Explaining Robust Additive Decision Models: Generation of Mixed Preference-Swaps by Using MILP

Manuel Amoussou, Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane

Explaining Robust Additive Decision Models: Generation of Mixed Preference-Swaps by Using MILP

Manuel Amoussou[1], Khaled Belahcene[2], Christophe Labreuche[3], Nicolas Maudet[4], Vincent Mousseau[1], Wassila Ouerdane[1]

Abstract. In this paper, we are interested in the question of explanation in Multicriteria Decision Aiding (MCDA) in general, and the explanation of the robust additive model in particular. To this end, a previous work has laid the foundations for explaining the necessary preference relation through a sequence of preference swaps. We propose to extend this work by introducing the concept of "mixed explanation" where the computation of its components is done through the resolution of a Mixed-Integer Linear Program. With the help of several examples, we motivate the interest of such an extension and open a discussion toward several promising further questions.

Keywords : MCDA, Additive utility explanation, necessary preference relation, mixed sequence of preference swaps, MILP.

	1	2	3	4	5	6
c	+	+	−	−	+	+
o	+	−	+	+	+	−
n	−	+	+	+	+	−
a	+	+	+	−	−	+
v	+	+	+	+	−	−
i	+	−	+	−	+	+
r	−	−	+	+	+	+
u	+	+	−	+	+	−
s	−	+	+	+	−	+

Table 1. Motivating example: Performance table

1 Introduction

This work is concerned by generation of explanation patterns of the outcomes of Multi-Criteria Decision Aiding (MCDA) models. Only a few works deal with this question [9, 10, 11, 1, 3, 4]. The generation of explanations in this context is not a straightforward task, because different criteria are at stake, the user is not necessarily able to fully assess their importance or to understand how they interact. Moreover, once the user is confronted to the result and the explanation, she may realize that it is not exactly what she expected. Thus, beyond acceptance facility, presenting an explanation may have an impact on the representation of the user's mode of reasoning that is the basis of building the recommendation.

To illustrate a MCDA situation let consider the following example. In the context of the Covid pandemia, a corporation is willing to secure the supply of masks for the protection of its employees. The board of directors received 9 responses to the call for tender, and is willing to select the 4 best mask suppliers, each of which will obtain 25% of the market. Each supplier is evaluated on the characteristics of its product and also on his reputation. The analyst in collaboration with the logistics manager (decision-maker) has defined the following characteristics/criteria: **(1)** customizable: "yes" $(+)$ or "no" $(-)$, **(2)** washable: "yes" $(+)$ or "no" $(-)$, **(3)** delivery time: "1 - 14 days" $(+)$ or "15 - 30 days" $(-)$, **(4)** quality: "high" $(+)$ or "good" $(-)$, **(5)** affordability: "acceptable" $(+)$ or "expensive" $(-)$, and **(6)** provider reputation: "good" $(+)$ or "fair" $(-)$

The performance table (see Table 1) describing the evaluations of the 9 suppliers on the 6 criteria is provided in Table 1. Each of the 6 criteria is described on bi-levels scales, which facilitate the symbolic representation of the 9 alternatives . Moreover, for each of the 6 criteria, the value symbolized by $+$ is more desirable than the value symbolized by $-$.

After a thorough discussion with the logistics manager, the analyst felt that her preferences are representable by an additive value model. In addition, these preferences will be expressed through holistic subjective pairwise judgments on the set of alternatives, through a preference elicitation process. As previously mentioned, the aim is to select 4 best suppliers. The analyst will therefore potentially have to justify, as he collects preferential information, why a specific supplier (assume supplier z) should be chosen. To do so, he will have to justify or explain why z is preferred over at least 5 other suppliers. It is obvious that an explanation is not required if the preference of z over any other suppliers has been explicitly expressed by the logistics manager or deduced by transitivity from his previous statements. Indeed, the deductions (binary preference comparisons between suppliers) which back up the recommendation and that are subject to explanation are the ones derived from the $\mathcal{DM}$ statements but not easy to grasp. So, what kind of explanation the analyst should give to the manager, under the assumption that a robust additive value model is used and that the manager has provided some preference information on the decision situation?

A first work, by [4], has proposed explanation schemes for the robust additive model under the form of a sequence of "preference-swaps" (inspired by the *even-swaps* concept [8]). More precisely, the robust additive utility model is a necessary preference relation [7], [12] which makes minimal assumptions, while handling a collection of compatible utility functions, which are impossible to exhibit to the user. The proposed explanation engine presents an explanation for a necessary preference as a sequence of pairwise comparisons such that the compared alternatives may only differ at most two criteria. However, such an explanation is not always easy to construct

[1] MICS, CentraleSupélec, Université Paris-Saclay, Gif-Sur-Yvette, France
[2] Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, Compiègne, France
[3] Thales TRT, Palaiseau, France
[4] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, France

and even in some situation does not exist. Therefore, in this paper we propose to alleviate some of the *preference-swaps* explanation engine constraints to arrive at what we will call a *mixed explanation* composed of "elementary" elements belonging to both necessary and possible preference relations.

The paper is organised as follows. Section 2 is dedicated to present the additive value model, the necessary relation and the computation engine of preference-swap sequences. Section 3 is devoted to highlight the limitations of the preference swaps engine supporting the extension proposed in this paper and to present the Mixed-Integer Linear Programming (MILP) as a tool for computing the mixed preference swaps. We end the paper by a conclusion and some perspectives in Section 4.

