Explanation of the robust additive preference model by even swap sequences

Christophe Labreuche[1] and **Nicolas Maudet**[2] and **Vincent Mousseau**[3] and **Wassila Ouerdane**[4]

Abstract. The even swap method is an interesting approach for identifying the best alternative among several options [5]. This constructive method is intuitively attracting: only two attributes are involved in even swaps, and utilities are never explicitely mentioned to the Decision Maker (DM). The aim of this paper is to investigate whether this approach can be generalized to robust preference relations and used to generate convincing explanations.

1 Introduction

The problem of constructing or providing convincing explanations to a Decision Maker (DM) in order to justify recommended decisions is a central concern for decision-aiding tools (see for instance [1, 8, 11]). This issue raises many questions. What is an explanation? How to construct an explanation? What is the information, beyond the utility functions, that is useful or necessary to get to construct a "good" explanation? Roughly speaking, the aim is to increase the user's acceptance of the recommended choice, by providing supporting evidence that this choice is justified [6].

One of the difficulties of this question lies on the fact that the relevant concept of an explanation may be different, depending on the decision problem at hand and on the targeted audience. Depending on the situations, explanations may be required to be precise (like a proof), or instead to be only convincing arguments. Also, the information that may be put forward to generate an explanation may greatly vary: a convincing explanation for the decision analyst may be impossible to understand for the DM, simply because their level of understanding of the problem differ. This problem is especially difficult in the context of multi-attribute models [9, 10], where different criteria are at stake, where the DM is not necessarily able to fully assess how important are criteria or to understand the way criteria interact. In such models, explaining the result is certainly not an easy task.

In this paper we shall concentrate on the basic additive utility model. This well-known (quantitative) model assumes independence among criteria, although of course different criteria may have different weights. In other words, no synergy (either positive or negative) occurs between the different criteria. In this model, the basic approach is to construct, by elicitation techniques, a so-called *utility function* which hopefully captures the DM preferences. A natural way to generate an explanation would thus be to use this constructed function and to justify the decision by exploiting this function. Unfortunately, this constructed function is often not very meaningful to the DM.

Within the additive model, an alternative interesting approach for identifying the best decision is based on so-called *even swaps* [5]. This is basically an elimination process based on trade-offs between *pairs* of attributes (hence the name *even swaps*). Broadly speaking, in such a swap, the DM changes the consequence (or score) of an alternative on one attribute, and compensates this change with one on another attribute, so that the new alternative is equally preferred in the end. What is the point of making such swaps? Suppose you want to compare two options but that none dominates (in the Pareto sense) the other one. By replacing one option with a different but equally preferred one, the hope is that dominance will occur. The process is thus repeated until dominance can be shown to hold, allowing to progressively eliminate options. An intuitive interpretation of this method is thus to see it as a scattered exploration of the iso-preference curve (the curve where lies, even virtually, the alternatives equally preferred) of the DM. This constructive method is quite intuitive as only two attributes are involved in even swaps, and utilities are never explicitly mentioned to the DM. The idea is then that it may constitute a good starting point to justify a recommendation without refereeing explicitely to the utility function or the model used to get the solution. However this approach suffers from a limitation: by requiring each new generated option to be equally preferred to the initial one, this makes the technique poorly adapted to the context of incomplete preferences where such equivalence virtually never hold.

When utility functions are only partially known, a conservative approach consists in relying on a robust (or necessary) preference relation. In words, the relation holds if *any* possible completion of the available preferential information yields the preferential statement. The aim of this paper is to investigate whether this approach can be generalized to robust preference relations and used to generate convincing explanations. In fact, the sequence of swaps obtained at the end of the process can be seen as the reasoning steps allowing to highlight why an alternative is the best choice.

The remainder of the paper is as follows. In the next section, we provide the necessary background notions and concepts that we shall use for formulating explanations. In Section 3 we describe what is an explanation based on sequence of preference-swaps and we address the problem of its length in Section 4. In general, we argue that the simplicity of an explanation is not directly captured by its length. However, we focus on a specific case where such a simplified view is possible, and provide first results. Section 5 discusses related works.

[1] Thales Research & Technology, 91767 Palaiseau Cedex, France, email: christophe.labreuche@thalesgroup.com

[2] LIP6, Université Paris-6, 75006 Paris Cedex 06, France, email: nicolas.maudet@lip6.fr

[3] LGI, Ecole Centrale de Paris, Chatenay Malabry, France, email: vincent.Mousseau@ecp.fr

[4] LGI, Ecole Centrale de Paris, Chatenay Malabry, France, email: wassila.ouerdane@ecp.fr

2 Background and basic definitions

We consider a finite set $N = \{1, \dots, n\}$ of criteria. Each criterion $i \in N$ is described by an attribute X_i. We assume that all attributes are numerical. For discrete attributes, X_i represents integers. For continuous attributes, X_i is an interval (possibly infinite). Alternatives are considered as elements of the Cartesian product of the attributes: $X = X_1 \times \cdots \times X_n$.

