Modelling and Analysing Interval Data in R

Paula Brito(1), A. Pedro Duarte Silva(2)

(1)Fac. Economia,Universidade do Porto & LIAAD-INESC TEC
(2)Católica Porto Business School & CEGE, Universidade Católica Portuguesa

WhyR 2019
Warsaw, Poland, 29th September 2019

Outline

Outline

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)

	Nb. passengers	Delay (min)	Airline	Distance
Flight 1	200	20	Air France	Long
Flight 2	120	0	Ryanair	Short
Flight 3	100	10	Lufthansa	Medium

The data

Symbolic Data Analysis :
to take into account **variability** inherent to the data
Variability occurs when we have

- Descriptors on flights, but: analyse the airline companies - not each individual flight
- Data on individual purchases, but: analyse the clients
- Official statistics - Descriptors on citizens, but: analyse cities, the regions, sociographic groups - not the individual citizens

$\Longrightarrow$(symbolic) variable values are

sets, intervals

distributions on an underlying set of sub-intervals or categories

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

The data

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

Flight	Airline	Nb. Passengers	Delay (min)	Aircraft
1	A	180	10	Boeing
2	B	120	0	Boeing
3	A	200	20	Airbus
4	C	80	15	Embraer
5	B	100	5	Embraer
6	A	300	35	Airbus
7	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]

Sources of symbolic data: Concept description

Description of the species "Dog" - not "my dog" !

Species	coat	vision range (m)	hearing frequency (Hz)	smell receptors (millions)
Dog	$\{single, double\}$	$[500, 900]$	$[40, 60000]$	$[125, 220]$

Outline

1. Variability in Data
2. Symbolic Variables
   - Interval-Valued Variables
3. Parametric models for Interval Data
   - Robust estimation and Outlier detection
4. Methods for Multivariate Analysis of Interval Data
   - Analysis of Variance
   - Discriminant Analysis
   - Model-Based Clustering
5. R Package
6. Concluding Remarks

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 entities 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 on O_j

Interval-Valued Variables

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

Interval-valued variable :

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

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	$[10, 20]$

Number of passengers in flights :

	Nb. Passengers
Airline A	$[150, 200]$
Airline B	$[180, 300]$
Airline C	$[200, 400]$

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,...

Outline

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	$[l_{n1}, u_{n1}]$	$\ldots$	$[l_{nj}, u_{nj}]$	$\ldots$	$[l_{np}, u_{np}]$

Parametric models for interval data

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

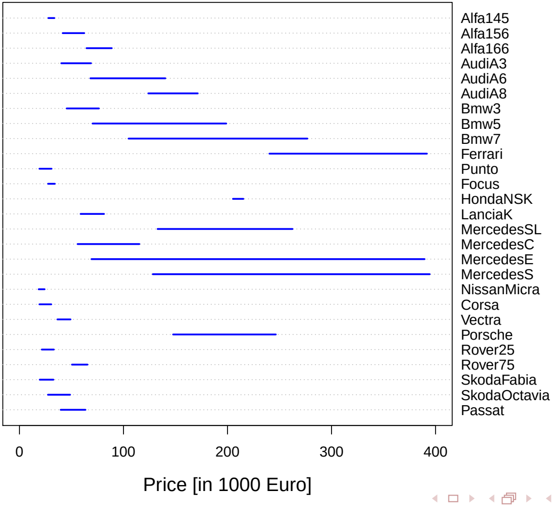
For each s_i, $Y_j(s_i) = I_{ij} = [l_{ij}, u_{ij}]$ is naturaly defined by the lower and upper bounds l_{ij} and u_{ij}

For modelling purposes $\rightarrow$ preferable equivalent parametrization:

