Package 'pgmm'

December 4, 2019

Type Package

Title Parsimonious Gaussian Mixture Models

Version 1.2.4

Date 2019-12-03

Author Paul D. McNicholas [aut, cre],
Aisha ElSherbiny [aut],
K. Raju Jampani [ctb],
Aaron F. McDaid [aut],
T. Brendan Murphy [aut],
Larry Banks [ctb]

Maintainer Paul D. McNicholas <mcnicholas@math.mcmaster.ca>

Description Carries out model-based clustering or classification using parsimonious Gaussian mixture models. McNicholas and Murphy (2008) <doi:10.1007/s11222-008-9056-0>, McNicholas (2010) <doi:10.1016/j.jspi.2009.11.006>, McNicholas and Murphy (2010) <doi:10.1093/bioinformatics/btq498>, McNicholas et al. (2010) <doi:10.1016/j.csda.2009.02.011>.

License GPL (>= 2)

LazyLoad yes

Repository CRAN

NeedsCompilation yes

Date/Publication 2019-12-04 16:10:02 UTC

R topics documented:

coffee *Coffee*

Description

Data on the chemical composition of coffee samples collected from around the world, comprising 43 samples from 29 countries. Each sample is either of the Arabica or Robusta variety. Twelve of the thirteen chemical constituents reported in the study are given. The omitted variable is total chlorogenic acid; it is generally the sum of the chlorogenic, neochlorogenic and isochlorogenic acid values.

Usage

```
data(coffee)
```

Format

A data frame with 43 observations and 14 columns. The first two columns contain `Variety` and `Country`, respectively, while the remaining 12 columns contain the chemical properties. The `Variety` is either (1) Arabica or (2) Robusta.

Note

The German to English translations of the variable names were carried out by Dr. Sharon M. McNicholas.

Source

Streuli, H. (1973). Der heutige stand der kaffeechemie. In *Association Scientifique International du Cafe, 6th International Colloquium on Coffee Chemisrty*, Bogata, Columbia, pp. 61–72.

olive *Italian Olive Oil*

Description

Data on the percentage composition of eight fatty acids found by lipid fraction of 572 Italian olive oils. The data come from three regions: Southern Italy, Sardinia, and Northern Italy. Within each region there are a number of different areas. Southern Italy comprises North Apulia, Calabria, South Apulia, and Sicily. Sardinia is divided into Inland Sardinia and Costal Sardinia. Northern Italy comprises Umbria, East Liguria, and West Liguria.

Usage

```
data(olive)
```

Format

A data frame with 572 observations and 10 columns. The first column gives the region: (1) Southern Italy, (2) Sardinia, or (3) Northern Italy. The second column gives the area: (1) North Apulia, (2) Calabria, (3) South Apulia, (4) Sicily, (5) Inland Sardinia, (6) Costal Sardinia, (7) East Liguria, (8) West Liguria, and (9) Umbria. The other eight columns contain the variables.

Source

These data are available within the GGobi software (Swayne et al., 2006).

References

Forina, M., Armanino, C., Lanteri, S. and Tiscornia, E. (1983). Classification of olive oils from their fatty acid composition. In *Food Research and Data Analysis*, pp. 189–214. Applied Science Publishers, London.

Forina, M. and Tiscornia, E. (1982). Pattern recognition methods in the prediction of Italian olive oil origin by their fatty acid content. *Annali di Chimica* **72**, 143–155.

Swayne, D.F., Cook, D., Buja, A., Lang, D.T., Wickham, H. and Lawrence, M. (2006). *GGobi Manual*. Sourced from www.ggobi.org/docs/manual.pdf.

---

pgmm *Parsimonious Gaussian Mixture Models*

Description

An implementation of model-based clustering and model-based classification using parsimonious Gaussian mixture models.

Details

Package:	pgmm
Type:	Package
Version:	1.2.4
Date:	2019-12-03
License:	GPL (>=2)

This package contains the function `pgmmEM` plus three data sets.

Author(s)

Paul D. McNicholas [aut, cre], Aisha ElSherbiny [aut], K. Raju Jampani [ctb], Aaron McDaid [aut], Brendan Murphy [aut], Larry Banks [ctb].

Maintainer: Paul D. McNicholas <mcnicholas@math.mcmaster.ca>

See Also

Details, examples, and references are given under pgmmEM.

