Package 'clustMD'

October 12, 2022

Title Model Based Clustering for Mixed Data

Version 1.2.1

Description Model-based clustering of mixed data (i.e. data which consist of continuous, binary, ordinal or nominal variables) using a parsimonious mixture of latent Gaussian variable models.

Depends R (>= 3.3.2)

Imports ggplot2, mclust, reshape2, MASS, msm, mvtnorm, parallel, truncnorm, viridis, stats

License GPL-2

LazyData true

RoxygenNote 6.0.1

NeedsCompilation no

Author Damien McParland [aut, cre], Isobel Claire Gormley [aut]

Maintainer Damien McParland <damien.mcparland@ucd.ie>

Repository CRAN

Date/Publication 2017-05-08 17:19:20 UTC

R topics documented:

`clustMD-package` *Model based clustering for mixed data: clustMD*

Description

Model-based clustering of mixed data (i.e. data that consist of continuous, binary, ordinal or nominal variables) using a parsimonious mixture of latent Gaussian variable models.

Author(s)

Damien McParland

Damien McParland <damien.mcparland@ucd.ie> Isobel Claire Gormley <claire.gormley@ucd.ie>

References

McParland, D. and Gormley, I.C. (2016). Model based clustering for mixed data: clustMD. Advances in Data Analysis and Classification, 10 (2):155-169.

See Also

`clustMD`

`Byar` *Byar prostate cancer data set.*

Description

A data set consisting of variables of mixed type measured on a group of prostate cancer patients. Patients have either stage 3 or stage 4 prostate cancer.

Usage

```
Byar
```

Format

A data frame with 475 observations on the following 15 variables.

`Age` a numeric vector indicating the age of the patient.

`Weight` a numeric vector indicating the weight of the patient.

`Performance.rating` an ordinal variable indicating how active the patient is: 0 - normal activity, 1 - in bed less than 50% of daytime, 2 - in bed more than 50% of daytime, 3 - confined to bed.

`Cardiovascular.disease.history` a binary variable indicating if the patient has a history of cardiovascular disease: 0 - no, 1 - yes.

`Systolic.Blood.pressure` a numeric vector indicating the systolic blood pressure of the patient in units of ten.

`Diastolic.blood.pressure` a numeric vector indicating the diastolic blood pressure of the patient in units of ten.

`Electrocardiogram.code` a nominal variable indicating the electorcardiogram code: 0 - normal, 1 - benign, 2 - rythmic disturbances and electrolyte changes, 3 - heart blocks or conduction defects, 4 - heart strain, 5 - old myocardial infarct, 6 - recent myocardial infarct.

`Serum.haemoglobin` a numeric vector indicating the serum haemoglobin levels of the patient measured in g/100ml.

`Size.of.primary.tumour` a numeric vector indicating the estimated size of the patient's primary tumour in centimeters squared.

`Index.of.tumour.stage.and.histolic.grade` a numeric vector indicating the combined index of tumour stage and histolic grade of the patient.

`Serum.prostatic.acid.phosphatase` a numeric vector indicating the serum prostatic acid phosphatase levels of the patient in King-Armstong units.

`Bone.metastases` a binary vector indicating the presence of bone metastasis: 0 - no, 1 - yes.

`Stage` the stage of the patient's prostate cancer.

`Observation` a patient ID number.

`SurvStat` the post trial survival status of the patient: 0 - alive, 1 - dead from prostatic cancer, 2 - dead from heart or vascular disease, 3 - dead from cerebrovascular accident, 3 - dead form pulmonary ebolus, 5 - dead from other cancer, 6 - dead from respiratory disease, 7 - dead from other specific non-cancer cause, 8 - dead from other unspecified non-cancer cause, 9 - dead from unknown cause.

Source

Byar, D.P. and Green, S.B. (1980). The choice of treatment for cancer patients based on covariate information: applications to prostate cancer. Bulletin du Cancer 67: 477-490.