2 Inferring and Explaining Necessary Preference Relations

We are considering a decision situation in which a Decision Maker ($\mathcal{DM}$) and a Decision Analyst ($\mathcal{DA}$) are engaged in an interaction from which a recommendation should arise. This recommendation whatever its kind (choosing, ranking or sorting) will be built upon a finite and non-empty set of alternatives (actions) $\mathbb{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \dots\}$ evaluated on a family of n discrete criteria. We denote by $\mathbb{X}_i$ the finite set of possible levels for criterion $i \in \{1, \dots, n\}$. We assume that the set $\mathbb{X}_i = \{x_i^1, x_i^2, \dots, x_i^{r_i-1}, x_i^{r_i}\}$ is ordered and without loss of generality, we consider that : $x_i^{r_i} \succsim_i x_i^{r_i-1} \succsim_i \cdots \succsim_i x_i^2 \succsim_i x_i^1$ for all $i \in \{1, \dots, n\}$, with $\succsim_i$ denoting the marginal preference order on criteria i. Hence, each alternative in $\mathbb{A}$ is described by a tuple $x \in \mathbb{X} = \prod_{i=1}^n \mathbb{X}_i$, and in general $\mathbb{A} \subset \mathbb{X}$.

Moreover, the $\mathcal{DM}$'s preferences (expressed through holistic pairwise statements) denoted by $\succsim$ is assumed to be representable by an additive multi-attribute value function. Under this assumption, there exists a function $\mathcal{U}$ defined on $\mathbb{X}$ such that, for all $a,\ b \in \mathbb{X}$:

$$a \succsim b \Leftrightarrow \mathcal{U}(a) \geq \mathcal{U}(b) \tag{1}$$

such that, $\mathcal{U}(x) = \sum_{i=1}^n u_i(x_i)$ and $\mathcal{U}(y) = \sum_{i=1}^n u_i(y_i)$ and u_i is a function mapping $\mathbb{X}_i$ into $\mathbb{R}$ for all i and x_i (resp. y_i) is the $i-$th component of x (resp. y).

In the rest of this document, we will focus on the case where for all $i \in \{1, \dots, n\}$, $r_i = 2$ (as in our example, see Section1), and without loss of generality, we will designate by $+$ (resp. $-$) any $i-$th component of any alternative x such that $x_i = x_i^1$ (resp. $x_i = x_i^0$). The case of $r_i > 2$ is left for further investigation in next work.

2.1 Inference of Necessary Preference Relations

The concept of *necessary preference relation* in the context of the additive value model, given a set of pairwise holistic preference statements (denoted by $\mathbb{PI}$), refers to an idea of robustness according to which during the preference learning process the derivation of recommendation should take all $\mathbb{PI}-$compatible value functions into account (see [6]). The *necessary preference relation* has been addressed in [7], where its fundamentals have been formalized and its characterization with a linear program has been proposed. In what follows the notations and the main results of [7] are adapted to our case where n components are representable on discrete scales.

Definition 2.1 (Necessary preference relation [7])**.** *Let $x,\ y \in \mathbb{X}$, $\mathbb{PI} \subset \mathbb{X} \times \mathbb{X}$. x is necessarily preferred to y (noted $(x, y) \in \mathcal{N}_{\mathbb{PI}}$) if $\mathcal{U}(x) = \sum_{i=1}^n u_i(x_i) \geq \sum_{i=1}^n u_i(y_i) = \mathcal{U}(y)$ holds for every function $\mathcal{U} \in \mathbb{X} \to \mathbb{R}$ additively compatible with $\mathbb{PI}$ i.e for all $(a,\ b) \in \mathbb{PI}$, $\mathcal{U}(a) \geq \mathcal{U}(b)$.*

Proposition 2.1 ([7])**.** *Given $\mathbb{PI} \subset \mathbb{X} \times \mathbb{X}$ and $x, y \in \mathbb{X}$; $(x, y) \in \mathcal{N}_{\mathbb{PI}}$ if and only if, the following linear program has a non-negative solution:*

$$\text{Min} \sum_{i=1}^n u_i(x_i) - \sum_{i=1}^n u_i(y_i)$$

$$s.t. \begin{cases} \sum_{i=1}^n u_i(a_i) \geq \sum_{i=1}^n u_i(b_i) & \forall (a,b) \in \mathbb{PI} \\ u_i(z_i^{r_i}) \geq u_i(z_i^{r_i-1}) & \forall\, i \in \{1, \dots, n\}, \forall\, r_i \geq 2 \end{cases}$$

We note that in the case where for each $\mathbb{X}_i$, $r_i = 2$ ($i \in \{1 \dots n\}$), this model boils down defining weights $w_i = u_i(+) - u_i(-)$, $i \in \{1, \dots, n\}$.

Under the assumption that the $\mathcal{DM}$ preferences are representable by an additive model, the inference of a necessary preference relation between two alternatives $x, y \in \mathbb{A}$ ($(x, y) \in \mathcal{N}_{\mathbb{PI}} \setminus \mathbb{PI}$) can be seen as a consequence of the $\mathcal{DM}$ holistic statements. This element can be used by the $\mathcal{DA}$ as a feedback to confront the $\mathcal{DM}$ with his or her subjective judgements.

