JIAF 2022

Step-wise Explanations for the Additive Model

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

[1]MICS, CentraleSupélec, Université Paris-Saclay
[2] Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc
[3] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6

Step-wise Explanations for the Additive Model

What ?

Explain methodically the comparative statement

$$X \succsim_\omega Y$$

where :

$\succsim_\omega$ is a binary relation on $2^{[m]}$

represented by the score function $\omega : \langle \omega_i \rangle_{i \in [m]}$ with $\omega_i \in \mathbb{N}$

$[m]$ is a finite set of indivible items/criteria

$$X \succsim_\omega Y \Leftrightarrow \omega(X) \geq \omega(Y)$$

$$\omega(X) = \sum_{i \in X} \omega_i \text{ and } \omega(Y) = \sum_{j \in Y} \omega_j$$

How ?

Throughout the use of **argument schemes**:

operators tying a sequence of statements, called the premise, satisfying some conditions, into another statement called the conclusion

For ?

Elicitation procedures improvement and **Justification of decision models outcomes**

in Decision Theory and Theory of measurement

Outline

1. Generalities
   - A motivating example
   - The Argument schemes & the Explanation problem
   - The Argument schemes & the Additive model properties
2. Technical aspects
   - The ceteris paribus scheme
   - The transitive scheme
   - The covering scheme
3. Numerical experimentation

Generalities

A motivating example

Criteria

(**a**) affordability: “acceptable” (+) or “expensive” (−),

(**b**) Metropole shipping fee included: “yes” (+) or “no” (−),

(**c**) Overseas shipping fee included: “yes” (+) or “no” (−),

(**d**) mask quality: “high” (+) or “good” (−),

(**e**) provider reputation: “good” (+) or “fair” (−),

(**f**) washable: “yes” (+) or “no” (−),

(**g**) customizable: “yes” (+) or “no” (−).

Performance table & score function

	a	**b**	**c**	**d**	**e**	**f**	**g**	Scores
W	−	+	−	−	−	+	+	204
X	−	−	+	+	+	−	−	188
Y	+	−	−	+	−	−	−	187
Z	−	−	+	−	+	−	+	166
ω	128	126	77	59	52	41	37	-

$$\mathsf{W} \equiv \mathbf{bfg} \mid \mathsf{X} \equiv \mathbf{cde} \mid \mathsf{Y} \equiv \mathbf{ad} \mid \mathsf{Z} \equiv \mathbf{ceg}$$

$$\omega(\mathsf{Z}) = \omega(\mathbf{ceg}) = 77 + 52 + 37 = 166$$

Explain why W is the best supplier

through the *step-wise* explanations of the comparative statements

$\mathsf{W} \succsim_\omega \mathsf{X}$, $\mathsf{W} \succsim_\omega \mathsf{Y}$ **and** $\mathsf{W} \succsim_\omega \mathsf{Z}$

ω is a hidden preference model

The Argument schemes & the Explanation problem

An argument scheme is an operator tying a sequence of statements, called the premise, satisfying some conditions, into another statement called the conclusion.

$\omega = \{\mathbf{a} : 128, \mathbf{b} : 126, \mathbf{c} : 77, \mathbf{d} : 59, \mathbf{e} : 52, \mathbf{f} : 41, \mathbf{g} : 37\}$

$\mathbf{bfg} \succsim_\omega \mathbf{cde}$

$\mathbf{ab} \succsim_\omega \mathbf{ad}$

$\mathbf{cde} \succsim_\omega \mathbf{ab}$

$\mathbf{bfg} \succsim_\omega \mathbf{cde}$

$\mathbf{a} \succsim_\omega \mathbf{c}$

$\mathbf{bc} \succsim_\omega \mathbf{ad}$

$\mathbf{bfg} \succsim_\omega \mathbf{ad}$

$\mathbf{b} \succsim_\omega \mathbf{de}$

$\mathbf{fg} \succsim_\omega \mathbf{c}$

$\mathbf{bfg} \succsim_\omega \mathbf{ceg}$

$\mathbf{bf} \succsim_\omega \mathbf{ce}$

$\mathbf{b} \succsim_\omega \mathbf{d}$

The Explanation problem

Inputs:

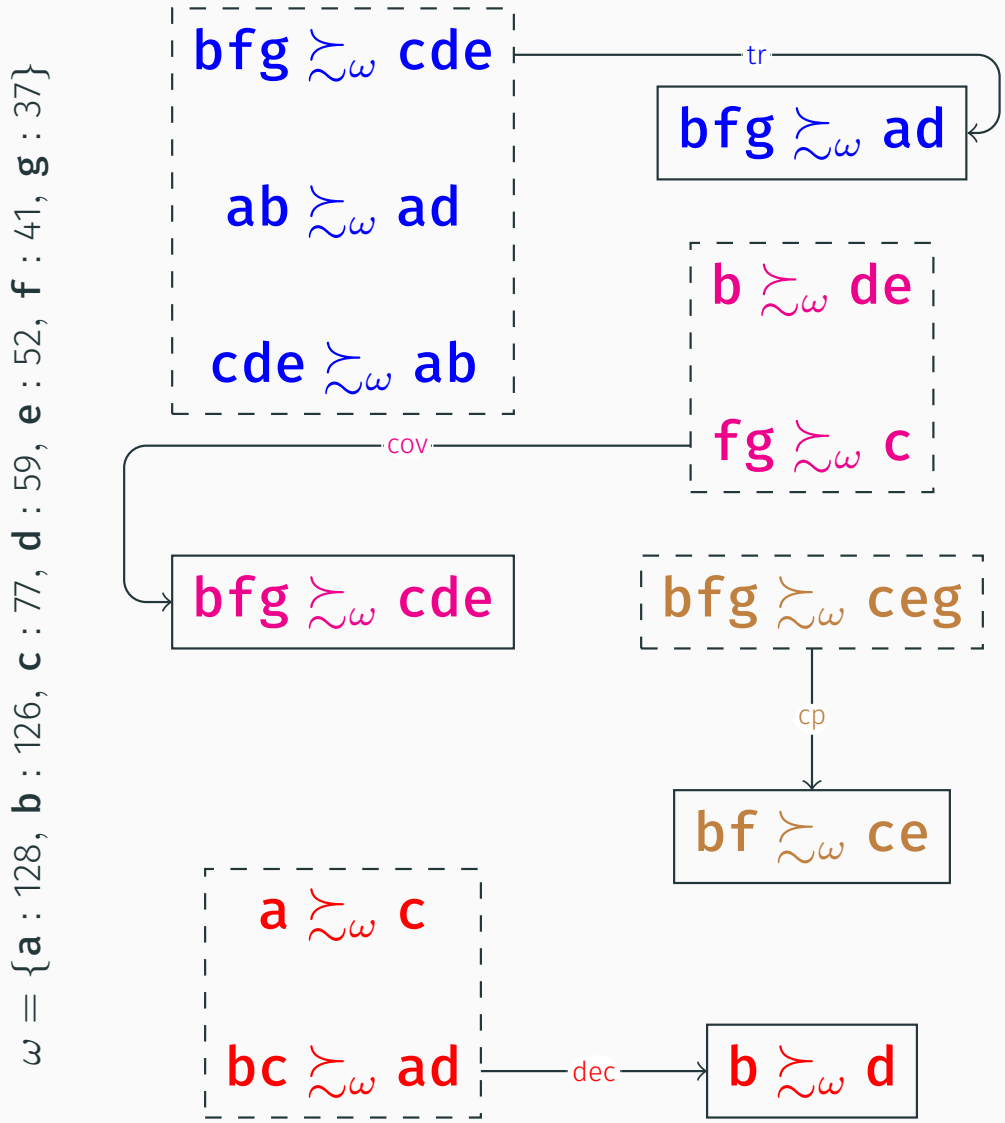
- The comparative statement $(A, B) \in 2^{[m]} \times 2^{[m]}$ to explain,
- The score function ω and the induced preference relation $\succsim_\omega$,
- $\mathcal{A}$: a set of statements belonging to $\succsim_\omega$,
- A set of argument schemes $\mathcal{S}$,
- a positive integer k.

Question:

Is there a positive integer $k' \leq k$, a list of length k' of statements $[(A_1, B_1), \ldots, (A_{k'}, B_{k'})]$ all belonging to $\mathcal{A}$, and a scheme $s \in \mathcal{S}$ such that $[(A_1, B_1), \ldots, (A_k, B_k)] \xrightarrow{s} (A, B)$?

The Argument schemes & the Explanation problem

An argument scheme is an operator tying a sequence of statements, called the premise, satisfying some conditions, into another statement called the conclusion.



The Explanation problem

Inputs:

- The comparative statement $(A,\ B) \in 2^{[m]} \times 2^{[m]}$ to explain,
- The score function ω and the induced preference relation $\succsim_\omega$,
- $\mathcal{A}$: a set of statements belonging to $\succsim_\omega$,
- A set of argument schemes $\mathcal{S}$,
- a positive integer k.

Question:

Is there a positive integer $k' \leq k$, a list of length k' of statements $[(A_1, B_1), \ldots, (A_{k'}, B_{k'})]$ all belonging to $\mathcal{A}$, and a scheme $s \in \mathcal{S}$ such that $[(A_1, B_1), \ldots, (A_k, B_k)] \xrightarrow{s} (A, B)$?

The additive model properties

Let $\succsim_\omega$ be the binary relation induced by the score function ω.

- $\succsim_\omega$ is ***transitive*** : $X \succsim_\omega Y$ and $Y \succsim_\omega Z$ imply $X \succsim_\omega Z$
- $\succsim_\omega$ satisfies ***1-order cancellation*** : $(X \setminus Y) \succsim_\omega (Y \setminus X) \iff X \succsim_\omega Y$
- $\succsim_\omega$ satisfies ***K-order cancellation*** ($K \geq 2$):

$$\left.\begin{array}{c} (X^{(1)}, \ldots, X^{(K)}) =_0 (Y^{(1)}, \ldots, Y^{(K)}) \\ \\ X^{(k)} \succsim_\omega Y^{(k)} \quad \text{for all } k < K \end{array}\right\} \Rightarrow \quad Y^{(K)} \succsim_\omega X^{(K)}$$

$(X^{(1)}, \ldots, X^{(K)}) =_0 (Y^{(1)}, \ldots, Y^{(K)})$ mean that for every $i \in [m]$, $|\{k : i \in X^{(k)}| = |\{k : i \in Y^{(k)}|$ i.e. each criterion i appears as many times in the subsets $X^{(k)}$ than in the subsets $Y^{(k)}$.

