Package ‘ks’

July 26, 2018

Version 1.11.3
Date 2018-07-24
Title Kernel Smoothing
Author Tarn Duong <tarn.duong@gmail.com>
Maintainer Tarn Duong <tarn.duong@gmail.com>
Depends R (>= 2.10.0)
Imports FNN (>= 1.1), kernlab, KernSmooth (>= 2.22), Matrix, mclust, mgcv, multicool, mvtnorm (>= 1.0-0)
Suggests maps, MASS, misc3d (>= 0.4-0), OceanView, oz, rgl (>= 0.66)
Description Kernel smoothers for univariate and multivariate data, including densities, density derivatives, cumulative distributions, clustering, classification, density ridges, significant modal regions, and two-sample hypothesis tests. Duong (2017) <doi:10.18637/jss.v021.i07>.
License GPL-2 | GPL-3
URL http://www.mvstat.net/mvksa
NeedsCompilation yes
Repository CRAN
Date/Publication 2018-07-25 22:40:03 UTC

R topics documented:

   ks-package .................................................. 3
   air ......................................................... 7
   binning ...................................................... 7
   cardio ....................................................... 8
   contour ..................................................... 9
   grevillea .................................................. 10
   Hbcv ....................................................... 11
   histde ..................................................... 12
   Hlscv ..................................................... 13
   Hmm ....................................................... 15
   Hns ....................................................... 16
R topics documented:

Hpi ........................................ 17
hsct ....................................... 18
Hscv ....................................... 19
ise.mixt ................................... 20
kcde ....................................... 22
kcopula ................................... 24
kda ......................................... 26
kdcd ....................................... 28
kdde ....................................... 30
kde ......................................... 32
kde.boundary ................................ 34
kde.local.test ............................ 36
kde.test .................................. 38
kde.truncate ................................ 39
kd ........................................... 40
kfe ......................................... 42
kfs ......................................... 44
kms ......................................... 46
kroc ........................................ 48
ksupp ...................................... 50
mixt ....................................... 51
plot.histde ................................ 52
plot.kcde .................................. 54
plot.kda ................................... 55
plot.kdcd .................................. 56
plot.kdde .................................. 58
plot.kde ................................... 60
plot.kde.loctest .......................... 60
plot.kde.part ................................ 61
plot.kfs ................................... 63
plot.kroc .................................. 64
plotmixt ................................... 65
pre.transform ............................. 66
quake ...................................... 67
rkde ....................................... 67
tempb ...................................... 69
unicef ..................................... 69
vector ..................................... 70
vkde ....................................... 70
worldbank .................................. 72

Index 73
Description

Kernel smoothing for data from 1- to 6-dimensions.

Details

There are three main types of functions in this package:

• computing kernel estimators - these function names begin with ‘k’
• computing bandwidth selectors - these begin with ‘h’ (1-d) or ‘H’ (>1-d)
• displaying kernel estimators - these begin with ‘plot’.

The kernel used throughout is the normal (Gaussian) kernel $K$. For 1-d data, the bandwidth $h$ is the standard deviation of the normal kernel, whereas for multivariate data, the bandwidth matrix $H$ is the variance matrix.

–For kernel density estimation, $kde$ computes

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_{H}(x - X_i).$$

The bandwidth matrix $H$ is a matrix of smoothing parameters and its choice is crucial for the performance of kernel estimators. For display, its plot method calls $plot.kde$.

–For kernel density estimation, there are several varieties of bandwidth selectors

• plug-in $hpi$ (1-d); $Hpi, Hpi.diag$ (2- to 6-d)
• least squares (or unbiased) cross validation (LSCV or UCV) $hlscv$ (1-d); $Hlscv, Hlscv.diag$ (2- to 6-d)
• biased cross validation (BCV) $Hbcv, Hbcv.diag$ (2- to 6-d)
• smoothed cross validation (SCV) $hscv$ (1-d); $Hscv, Hscv.diag$ (2- to 6-d)
• normal scale $hns$ (1-d); $Hns$ (2- to 6-d).

–For kernel density support estimation, the main function is $ksupp$ which is (the convex hull of)

$$\{x : \hat{f}(x) > \tau\}$$

for a suitable level $\tau$. This is closely related to the $\tau$-level set of $\hat{f}$.

–For truncated kernel density estimation, the main function is $kde.truncate$

$$\hat{f}(x)1\{x \in \Omega\} / \int_{\Omega} \hat{f}(x) \, dx$$

for a bounded data support $\Omega$. The standard density estimate $\hat{f}$ is truncated and rescaled to give unit integral over $\Omega$. Its plot method calls $plot.kde$. 


For boundary kernel density estimation where the kernel function is modified explicitly in the boundary region, the main function is `kde.boundary`

\[ n^{-1} \sum_{i=1}^{n} K_H^*(x - X_i) \]

for a boundary kernel \( K^* \). Its plot method calls `plot.kde`.

For variable kernel density estimation where the bandwidth is not a constant matrix, the main functions are `kde.balloon`

\[ \hat{f}_{\text{ball}}(x) = n^{-1} \sum_{i=1}^{n} K_{H(x)}(x - X_i) \]

and `kde.sp`

\[ \hat{f}_{\text{SP}}(x) = n^{-1} \sum_{i=1}^{n} K_{H(X_i)}(x - X_i) \]

For the balloon estimation \( \hat{f}_{\text{ball}} \) the bandwidth varies with the estimation point \( x \), whereas for the sample point estimation \( \hat{f}_{\text{SP}} \) the bandwidth varies with the data point \( X_i, i = 1, \ldots, n \). Their plot methods call `plot.kde`. The bandwidth selectors for `kde.balloon` are based on the normal scale bandwidth \( hns(\cdot, \text{deriv.order}=2) \) via the MSE minimal formula, and for `kde.sp` on \( hns(\cdot, \text{deriv.order}=4) \) via the Abramson formula.

For kernel density derivative estimation, the main function is `kdde`

\[ D^{\otimes r} \hat{f}(x) = n^{-1} \sum_{i=1}^{n} D^{\otimes r} K_H(x - X_i) \]

The bandwidth selectors are a modified subset of those for `kde`, i.e. `Hlscv`, `Hns`, `Hpi`, `Hscv` with `deriv.order>0`. Its plot method is `plot.kdde` for plotting each partial derivative singly.

For kernel summary curvature estimation, the main function is `kcurv`

\[ \hat{s}(x) = -1 \{ D^2 \hat{f}(x) < 0 \} \text{abs}(|D^2 \hat{f}(x)|) \]

where \( D^2 \hat{f}(x) \) is the kernel Hessian matrix estimate. It has the same structure as a kernel density estimate so its plot method calls `plot.kde`.

For kernel discriminant analysis, the main function is `kda` which computes density estimates for each the groups in the training data, and the discriminant surface. Its plot method is `plot.kda`. The wrapper function `hkda`, `Hkda` computes bandwidths for each group in the training data for `kde`, e.g. `hpi`, `Hpi`.

For kernel functional estimation, the main function is `kfe` which computes the \( r \)-th order integrated density functional

\[ \hat{\psi}_r = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} D^{\otimes r} K_H(X_i - X_j) \]

The plug-in selectors are `hpi.kfe` (1-d), `Hpi.kfe` (2- to 6-d). Kernel functional estimates are usually not required to computed directly by the user, but only within other functions in the package.
For kernel-based 2-sample testing, the main function is \texttt{kde.test} which computes the integrated $L^2$ distance between the two density estimates as the test statistic, comprising a linear combination of 0-th order kernel functional estimates:

$$
\hat{T} = \hat{\psi}_{0,1} + \hat{\psi}_{0,2} - (\hat{\psi}_{0,12} + \hat{\psi}_{0,21}),
$$

and the corresponding p-value. The $\psi$ are zero order kernel functional estimates with the subscripts indicating that 1 = sample 1 only, 2 = sample 2 only, and 12, 21 = samples 1 and 2. The bandwidth selectors are \texttt{hpi.kfe}, \texttt{Hpi.kfe} with \texttt{deriv.order=0}.

For kernel-based local 2-sample testing, the main function is \texttt{kde.local.test} which computes the squared distance between the two density estimates as the test statistic

$$
\hat{U}(x) = [\hat{f}_1(x) - \hat{f}_2(x)]^2
$$

and the corresponding local p-values. The bandwidth selectors are those used with \texttt{kde}, e.g. \texttt{hpi}, \texttt{Hpi}.

For kernel cumulative distribution function estimation, the main function is \texttt{kcde}

$$
\hat{F}(x) = n^{-1} \sum_{i=1}^{n} K_H(x - X_i)
$$

where $K$ is the integrated kernel. The bandwidth selectors are \texttt{hpi.kcde}, \texttt{Hpi.kcde}. Its plot method is \texttt{plot.kcde}. There exist analogous functions for the survival function $\hat{\bar{F}}$.

For kernel estimation of a ROC (receiver operating characteristic) curve to compare two samples from $\hat{F}_1, \hat{F}_2$, the main function is \texttt{kroc}

$$
\{\hat{F}_{Y_1}(z), \hat{F}_{Y_2}(z)\}
$$

based on the cumulative distribution functions of $\hat{Y}_j = \hat{F}_1(X_j), j = 1, 2$.

The bandwidth selectors are those used with \texttt{kcde}, e.g. \texttt{hpi.kcde}, \texttt{Hpi.kcde} for $\hat{F}_{Y_j}, \hat{F}_1$. Its plot method is \texttt{plot.kroc}.

For kernel estimation of a copula, the main function is \texttt{kcopula}

$$
\hat{C}(z) = \hat{F}(\hat{F}^{-1}_1(z_1), \ldots, \hat{F}^{-1}_d(z_d))
$$

where $\hat{F}^{-1}_j(z_j)$ is the $z_j$-th quantile of of the $j$-th marginal distribution $\hat{F}_j$. The bandwidth selectors are those used with \texttt{kcde} for $\hat{F}, \hat{F}_j$. Its plot method is \texttt{plot.kcde}.

For kernel mean shift clustering, the main function is \texttt{kms}. The mean shift recurrence relation of the candidate point $x$

$$
x_{j+1} = x_j + H\hat{f}(x_j)/\hat{f}(x_j),
$$

where $j \geq 0$ and $x_0 = x$, is iterated until $x$ converges to its local mode in the density estimate $\hat{f}$ by following the density gradient ascent paths. This mode determines the cluster label for $x$. The bandwidth selectors are those used with \texttt{kdde}, \texttt{deriv.order=1}.

For kernel density ridge estimation, the main function is \texttt{kdr}. The kernel density ridge recurrence relation of the candidate point $x$

$$
x_{j+1} = x_j + U_{(d-1)}(x_j) U_{(d-1)}(x_j)^T H\hat{f}(x_j)/\hat{f}(x_j),
$$
where \( j \geq 0 \), \( x_0 = x \) and \( U_{(d-1)} \) is the 1-dimensional projected density gradient, is iterated until \( x \) converges to the ridge in the density estimate. The bandwidth selectors are those used with \texttt{kdde(\texttt{deriv.order}=2)}.

– For kernel feature significance, the main function \texttt{kfs}. The hypothesis test at a point \( x \) is \( H_0(x) : Hf(x) < 0 \), i.e. the density Hessian matrix \( Hf(x) \) is negative definite. The test statistic is

\[
W(x) = \|S(x)^{-1/2} \text{vech} \hat{Hf}(x)\|^2
\]

where \( \hat{Hf} \) is the Hessian estimate, \text{vech} is the vector-half operator, and \( S \) is an estimate of the null variance. \( W(x) \) is approximately \( \chi^2 \) distributed with \( d(d+1)/2 \) degrees of freedom. If \( H_0(x) \) is rejected, then \( x \) belongs to a significant modal region. The bandwidth selectors are those used with \texttt{kdde(\texttt{deriv.order}=2)}. Its plot method is \texttt{plot.kfs}.

– For deconvolution density estimation, the main function is \texttt{kdcde}. A weighted kernel density estimation with the contaminated data \( W_1, \ldots, W_n \),

\[
\hat{f}_w(x) = n^{-1} \sum_{i=1}^{n} \alpha_i K_H(x - W_i),
\]

is utilised, where the weights \( \alpha_1, \ldots, \alpha_n \) are chosen via a quadratic optimisation involving the error variance and the regularisation parameter. The bandwidth selectors are those used with \texttt{kde}.

– Binned kernel estimation is an approximation to the exact kernel estimation and is available for \( d=1, 2, 3, 4 \). This makes kernel estimators feasible for large samples.

– For an overview of this package with 2-d density estimation, see \texttt{vignette(“kde”).}

– For \texttt{ks} \( \geq 1.11.1 \), the \texttt{misc3d} and \texttt{rgl} (3-d plot), \texttt{OceanView} (quiver plot), \texttt{oz} (Australian map) packages have been moved from Depends to Suggests. This was done to allow \texttt{ks} to be installed on systems where these latter graphical-based packages can’t be installed.

**Author(s)**


**References**


See Also

feature, sm, KernSmooth

---

### air

**Air quality measurements in an underground train station**

**Description**

This data set contains the hourly mean air quality measurements from 01 January 2013 to 31 December 2016 in the Chatelet underground train station in the Paris metro.

**Usage**

data(air)

**Format**

A matrix with 35039 rows and 8 columns. Each row corresponds to an hourly measurement. The first column is the date (yyyy-mm-dd), the second is the time (hh:mm), the third is the nitric oxide NO concentration (g/m3), the fourth is the nitrogen dioxide NO2 concentration (g/m3), the fifth is the concentration of particulate matter less than 10 microns PM10 (ppm), the sixth is the carbon dioxide concentration CO2 (g/m3), the seventh is the temperature (degrees Celsius), the eighth is the relative humidity (percentage).

**Source**


---

### binning

**Linear binning for multivariate data**

**Description**

Linear binning for 1- to 4-dimensional data.

**Usage**

binning(x, H, h, bgridsize, xmin, xmax, supp=3.7, w, gridtype="linear")
Arguments

- **x**  
  matrix of data values
- **h**  
  bandwidth matrix, scalar bandwidth
- **xmin, xmax**  
  vector of minimum/maximum values for grid
- **supp**  
  effective support for standard normal is [-supp, supp]
- **bgridsize**  
  vector of binning grid sizes
- **w**  
  vector of weights. Default is a vector of all ones.
- **gridtype**  
  not yet implemented

Details

As of ks 1.10.0, binning is available for unconstrained (non-diagonal) bandwidth matrices. Code is used courtesy of A. & J. Gramacki, and M.P. Wand. Default bgridsize are d=1: 401; d=2: rep(151, 2); d=3: rep(51, 3); d=4: rep(21, 4).