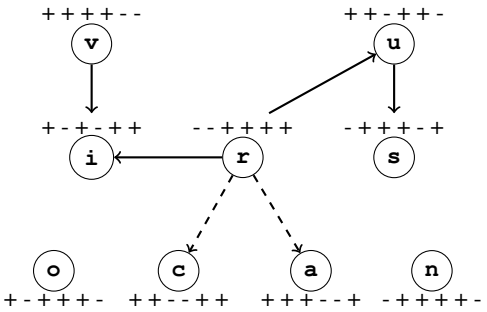

Figure 1. Representation of $\mathbb{PI}$ and some elements of $\mathcal{N}_{\mathbb{PI}}$

Example 2.1. *The decision aiding situation of our running example implies an interaction between the analyst and the manager to try to understand the problem and to collect preference information from the latter. At the beginning, the manager ($\mathcal{DM}$) provides three elements $\mathbb{PI} = \{\mathtt{r} \succsim \mathtt{i}, \mathtt{v} \succsim \mathtt{i}, \mathtt{u} \succsim \mathtt{s}\}$, representing his subjective judgement on a subset of the different alternatives.*

Based on this $\mathbb{PI}$, the Table 1, and the assumption that the manager follows an additive model in his reasoning, the 4-necessary best alternatives set (i.e. the recommendation) is empty. Indeed, the two following additive models compatible with $\mathbb{PI}$ imply two different rankings on the nine alternatives in which each of them does not belong to the set of the 4-necessary best alternatives

Models (w_i)	Implied ranking
$(0.20, 0.08, 0.25, 0.30, 0.12, 0.05)$	o, v, n, r, u, s, i, a, c
$(0.13, 0.30, 0.02, 0.15, 0.15, 0.25)$	c, u, s, a, n, v, r, i, o

After a while during the interaction, the $\mathcal{DM}$ states that $\mathtt{r}$ is preferred over $\mathtt{u}$, updating $\mathbb{PI} = \{\mathtt{r} \succsim \mathtt{i}, \mathtt{v} \succsim \mathtt{i}, \mathtt{u} \succsim \mathtt{s}, \mathtt{r} \succsim \mathtt{u}\}$. As a consequence, the 4- necessary best alternatives set is no longer empty. In fact, in addition to $(\mathtt{r}, \mathtt{i})$, $(\mathtt{r}, \mathtt{u})$ and $(\mathtt{r}, \mathtt{s})$ collected directly from the $\mathcal{DM}$, it is deduced the following : $(\mathtt{r}, \mathtt{a}) \in \mathcal{N}_{\mathbb{PI}}$ and $(\mathtt{r}, \mathtt{c}) \in \mathcal{N}_{\mathbb{PI}}$: Supplier $\mathtt{r}$ is then necessarily preferred over 5 other suppliers ($\mathtt{i}$, $\mathtt{u}$, $\mathtt{s}$, $\mathtt{a}$ and $\mathtt{c}$) as shown in Figure 1 where the preference information ($\mathbb{PI}$) provided by the $\mathcal{DM}$ are depicted

with solid arrows and the deductions made with dashes ones. Each alternative is represented by its tuple of evaluation on each criterion.

Although a linear program formulation is well suited to the definition of the *necessary preference relation*, we believe that in a decision aiding process and in situations where a proof (at least algebraic) is required for each of the inferences made, this type of formulation is not satisfying because it is mainly subordinated to the might of the device equipped with an optimization engine. In order to circumvent this impediment, [4] has proposed a toolbox for the inference of necessary preferences where the concept of *covector associated to a pair of alternatives* play a central role.

One of the main results indicates, among others, that the covector representing each $\mathcal{N}_{\mathbb{PI}}$ element can be written as a conical[5] combination of the covectors representing the preference information and the covectors of the dual base that represent Pareto dominance [4]. In our context we will rely only on convectors in the case of binary core (i.e. the preference information only reference two levels according to each point of view.) (see [2, 4]).

Definition 2.2 (Covector in the case of a binary core)**.** *Given a pair of alternatives $(x,\, y) \in \mathbb{X} \times \mathbb{X}$, the covector of $(x,\, y)$ denoted by $(x,\, y)^\star$ is the $n-$ component vector defined as follows :*

$$(x,\, y)^\star = (\lambda_i)_{1 \leq i \leq n}$$

where :

$$\lambda_i = \begin{cases} -1 & if\ y_i \succ x_i \\ 0 & if\ x_i \sim y_i \\ 1 & if\ x_i \succ y_i \end{cases}$$

Example 2.2 (Ex.2.1 Cont.)**.** *In the following, by relaying on the notion of covectors (see Def. 2.2) we show why the pairs (r,a) and (r,c) are included in $\mathcal{N}_{\mathbb{PI}}$. We already know that the following pairs are part of the necessary preference relation.*

$$\begin{array}{l} (r,i)^\star = (-1\ \ 0\ \ 0\ 1\ \ 0\ \ 0) \\ (v,i)^\star = (\ 0\ \ 1\ \ 0\ 1 -1 -1) \\ (u,s)^\star = (\ 1\ \ 0 -1\ 0\ \ 1 -1) \\ (r,u)^\star = (-1 -1\ \ 1\ 0\ \ 0\ \ 1) \end{array}$$

Therefore, we can write: $(r,a)^\star = (r,i)^\star + (r,u)^\star + (u,s)^\star$, as we can see in the following:

$$\begin{array}{rl} (-1\ \ 0\ \ 0\ 1\ 0\ \ 0) & = (r,i)^\star \\ +\,(-1 -1\ \ 1\ 0\ 0\ \ 1) & = (r,u)^\star \\ +\,(\ 1\ \ 0 -1\ 0\ 1 -1) & = (u,s)^\star \\ \hline (-1 -1\ \ 0\ 1\ 1\ \ 0) & = (r,a)^\star \end{array}$$

