Normalized k-means clustering of hyper-rectangles

M. Chavent

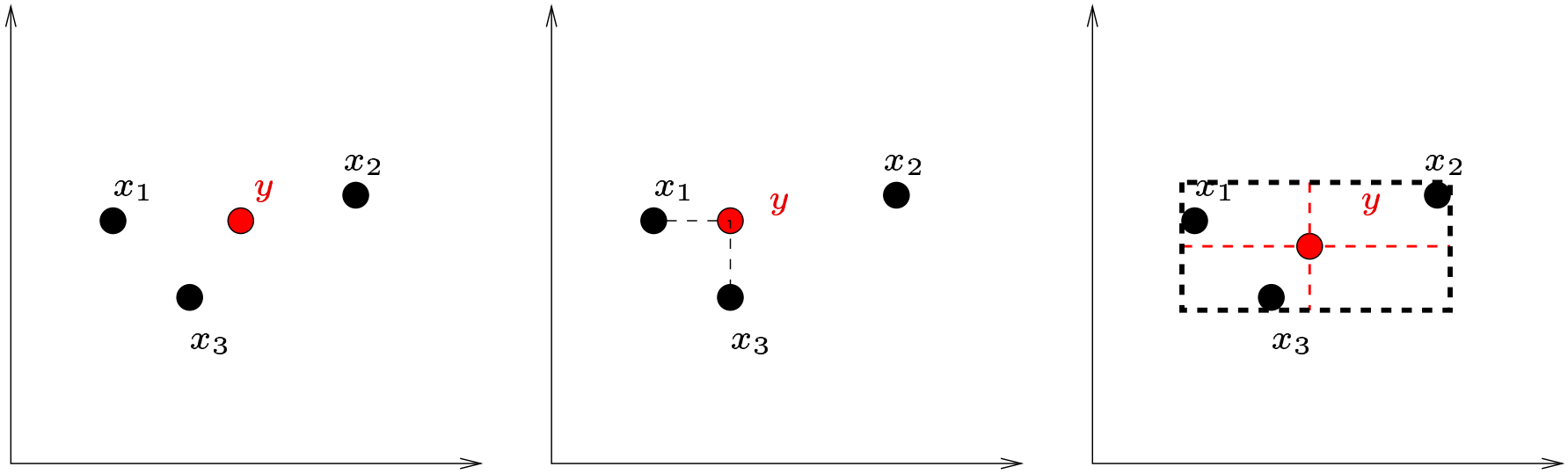
Laboratoire de Mathématiques Appliquées de Bordeaux, UMR CNRS 5466

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 :

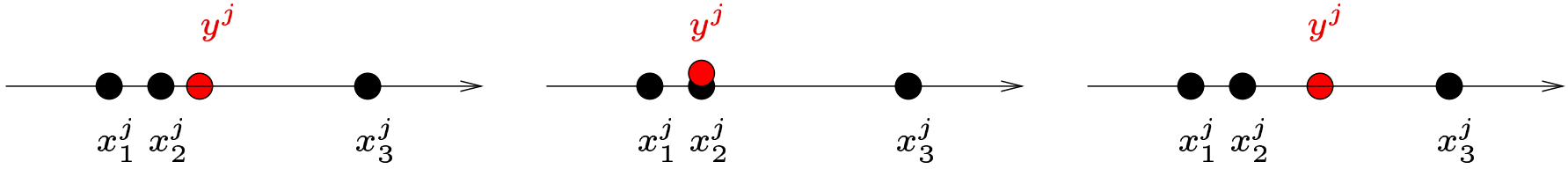


$\sum_{i=1}^{n} {d_2}^2(x_i, y)$ $\sum_{i=1}^{n} d_1(x_i, y)$ $\max_{i=1\ldots n} d_\infty(x_i, y)$

k-means Dynamical Clustering

Introduction

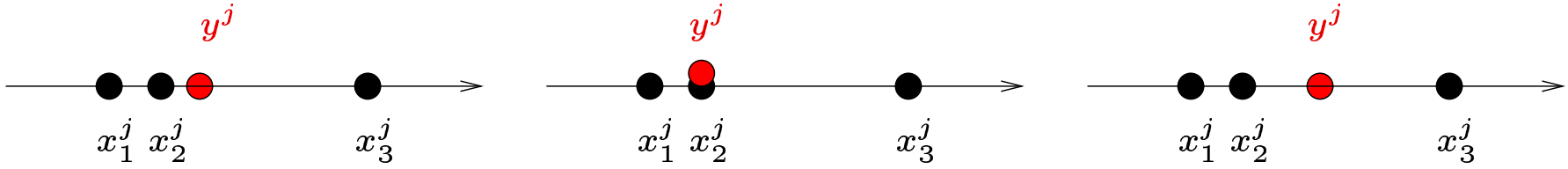
Three different measures of centrality y^j :



The mean | The median | The middle

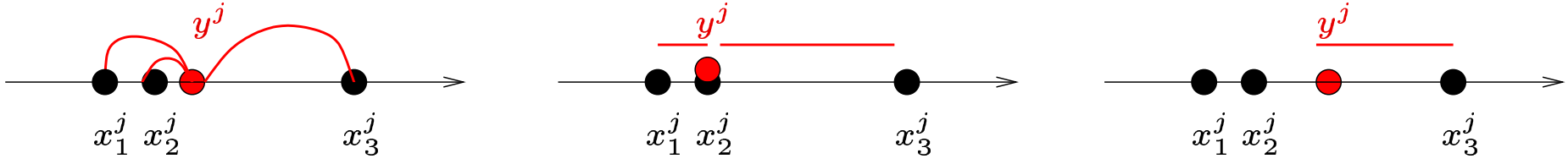
Introduction

Three different measures of centrality y^j :



The mean | The median | The middle

Three different measures of dispersion σ^j :



The square deviation | The absolute deviation | The max deviation

Introduction

Normalized L_α distance between two $\mathbb{R}^p$-points:

$$d_\alpha(x_1, x_2) = \left(\sum_{j=1}^{p} \frac{1}{\sigma^j} |x_1^j - x_2^j|^\alpha\right)^{\frac{1}{\alpha}}$$

Three normalized k-means algorithms:

Prototype	Distance	Measure of dispersion
The "mean" $\mathbb{R}^p$-point	L_2-Euclidean	The square deviation from the mean
The "median" $\mathbb{R}^p$-point	L_1-City-Block	The absolute deviation from the median
The "middle" $\mathbb{R}^p$-point	L_∞-Max	The max deviation from the middle

Plan

Part 1 Interval data

Part 2 Comparing hyper-rectangles

Part 3 Define a class prototype

Part 3 Normalization

PART 1

Interval data

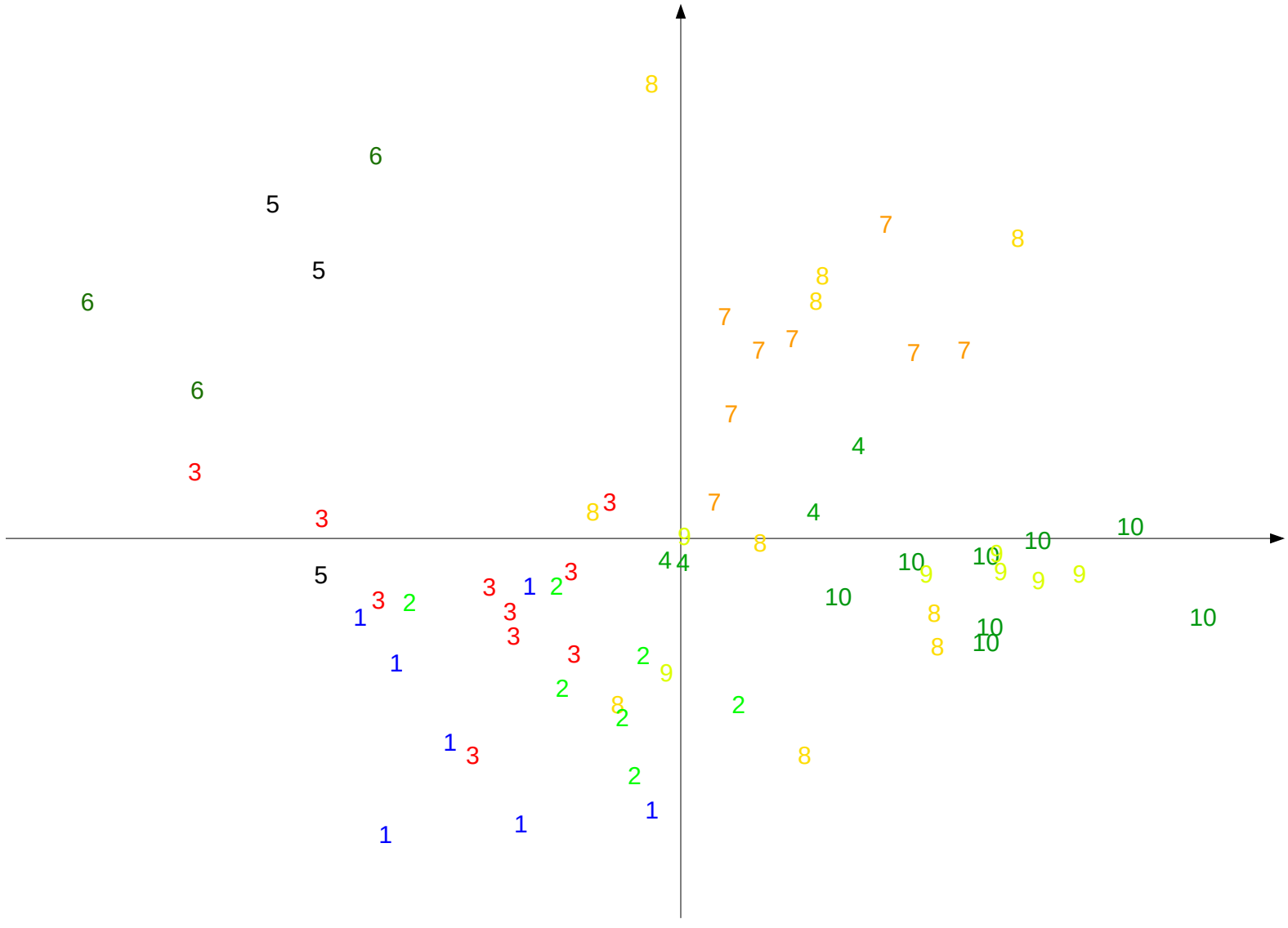
French Guyana Fish Example

- 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)

