Characterization of Régnier's matrices in classification

Olivier Hudry

Télécom ParisTech

olivier.hudry@telecom-paristech.fr

Structured Matrix Days 2019 - Limoges

TELECOM ParisTech

A clustering problem (Régnier's problem)

- A set $X = \{1, 2, \ldots, n\}$ of n objects.
- A collection Π, called a *profile*, of p equivalence relations (= partitions) defined on X:

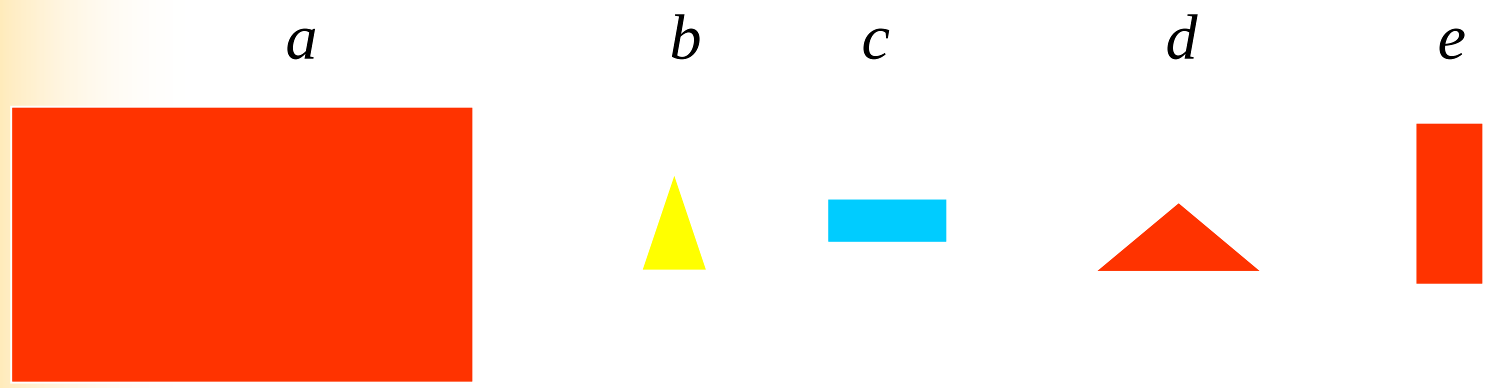
$$\Pi = (E_1, E_2, \ldots, E_p).$$

Each equivalence relation corresponds with a criterion gathering the objects sharing the same feature w.r.t. this criterion.

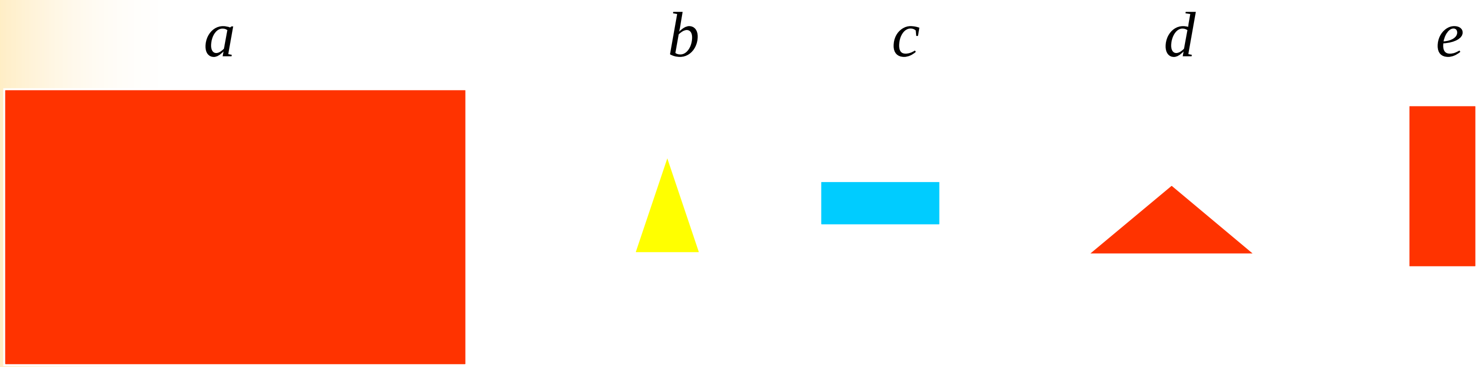
- We want to gather the n objects into clusters "as well as possible", i.e. so that the objects of any cluster look similar while the objects of two distinct clusters look dissimilar.