The same reasoning can be applied to $(r,c)^\star = 2 \times (r,u)^\star + (u,s)^\star + (v,i)^\star$.

$$\begin{array}{rl} (-2 -2\ \ 2\ 0\ \ 0\ \ 2) & = 2 \times (r,u)^\star \\ +\,(\ 1\ \ 0 -1\ 0\ \ 1 -1) & = (u,s)^\star \\ +\,(\ 0\ \ 1\ \ 0\ 1 -1 -1) & = (v,i)^\star \\ \hline (-1 -1\ \ 1\ 1\ \ 0\ \ 0) & = (r,c)^\star \end{array}$$

In the previous example, we exploited a way of reasoning about preferences based on cancelling out common values across statements. As a result, any conclusion drawn by means of this approach comes along with a justification. We refer the reader interested on how to exploit this mechanism to [5]

[5] A conical combination is a linear combination with non-negative coefficients.

2.2 Explanation with a Sequence of Preference Swaps

Preference-swaps have been introduced as a tool for explaining binary relations of weak preference $\mathcal{R} \subseteq \mathbb{X} \times \mathbb{X}$ satisfying the Pareto dominance, transitivity and optionally, first-order cancellation [4].

Given $(x,y) \in \mathcal{R}$, a *preference-swap* represents an elementary component $(e^{(j-1)}, e^{(j)}) \in \mathcal{R}$ (with $j \geq 1$) of the sequence (of length l) $x := e^{(0)} \succsim e^{(1)} \succsim \cdots \succsim e^{(l)} := y$. One of its characteristics is its *order* which can be briefly defined as the number of differing criteria between the alternatives $e^{(j-1)}$ and $e^{(j)}$. Since each of them is described on a set of n points of view, this number is noted k and we have $1 \leq k \leq n$. When k is equal to 1, it is a Pareto dominance.

In this paper, we will limit to values of k of the swap at most equal to 2 as done in the even-swaps method [8][6]. This restriction can be justified by the fact that on the one hand it is easily scriptable i.e. expressible in a natural language, and on the other hand it reduces to a strict minimum the number of points of view to be confronted and for which an explicit position of the $\mathcal{DM}$ is required. For these reasons we think that a preference-swap of order at most 2 can be used as a basic argument in the explanation of an element (x,y) of the relation $\mathcal{R}$. As a consequence, we consider that all 2-order elements of $\mathcal{R}$ are exempt from explanation. Moreover, as it was mentioned in Section 1, the relation $\mathcal{R}$ that we consider in this paper is the necessary preference relation (see Def. 2.1) under the assumption of additivity.

Definition 2.3 (Preference-swaps explanation [4])**.** *Given $\mathbb{PI}$ on $\mathbb{A}$, let $(x,\, y) \in \mathbb{A} \times \mathbb{A}$ and $(x,\, y) \in \mathcal{N}_{\mathbb{PI}}$, an explanation of $(x,\, y)$ is a sequence of swaps of length l ($l > 1$):*

$$x := e^{(0)} \succsim e^{(1)} \succsim \cdots \succsim e^{(l)} := y$$

where for all $j \in [\![1; l]\!]$, $(e^{(j-1)}, e^{(j)}) \in \mathcal{N}_{\mathbb{PI}}$ and is of order at most 2.

In the remainder, we will designate such swap by *necessary swaps* since they are in $\mathcal{N}_{\mathbb{PI}}$. Therefore, a *necessary explanation* or simply an *explanation* of $(x,\, y) \in \mathcal{N}_{\mathbb{PI}}$ will refer to a sequence of necessary swaps linking x to y.

Example 2.3. *We proved in Ex. 2.2 that: $(r,a) \in \mathcal{N}_{\mathbb{PI}}$. Its order is equal to 4 and it is one of the results that legitimize the selection of r among the 4-necessary best suppliers.*

Thus, in what follows we provide a necessary explanation to $(r,a) \in \mathcal{N}_{\mathbb{PI}}$, by using the two pairs of alternatives (r,s) and (s,a), both belonging to $\mathcal{N}_{\mathbb{PI}}$ and of order 2.

Indeed, we recall that: $\mathbb{PI} = \{r \succsim i, v \succsim i, u \succsim s, r \succsim u\}$, and , we have:

$$\begin{array}{l} (r,i)^\star = (-1\ \ 0\ \ 0\ 1\ \ 0\ \ 0) \\ (v,i)^\star = (\ 0\ \ 1\ \ 0\ 1 -1 -1) \\ (u,s)^\star = (\ 1\ \ 0 -1\ 0\ \ 1 -1) \\ (r,u)^\star = (-1 -1\ \ 1\ 0\ \ 0\ \ 1) \end{array}$$

On one hand, $(r,s)^\star = (r,u)^\star + (u,s)^\star$, as shown below:

$$\begin{array}{rl} (-1 -1\ \ 1\ 0\ 0\ \ 1) & = (r,u)^\star \\ +\,(\ 1\ \ 0 -1\ 0\ 1 -1) & = (u,s)^\star \\ \hline (\ 0 -1\ \ 0\ 0\ 1\ \ 0) & = (r,s)^\star \end{array}$$

[6] It is important to recall that one of its limitations is that it can be difficult to apply the even-swaps method in the absence of criteria defined on continuous scales on which the property of solvability (essential to establish indifference) naturally occurs. In this paper, each alternative is defined on a set of discrete criteria.

