Package 'FunChisq'

September 8, 2019

Type Package

Version 2.4.8-1

Date 2019-09-08

Title Model-Free Functional Chi-Squared and Exact Tests

Author Yang Zhang [aut],
Hua Zhong [aut] (<https://orcid.org/0000-0003-1962-2603>),
Hien Nguyen [aut],
Ruby Sharma [aut],
Sajal Kumar [aut],
Joe Song [aut, cre] (<https://orcid.org/0000-0002-6883-6547>)

Maintainer Joe Song <joemsong@cs.nmsu.edu>

Description Statistical hypothesis testing methods for
inferring model-free functional dependency using asymptotic
chi-squared or exact distributions. Functional test
statistics are asymmetric and functionally optimal, unique
from other related statistics. Tests in this package reveal
evidence for causality based on the
causality-by-functionality principle. They include
asymptotic functional chi-squared tests
('Zhang & Song' 2013) <arXiv:1311.2707> and an exact
functional test ('Zhong & Song' 2019)
<doi:10.1109/TCBB.2018.2809743>. The normalized functional
chi-squared test was used by Best Performer 'NMSUSongLab'
in HPN-DREAM (DREAM8) Breast Cancer Network Inference
Challenges ('Hill et al' 2016) <doi:10.1038/nmeth.3773>. A
function index ('Zhong & Song' in press) ('Kumar et al' 2018)
<doi:10.1109/BIBM.2018.8621502> derived from the
functional test statistic offers a new effect size measure
for the strength of functional dependency, a better
alternative to conditional entropy in many aspects. For
continuous data, these tests offer an advantage over
regression analysis when a parametric functional form
cannot be assumed; for categorical data, they provide a
novel means to assess directional dependency not possible

with symmetrical Pearson's chi-squared or Fisher's exact tests.

License LGPL (>= 3)

Encoding UTF-8

Depends R (>= 3.0.0)

Imports Rcpp, stats

LinkingTo BH, Rcpp

Suggests Ckmeans.1d.dp, testthat, knitr, rmarkdown

NeedsCompilation yes

URL https://www.cs.nmsu.edu/~joemsong/publications

LazyData TRUE

VignetteBuilder knitr

Repository CRAN

Date/Publication 2019-09-08 16:20:02 UTC

R topics documented:

`FunChisq-package` *Model-Free Functional Chi-Squared and Exact Tests*

Description

Statistical hypothesis testing methods for model-free functional dependency using asymptotic chi-squared or exact distributions. Functional chi-squared test statistics (Zhang and Song, 2013; Zhang, 2014; Nguyen, 2018; Zhong and Song, 2019) are asymmetric, functionally optimal, and model-free, unique from other related statistical measures.

Tests in this package reveal evidence for causality based on the causality-by-functionality principle (Simon and Rescher, 1966). The tests require data from two or more variables be formatted as a contingency table. Continuous variables need to be discretized first, for example, using the R package **Ckmeans.1d.dp**.

The package implements asymptotic functional chi-squared tests (Zhang and Song, 2013; Zhang, 2014), an exact functional test (Zhong and Song, 2019; Nguyen 2018), a comparative functional chi-squared test (Zhang, 2014), and also a comparative chi-squared test (Song et al., 2014; Zhang et al., 2015). The normalized functional chi-squared test was used by Best Performer NMSUSongLab in HPN-DREAM (DREAM8) Breast Cancer Network Inference Challenges (Hill et al., 2016).

A function index derived from the functional chi-squared offers a new effect size measure for the strength of function dependency. It is asymmetrically functionally optimal, different from the symmetric Cramer's V, also a better alternative to conditional entropy in many aspects.

A simulator is provided to generate functional, dependent non-functional, and independent patterns (Sharma et al., 2017).

For continuous data, these tests offer an advantage over regression analysis when a parametric form cannot be reliably assumed for the underlying function. For categorical data, they provide a novel means to assess directional dependency not possible with symmetrical Pearson's chi-squared test, G-test, or Fisher's exact test.

Details

Package:	FunChisq
Type:	Package
Current version:	2.4.8-1
Initial release version:	1.0
Initial release date:	2014-03-08
License:	LGPL (>= 3)

Author(s)

Yang Zhang, Hua Zhong, Hien Nguyen, Ruby Sharma, Sajal Kumar and Joe Song

References

Hill, S. M., Heiser, L. M., Cokelaer, T., Unger, M., Nesser, N. K., Carlin, D. E., Zhang, Y., Sokolov, A., Paull, E. O., Wong, C. K., Graim, K., Bivol, A., Wang, H., Zhu, F., Afsari, B., Danilova, L. V., Favorov, A. V., Lee, W. S., Taylor, D., Hu, C. W., Long, B. L., Noren, D. P., Bisberg, A. J., HPN-DREAM Consortium, Mills, G. B., Gray, J. W., Kellen, M., Norman, T., Friend, S., Qutub, A. A., Fertig, E. J., Guan, Y., Song, M., Stuart, J. M., Spellman, P. T., Koeppl, H., Stolovitzky, G., Saez-Rodriguez, J. and Mukherjee, S. (2016) Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods* **13**(4), 310–318.

Nguyen, H. H. (2018) *Inference of Functional Dependency via Asymmetric, Optimal, and Model-free Statistics.* Unpublished doctoral dissertation, Department of Computer Science, New Mexico State University, Las Cruces, USA.

Sharma, R., Kumar, S., Zhong, H. and Song, M. (2017) Simulating noisy, nonparametric, and multivariate discrete patterns. *The R Journal* **9**(2), 366–377. Retrieved from https://doi.org/10.32614/RJ-2017-053

Simon, H. A. and Rescher, N. (1966) Cause and counterfactual. *Philosophy of Science* **33**(4), 323–340.

