An Hausdorff distance between hyper-rectangles for clustering interval data

M. Chavent

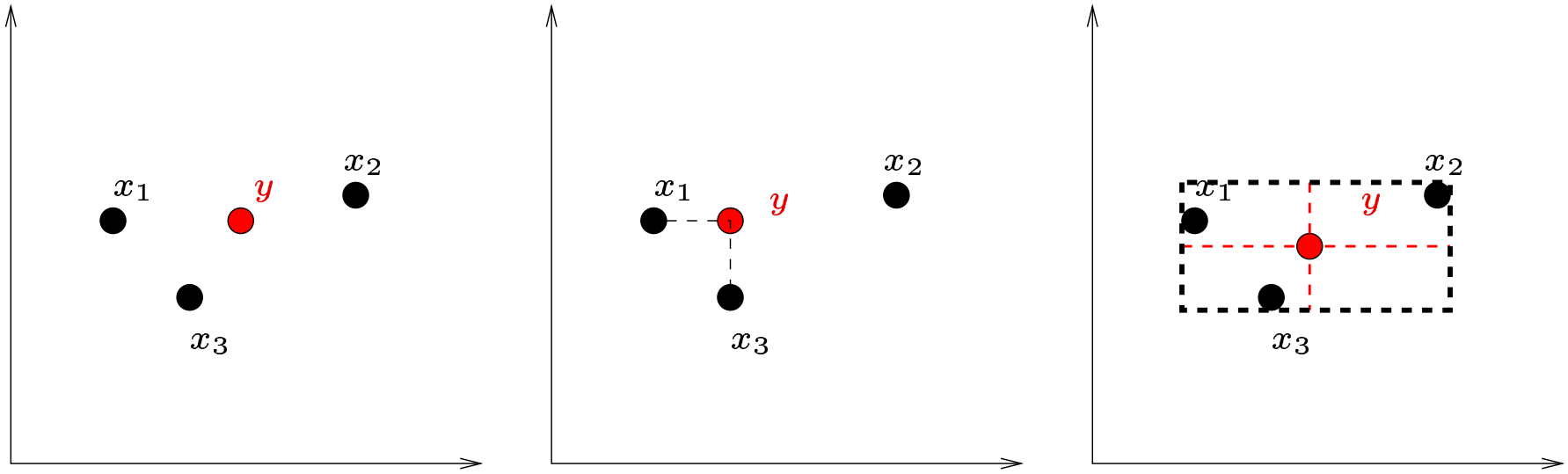
Laboratoire de Mathématiques Appliquées de Bordeaux, UMR CNRS 5466

Universités Bordeaux1 et 2, France

chavent@math.u-bordeaux1.fr

Introduction

Three different prototypes for points of $\mathbb{R}^p$ with L_1, L_2 or L_∞ distances :



$\sum_{j=1}^{p} {d_2}^2(x_i, y)$ $\sum_{j=1}^{p} d_1(x_i, y)$ $\max_{j=1\ldots p} d_\infty(x_i, y)$

PART 1

Interval data

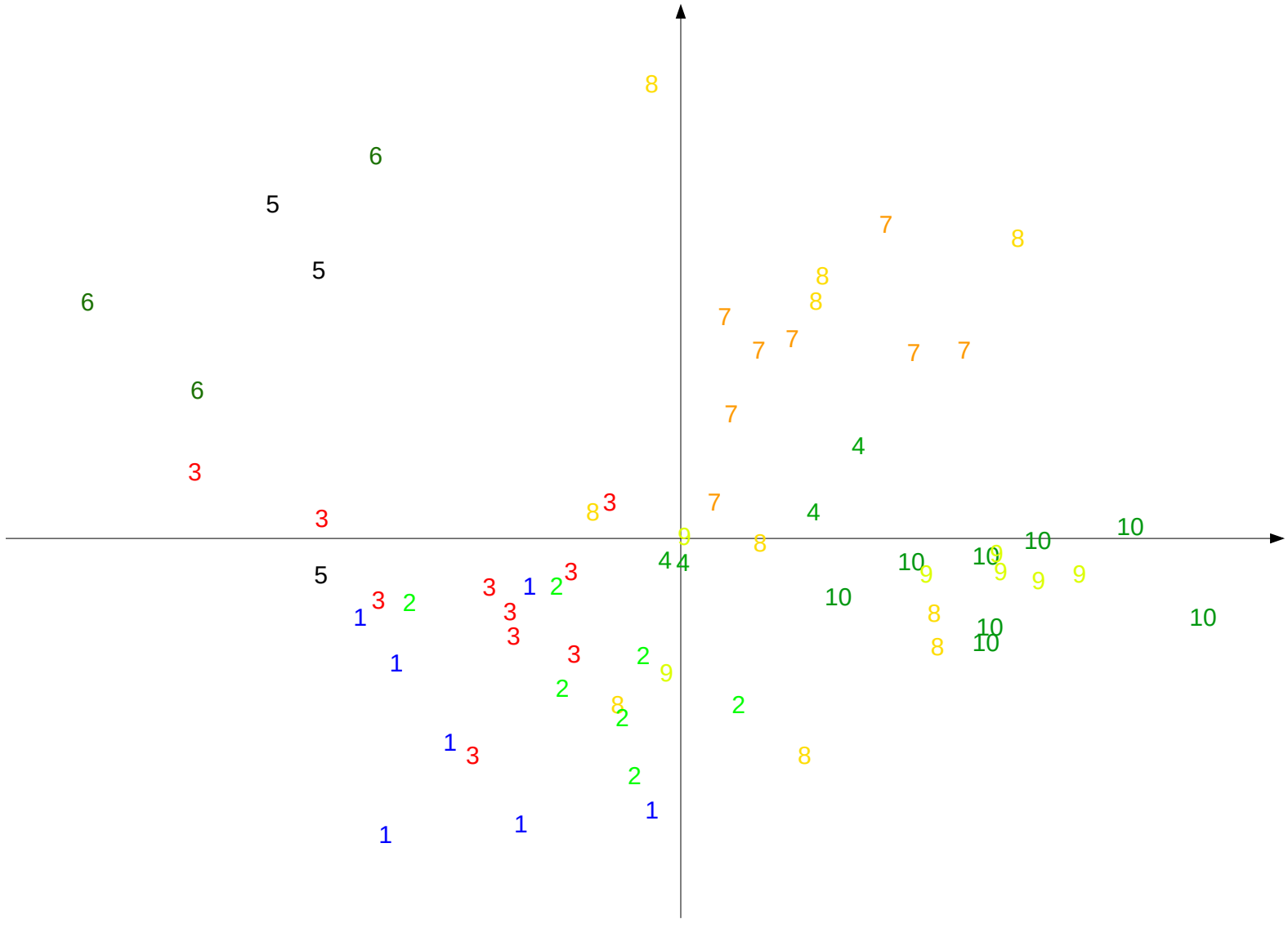
French Guyana Fish Example

- Mercury contamination in some Amerindian populations in French Guyana
- Data table : 67 fish of 10 different species described by 5 quantitative variables based on the mercury concentration (μg/g) in five organs (gills, liver, intestine, stomach, kidney)

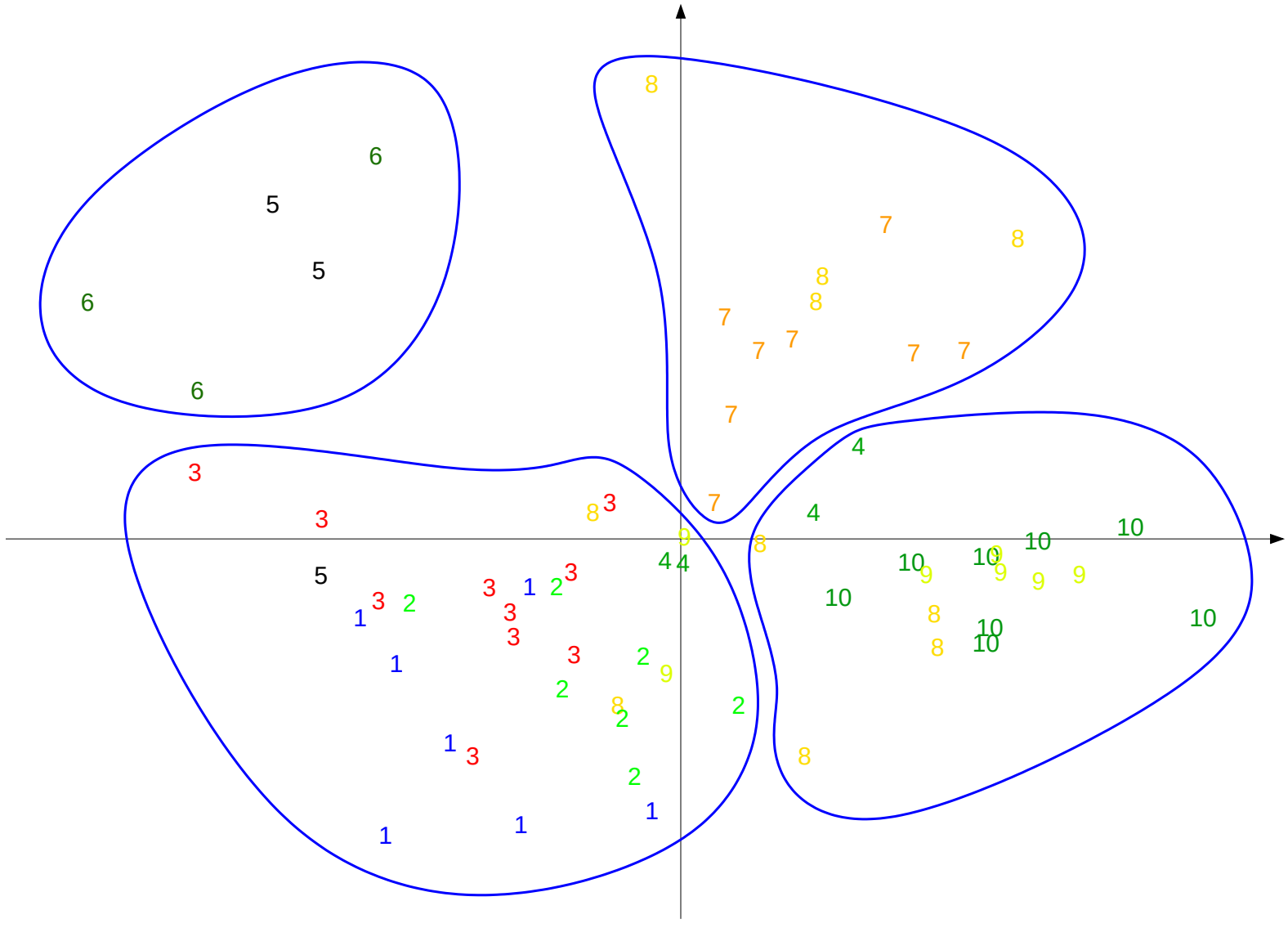
Fish	liver	kidney	gills	intestine	stomach	species
1	-0.116	0.352	-1.214	-1.147	NA	ageneiosus brevifi li
2	-0.083	-0.457	-1.881	-1.171	-1.485	ageneiosus brevifi li
...	...	...	...	...	...	...
8	1.416	0.684	-1.439	-1.554	-0.874	cynodon gibbus
9	0.115	-0.509	-1.910	NA	-1.610	cynodon gibbus
...	...	...	...	...	...	...
67	1.813	1.953	-2.251	0.390	-0.651	doras micopoeus

