Annals of Mathematics and Artificial Intelligence
https://doi.org/10.1007/s10472-024-09930-0

Learning preference representations based on Choquet integrals for multicriteria decision making

Margot Herin[1] · Patrice Perny[1] · Nataliya Sokolovska[2]

Accepted: 24 January 2024

Abstract
This paper concerns preference elicitation and learning of decision models in the context of multicriteria decision making. We propose an approach to learn a representation of preferences by a non-additive multiattribute utility function, namely a Choquet or bi-Choquet integral. This preference model is parameterized by one-dimensional utility functions measuring the attractiveness of consequences w.r.t. various point of views and one or two set functions (capacities) used to weight the coalitions and control the intensity of interactions among criteria, on the positive and possibly the negative sides of the utility scale. Our aim is to show how we can successively learn marginal utilities from properly chosen preference examples and then learn where the interactions matter in the overall model. We first present a preference elicitation method to learn spline representations of marginal utilities on every component of the model. Then we propose a sparse learning approach based on adaptive L_1-regularization for determining a compact Möbius representation fitted to the observed preferences. We present numerical tests to compare different regularization methods. We also show the advantages of our approach compared to basic methods that do not seek sparsity or that force sparsity a priori by requiring k-additivity.

Keywords Multicriteria decision making · Preference learning · Choquet and bi-Choquet integrals · Capacities · Möbius representations · Sparse learning

1 Introduction

Evaluation and decision making is often a matter of finding the most appropriate tradeoff between multiple and possibly conflicting criteria [1]. In the field of multicriteria decision

✉ Margot Herin
margot.herin@lip6.fr

Patrice Perny
patrice.perny@lip6.fr

Nataliya Sokolovska
nataliya.sokolovska@sorbonne-universite.fr

[1] LIP6, 4 Place Jussieu, Paris 75004, France

[2] LCQB, 4 Place Jussieu, Paris 75004, France

Published online: 07 February 2024

making, various evaluation and aggregation models have been proposed to evaluate and compare the alternatives of a decision problem [2]. These models generally use and combine objective and subjective information: on the one hand alternatives are described by consequence vectors representing their outcomes with respect to multiple points of views under consideration in the analysis of preferences. On the other hand, in order to go beyond straightforward preferences induced by Pareto dominance, more subjective preference parameters are used to model the value system of the Decision Maker (DM), e.g., the relative importance of criteria and their possible interactions in the evaluation process. Thus, a body of increasingly complex decision models is studied in decision theory to encompass an ever more sophisticated set of decision behaviors. This effort motivated by descriptive objectives comes at the cost of an additional complexity, both at the level of preference learning (fitting the parameters of the preference model to the DM value system to explain or predict her preferences) and at the recommendation level (finding an optimal alternative becomes computationally more difficult). In this paper, we address the first challenge and propose a methodology dedicated to the identification of utilities and capacities in decision models involving Choquet integrals.

The Choquet integral is a well known aggregation function used in multicriteria decision making to assign an overall score to any evaluation vector attached to an alternative [3]. It performs a kind of sophisticated weighted average where weights are defined for every subset of components. The Choquet integral is also used in machine learning to replace the linear function of variables which is commonly used in standard regression methods [4, 5]. For example, logistic regression was extended to Choquistic regression [6]. It is also used for learning to rank with the Choquet integral [7] where the data is provided with the labels which are preference degrees from an ordered categorical scale. Choquet integrals are also used to aggregate one-dimensional utility functions in order to define a non-additive multiattribute utility function. This preference model which is at the core of our paper will be referred to as the CIU model (Choquet Integral of Utilities) in the sequel.

The CIU model is based on two types of preference parameters: utility functions defining the attractiveness of consequences on every relevant criteria and a set function named *capacity*, monotonic with respect to set inclusion, assigning a weight to every subset of criteria. The Choquet integral was initially introduced in the framework of decision under uncertainty [8]; it has been generalized to be applied in multicriteria analysis [3, 9]. In this paper, we also consider the bipolar Choquet integral of utilities (bi-CIU) which is an extension of CIU using two capacities that cooperate in weighting criteria or subset of criteria; one applies to the positive part of the evaluation vector whereas the other applies to the negative part [10]. This extension inspired by Kanheman and Tversky's cumulative prospect theory (CPT) [11] allows the representation of decision behaviors that may vary depending on whether positive or negative consequences come into play. CPT was initially introduced in the context of decision making under risk and assumes that capacities are defined as monotone transforms of probability measures. The bi-CIU model under consideration here is more general than CPT by allowing any kind of (monotonic) capacity to weigh the subsets of criteria [10]. It can be used in multicriteria optimization when criteria scales and preferences are bipolar [12]. The bi-CIU model can be further generalized using bi-capacities in Choquet integrals [10] but this latter generalization is not considered here for the sake of simplicity.

Our focus on CIU and bi-CIU models is motivated by several reasons: first CIU is acknowledged as one of the most general monotone compromise aggregators since it includes various simpler decision models as special cases (e.g., additive utilities, weighted sums, OWA and WOWA aggregators [13, 14]). Therefore CIU includes a rich family of aggregation functions which provides a natural setting to study how model complexity can be fitted to the preference system we want to describe or implement. Moreover the use of possibly non-additive capacities in CIU may require the definition of 2^n weighting parameters in the worst case where n is the number of criteria under consideration (one weight per subsets of criteria). The multiplicity of these parameters obviously induces a significant gain of expressiveness. However, it also comes with an increase of model complexity and obviously raises the question of the parsimonious learning of the parameters defining the capacity. Then, considering bi-CIU is even more general and more powerful than CIU from a descriptive viewpoint. The bipolar version of the model being based on two capacities, it requires 2^{n+1} weighting parameters, beside utility functions, which raises even more crucially the need of methods to learn sparse representations of capacities. It might indeed prevent over-fitting of preference data and lead to more compact and more explainable decision models.

The definition of the CIU model requires learning the utilities and the capacity from preference information. For the determination of the utilities, the standard approaches proposed in the literature on multicriteria decision making rely on direct queries on attribute values and/or preference intensities using, e.g., the *Macbeth* method [15]. Another utility elicitation method based on the comparison of risky prospects has been proposed for the Choquet Expected Utility (CEU) model for decision-making under risk [16, 17]. Recently, the principle of the elicitation method has been exploited for the learning of the CEU utility function in the context of decision-making under uncertainty[18]. We will further extend this approach to cope with the multicriteria decision-making framework.

For the identification of capacities in CIU in multicrieria decision-making, standard methods either use a least square regression from examples of tuples labelled with their overall utility, or an ordinal regression method based on preference examples (pairwise comparisons) [9, 19], assuming the utilities have been elicitated beforehand. Other methods have been proposed to simultaneously learn utility and capacity [20–22]. More recently, an incremental elicitation method proceeding by successive reductions of the set of admissible capacities using well chosen preference queries was proposed in [23]. Also, an incremental Bayesian approach used to iteratively revise a probability density on the space of admissible capacities was proposed in [24]. In order to simplify the problem, all these contributions focus on simple instances of CIU where interactions are only allowed between pairs of criteria. Using this simplification, some direct methods simultaneously learn utilities and the capacity [20, 21], using for instance neuronal modules [22].

The question of learning where the most significant interactions take place in the general model and how a sparse representation of the capacity can be derived from preference examples is not directly addressed. For this reason, we want to study the potential of sparse learning to determine compact representations of capacities from preference data within the CIU or bi-CIU model. This problem is challenging due to the interplay of utilities and capacities in the computation of CIU values, making the learning of these two types of parameters interdependent. Another challenge comes from the fact that utilities and capacities are not directly observable and must be derived from preference statements (comparison of alternatives or possibly value judgments). In this paper, we propose an approach to learn utility functions and capacities in two successive steps: by properly selecting a first set of preference queries we learn a spline representation of the utility function for every criterion; then we learn a sparse representation of the capacities from a database of preference examples.

The paper is organized as follows: Section 2 introduces the CIU and bi-CIU models and some related concepts. In Section 3 we present an elicitation approach to learn marginal utility functions defined for each criterion. In Section 4 we propose an approach to learn sparse representation of the capacities in CIU and bi-CIU model. In Section 5, we present numerical tests to compare the performances of our preference learning approach compared to baseline methods.

2 Background on CIU and bi-CIU

We adopt the standard setting and notations for multiattribute or multicriteria decision making. Let $N = \{1, \dots, n\}$ be the set of criteria to be considered in a decision problem. Let $X = X_1 \times \dots \times X_n$ be the n-dimensional evaluation space where X_i is a bounded set of consequences. As usual in multicriteria problem, the elements of X_i are assumed to be weakly ordered by a preference relation denoted $\succsim_i$. For any $i \in N$, for any pair $(x_i, y_i) \in X_i$, $x_i \succ_i y_i$ (resp. $x_i \succsim_i y_i$) means that x_i is a better consequence than y_i (resp. better or equivalent). Within every set X_i we distinguish three reference elements denoted $-\mathbf{1}_i, \mathbf{0}_i$ and $\mathbf{1}_i$ representing the bottom level, the neutral level and the top level consequences respectively [9]. These levels must be obtained in close cooperation with the DM. The alternatives to be compared are seen as elements of X. Thus, every alternative $x \in X$ is described by its consequence vector $(x_1, \dots, x_n)$ where $x_i \in X_i$ is the consequence of x w.r.t. i, for $i = 1, \dots, n$. In this setting we consider n utility functions u_i defined on X_i and strictly increasing with preference $\succ_i$ for $i = 1, \dots, n$, such that $u_i(-\mathbf{1}_i) = -1$, $u_i(\mathbf{0}_i) = 0$ and $u_i(\mathbf{1}_i) = 1$. Utilities are used to quantify the attractiveness of consequences on a common scale [-1, 1]. Consequences above the neutral level receive a positive utility whereas consequences below the neutral level receive a negative utility.

2.1 CIU and bi-CIU models

We recall here the definition of models CIU and bi-CIU that use a Choquet integral to aggregate the utilities defined above. Let v denote a capacity defined on 2^N, i.e., a set function such that $v(\emptyset) = 0$, $v(N) = 1$ and $v(A) \leq v(B)$ for all $A, B \subseteq N$ such that $A \subseteq B$. The CIU model combines utilities u_i, $i = 1, \dots, n$ and the capacity v to define the value of any consequence vector $x = (x_1, \dots, x_n)$ by the discrete Choquet integral of the utility vector $u(x) = (u_1(x_1), \dots, u_n(x_n))$. Formally, the CIU model reads as follows:

$$f_v^u(x) = C_v(u(x)) = \sum_{i=1}^{n} \big[v(X_{(i)}) - v(X_{(i+1)})\big] u_{(i)}(x_{(i)}) \tag{1}$$

$$= \sum_{i=1}^{n} \big[u_{(i)}(x_{(i)}) - u_{(i-1)}(x_{(i-1)})\big] v(X_{(i)}) \tag{2}$$

where (.) is any permutation of N such that $u_{(i)}(x_{(i)}) \leq u_{(i+1)}(x_{(i+1)})$ and $X_{(i)} = \{j \in N : u_{(j)}(x_{(j)}) \geq u_{(i)}(x_{(i)})\}$, $i \in N$ with $u_{(0)}(x_{(0)}) = 0$ and $X_{(n+1)} = \emptyset$.

Example 1 If $N = \{1, 2, 3\}$ and $u_2(x_2) \leq u_1(x_1) \leq u_3(x_3)$ then $C_v(u(x_1, x_2, x_3)) = u_2(x_2)v(\{1, 2, 3\}) + [u_1(x_1) - u_2(x_2)]v(\{1, 3\}) + [u_3(x_3) - u_1(x_1)]v(\{3\})$ by (2). Similarly, if $u_3(x_3) \leq u_2(x_2) \leq u_1(x_1)$ then $C_v(u(x_1, x_2, x_3)) = u_3(x_3)v(\{1, 2, 3\}) + [u_2(x_2) - u_3(x_3)]v(\{1, 2\}) + [u_1(x_1) - u_2(x_2)]v(\{1\})$.

Then the preferences induced by CIU are obviously defined as follows: for any solutions $x, y \in X$, x is at least as good as y (denoted $x \succsim y$) if and only if $f_v^u(x) \geq f_v^u(y)$. Similarly, x is indifferent to y (denoted $x \sim y$) if and only if $f_v^u(x) = f_v^u(y)$. Let us recall that monotonicity of v w.r.t. set inclusion and the monotonicity of u_i w.r.t. $\succ_i$ are assumed to make sure that $f_v^u(x) \geq f_v^u(y)$ when $x_i \succsim_i y_i$ for all $i \in N$ (monotonicity of preference w.r.t. weak Pareto dominance).

Now, we consider the bi-CIU model that relies on the same utility functions than CIU but uses two capacities:

$$f_{v,w}^u(x) = C_v(u(x)^+) + C_w(-u(x)^-) \tag{3}$$

where $u(x)^+$ (resp. $u(x)^-$) is the utility vector $u(x)$ (resp. $-u(x)$) in which negative components are replaced by 0. It is well known that $C_w(z) = -C_{\bar{w}}(-z)$ for any utility vector z where $\bar{w}$ is the dual capacity of w defined by $\bar{w}(A) = 1 - w(N \setminus A)$ for all $A \subseteq N$. Therefore $f_{v,w}^u(x) = C_v(u(x)^+) - C_{\bar{w}}(u(x)^-)$. This latter formulation makes more explicit the balance between positive and negative arguments like in cumulative prospect theory. Moreover, if $v = w$, $f_{v,w}^u(x) = C_v(u(x)^+) + C_v(-u(x)^-) = C_v(u(x))$ and therefore the bi-CIU model boils down to CIU.

Example 1 (continued) If $u_2(x_2) \leq 0 \leq u_1(x_1) \leq u_3(x_3)$ then we have $f_{v,w}^u(x_1, x_2, x_3) = C_v(u_1(x_1), 0, u_3(x_3)) + C_w(0, u_2(x_2), 0)$. Equivalently $f_{v,w}^u(x_1, x_2, x_3)$ also reads as follows: $C_v(u_1(x_1), 0, u_3(x_3)) - C_{\bar{w}}(0, -u_2(x_2), 0)$.

2.2 Möbius inverse and sparsity

An alternative representation of capacities and the Choquet integral relies on the Möbius inverse of the capacity. The Möbius inverse of v is another set function m_v defined on N by: $m_v(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} v(B)$ for all $A \subseteq N$. The coefficients $m_v(A)$ are called Möbius masses, they completely characterize v. We indeed have $v(A) = \sum_{B \subseteq A} m_v(B)$. The values of m_v can be positive or negative but add up to 1 since $\sum_{B \subseteq N} m_v(B) = v(N) = 1$. It is interesting to note that CIU can be directly expressed from the Möbius inverse by [25]:

$$f_v^u(x) = \sum_{B \subseteq N} m_v(B) \min_{i \in B} \{u_i(x_i)\} \tag{4}$$

This formulation suggests that $C_v(u(x))$ might admit a compact representation when the Möbius inverse is sparse (i.e., when the vector of Möbius masses includes many zeros or small values that will not significantly impact the calculation). A frequent option used to obtain capacities having a sparse representation is to require that Möbius masses vanish for all subsets of criteria larger than a given $k < n$. In this case, the resulting capacity is said to be k-additive [26]. For instance, when the capacity is 1-additive then all Möbius masses are null except for some singletons (at least one) where they are positive due to monotonicity. In this case, (4) shows that CIU boils down to a simple additive utility function.

Considering only 2-additive capacities is a standard option to allow pairwise interactions while keeping a sparse model. One may also wish to relax 2-additivity for k-additivity ($2 < k < n$) with the aim of finding a better tradeoff between sparsity and expressivity. However reasoning about sparsity in terms of k-additivity is a drastic reduction that may significantly impact our ability to fit preference data with relevant CIU models. It may indeed happen that very sparse but still n-additive capacities are necessary to describe preference data, as shown hereafter:

Example 2 Let us consider a DM adopting an egalitarist attitude in the aggregation (focusing on the worse consequence) refined by an utilitarist criterion (using the sum of utilities) to break ties. Such a decision attitude can be obviously represented by the ϵ-min model $f_\epsilon(x) = (1 - \sum_{i=1}^n \epsilon_i) \min_{i \in N} u_i(x_i) + \sum_{i=1}^n \epsilon_i u_i(x_i)$ where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a vector of positive quantities chosen arbitrarily small. Clearly function f is an instance of CIU (see (4)) obtained for a capacity v whose Möbius inverse m_v is everywhere 0 excepted on singletons and on N ($m_v(\{i\}) = \epsilon_i$ for all $i \in N$ and $m_v(N) = 1 - \sum_{i=1}^n \epsilon_i$). We remark that due to the monotonicity w.r.t. set inclusion, function v is non-null on every subset since Möbius masses are positive and non-null on singletons. Despite the fact that v is never null, it admits a very sparse representation in terms of Möbius masses where only $n+1$ out of 2^n coefficients are non-null.

