Package 'coda.base'

July 3, 2025

Type Package

Title A Basic Set of Functions for Compositional Data Analysis

Version 1.0.3

Date 2025-07-02

Description A minimum set of functions to perform compositional data analysis using the log-ratio approach introduced by John Aitchison (1982). Main functions have been implemented in c++ for better performance.

URL https://mcomas.net/coda.base/, https://github.com/mcomas/coda.base

Depends R (>= 3.5)

Imports Rcpp (>= 0.12.12), stats, Matrix

LinkingTo Rcpp, RcppArmadillo

License GPL

Encoding UTF-8

LazyData true

NeedsCompilation yes

RoxygenNote 7.3.2

Suggests knitr, rmarkdown, testthat (>= 2.1.0), ggplot2, jsonlite

VignetteBuilder knitr

Author Marc Comas-Cufí [aut, cre] (ORCID: <https://orcid.org/0000-0001-9759-0622>)

Maintainer Marc Comas-Cufí <mcomas@imae.udg.edu>

Repository CRAN

Date/Publication 2025-07-02 22:10:09 UTC

Contents

`alimentation` *Food consumption in European countries*

Description

The alimentation data set contains the percentages of consumption of several types of food in 25 European countries during the 80s. The categories are: * RM: red meat (pork, veal, beef), * WM: white meat (chicken), * E: eggs, * M: milk, * F: fish, * C: cereals, * S: starch (potatoes), * N: nuts, and * FV: fruits and vegetables.

Usage

```
alimentation
```

Format

An object of class `data.frame` with 25 rows and 13 columns.

Details

Moreover, the dataset contains a categorical variable that shows if the country is from the North or a Southern Mediterranean country. In addition, the countries are classified as Eastern European or as Western European.

---

alr_basis *Additive log-ratio basis*

Description

Compute the transformation matrix to express a composition using the oblique additive log-ratio coordinates.

Usage

```
alr_basis(dim, denominator = NULL, numerator = NULL)
```

Arguments

`dim`	An integer indicating the number of components. If a dataframe or matrix is provided, the number of components is inferred from the number of columns. If a character vector specifying the names of the parts is provided the number of component is its length.
`denominator`	part used as denominator (default behaviour is to use last part)
`numerator`	parts to be used as numerator. By default all except the denominator parts are chosen following original order.

Value

matrix

References

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). 416p.

Examples

```
alr_basis(5)
# Third part is used as denominator
alr_basis(5, 3)
# Third part is used as denominator, and
# other parts are rearranged
alr_basis(5, 3, c(1,5,2,4))
```

`arctic_lake` *Arctic lake sediments at different depths*

Description

The arctic lake data set records the [sand, silt, clay] compositions of 39 sediment

Usage

```
arctic_lake
```

Format

An object of class `data.frame` with 39 rows and 5 columns.

`blood_mn` *The MN blood system*

Description

In humans the main blood group systems are the ABO system, the Rh system and the MN system. The MN blood system is a system of blood antigens also related to proteins of the red blood cell plasma membrane. The inheritance pattern of the MN blood system is autosomal with codominance, a type of lack of dominance in which the heterozygous manifests a phenotype totally distinct from the homozygous. The possible phenotypical forms are three blood types: type M blood, type N blood and type MN blood. The frequencies of M, N and MN blood types vary widely depending on the ethnic population. However, the Hardy-Weinberg principle states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. This implies that, in the long run, it holds that

$\frac{x_{MM}x_{NN}}{x_{MN}} = \frac{1}{4}$

where xM M and xN N are the genotype relative frequencies of MM and NN homozygotes, respectively, and xM N is the genotype relative frequency of MN heterozygotes. This principle was named after G.H. Hardy and W. Weinberg demonstrated it mathematically.

Usage

```
blood_mn
```

Format

An object of class `data.frame` with 49 rows and 5 columns.

`bmi_activity` *Physical activity and body mass index*

Description

The 'bmi_activity' data set records the proportion of daily time spent to sleep (sleep), sedentary behaviour (sedent), light physical activity (Lpa), moderate physical activity (Mpa) and vigorous physical activity (Vpa) measured on a small population of 393 children. Moreover the standardized body mass index (zBMI) of each child was also registered.

