Unsupervised aggregation of commensurate correlated attributes by means of the Choquet integral and entropy functionals *

Ivan Kojadinovic
LINA CNRS FRE 2729
Site école polytechnique de l'université de Nantes
Rue Christian Pauc, 44306 Nantes, France
ivan.kojadinovic@polytech.univ-nantes.fr

Version : April 2005

Abstract

In the framework of aggregation by the discrete Choquet integral, the *unsupervised* method for the identification of the underlying capacity initially put forward in [26] is presented and improvements are proposed. The suggested approach consists in replacing the subjective notion of *importance* of a subset of attributes by that of *information content* of a subset of attributes, which can be estimated from the set of profiles by means of an entropy measure. An example of the application of the proposed methodology is given : in the absence of *initial preferences*, the approach is applied to the evaluation of students.

1 Introduction

Consider a finite set of objects $\mathcal{O} := \{o_1, \ldots, o_n\}$ described by m cardinal attributes $A_1, \ldots, A_m$ defined on interval scales. Each object $o \in \mathcal{O}$ is identified with its *profile* $(a_1^o, \ldots, a_m^o) \in \mathbb{R}^m$ where, for any $i \in \{1, \ldots, m\}$, a_i^o is the value of attribute A_i for object o. For each object $o \in \mathcal{O}$, we shall further assume that the values $a_1^o, \ldots, a_m^o$ are given on the same scale, which implies that all the attributes are commensurate.

In numerous situations, it is useful to be able to associate with each object $o \in \mathcal{O}$ one unique value resulting from the merging of the values $a_1^o, \ldots, a_m^o$. In order to perform this step, an aggregation operator is to be used [9, 36].

In the presence of independent attributes, one of the most frequently used aggregation operator is the weighted arithmetic mean. The unique value $W_\omega(a)$ associated with a profile

*This paper is a revised and extended version with proofs of the conference paper [27].

$a = (a_1, \ldots, a_m)$ is then given by

$$W_\omega(a) := \sum_{i=1}^{m} \omega_i a_i,$$

where, for any $i \in \{1, \ldots, m\}$, ω_i is the *weight* of attribute A_i with $\omega_i \geq 0$ and $\sum_{i=1}^{m} \omega_i = 1$.

The assumption of independence among attributes is however rarely verified. In order to be able to model interaction phenomena among attributes, it has been recently proposed to substitute a monotone set function μ on $M := \{A_1, \ldots, A_m\}$ to the weight vector ω, thereby allowing to model not only the importance of each attribute but also the importance of each subset of attributes [14, 34]. Such a set function μ, called *(discrete) Choquet capacity* [6] or *fuzzy measure* [50], satisfies $\mu(\emptyset) = 0$, $\mu(M) = 1$ and $\mu(S) \leq \mu(T)$ whenever $S \subseteq T \subseteq M$.

A suitable aggregation operator that generalizes the weighted arithmetic mean when the attributes interact is the discrete Choquet integral with respect to (w.r.t) the capacity μ [14, 32, 34].

The use of a weighted arithmetic mean (resp. Choquet integral) as an aggregation operator first requires the definition of a weight vector ω (resp. a capacity μ). When the aggregation is to be carried out in the context of multicriteria decision making, additional information on the underlying problem given by a decision maker can be used to determine ω or μ. This additional knowledge, called *initial preferences* by Marchant [33], generally consists of preferences over the objects, intuitions about the importance of the attributes and the relationships among them, etc. In such a context, the aim is to model the preferences of the decision maker by means of the aggregation function. When the Choquet integral is to be used as aggregation operator, there exists several approaches that can be used to identify the capacity. More details can be found e.g. in [13, 18, 19, 38, 51]. In the sequel, we shall call these identification methods *supervised*.

The term *supervised* refers to the fact that some prior knowledge (*initial preferences*) has to be provided in order to fully determine the aggregation operator. The following question then naturally arises : what if the required knowledge cannot be easily given or, simply, is not available ? Such a situation may arise for instance in the context of sensor information fusion or aggregation of experts' points of view.

With these considerations in mind, an *unsupervised* identification method of the capacity based on the estimation of the interaction among attributes by means of information-theoretic functionals was proposed by the author in [26]. The suggested approach mainly consists in replacing the subjective notion of *importance* of a subset of attributes by that, probabilistic, of *information content* of a subset of attributes, which can be estimated from the set of profiles. Although it clearly does not pertain to the field of multicriteria decision making since it is unsupervised, the proposed identification method could still be considered as an alternative to the existing approaches developed in [13, 18, 19, 38, 51] when the prior knowledge they rely on cannot be provided. From a practical perspective, a sufficiently large number of profiles is necessary to obtain accurate estimates of the capacity coefficients and therefore of the Choquet integral. The proposed methodology has been implemented within the `kappalab` package [16] for the GNU R statistical system [42] (cf. § 4.3).

This paper is organized as follows. First, the Choquet integral is presented in the framework of aggregation and numerical indices that can be used to interpret its behavior are given. In the next section, it is shown how entropy measures can be used to define

capacities in a probabilistic context and what properties are satisfied by the resulting set functions. In the last but one section, the unsupervised identification method initially put forward in [26] is presented in details and improvements are proposed. Finally, under the hypothesis of absence of initial preferences, the suggested approach is applied to the evaluation of 89 first year students in Mathematics and Physics from University of Reunion Island (France).

2 Aggregation by the Choquet integral

After presenting the Choquet integral in a discrete setting, we recall the main indices that can be used to interpret its behavior in the framework of aggregation.

2.1 Choquet capacities and integral

As mentioned in the introduction, interaction phenomena among attributes can be modeled by a *(discrete) Choquet capacity* [6] also called *fuzzy measure* [50].

Let $\mathcal{P}(M)$ denote the power set of $M := \{A_1, \ldots, A_m\}$. A discrete Choquet capacity on M is a set function $\mu : \mathcal{P}(M) \to [0, 1]$ satisfying the following conditions :

(i) $\mu(\emptyset) = 0$,

(ii) for any $S, T \subseteq M$, $S \subseteq T \Rightarrow \mu(S) \leq \mu(T)$.

The capacity is further said to be :

- *normalized* when $\mu(M) = 1$,
- *additive* whenever $\mu(S \cup T) = \mu(S) + \mu(S)$ for all $S, T \subseteq M$ such that $S \cap T = \emptyset$.
- *subadditive* whenever $\mu(S \cup T) \leq \mu(S) + \mu(T)$ for all $S, T \subseteq N$ such that $S \cap T = \emptyset$.
- *submodular* whenever $\mu(S \cup T) + \mu(S \cap T) \leq \mu(S) + v(T)$ for all $S, T \subseteq N$.

Unless otherwise stated, we shall consider only normalized capacities in the sequel.

In the framework of aggregation, for each subset of attributes $S \subseteq M$, the number $\mu(S)$ can be interpreted as the *weight* or the *importance* of S. The monotonicity of μ means that the weight of a subset of attributes can only increase when new attributes are added to it.

When using a capacity to model the importance of the subsets of attributes, a suitable aggregation operator that generalizes the weighted arithmetic mean is the discrete Choquet integral [14, 32, 34]. The Choquet integral of a function $a : M \to \mathbb{R}$ represented by the profile $(a_1, \ldots, a_m)$ w.r.t a capacity μ on M can be defined by

$$C_\mu(a) := \sum_{i=1}^{m} a_{(i)}[\mu(B_{(i)}) - \mu(B_{(i+1)})], \tag{1}$$

where the notation $_{(.)}$ indicates a permutation such that $a_{(1)} \leq \cdots \leq a_{(m)}$, $B_{(i)} := \{A_{(i)}, \ldots, A_{(m)}\}$, for all $i \in \{1, \ldots, m\}$, and $B_{(m+1)} := \emptyset$.

Seen as an aggregation operator, the Choquet integral w.r.t μ can be regarded as taking into account interaction phenomena among attributes, that is, complementarity or redundancy effects among elements of M modeled by μ [34]. Indeed, complementarity (resp. redundancy) between two disjoints subsets of attributes A and B can be naturally modeled by the inequality $\mu(A \cup B) \geq \mu(A) + \mu(B)$ (resp. $\leq$). Note that if μ is a submodular capacity, only redundant interaction between two disjoints subsets of attributes can be modeled.

The Choquet integral generalizes the weighted arithmetic mean in the sense that, as soon as the capacity is additive (which intuitively coincides with the independence of the attributes), it collapses into a weighted arithmetic mean, i.e.

$$C_\mu(a) := \sum_{i=1}^{m} a_i \mu(\{A_i\}), \qquad \forall a = (a_1, \ldots, a_m) \in \mathbb{R}^m.$$

More details can be found e.g. [14, 34].

An intuitive presentation of the Choquet integral can be found in [41]. Note that an axiomatic characterization of the Choquet integral as an aggregation operator was proposed by Marichal [34].

2.2 The Möbius representation of a capacity

Any set function $\nu : \mathcal{P}(M) \to \mathbb{R}$ can be uniquely expressed in terms of its *Möbius representation* [15, 34, 45] by

$$\nu(T) = \sum_{S \subseteq T} m^\nu(S), \qquad \forall\, T \subseteq M,$$

where the set function $m^\nu : \mathcal{P}(M) \to \mathbb{R}$ is called the *Möbius transform* or *Möbius representation* of ν and is given by

$$m^\nu(S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} \nu(T), \qquad \forall\, S \subseteq M. \tag{2}$$

In terms of the Möbius representation of a capacity μ, Chateauneuf and Jaffray [5] showed that, for any $a = (a_1, \ldots, a_m) \in \mathbb{R}^m$, the Choquet integral of a w.r.t μ is given by

$$C_\mu(a) = \sum_{T \subseteq M} m^\mu(T) \bigwedge_{i : A_i \in T} a_i,$$

where the symbol $\wedge$ denotes the minimum operator.

