Package 'essHist'

March 10, 2019

Type Package

Title The Essential Histogram

Version 1.2.0

Date 2019-03-07

Author Housen Li [aut, cre],
Hannes Sieling [aut]

Maintainer Housen Li <housen.li@outlook.com>

Description Provide an optimal histogram, in the sense of probability density estimation and features detection, by means of multiscale variational inference. For details see Li, Munk, Sieling and Walther (2016) <arXiv:1612.07216>.

Depends R (>= 2.15.3)

License GPL-3

LazyData TRUE

Imports Rcpp (>= 0.12.5)

LinkingTo Rcpp

NeedsCompilation yes

Repository CRAN

Date/Publication 2019-03-10 10:52:47 UTC

R topics documented:

essHist-package 2
checkHistogram 3
Essential Histogram 5
Generate Intervals 7
Mixed normals 8
Multiscale Quantiles 9

Index **12**

essHist-package *The Essential Histogram*

Description

Provide an optimal histogram, in the sense of probability density estimation and features detection, by means of multiscale variational inference. For details see Li, Munk, Sieling and Walther (2016) <arXiv:1612.07216>.

Details

Package:	essHist
Type:	Package
Version:	1.2.0
Date:	2019-03-07
License:	GPL-3

Index:

```
essHistogram    Compute the essential histogram
checkHistogram  Check any estimator by the multiscale confidence set

genIntv         Generate the system of intervals
msQuantile      Simulate the quantile of multiscale statistics

dmixnorm        Compute density function of Gaussian mixtures
pmixnorm        Compute distribution function of Gaussian mixtures
rmixnorm        Generate random number of Gaussian mixtures
paramExample    Output detailed parameters for some famous examples
```

Author(s)

Housen Li [aut, cre], Hannes Sieling [aut]

Maintainer: Housen Li <housen.li@outlook.com>

References

Li, H., Munk, A., Sieling, H., and Walther, G. (2016). The essential histogram. arXiv:1612.07216

Examples

```
# Simulate data
set.seed(123)
n = 300
y = rnorm(n)
```

```
# Compute the essential histogram
eh = essHistogram(y, plot = FALSE)

# Plot results
#     compute oracle density
x  = seq(min(y), max(y), length.out = n)
od = dnorm(x)
#     compare with orcle density
plot(x, od, type = "l", xlab = NA, ylab = NA, col = "red")
lines(eh)
legend("topright", c("Oracle density", "Essential histogram"),
       lty = c(1,1), col = c("red", "black"))

############################################################

# Evaluate other method e.g. R default histogram function
# Data: mixture of Gaussians 1/3 N(0,0.5) + 1/3 N(5,1) + 1/3 N(15,2)
set.seed(123)
n = 300
y = rmixnorm(n, mean = c(0, 5, 15), sd = c(0.5, 1, 2))

# Oracle density
sy = sort(y)
ho = dmixnorm(sy, mean = c(0, 5, 15), sd = c(0.5, 1, 2))

# R default histogram
h  = hist(y, plot = FALSE)

# Check R default histogram to local multiscale constriants
b = checkHistogram(h, y)
lines(sy, ho, col = "red")
legend("topright", c("R-Histogram", "Truth"), col = c("black", "red"), lty = c(1,1))
```

`checkHistogram` *Check any histogram estimator by means of the multiscale confidence set*

Description

Give the locations (i.e. intervals) where the multiscale constraint is violated, and the change-points that are removable.

Usage

```
checkHistogram(h, y, alpha = 0.1, q = NA, intv = genIntv(length(y)),
               plot = TRUE, verbose = TRUE, xlim = range(y),
               ylim = NULL, xlab = "", ylab = "", yaxt = "n", ...)
```

