# Explaining robust additive utility models by sequences of preference swaps

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

# Explaining robust additive utility models by sequences of preference swaps

**K. Belahcene[1] · C. Labreuche[2] · N. Maudet[3] · V. Mousseau[1] · W. Ouerdane[1]**

**Abstract** As decision-aiding tools become more popular everyday—but at the same time more sophisticated—it is of utmost importance to develop their explanatory capabilities. Some decisions require careful explanations, which can be challenging to provide when the underlying mathematical model is complex. This is the case when recommendations are based on incomplete expression of preferences, as the decision-aiding tool has to infer despite this scarcity of information. This step is key in the process but hardly intelligible for the user. The robust additive utility model is a necessary preference relation which makes minimal assumptions, at the price of handling a collection of compatible utility functions, virtually impossible to exhibit to the user. This strength for the model is a challenge for the explanation. In this paper, we come up with an explanation engine based on sequences of preference swaps, that is, pairwise comparison of alternatives. The intuition is to confront the decision maker with "elementary" comparisons, thus building incremental explanations. Elementary

✉ N. Maudet
nicolas.maudet@lip6.fr

K. Belahcene
khaled.belahcene@centralesupelec.fr

C. Labreuche
christophe.labreuche@thalesgroup.com

V. Mousseau
vincent.mousseau@centralesupelec.fr

W. Ouerdane
wassila.ouerdane@centralesupelec.fr

1 LGI, CentraleSupélec, Université Paris-Saclay, Chatenay Malabry, France

2 Thales Research & Technology, 91767 Palaiseau Cedex, France

3 Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 75005 Paris, France

here means that alternatives compared may only differ on two criteria. Technically, our explanation engine exploits some properties of the necessary preference relation that we unveil in the paper. Equipped with this, we explore the issues of the existence and length of the resulting sequences. We show in particular that in the general case, no bound can be given on the length of explanations, but that in binary domains, the sequences remain short.

**Keywords** Multicriteria decision making · Explanation · Necessary preference relation

## 1 Introduction

A decision-aiding problem consists in formalizing the problem and eliciting the preferences of the decision maker (DM) to make recommendations. In many decision contexts, only providing recommendations based on the elicited preference model is insufficient. In fact, decision makers may want explanations which justify in a convincing way such recommendations. Indeed, justifying and explaining a rationale for a decision is almost as important as the recommendation itself. Building a convincing explanation is often required when the DM cannot be assumed to have any mathematical background, as in the case of online recommender systems, where it has been shown that explanations improve the acceptability of the recommended choice (Pu and Chen 2007; Symeonidis et al. 2009; O'Sullivan et al. 2007). But even experts of a domain can have huge difficulty to grasp with the mathematical models underlying some decision-aiding tools. In this case, it is not satisfactory to just put forward the preference model and the resulting recommendation. Although technically, of course, this model does contain all the information on which the recommendation is based, the format is unlikely to be suitable for presentation. Hence, the need for a synthetic, short and easy to understand explanation.

Depending on the setting considered, the nature of an explanation may greatly vary. Sometimes, even vague statements can prove effective to persuade a specific decision maker. But when the decision is important, or when the decision maker is accountable for the decision chosen (a situation where the decision needs to be justified to some other stakeholders who did not participate to the decision process), the explanation should be viable even under close scrutiny. Complete explanations provide some guarantees in that respect since they bring all the information required to reconstruct the rationale of the recommendation—in a sense they formally "prove" it.

In this paper, we shall thus concentrate on complete explanations in the context of decisions involving multiple criteria. More precisely, we propose to construct pieces evidence that support unambiguously a binary preference relation between two alternatives described along multiple attributes. Such a relation is very often not explicit but elicited by some algorithmic process from preference information stated by the decision maker. In our case, this initial information takes the form of pairwise comparisons of alternatives. This initial input may be scarce, in any case not sufficient to fully specify the preference relation of the DM. To deal with the incompleteness of the

expression of preferences, the decision-aiding method will make use of an inference step. It is usually an involved process, challenging for explanation.

Our explanation engine takes inspiration from the even-swaps method (Hammond et al. 1998), an elicitation procedure assuming an additive value model of preferences and based on trade-offs between pairs of attributes (hence the name even swaps). Broadly speaking, in each swap, the DM changes the score of an alternative on one attribute, and compensates this change with one another attribute, so that the new alternative is equally preferred. The process is repeated until dominance can be shown to hold, allowing to progressively eliminate attributes. The idea is to use similar sequences as explanations of a recommendation. The problem with such a process is that it requires each new generated option to be equally preferred to the initial one, which is poorly adapted to the context of incomplete preferences (as such an equivalence virtually never holds). To circumvent this issue, we propose a generalization of even swaps to preference swaps, and simply exhibit a comparison between alternatives. To keep the sequence as simple as possible, we aim at constructing a sequence of low-order preference swaps between two alternatives, in the sense that two successive alternatives in the sequence only differ on a few criteria. In the end, the resulting explanations can be appreciated through the number of swaps (length of an explanation) and the order of the most complex swap involved in the explanation (the number of differing attributes between the two alternatives). An interesting feature of this explanation engine is that it can be shown to operate on any value-based decision models satisfying some basic axiomatic properties.

We propose thereafter to instantiate the engine by relying on a robust additive utility model (Greco et al. 2008, 2010). The robust (necessary) relation is constructed according to preference information provided by the decision maker. However, contrary to the classical additive models, in the robust approach the relation holds if any possible completion of the available preferential information yields the preferential statement. In fact, in additive models, such as UTA (Additive UTility) (Jacquet-Lagrèze and Siskos 1982), the preferential information brought by the DM is not sufficient to uniquely specify the utility functions (utility functions are only partially known), but the multiplicity of the compatible utility is not taken into account. To provide a solid mechanism to construct explanations for necessary preference relations, we come up with a new characterization of the necessary preference relation, based on the notion of covectors, that facilitates its implementation in the explanation engine.

In a nutshell, our proposal is thus to decompose a robust preference into several simpler recommendations. This paper investigates this idea and tackles the following questions: are such explanations guaranteed to exist, in particular if we restrict the order of swaps? And if they do exist, can we exhibit upper bounds on their length? As we shall see, the answer to this question crucially depends on the number of distinct values referenced by the preference information. In binary domains, we provide an efficient algorithm which constructs such explanations.

The remainder of the paper is as follows. Section 2 presents the explanation engine which relies on the construction of sequence of preference swaps between two alternatives. In Sect. 3, we define and analyze the value-based robust preference relation. Section 4 proposes results concerning the construction of explanations when preference information is expressed using two levels on each criterion. Finally, Sect. 5

studies how our contributions relate to previous work and proposes extensions and further work.

## 2 The explanation engine

### 2.1 Presentation of the decision context

This article is set in the context of Multicriteria Decision Making, where a decision maker has to decide between several alternatives explicitly measured on several criteria. We call $N$ the set of criteria, so alternatives are represented by elements of a set $\mathbb{X} = \prod_{i \in N} \mathbb{X}_i$, where the attribute set $\mathbb{X}_i$ for criterion $i \in N$ is totally ordered by the relation $\succsim_i$ denoting preference.

*Example 1* You need to chose a hotel for a business trip, and you are undecided between four options described by the performance table (see below). Such options are evaluated according to four criteria.

- The room comfort, ranging from $*$ (low) to $*****$ (high).
- The presence of a restaurant on the premise, with yes preferred to no.
- The commute time to the convention center, the lower the better.
- The cost, the lower the better.

| Hotel | Comfort | Restaurant | Commute time (min) | Cost |
|---|---|---|---|---|
| $h_1$ | $5*$ | Yes | 10 | 160 $ |
| $h_2$ | $4*$ | Yes | 45 | 180 $ |
| $h_3$ | $3*$ | No | 15 | 60 $ |
| $h_4$ | $2*$ | No | 60 | 50 $ |

**Definition 1** (*ceteris paribus sets of pairs of alternatives*) for any partition of criteria $N = A \cup (N \backslash A)$ and corresponding partition of attributes $x_A \in \prod_{i \in A} \mathbb{X}_i$ and $x_{-A} \in \prod_{i \notin A} \mathbb{X}_i$, $(x_A, x_{-A})$ is an alternative belonging to $\mathbb{X}$. For $x_A, y_A \in \prod_{i \in A} \mathbb{X}_i$, we define the ceteris paribus set $(x_A, y_A)_{cp}$ as the set of every possible completions of the pair:

$$(x_A, y_A)_{cp} := \left\{ ((x_A, c_{-A}), (y_A, c_{-A})), c_{-A} \in \prod_{i \notin A} \mathbb{X}_i \right\}$$

When comparing two alternatives, the criteria may unanimously rank one alternative above the other.

**Definition 2** (*weak Pareto dominance*)

$$\forall (x, y) \in \mathbb{X} \times \mathbb{X}, \quad (x, y) \in \mathcal{D} \iff \forall i \in N, \quad x_i \succsim_i y_i$$

**Definition 3** (*sets of shared and differing attributes*)

$$\forall x, y \in \mathbb{X},\ N^{=}_{(x,y)} := \{i \in N : x_i = y_i\} \quad \text{and} \quad N^{\neq}_{(x,y)} := \{i \in N : x_i \neq y_i\}$$

Preferences of the decision maker make up a binary relation between alternatives $\mathcal{R} \subset \mathbb{X}^2$, so that $(x, y) \in \mathcal{R}$ denotes the (weak) preference of alternative $x$ over alternative $y$. More often than not, this relation is not explicit over $\mathbb{X}^2$, but elicited, extrapolated by some algorithmic process from preference information stated by the decision maker. In this context, an explanation of a statement $(x, y) \in \mathcal{R}$ is a piece of supportive evidence, enabling the decision maker to assert this preference. The explanation engine we develop in Sect. 4 assumes the relation $\mathcal{R}$ satisfies three core axioms:

Axiom 1 (compatibility to dominance) $\mathcal{D} \subset \mathcal{R}$
Axiom 2 (transitivity) $\forall x, y, z \in \mathbb{X} : (x, y) \in \mathcal{R} \wedge (y, z) \in \mathcal{R} \Rightarrow (x, z) \in \mathcal{R}$
Axiom 3 (cancelation) *For any ceteris paribus set of pairs $s$, if a pair of alternatives in $s$ is in relation $\mathcal{R}$, then every pair of alternatives in $s$ is in relation $\mathcal{R}$.*

*Example 2* (Ex. 1 cont.) Hotel $h_1$ dominates hotel $h_2$, as it is at the same time more comfortable, closer to the convention center, and cheaper, while being as good on the criterion presence of a restaurant. Thus, $(h_1, h_4) \in \mathcal{D}$, and $(h_1, h_4) \in \mathcal{R}$.

Hotels $h_3$ and $h_4$ share their absence of a restaurant on the premise. Thus, preference of one over the other ignores the criterion restaurant and is represented by the ceteris paribus set $\big((3*, \_\__r, 15\text{min}, 60\$), (2*, \_\__r, 60\text{min}, 50\$)\big)_{cp}$, where $\_\__r$ stands for any value in $\mathbb{X}_r$. As $\mathbb{X}_r$ contains two distinct values, there are two pairs in this set, and $(h_3, h_4) \in \mathcal{R} \iff \big((3*, \text{yes}, 15\text{min}, 60\$), (2*, \text{yes}, 60\text{min}, 50\$)\big) \in \mathcal{R}$.

Compatibility to dominance is a fundamental requirement to correctly model preference. Transitivity asks for the model to eschew Condorcet's paradox and to behave like a preorder relation. Cancelation implies the preferential independence of criteria, so that only differing attributes have a say in determining preference.

Many popular, value-based decision models fulfill these requirements, measuring the fitness of an alternative by combining its attributes in a single index, using the average, or weighted average of the attributes, or some carefully chosen separable, parametric value function of the attributes. So does the robust additive model, described in Sect. 3.

### 2.2 Sequences of low-order preference swaps

The explanation engine detailed in what follows is reminiscent of the even-swaps method (Hammond et al. 1998), an interactive and constructive elicitation procedure assuming an additive value model of preferences. This method aims at identifying, between two options $x$ and $y$, which one is preferred to the other, without explicitly constructing the utility functions. This is basically an elimination process based on trade-offs between pairs of attributes ("swaps"), that can be seen as a scattered exploration of the iso-preference curve of the decision maker (the curve where lies, even

virtually, the alternatives equally preferred).[1] Broadly speaking, in such a swap, the decision maker changes the consequence (or score) of an alternative on one attribute, and is asked to compensate for this change by acting on another attribute, so that the new alternative is equally preferred in the end ("even"). This creates a new fictitious alternative, that is indifferent to the previous one, with revised consequences. By replacing one option (say $x$) with a different but equally preferred one, the hope is that dominance will occur over $y$. The process is thus repeated allowing to progressively cancel irrelevant attributes, until dominance can be shown to hold, and building a sequence $x \sim e_1 \sim e_2 \ldots \sim e_{n-1}$, so that either $(e_{n-1}, y) \in \mathcal{D}$ or $(y, e_{n-1}) \in \mathcal{D}$.

Considered through the prism of explanation, even swaps have several very attractive features.

- Each swap involves only attributes on two criteria.
- The method entirely references alternatives inside the decision space $\mathbb{X}$, but not artifacts of the underlying decision model (such as utility functions), or relations between criteria.

