Package ‘KernSmooth’

July 10, 2023

Priority recommended
Version 2.23-22
Date 2023-07-10
Title Functions for Kernel Smoothing Supporting Wand & Jones (1995)
Note Maintainers are not available to give advice on using a package they did not author.
Depends R (>= 2.5.0), stats
Suggests MASS, carData
Description Functions for kernel smoothing (and density estimation) corresponding to the book:
License Unlimited
ByteCompile yes
NeedsCompilation yes
Author Matt Wand [aut], Cleve Moler [ctb] (LINPACK routines in src/d*), Brian Ripley [trl, cre, ctb] (R port and updates)
Maintainer Brian Ripley <ripley@stats.ox.ac.uk>
Repository CRAN
Date/Publication 2023-07-10 14:38:27 UTC

R topics documented:

bkde ................................................................. 2
bkde2D ............................................................ 3
bkfe ................................................................. 4
dpih ................................................................ 6
dpik ................................................................. 7
dpill .................................................................. 9
locpoly ............................................................. 10

Index 13

1
Compute a Binned Kernel Density Estimate

**Description**

Returns x and y coordinates of the binned kernel density estimate of the probability density of the data.

**Usage**

```r
bkde(x, kernel = "normal", canonical = FALSE, bandwidth, gridsize = 401L, range.x, truncate = TRUE)
```

**Arguments**

- **x**: numeric vector of observations from the distribution whose density is to be estimated. Missing values are not allowed.
- **bandwidth**: the kernel bandwidth smoothing parameter. Larger values of bandwidth make smoother estimates, smaller values of bandwidth make less smooth estimates. The default is a bandwidth computed from the variance of x, specifically the ‘oversmoothed bandwidth selector’ of Wand and Jones (1995, page 61).
- **kernel**: character string which determines the smoothing kernel. `kernel` can be: "normal" - the Gaussian density function (the default). "box" - a rectangular box. "epanech" - the centred beta(2,2) density. "biweight" - the centred beta(3,3) density. "triweight" - the centred beta(4,4) density. This can be abbreviated to any unique abbreviation.
- **canonical**: length-one logical vector: if `TRUE`, canonically scaled kernels are used.
- **gridsize**: the number of equally spaced points at which to estimate the density.
- **range.x**: vector containing the minimum and maximum values of x at which to compute the estimate. The default is the minimum and maximum data values, extended by the support of the kernel.
- **truncate**: logical flag: if `TRUE`, data with x values outside the range specified by `range.x` are ignored.

**Details**

This is the binned approximation to the ordinary kernel density estimate. Linear binning is used to obtain the bin counts. For each x value in the sample, the kernel is centered on that x and the heights of the kernel at each datapoint are summed. This sum, after a normalization, is the corresponding y value in the output.

**Value**

A list containing the following components:

- **x**: vector of sorted x values at which the estimate was computed.
- **y**: vector of density estimates at the corresponding x.
**Background**

Density estimation is a smoothing operation. Inevitably there is a trade-off between bias in the estimate and the estimate's variability: large bandwidths will produce smooth estimates that may hide local features of the density; small bandwidths may introduce spurious bumps into the estimate.

**References**


**See Also**

density, dpik, hist, ksmooth.

**Examples**

```r
data(geyser, package="MASS")
x <- geyser$duration
est <- bkde(x, bandwidth=0.25)
plot(est, type="l")
```

---

**bkde2D**  
*Compute a 2D Binned Kernel Density Estimate*

**Description**

Returns the set of grid points in each coordinate direction, and the matrix of density estimates over the mesh induced by the grid points. The kernel is the standard bivariate normal density.

**Usage**

`bkde2D(x, bandwidth, gridsize = c(51L, 51L), range.x, truncate = TRUE)`

**Arguments**

- `x`: a two-column numeric matrix containing the observations from the distribution whose density is to be estimated. Missing values are not allowed.
- `bandwidth`: numeric vector of length 2, containing the bandwidth to be used in each coordinate direction.
- `gridsize`: vector containing the number of equally spaced points in each direction over which the density is to be estimated.
- `range.x`: a list containing two vectors, where each vector contains the minimum and maximum values of `x` at which to compute the estimate for each direction. The default minimum in each direction is minimum data value minus 1.5 times the bandwidth for that direction. The default maximum is the maximum data value plus 1.5 times the bandwidth for that direction.
- `truncate`: logical flag: if TRUE, data with `x` values outside the range specified by `range.x` are ignored.
Value

a list containing the following components:

- **x1**: vector of values of the grid points in the first coordinate direction at which the estimate was computed.
- **x2**: vector of values of the grid points in the second coordinate direction at which the estimate was computed.
- **fhat**: matrix of density estimates over the mesh induced by x1 and x2.

Details

This is the binned approximation to the 2D kernel density estimate. Linear binning is used to obtain the bin counts and the Fast Fourier Transform is used to perform the discrete convolutions. For each x1,x2 pair the bivariate Gaussian kernel is centered on that location and the heights of the kernel, scaled by the bandwidths, at each datapoint are summed. This sum, after a normalization, is the corresponding fhat value in the output.

