

# Package ‘MFDA’

April 17, 2009

**Version** 1.1-1

**Date** 2007-10-30

**Title** Model Based Functional Data Analysis

**Author** Wenxuan Zhong <wenxuan@stat.harvard.edu>, Ping Ma <pingma@uiuc.edu>

**Maintainer** Wenxuan Zhong <wenxuan@stat.harvard.edu>

**Depends** R (>= 2.1.0), gss, mvtnorm

**Description** This is the package for doing model based functional clustering.

**License** GPL (>= 2)

**URL** <http://www.r-project.org>

**Repository** CRAN

**Date/Publication** 2009-02-04 09:57:30

## R topics documented:

check.start . . . . .	2
compute.center . . . . .	2
compute.M.all . . . . .	3
compute.M.pk . . . . .	4
compute.reject . . . . .	5
compute.weight . . . . .	5
em.bic . . . . .	6
em.clust . . . . .	7
Estep.tik . . . . .	8
Estep.tk . . . . .	9
MFclust . . . . .	10
MFclust.compute . . . . .	11
mkrandom . . . . .	12
testdata . . . . .	13

<b>Index</b>	<b>14</b>
--------------	-----------

---

<code>check.start</code>	<i>Check starting point</i>
--------------------------	-----------------------------

---

**Description**

Program used to check whether the starting point is a good starting point.

**Usage**

```
check.start(clust)
```

**Arguments**

<code>clust</code>	The initial starting point for the Markov chain
--------------------	---

**Value**

The logic value describe whether it is a good starting point.

**References**

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

**See Also**

...

---

<code>compute.center</code>	<i>Update parameters for cluster k in M-step for functional mixture models.</i>
-----------------------------	---

---

**Description**

Update and trace of the smoothing matrix for cluster k in the M step of EM algorithm for functional mixture models.

**Usage**

```
compute.center(x, weight)
```

**Arguments**

<code>x</code>	A matrix of observations with each row representing the observation and each column representing the experiment at each time point.
<code>weight</code>	A matrix with each element is the conditional posterior probability that functional observation belongs to cluster k.

**Value**

A list including the following components:

mu	A vector whose $k$ th row represent the center mean curve for the $k$ th cluster
zeta	A vector with each element representing the random effect $b_i$ for the $k$ th cluster.
varht	A vector with each element representing the estimated error variance.
trc	A vector with each element representing the trace of the smoothing matrix for the $k$ th cluster.

**References**

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

**See Also**

[compute.center](#)

---

compute.M.all	<i>M-step in the EM algorithm for functional mixture models.</i>
---------------	--

---

**Description**

Maximization step in the EM algorithm for functional mixture models.

**Usage**

```
compute.M.all(x, weight, my.label)
```

**Arguments**

x	A matrix, or data frame of observations, rows correspond to observations and columns correspond to variables.
weight	A matrix whose $[i, k]$ th entry is the conditional probability of the $i$ th observation belonging to the $k$ th component of the mixture.
my.label	A list whose $k$ th entry is the indicator value indicating whether the $i$ th functional observation is participant of M-step for cluster $k$ .

**Value**

A list including the following components:

mu	A matrix whose $k$ th column is the mean of the $k$ th component of the mixture model.
zeta	A numerical vector specifying the estimate of cluster precision parameters.
varht	A numerical vector specifying the estimate of cluster error variance.

mu	A matrix specifying the estimate of cluster mean profile.
zeta	A numerical vector specifying the estimate of cluster precision parameters.
trc	A numerical vector whose kth entry specifying the trace of smoothing matrix of cluster k.

## References

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

## See Also

[Estep.tk](#)

---

compute.M.pk	<i>Compute the cluster proportion.</i>
--------------	--

---

## Description

Compute cluster proportion in Gaussian mixture model.

## Usage

```
compute.M.pk(tk, nclust, alpha)
```

## Arguments

tk	An n by k matrix obtained from Estep.tk.
nclust	The number of cluster.
alpha	The parameters of a Dirichlet distribution.

## Value

A list including the following components:

p_k	A vector of updated cluster proportion.
-----	---

## References

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

---

compute.reject      *The rejection controlled EM.*

---

**Description**

Algorithm to estimate the weight in M-step.

**Usage**

```
compute.reject(tk, k, thres)
```

**Arguments**

tk	A matrix obtained from E-step.
k	The kth Gaussian component.
thres	The threshold for rejection-controlled EM.

**Value**

A list including the following components:

em.select	A vector.
sem.select	A vector.

**References**

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

---

compute.weight      *Compute the weight for the penalized likelihood in M-step.*

---

**Description**

Compute the weight for the penalized likelihood in M-step.

**Usage**

```
compute.weight(tk, thres)
```

**Arguments**

tk	A n by K matrix obtained from Estep.tk.
thres	A scaler specifying the threshold for the rejection step in RCEM.