An example

- $n = 5$, $X = \{a, b, c, d, e\}$: *a* is a large red rectangle; *b* is a small yellow triangle; *c* is a small blue rectangle; *d* is a small red triangle; *e* is a small red rectangle.



An example



- $p = 3$:

  * E_1 (geometrical shape): *a*, *c* and *e* together (they are rectangles), *b* and *d* together (they are triangles): $E_1 = a\ c\ e \mid b\ d$

  * E_2 (colour): *a*, *d* and *e* together (they are red), *b* alone (the only yellow form), *c* alone (the only blue form): $E_2 = a\ d\ e \mid b \mid c$

  * E_3 (size): *a* alone (the only large form); *b*, *c*, *d* and *e* together (they are small): $E_3 = a \mid c\ b\ d\ e$.

- How to gather *a*, *b*, *c*, *d* and *e*?

Median equivalence relation of Π

- To specify what "as well as possible" means, consider the *symmetric difference distance* δ between two binary relations R and S defined on X:

$$\delta(R, S) = |\{(x, y) \in X^2 \text{ with } [xRy \text{ and not } xSy] \text{ or } [\text{not } xRy \text{ and } xSy]\}|$$

$\rightarrow$ $\delta(R, S)$ measures the number of disagreements between R and S.

- Then define the *remoteness* $\rho_\Pi(R)$ of R from $\Pi = (E_1, E_2, \ldots, E_p)$ by:

$$\rho_\Pi(R) = \sum_{i=1}^{p} \delta(R, E_i)$$

$\rightarrow$ $\rho_\Pi(R)$ measures the total number of disagreements between R and Π.

TELECOM ParisTech

Median equivalence relation of Π

- A *median equivalence relation* (or *median partition*, or also a *central partition*) of Π is an equivalence relation E^* minimizing ρ_Π:

$$\rho_\Pi(E^*) = \min \rho_\Pi(E)$$

$$\text{for } E \in \{\text{equivalence relations defined on } X\}.$$

- What is the complexity of the computation of a median equivalence relation of a profile of equivalence relations (Régnier's problem, 1965)?

- Rk. The computation of a median equivalence relation of a profile of symmetric relations is known to be NP-hard (M. Krivanek, J. Moravek, 1986; Y. Wakabayashi, 1986)

TELECOM ParisTech

Computation of $\rho_\Pi(E)$

- Let $(e_{xy})_{(x,\,y)\in X^2}$ be the *characteristic matrix* of E:

  $e_{xy} = 1$ if E gathers x and y and $e_{xy} = 0$ otherwise.

- $p_{xy} = 2|\{i: 1 \leq i \leq p \text{ and } E_i \text{ gathers } x \text{ and } y\}| - p = p_{yx}$.

- Then: $\rho_\Pi(E) = C - \sum_{(x,y)\in X^2} p_{xy}\, e_{xy}$

with :

$\forall\, x \in X,\ e_{xx} = 1$ (reflexivity)

$\forall\, (x, y) \in X^2,\ e_{xy} = e_{yx}$ (symmetry)

$\forall\, (x, y, z) \in X^3,\ e_{xy} + e_{yz} - e_{xz} \leq 1$ (transitivity)

$\forall\, (x, y) \in X^2,\ e_{xy} \in \{0, 1\}$ (binarity)

Majority matrix of Π

- The quantities p_{xy} summarize Π utterly:

$$p_{xy} = 2 \times (|\{i: 1 \leq i \leq p \text{ and } E_i \text{ gathers } x \text{ and } y\}| - p/2).$$

  * $p_{xy} > 0$ means that x and y are rather similar, and $p_{xy} < 0$ means that x and y are rather dissimilar;

  * $p_{xy} = p_{yx}$;

  * $p_{xx} = p$;

  * $-p \leq p_{xy} \leq p$;

  * all the p_{xy} have the parity of p.

- The ***majority matrix*** of Π is the matrix $P = (p_{xy})_{x,y}$.

Example

- $E_1 = a\ c\ e \mid b\ d$; $E_2 = a\ d\ e \mid b \mid c$; $E_3 = a \mid c\ b\ d\ e$.

p_{xy}	***a***	***b***	***c***	***d***	***e***
a	3	–3	–1	–1	1
b	–3	3	–1	1	–1
c	–1	–1	3	–1	1
d	–1	1	–1	3	1
e	1	–1	1	1	3

- Median equivalence relation?

