

Package ‘minerva’

July 21, 2017

Version 1.4.7

Date 2017-07-05

Title Maximal Information-Based Nonparametric Exploration for Variable Analysis

Depends R (>= 2.14.0)

Imports parallel

Suggests testthat

Description Wrapper for 'minepy' implementation of Maximal Information-based Nonparametric Exploration statistics (MIC and MINE family). Detailed information of the ANSI C implementation of 'minepy' can be found at <<http://minepy.readthedocs.io/en/latest>>.

URL <https://www.r-project.org>, <http://minepy.readthedocs.io/en/latest>,
<http://www.exploredata.net>

License GPL-3

Author Michele Filosi [aut, cre], Roberto Visintainer [aut], Davide Albanese [aut], Samantha Riccadonna [ctb], Giuseppe Jurman [ctb], Cesare Furlanello [ctb]

Maintainer Michele Filosi <michele.filosi@unitn.it>

Repository CRAN

Date/Publication 2017-07-21 09:41:34 UTC

NeedsCompilation yes

R topics documented:

minerva-package	2
mine	3
Spellman	8
Index	9

minerva-package *The minerva package*

Description

Maximal Information-Based Nonparametric Exploration R Package for Variable Analysis. The package provides the `mine` function allowing the computation of Maximal Information-based Nonparametric Exploration statistics, firstly introduced in D. Reshef et al. (2011) *Detecting novel associations in large datasets*. Science 334, 6062 (<http://www.exploredata.net>). In particular, the package is an R wrapper for the C engine *cmine* (<http://minepy.readthedocs.io/en/latest/>).

Details

Summary:

```

Package:    minerva
Version:    1.4.3
Date:       2014-10-08
Depends:    R >= (2.14.0)
Enhances:   parallel
URL:        http://www.r-project.org,
            http://minepy.readthedocs.io/en/latest/,
            http://www.exploredata.net
License:    GPL-3

```

Index:

```

Spellman      Yeast Gene Expression Dataset
mine          MINE-family statistics
minerva-package The minerva package

```

Author(s)

Michele Filosi [aut, cre], Roberto Visintainer [aut], Davide Albanese [aut], Samantha Riccadonna [ctb], Giuseppe Jurman [ctb], Cesare Furlanello [ctb]

Maintainer: Michele Filosi <filosi@fbk.eu>

References

D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, P. Sabeti. (2011) *Detecting novel associations in large datasets*. Science 334, 6062 (<http://www.exploredata.net>).

D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, C. Furlanello. *cmine, minerva & minepy: a C engine for the MINE suite an its R and Python wrappers*. <http://minepy>.

readthedocs.io/en/latest/

minepy. Maximal Information-based Nonparametric Exploration in C and Python.
(<http://minepy.sourceforge.net>)

mine

MINE family statistics

Description

Maximal Information-Based Nonparametric Exploration (MINE) statistics. `mine` computes the MINE family measures between two variables.

Usage

```
mine(x, y=NULL, master=NULL, alpha=0.6, C=15, n.cores=1,
     var.thr=1e-5, eps=NULL, est="mic_approx", na.rm=FALSE, use="all.obs", ...)
```

Arguments

<code>x</code>	a numeric vector (of size n), matrix or data frame (which is coerced to matrix).
<code>y</code>	NULL (default) or a numeric vector of size n (<i>i.e.</i> , with compatible dimensions to <code>x</code>).
<code>master</code>	an optional vector of indices (numeric or character) to be given when <code>y</code> is not set, otherwise <code>master</code> is ignored. It can be either one column index to be used as reference for the comparison (versus all other columns) or a vector of column indices to be used for computing all mutual statistics. If not specified it is set to <code>1:ncol(x)</code> .
<code>alpha</code>	an optional number of cells allowed in the X -by- Y search-grid. Default value is 0.6 (see Details).
<code>C</code>	an optional number determining the starting point of the X -by- Y search-grid. When trying to partition the x -axis into X columns, the algorithm will start with at most CX <i>clumps</i> . Default value is 15 (see Details).
<code>n.cores</code>	optional number of cores to be used in the computations, when <code>master</code> is specified. It requires the parallel package, which provides support for parallel computing, released with R \geq 2.14.0. Defaults is 1 (<i>i.e.</i> , not performing parallel computing).
<code>var.thr</code>	minimum value allowed for the variance of the input variables, since <code>mine</code> can not be computed in case of variance close to 0. Default value is $1e-5$. Information about failed check are reported in <code>var_thr.log</code> file.
<code>eps</code>	integer in $[0,1]$. If 'NULL' (default) it is set to 1-MIC. It can be set to zero for noiseless functions, but the default choice is the most appropriate parametrization for general cases (as stated in Reshef et al. SOM). It provides robustness.

