Package 'Density.T.HoldOut'

February 19, 2015

Encoding UTF-8

Type Package

Title Density.T.HoldOut: Non-combinatorial T-estimation Hold-Out for density estimation

Version 2.00

Date 2014-07-11

Author Nelo Magalhães (Univ. Paris Sud 11 - INRIA team Select) and Yves Rozenholc (Univ. Paris Descartes - INRIA team Select)

Maintainer Nelo Magalhães <nelo.moltermagalhaes@gmail.com>

Description Implementation in the density framework of the non-combinatorial algorithm and its greedy version, introduced by Magalhães and Rozenholc (2014), for T-estimation Hold-Out proposed in Birgé (2006, Section 9). The package provide an implementation which uses several families of estimators (regular and irregular histograms, kernel estimators) which may be used alone or combined. As a complement, provides also a comparison with other Held-Out derived from least-squares and maximum-likelihood. This package implements also the T-estimation Hold-Out derived from the test introduced in Baraud (2011).

License GPL (>= 2)

Imports histogram

NeedsCompilation no

Repository CRAN

Date/Publication 2014-09-20 07:17:29

R topics documented:

`Density.T.HoldOut` *Density.T.HoldOut: Non-combinatorial T-estimation Hold-Out for density estimation*

Description

Implementation in the density framework of the non-combinatorial algorithm and its greedy version, introduced by Magalhães and Rozenholc (2014), for T-estimation Hold-Out proposed in Birgé (2006, Section 9). The package provide an implementation which uses several families of estimators (regular and irregular histograms, kernel estimators) which may be used alone or combined. As a complement, provides also a comparison with other Held-Out derived from least-squares and maximum-likelihood.

Details

Package:	Density.T.HoldOut
Type:	Package
Version:	1.00
Date:	2014-01-08
License:	GPL(>=2)

Author(s)

Nelo Magalhães and Yves Rozenholc.

Maintainer: Nelo Magalhães <nelo.moltermagalhaes@gmail.com>

References

L. Birgé, "Model selection via testing: an alternative to (penalized) maximum likelihood estimators.", Ann. Institut Henri Poincaré Probab. et Statist., 42, 273–325, (2006)

N. Magalhães and Y. Rozenholc, "A non-combinatorial algorithm for T-estimation.", (2014)

`DensityTestim` *Non-combinatorial T-estimation Hold-Out for density estimation.*

Description

Main function:

Estimation of the density of a given sample by a Hold-Out procedure derived from the T-estimation using the algorithms introduced in Magalhães and Rozenholc (2014).

The sample is divided into one training sample, used to build a set of potential estimators via the TBuildList function, and one validation sample, used to select one estimator from this set using T-estimation as introduced in Birgé(2006).

Usage

```
DensityTestim(X,p=1/2,family=NULL,test=c('birge','baraud'),theta=1/4,
last=c('full','training'),plot=TRUE,verbose=TRUE,wlegend=TRUE,kerneltab=NULL,
Dmax=NULL,bwtab=NULL,do.MLHO=FALSE,do.LSHO=FALSE,start=c('LSHO','MLHO'),csqrt=1,
H2dist=NULL,allImageX2=NULL,flist=NULL,...)
```

Arguments

X	numeric vector. The sample to which the density T-estimation Hold-Out procedure is applied.
p	proportion of the sample used in the training sample X[1:ceiling(p*n)] to build the family of estimators. Default 1/2.
family	estimator family name(s). If family is NULL (default), use family = c("Kernel", "RegularHisto", "IrregularHisto", "Parametric").
test	either 'birge' (default) or 'baraud'. Controls the test used in the T-estimation. Default value 'birge' implements T-estimation as introduced in Birgé (2006) while 'baraud' use its modified version using the test derived from Baraud (2011).
theta	parameter which controls the radius of test balls. Has to be smaller than 1/2 (cf. Magalhães and Rozenholc (2014)). Default 1/4.
last	either 'training' or 'full' (default) controlling if the resulting estimator is build with the training sample only or the full sample.
plot	logical (default TRUE), controls if plot are displayed.
verbose	logical (default TRUE), controls if the estimator description is printed.
wlegend	logical (default TRUE); controls if a legend is written on the plot.
kerneltab	vector of all desired kernel types. Only required when 'family' contains 'Kernel'. If NULL (default), use kerneltab = "epanechnikov".
Dmax	maximum number of bins. Only required when 'family' contains 'RegularHisto' or 'IrregularHisto'. If NULL (default), use Dmax=ceiling(n/log(n)).
bwtab	vector of bandwidth values. Only required when the family argument contains 'Kernel'. If NULL (default), use bwtab = diff(range(X))/2/(ceiling(n/log(n)):1).
do.MLHO	logical (default FALSE). If TRUE, the Maximum Likelihood Hold-Out is computed.
do.LSHO	logical (default FALSE). If TRUE, the Least-Squares Hold-Out is computed.
start	starting point of the algorithm, either 'LSHO' (default) or 'MLHO'.

csqrt	numeric (Default 1). If 0 the exact T-estimation is computed. Otherwise a faster but approximate T-estimator is computed based on estimators separated by an Hellinger distance larger than c/sqrt((1-p)*n). (See Magalhães and Rozenholc (2014) for more details)
H2dist	not documented. Only for simulation purpose.
allImageX2	not documented. Only for simulation purpose.
flist	not documented. Only for simulation purpose.
...	for other options when plot is TRUE, as in the plot function.

