Package 'coda.base'

March 4, 2026

Type Package

Title A Basic Set of Functions for Compositional Data Analysis

Version 1.0.5

Date 2026-03-03

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 2026-03-04 07:00:02 UTC

Contents

alimentation *Food consumption in European countries*

Description

The ‘alimentation‘ data set contains the percentage composition of food consumption in 25 European countries during the 1980s. The food 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,
- ‘FV‘: fruits and vegetables.

The data set also contains categorical variables indicating whether the country belongs to the North or South/Mediterranean group, and whether it is an Eastern or Western European country.

Usage

```
alimentation
```

Format

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

alr_basis *Additive log-ratio basis*

Description

Construct the transformation matrix associated with additive log-ratio (alr) coordinates.

Usage

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

Arguments

dim	Number of parts. It can be a single integer, a matrix or data frame, or a character vector of part names.
denominator	Part used as denominator. By default, the last part is used.
numerator	Parts used as numerators. By default, all parts except the denominator are used, preserving their original order.

Value

A matrix defining the alr coordinate system.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

Examples

```
alr_basis(5)
alr_basis(5, 3)
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 three-part composition $[sand, silt, clay]$ of 39 sediment samples collected at different water depths in an Arctic lake.

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 related to proteins of the red blood cell plasma membrane. Its inheritance pattern is autosomal with codominance, meaning that the heterozygous phenotype is distinct from both homozygous phenotypes.

The three phenotypes are M, N, and MN. Their frequencies vary across populations. Under the Hardy-Weinberg principle, allele and genotype frequencies remain constant across generations in the absence of evolutionary forces, implying that

$\frac{x_{MM}x_{NN}}{x_{MN}^2} = \frac{1}{4}$

where x_{MM} and x_{NN} are the genotype frequencies of the homozygotes and x_{MN} is the genotype frequency of heterozygotes.

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 in sleep (‘sleep‘), sedentary behaviour (‘sedent‘), light physical activity (‘Lpa‘), moderate physical activity (‘Mpa‘), and vigorous physical activity (‘Vpa‘) for 393 children. The standardized body mass index (‘zBMI‘) of each child is also included.

This data set was used in the example of Dumuid et al. (2019) to examine the expected differences in zBMI associated with reallocations of daily time between sleep, sedentary behaviour, and physical activity. Because the original data are confidential, ‘bmi_activity‘ contains simulated data that mimic the main features of the original study.

Usage

```
bmi_activity
```

Format

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

References

Dumuid, D., Pedisic, Z., Stanford, T. E., Martín-Fernández, J. A., Hron, K., Maher, C., Lewis, L. K., & Olds, T. S. (2019). *The Compositional Isotemporal Substitution 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), 846–857.

---

cc_basis *Canonical-correlation log-ratio basis*

Description

Construct an ilr basis rotated according to canonical correlations between a compositional response data set and an explanatory data set.

Usage

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

Arguments

Y	A compositional data set.
X	An explanatory data set.

Value

A matrix whose columns define a canonical-correlation-oriented ilr basis.

---

cdp_basis *CoDaPack default ilr basis*

Description

Construct the default isometric log-ratio basis used in CoDaPack.

Usage

```
cdp_basis(dim)
```

Arguments

dim	Number of parts. It can be a single integer, a matrix or data frame, or a character vector of part names.

Value

A matrix with D rows and $D - 1$ columns containing the CoDaPack default ilr basis.

Examples

```
cdp_basis(5)
cdp_basis(c("a", "b", "c", "d"))
```

`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

Construct the transformation matrix associated with centered log-ratio (clr) coordinates.

Usage

```
clr_basis(dim)
```

Arguments

dim	Number of parts. It can be a single integer, a matrix or data frame, or a character vector of part names.

Details

CLR coordinates are linearly dependent and lie in the $D - 1$ dimensional clr-plane.

Value

A square matrix defining the clr coordinate system.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

Examples

```
B <- clr_basis(5)
clr_coordinates <- coordinates(c(1, 2, 3, 4, 5), B)
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 of missing values and/or values below the detection limit in compositional data using an EM algorithm assuming normality on the simplex.

