Package ‘pmclust’

February 20, 2015

Version  0.1-6
Date     2014-02-02
Title    Parallel Model-Based Clustering
Depends  R (>= 3.0.0), methods, rlecuyer, pbdMPI (>= 0.2-2), MASS
Enhances MixSim, pbdSLAP (>= 0.1-7), pbdBASE (>= 0.3-0), pbdDMAT (>= 0.2-4)
LazyLoad yes
LazyData yes
Description The pmclust aims to utilize model-based clustering (unsupervised) for high dimensional and ultra large data, especially in a distributed manner. The package employs Rmpi to perform an expectation-gathering-maximization (EGM) algorithm for finite mixture Gaussian models. The unstructured dispersion matrices are assumed in the Gaussian models. The implementation is default in the single program multiple data (SPMD) programming model. The code can be executed through Rmpi and independent to most MPI applications. See the High Performance Statistical Computing (HPSC) website for more information, documents and examples.
License  GPL (>= 2)
URL     http://r-pbd.org/
BugReports http://group.r-pbd.org/
MailingList Please send questions and comments regarding pbdR to RBigData@gmail.com
Author  Wei-Chen Chen [aut, cre],
        George Ostrouchov [aut]
Maintainer Wei-Chen Chen <wccsnow@gmail.com>
NeedsCompilation yes
Repository CRAN
Date/Publication 2014-02-03 07:12:06
pmclust-package

R topics documented:

pmclust-package ......................................................... 2
as functions .......................................................... 4
assign.N.sample ....................................................... 5
EM-like algorithms ..................................................... 7
generate.basic ......................................................... 9
generate.MixSim ....................................................... 10
get.N.CLASS .......................................................... 12
Independent logL ....................................................... 13
Initialization ......................................................... 15
mb.print .............................................................. 16
One E-Step .......................................................... 17
One M-Step .......................................................... 19
One Step of EM algorithm ........................................ 20
pmclust and pkmeans .............................................. 21
print.object ........................................................ 23
Read Me First ....................................................... 24
Set Global Variables ............................................... 26
Set of CONTROL ................................................... 27
Set of PARAM ....................................................... 29
Update Class of EM or Kmeans Results ....................... 30

Index ............................................................................. 32

pmclust-package

Parallel Model-Based Clustering

Description

The pmclust aims to utilize model-based clustering (unsupervised) for high dimensional and ultra
large data, especially in a distributed manner. The package employs pbdMPI to perform a parallel
version of expectation and maximization (EM) algorithm for finite mixture Gaussian models. The
unstructured dispersion matrices are assumed in the Gaussian models. The implementation is de-
fault in the single program multiple data (SPMD) programming model. The code can be executed
through pbdMPI and independent to most MPI applications. See the High Performance Statistical
Computing (HPSC) website for more information, documents and examples.

Details

<table>
<thead>
<tr>
<th>Package:</th>
<th>pmclust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Package</td>
</tr>
<tr>
<td>License:</td>
<td>GPL</td>
</tr>
<tr>
<td>LazyLoad:</td>
<td>yes</td>
</tr>
</tbody>
</table>
The main function is `pmclust` implementing the parallel EM algorithm for mixture multivariate Gaussian models with unstructured dispersions. This function groups a data matrix \( X_{\text{gbd}} \) or \( X_{\text{spmd}} \) into \( K \) clusters where \( X_{\text{gbd}} \) or \( X_{\text{spmd}} \) is potentially huge and taken from the global environment \( .\text{GlobalEnv} \) or \( .\text{pmclustEnv} \).

Other main functions `em.step`, `aecm.step`, `apecm.step`, and `apecma.step` may provide better performance than the `em.step` in terms of computing time and convergent iterations.

`kmeans.step` provides the fastest clustering among above algorithms, but it is restricted by Euclidean distance and spherical dispersions.

**Author(s)**

Wei-Chen Chen &lt;wccsnow@gmail.com&gt; and George Ostrouchov

**References**


Programming with Big Data in R Website: [http://r-pbd.org/](http://r-pbd.org/)


**See Also**


**Examples**

```r
## Not run:
### Under command mode, run the demo with 2 processors by
### (Use Rscript.exe for windows system)
mpiexec -np 2 Rscript -e 'demo(gbd_em,"pmclust",ask=F,echo=F)'
mpiexec -np 2 Rscript -e 'demo(gbd_aecm,"pmclust",ask=F,echo=F)'
mpiexec -np 2 Rscript -e 'demo(gbd_apecm,"pmclust",ask=F,echo=F)'
mpiexec -np 2 Rscript -e 'demo(gbd_apecma,"pmclust",ask=F,echo=F)'
```
as functions

Convert between X.gbd (X.spmd) and X.dmat

Description

This function will convert a matrix of format from GBD row-major to ddmatrix vice versa.

Usage

```r
as.dmat(X.spmd, bldim = pbdDMAT::BLDIM, ICTXT = pbdDMAT::ICTXT,
      comm = .SPMD.CT$comm)

as.gbd(X.dmat, comm = .SPMD.CT$comm)
```

Arguments

- **X.spmd**: an input dataset of format GBD/SPMD row-major to be converted.
- **X.dmat**: an input dataset of format ddmatrix to be converted.
- **bldim**: block dimension, see `pbdBASE` and `pbdDMAT` for details.
- **ICTXT**: block context, see `pbdBASE` and `pbdDMAT` for details.
- **comm**: communicator, see `pbdMPI` for details.

Details

This function will provide a quick conversion to the input data `X.spmd` which is a default dataset to be clustered in `pmclust`. Usually, this matrix is distributed in GBD row-major format. For fully utilizing `pbdBASE` and `pbdDMAT`, it need to be converted to ddmatrix format.

