Package 'dbnorm'

December 3, 2019

Type Package

Title Drift across-batch normalization and visualization

Version 0.1.0

Author Nasim Bararpour

Maintainer Nasim Bararpour <nasimbararpour@gmail.com>

Description The dbnorm contains several functions applicable in a large-scale metabolomics as well as other high throughput experiments. Notably, it includes distinct functions for pre-processing of data and estimation of missing values, conventional functions for batch effect correction based on statistical models, as well as functions using advanced statistical tools to generate several diagnosis plots to inform users about their data structure. Several statistical models are included in the dbnorm such as two-stage procedure model as described by Giordan (2013) or empirical Bayes methods in two setting of parametric and non-parametric as described by Johnson et al.(2007), in order to give users the flexibility to choose one of those models which better fits to their data. By including advanced statistical tools, the dbnorm package allows user to inspect the structure and quality of multidimensional datasets both in macroscopic and microscopic scales, at the level of sample sets and metabolic features respectively.

License LGPL(>= 2)

Encoding UTF-8

Imports ber, ggfortify, factoextra, ggplot2, NormalizeMets, sva,MASS, base(>= 3.5.1)

Suggests limma,installr,impute,Biobase,pcaMethods,tibble,knitr, rmarkdown, processx, backports, fs, Rcpp,BiocParallel,genefilter,stats

Remotes r-lib/testthat

biocViews Software

LazyData yes

RoxygenNote 6.1.1

R topics documented:

`ACDdbnorm` *Adjusted coefficient of determination for a data normalized for across batch signal drift*

Description

This function gives a quick notification about the performance of the statistical models implemented in the `dbnorm` package such as *Giordan (2013)* and/or empirical Bayes methods in two setting of parametric and non-parametric as described by *Johnson et al.(2007)* and in *sva* package by *Leek et al.(2012)*. It calculates and plots adjusted coefficient of determination or *Adjusted R-Squared* for each variable estimated in a regression model for its dependency to the batch level in the raw data and treated data via either of those models. Immediately, a score calculated based on the maximum variability explained by the batch level presents the performance of applied models. This score notifies the consistency of a model performance for all detected features (variables), facilitating quick comparison of the models for selecting one of those models, which is more appropriate to the data structure. This function is suggested for less than 2000 features (variables) to keep maximum computational speed.

Usage

```
ACDdbnorm(m)
```

Arguments

`m`	A data frame in which rows define the independent experiments (samples) and columns the features (variables), with the batch levels in the first column.

Details

Zero and NA values are not allowed. Optionally missing value can be imputed by `emvf` or `emvd`, functions implemented in `dbnorm` package. Input data must be normalized prior.

Value

Several graphs compiled into a **PDF** file which are a *correlation* plot for each of applied models, a grouped *barplot* presenting the maximum variability associated with batch levels in the raw and the corrected datasets.
Files saved as **csv** in the working directory are a dataset corrected via either of applied models. Also, a two column matrix for Adjusted R-Square for raw and corrected datasets and a table summarizing the score values presented in *barplot*.

References

M.Giordan (2013) *https://link.springer.com/article/10.1007/s12561-013-9081-1*
Johnson et al. (2007) *http://www.ncbi.nlm.nih.gov/pubmed/16632515*
Leek et al. (2012) *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3307112/*

Examples

```
batch<- rep(gl(3,7,labels = c(1:3)),1)
y<- matrix(rnorm(2100),nrow=21)
m<-data.frame(batch,y)
```

`dbnormBer` *drift across batch normalization via ber- model and visualization base on unsupervised learning algorithm and regression analysis*

Description

This function allows you to adjust the data for across batch signal drift or batch effect using two-stage procedure approach as described by *M.Giordan (2013).dbnormBer* includes advanced statistical tools to inspect the structure and quality of high throughput experiment both in macroscopic and microscopic scales at the level of sample sets and metabolic feature, respectively. Notably, using this function users applied unsupervised learning algorithm to visualize the most variance explained by the two first components in the different set of samples analyzed in the entire experiment in the raw and corrected data. In parallel, linear association of feature (variable) and batch level has been estimated and visualized by a correlation plot. In fact, estimated *Adjusted- R squared* is considered to define the level of dependency of feature (variable) to the batch level in the raw and corrected datasets. Besides, for quick notification about the performance of the applied model a maximum variability detected in either of datasets is reported as a score. This score notify the consistency of model performance for all detected features (variables).

