Package ‘ToolsForCoDa’

January 9, 2025

Type Package

Title Multivariate Tools for Compositional Data Analysis

Version 1.1.0

Date 2025-01-09

Author Jan Graffelman [aut, cre]

Maintainer Jan Graffelman <jan.graffelman@upc.edu>

Depends R (>= 1.8.0), MASS, calibrate, Correlplot

Description Provides functions for multivariate analysis with compositional data. Includes a function for doing compositional canonical correlation analysis. This analysis requires two data matrices of compositions, which can be adequately transformed and used as entries in a specialized program for canonical correlation analysis, that is able to deal with singular covariance matrices. The methodology is described in Graffelman et al. (2017) <doi:10.1101/144584>. Functions for log-ratio principal component analysis with condition number computations and log-ratio discriminant analysis have been added to the package.

URL https://www.r-project.org, http://www-eio.upc.edu/~jan/

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2025-01-09 11:00:05 UTC

Contents

`Artificial` *Two sets of 3-part compositions*

Description

The list object `Artificial` contains two data frames of 3-part compositions. The data refer to the example in Section 3.1 of Graffelman et al. (2017)

Usage

```
data(Artificial)
```

Format

A list containing two data frames containing 100 observations.

Source

Laird, N. M. and Lange, C. Table 7.11, p. 124

References

Graffelman, J., Pawlowsky-Glahn, V., Egozcue, J.J. and Buccianti, A. (2017) Compositional Canonical Correlation Analysis.

`bentonites` *Isotopic and chemical compositions of bentonites*

Description

The data consists of 14 geological samples from the US with their major oxide composition (SiO2, Al2O3, Fe2O3, MnO, MgO, CaO, K2O, Na2O and H2O+) and delta Deuterium and delta-18-Oxysgen (dD,d18O).

Usage

```
data("bentonites")
```

Format

A data frame with 14 observations on the following 11 variables.

`Si` a numeric vector

`Al` a numeric vector

`Fe` a numeric vector

`Mn` a numeric vector

`Mg` a numeric vector

`Ca` a numeric vector

`K` a numeric vector

`Na` a numeric vector

`H20` a numeric vector

`dD` a numeric vector

`d18O` a numeric vector

Source

Cadrin, A.A.J (1995), Tables 1 and 2. Reyment, R. A. and Savazzi, E. (1999), pp. 220-222.

References

Cadrin, A.A.J., Kyser, T.K., Caldwell, W.G.E. and Longstaffe, F.J. (1995) Isotopic and chemical compositions of bentonites as paleoenvironmental indicators of the Cretaceous Western Interior Seaway Palaeogeography, Palaeoclimatology, Palaeoecology 119 pp. 301–320.

Reyment, R. A. and Savazzi, E. (1999) Aspects of Multivariate Statistical Analysis in Geology, Elsevier Science B.V., Amsterdam.

Examples

```
data(bentonites)
```

`canocov` *Canonical correlation analysis.*

Description

Function `canocov` performs a canonical correlation analysis. It operates on raw data matrices, which are only centered in the program. It uses generalized inverses and can deal with structurally singular covariance matrices.

Usage

```
canocov(X, Y)
```

Arguments

`X`	The n times p X matrix of observations
`Y`	The n times q Y matrix of observations

Details

`canocov` computes the solution by a singular value decomposition of the transformed between set covariance matrix.

Value

Returns a list with the following results

`ccor`	the canonical correlations
`A`	canonical weights of the X variables
`B`	canonical weights of the Y variables
`U`	canonical X variates
`V`	canonical Y variates
`Fs`	biplot markers for X variables (standard coordinates)
`Gs`	biplot markers for Y variables (standard coordinates)
`Fp`	biplot markers for X variables (principal coordinates)
`Gp`	biplot markers for Y variables (principal coordinates)
`Rxu`	canonical loadings, (correlations X variables, canonical X variates)
`Rxv`	canonical loadings, (correlations X variables, canonical Y variates)
`Ryu`	canonical loadings, (correlations Y variables, canonical X variates)
`Ryv`	canonical loadings, (correlations Y variables, canonical Y variates)
`Sxu`	covariance X variables, canonical X variates
`Sxv`	covariance X variables, canonical Y variates
`Syu`	covariance Y variables, canonical X variates
`Syv`	covariance Y variables, canonical Y variates
`fitRxy`	goodness of fit of the between-set correlation matrix
`fitXs`	adequacy coefficients of X variables
`fitXp`	redundancy coefficients of X variables
`fitYs`	adequacy coefficients of Y variables
`fitYp`	redundancy coefficients of Y variables

