## Package 'kerdiest'

June 23, 2025

Title Nonparametric Kernel Estimation of Distribution Function

Version 1.3-1

Date 2025-06-18

Author Alejandro Quintela-del-Río [aut], Graciela Estévez-Pérez [aut], Ignacio López-de-Ullibarri [cre]

Maintainer Ignacio López-de-Ullibarri <ignacio.lopezdeullibarri@udc.es>

**Depends** R (>= 4.5.0)

Suggests chron, date, evir

**Description** Nonparametric kernel distribution function estimation is performed. Three bandwidth selectors are implemented: the plug-in selectors of Altman and Leger and of Polansky and Baker, and the cross-validation selector of Bowman, Hall and Prvan. The exceedance function, the mean return period and the return level are also computed. For details, see Quintela-del-Río and Estévez-Pérez (2012) <doi:10.18637/jss.v050.i08>.

License GPL (>= 2)

LazyLoad yes

Encoding UTF-8

NeedsCompilation no

**Repository** CRAN

Date/Publication 2025-06-23 10:20:06 UTC

## Contents

rdiest-package
Lbw
/bw
e
τρ
/ip
Bbw
ltriver

## Index

kerdiest-package

## Description

Nonparametric kernel distribution function estimation for continuous random variables is performed. Three automatic bandwidth selection procedures for nonparametric kernel distribution function estimation are implemented: the plug-in method of Altman and Leger, the plug-in method of Polansky and Baker, and the modified cross-validation method of Bowman, Hall and Prvan. The exceedance function, the mean return period and the return level are also computed.

## Author(s)

Graciela Estévez Pérez and Alejandro Quintela del Río

Maintainer: Ignacio López-de-Ullibarri

## References

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

ALbw

Compute the Plug-in Bandwidth of Altman and Leger

#### Description

The bandwidth parameter for the distribution function kernel estimator is calculated, using the plugin method of Altman and Leger (1995). Four possible kernel functions can be used for the kernel estimator: "e" Epanechnikov, "n" Normal, "b" Biweight and "t" Triweight.

## Usage

ALbw(type\_kernel = "n", vec\_data)

## Arguments

type_kernel	The kernel function. You can use four types: "e" Epanechnikov, "n" Normal,
	"b" Biweight and "t" Triweight. The normal kernel is used by default.
vec_data	The data sample.

## ALbw

## Details

Altman and Leger (1995) recommend the use of the Epanechnikov kernel, because in this case the rate of convergence for the kernel derivative estimator is improved. For the sake of uniformity along the package, the gaussian kernel is used by default, but the user can obviously choose the Epanechnikov function.

## Value

A numeric value for the bandwidth parameter.

## Author(s)

Graciela Estévez Pérez and Alejandro Quintela del Río

## References

Altman, N. and Leger, C. (1995), "Bandwidth selection for kernel distribution function estimation", Journal of Statistical Planning and Inference, 46, 195-214.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

#### Examples

```
## Compute the plug-in bandwidth for a sample of 100 random N(0,1) data
x <- rnorm(100, 0, 1)
h_AL <- ALbw(type_kernel = "e", vec_data = x)</pre>
h_AL
## A quick plot of a distribution function estimate
x <- rnorm(1000)</pre>
h_AL <- ALbw(vec_data = x)
F_AL <- kde(vec_data = x, bw = h_AL)</pre>
plot(F_AL$grid, F_AL$Estimated_values, type = "1")
## Plotting the distribution function estimate, controling the grid points
## and the kernel function
ss <- quantile(x, c(0.05, 0.95))</pre>
## Number of points to be used in the representation of estimated distribution
## function
n_pts <- 100
y <- seq(ss[1], ss[2], length.out = n_pts)</pre>
F_AL <- kde(type_kernel = "e", x, y, h_AL)$Estimated_values</pre>
## Plot of the theoretical and estimated distribution functions
plot(y, F_AL, type = "l", lty = 2)
lines(y, pnorm(y), type = "1", lty = 1)
legend(-1, 0.8, c("Real", "Nonparametric"), lty = 1:2)
```

## Description

The bandwidth parameter for the distribution function kernel estimator is calculated using the modified cross-validation method of Bowman, Hall and Prvan (1998). Four possible kernel functions can be used: "e" Epanechnikov, "n" Normal, "b" Biweight and "t" Triweight. The cross-validation function involves an integral term, that is calculated using Simpson's rule.