Usage

```
dbnormBer(f)
```

Arguments

`f`	A data frame in which rows define the independent experiments (samples) and columns the features (variables), with the batch levels in the first column.

Details

Zero and NA values are not allowed. Optionally missing value can be imputed by functions implemented in `dbnorm` package. Input must be normalized and transformed prior.

Value

Several graphs compiled into a **PDF** file are a *PCA* score plot, *Scree* plot and a *correlation* plot estimated for raw and corrected data. Also, the *RLA* plot for each dataset visualized in the **Viewer** panel in the **rstudio** console.
Files saved as **csv** in the working directory are a dataset corrected by the applied model. Also, a two column matrix for Adjusted R-Square raw and corrected dataset and a table summarizing the maximum score.

References

M.Giordan (2013) *https://link.springer.com/article/10.1007/s12561-013-9081-1*

Examples

```
batch<- rep(gl(3,7,labels = c(1:3)),1)
y<- matrix(rnorm(2100),nrow=21)
f<-data.frame(batch,y)
```

`dbnormNPcom` *drift across batch normalization via nonParametric- ComBat model and visualization base on unsupervised learning algorithm and regression analysis*

Description

This function allows you to adjust the data for across batch signal drift or batch effect non-parametric Empirical Bayes approach as described by *Johnson et al.(2007)* and in *sva* package as explained by *Leek et al.(2012)*. emphdbnormNPcom includes advanced statistical tools to inspect the structure and quality of high throughput experiment both in macroscopic and microscopic scales at the level of sample sets and metabolic feature, respectively. Notably, using this function users applied unsupervised learning algorithm to visualize the most variance explained by the two first components in the different set of samples analyzed in the entire experiment in the raw and corrected data. In parallel, linear association of feature (variable) and batch level has been estimated and visualized by a correlation plot. In fact, estimated *Adjusted- R squared* is considered to define the level of dependency of feature (variable) to the batch level in the raw and corrected datasets. Besides, for quick notification about the performance of the applied model a maximum variability detected in either of datasets is reported as a score. This score notify the consistency of model performance for all detected features (variables).

Usage

```
dbnormNPcom(f)
```

Arguments

`f`	A data frame in which rows define the independent experiments (samples) and columns the features (variables), with the batch in the first column.

Details

Zero and NA values are not allowed. Optionally missing value can be imputed by `emvf` or `emvd`, functions implemented in `dbnorm` package. Input must be normalized and transformed prior.

Value

Several graphs compiled into a **PDF** file are a *PCA* score plot, *Scree* plot and a *correlation* plot for raw and corrected dataset. Also, the *RLA* plot for each dataset visualized in the **Viewer** panel in the **rstudio** console.
Files saved as **csv** in the working directory are a dataset corrected by the applied model. Also, a two column matrix for Adjusted R-Square raw and corrected dataset and a table summarizing the maximum score.

References

Johnson et al.(2007) *http://www.ncbi.nlm.nih.gov/pubmed/16632515*
Leek et al. (2012) *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3307112/*

Examples

```
batch<- rep(gl(3,7,labels = c(1:3)),1)
y<- matrix(rnorm(2100),nrow=21)
f<-data.frame(batch,y)
```

`dbnormPcom` *drift across batch normalization via Parametric- ComBat model and visualization base on unsupervised learning algorithm and regression analysis*

Description

This function allows you to adjust the data for across batch signal drift or batch effect parametric Empirical Bayes approach as described by *Johnson et al.(2007)* and in *sva* package as explained by *Leek et al.(2012)*. emphdbnormPcom includes advanced statistical tools to inspect the structure and quality of high throughput experiment both in macroscopic and microscopic scales at the level of sample sets and metabolic feature, respectively. Notably, using this function users applied unsupervised learning algorithm to visualize the most variance explained by the two first components in the different set of samples analyzed in the entire experiment in the raw and corrected data. In parallel, linear association of feature (variable) and batch level has been estimated and visualized by a correlation plot. In fact, estimated *Adjusted- R squared* is considered to define the level of dependency of feature (variable) to the batch level in the raw and corrected datasets. Besides, for quick notification about the performance of the applied model a maximum variability detected in either of datasets is reported as a score. This score notify the consistency of model performance for all detected features (variables).