Value

A ddmatrix is returned.
assign.N.sample

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/hpsc/
Programming with Big Data in R Website: http://r-pbd.org/

See Also

kmeans.step.dmat.

Examples

```r
## Not run:
# Examples can be found in the help pages of
# kmeans.step.dmat().

## End(Not run)
```

---

assign.N.sample  

Obtain a Set of Random Samples for X.spmd

Description

This utility function samples data randomly from \textit{X.spmd} to form a relatively small subset of original data. The EM algorithm on the smaller subset is topically performing fast and capturing rough structures of entire dataset.

Usage

```r
assign.N.sample(total.sample = 5000, N.org.spmd)
```

Arguments

- `total.sample` a total number of samples which will be selected from the original data \textit{X.spmd}.
- `N.org.spmd` the original data size, i.e. \textit{nrow}(\textit{X.spmd}).

Details

This utility function performs simple random sampling without replacement for the original dataset \textit{X.spmd}. Different random seeds should be set before calling this function.

Value

A list variable will be returned and containing:
assign.N.sample

N. spmd  sample size of given processor
N.allspmds  a collection of sample sizes for all S processors
ID. spmd  index of selected samples ranged from 1 to N.org.spmd

Note that N and N.allspmds are the same across all S processors, but N.spmd and ID.spmd are most likely all distinct. The lengths of these elements are 1 for N and N.spmd, S for N.allspmd, and N.spmd for ID.spmd.

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/hpsc/

Programming with Big Data in R Website: http://r-pbd.org/

See Also

set.global

Examples

## Not run:
# Save code in a file "demo.r" and run in 4 processors by
# > mpiexec -np 4 Rscript demo.r

### Setup environment.
library(pmclust, quiet = TRUE)
comm.set.seed(123)

### Generate an example data.
N.org.spmd <- 5000 + sample(1:1000, 1)
ret.spmd <- assign.N.sample(total.sample = 5000, N.org.spmd)
cat("Rank:", comm.rank(), " Size:", ret.spmd$N.spmd,
      "\n", sep = ")

### Quit.
finalize()

## End(Not run)
EM-like algorithms

EM-like Steps for GBD

Description

The EM-like algorithm for model-based clustering of finite mixture Gaussian models with unstructured dispersions.

*.dmat's are ddmatrix versions.

Usage

em.step(PARAM.org)
aecm.step(PARAM.org)
apecm.step(PARAM.org)
apecma.step(PARAM.org)
kmeans.step(PARAM.org)
em.step.dmat(PARAM.org)
kmeans.step.dmat(PARAM.org)

Arguments

PARAM.org an original set of parameters generated by set.global.

Details

A global variable called X.spmd should exist in the .pmclustEnv environment, usually the working environment. The X.spmd is the data matrix to be clustered, and this matrix has a dimension N.spmd by p.

A PARAM.org will be a local variable inside all EM-like functions em.step, aecm.step, apecm.step, apecma.step, and kmeans.step. This variable is a list containing all parameters related to models. This function also updates in the parameters by the EM-like algorithms, and return the convergent results. The details of list elements are initially generated by set.global.

Value

A convergent results will be returned the other list variable containing all new parameters which represent the components of models. See the help page of PARAM or PARAM.org for details.

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostroumov.
References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/hpsc/

Programming with Big Data in R Website: http://r-pbd.org/


See Also
set.global, mb.print, set.global.dmat.

Examples

```r
## Not run:
# Save code in a file "demo.r" and run in 4 processors by
# > mpiexec -np 4 Rscript demo.r

### Setup environment.
library(pmclust, quiet = TRUE)
comm.set.seed(123)

### Generate an example data.
N.allspmds <- rep(5000, comm.size())
N.spmd <- 5000
N.K.spmd <- c(2000, 3000)
N <- 5000 * comm.size()
p <- 2
K <- 2
data.spmd <- generate.basic(N.allspmds, N.spmd, N.K.spmd, N, p, K)
X.spmd <- data.spmd$X.spmd

### Run clustering.
PARAM.org <- set.global(K = K)  # Set global storages.
# PARAM.org <- initial.em(PARAM.org)  # One initial.
PARAM.org <- initial.RndEM(PARAM.org)  # Ten initials by default.
PARAM.new <- apecma.step(PARAM.org)  # Run APECMa.
em.update.class()  # Get classification.

### Get results.
```
Description

This function will generate a small set of data for testing algorithms.

Usage

```
generate.basic(N.allspmds, N.spmd, N.K.spmd, N, p, K)
```

Arguments

- `N.allspmds`: a collection of sample sizes for all `S` processors, i.e. a vector of length `S`.
- `N.spmd`: total sample size of given processor.
- `N.K.spmd`: sample size of each clusters given processor, i.e. sum over `N.K.spmd` is `N.spmd`, a vector of length `K`.
- `N`: total sample size across all `S` processors, i.e. sum over `N.spmd` is `N`.
- `p`: dimension of data `X.spmd`, i.e. `ncol(X.spmd)`.
- `K`: number of clusters.

Details

For all `S` processors, this function will generate in total `N` observations from `K` clusters in `p` dimensions.

The clusters centers and dispersions are generated automatically inside the code. Currently, it is not allowed for users to change, but it is not difficult to specify them by mimicking this code.

Value

A set of simulated data and information will be returned in a list variable including:

- `K`: number of clusters, as the input
- `p`: dimension of data `X.spmd`, as the input
- `N`: total sample size, as the input
- `N.allspmds`: a collection of sample sizes for all `S` processors, as the input
- `N.spmd`: total sample size of given processor, as the input
- `N.K.spmd`: sample size of each clusters given processor, as the input
- `X.spmd`: generated data set with dimension with dimension `N.spmd` * `p`
- `CLASS.spmd`: true id of each data, a vector of length `N.spmd` and has values from 1 to `K`
- `N.CLASS.spmd`: true sample size of each clusters, a vector of length `K`
generate.MixSim

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/hpsc/
Programming with Big Data in R Website: http://r-pbd.org/

See Also

generate.MixSim.