However, the even-swaps approach suffers from a severe limitation, as it requires each new generated option to be equally preferred to the initial one. This is a steep requirement, for several reasons.

- Indifference requires compensation between criteria (Krantz et al. 1971), barring the possibility that some difference in attributes on one criterion could be impossible to compensate for.
- Indifference requires solvability of the attribute scales (Krantz et al. 1971), which naturally occurs on continuous scales but rarely between discrete ones.
- Indifference imposes a high cognitive workload on the decision maker, as it repeatedly asks for cardinal information.
- Indifference is hardly a robust notion, especially in the context of incomplete preferences.[2]

Consequently, we propose a generalization of even swaps that avoids these issues, while retaining their simplicity and being well suited to the context of incomplete preference. In preference swaps, the assumption of indifference between consecutive alternatives in the sequence $e_0 := x, e_1, \ldots, e_n := y$ is relaxed and replaced by an assumption of (weak) preference: $(e_{j-1}, e_j) \in \mathcal{R}$. The following definitions extend the notion of swaps to pairs of alternatives differing on more than two criteria.

[1] Equally preferred, or indifferent, alternatives are pairs in the symmetric part of the relation $\mathcal{R}$ : $\forall x, y \in \mathbb{X}, x \sim y \iff \{(x, y), (y, x)\} \subset \mathcal{R}$.

[2] We note that [MH07,MH05] also propose to enrich the original even swaps method in a way that accounts for incomplete knowledge about the value function. They consider a "practical dominance" notion when the value of an alternative is at least as high as the value of another one with every feasible combination of parameters, this perspective being very close to the one developed in [GMS08] (see next section). However, this notion is only used for pre-processing dominated alternatives, and not integrated in the swap process, let alone used for explanatory purposes.

**Definition 4** (*preference swaps of order k*)

$$\forall k \in \mathbb{N}^{\star},\ \Delta_k = \begin{cases} \mathcal{D}, & if\, k = 1 \\ \{(x, y) \in \mathcal{R} \backslash \mathcal{D},\ |N^{\neq}_{(x,y)}| = k\}, & if\, k > 1 \end{cases}$$

This definition leverages two properties assumed for the relation $\mathcal{R}$. As $\mathcal{D} \subset \mathcal{R}$ (Axiom 1), $\mathcal{R} = \bigcup_{k \leq |N|} \Delta_k$: any pair in $\mathcal{R}$ is a swap, and we try to reflect its cognitive difficulty, in the context of explanation, by its order, the lower, the simpler. Dominance relations are deemed to be simple, and are given the lowest order. For relations requiring trade-offs between criteria, we define the order of a swap as the number of differing attributes between the two alternatives.

We can now define the notion of explanation by a sequence of preference swaps. This type of explanation transforms one single preference statement $(x, y) \in \mathcal{R}$ that the decision maker needs to understand to a sequence of several preference statements $(e_{j-1}, e_j) \in \mathcal{R}$. The idea is that the initial preference $(x, y)$ is complex to understand as the values of $x$ and $y$ differ on most (if not all) attributes, whereas each intermediate comparison $(e_{j-1}, e_j)$ is much easier to understand as it involves alternatives differing only on a few attributes.

**Definition 5** (*Explanation by preference swaps, order and length*) $\forall (x, y) \in \mathbb{X}^2, n \in \mathbb{N}$, an explanation of length $n$ of the pair $(x, y)$ for the relation $\mathcal{R}$ is a tuple $(e_0, e_1, \ldots, e_n) \in \mathbb{X}^n$ such that $e_0 = x$, $e_n = y$ and $\forall j \in \mathbb{N} : 1 \leq j \leq n$, $(e_{j-1}, e_j) \in \mathcal{R}$. The order of such explanation is the integer $k = \max\{k \in \mathbb{N} : \exists (j \in \mathbb{N} : 1 \leq j \leq n),\ (e_{j-1}, e_j) \in \Delta_k\}$.

As $\mathcal{R}$ is transitive (axiom 2), an explanation of a pair of alternatives is a proof that this pair belongs to $\mathcal{R}$. One can note that somehow we have two elements to appreciate the quality of the explanation. First, the number of comparisons (swaps) used to construct such an explanation. Second, its complexity which is defined by the most complex or difficult swap (with the highest order).

However, an important question regarding a pair $(x, y) \in \mathbb{X}^2$ is whether it is possible to find an explanation by preference swaps of the pair $(x, y)$. The answer obviously depends on the bound, if any, placed upon the order of the swaps linking the explanation chain, or the length of the explanation chain. In this article, we address this issue by first putting a cap on the order (the order of an explanation being the order of its most difficult link), then looking for the possibility of finding an explanation subject to this order constraint. Then, if explanations are available, we look for short ones.

**Definition 6** (*pairs explainable by low-order preference swaps*) $\forall k \in \mathbb{N}$, $\mathcal{E}_k(\mathcal{R})$ is the set of pairs $(x, y) \in \mathbb{X}^2$ for which there exists an explanation of any length and of order at most $k$.

There is a trade-off between the value of the cap placed upon the order of explanations and the set of pairs we are able to explain.

**Theorem 1** *(*hierarchy of binary relations)

$$\mathcal{D} = \mathcal{E}_1(\mathcal{R}) \subseteq \mathcal{E}_2(\mathcal{R}) \subseteq \cdots \subseteq \mathcal{E}_k(\mathcal{R}) \subseteq \cdots \subseteq \mathcal{E}_{|N|}(\mathcal{R}) = \mathcal{R}$$

*Proof* – For any $(x, y) \in \mathcal{E}_1(\mathcal{R})$, there is a tuple $(e_0, e_1, \ldots, e_n) \in \mathbb{X}^n$ such that $e_0 = x, e_n = y$ and $\forall j \in \mathbb{N} : 1 \leq j \leq n\ ,\ (e_{j-1}, e_j) \in \mathcal{D}$. As relation $\mathcal{D}$ is transitive, $(x, y) \in \mathcal{D}$, hence $\mathcal{D} \supseteq \mathcal{E}_1(\mathcal{R})$. Conversely, the sequence $e_0 := x, e_1 := y$ is an explanation of length one and of order one of any pair $(x, y) \in \mathcal{D}$ , hence $\mathcal{D} \subseteq \mathcal{E}_1(\mathcal{R})$. Finally, $\mathcal{D} = \mathcal{E}_1(\mathcal{R})$.

– For $k' \geq k$, an explanation of order at most $k$ is also an explanation of order at most $k'$, so $\mathcal{E}_k(\mathcal{R}) \subseteq \mathcal{E}_{k'}(\mathcal{R})$.

– The sequence $e_0 := x, e_1 := y$ is an explanation of length one and of order $|N^{\neq}_{(x,y)}|$ of any pair $(x, y) \in \mathcal{R}$. As $|N^{\neq}_{(x,y)}| \leq |N|$, $\mathcal{R} \subseteq \mathcal{E}_{|N|}(\mathcal{R})$. Conversely, an explanation (of any order and any length) of a pair $(x, y)$ is a proof by transitivity of $(x, y) \in \mathcal{R}$, thus $\mathcal{R} \supseteq \mathcal{E}_{|N|}(\mathcal{R})$. Finally, $\mathcal{R} = \mathcal{E}_{|N|}(\mathcal{R})$. □

### 2.3 Some technical challenges with explanation

In this section, we highlight a number of key issues affecting the feasibility (from a theoretical, algorithmic point of view), and the satisfaction of the decision maker, recipient of the explanation (from a practical point of view): the existence, or not, of an explanation, its length and the values of the attributes referenced in the sequences. In fact, throughout this work we investigate the conditions (in terms of order of swaps) under which an explanation may exist. Moreover, we show also that the length of an explanation depends on the number of values of the attributes in the sequence (see Sect. 4 for the binary case). However, many other interesting questions related to these issues remain open and are not addressed in this paper (see Sect. 5).

– *Existence of an explanation* The first point to consider in the construction of an explanation is to make sure there is one to be found. Without any additional assumption, for a low cap $k$ placed upon the order, it is quite possible that there are some statements that cannot be explained by preference swaps of order at most $k$. Technically, checking if we can explain a statement $(x, y)$ in $\mathcal{E}_k(\mathcal{R})$, can be seen as determining if the vertices $x$ and $y$ are connected in the directed graph of the relation $\bigcup_{1 \leq n \leq k} \Delta_n$. Of course, we have efficient algorithms to test if a graph is connected or not (Even and Tarjan 1975). However, it may be challenging to use them with regard to the size of the graph (possibly infinite, and, when finite, exponential in the number of criteria) in our context.

– *Length of an explanation* A second point that we address here is the length $n$ of the sequence. Indeed, keeping the explanation short has a great bearing on its ability to convince. Even if each elementary comparison $(e_{j-1}, e_j) \in \mathcal{R}$ is trivial for the decision maker, the overall sequence $(x, e_1, \ldots, e_{n-1}, y)$ cannot be seen as a convincing explanation if it is too long. One then looks for the shortest possible explanations, and hope for an upper bound on this minimal size. Finding the shortest explanation means resolving the problem of shortest path in the directed graph $\bigcup_{1 \leq n \leq k} \Delta_n$. Thus, the length of a shortest explanation is bounded by the

diameter of this graph.[3] Finding such a diameter is a classical problem in graph theory for which we have polynomial algorithm in terms of number, if finite, of vertices and edges [see for instance (Aingworth et al. 1996)]. Unfortunately, as soon as there are three criteria measured on infinite scales, this diameter has no upper bound, as expressed by the following theorem.

**Theorem 2** (long explanations) *For any integer $p$, if there is a subset $A \subseteq N : |A| = 3$ and $\forall i \in A,\ |\mathbb{X}_i| \geq p$, then there is a relation $\mathcal{R}$ satisfying axioms 1, 2 and 3, and a pair $(x, y) \in \Delta_3$ such that $(x, y) \in \mathcal{E}_2(\mathcal{R})$ and any explanation of $(x, y)$ by preference swaps of order at most 2 has a length greater than $2p$.*

*Proof* The proof requires instantiating the relation $\mathcal{R}$, and is presented in Appendix 1. We make use of the necessary preference relation introduced in the Sect. 3, for some carefully built preference information. □

– *Values of the terms in the sequence* Another point concerns the choice of the values of the intermediate alternatives $e_1, \ldots, e_{n-1}$ on the different attributes. If these values are not chosen carefully, we believe they can induce a cognitive load to the decision maker, when she analyzes the sequence. Several options may be considered for these values. A "dynamic" option is to restrict the values of the attributes of $e_1, \ldots, e_{n-1}$ to the value of the attributes of $x$ or $y$. This choice seems suitable to a decision context where there is only one statement $(x, y) \in \mathcal{R}$ to explain. However, the case may arise where the decision maker asks repeatedly for explanations for several statements, so that this policy would lead to intermediate alternatives having different values from one explained pair to the next. This issue may be solved considering a "static" option, where the values of the attributes $e_1, \ldots, e_{n-1}$ are restricted to a predefined list, independently of the pair $(x, y)$, so that the intermediate alternatives always reference the same values on the attributes, hopefully reducing the workload for the decision maker. One option or the other may prove more or less convincing, depending on the context (see Sect. 4).

## 3 Necessary preference relation

### 3.1 Presentation of the relation

In many decision-aiding contexts, the preference relation $\mathcal{R}$ is not explicitly specified. It is often elicited: some amount of preference information is stated by the decision maker, which is extended by an algorithmic process. We use a holistic representation of the preference information, described as a finite collection $\mathcal{P} \subset \mathbb{X}^2$ of preference statements: $(x, y) \in \mathcal{P}$ stating that $x$ is preferred to $y$.

[3] The diameter in the graph is the longest distance between two vertices in graph.

*Example 3* (ex. 1, continued) The preference information elicited from the decision maker can be expressed by three preference statements. $\mathcal{P} := \{\pi_1, \pi_2, \pi_3\}$, with

$$\begin{aligned}
\pi_1 &:= \big((4*, \text{no}, 15\text{ min}, 180\$), \quad (2*, \text{yes}, 45\text{ min}, 50\$)\big) \\
\pi_2 &:= \big((2*, \text{no}, 45\text{ min}, 50\$), \quad (2*, \text{yes}, 15\text{ min}, 180\$)\big) \\
\pi_3 &:= \big((2*, \text{yes}, 15\text{ min}, 180\$), \quad (4*, \text{no}, 45\text{ min}, 180\$)\big)
\end{aligned}$$

A model compatible with this preference information outputs a relation $\mathcal{R}_\mathcal{P} \supset \mathcal{P}$. For instance, a preference model can be built upon any value function $V \in \mathbb{R}^\mathbb{X}$ that assigns a value to each alternative, and gives precedence to the higher valued alternative.

**Definition 7** (*value models*) $\forall V \in \mathbb{R}^\mathbb{X},\ \mathcal{R}_V := \{(x, y) \in \mathbb{X}^2 : V(x) \geq V(y)\}$

Any value model is obviously transitive and satisfies Axiom 2 introduced in Sect. 2. To also satisfy Axioms 1 and 3, we require the value function to be separable.