In the above Example, we remark that the most important Möbius mass is put on N, which shows that preferences induced by f_ϵ could not be properly described by any k-additive capacity with $k < n$ despite the fact that it can be closely approximated by the min model involving a single non-null Möbius mass (attached to N). This shows that new approaches are needed to find sparse representations of capacities that best fit observed preferences, regardless of k-additivity. In this paper we propose a general approach to seek sparse Möbius representations of capacities and use it to learn simple instances of the CIU or bi-CIU model that best fit the preference data. This problem will be addressed in Section 4. We first present the learning of utility functions u_i.

3 Utility elicitation

In order to elicit utility functions we use indifference statements between carefully selected alternatives to obtain useful constraints (on difference of utilities) restricting the set of admissible utility functions u_i independently of the capacity. More precisely our approach consists in adapting the tradeoff method [16, 17] initially introduced in the context of cumulative prospect theory to the case of multicriteria evaluation to learn utility functions $u_i, i = 1, \dots, n$ within CIU or bi-CIU.

Let i be any element of N. The elicitation process to derive constraints on u_i involves tradeoffs between i and another element j of N that can be freely chosen. Starting from a solution x and considering a given modification of component x_j (sufficient to break indifference), the tradeoff query consists in asking which variation of x_i would exactly compensate the variation of x_j and restore the indifference. The existence of answers exactly achieving the indifference requires a certain richness of attribute X_i (solvability assumption). This assumption is formalized by the *restricted solvability* axiom well known in mathematical psychology [27]. For any two vectors x, y in X, let (x_i, y_{-i}) denote the vector derived from y by substituting the ith component by x_i. Then restricted solvability can be stated as follows:

Definition 1 (Restricted solvability) A preference relation $\succsim$ on X satisfies restricted solvability with respect to the ith component if for any $x \in X$, $a_i, b_i \in X_i$, $t_{-i} \in X_{-i}$ with $(a_i, t_{-i}) \succsim x \succsim (b_i, t_{-i})$, there exists y_i such that $x \sim (y_i, t_{-i})$. When this holds for all $i \in I$, the binary relation is said to satisfy restricted solvability.

Restricted solvability is not always satisfied, especially in the case of discrete attributes, as shown in the following example.

Example 3 Let $X_1 = \{0, 1\}$ and $X_2 = \{0, \frac{1}{2}, 1\}$ and define $\succsim$ on $X_1 \times X_2$ by $(x_1, x_2) \succsim (y_1, y_2)$ iff $x_1 + x_2 \geq y_1 + y_2$. We have $(1, 0) \succsim (0, \frac{1}{2}) \succsim (0, 0)$ but there is no $x_1 \in X_1$ such that $(x_1, 0) \sim (0, \frac{1}{2})$. Here restricted solvability does not hold w.r.t. the first component.

We present below the elicitation of utilities in the two cases (with and without restricted solvability). First, we consider the case where solvability holds and we derive equality constraints on u_i; then we consider the case where it does not hold and we derive inequality constraints on u_i.

3.1 Utility elicitation with restricted solvability

Let us present the elicitation process to derive constraints on u_i successively below and above the neutral level $\mathbf{0}_i$. In this part we assume that restricted solvability holds.

Utility elicitation below the neutral level

For any attribute $j \in N$, let r_j, $R_j \in X_j$, $x_i \in X_i$ such that $\mathbf{0}_j \precsim_j r_j \prec_j R_j$, (i.e., $0 \leq u_j(r_j) < u_j(R_j)$), $x_i \precsim_i \mathbf{0}_i$ (i.e., $u_i(x_i) \leq 0$) and $(\mathbf{0}_i, R_j, \mathbf{0}_{-ij}) \succsim (x_i, r_j, \mathbf{0}_{-ij}) \succsim (-\mathbf{1}_i, R_j, \mathbf{0}_{-ij})$, where $(\alpha_i, \beta_j, \mathbf{0}_{-ij})$ is a vector of neutral consequences everywhere excepted on components i and j where values are α_i and β_j. We consider the following query:

$Q^-_{ij}(x_i)$: What is the consequence y_i such that $(x_i, r_j, \mathbf{0}_{-ij}) \sim (y_i, R_j, \mathbf{0}_{-ij})$?

If we consider an instance of the restricted solvability axiom (Definition 1) obtained for $a_i = \mathbf{0}_i$, $b_i = -\mathbf{1}_i$, $t_{-i} = (R_j, \mathbf{0}_{-ij})$ and $x = (x_i, r_j, \mathbf{0}_{-ij})$ one can see that an answer $y_i \in X_i$ to question $Q^-_{ij}(x_i)$ is guaranteed to exist by the restricted solvability assumption.

We couple the observed indifference with a second one associated to the answer z_i to $Q^-_{ij}(h_i)$ for some h_i element of $X_i \setminus \{x_i\}$: $(h_i, R_j, \mathbf{0}_{-ij}) \sim (z_i, r_j, \mathbf{0}_{-ij})$

Assuming $(-\mathbf{1}_i, \mathbf{0}_{-i}) \prec \mathbf{0}$, i.e., $w(N \setminus \{i\}) < 1$, these indifferences imply that $u_i(x_i) - u_i(y_i) = u_i(h_i) - u_i(z_i)$ as shown in Proposition 1 in the Appendix Section A.1.1. Then, when h_i is chosen equal to y_i, we obtain the following simplified equation:

$$u_i(x_i) - u_i(y_i) = u_i(y_i) - u_i(z_i) \tag{5}$$

Figure 1 (left) represents the two indifference statements in the plan $X_i \times X_j$, used to obtain (5).

Utility elicitation above the neutral level

The process is symmetric to the one used to elicit utilities below the neutral level. Let $r_j, R_j \in X_j$ and $x_i \in X_i$ such that $r_j \prec_j R_j \precsim_j \mathbf{0}_j$ (i.e., $u_j(r_j) < u_j(R_j) \leq 0$), $x_i \succsim_i \mathbf{0}_i$ (i.e., $u_i(x_i) \geq 0$) and $(1_i, r_j, \mathbf{0}_{-ij}) \succsim (x_i, R_j, \mathbf{0}_{-ij}) \succsim (0_i, r_j, \mathbf{0}_{-ij})$, we consider the following query:

$Q^+_{ij}(x_i)$: What is the consequence y_i such that $(x_i, R_j, \mathbf{0}_{-ij}) \sim (y_i, r_j, \mathbf{0}_{-ij})$?

Here also, the existence of y_i is due to restricted solvability. Similarly to the elicitation under the neutral level, we couple this observed indifference with a second one associated to the answer z_i to a question $Q^+_{ij}(h_i)$ for any h_i element of $X_i \setminus \{x_i\}$: $(h_i, R_j, \mathbf{0}_{-ij}) \sim (z_i, r_j, \mathbf{0}_{-ij})$. Assuming $(\mathbf{1}_i, \mathbf{0}_{-i}) \succ \mathbf{0}$, i.e., $v(\{i\}) > 0$, these indifferences imply that $u_i(y_i) -$

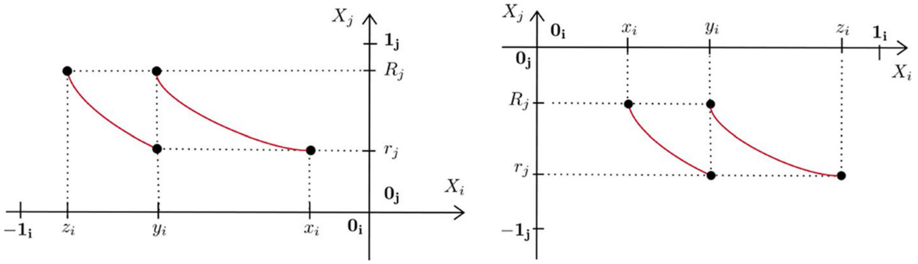

Fig. 1 Indifferences statements yielding to (5) (left) and (6) (right)

$u_i(x_i) = u_i(z_i) - u_i(h_i)$ as shown in Proposition 2 in the Appendix Section A.1.2. Then, when h_i is chosen equal to y_i, we obtain the following simplified equation:

$$u_i(y_i) - u_i(x_i) = u_i(z_i) - u_i(y_i) \tag{6}$$

Figure 1 (right) represents the two indifference statements in the plan $X_i \times X_j$, used to obtain (6).

3.2 Utility elicitation without restricted solvability

In this section we consider the case where restricted solvability w.r.t. component i does not hold, i.e., when exact answers to queries Q^+_{ij} or Q^-_{ij} do not necessarily exist. In particular, we consider the case of discrete attributes (the most common case where restricted solvability fails to hold). The elements of X_i are denoted $x_{i,k}$ and indexed according to their relative values: $x_{i,k} \precsim_i x_{i,k+1}$, for any k.

Utility elicitation below the neutral level

Let $r_j, R_j \in X_j$ and $x_i \in X_i$ such that $\mathbf{0}_j \precsim_j r_j \prec_j R_j$, i.e., $0 \leq u_j(r_j) < u_j(R_j)$ and $x_i \precsim_i \mathbf{0}_i$ i.e., $u_i(x_i) \leq 0$. We consider the two following queries:

- What is the lowest k such that $(x_i, r_j, \mathbf{0}_{-ij}) \precsim (x_{i,k+1}, R_j, \mathbf{0}_{-ij})$? Then we set $y^+_i = x_{i,k+1}$ and $y^-_i = x_{i,k}$.

Similarly, for any h_i element of $X_i \backslash \{x_i\}$ chosen such that $h_i \precsim \mathbf{0}_i$, we ask:

- What is the highest k such that $(h_i, r_j, \mathbf{0}_{-ij}) \succsim (x_{i,k}, R_j, \mathbf{0}_{-ij})$? Then we set $z^-_i = x_{i,k}$ and $z^+_i = x_{i,k+1}$.

Assuming $(-\mathbf{1}_i, \mathbf{0}_{-i}) \prec \mathbf{0}$, i.e., $w(N \setminus \{i\}) < 1$, we obtain (see Proposition 3 in Appendix Section A.2.1):

$$u_i(h_i) - u_i(z^-_i) \geq u_i(x_i) - u_i(y^+_i)$$
$$u_i(h_i) - u_i(z^+_i) < u_i(x_i) - u_i(y^-_i)$$

When h_i is chosen equal to y^+_i we obtain the following simplified inequations:

$$u_i(y^+_i) - u_i(z^-_i) \geq u_i(x_i) - u_i(y^+_i) \tag{7}$$
$$u_i(y^+_i) - u_i(z^+_i) < u_i(x_i) - u_i(y^-_i) \tag{8}$$

Then we overcome the solvability issue by deriving two inequality constraints on the utility function u_i, instead of a unique equality constraint.

Utility elicitation above the neutral level

Let $r_j, R_j \in X_j$ and $x_i \in X_i$ such that $r_j \prec_j R_j \precsim_j \mathbf{0}_j$, i.e., $u_j(r_j) < u_j(R_j) \leq 0$ and $x_i \succsim_i \mathbf{0}_i$, i.e., $u_i(x_i) \geq 0$. Consider the two following queries:

- What is the highest k such that $(x_i, R_j, \mathbf{0}_{-ij}) \succsim (x_{i,k}, r_j, \mathbf{0}_{-ij})$? Then we set $y^-_i = x_{i,k}$ and $y^+_i = x_{i,k+1}$.

Similarly, for any h_i element of $X_i \backslash \{x_i\}$ chosen such that $h_i \succsim \mathbf{0}_i$, we ask the following question:

- What is the lowest k such that $(h_i, R_j, \mathbf{0}_{-ij}) \precsim (x_{i,k+1}, r_j, \mathbf{0}_{-ij})$? Then we set $z_i^- = x_{i,k}$ and $z_i^+ = x_{i,k+1}$.

Assuming $(\mathbf{1}_i, \mathbf{0}_{-i}) \succ \mathbf{0}$, i.e., $v(\{i\}) > 0$, we obtain (see Proposition 4 in Appendix Section A.2.2):

$$u_i(h_i) - u_i(z_i^+) \leq u_i(x_i) - u_i(y_i^-)$$
$$u_i(h_i) - u_i(z_i^-) > u_i(x_i) - u_i(y_i^+)$$

By choosing h_i equal to y_i^-, we obtain the following simplified inequations:

$$u_i(y_i^-) - u_i(z_i^+) \leq u_i(x_i) - u_i(y_i^-) \tag{9}$$
$$u_i(y_i^-) - u_i(z_i^-) > u_i(x_i) - u_i(y_i^+) \tag{10}$$

In the next section, we use constraints of type (5)–(10) on the utility function u_i to derive the utility curve using spline regression.

3.3 Learning the utility curves

Let us present the learning of utility curves u_i under the restricted solvability assumption, for the sake of simplicity. It is based on preference information represented by linear equations over utility values, obtained as explained in Section 3.1. The learning procedure can be adapted to the other case (when retricted solvability does not hold) and will not be presented in details. It is basically sufficient to replace linear equalities constraining the admissible utilities by the linear inequalities obtained in Section 3.2.

In the original elicitation method [17], Q_{ij}^+ queries are involved in a recursive procedure known as *standard sequence* aiming to construct a sequence of points on the utility curve $(z_t, u_i(z_t))_{t=0}^q$ such that $z_t \in X_i$. The sequence is obtained as follows: $z_0 = \mathbf{0}_i$ and z_{t+1} is the answer to query $Q_{ij}^+(z_t)$. By construction this sequence is such that $z_t \prec_i z_{t+1}$ and this improving sequence stops at step q when consequence $\mathbf{1}_i$ is reached. Then we have $u_i(z_{t+1}) - u_i(z_t) = u_i(z_t) - u_i(z_{t-1})$ by (6), yielding the following recursive definition:

$$u_i(z_{t+1}) = 2u_i(z_t) - u_i(z_{t-1}) \tag{11}$$

This completely determines the sequence since $u_i(z_0) = 0$ and $u_i(z_q) = 1$. We indeed have $u_i(z_t) = t/q$ for $t = 1, \ldots, q$. A symmetric sequence can be implemented to construct points on the utility curve below the neutral level using Q^- queries.

However, such a method is known to be extremely sensitive to errors in the responses [28]. Indeed, if one considers that every answer is provided with some noise such that $u_i(z_{t+1}) = 2u_i(z_t) - u_i(z_{t-1}) + \epsilon_t$ for $t = 1, \ldots, q$ where $\epsilon_t \sim \mathcal{U}([-\epsilon_{max}, \epsilon_{max}])$ and $\epsilon_{max} > 0$, the estimation error on the points of the standard sequence can be large. Indeed $u_i(z_t) = \frac{t}{q} + \sum_{k=1}^t \epsilon_k - \frac{t}{q} \sum_{k=1}^q \epsilon_k$ for $t = 1, \ldots, q$ and the expected squared estimation error on a point z_t is: $\mathbb{E}[(u_i(z_t) - t/q)^2] = \mathbb{E}[(\sum_{k=1}^t \epsilon_k)^2] + t^2/q^2 \mathbb{E}[(\sum_{k=1}^q \epsilon_k)^2] - \frac{2t}{q} \mathbb{E}[\sum_{k_1=1}^t \sum_{k_2=1}^q \epsilon_{k_1}\epsilon_{k_2}] = \frac{2\epsilon_{max}^2}{3}(t - \frac{t^2}{q})$.

Then the maximum expected estimation error is reached on the middle of the sequence and is proportional to the length q of the sequence and to the squared noise intensity ϵ_{max}^2:

$$\max_t \mathbb{E}[(u_j(z_t) - t/q)^2] = q\epsilon_{max}^2/12$$

Regression based on short standard sequences Since the error increases with the length of the standard sequence, we propose an alternative approach that relies on multiple minimal length ($q = 2$) standard sequences of type (z_0, z_1, z_2). Multiplicity is obtained by varying the initial location z_0, the reference dimension j and the mesh (r_j, R_j). Note that if z_0 is below the neutral level we use decreasing sequences generated by Q^- queries. Putting all together, we obtain a set of p triplets $(z_0^\iota, z_1^\iota, z_2^\iota)_{\iota=1}^p$ each associated with a linear constraint on u_i given by (6). Then we define u_i as a I-spline function that best fits the resulting set of equalities.