2.1 Comparison of two alternatives with even swap sequences

We ground our work on the even swaps method [5] which relies on an additive utility function to compare multi-attribute alternatives $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$:

$$x \succsim y \quad \Leftrightarrow \quad \sum_{i \in N} u_i(x_i) \geq \sum_{i \in N} u_i(y_i)$$

So as to choose the best alternative, this method does not require to fully elicit the marginal utility functions, but only a limited number of trade-offs between pairs of attributes (swaps). In other words, the DM does not have to explicitly define the preferences over the attributes in general or to make any assumption about the form of the utility function. More precisely, the DM changes the consequence (or score) of an alternative on one attribute, and compensates this change with a preferentially equal change on another attribute. This creates a new fictitious alternative, that is indifferent to the previous one, with revised consequences. We use this alternative to try to eliminate the other ones. The aim of this process is to carry out even swaps that make either alternatives dominated or attributes irrelevant. To get an intuitive understanding of the process, consider the following example, largely inspired from the original example provided in [5].

Example 1. *You need to rent an office for your business and you have the choice between four alternatives $\{x, y, t, z\}$. Such options are evaluated, as it is depicted in the Table 1, on four criteria {commute (min), office services (A≻B≻C), size (m^2), cost (€) }. Of course you want to minimize the cost and commute time, while you seek to maximize the quality of service and the size. The problem is to identify the best option.*

Table 1. Evaluation of available offices

	$Commute$ (min)	$Service$	$Size$ (m^2)	$Cost$ (€)
x	25	B	700	1700
y	20	C	500	1500
z	25	A	950	1900
t	30	C	700	1750
y'	25	C	550	1500
y''	25	B	500	1750

First, we can observe that x dominates t, so t can be removed from the list of considered offices. As no more dominance exists among the remaining alternatives, the method proceeds by constructing a first trade-off (a swap), starting for instance with the office y, by asking the following question: *"What increase δ in $Size$ would exactly compensate a loss of 5 min on $Commute$?"*

This defines a new alternative $y' = (25, C, 500 + \delta, 1500)$ that is considered by the DM indifferent to y. Suppose, for instance that $\delta = 50$, then y can be substituted by $y' = (25, C, 550, 1500)$ in the analysis. As dominance cannot be applied, new trade-offs are assessed to neutralize the criterion $Service$ using $Cost$ as reference. This can be done in two ways as follows:

- what maximal increase in $Cost$ would you be prepared to pay to go from C to B on $Service$ for y' ? if the answer is 250 €, so y' is indifferent to $y'' = (25, B, 550, 1750)$
- what minimal decrease in $Cost$ would you ask if we go from A to B on service for z? if the answer is 100 €, so z is indifferent to $z' = (25, B, 950, 1800)$

Obtaining two new fictitious alternatives we can again check the dominance among the set of alternatives. We can observe that y'' is dominated by x, therefore y can be dropped. We continue the process by alternating phases of dominance and construction of trade-offs until obtaining the best alternative.

Originally, this method was designed to select the best alternative, by eliciting progressively the necessary swaps. What we can observe from this small example is that following the reasoning steps of the even swaps process we can deduce an intuitive and simple manner to explain the result to the decision maker. In fact, such an explanation will involve statements like "*an increase of δ_i on criterion i is compensated by a decrease of δ_j on criterion j*", together with dominance analysis, rather than utility computation. That is, we can simply rely on the sequence of even swaps used to identify the best alternative rather than discuss the parameters and values of the multi-attributes models.

For instance, in our example the statement $x \succ y$ can be explained in the following way: "*an increase in Size from 500m^2 to 550m^2 compensates a degradation from 20mn to 25mn in Commute, therefore office y=(20,C,500,1500) is indifferent to office y′=(25,C,550,1500). Moreover an improvement from C to B on Service is compensated by an increase in Cost from 1500€ to 1750€, therefore office y′=(25,C,550,1500) is indifferent to office y″=(25,B,550,1750)". Now observe that x is at least as good as y'' on all attributes, then x is preferred to y''. As y'' is indifferent to y, x is preferred to y.*"

Note that even if the utility functions u_i are known precisely and have been elicited with a technique different from the even swaps process, then it is possible from these utility functions to construct a sequence of even swaps that allows to show why x is preferred to y (assuming $x \succsim y$ of course). This sequence can be used as an explanation of why $x \succsim y$.

In a sense, the even swaps method already deals with some sort of incomplete preferences, as it does not require the full knowledge of the value function. The trick of the method is precisely to explore certain alternatives which stand on the same isopreference curve of the DM, until dominance occurs. However, at each step, the DM is required to answer *equivalence queries*, which may be difficult in practice. Consider again the question *"What increase in $Size$ would exactly compensate a loss of 5 min on $Commute$?"*. To such a question, the DM may be more comfortable to reply: *"I don't know, but 100 additional square meters would certainly compensate this additional commute time"*. Or, on the other hand, *"20 square meters are certainly not enough to compensate"*. By doing so however, the DM does not allow a proper even swap to occur. Instead of creating a fictitious alternative with the same utility, such statements generates a mere inequality constraint in the preferential information available. For instance, the last statement would create a new alternative $y''' = (25, C, 500 + 20, 1500)$, and it would be known that $y \succ y'''$.

