MAINT.Data: Modeling and Analyzing Interval Data in R

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

Paula Brito
Faculdade de Economia / LIAAD - INESC Porto LA
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 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

Symbolic data array

The dataset consists of information's about patients (adults) in healthcare centers, during one semester.

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

Interval Data

	Y_1	…	Y_j	…	Y_p
ω_1	$[l_{11}, u_{11}]$	…	$[l_{1j}, u_{1j}]$	…	$[l_{1p}, u_{1p}]$
…	…		…		…
ω_i	$[l_{i1}, u_{i1}]$	…	$[l_{ij}, u_{ij}]$	…	$[l_{ip}, u_{ip}]$
…	…		…		…
ω_n	$[l_{n1}, u_{n1}]$	…	$[l_{nj}, u_{nj}]$	…	$[l_{np}, u_{np}]$

Interval Data Representations

Original parametrisation : $I_{ij} = [l_{ij} , u_{ij}]$

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

$$c_{ij} = \frac{l_{ij} + u_{ij}}{2} \qquad r_{ij} = u_{ij} - l_{ij}$$

MAINT.Data:

Implements parametric inference methodologies

$\Rightarrow$

Assumes probabilistic models for interval variables

Normal Model

Let $R^* = \ell n(R)$

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 α, while trying to preserve some of its mathematical properties

Skew-Normal Model

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

$$f(y) = 2\phi_p(x - \xi; \Omega)\, \Phi_p\left(\alpha^t \omega^{-1}(x - \xi)\right)$$

ξ - p-dimensional vector of location parameters

α - p-dimensional vector of shape parameters

Ω - symmetric positive-definite matrix

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

ϕ_p , Φ_p - density and distribution function of a p-dimensional standard Gaussian vector

Skew-Normal Model

log-likelihood of a p dimensional Skew-Normal :

$$l = -\frac{1}{2} n \,\ell n|\Omega| - \frac{1}{2} \operatorname{tr}\left(\Omega^{-1} V\right) + \sum_i \zeta_0\left(\alpha^t \omega^{-1}(x_i - \xi_i)\right) \quad (*)$$

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

and $$\zeta_0(x) = \ell n\,(2\,\Phi(x))$$

Model Configurations

Model	Characterization	Σ
1	Non-restricted	Non-restricted
2	C_j not-correlated with R_l^* $l \neq j$	$\Sigma_{CR*} = \Sigma_{R*C}$ diagonal
3	Y_j's independent	$\Sigma_{CC}, \Sigma_{CR*} = \Sigma_{R*C}, \Sigma_{R*R*}$ all diagonal
4	C's not-correlated with R*'s	$\Sigma_{CR*} = \Sigma_{R*C} = 0$
5	All C's and R*'s are non-correlated	Σ diagonal

Maximum Likelihood Estimation: Normal Model

Maximum likelihood estimator for μ

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

Maximum likelihood estimator for Σ under Configuration 1:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^t =: \frac{1}{n}E$$

Maximum Likelihood Estimation: Normal Model

Maximum likelihood estimator for Σ

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

under configuration 2:
obtained by numerical maximization of

$$\ell n L(\hat{\mu}, \Sigma) = -np\, \ell n(2\pi) - \frac{n}{2} \ell n \left|\Sigma\right| - \frac{1}{2} \operatorname{tr} E \Sigma^{-1}$$

Maximum Likelihood Estimation: Skew-Normal Model

under configuration 1

Log-likelihood :

$$l = -\frac{1}{2} n \,\ell n|\Omega| - \frac{1}{2} tr\left(\Omega^{-1} V\right) + \sum_i \zeta_0\left(\alpha^t \omega^{-1}(x_i - \xi_i)\right) \quad (*)$$

maximized in two steps.

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

Then $\hat{\Omega} = V$.

The maximization with respect to η and ξ is then performed numerically.

Maximum Likelihood Estimation: Skew-Normal Model

under configurations 2-5

Given that $\Sigma = \Omega - \omega\mu_Z\mu_Z^t\omega$ 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 Σ.
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