Spline functions are piecewise polynomial functions of class C^k widely used for data interpolation or approximation due to their ability to smoothly approximate complex shapes. Moreover they allow for a compact representation of utilities. Indeed, a spline function can be expressed as a linear combination of basis functions and is thus characterized by the coefficients of the combination. Since utility functions are supposed to be increasing, we will use a basis $(I_l)_{l=1}^L$ of monotonically increasing spline functions, known as I-spline functions [29] weighted by positive coefficients (adding up to 1 so as to have $u_i(\mathbf{1}_i) = 1$). We use here cubic I-splines ($k = 3$) because they have matching first and second derivatives while preserving a local influence of every components. Note that we use a translation of the basis functions from $[0; 1]$ to $[-1; 1]$. Formally, u_i is defined by parameters $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,L}) \in [0, 1]^L$ such that:

$$\forall x_i \in [-\mathbf{1}_i, \mathbf{1}_i], \quad u_i(x_i) = 2\sum_{l=1}^L \alpha_{i,l} I_l(x) - 1 \tag{12}$$

Using (12), the problem of finding the utility that best fits the equalities can be formalized as a linear program with relaxed constraints:

$$\min \sum_{\iota=1}^p (\epsilon_\iota^+ + \epsilon_\iota^-)$$

$$\sum_{l=1}^L \alpha_{i,l}(2I_l(z_1^\iota) - I_l(z_0^\iota) - I_l(z_2^\iota)) + \epsilon_\iota^+ - \epsilon_\iota^- = 0, \quad \iota = 1 \ldots p \tag{13}$$

$$2\sum_{l=1}^L I_l(\mathbf{1}_i)\alpha_{l,j} - 1 = 1 \tag{14}$$

$$2\sum_{l=1}^L I_l(-\mathbf{1}_i)\alpha_{l,j} - 1 = -1 \tag{15}$$

$$2\sum_{l=1}^L I_l(\mathbf{0}_i)\alpha_{l,j} - 1 = 0 \tag{16}$$

$$\epsilon_\iota^+ \geq 0, \epsilon_\iota^- \geq 0, \alpha_{l,i} \geq 0.$$

Constraint (13) achieves the approximation of (11) for every triplet ι. Also, Constraint (14), (15), and (16) respectively represent the conditions $u_i(-\mathbf{1}_i) = -1$, $u_i(-\mathbf{0}_i) = 0$ and $u_i(\mathbf{1}_i) = 1$. A similar linear program can be considered in the case of non restricted solvability where preference statements replace indifference statements. It is sufficient to substitute linear inequalities used to approximate indifference judgements by linear inequality used to approximate preference judgements.

4 Learning sparse representations of capacities

Once utility functions u_i, $i \in N$ are known, the second-stage is to learn a sparse representation of capacity v. Given a set of preference statements $\{(x^\ell, y^\ell) \in X^2 : x^\ell \succsim y^\ell, \ell \in P\}$, a set of indifference statements $\{(x^\ell, y^\ell) \in X^2 : x^\ell \sim y^\ell, \ell \in I\}$ and utility functions u_i, $i \in N$, we want to find a capacity v having as many zero Möbius masses as possible and such that f_v^u well describes the observed preferences.

L_1-regularization [30, 31] is a well studied approach to control the model's complexity. Several options have been proposed to obtain sparse solutions for the non-additive integrals via the L_1 penalty term. For example, the sparsity inducing penalty was applied to the capacity [32, 33]. The L_1 penalty was also applied to capacities represented by interaction indices in [34]. Here, we also explore the L_1 penalty term to obtain a sparse solution, however, we focus on learning sparse Möbius representations. This choice is motivated by the fact that capacities are known to be less compact than their Möbius inverse (due to monotonicity); the same statement holds for interaction indices provided that Möbius masses are positive [18].

More precisely, our goal here is to fit preference examples with a simple model (i.e., with as few interactions as possible) but also to show the descriptive advantage of the Choquet integral compared to the standard linear aggregation model. The baseline model is therefore the weighted sum including all singletons as components (some possibly with a null coefficient). Then, when working with the Choquet integral, the aim of avoiding unnecessary interactions in the model naturally leads to include only the Möbius masses of subsets larger than 1 in the regularization term. This makes it possible to explore the tradeoff between simplicity and empirical error of the learned model, by progressively increasing the weight of the regularization term until it vanishes, yielding a linear model.

It is important to note that keeping the singletons in the regularization term (e.g., for criteria selection purposes) would raise an issue due to the normalization constraint ($v(N) = \sum_{B\subseteq N} m_v(B) = 1$). The regularization term $\|m_v\|_1$ is indeed bounded by 1 since we have: $\|m_v\|_1 = \sum_{B\subseteq N}|m_v(B)| \geq \sum_{B\subseteq N} m_v(B) = 1$. By noticing that this bound is reached for any positive Möbius transform, we have that increasing the level of regularization directly favors positive solutions, and can harm the ability to recover negative Möbius masses.

Besides, the L_1 regularization over the interaction factors might be impacted by the structural dependence that exists between the quantities of type $\min_{i\in B}\{u_i(x_i)\}$ involved in CIU (4). In particular, the statistical dependence can harm the ability to properly select the interaction factors. We thus propose to use a standard approach to correct this issue that consists in using a weighted L_1 regularization $\sum_{B\subseteq N} \lambda_B|m_v(B)|$ where the weights λ_B are adapted to preference data (see Section 4.2). Note that another weighting system based on the cardinality of factors has been used in Choquistic regression problems to favor the selection of small size factors [6]. However, this choice may prevent to recover preference systems where large coalitions are essential. For instance, this is the case of the ϵ-min model introduced in Example 2, and also of the so-called Hurwicz model [35] based on a convex combination of min and max factors taken on the grand coalition N.

For the sake of clarity, we first present the unweighted version of the L_1-regularized problem (Section 4.1). A more sophisticated version using a weighting system is presented in Section 4.2.

4.1 Unweighted L_1-regularization on the Möbius inverse

The unweighted version of the regression problem reads as follows:

$$(\mathcal{P}) \min \sum_{\ell\in I}|f_v^u(x^\ell) - f_v^u(y^\ell)| + \sum_{\ell\in P}(f_v^u(x^\ell) - f_v^u(y^\ell))^- + \lambda \sum_{B\subseteq N, |B|>1} |m_v(B)|$$

$$\textstyle\sum_{B\subseteq A, B\ni i} m_v(B) \geq 0, \quad \forall A \subseteq N, \forall i \in A \tag{17}$$

$$\textstyle\sum_{B\subseteq N} m_v(B) = 1 \tag{18}$$

where λ is a nonnegative hyper-parameter that controls the level of regularization. The objective function aims at minimizing the magnitude of violation of indifference and preference examples. Constraints (17) and (18) respectively ensure the monotonicity of the capacity w.r.t. set inclusion and its normalization. The monotonicity constraints, expressed in terms of Möbius masses, guarantee that for any criteria coalition $A \subseteq N$, removing a criterion i cannot increase the capacity value. In the optimization problem given above, the L_1-penalty allows sparse representations of capacities to be obtained by shrinking Möbius masses $m_v(B)$ towards zero (the intensity of the shrinkage depending on the level of regularization). Then, a selection of the criteria coalitions that actually play in the model is performed. As a consequence, it is of prime importance to assess the quality of such a selection. In the following, we give theoretical insights justifying the need for a more sophisticated L_1-regularization to perform a qualitative criteria coalition selection.

Issue due to interdependent components In order to make explicit a possible issue with L_1-regularization, let us consider a special case of Problem $\mathcal{P}$ wherein the database of learning examples is only made of indifference statements with specific pairs of examples (x^ℓ, y^ℓ), $\ell \in I$, chosen in such a way that y^ℓ has a constant utility vector (i.e., $u_i(y^\ell) = \theta^\ell$ for all $i \in N$ for some $\theta^\ell \in \mathbb{R}$). In such a case we have $f_v^u(y^\ell) = \theta^\ell$ whatever the capacity v. Therefore the indifference $x^\ell \sim y^\ell$ translates into the constraint $f_v^u(x^\ell) = \theta^\ell$. Hence, Problem $\mathcal{P}$ boils down to approximate values θ^ℓ by Choquet values $f_v^u(x^\ell)$, $\ell \in I$. We obtain the following regression problem:

$$\min \sum_{\ell \in I} |f_v^u(x^\ell) - \theta^\ell| + \lambda \sum_{B \subseteq N, |B|>1} |m_v(B)|$$

$$\text{s.t. (17), (18)}$$

This optimization problem is an instance of constrained least absolute deviation linear regression with L_1-regularization. Indeed, (4) presents f_v^u as a linear aggregator $\theta = \sum_{j=1}^p \beta_j \phi_j$ within a specific feature space of size p. The features ϕ_j are defined as the utility minima taken over every possible criteria coalition and the attached coefficients β_j are the Möbius masses. More formally, let us index the subsets of N in the lexicographical order. For any $B \subseteq N$, let $\rho(B)$ be the rank of subset B in this order and $\rho^{-1}(j)$ the subset positioned at rank j in the order. Then $\phi_j = \min_{i \in \rho^{-1}(j)}\{u_i(x_i)\}$ and $\beta_j = m_v(\rho^{-1}(j))$ for $j = 1, \ldots, 2^n$.

Often referred to as LAD-LASSO [36, 37], L_1-regularized least absolute deviation linear regression has been extensively studied in the statistical learning literature and, in particular, its properties concerning variable selection are now well understood. More precisely, consider a linear model $\theta^\ell = \sum_{j=1}^p \beta_j^* \phi_j^\ell + \epsilon^\ell$ where $\epsilon = (\epsilon^\ell)_{\ell \in I}$ is a vector of i.i.d centered random variables and where θ^ℓ and $\phi^\ell = (\phi_1^\ell, \ldots, \phi_p^\ell)$, $\ell \in I$, are respectively observations of the response θ and the p-dimensional predictor $\phi = (\phi_1, \ldots, \phi_p)$. Let $\Phi = (\phi^\ell)_{\ell \in I}$ denote the $|I| \times p$ design matrix and assume that $\lim\limits_{|I| \to \infty} \frac{\Phi\Phi^T}{|I|} = C$ where C is a positive definite covariance matrix. Suppose that the ground truth coefficient vector β^* is sparse in the sense that it contains few non-null components, and that it can be split in two parts A_1, A_2 with $A_1 = \{j | \beta_j^* \neq 0\}$ and $A_2 = \{j | \beta_j^* = 0\}$. Then it is known that, under mild assumptions, LAD-LASSO is able to select exactly the set of ground truth non-null coefficients A_1 only if a condition on the feature covariance matrix C known as the "Irrepresentable Condition" holds [37]. The condition reads as follows:

$$|C_{21} C_{11}^{-1} \operatorname{sign}(\beta_{A_1}^*)| < \mathbf{1} \tag{19}$$

where $\mathbf{1} = (1, \ldots, 1)$ and the inequality holds element-wise. Moreover, $\beta_{A_1} = (\beta_{j \in A_1})$, $C_{11} = (C_{ij})_{i,j \in A_1}$,$C_{21} = (C_{ij})_{i \in A_2, j \in A_1}$ and for any vector β, $\text{sign}(\beta)$ refers to its sign vector, i.e., $\text{sign}(\beta)_j = 1$ if $\beta_j > 0$, $\text{sign}(\beta)_j = -1$ if $\beta_j < 0$ and $\text{sign}(\beta)_j = 0$ otherwise. More formally, Condition (19) is necessary for guaranteeing that the probability of the existence of a λ value for which it correctly affects signs to coefficients goes towards 1 as the number of observations approaches infinity (general sign consistency), see [37] Theorem 4. A similar result [38, 39] is also available for least square L_1-penalized linear regression (a.k.a. LASSO [30, 31]).

In the case of CIU, the feature space is endowed with a very specific correlation structure since for any pair of criteria coalition $S_1, S_2 \subseteq N$ such that $S_1 \cap S_2 \neq \emptyset$, $\phi_{\rho(S_1)} = \min_{i \in S_1}\{u_i(x_i)\}$ and $\phi_{\rho(S_2)} = \min_{i \in S_2}\{u_i(x_i)\}$ are obviously statistically correlated due to the overlapping of the coalitions. Intuitively, the correlation is all the more important that the cardinal of the intersection is close to the cardinal of the union. This is well illustrated in Fig. 2 that compares the empirical correlation between $\min_{i \in S_1}\{u_i(x_i)\}$ and $\min_{i \in S_2}\{u_i(x_i)\}$ (left handside) and the ratio $R = |S_1 \cap S_2|/|S_1 \cup S_2|$ for any $S_1, S_2 \subseteq N$ (right handside). The number of criteria n is taken equal to 8 and for any $i \in N$, i.i.d. utility samples $(u_i(x_i^l))_{\ell \in I}$ of size $|I| = 500$ are simulated according to a uniform distribution within [0, 1] to compute the empirical correlations. The similarity of the patterns in both graphs suggests that the correlation scheme is indeed well described by ratio R introduced above.

In Example 4, we show that this correlation structure undermines the respect of Condition 19, and thus the ability of LAD-LASSO to recover a sparse ground truth model.

Example 4 Consider the ϵ-min CIU model (see Example 2) for $n = 3$ ($N = \{1, 2, 3\}$): $\theta = \epsilon_1 u_1(x_1) + \epsilon_2 u_2(x_2) + \epsilon_3 u_3(x_3) + (1 - \sum_{i=1}^{3} \epsilon_i) \min_{i \in N}\{u_i(x_i)\}$. Then $A_1 = \{\{1\}, \{2\}, \{3\}, \{1, 2, 3\}\}$, $A_2 = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ and $\text{sign}(\beta^*_{A_1}) = \mathbf{1}$. Suppose that the utilities $u_i(x_i)$, $i = 1, 2, 3$ follow a uniform distribution within [0, 1]. For any pair of criteria coalition $S_1, S_2 \subseteq N \setminus \emptyset$ of cardinals $s_1 = |S_1|$, $s_2 = |S_2|$ and $s_{12} = |S_1 \cap S_2| \neq 0$, denote $V^{s_{12}}_{s_1, s_2} = Cov(\phi_{\rho(S_1)}, \phi_{\rho(S_2)}) = C_{\rho(S_1), \rho(S_2)}$. Then C_{11} is positive definite and Condition

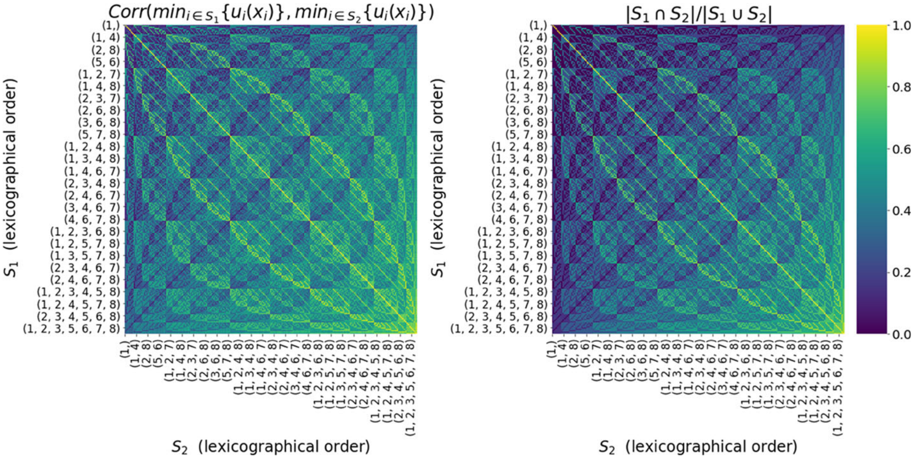

Fig. 2 Empirical correlation between $\min_{i \in S_1}\{u_i(x_i)\}$ and $\min_{i \in S_2}\{u_i(x_i)\}$ (right) and ratio $|S_1 \cap S_2|/|S_1 \cap S_2|$ (left) w.r.t. $S_1, S_2 \subseteq N$

(19) boils down to (see Propositon 8 in Appendix B):

$$|2V_{1,2}^1(V_{3,3}^3 - V_{1,3}^1) + V_{2,3}^2(V_{1,1}^1 - 3V_{1,3}^1)| < |V_{3,3}^3 V_{1,1}^1 - 3(V_{1,3}^1)^2| \tag{20}$$

Also, we show that for $n \leq 3$, the covariance $V_{s_1,s_2}^{s_{12}}$ reads as follows:

$$V_{s_1,s_2}^{s_{12}} = \sum_{k=1}^{3} g_k(s_{12})\gamma_k(s_1, s_2, s_{12}) - \frac{1}{(s_1+1)(s_2+1)} \tag{21}$$

with $g_k(s_{12}) = \frac{k!}{\prod_{i=1}^{k}(s_{12}+i)}$, $\gamma_1 = 1$, $\gamma_2(s_1, s_2, s_{12}) = -\frac{1}{2}((s_1 - s_{12})^+ + (s_2 - s_{12})^+)$ and $\gamma_3(s_1, s_2, s_{12}) = \frac{1}{4}((s_1 - s_{12})^+(s_2 - s_{12})^+) + \frac{1}{6}((s_1 - s_{12})^+(s_1 - s_{12} - 1)^+ + (s_2 - s_{12})^+(s_2 - s_{12} - 1)^+)$. The proof of this formula is provided in Propositon 7 of Appendix B. Using the numerical values provided in Table 1, we obtain that Condition (19) is equivalent to $\frac{19}{1440} < \frac{1}{800}$, which is obviously not true. Thus LAD-LASSO is not general sign consistent in the ϵ-min model recovery for $n = 3$.