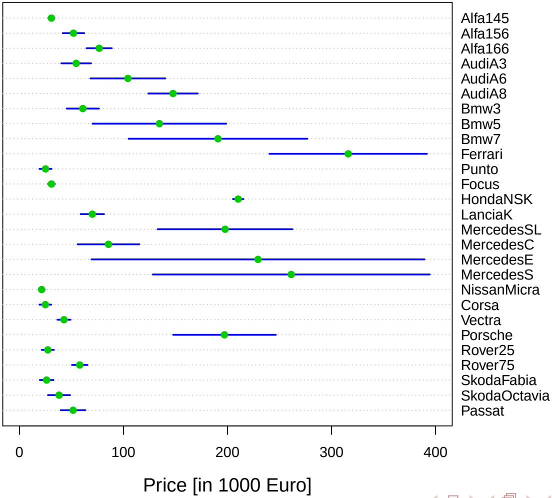
Represent $Y_j(s_i)$ by

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

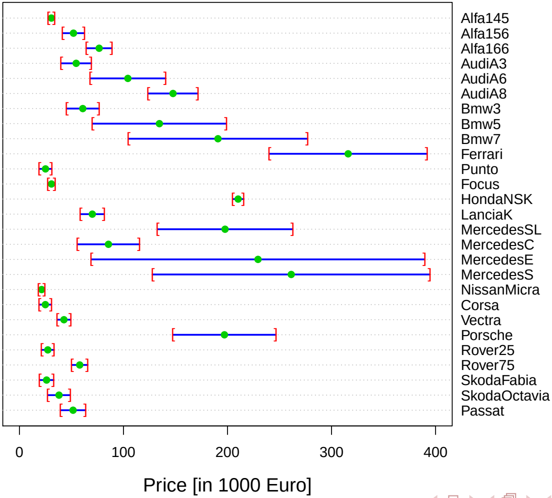
Example: Price of different car models



Example: Price of different car models



Example: Price of different car models



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

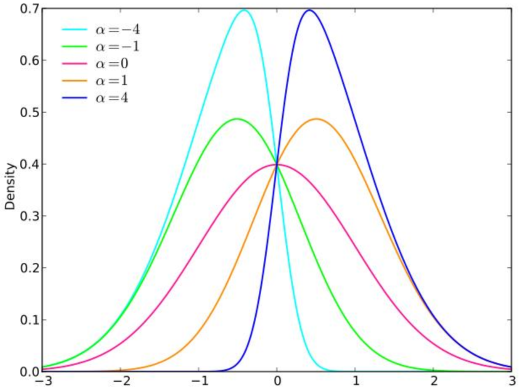
Skew-Normal model:

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


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
−3
−2
−1
0
1
2
3
$\alpha = -4$
$\alpha = -1$
$\alpha = 0$
$\alpha = 1$
$\alpha = 4$

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

Parametric 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;

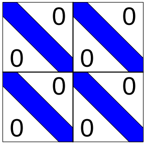
Parametric Models for interval data

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

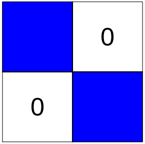
Parametric Models for interval data

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

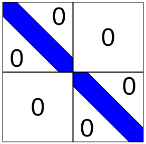
Configuration 1



Configuration 2



Configuration 3



Configuration 4

Parametric Models for interval data

- Configurations 2 and 3 are a particular case of 1
- Configuration 4 is a particular case of all the others

In cases 2, 3 and 4, 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,

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

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

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

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:

$l = \text{ constant } -\frac{1}{2} nln|\Omega| - \frac{n}{2} tr(\Omega^{-1}V) + \sum_i \zeta_0(\alpha^t \omega^{-1}(x_i - \xi))$ 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
$l = \text{ constant } -\frac{1}{2} nln|\Omega| - \frac{n}{2} tr(\Omega^{-1}V) + \sum_i \zeta_0(\eta(x_i - \xi))$

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:

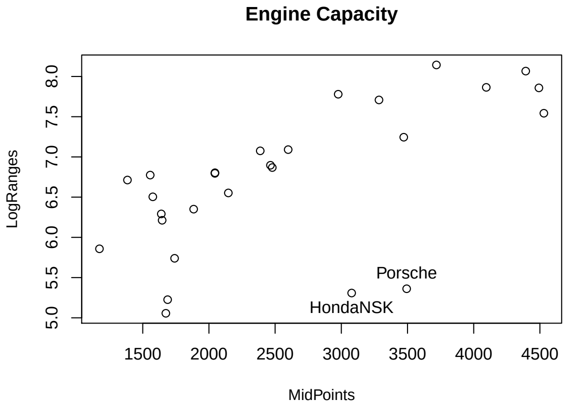
Let $\theta = (\xi, \Omega, \eta) = \theta(\psi)$ with $\psi = (\mu, \Sigma, \gamma_1)$

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

Brito, P., Duarte Silva, A. P. (2012).
Modelling Interval Data with Normal and Skew-Normal Distributions.
Journal of Applied Statistics, Volume 39, Issue 1, 3-20.