**Definition 8** (*additive value functions*) $\forall \mathcal{P} \subset \mathbb{X} \times \mathbb{X}$,

$$\begin{aligned}
\mathbb{V} &:= \left\{ V \in \mathbb{R}^\mathbb{X} : V(x) = \sum_{i \in N} v_i(x_i) \text{ and } \forall i \in N,\ v_i \in \mathbb{R}^{\mathbb{X}_i} \text{is non-decreasing} \right\} \\
\mathbb{V}_\mathcal{P} &:= \{V \in \mathbb{V} : \forall (x, y) \in \mathcal{P},\ V(x) \geq V(y)\}
\end{aligned}$$

**Proposition 1** (properties of additive value models) *(*Krantz et al. 1971*) For any value function $V \in \mathbb{V}$, the corresponding value model $\mathcal{R}_V$ satisfies Axioms 1, 2 and 3.*

Any additive value model can thus benefit from the explanation engine described in Sect. 2, as the conceits involved may prove difficult for a broad audience, especially when conclusions are drawn from the particular shape of the marginal value functions $v_i$.

The non-empty[4] set $\mathbb{V}_\mathcal{P}$ contains all the additive value functions compatible to $\mathcal{P}$, i.e., that correctly outputs each comparison in the preference information. While many decision frameworks, such as UTA, instantiate this model by specifying a single suitable function $V \in \mathbb{V}_\mathcal{P}$, the necessary preference relation (Greco et al. 2008) circumvents the arbitrary nature of the choice of a particular value function, by demanding that every value function compatible to $\mathcal{P}$ rates alternative $x$ higher than alternative $y$ to assess that $x$ is necessarily preferred to $y$.

**Definition 9** (*necessary preference relation inferred from $\mathcal{P}$*)

$$\forall \mathcal{P} \subset \mathbb{X} \times \mathbb{X}\,, \quad \mathcal{N}_\mathcal{P} := \{(x, y) \in \mathbb{X} \times \mathbb{X} : \forall V \in \mathbb{V}_\mathcal{P}\,, \quad V(x) \geq V(y)\}$$

[4] The set $\mathbb{V}_\mathcal{P}$ is not empty, as it contains at least all uniform value functions. It may sometimes come down to contain only these, if the preference information is somewhat inconsistent. Any uniform value function $V_{\text{uniform}}$ leads to a degenerated, complete relation $\mathcal{R}_{V_{\text{uniform}}} \equiv \mathbb{X}^2$.

We believe this extra layer of abstraction added on top of the modelling of preference by additive value functions requires some supportive evidence, the more down to earth the better. Fortunately, the necessary preference relation $\mathcal{N}_\mathcal{P}$ qualifies for the explanation engine developed in Sect. 2, as it satisfies all three axioms made on the relation to be explained.

**Theorem 3** : *The binary relation $\mathcal{N}_\mathcal{P}$ satisfies Axioms 1, 2 and 3*

*Proof* By definition, $\mathcal{N}_\mathcal{P} = \bigcap_{V \in \mathbb{V}_\mathcal{P}} \mathcal{R}_V$. By Theorem 1, every binary relation $\mathcal{R}_V$ satisfies Axiom 1 and is a superset of $\mathcal{D}$, and so is their intersection. Hence, $\mathcal{N}_\mathcal{P}$ satisfies Axiom 1.

Let $(x, y)$ and $(y, z)$ be two pairs in $\mathcal{N}_\mathcal{P}$. For any value function $V \in \mathbb{V}_\mathcal{P}$, both $(x, y)$ and $(y, z)$ are in $\mathcal{R}_V$ (by definition of the necessary preference relation), and the pair $(x, z)$ is in $\mathcal{R}_V$ (by transitivity of $\mathcal{R}_V$, see Theorem 1). As $(x, z) \in \mathcal{R}_V$ for any $V \in \mathbb{V}_\mathcal{P}$, the pair $(x, z)$ is in $\mathcal{N}_\mathcal{P}$, so $\mathcal{N}_\mathcal{P}$ is transitive and satisfies Axiom 2. It is straightforward to adapt this argument to prove $\mathcal{N}_\mathcal{P}$ also satisfies the cancelation axiom. □

In the remainder of this section, the preference information $\mathcal{P}$ is considered given once and for all, and we will omit the corresponding quantifier "$\forall \mathcal{P} \subset \mathbb{X}^2$".

### 3.2 The decision problem: basic principles

The inference of the relation $\mathcal{N}_\mathcal{P}$ from the preference information $\mathcal{P}$ amounts to solving many decision problems, queries of the form "is $x$ necessarily preferred to $y$?", for every pair $(x, y) \in \mathbb{X}^2$.

This issue has already been addressed by various techniques.

- In the wake of the original article (Greco et al. 2008) introducing the relation $\mathcal{N}_\mathcal{P}$, decision over a query requires solving a linear program (LP) minimizing $V(x) - V(y)$ subject to constraints ensuring the additive value function $V$ is compatible to both the preference information $\mathcal{P}$ and the Pareto dominance $\mathcal{D}$, then concluding that $x$ is indeed preferred to $y$ if and only if min $V(x) - V(y)$ is non-negative.
- Trying to write rule-based conditions on so-called positive and negative arguments for necessary preference of $x$ over $y$, as proposed by (Spliet and Tervonen 2014).

An issue sometimes mentioned [e.g., (Spliet and Tervonen 2014)] is that necessary preference is a tall order, often resulting to a quite small set $\mathcal{N}_\mathcal{P}$, so that most pairs $(x, y) \in \mathbb{X} \times \mathbb{X}$ end up being incomparable (that is, neither $(x, y)$ nor $(y, x)$ are in $\mathcal{N}_\mathcal{P}$). It should be noted though that $\mathcal{N}_\mathcal{P}$ is far from minimal:

- The transitive closure of $\mathcal{D} \cup \mathcal{P}$ does not generally satisfy Axiom 3, so it is usually a strict subset of $\mathcal{N}_\mathcal{P}$.
- $\mathcal{N}_\mathcal{P}$ is actually not minimal under Axioms 1, 2 and 3. Indeed, the necessary preference relation also satisfies an additional axiom of multiple cancelation, which will prove to be central in our setting.

To first illustrate the intuition behind this additional axiom, let us consider the following example:

*Example 4* (example 3 continued) For any $V \in \mathbb{V}_{\mathcal{P}}$, the following inequalities stand:

– From ((4*, no, 15 min, 180 $),(2*, yes, 45 min, 50 $)) $\in \mathcal{P}$ we derive:

$$u_*(4*) + u_r(\text{no}) + u_t(15\text{ min}) + u_\$(180\$) \geq u_*(2*) + u_r(\text{yes}) + u_t(45\text{ min}) + u_\$(50\$)$$

– From ((2*, no, 45 min, 50 $),(2*, yes, 15 min, 180 $)) $\in \mathcal{P}$ we derive:

$$u_*(2*) + u_r(\text{no}) + u_t(45\text{ min}) + u_\$(50\$) \geq u_*(2*) + u_r(\text{yes}) + u_t(15\text{ min}) + u_\$(180\$)$$

– From dominance for the criterion restaurant we derive:

$$u_r(\text{yes}) \geq u_r(\text{no})$$

Adding these three inequalities, and canceling terms appearing on both sides leads to:

$$\forall V \in \mathbb{V}_{\mathcal{P}},\ u_*(4*) + u_r(\text{no}) \geq u_*(2*) + u_r(\text{yes})$$

which in turn proves, for instance, the necessary preference of (4*, no, 15 min, 50 $) over (2*, yes, 15 min, 50 $).

Formally, this property is thus called multiple cancelation in the literature (Krantz et al. 1971; Fishburn 1997).[5] It has been established [see (Joel Michell 1988)] to be logically independent from the axiom of cancelation, and if $\mathbb{X}$ is large enough, there are relations in $\mathbb{X}^2$ that satisfy Axioms 1, 2 and 3, but not double cancelation.

Regarding our explanation objective, this principle is extremely attractive: it accounts for the inference of new pairs in $\mathcal{N}_{\mathcal{P}}$ by canceling arguments throughout multiple statements, as illustrated in the previous example, a feature that none of the other techniques offers. However, one can wonder if this situation, where a statement of $\mathcal{N}_{\mathcal{P}}$ is proven by combining a subset of the previously approved statements of $\mathcal{P}$ and $\mathcal{D}$, is the rule or a lucky exception. We now address this issue by introducing a new framework for the resolution of a query.

### 3.3 A novel technique to solve the decision problem

In this section, we present a decision framework for answering the query "is alternative $x$ necessarily preferred to alternative $y$?", given a set of preference statements $\mathcal{P}$: if

[5] $m$th-order cancelation axiom: consider $m+1$ alternatives $x^{(k)}$ in $\mathbb{X}$, $k \in \{0, 1, \ldots, m\}$. Let $y^{(k)}$ in $\mathbb{X}$, $k \in \{0, 1, \ldots, m\}$ $m+1$ alternatives such that, for every criterion $i \in N$, $(y_i^{(0)}, y_i^{(1)}, \ldots, y_i^{(m)})$ is a permutation of $(x_i^{(0)}, x_i^{(1)}, \ldots, x_i^{(m)})$. Then, $[(x^{(k)}, y^{(k)}) \in \mathcal{R}, \forall k \in \{0, 1, \ldots, m-1\}] \Rightarrow (y^{(m)}, x^{(m)}) \in \mathcal{R}$.

$(x, y)$ is an unbounded pair, as defined by Definition 13, then necessary preference does not hold (Theorem 4); else, we define covectors for the pair $(x, y)$ (see Definition 12) permitting to express three characterizations of a positive query (Theorem 5): the absence of solution to a linear system of inequalities; the expression of the covector expressing the query as a linear combination with non-negative coefficients of the covectors of the preference statements and of the covectors of the dual base; a slightly modified version of this linear combination, where the coefficients sought for are non-negative integers.

The preference information references a finite set of attributes for each criterion. We call core alternatives the finite set of alternatives combining these attributes.

**Definition 10** (*core alternatives*)

$$\mathbb{D}_i := \bigcup_{(x,y)\in\mathcal{P}} \{x_i, y_i\} := \{d_{i,1} \prec_i \cdots \prec_i d_{i,|\mathbb{D}_i|}\}\ ;\ \mathbb{D} := \prod_{i\in N} \mathbb{D}_i$$

*Example 5* (Example 3 continued)

$$\begin{array}{ll}
\mathbb{D}_* = \{a \prec_* A\} & \text{with } a := 2* \text{ and } A := 4* \\
\mathbb{D}_r = \{b \prec_r B\} & \text{with } b := \text{no and } B := \text{yes} \\
\mathbb{D}_t = \{c \prec_t C\} & \text{with } c := 45 \text{ min and } C := 15 \text{ min} \\
\mathbb{D}_\$ = \{d \prec_\$ D\} & \text{with } d := 180\$ \text{ and } D := 50\$
\end{array}$$

Consequently, the preference statements are: $\pi_1 = (AbCd, aBcD)$; $\pi_2 = (abcD, aBCd)$; $\pi_3 = (aBCd, Abcd)$ and there are 16 core alternatives: $\mathbb{D} = \{ABCD,\ ABCd,\ ABcD,\ ABcd,\ AbCD,\ AbCd,\ AbcD,\ Abcd,\ aBCD,\ aBCd,\ aBcD, aBcd, abCD, abCd, abcD, abcd\}$

In the remainder of this section, we often use interval semantics, where an interval designates all the attributes simultaneously higher than the lower bound and lower than the upper bound:

$$\forall i \in N, \forall a_i, b_i \in \mathbb{X}_i, [a_i, b_i] := \{z \in \mathbb{X}_i : a_i \precsim_i z \precsim_i b_i\}$$

In particular, core intervals $[d_{i,k}, d_{i,k+1}]$ play a key role. They are indexed by pairs $(i, k)$ conveniently grouped in an index set $\mathbb{I}$:

**Definition 11** (*indexes of core intervals*) The set $\mathbb{I} := \bigcup_{i\in N} \{(i, k)\ :\ k \in \mathbb{N} \text{ and } 1 \leq k \leq |\mathbb{D}_i| - 1\}$ contains the pairs $(i, k)$ indexing the core intervals $[d_{i,k}, d_{i,k+1}]$ and, consequently, the differences in marginal value between consecutive core levels $\Delta v_{(i,k)} := v_i(d_{i,k+1}) - v_i(d_{i,k})$.

We denote $\times$ the matrix multiplication, so that, for a (line) covector $v^\star$ and a (column) vector $w$ both taken in $\mathbb{R}^{\mathbb{I}}$, $v^\star \times w = \sum_{(i,k)\in\mathbb{I}} v^\star_{(i,k)} w_{(i,k)}$.

This collection of intervals $[d_{i,k}, d_{i,k+1}]$, $(i, k) \in \mathbb{I}$ is partitioned between pros, cons and neutral arguments of a pair of alternatives $(x, y)$.