#### Usage

```
CVbw(type_kernel = "n", vec_data, n_pts = 100, seq_bws = NULL)
```

## Arguments

type_kernel	The kernel function. You can use four types: "e" Epanechnikov, "n" Normal, "b" Biweight and "t" Triweight. The normal kernel is used by default.
vec_data	The data sample.
n_pts	The number of points used to approximate, by Simpson's rule, the integral term in the cross-validation function. Because this numeric method increases the computing time, you can check different numbers, depending on your sample size. By default, 100 points are used.
seq_bws	The sequence of bandwidths in which to compute the cross-validation function. By default, the procedure defines a sequence of 50 points, from the range of the data divided by 200 to the range divided by 2.

## Value

A list consisting of:

seq_bws	The sequence of bandwidths.
CV function	The values of the cross-validation function in the bandwidths grid.
bw	A numeric value for the cross-validation bandwidth.

## Author(s)

Graciela Estévez Pérez and Alejandro Quintela del Río

#### References

Bowman, A.W., Hall, P. and Prvan, T. (1998), "Cross-validation for the smoothing of distribution functions", Biometrika. 85, 799-808.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

#### Examples

ef

```
## Compute the cross-validation bandwidth for a sample of 100 random N(0,1)
## data
x <- rnorm(100, 0, 1)
num bws <- 50
seq_bws <- seq(((max(x) - min(x))/2)/50, (max(x) - min(x))/2, length = num_bws)
hCV <- CVbw(type_kernel = "e", vec_data = x, n_pts = 200, seq_bws = seq_bws)
hCV
## The CV function is plotted
h_CV <- CVbw(vec_data = x)
h_CV$bw
plot(h_CV$seq_bws, h_CV$CVfunction, type = "1")
## Plotting the distribution function estimate controlling the grid points
## and the kernel function
ss <- quantile(x, c(0.05, 0.95))</pre>
## Number of points to be used in the representation of estimated distribution
## function
n_pts <- 100
y <- seq(ss[1], ss[2], length.out = n_pts)</pre>
F_CV <- kde(type_kernel = "e", x, y, h_CV$bw)$Estimated_values</pre>
## Plot of the theoretical and estimated distribution functions
plot(y, F_CV, type = "l", lty = 2)
lines(y, pnorm(y), type = "1", lty = 1)
legend(-1, 0.8, c("Real", "Nonparametric"), lty = 1:2)
```

#### **Exceedance Function Estimation**

## Description

Computes the exceedance probability, i.e., the probability that a specified value c (a magnitude of a seismic event, a flow level...) will be exceeded in D time units.

## Usage

## Arguments

type_kernel	The kernel function. You can use four types: "e" Epanechnikov, "n" Normal, "b" Biweight and "t" Triweight. The Normal kernel is used by default.
vec_data	The data sample (earthquake magnitudes, flow levels, wind speed)
с	The concrete level in which we want to compute the exceedance probability.
bw	The bandwidth parameter. The plug-in method of Polansky and Baker (2000) is used by default.

5

Dmin	Minimum value for D time units (years, days ). Default is 0.
Dmax	Maximum value for D time units (years, days ). Default is 15.
size_grid	Length of a grid in which we compute the exceedance function. By default, 50.
lambda	The mean activity rate.

## Details

The exceedance function is usually calculated assuming that event occurrence follows a Poisson process. In this case, the exceedance function, i.e., the probability of an specific value c is calculated as

$$R(c, D) = 1 - exp(-\lambda D(1 - F_h(c))).$$

See, e.g., Orlecka-Sikora (2008) or Quintela-del-Rio (2010) for earthquake data applications.

