# Package 'mdqc'

October 9, 2013

**Type** Package

**Title** Mahalanobis Distance Quality Control for microarrays

**Version** 1.22.0

**Date** 2008-02-27

**Author** Justin Harrington

**Maintainer** Gabriela Cohen-Freue <gcohen@mrl.ubc.ca>

**Description** MDQC is a multivariate quality assessment method for microarrays based on quality control (QC) reports. The Mahalanobis distance of an array's quality attributes is used to measure the similarity of the quality of that array against the quality of the other arrays. Then, arrays with unusually high distances can be flagged as potentially low-quality.

**Depends** R (>= 2.2.1), cluster, MASS

**License** LGPL (>= 2)

**biocViews** Microarray, QualityControl

## R topics documented:

mdqc-package *MDQC: Mahalanobis Distance Quality Control*

## Description

MDQC is a multivariate quality assessment method for microarrays based on quality control (QC) reports.

## Details

| | |
|---|---|
| Package: | mdqc |
| Type: | Package |
| Version: | 1.0.0 |
| Date: | 2007-12-12 |
| License: | GPL |

## Author(s)

Justin Harrington <harringt@stat.ubc.ca> and Gabriela V. Cohen Freue <gcohen@stat.ubc.ca>.

## References

Cohen Freue, G. V. and Hollander, Z. and Shen, E. and Zamar, R. H. and Balshaw, R. and Scherer, A. and McManus, B. and Keown, P. and McMaster, W. R. and Ng, R. T. (2007) 'MDQC: A New Quality Assessment Method for Microarrays Based on Quality Control Reports'. *Bioinformatics* **23**, 3162 – 3169.

## See Also

mdqc

allQC *QC report for MLL.B*

## Description

A subset of arrays from a large acute lymphoblastic leukemia (ALL) study

## Usage

```
data(allQC)
```

## Format

A data frame with 20 observations on the following 11 variables.

`Scale Factor` a numeric vector

`Percent Present` a numeric vector

`Average Background` a numeric vector

`Minimum Background` a numeric vector

`Maximum Background` a numeric vector

`BioB` a numeric vector

`BioC` a numeric vector

`BioD` a numeric vector

`CreX` a numeric vector

`AFFX-HSAC07/X00351.3'/5'` a numeric vector

`AFFX-HUMGAPDH/M33197.3'/5'` a numeric vector

## Details

Contains the QC report obtained using Bioconductor's simpleaffy package for a subset of arrays from a large acute lymphoblastic leukemia (ALL) study (*Ross et al., 2004*). The QC report in allQC has been generated using the following R commands:

```
    library("affy")

    ## Get the raw data (see help("MLL.B") for further details)
    library("ALLMLL")
    data(MLL.B)

    ## Generate the QC metrics
    library("simpleaffy")
    data.all <- MLL.B[,1:20]
    all.qc <- qc(data.all)

    ## Select relevant information
    allQC <- cbind(sfs(all.qc),percent.present(all.qc)/100,
                   avbg(all.qc),minbg(all.qc),maxbg(all.qc),
                   spikeInProbes(all.qc),ratios(all.qc)[,c(1,3)])

    ## Specify row and column names
    colnames(allQC) <- c("Scale Factor","Percent Present",
"Average Background", "Minimum Background", "Maximum Background",
"BioB", "BioC", "BioD", "CreX", "AFFX-HSAC07/X00351.3'/5'",
"AFFX-HUMGAPDH/M33197.3'/5'")
    rownames(allQC) <- 1:20
```

Versions 1.16.0 of 'affy', 1.2.2 of 'ALLMLL', and 2.14.05 of 'simpleaffy' was used. Part of this dataset has been also studied by *Bolstad et al. (2005)* and *Brettschneider et al. (2007).*

## Source

Ross, M. E. and Zhou, X. and Song, G. and Shurtleff, S. A. and Girtman, K. and Williams, W. K. and Liu, H. and Mahfouz, R. and Raimondi, S. C. and Lenny, N. and Patel, A. and Downing, J. R. (2003) 'Classification of pediatric acute lymphoblastic leukemia by gene expression profiling.' *Blood* **102**, 2951–9.

All CEL files are freely available in http://www.stjuderesearch.org/data/ALL3/rawFiles.html.