This data set was used in the example of the article (Dumuid et al. 2019) to examine the expected differences in zBMI for reallocations of daily time between sleep, physical activity and sedentary behaviour. Because the original data is confidential, the data set BMIPhisActi includes simulated data that mimics the main features of the original data.

Usage

```
bmi_activity
```

Format

An object of class `data.frame` with 393 rows and 8 columns.

References

D. Dumuid, Z. Pedisic, T.E. Stanford, J.A. Martín-Fernández, K. Hron, C. Maher, L.K. Lewis and T.S. Olds, *The Compositional Isotemporal Sub- stitution Model: a Method for Estimating Changes in a Health Outcome for Reallocation of Time between Sleep, Sedentary Behaviour, and Physical Activity*. Statistical Methods in Medical Research **28**(3) (2019), 846–857

`cc_basis` *Isometric Log-Ratio Basis Based on Canonical Correlations*

Description

Constructs an isometric log-ratio (ilr) basis for a compositional dataset, optimized with respect to canonical correlations with an explanatory dataset.

Usage

```
cc_basis(Y, X)
```

Arguments

Y	A compositional dataset (matrix or data frame).
X	An explanatory dataset (matrix or data frame).

Value

A matrix representing the isometric log-ratio basis.

---

cdp_partition *CoDaPack's default binary partition*

Description

Compute the default binary partition used in CoDaPack's software

Usage

```
cdp_partition(ncomp)
```

Arguments

ncomp	number of parts

Value

matrix

Examples

```
cdp_partition(4)
```

---

center *Dataset center*

Description

Generic function to calculate the center of a compositional dataset

Usage

```
center(X, zero.rm = FALSE, na.rm = FALSE)
```

Arguments

X	compositional dataset
zero.rm	a logical value indicating whether zero values should be stripped before the computation proceeds.
na.rm	a logical value indicating whether NA values should be stripped before the computation proceeds.

Examples

```
X = matrix(exp(rnorm(5*100)), nrow=100, ncol=5)
g = rep(c('a','b','c','d'), 25)
center(X)
(by_g <- by(X, g, center))
center(t(simplify2array(by_g)))
```

---

clr_basis *Centered log-ratio basis*

Description

Compute the transformation matrix to express a composition using the linearly dependant centered log-ratio coordinates.

Usage

```
clr_basis(dim)
```

Arguments

dim	An integer indicating the number of components. If a dataframe or matrix is provided, the number of components is inferred from the number of columns. If a character vector specifying the names of the parts is provided the number of component is its length.

Value

matrix

References

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). 416p.

Examples

```
(B <- clr_basis(5))
# CLR coordinates are linearly dependant coordinates.
(clr_coordinates <- coordinates(c(1,2,3,4,5), B))
# The sum of all coordinates equal to zero
sum(clr_coordinates) < 1e-15
```

`coda.base` *coda.base*

Description

A minimum set of functions to perform compositional data analysis using the log-ratio approach introduced by John Aitchison (1982) <https://www.jstor.org/stable/2345821>. Main functions have been implemented in c++ for better performance.

Author(s)

Marc Comas-Cufí

See Also

Useful links:

- https://mcomas.net/coda.base/
- https://github.com/mcomas/coda.base

`coda_replacement` *Replacement of Missing Values and Below-Detection Zeros in Compositional Data*

Description

Performs imputation (replacement) of missing values and/or values below the detection limit (BDL) in compositional datasets using the EM-algorithm assuming normality on the Simplex. This function is designed to prepare compositional data for subsequent log-ratio transformations.