Song M., Zhang, Y., Katzaroff, A. J., Edgar, B. A. and Buttitta, L. (2014) Hunting complex differential gene interaction patterns across molecular contexts. *Nucleic Acids Research* **42**(7), e57. Retrieved from https://doi.org/10.1093/nar/gku086

Zhang, Y. (2014) *Nonparametric Statistical Methods for Biological Network Inference.* Unpublished doctoral dissertation, Department of Computer Science, New Mexico State University, Las Cruces, USA.

Zhang, Y., Liu, Z. L. and Song, M. (2015) ChiNet uncovers rewired transcription subnetworks in tolerant yeast for advanced biofuels conversion. *Nucleic Acids Research* **43**(9), 4393–4407. Retrieved from https://doi.org/10.1093/nar/gkv358

Zhang, Y. and Song, M. (2013) Deciphering interactions in causal networks without parametric assumptions. *arXiv Molecular Networks*, arXiv:1311.2707. Retrieved from https://arxiv.org/abs/1311.2707

Zhong, H. and Song, M. (2019) A fast exact functional test for directional association and cancer biology applications. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **16**(3), 818–826. Retrieved from https://doi.org/10.1109/TCBB.2018.2809743

See Also

For data discretization, an option is optimal univariate clustering. See package **Ckmeans.1d.dp**.

For symmetric dependency tests on discrete data, see Pearson's chi-squared test (`chisq.test`), Fisher's exact test (`fisher.test`), mutual information (package **entropy**), and G-test, implemented in packages **DescTools** and **RVAideMemoire**.

`add.noise` *Apply Noise to Discrete-Valued Tables*

Description

The function can apply two types of noise to contingency tables of discrete values. A house noise model is designed for ordinal variables; a candle noise model is for categorical variables. Noise is applied independently for each data point in a table.

Usage

```
add.noise(tables, u, noise.model, margin=0)
add.house.noise(tables, u, margin=0)
add.candle.noise(tables, u, margin=0)
```

Arguments

`tables`	a list of tables or one table. A table can be either a matrix or a data frame of integer values.

`u` a numeric value between 0 and 1 to specify the noise level to be applied to the input tables. See Details.

`noise.model` a character string indicating the noise model of either `"house"` for ordinal variables or `"candle"` for categorical variables. See Details.

`margin` a value of either 0, 1, or 2. Default is 0.

0: noise is applied along both rows and columns in a table. The sum of values in the table is the same before and after noise application.

1: noise is applied along each row. The sum of each row is the same before and after noise application.

2: noise is applied along each column. The sum of each column is the same before and after noise application.

Details

Each noise model defines a conditional probability function of a noisy version given an original discrete value and a noise level. In the house noise model for ordinal variables, defined in (Zhang et al., 2015), the probability decreases as the noisy version deviates from the original ordinal value. The shape of the function is like a pitched house roof. In the candle noise model for categorical variables, the probability of the noisy version for any value other than the original categorical value is the same given the noise level. The function shape is like a candle.

At a minimum level of 0, no noise is applied on the input table(s). A maximum level of 1 indicates that the original sample will be changed to some other values with a probability of 1. For a discrete random variable of two possible values, a noise level of 1 will flip the values and create a non-random pattern; a noise level of 0.5 creates the most random pattern.

Value

If `tables` is a list, the function returns a list of tables with noised applied. If `tables` is a numeric matrix or a data frame, the function returns one table with noise applied.

Author(s)

Hua Zhong, Yang Zhang and Joe Song.

References

Zhang, Y., Liu, Z. L. and Song, M. (2015) ChiNet uncovers rewired transcription subnetworks in tolerant yeast for advanced biofuels conversion. *Nucleic Acids Research* **43**(9), 4393–4407. Retrieved from https://nar.oxfordjournals.org/content/43/9/4393.long

See Also

`simulate_tables`.

Examples

```
# Example 1. Add house noise to a single table

# Create a 4x4 table
t <- matrix(c(3,0,0,0,
              0,2,2,0,
              0,0,0,4,
              3,3,2,0),
            nrow=4, ncol=4, byrow=TRUE)
# Two ways to apply house noise at level 0.1 along both rows
# and columns of the table:
add.noise(t, 0.1, "house", 0)
add.house.noise(t, 0.1, 0)


# Example 2. Add candle noise to a list of tables

# Create a list of tables
t.list <- list(t+5, t*10, t*2)
# Two ways to apply candle noise at level 0.2 along the rows
# of the table:
add.noise(t.list, 0.2, "candle", 1)
add.candle.noise(t.list, 0.2, 1)
```

`cond.fun.chisq.test` *Conditional Functional Chi-Squared Test for Model-Free Conditional Functional Dependency*

Description

Asymptotic chi-squared test to determine the model-free functional dependency of effect variable Y on a cause variable X, conditioned on a third variable Z.

Usage

```
cond.fun.chisq.test(x, y, z=NULL, data=NULL, log.p = FALSE,
                    method = c("fchisq", "nfchisq"))
```

Arguments

`x`	vector or character; either a discrete random variable (cause) represented as vector or a character column name in `data`.
`y`	vector or character; either a discrete random variable (effect) represented as vector or a character column name in `data`.
`z`	vector or character; either a discrete random variable (condition) represented as vector or a character column name in `data`. In case of `NULL` a `fun.chisq.test` on a contingency table, with `x` as row variable and `y` as column variable, is returned. See `?fun.chisq.test` for details. The default is `NULL`.

data	data.frame; a dataframe containing the three variables x, y and z. In case of NULL x, y and z should be vectors. The default is NULL.
log.p	logical; if TRUE, the p-value is given as log(p). Taking the log improves the accuracy when p-value is close to zero. The default is FALSE.
method	a character string to specify the method to compute the conditional functional chi-squared test statistic and its p-value. The options are "fchisq" (default) and "nfchisq". See Details.