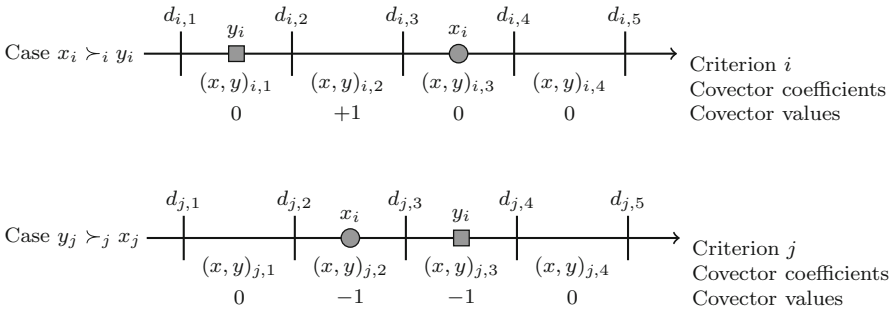



**Fig. 1** Covectors illustrated

**Definition 12** (*covector associated to a pair of alternatives*) $\forall (x, y) \in \mathbb{X}^2$, the covector $(x, y)^\star$ is a linear form operating on $\mathbb{R}^{\mathbb{I}}$. Its coefficient associated with criterion $i \in N$ and interval $[d_{i,k}, d_{i,k+1}] \subset \mathbb{X}_i$ is given by:

$$(x, y)^\star_{(i,k)} := \begin{cases} +1, & \text{if } [d_{i,k}, d_{i,k+1}] \subset [y_i, x_i] \\ -1, & \text{if } [d_{i,k}, d_{i,k+1}] \cap ]x_i, y_i[ \neq \emptyset \\ 0, & \text{else} \end{cases}$$

The canonical dual base is denoted $(\delta^\star_{(i,k)})_{(i,k)\in\mathbb{I}}$, where the covector $\delta^\star_{(i,k)}$ has all coefficients equal to zero, except for the coefficient associated to the interval indexed by $(i, k)$, which is equal to +1, so that $\delta^\star_{(i,k)} \times \Delta v = \Delta v_{(i,k)}$.

For alternatives $(x, y)$ in the core $\mathbb{D}^2$, for each criterion $i \in N$, intervals $[d_{i,k}, d_{i,k+1}]$ between $x_i$ and $y_i$ are taken into account, positively if $x_i \succ_i y_i$, and negatively if $y_i \succ_i x_i$. For alternatives $(x, y)$ outside the core, for some criterion $i \in N$, some attribute $x_i$, or $y_i$, or both, falls strictly between the values of $\mathbb{D}_i$, "breaking" some interval $[d_{i,k}, d_{i,k+1}]$. Because of the cautious nature of the relation $\mathcal{N}_\mathcal{P}$, "broken" intervals are rounded down: those that would support the preference of $x$ over $y$ is not taken into account and considered neutral, with coefficient 0, while "broken" intervals that would go against this preference are totally taken into account with coefficient −1. Figure 1 illustrates these notions.

*Example 6* As the preference information only refers two attributes level by criteria, there is exactly one core interval by criterion: from 2* to 4*, from no to yes, from 45 min to 15 min and from 180 $ to 50 $. Definition 12 is straightforward for core alternatives:

$$\pi_1 = (AbCd, aBcD); \qquad \pi_1^\star = (1, -1, 1, -1);$$
$$\pi_2 = (abcD, aBCd); \qquad \pi_2^\star = (0, -1, -1, 1);$$
$$\pi_3 = (aBCd, Abcd); \qquad \pi_3^\star = (-1, 1, 1, 0).$$

Alternatives outside the core demand a bit more effort: $(h_1,h_3)^\star = (0, 1, 0, -1)$, as:

– $h_1\,(5*)$ is more comfortable than $h_3\,(3*)$, but not strongly enough to warrant for a positive argument;
– $h_1$ is strongly better than $h_3$ on criterion restaurant;
– $h_1$ is weakly nearer than $h_3$;
– $h_3$ is weakly cheaper than $h_1$, and this counts as a fully negative argument.

We also find $(h_1,h_4)^\star = (1, 1, 1, -1)$, $(h_3, h_2)^\star = (-1, -1, 1, 1)$.

There is a class $\mathcal{U}_\mathcal{P}$ of unbounded queries $(x, y)$ for which covectors fail to account for arguments that are both negative (because $y_i \succ_i x_i$) and infinitely strong (because $x_i \prec_i \min \mathbb{D}_i$ or $y_i \succ_i \max \mathbb{D}_i$). In such a case, $x$ is clearly not necessarily preferred to $y$.

**Definition 13** (*unbounded pairs* $\mathcal{U}_\mathcal{P}$)

$$\forall x, y \in \mathbb{X}, (x, y) \in \mathcal{U}_\mathcal{P} \iff \exists i \in N : x_i < y_i \text{ and } [x_i, y_i[ \not\subseteq [\min \mathbb{D}_i, \max \mathbb{D}_i]$$

**Theorem 4** *:* $\mathcal{U}_\mathcal{P} \cap \mathcal{N}_\mathcal{P} = \emptyset$

*Proof* : see Appendix 1. □

*Example 7* (Example 5 continued) We see that $h_3$ is not necessarily preferred to $h_1$, as $(h_1)_* = 5*$ is better than both $(h_3)_* = 3*$ and the most comfortable hotel referenced by $\mathcal{P}$ ($\max \mathbb{D}_* = 4*$). No amount of positive arguments in favor of $h_3$ make up for such a high attribute within the cautious context of necessary preference.

Neither is $h_4$ preferred to $h_2$, as $(h_4)_t = 60$ min is worse than both $(h_2)_t = 35$ min and the farthest hotel referenced by $\mathcal{P}$ ($\min \mathbb{D}_t = 45$ min). No amount of arguments in favor of $h_4$ make up for such a low attribute.

For pairs outside the class $\mathcal{U}_\mathcal{P}$, we give three characterizations of the necessary preference of $x$ over $y$ using covectors.

**Theorem 5** (characterization of necessary preference using covectors) $\forall (x, y) \in \mathbb{X}^2 \backslash \mathcal{U}_\mathcal{P}$, *the following propositions are equivalent:*

1. *Necessary preference*

$$(x, y) \in \mathcal{N}_\mathcal{P}$$

2. *Linear feasibility problem*

$$\begin{cases} & (x, y)^\star \times \Delta v < 0 \\ \forall \pi \in \mathcal{P}, & \pi^\star \times \Delta v \geq 0 \\ \forall (i, k) \in \mathbb{I}, & \delta^\star_{(i,k)} \times \Delta v \geq 0 \end{cases} \quad \textit{has no solution} \Delta v \in \mathbb{R}^\mathbb{I}$$

3. *Combination of statements* $\exists \lambda \in [0, +\infty[^\mathcal{P}$, $\mu \in [0, +\infty[^\mathbb{I}$:

$$(x, y)^\star = \sum_{\pi \in \mathcal{P}} \lambda_\pi \pi^\star + \sum_{(i,k) \in \mathbb{I}} \mu_{(i,k)} \delta^\star_{(i,k)}$$

4. *Integral combination of statements* $\exists n \in \mathbb{N}^\star,\ \ell \in \mathbb{N}^{\mathcal{P}},\ m \in \mathbb{N}^{\mathbb{I}}$ :

$$n(x, y)^\star = \sum_{\pi \in \mathcal{P}} \ell_\pi \pi^\star + \sum_{(i,k) \in \mathbb{I}} m_{(i,k)} \delta^\star_{(i,k)}$$

*Proof:* see Appendix 1. □

Point 3 proves the situation depicted in example 4 is not a corner case, but a general one: every necessary preference statement results from basic arithmetic operations (namely multiplication by a positive number, addition and cancelation of terms) over fundamental inequalities expressing either the preference information, or dominance. The exploration of the different combinations of this grammar, to assess if an alternative is necessarily preferred to another, is a linear programming problem. Noticeably, when the pair $(x, y) \notin \mathcal{U}_{\mathcal{P}}$ changes, the constraints remain the same, and can be computed once and for all: two different queries differ only by their objective covector.

*Example 8* We use the fourth point of Theorem 5 to establish:

- $h_1$ is necessarily preferred to $h_4$, as $(h_1,h_4)^\star = \pi_1^\star + 2\delta^\star_{(2,1)}$;
- $h_3$ is necessarily preferred to $h_2$, as $(h_3,h_2)^\star = \pi_1^\star + 2\pi_2^\star + 2\pi_3^\star$;
- $h_1$ is not necessarily preferred to $h_3$, as there is no suitable linear combination.

Consequently, alternatives $h_1$ and $h_3$ are incomparable, as neither is preferred to the other.

We represent graphically the skeleton of the relation $\mathcal{N}_{\mathcal{P}} \cap \mathbb{D}^2$ (additional arcs resulting of the transitive closure of this skeleton are omitted in Fig. 2). For illustrative purpose, we show some example of the covectors associated to pairs involved in Example 4.

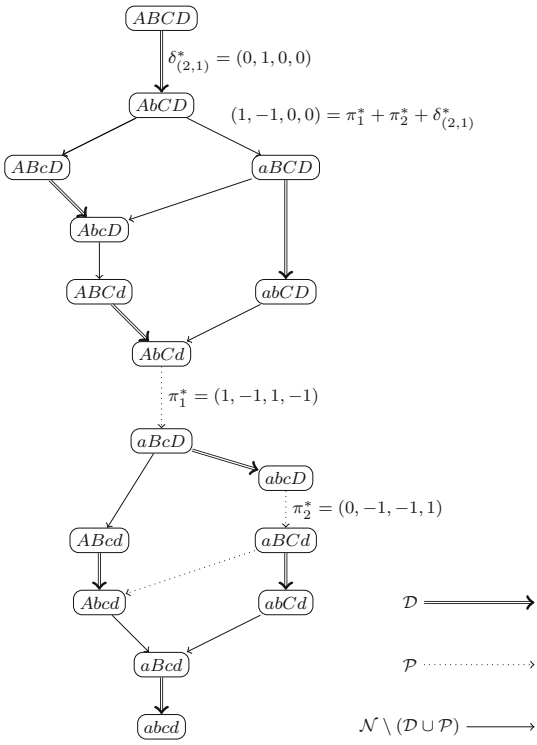
The integral version (point 4) is obviously less useful than the continuous one (point 3) for the actual decision of a query, as it implies the solving of an ILP, rather than an LP. It is nevertheless an important property that we shall leverage in the next section to derive insights into the problem of explaining a necessary preference relation statement $(x, y) \in \mathcal{N}_{\mathcal{P}}$ by low-order preference swaps, as introduced in Sect. 2.

## 4 Explanation of the necessary relation with binary reference scales

In this section, we bring together the main notions discussed in Sects. 2 and 3, connecting the explanation engine producing sequences of low-order preference swaps to the necessary preference relation. This coupling is made possible by Theorem 3, which ensures the necessary preference relation $\mathcal{N}_{\mathcal{P}}$ satisfies the requirement for the relation $\mathcal{R}$ explained by the explanation engine (i.e., we instantiate $\mathcal{R}$ as $\mathcal{N}_{\mathcal{P}}$). This coupling is also highly desirable, as the necessary preference relation makes minimal assumptions, handling a collection of compatible utility functions, virtually impossible to exhibit to the user.

To address some of the issues listed in Sect. 2.3, we make two additional assumptions. The first one concerns the number of distinct values referenced by the preference

**Fig. 2** Necessary preference relations




information $\mathcal{P}$ which serves as a basis for the inference of the necessary preference relation $\mathcal{N}_{\mathcal{P}}$, and is discussed in Sect. 4.1. The second one instantiates the cap on the order of the swaps linking the alternatives in the explaining sequence, and is discussed in Sect. 4.2. Under these assumptions, explanations have a core, term-by-term structure we expose in Sect. 4.3, followed by some resulting properties.

### 4.1 Binary reference scales

Binary reference scales are encountered when the preferences $\mathcal{P}$ expressed by the decision maker only reference two levels on each attribute.

**Definition 14** (*Binary reference scales*)

$$\forall i \in N,\ \mathbb{B}_i = \{\top_i \succsim_i \bot_i\},\ \mathbb{B} := \prod_{i \in N} \mathbb{B}_i$$

Besides luck, such a tight reference set is the consequence of one of these two situations :

– *Attributes are themselves binary*: present or absent features, passed or failed checks, etc. In addition, such binary attributes may result from any model relying

on subset comparisons. While they fall outside the scope of this article, we believe the explanation engine discussed here can address problems not necessarily resulting from an additive utility decision model (for instance, robust weighted majority decision models rely on subset comparisons between coalition of criteria, as do pan-balance comparisons encountered in extensive measurement problems).

– *When expressing preference statements, the decision maker is deliberately restricted to comparing between prototypical alternatives specifically chosen in* $\prod_{i \in N} \{\bot_i, \top_i\}$. This process is supposed to help the decision maker focusing on the main aspects of the preference problems, by limiting the number of changing parts between alternatives, and by referring to carefully chosen reference values, serving as anchors. This technique is used in the field of experimental design (yielding the one-factor-at-a-time or the factorial experiments methods), as well as in multicriteria decision aiding. For instance, the MACBETH method (Bana e Costa and Vansnick 1995; Bana e Costa et al. 2008) is based on binary alternatives: to assess hidden technical parameters (the weights of the various criteria), the decision maker is asked to express preference between prototypical alternatives, traditionally referencing a neutral level $\bot_i$ (for technological products, representing the attribute of a mid-range, available product), and a high-level $\top_i$ (representing the attribute of a luxury product, or a hypothetical performance demanding a technological breakthrough).