Value

Returns a list with 2 fields

- **counts**  
  linear binning counts
- **eval.points**  
  vector (d=1) or list (d>=2) of grid points in each dimension

References


Examples

```r
data(unicef)
ubinned <- binning(x=unicef)
```

```r
cardio
Foetal cardiotocograms
```

Description

This data set contains the cardiotocographic measurements from healthy, suspect and pathological foetuses.

Usage

```r
data(cardio)
```
Format

A matrix with 2126 rows and 8 columns. Each row corresponds to a foetal cardiotocogram. The class label for the foetal state is the last column: N = normal, S = suspect, P = pathological. Details for all variables are found in the link below.

Source


contour

Contours functions

Description

Contour levels and sizes.

Usage

contourLevels(x, ...)

## S3 method for class 'kde'
contourLevels(x, prob, cont, nlevels=5, approx=TRUE, ...)

## S3 method for class 'kda'
contourLevels(x, prob, cont, nlevels=5, approx=TRUE, ...)

contourSizes(x, abs.cont, cont=c(25,50,75), approx=TRUE)

Arguments

x an object of class kde or kda

prob vector of probabilities corresponding to highest density regions

cont vector of percentages which correspond to the complement of prob

abs.cont vector of absolute contour levels

nlevels number of pretty contour levels

approx flag to compute approximate contour levels. Default is TRUE.

... other parameters

Details

– For contourLevels, the most straightforward is to specify prob. Heights of the corresponding highest density region with probability prob are computed. The cont parameter here is consistent with cont parameter from plot.kde and plot.kda i.e. cont=(1-prob)*100%. If both prob and cont are missing then a pretty set of nlevels contours are computed.
–For contourSizes, the approximate Lebesgue measures are approximated by Riemann sums. These are rough approximations and depend highly on the estimation grid, and so should be interpreted carefully.

If approx=FALSE, then the exact KDE is computed. Otherwise it is interpolated from an existing KDE grid. This can dramatically reduce computation time for large data sets.

Value

–For contourLevels, for kde objects, returns vector of heights. For kda objects, returns a list of vectors, one for each training group.

–For contourSizes, an approximation of the Lebesgue measure of level set, i.e. length (d=1), area (d=2), volume (d=3), hyper-volume (d>4).

See Also

contour, contourLines

Examples

```r
set.seed(8192)
x <- rmvnorm.mixt(n=1000, mus=c(0,0), Sigmas=diag(2), props=1)
fhat <- kde(x=x, binned=TRUE)
contourLevels(fhat, cont=c(75, 50, 25))
contourSizes(fhat, cont=25, approx=TRUE)
  ## compare to approx circle of radius=0.75 with area=1.77
```

---

grevillea | Geographical locations of grevillea plants

Description

This data set contains the geographical locations of the specimens of *Grevillea uncinulata*, more commonly known as the Hook leaf grevillea, which is an endemic floral species to south Western Australia. This region is one of the 25 ‘biodiversity hotspots’ which are ‘areas featuring exceptional concentrations of endemic species and experiencing exceptional loss of habitat’.

Usage

data(grevillea)

Format

A matrix with 222 rows and 2 columns. Each row corresponds to a observed plant. The first column is the longitude (decimal degrees), the second is the latitude (decimal degrees).
Source


Biased cross-validation (BCV) bandwidth matrix selector for bivariate data

Description

BCV bandwidth matrix for bivariate data.

Usage

hbcv(x, whichbcv=1, Hstart, binned=FALSE, amise=FALSE, verbose=FALSE)
hbcv.diag(x, whichbcv=1, Hstart, binned=FALSE, amise=FALSE, verbose=FALSE)

Arguments

x          matrix of data values
whichbcv   1 = BCV1, 2 = BCV2. See details below.
Hstart     initial bandwidth matrix, used in numerical optimisation
binned     flag for binned kernel estimation. Default is FALSE.
amise      flag to return the minimal BCV value. Default is FALSE.
verbose    flag to print out progress information. Default is FALSE.

Details

Use hbcv for unconstrained bandwidth matrices and hbcv.diag for diagonal bandwidth matrices. These selectors are only available for bivariate data. Two types of BCV criteria are considered here. They are known as BCV1 and BCV2, from Sain, Baggerly & Scott (1994) and only differ slightly. These BCV surfaces can have multiple minima and so it can be quite difficult to locate the most appropriate minimum. Some times, there can be no local minimum at all so there may be no finite BCV selector.

For details about the advanced options for binned, Hstart, see hpi.

Value

BCV bandwidth matrix. If amise=TRUE then the minimal BCV value is returned too.

References

See Also

hlsvc, hpi, hscv

Examples

data(unicefd)
Hbcv(unicefd)
Hbcv.diag(unicefd)

histde

Histogram density estimate

Description

Histogram density estimate for 1- and 2-dimensional data.

Usage

histde(x, binw, xmin, xmax, adj=0)

## S3 method for class 'histde'
predict(object, ..., x)

Arguments

x matrix of data values
binw (vector) of binwidths
xmin, xmax vector of minimum/maximum values for grid
adj displacement of default anchor point, in percentage of 1 bin
object an object of class histde
... other parameters

Details

If binw is missing, the default binwidth is \( \hat{b}_i = 2 \cdot 3^{1/(d+2)} \pi^{d/(2d+4)} S_i n^{-1/(d+2)} \), the normal scale selector.

If xmin is missing then it defaults to the data minimum. If xmax is missing then it defaults to the data maximum.
Value

A histogram density estimate is an object of class histde which is a list with fields:

- `x` - data points - same as input
- `eval.points` - vector or list of points at which the estimate is evaluated
- `estimate` - density estimate at `eval.points`
- `binw` - (vector of) bandwidths
- `nbin` - (vector of) number of bins
- `names` - variable names

See Also

plot.histde

Examples

```r
## positive data example
set.seed(1992)
x <- 2*rnorm(100)
fhat <- histde(x=x)
plot(fhat, col=3)
points(c(0.5, 1), predict(fhat, x=c(0.5, 1)))

## large data example on a non-default grid
set.seed(1992)
x <- rmvnorm.minit(10000, mus=c(0,0), Sigmas=invvech(c(1,0.8,1)))
fhat <- histde(x=x, xmin=c(-5,-5), xmax=c(5,5))
plot(fhat)

## See other examples in ?plot.histde
```

Description

Least-squares cross-validation (LSCV) bandwidth matrix selector for multivariate data

Usage

```
Hlscv(x, Hstart, binned, bgridsize, amise=FALSE, deriv.order=0, verbose=FALSE, optim.fun="optim", trunc)
Hlscv.diag(x, Hstart, binned, bgridsize, amise=FALSE, deriv.order=0, verbose=FALSE, optim.fun="optim", trunc)
hlscv(x, binned=TRUE, bgridsize, amise=FALSE, deriv.order=0)
Hucv(...)
Hucv.diag(...)
hucv(...)
```
Arguments

- `x`: vector or matrix of data values
- `Hstart`: initial bandwidth matrix, used in numerical optimisation
- `binned`: flag for binned kernel estimation
- `bgridsize`: vector of binning grid sizes
- `amise`: flag to return the minimal LSCV value. Default is FALSE.
- `deriv.order`: derivative order
- `verbose`: flag to print out progress information. Default is FALSE.
- `optim.fun`: optimiser function: one of `nlm` or `optim`
- `trunc`: parameter to control truncation for numerical optimisation. Default is 4 for density.deriv>0, otherwise no truncation. For details see below.

Details

`hlscv` is the univariate LSCV selector of Bowman (1984) and Rudemo (1982). `hlscv` is a multivariate generalisation of this. Use `hlscv` for unconstrained bandwidth matrices and `hlscv.diag` for diagonal bandwidth matrices. `Hucv`, `Hucv.diag` and `hucv` are aliases with UCV (unbiased cross validation) instead of LSCV.

Truncation of the parameter space is usually required for the LSCV selector, for r > 0, to find a reasonable solution to the numerical optimisation. If a candidate matrix `h` is such that \( \text{det}(H) \) is not in \( [1/\text{trunc}, \text{trunc}]*\text{det}(H0) \) or \( \text{abs}(LSCV(H)) > \text{trunc}*\text{abs}(LSCV0) \) then the LSCV(H) is reset to LSCV0 where \( H0=Hns(x) \) and LSCV0=LSCV(H0).

From `ks` 1.11.1 onwards, the default optimisation function is `optim.fun="optim"`. For details about the advanced options for binned, `Hstart`, see `hpi`.

Value

LSCV bandwidth. If `amise=TRUE` then the minimal LSCV value is returned too.

References


See Also

`Hbcv`, `Hpi`, `Hscv`

Examples

```r
library(MASS)
data(forbes)
Hlscv(forbes)
hlscv(forbes$bp)
```
**Description**

Normal mixture bandwidth.

**Usage**

Hnm(x, deriv.order=0, G=1:9, subset.ind, mise.flag=FALSE, verbose, ...)
Hnm.diag(x, deriv.order=0, G=1:9, subset.ind, mise.flag=FALSE, verbose, ...)
hnm(x, deriv.order=0, G=1:9, subset.ind, mise.flag=FALSE, verbose, ...)

**Arguments**

- **x**: vector/matrix of data values
- **deriv.order**: derivative order
- **G**: range of number of mixture components
- **subset.ind**: index vector of subset of \( x \) for fitting
- **mise.flag**: flag to use MISE or AMISE minimisation. Default is FALSE.
- **verbose**: flag to print out progress information. Default is FALSE.
- **...**: other parameters for \texttt{mclust}

**Details**

The normal mixture fit is provided by the \texttt{Mclust} function in the \texttt{mclust} package. \texttt{Hnm} is then \texttt{Hmise.mixt} (if \texttt{mise.flag=TRUE}) or \texttt{Hamise.mixt} (if \texttt{mise.flag=FALSE}) with these fitted normal mixture parameters. Likewise for \texttt{Hnm.diag}, \texttt{hnm}.

**Value**

Normal mixture bandwidth. If \texttt{mise=TRUE} then the minimal MISE value is returned too.

**References**


**See Also**

\texttt{Hmise.mixt}, \texttt{Hamise.mixt}

**Examples**

```r
library(MASS)
data(forbes)
Hnm(forbes)
```
Hns

| Hns | Normal scale bandwidth |

### Description

Normal scale bandwidth.

### Usage

Hns(x, deriv.order=0)
Hns.diag(x)

hns(x, deriv.order=0)
Hns.kcde(x)

### Arguments

- **x**: vector/matrix of data values
- **deriv.order**: derivative order

### Details

Hns is equal to \(\frac{4}{n(d^2+2d+2)}\)\(^2\(\frac{2}{(d+2r+4)}\)\)\(\text{var}(x)\). \(n = \) sample size, \(d = \) dimension of data, \(r = \) derivative order. hns is the analogue of Hns for 1-d data. These can be used for density (derivative) estimators kde, kdde. The equivalents for distribution estimators kcde are Hns.kcde and hns.cde.

### Value

Normal scale bandwidth.

### References


### Examples

library(MASS)
data(forbes)
Hns(forbes, deriv.order=2)
hns(forbes$hp, deriv.order=2)
**Hpi**

*Plug-in bandwidth selector*

**Description**

Plug-in bandwidth for for 1- to 6-dimensional data.

**Usage**

```r
Hpi(x, nstage=2, pilot, pre="sphere", Hstart, binned, bgridsize, amise=FALSE, deriv.order=0, verbose=FALSE, optim.fun="optim")
Hpi.diag(x, nstage=2, pilot, pre="scale", Hstart, binned, bgridsize, amise=FALSE, deriv.order=0, verbose=FALSE, optim.fun="optim")
hpi(x, nstage=2, binned=TRUE, bgridsize, deriv.order=0)
```

**Arguments**

- **x**: vector or matrix of data values
- **nstage**: number of stages in the plug-in bandwidth selector (1 or 2)
- **pilot**: "amse" = AMSE pilot bandwidths
  "samse" = single SAMSE pilot bandwidth
  "unconstr" = single unconstrained pilot bandwidth
  "dscalar" = single pilot bandwidth for deriv.order >= 0
  "dunconstr" = single unconstrained pilot bandwidth for deriv.order >= 0
- **pre**: "scale" = pre.scale, "sphere" = preSphere
- **Hstart**: initial bandwidth matrix, used in numerical optimisation
- **binned**: flag for binned kernel estimation
- **bgridsize**: vector of binning grid sizes
- **amise**: flag to return the minimal scaled PI value
- **deriv.order**: derivative order
- **verbose**: flag to print out progress information. Default is FALSE.
- **optim.fun**: optimiser function: one of nlm or optim

**Details**

`hpi(. , deriv.order=0)` is the univariate plug-in selector of Wand & Jones (1994), i.e. it is exactly the same as `KernSmooth`’s dpik. For deriv.order>0, the formula is taken from Wand & Jones (1995). Hpi is a multivariate generalisation of this. Use Hpi for unconstrained bandwidth matrices and Hpi.diag for diagonal bandwidth matrices.

The default pilot is "samse" for d=2,r=0, and "dscalar" otherwise. For AMSE pilot bandwidths, see Wand & Jones (1994). For SAMSE pilot bandwidths, see Duong & Hazelton (2003). The latter is a modification of the former, in order to remove any possible problems with non-positive definiteness. Unconstrained and higher order derivative pilot bandwidths are from Chacon & Duong (2010).
For d=1, 2, 3, 4 and binned=TRUE, estimates are computed over a binning grid defined by bgridsize. Otherwise it's computed exactly. If Hstart is not given then it defaults to $h_{ns}(x)$.

From ks 1.11.1 onwards, the default optimisation function is optim.fun="optim". To reinstate the previous functionality, use optim.fun="nlm".

**Value**

Plug-in bandwidth. If amise=TRUE then the minimal scaled PI value is returned too.

**References**


**See Also**

*Hbcv, Hlscv, Hscv*

**Examples**

data(unicef)
\texttt{Hpi(unicef, pilot="dscalar")}
\texttt{hpi(unicef[,1])}

<table>
<thead>
<tr>
<th>hsct</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Haematopoietic stem cell transplant</em></td>
</tr>
</tbody>
</table>

**Description**

This data set contains the haematopoietic stem cell transplant (HSCT) measurements obtained a flow cytometer from mouse subjects. A flow cytometer measures the spectra of fluorescent signals from biological cell samples to study their properties.

**Usage**

data(hsct)
**Format**

A matrix with 39128 rows and 6 columns. The first column is the FITC-CD45.1 fluorescence (0-1023), the second is the PE-Ly65/Mac1 fluorescence (0-1023), the third is the PI-LiveDead fluorescence (0-1023), the fourth is the APC-CD45.2 fluorescence (0-1023), the fifth is the class label of the cell type (1, 2, 3, 4, 5), the sixth the mouse subject number (5, 6, 9, 12).