Details

Conditional functional chi-squared introduces the concept of conditional functional depedency, where the functional association between two variables (x and y) is tested conditioned on a third variable (z). Two methods are provided to compute the chi-squared statistic and its p-value. When method = "fchisq", the p-value is computed using the chi-squared distribution; when method = "nfchisq" a normalized statistic is obtained by shifting and scaling the original chi-squared statistic and a p-value is computed using the standard normal distribution (Box et al., 2005). The normalized test is more conservative on the degrees of freedom.

Value

A list with class "htest" containing the following components:

statistic	the conditional functional chi-squared statistic if method = "fchisq"; or the normalized conditional functional chi-squared statistic if method = "nfchisq".
parameter	degrees of freedom for the conditionalfunctional chi-squared statistic.
p.value	p-value of the conditional functional test. If method = "fchisq" the p-value is computed by an asymptotic chi-squared distribution; if method = "nfchisq" the p-value is computed by the standard normal distribution.
estimate	an estimate of the conditional function index between 0 and 1. The value of 1 indicates strong functional dependency between x and y, given z. It is asymmetrical with respect to whether x was chosen as the cause of effect y or vice versa.

Author(s)

Sajal Kumar and Mingzhou Song

References

Box, G. E., Hunter, J. S. and Hunter, W. G. (2005) *Statistics for Experimenters: Design, Innovation and Discovery*, 2nd ed., New York: Wiley-Interscience.

Zhang, Y. and Song, M. (2013) Deciphering interactions in causal networks without parametric assumptions. *arXiv Molecular Networks*, arXiv:1311.2707, https://arxiv.org/abs/1311.2707

Zhang, Y. (2014) *Nonparametric Statistical Methods for Biological Network Inference.* Unpublished doctoral dissertation, Department of Computer Science, New Mexico State University, Las Cruces, USA.

Zhong, H. and Song, M. (2018) A fast exact functional test for directional association and cancer biology applications. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* In press. https://doi.org/10.1109/TCBB.2018.2809743

Examples

```
# Generate a relationship between variables X and Z
xz = matrix(c(30,2,2, 2,2,40, 2,30,2),ncol=3,nrow=3,
            byrow = TRUE)
# Re-construct X
x = rep(c(1:nrow(xz)),rowSums(xz))
# Re-construct Z
z = c()
for(i in 1:nrow(xz))
  z = c(z,rep(c(1:ncol(xz)),xz[i,]))

# Generate a relationship between variables Z and Y
# Make sure Z retains its distribution
zy = matrix(c(4,30, 30,4, 4,40),ncol=2,nrow=3,
            byrow = TRUE)
# Re-construct Y
y = rep(0,length(z))
for(i in unique(z))
  y[z==i] = rep(c(1:ncol(zy)),zy[i,])

# Tables
table(x,z)
table(z,y)
table(x,y)

# Conditional functional dependency
# Y = f(X) | Z should be false
cond.fun.chisq.test(x=x,y=y,z=z)
# Z = f(X) | Y should be true
cond.fun.chisq.test(x=x,y=z,z=y)
# Y = f(Z) | X should be true
cond.fun.chisq.test(x=z,y=y,z=x)
```

`cp.chisq.test` *Comparative Chi-Squared Test for Association Heterogeneity*

Description

Comparative chi-squared tests on two or more contingency tables. This test does not consider functional dependencies.

Usage

```
cp.chisq.test(
  x, method=c("chisq", "nchisq", "default", "normalized"),
```

```
  log.p = FALSE
)
```

Arguments

x	a list of at least two matrices representing contingency tables of the same dimensionality.
method	a character string to specify the method to compute the chi-squared statistic and its p-value. The default is "chisq". See Details. Note: "default" and "normalized" are deprecated.
log.p	logical; if TRUE, the p-value is given as log(p). Taking the log improves the accuracy when p-value is close to zero. The default is FALSE.

Details

The comparative chi-squared heterogeneity test determines whether the patterns underlying multiple contingency tables are heterogeneous. Its null test statistic is proved to asymptotically follow the chi-squared distribution (Song et al., 2014; Zhang et al., 2015), different from the widely used chi-squared heterogeneity test (Zar, 2010).

Two methods are provided to compute the chi-squared statistic and its p-value. When method = "chisq" (or "default"), the p-value is computed using the chi-squared distribution; when method = "nchisq" (or "normalized") a normalized statistic is obtained by shifting and scaling the original chi-squared and a p-value is computed using the standard normal distribution (Box et al., 2005). The normalized test is more conservative on the degrees of freedom.

Value

A list with class "htest" containing the following components:

statistic	heterogeneity statistic if method = "chisq" (equivalent to "default"), or normalized statistic if method = "nchisq" (equivalent to "normalized").
parameter	degrees of freedom of the chi-squared statistic.
p.value	p-value of the comparative chi-squared test. By default, it is computed by the chi-squared distribution (method = "chisq" or "default"). If method = "nchisq" (or "normalized"), it is the p-value of the normalized chi-squared statistic using the standard normal distribution.

Author(s)

Joe Song

References

Box, G. E., Hunter, J. S. and Hunter, W. G. (2005) *Statistics for Experimenters: Design, Innovation and Discovery*, 2nd Ed., New York: Wiley-Interscience.

Song M., Zhang Y., Katzaroff A. J., Edgar B. A. and Buttitta L. (2014) Hunting complex differential gene interaction patterns across molecular contexts. *Nucleic Acids Research* **42**(7), e57. Retrieved from https://nar.oxfordjournals.org/content/42/7/e57.long

Zar, J. H. (2010) *Biostatistical Analysis*, 5th Ed., New Jersey: Prentice Hall.

