An overview of the Statistical Implicative Analysis (SIA) development

Régis Gras and Pascale Kuntz

Laboratoire d'Informatique de Nantes Atlantique
Equipe COnnaissances & Décision
Site Ecole Polytechnique de l'Université de Nantes
La Chantrerie — BP 50609 — 44306 Nantes cedex 3
regisgra@club-internet.fr, pascale.kuntz@univ-nantes.fr

Summary. This paper presents an overview of the Statistical Implicative Analysis which is a data analysis method devoted to the extraction and the structuration of quasi-implications. Originally developed by Gras [11] for applications in the didactics of mathematics, it has considerably evolved and has been applied to a wide range of data, in particular in data mining. This paper is a synthesis which both briefly presents the basic statistical framework of the approach and details recent developments.

Key words: quasi-implication, implication intensity, implicative graph, implicative hierarchy, typicality

1 Introduction

Two important components are involved in the operational human processes of knowledge acquisition: facts and rules between facts or between rules themselves. Through one's own culture and one's own personal experience, the learning process integrates a progressive elaboration of these knowledge forms. It can be faced with regressions, questions or changes which arise from decisive quashing, but the knowledge forms contribute to maintain a certain equilibrium. The rules formed inductively become quite stable when their success number -which depends on their explicative or inferential quality- reaches a certain level of confidence. At first, it is often difficult to replace an initial rule by another when few counter-examples appear. If they increase, the confidence in the rule can decrease and the rule can be reajusted or even rejected. However, when confirmations are numerous and counter-examples are rare, the rule is robust and can stay in our minds. For instance, let us consider the acceptable rule "All Ferraris are red". Even if one or two counter-examples happen this rule is maintained, and it will be even confirmed again by new examples.

Hence, contrary to what happens in mathematics where rules do not allow for any exception, the rules considered in human sciences are considered to be acceptable when the number of counter-examples remains "tolerable" in view of the number of situations where they are positive and efficient. In data analysis, the problem is to determine a consensus criterion which quantifies the confidence quality level of the rule according to the user's requirements. Our approach rests on three epistemological assumptions. The criterion is statistical. It is non linearly, robust to noise (i.e. not very influenced by the first counter-examples), and it becomes very low if the counter-examples often reappear. Our choice can be questioned, however it has been confirmed in various situations.

1.1 From didactics to data mining

"If a question is more complex than another, then each pupil who succeeds in the first one should also succeed in the second one". Every teacher knows that this situation shows exceptions whatever the complexity degree between questions. The evaluation and the structuration of such implicative relationships between didactic situations are the generic problems at the origin of the development of the Statistical Implicative Analysis (SIA) [11]. These problems, which have also drawn attention from psychologists interested in ability tests [5, 27], have known a significant renewed interest in the last decade in data mining.

Indeed, quasi-implications, also called association rules in this field, have become the major concept in data mining to represent implicative trends between itemset patterns. In data mining, the paradigmatic framework is the so-called basket analysis where a quasi-implication $T_i \rightarrow T_j$ means that if a transaction contains a set of items T_i then it is likely to contain a set of items T_j too. For simplicity's sake, let us now on call "rule" a quasi-implication. In data mining, rules are computed on large size databases. From the seminal work of Agrawal *et al.* [1], numerous algorithms have been proposed to mine such rules. Most of them attempt to extract a restricted set of relevant rules, easy to interpret for decision-making. Yet, comparative experiments have shown that results may vary with the choice of rule quality measures (e.g. [13, 25]). In the rich literature devoted to this problem, interestingness measures are often classified into two categories: the subjective (user-driven) ones and the objective (data-driven) ones. Subjective measures aim at taking into account unexpectedness and actionability relatively to prior knowledge, while objective measures give priority to statistical criteria. Among the latter, the most commonly used criterion for quantifying the quality of a rule $a \rightarrow b$ is the combination of the support (the frequency $f(a \wedge b)$) which indicates whether the items a and b occur reasonably often in the database, with the confidence (the conditional frequency). However, it is well-known that the confidence presents a major default: it is insensitive to the dilatation of $f(a)$, $f(b)$ and the database size. Other functions measure a link or an absence of

link between the items but, like χ^2, they do not clearly specify the direction of the relationship. Moreover, in addition to rule filtering, rule structuring is necessary to highlight relationships and makes rule interpretation both easier and more accurate.

The SIA provides a complete framework to evaluate the interestingness of the rules and to structure them in order to discover relationships at different granularity levels. The underlying objective is to highlight the emerging properties of the whole system which can not be deduced from a simple decomposition into sub-parts (e.g. [30]). All these properties, which emerge from complex interactions -probably non linear-, contribute to the interpretation of the global nature of the system.

1.2 Contents of the paper

Section 2 presents the statistical framework to measure the rule quality: we first remind the reader the definition of the implication intensity for binary variables and propose different properties. Section 3 presents the extensions of the basic definition for different types of variables (modal, frequential, interval), and an entropic version adapted to large datasets. The following sections are concerned with rule structuration. Section 4 defines the implicative graph. Section 5 generalizes the notion of rule to the notion of R-rule (rule of rule), and section 6 describes the combinatorial structure of an implicative hierarchy whose elements are R-rules. Aids for analyzing these complex structures are developed in section 7 (significative levels of the implicative hierarchy) and section 8 (supplementary individuals and variables). An illustration from a real data corpus coming from a survey on teacher's perception of training in mathematics is presented in section 9.

2 The implicative intensity for the binary case

2.1 The basic situation

Let us consider a population E of n objects or individuals described by a finite set V of binary variables (attributes, criteria, scores, . . .). We are here interested in the following question: "To what extent the variable b is true when the variable a is true" ? In other words, "do the subjects have a tendency for having b when we know that they have a? " In real-life situations — e.g. in human sciences— deductive theorems of the logical form $a \Rightarrow b$ are often difficult to establish because of the exceptions. Consequently, it is necessary to "mine" the dataset to extract rules reliable enough to conjecture causal relationships which structure the population. At the descriptive level, they allow to detect a certain stability in the structuration. And, at the predictive level, they allow to make assumptions. However, the rule mining processes require rigorous approaches which prevent a too flimsy empiricism.

2.2 The statistical framework

Our approach, based on the non-parametric test reasoning, is close to the Likelihood Linkage Analysis (LLA) developed by I.C. Lerman [26]. The quality measurement of an implicative relationship $a \to b$ is based on the unlikelihood of the counter-example number in the dataset i.e. cases where b is false when a is true [11, 12, 14]. To quantify this unlikelihood, we compare the deviation between the contingency and a theoretical model associated with a random drawing. In exploratory data analysis, we consider the deviation value and not just the H_0 acceptation/reject. This measure quantifies the "surprisingness" of the expert faced with a number of counter-examples improbably small for an independence assumed between the variables and for the cardinalities of the considered data.

More precisely, let us denote by $A \subset E$ the subset of individuals for which a is true, by $\overline{A}$ its complementary set and by $n_a = card\,(A)$ (resp. $n_{\overline{a}}$) the cardinal of A (resp. $\overline{A}$). The logical rule $A \Rightarrow B$ is true when $A \subset B$. However, this strict inclusion is exceptionally observed in real-life situations; in practice, it is quite common to observe a few subjects where a is true and b is false without having the general trend to have b when a is true contested. Consequently, we consider in the following quasi-rules —called rules for simplicity's sake— of the form $a \to b$.

2.3 Definitions

To accept or reject $a \to b$ it is quite common to consider the number $n_{a \wedge \overline{b}} = card\,(A \cap \overline{B})$ of counter-examples. However, to quantify the surprisingness of the rule, this number must be relativized according to n, n_a and n_b. Intuitively, it is all the more surprising to discover that a rule has a small number of counter-examples as the data set is large. The objective of the *implicative intensity* is precisely to express the unlikelihood of $n_{a \wedge \overline{b}}$ in E.

We compare the observed number of counter-examples $n_{a \wedge \overline{b}}$ with the number of expected counter-examples for an independence hypothesis. Like I.C. Lerman with the similarity in LLA [26], we randomly draw two subsets X and Y of, respectively, n_a and n_b elements.

Definition 1. *The rule $a \to b$ is said to be admissible for a given threshold α if the probability of having the observed number of counter-examples $card\,(A \cap \overline{B})$ greater than the expected number $card\,(X \cap \overline{Y})$ is smaller than α:*

$$\Pr\left(card\,(X \cap \overline{Y}) \leq card\,(A \cap \overline{B})\right) \leq \alpha$$

The distribution of $card\,(X \cap \overline{Y})$ depends on the drawing pattern. When X and Y are draw with throw-in the distribution is Binomial, otherwise it is Hypergeometric.

Remark 1. For a certain process of drawing, the random variable $card\left(X \cap \overline{Y}\right)$ follows a Poissonian distribution $P\left(\lambda\right)$ with $\lambda = \frac{n_a n_{\overline{b}}}{n}$.

Let us consider a process where the individuals dynamically arrive e.g. a flow of transactions which fill up a database. We stop the process when there are n_a individuals with a true and n_b individuals with b true. Let $card\left(X \cap \overline{Y}\right)$ be the random variable associated with the counter-example number during the process. We suppose that the process checks three hypotheses : *(i)* the waiting times for the events $(a$ and $\overline{b})$ are independent random variables, *(ii)* the distribution of the number of events which happen in the interval $[t, t+T]$ only depends on T, *(iii)* two events can not simultaneously happen.

