Package ‘nebula’

July 6, 2023

Type Package

Title Negative Binomial Mixed Models Using Large-Sample Approximation for Differential Expression Analysis of ScRNA-Seq Data

Version 1.4.2

Date 2023-07-03

Author Liang He

Maintainer Liang He <hyx520101@gmail.com>

Description A fast negative binomial mixed model for conducting association analysis of multi-subject single-cell data. It can be used for identifying marker genes, differential expression and co-expression analyses. The model includes subject-level random effects to account for the hierarchical structure in multi-subject single-cell data. See He et al. (2021) <doi:10.1038/s42003-021-02146-6>.

License GPL-3

Encoding UTF-8

LazyData true

Imports Rcpp (>= 1.0.7), nloptr, stats, Matrix, methods, RSpectra, Rfast, trust, doFuture (>= 0.12.2), future (>= 1.32.0), foreach (>= 1.5.2), doRNG (>= 1.8.6)

LinkingTo Rcpp, RcppEigen

Depends R (>= 4.1)

RoxygenNote 7.2.3

Suggests knitr, utils, rmarkdown

VignetteBuilder knitr

NeedsCompilation yes

Repository CRAN

Date/Publication 2023-07-05 23:20:07 UTC
Description

A fast negative binomial mixed model for conducting association analysis of multi-subject single-cell data. It can be used for identifying marker genes, differential expression and co-expression analyses. The model includes subject-level random effects to account for the hierarchical structure in multi-subject single-cell data. See He et al. (2021) <doi:10.1038/s42003-021-02146-6>.

Details

nebula is an R package for performing association analysis using a fast negative binomial mixed model for multi-subject single-cell data.

Author(s)

Liang He

Maintainer: Liang He <hyx520101@gmail.com>

References


Examples

library(nebula)
data(sample_data)
pred = model.matrix(~X1+X2+cc,data=sample_data$pred)
re = nebula(count=sample_data$count,id=sample_data$sid,pred=pred)
**Description**

Group cells according to subject IDs

**Usage**

```r
group_cell(count, id, pred = NULL, offset = NULL)
```

**Arguments**

- `count` A raw count matrix of the single-cell data. The rows are the genes, and the columns are the cells. The matrix can be a matrix object or a sparse dgCMatrix object.
- `id` A vector of subject IDs. The length should be the same as the number of columns of the count matrix.
- `pred` A design matrix of the predictors. The rows are the cells and the columns are the predictors. If not specified, an intercept column will be generated by default.
- `offset` A vector of the scaling factor. The values must be strictly positive. If not specified, a vector of all ones will be generated by default.

**Value**

- `count`: A reordered count matrix. If the cells are already grouped, the function returns NULL.
- `id`: A reordered subject ID vector.
- `pred`: A reordered design matrix of the predictors.
- `offset`: A reordered vector of the scaling factor.

**Examples**

```r
library(nebula)
data(sample_data)
pred = model.matrix(~X1+X2+cc, data=sample_data$pred)
df_order = group_cell(count=sample_data$count, id=sample_data$sid, pred=pred)
```
**nbresidual**

*Extract Pearson residuals from the results of NEBULA*

**Description**

Extract Pearson residuals from the results of NEBULA

**Usage**

```r
nbresidual(nebula, count, id, pred = NULL, offset = NULL, conditional = FALSE)
```

**Arguments**

- `nebula`: An object of the result obtained from running the function `nebula`.
- `count`: A raw count matrix of the single-cell data. The rows are the genes, and the columns are the cells. The matrix can be a matrix object or a sparse `dgCMatrix` object.
- `id`: A vector of subject IDs. The length should be the same as the number of columns of the count matrix.
- `pred`: A design matrix of the predictors. The rows are the cells and the columns are the predictors. If not specified, an intercept column will be generated by default.
- `offset`: A vector of the scaling factor. The values must be strictly positive. If not specified, a vector of all ones will be generated by default.
- `conditional`: A logical value. By default (FALSE), the function returns marginal Pearson residuals. If TRUE, the function will return conditional Pearson residuals.

**Value**

- `residuals`: A matrix of Pearson residuals. The number of columns is the number of cells in the count matrix. The rows correspond to gene IDs reported in the result from nebula.
- `gene`: Gene names corresponding to the row names of the count matrix.

