# Package 'dbnorm'

December 23, 2019

**Type** Package

**Title** Drift Across-Batches Normalization and Visualization

**Description** This package contains several functions applicable in a large-scale "Metabolomics" as well as other high throughput experiments. Notably, it includes distinct functions for preprocessing 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 (see "ber" , a package in R ) or empirical Bayes methods in two setting of parametric and non-parametric (see "sva", a package in bioconductor ), 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, in the dataset in which batch order considered in first column.

**Version** 0.1.0

**Maintainer** Nasim Bararpour <nasimbararpour@gmail.com>

**Encoding** UTF-8

**License** LGPL(>= 3)

**Imports** ber, ggfortify, factoextra, ggplot2, NormalizeMets, sva,base,graphics,grDevices,utils,MASS

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

**LazyData** yes

**RoxygenNote** 6.1.1

## R topics documented:


ACDdbnorm . . . . . 2
dbnormBer . . . . . 3
dbnormNPcom . . . . . 4
dbnormPcom . . . . . 5
emvd . . . . . 6
emvf . . . . . 7
profplotber . . . . . 8
profplotnpcom . . . . . 9
profplotpcom . . . . . 10
profplotraw . . . . . 11
Visdbnorm . . . . . 11

**Index** 13

ACDdbnorm *Adjusted Coefficient Of Determination for a data normalized for signal drift across batches*

## Description

This function gives a quick notification about the performance of the compiled statistical models namely two-stage regression procedure as described in M. Giordan 2013 and/or empirical Bayes methods in two setting of parametric and non-parametric as described in Johnson et al. 2007 ( see also "sva"), on accommodation of batch effect. Using this function users will estimate values of adjusted coefficient of determination (Adjusted R- Squared) which address the dependency of each feature (variable) to the batch order in each dataset. Immediately, a score calculated based on the maximum variability estimated by the regression analysis has been calculated and printed. This score notifies the consistency of a model performance for the detected features (variables), facilitating quick comparison of the models for selecting one of those models, which is more appropriate to the data structure.

## 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

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

```
## Not run:
batch<- rep(gl(5,10,labels = c(1:5)),1)
y<- matrix(rnorm(5000),nrow=50)
m<-data.frame(batch,y)
ACDdbnorm(m)
## End(Not run)
```

---

dbnormBer *Drift Across Batch Normalization via ber- model and visualization*

---

## Description

This function allows you to adjust the data for signal drift across multiple batches 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(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 the functions such as `emvf` and /or `emvd` implemented in the `'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

```
## Not run:
batch<- rep(gl(5,10,labels = c(1:5)),1)
y<- matrix(rnorm(5000),nrow=50)
m<-data.frame(batch,y)
dbnormBer(m)
## End(Not run)
```

---

dbnormNPcom *Drift Across Batch Normalization via Parametric- ComBat model and visualization*

---

## Description

This function allows you to adjust the data for signal drift across multiple batches or batch effect via non-parametric Empirical Bayes approach as described by Johnson et al. 2007 ( see also "sva" ). *dbnormNPcom* 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(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 the functions such as emvf and /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

```
## Not run:
batch<- rep(gl(5,10,labels = c(1:5)),1)
y<- matrix(rnorm(5000),nrow=50)
m<-data.frame(batch,y)
dbnormNPcom(m)
## End(Not run)
```

---

`dbnormPcom` *Drift Across Batch Normalization via Parametric- ComBat model and visualization*

---

## Description

This function allows you adjust the data for signal drift across multiple batches or batch effect via parametric Empirical Bayes approach as described by Johnson et al. 2007 ( see also "sva"). *dbnormPcom* 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(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 the functions such as `emvf` and /or `emvd` 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

```
## Not run:
batch<- rep(gl(5,10,labels = c(1:5)),1)
y<- matrix(rnorm(5000),nrow=50)
m<-data.frame(batch,y)
dbnormPcom(m)
## End(Not run)
```

---

`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.

### See Also

emvf, Visdbnorm, ACDdbnorm, profplotraw, profplotber, profplotpcom, profplotnpcom, dbnormBer, dbnormPcom, dbnormNPcom

### Examples

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

---

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

---

### Description

This function 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))
emvf(m)
```

profplotber *Profile Plot of Features (variables) in ber- corrected data*

## Description

This function allows you to adjust the data for batch effect using two-stage procedure approach as described by M. Giordan 2013 and informs users about the presence of batch effect or changes in the profile of detected features (variables) in the corrected data, determined by the shifted probability density function plots (*pdf* plots).

## 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 the functions such as emvf and/ or emvd implemented in the dbnorm. 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

