An Innovative Framework for Analysing Official Statistics: Symbolic Data Analysis

Paula Brito[(1)] & A. Pedro Duarte Silva[(2)]

[(1)]Fac. Economia, Univ. Porto & LIAAD-INESC TEC, Portugal
[(2)]Católica Porto Business School & CEGE, Univ. Católica Portuguesa, Portugal

11^{th} European Conference on Quality in Official Statistics
Estoril, 4-7 June 2024

Outline

Outline

Symbolic Data Analysis

Symbolic Data Analysis:

- Represent and analyse data with intrinsic variability
- In the form of sets, intervals, distributions
- Groups/concepts VS individuals
- Tackling data size

Relevant in Official Statistics:

- Aggregate data
- Confidentiality
- Combine surveys
- Cross-border comparison

Objective

- Analyse the **Portuguese Household Budget Survey**
- At aggregate level - based on location and income
- Check for structure among groups
- Typology based/connected to income level ?
- Typology connected to location and type (Rural/Urban) ?

Portuguese Household Budget Survey

- Data from 2015 (most recent)
  Proportion of total expenses

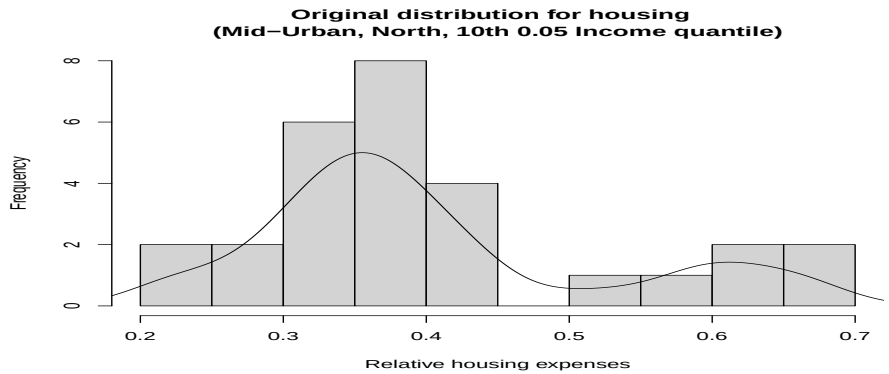
- Ten variables:
  - Food products and non-alcoholic beverages
  - Clothing and footwear
  - Housing, water, electricity, gas, and other fuels
  - Home accessories, household equipment, and routine household maintenance
  - Health
  - Transport
  - Communications
  - Leisure, recreation, and culture
  - Restaurants and hotels
  - Miscellaneous goods and services

Portuguese Household Budget Survey

Microdata were gathered on the basis of:

- Income class - 20 classes, based on equally-spaced quantiles
- Region - NUTS 2 (North, Centre, Lisbon Met Area, Alentejo, Algarve, Madeira, Azores)
- Type of area: Predominantly Rural (PRA), Medi-urban (MUA), Predominantly Urban (PUA)

- $20 \times 7 \times 3 = 420$ groups
- Each group described by the distribution of each of the ten variables



Outline

Parametric models for distributional data

Parametrization based on a central statistic and a given set of quantiles, $\psi_1, \ldots, \psi_k$

Represent each distribution $Y_j(s_i)$ by

- a central statistic C_{ij}, typically the Median Med_{ij} or the MidPoint $\frac{Max_{ij}+Min_{ij}}{2}$
- the $[Min, \psi_1[$ range: $R_{1ij} = \psi_{1ij} - Min_{ij}$
- the $[\psi_1, \psi_2[$ range: $R_{2ij} = \psi_{2ij} - \psi_{1ij}$
- ...
- the $[\psi_k, Max[$ range: $R_{mij} = Max_{ij} - \psi_{kij}$

Note: In the presence of strong outliers the Max (Min) may be replaced by high (low) quantiles

Parametric models for distributional data

Household Budget Survey data:

- Many zeros
- Upper outliers
- Therefore: Median & Min-Q40-Q60-Q80-Q99

Five (real-valued) indicators:

- Median
- R1 = Q40-Min
- R2 = Q60-Q40
- R3 = Q80-Q60
- R4 = Q99-Q80

Portuguese Household Budget Survey

	Food	...
MU-North-IncQnt3	$0.2369; \{[0.00, 0.22[, 0.4; [0.22, 0.24[, 0.2$ $[0.24, 0.28[, 0.2; [0.28, 0.42], 0.19\}$	...
MU-North-IncQnt4	$0.2379; \{[0.00, 0.17[, 0.4; [0.17, 0.24[, 0.2$ $[0.24, 0.30[, 0.2; [0.30, 0.62], 0.19\}$	...
...	...	...
PUA-Madeira-IncQnt20	$0.0980; \{[0.04, 0.09[, 0.4; [0.09, 0.10[, 0.2$ $[0.10, 0.13[, 0.2; [0.13, 0.25], 0.19\}$	...