Zhang, Y., Liu, Z. L. and Song, M. (2015) ChiNet uncovers rewired transcription subnetworks in tolerant yeast for advanced biofuels conversion. *Nucleic Acids Research* **43**(9), 4393–4407. Retrieved from https://nar.oxfordjournals.org/content/43/9/4393.long

See Also

For comparative *functional* chi-squared test, `cp.fun.chisq.test`.

Examples

```
## Not run:
x <- matrix(c(4,0,4,0,4,0,1,0,1), 3)
y <- t(x)
z <- matrix(c(1,0,1,4,0,4,0,4,0), 3)
data <- list(x,y,z)
cp.chisq.test(data)
cp.chisq.test(data, method="nchisq")

## End(Not run)
```

`cp.fun.chisq.test` *Comparative Chi-Squared Test for Model-Free Functional Heterogeneity*

Description

Comparative functional chi-squared tests on two or more contingency tables.

Usage

```
cp.fun.chisq.test(
  x, method = c("fchisq", "nfchisq", "default", "normalized"),
  log.p = FALSE
)
```

Arguments

`x`	a list of at least two matrices representing contingency tables of the same dimensionality.
`method`	a character string to specify the method to compute the functional chi-squared statistic and its p-value. The default is `"fchisq"` (equivalent to `"default"`). See Details. Note: `"default"` and `"normalized"` are deprecated.
`log.p`	logical; if `TRUE`, the p-value is given as `log(p)`. Taking the log improves the accuracy when p-value is close to zero. The default is `FALSE`.

Details

The comparative functional chi-squared test determines whether the patterns underlying the contingency tables are heterogeneous in a functional way. Specifically, it evaluates whether the column variable is a changed function of the row variable across the contingency tables.

Two methods are provided to compute the functional chi-squared statistic and its p-value. When `method = "fchisq"` (or `"default"`), the p-value is computed using the chi-squared distribution; when `method = "nfchisq"` (or `"normalized"`) a normalized statistic is obtained by shifting and scaling the original statistic and a p-value is computed using the standard normal distribution (Box et al., 2005). The normalized test is more conservative on the degrees of freedom.

Value

A list with class `"htest"` containing the following components:

`statistic`	functional heterogeneity statistic if `method = "fchisq"` (equivalent to `"default"`), or normalized statistic if `method = "nfchisq"` (equivalent to `"normalized"`).
`parameter`	degrees of freedom.
`p.value`	p-value of the comparative functional chi-squared test. By default, it is computed by the chi-squared distribution. If `method = "normalized"`, it is the p-value of the normalized statistic computed by the standard normal distribution.

Author(s)

Yang Zhang and Joe Song

References

Box, G. E., Hunter, J. S. and Hunter, W. G. (2005) *Statistics for Experimenters: Design, Innovation and Discovery*, 2nd Ed., New York: Wiley-Interscience.

Zhang, Y. (2014) *Nonparametric Statistical Methods for Biological Network Inference.* Unpublished doctoral dissertation, Department of Computer Science, New Mexico State University, Las Cruces, USA.

Zhang, Y. and Song, M. (2013) Deciphering interactions in causal networks without parametric assumptions. *arXiv Molecular Networks*, arXiv:1311.2707. https://arxiv.org/abs/1311.2707

See Also

For comparative chi-squared test that does not consider functional dependencies, `cp.chisq.test`.

Examples

```
## Not run:
x <- matrix(c(4,0,4,0,4,0,1,0,1), 3)
y <- t(x)
z <- matrix(c(1,0,1,4,0,4,0,4,0), 3)
data <- list(x,y,z)
cp.fun.chisq.test(data)
cp.fun.chisq.test(data, method="nfchisq")
```

```
## End(Not run)
```

`Exact Functional Test` *Exact Functional Test*

Description

Perform the exact functional test on a contingency table

Usage

```
EFTDP(nm)
EFTDQP(nm)
```

Arguments

`nm`	A matrix of nonnegative integers representing a contingency table.

Details

The functional exact test is performed using branch-and-bound with two algorithms (DP and DQP) to avoid re-calculation of bounds.

Value

The exact P-value of the test.

Note

The functions provide a direct entry into the C++ implementations of the exact functional test.

Author(s)

Hien Nguyen, Hua Zhong, Joe Song

References

Nguyen, H. H. (2018) *Inference of Functional Dependency via Asymmetric, Optimal, and Model-free Statistics.* Unpublished doctoral dissertation, Department of Computer Science, New Mexico State University, Las Cruces, USA.

See Also

`fun.chisq.test`

Examples

```
x = matrix(c(0, 6, 3, 0, 10, 5, 4, 4, 1), nrow=3)
EFTDQP(x)
EFTDQP(t(x))

EFTDP(x)
EFTDP(t(x))
```

`fun.chisq.test` *Model-Free Functional Chi-Squared and Exact Tests*

Description

Asymptotic chi-squared, normalized chi-squared or exact tests on contingency tables to determine model-free functional dependency of the column variable on the row variable.