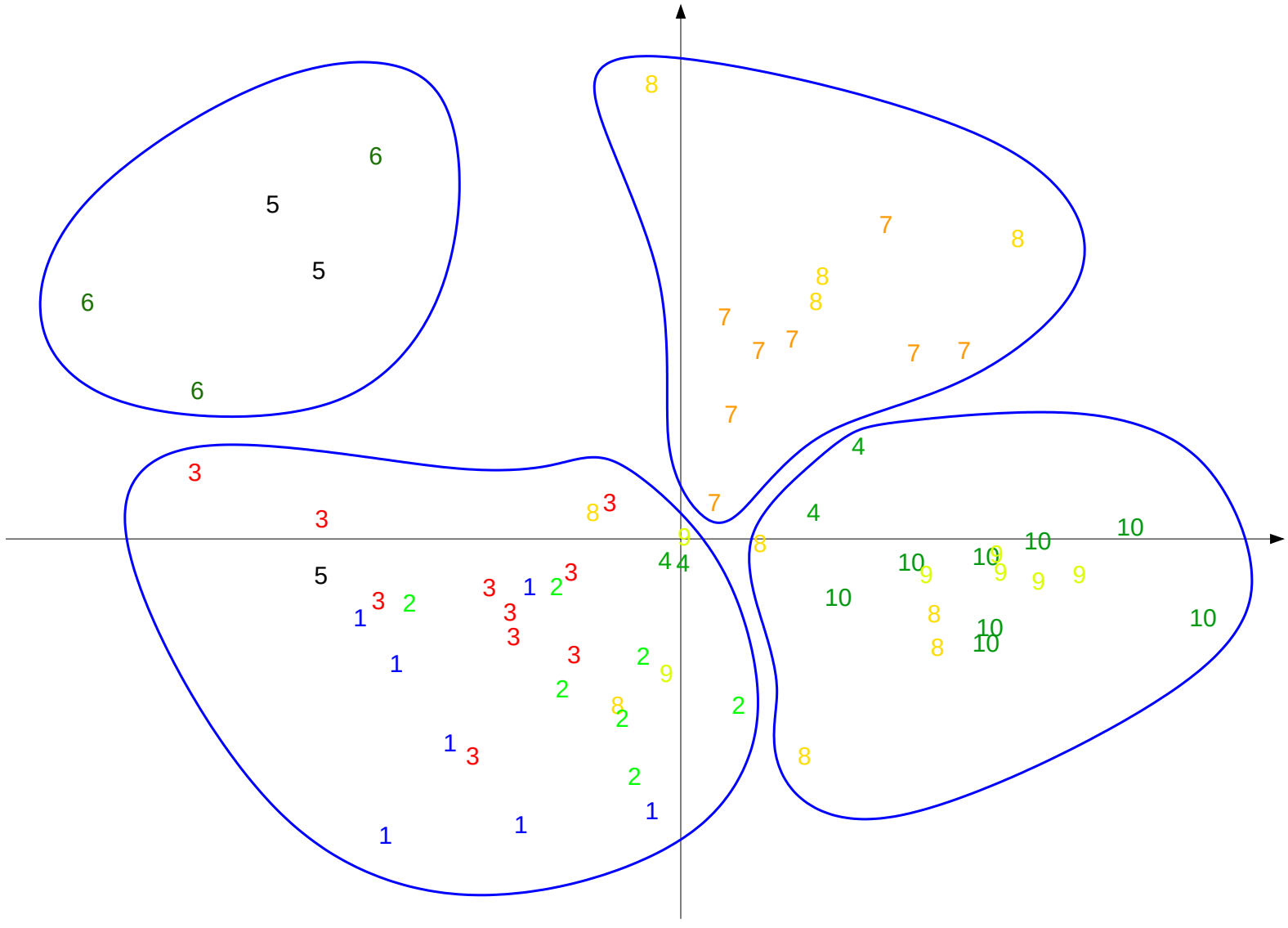
Fish	liver	kidney	gills	intestine	stomach	species
1	-0.116	0.352	-1.214	-1.147	NA	ageneiosus brevifi li
2	-0.083	-0.457	-1.881	-1.171	-1.485	ageneiosus brevifi li
...	...	...	...	...	...	...
8	1.416	0.684	-1.439	-1.554	-0.874	cynodon gibbus
9	0.115	-0.509	-1.910	NA	-1.610	cynodon gibbus
...	...	...	...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

- How to cluster the 10 species

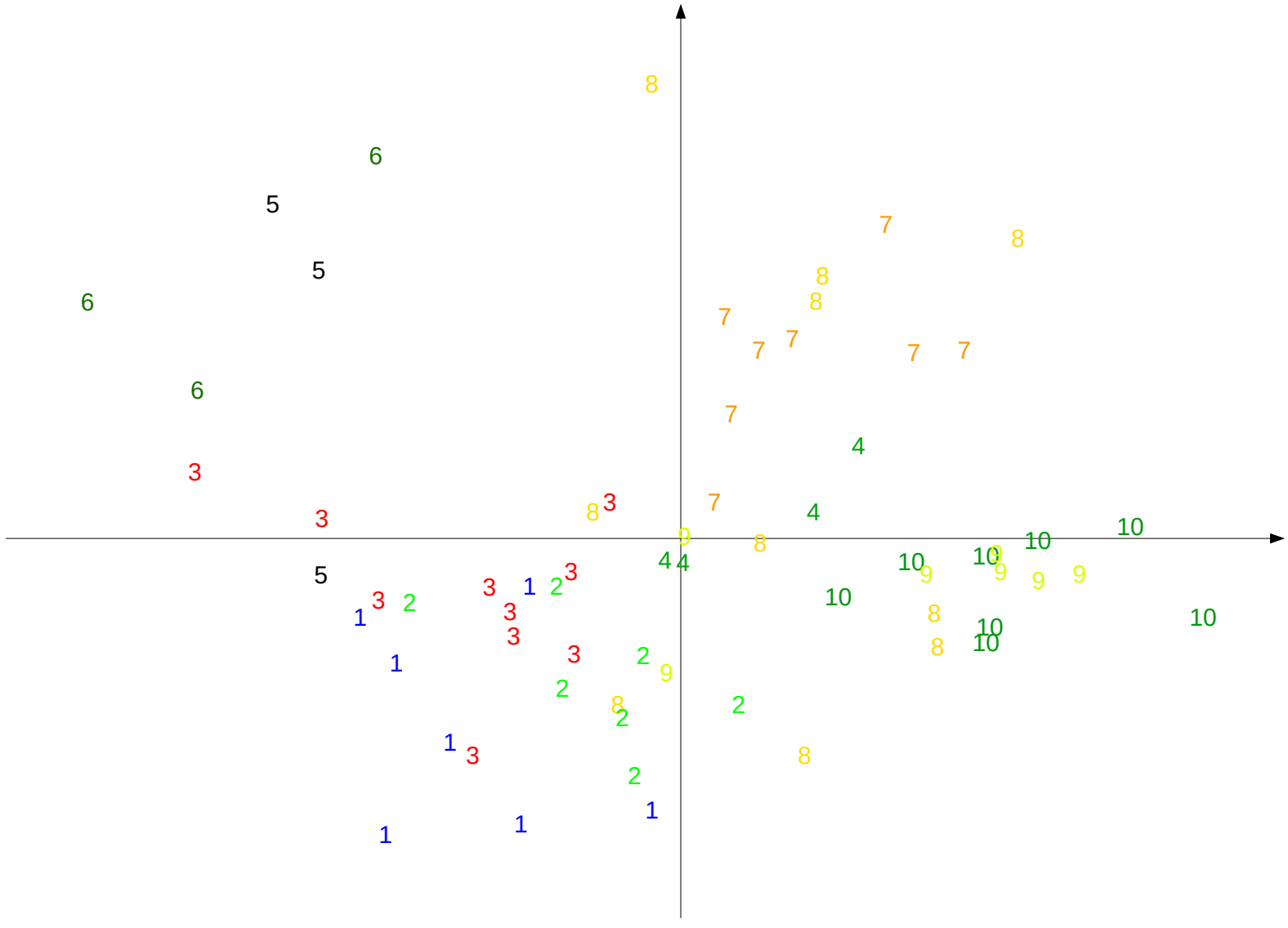
Classical or interval data representation



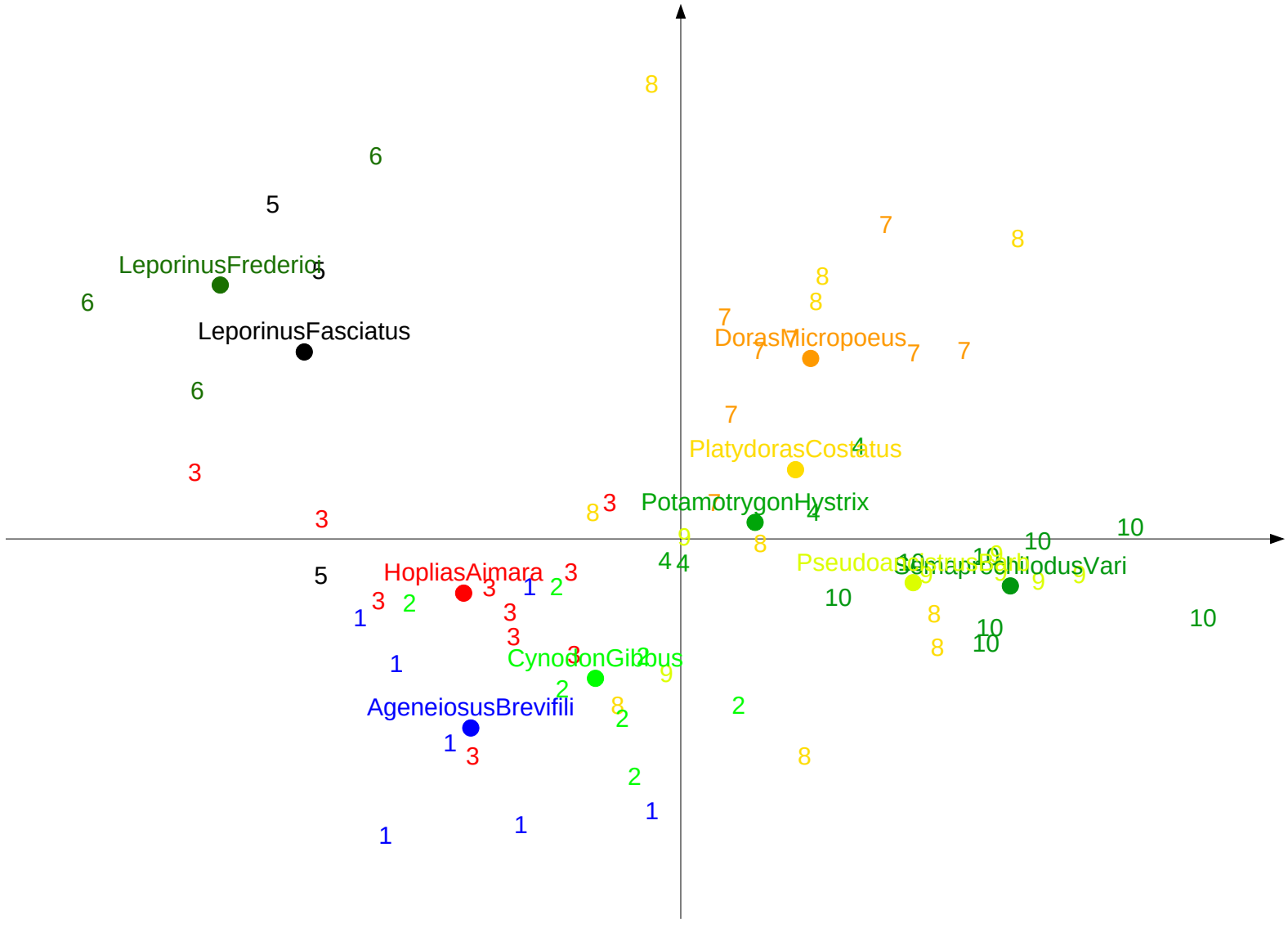
Classical or interval data representation