In the remainder of this paper, we investigate whether the swap principle can be extended in order to allow such a wider range of preference statements.

2.2 Robust relation with the additive utility model

In what follows, we assume the decision-maker (DM) provides us some Preferential Information (PI) denoted by $\mathcal{P}$. We may consider the following types of PI:

- The most classical one is a comparison of two alternatives x and y in X, which can take different forms $x \trianglerighteq y$ (x is at least as good as y), $x \triangleright y$ (x is strictly preferred to y) or $x \equiv y$ (x is indifferent to y)
- When X_i is an interval, one may also express the existence of saturation threshold. Value s_i^+ (resp. s_i^-) is an upper (resp. lower) saturation threshold on attribute i if for all $x_{-i} \in X_{-i}$ and all $x_i \in X_i$ with $x_i \geq s_i^+$ (resp. $x_i \leq s_i^-$), $(x_i, x_{-i}) \equiv (s_i^+, x_{-i})$ (resp. $(x_i, x_{-i}) \equiv (s_i^-, x_{-i})$).
- When X_i is an interval, the DM may also express that the preference over attribute X_i is concave or convex.

Given two alternatives $z, z' \in X$, we need to determine whether z is necessarily preferred (resp. similar) to z' given the previous PI [3, 4].

For each $i \in N$, we define $\widehat{V}_i$ as the set of values on attributes X_i appearing in the PI (i.e. the union of $\{x_i, y_i\}$ for all $[x \trianglerighteq y]$, $[x \triangleright y]$ or $[x \equiv y]$ in the PI $\mathcal{P}$, of the thresholds s_i^+, s_i^- of the PI). Moreover, we define $V_i = \widehat{V}_i \cup \{z_i, z'_i\}$. The elements of V_i are denoted by $v_i^1 < v_i^2 < \cdots < v_i^{p_i}$, where $p_i = |V_i|$. The unknown variables of the model are the utility $u_i(v_i^1), u_i(v_i^2), \ldots, u_i(v_i^{p_i})$ of these points.

Throughout this paper, we will assume that the utility functions are non-decreasing, so that

$$\forall i \in N \quad u_i(v_i^1) \leq u_i(v_i^2) \leq \cdots \leq u_i(v_i^{p_i}). \tag{1}$$

Concerning the PI on the comparison of two alternatives, we have the following constraints:

$$\text{If } [x \trianglerighteq y] \in \mathcal{P}, \text{ then} \quad \sum_{i \in N} u_i(x_i) \geq \sum_{i \in N} u_i(y_i) \tag{2}$$

$$\text{If } [x \triangleright y] \in \mathcal{P}, \text{ then} \quad \sum_{i \in N} u_i(x_i) > \sum_{i \in N} u_i(y_i) \tag{3}$$

$$\text{If } [x \equiv y] \in \mathcal{P}, \text{ then} \quad \sum_{i \in N} u_i(x_i) = \sum_{i \in N} u_i(y_i) \tag{4}$$

Now if the DM expresses an upper-saturation level on attribute i, the following constraint is added:

$$\forall v_i \in V_i \text{ with } v_i > s_i^+ \quad u_i(v_i) = u_i(s_i^+), \tag{5}$$

and if the DM expresses a lower-saturation level on attribute i, the following constraint is added:

$$\forall v_i \in V_i \text{ with } v_i < s_i^- \quad u_i(v_i) = u_i(s_i^-). \tag{6}$$

Finally, if the DM expresses that the utility function on criterion i is concave, then the following constraint is added:

$$\forall j \in \{3, \ldots, p_i\}$$
$$u_i(v_i^j) \leq u_i(v_i^{j-2}) + (u_i(v_i^{j-1}) - u_i(v_i^{j-2})) \, \frac{v_i^j - v_i^{j-2}}{v_i^{j-1} - v_i^{j-2}} \tag{7}$$

and if the DM expresses that the utility function on criterion i is convex, then the following constraint is added:

$$\forall j \in \{3, \ldots, p_i\}$$
$$u_i(v_i^j) \geq u_i(v_i^{j-2}) + (u_i(v_i^{j-1}) - u_i(v_i^{j-2})) \, \frac{v_i^j - v_i^{j-2}}{v_i^{j-1} - v_i^{j-2}} \tag{8}$$

Example 2. *(1, ctd.) For attribute X_1, we have $\widehat{V}_1 = \{20, 25, 30\}$. For attribute X_3, we have $\widehat{V}_3 = \{500, 550, 700, 950\}$. We have for instance $[y \triangleright y''']$. Finally, the DM could express an upper-saturation level by stating: "Frankly, I don't need more than 100 square meters."*