Technical aspects

Contents

Argument schemes and Additive model properties

The argument schemes : *Ceteris paribus scheme*

- derived from the ***1-order cancellation*** property.
- ***One premise :*** $(X \setminus Y) \succsim_\omega (Y \setminus X)$
  i.e. the comparison between the non common criteria of the comparative statement (X, Y) supported by ω.
- ***The conclusion :*** $X \succsim_\omega Y$.

e.g:

$$\omega = \{\mathbf{a} : 128, \mathbf{b} : 126, \mathbf{c} : 77, \mathbf{d} : 59, \mathbf{e} : 52, \mathbf{f} : 41, \mathbf{g} : 37\}$$

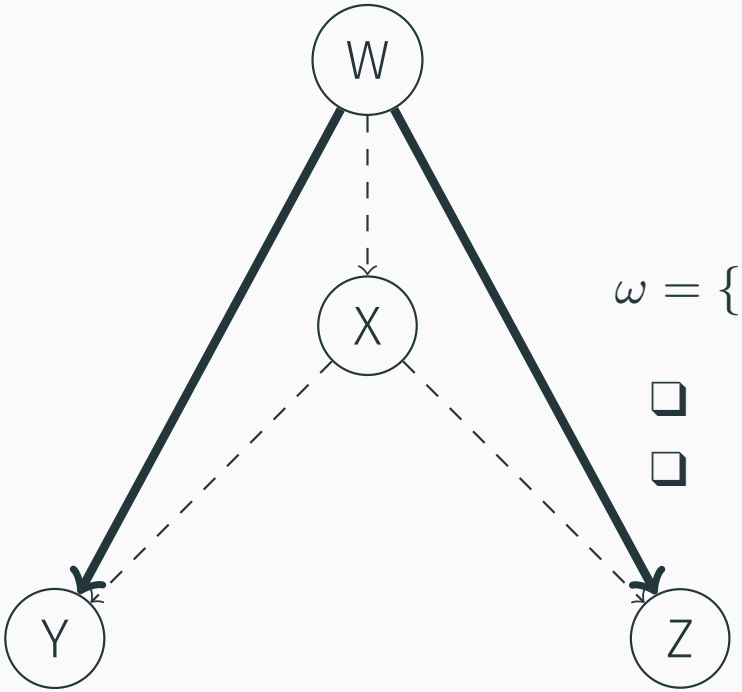
$$W \equiv \mathbf{bfg} \mid X \equiv \mathbf{cde} \mid Y \equiv \mathbf{ad} \mid Z \equiv \mathbf{ceg}$$

❑	$\mathbf{d} \succsim_\omega \mathbf{g}$	$\Rightarrow$	$X \succsim_\omega Z$
❑	$\mathbf{bf} \succsim_\omega \mathbf{ce}$	$\Rightarrow$	$W \succsim_\omega Z$
❑	$\mathbf{bfg} \succsim_\omega \mathbf{cde}$	$\Rightarrow$	$W \succsim_\omega X$

The argument schemes : *Transitive scheme*

- derived from the ***transitive*** property.
- ***Premises :*** $X^{(1)} \succsim_\omega Y^{(1)}, \ldots, X^{(k)} \succsim_\omega Y^{(k)}$ with $X^{(1)} = X$, $X^{(j)} = Y^{(j-1)}$ for all $j \geq 2$ and $Y^{(k)} = Y$.
- ***The conclusion :*** $X \succsim_\omega Y$.

e.g:



$\omega = \{\mathbf{a} : 128, \mathbf{b} : 126, \mathbf{c} : 77, \mathbf{d} : 59, \mathbf{e} : 52, \mathbf{f} : 41, \mathbf{g} : 37\}$

- $W \succsim_\omega X \quad \wedge \quad X \succsim_\omega Y \quad \Rightarrow \quad W \succsim_\omega Y$
- $W \succsim_\omega X \quad \wedge \quad X \succsim_\omega Z \quad \Rightarrow \quad W \succsim_\omega Z$

The argument schemes : *Covering scheme*

- derived from (a particular case of) the ***K-order cancellation*** property.

$$\left.\begin{array}{c}(X^{(1)},\ldots,X^{(K-1)},\ Y^{(K)}) =_0 (Y^{(1)},\ldots,Y^{(K-1)},\ X^{(K)}) \\ X^{(k)} \succsim_\omega Y^{(k)} \ \text{ for all } k < K\end{array}\right\} \Rightarrow \ X^{(K)} \succsim_\omega Y^{(K)}$$

- ***Premises :*** $X^{(k)} \succsim_\omega Y^{(k)}$ for all $k < K$ with
  - the subsets of criteria $X^{(1)},\ldots,X^{(K-1)}$ partitioning $X^{(K)}$.
  - the subsets of criteria $Y^{(1)},\ldots,Y^{(K-1)}$ partitioning $Y^{(K)}$.
  - $X \setminus Y = X^{(K)}$ and $Y \setminus X = Y^{(K)}$
- ***The conclusion :*** $X \succsim_\omega Y$.