On the other hand, we have :

$$(s, a)^\star = (-1\ 0\ 0\ 1\ 0\ 0) = (r, i)^\star$$

Therefore, as $(r, s) \in \mathcal{N}_{\mathbb{PI}}$ *and* $(s, a) \in \mathcal{N}_{\mathbb{PI}}$*, thanks to transitivity, we have* $(r, a) \in \mathcal{N}_{\mathbb{PI}}$*. Consequently, the explanation could be scripted as follows:*

You prefer the supplier r *over the supplier* a *because every thing else being equal:*

- *you prefer an affordable mask to a washable one and*
- *you prefer a high quality mask to a customizable one.*

As it was shown in the Example 2.3, the explanation involves different pieces of information that belong to $\mathcal{N}_{\mathbb{PI}}$. However, as it will be illustrated in the following Example2.4, it exists situations where it is not possible to build a necessary explanation to justify necessary preference information. In other terms, there not always exist a sequence involving *only* pairs of the necessary preference relation $\mathcal{N}_{\mathbb{PI}}$ to build the explanation.

Example 2.4. *We have seen that the pair* (r, c) *is in the necessary preference relation. However, this deduction can not be explained through a necessary explanation given* $\mathbb{PI}$*. If it were the case, according to [4, Theorem 6], at least we would have either* $(n, u) \in \mathcal{N}_{\mathbb{PI}}$ *or* $(i, c) \in \mathcal{N}_{\mathbb{PI}}$*. But this is not the case as it can be noticed through the two following additive models compatible with* $\mathbb{PI}$ *and which imply 2 rankings of the 9 alternatives where the supplier* u *is ranked before* n *and* c *is before* i *in the decreasing order of preference (see the following Table).*

Models (w_i)	Ranking implied
$(0.23, 0.03, 0.22, 0.30, 0.12, 0.10)$	o, v, r, u, i, n, s, a, c
$(0.04, 0.25, 0.24, 0.12, 0.29, 0.06)$	n, r, u, o, s, v, c, i, a

To overcome this difficulty to build a necessary explanation given some preference information ($\mathbb{PI}$) provided by the decision-maker, we introduce the notion of *possible swaps* which are composed of two alternatives z and z′ differing on exactly two criteria and such that $(\mathtt{z}, \mathtt{z}') \notin \mathcal{N}_{\mathbb{PI}}$ and $\mathbb{PI} \cup \{(\mathtt{z}, \mathtt{z}')\}$ is representable by an additive value function. We then propose the notion of *mixed explanation* (see Def. 3.3). This latter is composed of a mix of necessary swap(s) and possible swap(s). The next section is devoted to discuss the interests and the challenges behind a mixed explanation. Moreover, we propose a first tool to generate the components of a mixed explanation by using a MILP.

3 Explanation with Sequence of Mixed Preference Swaps

In [4], Belahcene et al. investigate the opportunity of providing *transitive explanations*. They highlight several challenges arising when trying to implement this type of explanations:

- *feasibility*– is it possible to find a transitive explanation for a given statement?
- *intelligibility*– what additional constraints should be put on the explaining sequences in order to be actually accepted as explanations? In particular, what order (i.e. upper bound on the Hamming distance (see Def. 3.2)) and length are acceptable, knowing that there are trade-offs between those parameters and the question of feasibility;
- *computation*– how to efficiently build those sequences?

The question of feasibility sometimes admit a negative answer, as illustrated in Ex. 2.4. We recall that a necessary preference relation makes minimal assumptions, while handling a collection of compatible utility functions. Consequently, the number of available *arguments* –here, necessary preference statements of order 2–is small, which greatly limits the feasibility of finding explanations. Thus, we propose to relax the constraint of using *only* necessary swaps and to support a statement by introducing in the reasoning *possible swaps* (*a subset* of compatible additive utility functions compatible with $\mathbb{PI}$). By doing so, we expect to be able to explain more pairs of the necessary relation. It is clear that providing a sequence composed of solely necessary swaps guarantees that the recipient of the explanation will accept and validate each swap without any doubt. However, using possible swaps offers a way to collect more additional preference information (valuable in a preference elicitation process) and thus enrich both $\mathbb{PI}$ and $\mathcal{N}_{\mathbb{PI}}$. Indeed, the decision-maker may accept or refute the possible swaps engaging him in an interaction towards the construction of a representation of its decision model (and thus the recommendation) [6]. Finally, our explanations offer a way of reasoning about preferences based on a chain of "elementary" elements (swaps) allowing the decision-maker to understand why an alternative is preferred to another one. We believe that confronted with this reasoning during an interaction; the $\mathcal{DM}$ could appropriate it and apply it by himself to a pair of alternatives, contributing to the preference elicitation process. By using possible swaps, we augment the chance to find an explanation and thus to enhance the contribution of the decision-maker.

Before formally introducing the notion of mixed explanation, let us consider the following example.