This provides a set of constraints resulting from the DM statements. Typically, a number of utilities will be compatible with these constraints. In order to compare the alternatives, we compute :

$$\underline{M} := \min \sum_{i \in N} (u_i(z_i) - u_i(z'_i))$$
$$\text{under (1) – (8)}$$

and

$$\overline{M} := \max \sum_{i \in N} (u_i(z_i) - u_i(z'_i))$$
$$\text{under (1) – (8)}$$

Definition 1 (see [3, 4]). *We say that z is* necessarily at least as good *as z' (noted $z \succsim_N z'$) if $\underline{M} \geq 0$. Likewise, z' is* necessarily at least as good *as z (noted $z' \succsim_N z$) if $\overline{M} \leq 0$.*

We say that z is necessarily preferred *to z' (noted $z \succ_N z'$) if $\underline{M} > 0$. Likewise, z' is* necessarily preferred *to z (noted $z' \succ_N z$) if $\overline{M} < 0$, and z is* necessarily similar *to z' (noted $z \sim_N z'$) if $\underline{M} = \overline{M} = 0$.*

Note that it is very unlikely that the necessarily similar relation holds when the PI is incomplete.

2.3 Objective of this paper

We denote by $\geq_{\text{Pareto}}$ the Pareto ordering on X. In Section 2.1, as we have seen, when the utilities are completely fixed, showing that an option z is at least as good as z' (denoted by $z \succsim z'$) consists in exhibiting a sequence $z[1], z[2], \ldots, z[q]$ in X with $z[1] = z$ such that

$$z[1] \sim z[2] \sim \ldots \sim z[q] \geq_{\text{Pareto}} z' \tag{9}$$

(where $\sim$ means indifference) and $z[l] \sim z[l+1]$ corresponds to an even-swap.

The generalization of this approach to robust preference relation is not simple. The main reason is that the robust indifference relation $\sim_N$ almost never occur in practice. To circumvent this issue, we propose to generalize (9) in two different ways.

First of all, the equivalence relation used in even-swaps needs to be relaxed into a preference relation.

Definition 2. *We say that $z[l] \succsim_N z[l+1]$ is a* preference-swap *if there exists $i, j \in N$ such that $z[l]_i > z[l+1]_i$, $z[l]_j < z[l+1]_j$ and $z[l]_k > z[l+1]_k$ for all $k \in N \setminus \{i, j\}$.*

Secondly, one can generalize even-swap to trade-offs among coalition of more than two criteria.

Definition 3. *We say that $z[l] \succsim_N z[l+1]$ is a* preference-swap of order *p (with $p \in \{2, \ldots, n\}$ if there exists $A \subset N$ with $|A| = n - p$, $\forall i \in A$, $x_i = y_i$ and $\forall i \in N \setminus A$, $x_i \neq y_i$.*

Explanations in our context will thus be sequences generalizing (9) and in particular consisting of preference swaps of order 2 or higher. This is described more formally in the next section.

3 Explanations based on preference-swaps

We introduce the following sets:

- Δ_0 is the set of pairs (x, y) in $X \times X$ such that $[x \trianglerighteq y] \in \mathcal{P}$. In other terms, it is the set of comparative preferential information given by the decision maker.
- Δ_1 is the set of pairs (x, y) in $X \times X$ such that there exists $(x', y') \in \Delta_0$ with $x \geq_{\text{Pareto}} x'$ and $y' \geq_{\text{Pareto}} y$.
- For $p \in \{2, \dots, n\}$, Δ_p is the set of pairs of alternatives (x, y) in $X \times X$ such that $x \succsim_N y$ is a preference-swap of order p.

$\Delta_0, \Delta_1, \dots, \Delta_n$ are in increasing complexity to understand them.

Let $\Delta := \Delta_0 \cup \Delta_1 \cup \ldots \cup \Delta_n$. Clearly Δ is the set of pairs satisfying the binary relation $\succsim_N$.

Definition 4. *An* explanation *of $z \succsim_N z'$ is a sequence $z[1], z[2], \dots, z[q]$ in X with $z[1] = z$ and $z[q] = z'$ such that $(z[k], z[k+1]) \in \Delta$ for all $k \in \{1, \dots, q-1\}$. Let Ex denote the set of explanations.*

In order to define orderings over explanations, we introduce the concept of complexity of an explanation.

Definition 5. *For $(z[1], \dots, z[q]) \in Ex$, The* complexity *of $(z[1], \dots, z[q]) \in Ex$ is*

$$comp(z[1], \dots, z[q]) := \Big(C_{(z[1],\dots,z[q])}(0), C_{(z[1],\dots,z[q])}(1), \dots, C_{(z[1],\dots,z[q])}(n) \Big)$$

where

$$C_{(z[1],\dots,z[q])}(k) = \Big| \{ j \in \{1, \dots, q-1\} \,,\, (z[j], z[j+1]) \in \Delta_k \} \Big|.$$

We now define several possible orderings over explanations (see Definitions 6 and 7).