e.g: $\omega = \{\mathbf{a} : 128,\ \mathbf{b} : 126,\ \mathbf{c} : 77,\ \mathbf{d} : 59,\ \mathbf{e} : 52,\ \mathbf{f} : 41,\ \mathbf{g} : 37\}$; $W \equiv \mathbf{bfg} \mid X \equiv \mathbf{cde}$

$$(\mathbf{b},\ \mathbf{fg},\ \overbrace{\mathbf{cde}}^{X}) =_0 (\mathbf{de},\ \mathbf{c},\ \overbrace{\mathbf{bfg}}^{W})$$

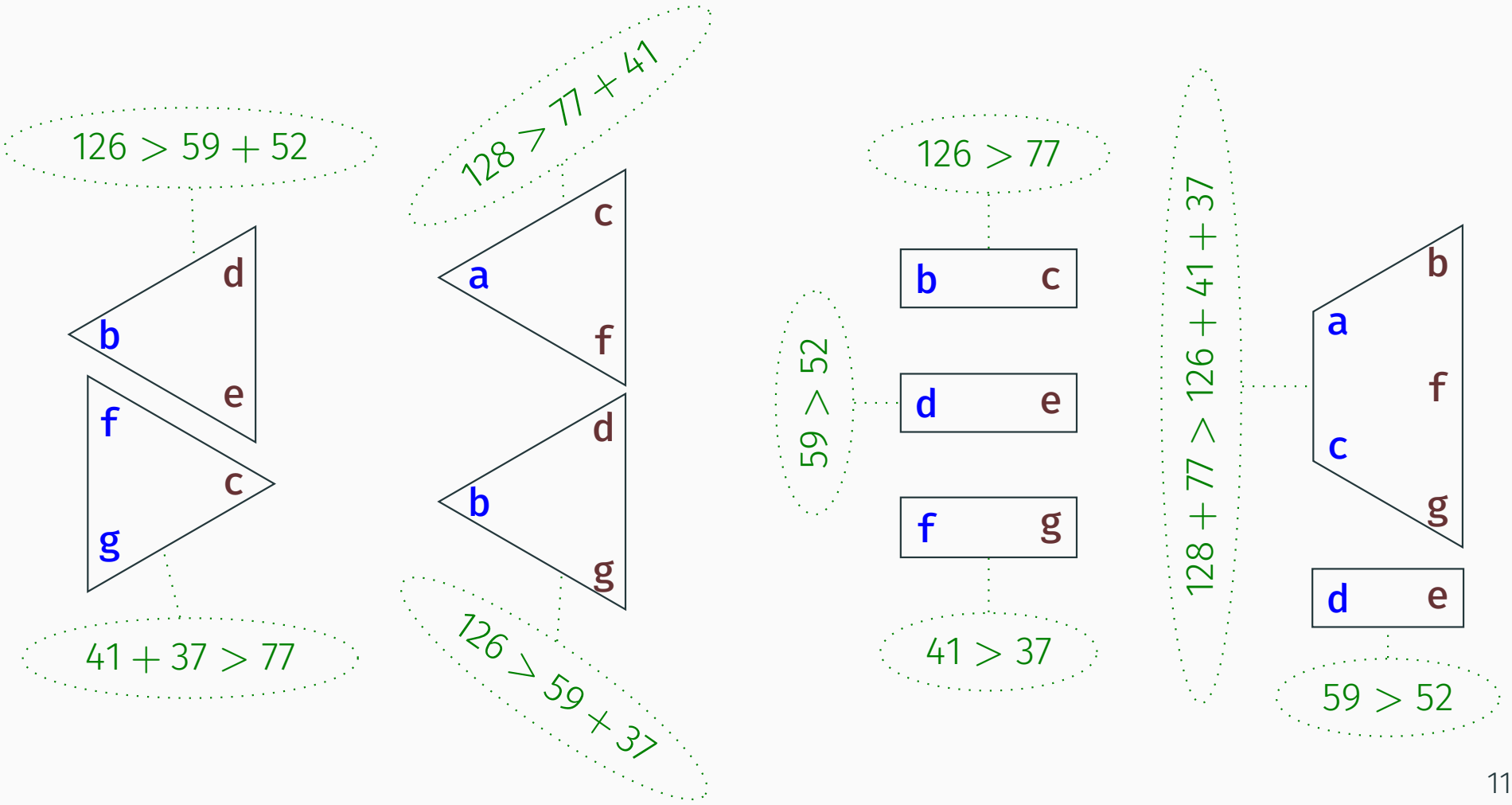
$$\mathbf{b} \succsim_\omega \mathbf{de}$$

$$\mathbf{fg} \succsim_\omega \mathbf{c}$$

The argument schemes : *Covering scheme*

$\omega = \{\mathbf{a} : 128, \mathbf{b} : 126, \mathbf{c} : 77, \mathbf{d} : 59, \mathbf{e} : 52, \mathbf{f} : 41, \mathbf{g} : 37\}$