Author(s)

Jan Graffelman <jan.graffelman@upc.edu>

References

Hotelling, H. (1935) The most predictable criterion. Journal of Educational Psychology (26) pp. 139-142.

Hotelling, H. (1936) Relations between two sets of variates. Biometrika (28) pp. 321-377.

Johnson, R. A. and Wichern, D. W. (2002) Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall.

See Also

cancor

Examples

```
set.seed(123)
X <- matrix(runif(75),ncol=3)
Y <- matrix(runif(75),ncol=3)
cca.results <- canocov(X,Y)
```

---

cen — *centring of a data matrix*

Description

centres the columns of a matrix to mean zero.

Usage

```
cen(X,w=rep(1,nrow(X))/nrow(X))
```

Arguments

X	a raw data matrix.
w	a vector of case weights.

Value

returns a matrix

Author(s)

Jan Graffelman (jan.graffelman@upc.edu)

Examples

```
X<-matrix(runif(10),ncol=2)
Y<-cen(X)
print(Y)
```

`clrmat` *Centred log-ratio transformation*

Description

Program `clrmat` calculates the centred log-ratio transformation for a matrix of compositions.

Usage

```
clrmat(X)
```

Arguments

`X`	A matrix of compositions

Value

A matrix containing the transformed data

Author(s)

Jan Graffelman <jan.graffelman@upc.edu>

Examples

```
data(Artificial)
Xsim.com <- Artificial$Xsim.com
Xclr <- clrmat(Xsim.com)
```

`largest.kappas` *Calculate condition indices for subcompositions*

Description

Function `largest.kappas` calculates the condition numbers for all subcompositions of a given size, for a particular compositional data set.

Usage

```
largest.kappas(Xcom, nparts = 3, sizetoplist = 10)
```

Arguments

`Xcom`	A data matrix with compositions in rows
`nparts`	The number of parts for the subcompositions to be analysed.
`sizetoplist`	The length of the list of the "best" subcompositions

Details

Log-ratio PCA is executed for each subcompostion, and the resulting eigenvalues and eigenvectors are stored.

Value

A data frame with an ordered list of subcompositions

Author(s)

Jan Graffelman (jan.graffelman@upc.edu)

Examples

```
X <- matrix(runif(600),ncol=6)
Xcom <- X/rowSums(X)
Results <- largest.kappas(Xcom)
```

---

lrcco *Logratio Canonical Correlation Analysis*

Description

Function `lrcco` is a wrapper function around `canocov`. It performs logratio canonical correlation analysis (LR-CCO) accepting two compositional data matrices as input.

Usage

```
lrcco(X, Y)
```

Arguments

`X`	The matrix of X compositions
`Y`	The matrix of Y compositions

Details

Matrices `X` and `Y` are assumed to contain positive elements only, and there rows sum to one.

Value

Returns a list with the following results

`ccor`	the canonical correlations
`A`	canonical weights of the X variables
`B`	canonical weights of the Y variables
`U`	canonical X variates

V	canonical Y variates
Fs	biplot markers for X variables (standard coordinates)
Gs	biplot markers for Y variables (standard coordinates)
Fp	biplot markers for X variables (principal coordinates)
Gp	biplot markers for Y variables (principal coordinates)
Rxu	canonical loadings, (correlations X variables, canonical X variates)
Rxv	canonical loadings, (correlations X variables, canonical Y variates)
Ryu	canonical loadings, (correlations Y variables, canonical X variates)
Ryv	canonical loadings, (correlations Y variables, canonical Y variates)
Sxu	covariance X variables, canonical X variates
Sxv	covariance X variables, canonical Y variates
Syu	covariance Y variables, canonical X variates
Syv	covariance Y variables, canonical Y variates
fitRxy	goodness of fit of the between-set correlation matrix
fitXs	adequacy coefficients of X variables
fitXp	redundancy coefficients of X variables
fitYs	adequacy coefficients of Y variables
fitYp	redundancy coefficients of Y variables

Author(s)

Jan Graffelman <jan.graffelman@upc.edu>

References

Hotelling, H. (1935) The most predictable criterion. Journal of Educational Psychology (26) pp. 139-142.

