Package ‘zebu’

August 25, 2023

Type Package
Title Local Association Measures
Version 0.2.2.0
Date 2023-08-25
Author Olivier M. F. Martin [aut, cre],
Michel Ducher [aut]
Maintainer Olivier M. F. Martin <oliviermfmartin@tutanota.com>
Description Implements the estimation of local (and global) association measures: Lewontin's D, Ducher's Z, pointwise mutual information, normalized pointwise mutual information and chi-squared residuals. The significance of local (and global) association is accessed using p-values estimated by permutations.
URL https://github.com/oliviermfmartin/zebu
BugReports https://github.com/oliviermfmartin/zebu/issues
License GPL-3
Imports ggplot2, data.table, reshape2, utils, Rcpp
Suggests knitr, rmarkdown, markdown, devtools, usethis, svs, pbapply, testthat
LinkingTo Rcpp
Encoding UTF-8
VignetteBuilder knitr
RoxygenNote 7.2.3
Config/testthat/edition 3
Depends R (>= 2.10)
NeedsCompilation yes
Repository CRAN
Date/Publication 2023-08-25 11:50:02 UTC
Chi-squared test

Description

Chi-squared test: statistical significance of (global) chi-squared statistic and (local) chi-squared residuals

Usage

chisqtest(x, p_adjust = "BH")

Arguments

x lassie S3 object.

p_adjust multiple testing correction method. (see p.adjust.methods for a list of methods).

Value

chisqtest returns an S3 object of class lassie and chisqtest. Adds the following to the lassie object x:

- global_p: global association p-value.
- local_p: array of local association p-values.

See Also

lassie
Examples

# Calling lassie on cars dataset
las <- lassie(cars, continuous = colnames(cars), measure = "chisq")

# Permutation test using default settings
chisqtest(las)

estimate_prob

Estimate marginal and multivariate probabilities

Description

Maximum-likelihood estimation of marginal and multivariate observed and expected independence probabilities. Marginal probability refers to probability of each factor per individual column. Multivariate probability refer to cross-classifying factors for all columns.

Usage

estimate_prob(x)

Arguments

x data.frame or matrix.

Value

List containing the following values:

• margins: a list of marginal probabilities. Names correspond to colnames(x).
• observed: observed multivariate probability array.
• expected: expected multivariate probability array

Examples

# This is what happens behind the curtains in the 'lassie' function
# Here we compute the association between the 'Girth' and 'Height' variables
# of the 'trees' dataset

# 'select' and 'continuous' take column numbers or names
select <- c('Girth', 'Height') # select subset of trees
continuous <- c(1, 2) # both 'Girth' and 'Height' are continuous

# equal-width discretization with 3 bins
breaks <- 3

# Preprocess data: subset, discretize and remove missing data
pre <- preprocess(trees, select, continuous, breaks)

# Estimates marginal and multivariate probabilities from preprocessed data.frame
prob <- estimate_prob(pre$pp)

# Computes local and global association using Ducher's Z
lam <- local_association(prob, measure = 'z')

---

format.lassie  

**Format a lassie object**

**Description**

Formats a lassie object for printing to console (see `print.lassie`) and for writing to a file (see `write.lassie`). Melts probability or local association measure arrays into a data.frame.

**Usage**

```r
## S3 method for class 'lassie'
format(x, what_x, range, what_range, what_sort, decreasing, na.rm, ...)
```

**Arguments**

- **x**
  - lassie S3 object.
- **what_x**
  - vector specifying values to be returned:
    - 'local': local association measure values (default).
    - 'obs': observed probabilities.
    - 'exp': expected probabilities.
    - 'local_p': p-value of local association (after running `permtest` or `chisqtest`).
- **range**
  - range of values to be retained (vector of two numeric values).
- **what_range**
  - character specifying what value range refers to (same options as what_x). By default, takes the first value in what_x.
- **what_sort**
  - character specifying according to which values should x be sorted (same options as what_x). By default, takes the first value in what_x.
- **decreasing**
  - logical value specifying sort order.
- **na.rm**
  - logical value indicating whether NA values should be stripped.
- **...**
  - other arguments passed on to methods. Not currently used.

**See Also**

lassie
**lassie**

**Local Association Measures**

**Description**

Estimates local (and global) association measures: Ducher’s Z, Lewontin’s D, pointwise mutual information, normalized pointwise mutual information and chi-squared residuals.