Details

More details about the algorithm and its implementation may be found in Magalhães and Rozenholc (2014).

Value

DensityTestim returns a list with components

THO	descriptor of the T-Hold-Out estimate.
MLHO	descriptor of the Maximum Likelihood Hold-Out estimate if do.MLHO=TRUE
LSHO	descriptor of the Least-Squares Hold-Out estimate if do.LSHO=TRUE
M	number of considered estimators
comput	number of tests needed to select the T-Hold-Out
total	M*(M-1)/2
H2dist	not documented. Only for simulation purpose.
allImageX2	not documented. Only for simulation purpose.
flist	not documented. Only for simulation purpose.

Moreover if plot=TRUE, the chosen estimator is plotted together with the one chosen by the LSHO (default).

Author(s)

Nelo Magalhães and Yves Rozenholc.

References

N. Magalhães and Y. Rozenholc, "A non-combinatorial algorithm for T-estimation Hold-Out" (2014)

L. Birgé, "Model selection via testing: an alternative to (penalized) maximum likelihood estimators.", Ann. Institut Henri Poincaré Probab. et Statist., 42, 273–325, (2006)

See Also

TBuildList, TBuildRegularHisto, TBuildIrregularHisto, TBuildKernel, TBuildParametric

Examples

```
## Not run:

### load the package
library(Density.T.HoldOut)

### Estimation of the beta density with parameters 5 and 2 from a sample of size 1000:
X=rbeta(1000,5,2)
DensityTestim(X)
x = seq(min(X),max(X),l=500)
lines(x,dbeta(x,5,2),col='green',lty=3)
title('T-estimation and Least-Squares Held-Out')


### Estimation of the lognormal density from a sample of size 500 via a set of regular
### histograms and parametric estimators build with 3/4 of the sample,
### provide as final estimator the one build with the training sample only:
X=rlnorm(500)
DensityTestim(X,p=3/4,family=c('RegularHisto','Parametric'),last=c('partial'))
x = seq(min(X),max(X),l=500)
lines(x,dlnorm(x),col='green',lty=3)
title('T-estimation and Least-Squares Held-Out')


### Estimation of the chi-square density with 5 degrees of freedom from a sample of
### size 250 via a set of regular and irregular histograms and kernel estimators with
### triangular and epanechnikov kernels, start from the maximum likelihood HO estimator:
X=rchisq(250,5)
DensityTestim(X,family=c('RegularHisto','IrregularHisto','Kernel'),
kerneltab=c('triangular','epanechnikov'),start=c('MLHO'))
x = seq(min(X),max(X),l=500)
lines(x,dchisq(x,5),col='green',lty=3)
title('T-estimation and Max. Likelihood Hold-Out')


### Estimation of a normal mixture from a sample of size 1000 via a set of kernel
### estimators, provide also the maximum likelihood HO estimator:
n=ceiling(runif(1)*1000)
X=c(rnorm(n,mean=5,sd=0.1),rnorm(1000-n))
DensityTestim(X,family=c('Kernel'),do.MLHO=TRUE)
x = seq(min(X),max(X),l=500)
lines(x,n/1000*dnorm(x,mean=5,sd=0.1)+(1000-n)/1000*dnorm(x),col='green',lty=3)
title('T-estimation, Least-Squares and Max. Likelihood Hold-Out')


### Estimation of the gaussian density from a sample of size 500 via a set of regular
### and irregular histograms estimators, start from the maximum likelihood HO estimator,
### uses the greedy version with constant 1/16:
X=rnorm(500)
DensityTestim(X,family=c('RegularHisto','IrregularHisto'),start=c('MLHO'),csqrt=1/16)
x = seq(min(X),max(X),l=500)
lines(x,dnorm(x),col='green',lty=3)
```

```
title('T-estimation and Max. Likelihood Hold-Out')

## End(Not run)
```

`TBuildIrregularHisto` *Irregular histogram estimator list constructor.*

Description

Given a sample X, builds the family of irregular histograms using the procedure described in the 'histogram' package.