**Source**


---

**Hscv**

Smoothed cross-validation (SCV) bandwidth selector

---

**Description**

SCV bandwidth for 1- to 6-dimensional data.

**Usage**

```r
Hscv(x, nstage=2, pre="sphere", pilot, Hstart, binned, bgridsize, amise=FALSE, deriv.order=0, verbose=FALSE, optim.fun="optim")
Hscv.diag(x, nstage=2, pre="scale", pilot, Hstart, binned, bgridsize, amise=FALSE, deriv.order=0, verbose=FALSE, optim.fun="optim")
```

**Arguments**

- `x` vector or matrix of data values
- `pre` "scale" = `pre.scale`, "sphere" = `pre.sphere`
- `pilot` "amse" = AMSE pilot bandwidths
  - "samse" = single SAMSE pilot bandwidth
  - "unconstr" = single unconstrained pilot bandwidth
  - "dscalar" = single pilot bandwidth for deriv.order>0
  - "dunconstr" = single unconstrained pilot bandwidth for deriv.order>0
- `Hstart` initial bandwidth matrix, used in numerical optimisation
- `binned` flag for binned kernel estimation
- `bgridsize` vector of binning grid sizes
- `amise` flag to return the minimal scaled SCV value. Default is FALSE.
- `deriv.order` derivative order
- `verbose` flag to print out progress information. Default is FALSE.
- `optim.fun` optimiser function: one of `nlm` or `optim`
- `nstage` number of stages in the SCV bandwidth selector (1 or 2)
- `plot` flag to display plot of SCV(h) vs h (1-d only). Default is FALSE.
Details


The default pilot is “samse” for d=2, r=0, and “dscalar” otherwise. For SAMSE pilot bandwidths, see Duong & Hazelton (2005). Unconstrained and higher order derivative pilot bandwidths are from Chacon & Duong (2011).

For d=1, the selector hscv is not always stable for large sample sizes with binning. Examine the plot from hscv(), plot=TRUE) to determine the appropriate smoothness of the SCV function. Any non-smoothness is due to the discretised nature of binned estimation.

From ks 1.11.1 onwards, the default optimisation function is optim.fun="optim".

For details about the advanced options for binned, Hstart, see Hpi.

Value

SCV bandwidth. If amise=TRUE then the minimal scaled SCV value is returned too.

References


See Also

Hbcv, Hlscv, Hpi

Examples

data(unicef)
Hscv(unicef)
hscv(unicef[,1])

ise.mixt

---

**Description**

The global errors ISE (Integrated Squared Error), MISE (Mean Integrated Squared Error) and the AMISE (Asymptotic Mean Integrated Squared Error) for 1- to 6-dimensional data. Normal mixture densities have closed form expressions for the MISE and AMISE. So in these cases, we can numerically minimise these criteria to find MISE- and AMISE-optimal matrices.
ise.mixt

Usage

Hamise.mixt(mus, Sigmas, props, samp, Hstart, deriv.order=0)
Hamise.mixt(mus, Sigmas, props, samp, Hstart, deriv.order=0)
Hamise.mixt.diag(mus, Sigmas, props, samp, Hstart, deriv.order=0)
Hamise.mixt.diag(mus, Sigmas, props, samp, Hstart, deriv.order=0)
hamise.mixt(mus, sigmas, props, samp, hstart, deriv.order=0)
hamise.mixt(mus, sigmas, props, samp, hstart, deriv.order=0)
amise.mixt(H, mus, Sigmas, props, samp, h, sigmas, deriv.order=0)
ise.mixt(x, H, mus, Sigmas, props, h, sigmas, deriv.order=0, binned=FALSE, bgridsize)
mise.mixt(H, mus, Sigmas, props, samp, h, sigmas, deriv.order=0)

Arguments

mus         (stacked) matrix of mean vectors (>1-d), vector of means (1-d)
Sigmas, sigmas (stacked) matrix of variance matrices (>1-d), vector of standard deviations (1-d)
props       vector of mixing proportions
samp        sample size
Hstart, hstart initial bandwidth (matrix), used in numerical optimisation
deriv.order derivative order
x           matrix of data values
H, h        bandwidth (matrix)
binned      flag for binned kernel estimation. Default is FALSE.
bgridsize   vector of binning grid sizes

Details

ISE is a random variable that depends on the data x. MISE and AMISE are non-random and don’t depend on the data. For normal mixture densities, ISE, MISE and AMISE have exact formulas for all dimensions.

Value

Unconstrained MISE- or AMISE-optimal bandwidth matrix. ISE, MISE or AMISE value.

References


Examples

x <- rmvnorm.mixt(100)
Hamise.mixt(samp=nrow(x), mus=rep(0,2), Sigmas=var(x), props=1, deriv.order=1)
Kernel cumulative distribution/survival function estimate

**Description**

Kernel cumulative distribution/survival function estimate for 1- to 3-dimensional data.

**Usage**

```r
kcde(x, H, h, gridsize, gridtype, xmin, xmax, supp=3.7, eval.points,
     binned, bgridsize, positive=FALSE, adj.positive, w, verbose=FALSE,
     tail.flag="lower.tail")
Hpi.kcde(x, nstage=2, pilot, Hstart, binned, bgridsize, amise=FALSE,
         verbose=FALSE, optim.fun="optim")
Hpi.diag.kcde(x, nstage=2, pilot, Hstart, binned, bgridsize, amise=FALSE,
             verbose=FALSE, optim.fun="optim")
hpi.kcde(x, nstage=2, binned, amise=FALSE)
```

```r
## S3 method for class 'kcde'
predict(object, ..., x)
```

**Arguments**

- `x`: matrix of data values
- `H,h`: bandwidth matrix/scalar bandwidth. If these are missing, then `Hpi.kcde` or `hpi.kcde` is called by default.
- `gridsize`: vector of number of grid points
- `gridtype`: not yet implemented
- `xmin,xmax`: vector of minimum/maximum values for grid
- `supp`: effective support for standard normal
- `eval.points`: vector or matrix of points at which estimate is evaluated
- `binned`: flag for binned estimation. Default is FALSE.
- `bgridsize`: vector of binning grid sizes
- `positive`: flag if 1-d data are positive. Default is FALSE.
- `adj.positive`: adjustment applied to positive 1-d data
- `w`: not yet implemented
- `verbose`: flag to print out progress information. Default is FALSE.
- `tail.flag`: "lower.tail" = cumulative distribution, "upper.tail" = survival function
- `nstage`: number of stages in the plug-in bandwidth selector (1 or 2)
- `pilot`: "dscalar" = single pilot bandwidth (default for `Hpi.diag.kcde`)
  "dunconstr" = single unconstrained pilot bandwidth (default for `Hpi.kcde`
- `Hstart`: initial bandwidth matrix, used in numerical optimisation
amise flag to return the minimal scaled PI value
optim.fun optimiser function: one of nlm or optim
object object of class kcde

Details

If tail.flag="lower.tail" then the cumulative distribution function \( \Pr(X \leq x) \) is estimated, otherwise if tail.flag="upper.tail", it is the survival function \( \Pr(X > x) \). For \( d>1 \), \( \Pr(X \leq x) \neq 1 - \Pr(X > x) \).

If the bandwidth \( h \) is missing in kcde, then the default bandwidth is the plug-in selector \( \text{hpi.kcde} \). Likewise for missing \( h \). No pre-scaling/pre-sphering is used since the \( \text{hpi.kcde} \) is not invariant to translation/dilation.

From ks 1.11.1 onwards, the default optimisation function is optim.fun="optim".

The effective support, binning, grid size, grid range, positive parameters are the same as \text{kde}.

Value

A kernel cumulative distribution estimate is an object of class kcde which is a list with fields:

- \( x \) data points - same as input
- eval.points vector or list of points at which the estimate is evaluated
- estimate cumulative distribution/survival function estimate at eval.points
- \( h \) scalar bandwidth (1-d only)
- \( H \) bandwidth matrix
- gridtype "linear"
- gridded flag for estimation on a grid
- binned flag for binned estimation
- names variable names
- \( w \) weights
- tail "lower.tail"=cumulative distribution, "upper.tail"=survival function

References


See Also

kde, plot.kcde
Examples

```r
library(MASS)
data(iris)
Fhat <- kcde(iris[,1:2])
predict(Fhat, x=as.matrix(iris[,1:2]))

## See other examples in ?plot.kcde
```

### kcopula

**Kernel copula (density) estimate**

**Description**

Kernel copula and copula density estimator for 2-dimensional data.

**Usage**

```r
kcopula(x, H, hs, gridsize, gridtype, xmin, xmax, supp=3.7, eval.points,
    binned, bgridsize, w, verbose=FALSE, marginal="kernel")
kcopula.de(x, H, gridsize, gridtype, xmin, xmax, supp=3.7,
    eval.points, binned, bgridsize, w, verbose=FALSE, compute.cont=TRUE,
    approx.cont=TRUE, marginal="kernel", boundary.supp, boundary.kernel="beta")
```

**Arguments**

- `x`  
  matrix of data values
- `H,hs`  
  bandwidth matrix. If these are missing, Hpi.kcde/Hpi or hpi.kcde/hpi is called by default.
- `gridsize`  
  vector of number of grid points
- `gridtype`  
  not yet implemented
- `xmin,xmax`  
  vector of minimum/maximum values for grid
- `supp`  
  effective support for standard normal
- `eval.points`  
  matrix of points at which estimate is evaluated
- `binned`  
  flag for binned estimation
- `bgridsize`  
  vector of binning grid sizes
- `w`  
  vector of weights. Default is a vector of all ones.
- `verbose`  
  flag to print out progress information. Default is FALSE.
- `marginal`  
  "kernel" = kernel cdf or "empirical" = empirical cdf to calculate pseudo-uniform values. Default is "kernel".
- `compute.cont`  
  flag for computing 1% to 99% probability contour levels. Default is TRUE.
- `approx.cont`  
  flag for computing approximate probability contour levels. Default is TRUE.
- `boundary.supp`  
  effective support for boundary region
- `boundary.kernel`  
  "beta" = 2nd form of beta boundary kernel, "linear" = linear boundary kernel
Details

For kernel copula estimates, a transformation approach is used to account for the boundary effects. If \( h \) is missing, the default is \( h_{pi} \). If \( h \) are missing, the default is \( h_{pi} \).

For kernel copula density estimates, for those points which are in the interior region, the usual kernel density estimator (\( \text{kde} \)) is used. For those points in the boundary region, a product beta kernel based on the boundary corrected univariate beta kernel of Chen (1999) is used (\( \text{kde boundary} \)). If \( h \) is missing, the default is \( h_{pi} \); if \( h \) are missing, the default is \( h_{pi} \).

The effective support, binning, grid size, grid range parameters are the same as for \( \text{kde} \).

Value

A kernel copula estimate, output from \( \text{kcopula} \), is an object of class \( \text{kcopula} \). A kernel copula density estimate, output from \( \text{kcopula.de} \), is an object of class \( \text{kde} \). These two classes of objects have the same fields as \( \text{kde} \) and \( \text{kde} \) objects respectively, except for

- \( x \) - pseudo-uniform data points
- \( x.\text{orig} \) - data points - same as input
- \( \text{marginal} \) - marginal function used to compute pseudo-uniform data
- \( \text{boundary} \) - flag for data points in the boundary region (\( \text{kcopula.de} \) only)

References


See Also

\( \text{kde}, \text{kde} \)

Examples

```r
library(MASS)
data(fgl)
x <- fgl[,c("RI", "Na")]
Chat <- kcopula(x=x)
plot(Chat, disp="persp", thin=3, col="white", border=1)
```
Kernel discriminant analysis for 1- to d-dimensional data.

Usage

kda(x, x.group, Hs, hs, prior.prob=NULL, gridsize, xmin, xmax, supp=3.7,
    eval.points, binned, bgridsize, w, compute.cont=TRUE, approx.cont=TRUE,
    kde.flag=TRUE)

HKda(x, x.group, Hstart, bw="plugin", ...)
HKda.diag(x, x.group, bw="plugin", ...)
hkda(x, x.group, bw="plugin", ...)

## S3 method for class 'kda'
predict(object, ..., x)

compare(x.group, est.group, by.group=FALSE)
compare.kda.cv(x, x.group, bw="plugin", prior.prob=NULL, Hstart, by.group=FALSE,
    verbose=FALSE, recompute=FALSE, ...)
compare.kda.diag.cv(x, x.group, bw="plugin", prior.prob=NULL, by.group=FALSE,
    verbose=FALSE, recompute=FALSE, ...)

Arguments

x matrix of training data values
x.group vector of group labels for training data
Hs, hs (stacked) matrix of bandwidth matrices/vector of scalar bandwidths. If these are missing, HKda or hkda is called by default.
prior.prob vector of prior probabilities
gridsize vector of grid sizes
xmin, xmax vector of minimum/maximum values for grid
supp effective support for standard normal
eval.points vector or matrix of points at which estimate is evaluated
binned flag for binned estimation
bgridsize vector of binning grid sizes
w vector of weights. Not yet implemented.
compute.cont flag for computing 1% to 99% probability contour levels. Default is TRUE.
approx.cont flag for computing approximate probability contour levels. Default is TRUE.
kde.flag flag for computing KDE on grid. Default is TRUE.
object object of class kda
bw        bandwidth: "plugin" = plug-in, "lscv" = LSCV, "scv" = SCV
hstart    (stacked) matrix of initial bandwidth matrices, used in numerical optimisation
est.group vector of estimated group labels
by.group  flag to give results also within each group
verbose   flag for printing progress information. Default is FALSE.
recompute flag for recomputing the bandwidth matrix after excluding the i-th data item
...       other optional parameters for bandwidth selection, see \texttt{Hpi}, \texttt{Hlscv}, \texttt{Hscv}

Details

If the bandwidths $h$s are missing from \texttt{kda}, then the default bandwidths are the plug-in selectors \texttt{hkda()}, \texttt{bw}="plugin"). Likewise for missing $h$s. Valid options for $bw$ are "plugin", "lscv" and "scv" which in turn call \texttt{Hpi}, \texttt{Hlscv} and \texttt{Hscv}.

The effective support, binning, grid size, grid range, positive parameters are the same as \texttt{kde}.

If prior probabilities are known then set \texttt{prior.prob} to these. Otherwise \texttt{prior.prob=NULL} uses the sample proportions as estimates of the prior probabilities.

As of \texttt{ks} 1.8.11, \texttt{kda.kde} has been subsumed into \texttt{kda}, so all prior calls to \texttt{kda.kde} can be replaced by \texttt{kda}. To reproduce the previous behaviour of \texttt{kda}, the command is \texttt{kda(, kde.flag=FALSE)}.