Hunt, L., Jorgensen, M. (1999). Mixture model clustering using the multimix program. Australia and New Zealand Journal of Statistics 41: 153-171.

`clustMD` *Model Based Clustering for Mixed Data*

Description

A function that fits the clustMD model to a data set consisting of any combination of continuous, binary, ordinal and nominal variables.

Usage

```
clustMD(X, G, CnsIndx, OrdIndx, Nnorms, MaxIter, model, store.params = FALSE,
  scale = FALSE, startCL = "hc_mclust", autoStop = FALSE, ma.band = 50,
  stop.tol = NA)
```

Arguments

X	a data matrix where the variables are ordered so that the continuous variables come first, the binary (coded 1 and 2) and ordinal variables (coded 1, 2, ...) come second and the nominal variables (coded 1, 2, ...) are in last position.
G	the number of mixture components to be fitted.
CnsIndx	the number of continuous variables in the data set.
OrdIndx	the sum of the number of continuous, binary and ordinal variables in the data set.
Nnorms	the number of Monte Carlo samples to be used for the intractable E-step in the presence of nominal data. Irrelevant if there are no nominal variables.
MaxIter	the maximum number of iterations for which the (MC)EM algorithm should run.
model	a string indicating which clustMD model is to be fitted. This may be one of: EII, VII, EEI, VEI, EVI, VVI or BD.
store.params	a logical argument indicating if the parameter estimates at each iteration should be saved and returned by the clustMD function.
scale	a logical argument indicating if the continuous variables should be standardised.
startCL	a string indicating which clustering method should be used to initialise the (MC)EM algorithm. This may be one of "kmeans" (K means clustering), "hclust" (hierarchical clustering), "mclust" (finite mixture of Gaussian distributions), "hc_mclust" (model-based hierarchical clustering) or "random" (random cluster allocation).
autoStop	a logical argument indicating whether the (MC)EM algorithm should use a stopping criterion to decide if convergence has been reached. Otherwise the algorithm will run for MaxIter iterations. If only continuous variables are present the algorithm will use Aitken's acceleration criterion with tolerance stop.tol. If categorical variables are present, the stopping criterion is based on a moving average of the approximated log likelihood values. Let t denote the current interation. The average of the ma.band most recent approximated log likelihood values is compared to the average of another ma.band iterations with a lag of 10 iterations. If this difference is less than the tolerance the algorithm will be said to have converged.
ma.band	the number of iterations to be included in the moving average calculation for the stopping criterion.
stop.tol	the tolerance of the (MC)EM stopping criterion.

Value

An object of class clustMD is returned. The output components are as follows:

model	The covariance model fitted to the data.
G	The number of clusters fitted to the data.
Y	The observed data matrix.
cl	The cluster to which each observation belongs.

tau	A N x G matrix of the probabilities of each observation blonging to each cluster.
means	A D x G matrix of the cluster means. Where D is the dimension of the combined observed and latent continuous space.
A	A G x D matrix containing the diagonal entries of the A matrix corresponding to each cluster.
Lambda	A G x D matrix of volume parameters corresponding to each observed or latent dimension for each cluster.
Sigma	A D x D x G array of the covariance matrices for each cluster.
BIChat	The estimated Bayesian information criterion for the model fitted.
ICLhat	The estimated integrated classification likelihood criterion for the model fitted.
paramlist	If store.params is TRUE then paramlist is a list of the stored parameter values in the order given above with the saved estimated likelihood values in last position.
Varnames	A character vector of names corresponding to the columns of Y
Varnames_sht	A truncated version of Varnames. Used for plotting.
likelihood.store	A vector containing the estimated log likelihood at each iteration.

References

McParland, D. and Gormley, I.C. (2016). Model based clustering for mixed data: clustMD. Advances in Data Analysis and Classification, 10 (2):155-169.