This tight set of core alternatives (see Definition 10) has bearing on the necessary preference relation. It increases the likelihood of single and multiple cancelation occurrence, thus enriching relation $\mathcal{N}_\mathcal{P}$ between core alternatives in $\mathbb{B}^2$. It aligns the individual technical arguments of the decision problem "is alternative $x$ necessarily preferred to alternative $y$?", the intervals between consecutive attributes of the core (see Definition 12), with the criteria themselves. This alignment has, in turn, consequences concerning explanations, as the criteria involved in a preference statement (precisely, their number) determine its order, which is a proxy for its cognitive complexity. Technically, with binary reference scales, the order of a swap $(x, y) \in \mathcal{N}_\mathcal{P} \backslash \mathcal{D}$ is exactly the number of non-zero coefficients of its covector $(x, y)^\star$.

### 4.2 Swaps of order two

While the assumption of binary scales is a favorable case for the joining of the explanation engine based on sequences of preference swaps and the necessary preference relation, we make the choice concerning the bound placed on the order of the swaps eligible for participating in the explanation. We restrict the explanation to swaps of order at most two, that is:

– either a dominance relation or
– a trade-off between exactly two criteria.

The concept of swaps is known in engineering. For instance, the Architecture Trade-off Analysis Method (ATAM) is used to assess software architectures according to "quality attribute goals" (Kazman et al. 2000). A trade-off point is an architecture

parameter affecting at least two quality attributes in different directions. For example, increasing the speed of the communication channel improves throughput in the system but reduces its reliability. Thus, the speed of that channel is a trade-off point. The concept of trade-off point in ATAM makes explicit the interdependencies between attributes. Even though trade-offs can be defined for any number of attributes, the examples of trade-offs that are provided by experts are almost always given on pairs of attributes. This is the case of the example provided above. It is thus a very reasonable assumption to restrict ourselves to swaps of order two.

### 4.3 Structure of an explanation

Our restriction to binary scales allows us to introduce a simpler notation, in terms of positive or negative arguments:

**Definition 15** (*pros and cons of a necessary preference statement*) If $\mathcal{P} \subset \mathbb{B}^2$, $\forall (x, y) \in \mathcal{N}_{\mathcal{P}}$,

$$(x, y)^+ := \{i \in N : (x, y)^{\star}_{(i,1)} = +1\} = \{i \in N : y_i \succsim_i \bot_i \prec_i \top_i \succsim_i x_i\}$$
$$(x, y)^- := \{i \in N : (x, y)^{\star}_{(i,1)} = -1\} = \{i \in N : \bot_i \succsim_i x_i \prec_i y_i \succsim_i \top_i\}$$

Assuming binary reference scales, the relation $\Delta_2 \subset \mathbb{X}^2$ between alternatives induces a relation between criteria $\tilde{\Delta}_2 \subset N^2$.

**Definition 16** (*criteria swaps*) If $\mathcal{P} \subset \mathbb{B}^2$,

$$\tilde{\Delta}_2 := \{(i, i') \in N^2 : ((\top_i, \bot_{i'}), (\bot_i, \top_{i'}))_{cp} \subset \Delta_2\}$$

Note the use of the ceteris paribus syntax here (see Definition 1). We emphasize though that this relation is not suitable to being presented directly as an explanation. The reason is that it could be interpreted, sometimes erroneously, as giving more importance to criterion $i$ than to criterion $i'$. While this interpretation seems practically correct in an elicitation framework similar to the MACBETH procedure (see Sect. 4.1), it is highly dependent of the values of $\mathbb{D}_i \times \mathbb{D}_{i'}$ referred by the preference information $\mathcal{P}$. To remain on the safe side, the relation $\tilde{\Delta}_2$ should only appear as a technical tool to produce an explanation.

*Example 9* The necessary preference relation deduced from the preference information given in Example 3 contains the following compact criteria swap statements, represented in Fig. 3.

$$\tilde{\Delta}_2 = \{(*, r), (t, r), (\$, *), (\$, r), (\$, t)\}.$$

For instance, the compact criteria swap statement $(\$, r)$, represented by the arrow from \$ to $r$, means that an alternative ranking higher than $D$ on attribute \$ and low on attribute $r$ is necessarily preferred to one ranking low on \$ (between $d$ and $D$) and high on $r$, attributes $*$ and $t$ being equal: $\big((\_\__*, b, \_\__t, D), (\_\__*, B, \_\__t, d)\big)_{cp} = \{\big((x_*, b, x_t, D), (x_*, B, x_t, d)\big), \forall x_* \in \mathbb{X}_*, \forall x_t \in \mathbb{X}_t\} \subset \mathcal{N}_{\mathcal{P}}$.

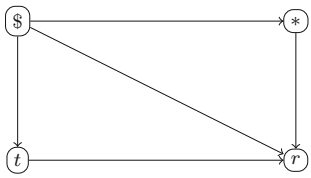



**Fig. 3** Binary relation between criteria

The following theorem reveals the core structure every explanation is built upon.

**Theorem 6** (Term-by-term explanation) *If* $\mathcal{P} \subset \mathbb{B}^2$, $\forall \sigma \in \mathcal{N}_\mathcal{P}$, *the following propositions are equivalent:*

1. $\sigma \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$
2. $\exists a \in \mathbb{N}^\star, \gamma_1, \ldots, \gamma_q \in \Delta_2, \ell_1, \ldots, \ell_q \in \mathbb{N}, m_1, \ldots, m_n \in \mathbb{N}$ :

$$a\sigma^\star = \sum_k \ell_k \gamma_k^\star + \sum_k m_k \delta^\star_{(k,1)}$$

3. *There is a matching of cardinality* $|\sigma^-|$ *in the graph of* $\tilde{\Delta}_2 \cap (\sigma^+ \times \sigma^-)$.
4. *There is an injection* $\phi : \sigma^- \to \sigma^+$ *such that* $\forall k \in \sigma^-, (\phi(k), k) \in \tilde{\Delta}_2$.

*Proof* See Appendix 1. □

In a nutshell, an explanation is a sequence where, at each step, a positive argument is used up to cancel an inferior negative argument and, eventually, every negative argument has been canceled. We highlight three consequences of this theorem:

– *If preferences only refer to swaps of order 2, then every necessary preference can be explained by swaps of order 2.* This is a potent existence result for explanations, and it provides a complete description of the necessary preference relation under the assumption of the decision maker expressing preferences between alternatives differing along two criteria only.

**Corollary 1** (case of 2-order preference statements) *If* $\mathcal{P} \subset \mathbb{B}^2$, *and* $\forall \pi \in \mathcal{P}, |N_\pi^{\neq}| = 2$ *then* $\mathcal{E}_2(\mathcal{N}_\mathcal{P}) = \mathcal{N}_\mathcal{P}$. *i.e., for any statement* $(x, y) \in \mathcal{N}_\mathcal{P}$, *there exists an explanation of it in* $\mathcal{E}_2(\mathcal{N}_\mathcal{P})$

*Proof* By Theorem 1, $\mathcal{E}_2(\mathcal{N}_\mathcal{P}) \subset \mathcal{N}_\mathcal{P}$. Reciprocally, if $(x, y) \in \mathcal{N}_\mathcal{P}$, the implication 1. $\Rightarrow$ 4. of Theorem 5 ensures the existence of a linear combination with integral, non-negative coefficients $n(x, y)^\star = \sum_{\pi \in \mathcal{P}} \ell_\pi \pi^\star + \sum_{(i,k) \in \mathbb{I}} m_{(i,k)} \delta^\star_{(i,k)}$. The assumption that $\forall \pi \in \mathcal{P}, |N_\pi^{\neq}| = 2$ entails $\mathcal{P} \subset \Delta_2$, so this linear combination satisfies proposition 2 of Theorem 6, thus $(x, y) \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$ by proposition 1.

– *Explanations can be kept short.* The next corollary proves that the size of the explanation is at most "half the number of criteria, rounded down, plus one", which appears manageable for the recipient of explanation.

**Corollary 2** (short explanations) *If $\mathcal{P} \subset \mathbb{B}^2$, for any statement $(x, y) \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$, there exists an explanation with a length at most $\lfloor \frac{|N|}{2} \rfloor + 1$, where $\lfloor m \rfloor$ denotes the integer part of $m$.*

The bound $\lfloor \frac{|N|}{2} \rfloor + 1$ basically comes from the fact that $|(x, y)^-| \leq \lfloor \frac{|N|}{2} \rfloor$, which follows directly from item 4 of Theorem 6. The main asset of this theorem is that it is constructive. The explanation sequence will be provided in the next section.

**Algorithm 1**: FINDEXPLANATION

**Data**: a statement $\sigma = (x, y)$ to be explained, a set of preference statements $\mathcal{P}$.
**Result**: a matching of each negative argument by a stronger positive one.

Compute $\sigma^+, \sigma^-$
**if** $|\sigma^+| < |\sigma^-|$ **then**
    return None
**if** $\sigma \notin \mathcal{N}_\mathcal{P}$ **then**
    return None
Build the graph of $\tilde{\Delta_2} \cap (\sigma^+ \times \sigma^-)$ :
Initialize $\mathcal{G}$ as a graph with nodes $\sigma^+ \cup \sigma^-$ and no edge.
**for** $i \in \sigma^+$ **do**
    **for** $j \in \sigma^-$ **do**
        **if** *the LP with* $|N| + |\mathcal{P}|$ *inequality constraints,* $|N|$ *equality constraints and* $|N| + |\mathcal{P}|$ *variables*
            $\forall p \in \mathcal{P}, \ell_p \geq 0$
            $\forall k \in N, m_k \geq 0$
            $\forall k \in N, \sum_{p \in \mathcal{P}} \ell_p \, p_k^\star + m_k = 1$ *if* $k = i$, -1 *if* $k = j$, 0 *else.*
        *is feasible* **then**
            add edge $(i, j)$ to $\mathcal{G}$
Find a matching $\phi$ of maximum cardinality $C$ in bipartite graph $\mathcal{G}$.
**if** $C < |\sigma^-|$ **then**
    return None
return $\phi$

– *Building an explanation, or ensuring there is none, is handled by an efficient algorithm (see Algorithm 1).* A quick inspection of the complexity reveals that in the first part of the algorithm, there are at most $\mathcal{O}(n^2)$ calls to a linear program (with $n$ the number of criteria). This is followed by the resolution of a matching problem, which runs in its simpler version in $\mathcal{O}(n^3)$. Note that in theory, the number of constraints and variables of the LP may be exponential in $n$, because of the number of preference statements can be. In practice, this is of course highly unrealistic as it is too demanding for the decision maker. Finally, for a polynomially bounded number of preference queries, the algorithm is efficient.

*Example 10* (Ex 8. ctd.) The pair ($h_1$,$h_4$) is in $\mathcal{N}_\mathcal{P}$. Its negative arguments are $(h_1,h_4)^- = \{*, r, \$\}$ and its positive arguments $(h_1, h_4)^+ = \{\$\}$. As there are more

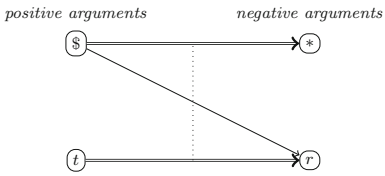



**Fig. 4** Matching returned by Algorithm 1 with data of Example 3

negative than positive arguments, the necessary preference of $h_1$ over $h_4$ cannot be explained by a sequence of preference swaps of order 1 or 2.

The pair ($h_3$,$h_2$) is also in $\mathcal{N}_{\mathcal{P}}$. $(h_3,h_2)^- = \{*, r\}$ and $(h_3, h_2)^+ = \{t, \$\}$.

Figure 4 shows the bipartite graph of the relation $\Delta_2$ restricted to pairs of positive–negative arguments of the statement ($h_3$,$h_2$). The double arrows highlight a matching of cardinality 2, covering the negative arguments, as returned by Algorithm 1: $\{(\$, *), (t, r)\} \subset \tilde{\Delta}_2$. Therefore, the statement ($h_3$,$h_2$) can be explained by a sequence of preference swaps of order 2 and dominance relations.

To explain that $h_3$ = (3∗, no, 15 min, 60 $) is necessarily preferred to $h_2$=(4∗, yes, 45 min, 180 $), several explanations can be considered:

- $h_3\ \Delta_2$ (4⋆, no, 15 min, 180$) $\Delta_2\ h_2$
- $h_3\ \Delta_2$ (3⋆, yes, 45 min, 180$) $\Delta_2\ h_2$
- $h_3\ \mathcal{D}\ (a, b, C, D)\ \Delta_2\ (A, b, C, d)\ \Delta_2\ (A, B, c, d)\ \mathcal{D}\ h_2$
- $h_3\ \mathcal{D}\ (a, b, C, D)\ \Delta_2\ (a, B, c, D)\ \Delta_2\ (A, B, c, d)\ \mathcal{D}\ h_2$

The first two explanations, which involve directly the attributes of the compared alternatives are shorter than the last two, which refer to core alternatives. It is interesting to observe how the two preference swaps (giving up cost for comfort and lengthening commute time to obtain access to a restaurant) can be presented in any order (since they do not have any criteria in common).