In the previous subsection, in the framework of aggregation, an interpretation of the coefficients $\mu(S)$, $S \subseteq M$, $S \neq \emptyset$, as *weights* of subsets of attributes has been given. As discussed in [26], in a similar way, the coefficients $\{m^\mu(S)\}$ can be assigned a particular meaning in that context : for any $S \subseteq M$, $|S| \geq 2$, the coefficient $m^\mu(S)$ can be interpreted as a measure of the *simultaneous interaction* among attributes in S. Should $m^\mu(S)$ be positive (resp. negative), the attributes in S can be regarded as simultaneously interacting in a *complementary* (resp. *redundant*) way. This interpretation is given even more weight by noticing that μ is additive if and only if $m^\mu(S) = 0$ for all $S \subseteq M$, $|S| \geq 2$, in which case, as mentioned before, the Choquet integral collapses into the weighted arithmetic mean.

2.3 Behavioral analysis of the aggregation

The behavior of the Choquet integral as an aggregation operator is generally difficult to understand. For a better comprehension of the interaction phenomena modeled by the underlying capacity, several numerical indices can be computed. In the sequel, we mention three of them. More details can be found e.g. in [35].

The global importance of an attribute A_i can be measured by means of its Shapley value [49], which is given by

$$\phi(\mu, A_i) := \sum_{T \subseteq M \setminus \{A_i\}} \frac{(m - |T| - 1)!|T|!}{m!} [\mu(T \cup \{A_i\}) - \mu(T)].$$

The Shapley value of an attribute A_i can be thought of as an average value of the *marginal contribution* $\mu(T \cup \{A_i\}) - \mu(T)$ of attribute A_i to a subset T not containing it.

The average interaction between two attributes A_i and A_j can be measured by means of their Shapley interaction index [15, 40] which can be computed by

$$I(\mu, \{A_i, A_j\}) := \sum_{T \subseteq M \setminus \{A_i, A_j\}} \frac{(m - |T| - 2)!|T|!}{(m-1)!} [\mu(T \cup \{A_i, A_j\}) - \mu(T \cup \{A_i\}) - \mu(T \cup \{A_j\}) + \mu(T)].$$

Similarly to the Shapley value, the Shapley interaction index between two attributes A_i and A_j can be regarded as an average value of their *marginal interaction*

$$\mu(T \cup \{A_i, A_j\}) - \mu(T \cup \{A_i\}) - \mu(T \cup \{A_j\}) + \mu(T)$$

in the presence of a subset T of attributes not containing them [17, 25].

The last index we shall mention is the Marichal entropy of a capacity, which is defined by

$$H_M(\mu) := -\sum_{i=1}^{m} \sum_{T \subseteq M \setminus \{A_i\}} \frac{(m - |T| - 1)!|T|!}{m!} [\mu(T \cup \{A_i\}) - \mu(T)] \ln[\mu(T \cup \{A_i\}) - \mu(T)].$$

In the context of aggregation by the Choquet integral w.r.t μ, $H_M(\mu)$ can be interpreted as the average degree of utilization of a profile [28, 29]. More precisely, should $H_M(\mu)$ be close to its maximum $\ln m$, all the partial evaluations $a_1, \ldots, a_m$ of a profile a will be involved almost equally on average in the calculation of $C_\mu(a)$, which will then be close to the arithmetic mean of the a_i's. On the contrary, should $H_M(\mu)$ be close to zero, $C_\mu(a)$ will be very close on average to one partial evaluation and C_μ shall therefore exhibit a very disjunctive or conjunctive behavior.

3 Capacities induced by entropy measures

The proposed unsupervised identification method is grounded on the notion of *entropy*. In this section, we present theoretical results necessary for the development of the suggested approach.

3.1 Shannon entropy

The fundamental concept of *entropy of a probability distribution* was initially proposed by Shannon [47, 48]. It can be interpreted as a measure of the *uncertainty* or the *information* or, equivalently, the *structure* contained in a probability distribution.

The Shannon entropy of a probability distribution p defined on a nonempty finite set N is defined by

$$H_S(p) := -\sum_{i \in N} p(i) \ln p(i)$$

with the convention that $0 \ln 0 := 0$. The quantity $H_S(p)$ is always non negative and zero if and only if p is a Dirac mass (*decisivity* property). It reaches its maximum value ($\ln |N|$) if and only if p is uniform (*maximality* property).

In a general non probabilistic setting, $H_S(p)$ is merely a measure of the uniformity (evenness) of p. In a probabilistic context, when p is associated with an $|N|$-state discrete stochastic system, it is naturally interpreted as a measure of its unpredictability and thus reflects the uncertainty associated with a future state of the system.

Several axiomatic characterizations of the Shannon entropy were proposed in the literature [3, 10, 21, 23], among which the most famous is probably Shannon's theorem [48].

Note that the Shannon entropy was also defined in a continuous setting [7, Chap. 9]. However, the Shannon entropy of a probability density, when it exists, is not necessarily non negative and cannot therefore be interpreted as an information measure anymore.

3.2 Mutual information

The Shannon entropy is closely linked with the concept of *mutual information* at the root of information theory [7].

Let us consider two discrete random vectors $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ taking a finite number of values. The *mutual information* between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ is defined as the *distance from independence* between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ measured by the Kullback and Leibler divergence [7, 30, 31, 52].

For two probability distributions p and q on N with same support, the Kullback and Leibler divergence is defined by

$$KL(p, q) := \sum_{i \in N} p(i) \ln \left(\frac{p(i)}{q(i)} \right) \tag{3}$$

with the convention that $0 \ln \frac{0}{0} := 0$.

Let us denote by $p_{(\vec{\mathbb{X}}, \vec{\mathbb{Y}})}$ the joint distribution of $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ and by $p_{\vec{\mathbb{X}}}$ and $p_{\vec{\mathbb{Y}}}$ respectively the marginal distributions of $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ respectively. The mutual information between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ is then defined by

$$I(\vec{\mathbb{X}}; \vec{\mathbb{Y}}) := KL(p_{(\vec{\mathbb{X}}, \vec{\mathbb{Y}})}, p_{\vec{\mathbb{X}}} \otimes p_{\vec{\mathbb{Y}}}),$$

where $p_{\vec{\mathbb{X}}} \otimes p_{\vec{\mathbb{Y}}}$ denotes the tensor product of $p_{\vec{\mathbb{X}}}$ and $p_{\vec{\mathbb{Y}}}$. From the above definition, we see that the mutual information is symmetric and, by applying the Jensen inequality to the Kullback and Leibler divergence, we obtain that the mutual information is always

non negative and zero if and only if $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ are stochastically independent. Furthermore, as discussed in [22], the "stronger the stochastic dependence" between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$, the higher $I(\vec{\mathbb{X}};\vec{\mathbb{Y}})$.

The mutual information can also be interpreted as the *uncertainty reduction measure* [8] obtained from the Shannon entropy. Indeed, w.r.t the Shannon entropy, the mutual information between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ can be easily rewritten as

$$I(\vec{\mathbb{X}};\vec{\mathbb{Y}}) = H(p_{\vec{\mathbb{X}}}) - E_{p_{\vec{\mathbb{Y}}}}[H(p_{\vec{\mathbb{X}}|\vec{\mathbb{Y}}=y})], \tag{4}$$

where $p_{\vec{\mathbb{X}}|\vec{\mathbb{Y}}=y}(x) := \frac{p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(x,y)}{p_{\vec{\mathbb{Y}}}(y)}$. By symmetry, we also have

$$I(\vec{\mathbb{X}};\vec{\mathbb{Y}}) = H(p_{\vec{\mathbb{Y}}}) - E_{p_{\vec{\mathbb{X}}}}[H(p_{\vec{\mathbb{Y}}|\vec{\mathbb{X}}=x})].$$

Hence, the mutual information can be interpreted as the reduction in the uncertainty of $\vec{\mathbb{X}}$ (resp. $\vec{\mathbb{Y}}$) due to the knowledge of $\vec{\mathbb{Y}}$ (resp. $\vec{\mathbb{X}}$) [52]. Rewriting the expectation in Eq. (4), we obtain

$$E_{p_{\vec{\mathbb{Y}}}}[H(p_{\vec{\mathbb{X}}|\vec{\mathbb{Y}}=y})] = H(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}) - H(p_{\vec{\mathbb{Y}}}), \tag{5}$$

and therefore

$$I(\vec{\mathbb{X}};\vec{\mathbb{Y}}) = H(p_{\vec{\mathbb{X}}}) + H(p_{\vec{\mathbb{Y}}}) - H(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}). \tag{6}$$

From the above formula, Abramson proposed a natural extension of the mutual information between more than two random vectors [1]. The mutual information among three random vectors $\vec{\mathbb{X}}$, $\vec{\mathbb{Y}}$ and $\vec{\mathbb{Z}}$ is defined by

$$I(\vec{\mathbb{X}};\vec{\mathbb{Y}};\vec{\mathbb{Z}}) := H(p_{\vec{\mathbb{X}}}) + H(p_{\vec{\mathbb{Y}}}) + H(p_{\vec{\mathbb{Z}}}) - H(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}) - H(p_{(\vec{\mathbb{X}},\vec{\mathbb{Z}})}) - H(p_{(\vec{\mathbb{Y}},\vec{\mathbb{Z}})}) + H(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}},\vec{\mathbb{Z}})}).$$

More generally, for $r \geq 2$ random vectors $\vec{\mathbb{X}}_1,\ldots,\vec{\mathbb{X}}_r$, the following definition was adopted by Abramson :

$$I(\vec{\mathbb{X}}_1;\ldots;\vec{\mathbb{X}}_r) := \sum_{k=1}^{r} \sum_{\{i_1,\ldots,i_k\}\subseteq\{1,\ldots,r\}} (-1)^{k+1} H(p_{(\vec{\mathbb{X}}_{i_1},\ldots,\vec{\mathbb{X}}_{i_k})}). \tag{7}$$

The mutual information among $r \geq 2$ random vectors $\vec{\mathbb{X}}_1,\ldots,\vec{\mathbb{X}}_r$ can be interpreted as a measure of their *simultaneous interaction* [53]. Should it be zero, we can say that the r random vectors do not interact "altogether". Note that the mutual information between more than two random vectors is not necessarily non negative [7].

