Symbolic Data Analysis
Taking Variability in Data into Account

Paula Brito

Fac. Economics, University of Porto & LIAAD - INESC TEC, Portugal

BIOSTAT 2025 - 25th SCHOOL OF BIOMETRICS
Varaždin, 12 June 2025

Outline

Outline

What is it ?

Symbolic data ?...

The data

Classical data analysis:
Data is represented in a $n \times p$ matrix
each of n individuals (in row) takes one single value
for each of p variables (in column)

Example: Abalone data

Abalone	Length	Diameter	...	Whole weight	Shell weight	Sex	Rings
Abalone 1	0.275	0.195	...	0.080	0.031	F	4
Abalone 2	0.290	0.210	...	0.275	0.113	F	6
Abalone 3	0.345	0.260	...	0.207	0.0775	F	11

The Data

What if ...

- Descriptors on individual Abalones, but: analyse the sex-rings classes - not each individual Abalone

The Data

What if ...

- Descriptors on individual Abalones, but: analyse the sex-rings classes - not each individual Abalone
- Descriptors on patients, but: analyse hospitals - not each individual patient

The Data

What if ...

- Descriptors on individual Abalones, but: analyse the sex-rings classes - not each individual Abalone
- Descriptors on patients, but: analyse hospitals - not each individual patient
- Descriptors on flights, but: analyse the airlines - not each individual flight

The Data

What if ...

- Descriptors on individual Abalones, but: analyse the sex-rings classes - not each individual Abalone
- Descriptors on patients, but: analyse hospitals - not each individual patient
- Descriptors on flights, but: analyse the airlines - not each individual flight
- Descriptors on purchases, but: analyse clients - not individual purchases

The Data

What if ...

- Descriptors on individual Abalones, but: analyse the sex-rings classes - not each individual Abalone
- Descriptors on patients, but: analyse hospitals - not each individual patient
- Descriptors on flights, but: analyse the airlines - not each individual flight
- Descriptors on purchases, but: analyse clients - not individual purchases
- Official statistics - Descriptors on citizens, but: analyse the cities, the regions - not the individual citizens

The data

Example: Abalone data

Abalone	Length	Diameter	...	Whole weight	Shell weight	Sex	Nb. Rings
Abalone 1	0.275	0.195	...	0.080	0.031	F	4
Abalone 2	0.290	0.210	...	0.275	0.113	F	6
Abalone 3	0.345	0.260	...	0.207	0.0775	F	11
...	...	...	...	...	...	...	...

The data

Example: Abalone data

Abalone	Length	Diameter	...	Whole weight	Shell weight	Sex	Nb. Rings
Abalone 1	0.275	0.195	...	0.080	0.031	F	4
Abalone 2	0.290	0.210	...	0.275	0.113	F	6
Abalone 3	0.345	0.260	...	0.207	0.0775	F	11
...	...	...	...	...	...	...	...

Abalone class	Length	Diameter	...	Whole weight	Shell weight
F 4-6	[0.28, 0.66]	[0.19, 0.47]	...	[0.15, 2.25]	[0.06, 1.26]
F 7-9	[0.31, 0.75]	[0.22, 0.58]	...	[0.08, 1.37]	[0.03, 0.64]
I 1-3	[0.08, 0.24]	[0.05, 0.17]	...	[0.00, 0.07]	[0.00, 0.03]
...	...	...	...	...	...

Symbolic Data Analysis (Diday (1988))

Symbolic Data Analysis:
to take explicitly into account **variability** inherent to the data

$\Longrightarrow$ (symbolic) variable values are

- sets
- intervals
- distributions on an underlying set of sub-intervals or categories

Micro-data $\longrightarrow$ **Macro-data**

The data

Data for three airline companies (e.g. arrival flights)

Airline	Nb Passengers	Delay (min)	Aircraft
A	180	10	Boeing
B	120	0	Boeing
A	200	20	Airbus
C	80	15	Embraer
B	100	5	Embraer
A	300	35	Airbus
C	70	30	Embraer
...	...	...	...

Temporal aggregation ↓

Airline	Nb. Passengers	Delay (min)	Aircraft
A	[180, 300]	{[0, 10[, 0.33; [10, 30[, 0.33; [30, 60], 0.33}	{Airbus(2/3), Boeing(1/3)}
B	[100, 120]	{[0, 10[, 1.0; [10, 30[, 0; [30, 60], 0}	{Boeing(1/2), Embraer(1/2)}
C	[70, 80]	{[0, 10[, 0; [10, 30[, 0.5; [30, 60[, 0.45; [60, 90], 0.05}	{Embraer(1)}

Sources of symbolic data: Aggregation of micro-data

Communityname	State	perCapInc	pctPoverty	persPerOccupHous	pctKids2Par
Aberdeencity	SD	11939	12,2	2,35	76,25
Aberdeencity	WA	11816	18,3	2,34	64,05
Aberdeentown	MD	13041	10,66	2,61	60,79
Aberdeentownship	NJ	19544	3,18	2,86	79,31
Adacity	OK	10491	22,93	2,21	63,11
Adriancity	MI	11006	20,65	2,61	61,92
AgouraHillscity	CA	27539	3,53	3,08	86,65
Aikencity	SC	15619	15,69	2,48	64,51
Akroncity	OH	12015	20,48	2,42	55,76
Alabastercity	AL	13645	5,65	2,94	80,57
Alamedacity	CA	19833	6,81	2,36	70,29
...	...	...	...	...	...

Contemporary aggregation $\downarrow$

State	perCapInc	pctPoverty	persPerOccupHous	pctKids2Par
ALabama	[5820, 39610]	[2, 44]	[2, 3]	[30, 90]
ARkansas	[7399, 15325]	[4, 42]	[2, 3]	[45, 81]
AriZona	[6619, 62376]	[3, 43]	[2, 4]	[57, 90]
CAlifornia	[5935, 63302]	[1, 32]	[2, 5]	[47, 90]

Symbolic Variable types

- Numerical (Quantitative) variables
  - Numerical single-valued variables
  - Numerical multi-valued variables
  - Interval variables
  - Histogram variables

- Categorical (Qualitative) variables:
  - Categorical single-valued variables
  - Categorical multi-valued variables
  - Categorical modal variables

Symbolic Variable types

$S = \{s_1, ..., s_n\}$: the set of n units to be analyzed
Let $Y_1, \ldots, Y_p$ be the variables, O_j the underlying domain of Y_j
B_j the observation space of $Y_j, j = 1, \ldots, p$

$$Y_j : S \longrightarrow B_j$$

- Y_j classical (numerical or categorical) single-valued variable : $B_j \equiv O_j$
- Y_j numerical or categorical multi-valued variable : $B_j = P(O_j)$
- Y_j interval variable : B_j set of intervals of O_j
- Y_j categorical modal or histogram variable : B_j set of distributions over O_j

Applications

In general : when it is wished to analyse data at a higher level (groups), rather than at individual level

- Official data: confidentiality issues $\rightarrow$ aggregation
- Survey data
- Big databases, e.g., purchases per client, phone calls per person, prescriptions per patient or per doctor
- Analysis of abstract concepts as such
- ...

Analysis Issues

To represent data taking into account internal variability within each observation, variables have been allowed to assume new forms.

- Are we still in the same framework when we allow for the variables to take multiple values?
- Are the definitions of basic statistical notions still so straightforward?
- What properties remain valid?

Analysis Issues

Quantitative variables:

- Evaluation of dispersion $\rightarrow$ consequences in the design of multivariate methods
- Clustering $\rightarrow$ standardization: different standardization techniques for interval-valued variables proposed
- Many methodologies rely on linear combinations and on the properties of dispersion measures under linear transformations
- How to define a linear combination of symbolic variables ?

Analysis Issues

- Different approaches considered by various authors
- Most existing methods for the analysis of such data rely on a non-parametric descriptive approach
- Some work with parametric models
  (Brito & Duarte Silva, JAS 2012 , ADAC 2025 ;
  Neto *et al*, JSCS 2011 ; Le-Rademacher and Billard, JSPI 2011)

Methods for multivariate data analysis

- Interval-valued variables: a special case of histogram-valued variables
- Methods first developed for interval-valued variables:
- Greater effort in addressing and designing methods for interval data

A large number of methods have to this day been developed for multivariate analysis, including :

- Clustering - Partitioning (crisp, fuzzy, adaptive), Hierarchical, SOM,...
- Classification - LDA, Decision tres, Neural networks,...
- Factorial analysis - PCA, Generalized canonical analysis
- Multiple Regression
- Time series analysis

Outline

Interval-Valued Variables

- $S = \{s_1, ..., s_n\}$: the set of n units to be analyzed
- $Y_1, \ldots, Y_p$: the descriptive variables

Interval-valued variable :

$Y_j : S \to B_j : \ Y_j(s_i) = [l_{ij}, u_{ij}], l_{ij} \leq u_{ij}$
B_j : the set of intervals of an underlying set $O_j \subseteq \boldsymbol{R}$

I : $n \times p$ matrix - values of p interval variables on S

Each $s_i \in S$: represented by vector of intervals,
$I_i = (I_{i1}, ..., I_{ip}), i = 1, ..., n, \ I_{ij} = [l_{ij}, u_{ij}], j = 1, \ldots, p$

Interval data

	Y_1	$\ldots$	Y_j	$\ldots$	Y_p
s_1	$[l_{11}, u_{11}]$	$\ldots$	$[l_{1j}, u_{1j}]$	$\ldots$	$[l_{1p}, u_{1p}]$
$\ldots$	$\ldots$		$\ldots$		$\ldots$
s_i	$[l_{i1}, u_{i1}]$	$\ldots$	$[l_{ij}, u_{ij}]$	$\ldots$	$[l_{ip}, u_{ip}]$
$\ldots$	$\ldots$		$\ldots$		$\ldots$
s_n	$[l_{n1}, u_{n1}]$	$\ldots$	$[l_{nj}, u_{nj}]$	$\ldots$	$[l_{np}, u_{np}]$

Examples

Albert, Barbara and Caroline are characterized by the amount of time (in minutes) they need to go to work, which varies from day to day :

	Time
Albert	$[15, 20]$
Barbara	$[25, 30]$
Caroline	$[10, 20]$

Abalones Length :

Abalone class	Length
F 4-6	$[0.28, 0.66]$
F 7-9	$[0.31, 0.75]$
I 1-3	$[0.08, 0.24]$

Interval data : Survey data application

Gender, Age, Level of Education, Job Category,
Income and debt variables - Household Income (HI), Debt to Income Ratio ($\times$ 100) (DIR), Credit Card Debt (in thousands) (CCD), Other Debts (OD)

5000 observations:

Gender	Age	Education	Job	HI	DIR	CCD	OD
Male	22	High school degree	Services	40	10	3	2
Male	45	College degree	Sales and Office	100	15	8	7
Female	30	Some college	Managerial and Professional	50	20	2	1

Individual observations aggregated on the basis of
Gender , Age Category , Level of Education and Job Category

Interval data: Survey data application

Group	HI	DIR	CCD	OD
Male, 18-24 High school degree, Service	[15, 61]	[0.1, 23.4]	[0.0, 6.57]	[0.02, 7.71]
Male, 35-49, College degree, Sales and Office	[19, 190]	[1.4, 20.4]	[0.04, 16.6]	[0.12, 15.39]
Female, 25-34, Some college Managerial and Professional	[17, 100]	[0.8, 31.7]	[0.05, 6.57]	[0.09, 7.65]

Native Interval Data

Temperatures and pluviosity measured in 283 meteorological stations in the USA:
temperature ranges in January and July, annual pluviosity range

Station	State	January Temperature	July Temperature	Annual Pluviosity
HUNTSVILLE	AL	$[32.3, 52.8]$	$[69.7, 90.6]$	$[3.23, 6.10]$
ANCHORAGE	AK	$[9.3, 22.2]$	$[51.5, 65.3]$	$[0.52, 2.93]$
NEW YORK (JFK)	NY	$[24.7, 38.8]$	$[66.7, 82.9]$	$[2.70, 4.13]$
...	...	...	...	...
SAN JUAN	PR	$[70.8, 82.4]$	$[76.9, 87.4]$	$[2.14, 6.17]$

Also: description of botanical species, specific diseases,...

Interval Data

Analysis of Interval Data: two main approaches

- Assuming a distribution within each observed interval - usually the Uniform, but not necessarily (e.g. Triangular)
- Represent an interval by two real numbers
  - the lower and upper bounds
  - the midpoint and (half) range

  and develop models and methods using these two values

The Abalone data set

Data comes from an original study:

Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn, and Wes B Ford (1994)

"The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait"

Original microdata: 4177 instances, 8 attributes (one nominal), no missing attribute values (UCI repository)

Units of interest: the 24 sex $\times$ nb rings classes, F 4-6, F 7-9, ...

Variables:

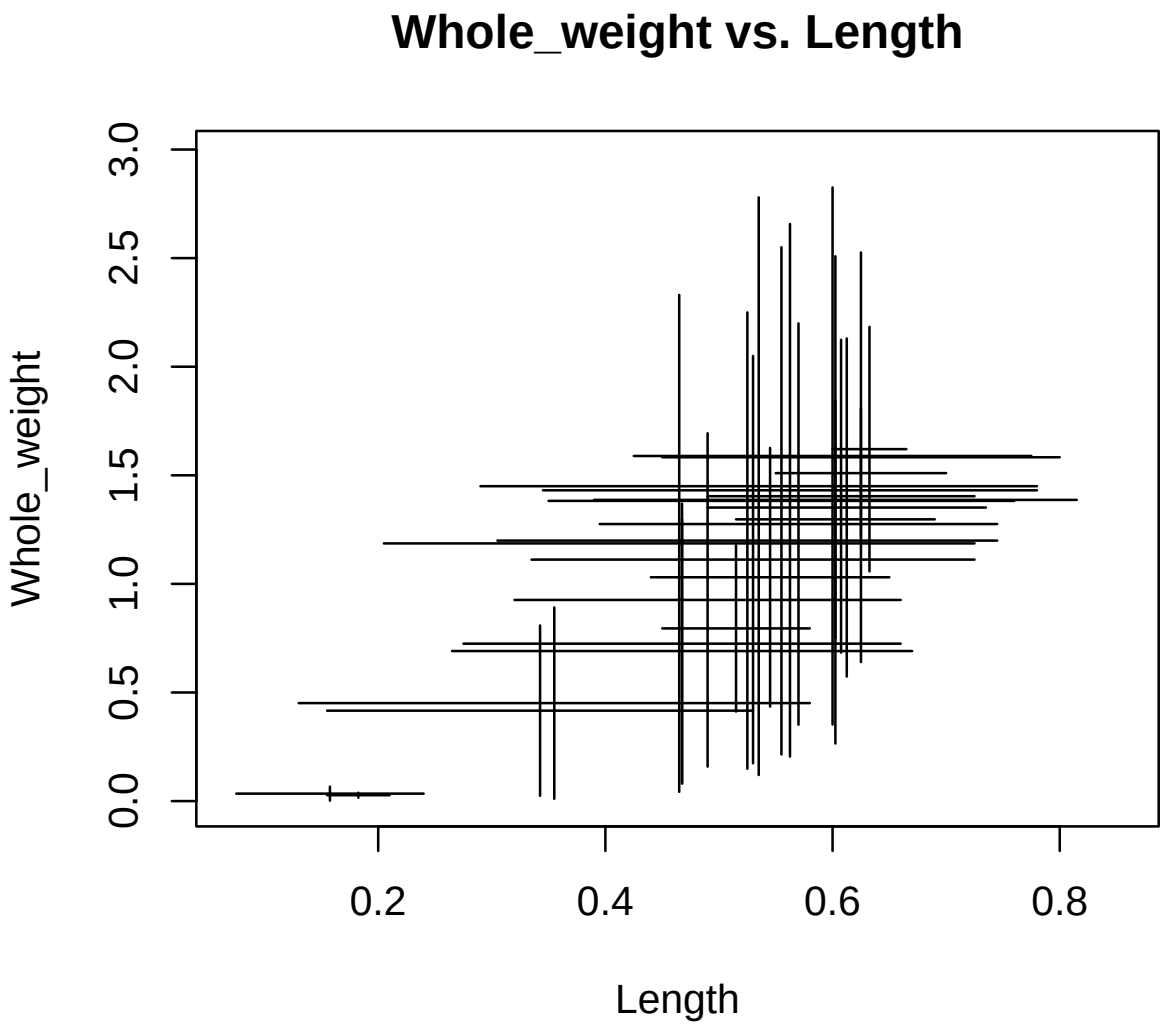
- Length
- Diameter
- Height
- Whole weight
- Shucked weight
- Viscera weight
- Shell weight