Arguments

h	a numeric vector specifying values of a histogram at sample points or a hitogram class object (i.e. the return value of hist).
y	a numeric vector containing the data.
alpha	significance level; if q is missing, q is chosen as the (1-alpha)-quantile of the null distribution of the multiscale statistic via Monte Carlo simulation, see also msQuantile.
q	threshold of the multiscale constraint.
intv	a data frame provides the system of intervals on which the multiscale statistic is defined. The data frame constains the following two columns left left index of an interval right right index of an interval By default, it is set to the sparse interval system proposed by Rivera and Walther (2013), see also Li et al. (2016).
plot	logical. If TRUE, the input estimator is potted, together with evaluation information. More precisely, at the very bottom, intervals where local constaints are violated are plotted. In the middle short vertical lines that indicate possibly removable change-points are drawn above a light blue horizontal line. Right below the light blue line, it plots a horizontal gray scale strap, the darkness of which reflects the number of violation intervals covering a given location, as a summary of violation information.
verbose	logical. If TRUE (default) it prints some details about the computation; otherwise nothing is printed.
xlim, ylim	numeric vectors of length 2 (default xlim = range(y), ylim = NULL): see plot.
xlab	a title for the x axis (default empty string): see title and plot.
ylab	a title for the y axis (default empty string): see title and plot.
yaxt	A character which specifies the y axis type (default "n"): see par.
...	further arguments and graphical parameters passed to plot (if plot = TRUE).

Value

A data frame provides the intervals where the corresponding local side constraint is violated; an empty data frame if there is no violation. The data frame constains the following four columns

leftIndex	left index of an interval
rightIndex	right index of an interval
leftEnd	left end point of an interval
rightEnd	right end point of an interval

Note

The argument intv is internally adjusted via function .validInterval to ensure it is reasonable, and contains no empty intervals in case of tied observations. Only the intervals on which the input histogram is constant will be checked!

References

Li, H., Munk, A., Sieling, H., and Walther, G. (2016). The essential histogram. arXiv:1612.07216.

See Also

essHistogram, genIntv, msQuantile

Examples

```
set.seed(123)
# Data: mixture of Gaussians 1/3 N(0,0.5) + 1/3 N(5,1) + 1/3 N(15,2)
n  = 500
y  = rmixnorm(n, mean = c(0, 5, 15), sd = c(0.5, 1, 2))

# Oracle density
sy = sort(y)
ho = dmixnorm(sy, mean = c(0, 5, 15), sd = c(0.5, 1, 2))

# R default histogram
h  = hist(y, plot = FALSE)

# Check R default histogram to local multiscale constriants
b = checkHistogram(h, y)
lines(sy, ho, col = "red")
legend("topright", c("R-Histogram", "Truth"), col = c("black", "red"), lty = c(1,1))
```

`Essential Histogram` *The Essential Histogram*

Description

Compute the essential histogram via (pruned) dynamic programming.

Usage

```
essHistogram(x, alpha = 0.5, q = NA, intv = genIntv(length(x)), plot = TRUE,
             verbose = TRUE, xname = deparse(substitute(x)), ...)
```

Arguments

`x`	a numeric vector containing the data.
`alpha`	significance level; if `q` is missing, `q` is chosen as the (1-`alpha`)-quantile of the null distribution of the multiscale statistic via Monte Carlo simulation, see also `msQuantile`.
`q`	threshold value.

intv	a data frame provides the system of intervals on which the multiscale statistic is defined. The data frame constains the following two columns left left index of an interval right right index of an interval By default, it is set to the sparse interval system proposed by Rivera and Walther (2013), see also Li et al. (2016).
plot	logical. If TRUE (default), a histogram is plotted, otherwise a list of breaks and counts is returned. In the latter case, a warning is used if (typically graphical) arguments are specified that only apply to the plot = TRUE case.
verbose	logical. If TRUE (default) it prints some details about the computation; otherwise nothing is printed.
xname	a character string with the actual x argument name.
...	further arguments and graphical parameters passed to plot.histogram and thence to title and axis (if plot = TRUE).

Details

The essential histogram is defined as the histogram with least blocks within the multiscale constraint. The one with highest likelihood is picked if there are more than one solutions. The essential histogram involves only one parameter q, the threshold of the multiscale constraint. Such a parameter can be chosen by means of the significance level alpha, which leads to nature statistical significance statements for the multiscale constraint. The computational complexity is often linear in terms of sample size, although the worst complexity bound is quadratic up to a log-factor in case of the sparse interval system. See Li et al. (2016) for further details.