Usage

```
fun.chisq.test(
  x,
  method = c("fchisq", "nfchisq",
             "exact", "exact.qp", "exact.dp", "exact.dqp",
             "default", "normalized", "simulate.p.value"),
  alternative = c("non-constant", "all"), log.p=FALSE,
  index.kind = c("conditional", "unconditional"),
  simulate.nruns = 2000,
  exact.mode.bound=TRUE
)
```

Arguments

`x`	a matrix representing a contingency table. The row variable represents the independent variable or all unique combinations of multiple independent variables. The column variable is the dependent variable.
`method`	a character string to specify the method to compute the functional chi-squared test statistic and its p-value. The options are `"fchisq"` (equivalent to `"default"`, the default), `"nfchisq"` (equivalent to `"normalized"`), `"exact"`, `"exact.qp"`, `"exact.dp"`, `"exact.dqp"` or `"simulate.p.value"`. See Details. Note: `"default"` and `"normalized"` are deprecated.
`alternative`	a character string to specify the alternative hypothesis. The options are `"non-constant"` (default, non-constant functions) and `"all"` (all types of functions including constant ones).
`log.p`	logical; if `TRUE`, the p-value is given as `log(p)`. Taking the log improves the accuracy when p-value is close to zero. The default is `FALSE`.
`index.kind`	a character string to specify the kind of function index xi.f to be estimated. The options are `"conditional"` (default) and `"unconditional"`. See Details.

`simulate.nruns` A number to specify the number of tables generated to simulate the null distribution. Default is `2000`. Only used when `method="simulate.p.value"`.

`exact.mode.bound`
logical; if `TRUE`, a fast branch-and-bound algorithm is used for the exact functional test (`method="exact"`). If `FALSE`, a slow brute-force enumeration method is used to provide a reference for runtime analysis. Both options provide the same exact p-value. The default is `TRUE`.

Details

The functional chi-squared test determines whether the column variable is a function of the row variable in contingency table `x` (Zhang and Song, 2013; Zhang, 2014). This function supports three hypothesis testing methods:

`index.kind` specifies the kind of function index to be computed. If the experimental design controls neither the row nor column marginal sums, `index.kind = "unconditional"` (default) is recommended; If the column marginal sums are controlled, `index.kind = "conditional"` is recommended. The choice of `index.kind` affects only the function index xi.f value, but not the test statistic or p-value.

When `method="fchisq"` (equivalent to `"default"`, the default), the test statistic is computed as described in (Zhang and Song, 2013; Zhang, 2014) and the p-value is computed using the chi-squared distribution.

When `method="nfchisq"` (equivalent to `"normalized"`), the test statistic is obtained by shifting and scaling the original test statistic (Zhang and Song, 2013; Zhang, 2014); and the p-value is computed using the standard normal distribution (Box et al., 2005). The normalized chi-squared, more conservative on the degrees of freedom, was used by the Best Performer NMSUSongLab in HPN-DREAM (DREAM8) Breast Cancer Network Inference Challenges.

When `method="exact"`, `"exact.qp"` (quadratic programming), `"exact.dp"` (dynamic programming), or `"exact.dqp"` (dynamic and quadratic programming), an exact functional test is performed. The option of `"exact"` uses `"exact.dqp"`, the fastest method. The methods compute an exact p-value, as described in (Zhong and Song, 2019; Nguyen, 2018).

For the `"exact.qp"` and `"exact.dp"` options, if the sample size is no more than 200 or the average cell count is less than five, and the table size is no more than 10 in either row or column, the exact test will not be called and the asymptotic functional chi-squared test (`method="fchisq"`) is used instead.

For `"exact.dqp"`, the exact functional test will always be performed.

For 2-by-2 contingency tables, the asymptotic test options (`method="fchisq"` or `"nfchisq"`) are recommended to test functional dependency, instead of the exact functional test.

When `method="simulate.p.value"`, a simulated null distribution is used to calculate `p-value`. The null distribution is a multinomial distribution that is the product of two marginal distributions. Like other Monte Carlo based methods, this method is slower but may be more accurate than other methods based on asymptotic distributions.

Value

A list with class `"htest"` containing the following components:

`statistic`	the functional chi-squared statistic if `method = "fchisq"`, `"default"`, or `"exact"`; or the normalized functional chi-squared statistic if `method = "nfchisq"` or `"normalized"`.
`parameter`	degrees of freedom for the functional chi-squared statistic.
`p.value`	p-value of the functional test. If `method = "fchisq"` (or `"default"`), it is computed by an asymptotic chi-squared distribution; if `method = "nfchisq"` (or `"normalized"`), it is computed by the standard normal distribution; if `method = "exact"`, it is computed by an exact hypergeometric distribution.
`estimate`	an estimate of function index between 0 and 1. The value of 1 indicates a strictly mathematical function. It is asymmetrical with respect to transpose of the input contingency table, different from the symmetrical Cramer's V based on the Pearson's chi-squared test statistic.

Author(s)

Yang Zhang, Hua Zhong and Joe Song

References

Box, G. E., Hunter, J. S. and Hunter, W. G. (2005) *Statistics for Experimenters: Design, Innovation and Discovery*, 2nd ed., New York: Wiley-Interscience.

Nguyen, H. H. (2018) *Inference of Functional Dependency via Asymmetric, Optimal, and Model-free Statistics.* Unpublished doctoral dissertation, Department of Computer Science, New Mexico State University, Las Cruces, USA.

Zhang, Y. and Song, M. (2013) Deciphering interactions in causal networks without parametric assumptions. *arXiv Molecular Networks*, arXiv:1311.2707, https://arxiv.org/abs/1311.2707

Zhang, Y. (2014) *Nonparametric Statistical Methods for Biological Network Inference.* Unpublished doctoral dissertation, Department of Computer Science, New Mexico State University, Las Cruces, USA.

Zhong, H. and Song, M. (2019) A fast exact functional test for directional association and cancer biology applications. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **16**(3), 818–826. Retrieved from https://doi.org/10.1109/TCBB.2018.2809743

See Also

For data discretization by optimal univariate k-means clustering, see **Ckmeans.1d.dp**.

