Package ‘NPRED’

July 25, 2024

Title Predictor Identifier: Nonparametric Prediction

Version 1.1.0

Author Ashish Sharma [aut] (<https://orcid.org/0000-0002-6758-0519>),
Raj Mehrotra [aut],
Sanjeev Jha [aut],
Jingwan Li [aut],
Ze Jiang [aut, cre] (<https://orcid.org/0000-0002-3472-0829>)

Maintainer Ze Jiang <ze.jiang@unsw.edu.au>

Description Partial informational correlation (PIC) is used to identify the meaningful predictors to the response from a large set of potential predictors. Details of methodologies used in the package can be found in Sharma, A., Mehrotra, R. (2014). <doi:10.1002/2013WR013845>, Sharma, A., Mehrotra, R., Li, J., & Jha, S. (2016). <doi:10.1016/j.envsoft.2016.05.021>, and Mehrotra, R., & Sharma, A. (2006). <doi:10.1016/j.advwatres.2005.08.007>.

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 3.4.0)

URL https://github.com/zejiang-unsw/NPRED#readme

BugReports https://github.com/zejiang-unsw/NPRED/issues

Imports stats

Suggests zoo, SPEI, WASP, knitr, ggplot2, synthesis, testthat, bookdown, rmarkdown

RoxygenNote 7.3.2

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2024-07-25 08:50:01 UTC

Contents

`calc.scaleSTDratio` *Calculate the ratio of conditional error standard deviations*

Description

Calculate the ratio of conditional error standard deviations

Usage

```
calc.scaleSTDratio(x, zin, zout)
```

Arguments

`x`	A vector of response.
`zin`	A matrix containing the meaningful predictors selected from a large set of possible predictors (z).
`zout`	A matrix containing the remaining possible predictors after taking out the meaningful predictors (zin).

Value

The STD ratio.

References

Sharma, A., Mehrotra, R., 2014. An information theoretic alternative to model a natural system using observational information alone. Water Resources Research, 50(1): 650-660.

`data.gen.ar1` *Generate predictor and response data.*

Description

Generate predictor and response data.

Usage

```
data.gen.ar1(nobs, ndim = 9)
```

Arguments

`nobs`	The data length to be generated.
`ndim`	The number of potential predictors (default is 9).

Value

A list of 2 elements: a vector of response (x), and a matrix of potential predictors (dp) with each column containing one potential predictor.

Examples

```
# AR1 model from paper with 9 dummy variables
data.ar1 <- data.gen.ar1(500)
stepwise.PIC(data.ar1$x, data.ar1$dp)
```

`data.gen.ar4` *Generate predictor and response data.*

Description

Generate predictor and response data.

Usage

```
data.gen.ar4(nobs, ndim = 9)
```

Arguments

`nobs`	The data length to be generated.
`ndim`	The number of potential predictors (default is 9).

Value

A list of 2 elements: a vector of response (x), and a matrix of potential predictors (dp) with each column containing one potential predictor.

Examples

```
# AR4 model from paper with total 9 dimensions
data.ar4 <- data.gen.ar4(500)
stepwise.PIC(data.ar4$x, data.ar4$dp)
```

---

data.gen.ar9 *Generate predictor and response data.*

Description

Generate predictor and response data.

Usage

```
data.gen.ar9(nobs, ndim = 9)
```

Arguments

nobs	The data length to be generated.
ndim	The number of potential predictors (default is 9).

Value

A list of 2 elements: a vector of response (x), and a matrix of potential predictors (dp) with each column containing one potential predictor.

Examples

```
# AR9 model from paper with total 9 dimensions
data.ar9 <- data.gen.ar9(500)
stepwise.PIC(data.ar9$x, data.ar9$dp)
```

`data1` *Sample data : AR9 model: x(i)=0.3*x(i-1)-0.6*x(i-4)-0.5*x(i-9)+eps*

Description

A dataset containing 500 rows (data length) and 16 columns. The first column is response data and the rest columns are possible predictors.

