Estimation of the weights of interacting criteria from the set of profiles by means of information-theoretic functionals

Ivan Kojadinovic

IREMIA, *Université de La Réunion, 15 avenue René Cassin - BP 7151, 97715 Saint-Denis messag cedex 9, Ile de La Réunion, France*

Abstract

Following the recent work of Marichal and Roubens [15] on fuzzy measure identification in multicriteria decision problems, we propose an alternative *unsupervised* identification method based on the estimation of the fuzzy measure coefficients by means of information-theoretic functionals.

Key words: Multiple criteria analysis; multicriteria decision making; Choquet integral; entropy; mutual information.

1 Introduction

Consider a finite set of *alternatives* $A = \{a_1, \ldots a_n\}$ and a finite set of *criteria* $N = \{c_1, \ldots, c_q\}$ in a multicriteria decision problem [14,15]. Each alternative $a \in A$ is associated with a *profile* $x^a = (x_1^a, \ldots, x_q^a) \in \mathbb{R}^q$, where, for all $i = 1, \ldots, q$, x_i^a is the partial score of a related to criterion c_i, with $x_i^a \in X_i \subseteq \mathbb{R}$. It is further assumed that all the partial scores are defined on the same interval scale, i.e. up to the same linear transformation [15].

To each profile can be associated a global score computed by means of an aggregation operator [13] which takes into account the weights of importance of the criteria. Based on the global scores, the alternatives can be ranked and the best alternative selected.

In presence of independent criteria, the most often used aggregation operators are the weighted arithmetic means [14]. The global score W_x^ω associated to a

Email address: Ivan.Kojadinovic@univ-reunion.fr (Ivan Kojadinovic).

profile $x = (x_1, \dots, x_q)$ is then given by

$$W_x^\omega = \sum_{i=1}^{q} \omega_i x_i$$

where, for all $i = 1, \dots, q$, $\omega_i \geq 0$ is the weight of criterion c_i and $\sum_{i=1}^{q} \omega_i = 1$.

To illustrate the different notions mentioned above, let us consider the following example taken from Marichal [14].

Example *The problem of evaluating n students $a_1, \dots, a_n$ (alternatives) with respect to three subjects (criteria): linear algebra (Al), probability (Pr) and statistics (St) is a multicriteria decision problem. The following table gives the marks of the students expressed on a scale from 0 to 20.*

student	*Al*	*Pr*	*St*	*global score*
a_1	*14*	*19*	*16*	*?*
a_2	*11*	*16*	*15*	*?*
$\vdots$		$\vdots$		$\vdots$
a_n	*19*	*19*	*12*	*?*

Under the assumption of independence among criteria, the calculation of the global scores by a weighted arithmetic mean obviously first requires the assignment of a weight to each criterion. This step is usually carried out by the decision maker and thus reflects her/his point of view on the multicriteria decision problem.

The assumption of independence among criteria is however rarely verified. In order to be able to model interaction phenomena among criteria, it has been recently proposed to substitute to the weight vector ω a monotonic set function μ on N allowing to model not only the importance of each criterion but also the importance of each coalition of criteria (see e.g. [6,7,14,15]). Such a monotonic set function μ, called *Choquet capacity* [3] or *fuzzy measure* [26], satisfies $\mu(\emptyset) = 0$, $\mu(N) = 1$ and $\mu(S) \leq \mu(T)$ whenever $S \subseteq T \subseteq N$.

A suitable aggregation operator that generalizes the weighted arithmetic mean when the criteria interact is then the discrete Choquet integral with respect to the fuzzy measure μ (see e.g. [7,14,18]).

Clearly, using the Choquet integral as an aggregation operator first requires the definition of the fuzzy measure μ. However, unlike for the weighted arithmetic means, it is rather unrealistic to assume that the $2^q - 2$ coefficients of a fuzzy measure on N can be provided by the decision maker especially when the number of criteria is large. In order to cope with the complexity of fuzzy measures, two *supervised* identification methods have been recently proposed [10,11,15]. The term *supervised* means that some prior knowledge on the multicriteria decision problem has to be provided by the decision maker before the methods can be applied. Obviously, the point of view of the decision maker will have a great influence on the results of the aggregation. The following questions then naturally arise: What if the complexity of the decision problem is such that the required knowledge cannot be easily given? What if the decision maker is not knowledgeable enough about the decision problem?

With these considerations in mind, we propose an *unsupervised* identification method of the fuzzy measure based on the estimation of the interaction among criteria by means of information-theoretic functionals. Our approach mainly consists in replacing the rather subjective notion of *importance* of a subset of criteria by that, probabilistic, of *information content* of a subset of criteria, which can be estimated from the set of profiles.

The proposed unsupervised identification method therefore appears as an alternative to the existing approaches developed in [10,11,15] when the prior knowledge they rely on cannot be provided. From a practical perspective, a sufficiently large number of profiles is obviously necessary to obtain accurate estimates of the fuzzy measure coefficients and therefore of the Choquet integral.

This paper is organized as follows. First, we recall the definitions of the concepts of fuzzy measure, Choquet integral, Möbius representation and interaction index in the framework of aggregation. Then, we show the analogy existing between some of these notions and those of entropy and mutual information at the root of information theory, which leads us to defining the weights of the subsets of criteria in terms of *information content*. Finally, we study the statistical properties of the natural estimate of the Choquet integral with respect to the fuzzy measure thus defined by means of information-theoretic functionals.

