MAINT.Data: Modeling and Analyzing Interval Data in R

1ères Rencontres R - Bordeaux juillet 2012

Pedro Duarte Silva
Faculdade de Economia e Gestão / CEGE, Universidade Católica Portuguesa, Porto, PORTUGAL

Paula Brito
Faculdade de Economia / LIAAD - INESC TEC, Universidade do Porto, PORTUGAL
Outline

- From Classical to Symbolic Data
- Parametric Modelization of Interval Data
  - Normal and Skew-Normal Models
  - Model configurations
- The MAINT.Data Package
  - The IData class and its basic methods
  - The IdtE classes and subclasses
  - The MANOVA, lda and qda methods for Interval Data
- Conclusions and Perspectives
From Classical to Symbolic Data

Symbolic variables: to take into account variability inherent to the data

Variability occurs when we have
Data about patients, but: analyse the healthcare centers - not the patients
Data about people, but: analyse the parishes, the cities - not the individual citizens

Variable values are sets, intervals, distributions on an underlying set of sub-intervals or categories

Micro-data → Macro-data
Symbolic data → new variable types:

- Set-valued variables: variable values are subsets of an underlying set
  - Interval variables
  - Categorical multi-valued variables

- Modal variables: variable values are distributions on an underlying set
  - Histogram variables
The dataset consists of information about patients (adults) in healthcare centers, during one semester.

<table>
<thead>
<tr>
<th>HealthCare Center</th>
<th>Age $Y_1$</th>
<th>Nb. Emergency consults $Y_2$</th>
<th>Pulse $Y_3$</th>
<th>Waiting time for consultation (min) $Y_4$</th>
<th>Education level $Y_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>[25,53]</td>
<td>{0,1,2}</td>
<td>[44,86]</td>
<td>([0,15] (0), [15,30] (0.25), [30,45] (0.5), [45,60] (0), $\geq$60 (0.25))</td>
<td>{9th grade, 1/2; Higher education, 1/2}</td>
</tr>
<tr>
<td>B</td>
<td>[33,68]</td>
<td>{1,4,5,10}</td>
<td>[54,76]</td>
<td>([0,15] (0.25), [15,30] (0.25), [30,45] (0.25), [45,60] (0.25), $\geq$60 (0))</td>
<td>{6th grade, 1/4; 9th grade, 1/4; 12th grade, 1/4; Higher education, 1/4}</td>
</tr>
<tr>
<td>C</td>
<td>[20,75]</td>
<td>{0,5,7}</td>
<td>[70,86]</td>
<td>([0,15] (0.33), [15,30] (0), [30,45] (0.33), [45,60] (0), $\geq$60 (0.33))</td>
<td>{4th grade, 1/3; 9th grade, 1/3; 12th grade 1/3}</td>
</tr>
</tbody>
</table>
## Interval Data

<table>
<thead>
<tr>
<th>( \omega_1 )</th>
<th>( Y_1 )</th>
<th>...</th>
<th>( Y_i )</th>
<th>...</th>
<th>( Y_D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( [l_{11}, u_{11}] )</td>
<td>...</td>
<td>( [l_{1j}, u_{1j}] )</td>
<td>...</td>
<td>( [l_{1p}, u_{1p}] )</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>( \omega_i )</td>
<td>( [l_{i1}, u_{i1}] )</td>
<td>...</td>
<td>( [l_{ij}, u_{ij}] )</td>
<td>...</td>
<td>( [l_{ip}, u_{ip}] )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>( \omega_n )</td>
<td>( [l_{n1}, u_{n1}] )</td>
<td>...</td>
<td>( [l_{nj}, u_{nj}] )</td>
<td>...</td>
<td>( [l_{np}, u_{np}] )</td>
</tr>
</tbody>
</table>
Interval Data Representations

Original parametrisation: \( I_{ij} = [l_{ij}, u_{ij}] \)

Alternative parametrisation: \((c_{ij}, r_{ij})\)

\[
c_{ij} = \frac{l_{ij} + u_{ij}}{2} \quad \quad r_{ij} = u_{ij} - l_{ij}
\]

MAINT.Data:

Implements parametric inference methodologies

⇒

Assumes probabilistic models for interval variables
Normal Model

Let \( R^* = \ln(R) \)

Assumption:

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

\[
\mu = \begin{bmatrix} \mu_C \\ \mu_{R^*} \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{bmatrix}
\]
Skew-Normal Model
(Azzalini 1985)

Normal model - imposes a symmetrical distribution on the midpoints and a specific relation between mean, variance and skewness for the ranges