Value

–For \texttt{kde.flag=TRUE}, a kernel discriminant analysis is an object of class \texttt{kda} which is a list with fields

\begin{itemize}
  \item \texttt{x} list of data points, one for each group label
  \item \texttt{estimate} list of density estimates at \texttt{eval.points}, one for each group label
  \item \texttt{eval.points} vector or list of points that the estimate is evaluated at, one for each group label
  \item \texttt{h} vector of bandwidths (1-d only)
  \item \texttt{H} stacked matrix of bandwidth matrices or vector of bandwidths
  \item \texttt{gridded} flag for estimation on a grid
  \item \texttt{binned} flag for binned estimation
  \item \texttt{w} weights
  \item \texttt{prior.prob} prior probabilities
  \item \texttt{x.group} group labels - same as input
  \item \texttt{x.group.estimate} estimated group labels. If the test data \texttt{eval.points} are given then these are classified. Otherwise the training data \texttt{x} are classified.
\end{itemize}

For \texttt{kde.flag=FALSE}, which is always the case for $d > 3$, then only the vector of estimated group labels is returned.

–The result from \texttt{hkda} and \texttt{hkda.diag} is a stacked matrix of bandwidth matrices, one for each training data group. The result from \texttt{hkda} is a vector of bandwidths, one for each training group.

–The \texttt{compare} functions create a comparison between the true group labels \texttt{x.group} and the estimated ones. It returns a list with fields
cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

cross

cross-classification table with the rows indicating the true group and the columns the estimated group

error

misclassification rate (MR)

In the case where the test data is independent of the training data, \texttt{compare} computes \( MR = \frac{\text{number of points wrongly classified}}{\text{total number of points}} \). In the case where the test data are not independent e.g. we are classifying the training data set itself, then the cross validated estimate of \( MR \) is more appropriate. These are implemented as \texttt{compare} \_ \texttt{kda} \_ \texttt{cv} (unconstrained bandwidth selectors) and \texttt{compare} \_ \texttt{kda} \_ \texttt{diag} \_ \texttt{cv} (for diagonal bandwidth selectors). These functions are only available for \( d > 1 \).

If \texttt{by.group=} \texttt{FALSE} then only the total \( MR \) rate is given. If it is set to \texttt{TRUE}, then the \( MR \) rates for each class are also given (estimated number in group divided by true number).

References


See Also

\texttt{plot.kda}

Examples

```r
set.seed(8192)
x <- \texttt{c(rnorm.mixt(n=100, mus=1), rnorm.mixt(n=100, mus=-1))}
x.gr <- \texttt{rep(c(1,2), times=c(100,100))}
y <- \texttt{c(rnorm.mixt(n=100, mus=1), rnorm.mixt(n=100, mus=-1))}
y.gr <- \texttt{rep(c(1,2), times=c(100,100))}
kda.gr <- \texttt{kda(x, x.gr)}
y.gr.est <- \texttt{predict(kda.gr, x=y)}
\texttt{compare(y.gr, y.gr.est)}

## See other examples in ? plot.kda
```

\begin{tabular}{ll}
\multicolumn{2}{l}{kdcde} \\
\multicolumn{2}{l}{\textit{Deconvolution kernel density derivative estimate}} \\
\end{tabular}

Description

Deconvolution kernel density derivative estimate for 1- to 6-dimensional data.

Usage

\texttt{kdcde(x, H, h, Sigma, sigma, reg, bgridsize, gridsize, binned, verbose=FALSE, \ldots)}
\texttt{dckde(\ldots)}
Arguments

- **x**: matrix of data values
- **H, h**: bandwidth matrix/scalar bandwidth. If these are missing, \( h_{\text{p}i} \) or \( h_{\text{pi}} \) is called by default.
- **Sigma, sigma**: error variance matrix
- **reg**: regularisation parameter
- **gridsize**: vector of number of grid points
- **binned**: flag for binned estimation
- **bggridsize**: vector of binning grid sizes
- **verbose**: flag to print out progress information. Default is FALSE.
- **...**: other parameters to \( \text{kde} \)

Details

A weighted kernel density estimate is utilised to perform the deconvolution. The weights are obtained from from a quadratic programming problem, and then input into \( \text{kde}(\cdot,w=) \). This weighted estimate also requires an estimate of the error variance matrix from repeated observations, and of the regularisation parameter. If the latter is missing, it is calculated internally using a 5-fold cross validation method. See Hazelton & Turlach (2009). \( \text{dckde} \) is an alias for \( \text{kdcde} \).

If the bandwidth \( H \) is missing from \( \text{kde} \), then the default bandwidth is the plug-in selector \( h_{\text{p}i} \). Likewise for missing \( h \).

The effective support, binning, grid size, grid range, positive parameters are the same as \( \text{kde} \).

Value

A deconvolution kernel density derivative estimate is an object of class \( \text{kde} \) which is a list with fields:

- **x**: data points - same as input
- **eval.points**: vector or list of points at which the estimate is evaluated
- **estimate**: density estimate at eval.points
- **h**: scalar bandwidth (1-d only)
- **H**: bandwidth matrix
- **gridtype**: "linear"
- **gridded**: flag for estimation on a grid
- **binned**: flag for binned estimation
- **names**: variable names
- **w**: weights
- **cont**: probability contour levels (if compute.cont=TRUE)

References

See Also

kdde

Examples

data(air)
air <- air[, c("date", "time", "co2", "pm10")]
air2 <- reshape(air, idvar="date", timevar="time", direction="wide")
air <- as.matrix(na.omit(air[,c("co2.20:00", "pm10.20:00")]))
Sigma.air <- diag(c(var(air2[,"co2.19:00"] - air2["co2.21:00"], na.rm=TRUE),
var(air2[,"pm10.19:00"] - air2[,"pm10.21:00"], na.rm=TRUE)))
fhat.air.dec <- kdcde(x=air, Sigma=Sigma.air, reg=0.00021, verbose=TRUE)
plot(fhat.air.dec, drawlabels=FALSE, display="filled.contour", lwd=1)

kdde

Kernel density derivative estimate

Description

Kernel density derivative estimate for 1- to 6-dimensional data.

Usage

kdde(x, H, h, deriv.order=0, gridsize, gridtype, xmin, xmax, supp=3.7,
     eval.points, binned, bgridsize, positive=FALSE, adj.positive, w,
     deriv.vec=TRUE, verbose=FALSE)
kcurv(fhat, compute.cont=TRUE)

## S3 method for class 'kdde'
predict(object, ..., x)

Arguments

x          matrix of data values
H, h       bandwidth matrix/scalar bandwidth. If these are missing, Hpi or hpi is called
            by default.
deriv.order derivative order (scalar)
gridsize   vector of number of grid points
gridtype   not yet implemented
xmin, xmax vector of minimum/maximum values for grid
supp       effective support for standard normal
eval.points vector or matrix of points at which estimate is evaluated
binned     flag for binned estimation
bgridsize  vector of binning grid sizes
**positive**  flag if data are positive (1-d, 2-d). Default is FALSE.

**adj.positive**  adjustment applied to positive 1-d data

**w**  vector of weights. Default is a vector of all ones.

**deriv.vec**  flag to compute all derivatives in vectorised derivative. Default is TRUE. If FALSE then only the unique derivatives are computed.

**verbose**  flag to print out progress information. Default is FALSE.

**compute.cont**  flag for computing 1% to 99% probability contour levels. Default is TRUE.

**fhat**  object of class kdde with deriv.order=2

**object**  object of class kdde

... other parameters

**Details**

For each partial derivative, for grid estimation, the estimate is a list whose elements correspond to the partial derivative indices in the rows of deriv.ind. For points estimation, the estimate is a matrix whose columns correspond to rows of deriv.ind.

If the bandwidth \( h \) is missing from kdde, then the default bandwidth is the plug-in selector \( \text{hpi} \). Likewise for missing \( h \).

The effective support, binning, grid size, grid range, positive parameters are the same as \( \text{kde} \).

The summary curvature is computed by kcurv, i.e.

\[
\hat{s}(x) = -1\{D^2\hat{f}(x) < 0\}\text{abs}(|D^2\hat{f}(x)|)
\]

where \( D^2\hat{f}(x) \) is the kernel Hessian matrix estimate. So \( \hat{s} \) calculates the absolute value of the determinant of the Hessian matrix and whose sign is the opposite of the negative definiteness indicator.

**Value**

A kernel density derivative estimate is an object of class kdde which is a list with fields:

**x**  data points - same as input

**eval.points**  vector or list of points at which the estimate is evaluated

**estimate**  density derivative estimate at eval.points

**h**  scalar bandwidth (1-d only)

**H**  bandwidth matrix

**gridtype**  "linear"

**gridded**  flag for estimation on a grid

**binned**  flag for binned estimation

**names**  variable names

**w**  weights

**deriv.order**  derivative order (scalar)

**deriv.ind**  each row is a vector of partial derivative indices
See Also

kde

Examples

set.seed(8192)
x <- rmvnorm.mixt(1000, mus=c(0,0), Sigmas=invvech(c(1,0.8,1)))
hat <- kde(x=x, deriv.order=1) ## gradient [df/dx, df/dy]
predict(hat, x=x[1:5,])

## See other examples in ? plot.kde

---

**kde**

*Kernel density estimate*

---

**Description**

Kernel density estimate for 1- to 6-dimensional data.

**Usage**

\[
kde(x, H, h, gridsize, gridtype, xmin, xmax, supp=3.7, eval.points,
\text{binned, bgridsize, positive=FALSE, adj.positive, w,}
\text{approx.cont=TRUE, approx.cont=TRUE, unit.interval=FALSE,}
\text{verbose=FALSE})
\]

## S3 method for class 'kde'
predict(object, ..., x, zero.flag=TRUE)

**Arguments**

- **x**: matrix of data values
- **H, h**: bandwidth matrix/scalar bandwidth. If these are missing, \( hpi \) or \( hpi \) is called by default.
- **gridsize**: vector of number of grid points
- **gridtype**: not yet implemented
- **xmin, xmax**: vector of minimum/maximum values for grid
- **supp**: effective support for standard normal
- **eval.points**: vector or matrix of points at which estimate is evaluated
- **binned**: flag for binned estimation.
- **bgridsize**: vector of binning grid sizes
- **positive**: flag if data are positive (1-d, 2-d). Default is FALSE.
- **adj.positive**: adjustment applied to positive 1-d data
- **w**: vector of weights. Default is a vector of all ones.
compute.cont  flag for computing 1% to 99% probability contour levels. Default is TRUE.
approx.cont  flag for computing approximate probability contour levels. Default is TRUE.
unit.interval  flag if 1-d data are bounded on unit interval [0,1]. Default is FALSE.
verbose  flag to print out progress information. Default is FALSE.
object  object of class kde
zero.flag  interpolated values \text{object$estimate of when x outside of object$eval.points=0} \text{ (if TRUE), = nearest object$estimate (if FALSE)}

...  other parameters

Details

For d=1, if \( h \) is missing, the default bandwidth is \( h_{pi} \). For d>1, if \( H \) is missing, the default is \( H_{pi} \).

For d=1, if \text{positive=TRUE} then \( x \) is transformed to \( \log(x+\text{adj.positive}) \) where the default \text{adj.positive} is the minimum of \( x \). This is known as a log transformation density estimate.

For d=1, 2, 3, and if \text{eval.points} is not specified, then the density estimate is computed over a grid defined by \text{gridsize} (if \text{binned=FALSE}) or by \text{bgridsize} (if \text{binned=TRUE}). This form is suitable for visualisation in conjunction with the \text{plot} method.

For d=4, 5, 6, and if \text{eval.points} is not specified, then the density estimate is computed over a grid defined by \text{gridsize}.

-If \text{eval.points} is specified, as a vector (d=1) or as a matrix (d=2, 3, 4), then the density estimate is computed at \text{eval.points}. This form is suitable for numerical summaries (e.g. maximum likelihood), and is not compatible with the \text{plot} method. Despite that the density estimate is returned only at \text{eval.points}, by default, a binned gridded estimate is calculated first and then the density estimate at \text{eval.points} is computed using the \text{predict} method. If this default intermediate binned grid estimate is not required, then set \text{binned=FALSE} to compute directly the exact density estimate at \text{eval.points}.

Binned kernel estimation is an approximation to the exact kernel estimation and is available for d=1, 2, 3, 4. This makes kernel estimators feasible for large samples. The default value of the binning flag \text{binned} is \text{n>1 (d=1), n>500 (d=2), n>1000 (d>=3)}.

The default \text{bgridsize, gridsize} are 401 (d=1), \text{rep(151, 2) (d=2), rep(51, 3) (d=3), rep(21,4) (d=4).}

The effective support for a normal kernel is where all values outside \([-\text{supp, supp}]^d\) are zero.

The default \text{xmin} is \text{min(x)+Hmax*supp} and \text{xmax} is \text{max(x)+Hmax*supp} where \( H_{\text{max}} \) is the maximum of the diagonal elements of \( H \). The grid produced is the outer product of \( c(x_{\text{min}[1]}, x_{\text{max}[1]}, ..., c(x_{\text{min}[d]}, x_{\text{max}[d])}.\)

Value

A kernel density estimate is an object of class \text{kde} which is a list with fields:

- \text{x}  data points - same as input
- \text{eval.points}  vector or list of points at which the estimate is evaluated
- \text{estimate}  density estimate at \text{eval.points}
- \text{h}  scalar bandwidth (1-d only)
Kernel density estimate for bounded data

**Description**

Kernel density estimate for bounded 2-dimensional data.

**Usage**

```r
kde.boundary(x, H, gridtype, xmin, xmax, supp=3.7, eval.points,
  gridsize, binned=FALSE, bgridsize, w, compute.cont=TRUE, approx.cont=TRUE,
  boundary.supp, boundary.kernel="beta", verbose=FALSE)
```
Arguments

- **x**: matrix of data values
- **H,h**: bandwidth matrix/scalar bandwidth. If these are missing, hpi or hpi is called by default.
- **gridsize**: vector of number of grid points
- **gridtype**: not yet implemented
- **xmin,xmax**: vector of minimum/maximum values for grid
- **supp**: effective support for standard normal
- **eval.points**: vector or matrix of points at which estimate is evaluated
- **binned**: flag for binned estimation.
- **bgridsize**: vector of binning grid sizes
- **w**: vector of weights. Default is a vector of all ones.
- **compute.cont**: flag for computing 1% to 99% probability contour levels. Default is TRUE.
- **approx.cont**: flag for computing approximate probability contour levels. Default is TRUE.
- **boundary.supp**: effective support for boundary region
- **boundary.kernel**: "beta" = 2nd form of beta boundary kernel, "linear" = linear boundary kernel
- **verbose**: flag to print out progress information. Default is FALSE.