Usage

```
dbnormPcom(f)
```

Arguments

f	A data frame in which rows define the independent experiments (samples) and columns the features (variables), with the batch levels in the first column.

Details

Zero and NA values are not allowed. Optionally missing value can be imputed by functions implemented in dbnorm package. Input must be normalized and transformed prior.

Value

Several graphs compiled into a **PDF** file are a *PCA* score plot, *Scree* plot and a *correlation* plot for raw and corrected data. Also, the *RLA* plots for each dataset visualized in the **Viewer** panel in the **rstudio** console.
Files saved as **csv** in the working directory are a dataset corrected by the applied model. Also, a two column matrix for Adjusted R-Square raw and corrected dataset and a table summarizing the maximum score.

References

Johnson et al. (2007) *http://www.ncbi.nlm.nih.gov/pubmed/16632515*
Leek et al. (2012) *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3307112/*

Examples

```
batch<- rep(gl(3,7,labels = c(1:3)),1)
y<- matrix(rnorm(2100),nrow=21)
f<-data.frame(batch,y)
```

---

emvd *Estimation of missing value data-based*

Description

It returns to a matrix of data in which missing values are estimated by the lowest detected value in the entire experiment. By this function, all NA values are replaced by Zero values, that of being ultimately replaced by the lowest value detected in the experiment. Ultimately, data matrix is transposed to restore original structure.

Usage

```
emvd(m)
```

Arguments

m	An array or a matrix

Details

empty entries are not allowed

Value

A matrix with estimated missing value.

Examples

```
m<- data.frame(x1=c(50,NA,6,10,30),x2=c(2,8,NA,15,0))
```

`emvf` *Estimation of missing value feature-based*

Description

It returns to a matrix of data in which missing values (Zero and/or NA values) are estimated. By this function, all Zero values are first replaced by NA values, which are then replaced by the lowest detected value on the column margin.

Usage

```
emvf(m)
```

Arguments

`m`	An array or a matrix

Details

empty entries are not allowed

Value

A matrix with estimated missing value.

Examples

```
m<- data.frame(x1=c(50,NA,6,10,30),x2=c(2,8,NA,15,0))
```

`profplotber` *Visualization of analytical heterogeneity on the profile of features (variables) in ber- corrected data*

Description

profplotber allows you to adjust the data for batch effect using two-stage procedure approach as describes by *Giordan (2013)* and informs you about the presence of across batch signal drift or batch effect in the treated data determined by the shifted probability density function plots (*pdf* plots) of features (variables) detected in an experiment.

Usage

```
profplotber(m)
```

Arguments

m	A data frame in which rows define the independent experiments (samples) and columns the features (variables), with the batch in the first column.

Details

Zero and NA values are not allowed. Optionally missing value can be imputed by functions implemented in dbnorm package. Input must be normalized and transformed prior.

Value

Original and adjusted datasets in **csv** format together with the series of profile plot for the variables(features) in the sample sets analyzed in the entire experiment provided by the *Scatter* plot,*Violin* plot and *pdf* plot compiled into **PDF** file.

References

M.Giordan (2013) *https://link.springer.com/article/10.1007/s12561-013-9081-1*

Examples

```
batch<- rep(gl(3,7,labels = c(1:3),1))
y<- matrix(rnorm(2100),nrow=21)
m<-data.frame(batch,y)
```

---

profplotnpcom — *Visualization of analytical heterogeneity on the profile of features (variables) in Non-Parametric ComBat corrected data*

Description

Visualization of analytical heterogeneity on the profile of features (variables) in Non-Parametric ComBat corrected data

Usage

```
profplotnpcom(m)
```

Arguments

m	A data frame in which rows define the independent experiments (samples) and columns the features (variables), with the batch in the first column.

Details

Zero and NA values are not allowed. Optionally missing value can be imputed by functions implemented in dbnorm package. Input must be normalized and transformed prior.

Value

Original and adjusted datasets in **csv** format together with the series of profile plot of the features (variables) in the sample sets provided by *Scatter* plot,*Violin* plot and *pdf* plot compiled into a **PDF** file.