3.3 Generalizations of the Shannon entropy : Rényi entropy of order α and Havrda and Charvat entropy of order β

Rényi [43] was the first to consider natural modifications of the axioms of Shannon. Given a strictly positive real α, the Rényi entropy of order α of a probability distribution p on N, is defined by

$$H_R^\alpha(p) := \begin{cases} \dfrac{1}{1-\alpha} \ln \sum_{i\in N} p(i)^\alpha, & \alpha \neq 1, \\ \\ H_S(p), & \alpha = 1. \end{cases}$$

The quantity $H_R^1(p)$ corresponds in fact to the limit of $H_R^\alpha(p)$ when α tends to 1.

The Rényi entropy of order α satisfies most of the properties satisfied by the Shannon entropy, like the decisivity and maximality properties. Furthermore, it is important to notice that the Rényi entropy of order α, $\alpha \neq 1$, is the only entropy measure, besides that of Shannon, to satisfy the *additivity* property, i.e.,

$$H_R^\alpha(p \otimes q) = H_R^\alpha(p) + H_R^\alpha(q), \tag{8}$$

where p is a probability distribution on a finite set $\mathcal{X}$, q is a probability distribution on a finite set $\mathcal{Y}$ and where $p \otimes q$ is the tensor product of p and q. Finally, note that, given a probability distribution p on N, $H_R^\alpha(p)$ is a decreasing function of α [43].

An axiomatic characterization of the Rényi entropy of order α was proposed by Aczél and Daróczy in [2].

Another well-known generalization of the Shannon entropy is the Havrda and Charvat entropy of order β [20] defined, for any strictly positive real β, and any probability distribution p on N, by

$$H_{HC}^\beta(p) := \begin{cases} \dfrac{1}{1-\beta}\left[\sum_{i \in N} p(i)^\beta - 1\right], & \beta \neq 1, \\ \\ H_S(p), & \beta = 1. \end{cases}$$

As the two previous entropies, the Havrda and Charvat entropy of order β satisfies the decisivity property and the maximality property (with the exception that its maximal value is not necessarily $\ln |N|$).

The axiomatic characterization of the entropy H_{HC}^β is very similar to that of the Shannon entropy proposed by Fadeev [12]. The only difference comes from the form of recursivity axiom [20].

Note that many other generalizations of the Shannon entropy were proposed in the literature. For an overview, see e.g. [11].

3.4 Capacities induced by the Shannon, Rényi and Havrda and Charvat entropies

In a probabilistic setting, entropy measures can be used to define capacities on a a set of discrete random variables.

Let $\aleph := \{X_1, \ldots, X_m\}$ be a set of discrete random variables. The subsets of $\aleph$ will be denoted by upper-case *black-board* letters, e.g. $\mathbb{X}$. Given a subset $\mathbb{X} \subseteq \aleph$ composed of r variables, $\vec{\mathbb{X}}$ will denote an r-dimensional random vector whose coordinates are distinct elements from $\mathbb{X}$. The probability distribution of a random vector $\vec{\mathbb{X}}$ will be denoted by $p_{\vec{\mathbb{X}}}$.

As discussed in [24, 26], in a probabilistic context, the *importance* of the subsets of $\aleph$ could be naturally quantified in terms of *probabilistic information*. A set function modeling the importance of the subsets of $\aleph$ can thus be defined by

$$h(\mathbb{X}) := \begin{cases} 0, & \text{if } \mathbb{X} = \emptyset, \\ \\ H(p_{\vec{\mathbb{X}}}), & \text{if } \mathbb{X} \neq \emptyset, \end{cases} \qquad \forall\, \mathbb{X} \subseteq \aleph, \tag{9}$$

where H is an entropy measure.

In the rest of this subsection, we shall study the set functions obtained from the Shannon entropy, the Rényi entropy of order α and the Havrda and Charvat entropy of order β. They will be denoted by h_S, h_R^α and h_{HC}^β respectively.

Proposition 3.1. *The set function h_S is additive if and only if $X_1, \ldots, X_m$ are stochastically mutually independent.*

Proof. The result is a direct consequence of the additivity property satisfied by H_S ; cf. Eq. (8). □

Proposition 3.2. *The set function h_S is monotonic.*

Proof. Let $\mathbb{X}$ and $\mathbb{Y}$ be two nonempty disjoint subsets of $\aleph$. From Eq. (5) and the non negativity of the Shannon entropy, we can write

$$H_S(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}) \geq H_S(p_{\vec{\mathbb{Y}}}),$$

which is equivalent to

$$h_S(\mathbb{X} \cup \mathbb{Y}) \geq h_S(\mathbb{Y}),$$

from where we obtain the desired result. □

The set function h_S is thus an unnormalized capacity on $\aleph$.

Proposition 3.3. *The Möbius representation of h_S is given by*

$$m^{h_S}(\mathbb{X}) := \begin{cases} 0, & \text{if } \mathbb{X} = \emptyset, \\ (-1)^{|\mathbb{X}|+1} I(X_1; \ldots; X_r), & \text{if } \mathbb{X} = \{X_1, \ldots, X_r\}, \end{cases} \qquad \forall\, \mathbb{X} \subseteq \aleph.$$

Proof. The result immediately follows from Eqs. (2) and (7). □

In other words, defining the importance of the subsets of $\aleph$ by means of the Shannon entropy is equivalent to measuring the simultaneous interaction among the random variables of $\aleph$ by means of the concept of mutual information (cf. § 3.2).

Proposition 3.4. *The capacity h_S is submodular.*

Proof. Let $\mathbb{X}$, $\mathbb{Y}$ and $\mathbb{Z}$ be three nonempty and pairwise disjoint subsets and let $p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})|\vec{\mathbb{Z}}=z}$, $p_{\vec{\mathbb{X}}|\vec{\mathbb{Z}}=z}$ and $p_{\vec{\mathbb{Y}}|\vec{\mathbb{Z}}=z}$ be the probability distributions respectively defined by

$$p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})|\vec{\mathbb{Z}}=z}(x,y) := \frac{p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}},\vec{\mathbb{Z}})}(x,y,z)}{p_{\vec{\mathbb{Z}}}(z)},\ p_{\vec{\mathbb{X}}|\vec{\mathbb{Z}}=z}(x) := \frac{p_{(\vec{\mathbb{X}},\vec{\mathbb{Z}})}(x,z)}{p_{\vec{\mathbb{Z}}}(z)} \text{ and } p_{\vec{\mathbb{Y}}|\vec{\mathbb{Z}}=z}(y) := \frac{p_{(\vec{\mathbb{Y}},\vec{\mathbb{Z}})}(y,z)}{p_{\vec{\mathbb{Z}}}(z)}.$$

Starting from Eq. (3), we have

$$E_{p_{\vec{\mathbb{Z}}}}[KL(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})|\vec{\mathbb{Z}}=z}, p_{\vec{\mathbb{X}}|\vec{\mathbb{Z}}=z} \otimes p_{\vec{\mathbb{Y}}|\vec{\mathbb{Z}}=z})] = E_{p_{\vec{\mathbb{Z}}}}[H_S(p_{\vec{\mathbb{X}}|\vec{\mathbb{Z}}=z})] + E_{p_{\vec{\mathbb{Z}}}}[H_S(p_{\vec{\mathbb{Y}}|\vec{\mathbb{Z}}=z})] - E_{p_{\vec{\mathbb{Z}}}}[H_S(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})|\vec{\mathbb{Z}}=z})],$$

which, using Eq. (5), can be rewritten as

$$\begin{aligned} E_{p_{\vec{\mathbb{Z}}}}[KL(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})|\vec{\mathbb{Z}}=z}, p_{\vec{\mathbb{X}}|\vec{\mathbb{Z}}=z} \otimes p_{\vec{\mathbb{Y}}|\vec{\mathbb{Z}}=z})] &= -H_S(p_{\vec{\mathbb{Z}}}) + H_S(p_{(\vec{\mathbb{X}},\vec{\mathbb{Z}})}) + H_S(p_{(\vec{\mathbb{Y}},\vec{\mathbb{Z}})}) - H_S(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}},\vec{\mathbb{Z}})}), \\ &= -h_S(\mathbb{Z}) + h_S(\mathbb{X} \cup \mathbb{Z}) + h_S(\mathbb{Y} \cup \mathbb{Z}) - h_S(\mathbb{X} \cup \mathbb{Y} \cup \mathbb{Z}). \end{aligned}$$

The Kullback and Leibler divergence being always non negative, we have

$$E_{p_{\vec{\mathbb{Z}}}}[KL(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})|\vec{\mathbb{Z}}=z}, p_{\vec{\mathbb{X}}|\vec{\mathbb{Z}}=z} \otimes p_{\vec{\mathbb{Y}}|\vec{\mathbb{Z}}=z})] \geq 0,$$

which is equivalent to the submodularity of h_S. □

In other terms, the capacity h_S can only model redundant interactions between two disjoint subsets of variables of $\aleph$. In certain situations, this behavior could be considered as too restrictive. A first way to obtain a set function having weaker properties than h_S consists in using the Rényi entropy of order α instead of the Shannon entropy. Indeed, the resulting set function h_R^α is not necessarily submodular because it is not subadditive in general [44]. However, as for the Shannon entropy, the additivity of h_R^α coincides with the stochastic mutual independence of the random variables of $\aleph$.

Proposition 3.5. *The set function h_R^α is additive if and only if $X_1, \ldots, X_m$ are stochastically mutually independent.*

Proof. The result is a direct consequence of the additivity property satisfied by H_R^α ; cf. Eq. (8). □

Proposition 3.6. *The set function h_R^α is monotonic.*

Proof. Consider two nonempty disjoint subsets $\mathbb{X}$ and $\mathbb{Y}$ of $\aleph$ and assume that the corresponding random vectors $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ take their values in the finite sets $\{x_1, \ldots, x_r\}$ and $\{y_1, \ldots, y_s\}$ respectively.

Recall that α is by definition a strictly positive real number different from 1.