Parametric Models for distributional data

Gaussian model:

Assume that the joint distribution of the central statistic C and the logs of the ranges $R_\ell^* = ln(R_\ell), \ell = 1, \ldots, m$, is multivariate Normal:

$$(C, R_1^*, \ldots, R_m^*) \sim N_{(m+1)p}(\mu, \Sigma)$$

$$\mu = \left[\mu_C^t, \mu_{R_1^*}^t, \ldots, \mu_{R_m^*}^t\right]^t ; \Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CR_1^*} & \ldots & \Sigma_{CR_m^*} \\ \Sigma_{R_1^*C} & \Sigma_{R_1^*R_1^*} & \ldots & \Sigma_{R_1^*R_m^*} \\ \ldots & \ldots & \ldots & \ldots \\ \Sigma_{R_m^*C} & \Sigma_{R_m^*R_1^*} & \ldots & \Sigma_{R_m^*R_m^*} \end{pmatrix}$$

μ_C and $\mu_{R_\ell^*}, \ell = 1, \ldots, m$ - p-dimensional column vectors of the mean values

$\Sigma_{CC}, \Sigma_{CR_\ell^*}, \Sigma_{R_\ell^*C}$ and $\Sigma_{R_{\ell 1}^* R_{\ell 2}^*}$ - $p \times p$ matrices

Parametric Models for distributional data

Model advantage:
Straightforward application of classical inference methods

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

Parametric Models for distributional data

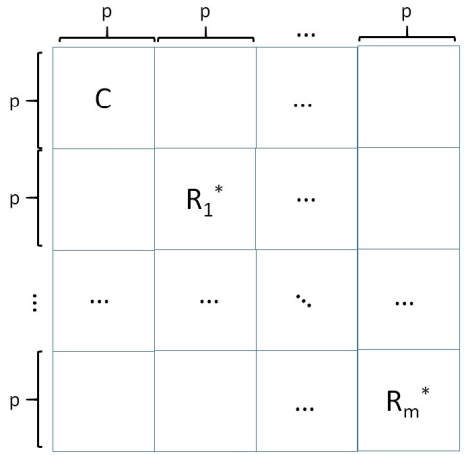
However, for distributional data:

Centre c_{ij} and Ranges $r_{\ell ij}$ of the value of an distributional-valued variable are quantities related to one only variable
$\rightarrow$ should not be considered separately

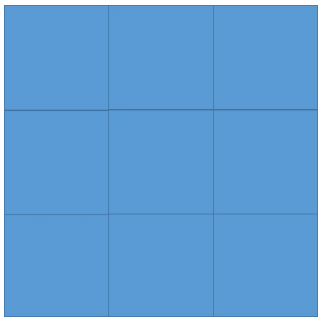
So: parametrizations of the global covariance matrix $\rightarrow$
take into account the link that may exist between centres and log-ranges of the same or different variables