---

pgmmEM *Model-Based Clustering & Classification Using PGMMs*

Description

Carries out model-based clustering or classification using parsimonious Gaussian mixture models. AECM algorithms are used for parameter estimation. The BIC or the ICL is used for model selection.

Usage

```
pgmmEM(x,rG=1:2,rq=1:2,class=NULL,icl=FALSE,zstart=2,cccStart=TRUE,loop=3,zlist=NULL,
modelSubset=NULL,seed=123456,tol=0.1,relax=FALSE)
```

Arguments

x	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
rG	The range of values for the number of components.
rq	The range of values for the number of factors.
class	If NULL then model-based clustering is performed. If a vector with length equal to the number of observations, then model-based classification is performed. In this latter case, the ith entry of class is either zero, indicating that the component membership of observation i is unknown, or it corresponds to the component membership of observation i. See Examples below.
icl	If TRUE then the ICL is used for model selection. Otherwise, the BIC is used for model selection.
zstart	A number that controls what starting values are used: (1) Random; (2) k-means; or (3) user-specified via zlist.
cccStart	If TRUE then random starting values are put through the CCC model and the resulting group memberships are used as starting values for the models specified in modelSubset. Only relevant for zstart=1. See Examples.
loop	A number specifying how many different random starts should be used. Only relevant for zstart=1.
zlist	A list comprising vectors of initial classifications such that zlist[[g]] gives the g-component starting values. Only relevant for zstart=3. See Examples.
modelSubset	A vector of strings giving the models to be used.
seed	A number giving the pseudo-random number seed to be used.

`tol` A number specifying the epsilon value for the convergence criteria used in the AECM algorithms. For each algorithm, the criterion is based on the difference between the log-likelihood at an iteration and an asymptotic estimate of the log-likelihood at that iteration. This asymptotic estimate is based on the Aitken acceleration and details are given in the References. Values of `tol` greater than the default are not accepted.

`relax` By default, the number of factors q must respect (p-q)^2 > p+q, where p is the number of variables (see Lawley & Maxwell, 1962). This is based on the values of q that will give data reduction in the factor analysis model or the mixture of factor analyzers model, i.e., model UUU. The same restriction applies for model CUU. However, for the other PGMM models, the restriction is a little different. The default `relax=FALSE` applies the constraint (p-q)^2 > p+q to the (maximum) value of q for all models. Setting `relax=TRUE` relaxes this constraint and allows q to take larger values; however, this option is not recommended for non-experts.

Details

The data `x` are either clustered using the PGMM approach of McNicholas & Murphy (2005, 2008, 2010) or classified using the method described by McNicholas (2010). In either case, all 12 covariance structures given by McNicholas & Murphy (2010) are available. Parameter estimation is carried out using AECM algorithms, as described in McNicholas et al. (2010). Either the BIC or the ICL is used for model-selection. The number of AECM algorithms to be run depends on the range of values for the number of components `rG`, the range of values for the number of factors `rq`, and the number of models in `modelSubset`. Starting values are very important to the successful operation of these algorithms and so care must be taken in the interpretation of results.

Value

An object of class `pgmm` is a list with components:

`map` A vector of integers, taking values in the range `rG`, indicating the maximum *a posteriori* classifications for the best model.

`model` A string giving the name of the best model.

`g` The number of components for the best model.

`q` The number of factors for the best model.

`zhat` A matrix giving the raw values upon which `map` is based.

`load` The factor loadings matrix (Lambda) for the best model.

`noisev` The Psi matrix for the best model.

`plot_info` A list that stores information to enable `plot`.

`summ_info` A list that stores information to enable `summary`.

In addition, the object will contain one of the following, depending on the value of `icl`.

`bic` A number giving the BIC for each model.

`icl` A number giving the ICL for each model.

Note

Dedicated `print`, `plot`, and `summary` functions are available for objects of class `pgmm`.

Author(s)

Paul D. McNicholas [aut, cre], Aisha ElSherbiny [aut], K. Raju Jampani [ctb], Aaron McDaid [aut], Brendan Murphy [aut], Larry Banks [ctb]

Maintainer: Paul D. McNicholas <mcnicholas@math.mcmaster.ca>

References

D. N. Lawley and A. E. Maxwell (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society: Series D* **12**(3), 209-229.