The violation of Condition (19) in Example 4 suggests some weaknesses in terms of criteria coalition selection for the unweighted L_1-regularization. In order to circumvent the criteria coalition selection problem, we investigate the benefit of an adaptive L_1-penalty, i.e., a weighted L_1-penalty with data-dependent weights.

4.2 Learning Möbius masses with adaptive L_1-penalty

The adaptive L_1-penalty is of the form $\sum_j \lambda_j |\beta_j|$ where the weights λ_j are data-dependent and adapted to each coefficient β_j, implying a two-stage algorithm where the first step is the weights computation. It has been introduced to correct LASSO and LAD-LASSO and guarantee better variable selection properties [39–43]. In particular, when the weights are the absolute values of the reciprocals of the L_2-penalized solution it is shown [39] that the adaptive LASSO selects the ground truth coefficients with a probability that tends to 1 when the number of observation goes toward the infinity (variable selection consistency). We thus propose to use this two-stage penalty in the learning of capacities. It yields the following approximation problem:

Problem$\mathcal{P}'$

$$\min \sum_{\ell \in I} |f_v^u(x^\ell) - f_v^u(y^\ell)| + \sum_{\ell \in P} (f_v^u(x^\ell) - f_v^u(y^\ell))^- + \sum_{B \subseteq N, |B|>1} \lambda_{\rho(B)} |m_v(B)|$$

$$\text{s.t. } (17), (18)$$

where $\lambda_{\rho(B)} = \lambda / (|\hat{m}_v(B)| + \epsilon)$ for all $B \subseteq N$ ($\epsilon > 0$ being introduced to avoid zero division) and $\hat{m}_v$ is the optimal solution of the following initialization problem:

Problem $\mathcal{P}_0'$

Table 1 Covariance numerical values

$V_{1,1}^1$	$V_{3,3}^3$	$V_{1,2}^1$	$V_{1,3}^1$	$V_{2,3}^2$
$\frac{1}{12}$	$\frac{3}{80}$	$\frac{1}{24}$	$\frac{1}{40}$	$\frac{1}{30}$

$$\min \sum_{\ell \in I} |f_v^u(x^\ell) - f_v^u(y^\ell)| + \sum_{\ell \in P} (f_v^u(x^\ell) - f_v^u(y^\ell))^- + \lambda_2 \|m_v\|_2^2$$
$$\text{s.t. (17), (18)}$$

Note that solving $\mathcal{P}_0'$ requires the setting of an additional hyperparameter λ_2. This can be done using cross-validation. Observing that $f_v^u(x^\ell) = \sum_{j=1}^{2^n} m_v(\rho^{-1}(j)) \min_{i \in \rho^{-1}(j)}\{u_i(x_i^\ell)\}$ by (4), we have $f_v^u(x^\ell) - f_v^u(y^\ell) = \sum_{j=1}^{2^n} m_v(\rho^{-1}(j))\Delta_j^\ell$ with $\Delta_j^\ell = \min_{t \in \rho^{-1}(j)}\{u_t(x_t^\ell)\} - \min_{t \in \rho^{-1}(j)}\{u_t(y_t^\ell)\}$. Then we solve Problem $\mathcal{P}'$ using linear programming by introducing auxiliary variables to linearize the objective function as follows:

$$\min \sum_{\ell \in I} (\epsilon_\ell^+ + \epsilon_\ell^-) + \sum_{\ell \in P} \epsilon_\ell + \sum_{j=n+1}^{2^n} \lambda_j (a_j + b_j)$$
$$\textstyle\sum_{j=1}^{2^n} (a_j - b_j)\Delta_j^\ell + \epsilon_\ell^+ - \epsilon_\ell^- = 0, \quad \ell \in I \tag{22}$$
$$\textstyle\sum_{j=1}^{2^n} (a_j - b_j)\Delta_j^\ell + \epsilon_\ell \geq 0, \quad \ell \in P \tag{23}$$
$$\textstyle\sum_{B \subseteq A, B \ni i} a_{\rho(B)} - b_{\rho(B)} \geq 0, \quad \forall A \subseteq N, \forall i \in A \tag{24}$$
$$\textstyle\sum_{j=1}^{2^n} a_j - b_j = 1 \tag{25}$$
$$\epsilon_\ell^+, \epsilon_\ell^-, \epsilon_\ell, a_j, b_j \geq 0$$

Equations 22 and 23 are flexible constraints used to approximate indifference and preference examples. The quantities $|m_v(\rho^{-1}(j))|$ are linearized using a standard reformulation of absolute values based on the fact that $|x|$ is the solution of minimizing $a + b$ subject to $x = a - b$ and $a, b \geq 0$. A similar linearization is used for terms $|f_v^u(x^\ell) - f_v^u(y^\ell)|$ and $(f_v^u(x^\ell) - f_v^u(y^\ell))^-$ using variables $\epsilon_\ell^+, \epsilon_\ell^-$ and ϵ_ℓ. Constraints (24) and (25) respectively impose the monotonicity and the normalization of the capacity. The weights λ_j are computed by solving $\mathcal{P}_0'$ with quadratic programming.

In order to derive a similar optimization problem for the learning of the bi-CIU model, let us reformulate $f_{v,w}^u$ from the Möbius inverses of capacities v and w. We obtain:

$$\begin{aligned} f_{v,w}^u(x) &= \sum_{B \subseteq N} m_v(B) \min_{i \in B}\{u_i(x_i)^+\} + \sum_{B \subseteq N} m_w(B) \min_{i \in B}\{-u_i(x_i)^-\} \\ &= \sum_{B \subseteq N} m_v(B) \min_{i \in B}\{u_i(x_i)^+\} - \sum_{B \subseteq N} m_w(B) \max_{i \in B}\{u_i(x_i)^-\} \end{aligned} \tag{26}$$

Using (26), we formulate the problem of learning sparse representations of the capacities v and w in bi-CIU as follows:

$$
\begin{aligned}
&\min \sum_{\ell \in I}(\epsilon_\ell^+ + \epsilon_\ell^-) + \sum_{\ell \in P}\epsilon_\ell + \sum_{j=n+1}^{2^n}\lambda_j^v(a_j+b_j) + \sum_{j=n+1}^{2^n}\lambda_j^w(c_j+d_j)\\
&\textstyle\sum_{j=1}^{2^n}(a_j-b_j)\Delta_j^\ell - \sum_{j=1}^{2^n}(c_j-d_j)\nabla_j^\ell + \epsilon_\ell^+ - \epsilon_\ell^- = 0, \quad \ell \in I\\
&\textstyle\sum_{j=1}^{2^n}(a_j-b_j)\Delta_j^\ell - \sum_{j=1}^{2^n}(c_j-d_j)\nabla_j^\ell + \epsilon_\ell \geq 0, \quad \ell \in P\\
&\textstyle\sum_{B\subseteq A, B\ni i}(a_{\rho(B)} - b_{\rho(B)}) \geq 0, \quad \forall A \subseteq N, \forall i \in A\\
&\textstyle\sum_{B\subseteq A, B\ni i}(c_{\rho(B)} - d_{\rho(B)}) \geq 0, \quad \forall A \subseteq N, \forall i \in A\\
&\textstyle\sum_{j=1}^{2^n}(a_j-b_j) = 1\\
&\textstyle\sum_{j=1}^{2^n}(c_j-d_j) = 1\\
&\epsilon_\ell^+, \epsilon_\ell^-, \epsilon_\ell, a_j, b_j, c_j, d_j \geq 0
\end{aligned}
$$

Where $\Delta_j^\ell = \min_{t\in\rho^{-1}(j)}\{u_t(x_t^\ell)^+\} - \min_{t\in\rho^{-1}(j)}\{u_t(y_t^\ell)^+\}$ and $\nabla_j^\ell = \max_{t\in\rho^{-1}(j)}\{u_t(x_t^\ell)^-\} - \max_{t\in\rho^{-1}(j)}\{u_t(y_t^\ell)^-\}$. The weights $(\lambda_j^v,\lambda_j^w)$ are computed beforehand with a quadratic program similar to $\mathcal{P}_0'$ but using a double L_2-penalization $\lambda_2(\|m_v\|_2^2 + \|m_w\|_2^2)$.

Another possible variant of standard L_1-regularization in the context of correlated variables is the Elastic Net [44]. This penalty is defined as a convex combination of L_1 and L_2 penalty : $\lambda\|m_v\|_1 + \lambda_2\|m_v\|_2^2$. However, this method tends to jointly select correlated features with uniformed coefficients values (grouping effect) as observed in [44]. This property is not desirable in our context. For instance, in the ϵ-min model, the importance of $m_v(N)$ in the ground truth model may thwart the elimination of sets having a large intersection with N. This is confirmed by our tests. In Section 5.3.3 we will show that adaptive L_1-penalty significantly outperforms the Elastic Net penalty.

5 Experiments

In this section we show the results of numerical experiments on synthetic and real-world preference data and we illustrate the advantage of our approach over some baseline methods.

5.1 Synthetic data

We generate the synthetic data as follows. First, random sparse Choquet models f_v^u or $f_{v,w}^u$ are created through the generation of n utility function u_i and one or two (depending on the choice for CIU or bi-CIU) capacities admitting a sparse Möbius representation. Sparse Möbius masses are first generated without requiring for monotonicity, then m_v or (m_v, m_w) are taken as the Möbius representations of the closest (in the sense of the L_1 norm) monotonic capacities (obtained by linear programming).

Then, for n given utility functions u_i and capacities v, w, we simulate Q-queries and their answers for the utility learning. Answers to Q-queries are provided with some random uniform noise $\epsilon \in [-\epsilon_{max}, \epsilon_{max}]$. For the capacity learning, we construct sets of preferences $\{(x^\ell, y^\ell) \in X^2 : x^\ell \succsim y^\ell, \ell \in P\}$ and of indifference statements $\{(x^\ell, y^\ell) \in X^2 : x^\ell \sim y^\ell, \ell \in I\}$ compatible with f_v^u or $f_{v,w}^u$. For this, pairs (x^ℓ, y^ℓ) are drawn uniformly within X^2.

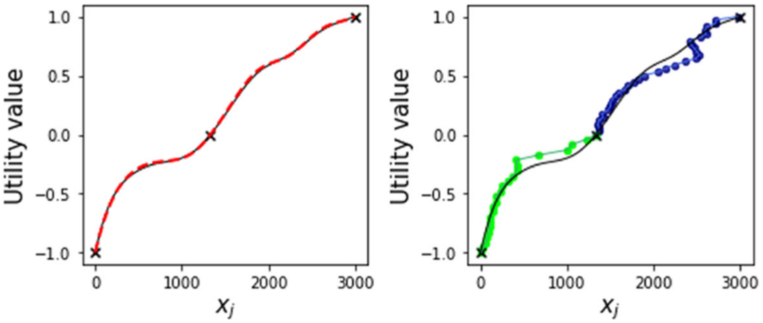

Fig. 3 Learned utility function with our method (left: red dotted line) and with standard sequences (right: green and blue points) along with the ground truth (black plain line)

In order to introduce noise, each example is associated with a preference statement $x^\ell \succsim y^\ell$ if $(f_v^u(x^\ell)+\sigma_x^\ell)-(f_v^u(y^\ell)+\sigma_y) \geq \sigma$ and $x^\ell \sim y^\ell$ if $|(f_v^u(x^\ell)+\sigma_x^\ell)-(f_v^u(y^\ell)+\sigma_y^\ell)| < \sigma$ (for CIU), where $\sigma_x^\ell, \sigma_y^\ell$ are noise values uniformly taken within an interval $[-\sigma, \sigma]$. This process is used to generate training sets of size $|P|+|I|$ which we vary in our experiments, and test sets of size $|P|+|I| = 1000$. The preference and indifference examples are in equal proportions. In the following, the generalizing performance (test error) of any learned model $\hat{f}$ is evaluated as the average absolute violation of preferences on a test set: $\frac{1}{|P|}\sum_{\ell\in P}(\hat{f}(y^\ell)-\hat{f}(x^\ell))_+ + \frac{1}{|I|}\sum_{\ell\in I}|\hat{f}(y^\ell)-\hat{f}(x^\ell)|$.

For all regularization methods, the hyper-parameter λ is chosen by cross-validation with the one-standard-error rule. It consists in cutting the training set in folds (here the number of folds is set to 3) and training the model as many times as the number of folds, each time reserving a different fold for evaluating the model (validation fold). Then, λ is selected as the highest value yielding a mean error on the validation folds lower than the minimum mean error over all λ plus the standard error associated to this minimum. A grid-search is performed over the second hyper-parameter λ_2 whenever it is needed (adaptive L_1-penalty and elastic net penalty).

5.2 Learning utilities

We conduct numerical tests on the utility learning. First, we generate a sparse model $f_{v,w}^u$ and learn the utility function u_i for some $i \in N$ with standard sequences and then with our method. Answers to Q-queries are simulated with a level of noise $\epsilon_{max} = 0.05$. Figure 3 displays the estimated utility functions for both methods along with the ground truth u_i. On the left, the estimation provided by our method perfectly matches the ground truth while on the right the estimation of the standard sequence clearly suffers from noise distortion.

We conducted the same experiment on 10 random sparse models, and obtained an average mean absolute error (MAE) of 0.084 ± 0.052 for the standard sequence method and 0.022 ± 0.008 for our approach. The MAE is the mean absolute difference between the ground truth and the estimated values on a fixed subdivision.

On Fig. 4 we represent the accuracy of both methods in terms of MAE as a function of the number of questions asked. The MAE are averaged on 10 simulations of random sparse models. Figure 4 shows the case $\epsilon_{max} = 0.05$ on the left and $\epsilon_{max} = 0.1$ on the right. The test confirms that long standard sequences constructed recursively (grey) lead to very poor

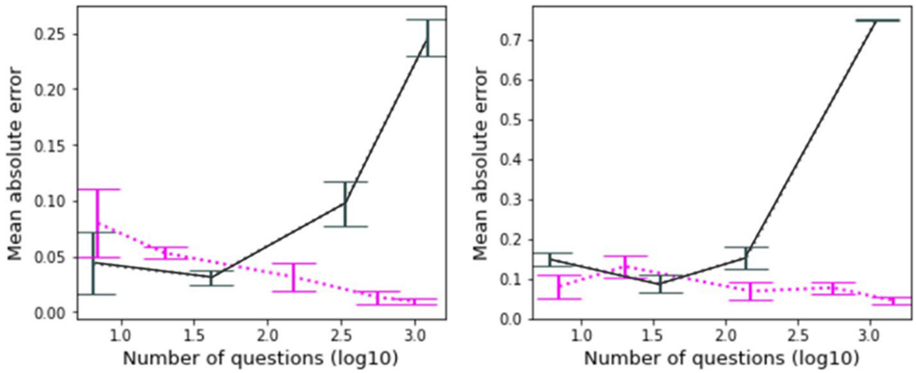

Fig. 4 MAE w.r.t. the number of questions for our method (dotted lines) and standard sequences (plain lines) over 10 simulations for $\epsilon_{max} = 0.05$ (left) and $\epsilon_{max} = 0.1$ (right)

results. Also, the difference between both graphs shows the impact of the increase of noise intensity on the estimation quality for both method. However, one can see that regardless the level of noise, our approach converges to a null MAE. It appears to be a more robust approach.

5.3 Learning Möbius representations of capacities

We first illustrate the process of learning a sparse Möbius representation of the capacity in the specific case of Example 2. Then, with other toy examples, we illustrate the benefit of adaptive L_1-regularization in terms of criteria coalition selection. Then we proceed to simulations to demonstrate the benefits of our approach in the general case of sparse synthetic data and real-world preference data.