For $\alpha > 1$ (resp. $\alpha < 1$), for any $i \in \{1, \ldots, r\}$, we have

$$\left(\sum_{j=1}^{s} p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(x_i, y_j)\right)^\alpha \geq \sum_{j=1}^{s} p^\alpha_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(x_i, y_j) \text{ (resp. } \leq),$$

which implies that

$$\sum_{i=1}^{r}\left(\sum_{j=1}^{s} p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(x_i, y_j)\right)^\alpha \geq \sum_{i=1}^{r}\sum_{j=1}^{s} p^\alpha_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(x_i, y_j) \text{ (resp. } \leq),$$

Using the decreasingness (resp. increasingness) of $x \mapsto x^{\frac{1}{1-\alpha}}$ on $[0, \infty[$, for any $\alpha \neq 1$, we obtain

$$\left(\sum_{i=1}^{r}\left(\sum_{j=1}^{s} p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(x_i, y_j)\right)^\alpha\right)^{\frac{1}{1-\alpha}} \leq \left(\sum_{i=1}^{r}\sum_{j=1}^{s} p^\alpha_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(x_i, y_j)\right)^{\frac{1}{1-\alpha}},$$

which is equivalent to

$$\ln\left(\sum_{i=1}^{r} p^\alpha_{\vec{\mathbb{X}}}(x_i)\right)^{\frac{1}{1-\alpha}} \leq \ln\left(\sum_{i=1}^{r}\sum_{j=1}^{s} p^\alpha_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(x_i, y_j)\right)^{\frac{1}{1-\alpha}},$$

form which we have

$$H_R^\alpha(p_{\vec{\mathbb{X}}}) \leq H_R^\alpha(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})})$$

and thus the desired result. □

If it is not required that the additivity of the induced capacity coincides with the mutual stochastic independence of the variables of $\aleph$, the Havrda and Charvat entropy of order β can be used to define the importance of the subsets of $\aleph$.

Proposition 3.7. *The set function h^{β}_{HC} is monotonic.*

Proof. Consider two nonempty disjoint subsets $\mathbb{X}$ and $\mathbb{Y}$ of $\aleph$. From the definition of the Havrda and Charvat entropy of order β, $\beta \neq 1$, Ullah [52] showed that

$$H^{\beta}_{HC}(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}) = H^{\beta}_{HC}(p_{\vec{\mathbb{X}}}) + E_{p_{\vec{\mathbb{X}}}}[p^{\beta-1}_{\vec{\mathbb{X}}} H^{\beta}_{HC}(p_{\vec{\mathbb{Y}}|\vec{\mathbb{X}}=x})].$$

The non negativity of the entropy of Havrda and Charvat of order β, $\beta \neq 1$, implies that

$$H^{\beta}_{HC}(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}) \geq H^{\beta}_{HC}(p_{\vec{\mathbb{X}}}),$$

which is immediately equivalent to

$$h^{\beta}_{HC}(\mathbb{X} \cup \mathbb{Y}) \geq h^{\beta}_{HC}(\mathbb{X}),$$

from which we obtain the desired result. □

4 A probabilistic view of the identification problem

In the context of aggregation by the Choquet integral and in the absence of initial preferences, the only available information from which the weights of the subsets of attributes (i.e. the underlying capacity) could be identified is the set of profiles. In such an unsupervised context, as mentioned in the introduction, the problem of the identification of the capacity can be regarded as an estimation problem. Hence, with each attribute A_i is uniquely associated a random variable X_i such that, for any object $o \in O$, the value a^o_i is seen as a realization of X_i. The set of m random variables associated with the set of attributes M is denoted $\aleph = \{X_1, \ldots, X_m\}$ as in the previous section. Every profile $a^o = (a^o_1, \ldots, a^o_m)$ can thus be seen as a realization of the random vector $\vec{\aleph} = (X_1, \ldots, X_m)$.

4.1 Definition of the weights

In [26], it was suggested to define the weights of the nonempty subsets of attributes by means of a capacity induced by an entropy measure, thereby replacing the subjective notion of importance by that of *information content*. In order to ensure that the resulting set function is monotonic, we further assume that the random variables $X_1, \ldots, X_m$ are discrete (cf. § 3.4) and take their values in the finite sets $\mathcal{X}_1, \ldots, \mathcal{X}_m$. From a practical perspective, as we shall see in § 5.2, this may require a prior discretization of the available profiles before the estimation of the weights. The trivial situation where the joint probability distribution $p_{(X_1,\ldots,X_m)}$ of $X_1, \ldots, X_m$ is a Dirac mass is also excluded. Indeed, in that case, all the profiles would necessarily be equal and further aggregation would make little sense. The

weight of every subset $S \subseteq M$ of attributes can then be defined by

$$\mu(S) := \begin{cases} 0, & \text{if } S = \emptyset, \\ \dfrac{h(\{X_{i_1}, \ldots, X_{i_s}\})}{h(\aleph)}, & \text{if } S = \{A_{i_1}, \ldots, A_{i_s}\}, \end{cases} \tag{10}$$

where h is the set function generically defined by Eq. (9). Notice that μ is well-defined provided the underlying entropy measure satisfies the decisivity property. Indeed, in this case $h(\aleph) \neq 0$ since it was assumed that $p_{(X_1,\ldots,X_m)}$ is not a Dirac mass.

The choice of the unnormalized capacity h directly influences the properties of μ. Among the three capacities studied in § 3.4, h_S may appear as the most appropriate one since a submodular behavior is natural in the considered context. Indeed, in a unsupervised setting, two attributes should be able to interact only in a redundant way since, in order to detect complementarity effects between two attributes, initial preferences would be necessary. The use of capacities h_R^α and h_{HC}^β should not however be excluded since they may offer a more flexible alternative. Indeed, parameters α and β could be used to tune the resulting set functions.

In the sequel, the resulting capacities shall be denoted μ_S (resp. μ_R^α, μ_{HC}^β) when obtained from h_S (resp. h_R^α, h_{HC}^β).

The random variables $X_1, \ldots, X_m$ being discrete and taking their values in the finite sets $\mathcal{X}_1, \ldots, \mathcal{X}_m$ respectively, μ_S, μ_R^α and μ_{HC}^β are therefore (normalized) capacities on M.

4.2 Estimation

The weights of the subsets of attributes all depend on the joint distribution $p_{(X_1,\ldots,X_m)}$. Indeed, for any $S = \{A_{i_1}, \ldots, A_{i_s}\} \subseteq M$, from the definitions of μ and h, we have

$$\mu(S) = \frac{h(\{X_{i_1}, \ldots, X_{i_s}\})}{h(\aleph)} = \frac{H(p_{(X_{i_1},\ldots,X_{i_s})})}{H(p_{(X_1,\ldots,X_m)})}.$$

The distribution of $(X_{i_1}, \ldots, X_{i_s})$ can in turn be immediately deduced from that of $\vec{\aleph} = (X_1, \ldots, X_m)$. Indeed, for any $(x_{i_1}, \ldots, x_{i_s}) \in \mathcal{X}_{i_1} \times \cdots \times \mathcal{X}_{i_s}$, we have

$$p_{(X_{i_1},\ldots,X_{i_s})}(x_{i_1}, \ldots, x_{i_s}) = \sum_{x_{i_{s+1}} \in \mathcal{X}_{i_{s+1}}} \cdots \sum_{x_{i_m} \in \mathcal{X}_{i_m}} p_{(X_1,\ldots,X_m)}(x_1, \ldots, x_m), \tag{11}$$

where $\{A_{i_{s+1}}, \ldots, A_{i_m}\} = M \setminus S$.

Hence, the coefficients $\mu(S)$, $S \subseteq M$, $S \neq \emptyset$, are all clearly functions of $p_{(X_1,\ldots,X_m)}$. The same is therefore true for the Choquet integral of a profile $a = (a_1, \ldots, a_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ w.r.t μ. For the subsequent developments, it is convenient to rewrite $C_\mu(a)$ as

$$C_\mu(a) = \sum_{i=1}^{m} (a_{(i)} - a_{(i-1)}) \mu(B_{(i)})$$

where $(.)$ is a permutation of M such that $a_{(1)} \leq \cdots \leq a_{(m)}$, $B_{(i)} := \{A_{(i)}, \ldots, A_{(m)}\}$, and $a_{(0)} := 0$ by convention.

The weights of the subsets of attributes (and therefore the Choquet integral of a profile a w.r.t μ) can thus be estimated from realizations of $\vec{\aleph} = (X_1, \ldots, X_m)$. Indeed, given a random sample $\vec{\aleph}_1, \ldots, \vec{\aleph}_n$ of $\vec{\aleph}$, a natural estimator of the weight of a subset $S = \{A_{i_1}, \ldots, A_{i_s}\} \subseteq M$ is immediately given by

$$\hat{\mu}(S) = \frac{H(\hat{p}_{(X_{i_1},\ldots,X_{i_s})})}{H(\hat{p}_{(X_1,\ldots,X_m)})}, \tag{12}$$

where $\hat{p}_{(X_1,\ldots,X_m)}$ is the classical maximum likelihood estimator defined, for any $(x_1, \ldots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$, by

$$\hat{p}_{(X_1,\ldots,X_m)}(x_1, \ldots, x_m) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{(x_1,\ldots,x_m)\}}(\vec{\aleph}_j),$$

where $\mathbf{1}_{\{(x_1,\ldots,x_m)\}}(\vec{\aleph})$ is the indicator function of event $\{\vec{\aleph} = (x_1, \ldots, x_m)\}$. It is easy to verify that, for any $S = \{A_{i_1}, \ldots, A_{i_s}\} \subseteq M$, estimator $\hat{p}_{(X_{i_1},\ldots,X_{i_s})}$ is linked to estimator $\hat{p}_{(X_1,\ldots,X_m)}$ by an equality similar to Eq. (11).