2 Aggregation by the Choquet integral

The Choquet integral with respect to a discrete fuzzy measure μ arises as a natural extension of the weighted arithmetic mean in the sense that it takes into account the interaction among criteria [14].

As mentioned in the introduction, interaction phenomena among criteria can be modeled by a *discrete fuzzy measure* [26].

Let $\mathcal{P}(N)$ denote the power set of N.

Definition 1 *A discrete fuzzy measure on N is a set function $\mu : \mathcal{P}(N) \to [0,1]$ satisfying the following conditions :*

(i) *$\mu(\emptyset) = 0$, $\mu(N) = 1$,*
(ii) *for all $S, T \subseteq N$, $S \subseteq T \Rightarrow \mu(S) \leq \mu(T)$.*

For each subset of criteria $S \subseteq N$, $\mu(S)$ can then be interpreted as the *weight* or the *importance* of the coalition S. The monotonicity of μ means that the weight of a subset of criteria can only increase when one adds new criteria to it.

When using a fuzzy measure to model the importance of each subset of criteria, a suitable aggregation operator that generalizes the weighted arithmetic mean is the discrete Choquet integral (see e.g. [7,13,18]) which is defined as follows.

Definition 2 *Let μ be a fuzzy measure on N. The discrete Choquet integral of a function $x : N \to \mathbb{R}$ with respect to μ is defined by*

$$C_x^\mu := \sum_{i=1}^{q} (x_{(i)} - x_{(i-1)})\mu(A_{(i)}),$$

where $(.)$ indicates a permutation such that $x_{(1)} \leq \cdots \leq x_{(q)}$ with the convention that $x_{(0)} = 0$ and where $A_{(i)} = \{c_{(i)}, \ldots, c_{(q)}\}$ for all $i = 1, \ldots, q$.

The Choquet integral, as we shall see in the next subsection, takes into account the interaction among criteria by means of the fuzzy measure μ. As soon as μ is additive, that is, as soon as the criteria are independent, the Choquet integral collapses into the weighted arithmetic mean, i.e.,

$$C_x^\mu = \sum_{i=1}^{q} \mu(\{c_i\})x_i.$$

An intuitive presentation of the Choquet integral can be found in [18]. In the framework of aggregation, its behavior is briefly illustrated in the following example.

Example (Continued) *Consider a fuzzy measure μ such that $\mu(\{Pr\}) = 0.3$, $\mu(\{St\}) = 0.4$ and $\mu(\{Pr, St\}) = 0.5$. Note that μ is not additive, i.e., that* *the criteria interact since* $\mu(\{Pr, St\}) \neq \mu(\{Pr\}) + \mu(\{Pr\})$. *The Choquet integral of* $x^{a_1} = (14, 19, 16)$ *with respect to* μ *is then given by*

$$C^{\mu}_{x^{a_1}} = 14\mu(\{Al, Pr, St\}) + (16 - 14)\mu(\{Pr, St\}) + (19 - 16)\mu(\{Pr\}) = 15.9.$$

It should be noted that an axiomatic characterization of the Choquet integral as an aggregation operator has been recently proposed by Marichal [14].

2.2 The Möbius representation of a fuzzy measure

Any fuzzy measure μ on N can uniquely be expressed in terms of its *Möbius representation* [8,14,21] by

$$\mu(T) = \sum_{S \subseteq T} a^{\mu}(S) \text{ for all } T \subseteq N,$$

where the set function $a^{\mu} : \mathcal{P}(N) \to \mathbb{R}$ is called the *Möbius transform* or *Möbius representation* of μ and is given by

$$a^{\mu}(S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} \mu(T) \text{ for all } S \subseteq N. \tag{1}$$

In terms of the Möbius representation of μ, Chateauneuf and Jaffray [2] showed that, for all $x \in \mathbb{R}^q$, the Choquet integral of x with respect to μ is given by

$$C^{\mu}_{x} = \sum_{T \subseteq N} a^{\mu}(T) \bigwedge_{i : c_i \in T} x_i, \tag{2}$$

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

In the previous subsection, an interpretation of the coefficients $\mu(S)$, $S \subseteq N$, as *weights* of subsets of criteria has been given. In a similar way, the coefficients $a^{\mu}(S)$, $S \subseteq N$, can be assigned a particular meaning in the framework of aggregation.

Notice from Eq. (1) that a fuzzy measure μ and its Möbius representation a^{μ} coincide on singletons. For pairs $\{c_i, c_j\} \subseteq N$, we have $a^{\mu}(\{c_i, c_j\}) = \mu(\{c_i, c_j\}) - [\mu(\{c_i\}) + \mu(\{c_j\})]$. How could we interpret this difference between the importance of the pair $\{c_i, c_j\}$ and the sum of the importance of criteria c_i and c_j?

Following Marichal [14], it seems natural that if there is no interaction between c_i and c_j, the importance of the pair $\{c_i, c_j\}$ be equal to the sum of the importance of c_i and c_j i.e. $\mu(\{c_i, c_j\}) = \mu(\{c_i\}) + \mu(\{c_j\})$ and therefore

$a^\mu(\{c_i, c_j\}) = 0$. If c_i and c_j interact in a *positive* or *complementary* way, the importance of the pair $\{c_i, c_j\}$ should be superior to the sum of the importance of c_i and c_j, that is $\mu(\{c_i, c_j\}) > \mu(\{c_i\}) + \mu(\{c_j\})$ and consequently $a^\mu(\{c_i, c_j\}) > 0$. In the same way, in case of *negative* interaction or *redundancy* between c_i and c_j, we should have $\mu(\{c_i, c_j\}) < \mu(\{c_i\}) + \mu(\{c_j\})$ i.e. $a^\mu(\{c_i, c_j\}) < 0$.