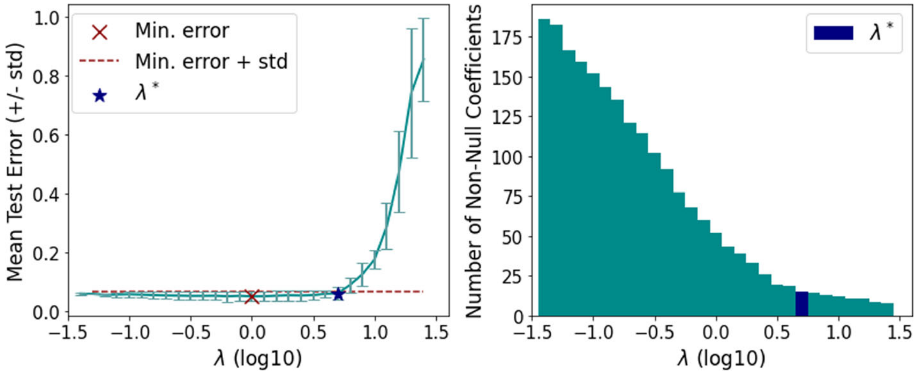

Fig. 5 Selection of the hyperparameter λ with cross validation: mean test error on the 3 tests folds (left) and L_0-norm of the learned models according λ (right)

5.3.1 Learning the ϵ-min example (Example 2)

We generate preference data according to the ϵ-min model of Example 2 instantiated with $n = 8$ and $\epsilon = (0.03, 0.03, 0.05, 0.05, 0.02, 0.02, 0.05, 0.05)$. More precisely, we generate a training set of size $|P| + |I| = 250$ and introduce noise using $\sigma = 0.03$. We compare our method based on adaptive L_1-regularization to some baselines, such as the standard L_1-regularization, the unpenalized regression and the use of 2-additivity constraints for an alternative control of model complexity. In Fig. 5 we illustrate the one-standard-error-rule used to select the optimal value λ^* of the regularization hyper-parameter λ (here $\lambda_2 = 0.05$). On the left of Fig. 5 we show the average generalizing performance (mean test error) obtained through cross-validation for different values of λ, and λ^* is highlighted. One can observe (Fig. 5 on the right) that the number of non-null coefficients decreases as λ increases, and λ^* corresponds to the optimal tradeoff between compactness and generalizing performance. On Fig. 6 we show the learned Möbius masses for the adaptive L_1-penalty with λ^* (top left), the L_1-penalty also with optimal regularization parameter λ (top right), the unpenalized regression (bottom left) and the use of 2-additivity constraints (bottom right). For each method, the learned model is superposed to the ground truth model. It is clear that the regression without any penalty term fails to recover the ϵ-min model; it does not find any compact representation either. It achieves, however, a reasonable generalizing performance on the test set (test error of 0.066). The 2-additive model, while being compact, is far from the ground truth and does not capture interactions involving a large number of attributes, leading to a poor generalizing performance (test error of 0.535). Our approach combines both advantages of the baselines: compactness and optimal generalizing performance (test error of 0.039). In fact, one can see that the ground truth model is exactly recovered. This is not the case with the standard L_1-penalty that includes other coefficients than the non-null ground truth coefficients in the estimated model. This directly illustrates the impact of the violation of Condition (19) for the ϵ-min model, as demonstrated in Example 4 for $n = 3$.

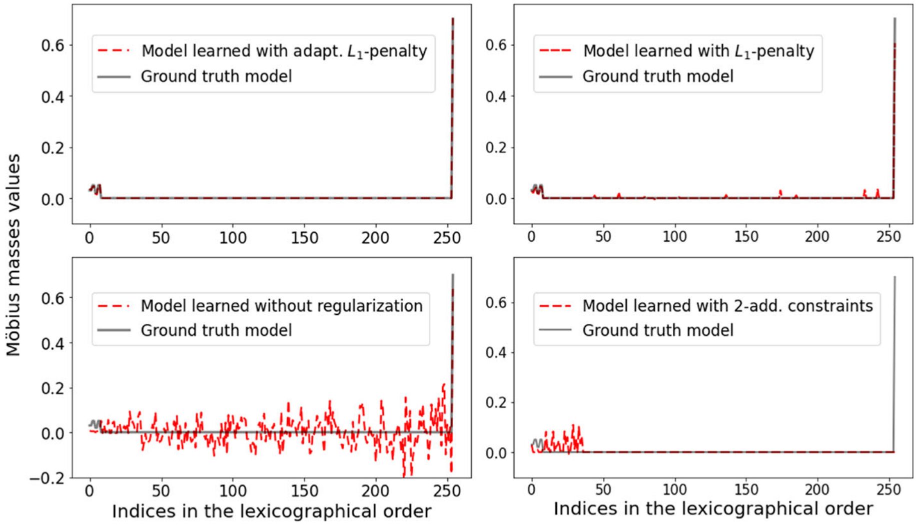

Fig. 6 Learned models and hidden model (ϵ-min model of Example 2)

5.3.2 Benefit of the adaptive L_1-penalty : illustrative example

In this section, we provide a second illustration of the benefit of adaptive L_1-regularization compared to standard L_1-regularization. To this end, we consider a model ($n = 6$) including 5 interaction terms attached to overlapping groups of criteria. The model is given by the following Möbius masses vector: $m_v(\{i\}) = \frac{\epsilon}{n}$ for any $i \in N$, $m_v(B) = \frac{1-\epsilon}{5}$ for any $B \in \{\{1,2\},\{1,2,3\},\{1,2,3,5,6\},\{1,3,4,5,6\},\{1,2,3,4,5,6\}\}$ and $m_v(B) = 0$ everywhere else, with $\epsilon = 0.2$.

We generate a training set of size $|P| + |I| = 120$ and observe the effect of the increase of the hyper-parameter λ for both standard and adaptive L_1-penalization. In Figs. 7 and 8 we represent the regularization paths i.e., the evolution of the learned Möbius masses w.r.t. λ, for both methods (for the adaptive penalty we take $\lambda_2 = 1$). The non-null coefficients of the hidden model are highlighted with blue star markers while the null coefficients are displayed with black plain lines. At first glance, the standard L_1-penalization does not succeed to efficiently distinguish ground truth non-null coefficients from null coefficients while the adaptive penalization provides a clear distinction for $\lambda \approx 10^{-0.25}$. Note that for high values of λ, the Möbius masses of singletons remain non-null for both methods. This is quite normal since they are not included in the penalization term (the aim of regularization being only to avoid unecessary non-linearities in the model).

In order to further evaluate and compare the quality of criteria coalition selection in both methods we compute the false discovery rate (FDR), i.e., the proportion of selected coefficients that are not actually in the ground truth model. We also compute the false exclusion rate (FER) which is the proportion of not selected coefficients that are actually in the ground truth model. Figure 9 shows the results for standard (left) and adaptive (right) L_1-regularization according to λ. Contrarily to standard L_1-regularization, adaptive L_1-penalty reaches 0% of false discovery rate (FDR) and 0% of false exclusion rate (FER) for $\lambda \in [10^{-1.35}, 10^{-1.1}]$. Thus adaptive L_1-penalty exactly recovers the set of non-null ground truth coefficients. Standard L_1 regularization appears to be less effective, the reduction of the false discovery rate comes at the expense of its false exclusion rate.

Stability study In the previous tests, we assessed the ability of adaptive L_1-regularization to efficiently recover a ground truth model. Now, with another illustrative example, we study the stability of the learned models w.r.t. the variability of the training preference data. We use a 5-dimensional CIU model with sparse Möbius transform and generate training sets

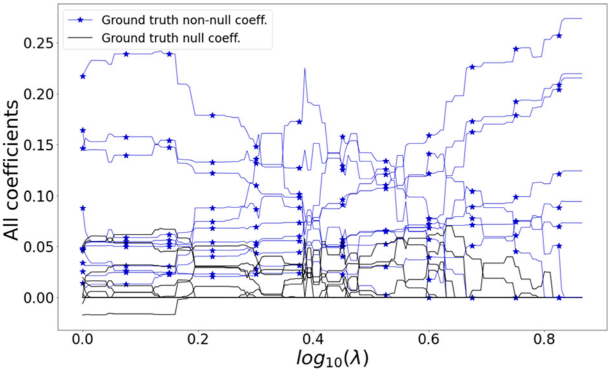

Fig. 7 Regularization path for standard L_1-penalty

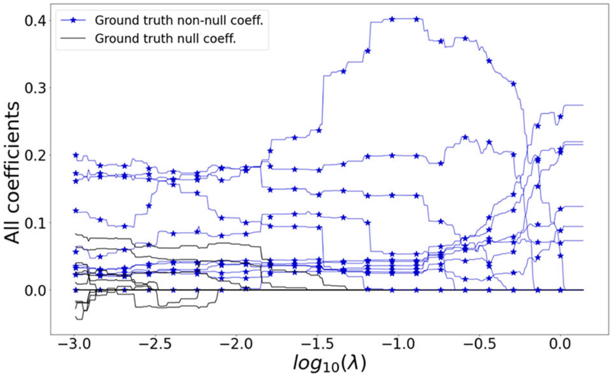

Fig. 8 Regularization path for adaptive L_1-penalty ($\lambda_2 = 1$)

of preference examples of size $|P| + |I| = 100$ with an increasing level of noise σ. In Fig. 10 are presented in boxplots the learned Möbius masses with adaptive L_1-regularization obtained for 10 random generations of preference data. From top to bottom are represented the results for increasing values of noise level $\sigma \in \{0, 0.03, 0.05, 0.1\}$. The ground truth model is highlighted with grey bars. For $\sigma = 0$ (top), the exact ground truth model is always recovered over the 10 simulations. Then, increasing the level of noise induces some variability in the learned models. However, for $\sigma = 0.03$ (second from top), very few coefficients that are not in the ground truth model are included in the learned model and the ground truth coefficients are recovered with a nearly constant amplitude. Finally, when the level of noise is high, i.e., $\sigma = 0.1$ (bottom), spurious coefficients such as the grand coalition are included in the learned model and the Möbius masses values are highly variable.

5.3.3 Comparative performance on arbitrary sparse models

We observed that CIU used with a 2-additive capacity can fail to properly approximate preference data when the underlying preferences contains higher-order interactions. Also, a more sophisticated L_1 regularization is sometimes needed to proceed to a good model selection. In this section, we provide broader tests on synthetic preference data and extend our compar-

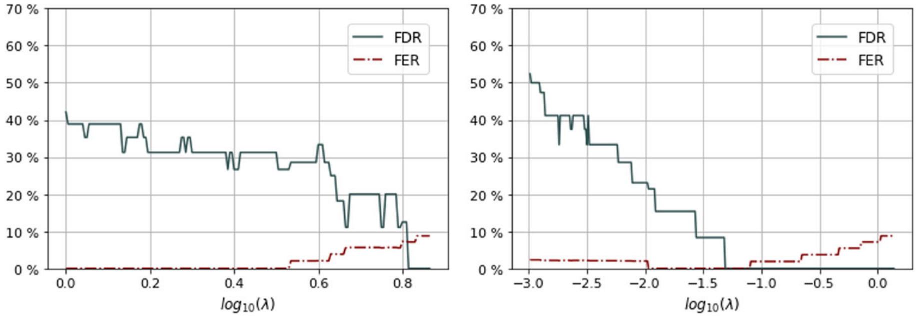

Fig. 9 False discovery rate (FDR) and false exclusion rate (FER) for standard (left) and adaptive (right) L_1-regularization w.r.t. λ

Table 2 Evaluation of ADA-L1 and the baseline methods with sparse CIU hidden models

	Test error	L_0-norm	$\|\hat{m} - m^*\|_2^2$	FDR	FER
ADA-L1	**0.07 ± 0.02**	**16.1 ± 8.6**	**0.05 ± 0.06**	**0.74 ± 0.09**	0.01 ± 0.01
L1	**0.07 ± 0.01**	25.1 ± 10.9	0.07 ± 0.10	0.82 ± 0.06	0.01 ± 0.01
E-Net	**0.07 ± 0.02**	27.3 ± 9.9	0.08 ± 0.12	0.83 ± 0.07	0.01 ± 0.01
No Reg.	0.09 ± 0.03	206.7 ± 27.1	1.57 ± 2.48	0.97 ± 0.02	**0.00 ± 0.01**
2-ADD	0.37 ± 0.18	21.9 ± 2.7	0.61 ± 0.49	0.95 ± 0.05	0.02 ± 0.01
4-ADD	0.13 ± 0.05	148.2 ± 4.8	0.45 ± 0.46	0.98 ± 0.02	0.02 ± 0.01
k^*-ADD	0.09 ± 0.03	147.0 ± 54.4	0.23 ± 0.24	0.97 ± 0.03	0.02 ± 0.02

isons to the use of k-additive models for $k = 1, \ldots, n-1$, and for an optimal k^* (chosen by cross-validation). Also, we compare the adaptive L_1-penalty to different penalizations such as the standard L_1-penalty and the elastic net. We finally compare the results of our method to the unpenalized regression method.

First, we generate 20 hidden arbitrary sparse CIU models (with 10 non-null coefficients in average) for $n = 8$. We also generate associated training preference datasets of size $|P| + |I| = 250$ (with $\sigma = 0.03$) and test sets of size $|P| + |I| = 1000$. The generalizing performance (test error) of our approach (ADA-L1) on test sets is averaged and displayed in Table 2 along with the average sparsity of the learned models (L_0 norm). The quality of the ground truth model retrieval is further assessed with the average gap to the ground truth model ($\|\hat{m} - m^*\|_2^2$) and the false discovery rate (FDR) and false exclusion rate (FER). We also present the results for the baseline methods: the standard L_1-penalty (L1), the Elastic Net penalty (E-Net), the unpenalized regression (No reg.) and methods that use k-additivity constraints for $k = 2, 4$ and k^*.

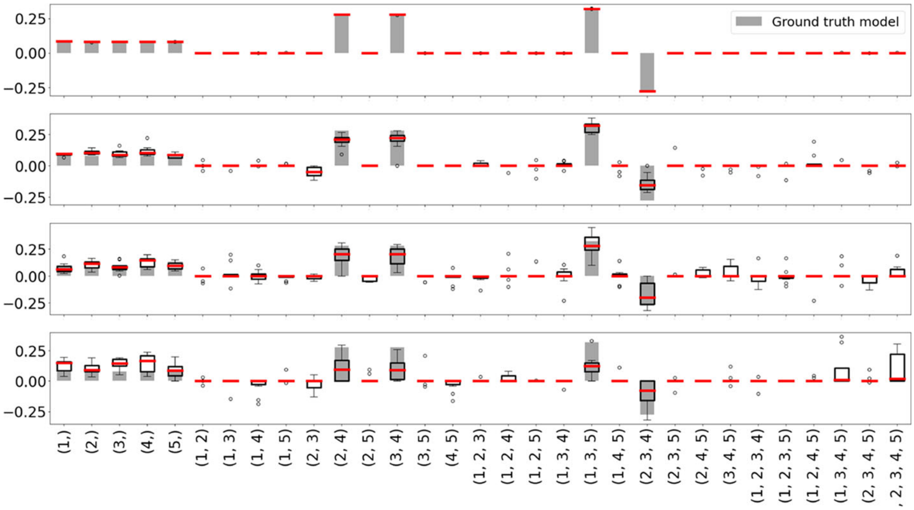

Fig. 10 Ground truth model (grey bar) and boxplots of the learned Möbius masses over 10 simulations for an increasing noise amplitude $\sigma \in \{0, 0.03, 0.05, 0.1\}$ from top to bottom

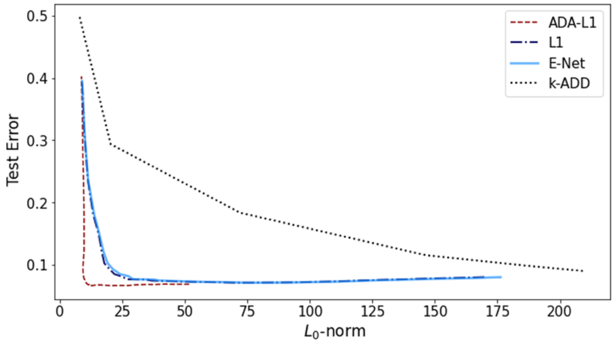

Fig. 11 Tradeoff between test error and L_0-norm depending on λ (for ADA-L1, L1 and E-Net) and k (for k-ADD)

Our approach (ADA-L1) clearly outperforms all the methods in terms of compactness, distance to the ground truth model and false discovery rate. Concerning generalizing performance, ADA-L1 outperforms the methods based on k-additive models, especially for $k = 2$ which performs very poorly. The other regularization methods (E-Net and L1) maintain competitive generalizing performance but incorporate non-null ground truth coefficients in the model as the higher falser discovery rate and L_0-norm suggest it. Note that, while having a generalizing performance close to the optimum, the unpenalized regression (No Reg.) provides a dense model and thus is unable to recover an underlying sparse model. As a consequence this method yields a null false exclusion rate. On Fig. 11 we show the evaluations obtained for each method using both the generalizing performance (test error) and the number of non-null Möbius masses (L_0-norm). Each curve represents various possible tradeoffs between the test error and the L_0-norm obtained for different values of the regularization hyperparameter λ (for ADA-L1, L1, E-Net) or for different values of k (for k-ADD). For the methods ADA-L1 and E-Net, λ_2 has been priorely set to its best value. We observe that our approach with adaptive L_1-penalty provides significantly better compromises than all the other methods. Moreover, k-additive models perform very poorly, providing models with high L_0-norm and high test error.