- How to cluster the 10 species

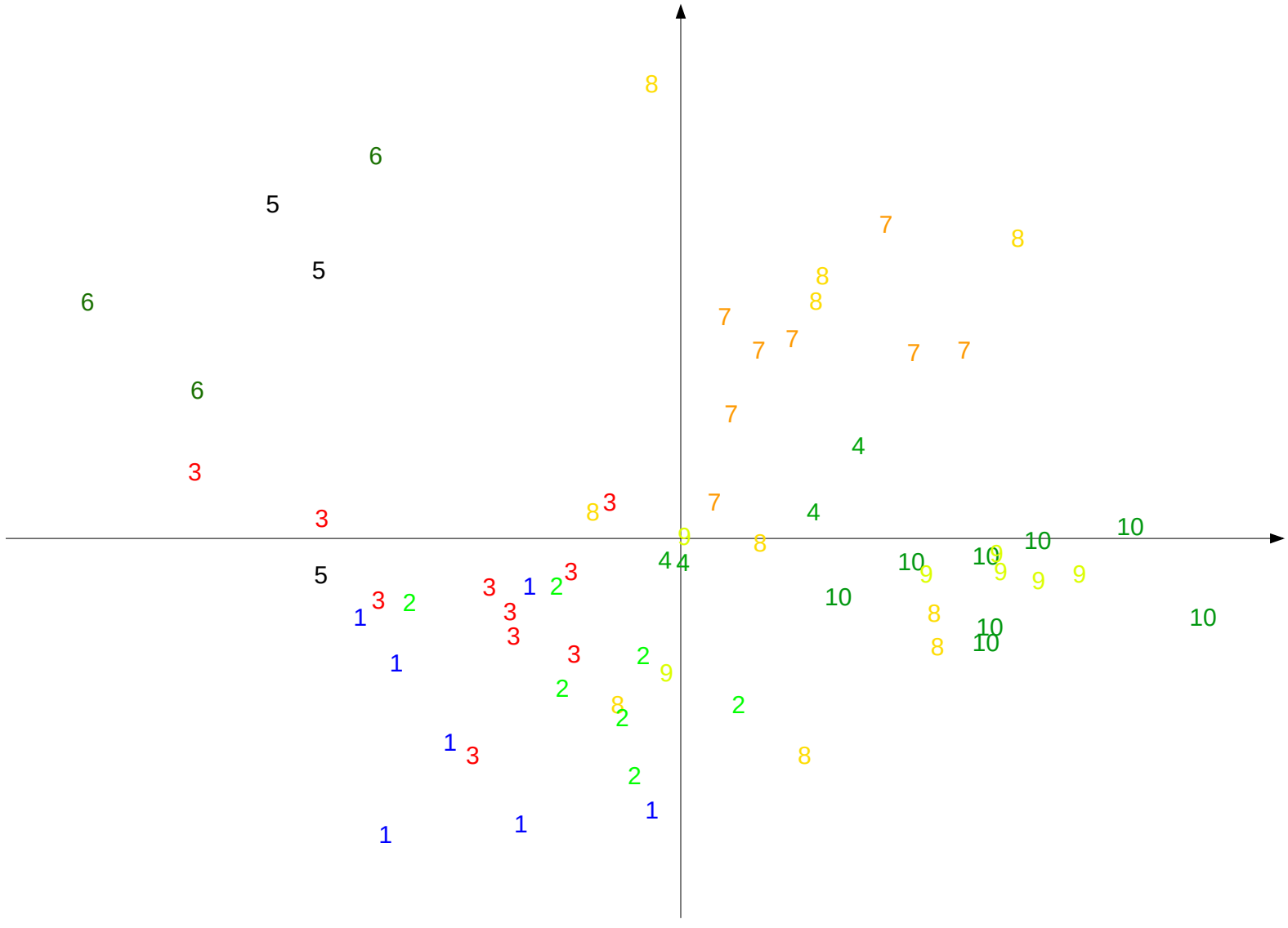
Classical or interval data representation ?



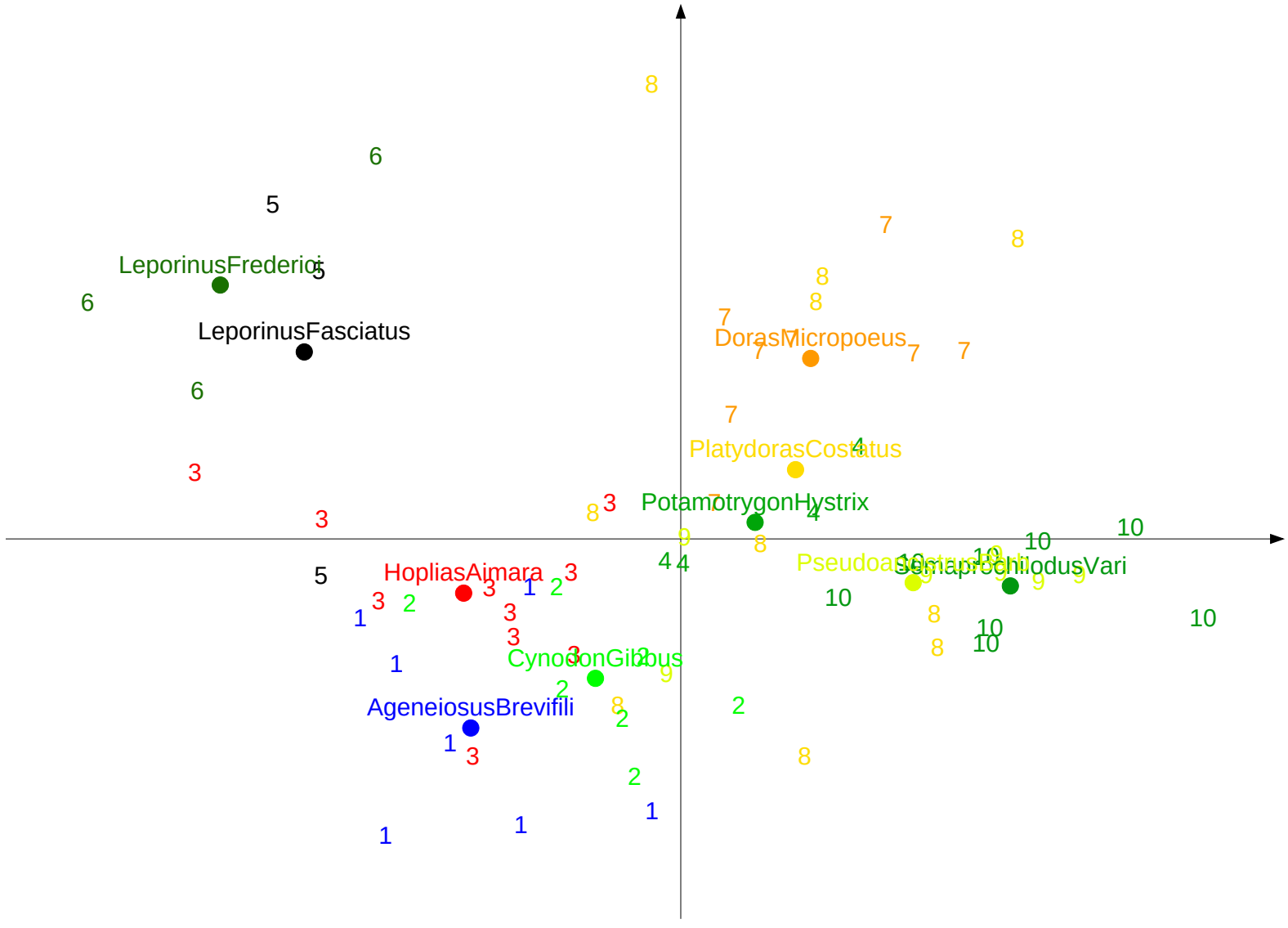
Classical or interval data representation ?



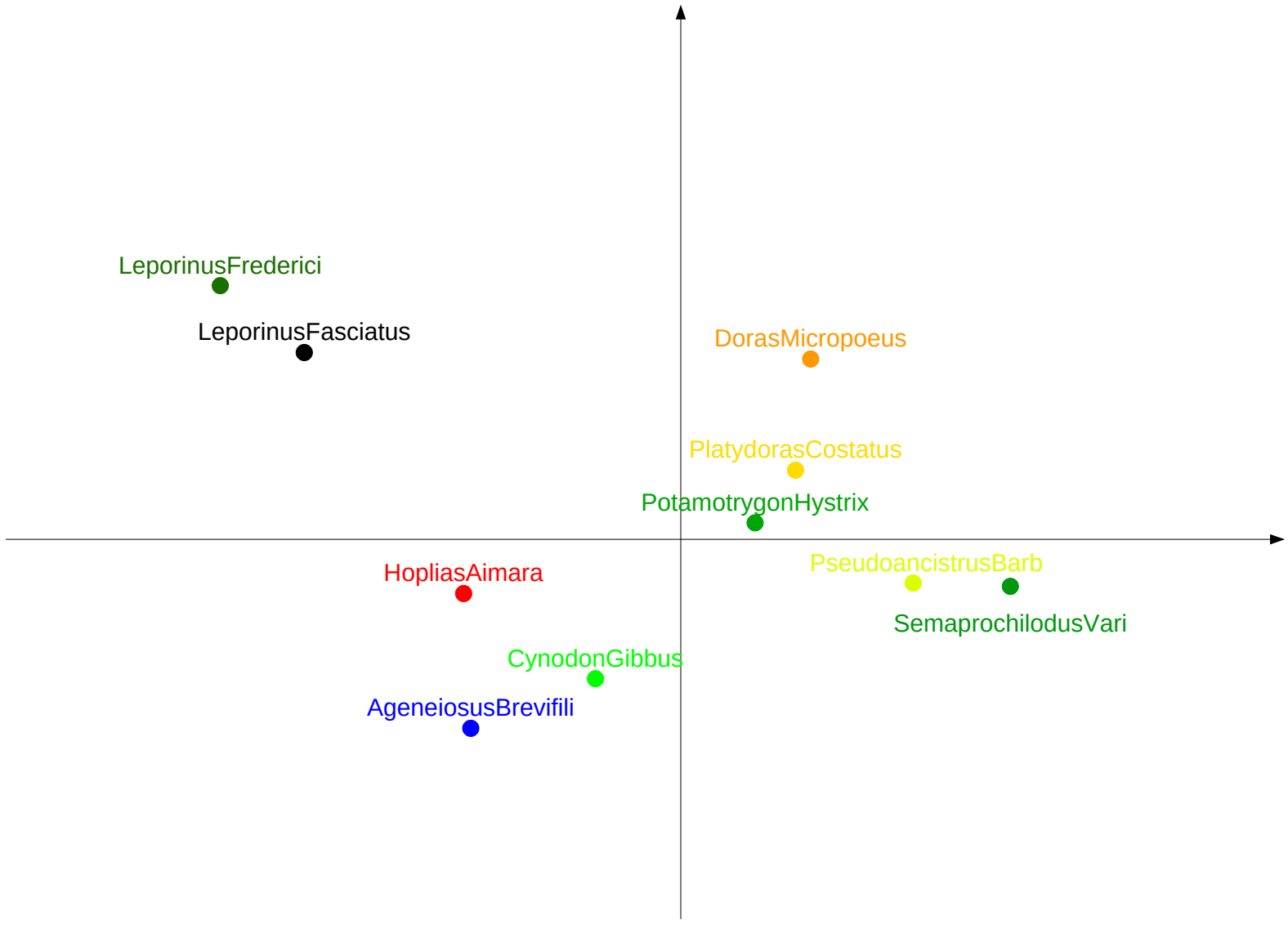
Classical or interval data representation ?



