Z_{ij} \tag{35}$$

We have highlighted the RA approach for biclustering tasks in the particular case of 0/1 data type but, the proposed method can be straightforwardly extended to other types of data such as real continuous quantitative data. In that case, matrices A and $\overline{A}$[18] are required.

[18] $\overline{A}$ could be taken from A.

Block seriation models, through their relational formalism, have generated many research works, both for theoretical and practical purposes. We can quote as an illustrative example C. Bédécarax's Phd Thesis (9) where is defined a more general framework called "quadri-decomposition". From that general model, the above mentioned clustering problems[19] are in fact structural derivatives. Besides, "quadri-decomposition" modeling was successfully applied to the automatic building of dictionaries in computational linguistics see (10). Furthermore, in order to take into account large amount of data, several heuristics have been developed and the interested reader could consult (43) for such algorithms and also (16) for an application to production management optimization.

6. Other results of the RA method in the context of clustering problems

In the analysis of equivalence relations, the RA approach has allowed other interesting contributions.

The first one, concerns the study of numerous "association criteria" crossing categorical variables[20] such as Belson, Lerman, χ^2 of Tchuprow, Jordan, Rand and Janson and Vegelius indexes for instance. Suppose that we have at our disposal two categorical variables V^k and V^l with respectively p_k and p_l categories. Then the previous association criteria are basically defined using the $(p_k \times p_l)$ contingency table $\mathbf{n}^{kl}$ where $\forall (u,v) \in \{1,\ldots,p_k\} \times \{1,\ldots,p_l\}$:

$$\mathbf{n}^{kl}_{uv} = \left\{ \begin{array}{l} \text{Nb of objects that have both category } D^u \text{ of } V^k \\ \text{and category } D^v \text{ of } V^l \end{array} \right. \tag{36}$$

Relational matrices such as C^k and C^l, are other ways to encode categorical variables. Following some previous contributions from M.G. Kendall (24), J.F. Marcotorchino in (28), (29), (30), (31), developed correspondence or transfer formulas that allow one to express the association criteria using relational matrices C^k and C^l. Some of the main correspondence formulas are given below:

$$\sum_{u=1}^{p_k} \sum_{v=1}^{p_l} (\mathbf{n}^{kl}_{uv})^2 = \sum_{i=1}^{N} \sum_{i'=1}^{N} C^k_{ii'} C^l_{ii'} \tag{37}$$

$$\sum_{u=1}^{p_k} (\mathbf{n}^{kl}_{u.})^2 = \sum_{i=1}^{N} \sum_{i'=1}^{N} C^k_{ii'} \tag{38}$$

The reformulation of association criteria into the RA formalism, allows us to model coefficients like Belson, Rand, χ^2 of Tchuprow, Janson and Vegelius..., as particular cases of a general Coefficient, which is nothing but a simple variant of Bravais-Pearson's correlation coefficient $\Delta(C^k, C^l, f, \mu^k, \mu^l)$ see (28), (38), (44), (2), (3).
The RA formalism has allowed to get a deeper understanding of the main differences between several association criteria: in fact, the latent differences between the above mentioned association criteria can be expressed, according to 3 parameters (f, μ^k, μ^l):

- f is a function that transforms the general term of each relational matrix[21]

[19] ie clustering $\mathbb{O}$ or $\mathbb{D}$ or both.

[20] Two way contingency tables analysis.

[21] For example, among the transformation functions that occurred in the relational formalism, Torgerson's transformation is the one related to the Belson criterion.

- μ^k is a central trend (playing the role of a mean) corresponding to C^k
- μ^l is a central trend (playing the role of a mean) corresponding to C^l