Finally, we conducted the same experiment with sparse bi-CIU hidden models and results are presented in Table 3. The results for the learning of both capacities m_v and m_w are averaged producing a unique result. Here again ADA-L1 produces significantly better results than the other methods in terms of generalizing performance, compactness and false discovery

Table 3 Evaluation of ADA-L1 and the baseline methods with sparse bi-CIU hidden models

	Test error	L_0-norm	$\|\hat{m} - m^*\|_2^2$	FDR	FER
ADA-L1	**0.06 ± 0.01**	**16.4 ± 12.9**	2.06 ± 0.82	**0.73 ± 0.11**	**0.02 ± 0.01**
L1	0.07 ± 0.02	25.8 ± 15.8	2.07 ± 0.81	0.84 ± 0.06	**0.02 ± 0.01**
E-Net	0.07 ± 0.01	35.8 ± 16.7	**1.15 ± 0.86**	0.85 ± 0.04	**0.02 ± 0.01**
No Reg.	0.09 ± 0.02	217.1 ± 36.3	58.74 ± 117.65	0.98 ± 0.02	**0.02 ± 0.01**
2-ADD	0.30 ± 0.16	20.6 ± 3.8	2.55 ± 0.78	0.95 ± 0.06	**0.02 ± 0.01**
4-ADD	0.12 ± 0.05	243.9 ± 72.1	6.94 ± 5.97	0.98 ± 0.02	**0.02 ± 0.02**
k*-ADD	0.09 ± 0.04	215.8 ± 141.5	13.03 ± 14.62	0.94 ± 0.07	**0.02 ± 0.01**

Table 4 Average L_0-norm for ADA-L1 and for the baselines on real datasets

	ESL	CITY	CPU	MPG
ADA-L1	**5.42 ± 2.38**	6.14 ± 3.82	**6.11 ± 1.83**	7.69 ± 2.48
L1	5.73 ± 2.81	6.69 ± 4.29	7.81 ± 3.45	**7.58 ± 2.8**
E-Net	5.93 ± 2.93	6.99 ± 5.67	17.44 ± 13.16	23.44 ± 12.51
No Reg.	12.71 ± 1.60	23.26 ± 4.77	42.04 ± 9.15	55.83 ± 14.69
2-ADD	7.80 ± 1.19	9.09 ± 1.61	9.73 ± 1.84	8.21 ± 1.58
4-ADD	12.71 ± 1.60	22.58 ± 4.42	36.73 ± 7.47	36.54 ± 12.29
k^*-ADD	5.77 ± 2.77	**5.97 ± 3.52**	12.05 ± 9.94	12.79 ± 12.12

rate. Concerning distance to the ground truth, the elastic net penalty provides slightly better results. Remark that all methods perform equally in terms of false exclusion rate.

5.3.4 Real data

In this subsection, we test our method for learning sparse Möbius capacity representations on real preference datasets. For this, we use standard monotonic multicriteria decision-making datasets containing overall evaluations of alternatives described by continuous or discrete criteria. Using these datasets, we make the assumption that the learning examples are directly expressed in terms of utilities.

We use the dataset *Employee Selection* (ESL) from the Weka repository [1], the datasets CPU[2] and Car MPG[3] (MPG) from the UCI repository and the *Movehub city ranking*[4](CITY) dataset. Below, we briefly describe the four datasets:

- ESL: psychologists evaluations on $n = 4$ criteria of some candidates (488) and overall suitability to a position.
- CITY: overall evaluations of quality of life in some cities (216) and $n = 5$ associated descriptors, e.g., purchase power, quality and access to health care.
- CPU: relative performance of some CPUs (209) and $n = 6$ associated technical characteristics, e.g., machine cycle time in nanoseconds, cache memory in kilobytes.
- MPG: city-cycle fuel consumption in miles per gallon of some cars (398) and $n = 7$ associated technical characteristics, e.g., weight, acceleration, model year.

These datasets of overall evaluations are turned into datasets of preference and indifference statements by randomly drawing pairs of alternatives (without replacing them) and comparing their global scores. The criteria associated with a decreasing monotonicity are multiplied by -1 and the utility values are made commensurate by means of linear normalization.

We compare ADA-L1 and the baseline methods in terms of test error (average magnitude of preference violation on the test sets) and number of non-null coefficients of the learned models (L_0-norm). The results are averaged over 100 simulations for each dataset. For each simulation, the models are trained on 80% of the dataset and tested over the 20% left with a random split. In Table 4 are presented the average L_0-norm of the learned models for the

[1] https://www.openml.org

[2] https://archive.ics.uci.edu/dataset/29/computer+hardware

[3] https://archive.ics.uci.edu/dataset/9/auto+mpg

[4] https://www.kaggle.com/datasets/blitzr/movehub-city-rankings

Table 5 Average test error for ADA-L1 on real datasets

ESL	CITY	CPU	MPG
0.22 ± 0.04	0.05 ± 0.03	0.12 ± 0.05	0.15 ± 0.07

different methods. The results leading to the smaller L_0-norms are highlighted in bold and the second-best results are underlined. ADA-L1 provides very sparse models with significantly lower L_0-norms than the one obtained with the baseline methods. This model compacity is obtained at no cost in terms of generalizing performance since ADA-L1 provides test errors similar to the baseline methods. We indeed performed pairwise t-tests to test the significance of the difference in test error between all the methods and we obtained p-values of magnitude 0.5. The test error numerical values obtained for ADA-L1 are provided in Table 5. This suggests that ADA-L1 is able to identify the few criteria coalitions that really matter in the preference value system underlying each dataset.

6 Conclusion

We have introduced a new approach to learn both utilities and capacities in CIU and bi-CIU models in the context of multicriteria decision making. We first proposed a variant of the tradeoff method to learn one-dimensional utility functions which appears to be more robust than usual elicitation methods based on standard sequences. Then we presented a method to learn compact representations of capacities in terms of Möbius masses using adaptive L_1-regularization. It determines where are the Möbius masses that really matter to define the capacity. This reveals those interacting subsets of criteria that must be kept in the general Choquet model to fit the observed preferences. One important advantage of this approach is that interacting subsets of any size can be included in the model. No prior restriction on the size of interaction factors is made, they are derived from the database of preference examples.

An important aspect concerns the complexity of the learning task. The linear reformulation of problem $\mathcal{P}'$ introduced in Section 4 includes $2^{n+1} + 2|I| + |P|$ variables and $\sum_{k=1}^{n} k\binom{n}{k} + |I| + |P| + 1$ constraints. Therefore the problem to be solved grows exponentially with the number of criteria. It remains tractable up to a dozen of criteria which covers most of practical cases[5]. In order to improve scalability of the method, several options could be investigated but this goes beyond the scope of this paper. First, a hierarchical structure over criteria can be used which may drastically reduce the number of criteria to be aggregated at every level and therefore the size of the learning problem. This idea was implemented in [22] to learn 2-additive capacities and could be extended to learn general capacities [45]. Another option would be to use the dual formulation of the optimization problem $\mathcal{P}'$ as in kernel-based machine learning methods. Some recent attempts in this direction are proposed in [46].

Beside scalability, several natural extensions of this work could be considered. First, the construction of compact representations of CIU is based on (4) that combines Möbius masses and terms of type $\min_{i \in B}\{u_i\}$. Alternatives representations exist for CIU and bi-CIU, combining Möbius masses and factors of type $\max_{i \in B}\{u_i\}$. They could lead to compact representations as well. This suggests extending our approach and combining min and max factors to produce even more compact representations of capacities. Another extension could

[5] criteria represent the evaluations dimensions and must not be confused with attributes describing the alternatives that may be more numerous. It is generally the case that several attributes contribute to the definition of one criterion

be to adapt our approach to other decision models allowing interacting criteria. For example, the multilinear utility model [47] admits a representation in terms of Möbius masses similar to (4) where min factors are substituted by products $\prod_{i \in B} u_i$. Clearly, the learning approach we have proposed here for the capacity identification also applies to this model with very minor modifications. Finally, interactions terms occurring in the model may be more general than min, max or product of criterion values, and could also be learned from preference data. This would be helpful to learn GAI models that are, by definition, decomposable as the sum of utility factors defined on a collection of non-necessarily disjoint subsets of attributes [48–50]. It is likely that the approach proposed here to learn the active interacting coalitions may also be applied to the determination of the relevant factors in a GAI model.

Appendix A

In this appendix we explain how equations on utilities (5)–(6) and (7)–(10) are derived from preference and indifference statements. The explanation is directly given for the bi-CIU model but can be adapted to the particular case of the CIU model by taking $w = v$.

A.1 Utility elicitation with solvability

A.1.1 Utility elicitation below the neutral level with solvability assumption

Proposition 1 *Let $x_i, h_i \in X_i$ such that $x_i \precsim_i \mathbf{0}_i$ and $h_i \precsim_i \mathbf{0}_i$ and $r_j, R_j \in X_j$ such that $\mathbf{0}_j \precsim_j r_j \prec_j R_j$. Suppose that $(x_i, r_j, \mathbf{0}_{-ij}) \sim (y_i, R_j, \mathbf{0}_{-ij})$ and $(h_i, r_j, \mathbf{0}_{-ij}) \sim (z_i, R_j, \mathbf{0}_{-ij})$. Assuming $(-\mathbf{1}_i, \mathbf{0}_{-i}) \prec \mathbf{0}$, we have:*

$$u_i(y_i) - u_i(x_i) = u_i(z_i) - u_i(y_i)$$

Proof From $(x_i, r_j, \mathbf{0}_{-ij}) \sim (y_i, R_j, \mathbf{0}_{-ij})$ we have: $f^u_{v,w}(x_i, r_j, \mathbf{0}_{-ij}) = f^u_{v,w}(y_i, R_j, \mathbf{0}_{-ij})$. Also, since $r_j \prec_j R_j$ we have $y_i \prec_i x_i$. Moreover, since $x_i \precsim_i \mathbf{0}_i$ and $\mathbf{0}_j \precsim_j r_j \prec_j R_j$, we have $u_i(y_i) < u_i(x_i) \leq 0 \leq u_j(r_j) < u_j(R_j)$ and therefore $f^u_{v,w}(x_i, r_j, \mathbf{0}_{-ij}) = u_j(r_j)v(\{j\}) + u_i(x_i)(1 - w(N \setminus \{i\}))$.

Similarly, we have: $f^u_{v,w}(y_i, R_j, \mathbf{0}_{-ij}) = u_j(R_j)v(\{j\}) + u_i(y_i)(1 - w(N \setminus \{i\}))$. Hence we have: $(u_i(x_i) - u_i(y_i))(1 - w(N \setminus \{i\})) = (u_j(R_j) - u_j(r_j))v(\{j\})$.

Moreover, using the second indifference $(h_i, r_j, \mathbf{0}_{-ij}) \sim (z_i, R_j, \mathbf{0}_{-ij})$, we obtain $(u_i(h_i) - u_i(z_i))(1 - w(N \setminus \{i\})) = (u_j(R_j) - u_j(r_j))v(\{j\})$. Then $(u_i(x_i) - u_i(y_i))(1 - w(N \setminus \{i\})) = (u_i(h_i) - u_i(z_i))(1 - w(N \setminus \{i\}))$. Assuming $(-\mathbf{1}_i, \mathbf{0}_{-i}) \prec \mathbf{0}$, i.e., $w(N \setminus \{i\}) < 1$ we obtain:

$$u_i(x_i) - u_i(y_i) = u_i(h_i) - u_i(z_i)$$

□

A.1.2 Utility elicitation above the neutral level with solvability assumption

Proposition 2 *Let $x_i, h_i \in X_i$ such that $x_i \succsim_i \mathbf{0}_i$ and $h_i \succsim_i \mathbf{0}_i$ and $r_j, R_j \in X_j$ and $x_i \in X_i$ such that $r_j \prec_j R_j \precsim_j \mathbf{0}_j$. Suppose that $(x_i, R_j, \mathbf{0}_{-ij}) \sim (y_i, r_j, \mathbf{0}_{-ij})$ and $(h_i, R_j, \mathbf{0}_{-ij}) \sim (z_i, r_j, \mathbf{0}_{-ij})$. Assuming $(\mathbf{1}_i, \mathbf{0}_{-i}) \succ \mathbf{0}$, we have:*

$$u_i(x_i) - u_i(y_i) = u_i(h_i) - u_i(z_i)$$

Proof From $(x_i, R_j, \mathbf{0}_{-ij}) \sim (y_i, r_j, \mathbf{0}_{-ij})$, we have: $f^u_{v,w}(x_i, R_j, \mathbf{0}_{-ij}) = f^u_{v,w}(y_i, r_j, \mathbf{0}_{-ij})$. Since $r_j \prec_j R_j$ we have $x_i \prec_i y_i$. Moreover, since $x_i \succsim_i \mathbf{0}_i$ and $r_j \prec_j R_j \precsim_j \mathbf{0}_j$, we have $u_j(r_j) < u_j(R_j) \leq 0 \leq u_i(x_i) < u_i(y_i)$ and therefore $f^u_{v,w}(x_i, R_j, \mathbf{0}_{-ij}) = u_i(x_i)v(\{i\}) + u_j(R_j)(1 - w(N \setminus \{j\}))$.

Similarly $f^u_{v,w}(y_i, r_j, \mathbf{0}_{-ij}) = u_i(y_i)v(\{i\}) + u_j(r_j)(1 - w(N \setminus \{j\}))$. Hence we have: $(u_i(y_i) - u_i(x_i))v(\{i\}) = (u_j(R_j) - u_j(r_j))(1 - w(N \setminus \{j\}))$.

Moreover, using the second indifference $(h_i, R_j, \mathbf{0}_{-ij}) \sim (z_i, r_j, \mathbf{0}_{-ij})$, we obtain $(u_i(z_i) - u_i(h_i))v(\{i\}) = (u_j(R_j) - u_j(r_j))(1 - w(N \setminus \{j\}))$. Then $(u_i(y_i) - u_i(x_i))v(\{i\}) = (u_i(z_i) - u_i(h_i))v(\{i\})$. Assuming $(\mathbf{1}_i, \mathbf{0}_{-i}) \succ \mathbf{0}$, i.e., $v(\{i\}) > 0$ we obtain:

$$u_i(y_i) - u_i(x_i) = u_i(z_i) - u_i(h_i)$$

□

A.2 Utility elicitation without solvability

A.2.1 Utility elicitation below the neutral level without solvability assumption

Proposition 3 *Assume that the elements of X_i are denoted $x_{i,k}$ and indexed according to their relative values: $x_{i,k} \precsim_i x_{i,k+1}$, for any k. Let $x_i, h_i \in X_i$ such that $x_i \precsim_i \mathbf{0}_i$ and $h_i \precsim_i \mathbf{0}_i$ and $r_j, R_j \in X_j$ such that $\mathbf{0}_j \precsim_j r_j \prec_j R_j$. Let k be the lower integer such that $(x_i, r_j, \mathbf{0}_{-ij}) \precsim (x_{i,k+1}, R_j, \mathbf{0}_{-ij})$ and k' the highest integer such that $(h_i, r_j, \mathbf{0}_{-ij}) \succsim (x_{i,k'}, R_j, \mathbf{0}_{-ij})$. Assuming $(-\mathbf{1}_i, \mathbf{0}_{-i}) \prec \mathbf{0}$, we have:*

$$u_i(h_i) - u_i(z_i^-) \geq u_i(x_i) - u_i(y_i^+)$$
$$u_i(h_i) - u_i(z_i^+) < u_i(x_i) - u_i(y_i^-)$$

where $y_i^+ = x_{i,k+1}$, $y_i^- = x_{i,k}$, $z_i^- = x_{i,k'}$ and $z_i^+ = x_{i,k'+1}$.