Usage

```
coda_replacement(
  X,
  DL = NULL,
  dl_prop = 0.65,
  eps = 1e-04,
  parameters = FALSE,
  debug = FALSE
)
```

Arguments

X	A compositional dataset: numeric matrix or data frame where rows represent observations and columns represent parts.
DL	An optional matrix or vector of detection limits. If NULL, the minimum non-zero value in each column of X is used.
dl_prop	A numeric value between 0 and 1, used for initialization in the EM algorithm (default is 0.65).
eps	A small positive value controlling the convergence criterion for the EM algorithm (default is 1e-4).
parameters	Logical. If TRUE, returns additional output including estimated multivariate normal parameters (default is FALSE).
debug	Logical. Show the log-likelihood in every iteration.

Details

- Missing values are imputed based on a multivariate normal model on the simplex. - Zeros are treated as censored values and replaced accordingly. - The EM algorithm iteratively estimates the missing parts and model parameters. - To initialize the EM algorithm, zero values (considered below the detection limit) are replaced with a small positive value. Specifically, each zero is replaced by dl_prop times the detection limit of that part (column). This restrictions is imposed in the geometric mean of the parts with zeros against the non-missing positive values, helping to preserve the compositional structure in the simplex.

Value

If parameters = FALSE, returns a numeric matrix with imputed values. If parameters = TRUE, returns a list with two components:

X_imp The imputed compositional data matrix.

info A list containing information about the EM algorithm parameters and convergence diagnostics.

Examples

```
# Simulate compositional data with zeros
set.seed(123)
X <- abs(matrix(rnorm(100), ncol = 5))
X[sample(length(X), 10)] <- 0  # Introduce some zeros
X[sample(length(X), 10)] <- NA  # Introduce some NAs
# Apply replacement
summary(X/rowSums(X, na.rm=TRUE))
summary(coda_replacement(X))
```

composition *Get composition from coordinates w.r.t. an specific basis*

Description

Calculate a composition from coordinates with respect a given basis

Usage

```
composition(H, basis = "ilr")

comp(H, basis = "ilr")
```

Arguments

H	coordinates of a composition. Either a matrix, a data.frame or a vector
basis	basis used to calculate the coordinates

Value

coordinates with respect the given basis

See Also

See functions ilr_basis, alr_basis, clr_basis, sbp_basis to define different compositional basis. See function coordinates to obtain details on how to calculate coordinates of a given composition.

coordinates *Get coordinates from compositions w.r.t. an specific basis*

Description

Calculate the coordinates of a composition with respect a given basis

Usage

```
coordinates(X, basis = "ilr")

coord(..., basis = "ilr")

alr_c(X)

clr_c(X)

ilr_c(X)

olr_c(X)
```

Arguments

X	compositional dataset. Either a matrix, a data.frame or a vector
basis	basis used to calculate the coordinates. basis can be either a string or a matrix. Accepted values for strings are: 'ilr' (default), 'clr', 'alr', 'pw', 'pc', 'pb' and 'cdp'. If basis is a matrix, it is expected to have log-ratio basis given in columns.
...	components of the compositional data

Details

coordinates function calculates the coordinates of a compositiona w.r.t. a given basis. 'basis' parameter is used to set the basis, it can be either a matrix defining the log-contrasts in columns or a string defining some well-known log-contrast: 'alr' 'clr', 'ilr', 'pw', 'pc', 'pb' and 'cdp', for the additive log-ratio, centered log-ratio, isometric log-ratio, pairwise log-ratio, clr principal components, clr principal balances or default's CoDaPack balances respectively.

Value

Coordinates of composition X with respect the given basis.

See Also

See functions ilr_basis, alr_basis, clr_basis, sbp_basis to define different compositional basis. See function composition to obtain details on how to calculate a compositions from given coordinates.

Examples

```
# Default ilr given by ilr_basis(5) is given
coordinates(1:5)
B = ilr_basis(5)
coordinates(1:5, B)
```

---

dist *Distance Matrix Computation (including Aitchison distance)*

Description

This function overwrites dist function to contain Aitchison distance between compositions.

Usage

```
dist(x, method = "euclidean", ...)
```

Arguments

x	compositions method
method	the distance measure to be used. This must be one of "aitchison", "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski". Any unambiguous substring can be given.
...	arguments passed to dist function

Value

dist returns an object of class "dist".

See Also

See functions dist.