```
## Not run:
batch<- rep(gl(5,10,labels = c(1:5),1))
y<- matrix(rnorm(5000),nrow=50)
m<-data.frame(batch,y)
profplotber(m)
## End(Not run)
```

`profplotnpcom` *Profile Plot of Features (variables) in corrected data via NonParametric ComBat*

## Description

This function allows users to adjust the data for batch effect based on Non-Parametric Empirical Bayes approach as described by Johnson et al. 2007 ( see also "sva"). *profplotnpcom* informs users about the presence of batch effect or changes in the profile of detected features (variables) in the corrected data, determined by the shifted probability density function plots (*pdf* plots).

## 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 the functions such as `emvf` and/ or `emvd` implemented in the `dbnorm`. 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

```
## Not run:
batch<- rep(gl(2,3,labels=(1:2)),2)
y<- matrix(rnorm(6000), nrow=12)
m<- data.frame (batch,y)
profplotnpcom(m)

## End(Not run)
```

`profplotpcom` *Profile Plot of Features (variables) in corrected data via Parametric ComBat*

## Description

This function allows users to adjust the data for batch effect using Parametric Empirical Bayes approach as described by Johnson et al. 2007 ( see also "sva"). *profplotpcom* informs users about the presence of batch effect or changes in the profile of detected features (variables) in the corrected data, determined by the shifted probability density function plots (*pdf* plots).

## 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 by the functions such as `emvf` and/ or `emvd` implemented in the `'dbnorm'`. 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

```
## Not run:
batch<- rep(gl(2,3,labels=(1:2)),2)
y<- matrix(rnorm(6000), nrow=12)
m<- data.frame (batch,y)
profplotpcom(m)

## End(Not run)
```

## profplotraw *Profile Plot 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 such as `emvf` or `emvd` Compiled in the `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

```
## Not run:
batch<- rep(gl(5,10,labels = c(1:5),1))
y<- matrix(rnorm(5000),nrow=50)
m<-data.frame(batch,y)
profplotraw(m)
## End(Not run)
```

## Visdbnorm *Visualization and normalization of signal drift across batches*

### Description

This function performs batch effect adjustment via three statistical models, namely two-stage procedure as described by M. Giordan 2013 and/or empirical Bayes methods in two setting of parametric and non-parametric as described by Johnson et al. 2007 ( see also "sva") . 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 sample sets and across multiple batches.

## 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 and /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

```
## Not run:
batch<- rep(gl(5,10,labels = c(1:5)),1)
y<- matrix(rnorm(5000),nrow=50)
f<-data.frame(batch,y)
Visdbnorm(f)
## End(Not run)
```

# Index


∗Topic **Adjusted**
ACDdbnorm, 2
∗Topic **ComBat-Parametric**
dbnormPcom, 5
∗Topic **ComBat**
dbnormNPcom, 4
profplotnpcom, 9
profplotpcom, 10
∗Topic **Missing**
emvd, 6
emvf, 7
∗Topic **NON-Parametric**
dbnormNPcom, 4
∗Topic **Non**
profplotnpcom, 9
∗Topic **Parametric**
profplotpcom, 10
∗Topic **R-squared**
ACDdbnorm, 2
∗Topic **Unsupervised**
dbnormBer, 3
∗Topic **Visualization**
Visdbnorm, 11
∗Topic **across**
Visdbnorm, 11
∗Topic **analysis**
dbnormBer, 3
dbnormNPcom, 4
dbnormPcom, 5
∗Topic **and**
dbnormBer, 3
dbnormNPcom, 4
dbnormPcom, 5
profplotber, 8
profplotpcom, 10
Visdbnorm, 11
∗Topic **batch**
Visdbnorm, 11
∗Topic **ber-model**
dbnormBer, 3
∗Topic **ber**
profplotber, 8
∗Topic **correction**
profplotber, 8
∗Topic **data**
dbnormBer, 3
dbnormNPcom, 4
dbnormPcom, 5
profplotraw, 11
∗Topic **estimation**
emvd, 6
emvf, 7
∗Topic **for**
dbnormBer, 3
dbnormNPcom, 4
dbnormPcom, 5
∗Topic **normalization**
Visdbnorm, 11
∗Topic **normalized**
dbnormBer, 3
dbnormNPcom, 4
dbnormPcom, 5
∗Topic **parametric**
profplotnpcom, 9
∗Topic **plot**
profplotber, 8
profplotnpcom, 9
profplotpcom, 10
profplotraw, 11
∗Topic **profile**
profplotber, 8
profplotnpcom, 9
profplotpcom, 10
profplotraw, 11
∗Topic **raw**
profplotraw, 11
∗Topic **regression**
dbnormBer, 3
dbnormNPcom, 4
dbnormPcom, 5
∗Topic **unsupervised**
dbnormNPcom, 4
dbnormPcom, 5
∗Topic **value**
emvd, 6
emvf, 7

∗Topic **via**
dbnormBer, 3
dbnormNPcom, 4
dbnormPcom, 5

ACDdbnorm, 2

dbnormBer, 3
dbnormNPcom, 4
dbnormPcom, 5

emvd, 6
emvf, 7

profplotber, 8
profplotnpcom, 9
profplotpcom, 10
profplotraw, 11

Visdbnorm, 11