Proof By construction y_i^- necessarily verifies the following strict preference: $(x_i, r_j, \mathbf{0}_{-ij}) \succ (y_i^-, R_j, \mathbf{0}_{-ij})$. Hence, with $(x_i, r_j, \mathbf{0}_{-ij}) \precsim (y_i^+, R_j, \mathbf{0}_{-ij})$, we obtain the following inequations: $(u_i(x_i) - u_i(y_i^+))(1 - w(N \setminus \{i\})) \leq (u_i(R_j) - u_i(r_j))v(\{j\})$ and $(u_i(x_i) - u_i(y_i^-))(1 - w(N \setminus \{i\})) > (u_i(R_j) - u_i(r_j))v(\{j\})$.

Similarly, z_i^+ verify $(h_i, r_j, \mathbf{0}_{-ij}) \prec (z_i^+, R_j, \mathbf{0}_{-ij})$. Hence, with $(h_i, r_j, \mathbf{0}_{-ij}) \succsim (z_i^-, R_j, \mathbf{0}_{-ij})$ we obtain the following inequations: $(u_i(h_i) - u_i(z_i^-)(1 - w(N \setminus \{i\})) \geq (u_i(R_j) - u_i(r_j))v(\{j\})$ and $(u_i(h_i) - u_i(z_i^+))(1 - w(N \setminus \{i\})) < (u_i(R_j) - u_i(r_j))v(\{j\})$. Hence we have $(u_i(x_i) - u_i(y_i^+))(1 - w(N \setminus \{i\})) \leq (u_i(R_j) - u_i(r_j))v(\{j\}) \leq (u_i(h_i) - u_i(z_i^-))(1 - w(N \setminus \{j\}))$.

Moreover, $(u_i(h_i) - u_i(z_i^+)(1 - w(N \setminus \{i\})) < (u_i(R) - u_i(r))v(\{j\}) < (u_i(x_i) - u_i(y_i^-)(1 - w(N \setminus \{i\}))$. Assuming $(-\mathbf{1}_i, \mathbf{0}_{-i}) \prec \mathbf{0}$, i.e., $w(N \setminus \{i\}) < 1$, we obtain:

$$u_i(h_i) - u_i(z_i^-) \geq u_i(x_i) - u_i(y_i^+)$$
$$u_i(h_i) - u_i(z_i^+) < u_i(x_i) - u_i(y_i^-)$$

□

A.2.2 Utility elicitation above the neutral level without solvability assumption

Proposition 4 *Assume that the elements of X_i are denoted $x_{i,k}$ and indexed according to their relative values: $x_{i,k} \precsim_i x_{i,k+1}$, for any k. Let $x_i, h_i \in X_i$ such that $x_i \succsim_i \mathbf{0}_i$ and*

$h_i \succsim_i \mathbf{0}_i$ and r_j, $R_j \in X_j$ such that $r_j \prec_j R_j \precsim_j \mathbf{0}_j$. Let k be the higher integer such that $(x_i, R_j, \mathbf{0}_{-ij}) \succsim (x_{i,k}, r_j, \mathbf{0}_{-ij})$ and k' the lower integer such that $(h_i, R_j, \mathbf{0}_{-ij}) \precsim (x_{i,k+1}, r_j, \mathbf{0}_{-ij})$. Assuming $(\mathbf{1}_i, \mathbf{0}_{-i}) \succ \mathbf{0}$, we have:

$$\begin{aligned} u_i(h_i) - u_i(z_i^+) &\le u_i(x_i) - u_i(y_i^-) \\ u_i(h_i) - u_i(z_i^-) &> u_i(x_i) - u_i(y_i^+) \end{aligned}$$

where $y_i^+ = x_{i,k+1}$, $y_i^- = x_{i,k}$, $z_i^- = x_{i,k'}$ and $z_i^+ = x_{i,k'+1}$.

Proof By construction y_i^+ necessarily verifies the following strict preference: $(x_i, R_j, \mathbf{0}_{-ij}) \prec (y_i^+, r_j, \mathbf{0}_{-ij})$. Hence, with $(x_i, R_j, \mathbf{0}_{-ij}) \succsim (y_i^-, r_j, \mathbf{0}_{-ij})$, we obtain the following inequations: $(u_i(x_i) - u_i(y_i^-))v(\{i\}) \ge (u_i(r_j) - u_i(R_j))(1 - w(N \setminus \{j\}))$ and $(u_i(x_i) - u_i(y_i^+))v(\{i\}) < (u_i(r_j) - u_i(R_j))(1 - w(N \setminus \{j\}))$.

Similarly z_i^- verify $(h_i, R_j, \mathbf{0}_{-ij}) \succ (z_i^-, r_j, \mathbf{0}_{-ij})$. Hence, with $(h_i, R_j, \mathbf{0}_{-ij}) \precsim (z_i^+, r_j, \mathbf{0}_{-ij})$, we obtain the following inequations: $(u_i(h_i) - u_i(z_i^+))v(\{i\}) \le (u_i(r_j) - u_i(R_j))(1 - w(N \setminus \{j\}))$ and $(u_i(h_i) - u_i(z_i^-))v(\{i\}) > (u_i(r_j) - u_i(R_j))(1 - w(N \setminus \{j\}))$. Hence we have $(u_i(h_i) - u_i(z_i^+))v(\{i\}) \le (u_i(R_j) - u_i(r_j))(1 - w(N \setminus \{j\})) \le (u_i(x_i) - u_i(y_i^-))v(\{i\})$. Moreover, $(u_i(x_i) - u_i(y_i^+))v(\{i\}) < (u_i(R_j) - u_i(r_j))(1 - w(N \setminus \{j\})) < (u_i(h_i) - u_i(z_i^-))v(\{i\})$. Assuming $(\mathbf{1}_i, \mathbf{0}_{-i}) \succ \mathbf{0}$, i.e., $v(\{i\}) > 0$, we obtain:

$$\begin{aligned} u_i(h_i) - u_i(z_i^+) &\le u_i(x_i) - u_i(y_i^-) \\ u_i(h_i) - u_i(z_i^-) &> u_i(x_i) - u_i(y_i^+) \end{aligned}$$

□

Appendix B

In the following, we use the convention that for any subset $B \subseteq N$, $\int_{\mathcal{I}_B} f(z_1, \dots, z_n) dz_B$ denotes the multiple integral of the function f w.r.t. the arguments z_i, $i \in B$ on the hypercube $\mathcal{I}_B = [0, 1]^{|B|}$.

Lemma 5 *Let $B \subseteq N \setminus \emptyset$ and $k \in \mathbb{N}^*$, then the following equality holds:*

$$\int_{\mathcal{I}_B} \min_{i \in B}\{z_i\}^k dz_B = \frac{k!}{\prod_{i=1}^{k}(|B| + i)}$$

Proof Consider $|B|$ random variables $(Z_i)_{i \in B}$ independent and identically distributed according a uniform distribution within $[0, 1]$. It can be easily shown that the random variable $Y = \min_{i \in B}\{Z_i\}$ admits the following density function:

$$f_Y(y) = \begin{cases} |B|(1 - y)^{|B|-1} \text{ if } y \in [0, 1] \\ 0 \text{ else.} \end{cases}$$

Then we obtain:

$$
\begin{aligned}
\mathbb{E}[Y^k] = \int_0^1 y^k|B|(1-y)^{|B|-1}dy &= k\int_0^1 y^{k-1}(1-y)^{|B|}dy \\
&= k(k-1)\int_0^1 y^{k-2}\frac{(1-y)^{|B|+1}}{|B|+1}dy \\
&= \ldots \\
&= k!\int_0^1 y^{k-k}\frac{(1-y)^{|B|+k-1}}{\prod_{i=1}^{k-1}(|B|+i)}dy \\
&= \frac{k!}{\prod_{i=1}^{k-1}(|B|+i)}\frac{1}{(|B|+k)} = \frac{k!}{\prod_{i=1}^{k}(|B|+i)}
\end{aligned}
$$

Finally, we conclude:

$$
\int_{\mathcal{I}_B}\min_{i\in B}\{z_i\}^k dz_B = \mathbb{E}[\min_{i\in B}\{Z_i\}^k] = \mathbb{E}[Y^k] = \frac{k!}{\prod_{i=1}^{k}(|B|+i)}
$$

□

Lemma 6 *Let $n \leq 3$ and $B_1, B_2 \subseteq N$ such that $B_1 \cap B_2 = \emptyset$ and $B_2 \neq \emptyset$. For any vector $(z_j)_{j\in B_2}$ taking values in $[0,1]$, the following equality holds:*

$$
\int_{\mathcal{I}_{B_1}}\min_{i\in B_2\cup B_1}\{z_i\}dz_{B_1} = \wedge_{B_2} - \frac{|B_1|\wedge_{B_2}^2}{2} + \frac{|B_1|(|B_1|-1)^+\wedge_{B_2}^3}{6}
$$

with $\wedge_{B_2} = \min_{i\in B_2}\{z_i\}$ and $x^+ = \max\{0, x\}$.

Proof Firstly, for any $A \subseteq N\setminus\emptyset$, any vector $(z_i)_{i\in A}$ valued in $[0,1]$ and any $k\in\mathbb{N}$, we have:

$$
\begin{aligned}
\int_0^1 \min\{x, \wedge_A\}^k dx &= \int_0^{\wedge_A}\min\{x,\wedge_A\}^k dx + \int_{\wedge_A}^1 \min\{x,\wedge_A\}^k dx \\
&= \int_0^{\wedge_A} x^k dx + \int_{\wedge_A}^1 \wedge_A^k = \frac{\wedge_A^{k+1}}{k+1} + (1-\wedge_A)\wedge_A^k \\
&= \wedge_A^k - \frac{k}{k+1}\wedge_A^{k+1} \qquad \text{(B1)}
\end{aligned}
$$

where $\wedge_A = \min_{i\in A}\{z_i\}$.

Remark that for $B_1 = \emptyset$, the left-hand term boils down to $\wedge_{B_2}$ which equals the right-hand term for $|B_1| = 0$. Suppose now that $B_1 \neq \emptyset$. Since $n \leq 3$, $B_2 \neq \emptyset$ and $B_1 \cap B_2 = \emptyset$, B_1 is necessarily a singleton or a pair, then we have: $|B_1| \in \{1, 2\}$.

Then let $(\pi_1, \ldots, \pi_{|B_1|})$ be any ordering of the elements of B_1. Using (B1) with $A = (B_1 \cup B_2)\setminus\{\pi_1\}$, $x = z_{\pi_1}$ and $k = 1$, we have:

$$
\begin{aligned}
\int_{\mathcal{I}_{B_1}}\min_{i\in B_2\cup B_1}\{z_i\}dz_{B_1} &= \int_{\mathcal{I}_{B_1\setminus\{\pi_1\}}}\Big(\int_0^1 \min_{i\in B_2\cup B_1}\{z_i\}dz_{\pi_1}\Big)dz_{B_1\setminus\{\pi_1\}} \\
&= \int_{\mathcal{I}_{B_1\setminus\{\pi_1\}}}\Big(\wedge_{(B_1\cup B_2)\setminus\{\pi_1\}} - \frac{\wedge^2_{(B_1\cup B_2)\setminus\{\pi_1\}}}{2}\Big)dz_{B_1\setminus\{\pi_1\}}
\end{aligned}
$$

Then if $|B_1| = 1$, we have $B_1 \setminus \{\pi_1\} = \emptyset$ and therefore we obtain:

$$\begin{aligned}\int_{\mathcal{I}_{B_1}} \min_{i\in B_2\cup B_1}\{z_i\}dz_{B_1} &= \wedge_{(B_1\cup B_2)\setminus\{\pi_1\}} - \frac{\wedge^2_{(B_1\cup B_2)\setminus\{\pi_1\}}}{2}\\ &= \wedge_{B_2} - \frac{|B_1|\wedge^2_{B_2}}{2} + \frac{|B_1|(|B_1|-1)^+\wedge^3_{B_2}}{6}\end{aligned}$$

Finally, if $|B_1| = 2$, we have $B_1\setminus\{\pi_1\} = \{\pi_2\}$ and using (B1) for $A = (B_1\cup B_2)\setminus\{\pi_1, \pi_2\} = B_2$, $x = z_{\pi_2}$ and $k \in \{1, 2\}$, we obtain:

$$\begin{aligned}\int_{\mathcal{I}_{B_1}} \min_{i\in B_2\cup B_1}\{z_i\}dz_{B_1} &= \int_0^1 \Big(\wedge_{(B_1\cup B_2)\setminus\{\pi_1\}} - \frac{\wedge^2_{(B_1\cup B_2)\setminus\{\pi_1\}}}{2}\Big)dz_{\pi_2}\\ &= \int_0^1 \Big(\min_{i\in B_2\cup\{\pi_2\}}\{z_i\} - \frac{\min_{i\in B_2\cup\{\pi_2\}}\{z_i\}^2}{2}\Big)dz_{\pi_2}\\ &= \wedge_{B_2} - \frac{|B_1|\wedge^2_{B_2}}{2} + \frac{|B_1|(|B_1|-1)^+\wedge^3_{B_2}}{6}\end{aligned}$$

□

Proposition 7 *Let $n \leq 3$. Assume that the features $(\phi_j)_{j=1}^{2^n}$ are defined such that $\phi_j = \min_{i\in\rho^{-1}(j)}\{Z_i\}$ where $(Z_i)_{i\in N}$ are i.i.d. random variables such that $Z_i \sim \mathcal{U}([0,1])$ for any $i \in N$. Then, for any pair of criteria coalition $S_1, S_2 \subseteq N\setminus\emptyset$ of cardinals $|S_1| = s_1$,$|S_2| = s_2$ and $|S_1 \cap S_2| = s_{12} \neq 0$, the covariance between $\phi_{\rho(S_1)}$ and $\phi_{\rho(S_2)}$ reads as follows:*

$$Cov(\phi_{\rho(S_1)}, \phi_{\rho(S_2)}) = \sum_{k=1}^{3} g_k(s_{12})\gamma_k(s_1, s_2, s_{12}) - \frac{1}{(s_1+1)(s_2+1)}$$

with $g_k(s_{12}) = \frac{k!}{\prod_{i=1}^{k}(s_{12}+i)}$, $\gamma_1 = 1$, $\gamma_2(s_1, s_2, s_{12}) = -\frac{1}{2}((s_1 - s_{12})^+ + (s_2 - s_{12})^+)$ and $\gamma_3(s_1, s_2, s_{12}) = \frac{1}{4}((s_1-s_{12})^+(s_2-s_{12})^+) + \frac{1}{6}((s_1-s_{12})^+(s_1-s_{12}-1)^+ + (s_2-s_{12})^+(s_2-s_{12}-1)^+)$ where $x^+ = \max\{x, 0\}$.