Examples

```
data(Byar)
    # Transformation skewed variables
    Byar$Size.of.primary.tumour <- sqrt(Byar$Size.of.primary.tumour)
    Byar$Serum.prostatic.acid.phosphatase <- log(Byar$Serum.prostatic.acid.phosphatase)

    # Order variables (Continuous, ordinal, nominal)
    Y <- as.matrix(Byar[, c(1, 2, 5, 6, 8, 9, 10, 11, 3, 4, 12, 7)])

    # Start categorical variables at 1 rather than 0
    Y[, 9:12] <- Y[, 9:12] + 1

    # Standardise continuous variables
    Y[, 1:8] <- scale(Y[, 1:8])

    # Merge categories of EKG variable for efficiency
    Yekg <- rep(NA, nrow(Y))
    Yekg[Y[,12]==1] <- 1
    Yekg[(Y[,12]==2)|(Y[,12]==3)|(Y[,12]==4)] <- 2
    Yekg[(Y[,12]==5)|(Y[,12]==6)|(Y[,12]==7)] <- 3
    Y[, 12] <- Yekg

    ## Not run:
    res <- clustMD(X = Y, G = 3, CnsIndx = 8, OrdIndx = 11, Nnorms = 20000,
    MaxIter = 500, model = "EVI", store.params = FALSE, scale = TRUE,
```

```
    startCL = "kmeans", autoStop= TRUE, ma.band=30, stop.tol=0.0001)

## End(Not run)
```

---

clustMDlist *Model Based Clustering for Mixed Data*

Description

A function that fits the clustMD model to a data set consisting of any combination of continuous, binary, ordinal and nominal variables. This function is a wrapper for clustMD that takes arguments as a list.

Usage

```
clustMDlist(arglist)
```

Arguments

arglist	a list of input arguments for clustMD. See clustMD.

Value

A clustMD object. See clustMD.

References

McParland, D. and Gormley, I.C. (2016). Model based clustering for mixed data: clustMD. Advances in Data Analysis and Classification, 10 (2):155-169.

See Also

clustMD

Examples

```
data(Byar)

# Transformation skewed variables
Byar$Size.of.primary.tumour <- sqrt(Byar$Size.of.primary.tumour)
Byar$Serum.prostatic.acid.phosphatase <-
log(Byar$Serum.prostatic.acid.phosphatase)

# Order variables (Continuous, ordinal, nominal)
Y <- as.matrix(Byar[, c(1, 2, 5, 6, 8, 9, 10, 11, 3, 4, 12, 7)])

# Start categorical variables at 1 rather than 0
Y[, 9:12] <- Y[, 9:12] + 1
```

```
    # Standardise continuous variables
    Y[, 1:8] <- scale(Y[, 1:8])

    # Merge categories of EKG variable for efficiency
    Yekg <- rep(NA, nrow(Y))
    Yekg[Y[,12]==1] <- 1
    Yekg[(Y[,12]==2)|(Y[,12]==3)|(Y[,12]==4)] <- 2
    Yekg[(Y[,12]==5)|(Y[,12]==6)|(Y[,12]==7)] <- 3
    Y[, 12] <- Yekg

    argList <- list(X=Y, G=3, CnsIndx=8, OrdIndx=11, Nnorms=20000,
    MaxIter=500, model="EVI", store.params=FALSE, scale=TRUE,
    startCL="kmeans", autoStop=FALSE, ma.band=50, stop.tol=NA)

    ## Not run:
    res <- clustMDlist(argList)

## End(Not run)
```

`clustMDparallel` *Run multiple clustMD models in parallel*

Description

This function allows the user to run multiple clustMD models in parallel. The inputs are similar to `clustMD()` except `G` is now a vector containing the the numbers of components the user would like to fit and `models` is a vector of strings indicating the covariance models the user would like to fit for each element of G. The user can specify the number of cores to be used or let the function detect the number available.