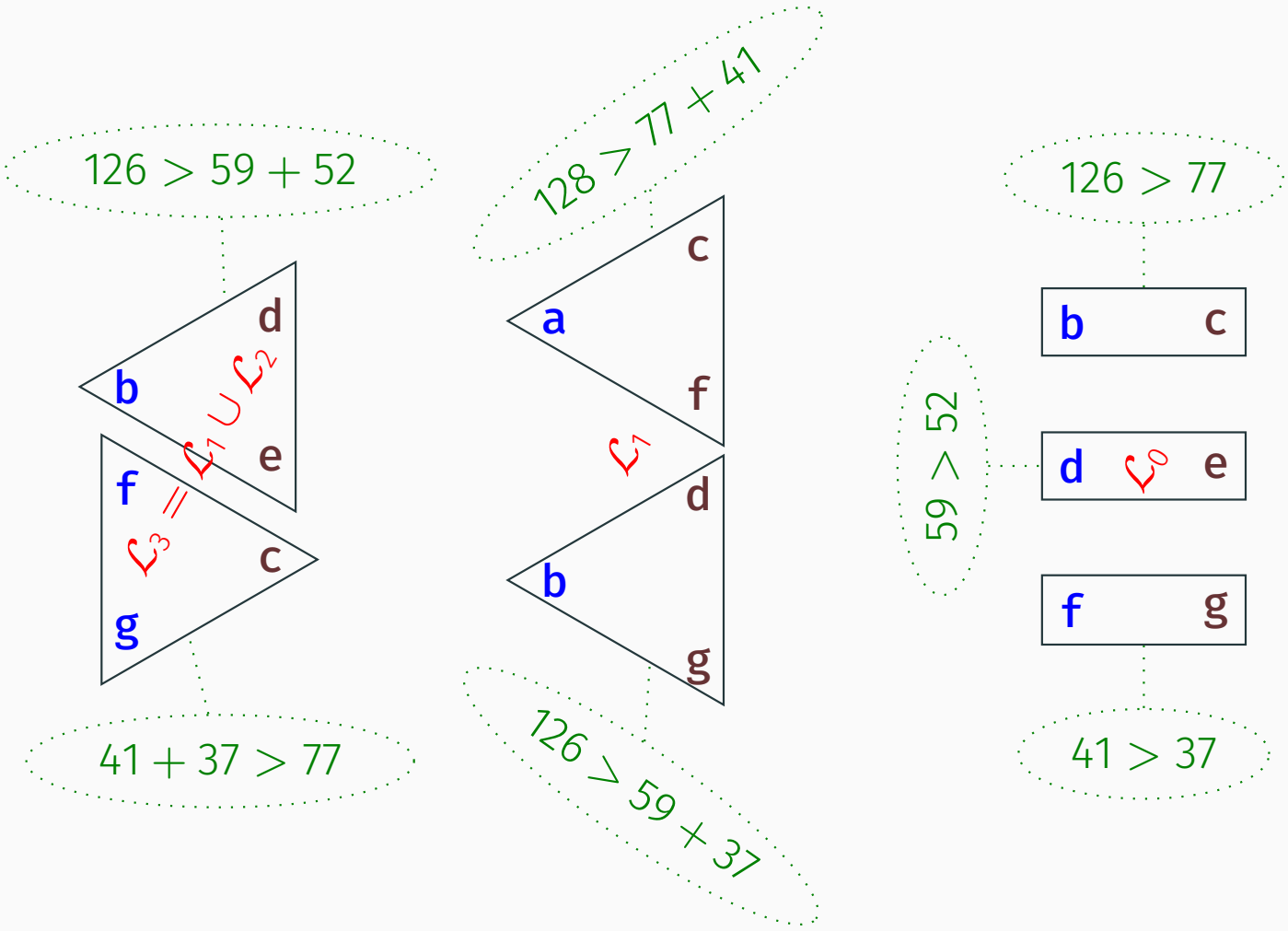
pro criteria vs. con criteria



The argument schemes : *Covering scheme*

$\omega = \{\mathbf{a} : 128, \mathbf{b} : 126, \mathbf{c} : 77, \mathbf{d} : 59, \mathbf{e} : 52, \mathbf{f} : 41, \mathbf{g} : 37\}$

pro criteria vs. con criteria

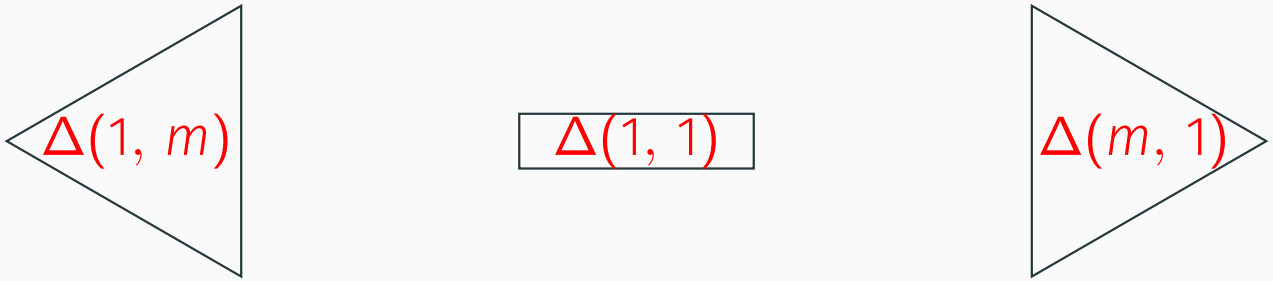


Decomposition Languages

- $\mathcal{L}_0 = \Delta(1, 1)$
- $\mathcal{L}_1 = \Delta(1, m)$
- $\mathcal{L}_2 = \Delta(m, 1)$
- $\mathcal{L}_3 = \mathcal{L}_1 \cup \mathcal{L}_2$