Consequently, the number of events which happen during a fixed period follows a Poissonian distribution $P\left(\lambda\right)$ where λ is the cadence of the event arrival.

The probability of the event $(a = 1)$ (resp. $(b = 0)$) is estimated by the frequency n_a/n (resp. $n_{\overline{b}}/n$). Then, the probability of the joint event $(a = 1$ and $b = 0)$ is estimated by

$$\frac{n_a}{n} \cdot \frac{n_{\overline{b}}}{n}$$

Hence, for a flow of n individual, the arrivals of the event $(a = 1$ and $b = 0)$ follow a Poissonian distribution with parameter $\lambda = \frac{n_a n_{\overline{b}}}{n}$.

Consequently,

$$\Pr\left(card\left(X \cap \overline{Y}\right) = s\right) = e^{-\lambda} \frac{\lambda^s}{s!}$$

and the probability that the chance leads to a greater number of counter-examples than those observed is defined by

$$\Pr\left(card\left(X \cap \overline{Y}\right) \leq card\left(A \cap \overline{B}\right)\right) = \sum_{s=0}^{card\left(A \cap \overline{B}\right)} \frac{\lambda^s}{s!} e^{-\lambda}$$

In the following, we consider the Poissonian distribution. In the classical approximation conditions, the other distributions converge on the Poissonian type.

Let us consider, for $n_{\overline{b}} \neq 0$, the standardized random variable $Q\left(a, \overline{b}\right)$:

$$Q\left(a, \overline{b}\right) = \frac{card\left(X \cap \overline{Y}\right) - \frac{n_a n_{\overline{b}}}{n}}{\sqrt{\frac{n_a n_{\overline{b}}}{n}}}$$

We denote by $q\left(a, \overline{b}\right)$ the observed value of $Q\left(a, \overline{b}\right)$ in the experimental realization. It is defined by

$$q\left(a, \overline{b}\right) = \frac{n_{a \wedge \overline{b}} - \frac{n_a n_{\overline{b}}}{n}}{\sqrt{\frac{n_a n_{\overline{b}}}{n}}}$$

This value measures a deviation between the contingency and the expected value when a and b are independent.

When the approximation is justified (e.g. $\lambda > 4$) the random variable $Q\left(a, \overline{b}\right)$ is approximatively $N(0, 1)$-distributed.

Definition 2. *The implication intensity $\varphi(a, b)$ of the rule $a \rightarrow b$ is defined by*

$$\varphi(a, b) = 1 - \Pr\left(Q\left(a, \overline{b}\right) \leq q\left(a, \overline{b}\right)\right) = \frac{1}{\sqrt{2\pi}} \int_{q\left(a, \overline{b}\right)}^{\infty} e^{-\frac{t^2}{2}} dt$$

if $n_b \neq n$, and $\varphi(a, b) = 0$ otherwise.

Definition 3. *The implication intensity $\varphi(a, b)$ is admissible for a given threshold α if $\varphi(a, b) \geq 1 - \alpha$.*

The implication intensity measures the surprisingness to observe a small number of counter-examples. It is an inductive and informative quality measure. Consequently, if the rule is trivial —-i.e. when B is small or equal to E— this surprisingness is small.

Proposition 1. *[12] Let us suppose that n_a is fixed and $A \subset B$. If n_b tends towards n, then $\varphi(a, b)$ tends towards 0.*

We set $\varphi(a, b) = 0$ if $n_b = n$ by continuity (consequence of the property 1). If $A \subset B$ then $\varphi(a, b)$ can be smaller than 1 when the surprisingness is not sufficient.

2.4 Comparison with some classical measures

The observed quasi-implication $q\left(a, \overline{b}\right)$ is not symmetrical. It is different from the Pearson's correlation coefficient $\rho(a, b)$ which measures the linkage between a and b.

Proposition 2. *[12] Let $\rho(a, b)$ be the value of the Pearson's correlation between the binary variables a and b. If $q\left(a, \overline{b}\right) \neq 0$ then*

$$\frac{\rho(a, b)}{q\left(a, \overline{b}\right)} = -\sqrt{\frac{n}{n_b n_{\overline{a}}}}$$

The variation of the implication intensity is different from the Loevinger's coefficient [27] and from the confidence $conf(a, b) = n_{a \wedge b}/n_a$. It increases non linearly with the increasing of E, A and B, and it decreases with the trivial situations. Moreover, the maximal intensity is not necessarily reached for the inclusion $A \subset B$; indeed, the inductive quality may be quite low, whereas $conf(a, b) = 1$ [13].

2.5 Stability of the implication intensity

The implication intensity is noise-resistant in the neighbourhood of $n_{a \wedge \bar{b}} = 0$ [13].

In the following, we study the sensitivity of $\varphi(a, b)$ for small variations of the parameters n, n_a, n_b and $n_{a \wedge \bar{b}}$. Previous numerical experiments have confirmed the influence of the parameter variations on φ [10, 13]. Here, we study the differentiation of q.

Let us consider the parameters n, n_a, n_b and $n_{a \wedge \bar{b}}$ as real numbers which satisfy the following inequalities: $n_{a \wedge \bar{b}} \leq \inf(n_a, n_b)$ et $\sup(n_a, n_b) \leq n$. In this case, q can be considered as a continuous differentiable function:

$$dq = \frac{\partial q}{\partial n} dn + \frac{\partial q}{\partial n_a} dn_a + \frac{\partial q}{\partial n_b} dn_b + \frac{\partial q}{\partial n_{a \wedge \bar{b}}} dn_{a \wedge \bar{b}}$$

To study the variability of q depending on $n_{\bar{b}}$, we replace $n_{\bar{b}}$ by $n - n_{\bar{b}}$, and consequently the sign in the partial derivative.

Example 2. Let us suppose that n_a is constant, and that n_b and $n_{a \wedge \bar{b}}$ may vary. Then,

$$\frac{\partial q}{\partial n_b} = \frac{1}{2} n_{a \wedge \bar{b}} \left(\frac{n_a}{n}\right)^{1/2} (n - n_b)^{-3/2} + \frac{1}{2} \left(\frac{n_a}{n}\right)^{1/2} (n - n_b)^{-1/2}$$
$$\frac{\partial q}{\partial n_{a \wedge \bar{b}}} = \frac{1}{\sqrt{\frac{n_{a \wedge \bar{b}}}{n}}}$$
$$\frac{\partial q}{\partial n_a} = 0$$

Consequently, if Δn_b and $\Delta n_{a \wedge \bar{b}}$ are positive, then $\Delta q(a, \bar{b})$ is positive. This property can be interpreted as follows: for fixed n and n_a, the implication intensity decreases when the numbers of the b examples and the $a \Rightarrow b$ counter-examples increase. The implication intensity is maximal for the observed values n_b and $n_{a \wedge \bar{b}}$, and minimal for $n_b + \Delta n_b$ and $n_{a \wedge \bar{b}} + \Delta n_{a \wedge \bar{b}}$.

To examine the sensibility of the implication intensity, we consider φ as a function of q:

$$\varphi(q) = \frac{1}{\sqrt{2\pi}} \int_q^\infty e^{-t^2} dt$$

By differentiation, we obtain

$$\frac{d\varphi}{dq} = -\frac{1}{\sqrt{2\pi}} e^{-q^2} < 0$$

This result confirms that the implication intensity decreases with q, and it gives the speed of the variation.

With a similar approach, let us compare the stability of φ with the stability of the confidence $conf(a, b)$. The sensibility of $conf$ to the variation of the counter-examples is defined by

$$\frac{\partial c}{\partial n_{a \wedge \overline{b}}} = -\frac{1}{n_a}$$

Consequently, as expected, the confidence increases when $n_{a \wedge \overline{b}}$ decreases. However, the variation of the decreasing speed is constant whatever n and n_b. This situation highlights the limits of the parameter role in the sensitivity of the measure.

3 Extensions to different types of variables

3.1 Modal and frequential variables

The basic situation

The first applicative framework of this research was concerned with the representation that the teachers have of their own practice [3]. In a survey, a set of teachers has been asked to order a list of significative words depending on their importance. The resulting implications were: "if I select a word x with the importance i_x then I select the word y with the importance $i_y \geq i_x$".

In this case, we consider modal variables $a \in [0, 1]$ which describe satisfaction degrees. A similar case appears in situations where the variable frequency can be interpreted as a pre-order on the set of the values given by the subjects. Such situation appears in didactics when we study the success frequency for a test composed of questions coming from different domains.

Formalization

Let us denote by $a(i)$ and $\overline{b}(i)$ the values of $i \in E$ for the modal variables a and $\overline{b}$, and by s_a and s_b their empirical standard-deviations.

Definition 4. *[24]. For a pair (a, b) of modal variables, the implication intensity, called the propension index, is defined by*

$$q_p\left(a, \overline{b}\right) = \frac{\sum_{i \in E} a(i)\,\overline{b}(i) - \frac{n_a n_{\overline{b}}}{n}}{\sqrt{\frac{\left(n^2 s_a^2 + n_a^2\right)\left(n^2 s_{\overline{b}}^2 + n_{\overline{b}}^2\right)}{n^3}}}$$

Proposition 3. *When a and b are binary variables then $q_p\left(a, \overline{b}\right) = q\left(a, \overline{b}\right)$.*

In this case, it is easy to prove that $n^2 s_a^2 + n_a^2 = n n_a$, $n^2 s_{\overline{b}}^2 + n_{\overline{b}}^2 = n n_{\overline{b}}$ and $\sum_{i \in E} a(i)\,\overline{b}(i) = n_{a \wedge \overline{b}}$.