Usage

```
data(data1)
```

`data2` *Sample data : AR4 model: x(i)=0.6*x(i-1)-0.4*x(i-4)+eps*

Description

A dataset containing 500 rows (data length) and 16 columns. The first column is response data and the rest columns are possible predictors.

Usage

```
data(data2)
```

`data3` *Sample data : AR1 model: x(i)=0.9*x(i-1)+0.866*eps*

Description

A dataset containing 500 rows (data length) and 16 columns. The first column is response data and the rest columns are possible predictors.

Usage

```
data(data3)
```

knn *Modified k-nearest neighbour conditional bootstrap or regression function estimation with extrapolation*

Description

Modified k-nearest neighbour conditional bootstrap or regression function estimation with extrapolation

Usage

```
knn(
  x,
  z,
  zout,
  k = 0,
  pw,
  reg = TRUE,
  nensemble = 100,
  tailcorrection = TRUE,
  tailprob = 0.25,
  tailfac = 0.2,
  extrap = TRUE
)
```

Arguments

x	A vector of response.
z	A matrix of existing predictors.
zout	A matrix of predictor values the response is to be estimated at.
k	The number of nearest neighbours used. The default value is 0, indicating Lall and Sharma default is used.
pw	A vector of partial weights of the same length of z.
reg	A logical operator to inform whether a conditional expectation should be output or not nensemble, Used if reg=F and represents the number of realisations that are generated Value.
nensemble	An integer the specifies the number of ensembles used. The default is 100.
tailcorrection	A logical value, T (default) or F, that denotes whether a reduced value of k (number of nearest neighbours) should be used in the tails of any conditioning plane. Whether one is in the tails or not is determined based on the nearest neighbour response value.
tailprob	A scalar that denotes the p-value of the cdf (on either extreme) the tailcorrection takes effect. The default value is 0.25.

tailfac	A scalar that specifies the lowest fraction of the default k that can be used in the tails. Depending on the how extreme one is in the tails, the actual k decreases linearly from k (for a p-value greater than tailprob) to tailfac*k proportional to the actual p-value of the nearest neighbour response, divided by tailprob. The default value is 0.2.
extrap	A logical value, T (default) or F, that denotes whether a kernel extraplation method is used to predict x.

Value

A matrix of responses having same rows as zout if reg=T, or having nensemble columns is reg=F.

References

Sharma, A., Tarboton, D.G. and Lall, U., 1997. Streamflow simulation: A nonparametric approach. Water resources research, 33(2), pp.291-308.

Sharma, A. and O'Neill, R., 2002. A nonparametric approach for representing interannual dependence in monthly streamflow sequences. Water resources research, 38(7), pp.5-1.

Examples

```
data(data1) # AR9 model   x(i)=0.3*x(i-1)-0.6*x(i-4)-0.5*x(i-9)+eps
x <- data1[, 1] # response
py <- data1[, -1] # possible predictors
ans.ar9 <- stepwise.PIC(x, py) # identify the meaningful predictors and estimate partial weights
z <- py[, ans.ar9$cpy] # predictor matrix
pw <- ans.ar9$wt # partial weights

# vector denoting where we want outputs, can be a matrix representing grid.
zout <- apply(z, 2, mean)

knn(x, z, zout, reg = TRUE, pw = pw) # knn regression estimate using partial weights.

knn(x, z, zout, reg = FALSE, pw = pw) # alternatively, knn conditional bootstrap (100 realisations).
# Mean of the conditional bootstrap estimate should be
# approximately the same as the regression estimate.

zout <- ts(data.gen.ar9(500, ndim = length(ans.ar9$cpy))$dp) # new input
xhat1 <- xhat2 <- x
xhat1 <- NPRED::knn(x, z, zout, k = 5, reg = TRUE, extrap = FALSE) # without extrapolation
xhat2 <- NPRED::knn(x, z, zout, k = 5, reg = TRUE, extrap = TRUE) # with extrapolation

ts.plot(ts(x), ts(xhat1), ts(xhat2),
  col = c("black", "red", "blue"), ylim = c(-5, 5),
  lwd = c(2, 2, 1)
)
```

knnregl1cv *Leave one out cross validation.*

Description

Leave one out cross validation.