Examples

```
X = exp(matrix(rnorm(10*50), ncol=50, nrow=10))

(d <- dist(X, method = 'aitchison'))
plot(hclust(d))

# In contrast to Euclidean distance
dist(rbind(c(1,1,1), c(100, 100, 100)), method = 'euc') # method = 'euclidean'
# using Aitchison distance, only relative information is of importance
dist(rbind(c(1,1,1), c(100, 100, 100)), method = 'ait') # method = 'aitchison'
```

---

eurostat_employment *Employment distribution in EUROSTAT countries*

Description

According to the three–sector theory, as a country's economy develops, employment shifts from the primary sector (raw material extraction: farming, hunting, fishing, mining) to the secondary sector (industry, energy and construction) and finally to the tertiary sector (services). Thus, a country's employment distribution can be used as a predictor of economic wealth.

The 'eurostat_employment' data set contains EUROSTAT data on employment aggregated for both sexes, and all ages distributed by economic activity (classification 1983-2008, NACE Rev. 1.1) in 2008 for the 29 EUROSTAT member countries, thus reflecting reality just before the 2008 financial crisis. Country codes in alphabetical order according to the country name in its own language are: Belgium (BE), Cyprus (CY), Czechia (CZ), Denmark (DK), Deutchland–Germany (DE), Eesti–Estonia (EE), Eire–Ireland (IE), España–Spain (ES), France (FR), Hellas-Greece (GR), Hrvatska–Croatia (HR), Iceland (IS), Italy (IT), Latvia (LV), Lithuania (LT), Luxembourg (LU), Macedonia (MK), Magyarország-Hungary (HU), Malta (MT), Netherlands (NL), Norway (NO), Österreich–Austria (AT), Portugal (PT), Romania (RO), Slovakia (SK), Suomi–Finland (FI), Switzerland (CH), Turkey (TR), United Kingdom (GB).

A key related variable is the logarithm of gross domestic product per person in EUR at current prices ("logGDP"). For the purposes of exploratory data analyses it has also been categorised as a binary variable indicating values higher or lower than the median ("Binary GDP"). The employment composition (D = 11) is:

* Primary sector (agriculture, hunting, forestry, fishing, mining, quarrying) * Manufacturing * Energy (electricity, gas and water supply) * Construction * Trade repair transport (wholesale and retail trade, repair, transport, storage, communications) * Hotels restaurants * Financial intermediation * Real estate (real estate, renting and business activities) * Educ admin defense soc sec (education, public administration, defence, social security) * Health social work * Other services (other community, social and personal service activities)

Usage

```
eurostat_employment
```

Format

An object of class `data.frame` with 29 rows and 17 columns.

`foraminiferals` *Paleocological compositions*

Description

The foraminiferal data set (Aitchison, 1986) is a typical example of paleocological data. It contains compositions of 4 different fossils (Neogloboquadrina atlantica, Neogloboquadrina pachyderma, Globorotalia obesa, and Globigerinoides triloba) at 30 different depths. Due to the rounded zeros present in the data set we will apply some zero replacement techniques to impute these values in advance. After data preprocessing, the analysis that should be undertaken is the association between the composition and the depth.

Usage

```
foraminiferals
```

Format

An object of class `data.frame` with 30 rows and 5 columns.

gmean *Geometric Mean*

Description

Generic function for the (trimmed) geometric mean.

Usage

```
gmean(x, zero.rm = FALSE, trim = 0, na.rm = FALSE)
```

Arguments

x	A nonnegative vector.
zero.rm	a logical value indicating whether zero values should be stripped before the computation proceeds.
trim	the fraction (0 to 0.5) of observations to be trimmed from each end of x before the mean is computed. Values of trim outside that range are taken as the nearest endpoint.
na.rm	a logical value indicating whether NA values should be stripped before the computation proceeds.

See Also

center

household_budget *Household budget patterns*

Description

In a sample survey of single persons living alone in rented accommodation, twenty men and twenty women were randomly selected and asked to record over a period of one month their expenditures on the following four mutually exclusive and exhaustive commodity groups: * Hous: Housing, including fuel and light. * Food: Foodstuffs, including alcohol and tobacco. * Serv: Services, including transport and vehicles. * Other: Other goods, including clothing, footwear and durable goods.