## Value

Returns a list containing:

Estimated\_values

	Vector containing the estimated function.
grid	The used grid.
bw	Value of the bandwidth.

## Author(s)

Graciela Estévez Pérez and Alejandro Quintela del Río

## References

Orlecka-Sikora, B. (2008), "Resampling methods for evaluating the uncertainty of the nonparametric magnitude distribution estimation in the probabilistic seismic hazard analysis", Tectonophysics 456, 38-51.

Quintela-del-Rio, A. (2010), "On nonparametric techniques for area-characteristic seismic hazard parameters", Geophysical Journal International, 180, 339-346.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

## Examples

```
## Working with earthquake data. We use the catalogue of the National
## Geographic Institute (IGN) of Spain and select the data of the Northwest
## of the Iberian Peninsula.
data(nwip)
require(chron)
require(date)
## The data with magnitude greater than 3 are considered
mg <- nwip$magnitude[nwip$magnitude > 3.0]
x1 <- nwip$year</pre>
```

```
x2 <- nwip$month
x3 <- nwip$day
ys <- paste(x1, x2, x3)
earthquake_date <- as.character(ys)
y1s <- as.date(earthquake_date, order = "ymd")
## Computation of the total number of years
y2s <- as.POSIXct(y1s)
z <- years(y2s)
n.years <- length(levels(z))
## Mean rate of earthquakes per year
lambda <- length(mg)/n.years
## Estimation of the exceedance probability for magnitude = 4
est <- ef(vec_data = mg, c = 4, lambda = lambda)
plot(est$grid, est$Estimated_values, type = "l", xlab = "Years",
ylab = "Probability of Exceedance")
```

kde

Kernel Estimation of the Distribution Function

## Description

Computes the value of the kernel estimator of the distribution function, in a single value or in a grid. Four possibilities for the kernel function are implemented, and the bandwidth parameter can be directly calculated by the plug-in method of Polansky and Baker (2000).

## Usage

## Arguments

type_kernel	The kernel function. You can use four types: "e" Epanechnikov, "n" Normal, "b" Biweight and "t" Triweight. The normal kernel is used by default.
vec_data	The data sample.
У	The single value or the grid vector where the distribution function is estimated. By default, a grid of 100 equidistant points from the minimum to the maximum of the data sample is selected.
bw	The bandwidth used. If it is not provided, the plug-in bandwidth of Polansky and Baker (2000) is computed.

#### Value

A list	ontaining:	
Estimated_values		
	Vector containing the estimated function in the grid values.	
grid	The used grid.	
bw	Value of the bandwidth.	

7

#### Author(s)

Graciela Estévez Pérez and Alejandro Quintela del Río

#### References

Reiss, R.D. (1981), "Nonparametric estimation of smooth distribution functions", Scandinavian Journal of Statistics, 8, 116-119.

Simonoff, J. (1996), "Smoothing Methods in Statistics", Springer, New York.

Polansky, A.M. and Baker, E.R. (2000), "Multistage plug-in bandwidth selection for kernel distribution function estimates", Journal of Statistical Computation and Simulation, 65, 63-80.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

## Examples

```
## Comparison of three bandwidth selection methods
x <- rnorm(100)
## The bandwidths by cross-validation, plug-in of Altman and Leger
## and plug-in of Polansky and Baker are computed using a normal kernel
## and a standard setting of parameters
h_CV <- CVbw(vec_data = x)$bw
h_CV
## Plug-in of Altman and Leger
h_AL <- ALbw(vec_data = x)
h_AL
## Plug-in of Polansky and Baker
h_PB <- PBbw(vec_data = x)
h PB
## Plot of the three estimates together with the real distribution
F_CV \le kde(vec_data = x, bw = h_CV)
F_AL <- kde(vec_data = x, bw = h_AL)</pre>
F_PB <- kde(vec_data = x, bw = h_PB)</pre>
y <- F_CV$grid
Ft <- pnorm(y)</pre>
plot(y, Ft, ylab = "Distribution", xlab = "Data", type = "l", lty = 1)
lines(y, F_CV$Estimated_values, type = "1", lty = 2)
lines(y, F_AL$Estimated_values, type = "1", lty = 3)
lines(y, F_PB$Estimated_values, type = "1", lty = 4)
legend(1, 0.4, c("Real", "F_CV", "F_AL", "F_PB"), lty = 1:4)
```