The argument schemes : *Covering scheme*

The chosen decomposition languages



+ Cognitively easy to grasp.
+ Easily scriptable in a natural language.
- Not complete.

$$\omega = \{\mathbf{a} : 128,\ \mathbf{b} : 126,\ \mathbf{c} : 77,\ \mathbf{d} : 59,\ \mathbf{e} : 52,\ \mathbf{f} : 41,\ \mathbf{g} : 37\}$$

$$\mathbf{cf} \succsim_{\omega} \mathbf{de}$$

is not decomposable !!!

Covering scheme: Computational complexity

Instance:

- A set $\mathbb{PRO}$ of pro criteria and a set $\mathbb{CON}$ of con criteria
- The preference model representation ω

$\mathcal{L}_0$

Question: Is there a bijective function f from $\mathbb{CON}$ to $\mathbb{PRO}$ such that for all $j \in \mathbb{CON}$, $\omega(f(j)) \geq \omega(j)$?

Complexity class: P

$\mathcal{L}_1$

Question: Is there a function f from $\mathbb{PRO}$ to $2^{\mathbb{CON}}$ such that :

1. $f(i) \cap f(i') = \emptyset$ if $i \neq i'$ with $i,\ i' \in \mathbb{PRO}$
2. $\bigcup_{i \in \mathbb{PRO}} f(i) = \mathbb{CON}$
3. $\omega(i) \geq \sum_{j \in f(i)} \omega(j)$ for all $i \in \mathbb{PRO}$

Complexity class: NP-hard

$\mathcal{L}_2$

Question: Is there a application g from $\mathbb{CON}$ to $2^{\mathbb{PRO}} \setminus \{\emptyset\}$ such that :

1. $g(j) \cap g(j') = \emptyset$ if $j \neq j'$ with $j,\ j' \in \mathbb{CON}$
2. $\sum_{i \in g(j)} \omega(i) \geq \omega(j)$ for all $j \in \mathbb{CON}$

Complexity class: NP-hard

$\mathcal{L}_3$

Question: Are there two disjoint subsets $\mathbb{PRO}^1$ and $\mathbb{PRO}^2$ of $\mathbb{PRO}$, two disjoint subsets $\mathbb{CON}^1$ and $\mathbb{CON}^2$ of $\mathbb{CON}$, a function f from $\mathbb{PRO}^1$ to $2^{\mathbb{CON}^1}$ and a application g from $\mathbb{CON}^2$ to $2^{\mathbb{PRO}^2} \setminus \{\emptyset\}$ such that:

1. $\mathbb{PRO}^1 \cup \mathbb{PRO}^2 = \mathbb{PRO}$ and $\mathbb{CON}^1 \cup \mathbb{CON}^2 = \mathbb{CON}$
2. $f(i) \cap f(i') = \emptyset$ if $i \neq i'$ with $i,\ i' \in \mathbb{PRO}^1$
3. $\bigcup_{i \in \mathbb{PRO}^1} f(i) = \mathbb{CON}^1$
4. $\omega(i) \geq \sum_{j \in f(i)} \omega(j)$ for all $i \in \mathbb{PRO}^1$
5. $g(j) \cap g(j') = \emptyset$ if $j \neq j'$ with $j,\ j' \in \mathbb{CON}^2$
6. $\sum_{i \in g(j)} \omega(i) \geq \omega(j)$ for all $j \in \mathbb{CON}^2$

Complexity class: NP-hard

$\mathcal{L}_3$—*Covering scheme*: Computation using ILP

Inputs

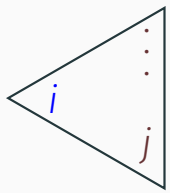
- The score function $\omega : \langle \omega_i \rangle_{i \in [m]}$.
- The pair (X, Y) to explain.