The Abalone data set

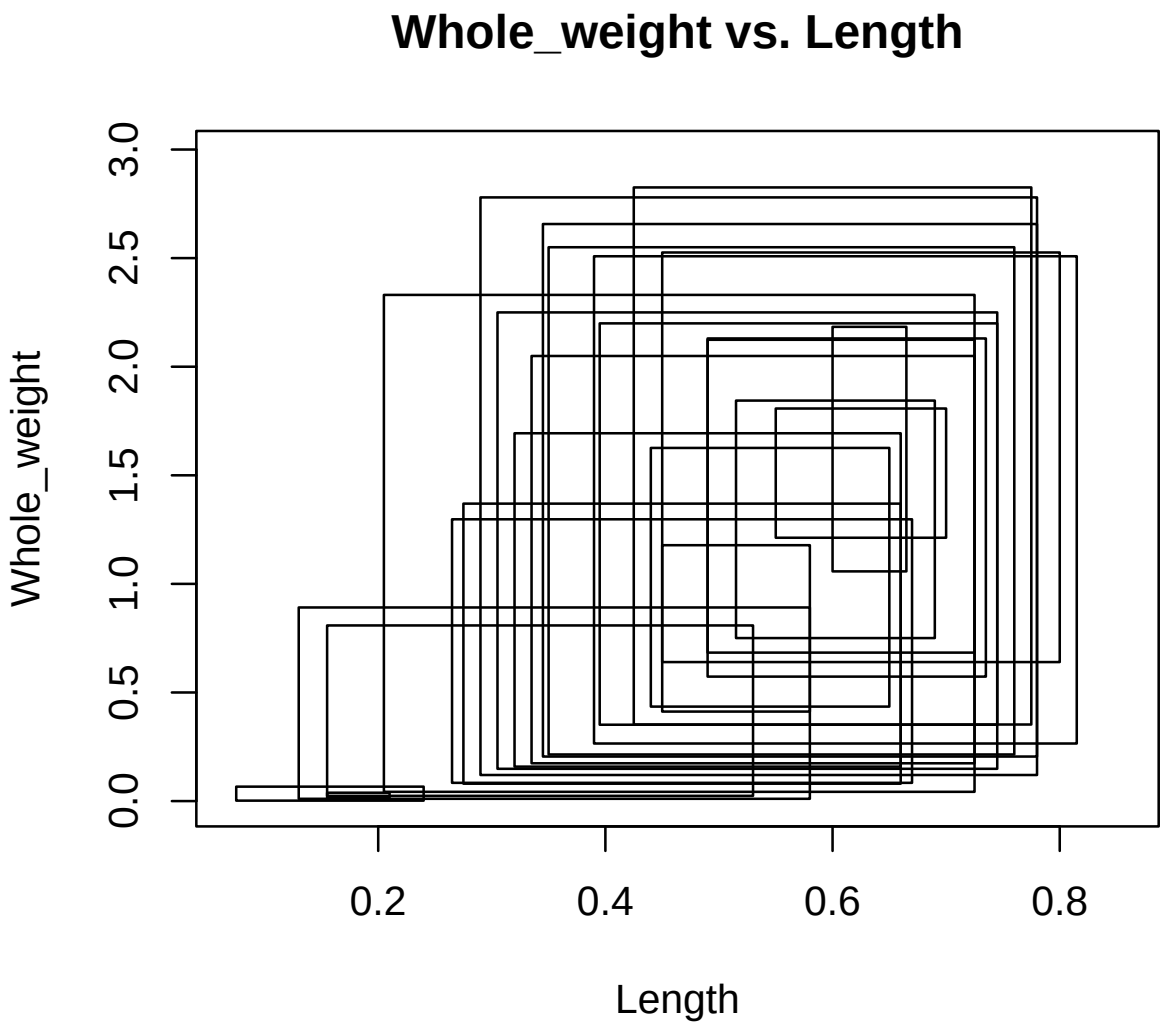
The values of these variables are aggregated in the form of intervals for each Abalone class, defined by sex $\times$ ring-class - now the units of interest

Abalone class	Length	Diameter	...	Whole weight	Shell weight
F 4-6	[0.28, 0.66]	[0.19, 0.47]	...	[0.15, 2.25]	[0.06, 1.26]
F 7-9	[0.31, 0.75]	[0.22, 0.58]	...	[0.08, 1.37]	[0.03, 0.64]
I 1-3	[0.08, 0.24]	[0.05, 0.17]	...	[0.00, 0.07]	[0.00, 0.03]
...	...	...	...	...	...

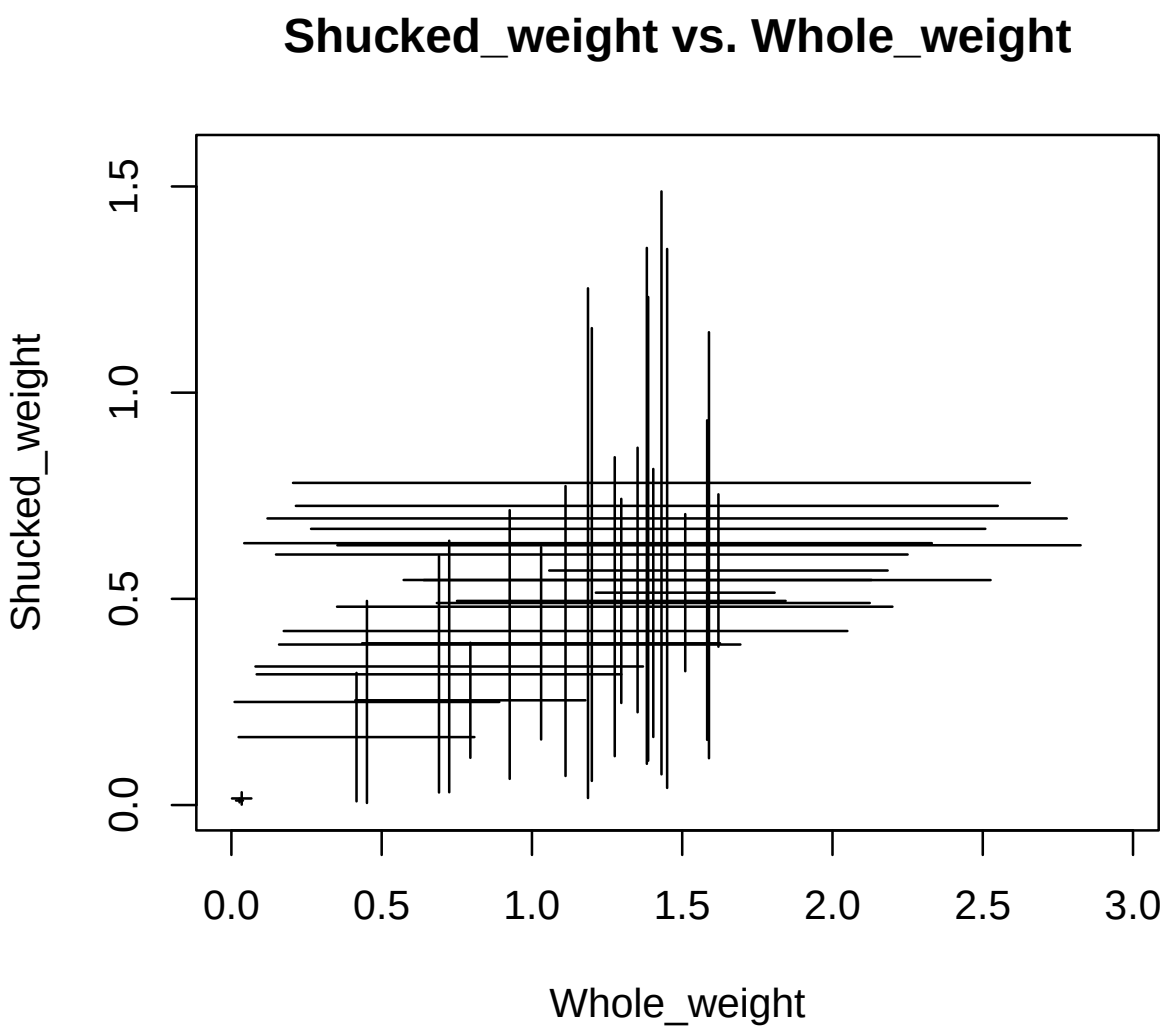
The Abalone data set



The Abalone data set

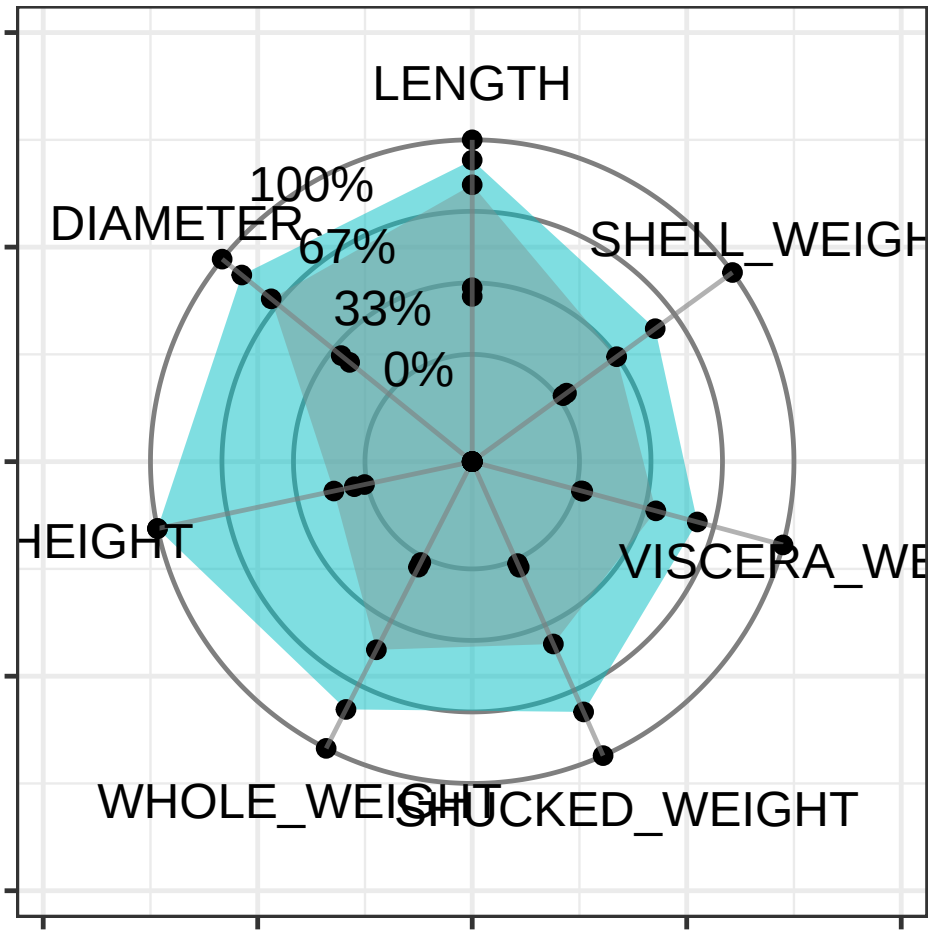


The Abalone data set



The Abalone data set

Radar plot



Descriptive Statistics for Interval variables

Assumming an Uniform distribution within each interval $Y_k(s_i)$, $i = 1, \ldots, n$, $l_{ik} = [l_{ik}, u_{ik}]$, $k = j, j'$ we have

- **Symbolic sample mean** :

$$\overline{Y_k} = \frac{1}{2n} \sum_{i=1}^{n} (l_{ik} + u_{ik}) = \frac{1}{n} \sum_{i=1}^{n} c_{ik}$$

- **Symbolic sample variance** :

$$S^2_{Y_k} = \frac{1}{3n} \sum_{i=1}^{n} (l^2_{ik} + l_{ik} u_{ik} + u^2_{ik}) - \overline{Y_k}^2$$

Bertrand and Goupil (2000)

Descriptive Statistics for Interval variables

Covariance

Billard & Diday (2003), obtained from the empirical joint density function:

$$Cov(Y_j, Y_{j'}) = \frac{1}{4n}\sum_{i=1}^{n}(l_{ij} + u_{ij})(l_{ij'} + u_{ij'}) - \overline{Y_j}.\overline{Y_{j'}}$$

New definition in Billard (2008), considering a decomposition into Within observations Sum of Products (WithinSP) and Between observations Sum of Products (BetweenSP):

Descriptive Statistics for Interval variables

$$Cov(Y_j, Y_{j'}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{(u_{ij}-l_{ij})(u_{ij'}-l_{ij'})}{12}}_{\text{WithinSP}} +$$

$$+\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{l_{ij}+u_{ij}}{2}-\overline{Y_j}\right)\left(\frac{l_{ij'}+u_{ij'}}{2}-\overline{Y_{j'}}\right)}_{\text{BetweenSP}}$$

$$= \frac{1}{6n}\sum_{i=1}^{n}[2(l_{ij}-\overline{Y_j})(l_{ij'}-\overline{Y_{j'}}) + (l_{ij}-\overline{Y_j})(u_{ij'}-\overline{Y_{j'}}) + (u_{ij}-\overline{Y_j})(l_{ij'}-\overline{Y_{j'}}) + 2(u_{ij}-\overline{Y_j})(u_{ij'}-\overline{Y_{j'}})]$$

Descriptive Statistics for Interval variables

Abalone data:

mean, standard deviation, coefficient of variation

$$\mathbf{m} = \begin{pmatrix} L & D & H & WW & ShW & VW & SW \\ 0.507 & 0.396 & 0.167 & 1.08 & 0.456 & 0.224 & 0.337 \end{pmatrix}$$

$$\mathbf{s} = \begin{pmatrix} L & D & H & WW & ShW & VW & SW \\ 0.1606 & 0.1327 & 0.1265 & 0.6663 & 0.3069 & 0.1449 & 0.2159 \end{pmatrix}$$

$$\mathbf{cv} = \begin{pmatrix} L & D & H & WW & ShW & VW & SW \\ 31.68\% & 33.50\% & 75.74\% & 61.70\% & 67.31\% & 64.69\% & 64.06\% \end{pmatrix}$$

Interval-valued variables: Distance measures

Many measures proposed in the literature to compare two intervals I_i, I_j

- **Hausdorff distance**
  $d_H(I_i, I_j) = \max\{\{|l_i - l_j|, |u_i - u_j|\}$
- **Euclidean distance**
  $d_2(I_i, I_j) = \sqrt{(l_i - l_j)^2 + (u_i - u_j)^2}$
- **City-Block distance**
  $d_1(I_i, I_j) = |l_i - l_j| + |u_i - u_j|$

Considering a distribution F within the interval, with $\Psi = F^{-1}$:

- **Wasserstein distance**
  $D_W(I_i, I_j) = \int_0^1 \left|\Psi_{I_i}(t) - \Psi_{I_j}(t)\right| dt$
- **Mallows distance**
  $D_M(I_i, I_j) = \sqrt{\int_0^1 (\Psi_{I_i}(t) - \Psi_{I_j}(t))^2 dt}$

Clustering Interval-valued data

- K-means-like approaches - De Carvalho and co-workers (2004 - ...)
  - Different distances considered
  - Also: Adaptive distances
  - Also: Multiple dissimilarity matrices
  - Using Hausdorff distance - Chavent & Lechevalllier (2002)
- Fuzzy clustering
  - El-Sonbaty, Ismail (1998)
  - Yang, Hwang and Chen (2004)
  - D'Urso and Giordani (2006)
  - De Carvalho *et al* (2007, 2010)
  - Jeng, Chuang, Tseng and Juan (2010)
- SOM approaches:
  - Bock *et al* (2002)
  - De Carvalho *et al* (2011)
  - Hajjar and Hamdan (2011)
  - Yang, Hung, Chen (2012)

SCLUST: Dynamical clustering for symbolic data (De Carvalho et al. (2008))

SCLUST : non-hierarchical clustering on symbolic data, using a k-means - or dynamical clustering - like method

- Starting from a partition on a pre-fixed number of clusters
- alternates an assignment step (based on minimum distance to cluster prototypes)
- and a representation step (which determines new protoypes in each cluster)
- until convergence is achieved (or a pre-fixed number of iterations is reached)

SCLUST: Dynamical clustering for symbolic data

- The method locally optimizes a criterion that measures the fit between cluster propotypes and cluster members
- which is additive, and based on the assignment-distance function
- The method allows for all types of variables in the input data
- Selects the distances for the assigning step accordingly:
  - Quantitative real-valued data: Euclidean distance
  - Interval and quantitative multi-valued data: Hausdorff distance
  - Categorical single-valued data: χ-square distance
  - Categorical multi-valued data: De Carvalho distance
  - Distributional data: a classical ϕ^2 distance

SCLUST includes functions for the determination of the appropriate number of clusters, based on classical indices
(see Hardy, (2008))

Clustering Interval-valued data: SCLUST

Hausdorff distance

Partition in three clusters

Class: 1 Cardinal : 4

I 1-3 , I 4-6 , M 1-3 , M 4-6

Class: 2 Cardinal : 14

F 7-9 , F 10-12 , F 13-15 , F 16-18 , F 22-24 , F 19-21, F 25-29 , M 7-9 , M 10-12 , M 13-15, M 16-18 , M 19-21 , M 22-24 , M 25-29

Class: 3 Cardinal : 6

I 7-9 , I 13-15 , I 10-12 , I 16-18 , I 19-21

Principal Component Analysis of Interval data

- Factorial method for interval data
- Graphical representation by rectangles in the factorial planes
- Interpretation indices

Principal Component Analysis of Interval data

Direct solution:

- use covariance / correlation values obtained for the interval-valued variables
- proceed as for real (classical) data
- obtain interval representations in the new variables' space - applying the linear combinations to lower and upper bounds of the observed interval data

Principal Component Analysis of Interval data

Two first methods

(Cazes, P., Chouakria, A., Diday, E., Schektman, Y. (1997)) :

- Vertices method:
  Analysis of the data array with $n \times 2^p$ rows and p columns containing all vertices of the hyper-rectangles in the interval data-array
- Centres method:
  Represents each interval by its midpoint

See also: Chouakria, A., Billard, L., Diday, E. (2011)

Principal Component Analysis of Interval data: Vertices method

Example:
if s_i has the description $([a, b], [c, d])$, s_i will be "transformed" in :

$$\begin{bmatrix} a & c \\ b & c \\ a & d \\ b & d \end{bmatrix}$$

A classical PCA of the new $(n \times 2^p)$ rows $\times p$ columns data array is performed.

