Package 'PUGMM'

January 20, 2025

Version 0.1.0

Title Parsimonious Ultrametric Gaussian Mixture Models

Description Finite Gaussian mixture models with parsimonious extended ultrametric covariance structures estimated via a grouped coordinate ascent algorithm, which is equivalent to the Expectation-Maximization algorithm. The thirteen ultrametric covariance structures implemented allow for the inspection of different hierarchical relationships among variables. The estimation of an ultrametric correlation matrix is included as a function. The methodologies are described in Cavicchia, Vichi, Zaccaria (2024) <doi:10.1007/s11222-024-10405-9>, Cavicchia, Vichi, Zaccaria (2022) <doi:10.1007/s11634-021-00488-x> and Cavicchia, Vichi, Zaccaria (2020) <doi:10.1007/s11634-020-00400-z>.

Depends R (>= 4.0)

Imports ClusterR, doParallel, foreach, igraph, MASS, Matrix, mclust, mcompanion, ppclust

License MIT + file LICENSE

URL https://github.com/giorgiazaccaria/PUGMM

BugReports https://github.com/giorgiazaccaria/PUGMM/issues

NeedsCompilation no

LazyData yes

Encoding UTF-8

RoxygenNote 7.3.1

Maintainer Giorgia Zaccaria <giorgia.zaccaria@unimib.it>

Author Giorgia Zaccaria [aut, cre] (<https://orcid.org/0000-0001-9119-9104>), Carlo Cavicchia [aut] (<https://orcid.org/0000-0003-1816-3521>), Lorenzo Balzotti [aut] (<https://orcid.org/0000-0001-6191-9801>)

Repository CRAN

Date/Publication 2024-05-10 13:40:02 UTC

Contents

`penguins` *Penguins*

Description

The data set contains five measurements made on 342 penguins which are classified into three species.

Usage

```
data(penguins)
```

Format

A data frame with 342 observations and 5 variables, which are described as follows.

species Penguin species (Chinstrap, Adélie, or Gentoo)

culmen_length_mm Culmen length (mm)

culmen_depth_mm Culmen depth (mm)

flipper_length_mm Flipper length (mm)

body_mass_g Body mass (g)

Details

Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network. The categorical variables 'island' and 'sex' have been removed from the original dataset, as well as the incomplete observations on the five variables reported herein.

Source

Dataset downloaded from Kaggle https://www.kaggle.com/code/parulpandey/penguin-dataset-the-new-iris.

References

Gorman, K.B., Williams T.D., Fraser W.R. (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). *PLoS ONE*, 9(3), e90081.

Examples

```
data(penguins)
```

`plot.pugmm` *Plotting method for* `pugmm` *object*

Description

Plots for Parsimonious Ultrametric Gaussian Mixture Models results, such as BIC and path diagrams.

Usage

```
## S3 method for class 'pugmm'
plot(x, what = NULL, nrow = NULL, ncol = NULL, cluster_names = NULL, ...)
```

Arguments

`x`	Output from `pugmm`.
`what`	A string specifying the type of graph requested. Available choices are: `"BIC"` Plot of BIC values for the fitted models. For each G, the best BIC among the ones corresponding to different m is displayed. `"Path Diagram"` Path diagram representation of the extended ultrametric covariance matrix per component for the best model.
`nrow`	Number of rows in the graphical window. A new graphical window is opened every 6 plots, i.e., components of `pugmm`.
`ncol`	Number of columns in the graphical window. A new graphical window is opened every 6 plots, i.e., components of `pugmm`.
`cluster_names`	String of dimension G with the clusters/components' name.
`...`	Other graphics parameters.

Value

No return value since this is a plot method.

See Also

`pugmm()`

Examples

```
data(penguins)
x <- scale(penguins[, 2:5])
pugmm.penguins <- pugmm(x, 3, 1)
plot.pugmm(pugmm.penguins, what = c("BIC", "Path Diagram"))
```

---

pugmm *Parsimonious Ultrametric Gaussian Mixture Models*

Description

Model-based clustering via Parsimonious Ultrametric Gaussian Mixture Models. Hierarchical relationships among variables within and between clusters are inspected. The grouped coordinate ascent algorithm is used for the parameter estimation. The optimal model is selected according to BIC.