Usage

```
TBuildIrregularHisto(X,n=length(X),Dmax=NULL,greedyfirst=TRUE,
grid=c("data","regular","quantiles"),breaks=NULL,verbose=FALSE)
```

Arguments

`X`	numeric vector. The sample to build the irregular histograms.
`n`	size of the sample.
`Dmax`	maximum number of bins allowed.
`greedyfirst`	logical; if TRUE (default), a greedy procedure is used to recursively build a finest partition as provided by 'histogram'.
`grid`	if type="irregular", grid chooses the set of possible partitions of the data range. The default value "data" gives a set of partitions constructed from the data points, "regular" uses a fine regular grid of points as possible break points. A regular quantile grid can be chosen using "quantiles".
`breaks`	controls the maximum number of bins allowed in a regular histogram, or the size of the finest grid in an irregular histogram when grid is set to "regular" or "quantiles".
`verbose`	logical; if TRUE (default), some information is given during histogram construction and the resulting histogram object is printed.

Details

We refer to the 'histogram' package for more details about the construction of these irregular histograms and for the different options available.

Value

TBuildIrregularHisto returns a list with components:

`f`	gives the explicit expression of the corresponding density function,
`cuts`	gives the vector of break points (the same as breaks in histo),
`descript`	list containing:

- 'histo'nature of the estimator,
- Dnumber of bins of the estimator,
- breaksvector of break points.

Author(s)

Nelo Magalhães and Yves Rozenholc.

References

'histogram' R-package: http://cran.r-project.org/web/packages/histogram/index.html.

Y. Rozenholc, T. Mildenberger and U. Gather: "Combining Regular and Irregular Histograms by Penalized Likelihood", Computational Statistics and Data Analysis, 54(12), 3313-3323 (2010).

N. Magalhães and Y. Rozenholc, "A non-combinatorial algorithm for T-estimation", (2014).

See Also

DensityTestim, TBuildList, TBuildRegularHisto, TBuildKernel, TBuildParametric

Examples

```
## Not run:
## build irregular histograms for a sample of the normal density:
TBuildIrregularHisto(X=rnorm(1000))

## End(Not run)
```

TBuildKernel *Kernel estimator list constructor.*

Description

Given a sample X, builds the family of kernel estimators using generic function 'density'.

Usage

```
TBuildKernel(X, n = length(X), bwtab = NULL, kerneltab = NULL)
```

Arguments

X	numeric vector. The sample to build the kernel estimators.
n	size of the sample.
bwtab	numeric vector of bandwidths. If NULL (default) bwtab = diff(range(X))/2/(ceiling(n/log(n)):1).
kerneltab	vector of kernel types. Should be available types of the generic function 'density'. If NULL (default) kerneltab = "epanechnikov".

Details

We refer to the 'density' function in the 'stats' package for more details about the different options available.

Value

TBuildKernel returns a list with components

`f`	gives the explicit expression of the corresponding density function,
`cuts`	gives the vector of break points,
`descript`	list containing:

- 'kernel'nature of the estimator,
- kerneltype of kernel,
- bwvalue of the bandwidth.

Author(s)

Nelo Magalhães and Yves Rozenholc.

References

'density' generic function in the 'stats' library: http://127.0.0.1:14946/library/stats/html/density.html

N. Magalhães and Y. Rozenholc, "A non-combinatorial algorithm for T-estimation", (2014).

See Also

DensityTestim, TBuildList, TBuildRegularHisto, TBuildIrregularHisto, TBuildParametric

Examples

```
## Not run:
## build epanechnikov and triangular kernel estimators for a sample of the gamma density:
## with parameters 2 and 3:
TBuildKernel(X=rgamma(1000,2,3),kerneltab=c('epanechnikov','triangular'))

## End(Not run)
```

TBuildList *Construction of a list of regular and irregular histograms, kernel and parametric estimators.*

Description

Given a sample X, builds a list containing all the desired estimators with specified smoothing parameters.

Usage

```
TBuildList(X,family=c("Kernel","RegularHisto","IrregularHisto","Parametric"),
kerneltab=NULL,bwtab=NULL,Dmax=NULL,Dtab=NULL)
```

Arguments

X	numeric vector. The sample to build the estimators.
family	vector of estimator types. Default family=c("Kernel", "RegularHisto", "IrregularHisto", "Parametric"). Has to be a subset of the default value.
kerneltab	vector of kernel types, required when the family argument contains 'Kernel'. Should be available types of the generic function 'density'. If NULL (default) kerneltab="epanechnikov".
bwtab	numeric vector of bandwidths, required when the family argument contains 'Kernel'. If NULL (default) bwtab = diff(range(X))/2/(ceiling(n/log(n)):1).
Dmax	maximum number of bins, required when the family argument contains 'RegularHisto' or 'IrregularHisto'. If NULL (default) Dmax=ceiling(n/log(n)).
Dtab	vector of number of bins. See TBuildRegularHisto. If NULL (default) Dtab=1:Dmax.