Weight of $s_i = \frac{p_i}{2^p}$ (p_i : original weight of s_i)

Principal Component Analysis of Interval data: Vertices method

Let Y_j^* be the les q first principal components, $j = 1, \ldots, q$

For each principal component Y_j^*:

the minimum value $min_j(s_i)$ and the maximum value $max_j(s_i)$ of Y_j^* are determined, across the 2^p rows representing each entity s_i

We then define: $Y_j^*(s_i) = [min_j(s_i), max_j(s_i)]$

Principal Component Analysis of Interval data: Vertices method

Interpretation indices: Quality of representation

G : centre of gravity of the $n \times 2^p$ points ; d : Euclidean distance
L_i : 2^p points representing s_i

Average squared cosines of the angles between the vertices in L_i and the axis of the j^{th} principal component:

$$COR^1(i, v_j) = \sum_{k \in L_i} \frac{Y_{kj}^{*2}}{d^2(k, G)}$$

Ratio between the contribution of all the vertices in L_i to the variance λ_j of the j^{th} principal component and their contribution to the total inertia (or total variance).

$$COR^2(i, v_j) = \frac{1}{2^p} \frac{\sum_{k \in L_i} Y_{kj}^{*2}}{\sum_{k \in L_i} d^2(k, G)}$$

Principal Component Analysis of Interval data: Vertices method

Interpretation indices: Contributions

Contribution of s_i to the variance of the j^{th} principal component:

$$CTR(i, v_j) = \frac{\sum_{k \in L_i} q_k Y_{kj}^{*2}}{\lambda_j} = \frac{p_i}{2^p \lambda_j} \sum_{k \in L_i} Y_{kj}^{*2}$$

Contribution of s_i to the total inertia:

$$INR(i) = \frac{p_i}{2^p} \frac{\sum_{k \in L_i} d^2(k, G)}{\sum_j \lambda_j}$$

Principal Component Analysis of Interval data: Centres method

Only the intervals' midpoints are used:

if s_i has the description $([a, b], [c, d])$, s_i will be "transformed" in :

$$c_i = \left(\frac{a+b}{2}, \frac{c+d}{2} \right)$$

A classical PCA of the new $n \times p$ data array is performed

x_{ij}^c : midpoint of variable Y_j for entity s_i (midpoint c_i)

$\overline{x_j^c}$: mean of the midpoints of variable Y_j

$v_\ell = (v_{1\ell}, \ldots, v_{p\ell})$: the ℓ^{th} eigenvector

The ℓ^{th} principal component of centre c_i is given by:

$$Y_{i\ell}^{*c} = \sum_{j=1}^{p} (x_{ij}^c - \overline{x_j^c}) v_{j\ell}$$

Principal Component Analysis of Interval data: Centres method

I_{ij} : interval corresponding to unit s_i and variable Y_j

Interval corresponding to s_i for the ℓ^{th} principal component:

$$I_{ij}^* = [\underline{I_{ij}^*}, \overline{I_{ij}^*}]$$

$$\underline{I_{ij}^*} = \sum_{j=1}^{p} \text{Min}_{x \in Iij}\{(x - \overline{x_j^c})v_{j\ell}\}$$

$$\overline{I_{ij}^*} = \sum_{j=1}^{p} \text{Max}_{x \in Iij}\{(x - \overline{x_j^c})v_{j\ell}\}$$

CIPCA - Complete Information Principal Component Analysis of Interval data

Wang *et al* (2012), Neurocomputing

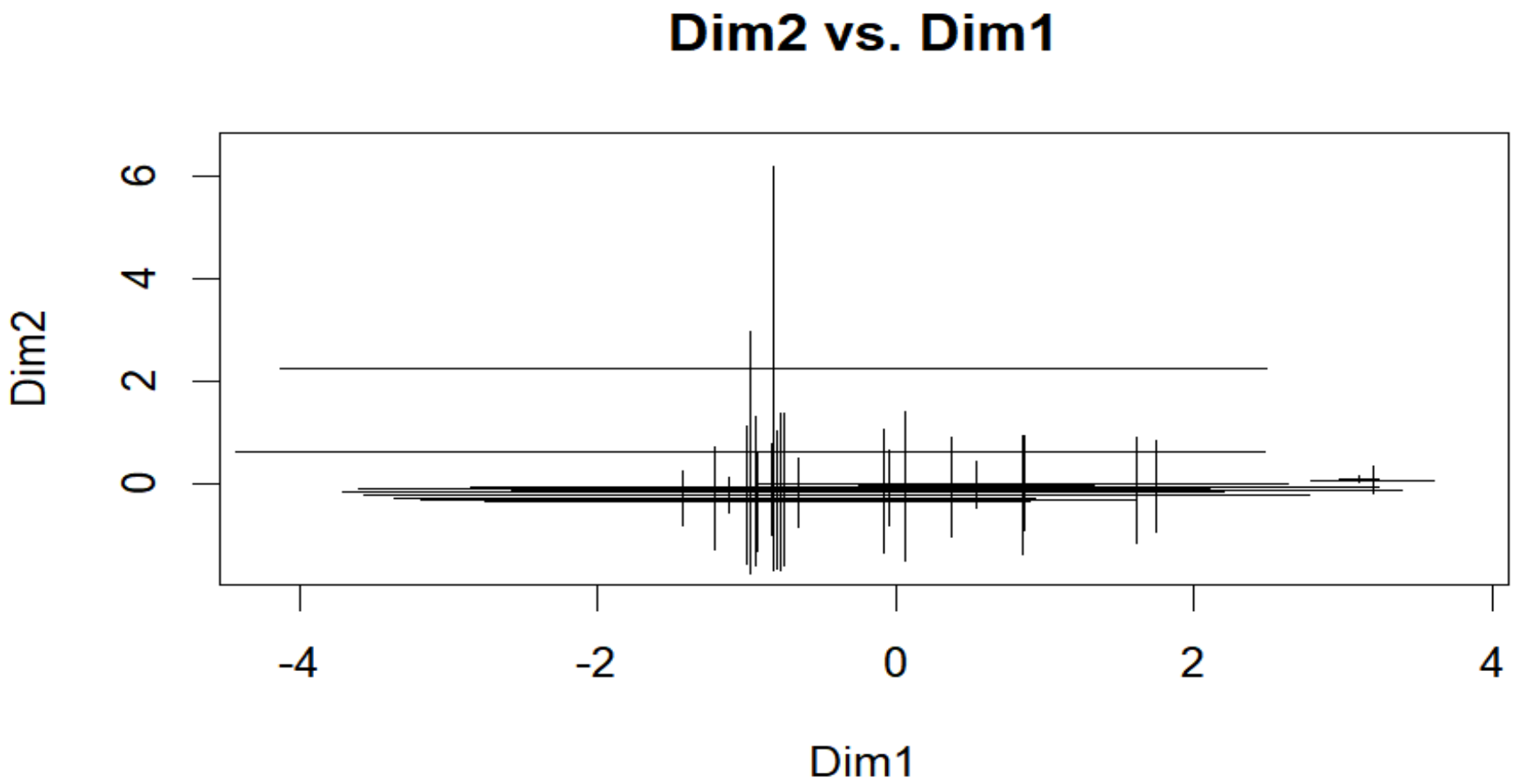
Objective : use the interval as such, and not only boundary points

Relies on assuming a Uniform distribution within each observed interval

- compute the covariance / correlation matrix for the interval-valued variables
- determine eigenvalues and eigenvectors
- obtain interval representations in the new variables' space - applying the linear combinations defined by interval arithmetic (Moore, 1966)

Principal Component Analysis of Interval data

Vertices method



Principal Component Analysis of Interval data

Other methods for PCA of Interval data:

- Lauro, C., Palumbo, F. (2000). Principal component analysis of interval data: a symbolic data analysis approach.
  Introduce a label matrix to code vertices belonging to the same hypercube
- Palumbo, F., Lauro, C. (2003). A PCA for interval valued data based on midpoints and radii.
- D'Urso, P., Giordani, P. (2004). A least squares approach to principal component analysis for interval valued data.
  PCA on Midpoints and Radii, using a least squares approach.

Principal Component Analysis of Interval data

- Gioia, F., Lauro, C. (2006). Principal component analysis on interval data.
  Based on Interval algebra and optmization ; uses interval-valued covariance matrix.
- Le-Rademacher, J., Billard, L. (2012). Symbolic covariance principal component analysis and visualization for interval-valued data
  Use symbolic covariance (based on within and between SUMSQ), new visualization by polytopes (from the convex hull).

Linear Regression for Interval-valued variables

- Methods based in symbolic covariance definitions (Billard and Diday, 2000; 2006; Xu, 2010)
- Minmax Method (Billard and Diday, 2002)
- Center and Range Method (Lima Neto and De Carvalho, 2008)
- Center and Range Least Absolute Deviation Regression Method (Maia and De Carvalho, 2008)
- Constrained Center and Range Method (Lima Neto and De Carvalho, 2010)
- Lasso IR Method (Giordani, 2014)
- Bivariante Symbolic Regression Models (Lima Neto *et al*, 2011)
- Linear Regression Models for Symbolic Interval Data Using Pso Algorithm (Yang *et al*, 2011)
- Monte Carlo Method (Ahn et al, 2012)
- Radial Basis Function Networks (Su *et al*, 2012)
- Copula Interval Regression Method (Neto *et al*, 2012)
- Interval Distributional Model (Dias and Brito, 2017)

Parametric models for interval data

Most existing methods: non-parametric descriptive approaches
Our goal: parametric inference methodologies
$\longrightarrow$ probabilistic models for interval variables

For each s_i, $Y_j(s_i) = I_{ij} = [l_{ij}, u_{ij}]$ is naturaly defined
by the lower and upper bounds l_{ij} and u_{ij}

For modeling purposes $\rightarrow$ preferable equivalent parametrization:

Represent $Y_j(s_i)$ by

- the midpoint $c_{ij} = \dfrac{l_{ij} + u_{ij}}{2}$
- the range $r_{ij} = u_{ij} - l_{ij}$

Parametric Models for interval data

Gaussian model:

Assume that the joint distribution of the midpoints C and the logs of the ranges R is multivariate Normal:

$$R^* = ln(R), (C, R^*) \sim N_{2p}(\mu, \Sigma)$$

$$\mu = [\mu_C^t, \mu_{R^*}^t]^t \; ; \; \Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{pmatrix}$$

μ_C and μ_{R^*} - p-dimensional column vectors of the mean values

$\Sigma_{CC}, \Sigma_{CR^*}, \Sigma_{R^*C}$ and $\Sigma_{R^*R^*}$ - $p \times p$ matrices

Model advantage:
Straightforward application of classical inference methods

Parametric Models for interval data

- Intervals' midpoints: location indicators $\rightarrow$ assuming a joint Normal distribution corresponds to the usual Gaussian assumption
- Log transformation of the ranges $\rightarrow$ to cope with their limited domain

This model implies:

- marginal distributions of the midpoints are Normals
- marginal distributions of the ranges are Log-Normals
- specific relation between mean, variance and skewness for the ranges

Parametric Models for interval data

More general models that try to alleviate limitations of the multivariate Normal distribution

Skew-Normal model:

Assume that the joint distribution of the midpoints C and the logs of the ranges R is multivariate Skew-Normal:

$(C, R^*) \sim SN_{2p}(\xi, \Omega, \alpha)$

Skew-Normal distribution (Azzalini, 1985):

- Generalizes the Gaussian distribution
- Introducing additional shape parameters
- Preserves some of its mathematical properties
- Alternative parametrization (traditional moments): $SN_{2p}(\mu, \Sigma, \gamma_1)$ (Arellano-Valle & Azzalini, 2008)

Density of a p-dimensional Skew-Normal distribution:

$$f(y;\alpha,\xi,\Omega) = 2\phi_p(x-\xi;\Omega)\Phi(\alpha^t\omega^{-1}(x-\xi)), x \in \boldsymbol{R}^p$$

ξ and α are p-dimensional vectors, Ω is a symmetric $p \times p$ positive-definite matrix,

ω is a diagonal matrix formed by the square-roots of the diagonal elements of Ω

ϕ_p is the density of a p-dimensional standard Gaussian vector
Φ is the distribution function of a standard normal variable

Parametric Models for interval data

However, for interval data:

Midpoint c_{ij} and Range r_{ij} of the value of an interval-valued variable are two quantities related to one only variable
$\rightarrow$ should not be considered separately

So: parameterizations of the global covariance matrix $\rightarrow$
take into account the link that may exist between midpoints and log-ranges of the same or different variables

Models for interval data

Most general formulation: allow for non-zero correlations among all midpoints and log-ranges; other cases of interest:

- The interval variables Y_j are non-correlated, but for each variable, the MidPoint may be correlated with its Log-Range;
- MidPoints (respectively, Log-Ranges) of different variables may be correlated, but no correlation between MidPoints and Log-Ranges is allowed;
- All Midpoints and Log-Ranges are non-correlated.