Parametric Models for distributional data

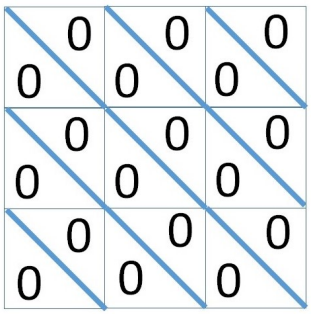
$\Sigma =$



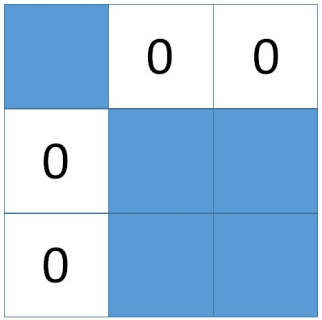
For $m = 2$



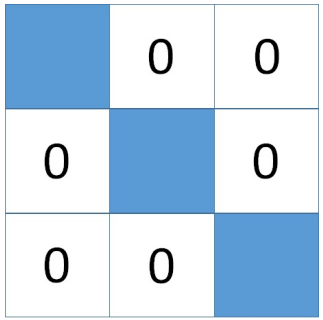
Configuration 1



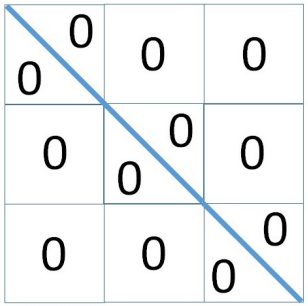
Configuration 2



Configuration 3



Configuration 4



Configuration 5

Models for distributional data

- Configurations 2 and 3 are particular cases of 1
- Configuration 4 is a particular case of of 3
- Configuration 5 is a particular case of all the others

In cases 2, 3, 4 and 5, Σ can be written as a block diagonal matrix

- Configuration 2 : there are p blocks, all $(m+1) \times (m+1)$
- Configuration 3 : there are 2 blocks, one is $p \times p$, and the other is $mp \times mp$
- Configuration 4 : there are $m+1$ blocks , all $p \times p$
- Configuration 5 : the $(m+1)p$ blocks are single real elements

Household Budget Survey Data

- Original microdata with 11398 observations
- $n0 = 420$ units $=$ 20 Income classes $\times$ 7 NUTS $\times$ 3 Area types
- But: 133 units with degenerate intervals discarded $\longrightarrow n = 287$
- Ten distributional-valued variables
- Analysed from Minimum to 0.99 quantile
- Location measure: Median
- Three intermediate quantiles: Q40, Q60, Q80
- Therefore: $p = 10$ variables, $m = 4$ intervals, 5 indicators, μ is a 50-dim vector, Σ is 50×50