Robust estimation and Outlier detection

Example: Different car models

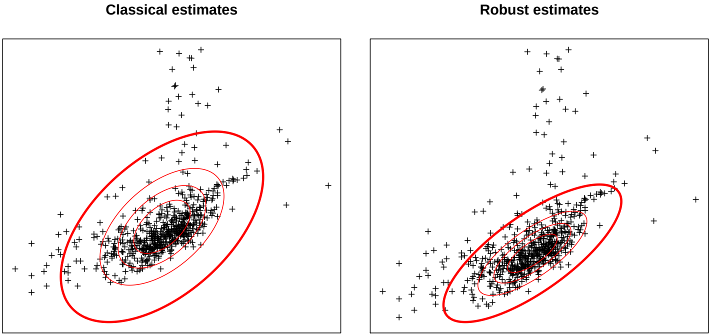


Multivariate outlier detection

Outlier detection:

Outliers will typically have large distance. If multivariate normal distribution is assumed, MD_i^2 is approx. χ_d^2 distributed.

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



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.

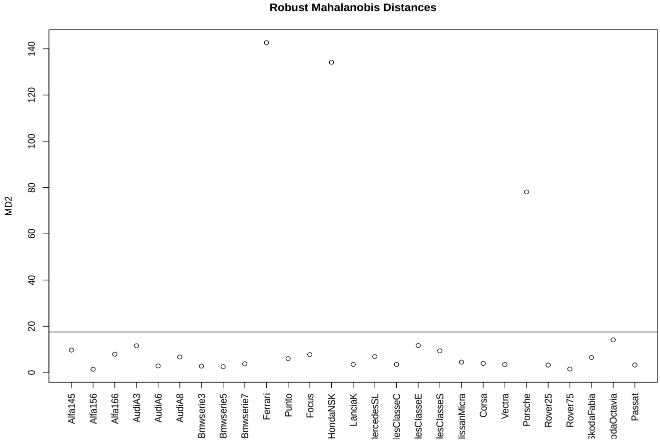
Afterwards: compute **robust** Mahalanobis distances based on these estimates.

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 assymptotic approximations OR
  - F and Beta finite sample approximations (Hardin & Rocke (2005); Cerioli (2010))

Multivariate outlier detection



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

Outline

(M)ANOVA : Objectives

Comparison of means of one or more numerical variables
in two or more populations,
from which random samples were drawn

Example : compare the mean value of the sales of a given product in different shops.

Factor levels			
1	2	...	k
x_{11}	x_{12}	...	x_{1k}
x_{21}	x_{22}	...	x_{2k}
...	...	...	...
$x_{n_1 1}$	$x_{n_2 2}$	...	$x_{n_k k}$

$H0 : \mu_1 = \mu_2 = \ldots = \mu_k$ - population mean values are all equal
$H1 : \exists j, \ell : \mu_j \neq \mu_\ell$ - there are at least two populations with different mean values

(M)ANOVA for interval-valued variables

$\rightarrow$ **Likelihood ratio approach**

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^*)

Gaussian and Skew-Normal model:

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

In all cases, under the null hypothesis, the likelihood ratio statistics follows asymptotically a chi-square distribution

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

(M)ANOVA for interval-valued variables

Brito, P., Duarte Silva, A. P. (2012).
Modelling Interval Data with Normal and Skew-Normal Distributions.
Journal of Applied Statistics, Volume 39, Issue 1, 3-20.

(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

Temperatures measured in meteorological stations in northern China.
Data : intervals of observed temperatures (Celsius scale) in each of the four quarters, Q_1 to Q_4, of the years 1974 to 1988 in 22 stations.