In order to illustrate those results, we give as an example, the different formulations of the χ^2 of Tchuprow criterion:

$$\text{Tchuprow}(V^k, V^l) = \frac{\sum_{u,v} \frac{1}{\mathbf{n}_{u.}^{kl}\mathbf{n}_{.v}^{kl}} \left(\mathbf{n}_{uv}^{kl} - \frac{\mathbf{n}_{u.}^{kl}\mathbf{n}_{.v}^{kl}}{N}\right)^2}{\sqrt{(p_k - 1)(p_l - 1)}} \tag{39}$$

$$= \frac{\sum_{i,i'} \left(\frac{C_{ii'}^k}{C_{i.}^k} - \frac{1}{N}\right)\left(\frac{C_{ii'}^l}{C_{i.}^l} - \frac{1}{N}\right)}{\sqrt{\sum_{i,i'} \left(\frac{C_{ii'}^k}{C_{i.}^k} - \frac{1}{N}\right)^2 \sum_{i,i'} \left(\frac{C_{ii'}^l}{C_{i.}^l} - \frac{1}{N}\right)^2}} \tag{40}$$

where $C_{i.}^k = \sum_{i'} C_{ii'}^k$ gives the number of objects that belong to the same cluster[22] of O^i according to V^k. Here, the parameters (f, μ^k, μ^l) corresponding to the χ^2 of Tchuprow is $(f(C_{ii'}) = C_{ii'}/C_{i.}, \mu^k = 1/N, \mu^l = 1/N)$. Another example is the (modified) Rand index, which can also be expressed using this general coefficient Δ: it is linked to the particular coefficient given by $(f(C_{ii'}) = C_{ii'}, \mu^k = 1/2, \mu^l = 1/2)$.

Concerning the relational expression of association criteria, we can also quote the "Maximal Association model" defined in (32) and (48) which is aimed at defining partitioning criteria by aggregating association criteria between relations and X. More precisely, suppose that we have M relational matrices[23], $C^k; k = 1, \ldots, M$, and we want to find out a consensual equivalence relation X. In that case, we can use a particular association criterion $\Delta(C^k, X, f, \mu^k, \mu^l)$ in order to measure the correlation between a given partition C^k and an unknown relational matrix X representing the consensus partition. Then, one can consider to determine X such that it maximizes[24] the following partitioning criterion[25]:

$$\sum_{k=1}^{M} \Delta(C^k, X, f, \mu^k, \mu^l) \tag{41}$$

We can see that the Maximal Association model for partitions gives many solutions to the initial problem given by eq. (3).
Moreover these association criteria, can be interpreted as similarity measures between categorical variables. As a result, one can use those measures to partition categorical variables and apply these results in a dimension reduction purpose. This question was investigated in the context of the RA framework in (1).

The second contribution in data analysis that is worth noticing is related to the measure of similarity between objects. The approach developed in (11), (12), called "regularized similarity" consists in giving automatic weights to the initial variables according to particular models

[22] ie the number of objects that have the same category of O^i according to V^k.

[23] Derived from M categorical variables.

[24] Most of the association criterion's numerators given by particular $\Delta(C^k, X, f, \mu^k, \mu^l)$ lead to linear partitioning criteria according to X and can then be used with integer linear programming.

[25] With respect to the linear equations given in eq. (18).

of weighting, among them let us quote: the logical, the statistical and the probabilistic models.

One can represent a set of M categorical variables using the $(N \times P)$ indicator matrix $\mathbf{K}$ where $P = \sum_{k=1}^{M} p_k$. This matrix is a binary one and we have, $\forall i = 1,\ldots,N; \forall j = 1,\ldots,P$:

$$\mathbf{K}_{ij} = \begin{cases} 1 & \text{if object } O^i \text{ is in the category}^{26}\ D^j \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

We have previously defined $\mathbf{C}_{ii'}$ according to a logical approach in terms of equivalence relations aggregation. We can also have a more geometrical approach since we have:

$$\mathbf{C}_{ii'} = \sum_{j=1}^{P} \mathbf{K}_{ij}\mathbf{K}_{i'j} = \langle \vec{O}^i, \vec{O}^{i'} \rangle \tag{43}$$

where $\vec{O}^i = (\mathbf{K}_{i1},\ldots,\mathbf{K}_{ij},\ldots,\mathbf{K}_{iN})$ and $\langle .,. \rangle$ is the canonical scalar product.
From a geometrical standpoint, the regularized similarity method amounts to exhibiting a diagonal metric, for which the diagonal weights are computed from the categorical variables. The regularized similarity of type α denoted by $A^{\alpha}_{ii'}$ is given by:

$$A^{\alpha}_{ii'} = \sum_{j=1}^{P} \alpha_j \mathbf{K}_{ij}\mathbf{K}_{i'j} \tag{44}$$

For instance, "statistical regularized similarity", defined in (11), gives higher weights to infrequent categories and reciprocally, very low weights to those frequent categories. In this particular case, we have actually $\alpha_j = 1/\mathbf{K}_{.j}$. We can observe that the model supposes that if two objects have a rare category in common then their similarity should be higher than if their shared category were frequent.