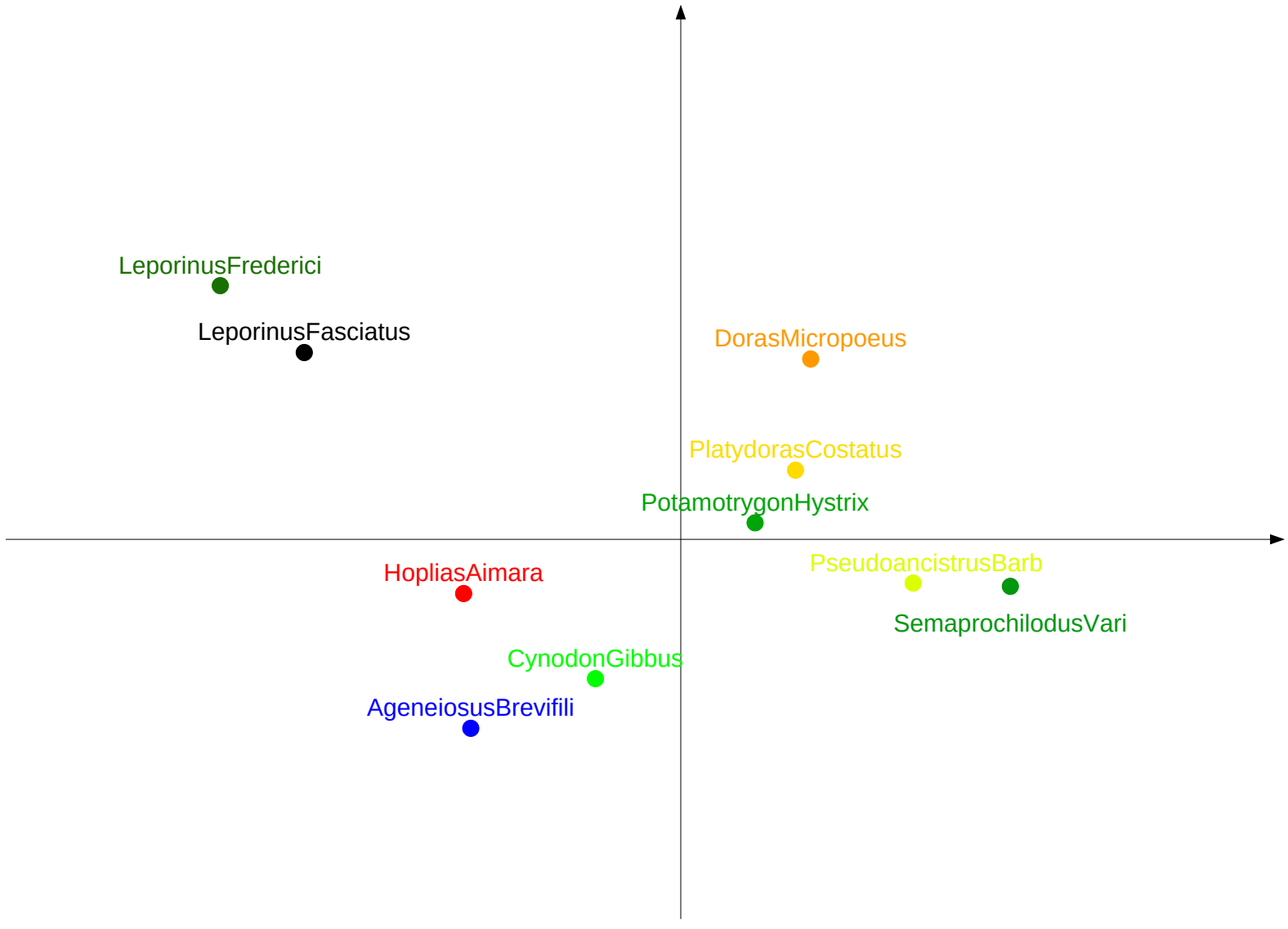
Classical or interval data representation



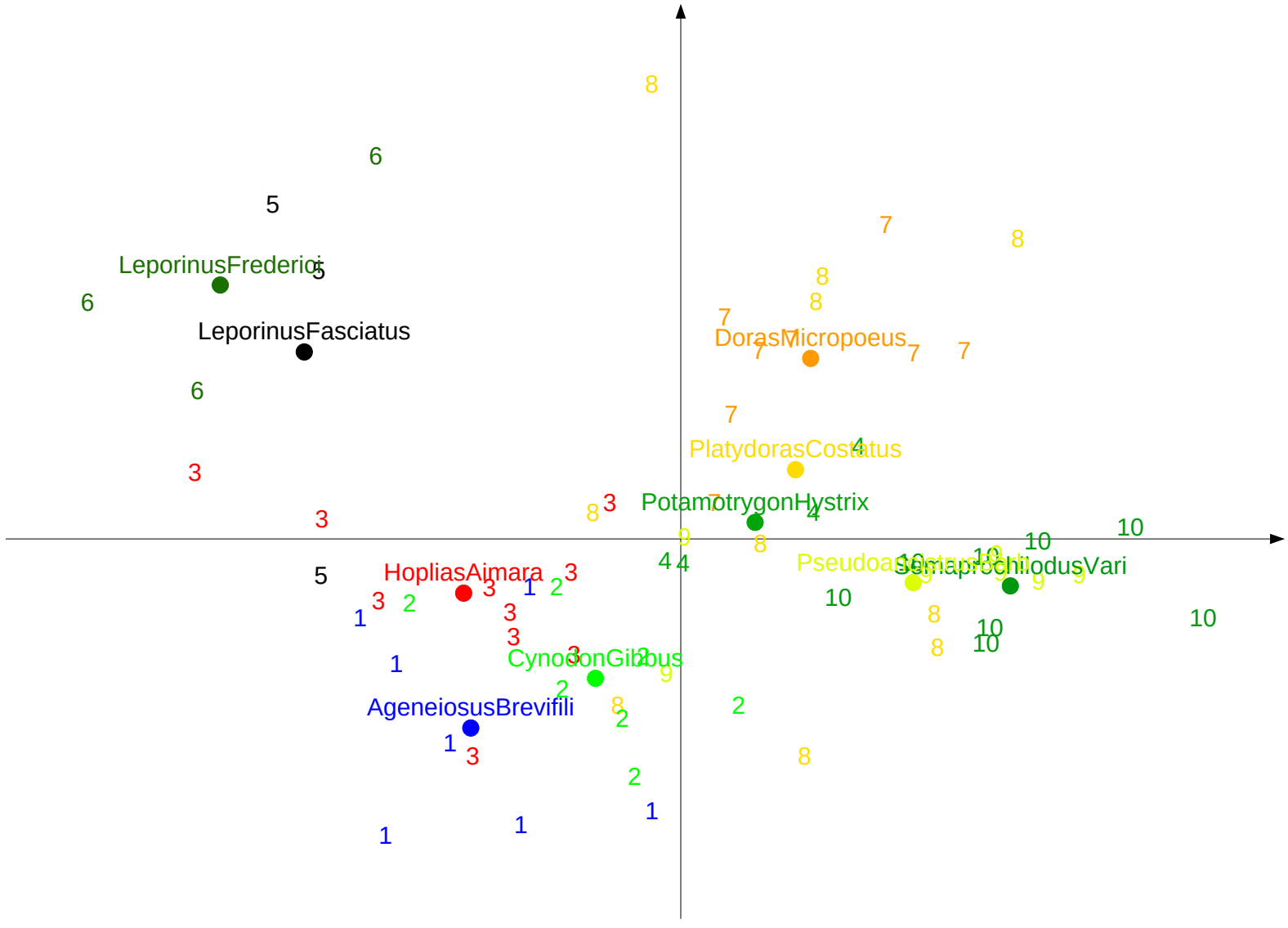
Classical or interval data representation



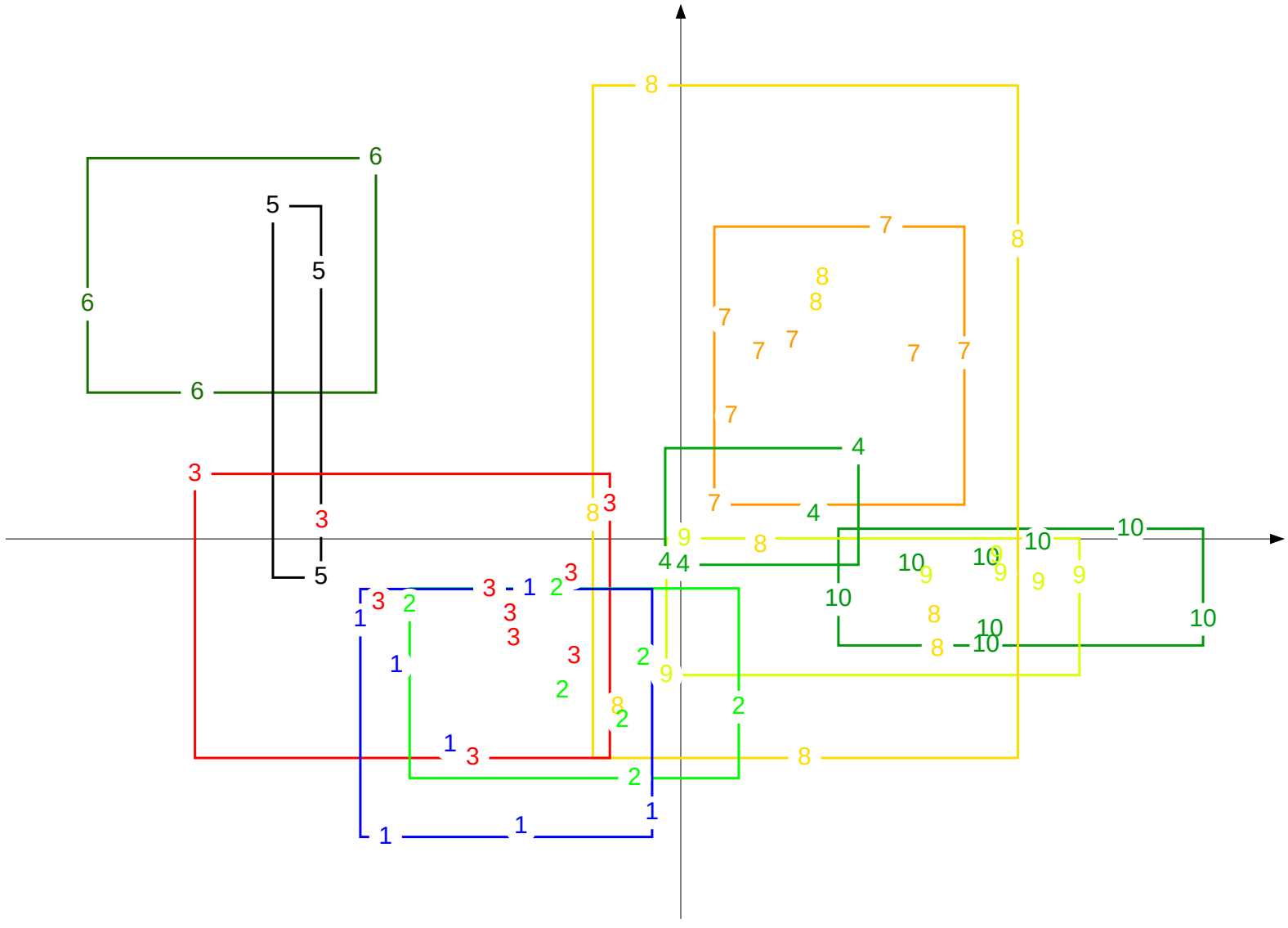
Classical or interval data representation