For symmetrical dependency tests on discrete data, see Pearson's chi-squared test `chisq.test`, Fisher's exact test `fisher.test`, and mutual information **entropy**.

Examples

```
## Not run:
# Example 1. Asymptotic functional chi-squared test
x <- matrix(c(20,0,20,0,20,0,5,0,5), 3)
fun.chisq.test(x) # strong functional dependency
fun.chisq.test(t(x)) # weak functional dependency

# Example 2. Normalized functional chi-squared test
```

```
x <- matrix(c(8,0,8,0,8,0,2,0,2), 3)
fun.chisq.test(x, method="nfchisq") # strong functional dependency
fun.chisq.test(t(x), method="nfchisq") # weak functional dependency

# Example 3. Exact functional chi-squared test
x <- matrix(c(4,0,4,0,4,0,1,0,1), 3)
fun.chisq.test(x, method="exact") # strong functional dependency
fun.chisq.test(t(x), method="exact") # weak functional dependency

# Example 4. Exact functional chi-squared test on a real data set
#            (Shen et al., 2002)
# x is a contingency table with row variable for p53 mutation and
#   column variable for CIMP
x <- matrix(c(12,26,18,0,8,12), nrow=2, ncol=3, byrow=TRUE)

# Test the functional dependency: p53 mutation -> CIMP
fun.chisq.test(x, method="exact")

# Test the functional dependency CIMP -> p53 mutation
fun.chisq.test(t(x), method="exact")

# Example 5. Asymptotic functional chi-squared test with simulated distribution
x <- matrix(c(20,0,20,0,20,0,5,0,5), 3)
fun.chisq.test(x, method="simulate.p.value")
fun.chisq.test(x, method="simulate.p.value", simulate.n = 1000)

## End(Not run)
```

`plot_table` *Plot a Contingency Table*

Description

The input table is visualized as matrix by shades of a given color for better understanding of the underlying pattern. The values in the table must be real. Thus non-integers or negative numbers are acceptable.

Usage

```
plot_table(table, xlab = "Column", ylab = "Row", col = "green3",
           xaxt = "n", yaxt = "n", main = NULL,
           show.value = TRUE, value.cex = 2,
           highlight=c("row.maxima", "none"),
           highlight.col=col,
           mgp=c(0.5,0,0), mar=c(2,2,3,1.5), ...)
```

Arguments

`table`	A data frame or a matrix.

xlab	The lable of the horizontal axis.
ylab	The lable of the vertical axis.
col	The color corresponding to the maximum value in the table.
xaxt	The style of the horizontal axis. See par.
yaxt	The style of the vertical axis. See par.
main	The title of the plot.
show.value	logical. Show the value of each cell in the table on the plot.
value.cex	Relative magnification factor if values are to be put in the cell.
...	Parameters acceptable to image function in the **graphics** package.
highlight	Specify to highlight row maxima or no highlight. When highlighted, a box is placed around each row maximum.
highlight.col	The color used to highlight a cell in the table.
mgp	The margin (in mex units) for the axis title, labels and line. See par.
mar	The margins of the four sides of the plot. See par.

Author(s)

Joe Song

Examples

```
opar <- par(mfrow=c(2,2))
plot_table(matrix(1:6, nrow=2), col="seagreen2")

plot_table(matrix(rnorm(20), nrow=5), col="orange", show.value=FALSE)

plot_table(matrix(rpois(16, 2), nrow=4), col="cornflowerblue", highlight="none")

plot_table(matrix(rbinom(15, 8, 0.5), nrow=3), col="sienna2", highlight="none")
par(opar)
```

`simulate_tables` *Simulate Noisy Contingency Tables to Represent Diverse Discrete Patterns*

Description

Generate random contingency tables representing various functional, non-functional, dependent, or independent patterns, without specifying a parametric model for the patterns.

Usage

```
simulate_tables(
  n=100, nrow=3, ncol=3,
  type = c("functional", "many.to.one",
           "discontinuous", "independent",
           "dependent.non.functional"),
  noise.model = c("house", "candle"), noise=0.0,
  n.tables=1,
  row.marginal=rep(1/nrow, nrow),
  col.marginal=rep(1/ncol, ncol)
)
```

Arguments

n	a positive integer specifying the sample size to be distributed in each table. For "functional", "many.to.one", and "discontinuous" tables, n must be no less than nrow. For "dependent.non.functional" tables, n must be no less than nrow*ncol. For "independent" tables, n must be a positive integer.
nrow	a positive integer specifying the number of rows in each table. The value must be no less than 2. For "many.to.one" tables, nrow must be no less than 3.
ncol	a positive integer specifying the number of columns in output table. ncol must be no less than 2.
type	a character string to specify the type of pattern underlying the table. The options are "functional" (default), "many.to.one", "discontinuous", "independent", and "dependent.non.functional". See Details.
noise.model	a character string indicating the noise model of either "house" for ordinal variables (Zhang et al., 2015) or "candle" for categorical variables. See add.noise for details.
noise	a numeric value between 0 and 1 specifying the noise level to be added to the table using function add.noise. The noise is applied along the rows of the table, except "independent" tables where noise is applied along both row and coloumn. See add.noise for details.
n.tables	a positive integer value specifying the number of tables to be generated.
row.marginal	a non-negative numeric vector of length nrow specifying row marginal probabilities. The vector is linearly scaled so that the sum is 1. The default is a uniform distribution. For "many.to.one" tables, the length of row.marginal vector must be no less than 3.
col.marginal	a non-negative numeric vector of length ncol specifying column marginal probabilities. The vector is linearly scaled so that the sum is 1. The vector is used only in generating "independent" tables. The default is a uniform distribution.