Example (Continued) *Because of the relationship existing between the subjects of probability and statistics, it is highly likely that students good at probability be also good at statistics and vice versa. The redundancy between the two criteria can therefore be modeled by a fuzzy measure* μ *such that* $\mu(\{Pr, St\}) < \mu(\{Pr\}) + \mu(\{St\})$ *i.e.* $a^\mu(\{Pr, St\}) < 0$.

For three criteria $c_i, c_j, c_k \in N$, we have

$$a^\mu(\{c_i, c_j, c_k\}) = \mu(\{c_i, c_j, c_k\}) - \mu(\{c_i, c_j\}) - \mu(\{c_i, c_k\}) - \mu(\{c_j, c_k\}) \\ + \mu(\{c_i\}) + \mu(\{c_j\}) + \mu(\{c_k\}),$$

which we can, for instance, rewrite as

$$a^\mu(\{c_i, c_j, c_k\}) = \mu(\{c_i, c_j, c_k\}) - \mu(\{c_i, c_k\}) - \mu(\{c_j, c_k\}) + \mu(\{c_k\}) \\ - a^\mu(\{c_i, c_j\}). \quad (3)$$

Could we give an interpretation of the expression $\mu(\{c_i, c_j, c_k\}) - \mu(\{c_i, c_k\}) - \mu(\{c_j, c_k\}) + \mu(\{c_k\})$ in the spirit of that of the Möbius coefficient $a^\mu(\{c_i, c_j\})$?

Note that, for any $T \subseteq N$, for any $S \subseteq N \setminus T$, the difference $\mu(T \cup S) - \mu(T)$ can be interpreted as the *marginal contribution* of subset S of criteria to subset T or as the *marginal importance* of S in the presence of T [14,25].

Thus, expression $\mu(\{c_i, c_j, c_k\}) - \mu(\{c_i, c_k\}) - \mu(\{c_j, c_k\}) + \mu(\{c_k\})$ can be regarded as the difference between the marginal contribution of c_j to the pair $\{c_i, c_k\}$ and the marginal contribution of c_j to criterion c_k alone. Having for instance $\mu(\{c_i, c_j, c_k\}) - \mu(\{c_i, c_k\}) > \mu(\{c_j, c_k\}) - \mu(\{c_k\})$ (resp. $<$) can then be interpreted as positive (resp. negative) interaction between c_i and c_j *in the presence* of c_k. Hence, following Marichal [14], the expression $\mu(\{c_i, c_j, c_k\}) - \mu(\{c_i, c_k\}) - \mu(\{c_j, c_k\}) + \mu(\{c_k\})$ will be called the *marginal interaction between* c_i *and* c_j *in the presence of* c_k and shall be denoted $a^\mu(\{c_i, c_j\}|\{c_k\})$ by analogy with the coefficient $a^\mu(\{c_i, c_j\})$ of the Möbius representation of μ.

Going back to Eq. (3), it is easy to check that $a^\mu(\{c_i, c_j, c_k\})$ can be rewritten as

$$\begin{aligned} a^\mu(\{c_i, c_j, c_k\}) &= a^\mu(\{c_i, c_j\}|\{c_k\}) - a^\mu(\{c_i, c_j\}), \\ &= a^\mu(\{c_i, c_k\}|\{c_j\}) - a^\mu(\{c_i, c_k\}), \\ &= a^\mu(\{c_j, c_k\}|\{c_i\}) - a^\mu(\{c_j, c_k\}). \end{aligned}$$

Assume now for instance that $a^\mu(\{c_i, c_j\}|\{c_k\}) = a^\mu(\{c_i, c_j\})$ i.e. in others words that the presence of c_k does not affect the interaction between criteria c_i and c_j. From the equations above, this is equivalent to

$$\begin{cases} a^\mu(\{c_i, c_k\}|\{c_j\}) = a^\mu(\{c_i, c_k\}), \\ a^\mu(\{c_j, c_k\}|\{c_i\}) = a^\mu(\{c_j, c_k\}), \\ a^\mu(\{c_i, c_j, c_k\}) = 0. \end{cases}$$

Thus, we see that if the presence of one of the three criteria does not affect the interaction between the two others, then, by symmetry, the interaction between any pair of criteria is not affected by the presence of the remaining criterion. In this case, it seems natural to say that the three criteria *do not interact*. Similarly, if, for example, $a^\mu(\{c_i, c_j\}|\{c_k\}) > a^\mu(\{c_i, c_j\})$ (resp. $<$), we shall say that c_i, c_j, c_k interact *positively* (resp. *negatively*).

By extension, for all $S \subseteq N$, $|S| \geq 2$, the coefficient $a^\mu(S)$ can be interpreted as the *interaction* among criteria in S. Should it be positive (resp. negative), then the criteria in S interact *positively* (resp. *negatively*).

This interpretation is given even more weight by noticing that μ is additive if and only if $a^\mu(S) = 0$ for all $S \subseteq N$, $|S| \geq 2$, in which case, as mentioned before, the Choquet integral collapses into the weighted arithmetic mean.

2.3 Importance and interaction indices

Although fuzzy measures form powerful tools for modeling interaction phenomena among criteria, their behavior, and therefore that of the Choquet integral, is generally difficult to understand. For a better understanding of these interaction phenomena, it is possible to compute global indices of importance and interaction which can be seen as an *identity card* of a fuzzy measure [8,9,14,22].