Station	Region	$Q1$	$Q2$	$Q3$	$Q4$
Beijing-1974	North	$[-9.5, 10.6]$	$[6.5, 29.8]$	$[12.6, 29.6]$	$[-10.44, 9.06]$
Beijing-1975	North	$[-8.6, 12.9]$	$[7.9, 30.2]$	$[15.0, 31.6]$	$[-7.0, 19.2]$
...	...	...	...	...	...
ZhangYe-1988	Northwest	$[-15.4, 7.2]$	$[2.3, 26.4]$	$[8.6, 30.2]$	$[-12.0, 15.1]$

The full table comprises $n = 22 \times 15 = 330$ rows and 4 columns.

(M)ANOVA: Example

The 22 meteorological stations belong to 3 different regions in China (North, Northwest, Northeast):
MANOVA performed to assess whether the regions are different

MODEL	-2 lnλ	DF	P-VALUE
NORM 1	480.2475	16	$<$ 1E-10
NORM 2	989.9340	16	$<$ 1E-10
NORM 3	529.2541	16	$<$ 1E-10
NORM 4	1057.9210	16	$<$ 1E-10
SkN 1	**447.4244**	**16**	$<$ **1E-10**
SkN 2	974.0240	16	$<$ 1E-10
SkN 3	530.3980	16	$<$ 1E-10
SkN 4	1110.3840	16	$<$ 1E-10

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 :

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

Quadratic :

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

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 :

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

General :

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

Discriminant Analysis

Experimental results

- Parametric rules generally outperform distance-based ones
- Homocedastic problems: linear discriminant rules perform best
- Large training samples and heterocedastic conditions quadratic methods are usually superior
- Small training samples in heterocedastic problems: restricted quadratic rules are preferable
  - even in some cases where the model assumed is not true

Restricted configurations 2 - 4:

- provide a natural way of imposing constraints
- are effective in reducing expected error rates
- for heterocedastic problems with small or moderate training samples

Discriminant Analysis

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$= -2\ell(\hat{\varphi}; \mathbf{x}) + d_\varphi ln(n)$

Labour force survey

Data from Portuguese Labour Force Survey, 1^{st} semester of 2008

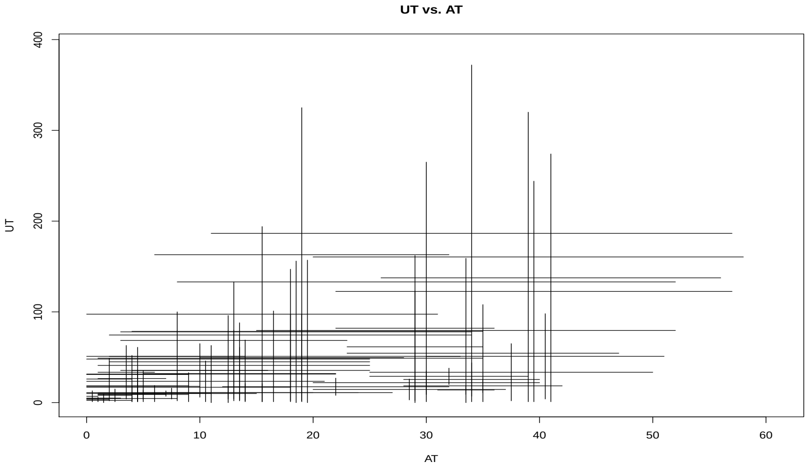
1540 cases : people who were unemployed at the time of the survey

Two variables:
Activity Time, in years (AT)
Unemployment Time, in months (UT)

Micro data were gathered on the basis of
Gender, Region, Age-Group and Education $\rightarrow$
58 sociological groups

Lowest BIC value:
solution in 5 components, heterocedastic setup,
Case 2 - independent interval-valued variables