Proof Let $S_1, S_2 \subseteq N \setminus \emptyset$ such that $S_1 \cap S_2 \neq 0$. Firstly, recall that:

$$Cov(\phi_{\rho(S_1)}, \phi_{\rho(S_2)}) = \mathbb{E}[\phi_{\rho(S_1)}\phi_{\rho(S_2)}] - \mathbb{E}[\phi_{\rho(S_1)}]\mathbb{E}[\phi_{\rho(S_2)}]$$

Also, since the variables $(Z_i)_{i\in N}$ are independent and identically distributed according a uniform distribution within [0, 1], for any $S \subseteq N$, we have:

$$\mathbb{E}[\phi_{\rho(S)}] = \int_{\mathcal{I}_S} \min_{i\in S}\{z_i\}dz_S$$

Then using Lemma 5 for $k = 1$ and $B = S$, we obtain:

$$\mathbb{E}[\phi_{\rho(S)}] = \frac{1}{|S|+1} \tag{B2}$$

Moreover, since $S_1 \cap S_2 \neq \emptyset$, we have:

$$\begin{aligned}\mathbb{E}[\phi_{\rho(S_1)}\phi_{\rho(S_2)}] &= \mathbb{E}[\min_{i\in S_1}\{Z_i\}\min_{i\in S_2}\{Z_i\}]\\ &= \int_{\mathcal{I}_{S_1\cup S_2}} \min_{i\in S_1}\{z_i\}\min_{i\in S_2}\{z_i\} dz_{S_1\cup S_2}\\ &= \int_{\mathcal{I}_{S_2}} \min_{i\in S_2}\{z_i\}\Big(\int_{\mathcal{I}_{S_1\setminus S_1\cap S_2}} \min_{i\in S_1}\{z_i\} dz_{S_1\setminus(S_1\cap S_2)}\Big) dz_{S_2}\end{aligned}$$

Then, using Lemma 6 sequentially for $B_1 = S_1 \setminus (S_1 \cap S_2)$, $B_2 = S_1 \cap S_2$ and $B_1 = S_2 \setminus (S_1 \cap S_2)$, $B_2 = S_1 \cap S_2$, we obtain:

$$\begin{aligned}\mathbb{E}[\phi_{\rho(S_1)}\phi_{\rho(S_2)}] &= \int_{\mathcal{I}_{S_1\cap S_2}} \Big(\int_{\mathcal{I}_{S_2\setminus(S_1\cap S_2)}} \min_{i\in S_2}\{z_i\}\Big(\wedge_{S_1\cap S_2} - \frac{(s_1-s_{12})^+\wedge^2_{S_1\cap S_2}}{2}\\ &+ \frac{(s_1-s_{12})^+(s_1-s_{12}-1)^+\wedge^3_{S_1\cap S_2}}{6}\Big) dz_{S_2\setminus(S_1\cap S_2)}\Big) dz_{S_1\cap S_2}\\ &= \int_{\mathcal{I}_{S_1\cap S_2}} \Big(\wedge_{S_1\cap S_2} - \frac{(s_1-s_{12})^+\wedge^2_{S_1\cap S_2}}{2} + \frac{(s_1-s_{12})^+(s_1-s_{12}-1)^+\wedge^3_{S_1\cap S_2}}{6}\Big)\\ &\Big(\wedge_{S_1\cap S_2} - \frac{(s_2-s_{12})^+\wedge^2_{S_1\cap S_2}}{2} + \frac{(s_2-s_{12})^+(s_2-s_{12}-1)^+\wedge^3_{S_1\cap S_2}}{6}\Big) dz_{S_1\cap S_2}\end{aligned}$$

where $\wedge_{S_1\cap S_2} = \min_{i\in S_1\cap S_2}\{z_i\}$ for any vector $(z_i)_{i\in S_1\cap S_2}$ valued in $[0, 1]$. This expression can be simplified remarking that since $n \leq 3$ and $S_1 \cap S_2 \neq \emptyset$, we have that the cross products $(s_2-s_{12})^+(s_2-s_{12}-1)^+(s_1-s_{12})^+(s_1-s_{12}-1)^+$, $(s_2-s_{12})^+(s_2-s_{12}-1)^+(s_1-s_{12})^+$ and $(s_1-s_{12})^+(s_1-s_{12}-1)^+(s_2-s_{12})^+$ necessarily equal zero. Finally, using Lemma 5 for $B = S_1 \cap S_2$ and $k \in \{2, 3, 4\}$, we obtain that:

$$\begin{aligned}\mathbb{E}[\phi_{\rho(S_1)}\phi_{\rho(S_2)}] &= g_2(s_{12}) - g_3(s_{12})\frac{1}{2}((s_1-s_{12})^+ + (s_2-s_{12})^+)\\ &+ g_4(s_{12})(\frac{1}{4}((s_1-s_{12})^+(s_2-s_{12})^+)\\ &+ \frac{1}{6}((s_1-s_{12})^+(s_1-s_{12}-1)^+ + (s_2-s_{12})^+(s_2-s_{12}-1)^+)) \qquad \text{(B3)}\end{aligned}$$

with $g_k(s_{12}) = \frac{k!}{\prod_{i=1}^k(s_{12}+i)}$. We obtain the final result by combining (B3) and (B2). □

Proposition 8 *Suppose that $n = 3$ and that the underlying model β^* is such that $A_1 = \{\{1\}, \{2\}, \{3\}, \{1, 2, 3\}\}$, $A_2 = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ and $\text{sign}(\beta^*_{A_1}) = \mathbf{1}$. Assume also that the features $(\phi_j)_{j=1}^{2^n}$ are defined such that $\phi_j = \min_{i\in\rho^{-1}(j)}\{Z_i\}$ where $(Z_i)_{i\in N}$ are i.i.d. random variables such that $Z_i \sim \mathcal{U}([0, 1])$ for any $i \in N$. Then, Condition (19) boils down to:*

$$|2V^1_{1,2}(V^3_{3,3} - V^1_{1,3}) + V^2_{2,3}(V^1_{1,1} - 3V^1_{1,3})| < |V^3_{3,3}V^1_{1,1} - 3(V^1_{1,3})^2|$$

where $V^l_{j,k} = Cov(\phi_{\rho(S_1)}, \phi_{\rho(S_2)}) = C_{\rho(S_1),\rho(S_2)}$ for any pair of criteria coalition $S_1, S_2 \subseteq N$ such that $j = |S_1|$, $k = |S_2|$ and $l = |S_1 \cap S_2|$.

Proof Since the random variables $(Z_i)_{i\in N}$ are supposed independent, for any pair of criteria coalition $S_1, S_2 \subseteq N$ such that $S_1 \cap S_2 = \emptyset$, we have: $Cov(\phi_{\rho(S_1)}, \phi_{\rho(S_2)}) = 0$. Also, indexing the columns and rows of C_{11} and C_{21} in the lexicographical order, we have:

$$C_{11} = \begin{pmatrix} V^1_{1,1} & 0 & 0 & V^1_{1,3} \\ 0 & V^1_{1,1} & 0 & V^1_{1,3} \\ 0 & 0 & V^1_{1,1} & V^1_{1,3} \\ V^1_{1,3} & V^1_{1,3} & V^1_{1,3} & V^3_{3,3} \end{pmatrix} \text{ and } C_{21} = \begin{pmatrix} V^1_{1,2} & V^1_{1,2} & 0 & V^2_{2,3} \\ V^1_{1,2} & 0 & V^1_{1,2} & V^2_{2,3} \\ 0 & V^1_{1,2} & V^1_{1,2} & V^2_{2,3} \end{pmatrix}$$

Then C_{11} can be rewritten as a block-matrix as follows:

$$C_{11} = \begin{pmatrix} M_1 & M_2^T \\ M_2 & M_3 \end{pmatrix} \text{ with } M_1 = \begin{pmatrix} V^1_{1,1} & 0 & 0 \\ 0 & V^1_{1,1} & 0 \\ 0 & 0 & V^1_{1,1} \end{pmatrix}, M_2 = \begin{pmatrix} V^1_{1,3} & V^1_{1,3} & V^1_{1,3} \end{pmatrix}, M_3 = \begin{pmatrix} V^3_{3,3} \end{pmatrix}$$

Since C_{11} is the covariance matrix of the random variables $(\phi_j)_{j \in A_1}$ it is a positive semi-definite matrix. Also, remarking that M_1 is inversible and using Proposition 7, one can compute the Schur complement $S = M_3 - M_2^T M_1^{-1} M_2 = \frac{V^3_{3,3} V^1_{1,1} - 3 V^1_{1,3}}{V^1_{1,1}} \neq 0$ and obtain that C_{11} is a positive definite matrix the inverse of which reads as follows:

$$\begin{aligned} C_{11}^{-1} &= \begin{pmatrix} M_1^{-1} + M_1^{-1} M_2^T S^{-1} M_2 M_1^{-1} & -M_1^{-1} M_2^T S^{-1} \\ -S^{-1} M_2 M_1^{-1} & S^{-1} \end{pmatrix} \\ &= \frac{1}{S V^1_{1,1}} \begin{pmatrix} S + \frac{(V^1_{1,3})^2}{V^1_{1,1}} & \frac{(V^1_{1,3})^2}{V^1_{1,1}} & \frac{(V^1_{1,3})^2}{V^1_{1,1}} & -V^1_{1,3} \\ \frac{(V^1_{1,3})^2}{V^1_{1,1}} & S + \frac{(V^1_{1,3})^2}{V^1_{1,1}} & \frac{(V^1_{1,3})^2}{V^1_{1,1}} & -V^1_{1,3} \\ \frac{(V^1_{1,3})^2}{V^1_{1,1}} & \frac{(V^1_{1,3})^2}{V^1_{1,1}} & S + \frac{(V^1_{1,3})^2}{V^1_{1,1}} & -V^1_{1,3} \\ -V^1_{1,3} & -V^1_{1,3} & -V^1_{1,3} & V^1_{1,1} \end{pmatrix} \end{aligned}$$

Then we finally obtain:

$$C_{21} C_{11}^{-1} \operatorname{sign}(\beta^*_{A_1}) = \frac{2 V^1_{1,2}(V^3_{3,3} - V^1_{1,3}) + V^2_{2,3}(V^1_{1,1} - 3 V^1_{1,3})}{V^3_{3,3} V^1_{1,1} - 3(V^1_{1,3})^2} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Therefore Condition (19) is satisfied if and only if $|2 V^1_{1,2}(V^3_{3,3} - V^1_{1,3}) + V^2_{2,3}(V^1_{1,1} - 3 V^1_{1,3})| < |V^3_{3,3} V^1_{1,1} - 3(V^1_{1,3})^2|$. □

Acknowledgements We wish to thank the reviewers for their helpful comments on an earlier draft of the text.

Data Availability In this study, we mainly used synthetic datasets and we have explained in details how they have been generated ; it is not necessary to store them. For this reason, the datasets generated during the current study are not publicly available but are available from the corresponding author on reasonable request. All real-world datasets we use are publicly available.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Roy, B.: Multicriteria Methodology for Decision Aiding, vol. 12. Springer Science & Business Media (1996)
2. Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E.: Aggregation Functions, vol. 127. Cambridge University Press (2009)
3. Grabisch, M.: The application of fuzzy integrals in multicriteria decision making. European Journal of Operational Research **89**(3), 445–456 (1996)
4. Gagolewski, M., James, S., Beliakov, G.: Supervised learning to aggregate data with the Sugeno integral. IEEE Transactions on Fuzzy Systems **27**(4), 810–815 (2019)
5. Beliakov, G., Divakov, D.: On representation of fuzzy measures for learning Choquet and Sugeno integrals. Knowl. Based Syst. **189**, 105134 (2020)
6. Tehrani, A.F., Hüllermeier, E.: Ordinal Choquistic regression. In: 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-13), pp. 842–849 (2013)
7. Tehrani, A.F., Cheng, W., Hüllermeier, E.: Preference learning using the Choquet integral: the case of multipartite ranking. IEEE Transactions on Fuzzy Systems **20**(6), 1102–1113 (2012)
8. Schmeidler, D.: Subjective probability and expected utility without additivity. Econometrica **57**(3), 571–587 (1989)
9. Grabisch, M., Labreuche, C.: A decade of application of the Choquet and Sugeno integrals in multicriteria decision aid. Annals of Operations Research **175**(1), 247–286 (2010)
10. Labreuche, C., Grabisch, M.: Generalized Choquet-like aggregation functions for handling bipolar scales. European Journal of Operational Research **172**(3), 931–955 (2006)
11. Tversky, A., Kahneman, D.: An analysis of decision under risk. Econometrica **47**(2), 263–292 (1979)
12. Martin, H., Perny, P.: New computational models for the Choquet integral. In: ECAI 2020, pp. 147–154 (2020)
13. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans. Syst. Man Cybern. Syst. **18**(1), 183–190 (1988)
14. Torra, V.: The weighted OWA operator. International Journal of Intelligent Systems **12**(2), 153–166 (1997)
15. Bana e Costa, C.A., Vansnick, J.-C.: A theoretical framework for measuring attractiveness by a categorical based evaluation technique (MACBETH). In: Multicriteria Analysis: Proceedings of the XIth International Conference on MCDM, pp. 15–24 (1997)
16. Wakker, P., Deneffe, D.: Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. Manag. Sci. **42**(8), 1131–1150 (1996)
17. Abdellaoui, M.: Parameter-free elicitation of utility and probability weighting functions. Manag. Sci. **46**(11), 1497–1512 (2000)
18. Herin, M., Perny, P., Sokolovska, N.: Learning sparse representations of preferences within choquet expected utility theory. In: Uncertainty in Artificial Intelligence, pp. 800–810 (2022). PMLR
19. Grabisch, M., Kojadinovic, I., Meyer, P.: A review of methods for capacity identification in Choquet integral based multi-attribute utility theory: Applications of the Kappalab R package. Eur. J. Oper. Res. **186**(2), 766–785 (2008)
20. Tehrani, A.F., Labreuche, C., Hüllermeier, E.: Choquistic utilitaristic regression. In: DA2PL, pp. 35–42 (2014)
21. Galand, L., Mayag, B.: A heuristic approach to test the compatibility of a preference information with a Choquet integral model. In: ADT, pp. 65–80 (2017)
22. Bresson, R., Cohen, J., Hüllermeier, E., Labreuche, C., Sebag, M.: Neural representation and learning of hierarchical 2-additive Choquet integrals. In: IJCAI, pp. 1984–1991 (2020)
23. Benabbou, N., Perny, P., Viappiani, P.: Incremental elicitation of Choquet capacities for multicriteria choice, ranking and sorting problems. Artif. Intell. **246**, 152–180 (2017)
24. Bourdache, N., Perny, P., Spanjaard, O.: Incremental elicitation of rank-dependent aggregation functions based on Bayesian linear regression. In: IJCAI-19-Twenty-Eighth International Joint Conference on Artificial Intelligence, pp. 2023–2029 (2019)
25. Chateauneuf, A., Jaffray, J.-Y.: Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. Math. Soc. Sci. **17**(3), 263–283 (1989)
26. Grabisch, M.: K-order additive discrete fuzzy measures and their representation. Fuzzy Sets Syst. **92**(2), 167–189 (1997)
27. Krantz, D.H., Tversky, A.: Conjoint-measurement analysis of composition rules in psychology. Psychol. Rev. **78**(2), 151 (1971)
28. Blavatskyy, P.: Error propagation in the elicitation of utility and probability weighting functions. Theory Decis. **60**(2), 315–334 (2006)
29. Ramsay, J.O.: Monotone regression spline in action. Stat. Sci., 425–441 (1988)

30. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B. Stat. Methodol. (Methodological) **58**(1), 267–88 (1996)
31. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical learning with sparsity. Monogr. Stat. Appl. Probab. **143** (2015)
32. Anderson, D.T., Price, S.R., Havens, T.C.: Regularization-based learning of the Choquet integral. In: FUZZ-IEEE, pp. 2519–2526 (2014)
33. Adeyeba, T.A., Anderson, D.T., Havens, T.C.: Insights and characterization of l1-norm based sparsity learning of a lexicographically encoded capacity vector for the Choquet integral. In: FUZZ-IEEE, pp. 1–7 (2015)
34. de Oliveira, H.E., Duarte, L.T., Romano, J.M.T.: Identification of the Choquet integral parameters in the interaction index domain by means of sparse modeling. Expert Syst. Appl. **187** (2022)
35. Hurwicz, L.: The generalized Bayes minimax principle: a criterion for decision making under uncertainty. Cowles Comm. Discuss. Paper Stat. **335**, 1950 (1951)
36. Wang, H., Li, G., Jiang, G.: Robust regression shrinkage and consistent variable selection through the LAD-lasso. J. Bus. Econ. Stat. **25**(3), 347–355 (2007)
37. Gao, X., Huang, J.: Asymptotic analysis of high-dimensional LAD regression with lasso. Stat. Sin., 1485–1506 (2010)
38. Zhao, P., Yu, B.: On model selection consistency of lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)
39. Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**(476), 1418–1429 (2006)
40. van de Geer, S.: $\ell 1$-regularization in high-dimensional statistical models. In: Proceedings of the International Congress of Mathematicians 2010 (ICM 2010), pp. 2351–2369 (2010)
41. Wu, X., Liang, R., Yang, H.: Penalized and constrained LAD estimation in fixed and high dimension. Stat. Papers, 1–43 (2022)
42. Zheng, Q., Gallagher, C., Kulasekera, K.: Robust adaptive lasso for variable selection. Commun. Stat. Theory Methods **46**(9), 4642–4659 (2017)
43. Xu, J., Ying, Z.: Simultaneous estimation and variable selection in median regression using lasso-type penalty. Ann. Inst. Stat. Math. **62**, 487–514 (2010)
44. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc., B: Stat. (methodological) **67**(2), 301–320 (2005)
45. Bresson, R.: Neural learning and validation of hierarchical multi-criteria decision aiding models with interacting criteria. PhD thesis, Université Paris-Saclay (2022)
46. Herin, M., Perny, P., Sokolovska, N.: Learning preference models with sparse interactions of criteria. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23, pp. 3786–3794 (2023)
47. Keeney, R.L., Raiffa, H., Meyer, R.F.: Decisions with Multiple Objectives: Preferences and Value Trade-offs, Cambridge University Press (1993)
48. Fishburn, P.C.: Interdependence and additivity in multivariate, unidimensional expected utility theory. Int. Econ. Rev. **8**(3), 335–342 (1967)
49. Gonzales, C., Perny, P.: GAI networks for utility elicitation. KR **4**, 224–234 (2004)
50. Braziunas, D., Boutilier, C.: Local utility elicitation in GAI models. In: Proceedings of UAI, pp. 42–49 (2005)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.