Example

- $E_1 = a\ c\ e \mid b\ d$; $E_2 = a\ d\ e \mid b \mid c$; $E_3 = a \mid c\ b\ d\ e$.

p_{xy}	**a**	**e**	**b**	**d**	**c**
a	**3**	**1**	–3	–1	–1
e	**1**	**3**	–1	1	1
b	–3	–1	**3**	**1**	–1
d	–1	1	**1**	**3**	–1
c	–1	1	–1	–1	**3**

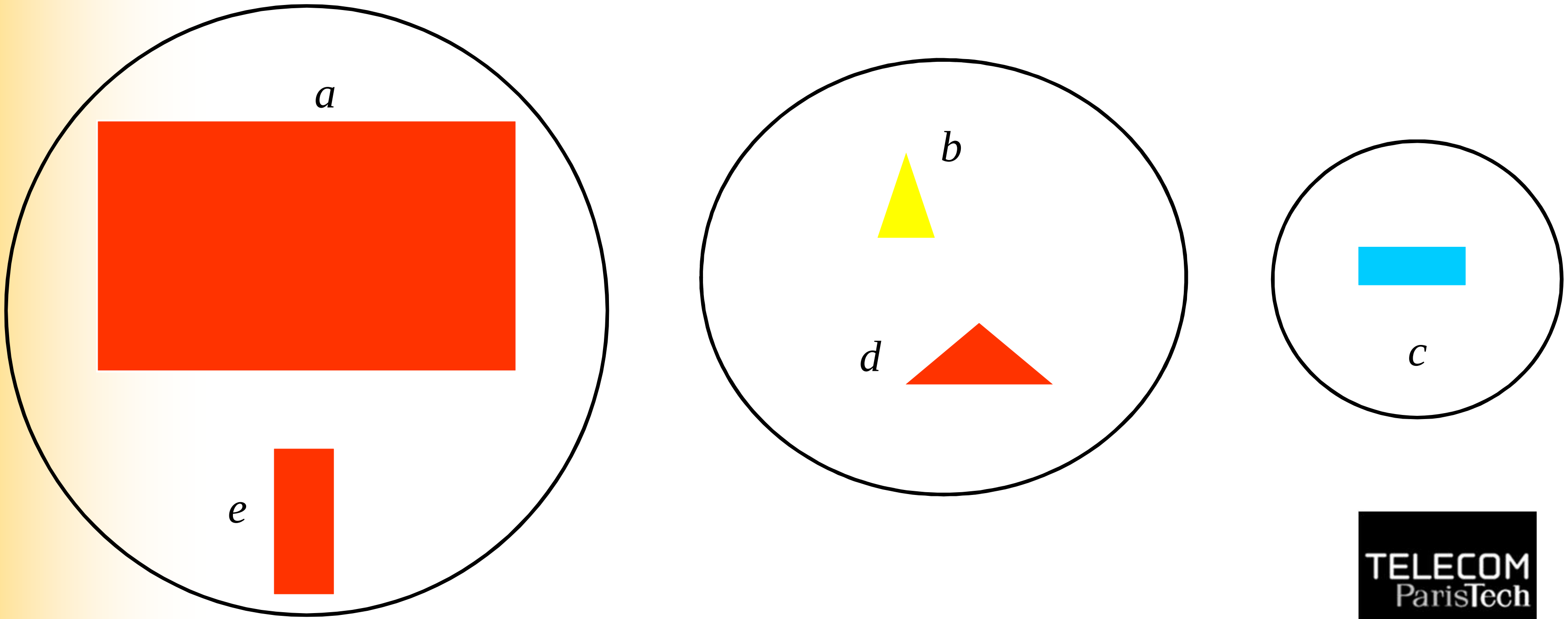
p_{xy}	**a**	**c**	**e**	**b**	**d**
a	**3**	**–1**	**1**	–3	–1
c	**–1**	**3**	**1**	–1	–1
e	**1**	**1**	**3**	–1	1
b	–3	–1	–1	**3**	**1**
d	–1	–1	1	**1**	**3**

- Then ***a e*** | ***b d*** | ***c*** or ***a c e*** | ***b d*** are median equivalence relations.