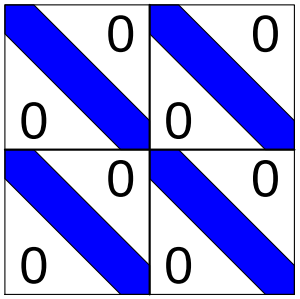
Models for interval data

Config.	Characterization	Σ
1	Non-restricted	Non-restricted
2	Y_j's non correlated	$\Sigma_{CC}, \Sigma_{CR^*} = \Sigma_{R^*C}, \Sigma_{R^*R^*}$ all diagonal
3	C's non-correlated with R^*'s	$\Sigma_{CR^*} = \Sigma_{R^*C} = 0$
4	All C's and R^*'s are non-correlated	Σ diagonal

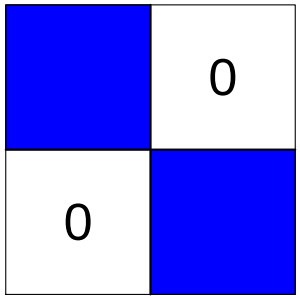
Parametric Models for interval data

$$\Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{pmatrix}$$

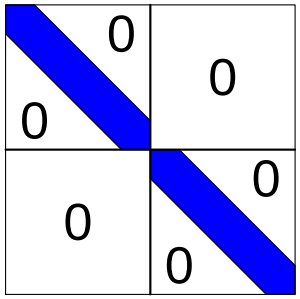
Configuration 1



Configuration 2



Configuration 3



Configuration 4

Models for interval data

- Configurations C2 and C3 are a particular case of C1
- Configuration C4 is a particular case of all the others

In cases C2, C3 and C4, Σ can be written as a block diagonal matrix

- Configuration 2: there are p 2×2 blocks
- Configuration 3: the matrix Σ is formed by two $p \times p$ blocks,
- Configuration 4: the $2p$ blocks are single real elements

Parametric analysis of interval data: ML estimation

Gaussian model:

For all configurations,

$$ln\ L(\mu, \Sigma) =$$
$$-np\ ln(2\pi) - \frac{n}{2}\ ln\left|\Sigma\right| - \frac{1}{2}\ trE\Sigma^{-1} - \frac{n}{2}\left(\bar{X} - \mu\right)^t \Sigma^{-1}\left(\bar{X} - \mu\right)$$

Σ^{-1} is symmetric positive definite $\Rightarrow$
maximum-likelihood estimate of the mean vector is always $\bar{X}$

Maximization of the likelihood function with respect to Σ reduces to maximizing

$$ln\ L(\mu, \Sigma) = \text{ constant } - \frac{n}{2}\ ln\left|\Sigma\right| - \frac{1}{2}\ trE\Sigma^{-1}$$

Parametric analysis of interval data: ML estimation

Configurations 2, 3 and 4, Σ is subject to the constraints

In these cases Σ can be written as a block diagonal matrix, after a possible rearrangement of rows and columns

The maximum can be obtained by separately maximizing with respect to each block of Σ

Parametric analysis of interval data: ML estimation

Skew-Normal model:

Log-likelihood of a p-dimensional Skew-Normal distribution:

$l = \text{ constant } - \frac{1}{2} n ln\,|\Omega| - \frac{n}{2} tr(\Omega^{-1} V) + \sum_i \zeta_0(\alpha^t \omega^{-1}(x_i - \xi))$ where
$V = n^{-1} \sum_i (x_i - \xi)(x_i - \xi)^t$ and $\zeta_0(x) = ln(2\Phi(x))$

Configuration 1 (Azzalini and Capitanio, 1999):

the log-likelihood can be re-parametrized as
$l = \text{ constant } - \frac{1}{2} n ln\,|\Omega| - \frac{n}{2} tr(\Omega^{-1} V) + \sum_i \zeta_0(\eta(x_i - \xi))$

Then, for each ξ and η the log-likelihood is maximized on Ω by $\hat{\Omega} = V$

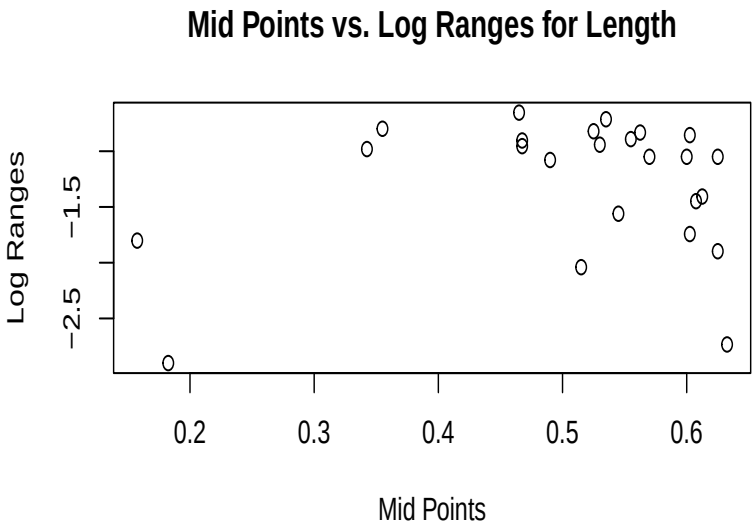
Parametric analysis of interval data: ML estimation

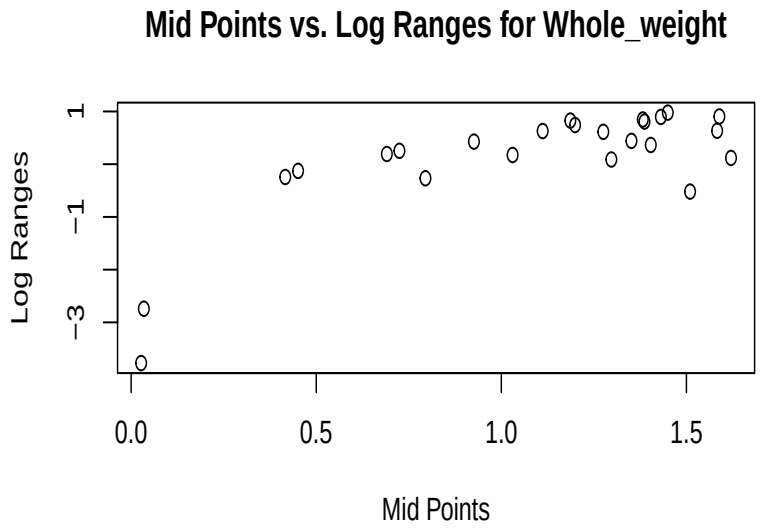
Other configurations:

Let $\theta = (\xi, \Omega, \eta) = \theta(\psi)$ with $\psi = (\mu, \Sigma, \gamma_1)$

We maximize numerically the log-likelihood of $\theta(\psi)$ using as arguments the free elements of μ, Σ, γ_1, subject to admissibility restrictions

The Abalone data set





(a) MidPoints vs Log-Ranges of Length
(b) MidPoints vs Log-Ranges of Whole weight

Parametric analysis of Interval data: ML estimation for the Abalone data set

Model Config.	Normal C1	Normal C2	Normal C3	Normal C4	SkN C1	SkN C2	SkN C3	SkN C4
BIC	**-338.30**	308.72	-261.52	386.28	-322.40	299.00	-217.03	379.03

$$\hat{\mu}^{\mathbf{t}} = \left(\begin{array}{ccccccc|ccccccc} & & & C & & & & & & & R^* & & & \\ L & D & H & WW & ShW & VW & SW & L & D & H & WW & ShW & VW & SW \\ 0.51 & 0.40 & 0.17 & 1.08 & 0.46 & 0.22 & 0.34 & -1.30 & -1.54 & -2.24 & 0.10 & -0.70 & -1.44 & -1.18 \end{array} \right)$$

$$\hat{\sigma}^{\mathbf{t}} = \left(\begin{array}{ccccccc|ccccccc} & & & C & & & & & & & R^* & & & \\ L & D & H & WW & ShW & VW & SW & L & D & H & WW & ShW & VW & SW \\ 0.13 & 0.11 & 0.10 & 0.46 & 0.20 & 0.10 & 0.16 & 0.60 & 0.65 & 0.87 & 1.10 & 1.14 & 1.11 & 1.23 \end{array} \right)$$

Models for interval data

These models allow for multivariate parametric analysis of interval-valued data:

- Robust estimation and Outlier detection (Gaussian model)
- (M)ANOVA
- Linear and Quadratic Discriminant Analysis
- Model-based Clustering (Gaussian model)

- R package MAINT-Data - available at CRAN

- Brito, P., Duarte Silva, A. P. (2012). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, 39(1), 3-20.
- Duarte Silva, A.P., Brito, P., Filzmoser, P., Dias, J. (2021). MAINT. Data: Modelling and analysing interval data in R. *R Journal*, 13(2), 336-364.

Multivariate outlier detection

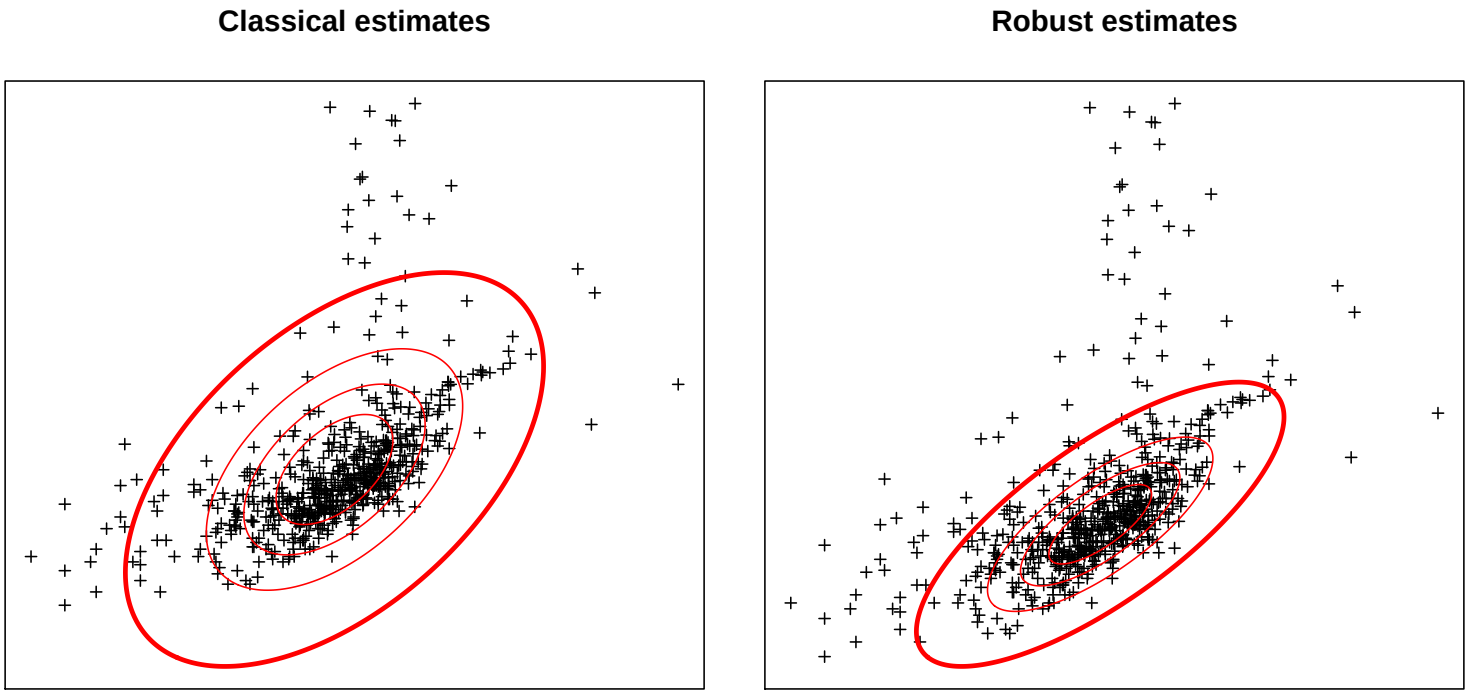
Outlier detection:

Outliers will typically have large distance:

If multivariate normal distribution is assumed

$\Longrightarrow$ MD_i^2 is approx. χ^2_d distributed.

$\Longrightarrow$ suspect observations: $MD_i^2 > \chi^2_{d,0.975}$



Methodology

Outliers in (multivariate) interval data can be identified by:

- Represent interval data as **midpoints** C and **ranges** R.
- Assume $(C, \ln(R)) \sim N(\mu, \Sigma)$; possibly restrict Σ.
- Use robust parameter estimation $\longrightarrow \hat{\mu}$, $\hat{\Sigma}$
- Compute robust Mahalanobis distances based on $\hat{\mu}$, $\hat{\Sigma}$
- Interpret multivariate outliers based on EDA graphics.

Robust parameter estimation

Idea: use a **trimmed version** of the complete-data log likelihood, i.e. replace $\sum_{i=1}^{n}$ by a trimmed sum,
using **Trimmed Likelihood Estimators (TLE)**.

Basic idea behind trimming: removal of those observations whose values would be highly unlikely to occur if the fitted model was true.

Gaussian data:
Minimum Covariance Determinant (MCD) method and Weighted Trimmed Likelihood lead to the same estimators of covariance

Robust Mahalanobis Distance

The TLE is applied to each of the **specific configuration** (structures of Σ), resulting in robust estimates of μ and Σ.

Afterwards: compute **robust** Mahalanobis distances based on these estimates.

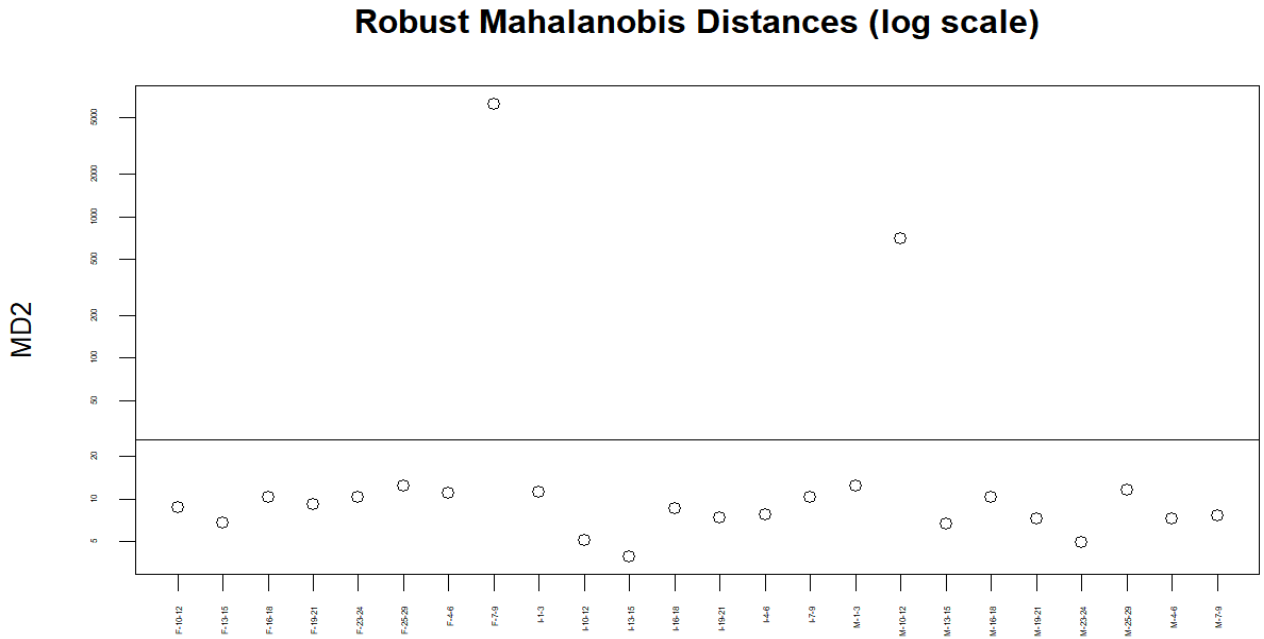
Refinements of TLE

- One step re-weighted estimation of location and scatter
- Small sample covariance-bias correction

  Pison *et al* (2002) approach replicated for all covariance configurations

- Automatic selection of trimming parameter, using a two-step procedure
- Mahalanobis distances distributions assumed as:
  - Classical chi-square asymptotic approximations OR
  - F and Beta finite sample approximations (Hardin & Rocke (2005); Cerioli (2010))

Multivariate outlier detection



Outliers: F-7-9 ; M-10-12

Duarte Silva A.P., Filzmoser, P., Brito, P. (2018)
Outlier Detection in Interval Data. *ADAC*, 12(3), 785–822.

(M)ANOVA for interval-valued variables

Each interval-valued variable Y_j is modelled by the pair $(C_j, R_j^*) \Rightarrow$ analysis of variance of Y_j : two-dimensional MANOVA of (C_j, R_j^*)

Simultaneous analysis of all the Y's may be accomplished by a $2p$ dimensional MANOVA, following the same procedure

Gaussian and Skew-Normal model

(M)ANOVA for interval-valued variables

General case :

$\rightarrow$ **Likelihood ratio approach**

Maximize the log-likelihood for the null (mean/location vectors equal across groups) and the alternative hypothesis

Likelihood ratio statistic

$$\lambda = \frac{L_{null}}{L_{alt}}$$

L_{null} and L_{alt} are the maximum log-likelihoods
under the null (mean/location vectors equal across groups)
and alternative (mean/location vectors different) hypothesis

In all cases, under the null hypothesis, $-2ln\lambda$ follows asymptotically a chi-square distribution

(M)ANOVA for interval-valued variables

Simulation study:

When sample sizes are not too small:

- Tests have good power
- True significance level approaches nominal levels when the constraints assumed for the model are respected
- Method assuming data is Normal with configuration 1 (non-restricted) never performs worse than any other method when data is indeed Normal
- Skew Normal model requires large samples

(M)ANOVA: Example

The 22 (24 minus the 2 outliers) abalone classes have been gathered in 2 groups according to the number of rings :

- Young: Nb. Rings ≤ 15
- Adult: Nb. Rings > 15

Global MANOVA to assess whether the two groups are different

MODEL	-2 lnλ	DF	P-VALUE
NORM 1	42.74	14	9.41×10^{-5}

Small-sample size $\rightarrow$ **Permutation test:** p-value $= 0.065$

(M)ANOVA for variable "Length"

MODEL	-2 lnλ	DF	P-VALUE
NORM 1	12.69	2	1.79×10^{-5}

Permutation test: p-value ≈ 0.0

Discriminant Analysis

Gaussian model:

For each configuration, an estimate of the optimum classification rule can be obtained with the corresponding Σ

Direct generalisation of the classical linear and quadratic discriminant classification rules:

Linear :

$$Y = argmax_g(\hat{\mu_g}^t \hat{\Sigma}^{-1} X - \frac{1}{2} \hat{\mu_g}^t \hat{\Sigma}^{-1} \hat{\mu_g} + log\ \hat{\pi_g})$$

Quadratic :

$$Y = argmax_g(-\frac{1}{2} X^t \hat{\Sigma_g}^{-1} X + \hat{\mu_g}^t \hat{\Sigma_g}^{-1} X + log\ \hat{\pi_g} - \frac{1}{2}(log\ det \hat{\Sigma_g} + \hat{\mu_g}^t \hat{\Sigma_g}^{-1} \hat{\mu_g}))$$

Discriminant Analysis

Skew-Normal model:

Three different alternatives may be considered:

1. the groups differ only in terms of μ;
2. the groups differ in terms of both μ and Σ;
3. the groups differ in terms of μ, Σ and γ_1.