MAINT.Data: plot

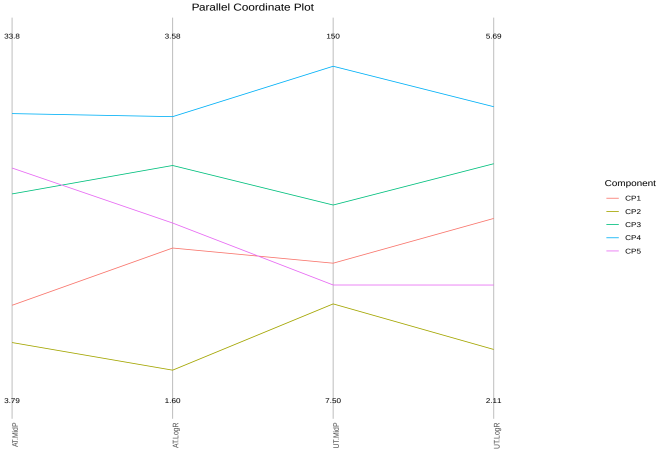


Labour force survey

Unemployement data
Component Proportions, Mean-Vectors and Variances

		C1	C2	C3	C4	C5
	Proportions	0.271	0.206	0.227	0.103	0.191
Mean Values	AT MidP	8.662	3.785	23.237	33.750	26.627
	AT LogR	2.553	1.600	3.197	3.578	2.748
	UT MidP	31.985	7.495	66.990	150.500	18.869
	UT LogR	4.042	2.110	4.849	5.690	3.060
Variances	AT MidP	17.065	4.963	73.153	57.479	78.060
	AT LogR	0.340	0.544	0.119	0.040	0.182
	UT MidP	101.545	7.841	230.649	468.583	54.774
	UT LogR	0.113	0.685	0.057	0.021	0.225

Labour force survey



Labour force survey

- Although the number of observations is relatively low
  $\rightarrow$ a restricted though heterocedastic model has been identified as best fit
- The method chose the best parameters for clustering, preferring a heterocedastic model to a "lighter" homocedastic one $\rightarrow$ picking up a restricted configuration for the variance-covariance matrix
- Choosing Case 2 as opposed to Case 3 $\rightarrow$ correlation between the two parts of the interval-variables is considered more important than correlation between different variables
- Components can only be separated by considering simultaneously Midpoints and Log-Ranges

USA meteorological data application

This dataset records temperatures and pluviosity measured in 282 meteorological stations in the USA.

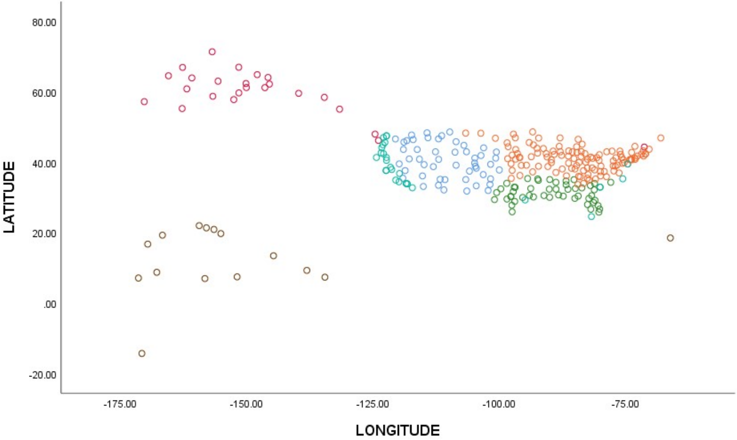
Three interval-valued variables, measured in each station:

- Temperature ranges in January
- Temperature ranges in July
- Annual pluviosity ranges

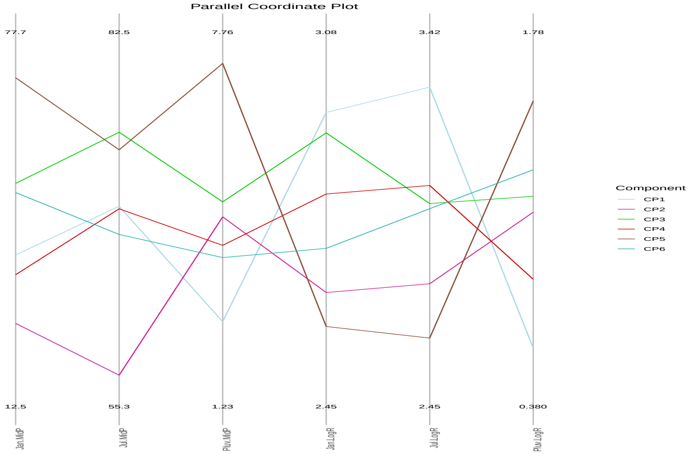
The lowest BIC value is observed for the unrestricted (Case 1) heterocedastic solution with 6 natural clusters:

1: Arid Inland West, 2: Alaska,
3: Southeast, 4: Northeast and Midwest,
5: Pacific Islands and Puerto Rico, 6: Pacific Coast

USA meteorological data application



USA meteorological data



USA meteorological data application

- Clusters are differentiated not only by the MidPoint but also by the Log-Range variables
- Moreover, clusters display highly different variances
  - Alaska cluster presents very high variance for the January MidPoint variable, while Arid Inland West and Pacific Coast clusters present high variances for the July MidPoint variable
  - the Alaska cluster has a high variance for the Log-Range of the pluviosity and the Pacific Coast cluster for the Log-Range of the temperature in July
- This stark difference illustrates well the need of a heterocedastic setup for these data

Model-Based Clustering : Conclusions

- Proposed modelling successfully applied to real data sets of different nature and size
- Adopting configurations adapted to interval data proved to be the adequate approach
- Important to consider both the information about
  - position - conveyed by the MidPoints
  - intrinsic variability - conveyed by the LogRanges

  when analysing interval data
- Flexibility of the model in identifying heterocedastic models

Brito, P., Duarte Silva A.P., Dias, J.G. (2015).
Probabilistic Clustering of Interval Data.
Intelligent Data Analysis, vol. 19, no. 2.

Outline

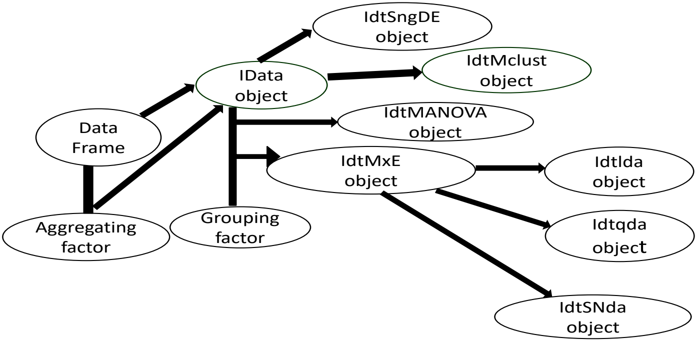
R Package: MAINT.Data

MAINT-Data : Modelling and Analysing Interval Data

$\rightarrow$ Available at CRAN

- Specialized data classes for interval-data
- Microdata aggregation
  - Min-Max
  - User defined quantiles
- Methods for Maximum Likelihood Estimation
- Robust estimation and Outlier detection (Gaussian model)
- ANOVA and MANOVA
- Discriminant Analysis
- Model-based Clustering (Gaussian model)

MAINT.Data overview



MAINT.Data overview

Idata

MidP: data.frame
LogR: data.frame
Obsnames: character
Varnames: character
Nobs: numeric
NIVar: numeric

summary(): summaryIData
print(): Idata
ncol(): numeric
colnames(): character
cbind(): Idata
plot(): void
...
mle(): IDataE
fasttle(): IDtSngNDRE
fulltle(): IDtSngNDRE
MANOVA(): IdtMANOVA
RobMxtDEst(): IdtMxNDRE
lda(): Idtlda
qda(): Idtqda
snda(): IdtSNda
Idtmclust(): IdtMclust

Outline

Summary

- From Micro-data to Macro-data: Interval-valued variables
- Take variability into account
- Several methodologies already developed
  that do take into account the variability of the data
- Parametric models specific for interval-valued variables $\longrightarrow$
  model-based multivariate analysis of interval data
- R implementation: Package MAINT.Data, available at CRAN
- New problems / challenges for the 21^{st} century:
  intervals are not real numbers !

SDA: Books and Main Papers

Books:

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

Survey Papers:

- 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.