TELECOM ParisTech

Example

- $n = 5$, $X = \{a, b, c, d, e\}$, for ***a e* | *b d* | *c***
  (*a* is a large red rectangle; *b* is a small yellow triangle; *c* is a small blue rectangle; *d* is a small red triangle; *e* is a small red rectangle).



Building a profile from a matrix?

- **Theorem 1.** Let $P = (p_{xy})_{x,y}$ be a symmetric matrix with nonnegative or nonpositive integers p_{xy} such that:

1. all the p_{xy}'s have the same parity;
2. all the p_{xx}'s have the same value p and this value is positive;
3. p is large enough w.r.t. to the other entries p_{xy}.

   Then there exists a profile of equivalence relations with P as its majority matrix.

Rk. If the p_{xy} are bounded by a constant, p is about n^3.

Sketch of the proof for p even

Main steps:

1. For $x < y$, we build a profile Π^+_{xy} of two equivalence relations such that the entries of the majority matrix of Π^+_{xy} are equal to 0 except p_{xy}, p_{yx} and the diagonal entries p_{zz}, which are equal to 2.

2. For $x < y$, we build a profile Π^-_{xy} of $4n - 6$ equivalence relations such that the entries of the majority matrix of Π^-_{xy} are equal to 0 except p_{xy}, p_{yx}, which are equal to –2, and the diagonal entries p_{zz}, which are equal to $4n - 6$.

Sketch of the proof for p even

3. We obtain the profile Π associated to P as the concatenation, for all x and y with $x < y$, of $p_{xy}/2$ times Π^{+}_{xy} if p_{xy} is positive and of $|p_{xy}|/2$ times Π^{-}_{xy} if p_{xy} is negative.

The obtained profile Π contains

$$2\sum_{(x<y \text{ with } p_{xy}>0)} p_{xy}/2 + (4n-6)\sum_{(x<y \text{ with } p_{xy}<0)} |p_{xy}|/2$$

equivalence relations.

Example

p_{xy}	a	b	c	d	e
a	24	–2	0	0	4
b	–2	24	0	2	0
c	0	0	24	0	2
d	0	2	0	24	2
e	4	0	2	2	24

=

p_{xy}	a	b	c	d	e
a	10	0	0	0	4
b	0	10	0	2	0
c	0	0	10	0	2
d	0	2	0	10	2
e	4	0	2	2	10

+

p_{xy}	a	b	c	d	e
a	14	–2	0	0	0
b	–2	14	0	0	0
c	0	0	14	0	0
d	0	0	0	14	0
e	0	0	0	0	14

$$P = P^{+} + P^{-}$$

TELECOM ParisTech

Example

p_{xy}	a	b	c	d	e
a	10	0	0	0	4
b	0	10	0	2	0
c	0	0	10	0	2
d	0	2	0	10	2
e	4	0	2	2	10

P^+

$= 2 \times$

p_{xy}	a	b	c	d	e
a	2	0	0	0	2
b	0	2	0	0	0
c	0	0	2	0	0
d	0	0	0	2	0
e	2	0	0	0	2

$+$

p_{xy}	a	b	c	d	e
a	2	0	0	0	0
b	0	2	0	2	0
c	0	0	2	0	0
d	0	2	0	2	0
e	0	0	0	0	2

$+$

p_{xy}	a	b	c	d	e
a	2	0	0	0	0
b	0	2	0	0	0
c	0	0	2	0	2
d	0	0	0	2	0
e	0	0	2	0	2

$+$

p_{xy}	a	b	c	d	e
a	2	0	0	0	0
b	0	2	0	0	0
c	0	0	2	0	0
d	0	0	0	2	2
e	0	0	0	2	2

Example

p_{xy}	a	b	c	d	e
a	2	0	0	0	2
b	0	2	0	0	0
c	0	0	2	0	0
d	0	0	0	2	0
e	2	0	0	0	2

$\rightarrow$ $ae \mid b \mid c \mid d$
$abcde$

p_{xy}	a	b	c	d	e
a	2	0	0	0	0
b	0	2	0	2	0
c	0	0	2	0	0
d	0	2	0	2	0
e	0	0	0	0	2

$\rightarrow$ $bd \mid a \mid c \mid e$
$abcde$

p_{xy}	a	b	c	d	e
a	2	0	0	0	0
b	0	2	0	0	0
c	0	0	2	0	2
d	0	0	0	2	0
e	0	0	2	0	2

$\rightarrow$ $ce \mid a \mid b \mid d$
$abcde$

p_{xy}	a	b	c	d	e
a	2	0	0	0	0
b	0	2	0	0	0
c	0	0	2	0	0
d	0	0	0	2	2
e	0	0	0	2	2

$\rightarrow$ $de \mid a \mid b \mid c$
$abcde$

TELECOM ParisTech

Example

p_{xy}	a	b	c	d	e
a	10	0	0	0	4
b	0	10	0	2	0
c	0	0	10	0	2
d	0	2	0	10	2
e	4	0	2	2	10