## 5 Related works and extensions

Generating explanations to justify recommendation is a key challenge to decision-aiding systems. While we witness the emergence of highly sophisticated methods to elicit preferences and compute recommended alternatives, the question of explanation is often neglected. We believe this may hinder the development of such systems. As a matter of fact, real decision makers often prefer the use of a very basic model if its outcomes are transparent, rather that elaborate models that look as a black box for them.

Explanations can either be conceived as being complete or incomplete. While we clearly follow the first option in this paper, some papers assume that explanations can be effective without being formally sufficient to support the statement [this may indeed be absolutely appropriate in settings with low stakes, for instance for most recommender systems (Herlocker et al. 2000; Friedrich and Zanker 2011)]. In that case, explanations can be seen as positive evidence supporting the conclusion. In a multicriteria setting close to ours, the approaches of Klein (1994), Carenini and Moore (2006), Labreuche (2011), Nunes et al. (2014) fall into that category: they build upon

patterns (or anchors) that are used to present some sufficiently convincing evidence to the user. The idea in that case is for instance to identify which set of criteria should be highlighted in the explanation.

A second distinctive feature of explanation is whether it is data based or process based (Herlocker et al. 2000). The vast majority of approaches dealing with this concept and emanating from A.I. adopts a data-based approach: this is true in particular of the literature investigating explanations in diagnosis systems [see for instance (Eiter and Gottlob 1995)] or constraints [where the aim is to return a minimal subset of mutually incoherent constraints in case of infeasibility (Ulrich Junker 2004)]. Here the objective is to find a minimal subset of the data provided by the user which implies the conclusion. This assumes that the explanation is to be presented to a user who has no problem in understanding the process by which these data then lead to a given conclusion. This is not the case in our setting (as inference from the necessary preference relation is a difficult notion to handle), and our approach follows instead a process-based approach. We would like to point out though that these two approaches are by no means contradictory: in particular, it would be certainly relevant to incorporate some data-based consideration when building sequences of preference swaps, as was already alluded to in the paper. Giving priority to the statements presented by the user, or defining notions of proximity so that sequences of explanations can be evaluated with respect to their distance to the initial data is certainly a promising perspective.

In our setting the initial preference information is provided as comparisons between alternatives. Other form of input may justify the use of other decision models (and consequently, of explanation techniques). For instance, complete explanations have been investigated for (weighted) majority-based decision models, when ordinal rankings on alternatives are given as input (Labreuche et al. 2011, 2012). In that case, explanations also amounts to exhibit coalitions of criteria.

Each explanatory step produced by our approach is typically performed by focusing on trade-offs on a subset of criteria, assuming the other ones remain unchanged. This ceteris paribus principle, which lies at the heart of the initial even-swap technique, has also been exploited for its ability to compactly represent qualitative conditional preferences (Boutilier et al. 2004). This language was later extended to account for possible trade-offs among criteria (Brafman et al. 2006), and (Nic Wilson 2011) proposed an even more expressive language (allowing to capture also stronger semantics). The resulting statements are similar in spirit to the criteria swaps that we use in this paper as technical constructs. Interestingly, "flipping" or swapping sequences appear as proof-theoretical counterpart for the semantics of these logical theories. While such compact statements are certainly useful for users to express preferences, it is not clear whether they should be used per se in producing explanations, because they may be inappropriately interpreted, as discussed in Sect. 4.3. Investigating their relevance in our setting is nevertheless an interesting future work.

We conclude by mentioning some further perspectives of this work.

– There remain theoretical questions to be studied. We have investigated two extreme cases: in the first one, no assumption is made on the preference information (yielding a negative result in terms of the length of the explanation), while in the second one we assume a binary reference scale (and can guarantee the existence of a short

explanation). A natural but challenging question is whether the complexity of the reference scale can be more generally linked to the size of the explanations.
- We have provided an algorithm for the binary case only. It would be of practical interest to design and implement an algorithm finding the simplest (e.g., shortest) explanation in the general case.
- While we discuss good theoretical properties of explanations, an empirical validation remains to be conducted on other aspects mentioned (the sequencing of swaps, the choice of values, for instance). What makes the exercise difficult though is that this may highly depend on the context of use: a DM who needs to justify an important decision before a committee may not have the same expectations as a DM taking a decision for herself. Other issues are likely to emerge too: in particular, as we saw in Example 10, the same preference swaps can (sometimes) be presented in different orders. Are there good heuristics to select a given ordering?
- The framework may be smoothly extended to cater for more general situations. For instance, the nature of the preferential information may be different. The DM may use a more expressive language, and give some statements on the intensity of their preferences. A first step in that direction is to assume a quaternary relation, of the form "$o_1$ is more intensely preferred to $o_2$ than $o_3$ is preferred to $o_4$". While this would constitute a first step towards dealing with intensities, we are confident that this may still be handled within the framework described here.
- As a final suggestion on a possible extension of this framework, we note that this work makes the assumption that elicitation and explanation are dealt with separately. A certainly promising perspective is to extend the framework so that explanation and elicitation are actually intertwined. By putting forward an explanation, the system shows some evidence which can in turn trigger some reaction from the DM.

## Proofs

### Proof of theorem 2

For the sketch of the proof, we construct, for every $p$, a preference between $x = (0, 0, 0)$ and $y = (2p, -p, -p)$. Starting from alternative $(0, 0, 0)$, we begin with a preference swap between attributes 1 and 2 (adding value 1 on the first attribute, and subtracting 1 on the second one). Then we perform a preference swap between attributes 1 and 3 (adding value 1 on the first attribute, and subtracting 1 on the third one). We proceed then again by a preference swap between attributes 1 and 2, and so on (the sequence is depicted in Fig. 5).

*Proof* (Theorem 2) The proof is based on an instantiation of $\mathcal{R}$ with the necessary preference relation. This latter is inferred from information $\mathcal{P}$, and is denoted by $\mathcal{N}_{\mathcal{P}}$. Let $n = 3$, $p \in \mathbb{N}^*$. Assume that $\mathbb{X}_1 \supseteq \{0, 1, 2, \ldots, 2p\}$, $\mathbb{X}_2 \supseteq \{-p, -p + 1, \ldots, -1, 0\}$ and $\mathbb{X}_3 \supseteq \{-p, -p+1, \ldots, -1, 0\}$. Consider the following preference information $\mathcal{P}$:

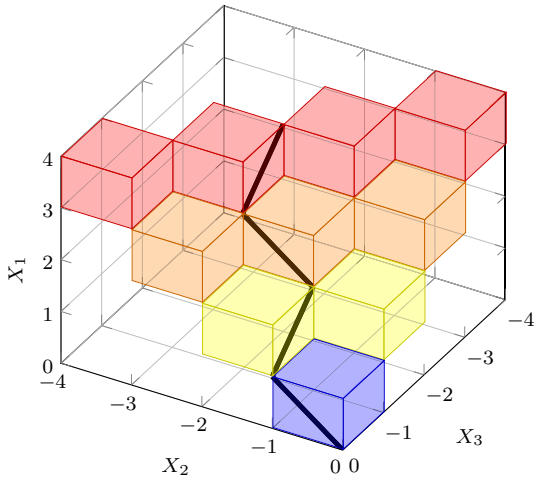



**Fig. 5** Description of the sequence

$$\forall j \in \{0, \ldots, p-1\} \quad (((2j)_1, (-j)_2), ((2j+1)_1, (-j-1)_2))_{cp} \subset \mathcal{P} \tag{1}$$

$$\forall j \in \{0, \ldots, p-1\} \quad (((2j+1)_1, (-j)_3), ((2j+2)_1, (-j-1)_3))_{cp} \subset \mathcal{P} \tag{2}$$

where (1) [resp. (2)] correspond to a ceteris paribus pair on attributes $\{1, 2\}$ (resp. $\{1, 3\}$). Hence, $\mathbb{D}_1 = \{0, 1, 2, \ldots, 2p\}$, $\mathbb{D}_2 = \{-p, -p+1, \ldots, -1, 0\}$ and $\mathbb{D}_3 = \{-p, -p+1, \ldots, -1, 0\}$.

We set $x = (0, 0, 0)$ and $y = (2p, -p, -p)$. With this $\mathcal{P}$, we clearly obtain the sequence

$$\begin{aligned}
&(x\,,\ (1, -1, 0)) \in \mathcal{P} \qquad \text{(by (1))}\\
&((1, -1, 0)\,,\ (2, -1, -1)) \in \mathcal{P}\,,\ \ldots \qquad \text{(by (2))}\\
&((2p-2, -(p-1), -(p-1))\,,\ (2p-1, -p, -(p-1))) \in \mathcal{P} \qquad \text{(by (1))}\\
&((2p-1, -p, -(p-1))\,,\ (2p, -p, -p)) \in \mathcal{P} \qquad \text{(by (2))}
\end{aligned}$$

so that $(x, y) \in \mathcal{R}$. This sequence is of length $2p$.

There remains to prove that this is the shortest explanation.

To this end, we first need to determine the form of $\Delta_2$. By Theorem 4, the necessary preference relation cannot hold outside the interval between the minimal and maximal elements of $\mathbb{D}$. Moreover, according to Theorem 5, the necessary preference relation between two alternatives $z, z'$ holds iff a linear problem involving the covector of $(z, z')$ is feasible. From these results, checking whether $(z, z') \in \mathcal{N}_{\mathcal{P}}$ is equivalent to checking boundness on $z$ and $z'$, and also checking whether $(t, t') \in \mathcal{N}_{\mathcal{P}}$ where $t, t' \in \mathbb{D}$ are appropriately chosen from $z$ and $z'$. Therefore, we need only to consider the elements in $\Delta_2$ that belong to $\mathbb{D}_1 \times \mathbb{D}_2 \times \mathbb{D}_3$ (The other ones can be deduced by Pareto dominance). The preference information (1) and (2) is very specific. In

particular, any value $k \in \mathbb{D}_1$ appears only in two examples—one in which $k$ appears in the left-hand side [in (1)] and the other one where $k$ appears in the right-hand side [in (2)]. Moreover, we notice that, in (1) and (2), the value on the first attribute is always increasing from the left-hand side to the right-hand side, and the value of the second and the third attributes is decreasing from the left-hand side to the right-hand side. Hence, the elements of $\Delta_2$ cannot be obtained by a combination of two or more preference information. They are obtained only from one preference information [(1), (2)] and Pareto dominance $\mathcal{D}$. More precisely, $\Delta_2$ is composed of the following pairs

$$\Big((i, j, k)\,,\ (i', j', k')\Big)$$

where either there exists $l$ such that $i = 2l$, $j = 2l + 1$, $j \geq -l > -l - 1 \geq j'$ and $k = k'$, or there exists $l$ such that $i = 2l + 1$, $j = 2l + 2$, $j = j'$ and $k \geq -l > -l - 1 \geq k'$. From this, one can readily see that the explanation of the preference of $x$ over $y$ described earlier is the shortest one. □

**Proof of Theorem 4 and Theorem 5**

*Proof of (1) $\iff$ (2)*

The belonging of a pair of alternatives to the necessary preference relation can be expressed as a mathematical program. We have to prove that when the pair is not unbounded, its constrains and objective function are linear and can be expressed using the proposed, fixed-length covectors.

Pairs of core alternatives, and in particular, preference statements, are never unbounded. We begin by introducing $\forall (i, k) \in \mathbb{I},\ \Delta v_{i,k} := v_i(d_{(i,k+1)}) - v_i(d_{(i,k)})$ and proving covectors, when applied to such a vector $\Delta_v$ of differences in value, correctly compute the difference of value between core alternatives.

We break down the Definition 12 by criterion:

$\forall i \in N, \quad \forall x_i, y_i \in \mathbb{X}_i, \quad$ let $(x_i, y_i) \in \mathbb{R}^{|\mathbb{D}_i|-1} : \forall k \in \mathbb{N} : 1 \leq k \leq |\mathbb{D}_i| - 1,$

$$(x_i, y_i)^\star_k := \begin{cases} +1, & \text{if } [d_{i,k}, d_{i,k+1}] \subset [y_i, x_i] \\ -1, & \text{if } [d_{i,k}, d_{i,k+1}] \cap ]x_i, y_i[ \neq \emptyset \\ 0, & \text{else} \end{cases}$$

So that $\forall x, y \in \mathbb{X}, \forall (i, k) \in \mathbb{I},\ (x, y)^\star_{(i,k)} = (x_i, y_i)^\star_k$.