Example 3.1 (Ex.2.3 Cont.)**.** *In Ex. 2.4, we concluded that it was not possible to explain* $(r, c) \in \mathcal{N}_{\mathbb{PI}}$ *via a necessary explanation given* $\mathbb{PI}$*. Thus, we propose to illustrate here two variants of explanation allowing to justify this pair, by introducing elements that are not belonging to the necessary preference relation, namely possible swaps.*

1. *Explanation variant #1 for* (r, c)*:*

   We have :
   $(r, c)^\star = (-1\ -1\ 1\ 1\ 0\ 0) = (0\ -1\ 1\ 0\ 0\ 0) + (-1\ 0\ 0\ 1\ 0\ 0)$.
   The first term of the decomposition can be interpreted as that everything else being equal the difference of utility between the top level (+) and the lowest level (−) on the third component is greater than the second one. The member at the left of the operator + *suggests that : everything else being equal the difference of utility between the top level (+) and the lowest level (−) on the third dimension is greater than the second one. The corresponding explanation sequence could then be :*

   $$r \succsim z \succsim c$$

   where : $r := (--++++)$*;* $c := (++--++)$ *and* z *is a fictitious alternative (*$\notin \mathbb{A}$*), with* $z := (-+-+++)$.
   We note that in this sequence the swap $z \succsim c$ *is such that* $(z,\ c) \in \mathcal{N}_{\mathbb{PI}}$*, since the right term the of the decomposition of* $(r,\ c)^\star$ *is equal to* $(r,\ i)^\star$ *and* $(r,\ i) \in \mathbb{PI}$*. However, swap* $r \succsim z$ *is such that* $(r,\ z) \notin \mathcal{N}_{\mathbb{PI}}$ *(see Ex.2.4). Thus, the mixed explanation can be scripted as follows : Every thing else being equal,*

   - *you* ***might*** *prefer a quick delivery of non-washable masks to a late delivery of washable masks and*
   - *you prefer non-customizable masks of high quality to customizable ones and of good quality.*

2. *Explanation variant #2 of* (r, c)*:*
By using the same reasoning based on the fact that : $(r, c)^\star$ = *(*$-1\ -1\ 1\ 1\ 0\ 0$*) = (*$-1\ 0\ 1\ 0\ 0\ 0$*) + (*$0\ -1\ 0\ 1\ 0\ 0$*), we derive the following sequence :*

$$r \succsim z' \succsim c$$

with : $r := (- - + + ++)$, $c := (+ + - - ++)$ *and* $z' := (+ - - + ++)$ *a fictitious alternative. Both swaps used in this explanation are not necessary and the corresponding mixed explanation could be scripted as follows: Every thing else being equal,*

- *you* ***might*** *prefer a quick delivery of non-customized masks to a late delivery of customized masks and*
- *you* ***might*** *prefer a non-washable mask of high quality to a washable mask of good quality.*

3.1 Towards a formal definition of Mixed Explanations

We would like to define a mixed explanation in the same manner as a fully necessary one, but instead of restricting ourselves to using transitive links that are pairwise preference statements that hold in every world compatible to the preference information, we simply specify that all the links must hold in at least one of these worlds. Nevertheless, this formal knowledge representation approach does not lead to a compelling notion of explanation.

In order to circumvent this obstacle, we begin by recalling formal definitions of concepts permitting to describe explanations.

Definition 3.1 (Pros and Cons of a Necessary Preference Statement [4])**.** *Given* $\mathbb{PI}$ *and* $(x,\ y) \in \mathcal{N}_{\mathbb{PI}}$*, we define :*

$$(x,\ y)^+ := \{i \in \{1, \ldots, n\} : (x,\ y)_i^\star = +1\}$$

$$(x,\ y)^- := \{i \in \{1, \ldots, n\} : (x,\ y)_i^\star = -1\}$$

In other words, $(x,\ y)^+$ *is the subset of criteria* i *on which* $x_i = +$ *and* $y_i = -$ *and* $(x,\ y)^-$*, the subset of criteria* i *on which* $x_i = -$ *and* $y_i = +$*.*

Transitive explanations implement the *divide and conquer* paradigm in order to break down the complexity of the explanadum (what needs to be explained) into smaller chunks deemed more palatable for the explainee (the recipient of the explanation). In the case of preference swaps, we assume that the Hamming distance between alternatives somehow reflects the cognitive difficulty to assess the trade-off of exchanging one against the other.

Definition 3.2 (Hamming distance between a pair of alternatives)**.** *Given* $(x,\ y) \in \mathbb{X} \times \mathbb{X}$*, the Hamming distance between* x *and* y *is the function* Φ *defined as follows:*

$$\begin{array}{ccc} \mathbb{X} \times \mathbb{X} & \longrightarrow & [\![0; n]\!] \\ x,\ y & \longmapsto & \Phi(x, y) = |\{i \in \{1, \ldots, n\} : (x,\ y)_i^\star \neq 0\}| \end{array}$$

where $|E|$ *designates the cardinality of the set* E

The challenges concerning transitive explanations detailed in [4] receive detailed answers in the specific case of the necessary relation under the assumption of additive preferences, and when the preference information is expressed using solely two values on each criterion. In this particular case, the problem of finding a transitive explanation of a necessary pairwise statement $(x, y) \in \mathcal{N}_{\mathbb{PI}}$ is shown to reduce, without loss of generality, to the problem of finding a matching in the graph induced by the preference relation $\mathcal{N}_{\mathbb{PI}}$ restricted to swaps of order 2 on the cartesian product $\mathcal{S} := (x,\ y)^+ \times (x,\ y)^-$ – effectively matching each *con* argument with a stronger *pro* argument.

When considering transitive chains consisting not only of links that are necessary preference statements, but also possible ones, this powerful result no longer applies directly. However, the facts that explanations do not need to refer to neutral arguments, and that each pro and con appear at most once have a normative appeal in terms of what constitutes a good explanation. It remains to be proven that this restriction can be made without loss of generality, i.e. that no necessary preference statement can be explained by means of a transitive chain, eventually with possible links, but not by one restricted to arguments matching a con with a possibly stronger pro. Until then, we propose a formal definition of mixed explanations based on these characteristics.