## References

Bolstad, B. M. and Collin, F. and Brettschneider, J. and Simpson, K. and Cope, L. and Irizarry R. A. and Speed T. P. (2005) 'Quality assessment of Affymetrix GeneChip data.' In Gentleman, R. and Carey, C. J. and Huber, W. and Irizarry, R. A. and Dudoit, S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* New York: Springer.

Brettschneider, J. and Collin, F. and Bolstad, B. M. and Speed, T. P. (2007) 'Quality assessment for short oligonucleotide arrays'. Forthcoming in *Technometrics (with Discussion)*.

## Examples

```
data(allQC)
```

---

mdqc *MDQC: Mahalanobis Distance Quality Control*

---

## Description

MDQC is a multivariate quality assessment method for microarrays based on quality control (QC) reports.

## Usage

```
mdqc(x, method=c("nogroups", "apriori", "global", "cluster", "loading"),
     groups=NULL, k=NULL, pc=NULL,
     robust=c("S-estimator","MCD", "MVE"), nsamp=10*nrow(x))
```

## Arguments

| | |
|---|---|
| x | a numeric matrix or data frame containing the quality measures (columns) for each array (rows). The number of rows must exceed the number of columns. |
| method | The Mahalanobis Distances (MDs) can be computed on all the quality measures in the QC report (this is the default method given by `method="nogroups"`), on the first k principal components resulting from a principal component analysis (PCA) of the QC report (`"global"`) or on subsets of quality measures in the QC report (`"apriori"`: groups defined by the user, `"cluster"`: groups resulting from a cluster analysis, or `"loading"`: groups resulting from a cluster analysis in the space of the loadings of a PCA). While the first two methods compute a |

| | |
|---|---|
| | single MD for each array, the last three compute one MD within each created group of quality measures. |
| groups | A list to specify the groups of quality measures when the "apriori" method is chosen. E.g. `groups = list(c(1,2), c(4,6))` puts column 1,2 as one group and 4,6 as a second. |
| k | An integer to specify the number of clusters (or groups) to be used in the cluster analysis when "cluster" or "loading" methods are chosen. |
| pc | An integer to specify the number of principal components analyzed from the PCA when "global" or "loading" methods are chosen. |
| robust | A robust multivariate location/spread estimator (choice of S-estimator, MCD or MVE). The default method uses S-estimators with a 25% breakdown point. |
| nsamp | The number of subsamples that the robust estimator should use. This defaults to 10 times the number of rows in the matrix. |

## Details

MDQC flags potentially low quality arrays based on the idea of outlier detection, that is, it flags those arrays whose quality attributes jointly depart from those of the bulk of the data.

This function computes a distance measure, the Mahalanobis Distance, to summarize the quality of each array. The use of this distance allows us to perform a multivariate analysis of the information in QC reports taking the correlation structure of the quality measures into account. In addition, by using robust estimators to identify the typical quality measures of good-quality arrays, the evaluation is not affected by the measures of outlying arrays.

MDQC can be based on all the quality measures simultaneously (using `method="nogroups"`), on subsets of them (using `method="apriori"`, `"cluster"`, or `"loading"`), or on a transformed space with a lower dimension (using `method="global"`).

In the "apriori" approach the user forms groups of quality measures on the basis of an a priori interpretation of them and according to the quality aspect they represent. The "cluster" and the "loading" methods are two data-driven methods to form the groups. The former groups the quality measures using clustering analysis, and the latter uses the loadings of a principal component analysis to identify the quality measures that contain similar information and group them. It is important to note that the "apriori", the "cluster", and the "loading" methods create groups of the original quality measures of the report and compute one MD within each group. Finally, the "global" method computes a single MD based on the reduced space of the first k principal components from a robust PCA. The number k of PCs can be chosen using a scree plot.

More details on each method are given in *Cohen Freue et al. (2007)*

## Value

An object of class "'mdqc"' (with associated plot, print and summary methods) with components

| | |
|---|---|
| ngroups | Number of groups in which the MDs have been computed |
| groups | column numbers corresponding to the quality measures in each group |
| mdqcValues | Mahalanobis Distance(s) for each array |
| x | dataset containing the numeric quality measures in the report |

| | |
|---|---|
| method | method used to group or transform the quality measures before computing the MD for each array |
| pc | number of principal components used in the robust PCA. |
| k | number of clusters used in the cluster analysis. |

## Note

We thank Christopher Croux for providing us a MATLAB code that we translated into R to compute the multivariate S-estimator

## Author(s)

Justin Harrington <harringt@stat.ubc.ca> and Gabriela V. Cohen Freue <gcohen@stat.ubc.ca>.