Details

There are two forms of density estimates which are suitable for bounded data, based on the modifying the kernel function. For **boundary.kernel="beta"**, the 2nd form of the Beta boundary kernel of Chen (1999) is employed. It is suited for rectangular data boundaries.

For **boundary.kernel="linear"**, the linear boundary kernel of Hazelton & Marshall (2009) is employed. It is suited for arbitrarily shaped data boundaries, though it is currently only implemented for rectangular boundaries.

Value

A kernel density estimate for bounded data is an object of class kde.

References


See Also

- **kde**
Examples

```r
data(worldbank)
w <- as.matrix(na.omit(worldbank[,c("internet", "ag.value")]))
fh <- kde(x=w)
fh.beta <- kde.boundary(x=w, xmin=c(0,0), xmax=c(100,100), boundary.kernel="beta")
fh.LB <- kde.boundary(x=w, xmin=c(0,0), xmax=c(100,100), boundary.kernel="linear")

plot(fh, xlim=c(0,100), ylim=c(0,100))
plot(fh.beta, add=TRUE, col=2)
rect(0,0,100,100, lty=2)
plot(fh, xlim=c(0,100), ylim=c(0,100))
plot(fh.LB, add=TRUE, col=3)
rect(0,0,100,100, lty=2)
```

---

**kde.local.test**  
*Kernel density based local two-sample comparison test*

**Description**

Kernel density based local two-sample comparison test for 1- to 6-dimensional data.

**Usage**

```r
kde.local.test(x1, x2, H1, H2, h1, h2, fhat1, fhat2, gridsize, binned,
               bgridsize, verbose = FALSE, supp = 3.7, mean.adj = FALSE, signif.level = 0.05,
               min.ESS, xmin, xmax)
```

**Arguments**

- `x1, x2`: vector/matrix of data values
- `H1, H2, h1, h2`: bandwidth matrices/scalar bandwidths. If these are missing, `hpi` or `hpi` is called by default.
- `fhat1, fhat2`: objects of class `kde`
- `binned`: flag for binned estimation
- `gridsize`: vector of grid sizes
- `bgridsize`: vector of binning grid sizes
- `verbose`: flag to print out progress information. Default is `FALSE`.
- `supp`: effective support for normal kernel
- `mean.adj`: flag to compute second order correction for mean value of critical sampling distribution. Default is `FALSE`. Currently implemented for `d<=2` only.
- `signif.level`: significance level. Default is `0.05`.
- `min.ESS`: minimum effective sample size. See below for details.
- `xmin, xmax`: vector of minimum/maximum values for grid
Details

The null hypothesis is \( H_0(x) : f_1(x) = f_2(x) \) where \( f_1, f_2 \) are the respective density functions. The measure of discrepancy is \( U(x) = \left( f_1(x) - f_2(x) \right)^2 \). Duong (2013) shows that the test statistic obtained, by substituting the KDEs for the true densities, has a null distribution which is asymptotically chi-squared with 1 d.f.

The required input is either \( x_1, x_2 \) and \( h_1, h_2 \), or \( \hat{f}_1, \hat{f}_2 \), i.e. the data values and bandwidths or objects of class kde. In the former case, the kde objects are created. If the \( H_1, H_2 \) are missing then the default are the plugin selectors \( \hat{h}_{pi} \). Likewise for missing \( h_1, h_2 \).

The \texttt{mean.adj} flag determines whether the second order correction to the mean value of the test statistic should be computed. \texttt{min.ESS} is borrowed from Godtliebsen et al. (2002) to reduce spurious significant results in the tails, though by it is usually not required for small to moderate sample sizes.

Value

A kernel two-sample local significance is an object of class \texttt{kde.local.test} which is a list with fields:

- \texttt{fhat1, fhat2} kernel density estimates, objects of class \texttt{kde}
- \texttt{chisq} chi squared test statistic
- \texttt{pvalue} matrix of local p-values at each grid point
- \texttt{fhat.diff} difference of KDEs
- \texttt{mean.fhat.diff} mean of the test statistic
- \texttt{var.fhat.diff} variance of the test statistic
- \texttt{fhat.diff.pos} binary matrix to indicate locally significant \( \hat{f}_1 > \hat{f}_2 \)
- \texttt{fhat.diff.neg} binary matrix to indicate locally significant \( \hat{f}_1 < \hat{f}_2 \)
- \texttt{n1, n2} sample sizes
- \texttt{H1, H2, h1, h2} bandwidth matrices/scalar bandwidths

References


See Also

\texttt{kde.test, plot.kde.loctest}

Examples

```r
library(MASS)
x1 <- crabs[crabs$sp="B", 4]
x2 <- crabs[crabs$sp="O", 4]
loct <- kde.local.test(x1=x1, x2=x2)
plot(loct)
```
kde.test

Kernel density based global two-sample comparison test

Description

Kernel density based global two-sample comparison test for 1- to 6-dimensional data.

Usage

kde.test(x1, x2, H1, H2, h1, h2, psi1, psi2, var.fhat1, var.fhat2, binned=FALSE, bgridsize, verbose=FALSE)

Arguments

x1,x2 vector/matrix of data values
H1,H2,h1,h2 bandwidth matrices/scalar bandwidths. If these are missing, Hpi.kfe, hpi.kfe is called by default.
psi1,psi2 zero-th order kernel functional estimates
var.fhat1, var.fhat2 sample variance of KDE estimates evaluated at x1, x2
binned flag for binned estimation. Default is FALSE.
bgridsize vector of binning grid sizes
verbose flag to print out progress information. Default is FALSE.

Details

The null hypothesis is \( H_0 : f_1 \equiv f_2 \) where \( f_1, f_2 \) are the respective density functions. The measure of discrepancy is the integrated squared error (ISE) \( T = \int [f_1(x) - f_2(x)]^2 \, dx \). If we rewrite this as \( T = \psi_{0,1} - \psi_{0,12} - \psi_{0,21} + \psi_{0,2} \) where \( \psi_{0,uv} = \int f_u(x) f_v(x) \, dx \), then we can use kernel functional estimators. This test statistic has a null distribution which is asymptotically normal, so no bootstrap resampling is required to compute an approximate p-value.

If \( H_1, H_2 \) are missing then the plug-in selector Hpi.kfe is automatically called by kde.test to estimate the functionals with kfe( deriv.order=0 ). Likewise for missing \( h_1, h_2 \).

As of ks 1.8.8, kde.test(binned=TRUE) invokes binned estimation for the computation of the bandwidth selectors, and not the test statistic and p-value.

## see examples in ? plot.kde.loctest
**Value**

A kernel two-sample global significance test is a list with fields:

- `tstat` T statistic
- `zstat` z statistic - normalized version of Tstat
- `pvalue` p-value of the double sided test
- `mean, var` mean and variance of null distribution
- `var.fhat1, var.fhat2` sample variances of KDE values evaluated at data points
- `n1, n2` sample sizes
- `H1, H2` bandwidth matrices
- `psi1, psi2, psi12, psi21, psi2` kernel functional estimates

**References**


**See Also**

[kde.local.test](#)

**Examples**

```r
set.seed(8192)
samp <- 1000
x <- rnorm.mixt(n=samp, mus=0, sigmas=1, props=1)
y <- rnorm.mixt(n=samp, mus=0, sigmas=1, props=1)
kde.test(x1=x, x2=y)$pvalue  ## accept H0: f1=f2

library(MASS)
data(crabs)
x1 <- crabs[crabs$sp=="B", c(4,6)]
x2 <- crabs[crabs$sp=="O", c(4,6)]
kde.test(x1=x1, x2=x2)$pvalue  ## reject H0: f1=f2
```

---

**kde.truncate**  *Truncated kernel density derivative estimate*

**Description**

Truncated kernel density derivative estimate for 2-dimensional data.
Usage

kde.truncate(fhat, boundary)
kdde.truncate(fhat, boundary)

Arguments

fhat          object of class kde or kdde
boundary      two column matrix delimiting the boundary for truncation
...           other parameters

Details

A simple truncation is performed on the kernel estimator. All the points in the estimation grid which are outside of the regions delimited by boundary are set to 0, and their probability mass is distributed proportionally to the density (derivative) value.

Value

A truncated kernel density (derivative) estimate inherits the same object class as the input estimate.

See Also

kde, kdde

Examples

library(oz)
data(grevillea)
wa.coast <- ozRegion(section=1)
wa.polygon <- cbind(wa.coast$lines[[1]]$x, wa.coast$lines[[1]]$y)
fhat <- kde(x=grevillea)
fhat <- kde.truncate(fhat, wa.polygon)
oz(section=1, xlim=c(114.75, 121.5), ylim=c(-33, -31.5))
plot(fhat, add=TRUE, cont=seq(10,90,by=10), col=2, drawlabels=FALSE, drawpoints=TRUE)

kdr

Kernel density ridge estimation

Description

Kernel density ridge estimation for 2- to 3-dimensional data.

Usage

kdr(x, y, H, p=1, max.iter=400, tol.iter, tol.seg, min.seg.size,
     keep.path=FALSE, gridsize, xmin, xmax, binned, bgridsize, w, fhat,
     density.cutoff, verbose=FALSE)
**Arguments**

- **x**: matrix of data values
- **y**: matrix of initial values
- **p**: dimension of density ridge
- **H**: bandwidth matrix/scalar bandwidth. If missing, \( \text{Hpi}(x, \text{deriv.order}=2) \) is called by default.
- **max.iter**: maximum number of iterations. Default is 400.
- **tol.iter**: distance under which two successive iterations are considered convergent. Default is 0.001*min marginal IQR of \( x \).
- **tol.seg**: distance under which two segments are considered to form one segment. Default is 0.01*max marginal IQR of \( x \).
- **min.seg.size**: minimum length of a segment of a density ridge. Default is \( \text{round}(0.001 \times \text{nrow}(y), 0) \).
- **keep.path**: flag to store the density gradient ascent paths. Default is FALSE.
- **gridsize**: vector of number of grid points
- **xmin,xmax**: vector of minimum/maximum values for grid
- **binned**: flag for binned estimation.
- **bgridsize**: vector of binning grid sizes
- **w**: vector of weights. Default is a vector of all ones.
- **fhat**: kde of \( x \). If missing, \( \text{kde}(x=x, w=w) \) is executed.
- **density.cutoff**: density threshold under which the \( y \) are excluded from the density ridge estimation. Default is \( \text{contourLevels}(\text{fhat}, \text{cont}=99) \).
- **verbose**: flag to print out progress information. Default is FALSE.
- **...**: other parameters to \( \text{kde} \)

**Details**

Kernel density ridge estimation is based on reduced dimension kernel mean shift. *** To do. ***

See Ozertem & Erdogmus (2011).

If \( y \) is missing, then it defaults to the grid of size gridsize spanning from \( \text{xmin} \) to \( \text{xmax} \).

If the bandwidth \( H \) is missing, then the default bandwidth is the plug-in selector for the density gradient \( \text{Hpi}(x, \text{deriv.order}=2) \). Any bandwidth that is suitable for the density Hessian is also suitable for the kernel density ridge.

**Value**

A kernel density ridge set is an object of class \( \text{kdr} \) which is a list with fields:

- **x,y**: data points - same as input
- **end.points**: matrix of final iterates starting from \( y \)
- **H**: bandwidth matrix
- **names**: variable names
tol.iter, tol.clust, min.seg.size
   tuning parameter values - same as input
binned       flag for binned estimation
names        variable names
w            weights
path         list of density gradient ascent paths where path[[i]] is the path of y[i,] (only if keep.path=TRUE)

References


See Also

kde

Examples

library(maps)
data(quake)
quake <- quake[quake$prof==1,]  ## Pacific Ring of Fire
quake$long[quake$long<0] <- quake$long[quake$long<0] + 360
quake <- quake[, c("long", "lat")]
data(plate)                     ## tectonic plate boundaries
plate <- plate[plate$long < -20 | plate$long > 20,]
plate$long[plate$long<0 & !is.na(plate$long)] <- plate$long[plate$long<0 & !is.na(plate$long)] + 360

dr.quake <- kdr(x=quake, xmin=c(70, -70), xmax=c(310, 80))
map("world2", xlim=c(85, 305), ylim=c(-70, 70), mar=c(0,0,0,0), interior=FALSE, lty=2)
lines(plate[,1:2], col=3, lwd=2)
points(dr.quake$end.points, cex=0.5, pch=16, col=2)

kfe

**Kernel functional estimate**

Description

Kernel functional estimate for 1- to 6-dimensional data.
Usage

kfe(x, G, deriv.order, inc=1, binned, bin.par, bgridsize, deriv.vec=TRUE, add.index=TRUE, verbose=FALSE)

Hpi.kfe(x, nstage=2, pilot, pre="sphere", Hstart, binned=FALSE, bgridsize, amise=FALSE, deriv.order=0, verbose=FALSE, optim.fun="optim")

Hpi.diag.kfe(x, nstage=2, pilot, pre="scale", Hstart, binned=FALSE, bgridsize, amise=FALSE, deriv.order=0, verbose=FALSE, optim.fun="optim")

hpi.kfe(x, nstage=2, binned=FALSE, bgridsize, amise=FALSE, deriv.order=0)

Arguments

x vector/matrix of data values
nstage number of stages in the plug-in bandwidth selector (1 or 2)
pilot "dscalar" = single pilot bandwidth (default)
       "dunconstr" = single unconstrained pilot bandwidth
pre "scale" = \code{pre.scale}, "sphere" = \code{pre_sphere}
Hstart initial bandwidth matrix, used in numerical optimisation
binned flag for binned estimation
bgridsize vector of binning grid sizes
amise flag to return the minimal scaled PI value
deriv.order derivative order
verbose flag to print out progress information. Default is FALSE.
optim.fun optimiser function: one of \code{nlm} or \code{optim}
G pilot bandwidth matrix
inc 0=exclude diagonal, 1=include diagonal terms in kfe calculation
bin.par binning parameters - output from \code{binning}
deriv.vec flag to compute duplicated partial derivatives in the vectorised form. Default is FALSE.
add.index flag to output derivative indices matrix. Default is true.

Details

\code{Hpi.kfe} is the optimal plug-in bandwidth for \(r\)-th order kernel functional estimator based on the unconstrained pilot selectors of Chacon & Duong (2010). \code{hpi.kfe} is the 1-d equivalent, using the formulas from Wand & Jones (1995, p.70).

\code{kfe} does not usually need to be called explicitly by the user.

Value

Plug-in bandwidth matrix for \(r\)-th order kernel functional estimator.
References


See Also

kde.test

kfs

Kernel feature significance

Description

Kernel feature significance for 1- to 6-dimensional data.