This extension remains valid for the frequential variables and the positive numerical variables when they are normalized: $\widetilde{a}(i) = a(i) / \max_{i \in E} a(i)$.

A similar measure has been recently introduced by Régnier and Gras [29] for ranking variables associated with a total order on a set of choices presented to a judge population. In this case, the considered implication is "if an object i is ordered by the judges at a place p_i then an object j is ordered by the same judges at a place $p_j > p_i$".

3.2 Variables on intervals

The basic situation

Let us consider a given set of biometric data. The considered implication is "if the weight of a male is between 65 and 70 kgs then his height is between 1.70 and 1.76 m".

More generally, let us consider two real variables a and b with a finite number of values in the respective intervals $A = [a_1, a_2]$ and $B = [b_1, b_2]$. Roughly speaking, the problem consists in finding implicative trends between representative unions of sub-intervals of A and B.

Main steps of the heuristic

The problem is decomposed in two steps. First, we partition the intervals A and B in a finite number of sub-intervals $\{A_1, A_2, \ldots, A_p\}$ and $\{B_1, B_2, \ldots, B_q\}$ which depend on the structure of the a and b distributions: there is an internal statistical homogeneity in each A_i (resp. B_i) and a high dispersion between each pair A_i, A_j (resp. B_i, B_j). Second, we compute the most significative implicative trends between unions of A_i and unions of B_j.

We have adapted the k-means algorithm for the interval partitioning problem [16]. The quality criteria of the partition are the intra-class and the inter-class inertia. Let $\pi(A)$ and $\pi(B)$ be two partitions obtained by this approach which respectively contain n_A and n_B elements. We denote by $\Omega(\pi(A))$ (resp. $\Omega(\pi(B))$) the set of the 2^{n_A-1} (resp. 2^{n_B-1}) partitions of A (resp. B) composed of the unions of elements of $\pi(A)$ (resp. $\pi(B)$) associated with adjacent intervals in A (resp. in B). For instance, if $\pi(A) = \{A_1, A_2, A_3, A_4\}$ s.t. $A = A_1 \cup A_2 \cup A_3 \cup A_4$ then

$$\Omega(\pi(A)) = \begin{matrix} \{\{\{A_1\}, \{A_2\}, \{A_3\}, \{A_4\}\}, \{\{A_1 A_2\}, \{A_3\}, \{A_4\}\}, \ldots, \\ \{A_1 A_2 A_3 A_4\}\} \end{matrix}$$

For each pair $(P_i, P_j) \in \Omega(\pi(A)) \times \Omega(\pi(B))$ (resp. $(P_j, P_i) \in \Omega(\pi(B)) \times \Omega(\pi(A))$) we compute the geometric mean of the implication intensities between each sub-interval of P_i (resp. P_j) and each sub-interval of P_j (resp. P_j). Let us denote by $\max_{AB}$ and $\max_{BA}$ the respective maximal values between $\Omega(\pi(A))$ and $\Omega(\pi(B))$ and between $\Omega(\pi(B))$ and $\Omega(\pi(A))$. The implication is optimal if there is a partitioning of A which corresponds to $\max_{AB}$ and a partitioning of interval of B which corresponds to $\max_{BA}$.

3.3 Interval variables

The basic situation

Let us consider a score distribution of a class for different subjects. The considered implication is "the sub-interval $[2; 5.5]$ in mathematics generally implies

the sub-interval $[4.25; 7.5]$ in physics". These two sub-intervals belong to an "optimal" partition -according to the inertia- of the definition domains $[1; 18]$ and $[3; 20]$ of the scores in mathematics and in physics.

Main steps of the heuristic

The previous approach can be adapted to the interval variables, which are symbolic data. Let us consider two variables a and b which are associated with a series of intervals due to the measure imprecision: I_i^a (resp. I_i^b) is the interval of a (resp. b) for the individual $i \in E$. Let I^a (resp. I^b) be the interval which contains all the a (resp. b) values. We can define on I^a and I^b a partition which optimizes a given criterium. The intersections between I_i^a and I^a and between I_i^b and I^b follow a distribution that takes into account the common parts. Consequently, the problem is similar to the computation of the rules between the on-interval variables (we refer to [16] for details).

3.4 The entropic version of the implication intensity

The limits of the basic implication intensity for large datasets

Pertinent results have been obtained with the implicative intensity φ for various applications where the data corpuses are relatively small ($n < 300$). However, in data mining, numerical experiments have highlighted two limits of φ for large datasets. First, it tends to be not discriminant enough when the size of E dramatically increases (e.g. [8]); its values are close to 1 even though the inclusion $A \subset B$ is far from being perfect. Second, like numerous measures proposed in the literature, it does not take into account the contrapositive $\bar{b} \Rightarrow \bar{a}$ which could allow to reinforce the affirmation of the good quality of the implicative relationship between a and b, and the capacity to estimate the causality between the variables.

The entropic implication intensity

To overcome these difficulties, we have proposed to modulate the value of the surprise quantified by the implication intensity by taking into account both the imbalance between $card\,(A \cap B)$ and $card\,(A \cap \overline{B})$ associated with $a \Rightarrow b$ and the imbalance between $card\,(A \cap \overline{B})$ and $card\,(\overline{A} \cap \overline{B})$ associated with the contrapositive $\bar{b} \Rightarrow \bar{a}$ [6, 7, 18]. We have introduced a new measure, called the entropic implication intensity, based on the Shannon's entropy to non-linearly quantify these differences.

More precisely, let us first consider a weighted version of the implication intensity $\phi\,(a, b) = (\varphi\,(a, b)\,.\tau\,(a, b))^{1/2}$ where $\tau\,(a, b)$ measures the imbalance between $n_{a \wedge b}$ and $n_{a \wedge \bar{b}}$ and the imbalance between $n_{a \wedge \bar{b}}$ and $n_{\bar{a} \wedge \bar{b}}$. Intuitively, the surprise measured by ϕ must be softened (resp. confirmed) when the

number of counter-examples $n_{a \wedge \bar{b}}$ is high (resp. small) for the rule and its contrapositive considering the observed numbers n_a and $n_{\bar{b}}$.

A well-known index for taking the imbalances into account non-linearly is the Shannon's conditional entropy. The conditional entropy $H_{b/a}$ of cases (a and b) and (a and $\bar{b}$) given a is defined by

$$H_{b/a} = -\frac{n_{a \wedge b}}{n_a} \log_2 \frac{n_{a \wedge b}}{n_a} - \frac{n_{a \wedge \bar{b}}}{n_a} \log_2 \frac{n_{a \wedge \bar{b}}}{n_a}$$

and, similarly, the conditional entropy $H_{\bar{b}/\bar{a}}$ of cases ($\bar{a}$ and $\bar{b}$) and (a and $\bar{b}$) given $\bar{b}$ is defined by

$$H_{\bar{a}/\bar{b}} = -\frac{n_{a \wedge \bar{b}}}{n_{\bar{b}}} \log_2 \frac{n_{a \wedge \bar{b}}}{n_{\bar{b}}} - \frac{n_{\bar{a} \wedge \bar{b}}}{n_{\bar{b}}} \log_2 \frac{n_{\bar{a} \wedge \bar{b}}}{n_{\bar{b}}}$$

We can here consider that these entropies measure the average uncertainty on the random experiments in which we check whether b (resp. $\bar{a}$) is realized when a (resp. $\bar{b}$) is observed. The complements of 1 for these uncertainties $I_{b/a} = 1 - H_{b/a}$ and $I_{\bar{b}/\bar{a}} = 1 - H_{\bar{b}/\bar{a}}$ can be interpreted as the average information collected by the realization of these experiments; the higher this information is, the stronger the guarantee of the quality of the implication and its contrapositive will be.

Intuitively, the expected behavior of the measure ϕ is determined by three phases:

1. a slow reaction to the first counter-examples (robustness to noise).
2. an acceleration of the reject in the neighborhood of the balance.
3. an increasing rejection beyond the balance -which was not guaranteed by the basic implication intensity φ.

Hence, in order to have the expected significance, our model must satisfy the following constraints:

1. Integrating both the information relative to $a \rightarrow b$ and that relative to $\bar{b} \rightarrow \bar{a}$ respectively measured by $I_{b/a}$ and $I_{\bar{a}/\bar{b}}$. A product $I_{b/a}.I_{\bar{a}/\bar{b}}$ is well-adapted to simultaneously highlight the quality of these two values.
2. Raising the conditional entropies to the power for a fixed number $\alpha > 1$ in the information definitions to reinforce the contrast between the different phases described below: $\left(\left(1 - H_{b/a}^{\alpha}\right) \cdot \left(1 - H_{\bar{a}/\bar{b}}^{\alpha}\right) \right)^{1/\beta}$ with $\beta = 2\alpha$ to remain of the same dimension as φ.
3. The need to consider that the implications have lost their inclusive meaning when the number of counter-examples is greater than half of the observation of a and $\bar{b}$. Beyond these values we consider that the terms $\left(1 - H_{b/a}^{\alpha}\right)$ and $\left(1 - H_{\bar{b}/\bar{a}}^{\alpha}\right)$ are equal to 0.