## References

Cohen Freue, G. V. and Hollander, Z. and Shen, E. and Zamar, R. H. and Balshaw, R. and Scherer, A. and McManus, B. and Keown, P. and McMaster, W. R. and Ng, R. T. (2007) 'MDQC: A New Quality Assessment Method for Microarrays Based on Quality Control Reports'. *Bioinformatics* **23**, 3162 – 3169.

Bolstad, B. M. and Collin, F. and Brettschneider, J. and Simpson, K. and Cope, L. and Irizarry R. A. and Speed T. P. (2005) 'Quality assessment of Affymetrix GeneChip data.' In Gentleman R. and Carey C. J. and Huber W. and Irizarry R. A. and Dudoit S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer.

Brettschneider, J. and Collin, F. and Bolstad, B. M. and Speed, T. P. (2007) 'Quality assessment for short oligonucleotide arrays'. Forthcoming in *Technometrics (with Discussion)*.

Ross, M. E. and Zhou, X. and Song, G. and Shurtleff, S. A. and Girtman, K. and Williams, W. K. and Liu, H. and Mahfouz, R. and Raimondi, S. C. and Lenny, N. and Patel, A. and Downing, J. R. (2003) 'Classification of pediatric acute lymphoblastic leukemia by gene expression profiling.' *Blood* **102**, 2951–9.

## See Also

prcomp.robust,pam, mahalanobis, allQC

## Examples

```
data(allQC)

## Contains the QC report obtained using Bioconductor's simpleaffy package
## for a subset of arrays from a large acute lymphoblastic leukemia (ALL)
## study (Ross et al., 2004).
## This dataset has been also studied by Bolstad et al. (2005) and
## Brettschneider et al. (2007).
## For further information see allQC.


#### No Groups method
```

```
# Figure 2 in Cohen Freue et al. (2007):
# Results of MDQC based on all measures of the QC report.

mdout <- mdqc(allQC, method="nogroups")
plot(mdout)
print(mdout)
summary(mdout)

#### A-Priori grouping method
# Figure 3 in Cohen Freue et al. (2007):
# Results of MDQC using the apriori grouping method.

mdout <- mdqc(allQC, method="apriori", groups=list(1:5, 6:9, 10:11))
plot(mdout)



#### Global PCA method
# Figure 4 in Cohen Freue et al.(2007):
# Results of MDQC using the global PCA method.

mdout <- mdqc(allQC, method="global", pc=4)
plot(mdout)



#### Clustering grouping method
# Figure 4 in Supplementary Material of Cohen Freue et al. (2007):
# Results of MDQC using a cluster analysis to form
# 3 groups of quality measures.

mdout <- mdqc(allQC, method="cluster", k=3)
plot(mdout)



#### Loading grouping method
# Figure 4 in Supplementary Material of Cohen Freue et al. (2007):
# Results of MDQC using a cluster analysis on the first
# k=4 loading vectors from a robust PCA to form 3 groups of quality measures.

mdout <- mdqc(allQC, method="loading", k=3, pc=4)
plot(mdout)


### To get the raw MD distances
mdout$mdqcValues
```

---

plot.mdqc *The plot method for a MDQC object*

---

## Description

The plot method for a MDQC object, which plots ...

## Usage

```
## S3 method for class 'mdqc'
plot(x, levels = c(0.9, 0.95, 0.99), xlab="", ylab="",
mfrow=NULL, mfcol=NULL, ...)
```

## Arguments

| | |
|---|---|
| x | An object of the class 'mdqc'. |
| levels | A vector or scalar between 0 and 1 for displaying critical values for outliers. See details. |
| xlab | The label for for x-axis. Note that when there are multiple plots, the same value of this argument is used for each one. |
| ylab | The label for the y-axis. Note that when there are multiple plots, the same value of this argument is used for each one. |
| mfrow | Specify the arrangement of plots on the page, by rows, or leave NULL to let the function work it out |
| mfcol | As for mcol, but arrange plots by column instead |
| ... | Other arguments passed to the default plot method. |

## Details

This plot method is for the output from the function mdqc, and plots the Mahalanobis distances for each array. The levels argument plots horizontal lines at critical values (based on the quantiles of a chi-squard distribution), and aids in identifying outliers.