Usage

```
household_budget
```

Format

An object of class data.frame with 40 rows and 6 columns.

house_expend *Household expenditures*

Description

From Eurostat (the European Union's statistical information service) the houseexpend data set records the composition on proportions of mean consumption expenditure of households expenditures on 12 domestic year costs in 27 states of the European Union. Some values in the data set are rounded zeros. In addition the data set contains the gross domestic product (GDP05) and (GDP14) in years 2005 and 2014, respectively. An interesting analysis is the potential association between expenditures compositions and GDP. Once a linear regression model is established, predictions can be provided.

Usage

```
house_expend
```

Format

An object of class `data.frame` with 27 rows and 15 columns.

ilr_basis *Isometric/Orthonormal Log-Ratio Basis for Log-Transformed Compositions*

Description

Builds an isometric log-ratio (ilr) basis for a composition with `k+1` parts, also called orthonormal log-ratio (olr) basis.

Usage

```
ilr_basis(dim, type = "default")

olr_basis(dim, type = "default")
```

Arguments

`dim`	An integer indicating the number of components. If a dataframe or matrix is provided, the number of components is inferred from the number of columns. If a character vector specifying the names of the parts is provided the number of component is its length.
`type`	Character string specifying the type of basis to generate. Options are `"pivot"`, `"cdp"`. Any other option will return the Helmert basis defined by Egozcue et al., 2013..

Details

The basis vectors are constructed as:

$$h_i = \sqrt{\frac{i}{i+1}} \log \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}$$

for $i = 1, \ldots, k$.

Setting the `type` parameter to `"pivot"` (pivot balances) or `"cdp"` (codapack balances) allows generating alternative ilr/olr bases.

Value

A matrix representing the orthonormal basis.

References

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). *Isometric logratio transformations for compositional data analysis*. Mathematical Geology, **35**(3), 279–300.

Examples

```
ilr_basis(5)
ilr_basis(alimentation[,1:9])
```

---

kilauea_iki *Chemical Composition of Volcanic Rocks from Kilauea Iki*

Description

This dataset contains the chemical composition of volcanic rocks sampled from the lava lake at Kilauea Iki (Hawaii). The data represents major oxide concentrations in fractional form.

Usage

```
kilauea_iki
```

Format

A data frame with 17 observations and 11 variables:

SiO2 Silicon dioxide (fraction)

TiO2 Titanium dioxide (fraction)

Al2O3 Aluminium oxide (fraction)

Fe2O3 Ferric oxide (fraction)

FeO Ferrous oxide (fraction)

MnO Manganese oxide (fraction)

MgO Magnesium oxide (fraction)

CaO Calcium oxide (fraction)

Na2O Sodium oxide (fraction)

K2O Potassium oxide (fraction)

P2O5 Phosphorus pentoxide (fraction)

Details

The variability in the oxide concentrations is attributed to magnesic olivine fractionation, starting from a single magmatic mass as suggested by Richter & Moore (1966).

Source

Richter, D.H., & Moore, J.G. (1966). Petrology of Kilauea Iki lava lake, Hawaii. *Geological Survey Professional Paper* 537-B.

`mammals_milk` *Mammal's milk*

Description

The mammalsmilk data set contains the percentages of five constituents (W: water, P: protein, F: fat, L: lactose, and A: ash) of the milk of 24 mammals. The data are taken from [Har75].

Usage

```
mammals_milk
```

Format

An object of class `data.frame` with 24 rows and 6 columns.

`milk_cows` *Milk composition study*

Description

In an attempt to improve the quality of cow milk, milk from each of thirty cows was assessed by dietary composition before and after a strictly controlled dietary and hormonal regime over a period of eight weeks. Although seasonal variations in milk quality might have been regarded as negligible over this period, it was decided to have a control group of thirty cows kept under the same conditions but on a regular established regime. The sixty cows were of course allocated to control and treatment groups at random. The 'milk_cows' data set provides the complete set of before and after milk compositions for the sixty cows, showing the protein (pr), milk fat (mf), carbohydrate (ch), calcium (Ca), sodium (Na) and potassium (K) proportions by weight of total dietary content.