Usage

```
coda_replacement(
  X,
  DL = NULL,
  dl_prop = 0.65,
  eps = 1e-04,
  parameters = FALSE,
  debug = FALSE,
  maxit = 500
)
```

Arguments

X	A compositional data set: 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.
eps	Convergence tolerance.
parameters	Logical; if 'TRUE', return additional estimated parameters.
debug	Logical; if 'TRUE', print the log-likelihood at each iteration.
maxit	Maximum number of iterations

Value

If 'parameters = FALSE', a numeric matrix with imputed values. If 'parameters = TRUE', a list with the estimated clr mean, clr covariance, and imputed clr coordinates.

---

composition *Compositions from coordinates with respect to a basis*

Description

Reconstruct a composition from coordinates with respect to a given basis.

Usage

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

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

Arguments

H	Coordinates of a composition. It can be a numeric matrix, a data frame, or a numeric vector.
basis	Basis used to interpret the coordinates. Either a character string naming a predefined basis or a matrix.

Value

A composition corresponding to the given coordinates.

See Also

coordinates, ilr_basis, alr_basis, clr_basis, sbp_basis

`conditional_obasis` *Conditional orthonormal basis*

Description

Compute orthonormal ilr bases associated with conditioning patterns on the parts of a composition.

Usage

```
conditional_obasis(C)
```

Arguments

`C`	A numeric matrix or data frame with one conditioning pattern per row. Columns correspond to parts. For each row, entries equal to '0' define one block and positive entries define the complementary block.

Details

Each row of 'C' defines one conditioning pattern. For a given row, the ilr basis is constructed by separating the parts marked with '0' from the parts marked with a positive value.

If a conditioning row contains 'nz' zeros, then:

- the first 'nz - 1' coordinates describe the internal log-ratio structure of the parts marked with '0',
- the coordinate 'nz' describes the balance between the block of parts marked with '0' and the block of parts marked with positive values,
- the remaining coordinates describe the internal log-ratio structure of the parts marked with positive values.

Thus, each basis preserves the split defined by the conditioning pattern and completes it to an orthonormal basis of the clr-plane.

Value

A three-dimensional array of dimension '(D - 1, D, nrow(C))', where 'D' is the number of parts. Each slice contains one orthonormal ilr basis.

Examples

```
C <- rbind(
  c(0, 0, 1, 1, 0),
  c(0, 1, 0, 1, 0)
)

conditional_obasis(C)

Cdf <- data.frame(
```

```
  a = c(0, 0),
  b = c(0, 1),
  c = c(1, 0),
  d = c(1, 1),
  e = c(0, 0)
)

conditional_obasis(Cdf)
```

`coordinates` *Coordinates of compositions with respect to a basis*

Description

Compute coordinates of a composition or a compositional data set with respect to a given log-ratio basis.

The ‘basis‘ argument can be either:

- a character string identifying a predefined coordinate system, or
- a matrix whose columns define a system of log-contrasts.

The predefined options are:

- ‘"ilr"‘: isometric log-ratio coordinates,
- ‘"olr"‘: orthonormal log-ratio coordinates,
- ‘"clr"‘: centered log-ratio coordinates,
- ‘"alr"‘: additive log-ratio coordinates,
- ‘"pw"‘: pairwise log-ratios,
- ‘"pc"‘: principal component log-ratio coordinates,
- ‘"pb"‘: principal balance coordinates,
- ‘"cdp"‘: CoDaPack default balances.

Usage

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

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

alr_c(X)

clr_c(X)

ilr_c(X)

olr_c(X)
```

Arguments

X	A compositional data set. It can be a numeric matrix, a data frame, or a numeric vector.
basis	Basis used to compute the coordinates. Either a character string naming a predefined basis or a matrix with log-ratio basis vectors in columns.
...	components of the composition

Value

Coordinates of 'X' with respect to the given 'basis'. The returned object has the same general type as the input when possible.

See Also

ilr_basis, alr_basis, clr_basis, sbp_basis, composition

Examples

```
coordinates(1:5)

B <- ilr_basis(5)
coordinates(1:5, B)