Notations

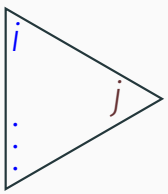
- *Pro* arguments set : $(X, Y)^+$
- *Con* arguments set : $(X, Y)^-$

(Binary) Variables

$$s^1_{ij} = \begin{cases} 1 & \text{if } j \in f(i). \\ 0 & \text{Otherwise.} \end{cases}$$



$$s^2_{ij} = \begin{cases} 1 & \text{if } i \in g(j). \\ 0 & \text{Otherwise.} \end{cases}$$



Constraints

- *Syntactic constraints*
  - For each *pro* argument i

$$s^1_{ij} + \sum_{j' \in (X, Y)^-} s^2_{ij'} \leq 1 \qquad \forall j \in (X, Y)^-$$

  - For each *con* argument j

$$\sum_{i' \in (X, Y)^+} s^1_{i'j} + s^2_{ij} \leq 1 \qquad \forall i \in (X, Y)^+$$

- ω—*compatibility constraints*
  - For each *pro* argument i

$$\omega_i \geq \sum_{j \in (X, Y)^-} \omega_j s^1_{ij}$$

  - For each *con* argument j

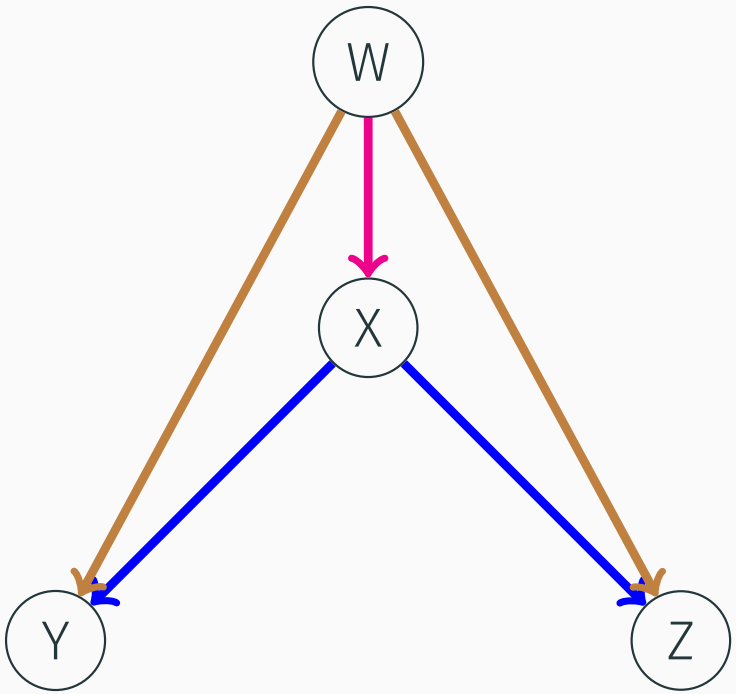
$$\sum_{i \in (X, Y)^+} (s^1_{ij} + s^2_{ij})\, \omega_i \geq \omega_j$$

Back to the motivating example

Combining Ceteris Paribus, Transitive and Covering schemes

$$\omega = \{\mathbf{a} : 128,\ \mathbf{b} : 126,\ \mathbf{c} : 77,\ \mathbf{d} : 59,\ \mathbf{e} : 52,\ \mathbf{f} : 41,\ \mathbf{g} : 37\}$$

$$W \equiv \mathbf{bfg} \mid X \equiv \mathbf{cde} \mid Y \equiv \mathbf{ad} \mid Z \equiv \mathbf{ceg}$$



$\mathcal{L}_3$—Covering Scheme

$$\star \quad \mathbf{b} \succsim_\omega \mathbf{de} \quad \wedge \quad \mathbf{fg} \succsim_\omega \mathbf{c} \quad \Rightarrow \quad W \succsim_\omega X$$

Transitive Scheme

$$\star \quad W \succsim_\omega X \quad \wedge \quad X \succsim_\omega Y \quad \Rightarrow \quad W \succsim_\omega Y$$
$$\star \quad W \succsim_\omega X \quad \wedge \quad X \succsim_\omega Z \quad \Rightarrow \quad W \succsim_\omega Z$$

Ceteris paribus Scheme

$$\star \quad \mathbf{ce} \succsim_\omega \mathbf{a} \quad \Rightarrow \quad X \succsim_\omega Y$$
$$\star \quad \mathbf{d} \succsim_\omega \mathbf{g} \quad \Rightarrow \quad X \succsim_\omega Z$$