Discriminant Analysis

Skew-Normal model:
Considering cases 1) and 3) :

Location :

$$Y = argmax_g(\hat{\xi_g}^t \hat{\Omega}^{-1} X - \frac{1}{2}\hat{\xi_g}^t \hat{\Omega}^{-1}\hat{\xi_g} + log\ \hat{\pi_g} + \zeta_0(\hat{\alpha}^t \hat{\omega}^{-1}(X - \hat{\xi_g})))$$

General :

$$Y = argmax_g(-\frac{1}{2}X^t \hat{\Omega_g}^{-1} X + \hat{\xi_g}^t \hat{\Omega_g}^{-1} X + log\ \hat{\pi_g} - \frac{1}{2}(log\ det\, \hat{\Omega_g} + \hat{\xi_g}^t \hat{\Omega_g}^{-1}\hat{\xi_g}) + \zeta_0(\hat{\alpha_g}^t \hat{\omega_g}^{-1}(X - \hat{\xi_g})))$$

Discriminant Analysis

Discriminating the 2 Abalone groups:

Error rate estimates of algorithm LDA (LOO)

Adult	Young	Global
0.50	0.25	0.36

Error rate estimates of algorithm QDA (LOO)

Adult	Young	Global
1.00	0.33	0.64

Duarte Silva A.P., Brito, P. (2015).
Discriminant Analysis of Interval Data: An Assessment of Parametric and Distance-Based Approaches.
Journal of Classification, Volume 32, Issue 3, 516-541.

Model-Based Clustering

Finite-mixture model:

$$f(x_i;\varphi) = \sum_{\ell=1}^{k} \pi_\ell f_\ell(x_i;\Theta_\ell),$$

Maximum likelihood (ML) parameter estimation $\rightarrow$ maximization of the log-likelihood function:

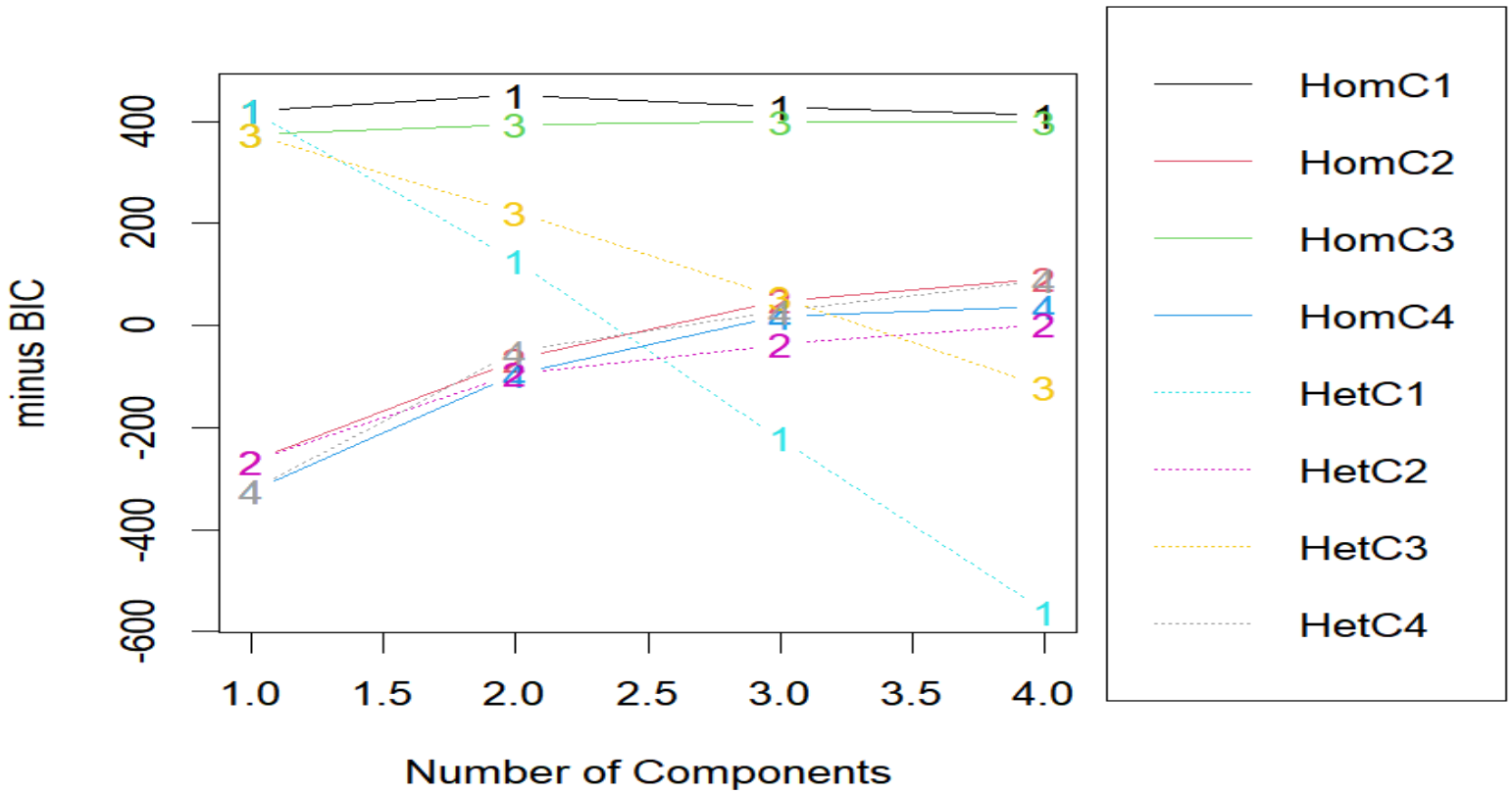
$$\ell(\varphi;\mathbf{x}) = \sum_{i=1}^{n} \ln f(\mathbf{x}_i;\varphi)$$

Expectation-Maximization (EM) algorithm

Trying to avoid local optima $\rightarrow$ each search of the EM algorithm is replicated from different starting points

Selection of the **model** and **number of components** (K) $\rightarrow$ Bayesian Information Criterion : BIC$= -2\ell(\hat{\varphi};\mathbf{x}) + d_\varphi ln(n)$

Clustering the Abalones



Clustering the Abalones

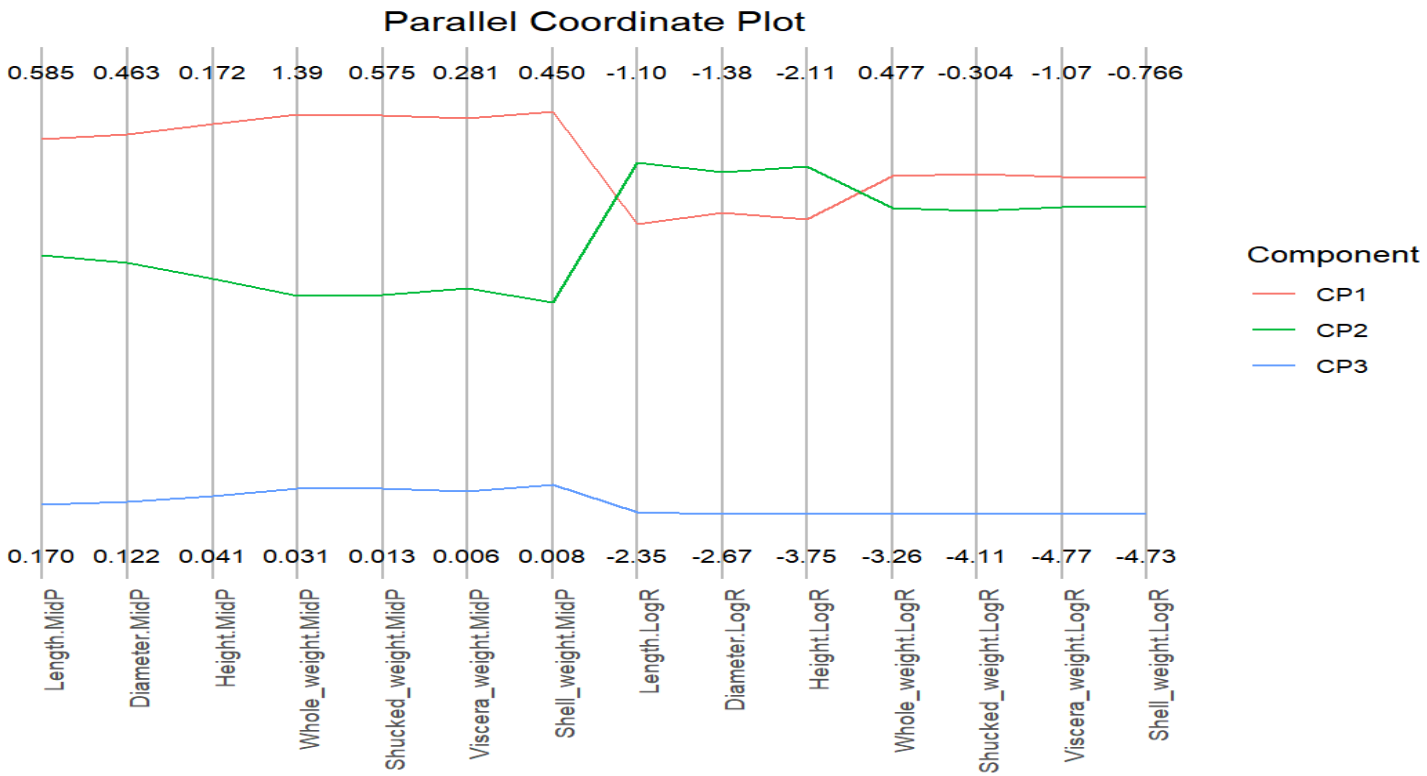
Component 1 - 20 Abalone classes
Component 2 - 2 Abalone classes : "I-1-3" "M-1-3"

Indicator	CP1	CP2
Length.MidP	0.539	0.17000
Diameter.MidP	0.422	0.12250
Height.MidP	0.153	0.04125
Whole-weight.MidP	1.158	0.03062
Shucked-weight.MidP	0.480	0.01350
Viscera-weight.MidP	0.237	0.00638
Shell-weight.MidP	0.371	0.00838
Length.LogR	-1.244	-2.35112
Diameter.LogR	-1.478	-2.66957
Height.LogR	-2.273	-3.75280
Whole-weight.LogR	0.354	-3.25668
Shucked-weight.LogR	-0.447	-4.10854
Viscera-weight.LogR	-1.188	-4.77291
Shell-weight.LogR	-0.883	-4.73262

Clustering the Abalones

In 3 components:

Component 1 - 13 Abalone classes
Component 2 - 27 Abalone classes
Component 3 - 2 Abalone classes : "I-1-3" "M-1-3"



Outline

Distribution-Valued Data

Keeping more information (requires more data at the micro level)

Example : Data for three airline companies

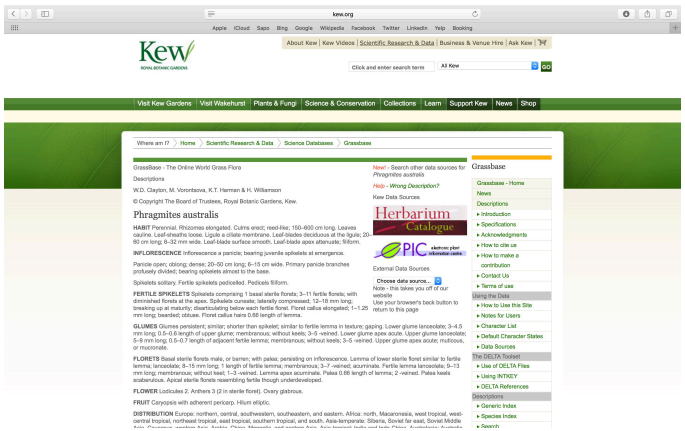
Airline	Delay (min)	Aircraft
A	$\{[0, 10[, 0.33; [10, 30[, 0.33; [30, 60], 0.33\}$	$\{\text{Airbus}(2/3), \text{Boeing}(1/3)\}$
B	$\{[0, 10[, 1.0; [10, 30[, 0; [30, 60], 0\}$	$\{\text{Boeing}(1/2), \text{Embraer}(1/2)\}$
C	$\{[0, 10[, 0; [10, 30[, 0.5; [30, 60[, 0.45; [60, 90], 0.05\}$	$\{\text{Embraer}(1)\}$

Study I - Relation between grass height and temperature

Joint work with M. Hoffmann, Martin Luther Univ., Halle-Wittenberg, Germany

Goal: Study if there are global morphological trends in grasses that relate to climate factors.
In particular, study if morphological characters like grass height are related with the temperature.

Data source: GrassBase in http://www.kew.org/data/grasses-db.html
Climate data - WorldClim data base



Original data and SDA contribution

Original data:

- The database reports the occurrences of more than 10,000 species in 295 regions of the two hemispheres of the world.
- Minimum and maximum values for the plant height were reported, e.g., the height of *Phragmites australis* ranges from 150 cm to 600 cm.
- Mean temperatures in countries or country-like regions.

Original data and SDA contribution

Original data:

- The database reports the occurrences of more than 10,000 species in 295 regions of the two hemispheres of the world.
- Minimum and maximum values for the plant height were reported, e.g., the height of *Phragmites australis* ranges from 150 cm to 600 cm.
- Mean temperatures in countries or country-like regions.

Ecologists - Classical approach:

- The interest is to know, e.g., what is the size distribution of height across the world and whether taller plants occur in the tropics.
- For this study the mid-range values and the centroid of the countries are considered. A significant amount of information is lost due to the application of mid-range values.

Original data and SDA contribution

Original data:

- The database reports the occurrences of more than 10,000 species in 295 regions of the two hemispheres of the world.
- Minimum and maximum values for the plant height were reported, e.g., the height of *Phragmites australis* ranges from 150 cm to 600 cm.
- Mean temperatures in countries or country-like regions.

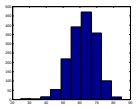
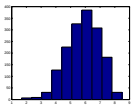
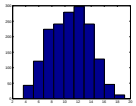
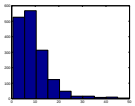
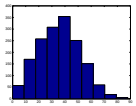
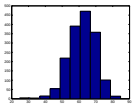
Ecologists - Classical approach:

- The interest is to know, e.g., what is the size distribution of height across the world and whether taller plants occur in the tropics.
- For this study the mid-range values and the centroid of the countries are considered. A significant amount of information is lost due to the application of mid-range values.

The contribution of SDA for Ecological studies:

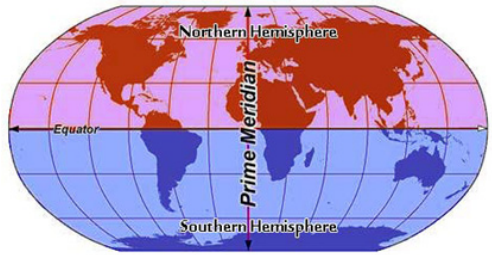
- For each country we recorded the min and max values of all grass species and built a histogram. Each species: two data points in the histogram.
- For each country or region: histogram of the mean temperatures.