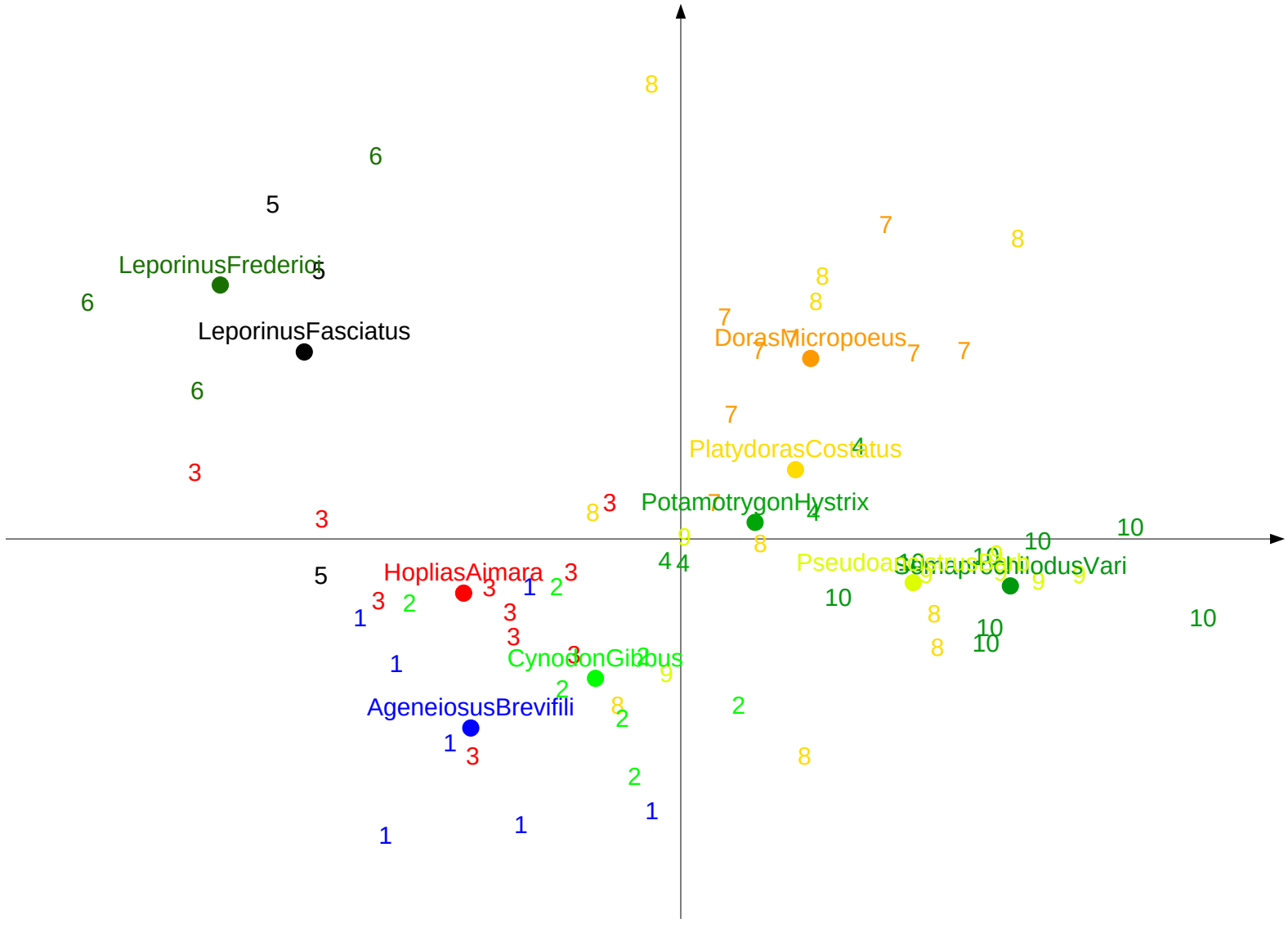
Classical or interval data representation ?



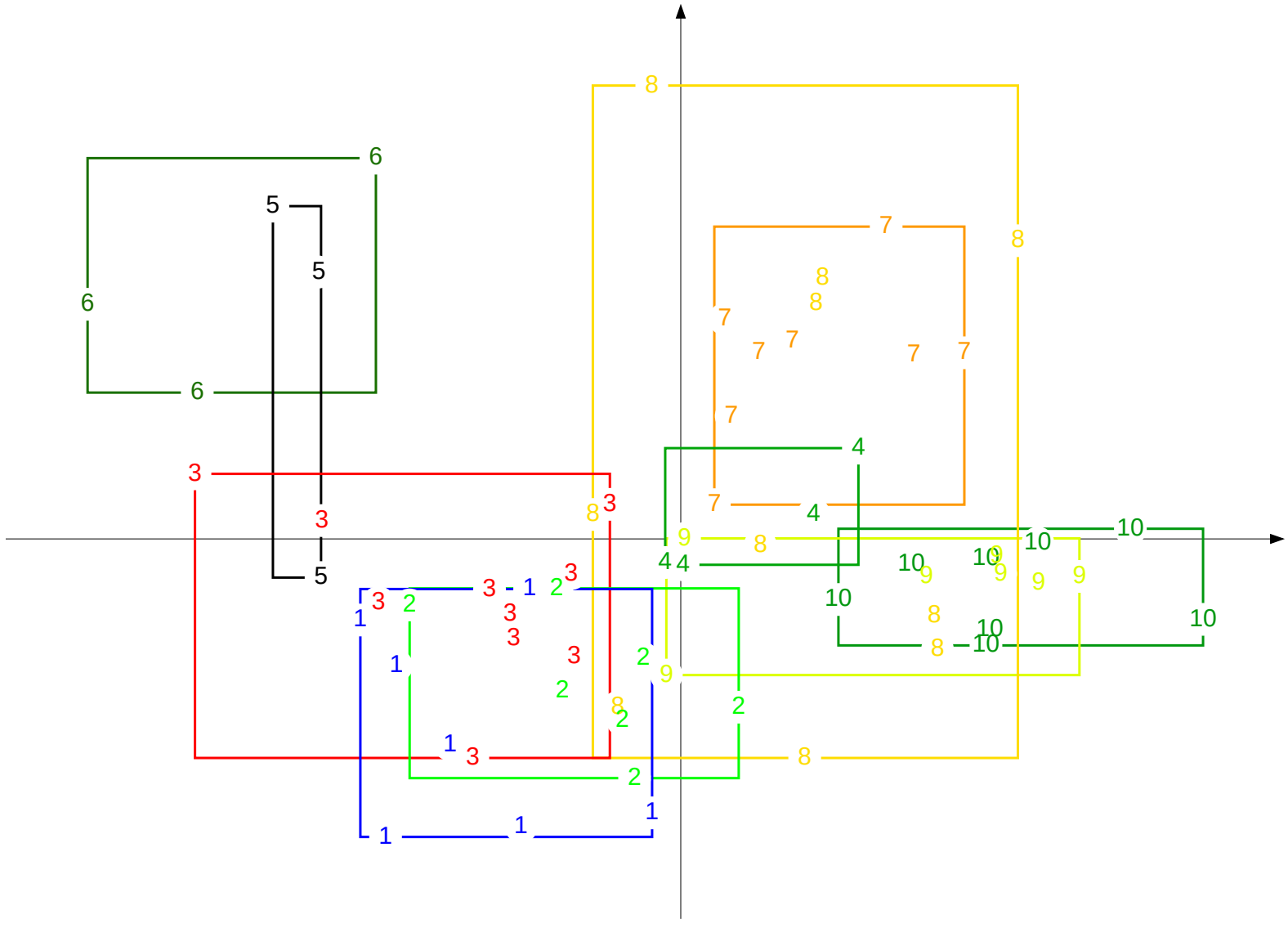
Classical or interval data representation ?



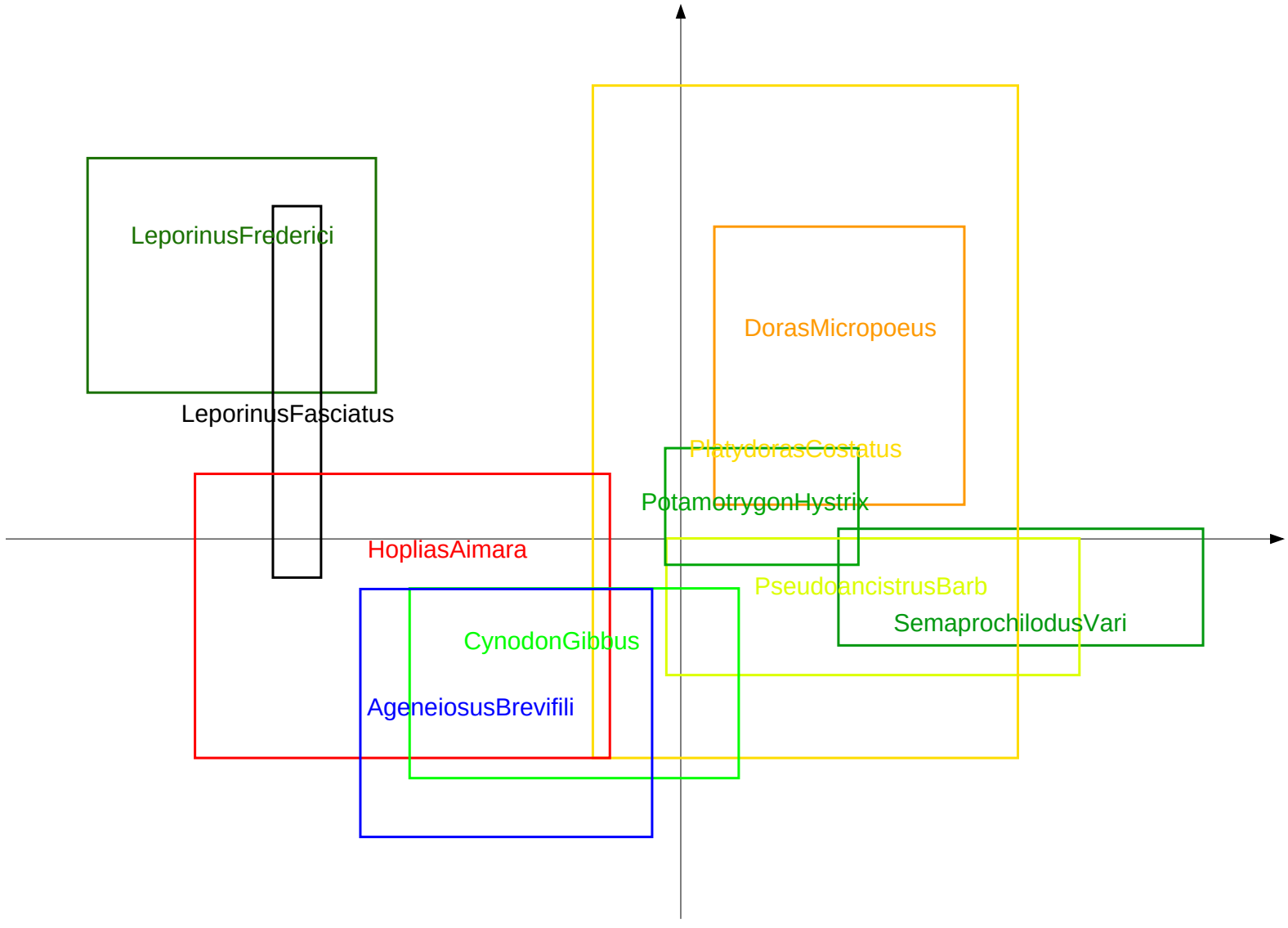
Classical or interval data representation ?



Classical or interval data representation ?



Classical or interval data representation ?