Usage

```
pugmm(
  X,
  G = NULL,
  m = NULL,
  normalization = NULL,
  model = NULL,
  maxiter = 500,
  tol = 1e-06,
  stop = "aitken",
  rndstart = 1,
  initG = "kmeans",
  initm = "ucms",
  gaussian = "mclust",
  parallel = FALSE
)
```

Arguments

X	$(n \times p)$ numeric matrix or data frame, where n and p represent the number of units and variables, respectively. Categorical variables are not allowed.
G	Integer (vector) specifying the number of mixture components (default: G = 1:5).
m	Integer (vector) specifying the number of variable groups (default: m = 1:5).
normalization	Character string specifying the data transformation. If NULL, no transformation is applied to the data matrix (default). Other options are: "standard" for the standardization; "center" for centering the data; "range" for the MinMax transformation; "SVD" for the Singular Value Decomposition transformation.

model	Vector of character strings indicating the model names to be fitted. If NULL, all the possible models are fitted (default). See the possible models using available_models().
maxiter	Integer value specifying the maximum number of iterations of the algorithm (default: maxiter = 500).
tol	Numeric value specifying the tolerance for the convergence criterion (default: tol = 1e-6).
stop	Character string specifying the convergence criteria. If "aitken", the Aitken acceleration-based stopping rule is used (default); if "relative", the relative log-likelihood in two sequential iterations is evaluated.
rndstart	Integer value specifying the number of random starts (default: rndstart = 1).
initG	Character string specifying the method for the initialization of the unit-component membership. If "kmeans", k-means via RcppArmadillo is used (default). Other options are: "random" for random assignment; "kmeansf" for fuzzy c-means (via the function fcm of the package ppclust).
initm	Character string specifying the method for the initialization of the variable-group membership. If "ucms", the multivariate model to be used for obtaining the variable-group membership estimated is the same model.name used for estimating the Parsimonious Ultrametric Gaussian Mixture Model (default); if "random", a random assignment is performed.
gaussian	Character string specifying the way to compute the log-likelihood. If "mclust", dmvnorm of mclust is used (default); if "canonical", the log-likelihood computation is based upon the canonical representation of an extended ultrametric covariance matrix.
parallel	A logical value, specifying whether the models should be run in parallel.

Details

The grouped coordinate ascent algorithm used for the estimation of PUGMMs parameters was demonstrated to be equivalent to an Expectation-Maximization algorithm in the GMM framework (Hathaway, 1986).

Value

An object of class pugmm containing the results of the optimal - according to BIC - Parsimonious Ultrametric Gaussian Mixture Model estimation.

call Matched call.

X Input data matrix.

G Number of components of the best model.

m Number of variable groups of the best model.

label Integer vector of dimension n, taking values in $\{1, \ldots, G\}$. It identifies the unit classification according to the maximum a posteriori of the best model.

pp Numeric vector of dimension G containing the prior probabilities for the best model.

mu ($G \times p$) numeric matrix containing the component mean vectors (by row) for the best model.

`sigma` List of dimension G containing the $(p \times p)$ numeric component extended ultrametric covariance matrices for the best model.

`V` List of dimension G containing the $(p \times m)$ binary variable-group membership matrices for the best model.

`Sv` List of dimension G containing the $(m \times m)$ numeric diagonal matrices of the group variances for the best model.

`Sw` List of dimension G containing the $(m \times m)$ numeric diagonal matrices of the within-group covariances for the best model.

`Sb` List of dimension G containing the $(m \times m)$ numeric hallow matrices of the between-group covariances for the best model.

`post` $(n \times G)$ numeric matrix containing the posterior probabilities for the best model.

`pm` Number of parameters of the best model.

`pm.cov` Number of covariance parameters of the best model.

`pm.free` Number of free parameters of the best model (`pm` - (constraints on V + `count.constr.SwSb` + `count.constr.SvSw`)).

`count.constr.SwSb` Number of times the constraint between `Sw` and `Sb` has been turned on for the best model.

`count.constr.SvSw` Number of times the constraint between `Sv` and `Sw` has been turned on for the best model.