Let $f_a = \frac{n_a}{n}$ (resp. $f_{\bar{b}} = \frac{n_{\bar{b}}}{n}$) be the frequency of a (resp. $\bar{b}$) on E and $f_{a \wedge \bar{b}}$ be the frequency of the counter-examples. The proposed adjustment of the previous informations $I_{b/a}$ and $I_{\bar{a}/\bar{b}}$ can be defined by-

$$\widehat{I^\alpha}_{b/a} = 1 - H^\alpha_{b/a} = 1 + \left(\left(1 - \frac{f_{a \wedge \bar{b}}}{f_a} \right) \log_2 \left(1 - \frac{f_{a \wedge \bar{b}}}{f_a} \right) + \frac{f_{a \wedge \bar{b}}}{f_a} \log_2 \left(\frac{f_{a \wedge \bar{b}}}{f_a} \right) \right)^\alpha$$

if $f_{a \wedge \bar{b}} \in \left[0, \frac{f_a}{2} \right[$; otherwise, $\widehat{I^\alpha}_{\bar{a}/\bar{b}} = 0$
and

$$\widehat{I^\alpha}_{\bar{a}/\bar{b}} = 1 - H^\alpha_{\bar{b}/\bar{a}} = 1 + \left(\left(1 - \frac{f_{a \wedge \bar{b}}}{f_{\bar{b}}} \right) \log_2 \left(1 - \frac{f_{a \wedge \bar{b}}}{f_{\bar{b}}} \right) + \frac{f_{a \wedge \bar{b}}}{f_{\bar{b}}} \log_2 \left(\frac{f_{a \wedge \bar{b}}}{f_{\bar{b}}} \right) \right)^\alpha$$

if $f_{a \wedge \bar{b}} \in \left[0, \frac{f_a}{2} \right[$; otherwise, $\widehat{I^\alpha}_{\bar{a}/\bar{b}} = 0$.

Definition 5. *The imbalances are measured by $\tau(a,b)$ —called the inclusion index— defined by*

$$\tau(a,b) = \left(\widehat{I^\alpha}_{b/a} . \widehat{I^\alpha}_{\bar{a}/\bar{b}} \right)^{1/2\alpha}$$

and, the weighted version of the implication intensity —called the entropic implication intensity— is given by

$$\phi(a,b) = (\varphi(a,b) . \tau(a,b))^{1/2}$$

Example 3.

	b	$\bar{b}$	$\sum$
a	200	400	600
$\bar{a}$	600	2800	3400
$\sum$	800	3200	4000

	b	$\bar{b}$	$\sum$
a	400	200	600
$\bar{a}$	1000	2400	3400
$\sum$	1400	2600	4000

	b	$\bar{b}$	$\sum$
a	40	20	60
$\bar{a}$	60	280	340
$\sum$	100	300	400

Table 1. Distribution examples (a, b and c).

For the table 1.a, the implicative intensity is $\varphi(a,b) = 0.9999$. The entropic functions are $H_{a/\bar{b}} = 0 = H_{\bar{b}/\bar{a}}$. The weighting coefficient is $\tau(a,b) = 0$. And,

$\phi(a, b) = 0$ whereas the confidence $c(a, b)$ is equal to 0.333. The entropic functions moderate the implication intensity when the inclusion is bad.

For the table 1.b, the implication intensity $\varphi(a, b) = 1$. The entropic functions are $H_{a/\overline{b}} = 0.918$ and $H_{\overline{b}/\overline{a}} = 0.391$. The weighting coefficient is $\tau(a, b) = 0.6035$. And, $\phi(a, b) = 0.777$ and the confidence $c(a, b) = 0.666$.

The table 1.c proves that the correspondance between φ and ϕ is not monotonous. The intensity implication is lower for the table 1.c than for the table 1.b. And, it is the contrary for $\phi(a, b)$. Let us remark that the confidence is the same for the two tables.

4 The implicative graph

When computing the implication intensities between all pairs of variables of V, we obtain a square matrix M of numbers in $[0, 1]$. The global structure of the relationships between the variables does not clearly appear. To highlight this structure we have associated a directed graph with M, called the *implicative graph* [2, 11].

Let Φ_α be the relationship defined on $V \times V$ by φ for a given threshold $\alpha \in [0, 1]$: $a\Phi_\alpha b$ if and only if $\varphi(a, b) \geq \alpha$. The threshold α, which controls the implicative quality of the rules, is chosen by the user. The relationship Φ_α is reflexive, not symmetric and not transitive. However, it is interesting to consider the partial order relationships between the subsets of V. Consequently, we extend the relationship Φ_α: if $a\Phi_\alpha b$ and $b\Phi_\alpha c$ then we accept the transitive closure $a\Phi_\alpha c$ if and only if $\varphi(a, c) \geq 0.5$ i.e. when the implicative trend of a on c is better than the neutrality.

Hence, for a given threshold α, the graph $G_{M,\alpha}$ is defined as follows: its vertices are the variables of V, and there is an arc between a pair of variables (a, b) if and only if $a\Phi_\alpha b$.

Different options of the software CHIC allows to easily interact with the drawing of the graph.

5 From rules to R-rules

5.1 The basic situation

In the didactics of mathematics, one of the fundamental question is to identify the source of the problems -both didactical and epistemological- the pupil is faced with during his learning processes. These obstacles are based on the conceptions the pupil is building up. These conceptions are structured by simple or complex rules which together allow to elaborate the basis of a cognitive model.

This structuration is neither a simple union of rules nor a classical hierarchical structure where the variable classes are fit into partitions which

are partially ordered by the relation "thiner than" which reflects the similarity between the class elements. To complete the information provided by the previous models, we have proposed the concept of R-rules (rules of rules) which are an extension of the quasi-implications: their premisses and their conclusions can be rules themselves [15, 17, 20, 23]. To guide the intuition a parallel can be drawn from the proof theory with the logical implication: $(X \Rightarrow Y) \Rightarrow (Z \Rightarrow W)$ describes an implication between the two theorems $X \Rightarrow Y$ and $Z \Rightarrow W$ previously established.

5.2 The R-rules and their interpretation

In the following, we consider binary variables. The R-rules are an extension of the classical binary rules $a \rightarrow b$ to rules of rules $R' \rightarrow R''$, which may be complex themselves. For instance $a \rightarrow (b \rightarrow c)$ is a R-rule between a variable a and a rule $(b \rightarrow c)$, and $(a \rightarrow b) \rightarrow (c \rightarrow d)$ is a R-rule between two rules $(a \rightarrow b)$ and $(c \rightarrow d)$. To indicate the complexity of the implication composition, we associate a complexity degree with each R-rule.

Definition 6. *The R-rules of degree 0 are variables of V. The R-rules of degree 1 are the simple quasi-implications of the form $a \rightarrow b$. A R-rule of degree i, $1 < i \leq p$, is a rule $R' \rightarrow R''$ between two R-rules R' and R'' whose respective degrees j and k satisfy $j + k = i - 1$.*

For instance, $a \rightarrow b$ is a R-rule of degree 1, $a \rightarrow (b \rightarrow c)$ a R-rule of degree 2 and $(a \rightarrow b) \rightarrow (c \rightarrow d)$ a R-rule of degree 3. When there is no ambiguity we denote by R a R-rule of degree greater or equal than 1.

The R-rules allow to express different levels of abstraction: (1) situation or object descriptions (conjunction of R-rules of degree 0), (2) implications between variables (R-rules of degree 1), and (3) implications between implications (some R-rules of degree greater than 1). Consequently, their interpretation may vary according to three typical cases:

1. when $R \rightarrow a$ then a may be interpreted as a quasi-consequence of R;
2. the R-rule $a \rightarrow R$ means that a R-rule R may be partially deduced from the observation of a. Moreover, although we here consider quasi-implications only, the intuition can be supported by Heyting algebra where an implication $a \Rightarrow (b \Rightarrow c)$ is equivalent to $(a \; AND \; b) \Rightarrow c$;
3. the R-rule $R' \rightarrow R''$ means that the property R'' is the quasi-corollary of a previous property R'

5.3 A measure of cohesion of the R-rules

The objective is to discover R-rules with a good implicative quality —called *cohesion* in the following— i.e. R-rules $R' \rightarrow R''$ with a strong implicative

relationship between the components of R' and those of R''. For instance, it seems natural to form a R-rule $(a \to b) \to (c \to d)$ if the implicative relationships $a \to c$, $a \to d$, $b \to c$ and $b \to d$ are significant enough. Intuitively, this means that they must contrast with the disorder of a random experience.

The entropy is well-suited to measure this disorder. Let us first consider a R-rule $a \to b$ of degree 1. And, let Y be the random indicator variable of the event $\left(Q\left(a, \overline{b}\right) \geq q\left(a, \overline{b}\right) \right)$. The distribution of Y is defined by $\Pr(Y = 1) = \varphi(a, b)$ and $\Pr(Y = 0) = 1 - \varphi(a, b)$. The entropy of this experience is $-p \log_2 p - (1 - p) \log_2 (1 - p)$ where $p = \varphi(a, b)$.

The extreme values are 0 if $\varphi(a, b) = 0$ (by setting $0 \log_2 0 = 0$) and 1 if $\varphi(a, b) = 0.5$. This last value is reached when $n_{a \wedge \overline{b}} = n_a n_{\overline{b}} / n$ i.e. when $n_{a \wedge \overline{b}}$ is equal to the expected mean. In this case, when $\varphi(a, b) < 0.5$, the meaning of the implication is lost and it seems natural to set te cohesion equal to 0.

Definition 7. *The cohesion $c(a, b)$ of a R-rule $a \to b$ of degree 1 is defined by*

$$c(a, b) = \left(1 - \left(-p \log_2 p - (1 - p) \log_2 (1 - p) \right)^2 \right)^{1/2}$$

if $p = \varphi(a, b) > 0.5$ and $c(a, b) = 0$ otherwise.