SODAS Software

- The data table obtained with DB2SO method of SODAS software

species	liver	kidney	gills	intestine	stomach
ageneiosus brevifi li	[-0.80:0.34]	[-1.50:0.35]	[-1.88:-1.21]	[-1.45:-0.48]	[-1.49:-1.05]
Cynodon Gibbus	[0.12:1.59]	[-0.51:1.18]	[-1.91:-1.44]	[-1.75:-0.68]	[-1.61:0.22]
Hoplias Aimara	[-0.44:0.90]	[-0.17:1.60]	[-1.98:-1.53]	[-2.17:-0.71]	[-2.36:-0.93]
Potamitrigon Hystrix	[0.66:2.01]	[0.77:2.15]	NA	[-0.50:0.23]	[-0.80:-0.69]
Leporinus Fasciatus	[-0.98:-0.58]	[-0.32:0.35]	[-3.00:-2.63]	NA	[-2.11:-2.76]
Leporinus Frederici	[-0.82:-0.04]	[-0.95:-0.19]	[-3.27:-2.55]	[-1.74:-1.42]	[-2.03:-0.55]
Doras Micropoeus	[1.34:2.12]	[1.47:2.69]	[-2.38:-2.21]	[-1.99:0.39]	[-1.45:-0.24]
Platidoras Costatus	[0.41:2.42]	[-0.02:2.75]	[-2.90:-1.27]	[-1.22:0.38]	[-1.41:-0.49]
Pseudoancistrus Barbatus	[1.26:2.84]	[-0.99:0.99]	NA	[-0.31:0.68]	[-0.71:0.12]
Semaprochilodus Vari	[2.70:3.96]	[1.11:1.91]	[-1.79:-1.40]	[-0.91:0.52]	[-0.74:0.22]

- Each of the $n = 10$ species i is an hyper-rectangle of $\mathbb{R}^p$ (here $p = 5$) noted:

$$x_i = \prod_{j=1}^{p} \underbrace{[a_i^j, b_i^j]}_{x_i^j}$$

PART 2

Comparing hyper-rectangles

Several approaches

- Simple Euclidean distance or more generally Minkowsky distance $\Rightarrow$ lower and upper bound are used independantly

species	Axis1	Axis2
AgeneiosusBrevifi li	[-1,957:-0,175]	[0,306:1,819]
CynodonGibbus	[-1,656:0,354]	[0,302:0,302]
DorasMicropoeus	[-2,967:-0,433]	[-0,397:1,337]
HopliasAimara	[-0,095:1,085]	[-0,554:0,158]
LeporinusFasciatus	[-2,491:-2,197]	[-2,031:0,236]
...	...	...

Axis1		Axis2	
-1,957	-0,175	0,306	1,819
-1,656	0,354	0,302	0,302
-2,967	-0,433	-0,397	1,337
-0,095	1,085	-0,554	0,158
-2,491	-2,197	-2,031	0,236
...	...	...	...

Several approaches

- Simple Euclidean distance or more generally Minkowsky distance $\Rightarrow$ lower and upper bound are used independantly

species	Axis1	Axis2
AgeneiosusBrevifi li	[-1,957:-0,175]	[0,306:1,819]
CynodonGibbus	[-1,656:0,354]	[0,302:0,302]
DorasMicropoeus	[-2,967:-0,433]	[-0,397:1,337]
HopliasAimara	[-0,095:1,085]	[-0,554:0,158]
LeporinusFasciatus	[-2,491:-2,197]	[-2,031:0,236]
...	...	...

Axis1		Axis2	
-1,957	-0,175	0,306	1,819
-1,656	0,354	0,302	0,302
-2,967	-0,433	-0,397	1,337
-0,095	1,085	-0,554	0,158
-2,491	-2,197	-2,031	0,236
...	...	...	...

- Elaborated distances taking into account both position and span of the intervals $\Rightarrow$ Explicit formulas for the optimum class prototype ?

The Hausdorff distance

The Hausdorff distance d_H between two sets $A, B \subset \mathbb{R}^p$ is :

$$d_H(A,B) = \max(h(A,B), h(B,A))$$

with

$$h(A,B) = \sup_{u \in A} \inf_{v \in B} d(u,v)$$

- Depends on the distance d chosen (L_1,L_2...)
- Here :

$$d_\infty(u,v) = \max_{j=1,\ldots,p} |u_j - v_j|$$

$\Rightarrow$ we are able to give an explicit formula of the optimum class prototype with $d_{H,\infty}$

Mathematical properties

Here A and B are two hyper-rectangles of $\mathbb{R}^p$ noted:

$$A = \prod_{j=1}^{p} A_j,\ B = \prod_{j=1}^{p} B_j$$

where $A_j = [a_j, b_j]$ and $B_j = [\alpha_j, \beta_j]$ are intervals of $\mathbb{R}$.

- *Property 1.* In the one dimensional case we can drop the ∞ subscript:

$$d_H(A_j, B_j) = \max(|a_j - \alpha_j|, |b_j - \beta_j|)$$

Mathematical properties

Here A and B are two hyper-rectangles of $\mathbb{R}^p$ noted:

$$A = \prod_{j=1}^{p} A_j,\ B = \prod_{j=1}^{p} B_j$$

where $A_j = [a_j, b_j]$ and $B_j = [\alpha_j, \beta_j]$ are intervals of $\mathbb{R}$.

- *Property 1.* In the one dimensional case we can drop the ∞ subscript:

$$d_H(A_j, B_j) = \max(|a_j - \alpha_j|, |b_j - \beta_j|)$$

- *Property 2.* With the L_∞ distance, we have the following relation between the Hausdorff distances d_H in p dimensions and in one dimension:

$$d_{H,\infty}(A, B) = \max_{j=1,\ldots,p} d_H(A_j, B_j)$$

Mathematical properties

Here A and B are two hyper-rectangles of $\mathbb{R}^p$ noted:

$$A = \prod_{j=1}^{p} A_j,\ B = \prod_{j=1}^{p} B_j$$

where $A_j = [a_j, b_j]$ and $B_j = [\alpha_j, \beta_j]$ are intervals of $\mathbb{R}$.

- *Property 1.* In the one dimensional case we can drop the ∞ subscript:

$$d_H(A_j, B_j) = \max(|a_j - \alpha_j|, |b_j - \beta_j|)$$

- *Property 2.* $\Rightarrow$ In the particular case of intervals reduced to single points, the L_∞ Hausdorff distance is the well-known L_∞ distance between $\mathbb{R}^p$ points

Example

num	species	Axis1	Axis2
1	AgeneiosusBrevifi li	[-1,957:-0,175]	[0,306:1,819]
2	CynodonGibbus	[-1,656:0,354]	[0,302:0,302]
3	DorasMicropoeus	[-2,967:-0,433]	[-0,397:1,337]
...	...	...	...

- Comparing species 1 and 2 in one dimension:

  $d_H(x_1^1, x_2^1) = \max(|-1,957+1,656|, |-0,175-0,354|) = 0,529$

  $d_H(x_1^2, x_2^2) = 1,517$

Example

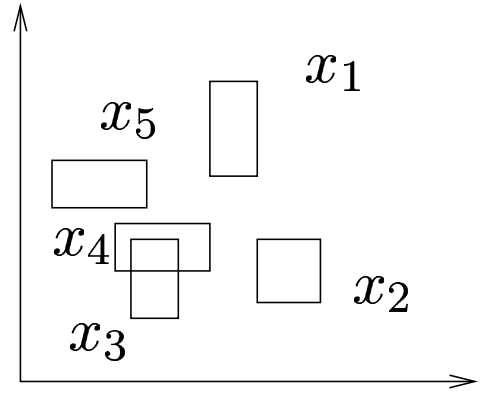
num	species	Axis1	Axis2
1	AgeneiosusBrevifi li	[1,957:-0,175]	[0,306:1,819]
2	CynodonGibbus	[-1,656:0,354]	[0,302:0,302]
3	DorasMicropoeus	[-2,967:-0,433]	[-0,397:1,337]
...	...	...	...

- Comparing species 1 and 2 in one dimension:
  $d_H(x_1^1, x_2^1) = \max(|-1,957+1,656|, |-0,175-0,354|) = 0,529$
  $d_H(x_1^2, x_2^2) = 1,517$
- Comparing species 1 and 2 in two dimensions:
  - with the L_∞ Hausdorff distance:
    $d_{H,\infty}(x_1, x_2) = \max(0,529, 1,517)$
  - with the distance used in (Chavent and Lechevallier, 02):
    $d(x_1, x_2) = 0,529+1,517$

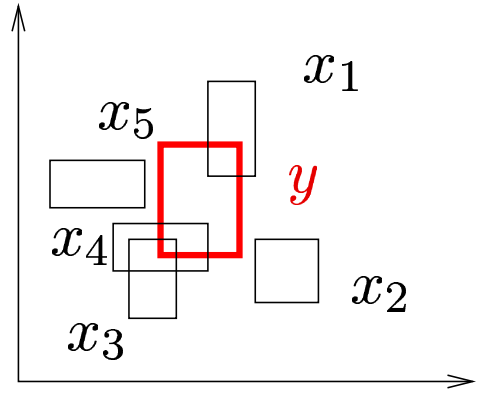
PART 3

Define a class prototype

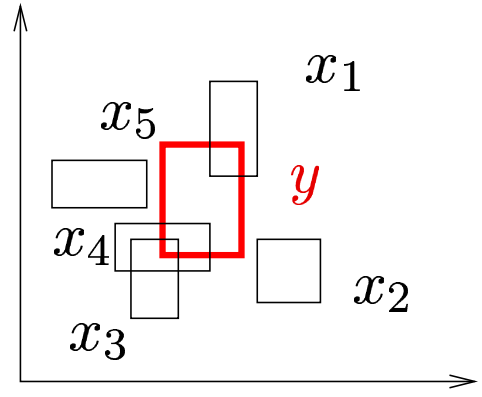
Define a class prototype



Define a class prototype

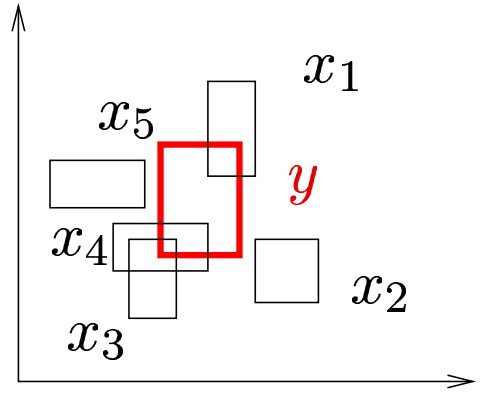


Define a class prototype



- Adequacy criterion f between y and $C = \{x_1, ..., x_5\}$
- A distance between hyper-rectangles y and x_i

Define a class prototype



- Adequacy criterion f between y and $C = \{x_1, ..., x_5\}$
- A distance between hyper-rectangles y and x_i

$\Rightarrow$ Find an explicit formula for the prototype y which optimizes f

IFCS Krakow 2002

- Distance between hyper-rectangles: $d(x_i, y) = \sum_{j=1}^{p} d_H(x_i^j, y^j)$

  $\Rightarrow$ Not an Hausdorff distance between $\mathbb{R}^p$-sets

IFCS Krakow 2002

- Distance between hyper-rectangles: $d(x_i, y) = \sum_{j=1}^{p} d_H(x_i^j, y^j)$

  $\Rightarrow$ Not an Hausdorff distance between $\mathbb{R}^p$-sets

- Adequacy criterion: $f(y) = \sum_{i \in C} d(x_i, y)$ ('The star')

IFCS Krakow 2002

- Distance between hyper-rectangles: $d(x_i, y) = \sum_{j=1}^{p} d_H(x_i^j, y^j)$

  $\Rightarrow$ Not an Hausdorff distance between $\mathbb{R}^p$-sets

- Adequacy criterion: $f(y) = \sum_{i \in C} d(x_i, y)$ ('The star')

- Explicit formula of the minimizer $\hat{y} = \prod_{j=1}^{p} [\hat{\mu}^j - \hat{\lambda}^j, \hat{\mu}^j + \hat{\lambda}^j]$:

$$\begin{aligned} \hat{\mu}^j &= median\{m_i^j \mid i \in C\} \\ \hat{\lambda}^j &= median\{l_i^j \mid i \in C\} \end{aligned}$$

  with m_i^j and l_i^j the midpoints and the half-lengths of the intervals x_i^j

IFCS Chicago 2004

- L_∞ Hausdorff distance: $d_{H,\infty}(x_i, y) = \max_{j=1...p} d_H(x_i^j, y^j)$

  $\Rightarrow$ "Real" Hausdorff distance between $\mathbb{R}^p$-sets

IFCS Chicago 2004

- L_∞ Hausdorff distance: $d_{H,\infty}(x_i, y) = \max_{j=1...p} d_H(x_i^j, y^j)$

