Package 'teigen'

August 19, 2014

Type Package

Title Model-based clustering and classification with the multivariate t-distribution

Version 2.0.7

Date 2014-08-19

Author Jeffrey L. Andrews, Paul D. McNicholas

Maintainer Jeffrey L. Andrews <jeffrey.andrews@macewan.ca>

Description Fits mixtures of multivariate t-distributions (with eigen-decomposed covariance structure) via the multi-cycle ECM algorithm under a clustering or classification paradigm.

License GPL (>= 2)

LazyLoad yes

Imports parallel

NeedsCompilation no

Repository CRAN

Date/Publication 2014-08-19 22:33:27

R topics documented:

`teigen-package` *teigen: Model-based clustering and classification with the multivariate t-distribution*

Description

Fits mixtures of multivariate t-distributions (with eigen-decomposed covariance structure) via the multi-cycle ECM algorithm under a clustering or classification paradigm.

Details

Package:	teigen
Type:	Package
Version:	2.0.7
Date:	2014-08-19
License:	GPL (>=2)
LazyLoad:	yes

Author(s)

Jeffrey L. Andrews, Paul D. McNicholas

Maintained by: Jeffrey L. Andrews <jeffrey.andrews@macewan.ca>

References

Andrews JL and McNicholas PD (2012). "Model-based clustering, classification, and discriminant analysis with the multivariate *t*-distribution: The *t*EIGEN family" *Statistics and Computing* 22(5), 1021–1029.

Andrews JL, McNicholas PD, and Subedi S (2011) "Model-based classification via mixtures of multivariate t-distributions" *Computational Statistics and Data Analysis* 55, 520–529.

See Also

`teigen` for main function

`plot.teigen` *plot.teigen: Plotting function for teigen objects*

Description

Outputs marginal contour or uncertainty plots to the graphics device for objects of class `teigen`.

Usage

```
## S3 method for class 'teigen'
plot(x, xmarg = 1, ymarg = 2, res = 200, levels = c(seq(0.01, 1, by = 0.01), 0.001),
    what = c("contour", "uncertainty"), main=NULL, xlab=NULL, legend=TRUE, ...)
```

Arguments

x	An object of class teigen
xmarg	Scalar argument giving the number of the variable to be used on the x-axis
ymarg	Scalar argument giving the number of the variable to be used on the y-axis
res	Scalar argument giving the resolution for the calculation grid required for the contour plot. Default is 200, which results in a 200x200 grid.
levels	Numeric vector giving the levels at which contours should be drawn. Default is to draw a contour in 0.01 steps, plus a contour at 0.001. This may result in more/less contours than desired depending on the resulting density.
what	Character vector stating which plots should be sent to the graphics device. Choices are "contour" or "uncertainty".
main	Optional character string for title of plot. Useful default if left as NULL.
xlab	Optional character string for x-axis label.
legend	Logical for a default generation of a legend in the top left corner of the plot(s).
...	Options to be passed to plot.

Details

"contour" plots the marginal distribution of the mixture distribution. For univariate data, or if ymarg is NULL, a univariate marginal is provided that includes the kernel density estimate from density(), the mixture distribution, and colour-coded component densities.

"uncertainty" plots the uncertainty of each observation's classification - the larger the point, the more uncertainty associated with that observation. Uncertainty in this context refers to the probability that the observation arose from the mixture component specified by the colour in the plot rather than the other components.

Author(s)

Jeffrey L. Andrews

See Also

teigen

print.teigen *print.teigen: Print function for teigen objects*

Description

Outputs summary information.

Usage

```
## S3 method for class 'teigen'
print(x, ...)
```

Arguments

x	An object of class teigen
...	Options to be passed to print.

Author(s)

Jeffrey L. Andrews

See Also

teigen

summary.teigen *summary.teigen: Summary function for teigen objects*

Description

Gives summary information.

Usage

```
## S3 method for class 'teigen'
summary(object, ...)
```

Arguments

object	An object of class teigen
...	Options to be passed to summary.

Author(s)

Jeffrey L. Andrews

See Also

teigen

---

teigen *teigen: Function for model-based clustering and classification with the multivariate t-distribution*

Description

Fits multivariate t-distribution mixture models (with eigen-decomposed covariance structure) to the given data within a clustering paradigm (default) or classification paradigm (by giving either training index or percentage of data taken to be known).

Usage

```
teigen(x, Gs = 1:9, models = "all", init ="kmeans", scale = TRUE, dfstart = 50,
     clas = 0, known= NULL, training = NULL, gauss = FALSE, dfupdate = "approx",
      eps = c(0.001, 0.1), verbose=TRUE, anneal=NULL, maxit=c(20,1000))
```