X <- rbind(1:5, 2:6)
coordinates(X, "clr")
```

---

dist *Distance Matrix Computation (including Aitchison distance)*

Description

Compute a distance matrix for compositional data, including the Aitchison distance as an extension of dist.

Usage

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

Arguments

x	A data matrix whose rows are compositions.
method	The distance measure to be used. This must be one of "aitchison", "euclidean", "maximum", "manhattan", "canberra", "binary", or "minkowski". Any unambiguous abbreviation can be given.
...	Additional arguments passed to dist.

Value

An object of class "dist".

See Also

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")

# Using Aitchison distance, only relative information is of importance
dist(rbind(c(1, 1, 1), c(100, 100, 100)), method = "ait")
```

---

eurostat_employment *Employment distribution in EUROSTAT countries*

Description

According to the three-sector theory, employment shifts from the primary sector (raw material extraction), to the secondary sector (industry, energy, and construction), and then to the tertiary sector (services) as economies develop. The 'eurostat_employment' data set contains EUROSTAT data on employment, aggregated for both sexes and all ages, distributed by economic activity in 2008 for 29 EUROSTAT member countries.

A related variable is the logarithm of gross domestic product per person in EUR at current prices ('logGDP'). For exploratory purposes, it is also categorised as a binary variable indicating values above or below the median ('Binary GDP').

The employment composition has 11 parts:

- Primary sector
- Manufacturing
- Energy
- Construction
- Trade repair transport
- Hotels restaurants
- Financial intermediation
- Real estate
- Educ admin defense soc sec
- Health social work
- Other services

Usage

```
eurostat_employment
```

Format

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

---

foraminiferals *Paleocological compositions*

Description

The ‘foraminiferals‘ data set (Aitchison, 1986) is a classical example of paleocological compositional data. It contains the composition of four fossil types (Neogloboquadrina atlantica, Neogloboquadrina pachyderma, Globorotalia obesa, and Globigerinoides triloba) at 30 different depths.

Because the data contain rounded zeros, zero-replacement techniques are typically required before analysis. A natural goal is then to study the association between fossil composition and depth.

Usage

```
foraminiferals
```

Format

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

---

gen_coda_with_zeros_and_missings

Generate compositional data with zeros and missing values

Description

Simulate compositional data and optionally introduce structural zeros (interpreted as values below a detection limit) and missing values.

The function first generates a compositional data set ‘X0‘, then creates a modified version ‘X‘ by:

- replacing values below ‘dl_par‘ by zero, if ‘zeros = TRUE‘,
- introducing missing values at random, if ‘missings = TRUE‘.

A matrix of detection limits ‘DL‘ is also returned. It contains ‘dl_par‘ in the positions that were censored to zero, and ‘0‘ elsewhere.

Usage

```
gen_coda_with_zeros_and_missings(
  n,
  d,
  missings = TRUE,
  zeros = TRUE,
  dl_par = 0.05,
  na_p = 0.15
)
```

Arguments

n	Number of observations.
d	Dimension of the latent coordinate space used to generate the compositions.
missings	Logical; if 'TRUE', introduce missing values at random.
zeros	Logical; if 'TRUE', replace values below 'dl_par' by zero.
dl_par	Detection-limit threshold used to generate zeros.
na_p	Probability that any entry is replaced by 'NA' when 'missings = TRUE'.

Details

Compositions are generated from multivariate normal coordinates and mapped to the simplex through 'composition()'. The eigenvector rotation is included to induce a non-trivial covariance structure in the generated coordinates.

Missing values are introduced completely at random, independently for each cell, with probability 'na_p'.

Value

A list with three components:

X The generated compositional data set with simulated zeros and/or missing values.

DL A matrix of detection limits, with 'dl_par' in censored positions and '0' elsewhere.

X0 The original simulated compositional data set before introducing zeros or missing values.

Examples

```
set.seed(123)
sim <- gen_coda_with_zeros_and_missings(100, 4)

str(sim)
summary(sim$X0)
summary(sim$X)
table(sim$X == 0, useNA = "ifany")
```

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 their expenditure over one month in 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

The 'house_expend' data set, obtained from Eurostat, records the composition of mean household consumption expenditure across 12 expenditure categories in 27 European Union countries. Some values are rounded zeros.

In addition, the data set contains gross domestic product values for 2005 ('GDP05') and 2014 ('GDP14'). A relevant analysis is the relationship between expenditure compositions and GDP.