Hotelling, H. (1936) Relations between two sets of variates. Biometrika (28) pp. 321-377.

Johnson, R. A. and Wichern, D. W. (2002) Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall.

Graffelman, J. and Pawlowsky-Glahn, V. and Egozcue, J.J. and Buccianti, A. (2018) Exploration of geochemical data with compositional canonical biplots, Journal of Geochemical Exploration 194, pp. 120–133. doi:10.1016/j.gexplo.2018.07.014

See Also

cancor,canocov

Examples

```
set.seed(123)
X  <- matrix(runif(75),ncol=3)
Y  <- matrix(runif(75),ncol=3)
Xc <- X/rowSums(X) # create compositions by closure
Yc <- Y/rowSums(Y)
out.lrcco <- lrcco(X,Y)
```

`lrlda` *Logratio Linear Discriminant Analysis*

Description

Function `lrlda` implements logratio linear discriminant analysis for compositional data, using the centred logratio transformation (clr)

Usage

```
lrlda(Xtrain, group, Xtest = NULL, divisorn = FALSE, verbose = FALSE)
```

Arguments

`Xtrain`	A compositional data set, the training data for logratio-LDA.
`group`	A categorical variable defining the groups.
`Xtest`	A compositional data set for which group prediction is sought (the test data). If no test data is supplied, the training data itself is classified.
`divisorn`	Use divisor "n" (`divisorn=TRUE`) in the calculation of covariance or use "n-1" (`divisorn=TRUE`)
`verbose`	Print output (`verbose = TRUE`) or not.

Details

Function `lrlda` uses the centred logratio transformation, which produces a singular covariance matrix. This singularity is dealt with by using a generalized inverse. When test data is supplied via argument `Xtest`, the scores of the linear classifier, the poster probabilities and the predicted classes are calculated for the test data. If no test data is supplied, these quantities are calculated for the training data.

Value

`LD`	Scores on the linear classifier for the test observations. These are also the biplot coordinates of the individuals.
`Fp`	Biplot coordinates of the group means.
`Gs`	Biplot coordinates of the variables.
`Sp`	Pooled covariance matrix.
`Mc`	Matrix of centred clr mean vectors, one row for each group.
`S.list`	Covariance matrices of each group.
`la`	Vector of eigenvalues.
`pred`	Predicted class for the test observations.
`CM`	The confusion matrix.
`gsize`	Sample size of each group.

Mclr	Matrix of mean vectors for clr coordinates, one row for each group.
prob.posterior	Vector of posterior probabilities.
decom	Table with decomposition of variability as expressed by the eigenvalues.

Author(s)

Jan Graffelman (jan.graffelman@upc.edu)

See Also

lrpca,lrlda

Examples

```
data(Tubb)
sampleid <- Tubb$Sample
site     <- factor(Tubb$site)
Oxides   <- as.matrix(Tubb[,2:10])
rownames(Oxides) <- sampleid
Oxides   <- Oxides/rowSums(Oxides)
out.lda  <- lrlda(Oxides,site,verbose=FALSE)
```

---

lrpca *Logratio principal component analysis with condition indices*

Description

Function lrpca performs logratio principal component analysis. It returns the variance decomposition, principal components, biplot coordinates and a table with condition indices.

Usage

```
lrpca(Xcom)
```

Arguments

Xcom	A matrix with compositions in its rows

Details

Calculations are based on the singular value decompositon of the clr transformed compositions.