The concept of *importance index*, which was initially proposed by Shapley [25] in game theory, can be used to measure the *overall importance* of a criterion

$c_i \in N$ with respect to a fuzzy measure μ. More formally, the *Shapley importance index* or *Shapley value* of a criterion c_i can be interpreted as a weighted average of the marginal contributions of c_i to coalitions T not containing c_i and is defined by

$$\mathcal{I}_{Sh}^{\mu}(\{c_i\}) := \frac{1}{q}\sum_{k=0}^{q-1}\frac{1}{\binom{q-1}{k}}\sum_{\substack{T\subseteq N\setminus\{c_i\}\\ |T|=k}}[\mu(T\cup\{c_i\})-\mu(T)].$$

Two basic properties of the Shapley importance index are $\sum_{i=1}^{q}\mathcal{I}_{Sh}^{\mu}(c_i) = 1$ and $\mathcal{I}_{Sh}^{\mu}(\{c_i\}) = \mu(\{c_i\})$ if μ is additive.

Following Shapley [25] and Owen [19], Murofushi and Soneda [17] have proposed the notion of *interaction index between two elements* as a way to measure the average marginal interaction between two criteria c_i and c_j. The *Shapley interaction index between two criteria c_i and c_j* is defined by

$$\begin{aligned}\mathcal{I}_{Sh}^{\mu}(\{c_i,c_j\}) := \frac{1}{q-1}\sum_{k=0}^{q-2}\frac{1}{\binom{q-2}{k}}\sum_{\substack{T\subseteq N\setminus\{c_i,c_j\}\\ |T|=k}}&[\mu(T\cup\{c_i,c_j\})\\ &-\mu(T\cup\{c_i\})-\mu(T\cup\{c_j\})+\mu(T)].\end{aligned}$$

These notions have been recently further generalized by Grabisch [8] and Roubens [22]. Thus, the Shapley interaction index among criteria in $S \subseteq N$ is defined by

$$\mathcal{I}_{Sh}^{\mu}(S) := \sum_{T\subseteq N\setminus S}\frac{(q-|T|-|S|)!|T|!}{(q-|S|+1)!}\sum_{L\subseteq S}(-1)^{|S|-|L|}\mu(L\cup T).$$

The set function $\mathcal{I}_{Sh}^{\mu} : \mathcal{P}(N) \to \mathbb{R}$ is called the *Shapley interaction representation* of the fuzzy measure μ [8,9].

3 Definition of the weights of interacting criteria using information-theoretic functionals

Before being able to use the Choquet integral as an aggregation operator in a multicriteria decision problem, it is obviously necessary to identify the *weights* of all subsets of criteria, or equivalently, the *interactions* among elements of all subsets of criteria. In other words, the problem is to identify the coefficients of a fuzzy measure μ on N or the coefficients of its Möbius representation a^{μ}.

However, unlike for weighted arithmetic means, it is rather unrealistic to assume that the 2^q coefficients of a fuzzy measure μ or of its Möbius representation a^μ can be provided by the decision maker especially if the number of criteria is large.

In order to cope with the exponentially increasing complexity of fuzzy measures, two *supervised* identification methods have been recently proposed. The first approach assumes that profiles along with their global scores are given by the decision maker [10]. The fuzzy measure coefficients are then obtained by solving a quadratic program. The second approach, proposed by Marichal and Roubens [15], uses knowledge in the form of a partial preorder over a reference set of alternatives with semantical considerations. The two approaches have been compared in [11].

3.1 A probabilistic view of the identification problem

Unlike the two *supervised* identification methods mentioned above, we aim at identifying μ or a^μ in an *unsupervised* way, that is only from the set $\{x^a, a \in A\}$ of profiles whose global scores are to be computed. Indeed, such an approach could be an interesting alternative for the decision maker when the prior knowledge required by the supervised methods proposed in [10,11,15] cannot be easily given.

The identification problem thus clearly appears as an *estimation* problem, which naturally leads us to adopting a probabilistic point of view. Hence, in the rest of the paper, we assume that each criterion $c_i \in N$ is a *random variable* taking its values in the set X_i. It follows that, for each alternative $a \in A$, the partial score x_i^a can be interpreted as a realization of the random variable c_i, or equivalently, that each profile x^a, $a \in A$, can be seen as a realization of the random vector $\vec{N} = (c_1, \ldots, c_q)$.

The fundamental assumption made in the introduction which states that all the partial scores are defined up to the same linear transformation [14] enables us to consider, without loss of generality, that, for all $i = 1, \ldots, q$, $X_i = X \subseteq \mathbb{R}$.

Furthermore, from a more practical perspective, in most aggregation problems criteria usually only take a finite number of values. Thus, for all $i = 1, \ldots, q$, c_i can be considered as a *discrete* random variable taking its values in the *finite* set X. Should the criteria happen to have a continuous nature, they can be straightforwardly transformed into discrete random variables by *discretizing* the real interval $[\min X, \max X]$.

3.2 Measuring interaction among random variables

As we have seen in Subsection 2.2, for all $S \subseteq N$, $|S| \geq 2$, the coefficient $a^{\mu}(S)$ of the Möbius representation of a fuzzy measure μ can be interpreted as a measure of interaction among the elements in S. One way of tackling our identification problem is therefore to try to answer the following question: How can we measure the interaction among several discrete random variables?