Usage

```
milk_cows
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 116 rows and 10 columns.

---

montana *Concentration of minor elements in carbon ashes*

Description

The montana data set consists of 229 samples of the concentration (in ppm) of minor elements [Cr, Cu, Hg, U, V] in carbon ashes from the Fort Union formation (Montana, USA), side of the Powder River Basin. The formation is mostly Palaeocene in age, and the coal is the result of deposition in conditions ranging from fluvial to lacustrine. All samples were taken from the same seam at different sites over an area of 430 km by 300 km, which implies that on average, the sampling spacing is 24 km. Using the spatial coordinates of the data, a semivariogram analysis was conducted for each chemical element in order to check for a potential spatial dependence structure in the data (not shown here). No spatial dependence patterns were observed for any component, which allowed us to assume an independence of the chemical samples at different locations.

The aforementioned chemical components actually represent a fully observed subcomposition of a much larger chemical composition. The five elements are not closed to a constant sum. Note that, as the samples are expressed in parts per million and all concentrations were originally measured, a residual element could be defined to fill up the gap to 10^6.

Usage

```
montana
```

Format

An object of class `data.frame` with 229 rows and 6 columns.

---

pairwise_basis *Pairwise log-ratio generator system*

Description

The function returns all combinations of pairs of log-ratios.

Usage

```
pairwise_basis(dim)
```

Arguments

dim	An integer indicating the number of components. If a dataframe or matrix is provided, the number of components is inferred from the number of columns. If a character vector specifying the names of the parts is provided the number of component is its length.

Value

matrix

---

parliament2017 *Results of catalan parliament elections in 2017 by regions.*

Description

Results of catalan parliament elections in 2017 by regions.

Usage

```
parliament2017
```

Format

A data frame with 42 rows and 9 variables:

com Region

cs Votes to Ciutadans party

jxcat Votes to Junts per Catalunya party

erc Votes to Esquerra republicana de Catalunya party

psc Votes to Partit socialista de Catalunya party

catsp Votes to Catalunya si que es pot party

cup Votes to Candidatura d'unitat popular party

pp Votes to Partit popular party

other Votes to other parties

Source

https://www.idescat.cat/tema/elecc

pb_basis *Isometric log-ratio basis based on Principal Balances.*

Description

Exact method to calculate the principal balances of a compositional dataset. Different methods to approximate the principal balances of a compositional dataset are also included.

Usage

```
pb_basis(
  X,
  method,
  constrained.criterion = "variance",
  cluster.method = "ward.D2",
  ordering = TRUE,
  ...
)
```

Arguments

X	compositional dataset
method	method to be used with Principal Balances. Methods available are: 'exact', 'constrained' or 'cluster'.
constrained.criterion	Criterion used to compare the partition and the principal balance. Either 'variance' (default) or 'angle'.
cluster.method	Method to be used with the hclust function (default: 'ward.D2') or any other method available in hclust function
ordering	should the principal balances found be returned ordered? (first column, first principal balance and so on)
...	parameters passed to hclust function

Value

matrix

References

Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado R. (2018). Advances in Principal Balances for Compositional Data. *Mathematical Geosciencies*, 50, 273-298.

Examples

```
set.seed(1)
X = matrix(exp(rnorm(5*100)), nrow=100, ncol=5)

# Optimal variance obtained with Principal components
(v1 <- apply(coordinates(X, 'pc'), 2, var))
# Optimal variance obtained with Principal balances
(v2 <- apply(coordinates(X,pb_basis(X, method='exact')), 2, var))
# Solution obtained using constrained method
(v3 <- apply(coordinates(X,pb_basis(X, method='constrained')), 2, var))
# Solution obtained using Ward method
(v4 <- apply(coordinates(X,pb_basis(X, method='cluster')), 2, var))