We square the entropy to reinforce the contrast between values in $[0, 1]$ and the square root to the complement to 1 allows to measure the cohesion on a same scale as the entropy.

The generalization of this definition to R-rules of higher degree is guided by the following requirement: the cohesion of $R' \to R''$ must take into account both the cohesion of R' and R'' as well as the implicative relationships between the attributes of R' and those of R''. Let $\prec_R$ be the left right reading order on the variables which composed a R-rule. For instance, for $(a \to b) \to (c \to d)$ the order on $\{a, b, c, d\}$ is defined by $a \prec_R b \prec_R c \prec_R d$.

Then, a simple way to satisfy the previous requirements is to take the mean of the cohesions of R' and R'' and of the cohesions if each ordered pairs composed of one attribute of R' and one attribute of R'' in accordance with the permutation orders. Here we favour the geometric mean as it is equal to 0 as soon as the cohesion of one ordered pair is equal to 0 (i.e. when an implication is low or without surprise) and it is close to 1 when the cohesions of all ordered pairs are high.

Definition 8. *Let R be a R-rule of the form $R' \to R''$ where R' and R'' are respectively associated with the orders $a'_1 \prec_{R'} a'_2 \prec_{R'} \ldots a'_k$ and $a''_1 \prec_{R''} a''_2 \prec_{R''} \ldots a''_h$. The cohesion of R is defined by*

$$c(R) = \left(\prod_{i=1, k-1; j=2, k} c\left(a'_i, a'_j\right) \cdot \prod_{i=1, h-1; j=2, h} c\left(a_i, a''_j\right) \cdot \prod_{i=1, k; j=1, h} c\left(a'_i, a''_j\right) \right)^{2/r(r-1)}$$

where $r = k + h$.

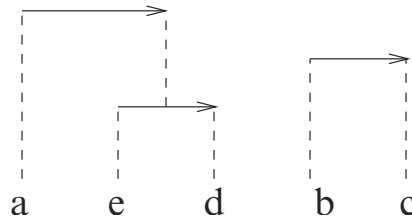# 6 The implicative hierarchy

6.1 The basic situation

Generally speaking, R-rules contribute to increasing the analysis richness. We do not solely extract facts or isolated behaviors, but more general conducts, revealing more global, less singular phenomena i.e. in didactics profound psychological representations. The different complexity degrees of the R-rules can be associated with a hierarchical structure which reflects the genesis of the "operating knowledge" developed by Piaget [28]. We go from one level to another by a process of reflecting abstraction: from object representation to representation of operations on the objects, then to representation of operations on the operations. This process involves a dynamical hierarchical point of view in contrast with the static point of view associated with a taxonomy. Hence, the individual description of the R-rules by aggregating simple rules is not sufficient. It is necessary to develop a global structure which reflects the emerging properties of the whole. Consequently, we have developed the concept of "implicative hierarchy" to structure the significant R-rules.

Let us introduce this notion by an example. A graphical representation of an implicative hierarchy on the variable set $V = \{a, b, c, d, e\}$ is given on figure 1. The elements of the implicative hierarchy $\overrightarrow{H}_V$ are R-rules:

$$\overrightarrow{H_V} = \{a, b, c, d, e, b \to c, e \to d, a \to (e \to d)\}$$

Fig. 1. Graphical representation of the implicative hierarchy $\overrightarrow{H}_V = \{a, b, c, d, e, b \to c, e \to d, a \to (e \to d)\}$

Note that contrary to hierarchies in classical hierarchical classification (HC) the tree associated with the implicative hierarchy is not necessarily connected. Intuitively, this means that it contains only significant R-rules according to the cohesion measure.

6.2 Definitions

The R-rules which composed an implicative hierarchy can be associated with k-permutations —called *classes* by analogy with the HC— that satisfy special

interlocking conditions. For instance, in the example given below, the R-rule $a \rightarrow (e \rightarrow d)$ is associated with the permutation aed. And, this is the only possible association as the R-rule $(a \rightarrow e)$ associated with the permutation ae is not in $\overrightarrow{H}_V$. The class set H_V associated with $\overrightarrow{H}_V$ is

$$H_V = \{a, b, c, d, e, bc, ed, aed\}$$

The R-rules are deduced by a recursive decomposition of the non elementary classes of H_V. The class aed is the unique amalgamation of $a \in H_V$ and $ed \in H_V$. Since the class ed is associated with $e \rightarrow d$, the class aed is associated with $a \rightarrow (e \rightarrow d)$.

More formally, let Ω_V be the set of all k-permutations on the variable set V, $k = 1, p$. The elements C of Ω_V are strings with distinct characters. Let $\prec$ be the left-right reading order on the variables of a permutation of Ω_V as we defined it previously. In order to compare and combine the elements of Ω_V to form an implicative hierarchy, we define three operators on Ω_V, whose appelations are inspired by the set theory:

- *Intersection.* The intersection $C' \widehat{\cap} C''$ of two strings of Ω_V is the largest sub-string of contigous variables common to C' and C''. In case of equality we keep the first sub-string of C' according to $\prec$. If $C' = acdb$ and $C'' = cdab$ then $C' \widehat{\cap} C'' = cd$, and if $C' = abcd$ and $C'' = cdab$ then $C' \widehat{\cap} C'' = ab$.
- *Union.* The union $C' \widehat{\cup} C''$ of two distinct strings C' and C'' s.t. $C' \widehat{\cap} C'' = \emptyset$ is the concatenation of C' and C'' with C' first according to $\prec$. If $C' = aceb$ and $C'' = fgh$ than $C' \widehat{\cup} C'' = acebfgh$.
- *Difference.* For three strings C, C' and C'' of Ω_V s.t. $C = C' \widehat{\cup} C''$, the difference $C \widehat{-} C'$ between C and C' is C'' and the difference $C \widehat{-} C''$ between C and C'' is C'. If $C = abc$, $C' = ab$ and $C'' = c$ then $C \widehat{-} C' = c$ and $C \widehat{-} C'' = ab$.

Definition 9. *An implicative hierarchy H_V is a subset of permutations of Ω_V satisfying the three following requirements:*

1. *H_V contains the variables of V, called elementary classes*
2. *for each pair C', $C'' \in H_V$, $C' \widehat{\cap} C'' = \{\emptyset, C', C''\}$*
3. *for each non elementary class $C \in H_V$, there is a single pair C', $C'' \in H_V$ s.t. $C = C' \widehat{\cup} C''$*

From the condition 2, a hierarchy is a partially ordered set with the inclusion relation $\widehat{\subset}$ defined on Ω_V by: $C' \widehat{\subset} C''$ if and only if $C' \widehat{\cap} C'' = C'$. The condition 3 is required to recover all the classes of the hierarchy.

The isolated interpretation of a class of the hierarchy is tricky since it is a k-permutation which does not state the implication composition. For instance, if we analyse the class $aed \in H_V$ all alone, we do not know the exact meaning of aed: it could be either $a \rightarrow (e \rightarrow d)$ or $(a \rightarrow e) \rightarrow d$. However, the

whole H_V class set allows to dispel ambiguity: $a \to (e \to d)$ is chosen as ed is a class of H_V.

Proposition 4. *[15] Each non elementary class C of an implicative hierarchy H_V can be associated with a unique R-rule.*

The R-rule set $\overrightarrow{H}_V$ associated with H_V can be graphically represented by a valuated binary directed tree:

- each of the elementary classes are located at a terminal node;
- each of the internal node is represented by an arrow which describes the R-rule subtended by the associated class;
- the height $h(C) \in R_+$ of each node C satisfies the following condition: for each node $C' \in H_V$ s.t. $C' \widehat{\subset} C$ then $h(C) > h(C')$.

6.3 Construction of an implicative hierarchy

The significant R-rules which form an implicative hierarchy are calculated by a incremental algorithm similar to the basic process of the classical HC. The amalgamation criterium is here the maximization of the cohesion.

At each level h_i of H_V, a new R-rule is built. It results from the amalgamation of two R-rules built at a previous level h_j, $0 < j < i$. More precisely,

- the initial level h_0 of H_V is composed of the variable set V;
- at h_1, two variables of V with the maximal cohesion are "grouped" together to form a R-rule of degree 1;
- at h_2, the R-rule is composed either of two variables not yet aggregated, called separate variables, or of the R-rule of degree 1 built at h_1 and a separate variable. The selected R-rule is the one with the maximal cohesion;
- at h_3, the R-rule may be of three types: a R-rule of degree 1 composed of two separate variables, a R-rule of degree 2 composed of a R-rule of degree 1 built at h_1 or h_2 and a separate variable, or a R-rule of degree 3 composed of the two R-rules of degree 1 built at h_1 and h_2.
- and so on. The process stops as soon each cohesion of the new potential R-rules is null.

For instance, for the implicative hierarchy of the figure 1, the process stops at h_3 if the cohesion is null for the R-rules $(a \to (d \to e)) \to (b \to c)$ and $(b \to c) \to (a \to (e \to d))$.

We refer to [15] for an algorithmic description of this algorithm and the analysis of its complexity.

The directed hierarchy H_V can be associated with a valuation which satisfies the ultrametric inequality.