  $\Rightarrow$ "Real" Hausdorff distance between $\mathbb{R}^p$-sets

- Adequacy criterion: $f(y) = \max_{i \in C} d_{H,\infty}(x_i, y)$ ('The radius')

IFCS Chicago 2004

- L_∞ Hausdorff distance: $d_{H,\infty}(x_i, y) = \max_{j=1\ldots p} d_H(x_i^j, y^j)$

  $\Rightarrow$ "Real" Hausdorff distance between $\mathbb{R}^p$-sets

- Adequacy criterion: $f(y) = \max_{i \in C} d_{H,\infty}(x_i, y)$ ('The radius')

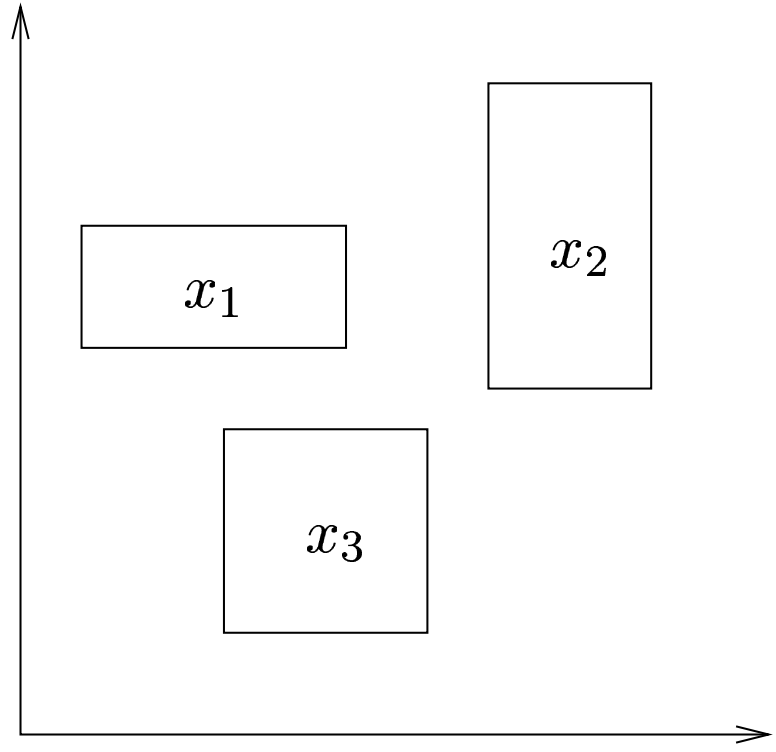
- Explicit formula of *a* minimizer $\hat{y} = \prod_{j=1}^{p} [\hat{\alpha}^j, \hat{\beta}^j]$:

$$\hat{\alpha}^j = \frac{\max_{i \in C} a_i^j + \min_{i \in C} a_i^j}{2}$$

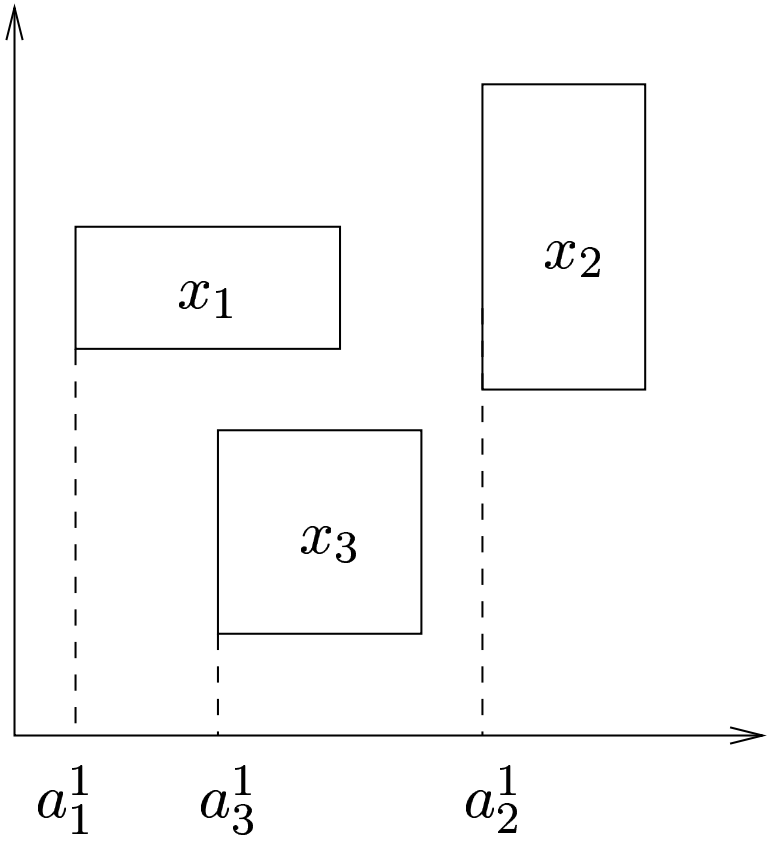
$$\hat{\beta}^j = \frac{\max_{i \in C} b_i^j + \min_{i \in C} b_i^j}{2}$$