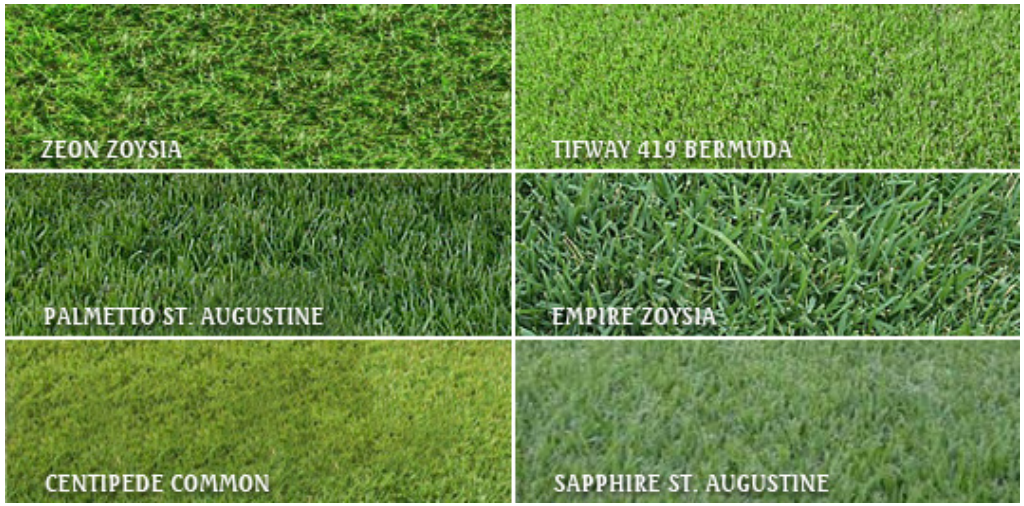
Symbolic data

Countries/Country-regions	*GrassHeight*	*Temp*
Albania		
Canada		
Portugal		
⋮	⋮	⋮

- For all observations the subintervals of each histogram have the same weight (equiprobable) with frequency 0.10.
- Units:

  235 countries/country-like regions of the northern hemisphere and 60 of the southern hemisphere

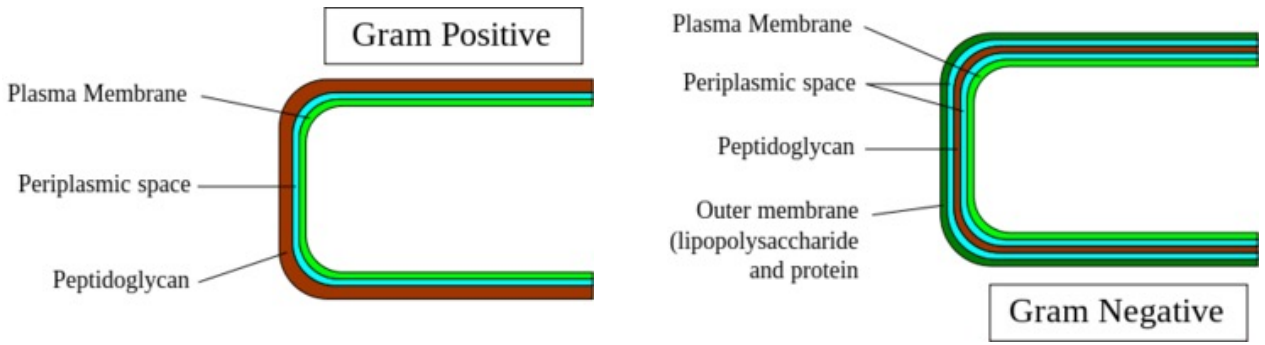




- How is it possible to study the linear relation/correlation between the grass height and the temperature, considering that the observations of these variables are histograms?
- Is the relation between grass height and temperature different in the northern and southern hemisphere?
- Is the grass height related to the atmospheric temperature of the country or country-like region?

Study II - Gram Positive *versus* Gram Negative bacteria

Goal: Classify a set of 23 species of bacteria considering the complete genome sequence of each species



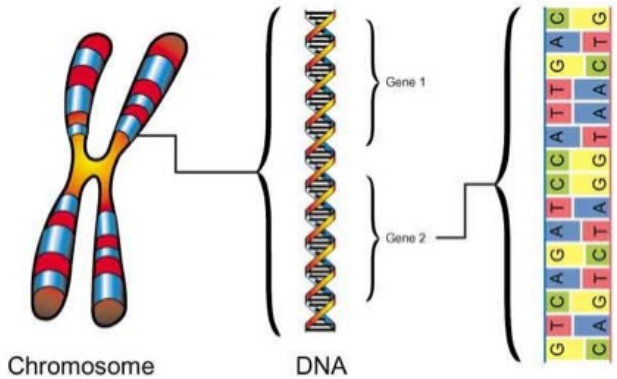
The structural difference - Gram-positive bacteria do not have an outer cell membrane found in Gram-negative bacteria.

Contribution: Classify the bacterial species according to the frequencies of each nucleotide in the genome or in a given position of the codon.

Microdata: Genomic data of 23 bacterial species

Joint work with Adelaide Freitas, Dep. Mathematics & CIDMA, Univ. Aveiro

Complete genome sequences of 11 Gram-positive and 12 Gram-negative bacteria (Sorimachi and Okayasu, 2004).
Downloaded from NCBI Genbank (https://www.ncbi.nlm.nih.gov/nuccore/).



Genoma: the collection of genetic information.
Gene: basic unit of genetic information. A sequence of codons composed by sets of three nucleotides: A - Adenine; C - Cytosine; G - Guanine; T- Thymine.

Complete genoma of 23 bacterial species

- Staphylococcus aureus Mu50 (2699 genes; Gram-positive)
- Streptococcus pyogenes M1 (1693 genes; Gram-positive)
- Bacillus subtilis (4185 genes; Gram-positive)
- Clostridium perfringens 13 (2660 genes; Gram-positive)
- Listeria monocytogenes (2846 genes; Gram-positive)
- Mycoplasma pulmonis (782 genes; mycoplasma)
- Mycoplasma genitalium (476 genes; mycoplasma)
- Mycoplasma pneumoniae (688 genes; mycoplasma)
- Ureaplasma urealyticum (646 genes; mycoplasma)
- Mycobacterium tuberculosis (4189 genes; Gram-positive)
- Mycobacterium leprae (1605 genes; Gram-positive)
- Rickettsia prowa-zekii (834 genes; Gram-negative)
- Borrelia burgdorferi (753 genes; Gram-negative)
- Campylobacter jejuni (1653 genes; Gram-negative)
- Helicobacter pylori 26695 (1566 genes; Gram-negative)
- Helicobacter pylori J99 (1491 genes; Gram-negative)
- Escherichia coli (4140 genes; Gram-negative)
- Salmonella typhi (4395 genes; Gram-negative)
- Vibrio cholerae (3828 genes; Gram-negative)
- Yersinia pestis (3885 genes; Gram-negative)
- Neisseria meningitides (1909 genes; Gram-negative)
- Haemophilus influenzae (1709 genes; Gram-negative)
- Treponema pallidum (1031 genes; Gram-negative)

Gene sequencing of two bacteria

Escherichia coli

```
>gi|16445223|ref|NC_002655.2| Escherichia coli O157:H7 str. EDL933 chromosome, complete
genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTCTCTGACAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCACCACCATCACCATTACCATTACCACAGGTAACGGTGCGGGCTGACGCGTAC
AGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTCGACCAAAGGTAACGAGGTAACA
ACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGGGTTGCCGATA
TTCTGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCA
CCTGGTGGCGATGATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGT
ATTTTTGCCGAACTTCTGACGGGACTCGCCGCCGCCCAGCCGGGGATTCCCGCTGGCGCAATTGAAAACTT
TCGTCGACCAGGAATTTGCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTAGGGCAGTGCCCGGA
TAGCATTAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAA
GCGCGCGGTCACAACGTTACCGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAAT
CTACTGTCGATATTGCAGAGTCCACCCGCCGTATTGCGGCAAGTCGTATTCCGGCTGATCACATGGTGCT
GATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTACTTGGACGCAACGGTTCCGACTAC
TCCGCGGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTAT
ATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAATCGATGTCCTACCAGGAAGCGATGGA
GCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCT
```

Bacillus cereus

```
>gi|30018278|ref|NC_004722.1| Bacillus cereus ATCC 14579, complete genome
TAGCCACTTTTTTTTGATATTATAGTTGTGTTTTCACTTTGAATAAGTTTTCCACATCTTTATCTTATCC
ACAATTTGTGTATAACATGTGGACAGTTTTAATCACATGTGGGTAAATAGTTGTCCACATTTGCTTTTTT
TGTCGAAAACCCTTTCTCAATACAAACGACGTTTTAGGTTTTAAAATCGTTTTCGTATAAATATACATTT
TAATTTATTAGGTTGTACATTTGTTGCACAACCTTTATTCTTTTACCAACTTAGTAAAGGAGGGACACCT
TTGGAAAATATCTCTGATTTATGGAATAGTGCCTTAAAAGAATTAGAAAAAAAGGTAAGCAAGCCTAGTT
ATGAGACATGGTTAAAATCCACAACGGCTCATAACTTGAAGAAAGACGTATTAACGATTACAGCTCCGAA
TGAATTTGCTCGTGACTGGCTAGAATCTCATTACTCCGAACTAATTTCAGAAACACTATACGATTTAACA
GGGGCAAAATTAGCAATTCGCTTTATTATTCCCCAAAGTCAAGCTGAAGAGGACATTGATCTGCCTCCAG
TTAAGCCGAATCCAGCACAAGATGATTCAGCTCATTTACCACAGAGCATGTTAAATCCAAAATATACATT
CGATACATTTGTTATTGGCTCTGGTAACCGTTTTGCCCATGCAGCTTCATTAGCTGTAGCTGAGGCGCCA
GCTAAAGCGTATAATCCACTCTTTATTTACGGGGGAGTTGGACTTGGAAAGACACATTTAATGCACGCAA
TTGGTCATTATGTAATTGAACATAATCCAAATGCAAAAGTTGTATATTTATCATCAGAAAAATTTACAAA
TGAATTTATTAACTCTATTCGTGATAATAAAGCTGTTGATTTTCGTAATAAATATCGTAACGTAGATGTT
TTATTGATAGATGATATTCAATTTCTTGCTGGAAAAGAACAGACTCAAGAAGAGTTTTTCCATACATTTA
ACGCATTACACGAAGAAAGTAAACAAATTGTAATTTCTAGTGACCGACCACCAAAAGAAATTCCAACTTT
```

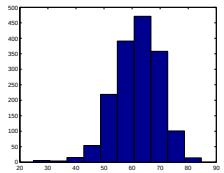
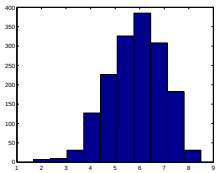
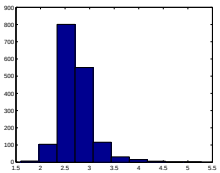
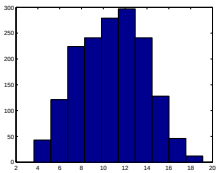
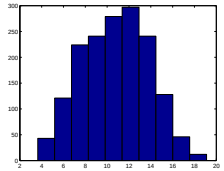
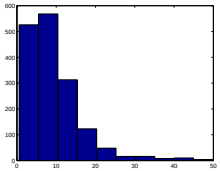
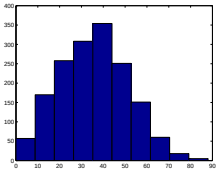
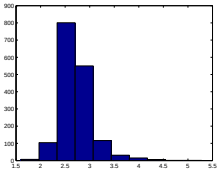
Microdata

Species	Genes	Gene sequencing	$\%A$	$\%T$	$\%G$	$\%C$
Staphylococcus aureus	Gene 1	$ATC\ AAG\ TCT\cdots$	30	20	60	40
	Gene 2	$GTC\ TAG\ ACG\cdots$	25	15	40	30
	⋮	⋮	⋮	⋮	⋮	⋮
	Gene n_1	$CCT\ GAG\ ATG\ldots\ldots$	12	68	12	8
Streptococcus pyogenes	Gene 1	$CCT\ GAG\ ATT\cdots$	32	25	18	25
	Gene 2	$AGC\ CAA\ CTG\cdots$	15	20	60	5
	⋮	⋮	⋮	⋮	⋮	⋮
	Gene n_2	$GGA\ TCG\ TTG\ldots$	30	44	6	20
Bacillus subtilis	Gene 1	$ACT\ TTG\ GAT\ldots$	12	18	60	10
	Gene 2	$TCG\ GCA\ ATG\cdots$	21	29	15	35
	⋮	⋮	⋮	⋮	⋮	⋮
	Gene n_3	$AGA\ CGT\ ACT\ldots$	43	12	35	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮

First level units: thousands of genes of the 23 bacterial species

Observations associated with each first level unit: The frequency of each nucleotide associated with each genes in each bacterial specie.

Macrodata - Case I

Species	X_1	X_2	X_3	X_4
Staphylococcus aureus				
Streptococcus pyogenes				
⋮	⋮	⋮	⋮	⋮

Case I: classify the bacteria as Gram + *vs* Gram - from the distributions of the four nucleotide frequencies in the genes.

- X_1 : distribution of the frequencies of nucleotide A in each species
- X_2 : distribution of the frequencies of nucleotide C in each species
- X_3 : distribution of the frequencies of nucleotide G in each species
- X_4 : distribution of the frequencies of nucleotide T in each species

Macrodata: Case II

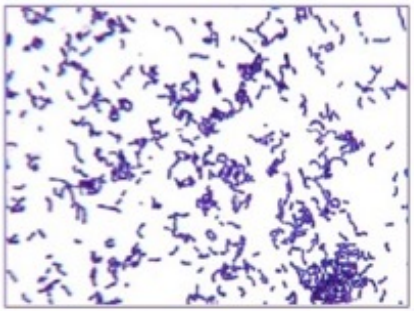
Case II (for each position - 3 studies): classify the bacteria as Gram + *vs* Gram - from the distributions of the four nucleotide frequencies at the first/second/third codon position.

Y_1 : distribution of the frequencies of nucleotide A in first/second/third codon position in each species

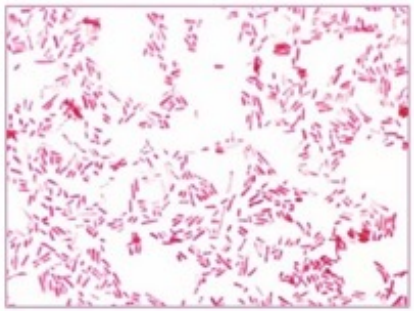
Y_2 : distribution of the frequencies of nucleotide C in first/second/third codon position in each species

Y_3 : distribution of the frequencies of nucleotide G in first/second/third codon position in each species

Y_4 : distribution of the frequencies of nucleotide T in first/second/third codon position in each species



Gram-Positive Bacteria



Gram-Negative Bacteria

Staphylococcus aureus	Rickettsia prowa-zekii
Streptococcus pyogenes	Borrelia burgdorferi
Bacillus subtilis	Campylobacter jejuni
Clostridium perfringens	Helicobacter pylori 26695
Listeria monocytogenes	Helicobacter pylori J99
Mycobacterium tuberculosis	Escherichia coli
Mycobacterium leprae	Salmonella typhi
Mycoplasma pulmonis	Vibrio cholerae
Mycoplasma genitalium	Yersinia pestis
Mycoplasma pneumoniae	Neisseria meningitides
Ureaplasma urealyticum	Haemophilus influenzae
	Treponema pallidum

- Which study better allows classifying the bacteria as Gram + or Gram -?
- How is the discriminant symbolic function used to calculate the scores?
- What is the distance used in the classification process of the bacteria in Gram + or Gram - ?