In the case of two random variables, many measures of *dependence* have been proposed in the literature (see e.g. [12,23]). Note that the term *dependence* traduces the fact that two random variables can interact only in a *negative* or *redundant* way.

For more than two random variables, the only notion measuring interaction in the spirit of the Möbius representation, to our knowledge, is that of *mutual information* [1,4].

3.3 Entropy and mutual information

The notions of *entropy* and *mutual information* are at the root of information theory [1,4,24] and arise as natural tools for measuring the *information content* of a random variable and the *information common to two random variables* respectively.

The *entropy* can be equivalently interpreted as a measure of the *uncertainty* contained in a discrete probability distribution [4,24] and is defined as follows.

Definition 3 *Let p be a discrete probability distribution on a set Θ. The (Shannon) entropy of p is defined by*

$$H(p) := -\sum_{\theta\in\Theta} p(\theta)\ln p(\theta).$$

The quantity $H(p)$ can also be seen as measure of the *uniformity* of the discrete probability distribution p. Indeed, it is always non-negative, it is zero if and only if p is a Dirac mass and it reaches its maximum value if and only if p is uniform. Furthermore, if a discrete probability distribution p_1 is obtained from another discrete probability distribution p_2 by combining masses, then $H(p_1) \leq H(p_2)$. Other properties of the entropy of a discrete probability distribution can be found in [4].

By convention, the entropy of a random vector is defined as the entropy of its probability distribution [4]. More formally, the entropy of a random vector

$(x_1, \ldots, x_r)$ with probability distribution p is given by

$$H(x_1, \ldots, x_r) := H(p).$$

It is easy to check that the $r!$ random vectors composed of the random variables $x_1, \ldots, x_r$ have the same entropy. Hence, in the sequel, we define the *entropy of the set* $\{x_1, \ldots, x_r\}$ as the entropy of any of the $r!$ random vectors composed of all the random variables $x_1, \ldots, x_r$ and we denote it by $h(\{x_1, \ldots, x_r\})$.

Following [4], the conditional entropy of a set of random variables V_1 given a set of random variables V_2 is then defined by

$$h(V_1|V_2) := h(V_1 \cup V_2) - h(V_2).$$

In terms of the entropy h, the *mutual information* between two discrete random variables x_1 and x_2 is given by

$$I(\{x_1, x_2\}) := h(\{x_1\}) + h(\{x_2\}) - h(\{x_1, x_2\}),$$

and can be interpreted as a measure of *general dependence* between x_1 and x_2. Indeed, it is always non-negative and it is zero if and only if x_1 and x_2 are statistically independent. Furthermore, $I(\{x_1, x_2\}) \le \min[h(\{x_1\}), h(\{x_2\})]$ with equality if and only if x_1 is a function of x_2 or x_2 is a function of x_1 [12].

As for the entropy, a conditional version of the mutual information can be straightforwardly defined by replacing entropies by conditional entropies [4]. Thus, the conditional mutual information between two discrete random variables x_1 and x_2 given a set of random variables V is simply given by

$$I(\{x_1, x_2\}|V) := h(\{x_1\}|V) + h(\{x_2\}|V) - h(\{x_1, x_2\}|V)$$

and satisfies the same properties as the non-conditional mutual information.

In [1], the mutual information among three random variables x_1, x_2 and x_3 is defined by

$$\begin{aligned} I(\{x_1, x_2, x_3\}) := {} & h(\{x_1\}) + h(\{x_2\}) + h(\{x_3\}) \\ & - h(\{x_1, x_2\}) - h(\{x_1, x_3\}) - h(\{x_2, x_3\}) + h(\{x_1, x_2, x_3\}). \end{aligned}$$

It is easy to check that it can be rewritten as

$$\begin{aligned} I(\{x_1, x_2, x_3\}) &= I(\{x_1, x_2\}) - I(\{x_1, x_2\}|\{x_3\}), \\ &= I(\{x_1, x_3\}) - I(\{x_1, x_3\}|\{x_2\}), \\ &= I(\{x_2, x_3\}) - I(\{x_2, x_3\}|\{x_1\}). \end{aligned}$$

Thus, following the same reasoning as in Subsection 2.2, we shall say that the random variables x_1, x_2 and x_3 interact if and only if $I(\{x_1, x_2, x_3\}) \neq 0$.

Unlike the (conditional) mutual information between two random variables, $I(\{x_1, x_2, x_3\})$ is not necessarily non-negative [4] which implies that three random variables can interact in a *positive* or *complementary* way.

By analogy with the mutual information among two and three random variables and with the convention that $h(\emptyset) = 0$, the mutual information among the r random variables $x_1, \ldots, x_r$ is defined by

$$I(\{x_1, \ldots, x_r\}) := \sum_{S \subseteq \{x_1, \ldots, x_r\}} (-1)^{|S|+1} h(S).$$

Note that this definition implies that h and I coincide on singletons.

By extension, the random variables $x_1, \ldots, x_r$ will be said to interact if and only if $I(\{x_1, \ldots, x_r\}) \neq 0$.

Thus, the notion of mutual information clearly appears as a measure of interaction among random variables which is in the spirit of the Möbius representation of a fuzzy measure. Furthermore, it is interesting to note that the mutual information I plays for h a role similar to that played by a^μ for μ .

