Package ‘SUITOR’

January 10, 2024

Title  Selecting the number of mutational signatures through cross-validation

Version  1.4.0

Date  2022-05-10

Description  An unsupervised cross-validation method to select the optimal number of mutational signatures. A data set of mutational counts is split into training and validation data. Signatures are estimated in the training data and then used to predict the mutations in the validation data.

Imports  stats, utils, graphics, ggplot2, BiocParallel

Depends  R (>= 4.2.0)

License  GPL-2

biocViews  Genetics, Software, SomaticMutation

Suggests  devtools, MutationalPatterns, RUnit, BiocManager, BiocGenerics, BiocStyle, knitr, rmarkdown

NeedsCompilation  yes

BugReports  https://github.com/wheelerb/SUITOR/issues

VignetteBuilder  knitr

git_url  https://git.bioconductor.org/packages/SUITOR

git_branch  RELEASE_3_18

git_last_commit  f2a596a

git_last_commit_date  2023-10-24

Repository  Bioconductor 3.18

Date/Publication  2024-01-09

Author  DongHyuk Lee [aut], Bin Zhu [aut], Bill Wheeler [cre]

Maintainer  Bill Wheeler <wheelerb@imsweb.com>
**R topics documented:**

- SUITOR-package .................................................. 2
- getSummary ......................................................... 3
- plotData ............................................................... 4
- plotErrors ............................................................ 4
- results ................................................................. 5
- SimData ............................................................... 5
- suitor ................................................................. 6
- suitorExtractWH ................................................... 7

**Description**

To select the number of mutational signatures through cross-validation.

**Details**

SUITOR (Selecting the number of mutational signatures through cross-validation), an unsupervised cross-validation method that requires little assumptions and no numerical approximations to select the optimal number of signatures without overfitting the data. The full dataset of mutation counts is split into a training set and a validation set; for a given number of signatures, these signatures are estimated in the training set and then they are used to predict the mutations in the validation set. Multiple candidate numbers of signatures are considered; and the number of signatures which predicts most closely the mutations in the validation set is selected.

The two main functions in this package are `suitor` and `suitorExtractWH`.

**Author(s)**

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

**References**

**getDescription**

*Compute summary results*

**Description**

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

**Usage**

```r
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 $\sqrt{\frac{\text{fold1} + \ldots + \text{foldK}}{\text{nrow(data)}*\text{ncol(data)}}}$

**Author(s)**

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

**See Also**

`plotErrors`

**Examples**

```r
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 columns 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>
results

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

results    suitors return object

Description
    An object returned from the suitors function for examples

See Also
    suitors

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
    suitors

Examples
    data(SimData, package="SUITOR")
    # Display a subset of data objects
    SimData[1:5, 1:5]
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:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Default Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>em.eps</td>
<td>EM algorithm stopping tolerance</td>
<td>1e-5</td>
</tr>
<tr>
<td>get.summary</td>
<td>0 or 1 to create summary results</td>
<td>1</td>
</tr>
<tr>
<td>k.fold</td>
<td>Number of folds</td>
<td>10</td>
</tr>
<tr>
<td>max.iter</td>
<td>Maximum number of iterations in EM algorithm</td>
<td>2000</td>
</tr>
<tr>
<td>max.rank</td>
<td>Maximum rank</td>
<td>10</td>
</tr>
<tr>
<td>min.rank</td>
<td>Minimum rank</td>
<td>1</td>
</tr>
<tr>
<td>min.value</td>
<td>Minimum value of matrix before factorizing</td>
<td>1e-4</td>
</tr>
<tr>
<td>BPPARAM</td>
<td>See BiocParallelParam</td>
<td>NULL</td>
</tr>
<tr>
<td>n.starts</td>
<td>Number of starting points</td>
<td>30</td>
</tr>
<tr>
<td>plot</td>
<td>0 or 1 to produce an error plot</td>
<td>1</td>
</tr>
<tr>
<td>print</td>
<td>0 or 1 to print info</td>
<td>1</td>
</tr>
<tr>
<td>kfold.vec</td>
<td>Vector of values in 1:k.fold when running on a cluster</td>
<td>NULL</td>
</tr>
</tbody>
</table>

Parallel computing

The BiocParallel package is used for parallel computing. If BPPARAM = NULL, then BPPARAM will be set to SerialParam.

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 n.starts to 1.
suitorExtractWH

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 $\sqrt{\frac{\sum\text{fold1} + ... + \sum\text{foldK}}{\text{nrow(data)}*\text{ncol(data)}}}$

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)

suitorExtractWH

Description

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

Usage

suitorExtractWH(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`:
### Parallel computing

The BiocParallel package is used for parallel computing. If BPPARAM = NULL, then BPPARAM will be set to SerialParam.

### 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

```r
data(SimData, package="SUITOR")
suitorExtractWH(SimData, 2)
```
Index

* NMF, cross-validation, mutational signatures
  getSummary, 3
  plotErrors, 4
  suitor, 6
  suitorExtractWH, 7

* data
  plotData, 4
  results, 5
  SimData, 5

* package
  SUITOR-package, 2

BiocParallelParam, 6, 8

getSummary, 3, 4, 7

plotData, 4
plotErrors, 3, 4, 7

results, 5

SerialParam, 6, 8
SimData, 5
SUITOR (SUITOR-package), 2
suitor, 2–5, 6, 8
SUITOR-package, 2
suitorExtractWH, 2, 7