Definition 6. *For $(z[1], \dots, z[q]), (t[1], \dots, t[q']) \in Ex$, $(z[1], \dots, z[q]) \rhd_{Ex} (t[1], \dots, t[q'])$ iff $comp(z[1], \dots, z[q]) \succ_{\text{lex}} comp(t[1], \dots, t[q'])$, where $\succ_{\text{lex}}$ is the lexicographic ordering. $(a_0, \dots, a_n) \succ_{\text{lex}} (b_0, \dots, b_n)$ if there exists $i \in \{0, \dots, n\}$ such that $a_i > b_i$ and $a_j = b_j$ for all $j \in \{0, \dots, n\}$ with $j > i$.*

Definition 7. *For $(z[1], \dots, z[q]), (t[1], \dots, t[q']) \in Ex$,*

$$(z[1], \dots, z[q]) \rhd'_{Ex} (t[1], \dots, t[q']) \quad \textit{iff}$$
$$\Big(C_{(z[1],\dots,z[q])}(0) \cup C_{(z[1],\dots,z[q])}(1) \cup C_{(z[1],\dots,z[q])}(2), C_{(z[1],\dots,z[q])}(3), \dots, C_{(z[1],\dots,z[q])}(n) \Big)$$
$$\succ_{\text{lex}} \Big(C_{(z'[1],\dots,z'[q])}(0) \cup C_{(z'[1],\dots,z'[q])}(1) \cup C_{(z'[1],\dots,z'[q])}(2), C_{(z'[1],\dots,z'[q])}(3), \dots, C_{(z'[1],\dots,z'[q])}(n) \Big).$$

According to Definition 6, Δ_0 is the less complex elements of Δ, Δ_1 are the second less complex elements, ..., and Δ_n are the most complex elements. On the other hand, in Definition 7, the three sets Δ_0, Δ_1 and Δ_2 are of the same complexity and are combined.

We then look for a minimal explanation in the sense of $\rhd_{Ex}$ or $\rhd'_{Ex}$.

Example 3. *Consider the following PI on a set of 3 attributes*

$$(10, 100, 1000) \trianglerighteq (20, 80, 900) \tag{10}$$
$$(20, 70, 900) \trianglerighteq (15, 100, 1000) \tag{11}$$
$$(0, 85, 700) \trianglerighteq (30, 90, 500) \tag{12}$$
$$(30, 80, 500) \trianglerighteq (0, 85, 600) \tag{13}$$

We wish to compare $z = (10, 70, 700)$ with $z' = (15, 90, 600)$. From (10) and (11) we get

$$(10, 70, 900) \succsim_N (15, 80, 900)$$

and thus from the independence property of the model:

$$(10, 70, 700) \succsim_N (15, 80, 700). \tag{14}$$

From (12) and (13) we get

$$(30, 80, 700) \succsim_N (30, 90, 600)$$

and thus from the independence property of the model:

$$(15, 80, 700) \succsim_N (15, 90, 600). \tag{15}$$

From (14) and (15), we get the sequence

$$z = (10, 70, 700) \succsim_N (15, 80, 700) \succsim_N (15, 90, 600) = z'.$$

Hence

$$comp((10, 70, 700), (15, 80, 700), (15, 90, 600)) = (0, 0, 2, 0)$$
$$comp((10, 70, 700), (15, 90, 600)) = (0, 0, 0, 1)$$

and

$$comp((10, 70, 700), (15, 80, 700), (15, 90, 600)) \succ_{\text{lex}} comp((10, 70, 700), (15, 90, 600)).$$

The explanation $((10, 70, 700), (15, 80, 700), (15, 90, 600))$ is simpler than $((10, 70, 700), (15, 90, 600))$ in the sense of $\rhd_{Ex}$ or $\rhd'_{Ex}$.

This shows that the simplicity of an explanation is not directly captured by the length of the sequence. Short sequences involving preference-swaps of high order may not be desirable. However, if we restrict our attention to preference-swaps of order 2, the length of the sequence becomes very important to consider.

4 On the length of preference-swap sequences

We consider an explanation of $z \succsim_N z'$, that is a sequence $z[1], z[2], \dots, z[q]$ (see Definition 4).

Definition 8. *The* length *of an explanation $(z[1], \dots, z[q]) \in Ex$ is its number of elements, that is q.*

Furthermore, we assume that $(z[k], z[k+1]) \in \Delta_1 \cup \Delta_2$ (Pareto ordering and preference-swap of order 2) for all $k \in \{1, \dots, q-1\}$.

In the case of sequences of even-swaps, it is easy to see that the length of the sequence is at most n. Let us show in an example that this is not the case with sequences of preference-swaps.