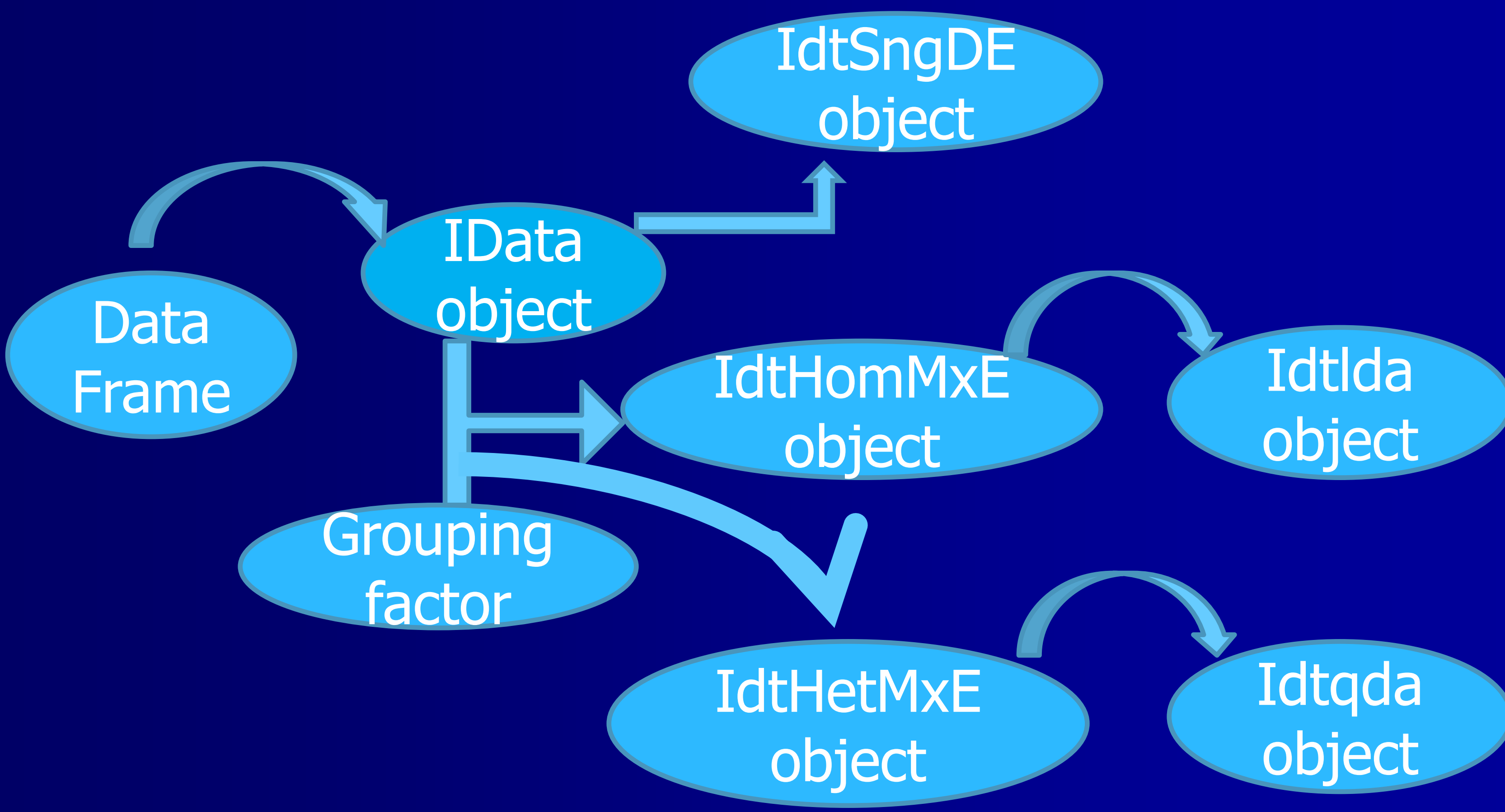
under configurations 2-5

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

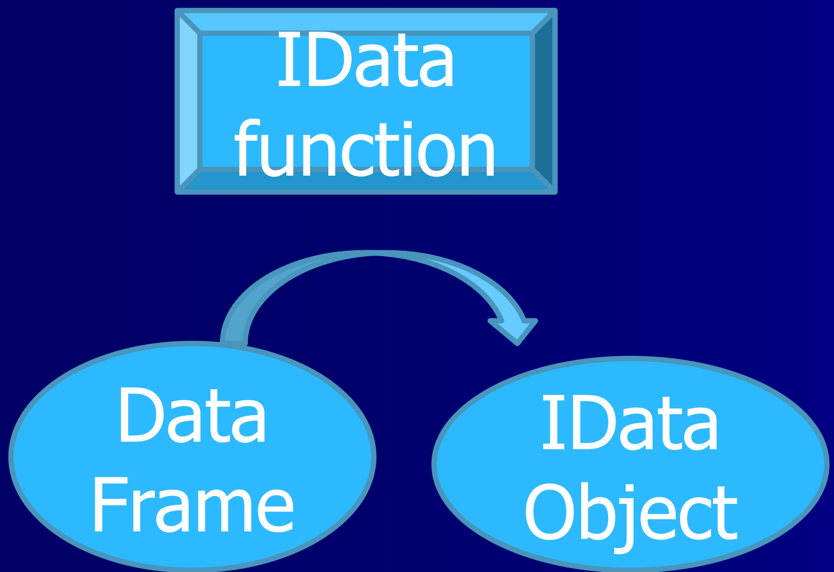
- Analytical gradients found by the chain rule and implicit function theorem
- Randomly generated multiple starting points to avoid local optima

Brito, P., Duarte Silva, A. P. (2011): "Modelling Interval Data with Normal and Skew-Normal Distributions”.
Journal of Applied Statistics (in press).