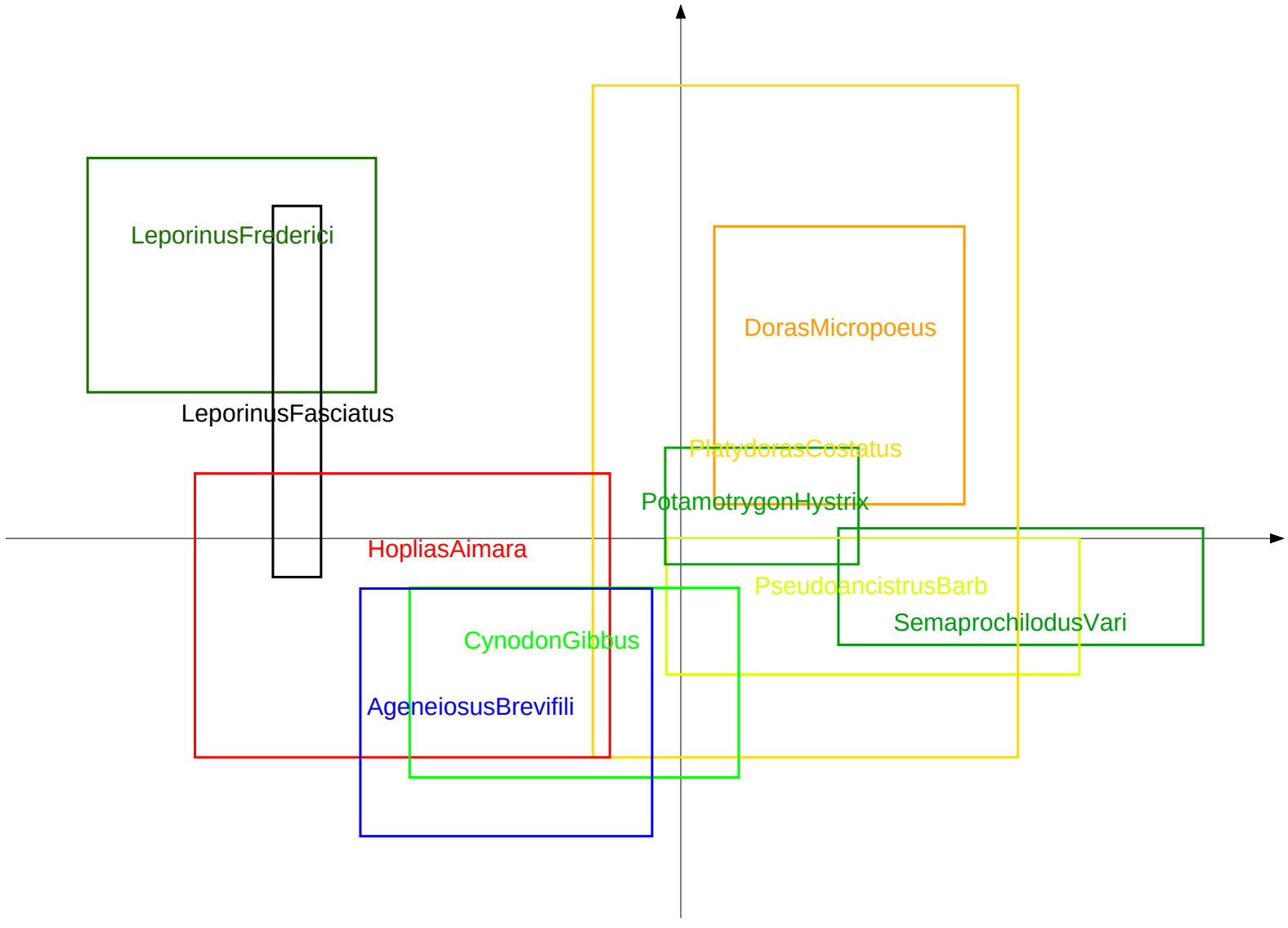
Classical or interval data representation



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

$$d_{H,\alpha}(A, B) = \max(h(A, B), h(B, A))$$

with

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

$\Rightarrow$ Depends on the distance d_α (L_1,L_2...L_∞) chosen to compare two points of $\mathbb{R}^p$

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 space we can drop the subscript α and:

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

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 space we can drop the subscript α and:

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

- *Property 2.* With the L_∞ distance, we have the following relation between the Hausdorff distance $d_{H,\infty}$ in p dimensions and d_H in one dimension:

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

Two distances between hyper-rectangles

We are able to give an explicit formula of the optimum class prototype with:

- The L_∞ Hausdorff distance:
  $$d_{H,\infty}(A,B) = \max_{j=1,\ldots,p} \underbrace{\max(|a_j - \alpha_j|, |b_j - \beta_j|)}_{d_H(A_j, B_j)}$$
  $\Rightarrow$ 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

Two distances between hyper-rectangles

We are able to give an explicit formula of the optimum class prototype with:

- The L_∞ Hausdorff distance:
$$d_{H,\infty}(A,B) = \max_{j=1,\ldots,p} \underbrace{\max(|a_j - \alpha_j|, |b_j - \beta_j|)}_{d_H(A_j,B_j)}$$
$\Rightarrow$ 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

- The Hausdorff-based distance:
$$d(A,B) = \sum_{j=1}^{p} \underbrace{\max(|a_j - \alpha_j|, |b_j - \beta_j|)}_{d_H(A_j,B_j)}$$
$\Rightarrow$ In the particular case of intervals reduced to single points, this distance is the well-known L_1 City-Block 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

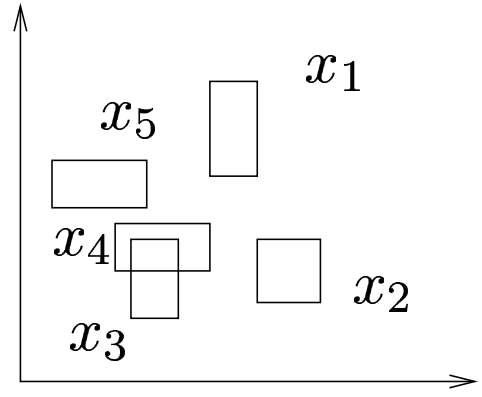
num	species	Axis1	Axis2
1	AgeneiosusBrevifi li	[1,957:-0,175]	[0,306:1,819]
2	CynodonGibbus	[-1,656:0,354]	[0,302:0,302]
3	DorasMicropoeus	[-2,967:-0,433]	[-0,397:1,337]
...	...	...	...

- 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 Hausdorff-based distance:
    $d(x_1, x_2) = 0,529+1,517$

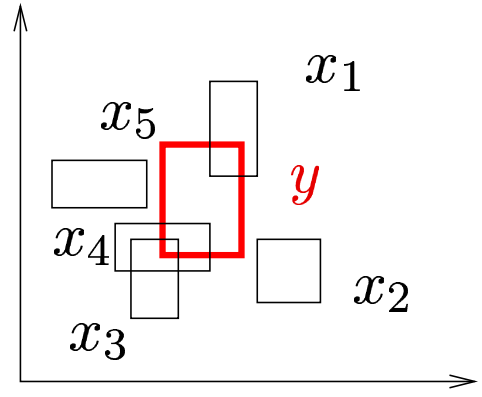
PART 3

Define a class prototype

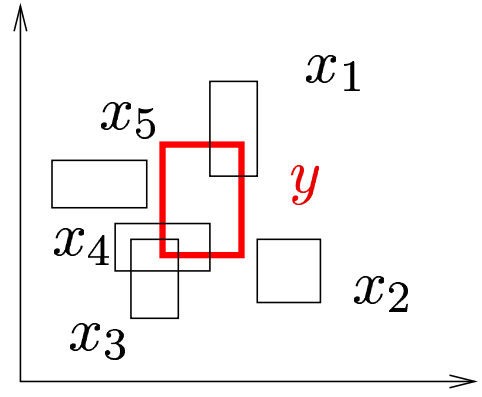
Define a class prototype



Define a class prototype

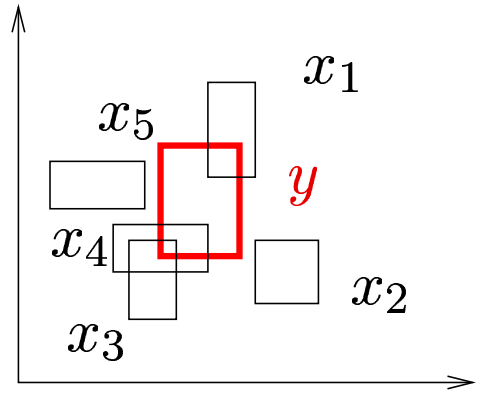


Define a class prototype



- 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

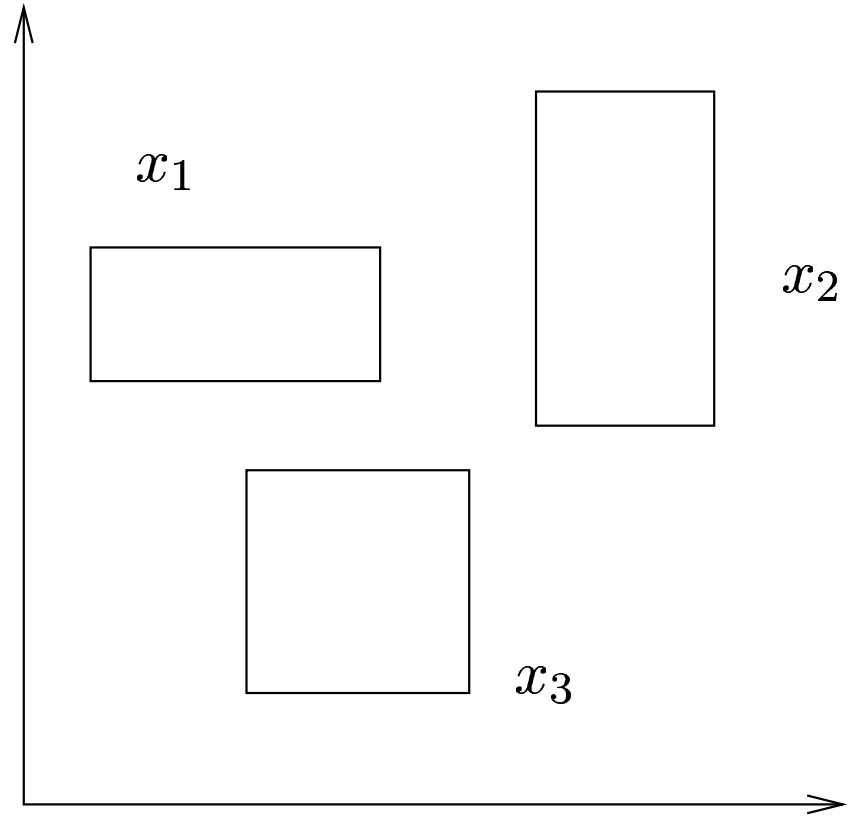
- Adequacy criterion: $f(y) = \sum_{i \in C} d(x_i, y)$ ('The star')

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

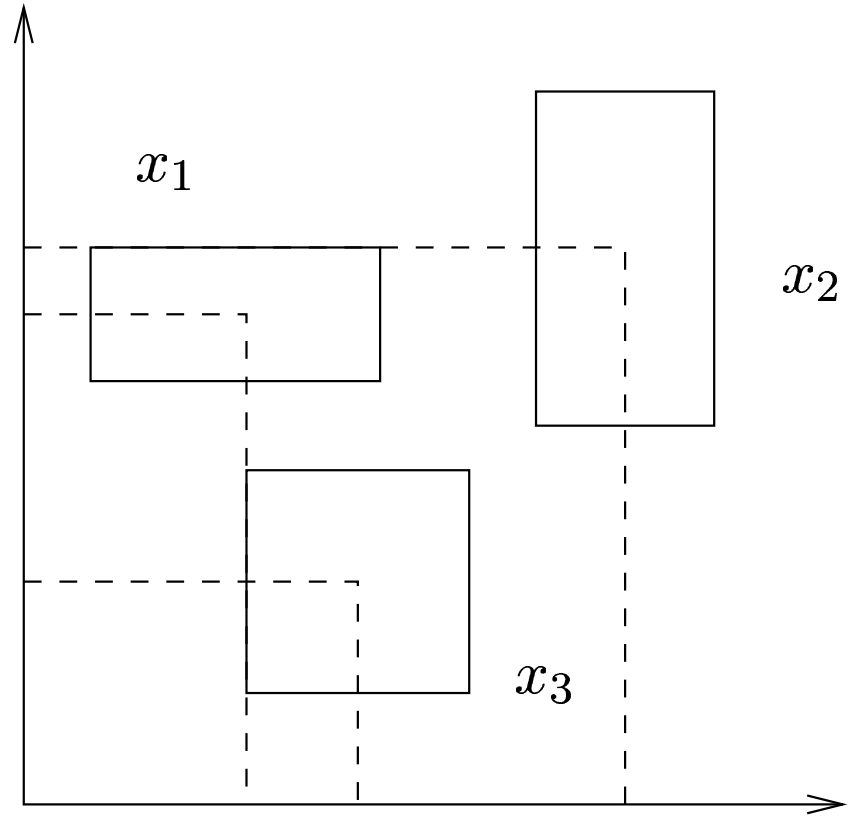
$$\begin{aligned} \hat{\mu}^j &= median\{m_i^j \mid i \in C\} \\ \hat{\lambda}^j &= median\{l_i^j \mid i \in C\} \end{aligned}$$