Examples

```r
## Not run:
# Examples can be found in the help pages of em.step(),
# aecm.step(), apecm.step(), and apecma.step().
## End(Not run)
```

generate.MixSim

Generate MixSim Examples for Testing

Description

This function utilizes MixSim to generate sets of data for testing algorithms.

Usage

generate.MixSim(N, p, K, MixSim.obj = NULL, MaxOmega = NULL, BarOmega = NULL, PiLow = 1.0, sph = FALSE, hom = FALSE)

Arguments

- **N** total sample size across all \( S \) processors, i.e. sum over \( N \). spmd is \( N \).
- **p** dimension of data \( X \). spmd, i.e. ncol(\( X \). spmd).
- **K** number of clusters.
- **MixSim.obj** an object returned from MixSim.
- **MaxOmega** maximum overlap as in MixSim.
- **BarOmega** averaged overlap as in MixSim.
- **PiLow** lower bound of mixture proportion as in MixSim.
- **sph** sph as in MixSim.
- **hom** hom as in MixSim.
Details

If `MixSim.obj` is NULL, then `BarOmega` and `MaxOmega` will be used in `MixSim` to obtain a new `MixSim.obj`.

Value

A set of simulated data and information will be returned in a list variable including:

- `K`: number of clusters, as the input
- `p`: dimension of data `X.spmd`, as the input
- `N`: total sample size, as the input
- `N.allspmds`: a collection of sample sizes for all `S` processors, as the input
- `N.spmd`: total sample size of given processor, as the input
- `X.spmd`: generated data set with dimension with dimension `N.spmd * p`
- `CLASS.spmd`: true id of each data, a vector of length `N.spmd` and has values from 1 to `K`
- `N.CLASS.spmd`: true sample size of each clusters, a vector of length `K`
- `MixSim.obj`: the true model where data `X.spmd` generated from

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References


Programming with Big Data in R Website: [http://r-pbd.org/](http://r-pbd.org/)

See Also

generate.basic.

Examples

```r
## Not run:
# Save code in a file "demo.r" and run in 4 processors by
# > mpiexec -np 4 Rscript demo.r

### Setup environment.
library(pmcclus, quiet = TRUE)

### Generate an example data.
N <- 5000
p <- 2
K <- 2
data.spmd <- generate.MixSim(N, p, K, BarOmega = 0.01)
X.spmd <- data.spmd$X.spmd
```
### Run clustering.

```r
PARAM.org <- set.global(K = K)  # Set global storages.
# PARAM.org <- initial.em(PARAM.org)  # One initial.
PARAM.org <- initial.RndEM(PARAM.org)  # Ten initials by default.
PARAM.new <- apecma.step(PARAM.org)  # Run APECma.
em.update.class()  # Get classification.
```

### Get results.

```r
N.CLASS <- get.N.CLASS(K)
comm.cat("# of class:", N.CLASS, "n")
comm.cat("# of class (true):", data.spmd$N.CLASS.spmd, "n")
```

### Quit.

```r
finalize()
```

```r
## End(Not run)
```

---

#### get.N.CLASS  
*Obtain Total Elements for Every Clusters*

**Description**

This function will collect the total elements for every clusters from all processors that the all reduced calls with the sum operation will be performed.

get.N.CLASS.dmat is a ddmatrix version of get.N.CLASS.

The get.CLASS returns class ids.

**Usage**

```r
get.N.CLASS(K)
get.N.CLASS.dmat(K)
get.CLASS(PARAM)
```

**Arguments**

- **K**  
  the total number of clusters.

- **PARAM**  
  a set of parameters.

**Details**

The final results are distributed in all processors including the total elements for each cluster. The global variable CLASS.spmd stores the identification for each observation on each processors. This function will first summary CLASS.spmd in K categories, then use the all reduce function with the sum operation to add the numbers by clusters. The COMM.RANK 0 will be used to take care the printing.
Description

This function is for debugging only and for checking if the observed data log likelihood is consistent for each EM iteration.

`indep.logL` is a ddmatrix version of `indep.logL`.

Usage

```r
indep.logL(PARAM)
```

```r
indep.logL.dmat(PARAM)
```
## Arguments

**PARAM**

a set of parameters.

## Details

This function will provide an observed data log likelihood based on the current parameter **PARAM**. This function will take in information from global, but no global variables will be updated by this function.

This function also don’t take care the numerical issues, so the return value may be inaccurate sometimes.

## Value

An observed data log likelihood will be returned. This value can quickly compare with the log likelihood computed inside `em.onestep`. Small difference is allowed, but large difference indicates bugs of code or illness of data.

## Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

## References


Programming with Big Data in R Website: [http://r-pbd.org/](http://r-pbd.org/)

## See Also

`set.global`, `em.onestep`

## Examples

```r
## Not run:
# This is a core function for em.estep() # see the source code for details.
# Reset .pmclustEnv$CONTROL$debug to turn on this function # automatically for each EM iteration.

## End(Not run)
```
Initialization

Description

These functions implement initialization of EM-like algorithms for model-based clustering based on \( X.\text{spmd} \), and initialization of K-means algorithm by randomly picking samples from data based on \( X.\text{spmd} \).

\*.dmat's are ddmatrix versions.