Details

This function generates five types of table representing different interaction patterns between row and column discrete random variables X and Y. Three of the five types are non-constant functional patterns (Y is a non-constant function of X):

type="functional": Y is a function of X but X may or may not be a function of Y. The samples are distributed using the given row marginal probabilities.

type="many.to.one": Y is a many-to-one function of X but X is not a function of Y. The samples are distributed on the basis of row probabilities.

type="discontinuous": Y is a function of X, where the function value of X must differ from its neighbors. X may or may not be a function of Y. The samples are distributed using the given row marginal probabilities. A discontinuous function forms a contrast with those that are close to constant functions.

The fourth type "dependent.non.functional" is non-functional patterns where X and Y are statistically dependent but not function of each other.

The fifth type "independent" represents patterns where X and Y are statistically independent whose joint probability mass function is the product of their marginal probability mass functions.

Random noise can be optionally applied to the tables using either the house or the candle noise model. See add.noise for details.

The paper by Sharma et al. (2017) provides full mathematical and statistical details of the simulation strategies for the above table types except the "discontinuous" type.

Value

A list containing the following components:

pattern.list	a list of tables containing binary patterns in 0's and 1's. Each table is created by setting all non-zero entries in the corresponding sampled contingency table from sample.list to 1. Each table strictly satisfies the mathematical relationship required for a given pattern type requested, but it does not meet the statistical requirements. As each table represents the truth regarding the mathematical relationship between the row and column variables, they can be used as the ground truth or gold standard for benchmarking.
sample.list	a list of tables satisfying both the mathematical and statistical requirements. These tables are noise free.
noise.list	a list of tables after applying noise to the corresponding tables in sample.list. Each table is the noisy version of the corresponding sampled contingency table. Due to the added noise, each table may no longer strictly satisfy the required mathematical or statistical relationships. These tables are the main output to be used for the evaluation of a discrete pattern discovery algorithm.
pvalue.list	a list of p-values reporting the statistical significance of the generated tables for the required type. When the pattern type specifies a functional relationship, the p-values are computed by the functional chi-square test (Zhang and Song, 2013); otherwise, the Pearson's chi-square test of independence is used to calculate the p-value.

Author(s)

Ruby Sharma, Sajal Kumar, Hua Zhong and Joe Song

References

Sharma, R., Kumar, S., Zhong, H. and Song, M. (2017) Simulating noisy, nonparametric, and multivariate discrete patterns. *The R Journal* **9**(2), 366–377. Retrieved from https://journal.r-project.org/archive/2017/RJ-2017-053/index.html

Zhang, Y., Liu, Z. L. and Song, M. (2015) ChiNet uncovers rewired transcription subnetworks in tolerant yeast for advanced biofuels conversion. *Nucleic Acids Research* **43**(9), 4393–4407. Retrieved from https://nar.oxfordjournals.org/content/43/9/4393.long

Zhang, Y. and Song, M. (2013) Deciphering interactions in causal networks without parametric assumptions. *arXiv Molecular Networks*, arXiv:1311.2707, https://arxiv.org/abs/1311.2707

See Also

add.noise for details of the noise model.

Examples

```
## Not run:
# In all examples, x is the row variable and y is the column
#    variable of a table.

# Example 1. Simulating a noisy function where y=f(x),
#            x may or may not be g(y)

tbls <- simulate_tables(n=100, nrow=4, ncol=5, type="functional",
                noise=0.2, n.tables = 1,
                row.marginal = c(0.3,0.2,0.3,0.2))

par(mfrow=c(2,2))
plot_table(tbls$pattern.list[[1]], main="Ex 1. Functional pattern")
plot_table(tbls$sample.list[[1]], main="Ex 1. Sampled pattern (noise free)")
plot_table(tbls$noise.list[[1]], main="Ex 1. Sampled pattern with 0.2 noise")
plot.new()

# Example 2. Simulating a noisy functional pattern where
#            y=f(x), x may or may not be g(y)

tbls <- simulate_tables(n=100, nrow=4, ncol=5, type="functional",
                noise=0.5, n.tables = 1,
                row.marginal = c(0.3,0.2,0.3,0.2))

par(mfrow=c(2,2))
plot_table(tbls$pattern.list[[1]], main="Ex 2. Functioal pattern", col="seagreen2")
plot_table(tbls$sample.list[[1]], main="Ex 2. Sampled pattern (noise free)", col="seagreen2")
plot_table(tbls$noise.list[[1]], main="Ex 2. Sampled pattern with 0.5 noise", col="seagreen2")
plot.new()


# Example 3. Simulating a noise free many.to.one function where
#            y=f(x), x!=f(y).

tbls <- simulate_tables(n=100, nrow=4, ncol=5, type="many.to.one",
```

```
                noise=0.2, n.tables = 1,
                row.marginal = c(0.4,0.3,0.1,0.2))
par(mfrow=c(2,2))
plot_table(tbls$pattern.list[[1]], main="Ex 3. Many-to-one pattern", col="limegreen")
plot_table(tbls$sample.list[[1]], main="Ex 3. Sampled pattern (noise free)", col="limegreen")
plot_table(tbls$noise.list[[1]], main="Ex 3. Sampled pattern with 0.2 noise", col="limegreen")
plot.new()

# Example 4. Simulating noise-free discontinuous
#   pattern where y=f(x), x may or may not be g(y)

tbls <- simulate_tables(n=100, nrow=4, ncol=5,
                type="discontinuous", noise=0.2,
                n.tables = 1, row.marginal = c(0.2,0.4,0.2,0.2))

par(mfrow=c(2,2))
plot_table(tbls$pattern.list[[1]], main="Ex 4. Discontinuous pattern", col="springgreen3")
plot_table(tbls$sample.list[[1]], main="Ex 4. Sampled pattern (noise free)", col="springgreen3")
plot_table(tbls$noise.list[[1]], main="Ex 4. Sampled pattern with 0.2 noise", col="springgreen3")
plot.new()


# Example 5. Simulating noise-free dependent.non.functional
#            pattern where y!=f(x) and x and y are statistically
#            dependent.

tbls <- simulate_tables(n=100, nrow=4, ncol=5,
                type="dependent.non.functional", noise=0.3,
                n.tables = 1, row.marginal = c(0.2,0.4,0.2,0.2))

par(mfrow=c(2,2))
plot_table(tbls$pattern.list[[1]], main="Ex 5. Dependent.non.functional pattern",
col="sienna2", highlight="none")
plot_table(tbls$sample.list[[1]], main="Ex 5. Sampled pattern (noise free)",
col="sienna2", highlight="none")
plot_table(tbls$noise.list[[1]], main="Ex 5. Sampled pattern with 0.3 noise",
col="sienna2", highlight="none")
plot.new()

# Example 6. Simulating a pattern where x and y are
#            statistically independent.

tbls <- simulate_tables(n=100, nrow=4, ncol=5, type="independent",
                noise=0.3, n.tables = 1,
                row.marginal = c(0.4,0.3,0.1,0.2),
                col.marginal = c(0.1,0.2,0.4,0.2,0.1))

par(mfrow=c(2,2))
plot_table(tbls$pattern.list[[1]], main="Ex 6. Independent pattern",
col="cornflowerblue", highlight="none")
plot_table(tbls$sample.list[[1]], main="Ex 6. Sampled pattern (noise free)",
col="cornflowerblue", highlight="none")
plot_table(tbls$noise.list[[1]], main="Ex 6. Sampled pattern with 0.3 noise",
```

```
col="cornflowerblue", highlight="none")
plot.new()


## End(Not run)
```