Value

An object of class "histogram", which is of the same class as returned by function hist.

Note

The argument intv is internally adjusted via function .validInterval to ensure it is reasonable, and contains no empty intervals in case of tied observations. The first block is a closed interval, and the rest blocks are left open right closed intervals.

References

Li, H., Munk, A., Sieling, H., and Walther, G. (2016). The essential histogram. arXiv:1612.07216.

Rivera, C., & Walther, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. Scand. J. Stat. 40, 752–769.

See Also

checkHistogram, genIntv, hist, msQuantile

Examples

```
# simulate data
set.seed(123)
n = 300
x = sort(rnorm(n))

# compute the essential histogram
eh = essHistogram(x, xname = "Gauss")
lines(x, dnorm(x), col='red')
legend('topright',c('Essential Histogram', 'True Desity'),
       col=c('black','red'), lty = c(1,1))
```

`Generate Intervals` *Generate the system of intervals*

Description

Generate the system of intervals on which the multiscale statistic is defined, see Li et al. (2016).

Usage

```
genIntv(n, type = c("Sparse", "Full"))
```

Arguments

`n`	number of observations.
`type`	type of interval system. `type = "Sparse"` (default) is the sparse system proposed by Rivera and Walther (2013), see also Li et al. (2016). `type = "Full"` is the system of all possible intervals with end-index ranging from 1 to `n`.

Value

A data frame provides the system of intervals, and consists two columns

`left`	left index of an interval
`right`	right index of an interval

References

Li, H., Munk, A., Sieling, H., and Walther, G. (2016). The essential histogram. arXiv:1612.07216.

Rivera, C., & Walther, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. Scand. J. Stat. 40, 752–769.

See Also

`checkHistogram`, `essHistogram`, `msQuantile`

Examples

```
n    = 5
intv = genIntv(n,"Full")
print(intv)
```

Mixed normals *The mixture of normal distributions*

Description

Density, distribution function and random generation for the mixture of normals with each component specified by mean and sd, and mixture weights by prob. paramExample gives detailed parameters for some examples specified by type.

Usage

```
dmixnorm(x, mean, sd, prob = rep(1/length(mean),length(mean)), type = NULL, ...)
pmixnorm(x, mean, sd, prob = rep(1/length(mean),length(mean)), type = NULL, ...)
rmixnorm(n, mean, sd, prob = rep(1/length(mean),length(mean)), type = NULL)
paramExample(type)
```

Arguments

x	vector of locations.
n	integer; number of observations.
mean	vector of means for each mixture component.
sd	vector of standard deviations for each mixture component.
prob	vector of prior probability for each mixture component (i.e. mixture weights).
type	a (case insensitive) character string of example name; It includes examples from Marron & Wand (1992): "MW1", ..., "MW15", or equivalently "guass", "skewed_unimodal", "strong_skewed", "kurtotic_unimodal", "outlier", "bimodal", "separated_bimodal", "skewed_bimodal", "trimodal", "claw", "double_claw", "asymmetric_claw", "asymmetric_double_claw", "smooth_comb", "discrete_comb"; It also includes "harp" example from Li et al. (2016).
...	further arguments passed to dnorm and pnorm.

Details

Users have to either provide mean, sd and optionally prob; or type. In case of providing type, the values of mean, sd and prob are ignored.

If prob is not specified it assumes the default value of equal weights. Each component is computed via dnorm, pnorm and rnorm.

Value

`dmixnorm` gives the density, `pmixnorm` gives the distribution function, and `rmixnorm` generates random deviates.

The length of the result is determined by `n` for `rmixnorm`, and is the length of `x` for `dmixnorm` and `pmixnorm`.

`paramExample` gives a data frame with components `mean`, `sd` and `prob`.

References

Li, H., Munk, A., Sieling, H., and Walther, G. (2016). The essential histogram. arXiv:1612.07216.

Marron, J. S., & Wand, M. P. (1992). Exact mean integrated squred error. The Annals of Statistics, 20(2), 712–736.

See Also

Normal for standard normal distributions; Distributions for other standard distributions.