Paul D. McNicholas and T. Brendan Murphy (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics* **26**(21), 2705-2712.

Paul D. McNicholas (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference* **140**(5), 1175-1181.

Paul D. McNicholas, T. Brendan Murphy, Aaron F. McDaid and Dermot Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis* **54**(3), 711-723.

Paul D. McNicholas and T. Brendan Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing* **18**(3), 285-296.

Paul D. McNicholas and T. Brendan Murphy (2005). Parsimonious Gaussian mixture models. Technical Report 05/11, Department of Statistics, Trinity College Dublin.

Examples

```
  ## Not run:
# Wine clustering example with three random starts and the CUU model.
 data("wine")
 x<-wine[,-1]
 x<-scale(x)
 wine_clust<-pgmmEM(x,rG=1:4,rq=1:4,zstart=1,loop=3,modelSubset=c("CUU"))
 table(wine[,1],wine_clust$map)

# Wine clustering example with custom starts and the CUU model.
 data("wine")
 x<-wine[,-1]
 x<-scale(x)
 hcl<-hclust(dist(x))
 z<-list()
 for(g in 1:4){
 z[[g]]<-cutree(hcl,k=g)
 }
 wine_clust2<-pgmmEM(x,1:4,1:4,zstart=3,modelSubset=c("CUU"),zlist=z)
 table(wine[,1],wine_clust2$map)
 print(wine_clust2)
 summary(wine_clust2)
```

```
# Olive oil classification by region (there are three regions), with two-thirds of
# the observations taken as having known group memberships, using the CUC, CUU and
# UCU models.
 data("olive")
 x<-olive[,-c(1,2)]
 x<-scale(x)
 cls<-olive[,1]
 for(i in 1:dim(olive)[1]){
 if(i%%3==0){cls[i]<-0}
 }
 olive_class<-pgmmEM(x,rG=3:3,rq=4:6,cls,modelSubset=c("CUC","CUU",
  "CUCU"),relax=TRUE)
 cls_ind<-(cls==0)
 table(olive[cls_ind,1],olive_class$map[cls_ind])

# Another olive oil classification by region, but this time suppose we only know
# two-thirds of the labels for the first two areas but we suspect that there might
# be a third or even a fourth area.
 data("olive")
 x<-olive[,-c(1,2)]
 x<-scale(x)
 cls2<-olive[,1]
 for(i in 1:dim(olive)[1]){
   if(i%%3==0||i>420){cls2[i]<-0}
 }
 olive_class2<-pgmmEM(x,2:4,4:6,cls2,modelSubset=c("CUU"),relax=TRUE)
 cls_ind2<-(cls2==0)
 table(olive[cls_ind2,1],olive_class2$map[cls_ind2])

# Coffee clustering example using k-means starting values for all 12
# models with the ICL being used for model selection instead of the BIC.
 data("coffee")
 x<-coffee[,-c(1,2)]
 x<-scale(x)
 coffee_clust<-pgmmEM(x,rG=2:3,rq=1:3,zstart=2,icl=TRUE)
 table(coffee[,1],coffee_clust$map)
 plot(coffee_clust)
 plot(coffee_clust,onlyAll=TRUE)

## End(Not run)

# Coffee clustering example using k-means starting values for the UUU model, i.e., the
# mixture of factor analyzers model, for G=2 and q=1.
 data("coffee")
 x<-coffee[,-c(1,2)]
 x<-scale(x)
 coffee_clust_mfa<-pgmmEM(x,2:2,1:1,zstart=2,modelSubset=c("UUU"))
 table(coffee[,1],coffee_clust_mfa$map)
```

wine *Italian Wine*

Description

Data on 27 chemical and physical properties of three types of wine (Barolo, Grignolino, Barbera) from the Piedmont region of Italy. The study did include one further variable but the sulphur measurements were not available.

Usage

```
data(wine)
```

Format

A data frame with 178 observations and 28 columns. The first column gives the type of wine: (1) Barolo, (2) Grignolino, or (3) Barbera. The other 27 columns contain the variables.

Source

Forina, M., Armanino, C., Castino, M. and Ubigli, M. (1986). Multivariate data analysis as a discriminating method of the origin of wines. *Vitis* **25**, 189–201.

Index