Package ‘represent’

November 3, 2023

Type Package

Title Determine How Representative Two Multidimensional Data Sets are

Version 1.0.1

Date 2023-11-02

Author Harmen Draisma [aut, cre]

Maintainer Harmen Draisma <h.h.m.draisma@gmail.com>

Description Compute the values of various parameters evaluating how similar two multidimensional datasets' structures are in multidimensional space, as described in: Jouan-Rimbaud, D., Massart, D. L., Saby, C. A., Puel, C. (1998), <doi:10.1016/S0169-7439(98)00005-7>. The computed parameters evaluate three properties, namely, the direction of the data sets, the variance-covariance of the data points, and the location of the data sets' centroids. The package contains workhorse function jrparams(), as well as two helper functions Mboxtest() and JRsMahaldist(), and four example data sets.

License GPL-3

Repository CRAN

Date/Publication 2023-11-03 16:40:02 UTC

NeedsCompilation no

R topics documented:

represent-package *Determine the representativity of two multidimensional data sets*

Description

This package contains the workhorse function jrparams(), as well as two helper functions Mboxtest() and JRsMahaldist(), and four example data sets. The jrparams() function computes the values of three types of parameters that assess the representativity of two multidimensional data sets. These parameters and the example data sets are described in a publication by Jouan-Rimbaud et al (1998).

Details

Package:	represent
Type:	Package
Version:	1.0.1
Date:	2023-11-02
License:	GPL-3

#

Author(s)

Harmen Draisma

Maintainer: Harmen Draisma <h.h.m.draisma@gmail.com>

References

Jouan-Rimbaud D, Massart DL, Saby CA, Puel C: Determination of the representativity between two multidimensional data sets by a comparison of their structure. Chemometrics and Intelligent Laboratory Systems 40 (1998) 129-144.

DATASET.1 *A 50 x 5 data set*

Description

One of two 50 x 5 data sets as mentioned on p. 135 of Jouan-Rimbaud et al (1998). See DATASET.2 for the other 50 x 5 data set.

Usage

```
data(DATASET.1)
```

Format

The format is: num [1:50, 1:5] 19.851 7.526 2.123 0.945 0.726 ...

Details

Variable 1: 50 values uniformly distributed between 0 and 20 + noise. Variable 2: 50 values uniformly distributed between 2 and 20. Variable 3: 50 values uniformly distributed between 10 and 20. Variable 4: 50 values uniformly distributed between 5 and 20. Variable 5: 50 values uniformly distributed between 6 and 20.

Source

Page 135 of: Jouan-Rimbaud D, Massart DL, Saby CA, Puel C: Determination of the representativity between two multidimensional data sets by a comparison of their structure. Chemometrics and Intelligent Laboratory Systems 40 (1998) 129-144.

Examples

```
data(DATASET.1)
```

DATASET.2 *A 50 x 5 data set*

Description

One of two 50 x 5 data sets as mentioned on p. 135 of Jouan-Rimbaud et al (1998). See DATASET.1 for the other 50 x 5 data set.

Usage

```
data(DATASET.2)
```

Format

The format is: num [1:50, 1:5] 2.72 12.05 6.5 12.27 16.03 ...

Details

Variable 1: 50 values uniformly distributed between 0 and 20 + noise. Variable 2: 50 values uniformly distributed between 2 and 20. Variable 3: 50 values uniformly distributed between 10 and 20. Variable 4: 50 values uniformly distributed between 5 and 20. Variable 5: 50 values uniformly distributed between 6 and 20.

Source

Page 135 of: Jouan-Rimbaud D, Massart DL, Saby CA, Puel C: Determination of the representativity between two multidimensional data sets by a comparison of their structure. Chemometrics and Intelligent Laboratory Systems 40 (1998) 129-144.

Examples

```
data(DATASET.2)
```

DATASET.3 *A 50 x 10 data set*

Description

One of two 50 x 10 data sets as mentioned on p. 135 of Jouan-Rimbaud et al (1998). See DATASET.4 for the other 50 x 10 data set.