<code>est</code>	Default value is "mic_approx". With <code>est="mic_approx"</code> the original MINE statistics will be computed, with <code>est="mic_e"</code> the equicharacteristic matrix is evaluated and the <code>mic()</code> and <code>tic()</code> methods will return <code>MIC_e</code> and <code>TIC_e</code> values respectively.
<code>na.rm</code>	boolean. This variable is passed directly to the cor-based functions. See <code>cor</code> for further details.
<code>use</code>	Default value is "all.obs". This variable is passed directly to the cor-based functions. See <code>cor</code> for further details.
<code>...</code>	currently not used.

Details

`mine` is an R wrapper for the C engine *cmine* (<http://minepy.readthedocs.io/en/latest/>), an implementation of Maximal Information-Based Nonparametric Exploration (MINE) statistics. The MINE statistics were firstly detailed in D. Reshef et al. (2011) *Detecting novel associations in large datasets*. Science 334, 6062 (<http://www.exploredata.net>).

Here we recall the main concepts of the MINE family statistics. Let $D = (x, y)$ be the set of n ordered pairs of elements of x and y . The data space is partitioned in an X -by- Y grid, grouping the x and y values in X and Y bins respectively.

The **Maximal Information Coefficient (MIC)** is defined as

$$\text{MIC}(D) = \max_{XY < B(n)} M(D)_{X,Y} = \max_{XY < B(n)} \frac{I^*(D, X, Y)}{\log(\min X, Y)},$$

where $B(n) = n^\alpha$ is the search-grid size, $I^*(D, X, Y)$ is the maximum mutual information over all grids X -by- Y , of the distribution induced by D on a grid having X and Y bins (where the probability mass on a cell of the grid is the fraction of points of D falling in that cell). The other statistics of the MINE family are derived from the mutual information matrix achieved by an X -by- Y grid on D .

The **Maximum Asymmetry Score (MAS)** is defined as

$$\text{MAS}(D) = \max_{XY < B(n)} |M(D)_{X,Y} - M(D)_{Y,X}|.$$

The **Maximum Edge Value (MEV)** is defined as

$$\text{MEV}(D) = \max_{XY < B(n)} \{M(D)_{X,Y} : X = 2 \text{ or } Y = 2\}.$$

The **Minimum Cell Number (MCN)** is defined as

$$\text{MCN}(D, \epsilon) = \min_{XY < B(n)} \{\log(XY) : M(D)_{X,Y} \geq (1 - \epsilon)\text{MIC}(D)\}.$$

More details are provided in the supplementary material (SOM) of the original paper.

The MINE statistics can be computed for two numeric vectors x and y . Otherwise a matrix (or data frame) can be provided and two options are available according to the value of `master`. If `master` is a column identifier, then the MINE statistics are computed for the *master* variable versus

the other matrix columns. If `master` is a set of column identifiers, then all mutual MINE statistics are computed among the column subset. `master`, `alpha`, and `C` refers respectively to the `style`, `exp`, and `c` parameters of the original `java` code. In the original article, the authors state that the default value $\alpha = 0.6$ (which is the exponent of the search-grid size $B(n) = n^\alpha$) has been empirically chosen. It is worthwhile noting that `alpha` and `C` are defined to obtain an heuristic approximation in a reasonable amount of time. In case of small sample size (n) it is preferable to increase `alpha` to 1 to obtain a solution closer to the theoretical one.

Value

The Maximal Information-Based Nonparametric Exploration (MINE) statistics provide quantitative evaluations of different aspects of the relationship between two variables. In particular `mine` returns a list of 5 statistics:

MIC	Maximal Information Coefficient. It is related to the relationship strenght and it can be interpreted as a correlation measure. It is symmetric and it ranges in $[0,1]$, where it tends to 0 for statistically independent data and it approaches 1 in probability for noiseless functional relationships (more details can ben found in the original paper).
MAS	Maximum Asymmetry Score. It captures the deviation from monotonicity. Note that $MAS < MIC$. <i>Note:</i> it can be useful for detecting periodic relationships (unknown frequencies).
MEV	Maximum Edge Value. It measures the closeness to being a function. Note that $MEV \leq MIC$.
MCN	Minimum Cell Number. It is a complexity measure.
MIC-R2	It is the difference between the MIC value and the Pearson correlation coefficient.

When computing `mine` between two numeric vectors `x` and `y`, the output is a list of 5 numeric values. When `master` is provided, `mine` returns a list of 5 matrices having `ncol` equal to `m`. In particular, if `master` is a single value, then `mine` returns a list of 5 matrices having 1 column, whose rows correspond to the MINE measures between the `master` column versus all. Instead if `master` is a vector of `m` indices, then `mine` output is a list of 5 `m-by-m` matrices, whose element `i,j` corresponds to the MINE statistics computed between the `i` and `j` columns of `x`.