The MAINT.Data Package: Overview



The MAINT.Data Package: The Idata class

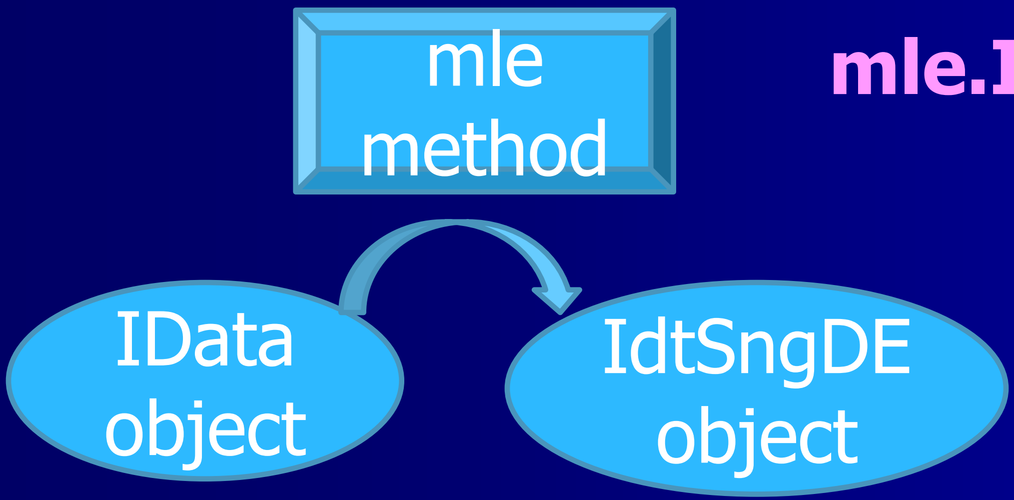


IData Methods

- print
- summary
- indexing
- assignment

  …
- mle
- MANOVA

The MAINT.Data Package: The IdtE classes I -- Single Dist.



mle.IData arguments

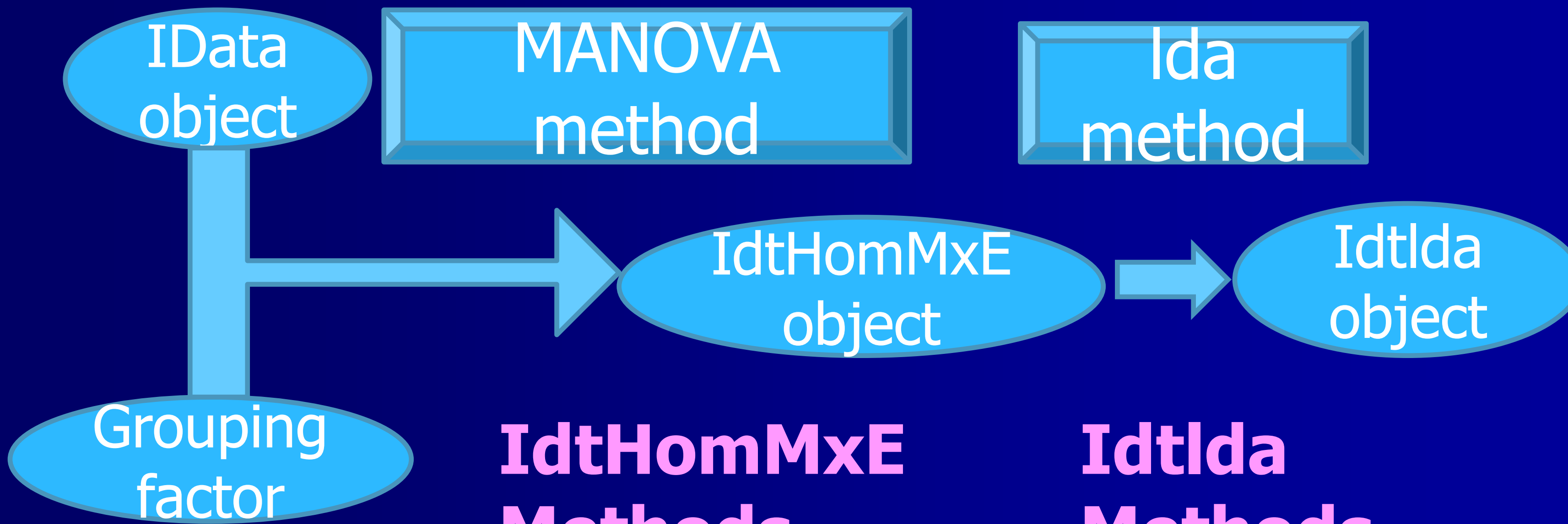
- Model
- Config
- SelCrit

IdtSngE Methods

- print
- summary
- coef
- stdEr
- testMod

...

