Package 'SUITOR'

January 14, 2022

Title Selecting the number of mutational signatures through cross-validation

Version 1.0.0

Date 2022-01-14

Author Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

Description Selects the number of mutational signatures through cross-validation.

Maintainer Bill Wheeler <wheelerb@imsweb.com>

Depends R (>= 3.5.0), doParallel, foreach, parallel, ggplot2

License GPL-2

NeedsCompilation yes

R topics documented:

`SUITOR-package` *Selecting the number of mutational signatures through cross-validation*

Description

Selecting the number of mutational signatures through cross-validation

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

`getSummary` *Compute summary results*

Description

Compute summary results and the optimal rank from the matrix containing all results

Usage

```
getSummary(obj, NC, NR=96)
```

Arguments

`obj`	Matrix containing all results in the return list from `suitor`.
`NC`	The number of columns in `data` when `suitor` was called.
`NR`	The number of rows in `data` when `suitor` was called. The default is 96.

Details

The input matrix `obj` must have column 1 as the rank, column 2 as the value of `k` in `1:k.fold`, column 4 as the training errors, and column 5 as the testing errors.

Value

A list containing the objects:

- `rank`: The optimal rank
- `all.results`: Matrix containing training and testing errors for all values of seeds, ranks, folds. NA values appear for runs in which the EM algorithm did not converge.
- summary: Data frame of summarized results for each possible rank created from `all.results`. The `MSErr` column is defined as sqrt({fold1 + ... +foldK}/{nrow(data)*ncol(data)})

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

See Also

`plotErrors`

Examples

```
data(SimData, package="SUITOR")
data(results, package="SUITOR")
ret <- getSummary(results$all.results, ncol(SimData))
ret$summary
ret$rank
```

plotData *Example data for plotting*

Description

A data frame with coumns Rank, Type, and MSErr

See Also

suitor

Examples

```
data(plotData, package="SUITOR")

plotData
```

plotErrors *Plot train and test errors*

Description

Plot train and test errors

Usage

```
plotErrors(x)
```

Arguments

x	Data frame of summary results in the return list from suitor or from getSummary, or a data frame with columns Rank, Type, and MSErr.

Details

The optimal rank is the minimum at which the test error is attained, and appears as a red dot on the graph.

Value

NULL

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

Examples

```
data(plotData, package="SUITOR")
plotErrors(plotData)
```

results *SUITOR return object*

Description

An object returned from the suitor function for examples

See Also

suitor

Examples

```
data(results, package="SUITOR")

results
```

SimData *Data for examples*

Description

Example input data and results

Details

Contains an example input data object of size 96 by 300. It is generated by `rpois` with mean WH where W (96 by 8) is profile of 8 signatures (SBS 4, 6, 7a, 9, 17b, 22, 26, 39) obtained from https://cancer.sanger.ac.uk/cosmic/signatures/SBS and H (8 by 300) is rounded integer generated from a uniform distribution between 0 and 100 with some randomly selected cells being set to zero.

See Also

suitor

Examples

```
data(SimData, package="SUITOR")

# Display a subset of data objects
SimData[1:5, 1:5]
```

suitor *suitor*

Description

Selecting the number of mutational signatures through cross-validation

Usage

```
suitor(data, op=NULL)
```

Arguments

data	Data frame or matrix containing mutational signatures. This object must contain non-negative values
op	List of options (see details). The default is NULL.

Details

The algorithm finds the optimal rank by applying k-fold cross validation.

Options list op:

Name	Description	Default Value
em.eps	EM algorithm stopping tolerance	1e-5
get.summary	0 or 1 to create summary results	1
k.fold	Number of folds	10
max.iter	Maximum number of iterations in EM algorithm	2000
max.rank	Maximum rank	10
min.rank	Minimum rank	1
min.value	Minimum value of matrix before factorizing	1e-4
n.cores	Number of cores to use for parallel computing	1
n.seeds	Number of seeds (starting points)	30
plot	0 or 1 to produce an error plot	1
print	0 or 1 to print info	1
seeds	Vector of seeds (takes precendence over n.seeds)	NULL
type	Socket type in `makeCluster` for parallel computing	NULL
kfold.vec	Vector of values in 1:k.fold when running on a cluster	NULL

Parallel computing

If `n.cores > 1` and `type = NULL`, then `type` will be set to "FORK" when running on unix, and set to "PSOCK" otherwise.

NOTE: The R_LIBS_USER environment variable may need to be set to the path where the SUITOR R package was installed, especially if it was not installed in the default location. Examples of setting the R_LIBS_USER environment variable on unix and windows are below.

(unix) export R_LIBS_USER=/path/to/SUITOR/package

(windows) set R_LIBS_USER=/path/to/SUITOR/package

Utilizing a cluster

When running on a cluster, the option `get.summary` should be set to 0. For fastest running jobs, set the options `min.rank = max.rank`, `kfold.vec` to a single integer in `1:k.fold`, and `seeds` to a

single value.

Value

A list containing the objects:

- `rank`: The optimal rank
- `all.results`: Matrix containing training and testing errors for all values of seeds, ranks, folds.
- summary: Data frame of summarized results for each possible rank created from `all.results`. The `MSErr` column is defined as sqrt({fold1 + ... +foldK}/{nrow(data)*ncol(data)})

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

See Also

`getSummary`, `plotErrors`

Examples

```
data(SimData, package="SUITOR")

# Using the default options will take several minutes to run
#ret <- suitor(SimData)
```

`suitor_extract_WH` *suitor_extract_WH*

Description

Extract the matrix of activities (exposures) and matrix of signatures

Usage

```
suitor_extract_WH(data, rank, op=NULL)
```

Arguments

`data`	Data frame or matrix containing mutational signatures. This object must contain non-negative values
`rank`	Integer > 0
`op`	List of options (see details). The default is NULL.

Details

Options list op:

Name	Description	Default Value
min.value	Minimum value of matrix before factorizing	1e-4
n.cores	Number of cores to use for parallel computing	1
n.seeds	Number of seeds (starting points)	30
print	0 or 1 to print info	1
seeds	Vector of seeds (takes precendence over n.seeds)	NULL
type	Socket type in `makeCluster` for parallel computing	NULL

Parallel computing
If `n.cores > 1` and `type = NULL`, then `type` will be set to "FORK" when running on unix, and set to "PSOCK" otherwise.
NOTE: The R_LIBS_USER environment variable may need to be set to the path where the SUITOR R package was installed, especially if it was not installed in the default location. Examples of setting the R_LIBS_USER environment variable on unix and windows are below.
(unix) export R_LIBS_USER=/path/to/SUITOR/package
(windows) set R_LIBS_USER=/path/to/SUITOR/package

Value

A list containing the objects:

- `H`: Matrix of activities (exposures)
- `W`: Matrix of signatures

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

See Also

`suitor`

Examples

```
data(SimData, package="SUITOR")

suitor_extract_WH(SimData, 2)
```

Index