Usage

kfs(x, H, h, deriv.order=2, gridsize, gridtype, xmin, xmax, supp=3.7, eval.points, binned, bgridsize, positive=FALSE, adj.positive, w, verbose=FALSE, signif.level=0.05)

Arguments

x matrix of data values
H,h bandwidth matrix/scalar bandwidth. If these are missing, Hpi or hpi is called by default.
deriv.order derivative order (scalar)
gridsize vector of number of grid points
gridtype not yet implemented
xmin,xmax vector of minimum/maximum values for grid
supp effective support for standard normal
eval.points vector or matrix of points at which estimate is evaluated
binned flag for binned estimation
bgridsize vector of binning grid sizes
positive flag if 1-d data are positive. Default is FALSE.
adj.positive adjustment applied to positive 1-d data
w vector of weights. Default is a vector of all ones.
verbose flag to print out progress information. Default is FALSE.
signif.level overall level of significance for hypothesis tests. Default is 0.05.
Details

Feature significance is based on significance testing of the gradient (first derivative) and curvature (second derivative) of a kernel density estimate. Only the latter is currently implemented, and is also known as significant modal regions.

The hypothesis test at a grid point \( x \) is \( H_0(x) : Hf(x) < 0 \), i.e. the density Hessian matrix \( Hf(x) \) is negative definite. The \( p \)-values are computed for each \( x \) using that the test statistic is approximately chi-squared distributed with \( d(d+1)/2 \) d.f. We then use a Hochberg-type simultaneous testing procedure, based on the ordered \( p \)-values, to control the overall level of significance to be \( \text{signif.level} \). If \( H_0(x) \) is rejected then \( x \) belongs to a significant modal region.

The computations are based on \( \text{kdde}(x, \text{deriv.order}=2) \) so \( \text{kfs} \) inherits its behaviour from \( \text{kdde} \). If the bandwidth \( h \) is missing from \( \text{kfs} \), then the default bandwidth is the plug-in selector \( \text{hpi}(,\text{deriv.order}=2) \). Likewise for missing \( h \). The effective support, binning, grid size, grid range, positive parameters are the same as \( \text{kde} \).

This function is similar to the \( \text{featureSignif} \) function in the \texttt{feature} package, except that it accepts unconstrained bandwidth matrices.

Value

A kernel feature significance estimate is an object of class \( \text{kfs} \) which is a list with fields

- \( x \) data points - same as input
- \( \text{eval.points} \) vector or list of points at which the estimate is evaluated
- \( \text{estimate} \) binary matrix for significant feature at \( \text{eval.points} \): 0 = not signif., 1 = signif.
- \( h \) scalar bandwidth (1-d only)
- \( H \) bandwidth matrix
- \( \text{gridtype} \) "linear"
- \( \text{gridded} \) flag for estimation on a grid
- \( \text{binned} \) flag for binned estimation
- \( \text{names} \) variable names
- \( w \) weights
- \( \text{deriv.order} \) derivative order (scalar)
- \( \text{deriv.ind} \) each row is a vector of partial derivative indices.

This is the same structure as a \( \text{kdde} \) object, except that \( \text{estimate} \) is a binary matrix rather than real-valued.

References


See Also

kdde, plot.kfs

Examples

## see example is ? plot.kfs

#### kms

**Kernel mean shift clustering**

**Description**

Kernel mean shift clustering for 2- to 6-dimensional data.

**Usage**

```r
kms(x, y, H, max.iter=400, tol.iter, tol.clust, min.clust.size, merge=TRUE,
    keep.path=FALSE, verbose=FALSE)
```

**Arguments**

- `x` matrix of data values
- `y` matrix of candidate data values for which the mean shift will estimate their cluster labels. If missing, `y=x`.
- `H` bandwidth matrix/scalar bandwidth. If missing, `Hpi(x, deriv, order=1)` is called by default.
- `max.iter` maximum number of iterations. Default is 400.
- `tol.iter` distance under which two successive iterations are considered convergent. Default is `0.001*min marginal IQR of x`.
- `tol.clust` distance under which two cluster modes are considered to form one cluster. Default is `0.01*max marginal IQR of x`.
- `min.clust.size` minimum cluster size (cardinality). Default is `0.01*nrow(y)`.
- `merge` flag to merge clusters which are smaller than `min.clust.size`. Default is TRUE.
- `keep.path` flag to store the density gradient ascent paths. Default is FALSE.
- `verbose` flag to print out progress information. Default is FALSE.
- `object` object of class `kms`
- `...` other parameters
Details

Mean shift clustering belongs to the class of modal or density-based clustering methods. The mean shift recurrence of the candidate point \( x \) is \( x_{j+1} = x_j + H \frac{\hat{f}(x_j)}{\hat{f}'(x_j)} \) where \( j \geq 0 \) and \( x_0 = x \). The sequence \( \{x_0, x_1, \ldots\} \) follows the density gradient ascent paths to converge to a local mode of the density estimate \( \hat{f} \). Hence \( x \) is iterated until it converges to its local mode, and this determines its cluster label.

The mean shift recurrence is terminated if successive iterations are less than \( \text{tol.iter} \) or the maximum number of iterations \( \text{max.iter} \) is reached. Final iterates which are less than \( \text{tol.cluster} \) distance apart are considered to form a single cluster. If \( \text{merge} = \text{TRUE} \) then the clusters whose cardinality is less than \( \text{min.cluster.size} \) are iteratively merged with their nearest cluster.

If the bandwidth \( H \) is missing, then the default bandwidth is the plug-in selector for the density gradient \( H_{\text{pi}}(x, \text{deriv.order}=1) \). Any bandwidth that is suitable for the density gradient is also suitable for the mean shift.

Value

A kernel mean shift clusters set is an object of class \( \text{kms} \) which is a list with fields:

- \( x, y \) data points - same as input
- \( \text{end.points} \) matrix of final iterates starting from \( y \)
- \( H \) bandwidth matrix
- \( \text{label} \) vector of cluster labels
- \( n\text{clust} \) number of clusters
- \( n\text{clust.table} \) frequency table of cluster labels
- \( \text{mode} \) matrix of cluster modes
- \( \text{names} \) variable names
- \( \text{tol.iter, tol.cluster, min.cluster.size} \) tuning parameter values - same as input
- \( \text{path} \) list of density gradient ascent paths where \( \text{path}[[]] \) is the path of \( y[1,] \) (only if \( \text{keep.path}=\text{TRUE} \))

References


See Also

- \( \text{kde} \)
Examples

```r
library(MASS)
data(crabs)
kms.crabs <- kms(x=crabs[,c("FL", "CW")])
plot(kms.crabs$x, col=kms.crabs$label)
summary(kms.crabs)
```

kroc

**Kernel receiver operating characteristic (ROC) curve**

Description

Kernel receiver operating characteristic (ROC) curve for 1- to 3-dimensional data.

Usage

```r
kroc(x1, x2, H1, h1, hy, gridsize, gridtype, xmin, xmax, supp=3.7, eval.points,
    binned, bgridsize, positive=FALSE, adj.positive, w, verbose=FALSE)
```

## S3 method for class 'kroc'
predict(object, ..., x)

## S3 method for class 'kroc'
summary(object, ...)

Arguments

- `x, x1, x2`: vector/matrix of data values
- `H1, h1, hy`: bandwidth matrix/scalar bandwidths. If these are missing, `hpi.kcde, hpi.kcde` is called by default.
- `gridsize`: vector of number of grid points
- `gridtype`: not yet implemented
- `xmin, xmax`: vector of minimum/maximum values for grid
- `supp`: effective support for standard normal
- `eval.points`: not yet implemented
- `binned`: flag for binned estimation
- `bgridsize`: vector of binning grid sizes
- `positive`: flag if 1-d data are positive. Default is FALSE.
- `adj.positive`: adjustment applied to positive 1-d data
- `w`: vector of weights. Default is a vector of all ones.
- `verbose`: flag to print out progress information. Default is FALSE.
- `object`: object of class kroc, output from kroc
- `...`: other parameters
Details

In this set-up, the values in the first sample \(x_1\) should be larger in general than those in the second sample \(x_2\). The usual method for computing 1-d ROC curves is not valid for multivariate data. Duong (2014), based on Lloyd (1998), develops an alternative formulation \((F_{Y_1}(z), F_{Y_2}(z))\) based on the cumulative distribution functions of \(Y_j = \bar{F}_1(X_j), j = 1, 2\).

If the bandwidth \(h_1\) is missing from \(kroc\), then the default bandwidth is the plug-in selector \(hpi.kcde\). Likewise for missing \(h_1, hy\). A bandwidth matrix \(H_1\) is required for \(x_1\) for \(d>1\), but the second bandwidth \(hy\) is always a scalar since \(Y_j\) are 1-d variables.

The effective support, binning, grid size, grid range, positive parameters are the same as \(kde\).

--The summary method for \(kroc\) objects prints out the summary indices of the ROC curve, as contained in the \(indices\) field, namely the AUC (area under the curve) and Youden index.

Value

A kernel ROC curve is an object of class \(kroc\) which is a list with fields:

- \(x\) : list of data values \(x_1, x_2\) - same as input
- \(eval.points\) : vector or list of points at which the estimate is evaluated
- \(estimate\) : ROC curve estimate at \(eval.points\)
- \(gridtype\) : "linear"
- \(gridded\) : flag for estimation on a grid
- \(binned\) : flag for binned estimation
- \(names\) : variable names
- \(w\) : weights
- \(tail\) : "lower.tail"
- \(h1\) : scalar bandwidth for first sample (1-d only)
- \(H1\) : bandwidth matrix for first sample
- \(hy\) : scalar bandwidth for ROC curve
- \(indices\) : summary indices of ROC curve.

References


See Also

- \(kcde\)
Examples

samp <- 1000
x <- rnorm.mixt(n=samp, mus=0, sigmas=1, props=1)
y <- rnorm.mixt(n=samp, mus=0.5, sigmas=1, props=1)
Rhat <- kroc(x1=x, x2=y)
summary(Rhat)
predict(Rhat, x=0.5)

ksupp

Kernel support estimate

Description

Kernel support estimate for 2-dimensional data.

Usage

ksupp(fhat, cont=95, abs.cont, convex.hull=FALSE)

Arguments

fhat object of class kde
cont percentage for contour level curve. Default is 95.
abs.cont absolute density estimate height for contour level curve
convex.hull flag to compute convex hull of contour level curve. Default is FALSE.

Details

The kernel support estimate is the level set of the density estimate that exceeds the cont percent contour level. If this level set is a simply connected region, then this can suffice to be a conservative estimate of the density support. Otherwise, the convex hull of the level set is advised.

Value

A kernel support estimate is a 2-column matrix which delimits the (convex hull of the) level set of the density estimate fhat.

See Also

kde
Examples

```r
library(oz)
data(grevillea)
fhat <- kde(x=grevillea)
fhat.supp <- ksupp(fhat, convex.hull=TRUE)
plot(fhat, display="filled.contour", cont=seq(10,90,by=10), drawlabels=FALSE)
plot(fhat, cont=95, add=TRUE)
polygon(fhat.supp, lty=2)
```

Description

Random generation and density values from normal and t-mixture distributions.

Usage

- `dnorm.mixt(x, mus=0, sigmas=1, props=1)`
- `rnorm.mixt(n=100, mus=0, sigmas=1, props=1, mixt.label=FALSE)`
- `dmvnorm.mixt(x, mus, Sigmas, props=1, verbose=FALSE)`
- `rmvnorm.mixt(n=100, mus=c(0,0), Sigmas=diag(2), props=1, mixt.label=FALSE)`
- `rmvt.mixt(n=100, mus=c(0,0), Sigmas=diag(2), dfs=7, props=1)`
- `dmvt.mixt(x, mus, Sigmas, dfs, props)`
- `mvnorm.mixt.mode(mus, Sigmas, props=1, verbose=FALSE)`

Arguments

- `n` number of random variates
- `x` matrix of quantiles
- `mus` (stacked) matrix of mean vectors (>1-d) or vector of means (1-d)
- `Sigmas` (stacked) matrix of variance matrices (>1-d)
- `sigmas` vector of standard deviations (1-d)
- `props` vector of mixing proportions
- `mixt.label` flag to output numeric label indicating mixture component. Default is FALSE.
- `verbose` flag to print out progress information. Default is FALSE.
- `dfs` vector of degrees of freedom

Details

`rmvnorm.mixt` and `dmvnorm.mixt` are based on the `rmvnorm` and `dmvnorm` functions from the `mvtnorm` package. Likewise for `rmvt.mixt` and `dmvt.mixt`.

For the normal mixture densities, `mvnorm.mixt.mode` computes the local modes: these are usually very close but not exactly equal to the component means.
plot.histde

Plot for histogram density estimate

Description

Plot for histogram density estimate for 1- and 2-dimensional data.

Usage

## S3 method for class 'histde'
plot(x, ...)

Arguments

x 
an object of class histde (output from histde)
...
other graphics parameters:
col plotting colour for density estimate
col.fun plotting colour function for levels
col.pt plotting colour for data points
jitter flag to jitter rug plot (1-d). Default is TRUE.
xlim,ylim axes limits
xlab,ylab axes labels
add flag to add to current plot. Default is FALSE.
drawpoints flag to draw data points on density estimate. Default is FALSE.
bbreaks vector of break values of density estimate. Default is an nbreaks equi-
linear sequence over the data range.
nbreaks number of breaks in breaks sequence
lty.rect,lwd.rect line type/width for histogram box lines (2-d)
plot.histde

border  colour of histogram box lines (2-d)
col.rect colour of histogram bars (1-d)
add.grid flag to add histogram grid (2-d). Default is TRUE.

Details

For histde objects, the function headers for the different dimensional data are

```r
## univariate
plot(fhat, xlab, ylab="Density function", add=FALSE, drawpoints=FALSE,
     col.pt="blue", jitter=FALSE, border=1, ...)

## bivariate
plot(fhat, breaks, nbreaks=11, xlab, ylab, zlab="Density function", cex=1,
     pch=1, add=FALSE, drawpoints=FALSE, col, col.fun, col.pt="blue",
     lty.rect=2, cex.text=1, border, lwd.rect=1, col.rect="transparent",
     add.grid=TRUE, ...)
```

The 1-d plot is a standard plot of a histogram generated by hist. If drawpoints=TRUE then a rug plot is added.

The 2-d plot is similar to the display="filled.contour" option from `plot.kde` with the default nbreaks=11 contour levels. Default colours are the default from the `image` command.

Value

Plots for 1-d and 2-d are sent to graphics window.