Definition 3.3 (Mixed explanation)**.** *Given* $\mathbb{PI}$*, let* x *and* y *such that* $(x,\ y) \in \mathbb{A} \times \mathbb{A}$*,* $\Phi(x, y) > 2$ *and* $(x,\ y) \in \mathcal{N}_{\mathbb{PI}}$*. A mixed explanation corresponds to a sequence:*

$$x := e^{(0)} \succsim_{\mathcal{U}} e^{(1)} \succsim_{\mathcal{U}} \cdots \succsim_{\mathcal{U}} e^{(l)} := y \tag{2}$$

where :

- $\succsim_{\mathcal{U}}$ *is the binary relation induced by* $\mathcal{U}$ *on* $\mathbb{X}$*, where* $\mathcal{U}$ *is an additive multi-attribute value function* $\mathcal{U}$ *compatible with* $\mathbb{PI}$*.*
- *all the alternatives of the set* $\mathcal{E} = \{e^{(0)}, \ldots, e^{(l)}\}$ *are identical on the criteria of the set* $\mathcal{I} = \{i \in \{1, \ldots, n\} : (x,\ y)_i^\star = 0\}$*,*
- *for all* $m \in [\![1, l]\!]$*,* $\Phi(e^{(m-1)}, e^{(m)}) \leq 2$*,*
- *each advantage of alternative* y *over alternative* x *should be compensated exactly once, i.e.*

$$\forall j \in (x,y)^-, \left|\left\{k \in [\![1, l]\!] : (e^{(k-1)}, e^{(k)})^- = \{j\}\right\}\right| = 1; \tag{3}$$

- *any advantage of alternative* x *over alternative* y *with respect to criterion* $i \in (x,\ y)^+$ *can be used at most once in the compensation of an advantage of alternative* y *over* x *with respect to another criterion in* $(x,\ y)^-$*, i.e.*

$$\forall i \in (x,y)^+, \left|\left\{k \in [\![1, l]\!] : (e^{(k-1)}, e^{(k)})^+ = \{i\}\right\}\right| \leq 1; \tag{4}$$

and

- *there is a (potentially empty) set* $\mathcal{M} \subset \{1, ..., l\}$ *such that for all* $m \in \mathcal{M}$*, we have* $(e^{(m-1)}, e^{(m)}) \in \mathcal{N}_{\mathbb{PI}}$

3.2 Computing Mixed Explanations

Feasibility In this section, we will expose how a *mixed explanation* of $(x,\ y) \in \mathcal{N}_{\mathbb{PI}}$ (with $\Phi(x, y) > 2$) could be deduced from the resolution of a Mixed-Integer Linear Program. As previously stated, this program has to produce the *swaps* of *order* 2 which compose the explanation sequence linking x to y (see Def.3.3). We recall that such swaps can be represented as elements $(i,\ j)$ of the set $(x,\ y)^+ \times (x,\ y)^-$ symbolizing a double switch of criteria i and j (from $+$ to $-$ on criterion i and from $-$ to $+$ on criterion j) and ensuring that $[u_i(+) - u_i(-)] + [u_j(+) - u_j(-)] \geq 0$.

In the remainder, we will denote the *swaps space* $(x,\ y)^+ \times (x,\ y)^-$ by

$$\mathbb{S} := (x,\ y)^+ \times (x,\ y)^- \tag{5}$$

3.2.1 Variables

The variables of this MILP are of two kinds :

- Two positive real variables $u_i(+)$ and $u_i(-)$, for each criterion $i \in \{1 \dots n\}$,
- A binary variable b_s such that

$$b_s = \begin{cases} 1 & \text{iff } s \in \mathcal{S} \\ 0 & \text{Otherwise.} \end{cases}$$

3.2.2 Constraints

We will distinguish six kinds of constraints:

- In order to take into account the preference information provided, we have :

$$\sum_{i=1}^{n} u_i(a_i) \geq \sum_{i=1}^{n} u_i(b_i) \text{ for all } (a,\, b) \in \mathbb{PI} \tag{6}$$

  where a_i is the $i-$th component of alternative a.
- Normalization of the marginal value functions:

$$\begin{cases} u_i(-) = 0 \text{ for all } i \in \{1, \dots, n\} \\ \sum_{i=1}^{n} u_i(+) = 1 \end{cases} \tag{7}$$

- Having assumed that each set $\mathbb{X}_i$ is ordered and considering that the value $+$ on each criteria i is more desirable than the value $-$, we have :

$$u_i(+) - u_i(-) \geq 0 \;\text{ for all } i \in \{1, \dots, n\}. \tag{8}$$

- The next constraint expresses the condition (3).

$$\sum_{s=(i,\,j)\in\mathbb{S}} b_s = 1 \;\; \forall\, j \in (x,\, y)^- \tag{9}$$

- The following constraint expresses the condition (4).

$$\sum_{s=(i,\,j)\in\mathbb{S}} b_s \leq 1 \;\; \forall\, i \in (x,\, y)^+ \tag{10}$$

- With the last constraint we ensure the compatibility of the swaps used with the $\mathbb{PI}$ and the representativeness of the $\mathcal{DM}$ preferences by an additive model.

$$\text{For all } s = (i,\, j) \in \mathbb{S}, [u_i(+) - u_i(-)] - [u_j(+) - u_j(-)] \geq b_s - 1 \tag{11}$$

This MILP is feasible if, and only if, there exists a mixed explanation of $(x,\, y) \in \mathcal{N}_{\mathbb{PI}}$, and any solution pinpoints a "bag" of swaps that can be used as arguments in order to build a proper explanative sequence, when correctly assembled. Adding a relevant objective function can help to address the issues of intelligibility.