This particular similarity measure is related to the χ^2 metric used in Correspondence Factor Analysis methods (20). We can mention here, the following paper (34), where the latent link between the RA method and Factor Analysis methods is explained.
Basically, the Condorcet's criterion, while we use the statistical regularized similarity, becomes highly related to "Inertial criteria". More precisely, J.F. Marcotorchino showed that the Condorcet's criterion associated to the similarity matrix of general term $A^{\alpha}_{ii'}$, and the dissimilarity matrix of general term $\overline{A}^{\alpha}_{ii'} = \frac{A^{\alpha}_{ii} + A^{\alpha}_{i'i'}}{2} - A^{\alpha}_{ii'}$, is the non trivial partitioning criterion strongly relevant to the family of criteria based upon inertia or "variance". This result led to the design of a full methodology called "Relational Factor Analysis method" (34) that consists in coupling the representations of clusters in terms of "bubbles", resulting from the RA[27] method, with the projection of objects on a 2D space obtained after applying the Factor Analysis method. Both approaches complement each other, because they resectively maximize objective criteria that are very close.

The third contribution that we can mention finally, is based on the extension of the RA method which is well-designed for clustering categorical data, to deal with objects described by real continuous quantitative variables. Indeed, we can notice that similarity measures can be expressed through scalar products or in a general manner by kernels. Then if we take

[27] With the statistical regularized similarity.

$A_{ii'} = \langle \vec{O}^i, \vec{O}^{i'} \rangle$ and $\overline{A}_{ii'} = \frac{\langle \vec{O}^i, \vec{O}^i \rangle + \langle \vec{O}^{i'}, \vec{O}^{i'} \rangle}{2} - \langle \vec{O}^i, \vec{O}^{i'} \rangle$ which is equal to $\frac{1}{2}||\vec{O}^i - \vec{O}^{i'}||$, we have the following simplified Condorcet's criterion:

$$\text{Condorcet}(A, X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\langle \vec{O}^i, \vec{O}^{i'} \rangle - \frac{1}{2} \left(\frac{\langle \vec{O}^i, \vec{O}^i \rangle + \langle \vec{O}^{i'}, \vec{O}^{i'} \rangle}{2} \right) \right) X_{ii'} \tag{45}$$

Following previous results given in (34), we can show that the criterion based upon the inertial difference can be expressed using the RA formalism as follows:

$$\begin{array}{rcl} \mathbb{IB}(X) - \mathbb{IW}(X) & = & \frac{2}{N} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\langle \vec{O}^i, \vec{O}^{i'} \rangle - \frac{1}{2} \left(\frac{\langle \vec{O}^i, \vec{O}^i \rangle + \langle \vec{O}^{i'}, \vec{O}^{i'} \rangle}{2} \right) \right) \left(\frac{X_{ii'}}{X_{i.}} \right) \\ & & - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \langle \vec{O}^i, \vec{O}^{i'} \rangle \end{array} \tag{46}$$

$\mathbb{IB}(X)$ is the "between classes inertia" related to a partition represented by its relational matrix X, $\mathbb{IW}(X)$ is the "within classes inertia[28]" and $\langle .,. \rangle$ is a scalar product (kernel).

In the formula (46), if we look at the subpart depending on X, we can observe that the main difference between Condorcet's criterion and the inertial difference criterion resides in the fact that the first one does not weight the general term $X_{ii'}$, whereas the second one integrates a weight[29], $1/X_{i.}$, to the general term $X_{ii'}$.

Finally, by considering eq. (46) and the $0/1$ integer linear program described in eq. (18) we can thus extend the RA approach for clustering problems to the more general case of objects described by real continuous quantitative variables.