mrp

Mean Return Period Estimation

#### Description

This functions computes an estimate of the time between two values of a concrete level (size of an earthquake, flow lewel, wind speed...).

mrp

## Usage

## Arguments

type_kernel	The kernel function. You can use four types: "e" Epanechnikov, "n" Normal, "b" Biweight and "t" Triweight. The normal kernel is used by default.
vec_data	The data sample (earthquake magnitudes, flow levels, wind speed).
У	A grid or a singular value where the estimator is computed. By default, a grid of 50 values between the minimum and the maximum of the data is computed.
bw	The bandwidth parameter. By default, the plug-in method of Polansky and Baker (2000) is used.
lambda	The mean activity rate.

## Details

The mean return period is usually calculated assuming that event occurrence follows a Poisson process. In this case, the mean return period of events of size c is calculated as

$$T(c) = \frac{1}{\lambda(1 - F_h(c))}.$$

In Orlecka-Sikora (2008) or Quintela-del-Rio (2010) an application to earthquake data is made. In hydrological applications, if we work with annual maxima data, the parameter of the Poisson variable is 1 (one maximum per year). The mean return period between flow levels of value c is calculated as

$$T(c) = \frac{1}{1 - F_h(c)}.$$

See, for instance, Quintela-del-Rio (2011), for an application to data of Salt River near Roosevelt, AZ, USA (saltriver data).

## Value

A list containing:

Estimated\_values

Estimated_values	
	Vector containing the estimated function.
grid	The used grid.
bw	Value of the bandwidth.

#### Author(s)

Graciela Estévez Pérez and Alejandro Quintela del Río

#### References

Orlecka-Sikora, B. (2008), "Resampling methods for evaluating the uncertainty of the nonparametric magnitude distribution estimation in the probabilistic seismic hazard analysis", Tectonophysics, 456, 38-51.

Quintela-del-Rio, A. (2010), "On nonparametric techniques for area- characteristic seismic hazard parameters", Geophysical Journal International, 180, 339-346.

Quintela-del-Rio, A. (2011), "On bandwidth selection for nonparametric estimation in flood frequency analysis", Hydrological Processes, 25, 671-678.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

#### Examples

```
## Working with earthquake data. We use the catalogue of the National
## Geographic Institute (IGN) of Spain and select the data of the Northwest
## of the Iberian Peninsula.
data(nwip)
require(chron)
require(date)
## Data with magnitude greater than 3 are considered
mg <- nwip$magnitude[nwip$magnitude > 3.0]
x1 <- nwip$year
x2 <- nwip$month
x3 <- nwip$day
ys <- paste(x1, x2, x3)</pre>
earthquake_date <- as.character(ys)</pre>
y1s <- as.date(earthquake_date, order = "ymd")</pre>
## Computation of the total number of years
y2s <- as.POSIXct(y1s)</pre>
z <- years(y2s)</pre>
n.years <- length(levels(z))</pre>
## Mean rate of earthquakes per year
lambda <- length(mg)/n.years</pre>
## Estimation of the mean return period (in years) between earthquakes of
## the same magnitude
est2 <- mrp(vec_data = mg, lambda = lambda)</pre>
plot(est2$grid, est2$Estimated_values, type = "1", xlab = "Magnitude",
  ylab = "Mean return period (years)")
## Working with hydrological data: annual peak instantaneous flow of the
## Salt River near Roosevelt, AZ, USA, for 1924-2009.
data(saltriver)
peak <- saltriver$peakflow</pre>
year <- saltriver$year</pre>
plot(year, peak, type = "1", xlab = "Year", ylab = "Annual peak flow")
## Mean return period for the Saltriver data
rp <- mrp(type_kernel = "n", vec_data = peak, lambda = 1)</pre>
plot(rp$grid, rp$Estimated_values, type = "1", xlab = "Flow level",
  ylab = "Years ", main = "Mean return period")
```