**Usage**

```r
lassie(x, select, continuous, breaks, measure = "chisq", default_breaks = 4)
```

**Arguments**

- `x` data.frame or matrix.
- `select` optional vector of column numbers or column names specifying a subset of data to be used. By default, uses all columns.
- `continuous` optional vector of column numbers or column names specifying continuous variables that should be discretized. By default, assumes that every variable is categorical.
- `breaks` numeric vector or list passed on to `cut` to discretize continuous variables. When a numeric vector is specified, break points are applied to all continuous variables. In order to specify variable-specific breaks, lists are used. List names identify variables and list values identify breaks. List names are column names (not numbers). If a continuous variable has no specified breaks, then `default_breaks` will be applied.
- `measure` name of measure to be used:
  - 'chisq': Chi-squared residuals.
  - 'd': Lewontin’s D.
  - 'z': Ducher’s 'z'.
  - 'pmi': Pointwise mutual information (in bits).
  - 'npmi': Normalized pointwise mutual information (Bouma).
  - 'npmi2': Normalized pointwise mutual information (Multivariate).
- `default_breaks` default break points for discretizations. Same syntax as in `cut`.

**Value**

An instance of S3 class `lassie` with the following objects:

- `data`: raw and preprocessed data.frames (see `preprocess`).
- `prob` probability arrays (see `estimate_prob`).
- `global` global association (see `local_association`).
- `local` local association arrays (see `local_association`).
- `lassie_params` parameters used in lassie.
See Also

Results can be visualized using `plot.lassie` and `print.lassie` methods. `plot.lassie` is only available in the bivariate case and returns a tile plot representing the probability or local association measure matrix. `print.lassie` shows an array or a data.frame.

Results can be saved using `write.lassie`.

The `permtest` function accesses the significance of local and global association values using p-values estimated by permutations.

The `chisqtest` function accesses the significance in the case of two dimensional chi-squared analysis.

Examples

```r
# In this example, we will use the 'mtcars' dataset

# Selecting a subset of mtcars.
# Takes column names or numbers.
# If nothing was specified, all variables would have been used.
select <- c('mpg', 'cyl') # or select <- c(1, 2)

# Specifying 'mpg' as a continuous variables using column numbers
# Takes column names or numbers.
# If nothing was specified, all variables would have been used.
continuous <-'mpg' # or continuous <- 1

# How should breaks be specified?
# Specifying equal-width discretization with 5 bins for all continuous variables ('mpg')
# breaks <- 5

# Specifying user-defined breakpoints for all continuous variables.
# breaks <- list('mpg' = 5)
# breaks <- list('mpg' = c(10, 15, 25, 30))

# Calling lassie
# Not specifying breaks means that the value in default_breaks (4) will be used.
las <- lassie(mtcars, select = c(1, 2), continuous = 1)

# Print local association to console as an array
print(las)

# Print local association and probabilities
# Here only rows having a positive local association are printed
# The data.frame is also sorted by observed probability
print(las, type = 'df', range = c(0, 1), what_sort = 'obs')

# Plot results as heatmap
```
lassie_get

plot(las)

# Plot observed probabilities using different colors
plot(las, what_x = 'obs', low = 'white', mid = 'grey', high = 'black', text_colour = 'red')

---

lassie_get  

Return the value of 'lassie' object

Description

Subroutine for lassie methods. Tries to retrieve a value from a lassie object and gives an error if value does not exist.

Usage

lassie_get(x, what_x)

Arguments

x  lassie S3 object.

what_x  vector specifying values to be returned:

• 'local': local association measure values (default).
• 'obs': observed probabilities.
• 'exp': expected probabilities.
• 'local_p': p-value of local association (after running permtest or chisqtest).

Value

Corresponding array contained in lassie object.

Examples

las <- lassie(trees)
las_array <- lassie_get(las, 'local')
local_association  Local Association Measures

Description
Subroutines called by lassie to compute local and global association measures from a list of probabilities.

Usage
local_association(x, measure = "chisq", nr = 1)
lewontin_d(x)
duchers_z(x)

Arguments
x  list of probabilities as outputted by estimate_prob.
measure  name of measure to be used:
  • 'chisq': Chi-squared residuals.
  • 'd': Lewontin’s D.
  • 'z': Ducher’s 'z'.
  • 'pmi': Pointwise mutual information (in bits).
  • 'npmi': Normalized pointwise mutual information (Bouma).
  • 'npmi2': Normalized pointwise mutual information (Multivariate).
nr  number of rows/samples. Only used to estimate chi-squared residuals.
normalize  0 for pmi, 1 for npmi, 2 for npmi2

Value
List containing the following values:
  • local: local association array (may contain NA, NaN and Inf values).
  • global: global association numeric value.