Usage

\[
\begin{align*}
\text{initial.RndEM} & (\text{PARAM}) \\
\text{initial.em} & (\text{PARAM, MU = NULL}) \\
\text{initial.center} & (\text{PARAM, MU = NULL}) \\
\text{initial.RndEM.dmat} & (\text{PARAM}) \\
\text{initial.em.dmat} & (\text{PARAM, MU = NULL}) \\
\text{initial.center.dmat} & (\text{PARAM, MU = NULL})
\end{align*}
\]

Arguments

\begin{itemize}
  \item PARAM \hspace{1cm} an original set of parameters generated by \text{set.global}.
  \item MU \hspace{1cm} a center matrix with dim = \( p \times K \).
\end{itemize}

Details

For \text{initial.RndEM}, the procedure is implemented by randomly picking \( \text{pmclustEnv}\$\text{CONTROL}\$\text{RndEM.iter} \) starting points from data \( X.\text{spmd} \) and run one E-step to obtain the log likelihood. Then pick the starting point with the highest log likelihood as the best choice to pursue the MLEs in further EM iterations.

This function repeatedly run \text{initial.em} by \( \text{pmclustEnv}\$\text{CONTROL}\$\text{RndEM.iter} \) random starts and pick the best initializations from the random starts.

For \text{initial.em}, it takes \( X.\text{spmd} \) from the global environment and randomly pick \( K \) of them as the centers of \( K \) groups. If \( \text{MU} \) is specified, then this \( \text{MU} \) will be the centers. The default identity dispersion in \text{PARAM}\$\text{SIGMA} \) will be used. Then, one E-step will be called to obtain the log likelihood and new classification will be updated.

This function is used to implement the RndEM procedure for more elaborate initialization scheme in \text{initial.RndEM}. Potentially, several random starts should be tried before running EM algorithms. This can benefit in two aspects including: shorter convergent iterations and better classification results.

For \text{initial.center}, if \( \text{MU} \) is given, then the center will be assigned according.
Value

The best initial starting points PARAM will be returned among all random starting points. The number of random starting points is assigned by set.global to a list variable CONTROL. See the help page of initial.em and set.global for details.

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/hpsc/
Programming with Big Data in R Website: http://r-pbd.org/

See Also

set.global, em.step, aecm.step,
aecm.step, apecma.step, kmeans.step.
set.global.dmat, kmeans.step.dmat.

Examples

## Not run:
# Examples can be found in the help page of em.step(),
# aecm.step(), apecma.step(), apecma.step(), and kmeans.step().

# Examples for ddmatrix version can be found in the help pages of
# kmeans.step.dmat().

## End(Not run)
One E-Step

Arguments

PARAM a set of convergent parameters to be printed.
CHECK a set of checking parameters to be printed.

Details

This function will provide a quick summary from the PARAM and CHECK typically the output of clusterings when algorithms stop. The COMM.RANK 0 will be used to take care the printing.

Value

Summarized messages will print/cat on screen by default.

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/hpsc/
Programming with Big Data in R Website: http://r-pbd.org/

See Also

e.m.step, aecm.step, apecm.step,
apecma.step.

Examples

```r
## Not run:
# Examples can be found in the help pages of em.step(),
# aecm.step(), apecm.step(), and apecma.step().
## End(Not run)
```

One E-Step

Compute One E-step and Log Likelihood Based on Current Parameters

Description

This function will perform one E-step based on current parameters. This is a core function of em.onestep.

e.step.dmat is a ddmatrix version of e.step.
**Usage**

```r
e.step(PARAM, update.logL = TRUE)
```

```r
e.step.dmat(PARAM, update.logL = TRUE)
```

**Arguments**

- **PARAM**: A set of parameters.
- **update.logL**: `TRUE` for update observed data log likelihood.

**Details**

This function will base on the current parameter to compute the densities for all observations for all K components, and update the `Z.spmd` matrix. If the `update.logL` is true, then the log likelihood `W.spmd.rowSums` will be also updated before the end of this function.

Sum of `W.spmd.rowSums` of all processors will be the observed data log likelihood for the current iteration.

**Value**

Several global variables will be overwrote after this call including `Z.spmd`, `W.spmd.rowSums`, `W.spmd`, `U.spmd`, and `Z.colSums`.

**Computing Issues**

Since the clusters can be degenerated or highly flat, these cause very large positive or negative exponents in densities. The log likelihood will tend to be inaccurate (not finite). Since the mixture structures can be over fit, this also cause very tiny mixing proportions. The poster probabilities can also unstable (NaN).

These can be solved by rescaling the range of exponents carefully and adjust the scaling factor on the log values. See **CONTROL** for details about constrains on E- and M-steps.

**Author(s)**

Wei-Chen Chen `<wccsnow@gmail.com>` and George Ostrouchov.

**References**

- Programming with Big Data in R Website: [http://r-pbd.org/](http://r-pbd.org/)

**See Also**

- `set.global`, `em.onestep.m.step`
Examples

```r
## Not run:
# This is a core function for em.onestep()
# see the source code for details.

## End(Not run)
```

### Description

This function will perform one M-step based on current posterior probabilities. This is a core function of `em.onestep`. `m.step.dmat` is a `ddmatrix` version of `m.step`.

### Usage

```r
m.step(PARAM)
m.step.dmat(PARAM)
```

### Arguments

- **PARAM**: A set of parameters.

### Details

This function will base on the current posterior probabilities `Z.spmd` to estimate the parameters `PARAM` mainly including mixing proportions `ETA`, centers of clusters `MU`, and dispersions of clusters `SIGMA`.

### Value

Returning a new `PARAM` which maximizes the complete data log likelihood for the current iteration.

### Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

### References


Programming with Big Data in R Website: [http://r-pbd.org/](http://r-pbd.org/)
One Step of EM algorithm

See Also

set.global, em.onestep.e.step.

Examples

```r
## Not run:
# This is a core function for em.onestep()
# see the source code for details.

## End(Not run)
```

Description

One EM step only for model-based clustering of finite mixture Gaussian models with unstructured dispersions. This is a core function of `em.step`.

`em.onestep.dmat` is a `ddmatrix` version of `em.onestep`.