Skew-Normal - generalizes the Gaussian by introducing an additional shape parameter $\alpha$, while trying to preserve some of its mathematical properties
Skew-Normal Model

p-variate density (Azzalini, Dalla Valle 1996):

\[
f(y) = 2\phi_p(x - \xi; \Omega) \Phi_p(\alpha^t \omega^{-1}(x - \xi))
\]

\(\xi\) - p-dimensional vector of location parameters
\(\alpha\) - p-dimensional vector of shape parameters
\(\Omega\) - symmetric positive-definite matrix
\(\omega\) - diagonal matrix formed by the square-roots of the diagonal elements of \(\Omega\)
\(\phi_p, \Phi_p\) - density and distribution function of a p-dimensional standard Gaussian vector
Skew-Normal Model

**log-likelihood** of a $p$ dimensional Skew-Normal:

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

where

$$V = \frac{1}{n} \sum_i (x_i - \xi_i)(x_i - \xi_i)^t$$

and

$$\zeta_0 (x) = \ln (2 \Phi(x))$$
<table>
<thead>
<tr>
<th>Model</th>
<th>Characterization</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Non-restricted</td>
<td>Non-restricted</td>
</tr>
<tr>
<td>2</td>
<td>$C_j$ not-correlated with $R^*_l$, $l \neq j$</td>
<td>$\Sigma_{CR^*} = \Sigma_{R^*C}$ diagonal</td>
</tr>
<tr>
<td>3</td>
<td>$Y_j$'s independent</td>
<td>$\Sigma_{CC}, \Sigma_{CR^*} = \Sigma_{R^*C}, \Sigma_{R^<em>R^</em>}$ all diagonal</td>
</tr>
<tr>
<td>4</td>
<td>$C$'s not-correlated with $R^*$'s</td>
<td>$\Sigma_{CR^*} = \Sigma_{R^*C} = 0$</td>
</tr>
<tr>
<td>5</td>
<td>All $C$'s and $R^*$'s are non-correlated</td>
<td>$\Sigma$ diagonal</td>
</tr>
</tbody>
</table>
Maximum Likelihood Estimation: Normal Model

Maximum likelihood estimator for $\mu$

$$\hat{\mu} = \overline{X}$$

Maximum likelihood estimator for $\Sigma$ under Configuration 1:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^t = : \frac{1}{n} \Sigma$$
Maximum likelihood estimator for $\Sigma$

under configurations 3, 4 and 5:
obtained from the non-restricted estimators → replacing by zeros the null parameters in the model for $\Sigma$

under configuration 2:
obtained by numerical maximization of

$$\ln L(\beta, \Sigma) = -np \ln(2\pi) - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \text{tr} E \Sigma^{-1}$$
Maximum Likelihood Estimation: Skew-Normal Model

under configuration 1

Log-likelihood:

\[ l = -\frac{1}{2} n \ln |\Omega| - \frac{1}{2} \text{tr}(\Omega^{-1} V) + \sum_i \xi_0 (\alpha^t \omega^{-1} (x_i - \xi_i)) \]

maximized in two steps.

New parameter \( \eta = \omega^{-1} \alpha \)

Then \( \hat{\Omega} = V \).

The maximization with respect to \( \eta \) and \( \xi \) is then performed numerically.
Maximum Likelihood Estimation: Skew-Normal Model

under configurations 2-5

Given that \( \Sigma = \Omega - \omega \mu Z \mu^t \omega \) a null covariance \( \Sigma(j,j') \) implies that

\[
\Omega(j,j') = \Omega(j,j)^{1/2} \mu Z_j \Omega(j',j')^{1/2} \mu Z_{j'}
\]

or, equivalently

\[
\Sigma(j,j') = 0 \Rightarrow \Omega(j,j') = \frac{2}{\pi} \frac{\Omega_j^t \omega^{-1} \alpha \alpha^t \omega^{-1} \Omega_j}{1 + \alpha^t \omega^{-1} \Omega \omega^{-1} \alpha}
\]

For configurations 2 - 5, this condition is imposed for the corresponding null elements of \( \Sigma \).
It defines a system of non-linear equations on the \( \Omega(j,j') \), which may be solved by standard numerical procedures.
Maximum Likelihood Estimation: Skew-Normal Model

under configurations 2-5

The ML estimate is then found by a Quasi-Newton optimization algorithm with:

- Analytical gradients found by the chain rule and implicit function theorem
- Randomly generated multiple starting points to avoid local optima

The MAINT.Data Package: Overview

- IData object
- Data Frame
- Grouping factor
- IdtSngDE object
- IdtHomMxE object
- IdtHetMxE object
- Idtlda object
- Idtqda object
The MAINT.Data Package: The IData class