with a_i^j and b_i^j the lower and upper bounds of the intervals x_i^j

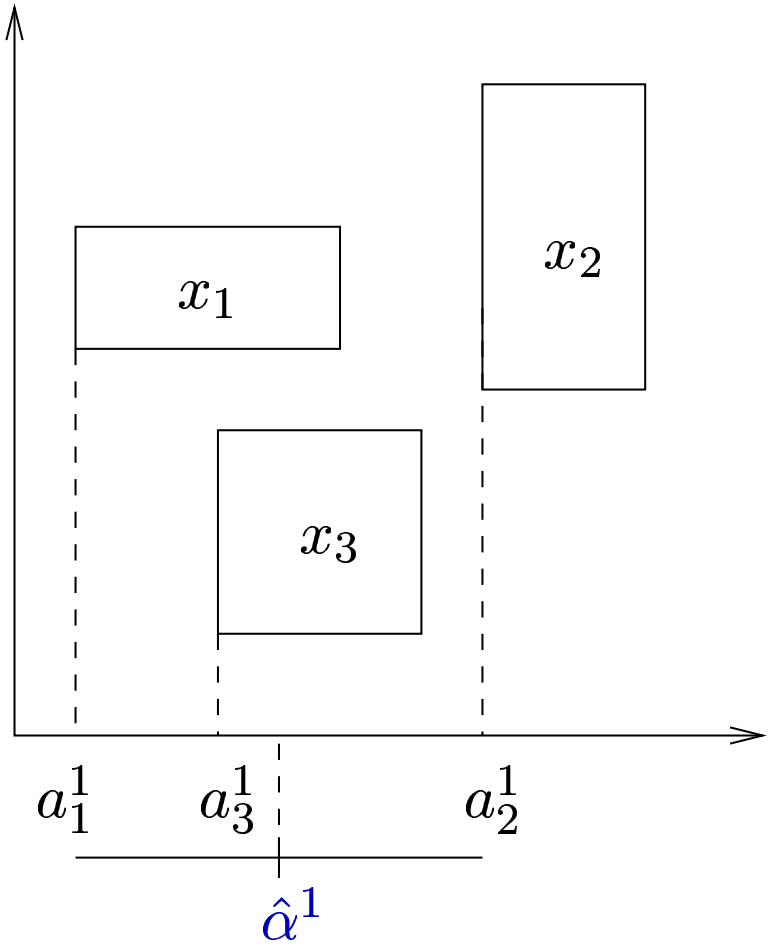
Construction of a prototype



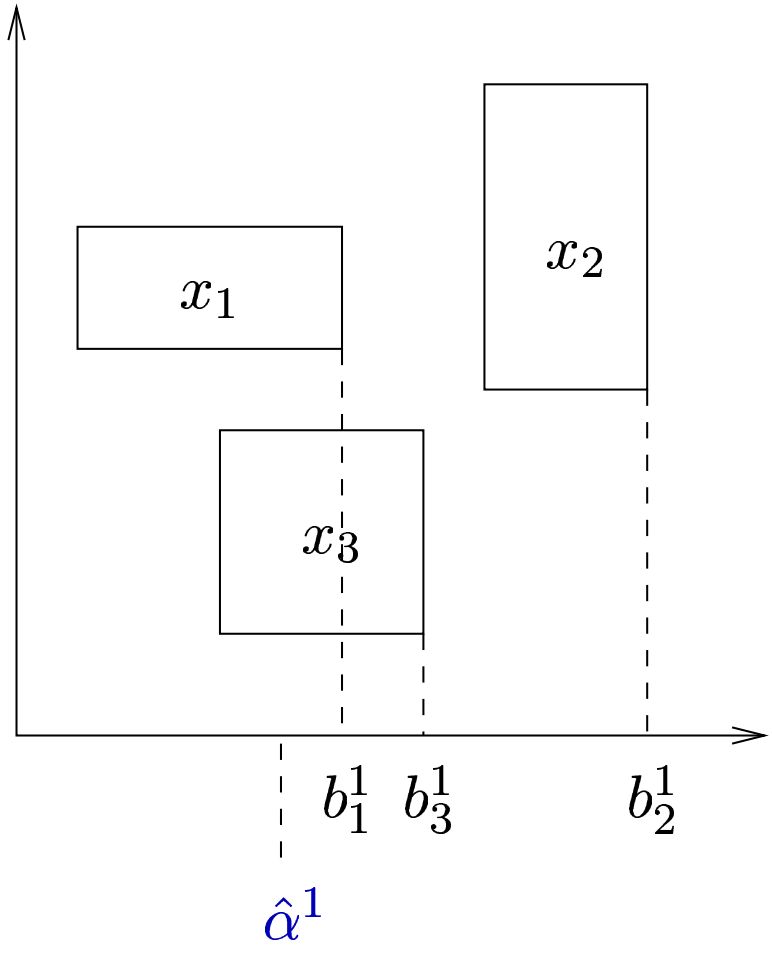
Construction of a prototype



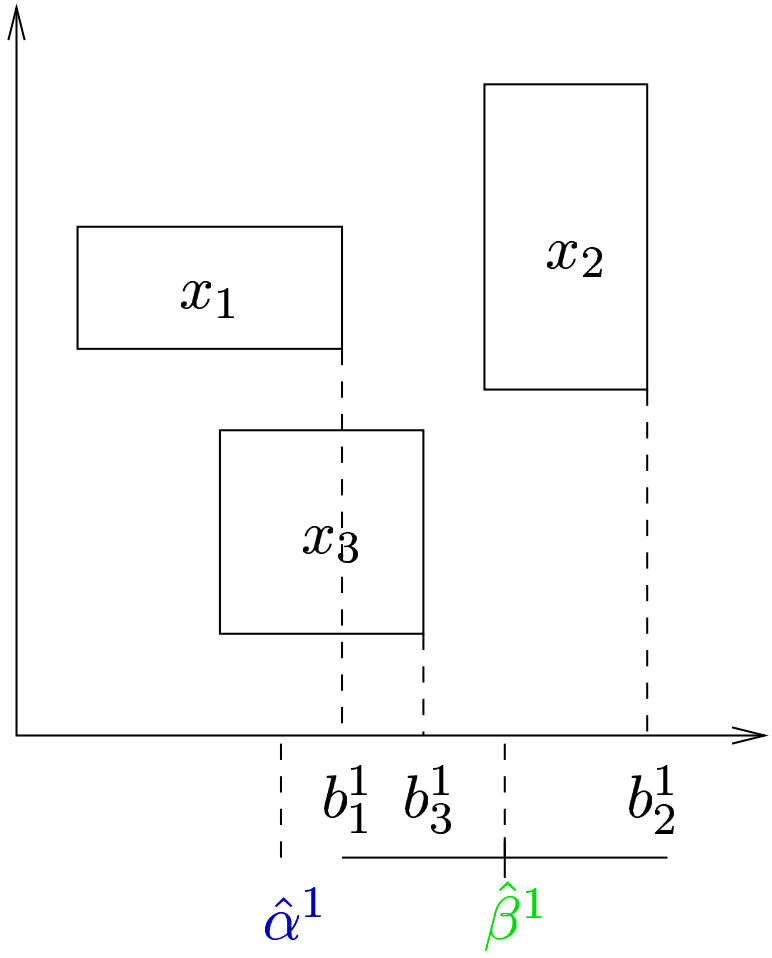
Construction of a prototype



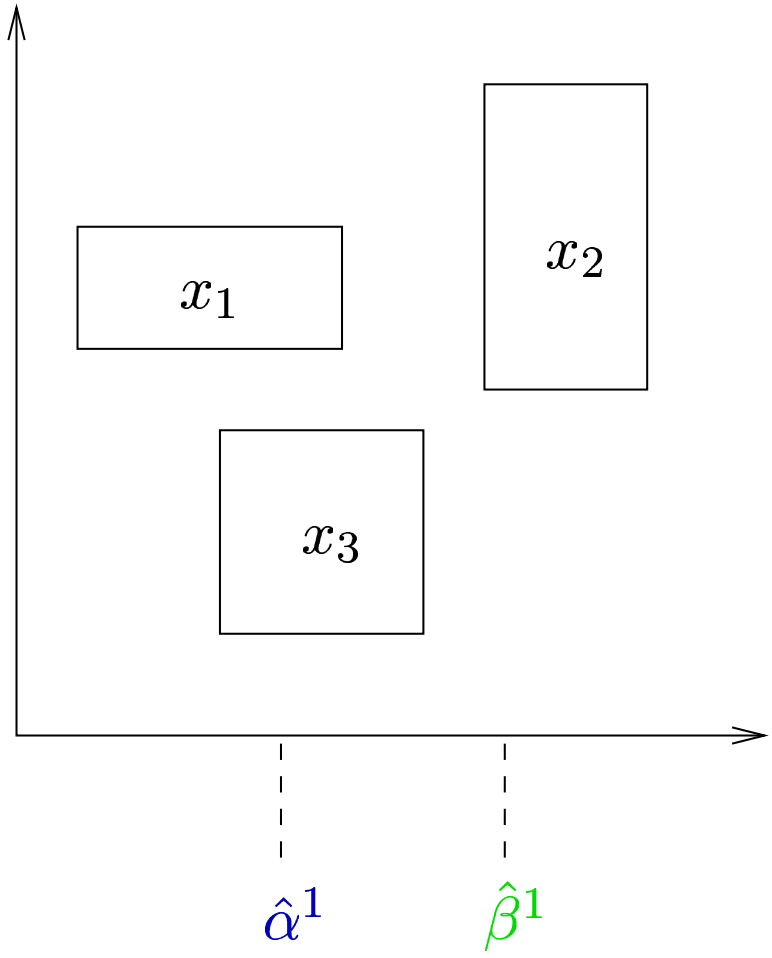
Construction of a prototype



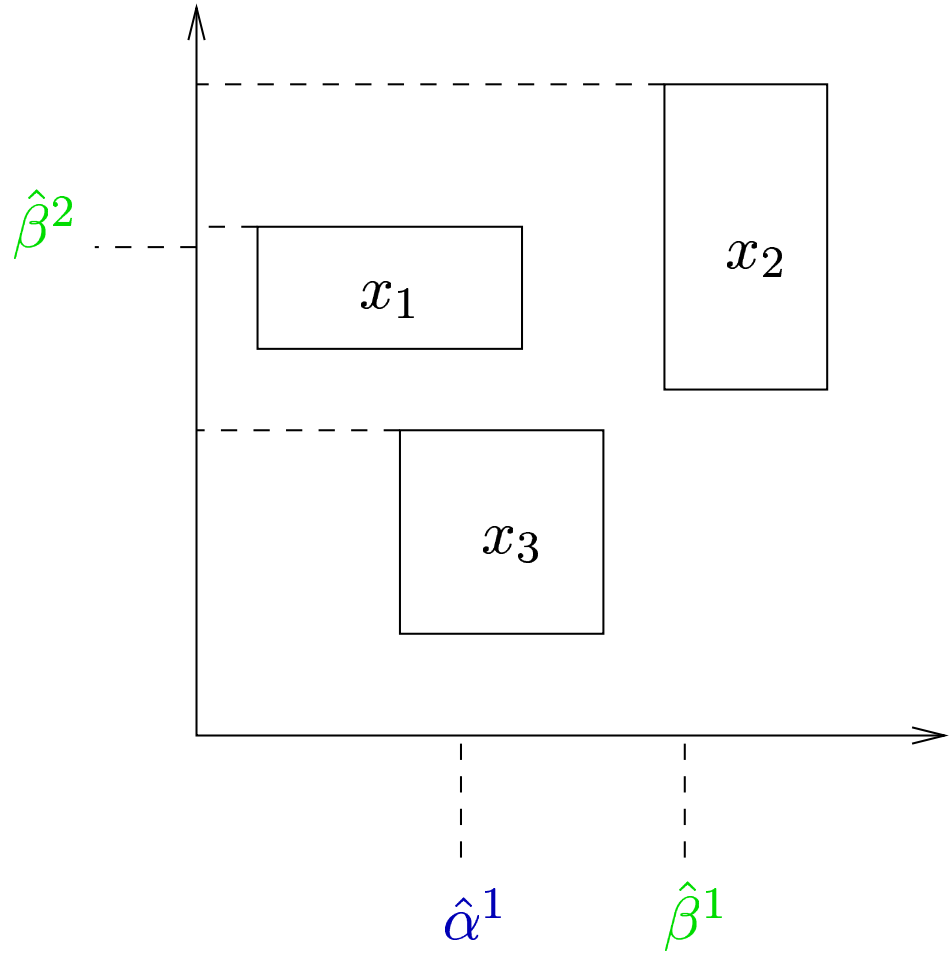
Construction of a prototype



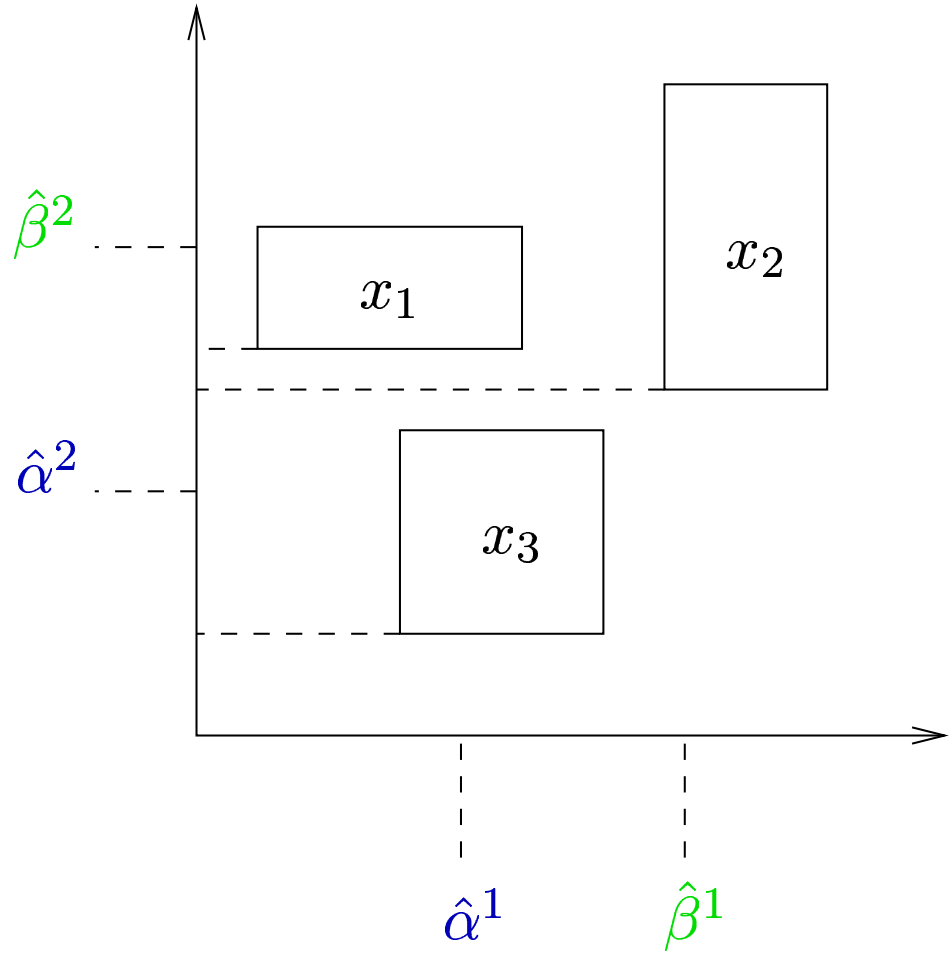
Construction of a prototype



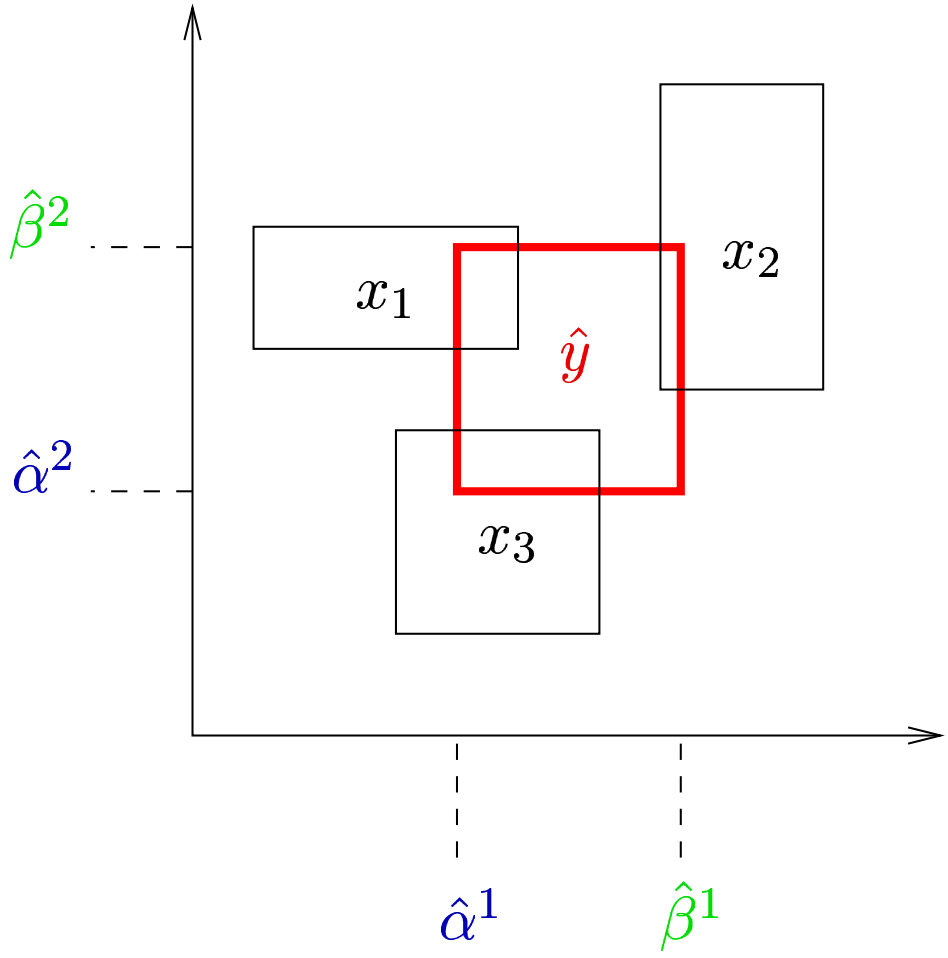
Construction of a prototype



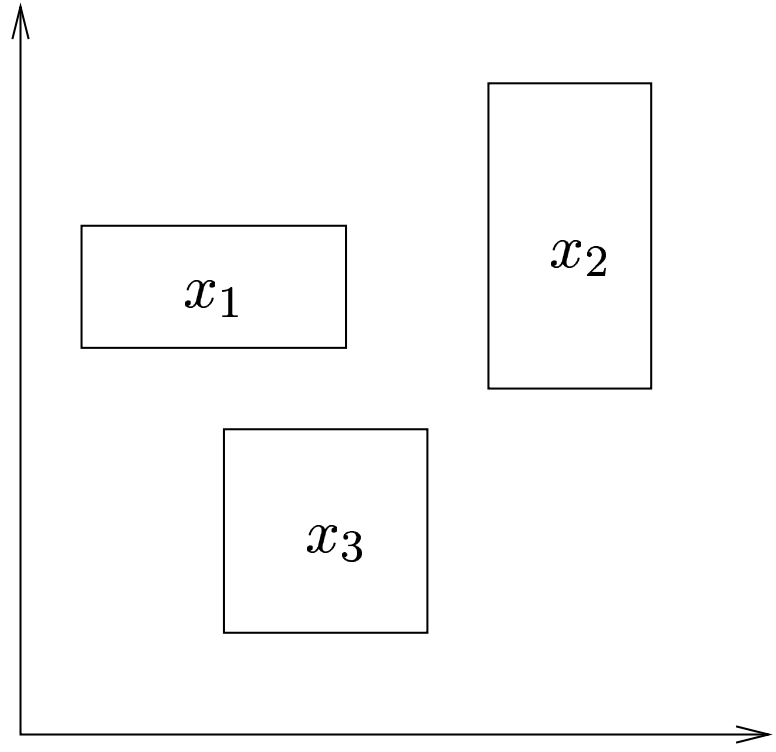
Construction of a prototype



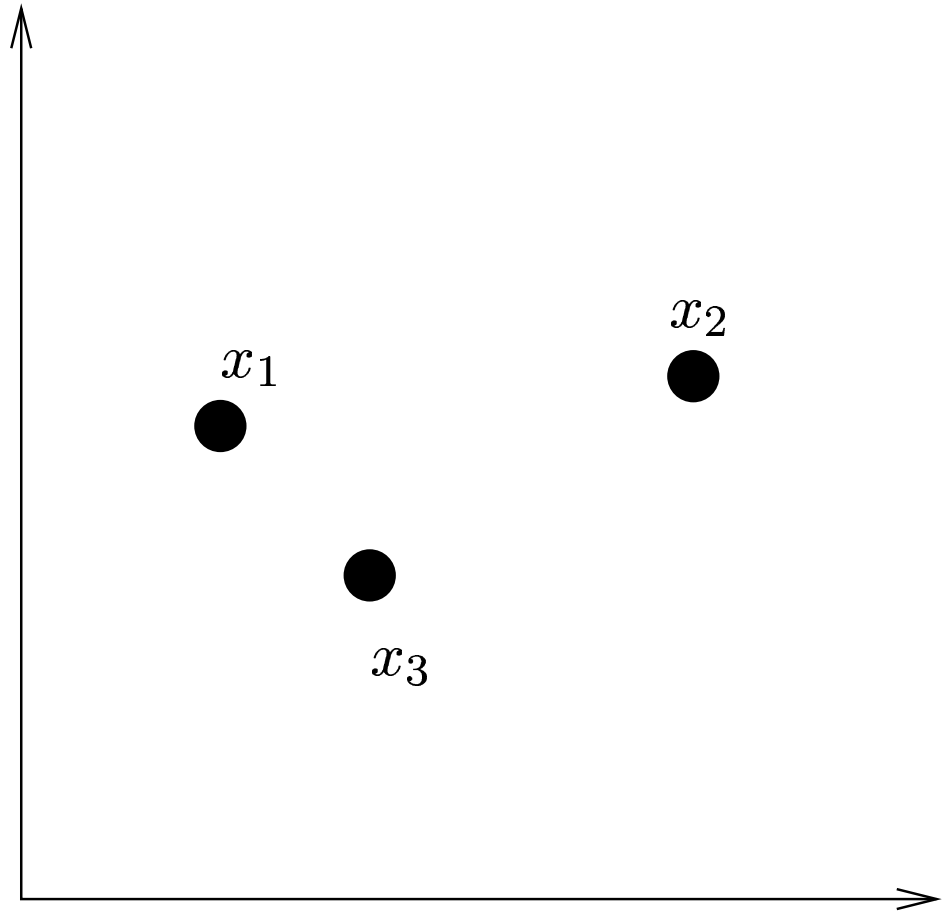
Construction of a prototype



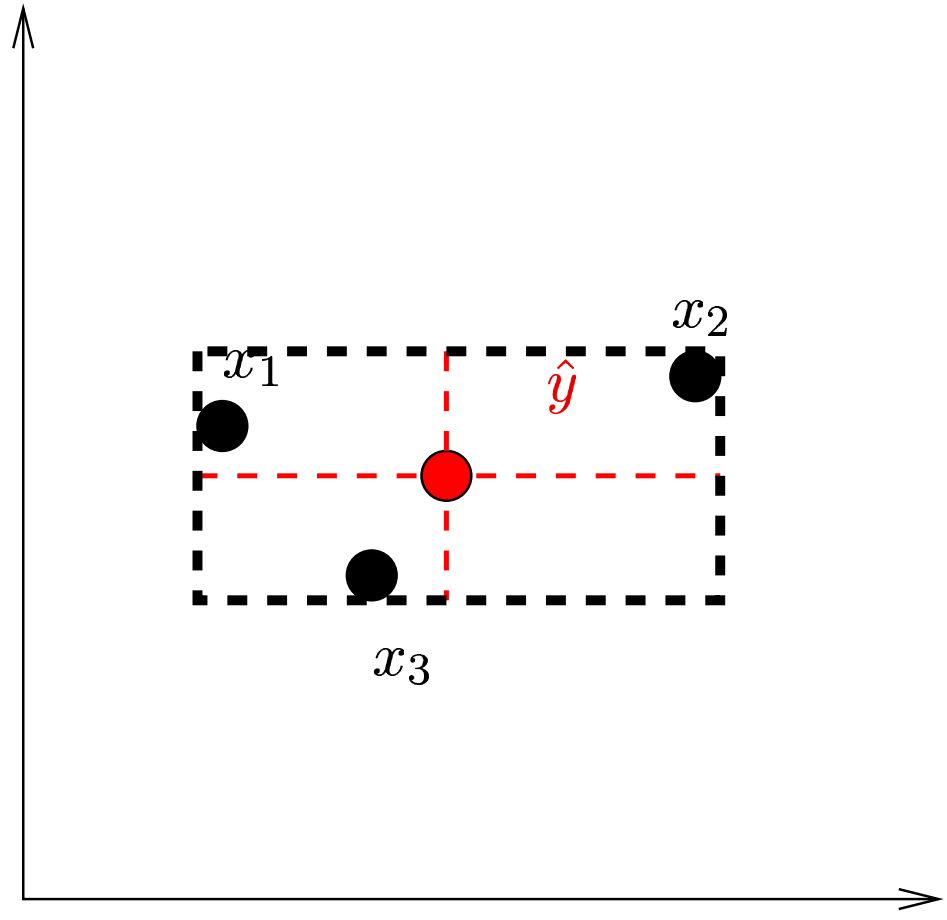
Construction of a prototype



Construction of a prototype



Construction of a prototype



Opimization problem

- The p-dimensional optimization problem is to find an hyper-rectangle y which minimizes:

$$\begin{aligned} f(y) &= \max_{i \in C} d_{H,\infty}(x_i, y) \\ &= \max_{j=1,\ldots,p} \underbrace{\max_{i \in C} d_H(x_i^j, y^j)}_{\tilde{f}^j(y^j)} \end{aligned}$$

- The one-dimensional optimization problem:
  $\Rightarrow$ find an interval $y^j = [\alpha^j, \beta^j]$ which minimizes:

$$\begin{aligned} \tilde{f}^j(y^j) &= \max_{i \in C} d_H(x_i^j, y^j) \\ &= max\{\max_{i \in C} |a_i^j - \alpha^j|, \max_{i \in C} |b_i^j - \beta^j|\} \end{aligned}$$

Opimization problem

- The p-dimensional optimization problem is to find an hyper-rectangle y which minimizes:

$$\begin{aligned} f(y) &= \max_{i \in C} d_{H,\infty}(x_i, y) \\ &= \max_{j=1,\ldots,p} \underbrace{\max_{i \in C} d_H(x_i^j, y^j)}_{\tilde{f}^j(y^j)} \end{aligned}$$

- The one-dimensional optimization problem:
  $\Rightarrow$ equivalent to:

$$\min_{\alpha^j \in \mathbb{R}} \max_{i \in C} |a_i^j - \alpha^j|$$

$$\min_{\beta^j \in \mathbb{R}} \max_{i \in C} |b_i^j - \beta^j|$$

Opimization problem

- The p-dimensional optimization problem is to find an hyper-rectangle y which minimizes:

$$\begin{aligned} f(y) &= \max_{i \in C} d_{H,\infty}(x_i, y) \\ &= \max_{j=1,\ldots,p} \underbrace{\max_{i \in C} d_H(x_i^j, y^j)}_{\tilde{f}^j(y^j)} \end{aligned}$$

- The one-dimensional optimization problem:
  $\Rightarrow$ has explicit solutions:

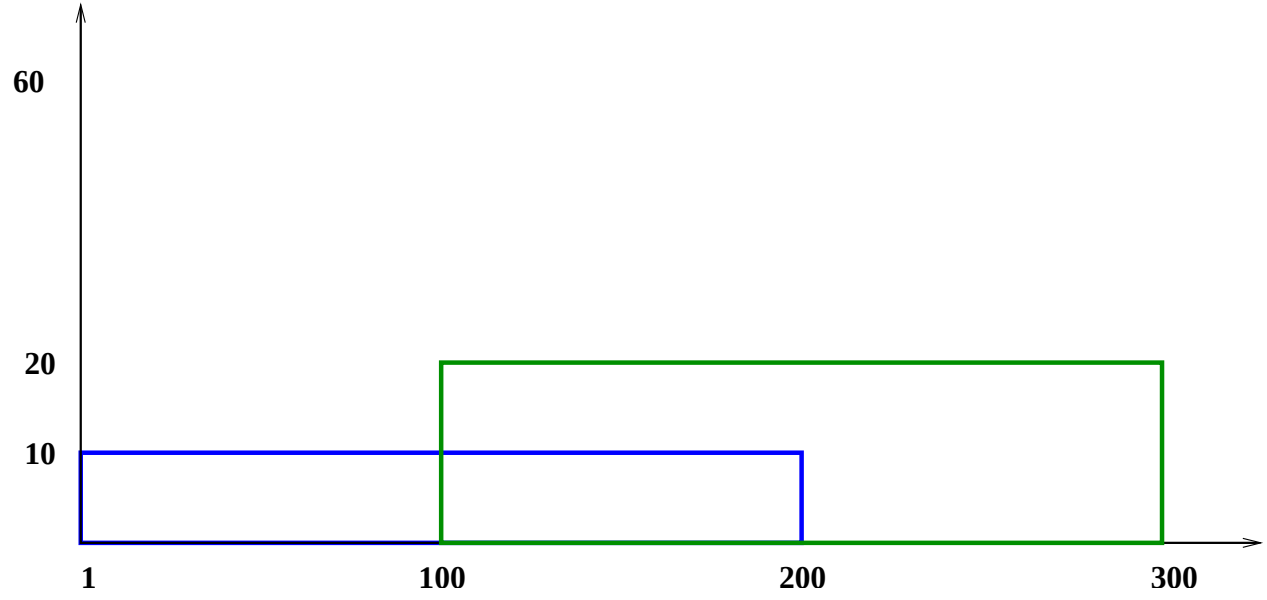
$$\begin{aligned} \hat{\alpha}^j &= \frac{\max_{i \in C} a_i^j + \min_{i \in C} a_i^j}{2} \\ \hat{\beta}^j &= \frac{\max_{i \in C} b_i^j + \min_{i \in C} b_i^j}{2} \end{aligned}$$

