Package 'SUITOR'

May 20, 2020

Title Selecting the number of mutational signatures through cross-validation

Version 0.0.3

Date 2020-05-20

Author Donghyuk Lee <donghyuk.lee@nih.gov> and Bin Zhu <bin.zhu@nih.gov>

Maintainer Bill Wheeler <wheelerb@imsweb.com>

Description Selecting the number of mutational signatures through cross-validation

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

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 <donghyuk.lee@nih.gov> and Bin Zhu <bin.zhu@nih.gov>

data *Data for examples*

Description

Example input data and results

Details

Contains an example input data object and an example matrix of results.

See Also

suitor

Examples

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

# Display a subset of data objects
data[1:5, 1:5]
results[1:10, ]
```

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 <donghyuk.lee@nih.gov> and Bin Zhu <bin.zhu@nih.gov>

See Also

`plotErrors`

Examples

```
data(data, package="SUITOR")
ret <- getSummary(results, 30)
ret$summary
ret$rank
```

`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`.

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 <donghyuk.lee@nih.gov> and Bin Zhu <bin.zhu@nih.gov>

Examples

```
data(data, package="SUITOR")
s <- getSummary(results, 30)
plotErrors(s$summary)
```

`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 non-Windows only)	1
n.seeds	Number of seeds (starting points)	30
plot	0 or 1 to produce an error plot	1
print	0-3 to print info (0=no printing)	1
seeds	Vector of seeds (takes precendence over n.seeds)	NULL
kfold.vec	Vector of values in 1:k.fold when running on a cluster	NULL

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. 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 <donghyuk.lee@nih.gov> and Bin Zhu <bin.zhu@nih.gov>

See Also

`getSummary`, `plotErrors`

Examples

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

# Options to run quick example
op  <- list(max.rank=1, seeds=123, k.fold=2, plot=0)
ret <- suitor(data, op=op)
ret$all.results
```

Index