Package 'kBET'

May 8, 2017

Type Package

Title k-nearest neighbour batch effect test

Version 0.1.0

Author Maren Büttner

Maintainer Maren Büttner <maren.buettner@helmholtz-muenchen.de>

Description This tool detects batch effects in high-dimensional data based on chi^2-test.

Imports FNN, RColorBrewer, ggplot2, cluster, stats

License GPL

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

Collate 'ExactMultinomialTest.R'
'addalpha.R'
'batch_sil.R'
'bisect.R'
'findVectors.R'
'kBET-utils.R'
'kBET.R'
'multinomial.test.R'
'pcRegression.R'

R topics documented:

addalpha *addalpha - add transparency info to colorset for plotting*

Description

addalpha adds transparency information to a vector in R color format

Usage

```
addalpha(colors, alpha = 1)
```

Arguments

colors	a hexadecimal string as commonly used in R
alpha	a transparency factor: 0 - completely invisible 1 - fully opaque

Value

a hexadecimal string of length colors

Examples

```
library(RColorBrewer)
set2 <- brewer.pal(5, 'Set2')
set2.light <- addalpha(set2, alpha=0.5)
```

batch_sil *batch_sil*

Description

Determine batch/bio effect using the silhouette coefficient (adopted from scone):

Usage

```
batch_sil(pca.data, batch, nPCs = 3)
```

Arguments

pca.data	a list as created by prcomp, batch_sil.R needs $x: the principal components (PCs, correctly: the rotated data)
batch	vector with the batch covariate (for each cell)
nPCs	the number of principal components to use (default: 3)

Value

The average silhouette width for all clusters. For batch effect, the smaller the better. For biological effect, the larger the better.

Examples

```
    ## Not run:
     pca.data <- prcomp(data, center=TRUE)
     batch.silhouette <- batch_sil(pca.data, batch)

## End(Not run)
```

`bisect` *bisect - a generic bisection function*

Description

provides recursive bisection algorithm for an arbitrary function It evaluates the function `foo` at the bounds and replaces one of the boundaries until a maximum is found or the interval becomes too small

Usage

```
bisect(foo, bounds, known = NULL, ..., tolx = 10, toly = 0.01)
```

Arguments

`foo`	a function mapping a one-dim argument to one-dim value
`bounds`	a vector of length 2 with real valued numbers (i.e. two arguments of `foo`)
`known`	tells for which of the arguments a value is known (defaults to NULL)
`...`	additional parameters for `foo`
`tolx`	break condition for argument (defaults to 10)
`toly`	break condition for value (defaults to 0.01)

`kBET` *kBET - k-nearest neighbour batch effect test*

Description

`kBET` runs a chi square test to evaluate the probability of a batch effect.

Usage

```
kBET(df, batch, k0 = NULL, knn = NULL, testSize = NULL, do.pca = TRUE,
  heuristic = FALSE, stats = 100, alpha = 0.05, addTest = FALSE,
  verbose = TRUE, plot = TRUE)
```

Arguments

df	dataset (rows: samples, columns: parameters)
batch	batch id for each cell or a data frame with both condition and replicates
k0	number of nearest neighbours to test on (neighbourhood size)
knn	a set of nearest neighbours for each cell (optional)
testSize	number of data points to test, (10 percent sample size default)
do.pca	perform a pca prior to knn search? (defaults to TRUE)
heuristic	compute an optimal neighbourhood size k
stats	to create a statistics on batch estimates, evaluate 'stats' subsets
alpha	significance level
addTest	perform an LRT-approximation to the multinomial test AND a multinomial exact test (if appropriate)
verbose	displays stages of current computation (defaults to TRUE)
plot	if stats > 10, then a boxplot of the resulting rejection rates is created

Value

list object

1. summary - a rejection rate for the data, an expected rejection rate for random labeling and the significance for the observed result
2. stats - extended test summary for every sample

Examples

```
batch.estimate <- kBET(data,batch)
```

---

pcRegression *pcRegression*

Description

pcRegression does a linear model fit of principal components and a batch (categorical) variable

Usage

```
pcRegression(pca.data, batch, tol = 1e-16)
```

Arguments

pca.data	a list as created by 'prcomp', pcRegression needs $x: the principal components (PCs, correctly: the rotated data) and $sdev: the standard deviations of the PCs)
batch	vector with the batch covariate (for each cell)
tol	truncation threshold for significance level, default: 1e-16

Value

List summarising principal component regression

- `maxVar` - the variance explained by principal component(s) that correlate(s) most with the batch effect
- `PmaxVar` - p-value (returned by linear model) for the respective principal components (related to `maxVar`)
- `R2Var` - sum over Var(PC_i)*r2(PC_i and batch) for all i
- `ExplainedVar` - explained variance for each PC
- `r2` - detailed results of correlation (R-Square) analysis

Index