Usage

```r
em.onestep(PARAM)
em.onestep.dmat(PARAM)
```

Arguments

`PARAM`  
an original set of parameters generated by `set.global`.

Details

A global variable called `X.spmd` should exist in the `.pmclustEnv` environment, usually the working environment. The `X.spmd` is the data matrix to be clustered, and this matrix has a dimension `N.spmd` by `p`.

The `PARAM` will be a local variable for the current iteration inside `em.onestep`, and this variable is a list containing all parameters related to models. This function also updates in the parameters by the EM algorithm, and return a new `PARAM` for the next iteration. The details of list elements are initially generated by `set.global`.

Value

This function is one EM step. The global variables will be updated and a new `PARAM` will be returned. See the help page of `PARAM` or `PARAM.org` for details.
pmclust and pkmeans

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/HPSC/
Programming with Big Data in R Website: http://r-pbd.org/

See Also

set.global, e.step, m.step.

Examples

```r
## Not run:
# This is a core function for em.step()
# see the source code for details.

## End(Not run)
```

pmclust and pkmeans

Parallel Model-Based Clustering and Parallel K-means Algorithm

Description

Parallel Model-Based Clustering and Parallel K-means Algorithm

Usage

```r
pmclust(X = NULL, K = 2, MU = NULL, 
algorithm = .PMC.CT$algorithm, RndEM.iter = .PMC.CT$RndEM.iter, 
CONTROL = .PMC.CT$CONTROL, method.own.X = .PMC.CT$method.own.X, 
rank.own.X = .SPMD.CT$rank.source, comm = .SPMD.CT$comm) 

pkmeans(X = NULL, K = 2, MU = NULL, 
algorithm = c("kmeans", "kmeans.dmat"), 
CONTROL = .PMC.CT$CONTROL, method.own.X = .PMC.CT$method.own.X, 
rank.own.X = .SPMD.CT$rank.source, comm = .SPMD.CT$comm)
```

Arguments

- `X` a GBD row-major matrix or a ddmatrix.
- `K` number of clusters.
- `MU` pre-specified centers.
- `algorithm` types of EM algorithms.
pmclust and pkmeans

RndEM.iter number of Rand-EM iterations.
CONTROL a control for algorithms, see CONTROL for details.
method.own.X how \(X\) is distributed.
rank.own.X who own \(X\) if method.own.X = "single".
comm MPI communicator.

Details

These are high-level functions for several functions in pmclust including: data distribution, setting global environment \(\text{.pmclustEnv}\), initializations, algorithm selection, etc.

The input \(X\) is either in \texttt{ddmatrix} or \texttt{gbd}. It will be converted in \texttt{gbd} row-major format and copied into \texttt{.pmclustEnv} for computation. By default, pmclust uses a GBD row-major format (\texttt{gbdr}). While \textit{common} means that \(X\) is identical on all processors, and \textit{single} means that \(X\) only exist on one processor \texttt{rank.own.X}.

Value

These functions return a list with class \texttt{pmclust} or \texttt{pkmeans}.

See the help page of \texttt{PARAM} or \texttt{PARAM.org} for details.

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References

High Performance Statistical Computing (HPSC) Website: \texttt{http://thirteen-01.stat.iastate.edu/snoweye/hpsc/}
Programming with Big Data in R Website: \texttt{http://r-pbd.org/}

See Also

\texttt{set.global.e.step.m.step}.
\texttt{set.global.dmat.e.step.dmat.m.step.dmat}.

Examples

```r
## Not run:
# Save code in a file "demo.r" and run in 4 processors by
# > mpiexec -np 4 Rscript demo.r

### Setup environment.
library(pmclust, quiet = TRUE)

### Load data
X <- as.matrix(iris[, -5])

### Distribute data
jid <- get.jid(nrow(X))
```
print.object

X.gbd <- X[jid,]

### Standardized
N <- allreduce(nrow(X.gbd))
p <- ncol(X.gbd)
mu <- allreduce(colSums(X.gbd / N))
X.std <- sweep(X.gbd, 2, mu, FUN = "-")
std <- sqrt(allreduce(colSums(X.std^2 / (N - 1))))
X.std <- sweep(X.std, 2, std, FUN = "/")

### Clustering
library(pmclust, quiet = TRUE)
comm.set.seed(123, diff = TRUE)

ret.mb1 <- pmclust(X.std, K = 3)
comm.print(ret.mb1)

ret.kms <- pkmeans(X.std, K = 3)
comm.print(ret.kms)

### Finish
finalize()

### End(Not run)

print.object

Functions for Printing or Summarizing Objects According to Classes

Description

Several classes are declared in `pmclust`, and these are functions to print and summary objects.

Usage

```r
## S3 method for class 'pmclust'
print(x, ...)
## S3 method for class 'pkmeans'
print(x, ...)
```

Arguments

- `x` an object with the class attributes.
- `...` other possible options.

Details

These are useful functions for summarizing.
Read Me First Function

Description
This function print the annotations of all variables used in this package.
*.dmat's are ddmatrix versions.

Usage
readme()
readme.dmat()

Details
This package is optimized in the way by pre-specifying several global variables in .pmclustEnv.
These variables will be overwrote by EM algorithms. Users should use these names to access the
results and utilize them with cautions.
readme.dmat is a ddmatrix version of readme.
Value

A readme message will print on screen by default and explain the global variables used in this package, including:

- **CHECK**  
  convergent checking

- **CLASS.spmd**  
  true id of each data, a vector of length N.spmd and has values from 1 to K

- **COMM.RANK**  
  rank of current processor, obtained from comm.rank of pbdMPI

- **COMM.SIZE**  
  total processors in MPI world, obtained from comm.size of pbdMPI

- **CONTROL**  
  controls for EM iterations

- **PARAM**  
  set or parameters

- **SAVE.param**  
  (debug only) save parameters for every iterations

- **SAVE.iter**  
  (debug only) save computing time for every iterations

- **U.spmd**  
  temporary storage for density

- **W.spmd**  
  temporary storage for eta * density

- **W.spmd.rowSums**  
  temporary storage for rowSums of W.spmd

- **X.spmd**  
  generated data set with dimension with dimension N.spmd * p

- **Z.colSums**  
  temporary storage for rowSums of Z.spmd

- **Z.spmd**  
  posterior probabilities

- **p.times.logtwopi**  
  p * log(2 * pi)

Each variable may contain several elements if it is a list, some variables are used for temporary storages in order to optimize computing, and some variables are used for constant variables. These variables may be restricted, and only generated by the function `set.global`.