7. Acknowledgements

Part of this work was supported by Cap Digital (french business cluster in the digital content creation and knowledge management area)

8. References

[1] Abdallah, H.: Application de l'analyse relationnelle pour classifier descripteurs et modalités en mode discrimination. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6, France) (1996)

[2] Ah-Pine, J.: Sur des aspects algébriques et combinatoires de l'analyse relationnelle. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6) (2007)

[3] Ah-Pine, J., Marcotorchino, J.F.: Statistical, geometrical and logical independences between categorical variables. In: Proceedings of Applied Stochastic Models and Data Analysis, Chania Crete, 2007 (2007)

[4] Arditti, D.: Un algorithme de recherche d'un ordre induit par des comparaisons par paires. Note technique CNET, Paris, France (1982)

[5] Arrow, K.J.: Social choice and individual values 2nd ed. Wiley, New-York (1963)

[6] Bansal, N., Blum, A., Chawla, S.: Correlation clustering. In: IEEE Symposium on Foundations of Computer Science (FOCS) (2002)

[28] See (49) for example for the definition of these quantities using euclidean distances.

[29] Which consists in dividing by the number of objects that belong to the same cluster as O^i.

[7] Barthélémy, J.P., Guenoche, A., Hudry, O.: Median linear orders: heuristics and a branch and bound algorithm. European Journal of Operational Research **41**, 313–325 (1989)
[8] Barthélémy, J.P., Monjardet, B.: The median procedure in cluster analysis and social choice theory. Mathematical social sciences **1**, 235–267 (1991)
[9] Bédécarrax, C.: Classification automatique en analyse relationnelle: la quadridécomposition et ses applications. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6, France) (1989)
[10] Bédécarrax, C., Warnesson, I.: Relational analysis and dictionnaries. In: Proceedings of Applied Stochastic Models and Data Analysis, pp. 131–151. Wiley, London, New-York (1989)
[11] Benhadda, H.: La similarité régularisée et ses applications en classification automatique. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6, France) (1998)
[12] Benhadda, H., Marcotorchino, J.F.: Introduction à la similarité régularisée en analyse relationnelle. Revue de statistique appliquée **46** (1998)
[13] Chah, S.: Nouvelles techniques de codage d'association et de classification. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6, France) (1986)
[14] Condorcet, M.J.A.: Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Paris (1785)
[15] Demaine, E., Immorlica, N.: Correlation clustering with partial information. In: International Workshop on Approximation Algorithms for Combinatorial Optimization (APPROX), Princeton, New Jersey (2003)
[16] Garcia, H., Proth, J.M.: Group technology in production management. Applied Stochastic Models and Data Analysis **1**, 25–34 (1985)
[17] Ghashghaie, S.: Agrégation relationnelle des données ordinales: généralisation de critères d'association. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6, France) (1990)
[18] Gionis, A., Mannila, H., Panayiotis, T.: Clustering aggregation. ACM Transactions on Knowledge Discovery from Data **1** (2007)
[19] Giraud, S.: Comparaison de règles d'agrégation d'ordres et études de règles multicritères d'aide à la décision. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6, France) (1985)
[20] Greenacre, J.: Theory and applications of correspondence analysis. Academic Press (1984)
[21] Grötschel, M., Jünger, M., Reinelt, G.: A cutting plane algorithm for the linear ordering problem. Operations research **32**, 1195–1220 (1984)
[22] Joachims, T., Hopcroft, J.: Error bounds for correlation clustering. In: Proceedings of the 22nd international conference on Machine learning (2005)
[23] Kemeny, J.G., Snell, J.L.: Mathematical models in the social sciences. MIT Press, Boston (1972)
[24] Kendall, M.G.: Rank correlation methods. Griffin Londres (1970)
[25] Labiod, L.: Contribution au formalisme relationnel des classifications simultanes de deux ensembles. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6, France) (2008(forthcoming))
[26] Leclerc, B., Monjardet, B.: Latticial theory of consensus. In: W. Barnett, H. Moulin, M. Salles, N.E. Schofield (eds.) Social Choice, Welfare, and Ethics, pp. 145–160. Cambridge University Press (1995)
[27] Lerman, I.C.: Classification et analyse ordinale de données. Dunod, Paris (1981)