Usage

```
data(DATASET.3)
```

Format

The format is: num [1:50, 1:10] 19.851 7.526 2.123 0.945 0.726 ...

Details

Variable 1: 50 values uniformly distributed between 0 and 20 + noise. Variable 2: 50 values uniformly distributed between 2 and 20. Variable 3: 50 values uniformly distributed between 10 and 20. Variable 4: 50 values uniformly distributed between 5 and 20. Variable 5: 50 values uniformly distributed between 6 and 20. Variables 6-10: five variables with values uniformly distributed between -0.5 and +0.5

Source

Page 135 of: Jouan-Rimbaud D, Massart DL, Saby CA, Puel C: Determination of the representativity between two multidimensional data sets by a comparison of their structure. Chemometrics and Intelligent Laboratory Systems 40 (1998) 129-144.

Examples

```
data(DATASET.3)
```

DATASET.4 *A 50 x 10 data set*

Description

One of two 50 x 10 data sets as mentioned on p. 135 of Jouan-Rimbaud et al (1998). See DATASET.3 for the other 50 x 10 data set.

Usage

```
data(DATASET.4)
```

Format

The format is: num [1:50, 1:10] 2.72 12.05 6.5 12.27 16.03 ...

Details

Variable 1: 50 values uniformly distributed between 0 and 20 + noise. Variable 2: 50 values uniformly distributed between 2 and 20. Variable 3: 50 values uniformly distributed between 10 and 20. Variable 4: 50 values uniformly distributed between 5 and 20. Variable 5: 50 values uniformly distributed between 6 and 20. Variables 6-10: five variables with values uniformly distributed between -0.5 and +0.5

Source

Page 135 of: Jouan-Rimbaud D, Massart DL, Saby CA, Puel C: Determination of the representativity between two multidimensional data sets by a comparison of their structure. Chemometrics and Intelligent Laboratory Systems 40 (1998) 129-144.

Examples

```
data(DATASET.4)
```

jrparams *Assess similarity of two multidimensional data sets*

Description

This function computes three types of parameters to assess the representativity of two multidimensional data sets by a comparison of their structure. Representativity is expressed as similarity of: I) principal component analysis (PCA) loadings patterns; II) variance-covariance matrix structures; III) data set centroid locations. All parameters are computed in principal component (PC) space. These parameters are described in a publication by Jouan-Rimbaud et al (1998).

Usage

```
jrparams(BLOCK.1,BLOCK.2,ncomp=min(c(dim(BLOCK.1),dim(BLOCK.2))),Cscrit=0.6,Rscrit=0.6)
```

Arguments

BLOCK.1	First multivariate data set (a numeric matrix)
BLOCK.2	Second multivariate data set (a numeric matrix), to be compared with the first
ncomp	The number of PCs to compute the parameter values for
Cscrit	The value of the "C*" parameter corresponding to the value of Box's M statistic being equal to its critical value
Rscrit	The value of the "R*" parameter corresponding to the Mahalanobis distance being equal to its critical value

Details

For argument 'ncomp', the default is based on the smallest number of rows or columns (whichever is smaller) in either of both data sets to be compared. This number should be a proxy for the minimum of the 'ranks' (i.e., the actual dimensionalities) of both data sets.

The default settings for the values of arguments 'Cscrit' and 'Rscrit' correspond to the values as recommended by Jouan-Rimbaud et al (1998) in their equations (9a) and (13a), respectively.

Value

A numeric matrix with rows containing the computed values for in total six parameters that are described in Jouan-Rimbaud et al (1998). The nomenclature for the parameters as in that publication has been adopted here. Hence, the first two rows ("P" and "P*") of the output are informative of the similarity of the PCA loadings patterns of both data sets. Rows 3 and 4 ("C" and "C*", respectively) are indicative of the similarity of the variance-covariance matrices. Finally, rows 5 and 6 ("R" and "R*") represent the similarity of the data set centroid locations. For all parameters, values equal to 1 indicate perfect similarity. The number of columns of the output matrix depends on the value of 'ncomp'.