Usage

```
knnregl1cv(x, z, k = 0, pw)
```

Arguments

x	A vector of response.
z	A matrix of predictors.
k	The number of nearest neighbours used. The default is 0, indicating Lall and Sharma default is used.
pw	A vector of partial weights of the same length of z.

Value

A vector of L1CV estimates of the response.

References

Lall, U., Sharma, A., 1996. A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. Water Resources Research, 32(3): 679-693.

Sharma, A., Mehrotra, R., 2014. An information theoretic alternative to model a natural system using observational information alone. Water Resources Research, 50(1): 650-660.

pic.calc *Calculate PIC*

Description

Calculate PIC

Usage

```
pic.calc(X, Y, Z = NULL)
```

Arguments

`X`	A vector of response.
`Y`	A matrix of new predictors.
`Z`	A matrix of pre-existing predictors that could be NULL if no prior predictors exist.

Value

A list of 2 elements: the partial mutual information (pmi), and partial informational correlation (pic).

References

Sharma, A., Mehrotra, R., 2014. An information theoretic alternative to model a natural system using observational information alone. Water Resources Research, 50(1): 650-660.

Galelli S., Humphrey G.B., Maier H.R., Castelletti A., Dandy G.C. and Gibbs M.S. (2014) An evaluation framework for input variable selection algorithms for environmental data-driven models, Environmental Modelling and Software, 62, 33-51, DOI: 10.1016/j.envsoft.2014.08.015.

`pw.calc` *Calculate Partial Weight*

Description

Calculate Partial Weight

Usage

```
pw.calc(x, py, cpy, cpyPIC)
```

Arguments

`x`	A vector of response.
`py`	A matrix containing possible predictors of x.
`cpy`	The column numbers of the meaningful predictors (cpy).
`cpyPIC`	Partial informational correlation (cpyPIC).

Value

A vector of partial weights(pw) of the same length of z.

References

Sharma, A., Mehrotra, R., 2014. An information theoretic alternative to model a natural system using observational information alone. Water Resources Research, 50(1): 650-660.

`stepwise.PIC` *Calculate stepwise PIC*

Description

Calculate stepwise PIC

Usage

```
stepwise.PIC(x, py, nvarmax = 100, alpha = 0.1)
```

Arguments

`x`	A vector of response.
`py`	A matrix containing possible predictors of x.
`nvarmax`	The maximum number of variables to be selected.
`alpha`	The significance level used to judge whether the sample estimate in Equation $$\hat{PIC} = sqrt(1 - exp(-2\hat{PI})$$ is significant or not. A default alpha value is 0.1.

Value

A list of 2 elements: the column numbers of the meaningful predictors (cpy), and partial informational correlation (cpyPIC).

References

Sharma, A., Mehrotra, R., 2014. An information theoretic alternative to model a natural system using observational information alone. Water Resources Research, 50(1): 650-660.

Examples

```
data(data1) # AR9 model   x(i)=0.3*x(i-1)-0.6*x(i-4)-0.5*x(i-9)+eps
x <- data1[, 1] # response
py <- data1[, -1] # possible predictors
stepwise.PIC(x, py)

data(data2) # AR4 model:  x(i)=0.6*x(i-1)-0.4*x(i-4)+eps
x <- data2[, 1] # response
py <- data2[, -1] # possible predictors
stepwise.PIC(x, py)

data(data3) # AR1 model  x(i)=0.9*x(i-1)+0.866*eps
x <- data3[, 1] # response
py <- data3[, -1] # possible predictors
stepwise.PIC(x, py)
```

`Sydney` *Sample data: Data over Sydney region*

Description

A dataset containing Rainfall (15 stations), NCEP and CSIRO (7 atmospheric variables).

Usage

```
data(Sydney)
```

Index