Outline

Model-Based Clustering

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

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

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

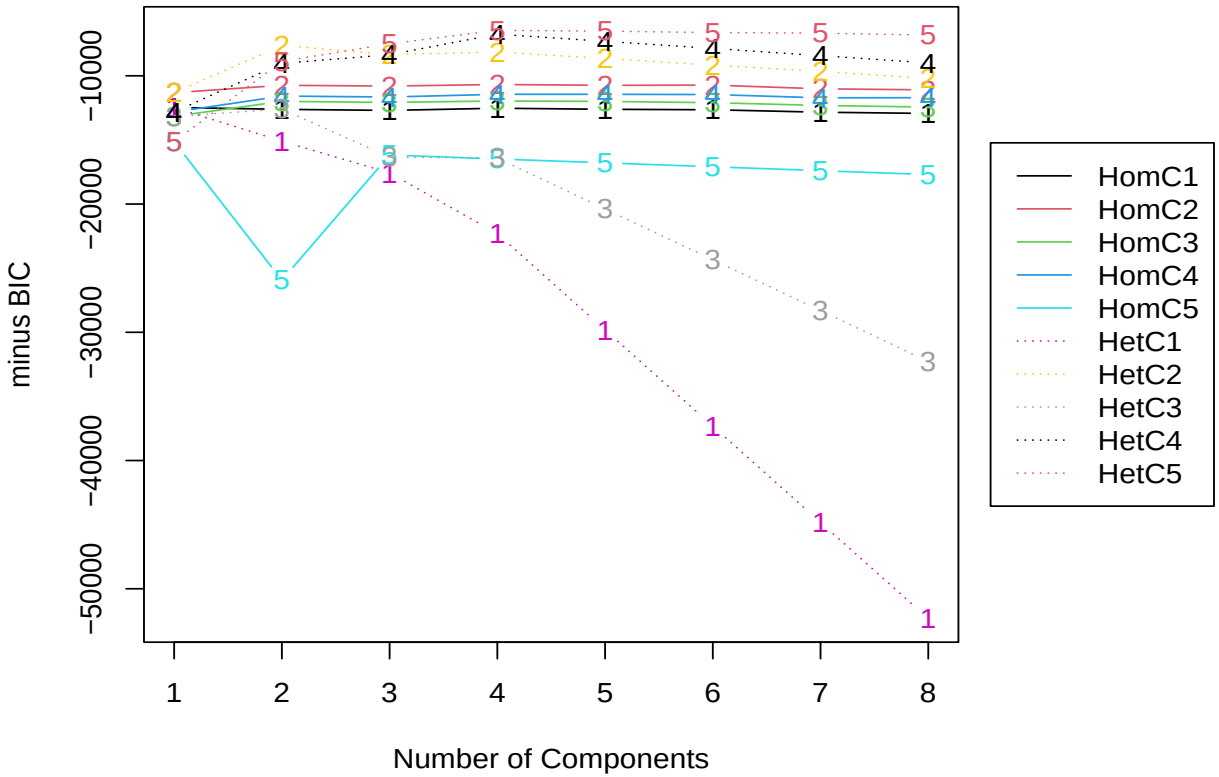
Expectation-Maximization (EM) algorithm

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

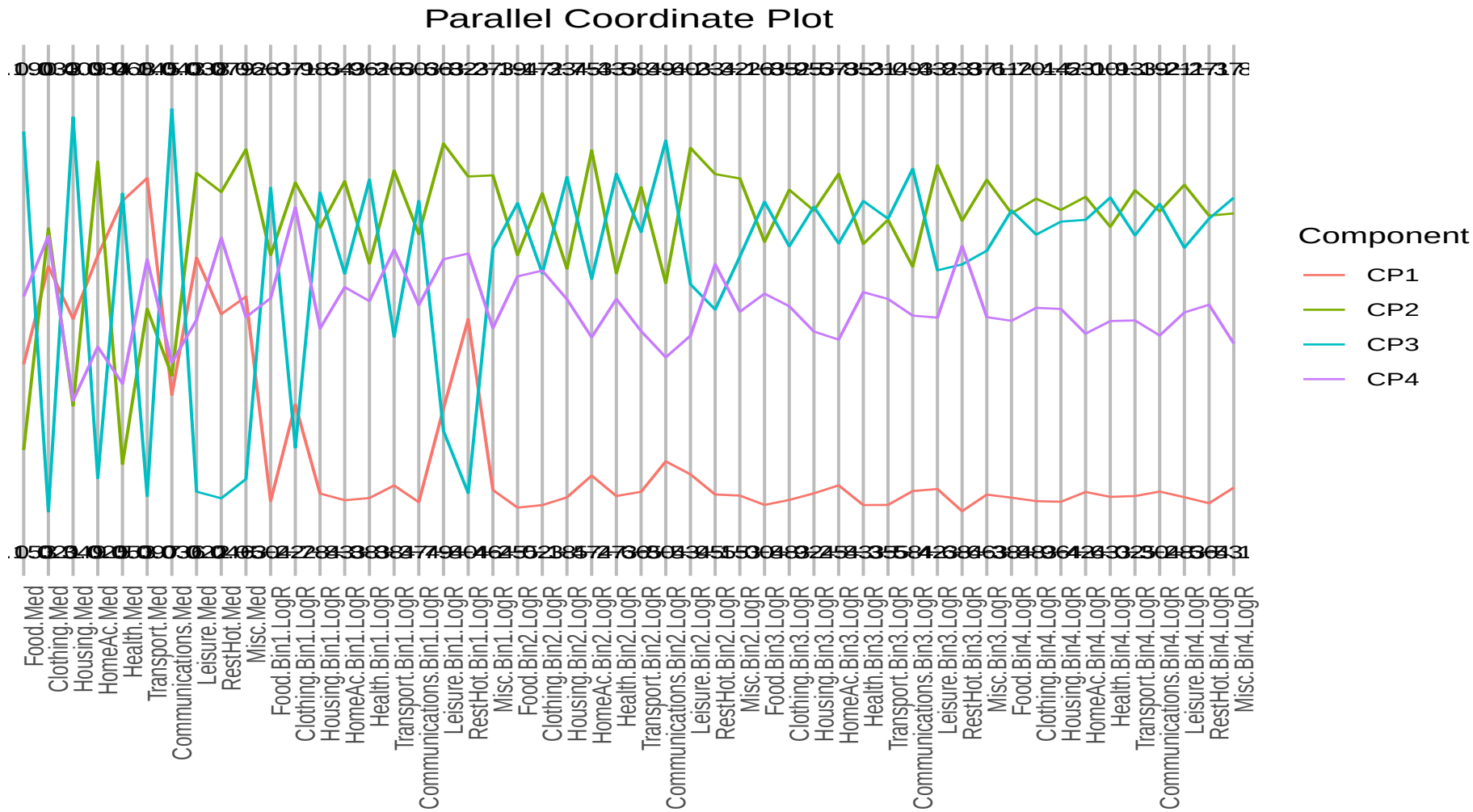
Model-Based Clustering of the HBS

BIC values, to decide on the model and number of components:

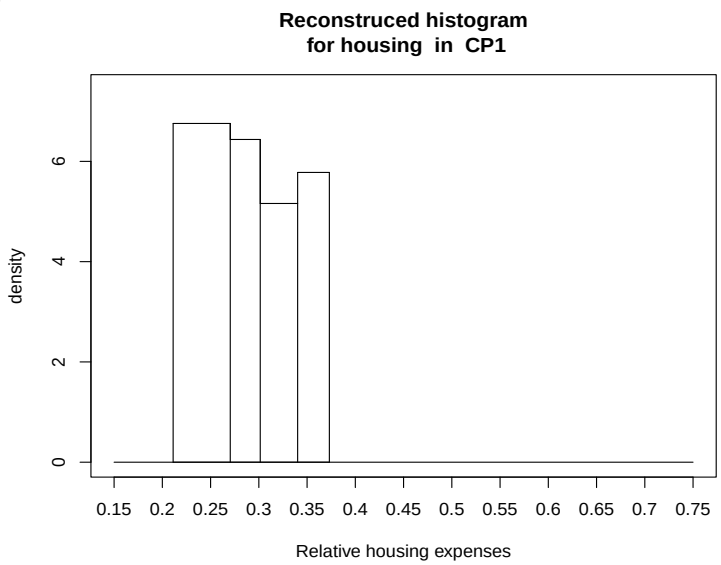


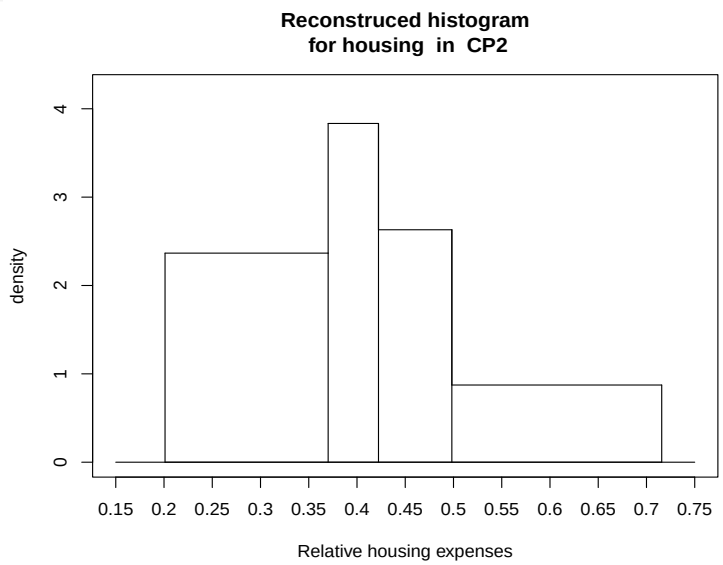
- 4 components, Config. 5 (Σ diagonal), Heterocedastic model