One can access these variables via the global environment `.pmclustEnv` such as `.pmclustEnv$CONTROL`.

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References


Programming with Big Data in R Website: [http://r-pbd.org/](http://r-pbd.org/)

See Also

- `set.global`, `set.global.dmat`

Examples

```r
## Not run:
readme()
readme.dmat()

## End(Not run)
```
Set Global Variables

Set Global Variables According to the global matrix X.gbd (X.spmd) or X.dmat

Description

This function will set several sets of variables globally in the environment .pmclustEnv according to the global matrix X.gbd/X.spmd.

set.global.dmat is a ddmatrix version of set.global.gbd.

Usage

```r
set.global.gbd(K = 2, X.gbd = NULL, PARAM = NULL, 
               algorithm = c("em", "aecm", "apecm", "apecma", "kmeans"), 
               RndEM.iter = 10)
```

```r
set.global.dmat(K = 2, X.dmat = NULL, PARAM = NULL, 
                algorithm = c("em.dmat", "kmeans.dmat"), 
                RndEM.iter = 10)
```

```r
set.global(K = 2, X.spmd = NULL, PARAM = NULL, 
           algorithm = c("em", "aecm", "apecm", "apecma", "kmeans"), 
           RndEM.iter = 10)
```

Arguments

- **K**: an original set of parameters generated by set.global.
- **X.gbd**: an input GBD matrix.
- **X.dmat**: an input ddmatrix.
- **X.spmd**: an input SPMD matrix.
- **PARAM**: an original set of parameters generated by set.global.
- **algorithm**: an original set of parameters generated by set.global.
- **RndEM.iter**: number of RndEM iterations.

Details

**WARNING**: A global variable named X.gbd/X.spmd should be set before calling set.global where X.gbd/X.spmd is a matrix containing data with dimension N.spmd * p. i.e. N.spmd observations and p variables.

**WARNING**: A global variable named X.dmat should be set before calling set.global.dmat where X.dmat is a ddmatrix (in block-cyclic format) containing data with dimension N * p.

X.gbd/X.spmd and X.dmat are supposed to exist in .GlobalEnv. If not, they should be as an input object and will be copied into .pmclustEnv which is less efficient.
Value
A new set of \texttt{PARAM} will be returned and several global variables will be set according to the data \texttt{X.gbd/X.spmd}.
Sets of global variables are store in the default environment \texttt{.pmclustEnv}.
Use \texttt{readme} to see all global variables set by this function.
Use \texttt{readme.dmat} to see all global variables of ddmatrix version set by this function.

Author(s)
Wei-Chen Chen <\texttt{wccsnow@gmail.com}> and George Ostrouchov.

References
High Performance Statistical Computing (HPSC) Website: \url{http://thirteen-01.stat.iastate.edu/snoweye/hpsc/}
Programming with Big Data in R Website: \url{http://r-pbd.org/}

See Also
\texttt{em.step, aecm.step, apecm.step, apecma.step}.

Examples
```r
## Not run:
# Examples can be found in the help pages of \texttt{em.step()},
# \texttt{aecm.step()}, \texttt{apecm.step()}, \texttt{apecma.step()}, and \texttt{kmeans.step()}.

# Examples for ddmatrix version can be found in the help pages of
# \texttt{kmeans.step.dmat()}.

## End(Not run)
```

Description
This set of controls are used to guide all algorithms implemented in this package.

Format
A list variable contains several parameters for computing.

Details
\texttt{.PMC.CT} stores all default controls for \texttt{pmclust} and \texttt{pkmeans} including
algorithm algorithms implemented
algorithm.gbd algorithms implemented for gbd/spmd
algorithm.dmat algorithms implemented for ddmatrix
method.own.X how X is distributed
CONTROL a CONTROL list as in next

The elements of CONTROL or .pmclustEnv$CONTROL are

- max.iter maximum number of iterations (1000)
- abs.err absolute error for convergence (1e-4)
- rel.err relative error for convergence (1e-6)
- debug debugging flag (0)
- RndEM.iter number of RndEM iterations (10)
- exp.min minimum exponent (log(.Machine$double.xmin))
- exp.max maximum exponent (log(.Machine$double.xmax))
- U.min minimum of diagonal of chol
- U.max maximum of diagonal of chol

These elements govern the computing including number of iterations, convergent criteria, ill conditions, and numerical issues. Some of them are machine dependent.

Currently, the algorithm could be em, aecm, apecm, apecma, and kmeans for GBD and ddmatrix (with *.dmat). The method.own.X could be gbdr, common, single, and ddmatrix.

Numerical Issues

For example, exp.min and exp.max will control the range of densities function before taking logarithm. If the density values were no in the range, they would be rescaled. The scaling factor will be also recorded for post adjustment for observed data log likelihood. This will provide more accurate posterior probabilities and observed data log likelihood.

Also, U.min and U.max will control the output of chol when decomposing SIGMA in every E-steps. If the diagonal terms were out of the range, a PARAM$U.check would be set to FALSE. Only the components with TRUE U.check will estimate and update the dispersions in M-steps for the rest of iterations.