See Also
lassie
Examples

# This is what happens behind the curtains in the 'lassie' function
# Here we compute the association between the 'Girth' and 'Height' variables
# of the 'trees' dataset

# 'select' and 'continuous' take column numbers or names
select <- c('Girth', 'Height') # select subset of trees
continuous <- c(1, 2) # both 'Girth' and 'Height' are continuous

# equal-width discretization with 3 bins
breaks <- 3

# Preprocess data: subset, discretize and remove missing data
pre <- preprocess(trees, select, continuous, breaks)

# Estimates marginal and multivariate probabilities from preprocessed data.frame
prob <- estimate_prob(pre$pp)

# Computes local and global association using Ducher's Z
lam <- local_association(prob, measure = 'z')

permtest

Permutation test for local and global association measures

Description

Permutation test: statistical significance of local and global association measures

Usage

permtest(x, nb = 1000L, group = as.list(colnames(x$data$pp)), p_adjust = "BH")

Arguments

x lassie S3 object.

nb number of resampling iterations.

group list of column names specifying which columns should be permuted together. This is useful for the multivariate case, for example, when there is many dependent variables and one independent variable. By default, permutes all columns separately.

p_adjust multiple testing correction method. (see p.adjust.methods for a list of methods).
Value

permtest returns an S3 object of class lassie and permtest. Adds the following to the lassie object x:

- global_p: global association p-value.
- local_p: array of local association p-values.
- global_perm: numeric global association values obtained with permutations.
- local_perm: matrix local association values obtained with permutations. Column number correspond to positions in local association array after converting to numeric (e.g. local_perm[, 1] corresponds to local[1]).
- perm_params: parameters used when calling permtest (nb and p_adjust).

See Also

lassie

Examples

# Calling lassie on cars dataset
las <- lassie(cars, continuous = colnames(cars))

# Permutation test using default settings
permtest(las, nb = 30) # keep resampling low for example

plot.lassie

Plot a lassie object

Description

Plots a lassie object as a tile plot using the ggplot2 package. Only available for bivariate association.

Usage

## S3 method for class 'lassie'
plot(
  x,
  what_x = "local",
  digits = 3,
  low = "royalblue",
  mid = "gainsboro",
  high = "firebrick",
  na = "purple",
  text_colour = "black",
)
Arguments

- **x**  
  lassie S3 object.

- **what_x**  
  vector specifying values to be returned:
  - 'local': local association measure values (default).
  - 'obs': observed probabilities.
  - 'exp': expected probabilities.
  - 'local_p': p-value of local association (after running permtest or chisqtest).

- **digits**  
  integer indicating the number of decimal places.

- **low**  
  colour for low end of the gradient.

- **mid**  
  colour for midpoint of the gradient.

- **high**  
  colour for high end of the gradient.

- **na**  
  colour for NA values.

- **text_colour**  
  colour of text inside cells.

- **text_size**  
  integer indicating text size inside cells.

- **limits**  
  limits of gradient.

- **midpoint**  
  midpoint of gradient.

- **...**  
  other arguments passed on to methods. Not currently used.

See Also

- lassie

Description

Subroutine called by lassie. Discretizes, subsets and remove missing data from a data.frame.

Usage

preprocess(x, select, continuous, breaks, default_breaks = 4)
Arguments

x | data.frame or matrix.
select | optional vector of column numbers or column names specifying a subset of data to be used. By default, uses all columns.
continuous | optional vector of column numbers or column names specifying continuous variables that should be discretized. By default, assumes that every variable is categorical.
breaks | numeric vector or list passed on to cut to discretize continuous variables. When a numeric vector is specified, break points are applied to all continuous variables. In order to specify variable-specific breaks, lists are used. List names identify variables and list values identify breaks. List names are column names (not numbers). If a continuous variable has no specified breaks, then default_breaks will be applied.
default_breaks | default break points for discretizations. Same syntax as in cut.

Value
List containing the following values:

- raw: raw subsetted data.frame
- pp: discretized, subsetted and complete data.frame
- select
- continuous
- breaks
- default_breaks

Examples

# This is what happens behind the curtains in the 'lassie' function
# Here we compute the association between the 'Girth' and 'Height' variables
# of the 'trees' dataset

# 'select' and 'continuous' take column numbers or names
select <- c('Girth', 'Height') # select subset of trees
continuous <- c(1, 2) # both 'Girth' and 'Height' are continuous

# equal-width discretization with 3 bins
breaks <- 3

# Preprocess data: subset, discretize and remove missing data
pre <- preprocess(trees, select, continuous, breaks)

# Estimates marginal and multivariate probabilities from preprocessed data.frame
prob <- estimate_prob(pre$pp)

# Computes local and global association using Ducher's Z
lam <- local_association(prob, measure = 'z')
print.lassie

### Description

Print a `lassie` object as an array or a data.frame.