Example 4. *Let us consider four criteria and the following PI:*

$$(1,0,\cdot,\cdot) \trianglerighteq (0,1,\cdot,\cdot) \tag{16}$$
$$(0,\cdot,1,\cdot) \trianglerighteq (1,\cdot,0,\cdot) \tag{17}$$
$$(\cdot,1,\cdot,0) \trianglerighteq (\cdot,0,\cdot,1) \tag{18}$$

where '$\cdot$' means that the value on this attribute does not matter provided that the alternatives on the left hand side and on the right hand side have the same value.

Consider now two alternatives $(1,0,1,0)$ *and* $(0,1,0,1)$*. It can be readily seen that* $(1,0,1,0) \succsim_N (0,1,0,1)$*. One can obtain the following sequence of only 5 comparisons from the PI* (Δ_0)*:*

$$(1,0,1,0) \underbrace{\succsim_N}_{\textit{from } (16)} (0,1,1,0) \underbrace{\succsim_N}_{\textit{from } (17)} (1,1,0,0)$$
$$\underbrace{\succsim_N}_{\textit{from } (18)} (1,0,0,1) \underbrace{\succsim_N}_{\textit{from } (16)} (0,1,0,1)$$

We restrict ourself in the rest of this paper to preference swaps of order 2. Can we get an upper bound on the length q?

4.1 Unboundedness of the length of the sequence

We start with a negative result. If we make no assumption on the values of the attributes taken by the alternatives in the PI, the length q is unbounded. This is shown by the following lemma when $n \geq 3$.

Lemma 1. *Consider* $n \geq 3$*. Let* $z, z' \in \mathbb{R}^N$ *where* $z_i \neq z_i$ *for at least three attributes* i_1, i_2, i_3 *with* $z_{i_1} < z'_{i_1}$*,* $z_{i_2} > z'_{i_2}$ *and* $z_{i_3} > z'_{i_3}$*. Assume that* $[\min(z_i, z'_i), \max(z_i, z'_i)] \subseteq X_i$ *for all* $i \in \{i_1, i_2, i_3\}$*. Then for every* $k \in \mathbb{N}^*$*, there exists some PI such that* $z \succsim_N z'$ *and the minimal length of the explanation in* $\Delta_0 \cup \Delta_1 \cup \Delta_2$ *is at least* k*.*

Note that we can also add Δ_1 on top of Δ_2 in the previous lemma.
Proof : Let $n \geq 3$, $k \in \mathbb{N}^*$ and $p = \lfloor \frac{k}{2} \rfloor$. Without loss of generality, take $i_1 = 1$, $i_2 = 2$, $i_3 = 3$, $z_1 = 0$, $z_2 = 0$, $z_3 = 0$, $z'_1 = 1$, $z'_2 = -1$ and $z'_3 = -1$. Assume that $X_1 \supseteq [0,1]$, $X_2 \supseteq [-1,0]$ and $X_3 \supseteq [-1,0]$. Consider the following PI:

$$\begin{aligned}
&\forall j \in \{0,\dots,p-1\}\\
&\qquad \Big(\frac{2j}{2p}, -\frac{j}{p}, -\frac{j}{p}, z_{-123}\Big) \trianglerighteq \Big(\frac{2j+1}{2p}, -\frac{j+1}{p}, -\frac{j}{p}, z_{-123}\Big)\\
&\forall j \in \{0,\dots,p-1\}\\
&\qquad \Big(\frac{2j+1}{2p}, -\frac{j+1}{p}, -\frac{j}{p}, z_{-123}\Big)\\
&\qquad\qquad \trianglerighteq \Big(\frac{2j+2}{2p}, -\frac{j+1}{p}, -\frac{j+1}{p}, z_{-123}\Big)\\
&\forall i \in \{4,\dots,n\}\\
&\qquad (z'_1, \dots z'_{i-1}, z_i, z_{i+1}, \dots, z_n)\\
&\qquad\qquad \equiv (z'_1, \dots z'_{i-1}, z'_i, z_{i+1}, \dots, z_n)
\end{aligned}$$

With this PI, we clearly obtain $z \succsim_N z'$ and the sequence

$$\begin{aligned}
z = &(0,0,0,z_{-123}) \succsim_N \Big(\frac{1}{2p}, -\frac{1}{p}, 0, z_{-123}\Big)\\
&\succsim_N \Big(\frac{2}{2p}, -\frac{1}{p}, -\frac{1}{p}, z_{-123}\Big) \succsim_N \cdots\\
&\succsim_N \Big(\frac{2p-2}{2p}, -\frac{p-1}{p}, -\frac{p-1}{p}, z_{-123}\Big)\\
&\succsim_N \Big(\frac{2p-1}{2p}, -1, -\frac{p-1}{p}, z_{-123}\Big) \succsim_N (1,-1,-1,z_{-123})\\
&\equiv_N (z'_1, \dots, z'_4, z_5\ \dots, z_n) \equiv_N \cdots \equiv_N z'.
\end{aligned}$$

This sequence is of length $(2\,p+1) + (n-3) \geq k + (n-3)$.