IData function

IData Methods

- print
- summary
- indexing
- assignment
- mle
- MANOVA
The MAINT.Data Package: The IdtE classes I -- Single Dist.

mle.IData arguments

- Model
- Config
- SelCrit

IdtSngE Methods

- print
- summary
- coef
- stdEr
- testMod
- ...

mle method

IData object

IdtSngDE object

- **IData object**
  - MANOVA method
  - IdtHomMxE object
  - Lda method

- **Grouping factor**
  - IdtHomMxE
  - IdtLda

**IdtHomMxE Methods**
- print
- summary
- coef
- stdEr

**IdtLda Methods**
- print
- summary
- Lda
- predict
- ...
- print
- summary
- predict
- ...

- **IData object**
- **MANOVA method**
- **qda method**
- **Grouping factor**
- **IdtHetMxE object**
- **Idtqda object**

**IdtHetMxE Methods**
- print
- summary
- coef
- stdEr

**Idtqda Methods**
- testMod
- qda
- print
- summary
- predict

...
Creating Idata Objects

ChinaT <- IData(ChinaTemp[1:8],
VarNames=c("Q1","Q2", "Q3","Q4"))

#Display the first three observations

head(ChinaT,n=3)

<table>
<thead>
<tr>
<th>AnQing_1974</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.673, 14.827]</td>
<td>[13.435, 28.465]</td>
<td>[19.821, 31.179]</td>
<td>[2.216,  9.984]</td>
<td></td>
</tr>
<tr>
<td>[0.673, 14.827]</td>
<td>[13.435, 28.465]</td>
<td>[19.821, 31.179]</td>
<td>[2.216,  9.984]</td>
<td></td>
</tr>
<tr>
<td>[2.319, 14.381]</td>
<td>[12.829, 28.471]</td>
<td>[23.192, 32.308]</td>
<td>[1.013, 10.987]</td>
<td></td>
</tr>
<tr>
<td>[0.906, 12.494]</td>
<td>[11.795, 28.405]</td>
<td>[19.680, 34.120]</td>
<td>[2.992, 10.308]</td>
<td></td>
</tr>
</tbody>
</table>
# MANOVA tests

```r
ManvChina <- MANOVA(ChinaT,ChinaTemp$GeoReg)
print(ManvChina)
```

Null Model Log likelihoods:

<table>
<thead>
<tr>
<th></th>
<th>NC1</th>
<th>NC2</th>
<th>NC3</th>
<th>NC4</th>
<th>NC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC1</td>
<td>-7336.254</td>
<td>-8331.416</td>
<td>-11564.904</td>
<td>-8390.351</td>
<td>-12648.760</td>
</tr>
</tbody>
</table>

Full Model Log likelihoods:

<table>
<thead>
<tr>
<th></th>
<th>NC1</th>
<th>NC2</th>
<th>NC3</th>
<th>NC4</th>
<th>NC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC1</td>
<td>-6209.280</td>
<td>-6820.555</td>
<td>-9049.276</td>
<td>-6857.536</td>
<td>-9450.228</td>
</tr>
</tbody>
</table>

Full Model Akaike Information Criteria:

<table>
<thead>
<tr>
<th></th>
<th>NC1</th>
<th>NC2</th>
<th>NC3</th>
<th>NC4</th>
<th>NC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC1</td>
<td>12586.56</td>
<td>13793.11</td>
<td>18234.55</td>
<td>13851.07</td>
<td>19012.46</td>
</tr>
</tbody>
</table>

Selected Model:

[1] "NC1"

Null Model log-likelihood: -7336.254
Full Model log-likelihood: -6209.28
Qui-squared statistic: 2253.949
degrees of freedom: 40
p-value ≈ 0
Chinalda <- lda(ManvChina)

PredRes <- predict(Chinalda,ChinaT)

# Estimate error rates by ten-fold cross-validation
CVlda <- DACrossVal(ChinaT,ChinaTemp$GeoReg,TrainAlg=lda, Config=BestModel(ManvChina@H1res),CVrep=1)
Conclusions and Perspectives

- Probabilistic Models proposed for Interval Variables

- Normal (and Skew-Normal) distributions (different configurations) for Midpoints and Log-Ranges

- Implemented as an R package based on Maximum-Likelihood Estimation S4 classes and methods
Conclusions and Perspectives

- Current version includes tools for:
  - Single distribution estimation and inference
  - ANOVA and MANOVA
  - Linear and Quadratic Discriminant Analysis

- Perspectives:
  - Extension to other multivariate methodologies (ex: Im method...)
  - Assume different distributions