### Usage

```r
## S3 method for class 'lassie'
print(x, type, what_x, range, what_range, what_sort, decreasing, na.rm, ...)
```

### Arguments

- **x**  
  *lassie* S3 object.

- **type**  
  print style: 'array' for array or 'df' for data.frame.

- **what_x**  
  vector specifying values to be returned:
  - 'local': local association measure values (default).
  - 'obs': observed probabilities.
  - 'exp': expected probabilities.
  - 'local_p': p-value of local association (after running `permtest` or `chisqtest`).

- **range**  
  range of values to be retained (vector of two numeric values).

- **what_range**  
  character specifying what value `range` refers to (same options as `what_x`). By default, takes the first value in `what_x`.

- **what_sort**  
  character specifying according to which values should `x` be sorted (same options as `what_x`). By default, takes the first value in `what_x`.

- **decreasing**  
  logical value specifying sort order.

- **na.rm**  
  logical value indicating whether NA values should be stripped.

- **...**  
  other arguments passed on to methods. Not currently used.

### See Also

`lassie, permtest, chisqtest`
write.lassie

Write a lassie object

Description

Writes lassie object to a file in a table structured format.

Usage

```r
write.lassie(
  x,
  file,
  sep = ",",
  dec = ",.",
  col.names = TRUE,
  row.names = FALSE,
  quote = TRUE,
  ...
)
```

Arguments

- **x** lassie S3 object.
- **file** character string naming a file.
- **sep** the field separator string. Values within each row of `x` are separated by this string.
- **dec** the string to use for decimal points in numeric or complex columns: must be a single character.
- **col.names** either a logical value indicating whether the column names of `x` are to be written along with `x`, or a character vector of column names to be written. See the section on ‘CSV files’ for the meaning of `col.names = NA`.
- **row.names** either a logical value indicating whether the row names of `x` are to be written along with `x`, or a character vector of row names to be written.
- **quote** a logical value (TRUE or FALSE) or a numeric vector. If TRUE, any character or factor columns will be surrounded by double quotes. If a numeric vector, its elements are taken as the indices of columns to quote. In both cases, row and column names are quoted if they are written. If FALSE, nothing is quoted.
- **...** other arguments passed on to write.table.

See Also

lassie, permtest, chisqtest
Description

The zebu package implements the estimation of local (and global) association measures: Ducher’s Z, Lewontin’s D, pointwise mutual information, normalized pointwise mutual information and chi-squared residuals. The significance of local (and global) association is accessed using p-values estimated by permutations.

Functions

lassie estimates local (and global) association measures: Ducher’s Z, Lewontin’s D, pointwise mutual information, normalized pointwise mutual information and chi-squared residuals.

permtest accesses the significance of local (and global) association values using p-values estimated by permutations.

chisqtest accesses the significance for two dimensional chi-squared analysis.

Author(s)

Maintainer: Olivier M. F. Martin <oliviermfmartin@tutanota.com>

Authors:

• Michel Ducher <michel.ducher@chu-lyon.fr>

See Also

Useful links:

• [https://github.com/oliviermfmartin/zebu](https://github.com/oliviermfmartin/zebu)
• Report bugs at [https://github.com/oliviermfmartin/zebu/issues](https://github.com/oliviermfmartin/zebu/issues)
### Index

chisq (local_association), 8  
chisqtest, 2, 2, 4, 6, 7, 11, 13–15  
class, 2, 5, 10  
cut, 5, 12  

duchers_z (local_association), 8  
estimate_prob, 3, 5, 8  
format.lassie, 4  
lassie, 2, 4, 5, 5, 7–11, 13–15  
lassie_get, 7  
lewontin_d (local_association), 8  
local_association, 5, 8  
p.adjust.methods, 2, 9  
permtest, 4, 6, 7, 9, 10, 11, 13–15  
plot.lassie, 6, 10  
pmi (local_association), 8  
preprocess, 5, 11  
print.lassie, 4, 6, 13  
write.lassie, 4, 6, 14  
zebu, 15  
zebu-package (zebu), 15