**Value**

A list including the following components:

<code>weight</code>	A matrix of the weight.
<code>my.label</code>	A list indicating whether the <i>i</i> th functional observation is a participant in penalized likelihood estimation in M-step.

**References**

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

**See Also**

[compute.M.pk](#)

---

em.bic

*BIC for Functional Mixture Gaussian Models*

---

**Description**

Compute the BIC (Bayesian Information Criterion) for functional mixture Gaussian models given the negative loglikelihood, the dimension of the data, and the trace of the smoothing matrix.

**Usage**

```
em.bic(likelihood, trc, n, m)
```

**Arguments**

<code>likelihood</code>	The negative loglikelihood for a data set with respect to the functional mixture model.
<code>trc</code>	The trace of the smoothing matrix.
<code>n</code>	The number of the functional data use to compute <code>loglik</code> .
<code>m</code>	The number of the repeated measurements in the data used to compute <code>loglik</code> .

**Value**

The BIC or Bayesian Information Criterion for the given input arguments.

**References**

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

**See Also**

[em.clust.](#)

---

`em.clust`*Model-Based Clustering for a single Markov chain*

---

**Description**

EM algorithm for functional Gaussian mixture models.

**Usage**

```
em.clust(x, clust, mu, zeta, varht, thres, iter.max, alpha)
```

**Arguments**

<code>x</code>	A numeric matrix, or data frame of observations, rows correspond to functional observations and columns correspond to the number of repeated measurements.
<code>clust</code>	An integer vector specifying the initial clustering membership.
<code>mu</code>	A matrix specifying the initial estimate of cluster mean profile.
<code>zeta</code>	A numerical vector specifying the initial estimate of cluster precision parameters.
<code>varht</code>	A numerical vector specifying the initial estimate of cluster error variance.
<code>thres</code>	A threshold value for rejection-controlled step.
<code>iter.max</code>	An integer limit on the number of EM iterations.
<code>alpha</code>	The prior for the cluster proportions $p$ .

**Value**

A list contains estimated cluster membership, estimated cluster mean profile, Bayesian Information Criterion, negative loglikelihood.

**References**

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

---

Estep.tik	<i>Compute the conditional probability of subject <math>i</math> in cluster <math>k</math> in E-step for functional mixture models.</i>
-----------	---

---

### Description

Compute the conditional probability of subject  $i$  in cluster  $k$  in the expectation step of EM algorithm for functional mixture models.

### Usage

```
Estep.tik(xx.i, old)
```

### Arguments

<code>xx.i</code>	A numeric vector corresponds to a functional observation.
<code>old</code>	The list of output from maximization step of EM algorithm.

### Value

A list including the following components:

<code>clust</code>	A vector whose $i$ th entry is the indicator value of the $i$ th observation belonging to the $k$ th component of the mixture.
<code>t.ik</code>	A vector whose $k$ th entry is the conditional probability of the $i$ th observation belonging to the $k$ th component of the mixture.
<code>loglike</code>	The negative loglikelihood for the data in the mixture model.

### References

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

### See Also

[Estep.tk](#)

---

Estep.tk	<i>Compute the cluster proportion in E-step for functional mixture models.</i>
----------	--

---

**Description**

Compute the cluster proportion in the expectation step of EM algorithm for functional mixture models.

**Usage**

```
Estep.tk(x, old, nclust)
```

**Arguments**

<code>x</code>	A numeric matrix, or data frame of observations, rows correspond to functional observations and columns correspond to the number of repeated measurements.
<code>old</code>	The list of output from maximization step of EM algorithm.
<code>nclust</code>	The number of clusters (componenets in mixture model).

**Value**

A list including the following components:

<code>clust</code>	A vector whose $i$ th entry is the indicator value of the $i$ th observation belonging to the $k$ th component of the mixture.
<code>tk</code>	A matrix whose $[i, k]$ th entry is the conditional probability of the $i$ th observation belonging to the $k$ th component of the mixture.
<code>loglikelihood</code>	The negative loglikelihood for the data in the mixture model.

**References**

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

**See Also**

[Estep.tik](#)

MFclust

*Model-Based Functional Data Clustering***Description**

Clustering via rejection-controlled EM initialized by kmeans clustering for functional Gaussian mixture models. The number of clusters and the clustering model is chosen to minimize the BIC.

