Package ‘MAI’

April 12, 2022

Type Package

Title Mechanism-Aware Imputation

Version 1.0.0

Description A two-step approach to imputing missing data in metabolomics.
Step 1 uses a random forest classifier to classify missing values as
either Missing Completely at Random/Missing At Random (MCAR/MAR) or Missing
Not At Random (MNAR). MCAR/MAR are combined because it is often difficult to
distinguish these two missing types in metabolomics data. Step 2 imputes the
missing values based on the classified missing mechanisms, using the
appropriate imputation algorithms. Imputation algorithms tested and
available for MCAR/MAR include Bayesian Principal Component Analysis (BPCA),
Multiple Imputation No-Skip K-Nearest Neighbors (Multi_nsKNN), and
Random Forest. Imputation algorithms tested and available for MNAR include
nsKNN and a single imputation approach for imputation of metabolites where
left-censoring is present.

License GPL-3

Encoding UTF-8

Imports caret, parallel, doParallel, foreach, e1071, future.apply,
future, missForest, pcaMethods, tidyverse, stats, utils,
methods, SummarizedExperiment, S4Vectors

biocViews Software, Metabolomics, StatisticalMethod, Classification

Suggests knitr, rmarkdown, BiocStyle, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

URL https://github.com/KechrisLab/MAI

BugReports https://github.com/KechrisLab/MAI/issues

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

git_branch RELEASE_3_14

git_last_commit 0953c88

git_last_commit_date 2021-10-26

Date/Publication 2022-04-12

Author Jonathan Dekermanjian [aut, cre],
Elin Shaddox [aut],
Debmalya Nandy [aut],
Debashis Ghosh [aut],
Katerina Kechris [aut]

Maintainer Jonathan Dekermanjian <Jonathan.Dekermanjian@CUAnschutz.edu>

R topics documented:

`MAI` *Mechanism-Aware Imputation*

Description

A two-step approach to imputing missing data in metabolomics. Step 1 uses a random forest classifier to classify missing values as either Missing Completely at Random/Missing At Random (MCAR/MAR) or Missing Not At Random (MNAR). MCAR/MAR are combined because it is often difficult to distinguish these two missing types in metabolomics data. Step 2 imputes the missing values based on the classified missing mechanisms, using the appropriate imputation algorithms. Imputation algorithms tested and available for MCAR/MAR include Bayesian Principal Component Analysis (BPCA), Multiple Imputation No-Skip K-Nearest Neighbors (Multi_nsKNN), and Random Forest. Imputation algorithms tested and available for MNAR include nsKNN and a single imputation approach for imputation of metabolites where left-censoring is present.

Usage

```
MAI(data_miss,
    MCAR_algorithm = c("BPCA", "Multi_nsKNN", "random_forest"),
    MNAR_algorithm = c("nsKNN", "Single"),
    n_cores = 1,
    assay_ix = 1,
    verbose = TRUE
    )
```

Arguments

`data_miss`	A matrix or dataframe, or a SummarizedExperiment containing missing values designated by "NA" to impute
`MCAR_algorithm`	The imputation algorithm you wish to use to impute MCAR predicted missing values. possible algorithms c("BPCA", "Multi_nsKNN", "random_forest")

`MNAR_algorithm`	The imputation algorithm you wish to use to impute MNAR predicted missing values. possible algorithms c("Single", "nsKNN")
`n_cores`	The number of cores you want to utilize. Default is 1 core. To use all cores specify n_cores = -1.
`assay_ix`	If data is a Summarized Experiment then this argument defines the index of the assay to impute. Default is set to the first assay.
`verbose`	A toggle to suppress console output. Default is TRUE

Value

When matrix or dataframe returns a list containing the following:

`Imputed Data`	Returns dataframes of MAI imputation
`Estimated Parameters`	Returns the estimated α, β, and γ parameters that define the missingness pattern in the data set. These parameters estimate the ratio of MCAR/MAR to MNAR in the data. The parameters α and β separate high, medium, and low average abundance metabolites, while the parameter γ is used to impose missingness in the medium and low abundance metabolites. A smaller α corresponds to more MCAR/MAR being present, while larger β and γ values imply more MNAR values being present. The returned estimated parameters are then used to impose known missingness in the complete subset of the input data. Subsequently, a random forest classifier is trained to classify the known missingness in the complete subset of the input data. Once the classifier is established it is applied to the unknown missingness of the full input data to predict the missingness. Finally, the missing values are imputed using a specific algorithm, chosen by the user, according to the predicted missingness mechanism.

When a Summarized Experiment returns:

`Imputed Assay`	Returns the imputed data in the specified assay based on the assay_ix assigned
`Estimated Parameters`	Returns estimated parameters in the metadata of the Summarized Experiment as a list

Examples

```
data(untargeted_LCMS_data)
MAI(data_miss=untargeted_LCMS_data,
    MCAR_algorithm = "BPCA",
    MNAR_algorithm="Single",
    n_cores = 1,
    assay_ix = 1,
    verbose = TRUE)
```

`untargeted_LCMS_data` *Example data set containing missing values*

Description

This data set is randomly generated. We impose 30 percent missing values using the Mixed missingness algorithm developed by Styczynski et al. Where the parameters alpha, beta, and gamma were chosen to be 30, 70, and 40 percent, respectively.

References

Lee JY, Styczynski MP. NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data. *Metabolomics*. 2018;14(12):153.

Index