Value

TBuildList returns a list containing all constructed estimators, each one consisting in a descriptor list.

Author(s)

Nelo Magalhães and Yves Rozenholc.

References

N. Magalhães and Y. Rozenholc, "A non-combinatorial algorithm for T-estimation", (2014).

See Also

DensityTestim, TBuildRegularHisto, TBuildIrregularHisto, TBuildKernel, TBuildParametric

Examples

```
## Not run:
## list of estimators containing the regular histograms with number of bins varying
## between 1 and 150, kernel estimators using triangular kernel and parametric estimators
## build from a normal sample of size 1000:
TBuildList(X=rnorm(1000),family = c("Kernel", "RegularHisto","Parametric"),
kerneltab = 'triangular',Dmax=150)

## list of estimators containing irregular histograms and kernel estimators
## with bandwidths 2^-j, j=1:,...,ceiling(log(length(X))), build from an exponential
## sample X of size 500:
TBuildList(X=rexp(500),family = c("IrregularHisto","Kernel"),
bwtab=2^-c(1:ceiling(log(length(X)))))

## End(Not run)
```

`TBuildParametric` *Parametric estimator list constructor.*

Description

Given a sample X, builds the family of the parametric estimators for the unif, norm, lnorm, beta, exp, gamma and chisq densities of the 'stats' library obtained from the method of moments.

Usage

```
TBuildParametric(X, namelist = NULL)
```

Arguments

`X`	numeric vector. The sample to build the parametric estimators.
`namelist`	list of all desired parametric estimators. If NULL (default) namelist=c("unif", "norm", "exp", "lnorm", "gamma","chisq", "beta"). Selected list has to be a subset of the default list.

Details

For a sample following an unknown density, all estimators are build in the following way. When: name='unif', the one-bin histogram is built, name='norm', the parametric guassian estimator is built, name='lnorm', the parametric log-normal estimator is built only if min(X)>0, name='exp', the parametric exponential estimator is built only if min(X)>0, name='gamma', the parametric gamma estimator is built only if min(X)>0, name='chisq', the parametric chi-square estimator is built only if min(X)>0, name='beta', the parametric beta estimator is built only if min(X)>0 and max(X)<1.

Value

TBuildParametric returns a list with components:

f	explicit expression of the corresponding density function,
cuts	range of the estimator, given by the 1e-6 -quantiles,
descript	list containing:

- 'parametric'nature of the estimator,
- namename of the estimator,
- parvalues of the estimated parameters.

Author(s)

Nelo Magalhães and Yves Rozenholc.

References

N. Magalhães and Y. Rozenholc, "A non-combinatorial algorithm for T-estimation", (2014).

See Also

DensityTestim, TBuildList, TBuildRegularHisto, TBuildIrregularHisto, TBuildKernel

Examples

```
## Not run:
## build log-normal and exponential parametric estimators for a sample of
## the log-normal density:
TBuildParametric(X=rlnorm(1000),namelist=c('exp','lnorm'))

## End(Not run)
```

TBuildRegularHisto *Regular histogram estimator list constructor.*

Description

Given a sample X, builds the family of regular histograms using the procedure described in the 'histogram' package.

Usage

```
TBuildRegularHisto(X, n = length(X), Dmax = NULL, Dtab = NULL)
```

Arguments

X	numeric vector. The sample to build the regular histograms.
n	size of the sample.
Dmax	maximum number of bins. If NULL (default) Dmax=ceiling(n/log(n)).
Dtab	vector of number of bins. If NULL (default) Dtab=1:Dmax.

Details

We refer to the 'histogram' package for more details about the different options available.

Value

TBuildRegularHisto returns a list with components:

f	gives the explicit expression of the corresponding density function,
cuts	gives the vector of break points (the same as breaks in histo),
descript	list containing:

- 'histo'nature of the estimator,
- Dnumber of bins of the estimator,
- breaksvector of break points.

Author(s)

Nelo Magalhães and Yves Rozenholc.

References

'histogram' R-package: http://cran.r-project.org/web/packages/histogram/index.html

Y. Rozenholc, T. Mildenberger and U. Gather: "Combining Regular and Irregular Histograms by Penalized Likelihood", Computational Statistics and Data Analysis, 54(12), 3313-3323 (2010).

N. Magalhães and Y. Rozenholc, "A non-combinatorial algorithm for T-estimation", (2014).

See Also

DensityTestim, TBuildList, TBuildIrregularHisto, TBuildKernel, TBuildParametric

Examples

```
## Not run:
## build regular histograms for a sample of the lognormal density:
TBuildRegularHisto(X=rlnom(1000))

## End(Not run)
```

Index