Arguments

x	A numeric matrix, data frame, or vector (for univariate data) .
Gs	A number or vector indicating the number of groups to fit. Default is 1-9.
models	A character vector giving the models to fit. See details for a comprehensive list of choices.
init	A list of initializing classification of the form that `init[[G]]` contains the initializing vector for all G considered (see example below). Alternatively, the user can use a character string indicating initialization method. Currently the user can choose from "kmeans" (default), 'hard' random - "hard", 'soft' random - "soft", and "uniform" (classification only).
scale	Logical indicating whether or not the function should scale the data. Default is TRUE and is the prescribed method — tEIGEN models are not scale invariant.
dfstart	The initialized value for the degrees of freedom. The default is 50.
clas	Value between 0-100 indicating the percentage of data taken to be known. Note that a vector of known classifications is needed. See next argument for an alternative. Default is 0 and performs clustering, otherwise the algorithm chooses the training index randomly (and will return it via index).
training	Optional indexing vector for the observations whose classification is taken to be known.
known	A vector of known classifications that can be numeric or character - optional for clustering, necessary for classification. Must be the same length as the number of rows in the data set. If using in a true classification sense, give samples with unknown classification the value NA within known (see training example below).

gauss	Logical indicating if the algorithm should use the gaussian distribution. If models="mclust" or "gaussian" then gauss=TRUE is forced.
dfupdate	Character string or logical indicating how the degrees of freedom should be estimated. The default is "approx" indicating a closed form approximation be used. Alternatively, "numeric" can be specified which makes use of uniroot. If FALSE, the value from dfstart is used and the degrees of freedom are not updated. If TRUE, "numeric" will be used for back-compatibility.
eps	Vector (of size 2) giving tolerance values for the convergence criterion. First value is the tolerance level for iterated M-steps. Second value is tolerance for the EM algorithm: convergence is based on Aitken's acceleration, see cited papers for more information.
verbose	Logical indicating whether the running output should be displayed. What is displayed depends on the width of the R window. With a width of 80 or larger: time run, estimated time remaining, percent complete are all displayed.
anneal	Optional vector giving the deterministic annealing schedule.
maxit	Vector (of size 2) giving maximum iteration number for the iterated M-steps and EM algorithm, respectively. A warning is displayed if either of these maximums are met.

Details

Model specification (via the models argument) follows either the nomenclature discussed in Andrews and McNicholas (2012), or via the nomenclature popularized in other packages. In both cases, the nomenclature refers to the decomposition and constraints on the covariance matrix:

$$\Sigma_g = \lambda_g D_g A_g D_g'$$

The nomenclature from Andrews and McNicholas (2012) gives four letters, each letter referring to (in order) λ, D, A, and the degrees of freedom. Possible letters are "U" for unconstrained, "C" for constrained (across groups), and "I" for when the parameter is replaced by the appropriately sized identity matrix (where applicable). As an example, the string "UICC" would refer to the model where $\Sigma_g = \lambda_g A$ with degrees of freedom held equal across groups.

The alternative nomenclature describes (in order) the volume (λ), shape (A), orientation (D), and degrees of freedom in terms of "V"ariable, "E"qual, or the "I"dentity matrix. The example model discussed in the previous paragraph would then be called by "VEIE".

Possible univariate models are c("univUU", "univUC", "univCU", "univCC") where the first capital letter describes "U"nconstrained or "C"onstrained variance and the second capital letter refers to the degrees of freedom. Once again, "V"ariable or "E"qual can replace U and C, but this time the orders match between the nomenclatures.

As many models as desired can be selected and ran via the vector supplied to models. More commonly, subsets can be called by the following character strings: "all" runs all 28 tEIGEN models (default), "dfunconstrained" runs the 14 unconstrained degrees of freedom models, "dfconstrained" runs the 14 constrained degrees of freedom models, "mclust" runs the 10 MCLUST models using the multivariate Gaussian distribution rather than the multivariate t, "gaussian" is similar but includes four further mixture models than MCLUST, "univariate" runs the univariate models - will automatically be called if one of the previous shortcuts is used on univariate data.

Note that adding "alt" to the beginning of those previously mentioned characters strings will run the same models, but return results with the V-E-I nomenclature.

Also note that for G=1, several models are equivalent (for example, UUUU and CCCC). Thus, for G=1 only one model from each set of equivalent models will be run.

Value

x	Data used for clustering/classification.
index	Indexing vector giving observations taken to be known (only available when clas is set greater than 0 or training is given).
classification	Vector of group classifications as determined by the BIC.
bic	BIC of the best fitted model.
modelname	Name of the best model according to the BIC.
allbic	Matrix of BIC values according to model and G. A value of -Inf is returned when the model did not converge.
bestmodel	Character string giving best model (BIC) details.
G	Value corresponding to the number of components chosen by the BIC.
tab	Classification table for BIC model (only available when known is given). When classification is used the "known" observations are left out of the table.
fuzzy	The fuzzy clustering matrix for the model selected by the BIC.
logl	The log-likelihood corresponding to the model with the best BIC.
parameters	List containing the fitted parameters: mean - matrix of means where the rows correspond to the component and the columns are the variables; sigma - array of covariance matrices (multivariate) or variances (univariate); lambda - vector of scale parameters, or constants of proportionality; d - eigenvectors, or orientation matrices; a - diagonal matrix proportional to eigenvalues, or shape matrices; df - vector containing the degrees of freedom for each component; weights - matrix of the expected value of the characteristic weights (used as an estimation of 'outlierness' by plot.teigen).
iclresults	List containing all the previous outputs, except x and index, pertaining to the model chosen by the best ICL (all under the same name except allicl and icl are the equivalent of allbic and bic, respectively).
info	List containing a few of the original user inputs, for use by other dedicated functions of the teigen class.