nwip

## Description

This data set corresponds to the earthquakes occurred in the Northwest of the Iberian Peninsula, from 25/Nov/1924 to 31/Jul/2010. The area is limited by the coordinates 41 N–44 N and 6 W–10 W, involving the autonomic region of Galicia (Spain) and northern Portugal.

#### Usage

data(nwip)

## Format

A data frame with 3491 observations on the following 10 variables, corresponding to the earthquake epicenters and time of ocurrence.

day Day month Month year Year hour Hour minute Minute second second latitude Latitude in degrees longitude Longitude in degrees depth Depth in km magnitude Magnitude in Richter Scale

## Source

The data catalogue has been obtained from the National Geographic Institute (IGN) of Spain. The web page is www.ign.es.

## References

Rueda, J., and J. Mezcua (2001), "Sismicidad, sismotectonica y peligrosidad sismica en Galicia", IGN Technical Publication, 35.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

## PBbw

## Description

The bandwidth parameter for the distribution function kernel estimator is calculated, using the plugin method of Polansky and Baker (2000). Four possible kernel functions can be used for the kernel estimator: "e" Epanechnikov, "n" Normal, "b" Biweight and "t" Triweight. Because kernel estimators of derivatives of order greater than two are required, only the normal kernel is used in this case.

#### Usage

```
PBbw(type_kernel = "n", vec_data, num_stage = 2)
```

## Arguments

type_kernel	The kernel function used. You can use four types: "e" Epanechnikov, "n" Nor- mal, "b" Biweight and "t" Triweight. The normal kernel is used by default.
vec_data	The data sample.
num_stage	The number of iterations in the Polansky and Baker method. The default, 2, is usually a good option. Values of 3 or 4 are also allowed.

## Value

A numeric value for the bandwidth parameter.

#### Author(s)

Graciela Estévez Pérez and Alejandro Quintela del Río

#### References

Polansky, A.M. and Baker, E.R. (2000), "Multistage plug-in bandwidth selection for kernel distribution function estimates", Journal of Statistical Computation and Simulation, 65, 63-80.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

#### Examples

```
## Compute the plug-in bandwidth for a sample of 100 random N(0,1) data
x <- rnorm(100, 0, 1)
h_PB <- PBbw(vec_data = x, num_stage = 4)
h_PB</pre>
```

## A quick plot of a distribution function estimate

rl

```
x <- rnorm(1000)
h_PB <- PBbw(vec_data = x)</pre>
F_PB <- kde(vec_data = x, bw = h_PB)</pre>
plot(F_PB$grid, F_PB$Estimated_values, type = "1")
## Plotting the distribution function estimate controling the grid points and
## the kernel function
ss <- quantile(x, c(0.05, 0.95))</pre>
## number of points to be used in the representation of the estimated
## distribution function
n_pts <- 100
y <- seq(ss[1], ss[2], length.out = n_pts)</pre>
F_PB <- kde(type_kernel = "e", x, y, h_PB)$Estimated_values</pre>
## Plot of the theoretical and estimated distribution functions
plot(y, F_PB, type = "l", lty = 2)
lines(y, pnorm(y), type = "1", lty = 1)
legend(-1.2, 0.8, c("Real", "Nonparametric"), lty = 1:2)
```

rl

## Return Level Estimation

## Description

The T-return level is defined as the value of the observed variable that can be expected to be once exceeded during a T-period of time. This is computed as the quantile of the distribution corresponding to the value  $F^{-1}(1 - \frac{1}{T})$ .