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

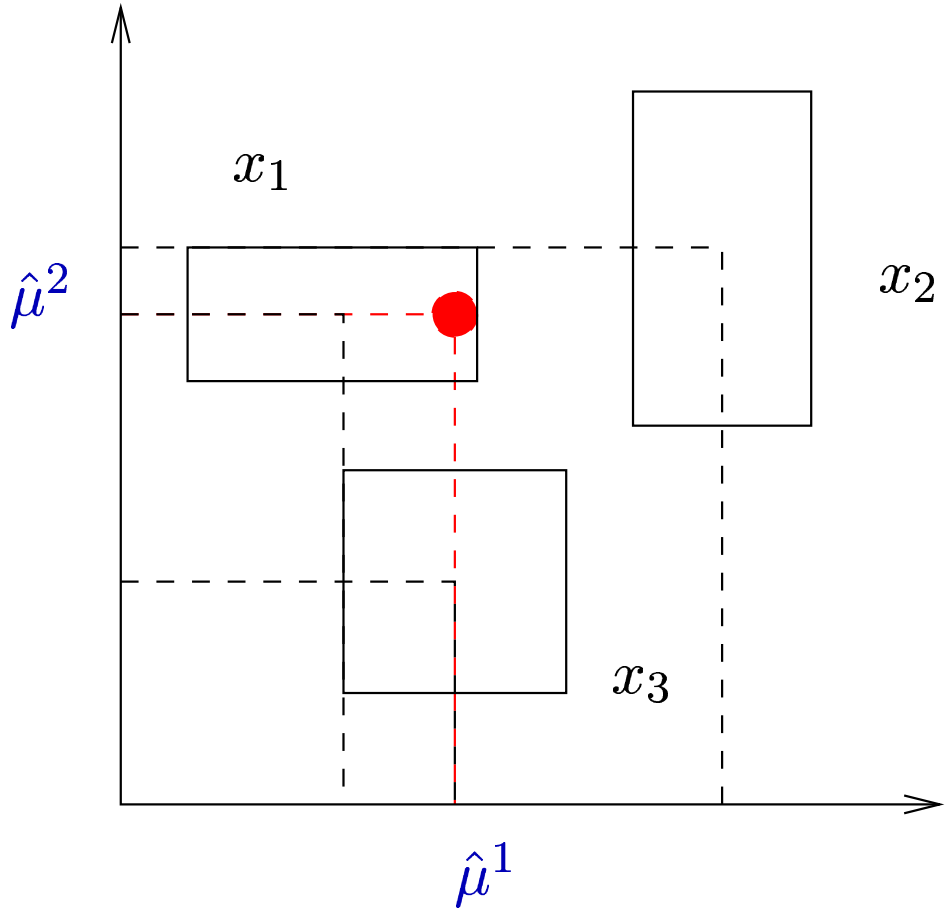
An example



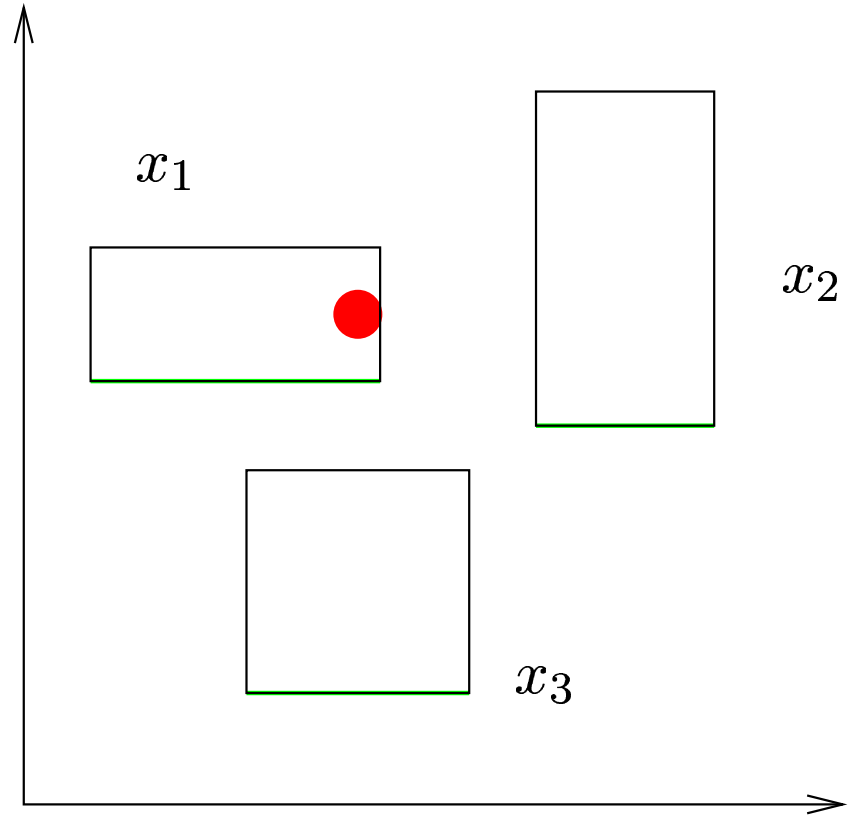
An example



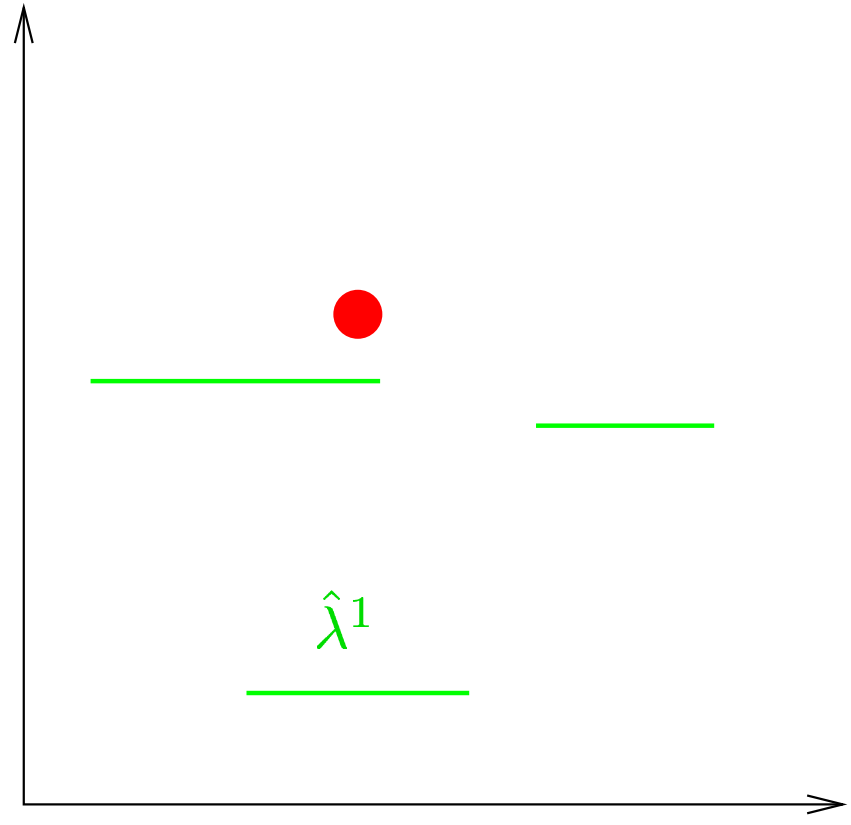
An example



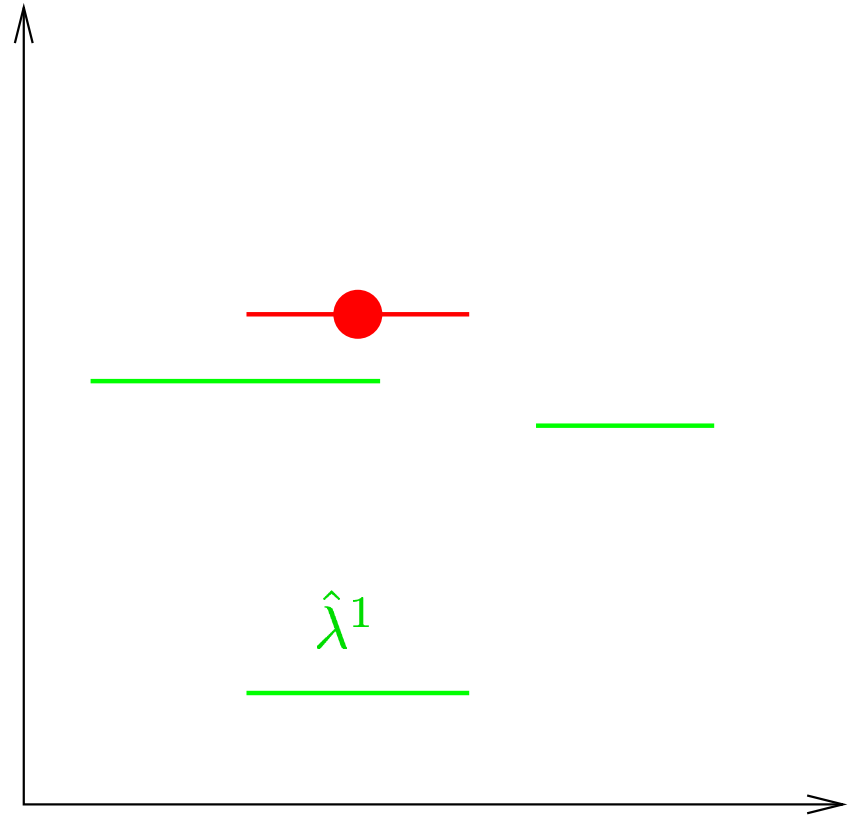
An example



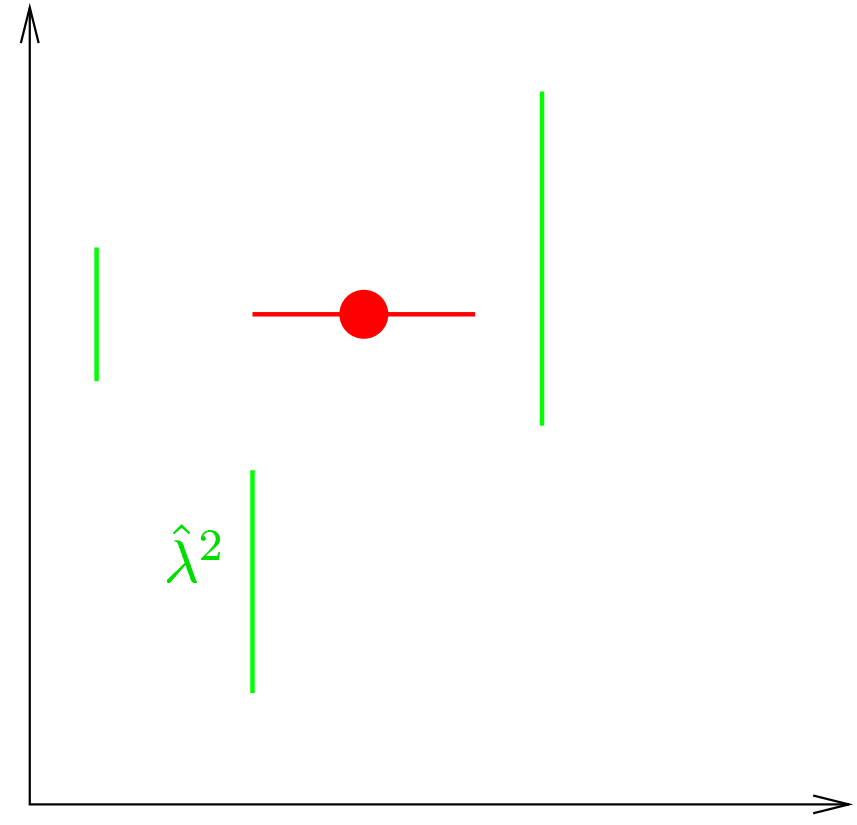
An example



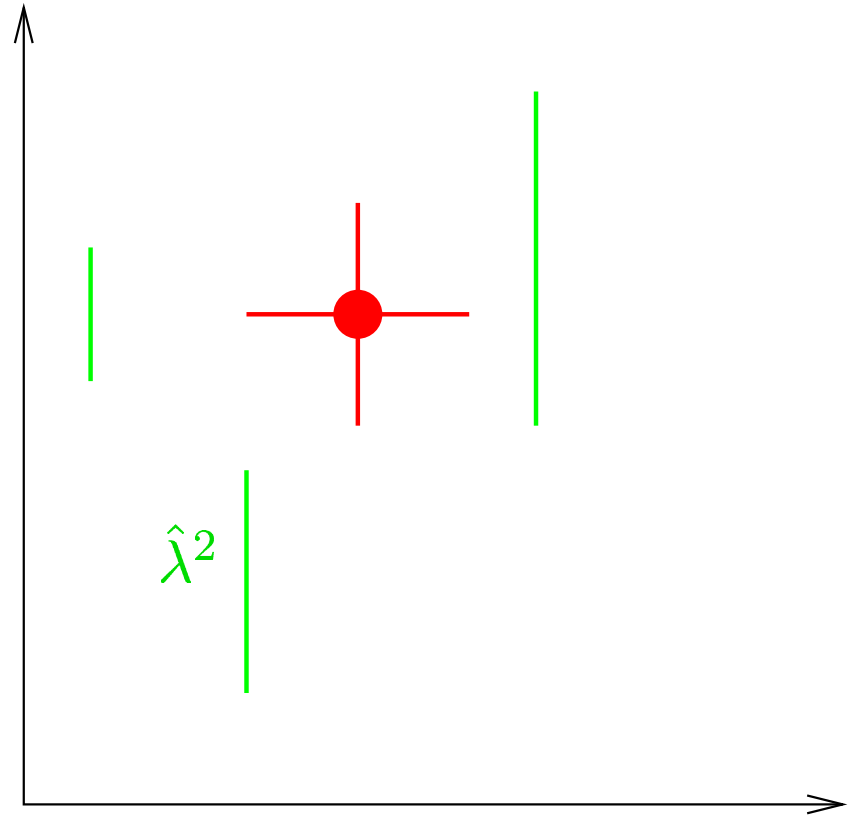
An example



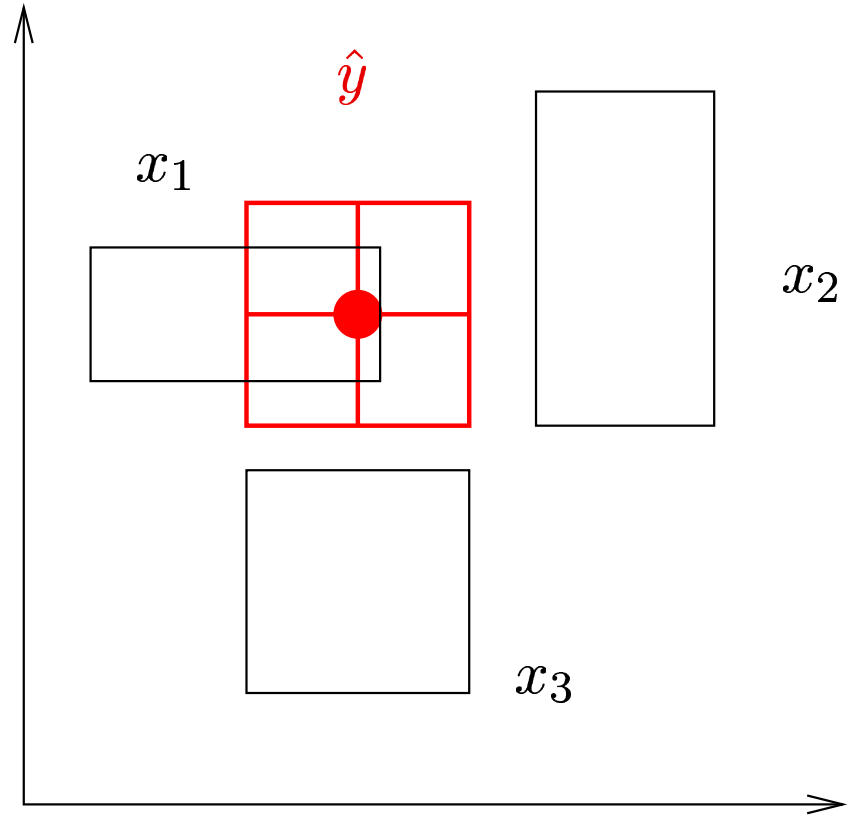
An example



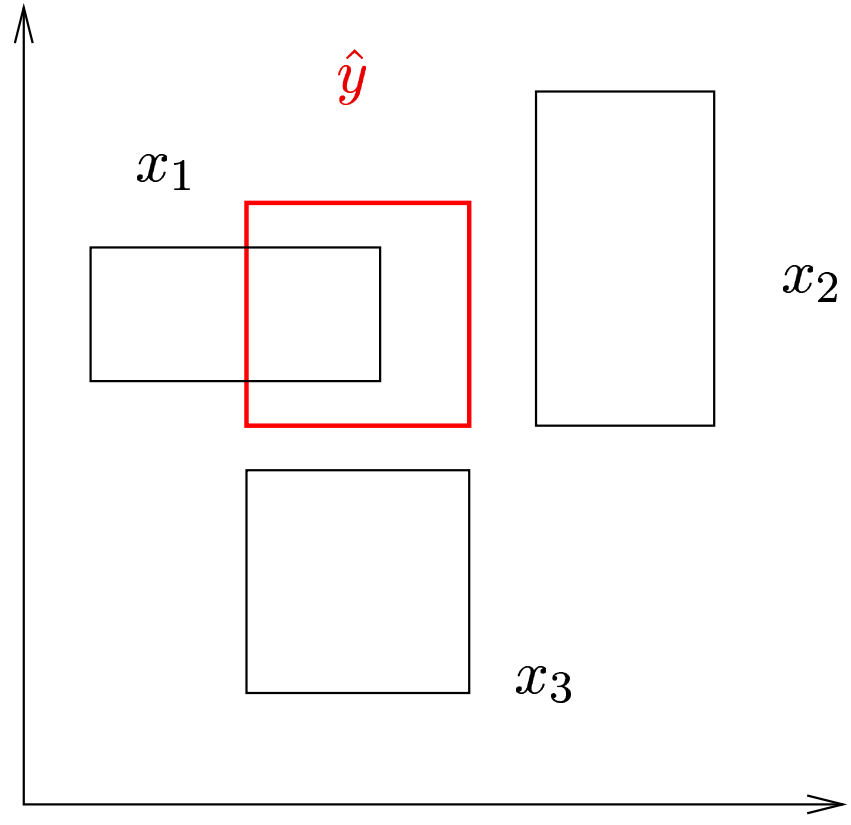
An example