`test.interactions` *Functional Chi-Squared Test of Functional Dependency among Many Variables in a Data Set*

Description

Apply functional chi-squared tests on many-to-one combinatorial relationships for functional dependency using multivariate discrete data.

Usage

```
test.interactions(
  x, list.ind.vars, dep.vars, var.names = rownames(x),
  index.kind = c("conditional", "unconditional")
)
```

Arguments

`x`	A numeric matrix or data frame of discrete values. Rows represent variables and columns represent samples. Thus, each row index is a variable index, used by `list.ind.vars` and `dep.vars`.
`list.ind.vars`	A list of numeric or integer vectors, each vector representing independent variable indices in one interaction. Each vector (parents) forms a pair with a dependent variable (child) of the same position in `dep.vars` to represent a many-to-one directional interaction.
`dep.vars`	A numeric vector representing indices of dependent variables (children) in multiple interactions.
`var.names`	Optional. A character vector specifying names of all variables (rows). If not provided, the default is the row names of `x`; or `1:nrow(x)` if `x` does not have row names.
`index.kind`	A character string to specify the kind of function index to return, identical to the same argument in `fun.chisq.test`. The value can be `"unconditional"` (default) or `"conditional"`.

Details

`test.interactions` tests functional dependencies in multiple directional interactions. Each interaction, either one-to-one or many-to-one, is a parents-child pair representing a relationship from

independent variables (parents) to a dependent variable (child). The parents-child pairs are specified in two input arguments `list.ind.vars` (a list of parents for each interaction) and `dep.vars` (vector of children in each interaction).

The function automatically creates contingency tables for interactions of interest, thus convenient to use on multivariate data sets. As the function is implemented in C++ and capable of testing multiple many-to-one interactions in one call, it is much faster than calling the R function `fun.chisq.test` multiple times.

`test.interactions` implements only the `method="fchisq"` option in `fun.chisq.test`.

When a contingency table is created for each interaction, all combinations of unique values of the independent variables (parents) form the rows and the unique values of dependent variable (child) form the columns in the contingency table. The table entries are the counts of the corresponding combination of parent and child values. Either rows or columns with all zero counts are removed from the contingency table before functional chi-squared test is applied.

Value

A data frame with five columns. Each row represents the testing result of each directional interaction. The 1st column is either the indices or names (if `var.names` is not `NULL`) of independent variables (parents); The 2nd column is the indices or names of the dependent variable (child); The 3rd column named `p.value` are p-values; The 4th column named `statistic` is chi-squared values; and the 5th column named `estimate` is the function indices for each interaction.

Author(s)

Hua Zhong and Joe Song

See Also

This function calls functional chi-squared test implemented in C++ and is thus much faster than the R version `fun.chisq.test`.

For data discretization by optimal univariate k-means clustering, see **Ckmeans.1d.dp**.

Examples

```
x <- matrix(
  c(0,0,1,0,1,
    1,0,2,1,0,
    2,2,0,0,0,
    1,2,1,1,2,
    1,0,2,1,2),
  nrow = 5, ncol = 5, byrow = TRUE)

list.ind.vars <-list(
  c(1),c(1),c(1),
  c(2),c(2),c(2),
  c(1,2), c(2,3),
  c(3,4), c(4,5))
dep.vars <- c(
  3,4,5,
```

```
  3,4,5,
  3,4,
  5,1)

# list.ind.vars and dep.vars together specify
#   the following ten interactions:
#   1 -> 3
#   1 -> 4
#   1 -> 5
#   2 -> 3
#   2 -> 4
#   2 -> 5
# 1,2 -> 3
# 2,3 -> 4
# 3,4 -> 5
# 4,5 -> 1

var.names <- paste0("var", 1:5)

test.interactions(
  x = x,
  list.ind.vars = list.ind.vars,
  dep.vars = dep.vars,
  var.names = var.names,
  index.kind = "unconditional")
```

Index