Theoretical concepts needed

- Develop a linear regression model that allows working with distribution/interval data
- Develop a discriminant method that allows working with distribution/interval data
- Use the discriminant function to classify observations in one of two groups

Histogram-valued variables

Histogram-valued variable : $Y : S \to B$

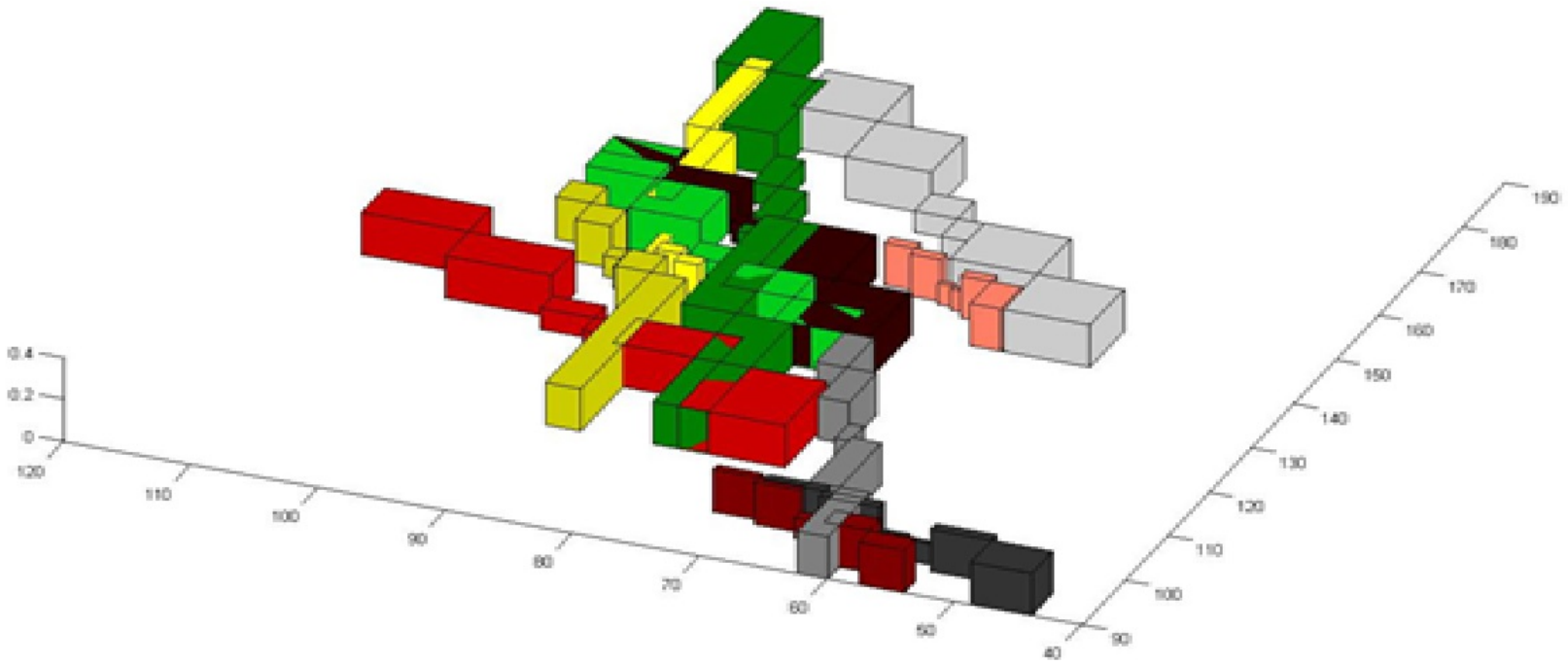
B: the set of all possible partitions of any compact of $\boldsymbol{R}$ and all possible distributions over the (finite set of) corresponding sub-intervals.

$$Y(s_i) = H_{Y(s_i)} = ([\underline{l}_{i1}, \overline{l}_{i1}[, p_{i1}; \ldots; [\underline{l}_{im_i}, \overline{l}_{im_i}], p_{im_i})$$

$p_{i\ell}$: probability or frequency associated to $l_{i\ell} = [\underline{l}_{i\ell}, \overline{l}_{i\ell}[$
$p_{i1} + \ldots + p_{im_i} = 1$

- Assumption: within each sub-interval $[\underline{l}_{i\ell}, \overline{l}_{i\ell}[$ the values of variable Y for unit s_i are uniformly distributed
- Interval-valued variables: particular case of histogram-valued variables: $Y(s_i) = [l_i, u_i] \to H_{Y(s_i)} = ([l_i, u_i], 1)$

Biplots for Histogram variables



Histogram-valued variables

$Y(s_i)$ can, alternatively, be represented by the inverse cumulative distribution function - quantile function

$\Psi : [0,1] \longrightarrow \mathbb{R}$

$$\Psi_{Y(i)}(t) = \begin{cases} c_{Y(i)_1} + \left(\frac{2t}{w_{i1}} - 1\right) r_{Y(i)_1} & \text{if} \quad 0 \leq t < w_{i1} \\ c_{Y(i)_2} + \left(\frac{2(t-w_{i1})}{w_{i2}-w_{i1}} - 1\right) r_{Y(i)_2} & \text{if} \quad w_{i1} \leq t < w_{i2} \\ \vdots & \\ c_{Y(i)_{m_i}} + \left(\frac{2(t-w_{i(m_i-1)})}{1-w_{i(m_i-1)}} - 1\right) r_{Y(i)_{m_i}} & \text{if} \quad w_{i(m_i-1)} \leq t \leq 1 \end{cases}$$

where $w_{ih} = \sum_{\ell=1}^{h} p_{i\ell}, h = 1, \ldots, m_i; r_{i\ell} = \overline{l}_{i\ell} - \underline{l}_{i\ell}$, for $\ell = \{1, \ldots, m_i\}$.

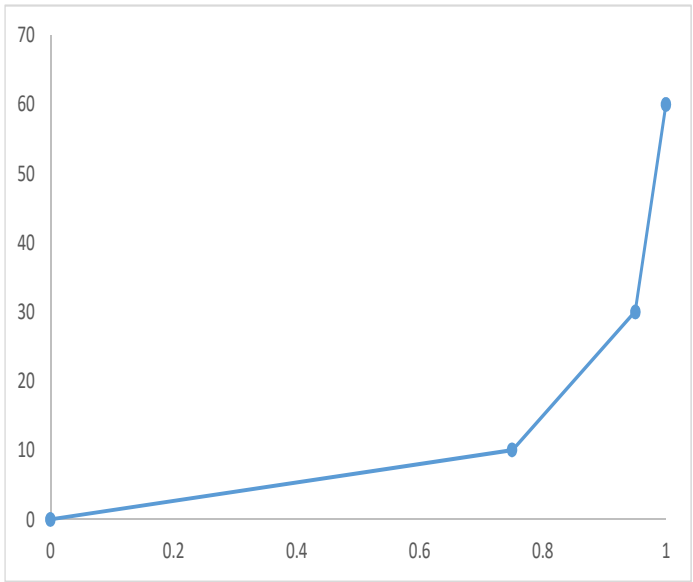
These are piecewise linear functions.

Histogram-valued variables: Example

Airline	Delay (minutes)
C	$\{[0, 10[\,, 0.75; [10, 30[\,, 0.20; [30, 60]\,, 0.05\}$

The associated quantile function is

$$\Psi(t) = \begin{cases} \frac{t}{0.75} \times 10 = \frac{40t}{3}\,, & 0 \leq t < 0.75 \\ 10 + \frac{t-0.75}{0.20} \times 20 = , & 0.75 \leq t < 0.95 \\ = 100t - 65 & \\ 30 + \frac{t-0.95}{0.05} \times 30 = , & 0.95 \leq t \leq 1 \\ = 600t - 540 & \end{cases}$$



Histogram-valued variables: Distance measures

- **Wasserstein distance** :
  $D_W(\Psi_{Y(i)}, \Psi_{Y(i')}) = \int_0^1 \left|\Psi_{Y(i)}(t) - \Psi_{Y(i')}(t)\right| dt$

- **Mallows distance**:
  $D_M(\Psi_{Y(i)}, \Psi_{Y(i')}) = \sqrt{\int_0^1 (\Psi_{Y(i)}(t) - \Psi_{Y(i')}(t))^2 dt}$

Under the uniformity hypothesis,
and considering a fixed weight decomposition
(same weights, different intervals),
we have (Irpino and Verde, 2006):

$$D_M^2(\Psi_{Y(i)}, \Psi_{Y(i')}) =$$
$$= \sum_{\ell=1}^{m} p_\ell \left[(c_{Y(i)} - c_{Y(i')})^2 + \frac{1}{3}(r_{Y(i)} - r_{Y(i')})^2 \right]$$

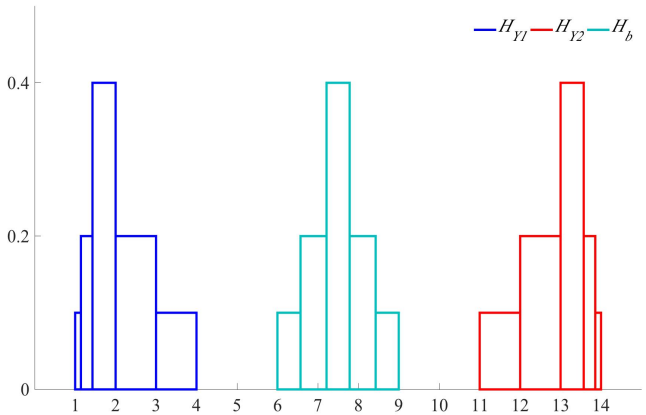
Descriptive Statistics for Histogram Variables: Barycenter

The Mallows **barycentric histogram** is the solution of the minimization problem

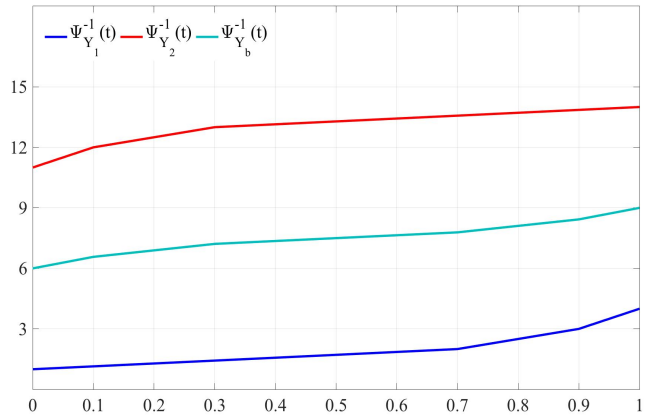
$$\min \quad \sum_{i=1}^{n} D_M^2(\Psi_{Y(i)}(t), \Psi_{Y_b}(t))$$

$\rightarrow$ quantile function where the centers and half ranges of each subinterval ℓ are the classical mean of the centers and half ranges of all observations

Histograms :



Quantile functions :



Histogram-valued variables: Mallows distance properties

Given a partition in K groups, the Mallows distance fulfils the Huygens theorem decomposition in Between and Within dispersion:

$$\sum_{i=1}^{n} D_M^2(\Psi_{s_i}(t), \overline{\Psi_S}(t)) =$$
$$\sum_{h=1}^{K} n_h D_M^2(\overline{\Psi_S}(t), \overline{\Psi_{C_h}}(t)) +$$
$$+ \sum_{h=1}^{K} \sum_{i \in C_h} D_M^2(\Psi_{s_i}(t), \overline{\Psi_{C_h}}(t))$$

where n_h is the number of units in group C_h

Irpino A., Verde R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. *Data Science and Classification, Proc. IFCS 2006.* Springer, 185-192

Descriptive Measures - based on the Mallows distance

Empirical Covariance (Verde, Irpino (2014))

Consider two histogram-valued variables X and Y with n units and the quantile functions $\Psi_{X(i)}(t)$ and $\Psi_{Y(i)}(t)$, for $i \in \{1, ..., n\}$:

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \left(\Psi_{X(i)}(t) - \overline{\Psi_X}(t)\right) \left(\Psi_{Y(i)}(t) - \overline{\Psi_Y}(t)\right) dt$$

Assuming the Uniform distribution across the subintervals:

$$cov(X, Y) = \sum_{i=1}^{n} \sum_{\ell=1}^{m} p_\ell \left[(c_{X(i)_\ell} - \overline{c}_{X_\ell})(c_{Y(i)_\ell} - \overline{c}_{Y_\ell}) + \frac{1}{3}(r_{X(i)_\ell} - \overline{r}_{X_\ell})(r_{Y(i)_\ell} - \overline{r}_{Y_\ell}) \right]$$

where
$c_{X(i)_\ell}, c_{Y(i)_\ell}, r_{X(i)_\ell}, r_{Y(i)_\ell}$: centers/half ranges of subinterval ℓ and unit i of X and Y
$\overline{c}_{X_\ell}, \overline{c}_{Y_\ell}, \overline{r}_{X_\ell}, \overline{r}_{Y_\ell}$: mean of the centers/half ranges of the subinterval ℓ of X and Y

Clustering

Methods based directly on dissimilarity measures: straightforward adaptation

- Hierarchical clustering : Complete linkage, Single linkage, Average linkage,...
- Non-hierarchical clustering : k-Medoïds

The Huyghens decomposition of the Mallows distance allows for the extension of other clustering methods :

- Ward hierarchical clustering (Irpino and Verde, 2006)
- Non-hierarchical dynamical clustering (k-Means like) (Irpino and Verde, 2006)
- Divisive clustering (Brito and Chavent, 2011, 2022)

First linear regression models

- First linear regression method for histogram-valued data due to Billard and Diday (2006)
  - Model based on the - real-valued - first and second-order moments for histogram-valued variables obtained previously
  - From these, the regression coefficients are derived
- Irpino and Verde (2008) developed a linear regression model
  - Minimizing the Mallows's distance between the observed and the derived quantile functions of the dependent variable
  - The method lies on the exploitation of the properties of a decomposition of the Mallows's distance
  - Used to measure the sum of squared errors and rewrite the model
  - Splitting the contribution of the predictors in a part depending on the averages of the distributions and another depending on the centered quantile distributions

Distribution and Symmetric Distribution Linear Regression model

- Dias and Brito (2015) propose a new Linear Regression model for histogram-valued variables
- Distributions are represented by their quantile functions
- The model includes both the quantile functions that represent the distributions that the independent histogram-valued variables take, and the quantile functions that represent the distributions that the respective symmetric histogram-valued variables take -
two terms per independent variable

Linear combination of quantile functions

The linear combination of quantile functions **is not defined** as:

$$\Psi_{Y(i)}(t) = a_1\Psi_{X_1(i)}(t) + a_2\Psi_{X_2(i)}(t) + \ldots + a_p\Psi_{X_p(i)}(t)$$

- Because when we multiply a quantile function by a negative number we do not obtain a non-decreasing function
- If non-negativity constraints are imposed on the parameters a_j, $j \in \{1, 2, \ldots, p\}$ a quantile function is always obtained.
  However, this solution forces a direct linear relation between $\Psi_{Y(i)}(t)$ and $\Psi_{X_j(i)}(t)$
- Dias and Brito (2015) proposed a definition for linear combination of quantile functions that solves the problem of the semi-linearity of the space of the quantile functions

Definition of linear combination

To allow for a direct and an inverse linear relation between the quantile functions, the linear combination includes:

- $\Psi_{X_j}(t)$ that represents the distributions of the histogram-valued variables X_j
- $-\Psi_{X_j}(1-t)$ the quantile function that represents the respective symmetric histograms.

Linear combination between quantile functions

The quantile function Ψ_Y may be expressed as a linear combination of $\Psi_{X_j}(t)$ and $-\Psi_{X_j}(1-t)$ as follows:

$$\Psi_Y(t) = \sum_{j=1}^{p} a_j \Psi_{X_j}(t) - \sum_{j=1}^{p} b_j \Psi_{X_j}(1-t) + \gamma$$

with $t \in [0,1]$; $a_j, b_j \geq 0$, $j \in \{1, 2, \ldots, p\}$, $\gamma \in \mathbb{R}$

Distribution and Symmetric Distribution Linear Regression model

- Non-negativity restrictions on the parameters do not imply a direct linear relationship
- Uses the Mallows distance to quantify the error
- Determination of the model requires solving a quadratic optimization problem, subject to non-negativity constraints on the unknowns

- Dias, S. and Brito, P. (2015), Linear Regression Model with Histogram-Valued Variables. *Statistical Analysis and Data Mining*, 8(2),75-113
- Dias, S. and Brito, P. (2017). Off the Beaten Track: A New Linear Model for Interval Data. *European Journal of Operational Research*, 258(3), 1118–1130.

Distribution and Symmetric Distribution Linear Regression model

The parameters of the model are an optimal solution of the minimization problem:

$$\text{Minimize} \qquad SE = \sum_{i=1}^{n} D_M^2(\Psi_{Y(i)}, \Psi_{\widehat{Y}(i)})$$

with $a_j, b_j \geq 0$, $j = \{1, 2, \ldots, p\}$ and $\gamma \in \mathbb{R}$

$\longrightarrow$ Kuhn Tucker optimality conditions allow defining a measure to evaluate the quality of fit of the model (determination coefficient):

$$\Omega = \frac{\sum_{i=1}^{n} D_M^2\left(\widehat{Y}_i, \overline{Y}\right)}{\sum_{i=1}^{n} D_M^2\left(Y_i, \overline{Y}\right)} \quad , \quad 0 \leq \Omega \leq 1$$

Linear relation between temperature and grass height

Histogram-valued variables in the model:

H: distribution of the min and max values of all grass species in each country or country like-region;

Temp: distribution of the atmospheric temperatures in each country or country like-region;

Northern Hemisphere:

$$\Psi_{H(i)}(t) = 2.32 + 1.35\Psi_{Temp(i)}(t) - 1.21\Psi_{Temp(i)}(1-t)$$

$$\Omega = 0.40$$

Southern Hemisphere:

$$\Psi_{H(i)}(t) = -2.78 + 1.54\Psi_{Temp(i)}(t) - 1.34\Psi_{Temp(i)}(1-t)$$

$$\Omega = 0.42$$

Conclusions: Relation between temperature and grass height

- The proportion of variation of the height distribution, which can be explained solely by the temperature distribution, is similar in both hemispheres;
- The simple linear relation between the height distribution and the temperature distribution is direct;
- Main conclusions of the study (in biological terms):
  The general and global pattern, as observed for other groups of organisms, can also be observed in grasses:
  Plants become smaller towards the Poles.