3.2.3 Objective functions

The normative conditions we chose to place on mixed explanations—taking the form of a matching of cons by pros—should already guarantee the optimality of the computed "bag of swaps" in terms of length. Also, the intelligibility of individual links is enforced through the definition of swaps as pairs of criteria. A powerful lever for ensuring that an explanation is accepted by the explainee is to try to maximize its plausibility. In a preliminary attempt, because we are (almost) totally confident in the veracity of the necessary swaps, but only partially convinced of the truthfulness of the others, we approximate this plausibility as a decreasing function of the number of possible-but-not-necessary swaps appearing in the explanation, or, equivalently, an increasing function of the number of necessary swaps. Thus, we consider using the following objective function:

$$\text{maximize} \sum_{s \in \mathbb{S}^{\mathcal{N}_{\mathbb{PI}}}} b_s \tag{12}$$

Here, we denote by $\mathbb{S}^{\mathcal{N}_{\mathbb{PI}}}$ the subset of *necessary swaps* in $\mathbb{S}$. In a more nuanced approach, we consider assessing the plausibility of a given possible swap through the ratio of the hypervolume of the set of worlds where the swap is true to the total hypervolume of the worlds compatible to the $\mathbb{PI}$, using the SMAA approach [12]. We would then aggregate the plausibilities of individual links into a compound plausibility of an explanation, by multiplying those ratio (with an implicit independence assumption). Therefore, we would pre-compute those ratios for every possible swap, and maximize the sum of their logarithms.

4 Concluding remarks

In this paper we investigated the question of providing explanations to justify the necessary preference relation under the assumption of an additive value model, by introducing the notion of mixed explanations. For this purpose, we proposed a first tool based on a MILP allowing to compute the components of this mixed explanation. However, as it was mentioned, different assumptions were taken to provide this first response. For instance, as it is illustrated through the following example, a first assumption was to always seek to keep the explanation as short as possible. Indeed, in Example 2.3, we provided an explanation for (`r`, `a`) based on two necessary swaps. Another way to explain the same pair is sketched in the following.

Example 4.1 (Ex. 2.3 Cont.)**.** *You prefer the supplier* `r` *over the supplier* `a` *because :*

- *you told me that you prefer* `r` *over* `u`*.*
- *you also told me that you prefer* `u` *over* `s` *and, every thing else being equal,*
- *you give more importance to a high quality but uncustomizable mask to a customizable but of just good quality one.*

At first sight, the main difference between the explanation of Ex.4.1 and the one in Ex.2.3 concerns the length (3 for the former and 2 for the later). One might think that a decision-maker would prefer the shorter one. However, if we take a look at the components of each explanation, we can see that the second one uses elements (preference information) provided by the decision-maker ($\mathcal{DM}$) himself. Consequently, we can assume that it would be difficult for the $\mathcal{DM}$ to reject his proper declarations unless he is contradicting the decision model under use, for instance. Thus, the compromise between the length of an explanation and its content deserves to be investigated. Another issue concerns the question of intelligibility of an explanation. In particular, what order (i.e. upper bound on the Hamming distance (see Def. 3.2)) and length are acceptable, knowing that there are trade-offs between those parameters and the question of feasibility.

References

[1] L. Amgoud and H. Prade, 'Using arguments for making and explaining decisions', *Artificial Intelligence*, **173**(3), 413–436, (2009).

[2] K. Belahcene, *Towards accountable decision aiding : explanations for the aggregation of preferences*, Ph.D. dissertation, 12 2018.
[3] K. Belahcene, Y. Chevaleyre, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane, 'Accountable approval sorting', in *Proceedings IJCAI'18*, pp. 70–76, (2018).
[4] K. Belahcene, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane, 'Explaining robust additive utility models by sequences of preference swaps', *Theory and Decision*, **82**(2), 151–183, (2017).
[5] K. Belahcene, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane, 'Comparing options with argument schemes powered by cancellation', in *Proc. of IJCAI'19*, pp. 1537–1543, (2019).
[6] D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukias, and P. Vincke, *Evaluation and Decision models with multiple criteria: stepping stones for the analyst*, volume 86, Springer, 2006.
[7] S. Greco, V. Mousseau, and R. Słowiński, 'Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions', *EJOR*, **191**(2), 416–436, (2008).
[8] J. Hammond, R. Keeney, and H. Raiffa, 'Even swaps: A rational method for making trade-offs', *Harvard business review*, **76**, 137–8, 143, (1998).
[9] C. Labreuche and S. Fossier, 'Explaining multi-criteria decision aiding models with an extended shapley value', in *Proc. of IJCAI'18*, pp. 331–339, (2018).
[10] C. Labreuche, N. Maudet, and W. Ouerdane, 'Minimal and complete explanations for critical multi-attribute decisions', in *Proc. of ADT*, pp. 121–134, Piscataway, NJ, USA, (2011).
[11] C. Labreuche, N. Maudet, and W. Ouerdane, 'Justifying dominating options when preferences are incomplete', in *Proceedings of ECAI*, pp. 486–491, Montpellier, France, (2012).
[12] R. Spliet and T. Tervonen, 'Preference inference with general additive value models and holistic pair-wise statements', *EJOR*, **232**, 607–612, (2014).