Usage

```
house_expend
```

Format

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

ilr_basis *Isometric and orthonormal log-ratio bases*

Description

Construct an isometric log-ratio (ilr) basis for a composition with D parts. The ilr basis is an orthonormal basis of the clr-plane and provides $D - 1$ coordinates. The same basis is sometimes referred to as an orthonormal log-ratio (olr) basis.

Usage

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

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

Arguments

`dim`	Number of parts. It can be: • a single integer, • a matrix or data frame, in which case the number of columns is used, • a character vector of part names, in which case its length is used.
`type`	Type of ilr basis to construct. Available options are: • '"default"': standard Helmert-type ilr basis, • '"pivot"': pivot balance basis, • '"cdp"': CoDaPack default basis.

Details

For ‘type = "default"‘, the function returns the standard Helmert-type ilr basis. Alternative constructions are available through ‘type = "pivot"‘ and ‘type = "cdp"‘.

The default basis vectors are:

$$h_i = \sqrt{\frac{i}{i+1}} \log \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \qquad i = 1, \ldots, D-1$$

Value

A matrix with D rows and $D - 1$ columns representing an orthonormal log-ratio 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])
ilr_basis(c("a", "b", "c", "d"), type = "pivot")
```

---

kilauea_iki *Chemical composition of volcanic rocks from Kilauea Iki*

Description

The ‘kilauea_iki‘ data set contains the chemical composition of volcanic rocks sampled from the lava lake at Kilauea Iki (Hawaii). The data represent major oxide concentrations in fractional form.

Usage

```
kilauea_iki
```

Format

A data frame with 17 observations and 11 variables:

SiO2 Silicon dioxide

TiO2 Titanium dioxide

Al2O3 Aluminium oxide

Fe2O3 Ferric oxide

FeO Ferrous oxide

MnO Manganese oxide

MgO Magnesium oxide

CaO Calcium oxide

Na2O Sodium oxide

K2O Potassium oxide

P2O5 Phosphorus pentoxide

Details

The variability in oxide concentrations is attributed to magnesian olivine fractionation from a single magmatic mass, as suggested by Richter and 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 *Mammals' milk*

Description

The 'mammals_milk' data set contains the percentages of five constituents of the milk of 24 mammals: $[W, P, F, L, A]$, where 'W' is water, 'P' is protein, 'F' is fat, 'L' is lactose, and 'A' is ash.

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 thirty cows was assessed before and after a controlled dietary and hormonal regime over eight weeks. A control group of thirty cows kept under the usual regime was also included.

The 'milk_cows' data set provides the complete before/after milk composition data for the sixty cows, with the proportions of protein ('pr'), milk fat ('mf'), carbohydrate ('ch'), calcium ('Ca'), sodium ('Na'), and potassium ('K').

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 coal ashes*

Description

The 'montana' data set contains 229 samples of the concentration (in ppm) of five minor elements $[Cr, Cu, Hg, U, V]$ in coal ashes from the Fort Union formation (Montana, USA), in the Powder River Basin.

The five measured elements form a fully observed subcomposition of a much larger chemical composition. Since the data are given in parts per million and all concentrations were measured, a residual component could in principle be added to close the compositions to 10^6.

Usage

```
montana
```

Format

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

---

pairwise_basis *Pairwise log-ratio generating system*

Description

Construct the system of all pairwise log-ratios between parts.

Usage

```
pairwise_basis(dim)
```

Arguments

`dim`	Number of parts. It can be a single integer, a matrix or data frame, or a character vector of part names.

Value

A matrix, or a sparse matrix for large dimensions, whose columns represent all pairwise log-ratio generators.

parliament2017 *Catalan Parliament election results in 2017 by region*

Description

The 'parliament2017' data set contains the results of the 2017 Catalan Parliament election aggregated by region.

Usage