Warning

Unexpected results might occur if the two data sets to be compared are of different rank, and the number of principal components to retain has not been passed to jrparams() as well (not tested).

Note

The function performs principal component analysis itself, so one can just input the original data sets (containing the original manifest variables). In general it is wise to compute the parameter values only for the significant principal components. Significance of principal components for both data sets to be compared can be assessed using e.g. scree plots, as available for instance in the 'psych' package.

Author(s)

Harmen Draisma

References

Jouan-Rimbaud D, Massart DL, Saby CA, Puel C: Determination of the representativity between two multidimensional data sets by a comparison of their structure. Chemometrics and Intelligent Laboratory Systems 40 (1998) 129-144.

Examples

```
#Load example data sets, 50 observations x 5 variables
data(DATASET.1)
data(DATASET.2)

#Assess representativity using all principal components
#(default; will be fine if both sets are of equal rank)
jrparams(DATASET.1, DATASET.2)

#Positive control: check similarity of DATASET.1 with itself
#(values for all parameters should be unity)
jrparams(DATASET.1, DATASET.1)
```

`JRsMahaldist` *Compute Mahalanobis distance*

Description

Computes Mahalanobis distance according to Section 2.5.1 in Jouan-Rimbaud et al (1998). The distance takes into account both the distance between the centroids of two data sets, as well as the dispersion of the data around these centroids. This function is used by the function jrparams() to assess representativity of two multidimensional data sets.

Usage

```
JRsMahaldist(DATA)
```

Arguments

`DATA`	Matrix containing the weighted scores on the principal components (PCs) that have been computed for the two data sets to be compared. The first column of this matrix should contain a group indicator variable, which has a value equal to 1 for the first data set and a value equal to 2 for the second data set. The remaining columns contain the weighted PC scores for the two data sets.

Value

A list type object containing one field named "Ds", a 1*1 matrix type object whose only element has the value of the Mahalanobis distance.

Note

The Mahalanobis distance is computed using the pooled variance-covariance matrix as defined in Section 2.4 of Jouan-Rimbaud et al (1998), and hence may differ somewhat from a 'regular' Mahalanobis distance as computed using e.g. the function mahalanobis() from the 'stats' package.

Author(s)

Harmen Draisma

References

Section 2.5.1 in: Jouan-Rimbaud D, Massart DL, Saby CA, Puel C: Determination of the representativity between two multidimensional data sets by a comparison of their structure. Chemometrics and Intelligent Laboratory Systems 40 (1998) 129-144.

See Also

jrparams(); MBoxtest()

`MBoxtest` *Compute Box's M-statistic*

Description

Computes Box's M-statistic according to Section 2.4 in Jouan-Rimbaud et al (1998). This statistic is used to compare the structures of the variance-covariance matrices of two multidimensional data sets. This function is used by the function jrparams().

Usage

```
MBoxtest(DATA, nmanvars)
```

Arguments

`DATA`	Matrix containing the weighted scores on the principal components (PCs) that have been computed for the two data sets to be compared. The first column of this matrix should contain a group indicator variable, which has a value equal to 1 for the first data set and a value equal to 2 for the second data set. The remaining columns contain the weighted PC scores for the two data sets.
`nmanvars`	Number of manifest variables in the original multidimensional data sets.

Value

A list with two elements:

`MB`	Box's M-statistic
`Sp`	Pooled covariance matrix

Author(s)

Harmen Draisma

References

Section 2.4 in: Jouan-Rimbaud D, Massart DL, Saby CA, Puel C: Determination of the representativity between two multidimensional data sets by a comparison of their structure. Chemometrics and Intelligent Laboratory Systems 40 (1998) 129-144.

See Also

jrparams()

Index