Proposition 5. *[15] For any class C of H_V, let us define the height h of C by $h(C) = 1 - c(C)$ if C is non elementary and $h(C) = 0$ otherwise, where $c(C)$ is the cohesion of the R-rule associated with C. Let u be a dissimilarity on $V \times V$ defined by*

- *$u(a, b) = 1$ if a and b are not amalgameted in H_V,*
- *$u(a, b) = h(C_{ab})$ otherwise*

where C_{ab} is the smallest class of H_V which contains both a and b. The dissimilarity u is symmetric, positive and satisfies the ultrametric inequality:

$$u(a, b) \leq Max\{u(a, c), u(b, c)\}$$

for any $a, b, c \in V$.

From the Benzécri-Johnson theorem [4, 22] this property a posteriory justifies our choice of the word "hierarchy".

7 The significative levels of the implicative hierarchy

7.1 The basic situation

Due to the multiplicity of the levels in the implicative hierarchy, it is necessary to highlight those which are the more relevant for the structuration process. In psycho-didactical or sociological applications, these levels seem to correspond to consistent and stable conceptions. Hence, they contribute to a finest interpretation of the set of the computed R-rules.

We have investigated two different approaches for this problem. The first one is based on a rank analysis used in HC by Lerman [26]: it compares the quality of the partitions obtained at each level of the hierarchy. The second one is more local [19]: it focusses on the quality of the R-rules built at each level. In the following, we present the first approach which is the only one to be implemented in the CHIC software [9].

7.2 A criterium to determine the significative levels

Let us note that the cohesion coefficient defined in the section 5.3 can be associated with a pre-ordering $\preceq_c$ on $P = V \times V - \{(a, a), (b, b), \ldots\}$:

$$(a, b) \preceq_c (c, d) \Leftrightarrow c(a, b) \leq c(c, d)$$

The idea consists in determining the levels of H_V which "better express" this pre-ordering. At each level h_k, two sets of variable pairs can be distinguished: the set A_k of the amalgameted variable pairs at h_k, and the set S_k of the separate variable pairs (not yet amalgameted to form a R-rule of degree ≥ 1). By construction, $A_k \cup S_k = P$.

Let $G_{\preceq_c}$ be the graph of $\preceq_c$. The set $G_{\preceq_c} \cap (A_k \times S_k)$ is composed of pairs of pairs which respect $\preceq_c$ at the level k. For instance, let us consider the variable set $V = \{a, b, e, f\}$ such that $c(a, b) < c(e, f)$. Let us suppose that at the level h_k the variables e, f and k are separate whereas the variables a and b are amalgameted in a class. Then, the pair $((e, f), (a, b)) \in G_{\preceq_c} \cap (A_k \times S_k)$.

The objective is now to measure the adequation between $G_{\preceq_c}$ and $A_k \times S_k$. Let us denote by Θ the set of all the pre-orderings on $\bar{P} = V \times V - \{(a, a), (b, b), \ldots\}$ with the same cardinality as $\preceq_c$. We consider the random preordering G^* on Θ -with a uniform distribution-. From the theorem of Wald and Wolfowitz [31] we can deduce that the theoretical mean of $G^* \cap (A_k \times S_k)$ is $\mu = 1/2 \, \mathrm{card}\,(A_k \times S_k)$ and its standard deviation is

$$\sigma = \frac{1}{12} \left(\mathrm{card}\,(A_k \times S_k)\,(\mathrm{card} G^* + 1) \right)$$

The adequation between $G_{\preceq_c}$ and $A_k \times S_k$ at the level h_k is measured by

$$s(\preceq_c, k) = \frac{\mathrm{card}\,(G_{\preceq_c} \cap (A_k \times S_k)) - \mu}{\sigma}$$

Definition 10. *A level h_k of the implicative hierarchy H_V is significative if it is a local maximum of $s(\preceq_c k)$: $s(\preceq_c, k - 1) < s(\preceq_c, k) < s(\preceq_c, k + 1)$*

If $G_{\preceq_c} \cap (A_k \times S_k) = A_k \times S_k$ then the partition $A_k \times S_k$ on $V \times V$ associated with the structuration at h_k is in total accordance with the pre-ordering induced by the cohesion.

8 Typicality and contributions

8.1 The basic situation

Like in factorial analysis, we introduce the notion of "additional variable": it does not contribute to the computation of the relationships involved in the implicative hierarchy, but it brings an additional information for its interpretation(e.g. age, sex, social-professional category).

Our objective is to identify individuals, or individual groups, and additional variables which contribute to class forming at each level of the implicative hierarchy.

8.2 A representation space

Let C be the class built at the level h_k of the hierarchy H_V. This class results from the amalgamation of two classes $C' \in H_V$ and $C'' \in H_v$ not amalgameted at the previous level h_{k-1}.

The variable pair (a, b) is a *generic pair* at h_k if $\varphi(a, b) \geq \varphi(i, j)$ for any $i \in C'$ and $j \in C''$. The *generic intensity* at h_k is denoted by $\varphi_k = \varphi(a, b)$. This pair characterizes the most noticeable implicative effect for a given class.

Moreover, the classes C' and C'' are themselves the results of an amalgamation at a lower level. Hence, at each level h_g, $g \leq k$, of H_V, we can determine a generic pair: the resulting vector $(\varphi_1, \varphi_2, \ldots, \varphi_k) \in [0, 1]^k$ is called the *implicative vector* of the class C built at h_k.

A similar representation can be used for evaluating the impact of an individual on the formation of a path on the implicative graph $G_{M,\alpha}$. Let us consider a path P of length k on $G_{M,\alpha}$ with a transitive closure (i.e. each arc is associated with a rule with an implication intensity greater than 0.5). Then, P contains $k(k-1)/2$ transitive arcs. A pair (a, b) of P is generic if $\varphi(a, b) \geq \varphi(i, j)$ for any $i, j \in P$.

The vectors $(\varphi_1, \varphi_2, \ldots, \varphi_k) \in [0, 1]^k$ form a representation space where the individuals can be projected. In the following, we precise the properties of this space for an implicative hierarchy. They could be similarly defined for an implicative graph.

8.3 Implicative power of an individual on a class

In this subsection, we define a dissimilarity on $E \times H_V$ to measure the "proximity" between an individual $i \in E$ and a class $C \in H_V$.

We first check if the individual i is in accordance with the implication of the generic pair (a, b) of C at the level h_k. Let us denote by $a(i)$ (resp. $b(i)$) the binary variable which characterizes the presence/absence of a (resp. b) for i. The contribution of i to the pair (a, b) is defined by

- $\varphi_{i,k} = 1$ if $a(i) = 1$ or 0 and $b(i) = 1$
- $\varphi_{i,k} = 0$ if $a(i) = 1$ and $b(i) = 0$
- $\varphi_{i,k} = p \in]0, 1[$ if $a(i) = b(i) = 0$

In practice, p is set to the neutral value 0.5.

Any individual i is associated with a k-dimensional vector $(\varphi_{i,1}, \varphi_{i,2}, \ldots, \varphi_{i,k})$ which characterizes its contribution to the k generic pairs of the class C buit at h_k. An individual whose components are equal to the implicative vector $(\varphi_1, \varphi_2, \ldots, \varphi_k)$ is called the *optimal typical individual*.

We measure the typicality of i in C by the χ^2 distance between the distributions $(1 - \varphi_g)$ and $(1 - \varphi_{i,g})$, for $g = 1, k$. In contrast with the usual Euclidean distance, it allows to compare $\varphi_g - \varphi_{i,g}$ to φ_g and to normalize the distance effect for large φ_g.

Definition 11. *The implicative distance $d_2(i, C)$ between an individual $i \in E$ and a class $C \in H_V$ built at the level h_k is defined by*

$$d_1\left(i,C\right) = \left(\frac{1}{k}\sum_{g=1}^{k}\frac{\left(\varphi_g - \varphi_{i,g}\right)^2}{1 - \varphi_g}\right)^{1/2}$$

If it exists g s.t. $\varphi_g = 0$ we set $\left(\varphi_g - \varphi_{i,g}\right)/\left(1 - \varphi_g\right) = 0$. In this case, the generic implication is maximal and thus it exists an excellent implicative relationship for all the individuals $i \in E$ $\left(\varphi_{i,g} = 1\right)$.

Remark 2. Let us consider a class $C \in H_V$ at the level h_k. We can define a metric space structure on E with

$$d_C\left(i,j\right) = \left(\frac{1}{k}\sum_{g=1}^{k}\frac{\left(\varphi_{i,g} - \varphi_{j,g}\right)^2}{1 - \varphi_g}\right)^{1/2}$$

for any $\left(i,j\right) \in E^2$.

The distance $d_C\left(i,j\right)$ measures the behavior difference between i and j considering C. It defines a discrete topological C-structure on E. Let us consider the vectors $\left(\varphi_{i,1}, \varphi_{i,2}, \ldots, \varphi_{i,k}\right)$ and the norm $\left\|\overrightarrow{i} - \overrightarrow{j}\right\| = d_C\left(i,j\right)$. This topology is equivalent to the previous one (similarly to the duality in correspondence analysis). The elements of the diagonal matrix of the symmetrical operator associated with the quadratic form which defines d_C are $\left(k\left(1 - \varphi_i\right)\right)^{-1}$ for $i = 1, k$. Let us remark that the semantic of the vector sum is not precised in the SIA. Nevertheless, it could be interesting to characterize the individuals which belong to a ball of a given diameter with a given center (e.g. the optimal individual).