References

Johnson et al. (2007) *http://www.ncbi.nlm.nih.gov/pubmed/16632515*
Leek et al. (2012) *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3307112/*

Examples

```
batch<- rep(gl(3,7,labels = c(1:3),1))
y<- matrix(rnorm(2100),nrow=21)
m<-data.frame(batch,y)
```

`profplotpcom` *Visualization of analytical heterogeneity on the profile of features (variables) in ComBat-Parametric -corrected data*

Description

profplotpcom allows you to adjust the data for batch effect using Parametric Empirical Bayes approach as described by *Johnson et al.(2007)* and via *sva* package as explained by *Leek et al.(2012)*, and informs you about the presence of across batch signal drift or batch effect in the treated data, determined by the shifted probability density function plots (*pdf* plots) of features (variables) detected in an experiment.

Usage

```
profplotpcom(m)
```

Arguments

`m`	A data frame in which rows define the independent experiments (samples) and columns the features (variables), with the batch in the first column.

Details

Zero and NA values are not allowed. Optionally missing value can be imputed functions implemented in `dbnorm` package. Input must be normalized and transformed prior.

Value

Original and adjusted datasets in **csv** format together with the series of profile plot for the features(variables) in the sample sets provided by the *Scatter* plot,*Violin* plot and *pdf* plot compiled into a **PDF** file.

References

Johnson et al. (2007) *http://www.ncbi.nlm.nih.gov/pubmed/16632515*
Leek et al. (2012) *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3307112/*

Examples

```
batch<- rep(gl(3,7,labels = c(1:3),1))
y<- matrix(rnorm(2100),nrow=21)
m<-data.frame(batch,y)
```

`profplotraw` *Visualization of analytical heterogeneity on the profile of features (variables) in raw data*

Description

This function informs you about the presence of across batch signal drift or batch effect in the raw data determined by the shifted probability density function plots (*pdf* plots) of features (variables) detected in an experiment.

Usage

```
profplotraw(m)
```

Arguments

`m`	A data frame in which rows define the independent experiments (samples) and columns the features (variables), with the batch level in the first column.

Details

Zero and NA values are not allowed. Optionally missing value can be imputed by functions implemented in `dbnorm` package. Input must be normalized and transformed prior.

Value

Original dataset in **csv** format together with the series of profile plot for the features (variables) in the sample sets analyzed in the entire experiment provided by the *Scatter* plot,*Violin* plot and *pdf* plot compiled into **PDF** file.

Examples

```
batch<- rep(gl(3,7,labels = c(1:3),1))
y<- matrix(rnorm(2100),nrow=21)
m<-data.frame(batch,y)
```

Visdbnorm *Visualization of signal drift across batch normalization*

Description

This function performs batch effect adjustment via three statistical models implemented in the dbnorm, namely two-stage procedure as described by *Giordan (2013)* and/or empirical Bayes methods in two setting of parametric and non-parametric as described by *Johnson et al.(2007)* and in *sva* package by *Leek et al.(2012)*. Meanwhile, the graphical inferences in the context of unsupervised learning algorithms create visual inspection to inform users about the spatial separation of the sample sets analyzed in the different analytical runs alongside the distribution of the features (variables) in the raw and treated datasets. This function is suggested for less than 2000 features (variables) to speed up the computational process.

Usage

```
Visdbnorm(f)
```

Arguments

f	A data frame in which rows define the independent experiments (samples) and columns the features (variables), with the batch in the first column.

Details

Zero and NA values are not allowed. Optionally missing value can be imputed by emvf or emvd, functions implemented in the dbnorm package. Input data must be normalized prior.

Value

Three datasets, adjusted by either of applied statistical algorithms prepared in **csv** and together with series of plot such as *PCA* plot and *Scree plot* compiled into a **PDF** file are saved in the working directory. *RLA* plots are represented in the **Viewer** panel of **rstudio**.

References

M.Giordan (2013) *https://link.springer.com/article/10.1007/s12561-013-9081-1*
Johnson et al. (2007) *http://www.ncbi.nlm.nih.gov/pubmed/16632515*
Leek et al. (2012) *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3307112/*

Examples

```
batch<- rep(gl(3,7,labels = c(1:3)),1)
y<- matrix(rnorm(2100),nrow=21)
f<-data.frame(batch,y)
```

Index