Author(s)

Jeffrey L. Andrews, Paul D. McNicholas

References

Andrews JL and McNicholas PD. "Model-based clustering, classification, and discriminant analysis with the multivariate *t*-distribution: The *t*EIGEN family" *Statistics and Computing* 22(5), 1021–1029.

Andrews JL, McNicholas PD, and Subedi S (2011) "Model-based classification via mixtures of multivariate t-distributions" *Computational Statistics and Data Analysis* 55, 520–529.

See Also

See package manual tEIGEN

Examples

```
###Note that only one model is run for each example
###in order to reduce computation time

#Clustering old faithful data with hard random start
tfaith <- teigen(faithful, models="UUUU", Gs=1:3, init="hard")
plot(tfaith)
summary(tfaith)

#Clustering old faithful with hierarchical starting values
initial_list <- list()
clustree <- hclust(dist(faithful))
for(i in 1:3){
initial_list[[i]] <- cutree(clustree,i)
}
tfaith <- teigen(faithful, models="CUCU", Gs=1:3, init=initial_list)
print(tfaith)

#Classification with the iris data set via percentage of data taken to have known membership
tiris <- teigen(iris[,-5], models="CUUU", init="uniform",clas=50,known=iris[,5])
tiris$tab

#Classification with the iris data set via training set
irisknown <- iris[,5]
#Introducing NAs is not required; this is to illustrate a `true' classification scenario
irisknown[134:150] <- NA
triris <- teigen(iris[,-5], models="CUUU", init="uniform", known=irisknown, training=1:133)
```

`teigen.parallel` *teigen.parallel: Parallelized implementation of the teigen function*

Description

Fits multivariate t-distribution mixture models under a clustering paradigm (default) or classification paradigm.

Usage

```
teigen.parallel(x, Gs = 1:9, numcores = NULL, models = "all", init = "kmeans",
 scale = TRUE, dfstart = 50, clas = 0, known= NULL, training = NULL,
 gauss = FALSE, dfupdate = "approx", eps = c(0.001, 0.1), anneal=NULL,
 maxit=c(20,1000) )
```

Arguments

x	A numeric matrix, data frame, or vector (for univariate data) .
Gs	A number or vector indicating the number of groups to fit. Default is 1-9.
numcores	Scalar argument giving how many cores for the function to utilize. If NULL (default), then the function discerns the number of cores available and uses all of them.
models	A character vector giving the models to fit. See teigen details for a comprehensive list of choices.
init	A list of initializing classification. See teigen for details .
scale	Logical indicating whether or not the function should scale the data. Default is TRUE and is the prescribed method — tEIGEN models are not scale invariant.
dfstart	The initialized value for the degrees of freedom. The default is 50.
clas	Value between 0-100 indicating the percentage of data taken to be known. Note that a vector of known classifications is needed. See next argument for an alternative. Default is 0 and performs clustering, otherwise the algorithm chooses the training index randomly (and will return it via index).
training	Optional indexing vector for the observations whose classification is taken to be known.
known	A vector of known classifications that can be numeric or character - optional for clustering, necessary for classification. Must be the same length as the number of rows in the data set. If using in a true classification sense, give samples with unknown classification the value NA within known (see training example below).
gauss	Logical indicating if the algorithm should use the gaussian distribution. If models="mclust" or "gaussian" then gauss=TRUE is forced.
dfupdate	Character string or logical indicating how the degrees of freedom should be estimated. The default is "approx" indicating a closed form approximation be used. Alternatively, "numeric" can be specified which makes use of uniroot. If FALSE, the value from dfstart is used and the degrees of freedom are not updated. If TRUE, "numeric" will be used for back-compatibility.
eps	Vector (of size 2) giving tolerance values for the convergence criterion. First value is the tolerance level for iterated M-steps. Second value is tolerance for the EM algorithm: convergence is based on Aitken's acceleration, see cited papers for more information.
anneal	Optional vector giving the deterministic annealing schedule.
maxit	Vector (of size 2) giving maximum iteration number for the iterated M-steps and EM algorithm, respectively. A warning is displayed if either of these maximums are met.

Details

This function is a parallelized wrapper of the teigen function. Please refer to details of that function.

Value

An object of class teigen.

Author(s)

Jeffrey L. Andrews

See Also

teigen

Examples

```
###Note: numcores set to 2 in order to comply
###with CRAN submission policies (set to higher
###number or NULL to automatically use all available cores)

#Clustering old faithful data with tEIGEN
tfaith <- teigen.parallel(faithful, models="UUUU",
numcores=2, Gs=1:3, init="hard")
plot(tfaith)

#Classification with the iris data set via percentage of
#data taken to have known membership
tiris <- teigen.parallel(iris[,-5], numcores=2, models="CUUU",
init="uniform",clas=50,known=iris[,5])
tiris$tab
```

Index