These problems may cause wrong posteriors and log likelihood due to the degenerate and inflated components. Usually, this is a sign of overestimate the number of components K, or the initialization do not provide good estimations for parameters. See e.step for more information about computing.

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.

References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/hpsc/
Programming with Big Data in R Website: http://r-pbd.org/
Set of PARAM

See Also

set.global.gbd, set.global.dmat, and set.global.

Examples

```r
## Not run:
# Use set.global() to generate one of this.
# X.spmd should be pre-specified before calling set.global().

## End(Not run)
```

---

Set of PARAM

A Set of Parameters in Model-Based Clustering.

Description

This set of parameters are used in initialization, EM iterations, and final convergent results. All share the same structure in a list variable.

Format

A list variable contains several parameters for computing.

Details

The elements of PARAM or PARAM.org are

- **N**: number of observations
- **p**: dimension of each observation, total number of variables
- **K**: number of clusters
- **ETA**: mixing proportion
- **log.ETA**: log of mixing proportion
- **MU**: centers, dim = \( p \times K \)
- **SIGMA**: dispersions, a list containing \( K \) elements, each element is a matrix, dim = \( p \times p \)
- **U**: Choleski of SIGMA, the same size of SIGMA
- **U.check**: checks of each elements of U, length \( K \)
- **logL**: log likelihood
- **min.N.CLASS**: minimum number of elements in a cluster (restrictions)

The model parameters are ETA, MU, and SIGMA, while log.ETA, U, U.check, and min.N.CLASS are only used in computing.

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.
Update Class of EM or Kmeans Results

References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/hpsc/
Programming with Big Data in R Website: http://r-pbd.org/

See Also

set.global.

Examples

```r
## Not run:
# Use set.global() to generate one of this.
# X.spmd should be pre-specified before calling set.global().

## End(Not run)
```

---

Update Class of EM or Kmeans Results

*Update CLASS.spmd Based on the Final Iteration*

Description

Update CLASS.spmd based on the final iteration of EM-like algorithms.
*.dmat's are ddmatrix versions.

Usage

```r
em.update.class()
kmeans.update.class()

em.update.class.dmat()
kmeans.update.class.dmat()
```

Details

This function takes Z.spmd from the global environment .pmclustEnv and update CLASS.spmd,
and provides the identification of groups for all data.

Value

CLASS.spmd will be updated.

Author(s)

Wei-Chen Chen <wccsnow@gmail.com> and George Ostrouchov.
Update Class of EM or Kmeans Results

References

High Performance Statistical Computing (HPSC) Website: http://thirteen-01.stat.iastate.edu/snoweye/hpsc/

Programming with Big Data in R Website: http://r-pbd.org/

See Also

em.step, aecm.step, apecm.step,
apecm.step, kmeans.step.
kmeans.step.dmat.

Examples

## Not run:
# Examples can be found in the help pages of em.step(),
# aecm.step(), apecm.step(), apecm.step(), and kmeans.step().

# Examples for ddmatrix version can be found in the help pages of
# kmeans.step.dmat().

## End(Not run)
Index

+Topic **algorithm**
  EM-like algorithms, 7
+Topic **core function**
  One E-Step, 17
  One M-Step, 19
  One Step of EM algorithm, 20
+Topic **debugging function**
  Independent logl, 13
+Topic **global variables**
  Read Me First, 24
  Set Global Variables, 26
  Set of CONTROL, 27
  Set of PARAM, 29
+Topic **high-level function**
  pmclust and pkmeans, 21
  print.object, 23
+Topic **initialization**
  Initialization, 15
+Topic **package**
  pmclust-package, 2
+Topic **programming**
  as functions, 4
  assign.N.sample, 5
  generate.basic, 9
  generate.MixSim, 10
  get.N.CLASS, 12
  mb.print, 16
  Update Class of EM or Kmenas Results, 30
  .PMC.CT (Set of CONTROL), 27
  .pmclustEnv, 3, 7, 20, 30
  .pmclustEnv (Set Global Variables), 26
  aecm.step, 3, 7, 13, 16, 17, 27, 31
  aecm.step (EM-like algorithms), 7
  apecm.step, 3, 7, 13, 16, 17, 27, 31
  apecm.step (EM-like algorithms), 7
  apecma.step, 3, 7, 13, 16, 17, 27, 31
  apecma.step (EM-like algorithms), 7
  as functions, 4
  as.dmat (as functions), 4
  as.gbd (as functions), 4
  as.spmd (as functions), 4
  assign.N.sample, 5
  CHECK (Read Me First), 24
  chol, 28
  CLASS.dmat (Read Me First), 24
  CLASS.spmd, 9, 11, 12, 30
  CLASS.spmd (Read Me First), 24
  COMM.RANK, 12, 17
  COMM.RANK (Read Me First), 24
  COMM.SIZE (Read Me First), 24
  CONTROL, 16, 18, 22, 25
  CONTROL (Set of CONTROL), 27
  e.step, 20–22, 28
  e.step (One E-Step), 17
  e.step.dmat, 22
  EM-like algorithms, 7
  em.onestep, 14, 17–20
  em.onestep (One Step of EM algorithm), 20
  em.step, 3, 7, 13, 16, 17, 20, 27, 31
  em.step (EM-like algorithms), 7
  em.step.dmat, 3
  em.update.class (Update Class of EM or Kmenas Results), 30
  ETA, 19
  ETA (Set of PARAM), 29

generate.basic, 9, 11
generate.MixSim, 10, 10
get.CLASS (get.N.CLASS), 12
get.N.CLASS, 12

indep.logL (Independent logL), 13
Independent logL, 13
initial.center (Initialization), 15
initial.em, 15, 16
INDEX