See Also

`plot.kde`

Examples

```r
library(MASS)
data(iris)

## univariate example
fhat <- histde(x=iris[,2])
plot(fhat, border=3, xlab="Sepal length")

## bivariate example
fhat <- histde(x=iris[,2:3])
plot(fhat, drawpoints=TRUE)
box()
```
plot.kcde

Plot for kernel cumulative distribution estimate

Description
Plot for kernel cumulative distribution estimate 1- to 3-dimensional data.

Usage
```r
## S3 method for class 'kcde'
plot(x, ...)  
```

Arguments
- `x` an object of class `kcde` (output from `kcde`)
- `...` other graphics parameters used in `plot.kde`

Details
For `kcde` objects, the function headers for the different dimensional data are

```r
## univariate
plot(Fhat, xlab, ylab="Distribution function", add=FALSE, drawpoints=FALSE,
     col.pt="blue", jitter=FALSE, ...)

## bivariate
plot(Fhat, display="persp", cont=seq(10,90, by=10), abs.cont, xlab, ylab,
     zlab="Distribution function", cex=1, pch=1, add=FALSE, drawpoints=FALSE,
     drawlabels=TRUE, theta=-30, phi=40, d=4, col.pt="blue", col, col.fun,
     lwd=1, border=NA, thin=1, ...)
```

Value
Plots for 1-d and 2-d are sent to graphics window. Plot for 3-d is sent to RGL window (not yet implemented).

See Also
- `plot.kde`

Examples
```r
library(MASS)
data(iris)
Fhat <- kcede(x=iris[,1])
plot(Fhat, xlab="Sepal.Length")
Fhat <- kcede(x=iris[,1:2])
plot(Fhat, thin=3)
```
plot.kda

Plot for kernel discriminant analysis

Description

Plot for kernel discriminant analysis for 1- to 3-dimensional data.

Usage

```r
## S3 method for class 'kda'
plot(x, y, y.group, ...)
```

Arguments

- `x` object of class kda (output from `kda`)
- `y` matrix of test data points
- `y.group` vector of group labels for test data points
- `...` other graphics parameters:
  - `rugsize` height of rug-like plot for partition classes (1-d)
  - `prior.prob` vector of prior probabilities
  - `col.part` vector of colours for partition classes (1-d, 2-d)
  - and those used in `plot.kde`

Details

For kda objects, the function headers for the different dimensional data are

```r
## univariate
plot(x, y, y.group, prior.prob=NULL, xlim, ylim, xlab="x",
     ylab="Weighted density function", drawpoints=FALSE, col, col.part,
     col.pt, lty, jitter=TRUE, rugsize, ...)

## bivariate
plot(x, y, y.group, prior.prob=NULL, cont=c(25,50,75), abs.cont,
     approx.cont=FALSE, xlim, ylim, xlab, ylab, drawpoints=FALSE,
     drawlabels=TRUE, col, col.part, col.pt, ...)

## trivariate
plot(x, y, y.group, prior.prob=NULL, cont=c(25,50,75), abs.cont,
     approx.cont=FALSE, colors, alphavec, xlab, ylab, zlab,
     drawpoints=FALSE, size=3, col.pt="blue", ...)
```
plot.kdde

Value

Plots for 1-d and 2-d are sent to graphics window. Plot for 3-d is sent to RGL window.

See Also

kda, kde

Examples

library(MASS)
data(iris)

### univariate example
ir <- iris[,1]
ir.gr <- iris[,5]
kda.fhat <- kda(x=ir, x.group=ir.gr, xmin=3, xmax=9)
plot(kda.fhat, xlab="Sepal length")

### bivariate example
ir <- iris[,1:2]
ir.gr <- iris[,5]
kda.fhat <- kda(x=ir, x.group=ir.gr)
plot(kda.fhat)

### trivariate example
ir <- iris[,1:3]
ir.gr <- iris[,5]
kda.fhat <- kda(x=ir, x.group=ir.gr)
plot(kda.fhat, drawpoints=TRUE, col.pt=c(2,3,4))
## colour=species, transparency=density heights

---

plot.kdde

Plot for kernel density derivative estimate

Description

Plot for kernel density derivative estimate for 1- to 3-dimensional data.

Usage

## S3 method for class 'kdde'
plot(x, ...)

Arguments

x       an object of class kdde (output from kde)
...
which.deriv.ind index of the partial derivative to be plotted (>1-d)
and those used in plot.kde
Details

For kdde objects, the function headers for the different dimensional data are

```r
# univariate
plot(fhat, ylab="Density derivative function", ...)
```

```r
# bivariate
plot(fhat, which.deriv.ind=1, cont=c(25,50,75), abs.cont, display="slice",
    zlab="Density derivative function", ...)
```

Value

Plots for 1-d and 2-d are sent to graphics window. Plot for 3-d is sent to RGL window.

In addition to the display options inherited from plot.kde, the first derivative has `display="quiver"`. This is a quiver plot where the size and direction of the arrow indicates the magnitude/direction of the density gradient. See quiver20 from the `OceanView` package for more details.

See Also

`plot.kde`

Examples

```r
## univariate example
data(tempb)
fhat1 <- kdde(x=tempb[,"tmin"], deriv.order=1)  # gradient [df/dx, df/dy]
plot(fhat1, xlab="Min. temp.")  # df/dx
points(20,predict(fhat1, x=20))

## bivariate example
fhat1 <- kdde(x=tempb[,,c("tmin", "tmax")], deriv.order=1)
plot(fhat1, display="quiver")  # gradient [df/dx, df/dy]

fhat2 <- kdde(x=tempb[,,c("tmin", "tmax")], deriv.order=2)
plot(fhat2, which.deriv.ind=2, display="persp", phi=15)
plot(fhat2, which.deriv.ind=2, display="filled.contour", col.fun=topo.colors)
  # d^2 f/(dx dy): purple=ve, green=zero, beige=-ve
s2 <- kcurv(fhat2)
plot(s2, display="filled.contour")
  # summary curvature
```
plot.kde

Plot for kernel density estimate

Description

Plot for kernel density estimate for 1- to 3-dimensional data.

Usage

```r
## S3 method for class 'kde'
plot(x, ...)  
```

Arguments

- `x`: an object of class kde (output from kde)
- `...`: other graphics parameters:
  - `display`: type of display, "slice" for contour plot, "persp" for perspective plot, "image" for image plot, "filled.contour" for filled contour plot (1st form), "filled.contour2" (2nd form) (2-d)
  - `cont`: vector of percentages for contour level curves
  - `abs.cont`: vector of absolute density estimate heights for contour level curves
  - `approx.cont`: flag to compute approximate contour levels. Default is FALSE.
  - `col`: plotting colour for density estimate (1-d, 2-d)
  - `col.cont`: plotting colour for contours
  - `col.fun`: plotting colour function for contours
  - `col.pt`: plotting colour for data points
  - `colors`: vector of colours for each contour (3-d)
  - `jitter`: flag to jitter rug plot (1-d). Default is TRUE.
  - `lwd.fc`: line width for filled contours (2-d)
  - `xlim,ylim,zlim`: axes limits
  - `xlab,ylab,zlab`: axes labels
  - `add`: flag to add to current plot. Default is FALSE.
  - `theta,phi,d,border`: graphics parameters for perspective plots (2-d)
  - `drawpoints`: flag to draw data points on density estimate. Default is FALSE.
  - `drawlabels`: flag to draw contour labels (2-d). Default is TRUE.
  - `alpha`: transparency value of plotting symbol (3-d)
  - `alphavec`: vector of transparency values for contours (3-d)
  - `size`: size of plotting symbol (3-d).
Details

For kde objects, the function headers for the different dimensional data are

```r
## univariate
plot(fhat, xlab, ylab="Density function", add=FALSE, drawpoints=FALSE,
    col.pt="blue", col.cont=1, cont.lwd=1, jitter=FALSE, cont, abs.cont,
    approx.cont=TRUE, ...)

## bivariate
plot(fhat, display="slice", cont=c(25,50,75), abs.cont, approx.cont=TRUE,
    xlab, ylab, zlab="Density function", cex=1, pch=1, add=FALSE,
    drawpoints=FALSE, drawlabels=TRUE, theta=-30, phi=40, d=4, col.pt="blue",
    col, col.fun, lwd=1, border=1, thin=3, lwd.fc=5, ...)

## trivariate
plot(fhat, cont=c(25,50,75), abs.cont, approx.cont=FALSE, colors,
    add=FALSE, drawpoints=FALSE, alpha, alphavec, xlab, ylab, zlab,
    size=3, col.pt="blue", ...)
```

The 1-d plot is a standard plot of a 1-d curve. If drawpoints=TRUE then a rug plot is added. If cont is specified, the horizontal line on the x-axis indicates the cont% highest density level set.

There are different types of plotting displays for 2-d data available, controlled by the display parameter. (a) If display="slice" then a slice/contour plot is generated using contour. (b) If display is "filled.contour" or "filled.contour2" then a filled contour plot is generated. The default contours are at 25%, 50%, 75% or cont=c(25,50,75) which are upper percentages of highest density regions. (c) If display= "persp" then a perspective/wire-frame plot is generated. The default z-axis limits zlim are the default from the usual persp command. (d) If display= "image" then an image plot is generated. Default colours are the default from the usual image command.

For 3-dimensional data, the interactive plot is a series of nested 3-d contours. The default contours are cont=c(25,50,75). The default colors are heat.colors and the default opacity alphavec ranges from 0.1 to 0.5.

To specify contours, either one of cont or abs.cont is required. cont specifies upper percentages which correspond to probability contour regions. If abs.cont is set to particular values, then contours at these levels are drawn. This second option is useful for plotting multiple density estimates with common contour levels. See contourLevels for details on computing contour levels.

If approx=FALSE, then the exact KDE is computed. Otherwise it is interpolated from an existing KDE grid. This can dramatically reduce computation time for large data sets.

Value

Plots for 1-d and 2-d are sent to graphics window. Plot for 3-d is sent to RGL window.

Examples

```r
library(MASS)
data(iris)
```
## Univariate Example

```r
fhat <- kde(x=iris[,2])
plot(fhat, cont=50, col.cont="blue", cont.lwd=2, xlab="Sepal length")
```

## Bivariate Example

```r
fhat <- kde(x=iris[,2:3])
plot(fhat, display="filled.contour", cont=seq(10,90,by=10))
plot(fhat, display="persp", thin=3, border=1, col="white")
```

## Trivariate Example

```r
fhat <- kde(x=iris[,2:4])
plot(fhat, drawpoints=TRUE)
```

---

**plot.kde.loctest**  
*Plot for kernel local significant difference regions*

### Description

Plot for kernel local significant difference regions for 1- to 3-dimensional data.

### Usage

```r
## S3 method for class 'kde.loctest'
plot(x, ...)
```

### Arguments

- `x`  
an object of class kde.loctest (output from `kde.local.test`)
- `...`  
other graphics parameters:
  - `lcol`  
    colour for KDE curve (1-d)
  - `col`  
    vector of 2 colours. Default is c("purple", "darkgreen"). First colour: sample 1>sample 2, second colour: sample 1<sample 2.
  - `add`  
    flag to add to current plot. Default is FALSE.
  - `rugsize`  
    height of rug-like plot (1-d)
  - `add.legend`  
    flag to add legend. Default is FALSE (1-d, 2-d).
  - `pos.legend`  
    position label for legend (1-d, 2-d)
  - `add.contour`  
    flag to add contour lines. Default is FALSE (2-d).
  - and those used in `plot.kde`

### Details

For kde.loctest objects, the function headers are

```r
## univariate
plot(x, lcol, col, add=FALSE, xlab="x", ylab, rugsize, add.legend=TRUE, pos.legend="topright", ...)
```
## bivariate
plot(x, col, add=FALSE, xlab="x", ylab="y", add.contour=FALSE,
     add.legend=TRUE, pos.legend="topright", ...)

## trivariate
plot(x, col, add=FALSE, xlab="x", ylab="y", zlab="z", box=TRUE, axes=TRUE,
     alphavec=c(0.5, 0.5), ...)

### Value

Plots for 1-d and 2-d are sent to graphics window. Plot for 3-d is sent to RGL window.

### See Also

kde.local.test

### Examples

```r
library(MASS)
data(crabs)
x1 <- crabs[crabs$sp=="B", c(4,6)]
x2 <- crabs[crabs$sp=="O", c(4,6)]
loct <- kde.local.test(x1=x1, x2=x2)
plot(loct)
```

### Description

Plot of partition for kernel density clustering for 2-dimensional data.

### Usage

```r
mvnorm.mixt.part(mus, Sigmas, props=1, xmin, xmax, gridsize, max.iter=100,
     verbose=FALSE)
kms.part(x, H, xmin, xmax, gridsize, verbose=FALSE, ...)
```

## S3 method for class 'kde.part'
plot(x, display="filled.contour", col, add=FALSE, ...)
Arguments

- mus: (stacked) matrix of mean vectors
- Sigmas: (stacked) matrix of variance matrices
- props: vector of mixing proportions
- xmin, xmax: vector of minimum/maximum values for grid
- gridsize: vector of number of grid points
- max.iter: maximum number of iterations
- verbose: flag to print out progress information. Default is FALSE.
- x: matrix of data values or an object of class kde.part
- H: bandwidth matrix. If missing, Hpi(x, deriv, order=1) is called by default.
- display: type of display, "slice" for contour plot, "persp" for perspective plot, "image" for image plot, "filled.contour" for filled contour plot (1st form), "filled.contour2" (2nd form)
- col: vector of plotting colours
- add: flag to add to current plot. Default is FALSE.
- ...: other parameters

Details

For 2-d data, kms.part and mvnorm.mixt.part produces a kde.part object whose values are the class labels, rather than probability density values.

Value

A kernel partition is an object of class kde.part which is a list with fields:

- x: data points - same as input
- eval.points: vector or list of points at which the estimate is evaluated
- estimate: density estimate at eval.points
- H: bandwidth matrix
- gridtype: "linear"
- gridded: flag for estimation on a grid
- binned: flag for binned estimation
- names: variable names
- w: weights
- cont: probability contour levels (if compute.cont=TRUE)
- end.points: matrix of final iterates starting from x
- label: vector of cluster labels
- mode: matrix of cluster modes
- nclust: number of clusters
- nclust.table: frequency table of cluster labels
- tol.iter, tol.clust, min.clust.size: tuning parameter values - same as input

Plot is sent to graphics window.
See Also

plot.kde, kms

Examples

## normal mixture partition
mus <- rbind(c(-1, 0), c(1, 2/sqrt(3)), c(-2/sqrt(3)))
Sigmas <- 1/25*rbind(invvech(c(9, 63/10, 49/4)), invvech(c(9, 0, 49/4)), invvech(c(9, 0, 49/4)))
props <- c(3, 3, 1)/7
nmixt.part <- mvnorm.mixt.part(mus=mus, Sigmas=Sigmas, props=props)
plot(nmixt.part, asp=1, xlim=c(-3, 3), ylim=c(-3, 3))

## Not run:
## kernel mean shift partition
set.seed(81928192)
x <- rmvnorm.mixt(n=1000, mus=mus, Sigmas=Sigmas, props=props)
msize <- round(151^2*0.05)
kms.nmixt.part <- kms.part(x=x, min.clust.size=msize)
plot(kms.nmixt.part, asp=1, xlim=c(-3, 3), ylim=c(-3, 3))
## End(Not run)

plot.kfs

Plot for kernel feature significance

Description

Plot for kernel significant regions for 1- to 3-dimensional data.