## Usage

#### Arguments

type_kernel	The kernel function used. You can use four types: "e" Epanechnikov, "n" Nor- mal, "b" Biweight and "t" Triweight. The normal kernel is used by default.
vec_data	The data sample (earthquake magnitudes, flow levels, wind speeds).
time	A time or a vector of times for T.
bw	The bandwidth parameter. By default, he plug-in method of Polansky and Baker (2000) is used.

#### Details

In several scientific fields, it is of interest to estimate quantiles corresponding to a probability of exceedance. E.g., in hydrology, the T-return level  $x_T$  is defined as the value of the observed flow that can be expected to be once exceeded during a T-period of time; i.e., the quantile

$$x_T = F^{-1}(1 - \frac{1}{T}).$$

It can be directly estimated by

$$\hat{x}_T = F_h^{-1}(1 - \frac{1}{T}).$$

See, e.g., Quintela-del-Rio (2011), for an application to data of Salt River near Roosevelt, AZ, USA.

#### Value

A single value or an array for the estimated quantiles.

## Author(s)

Graciela Estévez Pérez and Alejandro Quintela del Río

## References

Quintela-del-Rio, A. (2011), "On bandwidth selection for nonparametric estimation in flood frequency analysis". Hydrological Processes 25, 671-678.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

## Examples

```
data(saltriver)
peak <- saltriver$peakflow
year <- saltriver$peakflow
plot(year, peak, type = "1", ylab = "Annual peak flow")
## Calculating the return values for a period from 2 to 100 years
times <- seq(2,100, length.out = 100)
ret.lev <- rl(vec_data = peak, time = times)
plot(times, ret.lev, type = "1", xlab = "Years", ylab = "Flow (cumecs)",
main = "Return level Plot")</pre>
```

saltriver

Data from Salt River

#### Description

The annual peak instantaneous flow of the Salt River near Roosevelt, AZ, USA, for 1924-2009. Data are in cfs (0.028317 m3/s); water year October-September. The data were examined in several papers related with extreme values in hydrology. Among others, they were analyzed by Anderson and Meerschaert (1998) and Dettinger and Diaz (2000), where they were fitted to GEV and GPD distributions. In Quintela-del-Rio (2011), a nonparametric analysis for this data set is presented.

#### Usage

data(saltriver)

## saltriver

## Format

A data frame with 85 observations on the following 2 variables.

year Year

peakflow The annual observed maximum peak flow

## Source

US Geological Survey http://water.usgs.gov/nwis/peak.

## References

Anderson, P.L. and Meerschaert, M.M. (1998), "Modeling river flows with heavy tails", Water Resources Research, 34, 2271-2280.

Dettinger, M.D. and Diaz, H.F. (2000), "Global characteristics of stream flow seasonality and variability", Journal of Hydrometeorology, 1, 289-310.

Quintela-del-Rio, A. (2011), "On bandwidth selection for nonparametric estimation in flood frequency analysis". Hydrological Processes 25, 671-678.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012), "Nonparametric kernel distribution function estimation with kerdiest: an R package for bandwidth choice and applications", Journal of Statistical Software, 50(8), 1-21.

# Index

\* datasets nwip, 11 saltriver, 14 \* nonparametric ALbw, 2 CVbw, 4 ef, <mark>5</mark> kde, 7 mrp,8 PBbw, 12 rl, 13 \* smoothing ALbw, 2 CVbw, 4 ef, <mark>5</mark> kde, 7 mrp,<mark>8</mark> PBbw, 12 rl, 13 ALbw, 2CVbw, 4 ef, <mark>5</mark> kde, 7 kerdiest (kerdiest-package), 2 kerdiest-package, 2 mrp,<mark>8</mark> nwip, 11  $\mathsf{PBbw},\,\mathbf{12}$ rl, 13 saltriver, 14