```
parliament2017
```

Format

A data frame with 42 rows and 9 variables:

com Region

cs Votes for the Ciutadans party

jxcat Votes for the Junts per Catalunya party

erc Votes for the Esquerra Republicana de Catalunya party

psc Votes for the Partit Socialista de Catalunya party

catsp Votes for the Catalunya Sí que es Pot party

cup Votes for the Candidatura d'Unitat Popular party

pp Votes for the Partit Popular party

other Votes for other parties

Source

Idescat, statistics on Catalan Parliament elections.

pb_basis *Principal balance basis*

Description

Construct a basis of principal balances for a compositional data set.

Usage

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

Arguments

X	Compositional data set.
method	Method used to construct the principal balances. One of '"exact"', '"constrained"', or '"cluster"'.
constrained.criterion	Criterion used by the constrained method. Either '"variance"' (default) or '"angle"'.
cluster.method	Linkage criterion passed to hclust when 'method = "cluster"'.
ordering	Logical; if 'TRUE', reorder balances by decreasing explained variance.
...	Additional arguments passed to hclust.

Details

Several methods are available:

- '"exact"': exact computation of principal balances,
- '"constrained"': constrained approximation based on a target criterion,
- '"cluster"': approximation based on hierarchical clustering.

Value

A matrix whose columns are principal balances.

References

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

Examples

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

v1 <- apply(coordinates(X, "pc"), 2, var)
v2 <- apply(coordinates(X, pb_basis(X, method = "exact")), 2, var)
v3 <- apply(coordinates(X, pb_basis(X, method = "constrained")), 2, var)
v4 <- apply(coordinates(X, pb_basis(X, method = "cluster")), 2, var)

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` *Principal component log-ratio basis*

Description

Construct an ilr basis rotated according to the principal components of the log-ratio coordinates of a compositional data set.

Usage

```
pc_basis(X)
```

Arguments

`X`	Compositional data set.

Value

A matrix whose columns define a principal-component-oriented ilr basis.

`petrafm` *Calc-alkaline and tholeiitic volcanic rocks*

Description

The 'petrafm' data set contains 100 classified volcanic rock samples from Ontario (Canada). The three-part composition is

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

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

Usage

```
petrafm
```

Format

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

plot_balance *Plot a balance with node labels under horizontal branches*

Description

Plot a balance with node labels under horizontal branches

Usage

```
plot_balance(
  B,
  data = NULL,
  main = "Balance dendrogram",
  summary_fun = NULL,
  cex_node = 0.9,
  offset_node = 0.05,
  ...
)
```

Arguments

B	Balance basis matrix
data	Optional compositional data used to compute balance summaries
main	Plot title
summary_fun	Optional function applied to each balance coordinate vector. It must take a numeric vector and return a character string.
cex_node	Character expansion for node labels
offset_node	Vertical offset below the horizontal branch, relative to max height
...	Further arguments passed to plot

Value

Invisibly returns a data.frame with node coordinates and labels

Examples

```
X = waste[,5:9]
B = pb_basis(X, method = 'exact')

plot_balance(B)

plot_balance(B, data = X,
             summary_fun = function(x){
               q = quantile(x, probs = c(0.25, 0.5, 0.75))
               sprintf("%0.2f [%0.2f-%0.2f]", q[2], q[1], q[3])
             })
```

pollen *Pollen composition in fossils*

Description

The ‘pollen‘ data set contains 30 fossil pollen samples from three different locations (recorded in variable ‘group‘). The measured composition is the three-part composition $[pinus, abies, quercus]$.

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 contains the chemical composition of 45 specimens of Romano-British pottery. The measurements were obtained by atomic absorption spectrophotometry and include nine oxides: Al2O3, Fe2O3, MgO, CaO, Na2O, K2O, TiO2, MnO, and BaO.

The specimens 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 *Basis from a sequential binary partition*

Description

Construct a balance basis from a sequential binary partition (SBP) or from a more general collection of balances.

Usage

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

Arguments

sbp	A list of formulas or a matrix describing balances.
data	Optional compositional data set used to extract part names when 'sbp' is given as a list of formulas.
fill	Logical; if 'TRUE', complete the supplied balances to obtain a full basis.
silent	Logical; if 'FALSE', report whether the resulting balances form a basis, and whether they are orthogonal or orthonormal.

Details

The argument 'sbp' can be specified in two ways:

- as a list of formulas, where each formula defines the numerator and the denominator groups of a balance,
- as a matrix with one column per balance and one row per part. Positive entries indicate parts in the numerator, negative entries indicate parts in the denominator, and zeros indicate unused parts.

Value

A matrix whose columns are balances.

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
)

# Sequential SBP construction
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)

# Chain construction
sbp_basis(list(
  b1 = a ~ b,
  b2 = b1 ~ c,
  b3 = b2 ~ d,
  b4 = b3 ~ e,
  b5 = b4 ~ f,
  b6 = b5 ~ g
), data = X)

# Non-orthogonal system of balances
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)

# Direct construction from a contrast matrix
sbp_basis(cbind(
  c( 1,  1, -1, -1),
  c( 1, -1,  1, -1),
  c( 1, -1, -1,  1)
))
```

`serprot` *Serum proteins*

Description

The ‘serprot‘ data set records the percentages of four serum proteins from blood samples of 30 patients. Fourteen patients have one disease and sixteen have another.