[28] Marcotorchino, J.F.: Utilisation des comparaisons par paires en statistiques des contingences i. Etudes du Centre scientifique IBM France, Paris **F069** (1984)
[29] Marcotorchino, J.F.: Utilisation des comparaisons par paires en statistiques des contingences ii. Etudes du Centre scientifique IBM France, Paris **F071** (1984)
[30] Marcotorchino, J.F.: Utilisation des comparaisons par paires en statistiques des contingences iii. Etudes du Centre scientifique IBM France, Paris **F081** (1985)
[31] Marcotorchino, J.F.: Cross association measures and optimal clustering. In: Proceedings in Computational statistics, pp. 188–196. Physica-Verlag Heidelberg (1986)
[32] Marcotorchino, J.F.: Maximal association theory as a tool of research. In: W. Gaul, M.e. Schader (eds.) Classification as a tool of research, pp. 275–288. North Holland Amsterdam (1986)
[33] Marcotorchino, J.F.: An unified approach to the block seriation problems. Applied Stochastic Models and Data Analysis **3** (1987)
[34] Marcotorchino, J.F.: L'analyse factorielle relationnelle: parties i et ii. Etudes du CEMAP, IBM France, Paris **MAP-03** (1991)
[35] Marcotorchino, J.F.: Seriation problems: an overview. Applied stochastic models and Data Analysis **7**, 139–151 (1991)
[36] Marcotorchino, J.F., Michaud, P.: Optimisation en analyse ordinale des données. Masson Editions, Paris (1979)
[37] Marcotorchino, J.F., Michaud, P.: Heuristic approach of the similarity aggregation problem. Methods of operation research **43**, 395–404 (1981)
[38] Messatfa, H.: Unification relationnelle des critères et structures optimales des tables de contingences. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6, France) (1989)
[39] Michaud, P.: Opinions aggregation. In: New trends in data analysis and applications, pp. 5–27. North Holland Amsterdam (1983)
[40] Michaud, P.: Condorcet: A man of the avant-garde. Applied stochastic models and Data Analysis **3**, 173–189 (1987)
[41] Michaud, P., Marcotorchino, J.F.: Optimisation en analyse de données relationelles. In: Data Analysis and informatics. North Holland Amsterdam (1980)
[42] Mirkin, B.G.: Approximation problems in space of binary relations and analysis of non quantitative characteristics. Automatics and remote control (russian) **9**, 51–61 (1974)
[43] Mutel, B., De Guio, R., Bouzid, L.: Application of relational analysis techniques to production data structuring. In: Proceedings of Applied Stochastic Models and Data Analysis, Granad, Spain, pp. 493–507 (1991)
[44] Najah Idrissi, A.: Contribution à l'unification de critères d'association pour variables qualitatives. Ph.D. thesis, University of Pierre and Marie Curie (Paris 6, France) (2000)
[45] Owsinski, J.W.: Optimization in clustering: an approach and other approaches. Control and cybernetics **15**, 107–114 (1986)
[46] Régier, S.: Sur quelques aspects mathématiques des problèmes de classification automatique. ICC Bulletin **4**, 175–191 (1965)
[47] Reinelt, G.: The linear ordering problem: algorithms and applications. Research and Exposition in Mathematics **8** (1985)
[48] Saporta, G.: About maximal association criteria in linear analysis and in cluster analysis. In: H.H.e. Bock (ed.) Classification and related methods of Data Analysis, pp. 541–550. North Holland Amsterdam (1988)
[49] Saporta, G.: Probabilités, analyse des données et statistiques (2nd édition). Technip, Paris (2006)

[50] Schader, M., Tüshaus, U.: Analysis of qualitative data: a heuristic for finding a complete preorder. In: H.H.e. Bock (ed.) Classification and related methods of Data Analysis, pp. 341–346. North Holland Amsterdam (1988)
[51] Wakabayashi, Y.: On the complexity of computing medians of relations. Resenhas IMEUSP **3**, 323–349 (1998)
[52] Zahn, C.T.: Approximating symmetric relations by equivalence relations. SIAM Journal of applied mathematics **12**, 840–847 (1964)