# Plotting the variances
barplot(rbind(v1,v2,v3,v4), beside = TRUE, ylim = c(0,2),
        legend = c('Principal Components','PB (Exact method)',
                   'PB (Constrained)','PB (Ward approximation)'),
        names = paste0('Comp.', 1:4), args.legend = list(cex = 0.8), ylab = 'Variance')
```

---

pc_basis *Isometric log-ratio basis based on Principal Components.*

Description

Different approximations to approximate the principal balances of a compositional dataset.

Usage

```
pc_basis(X)
```

Arguments

X	compositional dataset

Value

matrix

petrafm *Calc-alkaline and tholeiitic volcanic rocks*

Description

This petrafm data set is formed by 100 classified volcanic rock samples from Ontario (Canada). The three parts are:

$[A : Na_2O + K_2O; F : FeO + 0.8998Fe_2O_3; M : MgO]$

Rocks from the calc-alkaline magma series (25) can be well distinguished from samples from the tholeiitic magma series (75) on an AFM diagram.

Usage

```
petrafm
```

Format

An object of class `data.frame` with 100 rows and 4 columns.

plot_balance *Plot a balance*

Description

Plot a balance

Usage

```
plot_balance(B, data = NULL, main = "Balance dendrogram", ...)
```

Arguments

`B`	Balance to plot
`data`	(Optional) Data used to calculate the statistics associated to a balance
`main`	Plot title
`...`	further arguments passed to plot

Value

Balance plot

pollen *Pollen composition in fossils*

Description

The pollen data set is formed by 30 fossil pollen samples from three different locations (recorded in variable group) . The samples were analysed and the 3-part composition [pinus, abies, quercus] was measured.

Usage

```
pollen
```

Format

An object of class `data.frame` with 30 rows and 4 columns.

pottery *Chemical compositions of Romano-British pottery*

Description

The pottery data set consists of data pertaining to the chemical composition of 45 specimens of Romano-British pottery. The method used to generate these data is atomic absorption spectophotometry, and readings for nine oxides (Al2O3, Fe2O3, MgO, CaO, Na2O, K2O, TiO2 , MnO, BaO) are provided. These samples come from five different kiln sites.

Usage

```
pottery
```

Format

An object of class `data.frame` with 45 rows and 11 columns.

read_cdp *Import data from a codapack workspace*

Description

Import data from a codapack workspace

Usage

```
read_cdp(fname)
```

Arguments

fname	cdp file name

sbp_basis *Isometric log-ratio basis based on Balances*

Description

Build an ilr_basis using a sequential binary partition or a generic coordinate system based on balances.

Usage

```
sbp_basis(sbp, data = NULL, fill = FALSE, silent = FALSE)
```

Arguments

sbp	parts to consider in the numerator and the denominator. Can be defined either using a list of formulas setting parts (see examples) or using a matrix where each column define a balance. Positive values are parts in the numerator, negative values are parts in the denominator, zeros are parts not used to build the balance.
data	composition from where name parts are extracted
fill	should the balances be completed to become an orthonormal basis? if the given balances are not orthonormal, the function will complete the balance to become a basis.
silent	inform about orthogonality

Value

matrix

Examples

```
X = data.frame(a=1:2, b=2:3, c=4:5, d=5:6, e=10:11, f=100:101, g=1:2)
sbp_basis(list(b1 = a~b+c+d+e+f+g,
               b2 = b~c+d+e+f+g,
               b3 = c~d+e+f+g,
               b4 = d~e+f+g,
               b5 = e~f+g,
               b6 = f~g), data = X)
sbp_basis(list(b1 = a~b,
               b2 = b1~c,
               b3 = b2~d,
               b4 = b3~e,
               b5 = b4~f,
               b6 = b5~g), data = X)
# A non-orthogonal basis can also be calculated.
sbp_basis(list(b1 = a+b+c~e+f+g,
              b2 = d~a+b+c,
              b3 = d~e+g,
              b4 = a~e+b,
              b5 = b~f,
              b6 = c~g), data = X)
```

---

serprot *Serum proteins*

Description

The 'serprot' data set records the percentages of the four serum proteins from the blood samples of 30 patients. Fourteen patients have one disease (1) and sixteen are known to have another different disease (2). The 4-compositions are formed by the proteins [albumin, pre-albumin, globulin A, globulin B].

Usage