Numerical experimentation

Numerical Experiments

Instance: $(\omega, \mathbb{A})$

- $|\mathbb{A}| = 10$
- $|\omega| = m \in [\![6; 15]\!]$

Sample: 500.000 instances

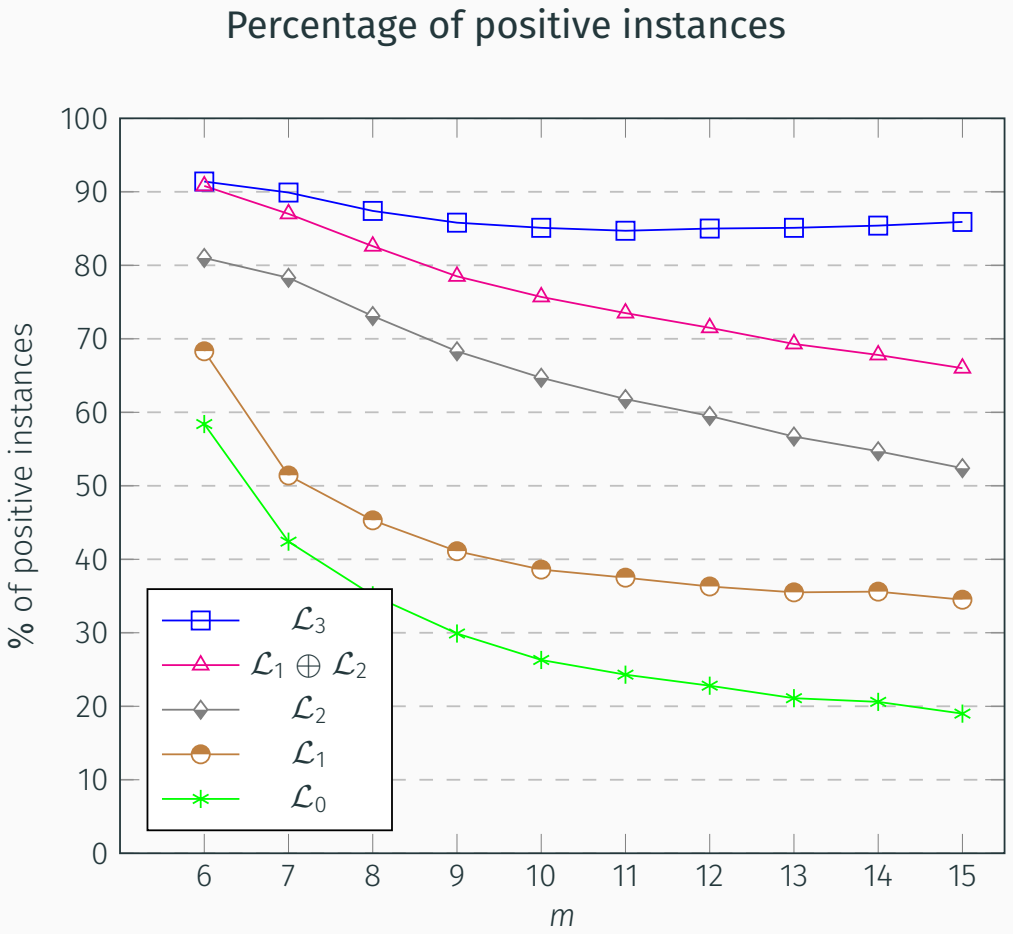
- 500 sets $\mathbb{A}$
- 1000 score functions ω

$\mathbb{A}$ set characteristics:

- no Pareto dominance
- significativity of all criteria

Argument Scheme used:

- Covering scheme



An instance is positive if and only if all "direct" statements (x^*, y) are $\mathcal{L}-$covering explainable.

m	6	7	8	9	10	11	12	13	14	15
$\mathcal{L}_3$	91.4%	89.9%	87.4%	85.8%	85.1%	84.7%	85.0%	85.1%	85.4%	85.9%
$\mathcal{L}_1 \oplus \mathcal{L}_2$	90.8%	87.0%	82.6%	78.5%	75.7%	73.5%	71.5%	69.3%	67.8%	66.0%
$\mathcal{L}_1$	81.0%	78.3%	73.1%	68.3%	64.7%	61.8%	59.1%	56.7%	54.7%	52.4%
$\mathcal{L}_2$	68.3%	51.4%	45.3%	41.1%	38.6%	37.5%	36.3%	35.5%	35.6%	34.5%
$\mathcal{L}_0$	58.4%	42.4%	35.1%	29.9%	26.3%	24.3%	22.8%	21.1%	20.6%	19.0%

Merci!

Bibliography i


Amgoud, L. and Prade, H. (2009).
Using arguments for making and explaining decisions.
Artificial Intelligence, 173(3):413–436.


Belahcene, K., Chevaleyre, Y., Labreuche, C., Maudet, N., Mousseau, V., and Ouerdane, W. (2018).
Accountable approval sorting.
In *Proceedings IJCAI'18*, pages 70–76.


Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., and Ouerdane, W. (2017).
Explaining robust additive utility models by sequences of preference swaps.
Theory and Decision, 82(2):151–183.

Bibliography ii


Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., and Ouerdane, W. (2019).
Comparing options with argument schemes powered by cancellation.
In *Proc. of IJCAI'19*, pages 1537–1543.


Belton, V. and Stewart, T. J. (2002).
Multiple criteria decision analysis - an integrated approach.
Springer.


Bouyssou, D., Marchant, T., Pirlot, M., Tsoukias, A., and Vincke, P. (2006).
***Evaluation and Decision models with multiple criteria: stepping stones for the analyst*, volume 86.**
Springer.

Bibliography iii


Greco, S., Mousseau, V., and Słowiński, R. (2008).
Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions.
EJOR, 191(2):416–436.


Hammond, J., Keeney, R., and Raiffa, H. (1998).
Even swaps: A rational method for making trade-offs.
Harvard business review, 76:137–8, 143.


Labreuche, C. and Fossier, S. (2018).
Explaining multi-criteria decision aiding models with an extended shapley value.
In *Proc. of IJCAI'18*, pages 331–339.

Bibliography iv

Labreuche, C., Maudet, N., and Ouerdane, W. (2011).
Minimal and complete explanations for critical multi-attribute decisions.
In *Proc. of ADT*, pages 121–134, Piscataway, NJ, USA.

Labreuche, C., Maudet, N., and Ouerdane, W. (2012).
Justifying dominating options when preferences are incomplete.

In *Proceedings of ECAI*, pages 486–491, Montpellier, France.