References


See Also

- `bkde`, `density`, `hist`.

Examples

```r
data(geyser, package="MASS")
x <- cbind(geyser$duration, geyser$waiting)
est <- bkde2D(x, bandwidth=c(0.7, 7))
contour(est$x1, est$x2, est$fhat)
persp(est$fhat)
```

---

**bkfe**

Compute a Binned Kernel Functional Estimate

Description

Returns an estimate of a binned approximation to the kernel estimate of the specified density functional. The kernel is the standard normal density.

Usage

```r
bkfe(x, drv, bandwidth, gridsize = 401L, range.x, binned = FALSE, truncate = TRUE)
```
**Arguments**

- **x**: numeric vector of observations from the distribution whose density is to be estimated. Missing values are not allowed.
- **drv**: order of derivative in the density functional. Must be a non-negative even integer.
- **bandwidth**: the kernel bandwidth smoothing parameter. Must be supplied.
- **gridsize**: the number of equally-spaced points over which binning is performed.
- **range.x**: vector containing the minimum and maximum values of \( x \) at which to compute the estimate. The default is the minimum and maximum data values, extended by the support of the kernel.
- **binned**: logical flag: if TRUE, then \( x \) and \( y \) are taken to be grid counts rather than raw data.
- **truncate**: logical flag: if TRUE, data with \( x \) values outside the range specified by \( \text{range.x} \) are ignored.

**Details**

The density functional of order \( \text{drv} \) is the integral of the product of the density and its \( \text{drv} \)th derivative. The kernel estimates of such quantities are computed using a binned implementation, and the kernel is the standard normal density.

**Value**

the (scalar) estimated functional.

**Background**

Estimates of this type were proposed by Sheather and Jones (1991).

**References**


**Examples**

```r
data(geyser, package="MASS")
x <- geyser$duration
est <- bkfe(x, drv=4, bandwidth=0.3)
```
**Description**

Uses direct plug-in methodology to select the bin width of a histogram.

**Usage**

```r
dpih(x, scalest = "minim", level = 2L, gridsize = 401L, range.x = range(x), truncate = TRUE)
```

**Arguments**

- `x`: numeric vector containing the sample on which the histogram is to be constructed.
- `scalest`: estimate of scale.
  - "stdev" - standard deviation is used.
  - "iqr" - inter-quartile range divided by 1.349 is used.
  - "minim" - minimum of "stdev" and "iqr" is used.
- `level`: number of levels of functional estimation used in the plug-in rule.
- `gridsize`: number of grid points used in the binned approximations to functional estimates.
- `range.x`: range over which functional estimates are obtained. The default is the minimum and maximum data values.
- `truncate`: if `truncate` is `TRUE` then observations outside of the interval specified by `range.x` are omitted. Otherwise, they are used to weight the extreme grid points.

**Details**

The direct plug-in approach, where unknown functionals that appear in expressions for the asymptotically optimal bin width and bandwidths are replaced by kernel estimates, is used. The normal distribution is used to provide an initial estimate.

**Value**

the selected bin width.

**Background**

This method for selecting the bin width of a histogram is described in Wand (1995). It is an extension of the normal scale rule of Scott (1979) and uses plug-in ideas from bandwidth selection for kernel density estimation (e.g. Sheather and Jones, 1991).
References


See Also

hist

Examples

data(geyser, package="MASS")
x <- geyser$duration
h <- dpik(x)
bins <- seq(min(x)-h, max(x)+h, by=h)
hist(x, breaks=bins)

dpik

Select a Bandwidth for Kernel Density Estimation

Description

Use direct plug-in methodology to select the bandwidth of a kernel density estimate.

Usage

dpik(x, scalest = "minim", level = 2L, kernel = "normal",
canonical = FALSE, gridsize = 401L, range.x = range(x),
truncate = TRUE)

Arguments

x numeric vector containing the sample on which the kernel density estimate is to be constructed.
scalest estimate of scale.
"stdev" - standard deviation is used.
"iqr" - inter-quartile range divided by 1.349 is used.
"minim" - minimum of "stdev" and "iqr" is used.
level number of levels of functional estimation used in the plug-in rule.
kernal character string which determines the smoothing kernel. kernel can be: "normal" - the Gaussian density function (the default). "box" - a rectangular box. "epanech" - the centred beta(2,2) density. "biweight" - the centred beta(3,3) density. "triweight" - the centred beta(4,4) density. This can be abbreviated to any unique abbreviation.
canonical logical flag: if TRUE, canonically scaled kernels are used
gridsize the number of equally-spaced points over which binning is performed to obtain kernel functional approximation.
range.x vector containing the minimum and maximum values of x at which to compute the estimate. The default is the minimum and maximum data values.
truncate logical flag: if TRUE, data with x values outside the range specified by range.x are ignored.

Details
The direct plug-in approach, where unknown functionals that appear in expressions for the asymptotically optimal bandwidths are replaced by kernel estimates, is used. The normal distribution is used to provide an initial estimate.

Value
the selected bandwidth.

Background
This method for selecting the bandwidth of a kernel density estimate was proposed by Sheather and Jones (1991) and is described in Section 3.6 of Wand and Jones (1995).