A natural estimator of the Choquet integral of a profile $a = (a_1, \ldots, a_m)$ w.r.t μ is then obtained by substitution, i.e.,

$$\hat{C}_\mu(a) = C_{\hat{\mu}}(a) = \sum_{i=1}^{m} (a_{(i)} - a_{(i-1)}) \hat{\mu}(B_{(i)}). \tag{13}$$

Estimator $\hat{C}_\mu(a)$ being clearly a function of $\hat{p}_{(X_1,\ldots,X_m)}$, its asymptotic properties can immediately be obtained using the *delta* method [39, 46, 4]. As shown in [26], we have

$$\lim_{n\to\infty} \hat{C}_\mu(a) = C_\mu(a) \text{ almost surely,}$$

and

$$\lim_{n\to\infty} n^{1/2}(\hat{C}_\mu(a) - C_\mu(a)) = N(0, \sigma_C^2) \text{ in distribution,}$$

where

$$\sigma_C^2 = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_m \in \mathcal{X}_m} \left(C'_{(x_1,\ldots,x_m)}\right)^2 p_{(X_1,\ldots,X_m)}(x_1, \ldots, x_m) - \left(\sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_m \in \mathcal{X}_m} C'_{(x_1,\ldots,x_m)} p_{(X_1,\ldots,X_m)}(x_1, \ldots, x_m) \right)^2,$$

and where, for any $(x_1, \ldots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$, $C'_{(x_1,\ldots,x_m)}$ denotes the partial derivative of $C_\mu(a)$ (seen as a function of $p_{(X_1,\ldots,X_m)}$) w.r.t $p_{(X_1,\ldots,X_m)}(x_1, \ldots, x_m)$. A classical calculation gives

$$C'_{(x_1,\ldots,x_m)} = \frac{1}{H(p_{(X_1,\ldots,X_m)})} \left[\left(\sum_{i=1}^{m} (a_{(i)} - a_{(i-1)}) \frac{\partial H(p_{(X_{(i)},\ldots,X_{(m)})})}{\partial p_{(X_1,\ldots,X_m)}(x_1, \ldots, x_m)} \right) - \frac{\partial H(p_{(X_1,\ldots,X_m)})}{\partial p_{(X_1,\ldots,X_m)}(x_1, \ldots, x_m)} C_\mu(a) \right].$$

In the case of the Shannon entropy for instance, we have

$$\frac{\partial H_S(p_{(X_{(i)},\ldots,X_{(m)})})}{\partial p_{(X_1,\ldots,X_m)}(x_1,\ldots,x_m)} = \ln\left(\sum_{x_{(1)}\in\mathcal{X}_{(1)}}\cdots\sum_{x_{(i-1)}\in\mathcal{X}_{(i-1)}} p_{(X_1,\ldots,X_m)}(x_1,\ldots,x_m)\right)+1.$$

The previous properties can be used for instance to obtain an approximate confidence interval for $C_\mu(a)$ [26].

Before ending this section, let us give an interpretation of $\hat{\mu}$. By considering Eq. (12), we can see that the weight of a nonempty subset of attributes directly depends on the uniformity of the corresponding estimated probability distribution : roughly speaking, the more discriminative among the alternatives a subset of attributes is, the more uniform the corresponding estimated probability distribution, the higher its weight, and reciprocally.

4.3 Practical implementation

The proposed approach was implemented within the `kappalab` package [16] for the `GNU R` statistical system [42]. The package is distributed as free software and should be soon downloadable from the `Comprehensive R Archive Network` (`http://cran.r-project.org`).

5 Application

In order to illustrate the application of the proposed identification method, let us consider the classical problem which consists in assigning global evaluations to students from their partial evaluations in different subjects. However, we shall here further assume that no initial preferences are available. The data correspond to marks of 89 first year students in Mathematics and Physics from University of Reunion Island (France) for five subjects : English (Eng), Computer Science (Com), Algebra (Alg), Analysis (Ana) and Physics (Phy). The marks on a 0 to 20 scale are given in Table 1. For each subject, the minimum, maximum, average mark and the standard deviation of the marks are given in Table 2 as well as the linear correlation matrix of the data.

The aim is to estimate the weights of the subsets of attributes (i.e. subjects) from the available profiles by means of the capacities $\hat{\mu}_S$, $\hat{\mu}_R^\alpha$ and $\hat{\mu}_{HC}^\beta$ previously defined and to compute the global evaluation of each student from its partial evaluations by means of the Choquet integral w.r.t these capacities. In the absence of initial preferences, the most natural aggregation operator for this task would be the simple arithmetic mean. Thus, in a second stage, the behavior of the Choquet integral will be compared with that of the simple arithmetic mean.

5.1 The problem of the non commensurateness of the partial evaluations

Before estimating the weights and the global evaluations of the students, it is fundamental to see whether the partial evaluations given in Table 1 can be considered as commensurate

Table 1: Marks of the 89 students to be evaluated.

N°	Eng	Com	Alg	Ana	Phy
1	7,7	14,3	3,9	3,7	9,4
2	11,0	12,0	1,1	4,6	4,4
3	14,3	1,3	4,9	3,6	7,6
4	11,2	9,5	2,2	3,3	3,3
5	6,3	7,3	8,7	6,3	11,5
6	9,5	5,6	4,4	3,3	4,7
7	9,7	9,7	1,5	2,0	1,7
8	9,7	8,9	9,8	7,9	8,4
9	13,0	1,3	5,4	5,4	7,0
10	6,7	5,6	3,7	4,0	5,0
11	11,7	6,4	5,0	6,3	6,4
12	14,7	11,4	7,0	5,9	1,8
13	12,3	6,6	8,0	6,8	1,4
14	13,7	9,6	5,3	4,9	8,7
15	1,0	7,7	11,0	7,6	6,8
16	15,7	2,0	3,2	3,9	6,3
17	7,0	5,9	7,5	4,9	5,1
18	14,7	13,6	11,6	8,6	1,4
19	11,3	4,7	5,2	3,1	2,7
20	17,0	17,8	9,3	6,5	7,8
21	5,0	1,6	12,2	9,9	6,3
22	7,7	1,6	4,3	5,3	7,5
23	14,3	7,2	9,5	9,9	7,6
24	1,3	13,8	4,5	2,3	12,2
25	16,8	6,6	3,8	1,8	7,6
26	11,3	5,4	8,5	6,3	4,4
27	3,3	16,5	4,8	3,8	3,7
28	5,8	1,3	0,7	1,0	1,6
29	13,0	13,9	8,2	6,5	8,2
30	4,0	7,6	3,2	2,5	1,4
31	9,3	12,5	2,5	3,1	3,1
32	16,3	2,0	12,4	1,2	7,9
33	5,3	9,9	6,0	3,2	4,3
34	13,3	11,9	5,2	5,8	5,2
35	11,3	9,9	9,5	6,9	8,1
36	12,3	13,2	9,4	6,8	9,1
37	15,7	13,7	13,2	1,8	12,9
38	8,5	7,1	6,1	5,6	5,2
39	12,5	4,9	6,1	5,7	9,2
40	12,0	7,7	5,9	5,7	7,0
41	15,7	16,2	14,5	1,1	13,1
42	12,0	1,2	5,8	6,3	6,5
43	1,0	8,6	3,6	5,4	7,8
44	16,3	9,9	3,9	5,0	7,3
45	14,7	2,0	7,3	5,8	9,6
46	15,0	14,0	11,5	8,5	7,0
47	11,3	11,4	5,4	2,8	5,5
48	3,0	4,5	0,2	3,5	1,7
49	11,0	6,6	5,3	5,7	1,1
50	9,0	6,9	0,8	3,0	4,1
51	14,0	1,7	1,5	4,1	8,7
52	14,0	1,4	5,3	6,4	8,2
53	15,8	11,8	6,0	5,4	7,5
54	9,7	14,9	11,5	5,6	5,9
55	16,0	15,4	14,3	8,5	7,5
56	7,7	6,9	1,5	4,3	8,3
57	15,0	11,7	12,6	9,6	2,0
58	13,3	8,3	6,4	4,6	5,7
59	8,0	8,8	4,6	5,5	8,2
60	16,7	11,3	1,3	4,2	7,9
61	11,0	11,9	8,1	5,5	6,9
62	1,0	9,5	2,9	4,8	4,7
63	8,7	6,8	4,4	4,9	4,3
64	11,0	6,8	2,2	0,2	1,8
65	1,0	3,6	7,2	5,0	5,7
66	13,3	1,0	8,8	5,8	7,0
67	6,3	9,5	5,5	2,8	1,2
68	18,7	12,7	13,0	6,2	5,5
69	16,3	11,2	9,4	5,8	5,7
70	9,0	1,7	4,9	4,7	6,7
71	15,3	8,8	7,3	6,9	9,3
72	12,3	5,8	5,3	3,3	2,9
73	7,0	8,4	8,2	7,2	8,5
74	11,7	13,7	9,0	6,7	7,6
75	15,0	8,4	5,2	4,1	5,7
76	8,0	1,7	0,5	3,0	2,2
77	12,0	9,5	6,6	7,3	8,8
78	1,7	5,9	5,1	2,5	7,7
79	12,3	12,2	11,0	9,2	7,9
80	11,3	11,6	2,5	1,9	1,4
81	15,7	8,6	2,8	4,2	3,5
82	1,5	13,8	1,9	1,8	6,9
83	3,3	5,7	1,8	9,2	1,5
84	12,7	7,2	3,4	4,5	5,6
85	12,3	14,6	7,6	8,8	6,4
86	16,0	12,4	1,3	10,0	8,7
87	13,0	18,1	8,5	5,9	8,0
88	7,5	7,8	7,1	7,8	7,6
89	9,0	12,6	3,9	4,3	5,6

Table 2: Statistical summary of the available marks and correlation matrix.

	Eng	Com	Alg	Ana	Phy
Minimum	1.0	1.0	0.2	0.2	1.1
Maximum	18.6	18.1	14.5	10.0	13.1
Average	10.6	8.6	6.1	5.1	6.1
Std. dev.	4.5	4.4	3.4	2.2	2.8

Com	Alg	Ana	Phy	
0.15	0.32	0.17	0,18	Eng
	0.30	0.12	0.15	Com
		0.46	0.35	Alg
			0.15	Ana

or not. The summary statistics given in Table 2 show that the marks in Mathematics and Physics are much lower on average than the marks in the other subjects. The rather large number of students suggests to consider that Mathematics and Physics are evaluated much more roughly then the other subjects and thus that the partial evaluations are not commensurate. In order to solve this problem, we state the following hypothesis : the 89 considered students form a representative sample of the student population. Under this hypothesis, it seems reasonable to consider that the available sample contains both very good and very bad students. We then suggest to linearly transform the available data such that, for each subject, the lowest mark be 0 and the highest 1. Although this may not be completely satisfactory, we shall assume in the sequel that the resulting partial evaluations are commensurate.