**Lemma 1** (expression of differences in value as a product)

$$\forall i \in N, \forall x_i, y_i \in \mathbb{D}_i, \forall V \in \mathbb{V}, \quad v_i(x_i) - v_i(y_i) = \sum_{k=1}^{|\mathbb{D}_i|-1} (x_i, y_i)^\star_k \Delta v_{(i,k)}$$

*Proof* First, we note that for any valid indexes $k_1 < k_2$, $\sum_{k=k_1}^{k_2} \Delta v_{(i,k)} = v_i(d_{i,k_2}) - v_i(d_{i,k_1})$

Second, we detail $\sum_{k=1}^{|\mathbb{D}_i|-1}(x_i, y_i)_k^\star \Delta v_{(i,k)}$, according to the sign of $x_i - y_i$ :

- If $x_i > y_i$, the interval $]x_i, y_i[$ is empty, so the case leading to a coefficient $(x, y)_{(i,k)}^\star = -1$ does not occur. Non-zero coefficients correspond to intervals $[d_{i,k}, k_{i,k+1}[$ partitioning $[y_i, x_i[$, so that $\sum_{k=1}^{|\mathbb{D}_i|-1}(x_i, y_i)_k^\star \Delta v_{(i,k)} = (+1)(v_i(x_i) - v_i(y_i))$
- If $x_i < y_i$, the interval $[y_i, x_i]$ is empty, so the case leading to a coefficient $(x, y)_{(i,k)}^\star = +1$ does not occur. Non-zero coefficients correspond to intervals $[d_{i,k}, k_{i,k+1}[$ partitioning $[x_i, y_i[$, so that $\sum_{k=1}^{|\mathbb{D}_i|-1}(x_i, y_i)_k^\star \Delta v_{(i,k)} = (-1)(v_i(y_i) - v_i(x_i)) = v_i(x_i) - v_i(y_i)$
- If $x_i = y_i$, the interval $[x_i, y_i]$ is trivial and the interval $]x_i, y_i[$ is empty, so every coefficient $(x, y)_{(i,k)}^\star$ is equal to zero. Consequently, $\sum_{k=1}^{|\mathbb{D}_i|-1}(x, y)_{(i,k)}^\star \Delta v_{(i,k)} = 0 = v_i(x_i) - v_i(y_i)$.

Thus, $\forall i \in N,\ v_i(x) - v_i(y) = \sum_{k=1}^{|\mathbb{D}_i|-1}(x, y)_{(i,k)}^\star \Delta v_{(i,k)}$. □

For any alternatives $x, y \in \mathbb{D}$, summing up these equalities over every criteria yields $V(x) - V(y) = (x, y)^\star \times \Delta_v$

Introducing $\forall x, y \in \mathbb{X},\ \Delta V_{\inf}(x, y) := \inf_{V \in \mathbb{V}_\mathcal{P}} V(x) - V(y) \in \mathbb{R} \cup \{-\infty\}$, Definition 9 states that

$$\forall x, y \in \mathbb{X},\ (x, y) \in \mathcal{N}_\mathcal{P} \iff \Delta V_{\inf}(x, y) \geq 0$$

In the case of pairs of core alternatives, the objective function as well as the constraints of the minimization problem $\Delta V_{\inf}(x, y)$ can be expressed using covectors and matrix multiplication, as permitted by Lemma 1, so that $\Delta V_{\inf}(x, y)$ is a linear program.

**Lemma 2** (query between core alternatives)

$$\forall x, y \in \mathbb{D},\ \Delta V_{\inf}(x, y) = \inf\ (x, y)^\star \times \Delta v \ \text{ s.t. } \Delta v \in \Omega_\mathcal{P} \cap \Omega_\mathcal{D}$$

*with* $\Omega_\mathcal{P} := \{\Delta v \in \mathbb{R}^\mathbb{I} : \forall \pi \in \mathcal{P},\ \pi^\star \times \Delta v \geq 0\}$ *and* $\Omega_\mathcal{D} := \{\Delta v \in \mathbb{R}^\mathbb{I} : \forall (i, k) \in \mathbb{I},\ \delta_{(i,k)}^\star \times \Delta v \geq 0\}$.

Generally, with alternatives $(x, y)$ not necessarily belonging to the core $\mathbb{D}$, it has been shown Greco et al. (2008) that minimizing $V(x) - V(y)$ over $V \in \mathbb{V}_\mathcal{P}$ is still a linear program, with additional decision variables accounting for the distinct values $\{x_i, y_i\} \notin \mathbb{D}_i$. The $v_i(x_i), v_i(y_i)$ are only constrained by the monotonicity of the marginal value functions, so the problem is separate:

$$\Delta V_{\inf} = \inf_{\Delta v \in \Omega_\mathcal{P} \cap \Omega_\mathcal{D}} \sum_{i \in N} \inf_{\substack{v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i}} v_i(x_i) - v_i(y_i)$$

with, $\forall i \in N$,

$$\begin{cases} UX_i := \{v_i(x_i) \in \mathbb{R} : \forall z_i \in \mathbb{D}_i \cup \{y_i\}, & z_i \succsim_i x_i \Rightarrow v_i(z_i) \geq v_i(x_i)\} \\ LX_i := \{v_i(x_i) \in \mathbb{R} : \forall z_i \in \mathbb{D}_i \cup \{y_i\}, & z_i \precsim_i x_i \Rightarrow v_i(z_i) \leq v_i(x_i)\} \\ UY_i := \{v_i(y_i) \in \mathbb{R} : \forall z_i \in \mathbb{D}_i \cup \{x_i\}, & z_i \succsim_i y_i \Rightarrow v_i(z_i) \geq v_i(y_i)\} \\ LY_i := \{v_i(y_i) \in \mathbb{R} : \forall z_i \in \mathbb{D}_i \cup \{x_i\}, & z_i \precsim_i y_i \Rightarrow v_i(z_i) \leq v_i(l_i)\} \end{cases}$$

Thus, it is possible to circumvent this augmentation of the decision space by:

- Considering a given criterion $i \in N$ and a given vector $\Delta v \in \Omega_{\mathcal{P}} \cap \Omega_{\mathcal{D}}$;
- Directly assigning the additional decision variables to their optimal values in the inner linear program

$$\inf_{v_i(x_i), v_i(y_i)} v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases};$$

- Checking this optimal case is correctly represented, either by an unbounded pair or in covector form.

We begin by focusing on the case where the values of $\mathbb{D}_i \cup \{x_i, y_i\}$ are all different. We sort these values in strictly ascending order, and we detail three cases according to the position of $x_i$ and $y_i$ amongst these $|\mathbb{D}_i| + 2$ values:

- The interval $[x_i, y_i]$ overflows the set $\mathbb{D}_i$, so that the pair $(x, y) \in \mathcal{U}_{\mathcal{P}}$ is unbounded. This case actually encompasses three subcases
- $x_i$ has no predecessor, when $x_i$ is the least element of $\mathbb{D}_i \cup \{x_i, y_i\}$. There is no constraints in $LX_i = \mathbb{R}$;
- $y_i$ has no successor, when $y_i$ is the highest element of $\mathbb{D}_i \cup \{x_i, y_i\}$. There are no constraints in $UY_i = \mathbb{R}$;
- Both preceding cases are simultaneously satisfied.

In any case,

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = -\infty,$$

thus $V_{\inf}(x, y) = -\infty$ and $(x, y) \notin \mathcal{N}_{\mathcal{P}}$, thus proving Theorem 4;

- $y_i$ is the predecessor of $x_i$, *so* $x_i$ is the successor of $y_i$. In this case, the constraints $UX_i, LX_i, UY_i, LY_i$ can all be replaced by the single equality $v_i(x_i) = v_i(y_i)$, which defines a solution both feasible and where the objective function is minimized with respect to the decision variables $v_i(x_i), v_i(y_i)$. Meanwhile, we consider the coefficients $(x, y)^{\star}_{(i,k)}$, $1 \leq k < |\mathbb{D}_i|$: the interval $[y_i, x_i]$ does not contain a single core value $d_{i,k} \in \mathbb{D}_i$, hence $(x, y)^{\star}_{(i,k)} \neq +1$; the interval $]x_i, y_i[$ is empty, hence $(x, y)^{\star}_{(i,k)} \neq -1$; finally $(x, y)^{\star}_{(i,k)} = 0$. This proves the identity:

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = \sum_{k=1}^{|\mathbb{D}_i|-1} (x, y)^{\star}_{(i,k)} \Delta u_{(i,k)},$$

  as both sides are equal to zero.
- $x_i$ has a predecessor *which is not* $y_i$, and $y_i$ has a successor which is not $x_i$. First, we rewrite $\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases}$ as a difference in marginal value between surrogate alternatives in the core $\mathbb{D}_i$. The predecessor $\underline{x_i}$ of $x_i$ is given by $\underline{x_i} := \max\{d \in \mathbb{D}_i, d \precsim_i x_i\}$, so that the constraints $UX_i, LX_i$ can

both be replaced by the single equality $v_i(x_i) = v_i(\underline{x_i})$, which defines a solution both feasible and where $v_i(x_i)$ is minimal with respect to the decision variable $v_i(x_i)$. The successor $\overline{y_i}$ of $y_i$ is given by $\overline{y_i} := \min\{d \in \mathbb{D}_i, d \succsim_i y_i\}$, so that the constraints $UY_i$, $LY_i$ can both be replaced by the single equality $v_i(y_i) = v_i(\overline{y_i})$, which defines a solution both feasible and where $v_i(y_i)$ is maximal, so the objective function is minimal, with respect to the decision variable $v_i(y_i)$.

Thus,

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = v_i(\underline{x_i}) - v_i(\overline{y_i})$$

Second, as both surrogate alternatives $\underline{x_i}$, $\overline{y_i}$ belong to $\mathbb{D}_i$, Lemma 1 ensures that

$$v_i(\underline{x_i}) - v_i(\overline{y_i}) = \sum_{k=1}^{|\mathbb{D}_i|-1} (\underline{x_i}, \overline{y_i})_k^\star \Delta u_{(i,k)}$$

Third, we check that the covector coefficients for criterion $i$ of the original pair match those of the surrogate pair, that is:

$$\forall k \in \mathbb{N} : 1 \leq k < |\mathbb{D}_i|, \ (x_i, y_i)_k^\star = (\underline{x_i}, \overline{y_i})_k^\star$$

The proof is straightforward:

- If $x_i \succ_i y_i$, then there is at least one attribute value $d \in \mathbb{D}_i$ between $x_i$ and $y_i$, so that the predecessor of $x_i$ and the successor of $y_i$ are in the same order, thus $\underline{x_i} \succsim_i \overline{y_i}$. Hence, the coefficient indexed by $(i,k)$ of their respective covectors are in $\{0, +1\}$, with value $+1$, respectively, when $y_i \precsim_i d_{i,k} \prec_i d_{i,k+1} \precsim_i x_i$ and when $\overline{y_i} \precsim_i d_{i,k} \prec_i d_{i,k+1} \precsim_i \underline{x_i}$. The definition of the surrogate pair ensures these conditions are equivalent.
- If $x_i \prec_i y_i$, then obviously $\underline{x_i} \precsim_i \overline{y_i}$. Hence, the coefficients of their respective covectors indexed by $(i,k)$ are in $\{0, -1\}$, with value 0, respectively, when $y_i \precsim_i d_{i,k}$ or $d_{i,k+1} \precsim_i x_i$, and when $\overline{y_i} \precsim_i d_{i,k}$ or $d_{i,k+1} \precsim_i \underline{x_i}$. The definition of the surrogate pair ensures these conditions are equivalent. Thus,

$$\inf v_i(x_i) - v_i(y_i) \text{ s.t. } \begin{cases} v_i(x_i) \in UX_i \cap LX_i \\ v_i(y_i) \in UY_i \cap LY_i \end{cases} = \sum_{k=1}^{|\mathbb{D}_i|-1} (x_i, y_i)_k^\star \Delta u_{(i,k)}$$

The cases where $|\mathbb{D}_i \cup \{x_i, y_i\}| = |\mathbb{D}_i| + 1$ are correctly handled in the discussion above: if overflow (when either $x_i \prec_i \min \mathbb{D}_i$ or $y_i \succ_i \max \mathbb{D}_i$) does not occur, the case $x_i = y_i$ extends the case where the optimal value of $v_i(x_i) - v_i(y_i)$ is zero; the case where $y_i \in \mathbb{D}_i$ leads to the introduction of $\overline{y_i} := y_i$, and the case where $x_i \in \mathbb{D}_i$ leads to $\underline{x_i} := x_i$.

Finally, for any pair $(x, y) \in \mathbb{X}^2$, we have proven that, in every case, either the pair is unbounded and not in the relation $\mathcal{N}_\mathcal{P}$, or it can be represented by a covector such that $\Delta V_{\inf}(x, y) = \inf_{\Delta v \in \mathbb{R}^\mathbb{I}} (x, y)^\star \times \Delta v$ s.t. $\begin{cases} \forall \pi \in \mathcal{P},\ \pi^\star \times \Delta v \geq 0 \\ \forall (i, k) \in \mathbb{I},\ \delta^\star_{(i,k)} \times \Delta v \geq 0 \end{cases}$

*Proof of (2) $\iff$ (3)*

By Farkas' lemma, the problem (2) has no solution if, and only if, the objective linear form $(x, y)^\star$ is a linear combination with non-negative coefficients of the constraint linear forms $\{\pi^\star, \pi \in \mathcal{P}\}$ and $\{\delta^\star_{i,k}, (i, k) \in \mathbb{I}\}$.

*Proof of (3) $\iff$ (4)*

Obviously, (4) $\Rightarrow$ (3). Conversely, as the covectors involved in (3) have integral coordinates, the non-negative coefficients $\{\lambda_\pi, \pi \in \mathcal{P}\}$ and $\{\mu_{(i,k)}, (i, k) \in \mathbb{I}\}$, if they exist, can be chosen in the field of rational numbers. Multiplying the relation by the common denominator $n \in \mathbb{N}^\star$ of these coefficients leads to (4).

**Proof of Theorem 6**

We prove Theorem 6 in four steps: (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4) $\Rightarrow$ (1).