3.4 Definition of the weights of interacting criteria

From the properties of the Shannon entropy, it is easy to check that h as a set function on N is always non-negative and monotone. Thus, h is a fuzzy measure on N that simply does not satisfy the boundary condition $h(N) = 1$. Furthermore, its Möbius transform is straighforwardly given by

$$a^h(S) = (-1)^{|S|+1} I(S) \text{ for all } S \subseteq N.$$

Having in mind the analogy between the notions of mutual information and Möbius representation in the aggregation framework, in a probabilistic context, its seems natural to define the weights of the subsets of criteria of N using the fuzzy measure ν given by

$$\nu(S) := \frac{h(S)}{h(N)} \text{ for all } S \subseteq N,$$

under the assumption that $h(N) \neq 0$.

The weight of a subset S of criteria is therefore defined as the quantity of *information* contained in the set S of random variables with respect to the *overall information* $h(N)$. Thus, in the considered probabilistic context, the rather subjective notion of *importance* is replaced by that of *information content*.

The trivial case $h(N) = 0$ occurs if and only if the probability distribution p^N of the random vector $\vec{N} = \{c_1, \ldots, c_q\}$ is a Dirac mass, which implies that all the realizations x^a, $a \in A$, of the random vector $\vec{N}$ would be necessarily equal. In such a case, it obviously makes no sense to calculate global scores.

In the rest of the paper, we shall therefore assume that $h(N) \neq 0$.

Since we know the expression of the Möbius transform of h and thus that of ν, we can therefore say that the criteria $c_1, \ldots, c_r$ interact *positively* (resp. *negatively*) if and only if

$$a^\nu(\{c_1, \ldots, c_r\}) = \frac{(-1)^{r+1} I(\{c_1, \ldots, c_r\})}{h(N)} > 0 \text{ (resp. } < 0).$$

Note that h and therefore ν satisfy the additional property of submodularity [8]. This is due to the fact that the marginal interaction between two criteria can be only negative, i.e, that the conditional mutual information between two random variables is necessarily non-negative [4].

4 Estimation of the Choquet integral from the set of profiles

Computing the Choquet integral of an alternative $x \in \mathbb{R}^q$ with respect to the fuzzy measure ν obviously requires the knowledge of the entropies $h(S)$, $S \subseteq N$, which, in turn, require the joint probability distribution p^N of the random variables $c_1, \ldots, c_q$ to be known.

The unknown probability distribution p^N can be easily estimated from the available profiles $\{x^a, a \in A\}$, which, as mentioned in the previous section, correspond to realizations of the random vector $\vec{N} = (c_1, \ldots, c_q)$. An estimate of the Choquet integral can then be straightforwardly obtained from the estimate of the probability distribution p^N.

In this section, based on the recent work of Morales et al. [16] on uncertainty measures, we study the statistical properties of the natural estimator of the Choquet integral of an alternative x with respect to the fuzzy measure ν.

4.1 Preliminaries

Throughout this section, we assume, without loss of generality, that $X = \{1, \dots, m\}$. To avoid trivialities, we further suppose that $m > 1$.

Having in mind that the criteria in N are considered as discrete random variables, to every subset $S = \{c_{i_1}, \dots, c_{i_r}\} \subseteq N$, we associate the discrete random vector $\vec{S} = (c_{(i_1)}, \dots, c_{(i_r)})$ where $(.)$ indicates a permutation of the indices such that $(i_1) < \cdots < (i_r)$.

Recall that the probability distribution of the random vector $\vec{N}$ has been denoted p^N. The probability of the event $\{\vec{N} = (k_1, \dots, k_q)\}$, $(k_1, \dots, k_q) \in \{1, \dots, m\}^q$, shall then be denoted $p^N_{k_1,\dots,k_q}$.

Moreover, in the sequel, the probability distribution p^N shall be equivalently represented by the m^q-dimensional vector

$$(p^N_{1,\dots,1}, \dots, p^N_{m,\dots,m})^t$$

where the symbol ^{t} denotes the vector or matrix transpose.

4.2 The Choquet integral as a function of p^N

Let $x \in \{1, \dots, m\}^q$ be an alternative whose global score is to be calculated. Without loss of generality, throughout the rest of this section we assume that x is such that $x_1 \le \cdots \le x_q$. The Choquet integral of x with respect to ν is then given by

$$C^x_\nu = \frac{\sum_{i=1}^q (x_i - x_{i-1}) h(A_i)}{h(N)},$$

where, for all $i = 1, \dots, q$, $A_i = \{c_i, \dots, c_q\}$ and with the convention that $x_0 = 0$.

The relationship between p^N and the marginal probability distribution p^{A_i} of the random vector $\vec{A_i}$ is obviously given by

$$p^{A_i}_{k_i,\dots,k_q} = \sum_{k_1,\dots,k_{i-1}} p^N_{k_1,\dots,k_q}$$

for $k_j = 1, \dots, m$, for $j = i, \dots, q$.

Thus, we can straightforwardly write $h(A_i) = H(p^{A_i})$ in terms of p^N as

$$H(p^{A_i}) = -\sum_{k_i,\dots,k_q} \left(\sum_{k_1,\dots,k_{i-1}} p^N_{k_1,\dots,k_q} \right) \ln \left(\sum_{k_1,\dots,k_{i-1}} p^N_{k_1,\dots,k_q} \right).$$

Hence, as expected, the Choquet integral of x with respect to ν can be written as a function of the probability distribution p^N. Thus, in the rest of this section, we denote it $C(p^N)$ and we study the statistical properties of its natural estimator $C(\hat{p}^N)$, where $\hat{p}^N$ is the standard maximum likelihood estimator of p^N.