5.2 Estimation of the weights of the subsets of attributes by means of the Shannon entropy

Recall that the weights of the subsets of attributes are defined by means of Eq. (10). We first consider the capacity obtained from the Shannon entropy. We then know that the resulting set function μ_S is a submodular capacity.

In order to be able to estimate μ_S, here, it is necessary to first discretize the available profiles. Given the rather low number of profiles w.r.t the dimension of the problem, we decide to first divide the domain of each attribute, i.e. the interval $[0, 1]$, into $d = 6$ classes : $[0, 1/6[$, $[1/6, 2/6[$, $[2/6, 3/6[$, $[3/6, 4/6[$, $[4/6, 5/6[$, and $[5/6, 1]$. This is equivalent to considering that the associated discrete random variables can take only six different values. The influence of parameter d on the estimation of the weights and on the Choquet integral will be studied in § 5.5.

Estimations of the weights of subsets can then be obtained using Eq. (12). Notice that, because of the way $\hat{\mu}_S$ was defined, the weight of a nonempty subset of subjects directly depends on the uniformity of the distribution of the marks for these subjects. In order to explain this point in more detail, consider the case of a subset reduced to a single subject. If most of the students have a similar mark for the considered subject, the weight of the subject will be low, which could be justified by the fact that it does not clearly discriminate between good and bad students. On the contrary, the more the marks are uniformly distributed, the higher the weight of the subject. The same reasoning can be applied to subsets containing more than one subject.

The estimated weight of each subject is given in Table 3. As one could have expected

from the submodularity of $\hat{\mu}_S$, the sum of the weights is (much) higher than 1, which indicates redundancy among subjects [24, 25].

Table 3: Estimated weights of the subjects.

Eng	Com	Alg	Ana	Phy
0.38	0.38	0.38	0.38	0.36

5.3 Behavioral analysis of the Choquet integral with respect $\hat{\mu}_S$

In order to study the behavior of the Choquet integral with respect $\hat{\mu}_S$, the Shapley importance index of each subject was computed (cf. § 2.3). The indices are given in Table 4. As one can notice, all the subjects have approximately the same global importance.

Table 4: Shapley importance indices and Shapley interaction indices between subjects.

Eng	Com	Alg	Ana	Phy
0.21	0.21	0.19	0.19	0.19

Com	Alg	Ana	Phy	
-0.10	-0.08	-0.08	-0.08	Eng
	-0.09	-0.07	-0.09	Com
		-0.12	-0.07	Alg
			-0.09	Ana

The average interactions between subjects can be evaluated by computing their Shapley interaction indices (cf. § 2.3). These indices are given in Table 4. Again, as one could have expected from the submodularity of $\hat{\mu}_S$, the interaction indices are all negative [15]. By considering Table 4, one can see that the two subjects that interact the most (negatively) are Analysis and Algebra. This redundancy effect implies that a high (resp. low) mark in Analysis is usually followed by a high (resp. low) mark in Algebra and *vice versa*.

Finally, the behavior of the Choquet integral w.r.t the capacity $\hat{\mu}_S$ can also be interpreted by means of the Marichal entropy of $\hat{\mu}_S$. As discussed in [29], the quantity $H_M(\hat{\mu}_S)$ can be seen as a measure of the average degree of utilization of a profile during the aggregation. The higher $H_M(\hat{\mu}_S)$, the closer the behavior of the Choquet integral to that of the simple arithmetic mean. On the contrary, the more *disjunctive* or *conjunctive* the Choquet integral is (i.e. close to the maximum or minimum resp.), the lower $H_M(\hat{\mu}_S)$.

In order to have an index in $[0, 1]$, H_M can be simply normalized by division by its maximum ($\ln 5$ for the considered problem). We obtain $H_M(\hat{\mu}_S)/\ln 5 = 0.84$, which could be considered as satisfying. More details can be found in [35].

5.4 Estimation of the global evaluations

Now that the weights of the nonempty subsets of attributes are estimated, the global evaluations of the students can be computed by means of the Choquet integral w.r.t $\hat{\mu}_S$. These

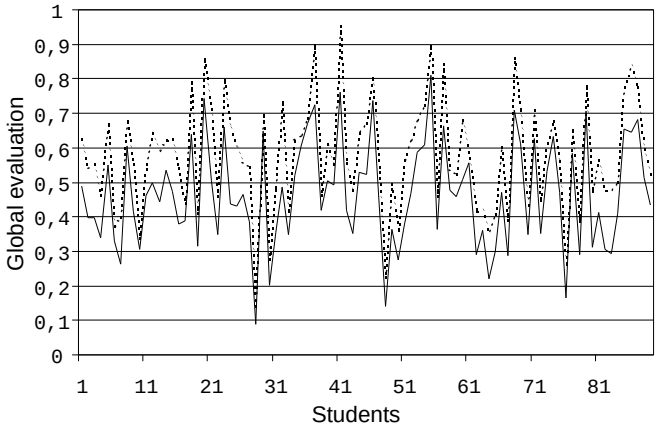

Figure 1: Global evaluations computed by the Choquet integral w.r.t $\hat{\mu}_S$ (dashed line) and the simple arithmetic mean (continuous line).

global evaluations are given in Figure 1 (dashed line). The continuous line corresponds to the global evaluations obtained by the simple arithmetic mean.

By considering Figure 1, one can notice that the global evaluations computed by the Choquet integral w.r.t $\hat{\mu}_S$ are always superior to the simple arithmetic mean of the marks. This *disjunctive* behavior of the Choquet integral is due to the strong redundancy among subjects modeled by $\hat{\mu}_S$.

In order to study the effects of the negative interaction phenomena among attributes modeled by $\hat{\mu}_S$, we compare the profile of the student designated best by the simple arithmetic mean to that designated best by the Choquet integral w.r.t $\hat{\mu}_S$. By observing Figure 1, it appears that, in the first case, the best student is student number 55 in Table 1 and that, in the second case, the best student is student number 41. In the sequel, the former will be called m, the latter c. Their profiles are given in Table 5.

Table 5: Profile of the student designated best by the simple arithmetic mean (m) and profile of the student designated best by the Choquet integral (c).

	Eng	Com	Alg	Ana	Phy
m	0.85	0.84	0.98	0.85	0.53
c	0.83	0.88	1.0	0.09	1.0

By observing Table 5, one can see that student m has good results *on average* but that the marks of student c are globally superior, except in Analysis where her/his mark is extremely low. The fact that student c is designated better than student m by the Choquet integral can be explained by c's high mark in Algebra and the disjunctive behavior of the Choquet integral due, among other things, to the strong negative interaction between Analysis and Algebra. In other terms, a high mark in Algebra or in Analysis is sufficient to significantly influence the global evaluation. In more academic terms, the extremely low mark of c in Analysis is interpreted as an "accident" in comparison to c's other marks and especially her/his mark in Algebra.

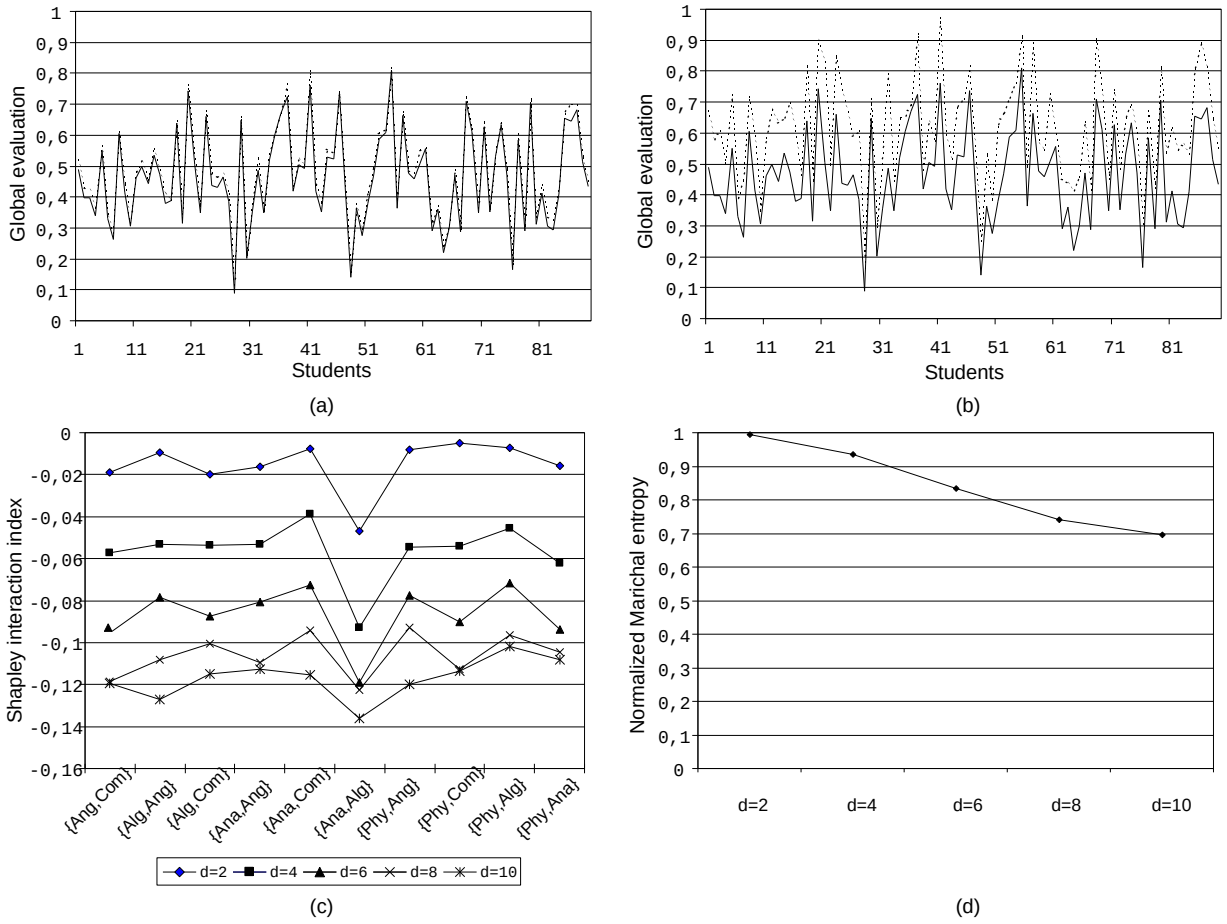

Figure 2: Influence of parameter d on the aggregation by the Choquet integral w.r.t $\hat{\mu}_S$.