– (1) $\Rightarrow$ (2): Assume a statement $\sigma := (x, y) \in \mathcal{E}_2(\mathcal{N}_\mathcal{P})$. By Theorem 1 and Definition 5, there is an integer $n$ and a tuple $(e_0, e_1, \ldots, e_n) \in \mathbb{X}^n$ such that $e_0 = x, e_n = y$ and $(e_j, e_{j+1}) \in \mathcal{D} \cup \Delta_2$ for any integer $j < n$. This transitive chain of dominance relations and swaps of order 2 can be transformed into the covector relation sought, by induction on the length of the explanation, as described by the following lemmas:

**Lemma 3** (covector representation of dominance relations)

$$\forall \rho \in \mathcal{D}, \exists q \in \{0, +1\}^I : \rho^\star = \sum_{(i,k) \in \mathbb{I}} q_{(i,k)} \delta^\star_{(i,k)}$$

*Proof* A dominance relation has no negative argument, so its covector coefficient, given by Definition 12, is in $\{0, +1\}$. □

**Lemma 4** (covector representation of transitivity relations)

$$\forall x, y, z \in \mathbb{X}, \exists q \in \mathbb{N}^\mathbb{I} : (x, z)^\star = (x, y)^\star + (y, z)^\star + \sum_{(i,k) \in \mathbb{I}} q_{(i,k)} \delta^\star_{(i,k)}$$

*Proof* For core alternatives $x, y, z \in \mathbb{D}$, for any separate value function $V \in \mathbb{V}$,

$$\begin{aligned} (x, z)^\star \times \Delta v &= V(x) - V(z) \\ &= (V(x) - V(y)) + (V(y) - V(z)) \end{aligned}$$

$$= (x, y)^\star \times \Delta v \; + \; (y, z)^\star \times \Delta v$$
$$= ((x, y)^\star + (y, z)^\star) \times \Delta v$$

As the relation above stands for any vector $\Delta v \in [0, +\infty[$, it yields $(x, z)^\star = (x, y)^\star + (y, z)^\star = (x, y)^\star + (y, z)^\star + \sum_{(i,k)\in\mathbb{I}} q_{(i,k)}\delta^\star_{(i,k)}$ with $q = 0$.

For alternatives not necessarily in the core, and for any criterion $i \in N$, the trivial cases where $y_i \in \{x_i, z_i\}$, the case where $x_i = z_i$, or the case where $x_i, y_i, z_i$ are all distinct, divided into 6 subcases considering the order of attributes $x_i, y_i, z_i$, all lead to $(x, z)^\star \geq (x, y)^\star + (y, z)^\star$ because of the rounding down of broken intervals occurring once in the LHS and twice in the RHS. As both sides are covectors with integer coefficients, the difference $(x, z)^\star - ((x, y)^\star + (y, z)^\star)$ is a covector with non-negative integer coefficients $q_{(i,k)}$. □

– (2) ⇒ (3): Suppose there exists integer coefficients $a, \ell_1, \ldots, \ell_q, m_1, \ldots, m_n$ and preference swaps of order 2: $\gamma_1, \ldots, \gamma_q$ such that

$$a\sigma^\star = \sum_k \ell_k \gamma_k^\star + \sum_k m_k \delta^\star_{(k,1)} \tag{3}$$

Multiplying both sides of the covector Equation (3) by the vector $(1, \ldots, 1)$, we obtain the relation:

$$M := a(|\sigma^+| - |\sigma^-|) = \sum m_k \geq 0$$

To homogenize the right-hand side, we represent the dominance relation thanks to a dummy criterion: $N' = N \cup \{0\}$ so that $\tilde{\Delta_1} := \{(i, 0), i \in N\} \subset N'^2$. Thus, relation $\mathcal{D} \cup \Delta_2$ is a graph with nodes in $N'$. Re-indexing coefficients $\ell_k$ by the positive and negative arguments of swap $\gamma_k$ (summing up duplicates if needed), and introducing $\ell_{k,0} := m_k$ :

$$a\,\sigma^\star = \sum_{\gamma\in\tilde{\Delta_1}\cup\tilde{\Delta_2}} \ell_{\gamma^+,\gamma^-}\gamma^\star \tag{4}$$

To complete the flow $\ell$, we introduce:

– A source $s$ supplying flow $\ell_{s,i} = a$ to the positive arguments $i \in \sigma^+$;
– A sink $t$ collecting flow $\ell_{j,t} = a$ from the negative arguments $j \in \sigma^-$, and $\ell_{0,t} = M$ from node 0.

Covector Equation (4) ensures $\ell$ defines a feasible flow on the graph $(N' \cup \{s, t\}, \tilde{\Delta_1} \cup \tilde{\Delta_2} \cup \{s\} \times \sigma^+ \cup \sigma^- \times \{t\} \cup \{(0, t)\})$, without capacity constraints, as projection on the $i^{th}$ coordinate ensures flow conservation for node $i \in N$. Flow $\ell$ can be decomposed as a superposition of:

– Cycles, involving necessary equivalence between the nodes, and not contributing to the value of the flow;

– Paths from the source $s$ to the sink $t$ passing through node 0, denoting a dominance relation. Their total contribution to the value of the flow is $M$;
– Paths from the source $s$ to the sink $t$ not passing through node 0, with an overall contribution of $a \times |\sigma^-|$ to the value of the flow. Each of these paths links a positive argument $i_1 \in \sigma^+$ to a negative argument $i_r \in \sigma^-$ through necessary preference swaps of order 2. Transitivity of the necessary preference relation entails that $i_1$ is necessarily preferred to $i_r$: the edge $(i_1, i_r)$ belongs to $\Delta_2 \cap (\sigma^+ \times \sigma^-)$.

We reduce the flow $\ell$ by ignoring the cycles and paths passing through node 0. In addition, the flow $a$ carried by the path from source to sink $s \to i_1 \to i_2 \to \cdots \to i_r \to t$ is redirected to edge $(i_1, i_r)$. As a result, we obtain a flow of value $a|\sigma^-|$ on the graph of the relation $\tilde{\Delta_2}$ restricted to $\sigma^+ \times \sigma^-$. This entails the existence of a matching of cardinality $|\sigma^-|$ in this graph, obtained by setting an upper capacity constraint of value 1 on each edge leaving the source $s$ and entering the sink $t$ (as a cut of capacity $C$ on the network with capacity constraints $c_{i,j} \in \{1, \infty\}$ is a cut of capacity $a \times C$ on the same network with capacity constraints $a \times c_{i,j}$).

- (3) $\Rightarrow$ (4) is simply a rewording.
- (4) $\Rightarrow$ (1): Let $\phi : \sigma^- \to \sigma^+$, injective, such that $\forall k \in \sigma^-, (\phi(k), k) \in \tilde{\Delta_2}$. Given any ordering $O$ of the negative argument set $\sigma^-$, we can build a sequence of alternatives of decreasing preference $e_0 := x, e_1, \ldots, e_{|\sigma^-|} \in V$ such that the $k^{th}$ statement $(e_{k-1}, e_k)$ matches the criteria swap $(\phi(O_k), O_k) \in \tilde{\Delta_2}$:

$$N^{\neq}_{(e_{k-1}, e_k)} := \{\phi(O_k), O_k\}\ ;\ N^{=}_{(e_k, y)} := N^{=}_{(e_{k-1}, y)} \cup \{\phi(O_k), O_k)\}$$

  Thus, the sequence of sets $(e_k, y)^-$ decreases from $\sigma^-$ to $\emptyset$, one element at a time, and the sequence of sets $(e_k \succsim y)^+$ also decreases from $\sigma^+$ to $\sigma^+ \backslash \phi[\sigma^-]$, one element at a time. If the set $\sigma^+ \backslash \phi[\sigma^-]$ is empty, $e_{|\sigma^-|} = y$, and the sequence $x = e_0, \ldots, e_{|\sigma^-|} = y$ is an explanation of $(x, y) \in \mathcal{N}_{\mathcal{P}}$ by preference swaps of order 2, of length $|\sigma^-|$. Else, $e_{|\sigma^-|} \neq y$ but $(e_{|\sigma^-|}, y)$ is a dominance statement, as its negative argument set is empty. Thus, the sequence $x = e_0, e_1, \ldots, e_{|\sigma^-|}, y$ is an explanation of $(x, y) \in \mathcal{N}_{\mathcal{P}}$ by preference swaps of order 2 and a dominance relation, of length $|\sigma^-| + 1$.

## References


Aingworth, D., Chekuri, C., & Motwani, R. 1996. Fast estimation of diameter and shortest paths (without matrix multiplication). In *Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '96, pp. 547–553

Bana e Costa, C. A., & Vansnick, J.C. (1995). General overview of the MACBETH approach. In Pardalos P. M., Siskos Y., & Zopounidis C., (Eds.) *Advances in Multicriteria Analysis*, pages 93–100. Dordrecht: Kluwer Academic Publishers

Bana e Costa, C. A., Lourencco J. C., Chagas, M. P. & Bana e Costa, J. C. (2008). Development of reusable bid evaluation models for the portuguese electric transmission company. *Decision Analysis*, 5(1):22–42

Boutilier, C., Brafman, R. I., Domshlak, C., Hoos, H. H., & Poole, D. (2004). Cp-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *J. Artif. Intell. Res. (JAIR)*, *21*, 135–191.

Brafman, R. I., Domshlak, C., & Shimony, S. E. (2006). On graphical modeling of preference and importance. *J. Artif. Intell. Res. (JAIR)*, *25*, 389–424.

Carenini, G., & Moore, J. D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence Journal*, *170*, 925–952.

Ch. Labreuche, Maudet N., & Ouerdane W. (2011). Minimal and complete explanations for critical multi-attribute decisions. In *Algorithmic Decision Theory (ADT)*, pp. 121–134, Piscataway, NJ, USA

Ch. Labreuche, Maudet, N., & Ouerdane, W. (2012). Justifying dominating options when preferences are incomplete. In *Proceedings of the European Conference on Artificial Intelligence*, *242*, pp 486–491, Montpellier, France, IOS Press

Eiter, T., & Gottlob, G. (1995). The complexity of logic-based abduction. *J. ACM*, *42*(1), 3–42.

Even, S., & Tarjan, R. E. (1975). Network flow and testing graph connectivity. *SIAM J. Comput.*, *4*(4), 507–518.

Fishburn, P.C. (1997). Cancellation conditions for multiattribute preferences on finite sets. In Mark, H. Karwan, J. S., & Jyrki W. (Eds.) *Essays In Decision Making*, pp. 157–167. Springer Berlin Heidelberg

Friedrich, G., & Zanker, M. (2011). A taxonomy for generating explanations in recommender systems. *AI Magazine*, *32*(3), 90–98.

Greco, S., Słowinski, R., Figueira, J., & Mousseau, V. (2010). Robust ordinal regression. In *Trends in Multiple Criteria Decision Analysis*, pp 241–284. Springer Verlag

Greco, S., Mousseau, V., & Słowinski, R. (2008). Ordinal regression revisited: Multiple criteria ranking with a set of additive value functions. *European Journal of Operational Research*, *191*, 416–436.

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

Herlocker, J. L., Konstan, J. A., & Riedl, J.(2000). Explaining collaborative filtering recommendations. In *Proceedings of the ACM conference on Computer Supported Cooperative Work*, pp. 241–250

Jacquet-Lagrèze, E., & Siskos, Y. (1982). Assessing a set of additive utility functions for multicriteria decision making: the UTA method. *European Journal of Operational Research*, *10*, 151–164.

Kazman, R., Klein, M., & Clements, P. (2000). *ATAM: Method for Architecture Evaluation*. TECHNICAL REPORT, CMU/SEI-2000-TR-004, http://www.sei.cmu.edu/reports/00tr004.pdf

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

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement*, volume 1: Additive and Polynomial Representations. Academic Press

Labreuche, Ch. (2011). A general framework for explaining the results of a multi-attribute preference model. *Artificial Intelligence Journal*, *175*, 1410–1448.

Michell, Joel. (1988). Some problems in testing the double cancellation condition in conjoint measurement. *Journal of Mathematical Psychology*, *32*(4), 466–473.

Nic, W. (2011). Computational techniques for a simple theory of conditional preferences. *Artificial Intelligence*, 175(7–8):1053–1091. Representing, Processing, and Learning Preferences: Theoretical and Practical Challenges.

Nunes, I., Miles, S., Luck, M., Barbosa, S., & Lucena, C. (2014) Pattern-based explanation for automated decisions. In *Proceedings of the 21st European Conference on Artificial intelligence*, pp. 669–674. IOS Press

O'Sullivan, B., Papadopoulos, A., Faltings, B., & Pu, P. (2007). Representative explanations for over-constrained problems. In *Proceedings of the 22nd national conference on Artificial intelligence*, pp. 323–328. AAAI Press

Pu, P., & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6):542 – 556. Special Issue On Intelligent User Interfaces.

Spliet, R., & Tervonen, T. (2014). Preference inference with general additive value models and holistic pair-wise statements. *European Journal of Operational Research*, *232*(3), 607–612.

Symeonidis, P., Nanopoulos, A., & Manolopoulos, Y. (2009). 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.

Ulrich, J. (2004) Quickxplain: Preferred explanations and relaxations for over-constrained problems. In *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 167–172, Menlo Park, California, 2004. AAAI Press /The MIT Press.