initial.em (Initialization), 15
initial.RndEM, 15
initial.RndEM (Initialization), 15
Initialization, 15
kmeans.step, 3, 7, 13, 16, 31
kmeans.step (EM-like algorithms), 7
kmeans.step.dmat, 3, 5, 13, 16, 31
kmeans.update.class (Update Class of EM or Kmenas Results), 30
m.step, 18, 21, 22
m.step (One M-Step), 19
m.step.dmat, 22
mb.print, 8, 16
MixSim, 10, 11
MU, 15, 19
MU (Set of PARAM), 29
One E-Step, 17
One M-Step, 19
One Step of EM algorithm, 20
p.times.logtwopi (Read Me First), 24
PARAM, 7, 14, 16, 19, 20, 22, 25, 27
PARAM (Set of PARAM), 29
PARAM.org, 7, 20, 22
pkmeans, 24
pkmeans (pmclust and pkmeans), 21
pmclust, 3, 24
pmclust (pmclust and pkmeans), 21
pmclust and pkmeans, 21
pmclust-package, 2
print.object, 23
print.pkmeans (print.object), 23
print.pmclust (print.object), 23
Read Me First, 24
readme, 27
readme (Read Me First), 24
readme.dmat, 27
SAVE.iter (Read Me First), 24
SAVE.param (Read Me First), 24
Set Global Variables, 26
Set of CONTROL, 27
Set of PARAM, 29
set.global, 6–8, 14–16, 18, 20–22, 25, 29, 30
set.global (Set Global Variables), 26
set.global.dmat, 8, 16, 22, 25, 29
set.global.gbd, 29
SIGMA, 19, 28
SIGMA (Set of PARAM), 29
U.dmat (Read Me First), 24
U.spmd, 18
U.spmd (Read Me First), 24
Update Class of EM or Kmenas Results, 30
W.dmat (Read Me First), 24
W.spmd, 18
W.spmd (Read Me First), 24
W.spmd.rowSums, 18
X.dmat (Set Global Variables), 26
X.gbd, 3
X.gbd (Set Global Variables), 26
X.spmd, 3, 5, 7, 9–11, 15, 20, 25
X.spmd (Set Global Variables), 26
Z.colSums, 18
Z.colSums (Read Me First), 24
Z.dmat (Read Me First), 24
Z.spmd, 18, 19, 30
Z.spmd (Read Me First), 24
set.Nglobal, 6–8, 14–16, 18, 20–22, 25, 29, 30
set.Nglobal (Set Global Variables), 26
set.Nglobal.dmat, 8, 16, 22, 25, 29
set.Nglobal.gbd, 29
SIGMA, 19, 28
SIGMA (Set of PARAM), 29
U.dmat (Read Me First), 24
U.spmd, 18
U.spmd (Read Me First), 24
Update Class of EM or Kmenas Results, 30
W.dmat (Read Me First), 24
W.spmd, 18
W.spmd (Read Me First), 24
W.spmd.rowSums, 18
X.dmat (Set Global Variables), 26
X.gbd, 3
X.gbd (Set Global Variables), 26
X.spmd, 3, 5, 7, 9–11, 15, 20, 25
X.spmd (Set Global Variables), 26
Z.colSums, 18
Z.colSums (Read Me First), 24
Z.dmat (Read Me First), 24
Z.spmd, 18, 19, 30
Z.spmd (Read Me First), 24
set.Nglobal, 6–8, 14–16, 18, 20–22, 25, 29, 30
set.Nglobal (Set Global Variables), 26
set.Nglobal.dmat, 8, 16, 22, 25, 29
set.Nglobal.gbd, 29
SIGMA, 19, 28
SIGMA (Set of PARAM), 29
U.dmat (Read Me First), 24
U.spmd, 18
U.spmd (Read Me First), 24
Update Class of EM or Kmenas Results, 30
W.dmat (Read Me First), 24
W.spmd, 18
W.spmd (Read Me First), 24
W.spmd.rowSums, 18
X.dmat (Set Global Variables), 26
X.gbd, 3
X.gbd (Set Global Variables), 26
X.spmd, 3, 5, 7, 9–11, 15, 20, 25
X.spmd (Set Global Variables), 26
Z.colSums, 18
Z.colSums (Read Me First), 24
Z.dmat (Read Me First), 24
Z.spmd, 18, 19, 30
Z.spmd (Read Me First), 24
set.Nglobal, 6–8, 14–16, 18, 20–22, 25, 29, 30
set.Nglobal (Set Global Variables), 26
set.Nglobal.dmat, 8, 16, 22, 25, 29
set.Nglobal.gbd, 29
SIGMA, 19, 28
SIGMA (Set of PARAM), 29
U.dmat (Read Me First), 24
U.spmd, 18
U.spmd (Read Me First), 24
Update Class of EM or Kmenas Results, 30
W.dmat (Read Me First), 24
W.spmd, 18
W.spmd (Read Me First), 24
W.spmd.rowSums, 18
X.dmat (Set Global Variables), 26
X.gbd, 3
X.gbd (Set Global Variables), 26
X.spmd, 3, 5, 7, 9–11, 15, 20, 25
X.spmd (Set Global Variables), 26
Z.colSums, 18
Z.colSums (Read Me First), 24
Z.dmat (Read Me First), 24
Z.spmd, 18, 19, 30
Z.spmd (Read Me First), 24
sigma, 19, 28
sigma (Set of Param), 29
uNdmat, 18
uNspmd, 18
uNspmd (Read Me First), 24
wNdmat, 18
wNspmd, 18
wNspmd (Read Me First), 24
xNgbd, 3
xNgbd (Set Global Variables), 26
xNspmd, 3, 5, 7, 9–11, 15, 20, 25
xNspmd (Set Global Variables), 26
zNdmat, 18
zNspmd, 18
zNspmd (Read Me First), 24
zNcolsums, 18
zNspmd (Set Global Variables), 26