An example



An example



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

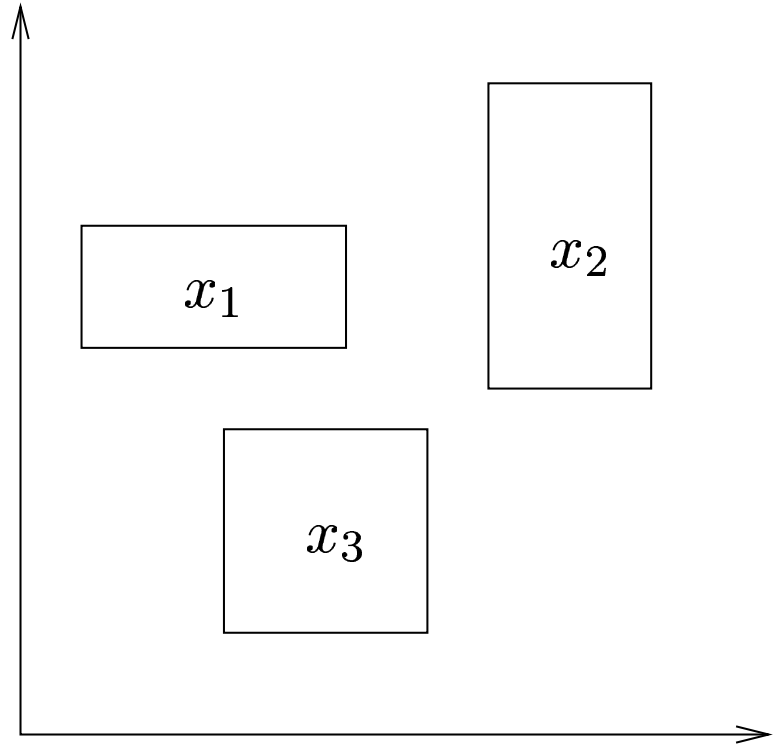
- Adequacy criterion: $f(y) = \max_{i \in C} d_{H,\infty}(x_i, y)$ ('The radius')

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

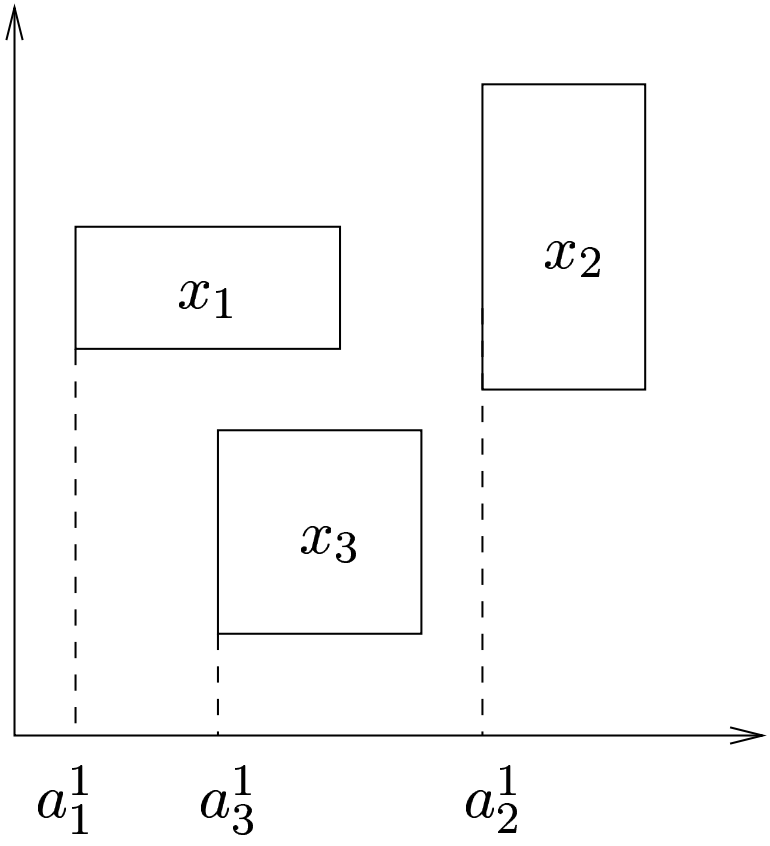
$$\begin{aligned} \hat{\alpha}^j &= \frac{\max_{i \in C} a_i^j + \min_{i \in C} a_i^j}{2} \\ \hat{\beta}^j &= \frac{\max_{i \in C} b_i^j + \min_{i \in C} b_i^j}{2} \end{aligned}$$

