Package 'deepgmm'

December 16, 2019

Type Package

Title Deep Gaussian Mixture Models

Version 0.1.59

Date 2019-12-16

Author Cinzia Viroli, Geoffrey J. McLachlan

Maintainer Suren Rathnayake <surenr@gmail.com>

Description Deep Gaussian mixture models as proposed by Viroli and McLachlan (2019) <doi:10.1007/s11222-017-9793-z> provide a generalization of classical Gaussian mixtures to multiple layers. Each layer contains a set of latent variables that follow a mixture of Gaussian distributions. To avoid overparameterized solutions, dimension reduction is applied at each layer by way of factor models.

URL https://github.com/suren-rathnayake/deepgmm

Imports mvtnorm, corpcor

Suggests testthat

License GPL (>= 3)

NeedsCompilation no

Repository CRAN

Date/Publication 2019-12-16 05:20:02 UTC

R topics documented:

deepgmm *Fits Deep Gaussian Mixture Models Using Stochastic EM algorithm.*

Description

Fits a deep Gaussian mixture model to multivariate data.

Usage

```
deepgmm(y, layers, k, r,
        it = 250, eps = 0.001, init = "kmeans", init_est = "factanal")
```

Arguments

y	A matrix or a data frame in which the rows correspond to observations and the columns to variables.
layers	The number of layers in the deep Gaussian mixture model. Limited to 1, 2 or 3.
k	A vector of integers of length layers containing the number of groups in the different layers.
r	A vector of integers of length layers containing the dimensions at the different layers. Dimension of the layers must be in decreasing size. See details.
it	Maximum number of EM iterations.
eps	The EM algorithm terminates if the relative increment of the log-likelihood falls below this value.
init	Procedure to obtain an initial partition of the observations. See Details.
init_est	Procedure for computing the initial parameter values for the given initial partition of the data. See Details.

Details

The deep Gaussian mixture model is an hierarchical model organized in a multilayered architecture where, at each layer, the variables follow a mixture of Gaussian distributions. This set of nested mixtures of linear models provides a globally nonlinear model that can model the data in a very flexible way. In order to avoid overparameterized solutions, dimension reduction by factor models can be applied at each layer of the architecture, thus resulting in deep mixtures of factor analyzers.

The data y must be a matrix or a data frame containing numerical values, with no missing values. The rows must correspond to observations and the columns to variables.

Presently, the maximum number of layers layers implemented is 3.

The ith element of k contain number of groups in the ith layer. Thus the length k must equal to layers.

The parameter vector r contains the latent variable dimension of each layer. Variables at different layers have progressively decreasing dimension, $r_1, r_2, \ldots, r_h$, where $p > r_1 > r_2 > \ldots > r_h \geq 1$.

The EM algorithm used by dgmm requires initialization. The initialization is done by partitioning the dataset, and then estimating the initial values for model parameters based on the partition. There

are three options available in dgmm for the initial partitioning of the data; random partitioning, clustering using the *k*-means algorithm and using the agglomerative hierarchical clustering. With the `init = "random"` the partitioning is done randomly, with `init = "kmeans"` the data are partitioned using the *k*-means algorithm of "Hartigan-Wong", while with `init = "hclass"` agglomerative hierarchical clustering is done to obtain initial partitioning. For this option the dissimilarity structure is obtained using the Euclidean distance measure.

After the initial partitioning has been chosen, initial values of the parameters in the component analyzers need to be calculated. There are two options available in `init_est`. The default option, `init_est = "factanal"` provides initial estimates of the parameters based on factor analysis. If `init_est = "ppca"` then mixtures of probabilistic principal component analyzers are fitted within each layer to provide initial estimates of the parameters.

Value

An object of class `"dgmm"` containing fitted values. It contains

`H`	A list in which the *i*th element is the loading matrix for the *i*th layer
`w`	A list containing mixing proportions for each layer. (i.e. the element w[[i]][j] contain the mixing proportion of the *j*th component in the *i* layer.)
`mu`	A list of matrices containing components means in the columns. (i.e. the element mu[[i]][, j] contain the component mean of the *j*th component in the *i* layer.)
`psi`	A list of arrays which contain covariance matrices for the random error components of each component (i.e. the element psi[[i]][j, ,,] contain the error covariance matrix for the *j*th component in the *i* layer.)
`lik`	The log-likelihood after each EM iteration
`bic`	The Bayesian information criterion for the model fit
`acl`	The Akaike information criterion for the model fit
`clc`	The Classification likelihood information criterion for the model fit
`icl.bic`	The integrated classification criterion for the model fit
`s`	Clustering of the observations

Author(s)

Cinzia Viroli, Geoffrey J. McLachlan

References

Viroli, C. and McLachlan, G.J. (2019). Deep Gaussian mixture models. Statistics and Computing 29, 43-51.

Examples

```
layers <- 2
k <- c(3, 4)
r <- c(3, 2)
it <- 50
```

```
eps <- 0.001
y <- scale(mtcars)

set.seed(1)
fit <-deepgmm(y = y, layers = layers, k = k, r = r,
                  it = it, eps = eps)
fit

summary(fit)
```

`model_selection` *Function to compare different models*

Description

Compares different models and return the best one selected according to criterion (BIC or AIC).

Usage

```
model_selection(y, layers, g, seeds = 3, it = 50, eps = 0.001,
        init = "kmeans", init_est = "factanal", criterion = "BIC")
```

Arguments

`y`	A matrix or a data frame in which rows correspond to observations and columns to variables.
`layers`	The number of layers in the deep Gaussian mixture model. Admitted values are 1, 2 or 3.
`g`	The number of clusters.
`seeds`	Numeric vector containing seeds to try.
`it`	Maximum number of EM iterations.
`eps`	The EM algorithm terminates the relative increment of the log-likelihod falls below this value.
`init`	Initial paritioning of the observations to determine initial parameter values. See Details.
`init_est`	To determine how the initial parameter values are computed. See Details.
`criterion`	Model selection criterion, either `"AIC"` of `"BIC"`.

Details

Compares different models and return the best one selected according to criterion (BIC or AIC). One can use diffefrent number of seeds.

Value

A list containing an object of class `"dgmm"` containing fitted values and list of BIC and AIC values.

References

Viroli, C. and McLachlan, G.J. (2019). Deep Gaussian mixture models. Statistics and Computing 29, 43-51.

Examples

```
layers <- 2
k <- c(3, 4)
r <- c(3, 2)
it <- 50
eps <- 0.001
y <- scale(mtcars)

sel <- model_selection(y, layers, 3, seeds = 1, it = 250, eps = 0.001)
sel

summary(sel)
```

Index