8.4 Individual and group typicalities

Definition 12. *The typicality $\gamma\left(i,C\right)$ of an individual $i \in E$ for a class $C \in H_V$ is defined by the ratio between the distance $d_1\left(i,C\right)$ and the maximal value of the distance on the individual set:*

$$\gamma\left(i,C\right) = 1 - \frac{d_1\left(i,C\right)}{Max_{j \in E}d_1\left(j,C\right)}$$

The maximal distance $d_1\left(j,C\right)$ is reached by the individuals with null or very low $\varphi_{i,k}$. They are contrasting with the generic rules. And, the typicality of i is large when i is different from these individuals.

A straightforward extension of the previous definition allows to define the typicality $\gamma\left(G,C\right)$ of a individual group $G \subset E$:

$$\gamma\left(G,C\right) = \frac{1}{card\left(G\right)}\sum_{i \in G}\gamma\left(i,C\right)$$

In practice, an operational tool is required to evaluate the statistical significancy of a group typicality. The basic idea consists in partitioning E in two opposite groups E_1 and E_2 with regards to their typicalities $\gamma(E_1, C)$ and $\gamma(E_2, C)$ in C. This dispersion can be measured by the inter-class inertia. The barycenter $\overline{\gamma}$ of the typicalities $\gamma(E_1, C)$ and $\gamma(E_2, C)$ is defined by

$$\overline{\gamma} = \frac{1}{n}\left(card(E_1)\gamma(E_1, C) + card(E_2)\gamma(E_2, C)\right)$$

By construction, $\overline{\gamma}$ is also the barycenter of all the individual typicalities in E. Consequently, the inter-class inertia is

$$V_E = \frac{card(E_1)}{n}\left(\gamma(E_1, C) - \overline{\gamma}\right)^2 + \frac{card(E_2)}{n}\left(\gamma(E_2, C) - \overline{\gamma}\right)^2$$

Definition 13. *An individual group $G_C^* \subset E$ is optimal for a class $C \in H_V$ if its typicality is greater than the typicality of its complementary set in E, and if it constitutes with this later a bi-partitioning which maximizes V_E. This partition is said to be significant.*

It is interesting to detect the group or the additional variable associated with the greatest typicality for the optimal group. We measure the surprisingness of the proportion of concerned individuals. Let $\{E_i\}_i$ be a given partition of E. It can be defined by an additional variable. For each class E_i, we consider the random variable X_i which is a random subset of E of cardinality $card(E_i)$, and the random variable Z_i defined by $Z_i = card(E_i \cap G_C^*)$. The variable Z_i follows a Binomial distribution with parameters $card(E_i)$ and $card(G_C^*)/n$ [21].

Definition 14. *The most typical group of the class C is the subset $E_i \subset E$ which minimizes the probability p_i on the set $\{p_i = \Pr(Z_i > card(X_i \cap G_C^*))\}_i$.*

The probability p_i is an error of the first kind: the risk of making a mistake when considering that the group is not typical.

8.5 Contribution

The contribution is different from the typicality: it measures the individual and additional variable responsabilities for the existence of a rule or a R-rule between the variables of V.

Let us consider two variables $a \in V$ and $b \in V$ linked by a rule $a \to b$ at the first level h_1 of the implicative hierarchy H_v. The contribution of an individual i to (a, b) is defined by $\varphi_{i,1}$. This notion can be extended to the formation of a class C at the level h_k.

Definition 15. *The distance $d_2(i,C)$ between an individual $i \in E$ and a class C at the level h_k of the implicative hierarchy H_V is defined by*

$$d_2(i,C) = \frac{1}{k} \sum_{g=1}^{k} (1 - \varphi_{i,g})^2$$

The contribution $\theta(i,C)$ of $i \in E$ to $C \in H_V$ is defined by $\theta(i,C) = 1 - d_2(i,C)$.

The maximal value of $\theta(i,C)$ is equal to 1; it is reached for an individual i whose components $\varphi_{i,g}$ are all equal to 1.

The concepts defined in the previous sections can be easily adapted to the distance d_2. In practice, the contribution is often easier to interpret than the typicality.

9 Illustration

We illustrate the applicative interest of the different concepts presented below on a data set stemming from a survey of the French Public Education Mathematical Teacher Society on the level in mathematics of pupils in the final year of secondary education and the perception of this subject [8]. In parallel with evaluation tests for students, a set of 311 teachers have been asked on the objectives of the training in mathematics (table 2 presents some items used in the following) and their opinions about commonly shared ideas on this subject (table 3). For each proposition, the teacher could answer "I agree with this idea" (positive opinion), "I disagree" (negative opinion) or "I partially agree".

The figure 3 presents a part of the directed hierarchy obtained on the set composed of the objectives and the different modalities for the opinions (51 items). The interpretation of the whole set of rules is far beyond the scope of this paper. Nevertheless, we have selected some of them, easy to interpret for a non specialist in education theory, to show the use of a directed hierarchy on a real-life corpus. As for the complementarity of this structure with a more classical approach based on the relationship representation by a graph, it is highlighted in figure 2. The vertex set V of this graph contains the same items as those selected for figure, and there is an arc between two vertices a_i and a_j of V if and only if $\varphi(a_i, a_j) \geq 0.5$ and for any $a_k \in V$, $\varphi(a_i, a_k) < 0.5$ and $\varphi(a_j, a_k) < 0.5$ (e.g. [2]). The choice of the threshold comes from the fact that beyond 0.5 the implicative tendency (e.g. $a_i \rightarrow a_j$) is better than neutrality. It is important to note that, due to the non transitivity of the relationship on A induced by φ, the existence of two arcs of the form (a_i, a_j) and (a_j, a_k) does not entail the existence of the arc (a_i, a_k). For instance, in figure, we can not deduce a relationship between the items E and $OP7$.
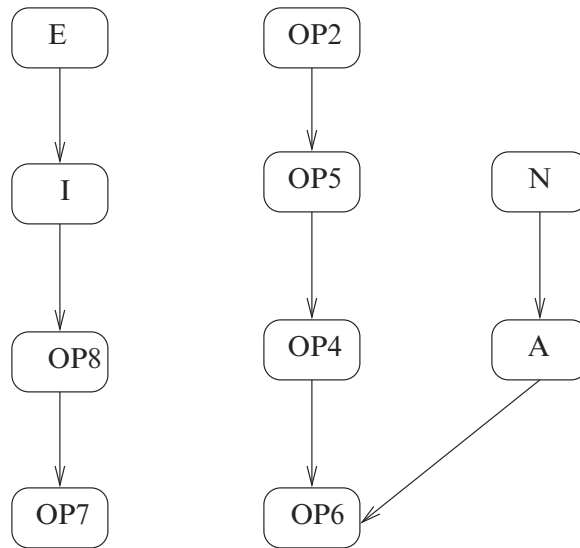
Fig. 2. A part of the implicative graph on the items of the survey on the training in mathematics
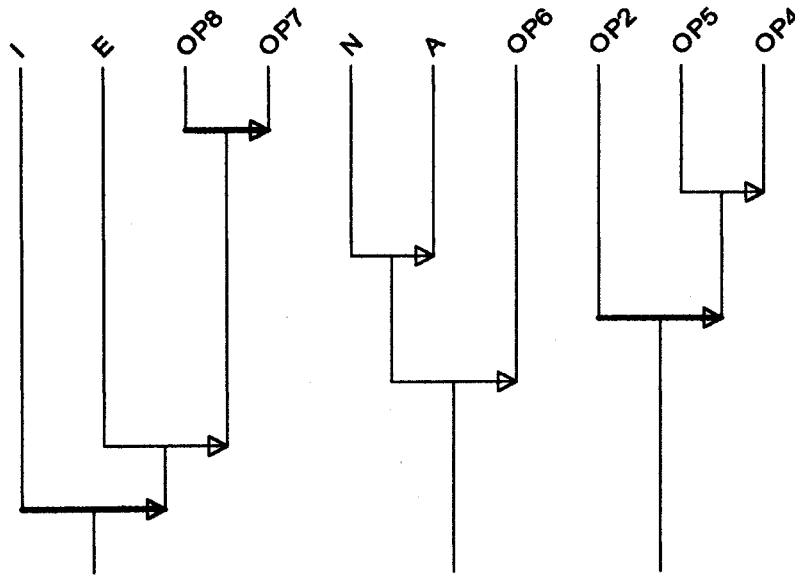
Fig. 3. A part of the directed hierarchy on the items of the survey on the training in mathematics

On the other hand, beside the binary rules, most of the R-rules of the total directed hierarchy involve three or four items. The interpretation of rules with more attributes are generally more difficult to interpret. Nevertheless, they provide more information than the set of the implied binary rules.

The R-rule $(N \to A) \to OP6$ has the following meaning: if know-how acquisition must be accompanied by knowledge acquisition, then the teacher ask for well-defined programs. In this case, focussing on knowledge requires a predefined charter from the institution. The R-rule allows to give a more synthetical interpretation than the binary rules: these are concerned with the behaviour, as seen within the behavioural framework, whereas the R-rule here describes a conduct of a higher order which determines the behaviour. Teachers who consider that the objective C (Preparation to civic and social life) is not relevant are mostly responsible for this R-rule. They have a very restrictive representation of the teaching of maths, focussed on the subject, and their teaching conforms to national standard without any questioning.

The R-rule $(OP2 \to (OP5 \to OP4))$ can be interpreted as follows: if I wish to keep up the complete problem for the A-level exam and if the importance given to the demonstration in maths is subordinated to a fixed scale of grading, then I conform to the national syllabus instructions. This rule corresponds to a class of teachers subjected to the institution and conservative in their educational choices. They consider that, in France, the land of Descartes, the demonstration is the foundation of the mathematical activity and that the complete problem at the exam is the evaluation criterion. For them, the syllabuses and the grading scales defined by the institution are essential to teaching and assessment. We find again a very classical teaching conception based on an explicit and unconditional support to the institution.