For further details, see *Cohen Freue et al. (2007)*

## Author(s)

Justin Harrington <harringt@stat.ubc.ca> and Gabriela V. Cohen Freue <gcohen@stat.ubc.ca>.

## References

Cohen Freue, G. V. and Hollander, Z. and Shen, E. and Zamar, R. H. and Balshaw, R. and Scherer, A. and McManus, B. and Keown, P. and McMaster, W. R. and Ng, R. T. (2007) 'MDQC: A New Quality Assessment Method for Microarrays Based on Quality Control Reports'. *Bioinformatics* **23**, 3162 – 3169.

## See Also

mdqc

## Examples

```
data(allQC)
mdout <- mdqc(allQC, method="cluster", k=3)
plot(mdout)

## Just one critical value
plot(mdout, levels=0.9)
```

---

`prcomp.robust` *Principal Components Analysis using Robust Estimators*

---

## Description

A function that performs PCA using the robust estimators "S-estimator","MCD" and "MVE".

## Usage

```
prcomp.robust(x, robust = c("S-estimator","MCD", "MVE"),
                        nsamp = 10*nrow(x), ...)
## S3 method for class 'robust'
prcomp(x, robust = c("S-estimator","MCD", "MVE"),
                        nsamp = 10*nrow(x), ...)
```

## Arguments

| | |
|---|---|
| `x` | a matrix. Contains the data to perform PCA on. |
| `robust` | The robust estimator to use. One of `"S-estimator"`, `"MCD"`, or `"MVE"`. The default robust estimator is the S-estimator with 25% breakdown point. |
| `nsamp` | The number of subsamples that the robust estimator should use. This defaults to 10 times the number of rows in the matrix. |
| `...` | Further arguments that can be passed to the robust estimator |

## Details

The calculation is done by a singular value decomposition of the robust centered and scaled data matrix, not by using `eigen` on the covariance matrix. This is generally the preferred method for numerical accuracy. The `print` method for the these objects prints the results in a nice format and the `plot` method produces a scree plot. The scree plot can be used to determine the number k of principal components preserved in the analysis, looking for the "elbow" or the first important bend in the line. A biplot can also be generated to represent the values of the first two principal components (PCs) and the contribution of each variable to these components in the same plot (see Supplementary Material of Cohen Freue et al. (2007)).

## Value

`prcomp.robust` returns a list with class `"prcomp"` containing the following components:

| | |
|---|---|
| `sdev` | the standard deviations of the principal components (i.e., the square roots of the eigenvalues of the covariance matrix calculated using the `robust` argument, though the calculation is actually done with the singular values of the data matrix). |
| `rotation` | the matrix of variable loadings (i.e., a matrix whose columns contain the eigenvectors). The function `princomp` returns this in the element `loadings`. |
| `x` | the value of the rotated data (the centered and scaled) data multiplied by the `rotation` matrix) is returned. |

## Author(s)

Justin Harrington <harringt@stat.ubc.ca> and Gabriela V. Cohen Freue <gcohen@stat.ubc.ca>.

## References

Cohen Freue, G. V. and Hollander, Z. and Shen, E. and Zamar, R. H. and Balshaw, R. and Scherer, A. and McManus, B. and Keown, P. and McMaster, W. R. and Ng, R. T. (2007) 'MDQC: A New Quality Assessment Method for Microarrays Based on Quality Control Reports'. *Bioinformatics* **23**, 3162 – 3169.

## See Also

mdqc, `prcomp`

## Examples

```
data(allQC) ## Loads the dataset allQC

prout <- prcomp.robust(allQC)
screeplot(prout, type="line")
biplot(prout)

prout <- prcomp.robust(allQC, robust="MCD")
screeplot(prout, type="line")
biplot(prout)

prout <- prcomp.robust(allQC, robust="MVE")
screeplot(prout, type="line")
biplot(prout)
```

# Index