**Examples**

```r
library(nebula)
data(sample_data)
pred = model.matrix(~X1+X2+cc, data=sample_data$pred)
re = nebula(count=sample_data$count, id=sample_data$sid, pred=pred)
resid = nbresidual(re, count=sample_data$count, id=sample_data$sid, pred=pred)
```
Association analysis of a multi-subject single-cell data set using a fast negative binomial mixed model

Description

Association analysis of a multi-subject single-cell data set using a fast negative binomial mixed model

Usage

```
nebula(
  count,  # A raw count matrix of the single-cell data. The rows are the genes, and the columns are the cells. The matrix can be a matrix object or a sparse dgCMatrix object.
  id,     # A vector of subject IDs. The length should be the same as the number of columns of the count matrix.
  pred = NULL,  # A design matrix of the predictors. The rows are the cells and the columns are the predictors. If not specified, an intercept column will be generated by default.
  offset = NULL,  # A vector of the scaling factor. The values must be strictly positive. If not specified, a vector of all ones will be generated by default.
  min = c(1e-04, 1e-04),  # Minimum values for the overdispersions parameters σ² and φ. Must be positive. The default is c(1e-4,1e-4).
  max = c(10, 1000),
  model = "NBGMM",  # The default is c(1e-4,1e-4).
  method = "LN",
  cutoff_cell = 20,
  kappa = 800,
  opt = "lbfgs",
  verbose = TRUE,
  cpc = 0.005,
  mincp = 5,
  covariance = FALSE,
  output_re = FALSE,
  reml = 0,
  ncore = 2,
  fmaxsize = Inf
)
```

Arguments

- `count`: A raw count matrix of the single-cell data. The rows are the genes, and the columns are the cells. The matrix can be a matrix object or a sparse dgCMatrix object.
- `id`: A vector of subject IDs. The length should be the same as the number of columns of the count matrix.
- `pred`: A design matrix of the predictors. The rows are the cells and the columns are the predictors. If not specified, an intercept column will be generated by default.
- `offset`: A vector of the scaling factor. The values must be strictly positive. If not specified, a vector of all ones will be generated by default.
- `min`: Minimum values for the overdispersions parameters σ² and φ. Must be positive. The default is c(1e-4,1e-4).
Maximum values for the overdispersions parameters \( \sigma^2 \) and \( \phi \). Must be positive. The default is \( c(10,1000) \).

NBGMM', 'PMM' or 'NBLMM'. 'NBGMM' is for fitting a negative binomial gamma mixed model. 'PMM' is for fitting a Poisson gamma mixed model. 'NGLMM' is for fitting a negative binomial lognormal mixed model (the same model as that in the lme4 package). The default is 'NBGMM'.

The data will be refit using NEBULA-HL to estimate both overdispersions if the product of the cells per subject and the estimated cell-level overdispersion parameter \( \phi \) is smaller than cutoff_cell. The default is 20.

Please see the vignettes for more details. The default is 800.

'lbfgs' or 'trust'. Specifying the optimization algorithm used in NEBULA-LN. The default is 'lbfgs'. If it is 'trust', a trust region algorithm based on the Hessian matrix will be used for optimization.

An optional logical scalar indicating whether to print additional messages. Default is FALSE.

A non-negative threshold for filtering low-expression genes. Genes with counts per cell smaller than the specified value will not be analyzed.

A positive integer threshold for filtering low-expression genes. A gene will not be analyzed if its number of cells that have a non-zero count is smaller than the specified value.

If TRUE, nebula will output the covariance matrix for the estimated log(FC), which can be used for testing contrasts.

If TRUE, nebula will output the subject-level random effects. Only effective for model='NBGMM' or 'NBLMM'.

Either 0 (default) or 1. If it is one, REML will be used to estimate the overdispersions.

The number of cores used for parallel computing.

The maximum allowed total size (in bytes) of global variables (future.globals.maxSize) when using parallel computing.

The estimated coefficient, standard error and p-value for each predictor.

The estimated cell-level and subject-level overdispersions \( \sigma^2 \) and \( \phi^{-1} \).

More information about the convergence of the algorithm for each gene. A value of -20 or lower indicates a potential failure of the convergence. A value of one indicates that the convergence is reached due to a sufficiently small improvement of the function value. A value of -10 indicates that the convergence is reached because the gradients are close to zero (i.e., the critical point) and no improvement of the function value can be found.

The algorithm used for analyzing the gene. More information can be found in the vignettes.

The covariance matrix for the estimated log(FC).

The subject-level random effects.
sample_data

Examples

library(nebula)
data(sample_data)
pred = model.matrix(~X1+X2+cc, data=sample_data$pred)
re = nebula(count=sample_data$count, id=sample_data$sid, pred=pred)

---

sample_data  An example data set for testing nebula

Description

A dataset containing a count matrix, subject IDs, a data frame of predictors and scaling factors.

Usage

sample_data

Format

A list of four objects:

- **count** A raw count matrix
- **sid** A vector of subject IDs
- **pred** A data frame of three predictors
- **offset** A vector of scaling factors
Index

* datasets
  sample_data, 7

* package
  nebula-package, 2

  group_cell, 3

nbresidual, 4
nebula, 5
nebula-package, 2

sample_data, 7