Usage

```
clustMDparallel(X, CnsIndx, OrdIndx, G, models, Nnorms, MaxIter, store.params,
  scale, startCL = "hc_mclust", Ncores = NULL, autoStop = FALSE,
  ma.band = 50, stop.tol = NA)
```

Arguments

`X`	a data matrix where the variables are ordered so that the continuous variables come first, the binary (coded 1 and 2) and ordinal variables (coded 1, 2,...) come second and the nominal variables (coded 1, 2,...) are in last position.
`CnsIndx`	the number of continuous variables in the data set.
`OrdIndx`	the sum of the number of continuous, binary and ordinal variables in the data set.
`G`	a vector containing the numbers of mixture components to be fitted.
`models`	a vector of strings indicating which clustMD models are to be fitted. This may be one of: `EII`, `VII`, `EEI`, `VEI`, `EVI`, `VVI` or `BD`.

Nnorms	the number of Monte Carlo samples to be used for the intractable E-step in the presence of nominal data.
MaxIter	the maximum number of iterations for which the (MC)EM algorithm should run.
store.params	a logical variable indicating if the parameter estimates at each iteration should be saved and returned by the clustMD function.
scale	a logical variable indicating if the continuous variables should be standardised.
startCL	a string indicating which clustering method should be used to initialise the (MC)EM algorithm. This may be one of "kmeans" (K means clustering), "hclust" (hierarchical clustering), "mclust" (finite mixture of Gaussian distributions), "hc_mclust" (model-based hierarchical clustering) or "random" (random cluster allocation).
Ncores	the number of cores the user would like to use. Must be less than or equal to the number of cores available.
autoStop	a logical argument indicating whether the (MC)EM algorithm should use a stopping criterion to decide if convergence has been reached. Otherwise the algorithm will run for MaxIter iterations. If only continuous variables are present the algorithm will use Aitken's acceleration criterion with tolerance stop.tol. If categorical variables are present, the stopping criterion is based on a moving average of the approximated log likelihood values. let t denote the current interation. The average of the ma.band most recent approximated log likelihood values is compared to the average of another ma.band iterations with a lag of 10 iterations. If this difference is less than the tolerance the algorithm will be said to have converged.
ma.band	the number of iterations to be included in the moving average stopping criterion.
stop.tol	the tolerance of the (MC)EM stopping criterion.

Value

An object of class clustMDparallel is returned. The output components are as follows:

BICarray	A matrix indicating the estimated BIC values for each of the models fitted.
results	A list containing the output for each of the models fitted. Each entry of this list is a clustMD object. If the algorithm failed to fit a particular model, the corresponding entry of results will be NULL.

References

McParland, D. and Gormley, I.C. (2016). Model based clustering for mixed data: clustMD. Advances in Data Analysis and Classification, 10 (2):155-169.

See Also

clustMD

Examples

```
data(Byar)

# Transformation skewed variables
Byar$Size.of.primary.tumour <- sqrt(Byar$Size.of.primary.tumour)
Byar$Serum.prostatic.acid.phosphatase <-
log(Byar$Serum.prostatic.acid.phosphatase)

# Order variables (Continuous, ordinal, nominal)
Y <- as.matrix(Byar[, c(1, 2, 5, 6, 8, 9, 10, 11, 3, 4, 12, 7)])

# Start categorical variables at 1 rather than 0
Y[, 9:12] <- Y[, 9:12] + 1

# Standardise continuous variables
Y[, 1:8] <- scale(Y[, 1:8])

# Merge categories of EKG variable for efficiency
Yekg <- rep(NA, nrow(Y))
Yekg[Y[,12]==1] <- 1
Yekg[(Y[,12]==2)|(Y[,12]==3)|(Y[,12]==4)] <- 2
Yekg[(Y[,12]==5)|(Y[,12]==6)|(Y[,12]==7)] <- 3
Y[, 12] <- Yekg

## Not run:
res <- clustMDparallel(X = Y, G = 1:3, CnsIndx = 8, OrdIndx = 11, Nnorms = 20000,
MaxIter = 500, models = c("EVI", "EII", "VII"), store.params = FALSE, scale = TRUE,
startCL = "kmeans", autoStop= TRUE, ma.band=30, stop.tol=0.0001)

res$BICarray

## End(Not run)
```

`getOutput_clustMDparallel` *Extracts relevant output from* `clustMDparallel` *object*

Description

This function takes a `clustMDparallel` object, a number of clusters and a covariance model as inputs. It then returns the output corresponding to that model. If the particular model is not contained in the `clustMDparallel` object then the function returns an error.