Linear discriminant function

The score for each unit i, $\Psi_{S(i)}(t)$, is defined as a linear combination of $\Psi_{X_j}(t)$ and $-\Psi_{X_j}(1-t)$:

$$S(i) = \Psi_{S(i)}(t) = \sum_{j=1}^{p} a_j \Psi_{X_j(i)}(t) - \sum_{j=1}^{p} b_j \Psi_{X_j(i)}(1-t)$$

For each subinterval ℓ this score is defined by

$$\sum_{j=1}^{p} \left(a_j c_{X_j(i)_\ell} - b_j c_{X_j(i)_{(m-\ell+1)}} \right) + \left(\frac{2(t - w_{\ell-1})}{w_\ell - w_{\ell-1}} - 1 \right) \sum_{j=1}^{p} \left(a_j r_{X_j(i)_\ell} + b_j r_{X_j(i)_{(m-\ell+1)}} \right).$$

with $t \in [0,1]$; $a_j, b_j \geq 0$, $j \in \{1, 2, \ldots, p\}$

Dias, S.; Brito, P. and Amaral, P. (2021). Discriminant analysis of distributional data via fractional programming. *European Journal of Operational Research*, 294(1), 206-218.

Symbolic Discriminant Function

Huygens theorem: Sum of Squares and Cross-Products =
= Sum between-groups + Sum within-groups

In matricial notation: $\gamma' T \gamma = \gamma' B \gamma + \gamma' W \gamma$,
with weight vector $\gamma = [a_1, b_1, ..., a_p, b_p] \geq 0$.

Optimization problem:

The optimal parameter vector is estimated as follows:

$$\gamma* = \arg\max_{\gamma} \frac{\gamma' \mathbf{B} \gamma}{\gamma' \mathbf{W} \gamma} \qquad \text{subject to: } \gamma \geq 0$$

Optimization of constrained rational quadratic functions

- Non-convex, hard optimization problem
- Easy to find a good solution
- Difficult to prove optimality
- Branch and Bound technique and Conic Optimization

Classification in two groups

Classification in two groups using the Mallows Distance

Considering two groups: C_1, C_2, a unit i and the respective quantile functions: $\overline{\Psi_{D_{C1}}}(t)$, $\overline{\Psi_{D_{C2}}}(t)$ and $\Psi_{D(i)}(t)$

- The unit i is assigned to Group $C1$ if

$$D_M^2\left(\Psi_{D(i)}(t), \overline{\Psi_{D_{G1}}}(t)\right) < D_M^2\left(\Psi_{D(i)}(t), \overline{\Psi_{D_{G2}}}(t)\right)$$

- The unit i is assigned to Group $C2$ if

$$D_M^2\left(\Psi_{D(i)}(t), \overline{\Psi_{D_{G2}}}(t)\right) < D_M^2\left(\Psi_{D(i)}(t), \overline{\Psi_{D_{G1}}}(t)\right)$$

A unit i is assigned to the group for which the Mallows distance between its score and the score of the corresponding barycentric histogram is minimum.

Gram positive/negative bacteria

Case I: classify the bacteria in Gram positive *vs* Gram negative from the distributions of the four nucleotide frequencies in the genes
Case II: classify the bacteria in Gram positive *vs* Gram negative from the distributions of the four nucleotide frequencies at the first/second/third codon position

Group 1: Gram positive - 11 bacterias
Group 2: Gram negative - 12 bacterias

For all observations the subintervals of each histogram have the same weight with frequency 0.25

Discriminant function

- The 23 species are considered to obtain the parameters of the discriminant function;
- Leave-One-Out (LOO) Cross-Validation is applied

Parameters: BARON optimization method

Gram Positive/Negative: classification

		without LOO	**with LOO**
Study I	% Well classified	83%	78%
	Obs. wrongly classified in G1	12; 13; 22	12; 13; 22
	Obs. wrongly classified in G2	6	6; 10
Study II - 1^{st} position	% Well classified	83%	74%
	Obs. wrongly classified in G1	12; 13; 22	12; 13; 22
	Obs. wrongly classified in G2	10	3; 10; 11
Study II - 2^{nd} position	% Well classified	74%	74%
	Obs. wrongly classified in G1	12; 13; 14	12; 13
	Obs. wrongly classified in G2	1; 3; 10	1; 3; 4; 10
Study II - 3^{rd} position	% Well classified	70%	65%
	Obs. wrongly classified in G1	12; 13; 14; 22	12; 13; 14; 22
	Obs. wrongly classified in G2	1;3;10	1; 3; 6; 10

Outline

Concluding remarks

- From micro-data to macro-data:
- Interval, Distribution-valued data
  - Non-parametric models
  - Parametric models
- Several methodologies already developed
  that do take into account the variability of the data
- New problems / challenges for the 21^{st} century:
  intervals and distributions are not real numbers !

Applications of SDA

Classically ...

- Official statistics
- Botany and Zoology: data have intrinsic variability

Emerging fields of application:

- Large surveys, e.g., at European level - analysed by region
- Text analysis - distributions over topics
- Internet traffic
- Demography
- Meta analysis
- Econometric studies
- Finance modelling

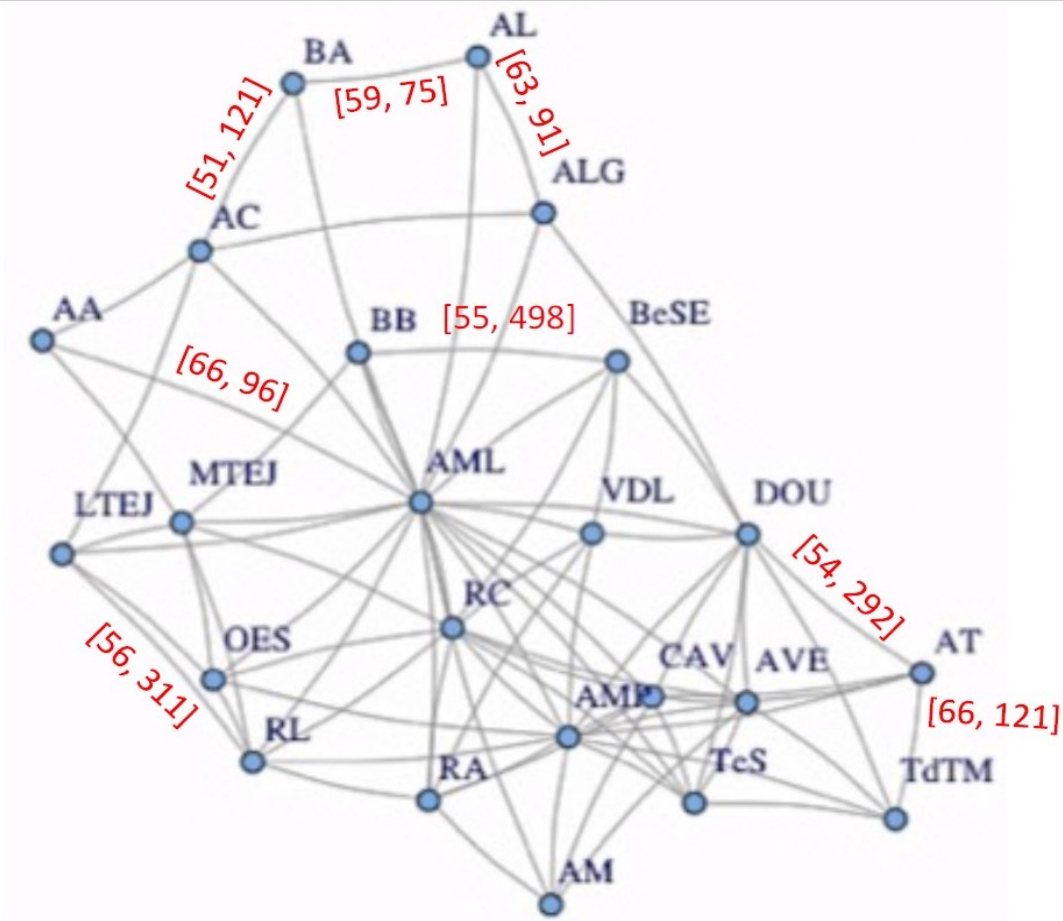
Applications of SDA

Complex data structures

- Social network analysis
  - Variability on the links' weights
  - Super nodes
  - Attributed networks
  - Multilayer networks
- Data streams
- Sensor data $\rightarrow$ "Smart statistics"

Applications of SDA

Social network analysis: Variability on the links' weights (Alves *et al*, 2022, 2023)



Applications of SDA

Recent approaches:

- Development of statistical methodologies
- to infer properties of underlying big datasets
- from higher-level symbolic summaries

(Sisson, Beranger, UNSW)

Future Perspectives

- Analysis based on marginal empirical distributions - the empirical distribution for each variable is considered separately

  $\longrightarrow$ need to go one step further, and consider the joint observed distributions in the data representation and analysis

- Data aggregation by histograms :
  Results depend heavily on the chosen partition or weights decomposition

  $\longrightarrow$ From histograms to densities - density-valued variables
  See: Functional Data Analysis

Future Perspectives

- Joint distributions
- Density-valued variables
- Spatial-temporal modelling and analysis
- Explore links with Multilevel statistical analysis
- Big Data analytics, e.g. deep learning approaches are still to be developed for symbolic data
- Latent Dirichlet allocation, then analysis of generated distributions

Concluding remarks

"Distributions are the numbers of the future"

(Schweizer, 1984)

Outline

Available Software

- R packages available in CRAN (increasing...):
  RSDA ; MAINT.Data ; HistDAWass ; iRegression ; symbolicDA ; ISDA.R

- SODAS: free, registration required - http://www.info.fundp.ac.be/asso/sodaslink.htm
- SYR: commercial software - https://www.symbad.co/le-logiciel-syr/

Symbolic Data Analysis

Symbolic Data Analysis was introduced by Prof. Edwin Diday in the late eighties of the 20^{th} century, in a seminal paper :

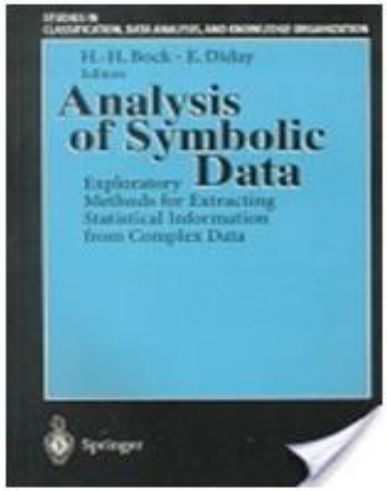
"The symbolic approach in clustering and relating methods of data analysis: The basic choices."
In : Proc. 1*st* IFCS Conference (Aachen, Germany, 1987), pp. 673-684, 1988.

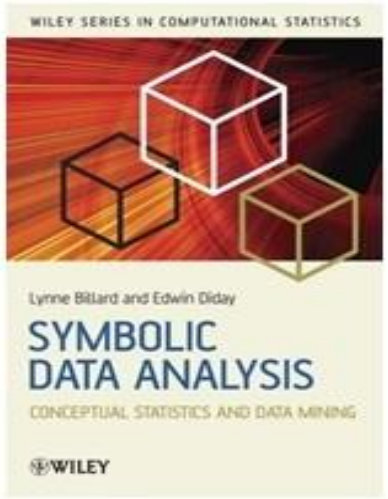
Symbolic Data Analysis

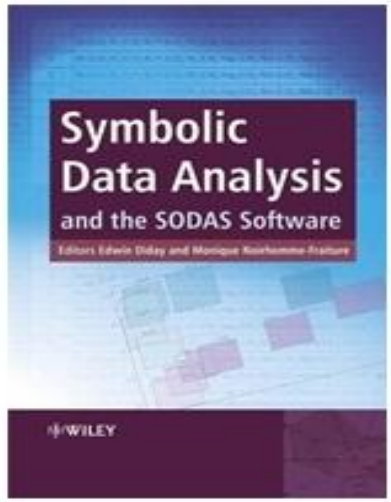
Nowadays :

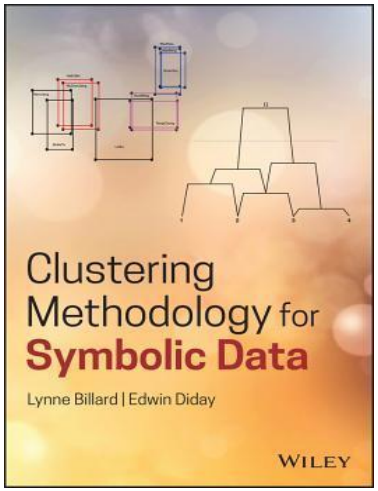
- Established Workshop series
- Increasing number of publications in a wide variety of journals
- Journals' special issues (SAD, ADAC)
- Website (wiki): http://vladowiki.fmf.uni-lj.si/doku.php?id=sda
- Linkedin SDA Group: for discussions, announcements...

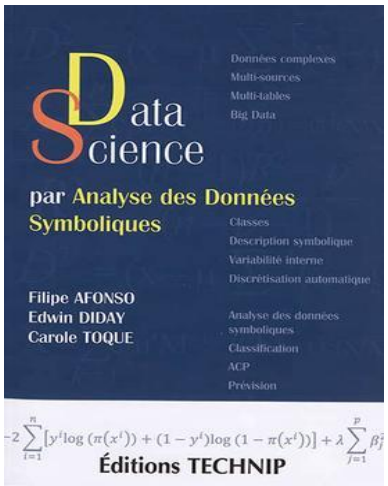
Books

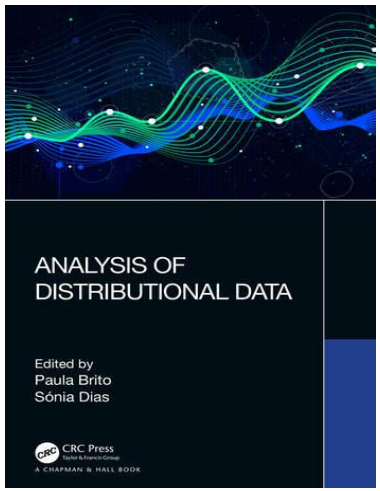

H.-H. Bock · E. Diday
Analysis of Symbolic Data
Exploratory Methods for Extracting Statistical Information from Complex Data
Springer


WILEY SERIES IN COMPUTATIONAL STATISTICS
Lynne Billard and Edwin Diday
SYMBOLIC DATA ANALYSIS
CONCEPTUAL STATISTICS AND DATA MINING
WILEY


Symbolic Data Analysis
and the SODAS Software
WILEY


Clustering Methodology for Symbolic Data
Lynne Billard | Edwin Diday
WILEY


Data Science
par Analyse des Données Symboliques
Données complexes
Multi-sources
Multi-tables
Big Data
Classes
Description symbolique
Variabilité interne
Discrétisation automatique
Filipe AFONSO
Edwin DIDAY
Carole TOQUE
Analyse des données symboliques
Classification
ACP
Prévision
Éditions TECHNIP


ANALYSIS OF DISTRIBUTIONAL DATA
Edited by
Paula Brito
Sónia Dias
CRC Press
A CHAPMAN & HALL BOOK

Books and Survey Papers


Bock, H.-H.; Diday, E. (2000): *Analysis of Symbolic Data: Exploratory methods for extracting statistical information from complex data.* Berlin-Heidelberg: Springer-Verlag.


Billard, L., Diday, E. (2007): *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* Wiley.


Diday, E., Noirhomme-Fraiture, M. (2008): *Symbolic Data Analysis and the SODAS Software.* Wiley.


Afonso, F., Diday, E., Toque, C. (2018). *Data Science par Analyse des Données Symboliques: une Nouvelle Façon d'Analyser les Données Classiques, Complexes et Massives à Partir des Classes: Applications avec Syr et R.* Editions Technip.


Billard, L., Diday, E. (2019). *Clustering Methodology for Symbolic Data.* John Wiley & Sons.


Brito, P., Dias, S. (2022). *Analysis of Distributional Data*. Chapman and Hall/CRC.

****


Billard, L., Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. *JASA*, 98 (462), 470–487.


Noirhomme-Fraiture, M., Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, 4(2), 157–170.


Brito, P. (2014). Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, 4 (4), 281–295.

Thank you very much for your attention !