```
serprot
```

Format

An object of class `data.frame` with 36 rows and 7 columns.

statistitian_time *A statistician's time budget*

Description

Time budgets –how a day or a period of work is divided up into different activities have become a popular source of data in psychology and sociology. To illustrate such problems we consider six daily activities undertaken by an academic statistician: teaching (T); consultation (C); administration (A); research (R); other wakeful activities (O); and sleep (S).

The 'statistician_time' data set records the daily time (in hours) devoted to each activity, recorded on each of 20 days, selected randomly from working days in alternate weeks so as to avoid possible carry-over effects such as a short-sleep day being compensated by make-up sleep on the succeeding day. The six activities may be divided into two categories: 'work' comprising activities T, C, A, and R, and 'leisure', comprising activities O and S. Our analysis may then be directed towards the work pattern consisting of the relative times spent in the four work activities, the leisure pattern, and the division of the day into work time and leisure time. Two obvious questions are as follows. To what extent, if any, do the patterns of work and of leisure depend on the times allocated to these major divisions of the day? Is the ratio of sleep to other wakeful activities dependent on the times spent in the various work activities?

Usage

```
statistitian_time
```

Format

An object of class `data.frame` with 20 rows and 7 columns.

variation_array *Variation array is returned.*

Description

Variation array is returned.

Usage

```
variation_array(X, include_means = FALSE)
```

Arguments

X	Compositional dataset
include_means	if TRUE logratio means are included in the lower-left triangle

Value

variation array matrix

Examples

```
set.seed(1)
X = matrix(exp(rnorm(5*100)), nrow=100, ncol=5)
variation_array(X)
variation_array(X, include_means = TRUE)
```

---

waste *The waste composition in Catalonia*

Description

The actual population residing in a municipality of Catalonia is composed by the census count and the so-called floating population (tourists, seasonal visitors, hostel students, short-time employees, and the like). Since actual population combines long and short term residents it is convenient to express it as equivalent full-time residents. Floating population may be positive if the + municipality is receiving more short term residents than it is sending elsewhere, or negative if the opposite holds (expressed as a percentage above –if positive– or below –if negative– the census count). The waste data set includes this information in the variable floating population. Floating population has a large impact on solid waste generation and thus waste can be used to predict floating population which is a hard to estimate demographic variable. This case study was presented in

Usage

```
waste
```

Format

An object of class `data.frame` with 215 rows and 10 columns.

Details

Tourists and census population do not generate the same volume of waste and have different consumption and recycling patterns (waste composition). The Catalan Statistical Institute (IDESCAT) publishes official floating population data for all municipalities in Catalonia (Spain) above 5000 census habitants. The composition of urban solid waste is classified into D = 5 parts: * x1 : non recyclable (grey waste container in Catalonia), * x2 : glass (bottles and jars of any colour: green waste container), * x3 : light containers (plastic packaging, cans and tetra packs: yellow container), * x4 : paper and cardboard (blue container), and * x5 : biodegradable waste (brown container).

References

G. Coenders, J.A.Martín-Fernández and B. Ferrer-Rosell, *When relative and absolute information matter: compositional predictor with a total in generalized linear models*. Statistical Modelling **17**(6) (2017), 494–512.

`weibo_hotels` *Hotel posts in social media*

Description

The 'weibo_hotels' data set aims at comparing the use of Weibo (Facebook equivalent in China) in hospitality e-marketing between small and medium accommodation establishments (private hostels, small hotels) and big and well-established business (such as international hotel chains or large hotels) in China. The 50 latest posts of the Weibo pages of each hotel (n = 10) are content-analyzed and coded regarding the count of posts featuring information on a 4-part composition [facilities, food, events, promotions]. Hotels were coded as large "L" or small "S" in the hotel size categorical variable.

Usage

```
weibo_hotels
```

Format

An object of class `data.frame` with 10 rows and 5 columns.

Index