**Usage**

```
MFclust(data, minG, maxG, nchain=NULL, thres=NULL, iter.max=NULL, my.alpha=NULL, ...)
```

**Arguments**

<code>data</code>	A numeric vector, matrix, or data frame of observations. Categorical variables are not allowed. If a matrix or data frame, rows correspond to observations and columns correspond to time.
<code>minG</code>	An integer vector specifying the minimum number of mixture components (clusters) to be considered. The default is 1 component.
<code>maxG</code>	An integer vector specifying the maximum number of mixture components (clusters) to be considered. The default is 9 components.
<code>nchain</code>	An integer specifying the number of Markov chains in RCEM. The default is 5 chains.
<code>thres</code>	A number between 0 and 1 specifying the threshold value of rejection step in RCEM. The default is 0.5.
<code>iter.max</code>	An integer specifying the maximum number of iteration in RCEM. The default is 10.
<code>my.alpha</code>	The prior for the cluster proportion. The default is 1.
<code>...</code>	The arguments to be part of the function.

**Value**

A list representing the best model (according to BIC) for the given range of numbers of clusters. The following components are included:

<code>BIC</code>	A vector giving the BIC value for each number of clusters.
<code>nclust</code>	A scalar giving the optimal number of clusters.
<code>clust</code>	A vector whose $i$ th entry is the indicator that observation $i$ belongs to the $k$ component in the model.
<code>clust.center</code>	A matrix whose rows are the means of each group.

**References**

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

**See Also**[MFclust.compute](#)**Examples**

```
data("testdata")
my.clust<-MFclust(testdata,minG=3,maxG=5,nchain=1,iter.max=1)
```

---

MFclust.compute      *Computational Component in Model-Based Functional Data Clustering*

---

**Description**

The computation component for clustering via rejection-controlled EM initialized by kmeans clustering for functional Gaussian mixture models. The number of clusters and the clustering model is chosen to minimize the BIC.

**Usage**

```
MFclust.compute(data, minG, maxG, nchain, thres, iter.max, my.alpha, ...)
```

**Arguments**

data	A numeric vector, matrix, or data frame of observations. Categorical variables are not allowed. If a matrix or data frame, rows correspond to observations and columns correspond to time.
minG	An integer vector specifying the minimum number of mixture components (clusters) to be considered. The default is 1 component.
maxG	An integer vector specifying the maximum number of mixture components (clusters) to be considered. The default is 9 components.
nchain	An integer specifying the number of Markov chains in RCEM. The default is 5 chains.
thres	A number between 0 and 1 specifying the threshold value of rejection step in RCEM. The default is 0.5.
iter.max	An integer specifying the maximum number of iteration in RCEM. The default is 10.
my.alpha	The prior for the cluster proportion. The default is 1.
...	The arguments to be part of the function.

**Value**

A list representing the best model (according to BIC) for the given range of numbers of clusters.

## References

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data, *Nucleic Acids Research*, 34 (4), 1261-1269.

## See Also

[MFclust](#)

---

mkrandom

*Generating Random Effects for cluster model*

---

## Description

Generate entries representing random effect matrix

## Usage

```
mkrandom(formula, data)
```

## Arguments

formula	Symbolic description of the random effects.
data	Data frame.

## Details

This function generates random effects terms from simple grouping variables, for use in nonparametric mixed-effect models as described in Gu and Ma (2003a, b).

## Value

A list of three components.

z	Z matrix.
sigma	Sigma matrix to be evaluated through <code>sigma\$fun(para, sigma\$env)</code> .
init	Initial parameter values.

## Note

This program intrinsic function for MFclust. User cannot specify the parameters.

## Author(s)

pingma@uiuc.edu

---

testdata	<i>Time course simulated data</i>
----------	-----------------------------------

---

**Description**

Simulated functional clustered data with each row represents a subject and each column representing a time point. Data are generated from 4 functions with 30 curve from  $f1 = -exp(tt)/1000$ , 40 curves from  $f2 = tan(tt)$ , 50 curves from  $f3 = 5 * (tt - 4)^2 / max((tt - 4)^2)$ , 30 curves from  $f4 = cos(tt)$

**Usage**

```
data(testdata)
```

**Format**

a matrix with 150 rows and 10 columns

**Source**

simulated data

**References**

...

# Index

## \*Topic **cluster**

- check.start, 1
- compute.center, 2
- compute.M.all, 3
- compute.M.pk, 4
- compute.reject, 4
- compute.weight, 5
- em.bic, 6
- em.clust, 6
- Estep.tik, 7
- Estep.tk, 8
- MFclust, 9
- MFclust.compute, 10

## \*Topic **datasets**

- testdata, 12

## \*Topic **models**

- mkrandom, 11

## \*Topic **regression**

- mkrandom, 11

- check.start, 1
- compute.center, 2, 3
- compute.M.all, 3
- compute.M.pk, 4, 5
- compute.reject, 4
- compute.weight, 5

- em.bic, 6
- em.clust, 6, 6
- Estep.tik, 7, 9
- Estep.tk, 3, 8, 8

- MFclust, 9, 11
- MFclust.compute, 10, 10
- mkrandom, 11

- testdata, 12