References

See Also
*bkde, density, ksmooth*

Examples
```r
data(geyser, package="MASS")
x <- geyser$duration
h <- dpik(x)
est <- bkde(x, bandwidth=h)
plot(est,type="l")
```
**dpill**  
*Select a Bandwidth for Local Linear Regression*

**Description**

Use direct plug-in methodology to select the bandwidth of a local linear Gaussian kernel regression estimate, as described by Ruppert, Sheather and Wand (1995).

**Usage**

```r
dpill(x, y, blockmax = 5, divisor = 20, trim = 0.01, proptrun = 0.05,  
gridsize = 401L, range.x, truncate = TRUE)
```

**Arguments**

- `x` numeric vector of x data. Missing values are not accepted.
- `y` numeric vector of y data. This must be same length as `x`, and missing values are not accepted.
- `blockmax` the maximum number of blocks of the data for construction of an initial parametric estimate.
- `divisor` the value that the sample size is divided by to determine a lower limit on the number of blocks of the data for construction of an initial parametric estimate.
- `trim` the proportion of the sample trimmed from each end in the x direction before application of the plug-in methodology.
- `proptrun` the proportion of the range of `x` at each end truncated in the functional estimates.
- `gridsize` number of equally-spaced grid points over which the function is to be estimated.
- `range.x` vector containing the minimum and maximum values of `x` at which to compute the estimate. For density estimation the default is the minimum and maximum data values with 5% of the range added to each end. For regression estimation the default is the minimum and maximum data values.
- `truncate` logical flag: if TRUE, data with x values outside the range specified by `range.x` are ignored.

**Details**

The direct plug-in approach, where unknown functionals that appear in expressions for the asymptotically optimal bandwidths are replaced by kernel estimates, is used. The kernel is the standard normal density. Least squares quartic fits over blocks of data are used to obtain an initial estimate. Mallow's $C_p$ is used to select the number of blocks.

**Value**

the selected bandwidth.
Warning

If there are severe irregularities (i.e. outliers, sparse regions) in the x values then the local poly-
nomial smooths required for the bandwidth selection algorithm may become degenerate and the
function will crash. Outliers in the y direction may lead to deterioration of the quality of the se-
lected bandwidth.

References


See Also

ksmooth, locpoly.

Examples

data(geyser, package = "MASS")
x <- geyser$duration
y <- geyser$waiting
plot(x, y)
h <- dpill(x, y)
fit <- locpoly(x, y, bandwidth = h)
lines(fit)

locpoly

Estimate Functions Using Local Polynomials

Description

Estimates a probability density function, regression function or their derivatives using local polyno-
mials. A fast binned implementation over an equally-spaced grid is used.

Usage

locpoly(x, y, drv = 0L, degree, kernel = "normal",
        bandwidth, gridsize = 401L, bwdisc = 25,
        range.x, binned = FALSE, truncate = TRUE)

Arguments

x numeric vector of x data. Missing values are not accepted.
bandwidth the kernel bandwidth smoothing parameter. It may be a single number or an
array having length gridsize, representing a bandwidth that varies according
to the location of estimation.
locpoly

\textbf{Value}

if \( y \) is specified, a local polynomial regression estimate of \( E[Y|X] \) (or its derivative) is computed. If \( y \) is missing, a local polynomial estimate of the density of \( x \) (or its derivative) is computed.

a list containing the following components:

\begin{itemize}
  \item \textbf{x} \hspace{1cm} \text{vector of sorted } x \text{ values at which the estimate was computed.}
  \item \textbf{y} \hspace{1cm} \text{vector of smoothed estimates for either the density or the regression at the corresponding } x.
\end{itemize}

\textbf{Details}

Local polynomial fitting with a kernel weight is used to estimate either a density, regression function or their derivatives. In the case of density estimation, the data are binned and the local fitting procedure is applied to the bin counts. In either case, binned approximations over an equally-spaced grid is used for fast computation. The bandwidth may be either scalar or a vector of length gridsize.

\textbf{References}


\textbf{See Also}

\texttt{bkde, density, dpill, ksmooth, loess, smooth, supsmu}. 
Examples

```r
data(geyser, package = "MASS")
# local linear density estimate
x <- geyser$duration
est <- locpoly(x, bandwidth = 0.25)
plot(est, type = "l")

# local linear regression estimate
y <- geyser$waiting
plot(x, y)
fit <- locpoly(x, y, bandwidth = 0.25)
lines(fit)
```
Index

* distribution
  bkde, 2
  bkde2D, 3
* regression
  locpoly, 10
* smooth
  bkde, 2
  bkde2D, 3
  bkfe, 4
  dpih, 6
  dpik, 7
  dpill, 9
  locpoly, 10

bkde, 2, 4, 8, 11
bkde2D, 3
bkfe, 4

density, 3, 4, 8, 11
dpih, 6
dpik, 3, 7
dpillar, 9, 11

hist, 3, 4, 7

ksmooth, 3, 8, 10, 11

locpoly, 10, 10
loess, 11

smooth, 11
supsmu, 11