`BIC` BIC values for all the fitted models. If BIC is NA, the model has not been computed since its structure is equal to another model, while if BIC is $-Inf$ the solution has a number of clusters $< G$.

`bic` BIC value of the best model.

`loglik` Log-likelihood of the best model.

`loop` Random start corresponding to the selected solution of the best model.

`iter` Number of iterations needed to estimate the best model.

`model.name` Character string denoting the PUGMM model name of the best model among the ones fitted.

`messages` Messages.

References

Cavicchia, C., Vichi, M., Zaccaria, G. (2024) Parsimonious ultrametric Gaussian mixture models. *Statistics and Computing*, 34, 108.

Cavicchia, C., Vichi, M., Zaccaria, G. (2022) Gaussian mixture model with an extended ultrametric covariance structure. *Advances in Data Analysis and Classification*, 16(2), 399-427.

Hathaway, R. (1986) Another interpretation of the EM algorithm for mixture distributions. *Statistics and Probability Letters*, 4(2), 53-56.

See Also

`pugmm_available_models()`, `plot.pugmm()`

Examples

```
data(penguins)
x <- scale(penguins[, 2:5])
pugmm.penguins <- pugmm(x, 3, 1)
table(penguins$species, pugmm.penguins$label)

pugmm.penguins <- pugmm(x)
pugmm.penguins$G
pugmm.penguins$m
pugmm.penguins$model.name
```

`pugmm_available_models` *PUGMM Model Names*

Description

Description of the model names used in the *PUGMM* package.

Usage

```
pugmm_available_models()
```

Details

The PUGMM model names in the *PUGMM* package are characterized by four letters:

- First letter: it refers to the variable-group membership matrix V, which can be equal (E) or free to vary (F) across components.
- Second, third, fourth letters: they refer to the matrices of the group variances Σ_V, the within-group covariances Σ_W and the between-group covariances Σ_B, respectively, by indicating if they are unique (U, i.e., equal within and across components), isotropic (I, i.e., equal within components), equal (E, i.e., equal across components) or free to vary across components (F).

Value

Available models in PUGMM, i.e., the thirteen extended ultrametric covariance structures of PUGMM.

References

Cavicchia, C., Vichi, M., Zaccaria, G. (2024) Parsimonious ultrametric Gaussian mixture models. *Statistics and Computing*, 34, 108.

See Also

`pugmm()`

Examples

```
pugmm_available_models()
```

`rand.member` *Random partition of objects into classes*

Description

Performs a random partition of objects into classes.

Usage

```
rand.member(n.obs, G)
```

Arguments

`n.obs`	Number of objects
`G`	Number of classes

Details

No empty classes can occur.

Value

A binary and row-stochastic matrix with $n.obs$ rows and G columns.

Examples

```
rand.member(10, 3)
```

`UCM` *Ultrametric Correlation Matrix*

Description

Fit an ultrametric correlation matrix on a nonnegative correlation one.

Usage

```
UCM(R, m, rndstart, maxiter = 100, eps = 1e-06)
```

Arguments

R	$(p \times p)$ nonnegative correlation matrix.
m	Integer specifying the number of variable groups.
rndstart	Integer value specifying the number of random starts.
maxiter	Integer value specifying the maximum number of iterations of the EM algorithm (default: maxiter = 100).
eps	Numeric value specifying the tolerance for the convergence criterion used in the coordinate descent algorithm (default: eps = 1e-6).

Value

A list with the following elements:

call Matched call.

V Optimal binary and row-stochastic $(p \times m)$ variable-group membership matrix.

Rt Optimal $(p \times p)$ ultrametric correlation matrix.

Rw Optimal $(m \times m)$ within-concept consistency (diagonal) matrix.

Rb Optimal $(m \times m)$ between-concept correlation matrix.

of Objective function corresponding to the optimal solution.

loop Random start corresponding to the optimal solution.

iter Number of iterations needed to obtain the optimal solution.

References

Cavicchia, C., Vichi, M., Zaccaria, G. (2020) The ultrametric correlation matrix for modelling hierarchical latent concepts. *Advances in Data Analysis and Classification*, 14(4), 837-853.

Examples

```
data(penguins)
R <- cor(penguins[, 2:5])
UCM(R, 4, 1)
```

Index