P^+

$\rightarrow$

$ae \mid b \mid c \mid d$
$ae \mid b \mid c \mid d$
$bd \mid a \mid c \mid e$
$ce \mid a \mid b \mid d$
$de \mid a \mid b \mid c$
$abcde$
$abcde$
$abcde$
$abcde$
$abcde$

Π^+

TELECOM ParisTech

Example

p_{xy}	a	b	c	d	e
a	14	–2	0	0	0
b	–2	14	0	0	0
c	0	0	14	0	0
d	0	0	0	14	0
e	0	0	0	0	14

P^{-}

$\rightarrow$

$a \mid bcde$
$b \mid acde$
$ab \mid c \mid d \mid e$
$ac \mid b \mid d \mid e$
$ad \mid b \mid c \mid e$
$ae \mid b \mid c \mid d$
$bc \mid a \mid d \mid e$
$bd \mid a \mid c \mid e$
$be \mid a \mid c \mid d$
$abcde \times 5$

Π^{-}

TELECOM ParisTech

Example

p_{xy}	a	b	c	d	e
a	24	–2	0	0	4
b	–2	24	0	2	0
c	0	0	24	0	2
d	0	2	0	24	2
e	4	0	2	2	24

$$P = P^+ + P^-$$

$\rightarrow$

$ae \mid b \mid c \mid d$
$ae \mid b \mid c \mid d$
$bd \mid a \mid c \mid e$
$ce \mid a \mid b \mid d$
$de \mid a \mid b \mid c$
$abcde \times 5$

$a \mid bcde$
$b \mid acde$
$ab \mid c \mid d \mid e$
$ac \mid b \mid d \mid e$
$ad \mid b \mid c \mid e$
$ae \mid b \mid c \mid d$
$bc \mid a \mid d \mid e$
$bd \mid a \mid c \mid e$
$be \mid a \mid c \mid d$
$abcde \times 5$

$$\Pi = \Pi^+ \cup \Pi^-$$

TELECOM ParisTech

Complexity of Régnier's problem

- Régnier's problem: given a profile Π of p equivalence relations, compute a median equivalence relation, i.e. an equivalence relation E minimizing $\rho_\Pi(E)$.
- Zahn's problem (1964): given a symmetric relation S, compute an equivalence relation E minimizing $\delta(S, E)$.
- **Theorem 2** (M. Krivanek, J. Moravek, 1986): Zahn's problem is NP-hard.

TELECOM ParisTech

Complexity of Régnier's problem

- **Theorem 3**: Régnier's problem is NP-hard.
- Sketch of the proof.

We transform Zahn's problem into Régnier's problem. For this, consider a symmetric relation S defined on X. Associate the majority matrix P with S: the entry p_{xy} is equal to 1 if x and y are in relation by S, or to –1 otherwise.

Complexity of Régnier's problem

- Example:

aSe

bSd

cSe

dSe

p_{xy}	*a*	*b*	*c*	*d*	*e*
a	1	–1	–1	–1	1
b	–1	1	–1	1	–1
c	–1	–1	1	–1	1
d	–1	1	–1	1	1
e	1	–1	1	1	1

Complexity of Régnier's problem

- We obtain a matrix P fulfilling the statement of Theorem 1.
- So, by Theorem 1, there exists a profile Π of equivalence relations s.t., for any equivalence relation E, $\rho_{\Pi}(E)$ is minimum if and only if $\delta(S, E)$ is minimum.
- The transformation is polynomial since, here, all the entries of P are –1 or 1.

Two open problems

- **Problem 1**:

  Given a majority matrix P, is it possible to design a construction of a profile of equivalence relations requiring less equivalence relations?

- **Problem 2**:

  What is the complexity of Régnier's problem if the number p of equivalence relations of the profile is a (large enough) constant?

TELECOM ParisTech

Thank you for your attention!

☺