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

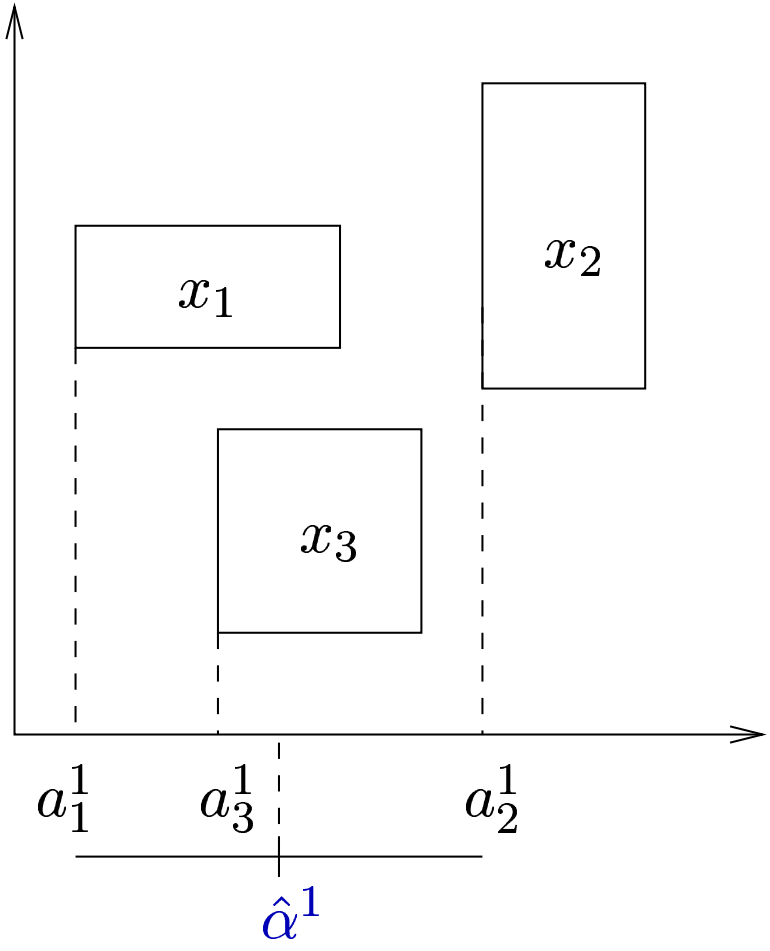
An example



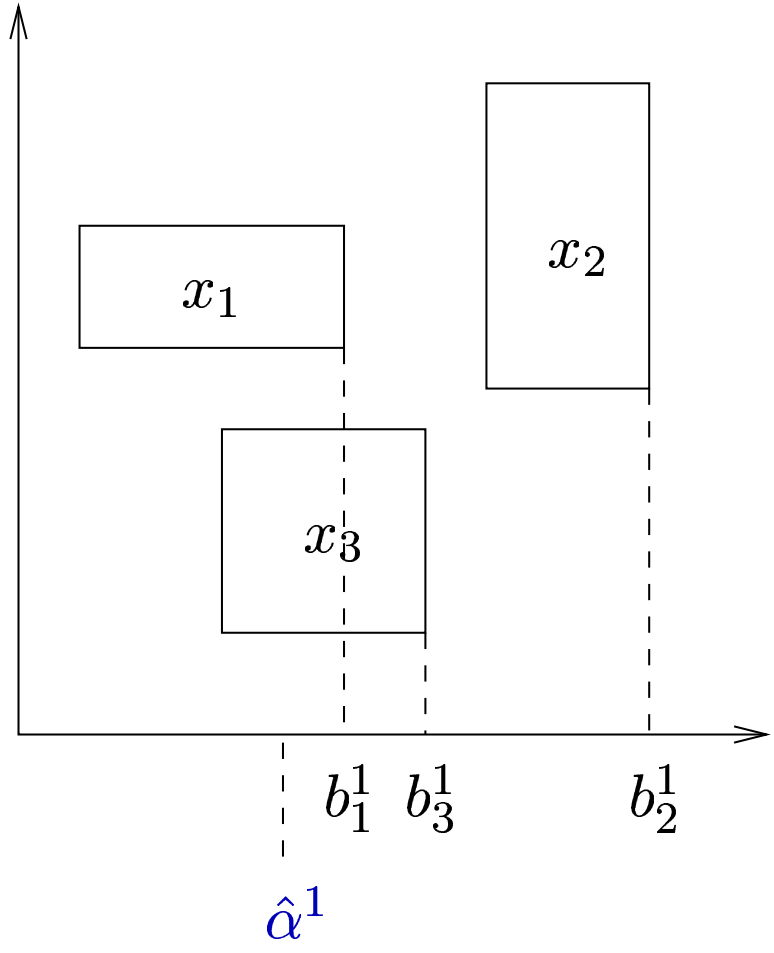
An example



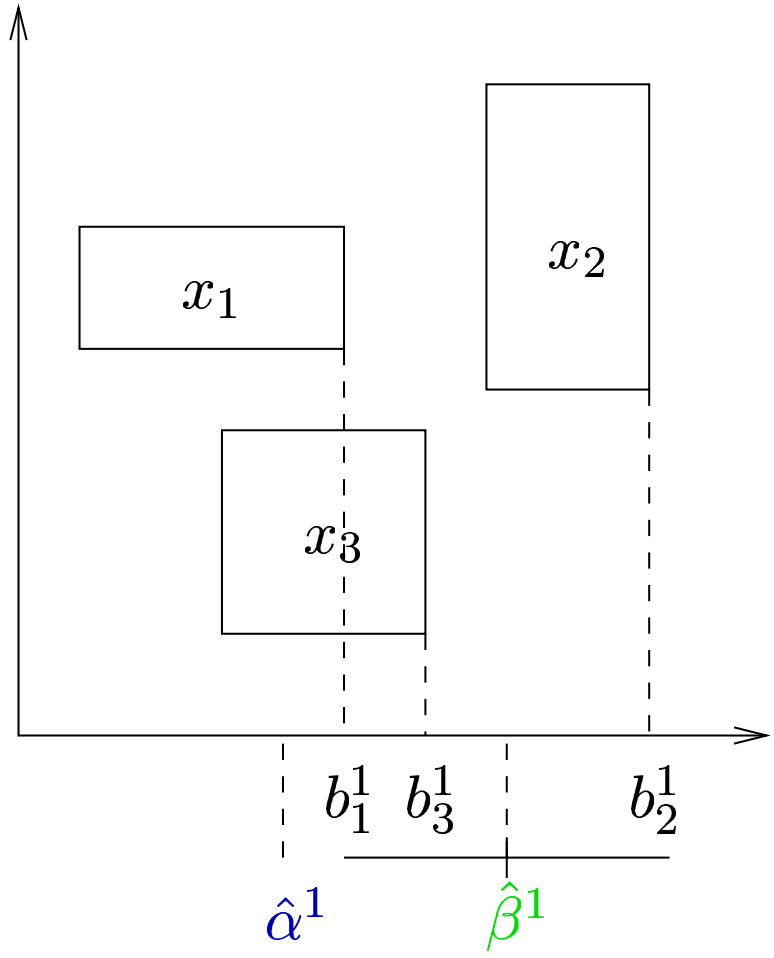
An example



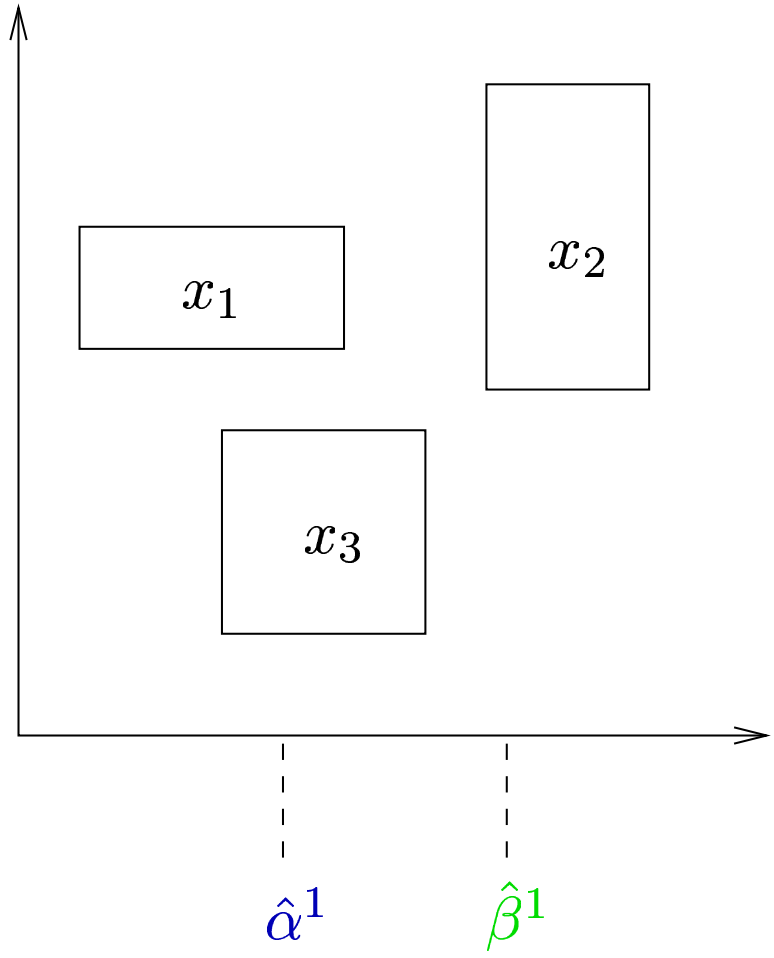
An example



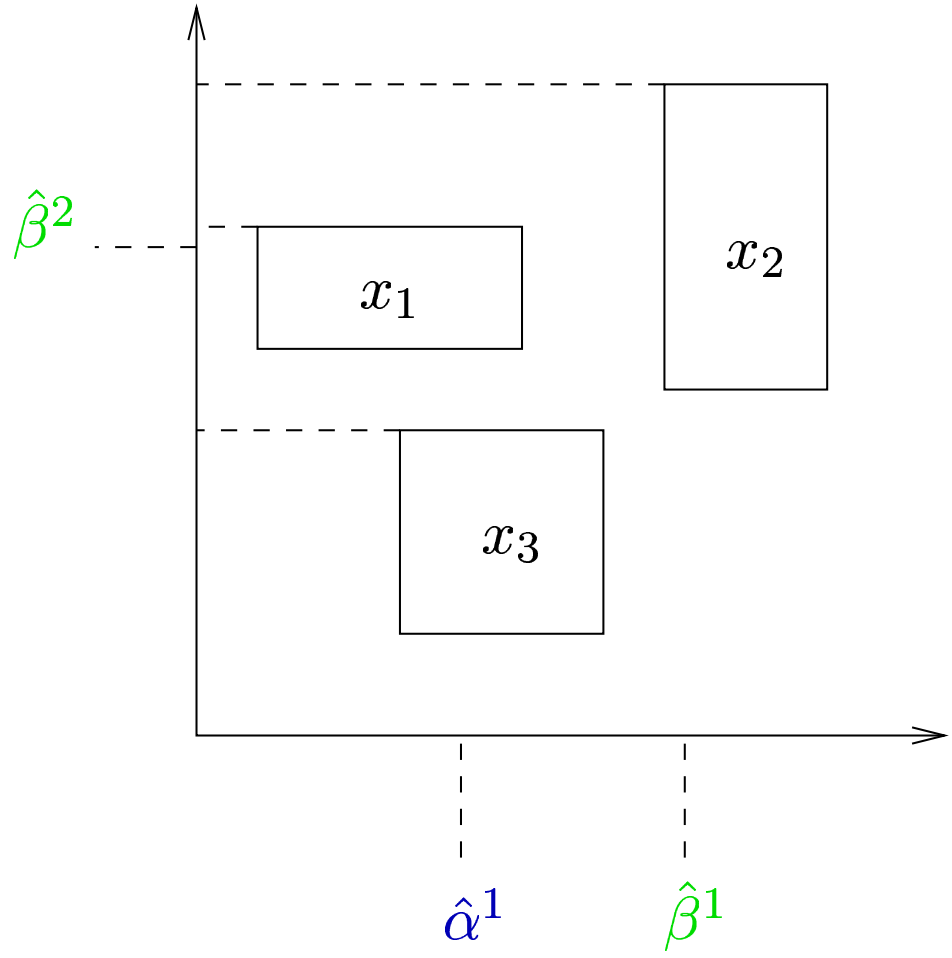
An example



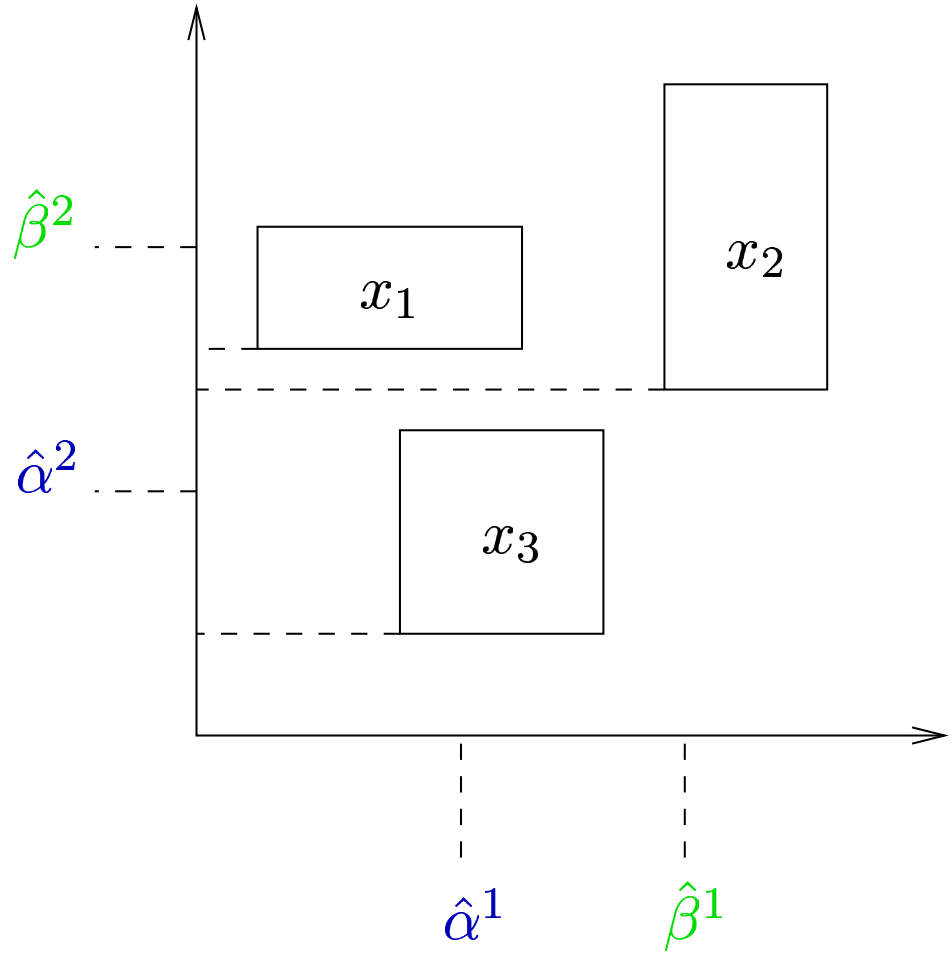
An example



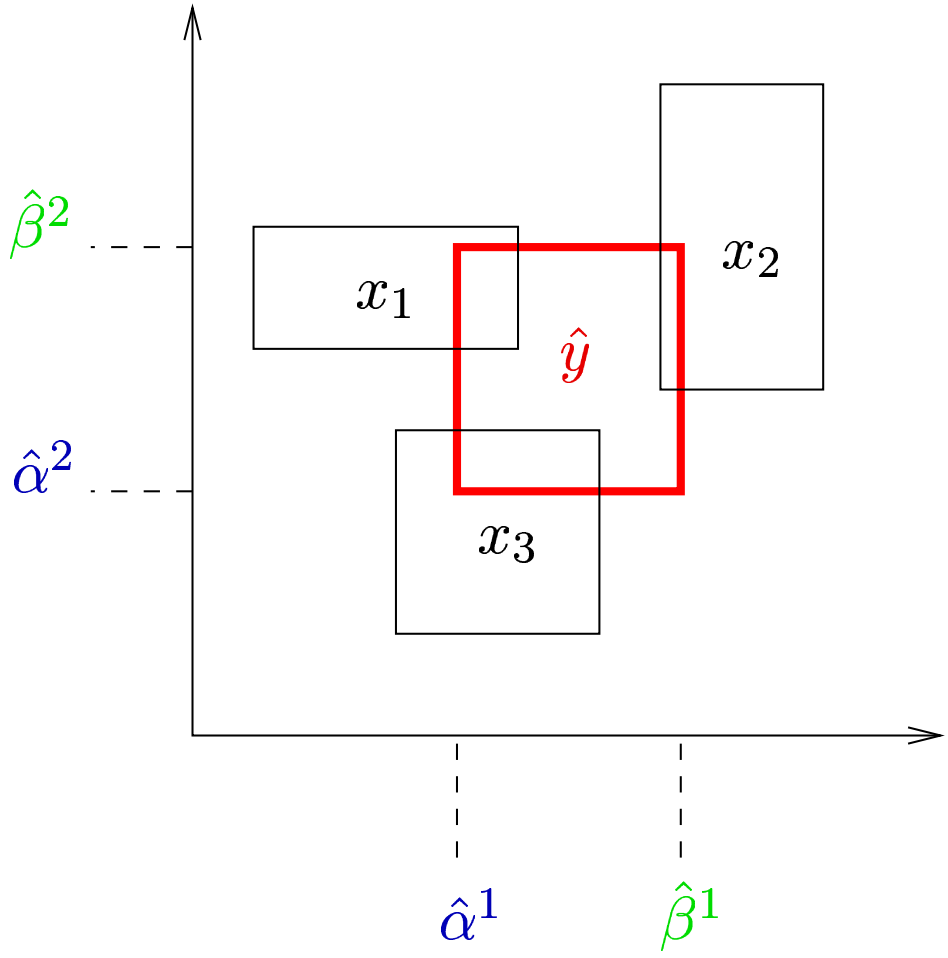
An example



An example



An example

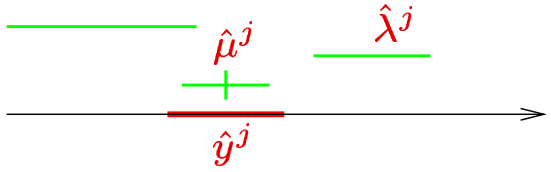


PART 4

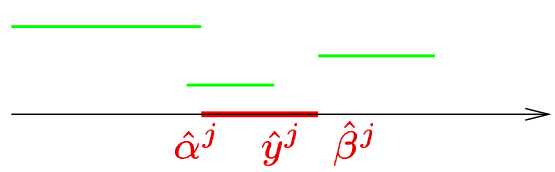
Normalization

Measure of centrality and dispersion

Two “central” intervals $\hat{y}^j$:



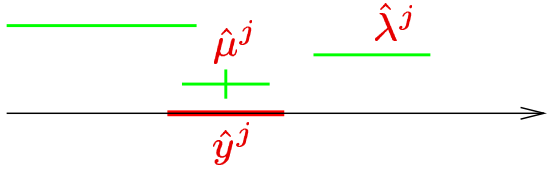
The “median” interval



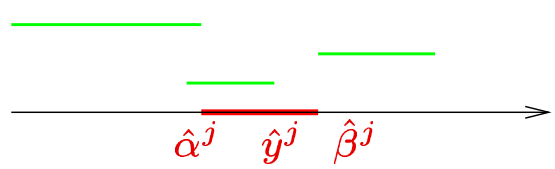
The “middle” interval

Measure of centrality and dispersion

Two "central" intervals $\hat{y}^j$:



The "median" interval



The "middle" interval

Two measures of dispersion σ^j from a "central" interval:

The "star":

$$\sum_{i=1}^{n} d_H(x_i^j, \hat{y}^j)$$

$$\Downarrow$$

$$\sigma^j = \sum_{i=1}^{n} \max(|a_i^j - \hat{\mu}^j + \hat{\lambda}^j|, |b_i^j - \hat{\mu}^j - \hat{\lambda}^j|)$$

The "radius":

$$\max_{i=1\ldots n} d_H(x_i^j, \hat{y}^j)$$

$$\Downarrow$$

$$\sigma^j = \max_{i=1\ldots n} \max(|a_i^j - \hat{\alpha}^j|, |b_i^j - \hat{\beta}$$

Normalized distance or data table

Initial data table $(x_i^j)_{n\times p}$

$$x_i^j = [a_i^j, b_i^j]$$

$$d_H(x_i^j, x_{i'}^j)$$

Normalized distance or data table

Initial data table $(x_i^j)_{n\times p}$

$$x_i^j = [a_i^j, b_i^j]$$

$$d_H(x_i^j, x_{i'}^j)$$

$$\Downarrow$$

Normalized distance

$$\sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j)$$

$$\max_{j=1\ldots p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j)$$

Normalized distance or data table

Initial data table $(x_i^j)_{n\times p}$

$$x_i^j = [a_i^j, b_i^j]$$

$$d_H(x_i^j, x_{i'}^j)$$

$$\Downarrow$$

Normalized distance

$$\sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j)$$

$$\max_{j=1\ldots p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j)$$

Normalized data table $(z_i^j)_{n\times p}$

$$z_i^j = [\frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j}]$$

$$d_H(z_i^j, z_{i'}^j)$$

Normalized distance or data table

Initial data table $(x_i^j)_{n\times p}$

$$x_i^j = [a_i^j, b_i^j]$$

$$d_H(x_i^j, x_{i'}^j)$$