4.3 Asymptotic properties of $\hat{p}^N$

Given a random sample $\vec{N}_1, \ldots, \vec{N}_n$ drawn according to the probability distribution p^N, the standard maximum likelihood estimator of p^N is defined by

$$\hat{p}^N(k_1, \ldots, k_q) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{(k_1,\ldots,k_q)\}}(\vec{N}_j) \text{ for all } (k_1, \ldots, k_q) \in \{1, \ldots, m\}^q,$$

where $\mathbf{1}_{\{(k_1,\ldots,k_q)\}}(N)$ is the indicator of the event $\{\vec{N} = (k_1, \ldots, k_q)\}$.

In order to determine the asymptotic distribution of $\hat{p}^N$, we define the m^q-dimensional random vector Y as

$$Y = (\mathbf{1}_{\{(1,\ldots,1)\}}(\vec{N}) - p^N_{1,\ldots,1}, \ldots, \mathbf{1}_{\{(m,\ldots,m)\}}(\vec{N}) - p^N_{m,\ldots,m})^t.$$

It is easy to check (see e.g [16,23]) that $E[Y] = 0$ and that $E[YY^t] = \Lambda = diag(p^N) - p^N(p^N)^t$ where $diag(p)$ denotes the diagonal $m^q \times m^q$ matrix whose diagonal elements are the coefficients $p^N_{1,\ldots,1}, \ldots, p^N_{m,\ldots,m}$.

The following proposition, which is a well-known result (see e.g [16,23]), gives the asymptotic properties of $\hat{p}^N$.

Proposition 4

$$\lim_{n\to\infty} \hat{p}^N = p^N \textit{ almost surely.}$$
$$\lim_{n\to\infty} n^{1/2}(\hat{p}^N - p^N) = N(0, \Lambda) \textit{ in distribution.}$$

Proof. Denote by Y_i the random vector corresponding to the random variable $\vec{N}_i$ in the random sample $\vec{N}_1, \ldots, \vec{N}_n$. Then, it is easy to check that

$$\hat{p}^N - p^N = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

The first relation therefore directly follows from the strong law of large numbers. The second relation is obtained by noting that $n(\hat{p}^N - p^N)$ is the sum of n

independent and identically distributed random vectors and therefore follows from the multidimensional central limit theorem. □

4.4 Asymptotic distribution of $C(\hat{p}^N)$

As we have seen in the previous subsections, the Choquet integral of x with respect to ν is a function of the probability distribution p^N. In the following, we study the asymptotic distribution of $C(\hat{p}^N)$ using the well-known *delta method* (see e.g [5,12,23]).

Recall that the support of p^N is defined by

$$\sup(p^N) = \{(k_1, \ldots, k_q) \in \{1, \ldots, m\}^q \text{ such that } p^N_{k_1,\ldots,k_q} \neq 0\}.$$

It is easy to check that, for all p^N such that $\sup(p^N) = \{1, \ldots, m\}^q$, $C(p^N)$ is differentiable and that its gradient $C'(p^N)$ is given by

$$C'(p^N) = (C'_{1,\ldots,1}(p^N), \ldots, C'_{m,\ldots,m}(p^N))^t,$$

where, for all $(k_1, \ldots, k_q) \in \{1, \ldots, m\}^q$, $C'_{k_1,\ldots,k_q}(p^N)$ is defined by

$$C'_{k_1,\ldots,k_q}(p^N) := \frac{\partial C(p^N)}{\partial p^N_{k_1,\ldots,k_q}}.$$

The partial derivate of $H(p^{A_i})$ with respect to $p^N_{k_1,\ldots,k_q}$ is straightforwardly given by

$$\frac{\partial H(p^{A_i})}{\partial p^N_{k_1,\ldots,k_q}} = \ln\left(\sum_{k_1,\ldots,k_{i-1}} p^N_{k_1,\ldots,k_q}\right) + 1,$$

from which we obtain the expression of the coefficient $C'_{k_1,\ldots,k_q}(p^N)$ of the gradient $C'(p^N)$ as

$$C'_{k_1,\ldots,k_q}(p^N) = \frac{1}{H(p^N)}\left[\left(\sum_{i=1}^{q}(x_i - x_{i-1})\frac{\partial H(p^{A_i})}{\partial p^N_{k_1,\ldots,k_q}}\right) - \frac{\partial H(p^N)}{\partial p^N_{k_1,\ldots,k_q}} C(p^N)\right],$$

where $A_i = \{c_i, \ldots, c_n\}$ and $x_0 = 0$ by convention.

Should $\sup(p^N) \neq \{1, \ldots, m\}^q$, following Morales et al. [16], we put the undefined components of the gradient, that is those corresponding to $(k_1, \ldots, k_q) \notin \sup(p^N)$, to be zero.

The asymptotic distribution of the $C(\hat{p}^N)$ is then given by the following proposition.

Proposition 5 *The estimator $C(\hat{p}^N)$ of $C(p^N)$ is consistent in the sense*

$$\lim_{n\to\infty} C(\hat{p}^N) = C(p^N) \text{ almost surely,}$$

and asymptotically normal in the sense

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

where

$$\begin{aligned}\sigma_C^2(p^N) &= C'(p^N)^t \Lambda C'(p^N),\\ &= \sum_{k_1,\ldots,k_q} C'_{k_1,\ldots,k_q}(p^N)^2 p^N_{k_1,\ldots,k_q} - \left(\sum_{k_1,\ldots,k_q} C'_{k_1,\ldots,k_q}(p^N) p^N_{k_1,\ldots,k_q}\right)^2.\end{aligned}$$

Proof. The first relation follows from the continuity of $C(p^N)$ and Proposition 4.