To conclude this subsection, we could say that, globally, the simple arithmetic mean tends to underestimate the students since it does not take in account the redundancy effects among subjects.

5.5 Influence of parameter d on the aggregation

Before considering other choices for μ, we synthetically present the results obtained for other values of the discretization parameter d. Recall that d corresponds to the number of subdivisions of $[0, 1]$. We have therefore estimated the capacity μ_S and computed the global evaluations of the 89 students for $d = 2$, 4, 6, 8 and 10. The obtained results show that the larger d, the more disjunctive the Choquet integral. In order to illustrate this behavior, the global evaluations computed by the Choquet integral w.r.t $\hat{\mu}_S$ for $d = 2$ and $d = 10$ are compared with the simple arithmetic mean of the marks. In Figure 2 (a), the global evaluations computed by the Choquet integral for $d = 2$ (dashed line) are compared with the simple arithmetic mean of the marks (continuous line). As one can notice, the two curves are almost superimposed. To our opinion, this is due to the fact that the low value of parameter d does not enable to fully highlight correlations among subjects. Figure 2 (b) shows a similar comparison for $d = 10$. In this case, the redundancy effects among subjects

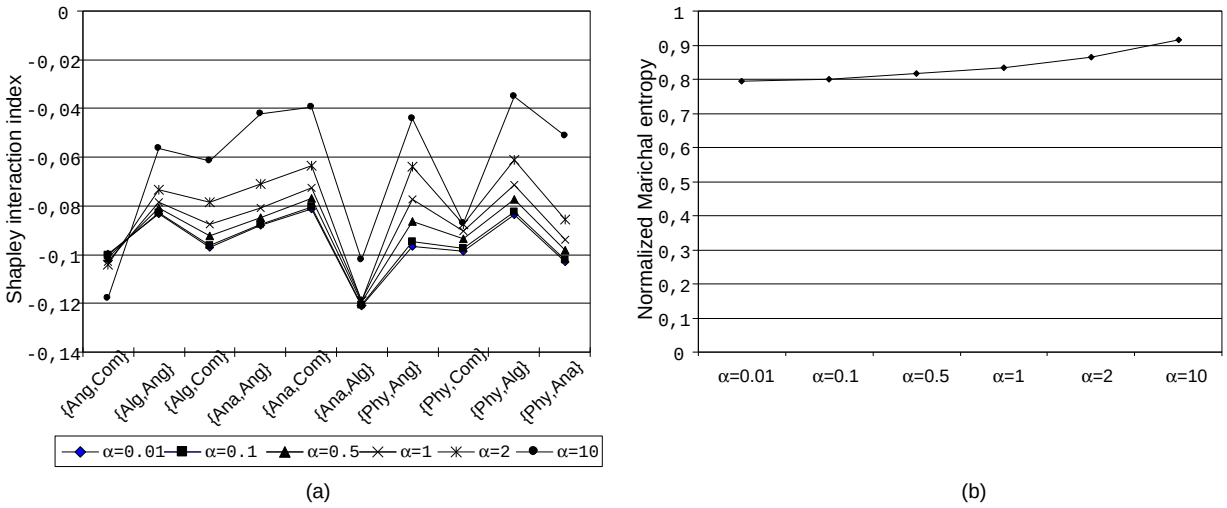

Figure 3: Influence of parameter α on the aggregation by the Choquet integral w.r.t $\hat{\mu}_R^\alpha$.

seem to have been taken into account since the Choquet integral w.r.t $\hat{\mu}_S$ shows a highly disjunctive behavior. This observation is strengthen by the evolution of the values of the Shapley interactions indices between subjects and of the Marichal entropy of $\hat{\mu}_S$ against parameter d, as can be noticed from Figures 2 (c) and 2 (d) respectively. Indeed, the higher d, the higher the redundancy between subjects and the lower $H_M(\hat{\mu}_S)$. However, the higher d, the larger the sample size necessary to obtain accurate estimates of $p_{(X_1,\ldots,X_m)}$. Indeed, as all multivariate statistical methods, the proposed approach suffers from the so-called *curse of dimensionality*. Hence, spurious redundancy effects could appear as d increases. This phenomena is illustrated in Figure 2 (c) : the higher d, the higher and more homogeneous the Shapley interaction indices.

5.5.1 Estimation of the weights and of the global evaluations by means of the Rényi entropy

As second study, we perform the computation of the global scores of the students by means the Choquet integral w.r.t μ_R^α. Contrarily to μ_S, we know from § 3.4 that μ_R^α is not necessarily submodular. However, the additivity of μ_R^α still coincides with the stochastic mutual independence of the random variables associated with the attributes.

In order to empirically study the influence of parameter α on the aggregation, we estimated μ_R^α and computed the global evaluations of the 89 students by means of the Choquet integral w.r.t $\hat{\mu}_R^\alpha$ for $d = 6$ and for $\alpha = 0.01$, 0.1, 0.5, 1, 2 and 10. As we can see from Figures 3 (a) and 3 (b), the values of the interaction indices between subjects and of the Marichal entropy of $\hat{\mu}_R^\alpha$ tend to (slowly) increase with α.

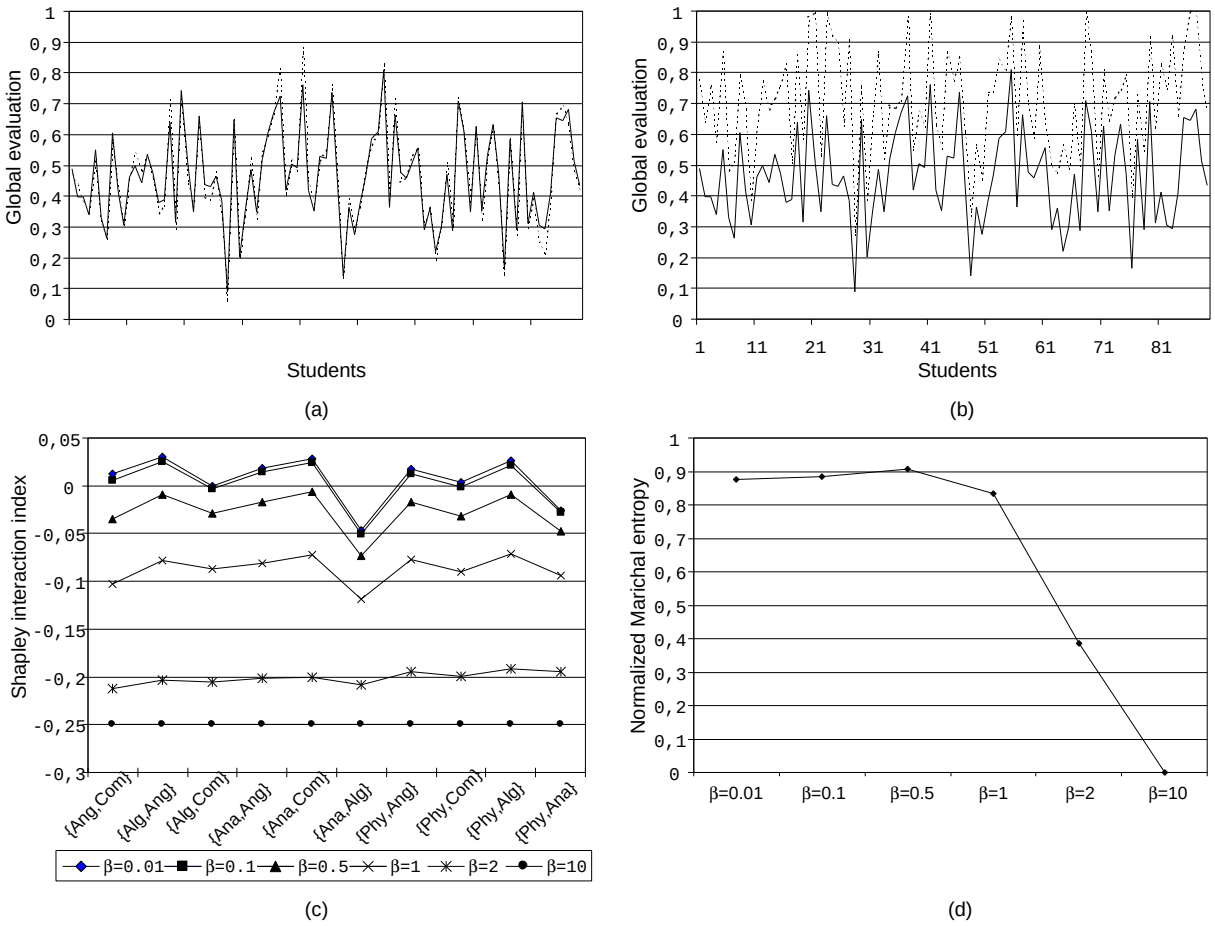

Figure 4: Influence of parameter β on the aggregation by the Choquet integral w.r.t $\hat{\mu}^{\beta}_{HC}$.

5.5.2 Estimation of the weights and of the global evaluations by means of the Havrda and Charvat entropy

Before ending this section, we perform the computation of the global evaluations by means the Choquet integral with to μ^{β}_{HC}. From § 3.4, we know that μ^{β}_{HC} is not necessarily submodular and that its additivity does not necessarily coincide with the stochastic mutual independence of the random variables associated with the attributes.