The four-part compositions are formed by $[albumin, pre\text{-}albumin, globulin\ A, globulin\ B]$.

Usage

```
serprot
```

Format

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

---

statistitian_time *A statistician's time budget*

Description

The ‘statistitian_time‘ data set records the daily time budget of an academic statistician across 20 working days. The six activities are teaching (‘T‘), consultation (‘C‘), administration (‘A‘), research (‘R‘), other wakeful activities (‘O‘), and sleep (‘S‘).

These activities may also be grouped into work (‘T‘, ‘C‘, ‘A‘, ‘R‘) and leisure (‘O‘, ‘S‘). The data allow investigation of the relationship between detailed time-allocation patterns and the broader division between work and leisure.

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, ml_covariance = FALSE)
```

Arguments

X	Compositional dataset
include_means	if TRUE logratio means are included in the lower-left triangle
ml_covariance	if TRUE Maximum-likelihood estimation of the covariance for the multivariate normal distribution is used (dividing the scatter matrix by n instead of n-1)

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 *Urban waste composition in Catalonia*

Description

The 'waste' data set studies the relationship between waste composition and floating population in Catalonia. The actual population of a municipality combines census population and floating population (tourists, seasonal visitors, temporary workers, and similar short-term residents), expressed as equivalent full-time residents.

The composition of urban solid waste is classified into five parts:

- 'x1': non-recyclable waste,
- 'x2': glass,
- 'x3': light containers,
- 'x4': paper and cardboard,
- 'x5': biodegradable waste.

Waste generation and composition are influenced by floating population, which makes waste composition a useful predictor of this difficult-to-measure demographic quantity.

Usage

```
waste
```

Format

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

References

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

---

weibo_hotels *Hotel posts in social media*

Description

The 'weibo_hotels' data set compares the use of Weibo (the Chinese equivalent of Facebook) in hospitality e-marketing between small and medium establishments and larger hotel businesses in China.

The 50 latest posts from the Weibo page of each hotel ($n = 10$) were content-analysed and coded into a four-part composition: $[facilities, food, events, promotions]$. Hotels were also classified by size as large ('L') or small ('S').

Usage

```
weibo_hotels
```

Format

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

---

zero_na_conditional_obasis *Conditional orthonormal basis for zeros and missing values*

Description

Compute orthonormal ilr bases adapted to patterns of missing values and structural zeros.

Usage

```
zero_na_conditional_obasis(X)
```

Arguments

`X`	A numeric matrix or data frame with observations in rows and parts in columns.

Details

Each row of ‘X‘ is treated as one observation. For each observation, parts are split into three ordered blocks:

- missing values (‘NA‘),
- zeros,
- strictly positive values.

The resulting basis is constructed so that:

- the first coordinates describe the internal structure of the ‘NA‘ block,
- the next coordinate contrasts the ‘NA‘ block with the positive block,
- the following coordinates describe the internal structure of the zero block,
- the next coordinate contrasts the zero block with the positive block,
- the remaining coordinates describe the internal structure of the positive block.

Value

A three-dimensional array of dimension ‘(D - 1, D, nrow(X))‘, where ‘D‘ is the number of parts. Each slice contains one orthonormal ilr basis.

Examples

```
X <- rbind(
  c(1, NA, 0, 2),
  c(NA, 3, 0, 4),
  c(1, 2, 3, 4)
)

zero_na_conditional_obasis(X)

Xdf <- data.frame(
  a = c(1, NA, 1),
  b = c(NA, 3, 2),
  c = c(0, 0, 3),
  d = c(2, 4, 4)
)

zero_na_conditional_obasis(Xdf)
```

Index