To prove the second relation, notice that $\sup(\hat{p}^N) = \sup(p)^N$ almost surely [16]. Hence, the Taylor expansion

$$C(\hat{p}^N) = C(p^N) + C'(p^N)^t(\hat{p}^N - p^N) + o_p(\|\hat{p}^N - p^N\|)$$

holds for every p^N on $\{1, \ldots, m\}^q$.

The desired result follows therefore from the above expansion and the Cramer-Wold theorem [23]. $\square$

4.5 Confidence interval for the Choquet integral

From the continuity of $\sigma_C(p^N)$ and from Proposition 4, we straightforwardly obtain that $\sigma_C(\hat{p}^N)$ converges towards $\sigma_C(p^N)$ almost surely. Therefore, provided $\sigma_C(p^N) \neq 0$, the statistic

$$\frac{n^{1/2}(C(\hat{p}^N) - C(p^N))}{\sigma_C(\hat{p}^N)}$$

converges in distribution to $N(0, 1)$.

Thus, if n and $\sigma_C(\hat{p}^N)$ are sufficiently large, an approximate α-level confidence interval for $C(p^N)$ is given by

$$[C(\hat{p}^N) - n^{-1/2}\sigma_C(\hat{p}^N)\Phi(1-\alpha/2), C(\hat{p}^N) + n^{-1/2}\sigma_C(\hat{p}^N)\Phi(1-\alpha/2)]$$

where $\Phi(\alpha)$ denotes the α-quantile of $N(0, 1)$.

Note that if $\sigma_C(p^N) = 0$ then $n^{1/2}(C(\hat{p}^N) - C(p^N))$ tends in distribution to the constant 0 and $C(p^N)$ is approximable by $C(\hat{p}^N)$ with a better accuracy than $n^{-1/2}$ [16].

4.6 Behavioral analysis of aggregation

As mentioned in Subsection 2.3, the notion of interaction index can help the decision maker to understand the behavior of the fuzzy measure ν and thus gain more knowledge about the decision problem.

As for the Choquet integral, a natural estimate of the interaction index $\mathcal{I}^{\nu}_{Sh}(S)$ among criteria in subset $S \subseteq N$ is straightforwardly obtained by replacing p^N by its estimate in the expression of $\mathcal{I}^{\nu}_{Sh}(S)$. Furthermore, proceeding as in the previous subsections, an approximate confidence interval for the estimate of the interaction index $\mathcal{I}^{\nu}_{Sh}(S)$ can be derived if necessary.

4.7 Testing for the independence of criteria

Using the Choquet integral as an aggregation operator instead of a weighted arithmetic mean obviously makes sense if and only if the criteria in N interact. In the considered probabilistic framework, the independence of the criteria could be tested using extensions of well-known independence tests based on divergence measures such as the well-known Pearson's χ^2-divergence [20].

5 Conclusion

In this paper, we have proposed an unsupervised method for the identification of weights of interacting criteria. Based on the analogy between the notion of Möbius representation in the aggregation framework and that of mutual information at the root of information theory, we have defined the weight of a subset of criteria in terms of its relative information content. The weight of each subset of criteria can then be easily estimated from the set of profiles whose global scores are to be computed. Provided the number of profiles is sufficiently large, an approximate confidence interval for the Choquet integral can be obtained.

The approach proposed in this paper can therefore be an interesting alternative for the decision maker when the prior knowledge required by the existing supervised methods [10,11,15] cannot be easily given.

From a practical perspective, as all multivariate statistical methods, the approach presented in this paper suffers from the so-called "curse of dimensionality". In order words, to obtain estimates of the global scores of a given accuracy, the number of profiles has to grow almost exponentially with the number of criteria. An empirical study would be necessary to explore the practical limits of the method.

References

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

[2] A. Chateauneuf and J.Y. Jaffray. Some charaterizations of lower probabilities and other monotone capacities through the use of möbius inversion. *Mathematical Social Sciences*, 17:263–283, 1989.

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

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

[5] A.C. Davison and D.V. Hinkley. *Bootsrap methods and their applications.* Cambridge University Press, 1997.

[6] M. Grabisch. Fuzzy integral in multicriteria decision making. *Fuzzy Sets and Systems*, 69:279–298, 1995.

[7] M. Grabisch. Alternative representations of discrete fuzzy measures for decison making. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 5(5):587–607, 1997.

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

[9] M. Grabisch, J.-L. Marichal, and M. Roubens. Equivalent representations of set functions. *Mathematics of Operations Research*, 25:157–178, 2000.

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

[11] 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.

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

[13] J-L. Marichal. *Aggregation operators for multicriteria decison aid.* PhD thesis, University of Liege, Liege, Belgium, 1998.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] G. Owen. Multilinear extension of games. *Management Sciences*, 18:64–79, 1972.

[20] L. Pardo. Generalized divergence measures: statistical applications. *Encyclopedia of Microcomputers*, 22, 1999. Supplement 1.

[21] G.C. Rota. On the foundation of combinatorial theory I. Theory of Möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2:340–368, 1964.

[22] M. Roubens. Interaction between criteria and definition of weights in mcda problems. In *44th meeting of the european working group "Multicriteria Aid for Decisions"*, Brussels, Belgium, october 1996.

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

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

[25] L.S. Shapley. A value for n-person game. In H.W. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games*, volume 2, pages 307–317. Princeton University Press, 1953.

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