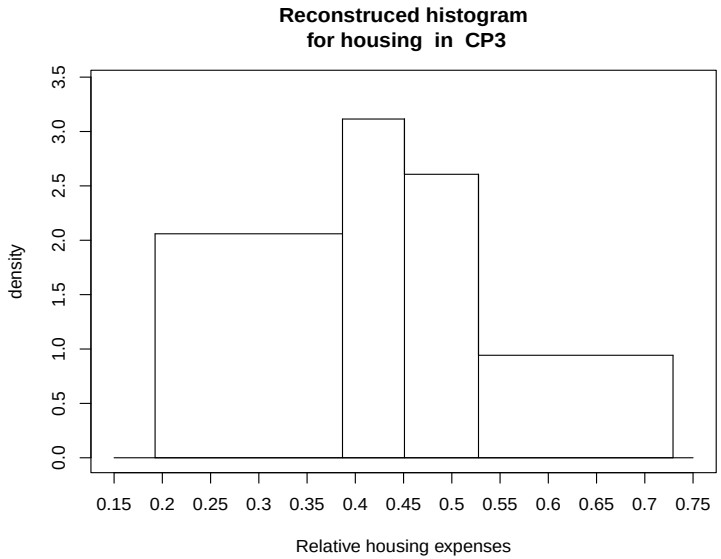
Model-Based Clustering of the HBS

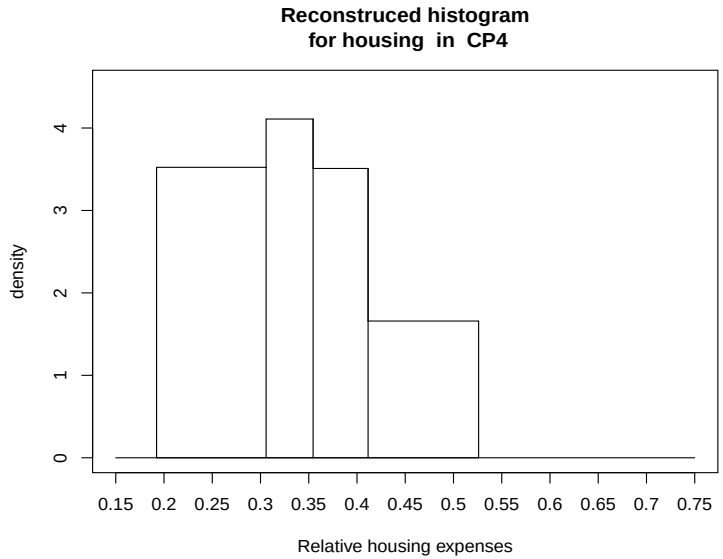


Model-Based Clustering of the HBS









Model-Based Clustering of the HBS

- **Comp. 1**: Urban Areas, from Lisbon Met Area, Algarve, and Madeira
  Low variation overall, High Median on Transports, Negative skweness on Leisure
- **Comp. 2**: Mainly Rural areas
  High variation overall, Relatively High Median on Home Acc, Leisure, Rest&Hotels
- **Comp. 3**: 63% Rural areas, mainly North, Centre, Alentejo
  High variation overall, Relatively High Median on Food, Housing, Communications
- **Comp. 4**: Urban Areas, except Lisbon Met Area and Madeira
  Medium variation overall

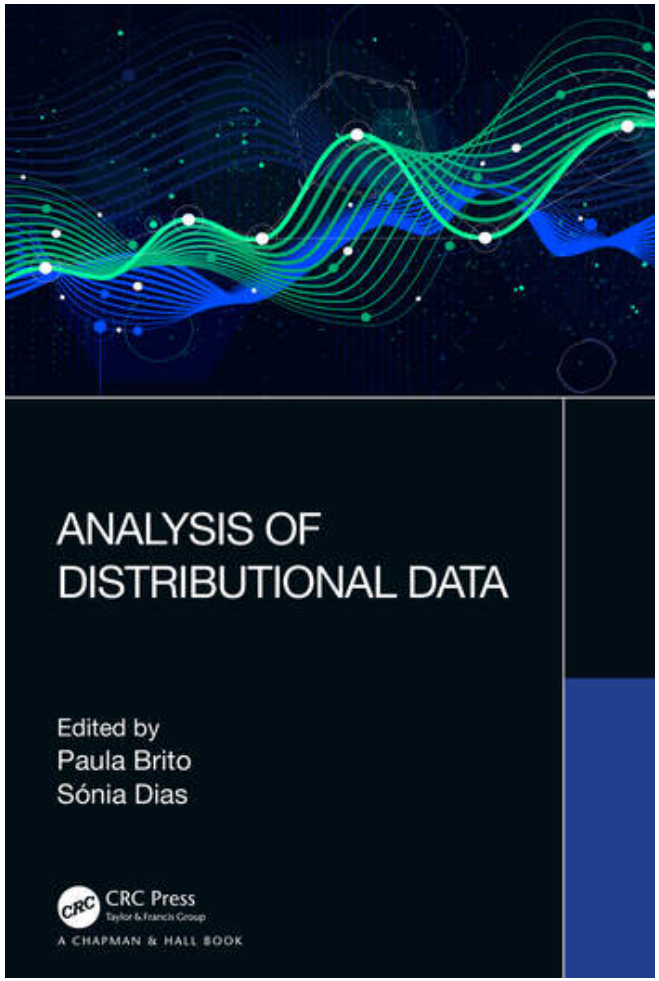
- Income classes similarly distributed among clusters

Concluding Remarks

- Parametric models specific for distributional-valued variables
- Multivariate analysis of numerical distributional data
  - Model-based clustering (finite-mixture modelling)
  - Experimental results show the pertinence and usefulness of the proposed approach

- Also being addressed:
  - Robust estimation and (distributional) outlier detection
  - Other multivariate methodologies: MANOVA, Discriminant Analysis,...
  - R Package under development

Recent Book



I Data Representation and Exploratory Analysis

II Clustering and Classification

III Dimension Reduction

IV Regression and Forecasting