We obtain the following constraints from the PI

$$\begin{aligned}
&\forall j \in \{0,\dots,p-1\}\\
&\qquad u_1\Big(\frac{2j}{2p}\Big) + u_2\Big(-\frac{j}{p}\Big) \geq u_1\Big(\frac{2j+1}{2p}\Big) + u_2\Big(-\frac{j+1}{p}\Big)\\
&\forall j \in \{0,\dots,p-1\}\\
&\qquad u_1\Big(\frac{2j+1}{2p}\Big) + u_3\Big(-\frac{j}{p}\Big) \geq u_1\Big(\frac{2j+2}{2p}\Big) + u_3\Big(-\frac{j+1}{p}\Big)\\
&\forall j \in \{0,\dots,p-1\}\\
&\qquad u_1\Big(\frac{2j}{2p}\Big) \leq u_1\Big(\frac{2j+1}{2p}\Big) \leq u_1\Big(\frac{2j+2}{2p}\Big)\\
&\forall j \in \{0,\dots,p-1\}\\
&\qquad u_2\Big(-\frac{j}{p}\Big) \geq u_2\Big(-\frac{j+1}{p}\Big) \text{ and } u_3\Big(-\frac{j}{p}\Big) \geq u_3\Big(-\frac{j+1}{p}\Big)\\
&\forall i \in \{4,\dots,n\} \qquad u_i(z_i) = u_i(z'_i)
\end{aligned}$$

As the alternatives appearing in the PI use different values on the attributes, the necessary relation is composed of $\trianglerighteq$ with the Pareto ordering. From this, one cannot skip any comparison in the sequence. Hence there is no explanation sequence of preference-swap of order 2 strictly shorter than $2\,p+1$. Note that the $n-3$ comparisons for the explanation of $(1,-1,-1,z_{-123}) \equiv_N z'$ can be done with only one Pareto comparison (depending on the values of z and z'). ■

The only hope to have an upper bound on the length q is thus to restrict the number of values that the PI can take on each attribute.

4.2 Some solution to bound the length of sequences

We take *binary alternatives* as an extreme case. Assume now that there exists two values on each attribute (denoted by 0 and 1) such that the alternatives appearing in the PI belong to $\{0,1\}^N \subseteq X$.

Lemma 2. *Let* $w_i := u_i(1) - u_i(0)$*. For every* $A, B \subseteq N$*, we have*

$$\begin{aligned}
&(1_A, 0_{-A}) \succsim_N (1_B, 0_{-B})\\
&\iff \sum_{i\in A\setminus B} w_i \geq \sum_{i \in B\setminus A} w_i \ \textit{ for all } w \textit{ compatible with the PI.}
\end{aligned}$$

Proof :

$$\begin{aligned}
&(1_A, 0_{-A}) \succsim_N (1_B, 0_{-B})\\
&\iff \sum_{i\in A} u_i(1) + \sum_{i\in N\setminus A} u_i(0) \geq \sum_{i\in B} u_i(1) + \sum_{i\in N\setminus B} u_i(0)\\
&\iff \sum_{i\in A\setminus B} u_i(1) + \sum_{i\in B\setminus A} u_i(0) \geq \sum_{i\in B\setminus A} u_i(1) + \sum_{i\in A\setminus B} u_i(0)\\
&\iff \sum_{i\in A\setminus B} w_i \geq \sum_{i\in B\setminus A} w_i
\end{aligned}$$

■

Let $\mathcal{W}$ be the set of $w \in \mathbb{R}_+^N$ s.t. $\sum_{i \in A \setminus B} w_i \geq \sum_{i \in B \setminus A} w_i$ for all PI $(1_A, 0_{-A}) \trianglerighteq (1_B, 0_{-B})$.

Lemma 3. *For $x, y \in \{0,1\}^N$. If $(x,y) \in \Delta_2$, there exist i and j are such that $x_i = 1$, $y_i = 0$, $x_j = 0$, $y_j = 1$ and $x_k = y_k$ for all $k \in N \setminus \{i,j\}$. Then we have*

$$(x,y) \in \Delta_2 \quad \Longleftrightarrow \quad w_i \geq w_j \text{ for all } w \in \mathcal{W}.$$

Proof : Clear from Lemma 2. ■

From Lemma 3, it is apparent that the length of sequences using only terms in Δ_2 is bounded. If the length of the sequence is large enough, one necessarily finds relations of the form $w_i \geq w_j$ and $w_j \geq w_k$ in the sequence. Clearly, these two relations can be replaced by the comparison $w_i \geq w_k$, by transitivity. We believe that this allows to keep the length of sequences under a given value (not provided in this paper). Let us illustrate this intuition on an example.