Contrary to the previous ones, the R-rule $(I \to (E \to (OP8 \to OP7)))$ can be interpreted as a sign of an openminded didactic conception. Indeed, it means that if a teacher lays the emphasis on the critical mind development and the imagination and creativity, then he considers that a personal training of the pupils in the search of examples and counter-examples is sufficient for discovering divisibility features by themselves. This R-rule reveals a relationship between the non-dogmatic behaviours of the teacher and the wish to place the pupil in a situation of personal research.

A	Knowledge acquisition
B	Preparation to professional life
C	Preparation to civic and social life
D	Preparation to examinations
E	Development of imagination and creativity
I	Development of critical mind
N	Know-how acquisition

Table 2. Some items from the list of the objectives of training in mathematics

OP1	It's true that maths are an element of selection
OP2	For the A-level exam, I prefer a complete problem with different parts rather than independent questions
OP3	In my grading system, I give more importance to the reasoning than to the result
OP3	When I correct, I prefer a very detailed grading system
OP5	The demonstration is the only rigourous way to do maths
OP6	I prefer well-defined programs precising what I must do and not do
OP7	In the last form of secondary education, a pupil should be able to recognize whether a number written in the base 10 is divisible by 4
OP8	In the last form of secondary education, a pupil should be able to give an example or counter-example of the following statement: if two applications f and g are strictly increasing on a given interval, then the product $f \times g$ is also increasing.
OPX	Individual estimation of a size (e.g. width, length)

Table 3. Some items from the list of the commonly shared ideas in the teaching of maths

We study now the additional information brought by the supplementary variable which defines the main option of the cursus: Scientific (S), Economic and Social (ES), Arts (A) and Technology (T). The observed distribution of the variable is: $S = 155$, $ES = 68$, $A = 22$ and $T = 66$.

Let us consider the class $C = (E \rightarrow (OP8 \rightarrow OP7)) \rightarrow OPX$. This rule corresponds to a class of teachers which give importance to imagination and personnal research. The most typical modality for this variable is S (scientific). Indeed, 116 teachers of the option S on 155 are in the optimal group G_C^* of cardinality 201. Let X be a random subset of same cardinality as S (155) and Z be random variable defined by the intersection of X and the optimal group G^*. Then, Z follows a binomial distribution of parameters 155 and $201/311 = 0.656$. The probability for Z to be greater than 116 is the risk 0.00393. The analysis of the series of the risks associated with the different options S, A and T shows that the most typical modality of the class C is S. The pair (S, C) is said to be mutually specific. Similarly, the most typical modality of the rule $B \rightarrow K$ is T; consequently, the pair $(T, (B, K))$ is mutually specific. It confirms that the teachers in technical cursus consider that the mathematics should be useful for the professional life (B) and consequently for the other disciplines.

The computation of the contribution of S to C shows that 111 teachers on 311 participate to the optimal group. The number of teachers of S has decreased (from 116 to 67), and its proportion in the optimal group is significantly lower than for the typicality computation. The teachers of S are the most typical, i.e. in accordance with the general behavior of the population. However, their contribution to the four involved variables is lower than the contribution of the teachers of the other cursus. The risk is equal to 0.0251: it

is more than 6 times greater than the typicality. This remark illustrates the nuances brought by the two concepts: typicality and contribution.

10 Conclusion

In this paper we have proposed an overview of the Statistical Implicative Analysis. Beyond the results, we have related the genesis of the considered problems which arise from questions of experts in different fields. The theoretical basis is quite simple, but the numerous questions on the original assumptions, which do not appear here, have lead to modifications and sometimes to deeper revisions. Fortunately, the proposed answers go beyond the original framework, and SIA is now a data analysis method, based on a non symmetrical approach, which has been shown to be relevant for various applications.

In the next future, we are planning to consider new problems: *(i)* the extension of SIA to vectorial data, *(ii)* and to fuzzy variables, *(iii)* the integration of missing data, *(iv)* the redundant rule reduction. We are also interested in the complementarity of SIA with other approaches, in particular with decision trees (see Ritschard's paper in this book). And, we will obviously carry on exploring real-life data sets and confronting our theoretical tools to experimental analysis to make them evolve.

References

1. R. Agrawal, T. Imielinsky, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD'93*, pages 679–696. AAAI Press, 1993.
2. M. Bailleul. Des réseaux implicatifs pour mettre en évidence des relations. *Mathématiques, Informatique et Sciences Humaines*, 154:31–46, 2001.
3. M. Bailleul and R. Gras. L'implication statistique entre variables modales. *Mathématiques et Sciences Humaines*, 128:41–57, 1995.
4. J.P Benzécri. *L'analyse des données (vol. 1): Taxonomie.* Dunod, Paris, 1973.
5. J.M. Bernard and S. Poitrenaud. L'analyse implicative bayesienne d'un questionnaire binaire : quasi-implications et treillis de galois simplifié. *Mathématiques, Informatique et Sciences Humaines*, 147:25–46, 1999.
6. J. Blanchard, P. Kuntz, F. Guillet, and R. Gras. Mesure de la qualité des règles d'association par l'intensité entropique. *Revue des Nouvelles Technologies de l'Information-Numéro spécial Mesures de qualité pour la fouille de données*, RNTI-E-1:33–44, 2004.
7. J. Blanchard, P. Kuntz, G. Guillet, and R. Gras. Implication intensity: From the basic definition to the entropic version - chapter 28. In *Statistical Data Mining and Knowledge Discovery*, pages 475–493. CRC Press - Chapman et al., 2003.
8. A. Bodin and R. Gras. Analyse du préquestionnaire enseignants. *Bulletin de l'Association des Professeurs de Mathématiques de l'Enseignement PUblic*, 425:772–786, 1999.

9. R. Couturier and R. Gras. C.h.i.c. : Traitement de données avec l'analyse implicative. *Revue des Nouvelles Technologies de l'Information*, RNTI-II:679–684, 2005.

10. L. Fleury. *Extraction de connaissances dans une base de données pour la gestion de ressources humaines*. PhD thesis, Université de Nantes, 1996.

11. R. Gras. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs en didactique des mathématiques*. PhD thesis, Université de Rennes 1, 1979.

12. R. Gras, S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, and A. Totohasina. *L'implication statistique - Nouvelle méthode exploratoire de données*. La Pensee Sauvage editions, France, 1996.

13. R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz, and P. Peter. Quelques critères pour une mesure de qualité de règles d'association. *Revue des Nouvelles Technologies de l'Information*, RNTI-E-1:197–202, 2004.

14. R. Gras, E. Diday, P. Kuntz, and R. Couturier. Variables sur intervalles et variables-intervalles en analyse statistique implicative. In *Proc. of Société Francophone de Classification*, pages 166–173. Université des Antilles-Guyane, 2001.

15. R. Gras and P. Kuntz. Discovering r-rules with a directed hierarchy. *Soft computing*, 1:46–58, 2005.

16. R. Gras, P. Kuntz, and H. Briand. Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et Sciences Humaines*, 154:9–29, 2001.

17. R. Gras, P. Kuntz, and H. Briand. Hiérarchie orientée de régles généralisées en analyse implicative. *Extraction des Connaissances et Apprentissage*, 17-3:145–157, 2003.

18. R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage*, 1-2:69–80, 2001.

19. R. Gras, P. Kuntz, and J.-C. Régnier. Significativité des niveaux d'une hiérarchie orientée en analyse statistique implicative. *Revue des Nouvelles Technologies de l'Information*, RNTI-C-1:39–50, 2004.

20. R. Gras and A. Larher. L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématiques, Informatique et Sciences Humaines*, 120:5–31, 1992.

21. R. Gras and H. Ratsimba-Rajohn. Analyse non symétrique de données par l'implication statistique. *RAIRO-Recherche opérationnelle*, 30-3:217–232, 1996.

22. S.C. Johnson. Hierarchical clustering scheme. *Psychometrika*, 32:241–254, 1967.

23. P. Kuntz, R. Gras, and J. Blanchard. Discovering extended rules with implicative hierarchies. In *Proc. of the new frontiers of statistical data mining and knowledge discovery*, pages 166–173. Knoxville, Tennesee, 2001.

24. J.B. Lagrange. Analyse implicative d'un ensemble de variables numériques: application au traitement d'un questionnaire aux réponses modales ordonnées. *Revue de statistique appliquée*, 46(1):71–93, 1998.

25. P. Lenca, P. Meyer, B. Vaillant, P. Picouet, and S. Lallich. Evaluation et analyse multi-critères des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information*, RNTI-E-1:219–246, 2004.

26. I.C. Lerman. *Classification et analyse ordinale des données*. Dunod, Paris, 1981.

27. J. Loevinger. A systemic approach to the construcion and evalation of tests of ability. *Psychological Monographs*, 61, 1947.

28. J. Piaget. *Le jugement et le raisonnement chez l'enfant.* Delachaux et Niestlé, 1967.
29. J.-C. Régnier and R. Gras. Statistique de rangs et analyse statistique implicative. *Revue de Statistique Appliquée*, LIII:5–38, 2005.
30. L. Seve. *Emergence, complexité et dialectique.* Odile Jacob, Paris, 2005.
31. A. Wald and J. Wolfowitz. Statistical tests based on permutations of the observations. *Ann. Math. Stat.*, 15, 1944.