$$\Downarrow$$

Normalized distance

$$\sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j)$$

$$\max_{j=1\ldots p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j)$$

Normalized data table $(z_i^j)_{n\times p}$

$$z_i^j = [\frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j}]$$

$$d_H(z_i^j, z_{i'}^j)$$

$$\Downarrow$$

Initial distance

$$\sum_{j=1}^{p} d_H(z_1^j, z_2^j)$$

$$\max_{j=1\ldots p} d_H(z_1^j, z_2^j)$$

Normalized distance or data table

Initial data table $(x_i^j)_{n\times p}$

$$x_i^j = [a_i^j, b_i^j]$$

$$\frac{1}{\sigma^j} \quad d_H(x_i^j, x_{i'}^j)$$

$$\Downarrow$$

Normalized distance

$$\sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j)$$

$$\max_{j=1\ldots p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j)$$

=

Normalized data table $(z_i^j)_{n\times p}$

$$z_i^j = [\frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j}]$$

$$d_H(z_i^j, z_{i'}^j)$$

$$\Downarrow$$

Initial distance

$$\sum_{j=1}^{p} d_H(z_1^j, z_2^j)$$

$$\max_{j=1\ldots p} d_H(z_1^j, z_2^j)$$

Normalized distance or data table

Initial data table $(x_i^j)_{n\times p}$

$$x_i^j = [a_i^j, b_i^j]$$

$$\frac{1}{\sigma^j} \quad d_H(x_i^j, x_{i'}^j) \quad = \quad d_H(z_i^j, z_{i'}^j)$$

$$\Downarrow \qquad\qquad\qquad \Downarrow$$

Normalized distance | Initial distance

$$\sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j) \quad = \quad \sum_{j=1}^{p} d_H(z_1^j, z_2^j)$$

$$\max_{j=1\ldots p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j) \quad = \quad \max_{j=1\ldots p} d_H(z_1^j, z_2^j)$$

Normalized data table $(z_i^j)_{n\times p}$

$$z_i^j = [\frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j}]$$

Conclusion

- General approach for the “normalization” of k-means (dynamical clustering) algorithms
- Two normalized k-means methods for hyper-rectangles clustering:

Prototype	Distance	Measure of dispersion
The “median” hyper-rectangle	Hausdorff-based	The “star” from the “median” hyper-rectang
The “middle” hyper-rectangle	L_∞-Hausdorff	The “radius” deviation from the “middle” hyper-re

- Explicit formula for prototypes with a real L_1 or L_2 Hausdorff distance between hyper-rectangles ?

Normalized k-means clustering of hyper-rectangles

M. Chavent

Laboratoire de Mathématiques Appliquées de Bordeaux, UMR CNRS 5466

Universités Bordeaux1 et 2, France

chavent@math.u-bordeaux1.fr