Author(s)

Michele Filosi and Roberto Visintainer

Special thanks to:

Davide Albanese, Giuseppe Jurman, Samantha Riccadonna

References

D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, P. Sabeti. (2011) *Detecting novel associations in large datasets*. Science 334, 6062

<http://www.exploredata.net>
 (SOM: Supplementary Online Material at <http://www.sciencemag.org/content/suppl/2011/12/14/334.6062.1518.DC1>)

minepy. Maximal Information-based Nonparametric Exploration in C and Python.
<http://minepy.sourceforge.net>

Examples

```
A <- matrix(runif(50),nrow=5)
mine(x=A, master=1)
mine(x=A, master=c(1,3,5,7,8:10))

x <- runif(10); y <- 3*x+2; plot(x,y,type="l")
mine(x,y)
# MIC = 1
# MAS = 0
# MEV = 1
# MCN = 2
# MIC-R2 = 0

set.seed(100); x <- runif(10); y <- 3*x+2+rnorm(10,mean=2,sd=5); plot(x,y)
mine(x,y)
# rounded values of MINE statistics
# MIC = 0.61
# MAS = 0
# MEV = 0.61
# MCN = 2
# MIC-R2 = 0.13

t <- seq(-2*pi,2*pi,0.2); y1 <- sin(2*t); plot(t,y1,type="l")
mine(t,y1)
# rounded values of MINE statistics
# MIC = 0.66
# MAS = 0.37
# MEV = 0.66
# MCN = 3.58
# MIC-R2 = 0.62

y2 <- sin(4*t); plot(t,y2,type="l")
mine(t,y2)
# rounded values of MINE statistics
# MIC = 0.32
# MAS = 0.18
# MEV = 0.32
# MCN = 3.58
# MIC-R2 = 0.31

# Note that for small n it is better to increase alpha
mine(t,y1,alpha=1)
# rounded values of MINE statistics
# MIC = 1
```

```
# MAS = 0.59
# MEV = 1
# MCN = 5.67
# MIC-R2 = 0.96

mine(t,y2,alpha=1)
# rounded values of MINE statistics
# MIC = 1
# MAS = 0.59
# MEV = 1
# MCN = 5
# MIC-R2 = 0.99

# Some examples from SOM
x <- runif(n=1000, min=0, max=1)

# Linear relationship
y1 <- x; plot(x,y1,type="l"); mine(x,y1)
# MIC = 1
# MAS = 0
# MEV = 1
# MCN = 4
# MIC-R2 = 0

# Parabolic relationship
y2 <- 4*(x-0.5)^2; plot(sort(x),y2[order(x)],type="l"); mine(x,y2)
# rounded values of MINE statistics
# MIC = 1
# MAS = 0.68
# MEV = 1
# MCN = 5.5
# MIC-R2 = 1

# Sinusoidal relationship (varying frequency)
y3 <- sin(6*pi*x*(1+x)); plot(sort(x),y3[order(x)],type="l"); mine(x,y3)
# rounded values of MINE statistics
# MIC = 1
# MAS = 0.85
# MEV = 1
# MCN = 4.6
# MIC-R2 = 0.96

# Circle relationship
t <- seq(from=0,to=2*pi,length.out=1000)
x4 <- cos(t); y4 <- sin(t); plot(x4, y4, type="l",asp=1)
mine(x4,y4)
# rounded values of MINE statistics
# MIC = 0.68
# MAS = 0.01
# MEV = 0.32
# MCN = 5.98
# MIC-R2 = 0.68
```

```
data(Spellman)
res <- mine(Spellman, master=1, n.cores=1)

## Not run: ## example of multicore computation
res <- mine(Spellman, master=1, n.cores=parallel::detectCores()-1)
## End(Not run)
```

Spellman

CDC15 Yeast Gene Expression Dataset

Description

The Spellman dataset provides the gene expression data measured (on a custom platform) in *Saccharomyces cerevisiae* cell cultures that have been synchronized at different points of the cell cycle by using a temperature-sensitive mutation (*cdc15-2*), which arrests cells late in mitosis at the restrictive temperature (it can cause heat-shock).

Usage

Spellman

Format

23 rows x 4382 columns: 4381 transcripts (columns 2:4382) measured at 23 timepoints (column 1).

Source

The original data were published by Spellman and colleagues in *Mol. Biol. Cell* (1998) as the Botstein dataset. Here we include the version of the dataset as processed by Reshef and colleagues for the MINE statistics original article published in *Science* (2011) (details are provided in the supplementary material).

References

- D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, P. Sabeti. (2011) *Detecting novel associations in large datasets*. *Science* 334, 6062 (<http://www.exploredata.net>).
- P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, B. Futcher. (1998) *Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization*. *Mol. Biol. Cell*, 9:12 3273–3297.

Index

*Topic **datasets**

Spellman, 8

*Topic **package**

minerva-package, 2

MAS (mine), 3

mas (mine), 3

MCN (mine), 3

mcn (mine), 3

MEV (mine), 3

mev (mine), 3

MIC (mine), 3

mic (mine), 3

MIC-R2 (mine), 3

mic-r2 (mine), 3

MINE (mine), 3

mine, 2, 3

minerva (minerva-package), 2

minerva-package, 2, 2

Spellman, 2, 8

spellman (Spellman), 8