Usage

```
getOutput_clustMDparallel(resParallel, nClus, covModel)
```

Arguments

resParallel	a clustMDparallel object.
nClus	the number of clusters in the desired output.
covModel	the covariance model of the desired output.

Value

A clustMD object containing the output for the relevant model.

---

plot.clustMD *Plotting method for objects of class* clustMD

Description

Plots a parallel coordinates plot and dot plot of the estimated cluster means, a barplot of the variances by cluster for diagonal covariance models or a heatmap of the covariance matrix for non-diagonal covariance structures, and a histogram of the clustering uncertainties for each observation.

Usage

```
## S3 method for class 'clustMD'
plot(x, ...)
```

Arguments

x	a clustMD object.
...	further arguments passed to or from other methods.

Value

Prints graphical summaries of the fitted model as detailed above.

References

McParland, D. and Gormley, I.C. (2016). Model based clustering for mixed data: clustMD. Advances in Data Analysis and Classification, 10 (2):155-169.

See Also

clustMD

plot.clustMDparallel *Summary plots for a clustMDparallel object*

Description

Produces a line plot of the estimated BIC values corresponding to each covariance model against the number of clusters fitted. For the optimal model according to this criteria, a parallel coordinates plot of the cluster means is produced along with a barchart or heatmap of the covariance matrices for each cluster and a histogram of the clustering uncertainties.

Usage

```
## S3 method for class 'clustMDparallel'
plot(x, ...)
```

Arguments

x	a clustMDparallel object.
...	further arguments passed to or from other methods.

Value

Produces a number of plots as detailed above.

print.clustMD *Print basic details of* clustMD *object.*

Description

Prints a short summary of a clustMD object to screen. Details the number of clusters fitted as well as the covariance model and the estimated BIC.

Usage

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

Arguments

x	a clustMD object.
...	further arguments passed to or from other methods.

Value

Prints summary details, as described above, to screen.

See Also

clustMD

`print.clustMDparallel` *Print basic details of* `clustMDparallel` *object*

Description

Prints basic details of `clustMDparallel` object. Outputs the different numbers of clusters and the different covariance structures fitted to the data. It also states which model was optimal according to the estimated BIC criterion.

Usage

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

Arguments

`x`	a `clustMDparallel` object.
`...`	further arguments passed to or from other methods.

Value

Prints details described above to screen.

See Also

clustMD

`summary.clustMD` *Summarise* `clustMD` *object*

Description

Prints a summary of a `clustMD` object to screen. Details the number of clusters fitted as well as the covariance model and the estimated BIC. Also prints a table detailing the number of observations in each cluster and a matrix of the cluster means.

Usage

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

Arguments

object	a clustMD object.
...	further arguments passed to or from other methods.

Value

Prints summary of clustMD object to screen, as detailed above.

See Also

clustMD

---

summary.clustMDparallel

Prints a summary of a clustMDparallel object to screen.

Description

Prints the different numbers of clusters and covariance models fitted and indicates the optimal model according to the estimated BIC criterion. The estimated BIC for the optimal model is printed to screen along with a table of the cluster membership and the matrix of cluster means for this optimal model.

Usage

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

Arguments

object	a clustMDparallel object.
...	further arguments passed to or from other methods.

Value

Prints a summary of the clustMDparallel object to screen, as detailed above.

See Also

clustMD

Index