Examples

```
## Example claw
type = "claw" # or equivalently "MW10"
#     generate random numbers
n = 500
Y = rmixnorm(n, type = type)
#     compute the density
x = seq(min(Y), max(Y), length.out = n)
f = dmixnorm(x, type = type)
#     compute the distribution
F = pmixnorm(x, type = type)
#     plots
op = par(mfrow = c(1,2))
plot(x, f, type = "l", main = "Claw Density")
points(Y, rep(0,n))
plot(x, F, type = "l", main = "Claw Distribution")
points(Y, rep(0,n))
par(op)
```

`Multiscale Quantiles` *Quantile of the multiscale statistics*

Description

Simulate quantiles of the multiscale statistics under any continuous distribution function.

Usage

```
msQuantile(n, alpha = c(0.1), nsim = 5e3, is.sim = (n < 1e4),
           intv = genIntv(n), verbose = TRUE, ...)
```

Arguments

n	number of observations.
alpha	significance level; the (1-alpha)-quantile of the null distribution of the multiscale statistic via Monte Carlo simulation.
nsim	numer of Monte Carlo simulations.
is.sim	logical. If TRUE (default if n < 10,000) the quantile is determined via Monte Carlo simulations, which might take a long time; otherwise (default if n >= 10,000) it uses the quantile with n = 10,000, which has been precomputed and stored.
verbose	logical. If TRUE (default) it prints some details about the computation; otherwise nothing is printed.
intv	a data frame provides the system of intervals on which the multiscale statistic is defined. The data frame constains the following two columns left left index of an interval right right index of an interval By default, it is set to the sparse interval system proposed by Rivera and Walther (2013), see also Li et al. (2016).
...	further arguments passed to function quantile.

Details

Empirically, it turns out that the (1-alpha)-quantile of the multiscale statistic converges fast to that of the limit distribution as the number of observations n increases. Thus, for the sake of computational efficiency, the quantile with n = 10,000 are used by default for that with n > 10,000, which has already been precomputed and stored. Of course, for arbitrary sample size n, one can always simulate the quantile by setting is.sim = TRUE, and use the precomputed value by setting is.sim = FALSE. For a given sample size n, simulations are once computed, and then automatically recorded in main memory for later usage.

Value

A vector of length length(alpha) is returned, the same structure as returned by funtion quantile. See Li et al. (2016) for further details.

References

Li, H., Munk, A., Sieling, H., and Walther, G. (2016). The essential histogram. arXiv:1612.07216.

Rivera, C., & Walther, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. Scand. J. Stat. 40, 752–769.

See Also

checkHistogram, essHistogram, genIntv

Examples

```
n    = 100 # number of observations
nsim = 100 # number of simulations

alpha = c(0.1, 0.9) # significance level
q     = msQuantile(n, alpha, nsim)

print(q)
```

Index

∗Topic **datagen**
    Mixed normals, 8
∗Topic **distribution**
    checkHistogram, 3
    Essential Histogram, 5
    essHist-package, 2
    Mixed normals, 8
    Multiscale Quantiles, 9
∗Topic **nonparametric**
    Essential Histogram, 5
    essHist-package, 2
    Generate Intervals, 7
    Multiscale Quantiles, 9
∗Topic **package**
    essHist-package, 2

axis, *6*

checkHistogram, 3, *6, 7*, *10*

Distributions, *9*
dmixnorm *(*Mixed normals*)*, 8
dnorm, *8*

Essential Histogram, 5
essHist *(*essHist-package*)*, 2
essHist-package, 2
essHistogram, *5*, *7*, *10*
essHistogram *(*Essential Histogram*)*, 5

Generate Intervals, 7
genIntv, *5, 6*, *10*
genIntv *(*Generate Intervals*)*, 7

hist, *4*, *6*

Mixed normals, 8
msQuantile, *4–7*
msQuantile *(*Multiscale Quantiles*)*, 9
Multiscale Quantiles, 9

Normal, *9*

par, *4*
paramExample *(*Mixed normals*)*, 8
plot, *4*
plot.histogram, *6*
pmixnorm *(*Mixed normals*)*, 8
pnorm, *8*

quantile, *10*

rmixnorm *(*Mixed normals*)*, 8
rnorm, *8*

title, *4*, *6*