Non uniqueness of the minimizer

- if $\hat{y}^j$ is the minimizer of $\tilde{f}^j$ for $j = 1, ..., p$
  $\Rightarrow f(\hat{y}) = \max_{j=1...p} \tilde{f}^j(\hat{y}^j)$ is a minimum of f
  $\Rightarrow \hat{y} = \prod_{j=1}^{p} \hat{y}^j$ is a minimizer of f
- for all indexes j such that $\tilde{f}^j(\hat{y}^j) < f(\hat{y})$, all intervals $\tilde{y}^j$ such that $\tilde{f}^j(\tilde{y}^j) \leq f(\hat{y})$ produce also optimal solutions

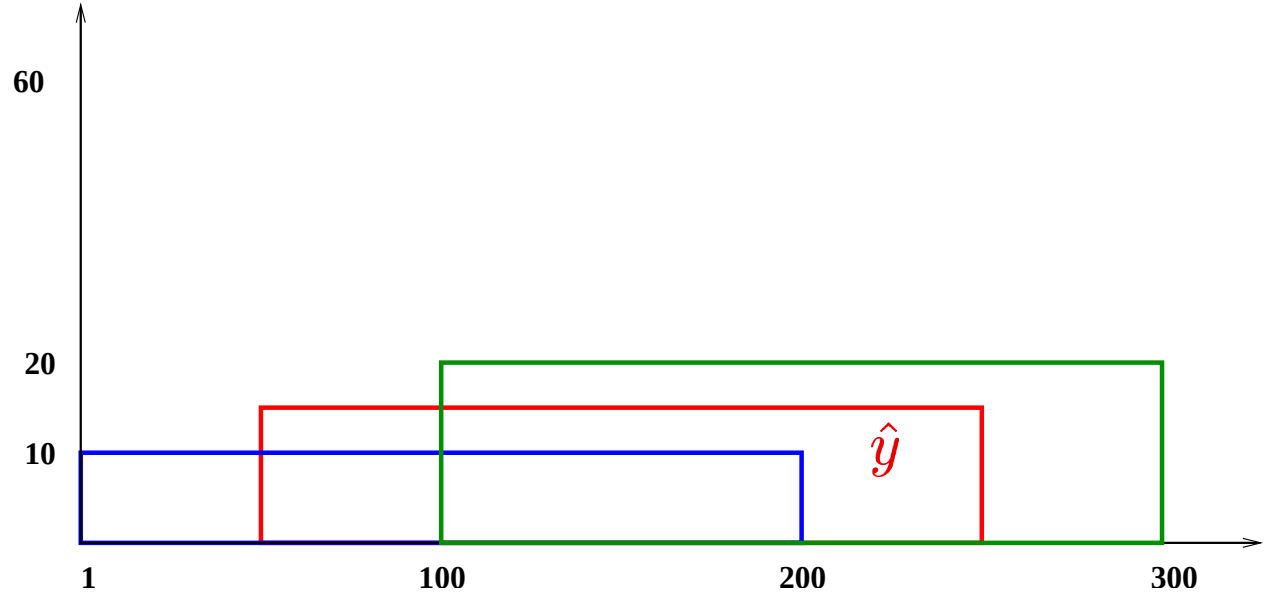
Non uniqueness of the minimizer

- if $\hat{y}^j$ is the minimizer of $\tilde{f}^j$ for $j = 1, ..., p$
  $\Rightarrow f(\hat{y}) = \max_{j=1...p} \tilde{f}^j(\hat{y}^j)$ is a minimum of f
  $\Rightarrow \hat{y} = \prod_{j=1}^{p} \hat{y}^j$ is a minimizer of f
- for all indexes j such that $\tilde{f}^j(\hat{y}^j) < f(\hat{y})$, all intervals $\tilde{y}^j$ such that $\tilde{f}^j(\tilde{y}^j) \leq f(\hat{y})$ produce also optimal solutions



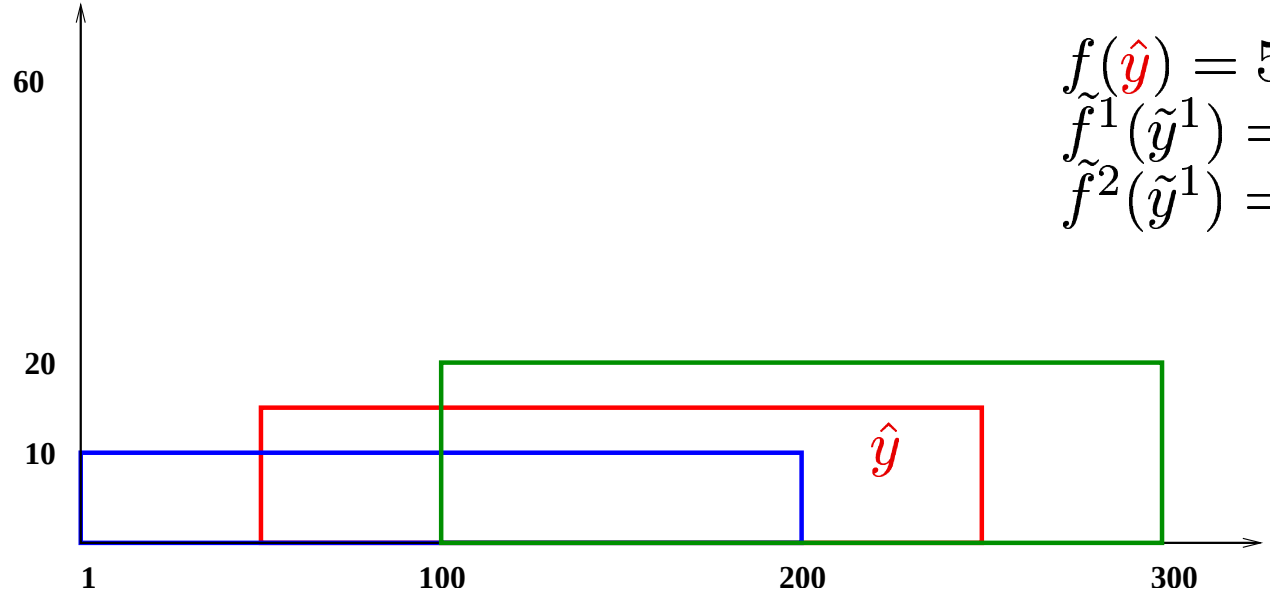
Non uniqueness of the minimizer

- if $\hat{y}^j$ is the minimizer of $\tilde{f}^j$ for $j = 1, ..., p$
  $\Rightarrow f(\hat{y}) = \max_{j=1...p} \tilde{f}^j(\hat{y}^j)$ is a minimum of f
  $\Rightarrow \hat{y} = \prod_{j=1}^{p} \hat{y}^j$ is a minimizer of f
- for all indexes j such that $\tilde{f}^j(\hat{y}^j) < f(\hat{y})$, all intervals $\tilde{y}^j$ such that $\tilde{f}^j(\tilde{y}^j) \leq f(\hat{y})$ produce also optimal solutions



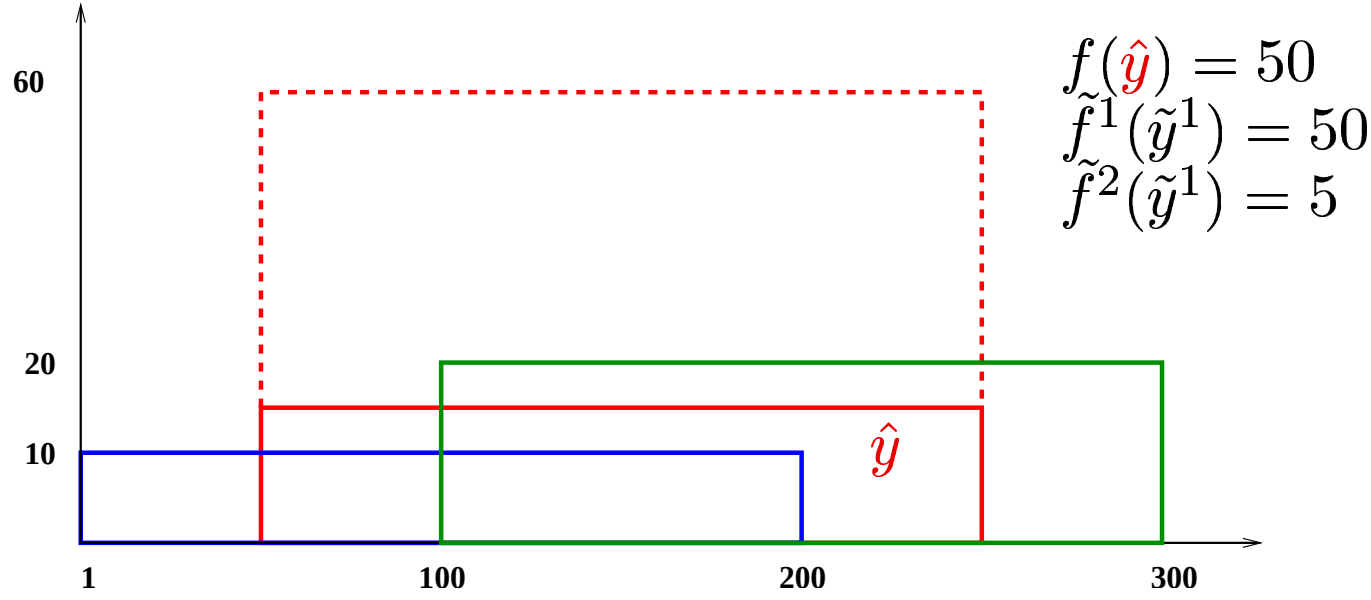
Non uniqueness of the minimizer

- if $\hat{y}^j$ is the minimizer of $\tilde{f}^j$ for $j = 1, ..., p$
  $\Rightarrow f(\hat{y}) = \max_{j=1...p} \tilde{f}^j(\hat{y}^j)$ is a minimum of f
  $\Rightarrow \hat{y} = \prod_{j=1}^{p} \hat{y}^j$ is a minimizer of f
- for all indexes j such that $\tilde{f}^j(\hat{y}^j) < f(\hat{y})$, all intervals $\tilde{y}^j$ such that $\tilde{f}^j(\tilde{y}^j) \leq f(\hat{y})$ produce also optimal solutions



Non uniqueness of the minimizer

- if $\hat{y}^j$ is the minimizer of $\tilde{f}^j$ for $j = 1, ..., p$
  $\Rightarrow f(\hat{y}) = \max_{j=1...p} \tilde{f}^j(\hat{y}^j)$ is a minimum of f
  $\Rightarrow \hat{y} = \prod_{j=1}^{p} \hat{y}^j$ is a minimizer of f
- for all indexes j such that $\tilde{f}^j(\hat{y}^j) < f(\hat{y})$, all intervals $\tilde{y}^j$ such that $\tilde{f}^j(\tilde{y}^j) \leq f(\hat{y})$ produce also optimal solutions



Conclusion

- with the optimum class prototype, at each iteration of the dynamical clustering algorithm any of the two clustering criteria decrease:

$$g(\{C_1, ..., C_K\}) = \sum_{k=1}^{K} \max_{i \in C_k} d_{H,\infty}(x_i, y_k)$$

$$g(\{C_1, ..., C_K\}) = \max_{k=1...K} \max_{i \in C_k} d_{H,\infty}(x_i, y_k)$$

  $\Rightarrow$ convergence of the algorithm

- sensitive to extreme values $\Rightarrow$ Explicit formula for prototypes with L_1 or L_2 Hausdorff distance ?

An Hausdorff distance between hyper-rectangles for clustering interval data

M. Chavent

Laboratoire de Mathématiques Appliquées de Bordeaux, UMR CNRS 5466

Universités Bordeaux1 et 2, France

chavent@math.u-bordeaux1.fr