As previously, in order to empirically study the influence of parameter β on the aggregation, we estimated μ^{β}_{HC} and computed the global evaluations of the 89 students for $d = 6$ and for $\beta = 0.01$, 0.1, 0.5, 1, 2 and 10. The obtained results show that, globally, the higher β, the more disjunctive the Choquet integral. In order to illustrate this phenomenon, we compare the global scores obtained from the Choquet integrals w.r.t $\hat{\mu}^{0.01}_{HC}$ and $\hat{\mu}^{10}_{HC}$. In Figure 4 (a), the global evaluations obtained from the Choquet integral w.r.t $\hat{\mu}^{0.01}_{HC}$ (dashed line) are compared to the means of the partial evaluations (continuous line). As we can notice, the Choquet w.r.t $\hat{\mu}^{0.01}_{HC}$ globally behaves like the simple arithmetic mean. In Figure 4 (b), the means of the partial evaluations (continuous line) are compared with the global scores computed by the Choquet integral w.r.t $\hat{\mu}^{10}_{R}$ (dashed line). As we can see, the Choquet w.r.t $\hat{\mu}^{10}_{HC}$ exhibits a strong disjunctive behavior. A simple computation shows that the Choquet

w.r.t $\hat{\mu}_{HC}^{10}$ behaves almost like the maximum operator. It therefore globally seems that the higher β, the higher the global scores of the students. This more and more disjunctive behavior of the Choquet integral is empirically confirmed, as we can see from Figures 4 (c) and 4 (d), by the evolution of the interactions indices and the normalized Marichal entropy w.r.t β. Indeed, globally, the higher β, the stronger the redundancy among subjects, and the lower $H_M(\hat{\mu}_S)$. Finally, notice from Figure 4 (c), that for low values of β, some interaction indices between subjects are positive, which corresponds to *complementarity* effects between attributes.

6 Conclusion

The unsupervised Choquet integral based aggregation method initially suggested in [26] has been presented from both a theoretical and a practical perspective, and improvements have been proposed. In the absence of initial preferences, the suggested methodology could be considered as more insightful than arbitrary parametrized weighted arithmetic mean based approaches that cannot not take redundancy effects among attributes into account. From a practical perspective, the proposed methodology might be useful in several fields where information fusion is necessary [9] such as sensor information fusion or aggregation of experts' points of view. Proceeding like Marichal *et al.* [37], it could also be used in an ordinal context. The use of parametric entropies, such as the Rényi or the Havrda and Charvat entropy, could also be of interest in order to control the amount of disjunctive behavior in the aggregation process. From a practical perspective, a sufficiently large number of profiles is necessary to obtain accurate estimates of the capacity coefficients and therefore of the Choquet integral. Furthermore, in the case of continuous attributes, a strategy for the choice of the discretization parameter d would need to be investigated.

References

[1] N. Abramson. *Information Theory and Coding*. McGraw Hill, New-York, 1963.

[2] J. Aczél and Z. Daróczy. Charakterisierung der entropien positiver ordung und der shannonschen entropie. *Acta Mathematica Academiae Scientiarium Hungaricae*, 14:95–121, 1963.

[3] J. Aczél and Z. Daróczy. *On measures of information and their characterizations*. Academic Press, New York–San Francisco–London, 1975.

[4] A. Agresti. *Categorical Data Analysis*. Wiley, 2002. Second edition.

[5] A. Chateauneuf and J.-Y. Jaffray. Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Math. Social Sci.*, 17(3):263–283, 1989.

[6] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1953.

[7] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, 1991.

[8] M. H. DeGroot. Uncertainty, information and sequential experiments. *Ann. Math. Statist.*, 33:404–419, 1962.

[9] D. Dubois and H. Prade. On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, 142(1):143–161, 2004.

[10] B. Ebanks, P. Sahoo, and W. Sander. *Characterizations of information measures.* World Scientific, Singapore, 1997.

[11] M. Esteban and D. Morales. A summary on entropy statistics. *Kybernetika*, 31(4):337–346, 1995.

[12] D. Fadeev. Zum begriff der entropie einer endlichen wahrscheinlichkeitschemes. *Arbeit zur Informationstheorie, Deutscher Verlag der Wissenschaften*, 1, 1957.

[13] D. Filev and R. Yager. On the issue of obtaining OWA operator weights. *Fuzzy Sets and Systems*, 94:157–169, 1998.

[14] M. Grabisch. The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89:445–456, 1992.

[15] M. Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2):167–189, 1997.

[16] M. Grabisch and I. Kojadinovic. `kappalab`*: Non additive measure and integral manipulation functions*, 2005. R package version 0.1-1.

[17] M. Grabisch, J.-L. Marichal, and M. Roubens. Equivalent representations of set functions. *Math. Oper. Res.*, 25(2):157–178, 2000.

[18] M. Grabisch, H. Nguyen, and E. Walker. *Fundamentals of uncertainty calculi with applications to fuzzy inference.* Kluwer Academic, Dordrecht, 1995.

[19] M. Grabisch and M. Roubens. Application of the Choquet intergral in multicriteria decision making. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals*, pages 349–374. Physica-Verlag, 2000.

[20] J. Havrda and F. Charvat. Quantification method in classification processes: concept of structural α-entropy. *Kybernetika*, 3:30–35, 1967.

[21] E. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

[22] H. Joe. Relative entropy measures of multivariate dependence. *J. Am. Statist. Assoc.*, 84:157–164, 1989.

[23] A. Khinchin. *Mathematical foundations of information theory.* Dover, 1957.

[24] I. Kojadinovic. *Modeling interaction phenomena using non additive measures : applications in data analysis.* PhD thesis, Université de La Réunion, France, 2002.

[25] I. Kojadinovic. Modeling interaction phenomena using fuzzy measures: on the notions of interation and independence. *Fuzzy Sets and Systems*, 135(3):317–340, 2003.

[26] I. Kojadinovic. Estimation of the weights of interacting criteria from the set of profiles by means of information-theoretic functionals. *European Journal of Operational Research*, 155:741–751, 2004.

[27] I. Kojadinovic. Unsupervised aggregation by the discrete Choquet integral based on entropy functionals : application to the evaluation of students. In *Modeling Decisions for Artificial Intelligence (MDAI 2004), Lecture Notes in Artificial Intelligence (LNAI 3131)*, pages 163–174, Barcelona, Spain, 2004. Springer-Verlag.

[28] I. Kojadinovic, J.-L. Marichal, and M. Roubens. An axiomatic approach to the definition of the entropy of a discrete Choquet capacity. In *9th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2002)*, pages 763–768, Annecy, France, July 2002.

[29] I. Kojadinovic, J.-L. Marichal, and M. Roubens. An axiomatic approach to the definition of the entropy of a discrete Choquet capacity. *Information Sciences*, 2005. In press.

[30] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

[31] V. Kus. *Divergences and Generalized Score Functions in Statistical Inference*. PhD thesis, Czech Technical University, Prague, Czech Republic, 1999.

[32] C. Labreuche and M. Grabisch. The Choquet integral for the aggregation of interval scales in multicriteria decision making. *Fuzzy Sets and Systems*, 137:11–16, 2003.

[33] T. Marchant. Towards a theory of MCDM : stepping away from social choice theory. *Mathematical Social Sciences*, 45:343–363, 2003.

[34] J.-L. Marichal. An axiomatic approach of the discrete Choquet integral as a tool to aggregate interacting criteria. *IEEE Transactions on Fuzzy Systems*, 8(6):800–807, 2000.

[35] J.-L. Marichal. Behavioral analysis of aggregation in multicriteria decision aid. In J. Fodor, B. D. Baets, and P. Perny, editors, *Preferences and Decisions under Incomplete Knowledge*, pages 153–178. Physica-Verlag, 2000.

[36] J.-L. Marichal. Entropy of discrete Choquet capacities. *European Journal of Operational Research*, 3(137):612–624, 2002.

[37] J.-L. Marichal, P. Meyer, and M. Roubens. Sorting multi-attribute alternatives : the TOMASO method. *Computers and Operations Research*, 32(4):861–877, 2004.

[38] J.-L. Marichal and M. Roubens. Determination of weights of interacting criteria from a reference set. *European Journal of Operational Research*, 124:641–650, 2000.

[39] D. Morales, L. Pardo, and I. Vajda. Uncertainty of discrete stochastic systems: general theory and statistical theory. *IEEE Trans. on System, Man and Cybernetics*, 26(11):1–17, 1996.

[40] T. Murofushi and S. Soneda. Techniques for reading fuzzy measures (iii): interaction index. In *9th Fuzzy System Symposium*, pages 693–696, Saporo, Japan, 1993.

[41] T. Murofushi and M. Sugeno. Fuzzy measures and fuzzy integrals. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals: Theory and Applications*, pages 3–41. Physica-Verlag, 2000.

[42] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3.

[43] A. Rényi. On the measures of entropy and information. In *4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561. Berkeley University Press, 1961.

[44] A. Rényi. On the foundations of information theory. *Review of the International Statistical Institute*, 33(1):1–14, 1965.

[45] G.-C. Rota. On the foundations of combinatorial theory. I. Theory of Möbius functions. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 2:340–368 (1964), 1964.

[46] G. Saporta. *Probabilités, Analyse de Données et Statistique*. Editions Technip, Paris, 1990.

[47] C. Shannon and W. Weaver. *A mathematical theory of communication*. University of Illinois, Urbana, 1949.

[48] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–623, 1948.

[49] L. S. Shapley. A value for n-person games. In *Contributions to the theory of games, vol. 2*, Annals of Mathematics Studies, no. 28, pages 307–317. Princeton University Press, Princeton, N. J., 1953.

[50] M. Sugeno. *Theory of fuzzy integrals and its applications*. PhD thesis, Tokyo Institute of Technology, Tokyo, Japan, 1974.

[51] A. Tanaka and T. Murofushi. A learning model using fuzzy measures and the Choquet integral. In *5th Fuzzy System Symposium*, pages 213–218, Kobe, Japan, 1989.

[52] A. Ullah. Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference*, 49:137–162, 1996.

[53] W. Wienholt and B. Sendhoff. How to determine the redundancy of noisy chaotic time series. *International Journal of Bifurcation and Chaos*, 6(1):101–117, 1996.