The MAINT.Data Package: The IdtE classes II – Hom. Mixt.



IdtHomMxE Methods

- print
- summary
- coef
- stdEr
- testMod
- lda

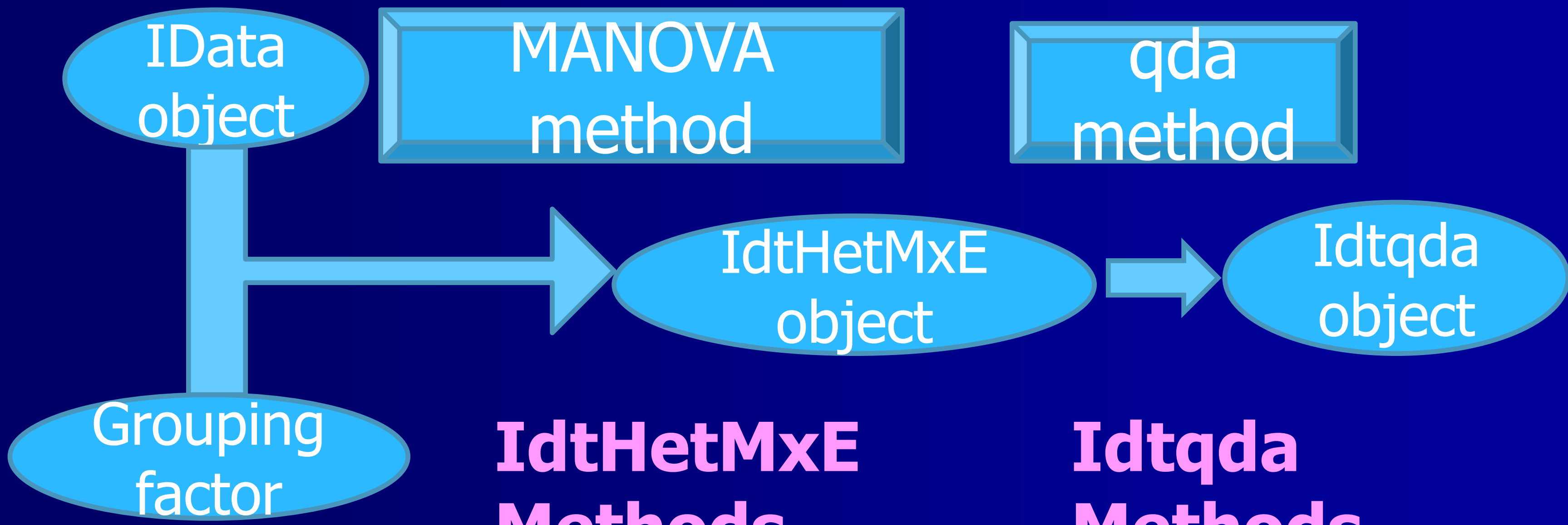
...

Idtlda Methods

- print
- summary
- predict

...

The MAINT.Data Package: The IdtE classes III – Het. Mixt.



IdtHetMxE Methods

- print
- summary
- coef
- stdEr
- testMod
- qda
- …

Idtqda Methods

- 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)

                    Q1                  Q2                 Q3               Q4
AnQing_1974  [ 0.673, 14.827]  [13.435, 28.465]  [19.821, 31.179]  [2.216,  9.984]
AnQing_1975  [ 2.319, 14.381]  [12.829, 28.471]  [23.192, 32.308]  [1.013, 10.987]
AnQing_1976  [ 0.906, 12.494]  [11.795, 28.405]  [19.680, 34.120]  [2.992, 10.308]
```

MANOVA tests

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

```
Null Model Log likelihoods:
      NC1          NC2            NC3          NC4         NC5
 -7336.254    -8331.416    -11564.904    -8390.351    -12648.760
Full Model Log likelihoods:
      NC1          NC2            NC3          NC4          NC5
-6209.280    -6820.555     -9049.276    -6857.536     -9450.228
Full Model Akaike Information Criteria:
     NC1          NC2          NC3        NC4          NC5
12586.56    13793.11    18234.55   13851.07    19012.46
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
```

Linear Discriminant Analysis

```
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: lm method…)
  - Assume different distributions

References

- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12, 171-178.
- Azzalini, A. and Dalla Valle, A. (1996). The multivariate SkN-normal distribution. *Biometrika* 83(4), 715-726.
- Bock, H.-H. and Diday, E. (2000). *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Heidelberg.
- Brito, P., Duarte Silva, A.P. (2011): Modelling Interval Data with Normal and Skew-Normal Distributions. Journal of Applied Statistics (in press).
- Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.