Value

Fp	matrix with principal components
Fs	matrix with standardized principal components
Gp	biplot markers for parts (principal coordinates)
Gs	biplot markers for parts (standard coordinates)
La	eigenvalues
D	singular values
decom	table with variance decomposition
kappalist	table with condition indices and eigenvectors

Author(s)

Jan Graffelman (jan.graffelman@upc.edu)

See Also

princomp

Examples

```
data(bentonites)
Ben <- bentonites[,1:8]
Ben.com <- Ben/rowSums(Ben)
out.lrpca <- lrpca(Ben.com)
```

PinotNoir *Chemical composition of Pinot Noir wines*

Description

Dataframe `PinotNoir` contains the composition of 17 chemical components for 37 Pinot Noir wines, as well as an Aroma evaluation.

Usage

```
data("PinotNoir")
```

Format

A data frame with 37 observations on the following 18 variables.

- `Cd` Cadmium
- `Mo` Molybdenum
- `Mn` Manganese
- `Ni` Nickel

`Cu` Copper

`Al` Aluminium

`Ba` Barium

`Cr` Chromium

`Sr` Strontium

`Pb` Lead

`B` Boron

`Mg` Magnesium

`Si` Silicon

`Na` Sodium

`Ca` Calcium

`P` Phosphorus

`K` Potassium

`Aroma` Aroma evaluation

Source

doi:10.1016/S00032670(00)842452

References

Frank, I.E. and Kowalski, B.R. (1984) Prediction of Wine Quality and Geographic Origin from Chemical Measurements by Partial Least-Squares Regression Modeling. Analytica Chimica Acta 162, pp. 241–251 doi:10.1016/S00032670(00)842452

Examples

```
data(PinotNoir)
```

`ternaryplot` *Create a Ternary Plot for three-part Compositions*

Description

Function `ternaryplot` accepts a matrix of three part compositions or non-negative counts and presents these in a ternary diagram.

Usage

```
ternaryplot(X, vertexlab = colnames(X), vertex.cex = 1, pch = 19, addpoints = TRUE,
            grid = FALSE, gridlabels = TRUE, ...)
```

Arguments

`X`	A matrix of counts or compositions with three columns
`vertexlab`	Labels for the vertices of the tenary diagram
`vertex.cex`	Character expansion factor for vertex labels
`pch`	Plotting character for the compositions
`addpoints`	Show the compositions `addpoints=TRUE` or not
`grid`	Place a grid over the ternary diagram
`gridlabels`	Place grid labels or not
`...`	Additional arguments for the `points` function

Value

NULL

Author(s)

Jan Graffelman (jan.graffelman@upc.edu)

Examples

```
data("Artificial")
Xsim.com <- Artificial$Xsim.com
colnames(Xsim.com) <- paste("X",1:3,sep="")
ternaryplot(Xsim.com)
```

`tr` *Compute the trace of a matrix*

Description

`tr` computes the trace of a matrix.

Usage

```
tr(X)
```

Arguments

`X`	a (square) matrix

Value

the trace (a scalar)

Author(s)

Jan Graffelman (jan.graffelman@upc.edu)

Examples

```
X <- matrix(runif(25),ncol=5)
print(X)
print(tr(X))
```

Tubb *Romano-British pottery oxides*

Description

A dataframe with the major oxide composition of pottery found at Romano-British kiln sites in Wales, Gloucester and the New Forest as determined by atomic absorption.

Usage

```
data("Tubb")
```

Format

A data frame with 48 observations on the following 11 variables.

`Sample` Sample identifier

`Al2O3` Aluminium oxide

`Fe2O3` Iron (III) oxide

`MgO` Magnesium oxide

`CaO` Calcium oxide

`Na2O` Sodium oxide

`K2O` Potassium oxide

`TiO2` Titaniium dioxide

`MnO` Manganese oxide

`BaO` Barium oxide

`site` Geographical region of the sample. G=Gloucester, NF=New Forest, W=Wales.

References

Tubb, A., Parker, A.J. and Nickless, G. (1980) The analysis of Romano-British pottery by atomic absorption spectrophotometry. Archaeometry 22(2) pp. 153–171.

Examples

```
data(Tubb)
```

Index