Example 5 (Example 4, ctd.). *By Lemma 3, (16) is equivalent to*

$$w_1 \geq w_2, \tag{19}$$

(17) is equivalent to

$$w_3 \geq w_1, \tag{20}$$

and (18) is equivalent to

$$w_2 \geq w_4. \tag{21}$$

Consider now the two alternatives $(1,0,1,0)$ and $(0,1,0,1)$ (note that $(1,0,1,0) \succsim_N (0,1,0,1)$). The explanation in Example 4 may seem simple for the user since it is based only on PI. However, it uses the first example (16) twice, giving the feeling that it is circular.

Actually, another explanation can be constructed. First of all, let us note that adding (19) and (20), we obtain

$$w_3 \geq w_2, \tag{22}$$

and adding (19) and (21), we get

$$w_1 \geq w_4. \tag{23}$$

Hence the following explanation is reached:

$$(1,0,1,0) \underbrace{\succsim_N}_{\textit{from } (22)} (0,0,1,1) \underbrace{\succsim_N}_{\textit{from } (23)} (0,1,0,1).$$

The length of this sequence is only 3 and it is composed of comparisons in Δ_2. This explanation is more direct than the previous one, and seams better for the user. It is better than the explanation of Example 5 in the sense of $\rhd'_{Ex}$.

In Example 5, the reduction of the length of the explanation is based on the trick described just before this example.

5 Related works

The idea of even swap can be found in negotiation in multi-agent systems [2]. There is a difference between the *concession* and *trade-off*. In a concession, the agent give up on something and it is ready to accept an offer which overall utility is smaller than a previous offer. By contrast, in a trade-off analysis, the agent explores the set of options that yield the same overall utility (a level-set), and one looks at balancing utilities among the criteria (while remaining on the same level curve) so that another agent will be better satisfied. This implies decreasing the expectation on a criterion while increasing the expectation on another criterion. For instance, a customer may be ready to pay more if the item is delivered faster. The main idea of [2] is to instantiate the idea of trade-off: the proposer at an iteration of the protocol shall propose, among all options that yield a given satisfaction to it, the option that is *best* for the other(s) agent(s). The main idea of the paper is to represent the preferences of an agent by a *similarity measure* to the last proposal made by this agent.

Another potentially fruitful connection to explore is with planning problems, where the objective is to find how to sequentially apply different operators so as to attain an objective state from an initial state. Preference-swaps can be seen as operators, and the objective state to reach is an alternative exhibiting the desired dominance.

6 Conclusion

This paper investigates the problem of providing minimal explanation by relying on an extension of the even swaps. A first contribution of the paper is to set up the framework allowing to generalize such an approach to more general preference statements, and to be used to generate convincing explanation (on the basis on so-called preference swaps). Another natural extension is to trade-off involving more than two criteria, although this may quickly be difficult to handle for the DM. The first result put forward in this paper is negative: it states that in the absence of any restriction on the size of the domain considered for the value of attributes (in particular when such a domain is an interval over the reals), the sequence of preference-swaps may not be bounded. This challenges the practical use of this technique in this case. It is thus natural to consider restricted domains: we show that in binary domains positive results (bounded sequence) can hold, and sketch possible solutions to reduce the length of explanations.

REFERENCES

[1] G. Carenini and J.D. Moore, 'Generating and evaluating evaluative arguments', *AIJ*, **170**, 925–952, (2006).

[2] P. Faratin, C. Sierra, and N.R. Jennings, 'Using similarity criteria to make issue trade-offs in automated negotiations', *AIJ*, **142**, 205–237, (2002).

[3] S. Greco, B. Matarazzo, and R. Słowinski, 'Ordinal regression revisited: Multiple criteria ranking with a set of additive value functions', *European Journal of Operational Research*, **191**, 416–436, (2008).

[4] S. Greco, R. Słowinski, J. Figueira, and V. Mousseau, 'Robust ordinal regression', in *Trends in Multiple Criteria Decision Analysis*, 241–284, Springer Verlag, (2010).

[5] J. Hammond, R. Keeney, and H. Raiffa, 'Even Swaps: a rational method for making trade-offs', *Harvard Business Review*, 137–149, (1998).

[6] J. L. Herlocker, J. A. Konstan, and J. Riedl, 'Explaining collaborative filtering recommendations', in *CSCW*, pp. 241–250, (2000).

[7] D.A. Klein, *Decision analytic intelligent systems: automated explanation and knowledge acquisition*, Lawrence Erlbaum Associates, 1994.

[8] Ch. Labreuche, 'A general framework for explaining the results of a multi-attribute preference model', *AIJ*, **175**, 1410–1448, (2011).

[9] Ch. Labreuche, N. Maudet, and W. Ouerdane, 'Justifying dominating options when preferences are incomplete', in *Proceedings of ECAI-12*, Montpellier, FR, (2012). To appear.

[10] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, 'MoviExplain: a recommender system with explanations', in *Proceedings of the third ACM conference on Recommender systems (RecSys'09)*, pp. 317–320, New York, NY, USA, (2009). ACM.