Usage

## S3 method for class 'kfs'
plot(x, display="filled.contour", col="orange", colors="orange", abs.cont, alphavec=0.4, add=FALSE, ...)

Arguments

x
an object of class kfs (output from kfs)
display
type of display, "slice" for contour plot, "persp" for perspective plot, "image" for image plot, "filled.contour" for filled contour plot (1st form), "filled.contour2" (2nd form) (2-d)
col, colors
colour for contour region (1-d, 2-d), (3-d)
abs.cont
absolute contour height. Default is 0.5.
alphavec
transparency value for contour (3-d)
add
flag to add to current plot. Default is FALSE.
...
other graphics parameters used in plot.kde
Value

Plots for 1-d and 2-d are sent to graphics window. Plot for 3-d is sent to RGL window.

See Also

plot.kde

Examples

library(MASS)
data(geyser)
geyser.fs <- kfs(geyser, binned=TRUE)
plot(geyser.fs)

plot.kroc  |
---
Plot for kernel receiver operating characteristic curve (ROC) estimate

Description

Plot for kernel receiver operating characteristic curve (ROC) estimate 1- to 3-dimensional data.

Usage

## S3 method for class 'kroc'
plot(x, add=FALSE, add.roc.ref=FALSE, ylab, xlab, ...)

Arguments

x          an object of class kroc (output from kroc)
add         flag to add to current plot. Default is FALSE.
add.roc.ref flag to add reference ROC curve. Default is FALSE.
xlab        x-axis label. Default is "False positive rate (bar(specificity))".
ylab        y-axis label. Default is "True positive rate (sensitivity)".
...         other graphics parameters used in plot.kde.

Value

Plots for 1-d and 2-d are sent to graphics window. Plot for 3-d is sent to RGL window.

See Also

plot.kde
Examples

```
library(MASS)
data(fgl)
x1 <- fgl[,c("RI","Na")]
x2 <- fgl[,c("Head","Na")]
Rhat <- kroc(x1=x1, x2=x2)
plot(Rhat, add.roc.ref=TRUE)
```

Description

Plot for 1- to 3-dimensional normal and t-mixture density functions.

Usage

```
plotmixt(mus, sigmas, Sigmas, props, dfs, dist="normal", draw=TRUE,
deriv.order=0, which.deriv.ind=1, binned=TRUE, ...)
```

Arguments

- **mus**: (stacked) matrix of mean vectors
- **sigmas**: vector of standard deviations (1-d)
- **Sigmas**: (stacked) matrix of variance matrices (2-d, 3-d)
- **props**: vector of mixing proportions
- **dfs**: vector of degrees of freedom
- **dist**: "normal" - normal mixture, "t" - t-mixture
- **draw**: flag to draw plot. Default is TRUE.
- **deriv.order**: derivative order
- **which.deriv.ind**: index of which partial derivative to plot
- **binned**: flag for binned estimation of contour levels. Default is TRUE.
- **...**: other graphics parameters, see `plot.kde`

Value

If draw=TRUE, the 1-d, 2-d plot is sent to graphics window, 3-d plot to RGL window. If draw=FALSE, then a kde-like object is returned.
Examples

```r
## bivariate
mus <- rbind(c(0, 0), c(-1, 1))
Sigma <- matrix(c(1, 0.7, 0.7, 1), nr=2, nc=2)
Sigmas <- rbind(Sigma, Sigma)
props <- c(1/2, 1/2)
plotmixt(mus=mus, Sigmas=Sigmas, props=props, display="filled.contour")

## trivariate
mus <- rbind(c(0, 0, 0), c(-1, 0.5, 1.5))
Sigma <- matrix(c(1, 0.7, 0.7, 0.7, 1, 0.7, 0.7, 0.7, 1), nr=3, nc=3)
Sigmas <- rbind(Sigma, Sigma)
props <- c(1/2, 1/2)
plotmixt(mus=mus, Sigmas=Sigmas, props=props, dfs=c(11, 8), dist="t")
```
Geographical locations of earthquakes

Description

The quake data set contains the geographical locations of severe earthquakes in the years 100 and 2016 inclusive. The plate data set contains the geographical locations of the tectonic plate boundaries.

Usage

data(quake)
data(plate)

Format

– For quake, a matrix with 5871 rows and 5 columns. Each row corresponds to an earthquake. The first column is the year (negative years indicate B.C.E.), the second is the longitude (decimal degrees), the third is the latitude (decimal degrees), the fourth is the depth beneath the Earth’s surface (km), the fifth is a flag for the location inside the circum-Pacific belt (aka Pacific Ring of Fire).

– For plate, a matrix with 3 columns and 6276 rows. Each row corresponds to an location of the tectonic plate boundaries. The first is the longitude, the second is the latitude, the third is the label of the tectonic plate.

Source


Derived quantities from kernel density estimates

Description

Derived quantities from kernel density estimates.
Usage

\[
dkde(x, fhat) \\
pkde(q, fhat) \\
qkde(p, fhat) \\
rkde(n, fhat, positive=FALSE)
\]

Arguments

- `x, q`: vector of quantiles
- `p`: vector of probabilities
- `n`: number of observations
- `positive`: flag to compute KDE on the positive real line. Default is FALSE.
- `fhat`: kernel density estimate, object of class `kde`.

Details

- `pkde` uses Simpson’s rule for the numerical integration.
- `rkde` uses Silverman (1986)’s method to generate a random sample from a KDE.

Value

For the 1-d kernel density estimate `fhat`, `pkde` computes the cumulative probability for the quantile `q`, `qkde` computes the quantile corresponding to the probability `p`, `dkde` computes the density value at `x`.

For any kernel density estimate, `rkde` computes a random sample of size `n`.

References

  London.

Examples

```r
set.seed(8192)
x <- rnorm.mixt(n=10000, mus=0, sigmas=1, props=1)
fhat <- kde(x=x, binned=TRUE)
p1 <- pkde(fhat=fhat, q=c(-1, 0, 0.5))
qkde(fhat=fhat, p=p1)
y <- rkde(fhat=fhat, n=100)

x <- rmvnorm.mixt(n=10000, mus=c(0,0), Sigmas=invvech(c(1,0.8,1)))
fhat <- kde(x=x, binned=TRUE)
y <- rkde(fhat=fhat, n=1000)

fhaty <- kde(x=y, binned=TRUE)
plot(fhat)
plot(fhaty, add=TRUE, col=2)
```
**Description**

This data set contains the daily minimum and maximum temperatures from the weather station in Badajoz, Spain, from 1 January 1955 to 31 December 2015.

**Usage**

```r
data(hsct)
```

**Format**

A matrix with 21908 rows and 5 columns. Each row corresponds to a daily measurement. The first column is the year (yyyy), the second is the month (mm), the third is the day (dd), the fourth is the minimum temperature (degrees Celsius), the fifth is the maximum temperature (degrees Celsius).

**Source**


---

**Description**

This data set contains the number of deaths of children under 5 years of age per 1000 live births and the average life expectancy (in years) at birth for 73 countries with GNI (Gross National Income) less than 1000 US dollars per annum per capita.

**Usage**

```r
data(unicef)
```

**Format**

A matrix with 2 columns and 73 rows. Each row corresponds to a country. The first column is the under 5 mortality rate and the second is the average life expectancy.

**Source**

Vector and vector half operators

Description

The vec (vector) operator takes a $d \times d$ matrix and stacks the columns into a single vector of length $d^2$. The vech (vector half) operator takes a symmetric $d \times d$ matrix and stacks the lower triangular half into a single vector of length $d(d + 1)/2$. The functions invvec and invvech are the inverses of vec and vech i.e. they form matrices from vectors.

Usage

vec(x, byrow=FALSE)
vec(x)
invvec(x, ncol, nrow, byrow=FALSE)
invvech(x)

Arguments

x vector or matrix
ncol, nrow number of columns and rows for inverse of vech
byrow flag for stacking row-wise or column-wise. Default is FALSE.

References


Examples

x <- matrix(1:9, nrow=3, ncol=3)
vec(x)
invvec(vec(x))

Variable kernel density estimate.

Description

Variable kernel density estimate for 2-dimensional data.

Usage

kde.balloon(x, H, h, gridsize, gridtype, xmin, xmax, supp=3.7, eval.points,
binned, bgridsize, w, compute.cont=TRUE, approx.cont=TRUE, verbose=FALSE)
kde.sp(x, H, h, gridsize, gridtype, xmin, xmax, supp=3.7, eval.points,
binned, bgridsize, w, compute.cont=TRUE, approx.cont=TRUE, verbose=FALSE)
Arguments

- **x**: matrix of data values
- **H**: bandwidth matrix. If this missing, hns is called by default.
- **h**: not yet implemented
- **gridsize**: vector of number of grid points
- **gridtype**: not yet implemented
- **xmin, xmax**: vector of minimum/maximum values for grid
- **supp**: effective support for standard normal
- **eval.points**: vector or matrix of points at which estimate is evaluated
- **binned**: flag for binned estimation.
- **bgridsize**: vector of binning grid sizes
- **w**: vector of weights. Default is a vector of all ones.
- **compute.cont**: flag for computing 1% to 99% probability contour levels. Default is TRUE.
- **approx.cont**: flag for computing approximate probability contour levels. Default is TRUE.
- **verbose**: flag to print out progress information. Default is FALSE.

Details

The balloon density estimate `kde.balloon` employs bandwidths which vary at each estimation point (Loftsgaarden & Quesenberry, 1965). There are as many bandwidths as there are estimation grid points. The default bandwidth is $h_{ns}(.,\text{deriv.order}=2)$ and the subsequent bandwidths are derived via a minimal MSE formula.

The sample point density estimate `kde.sp` employs bandwidths which vary for each data point (Abramson, 1982). There are as many bandwidths as there are data points. The default bandwidth is $h_{ns}(.,\text{deriv.order}=4)$ and the subsequent bandwidths are derived via the Abramson formula.

Value

A variable kernel density estimate for bounded data is an object of class `kde`.

References


See Also

- `kde`, `plot.kde`
Examples

data(worldbank)
wb <- as.matrix(na.omit(worldbank[,4:5]))
xmin <- c(-75,-25); xmax <- c(25,70)
fhat <- kde(x=wb, xmin=xmin, xmax=xmax)
fhat.sp <- kde.sp(x=wb, xmin=xmin, xmax=xmax)
plot(fhat, display="persp", box=TRUE, phi=20, zlim=c(0,max(fhat.sp$estimate)))
plot(fhat.sp, display="persp", box=TRUE, phi=20, zlim=c(0,max(fhat.sp$estimate)))
## Not run:

fhat.ball <- kde.balloon(x=wb, xmin=xmin, xmax=xmax)
plot(fhat.ball, display="persp", box=TRUE, phi=20, zlim=c(0,max(fhat.sp$estimate)))
## End(Not run)

worldbank

Development indicators from the World Bank Group

Description

This data set contains six development indicators for national entities for the year 2011, which is the latest year for which they are consistently available.

Usage

data(worldbank)

Format

A matrix with 7 columns and 218 rows. Each row corresponds to a country. The first column is the country, the second is the per capita carbon dioxide emissions (thousands Kg), the third is the per capita GDP (thousands of current USD), the fourth is the annual GDP growth rate (%), the fifth is the annual inflation rate (%), the sixth is the percentage of internet users in the population (%), the seventh is the added value agricultural production as a ratio of the total GDP (%).

Source

Index

*Topic algebra
  binning, 7
  pre.transform, 66
  vector, 70
*Topic cluster
  kdr, 40
  kms, 46
*Topic datasets
  air, 7
  cardio, 8
  grevillea, 10
  hsct, 18
  quake, 67
  tempb, 69
  unicef, 69
  worldbank, 72
*Topic distribution
  mixt, 51
*Topic hplot
  contour, 9
  plot.histde, 52
  plot.kcde, 54
  plot.kda, 55
  plot.kdde, 56
  plot.kde, 58
  plot.kde.loctest, 60
  plot.kde.part, 61
  plot.kfs, 63
  plot.kroc, 64
  plotmixt, 65
*Topic package
  ks-package, 3
*Topic smooth
  Hbcv, 11
  histde, 12
  Hscv, 13
  Hnm, 15
  Hns, 16
  Hpi, 17
  Hscv, 19
  ise.mixt, 20
  kcde, 22
  kcopula, 24
  kda, 26
  kdcde, 28
  kdde, 30
  kde, 32
  kde.boundary, 34
  kde.truncate, 39
  kfe, 42
  kfs, 44
  kroc, 48
  ksupp, 50
  rkde, 67
  vkde, 70
*Topic test
  kde.local.test, 36
  kde.test, 38
  air, 7
  amise.mixt (ise.mixt), 20
  binning, 7, 43
  cardio, 8
  compare (kda), 26
  contour, 9, 10
  contourLevels, 59
  contourLevels (contour), 9
  contourLines, 10
  contourSizes (contour), 9
  dckde (kdcde), 28
  dkde (rkde), 67
  dmnorm.mixt (mixt), 51
  dmvt.mixt (mixt), 51
  dnorm.mixt (mixt), 51
  grevillea, 10
INDEX

Hamise.mixt, 15
Hamise.mixt (ise.mixt), 20
hamise.mixt (ise.mixt), 20
Hamise.mixt.diag (ise.mixt), 20
Hbcv, 3, 11, 14, 18, 20
Hbcv.diag, 3
histde, 12, 52
Hkda, 4
Hkda (kda), 26
hkda, 4
hkda (kda), 26
Hkda.diag (kda), 26
Hlscv, 3, 4, 12, 13, 18, 20, 27
hlscv, 3
hlscv (Hlscv), 13
Hlscv.diag, 3
Hlscv.diag (Hlscv), 13
Hmise.mixt, 15
Hmise.mixt (ise.mixt), 20
hmise.mixt (ise.mixt), 20
Hmise.mixt.diag (ise.mixt), 20
Hnm, 15
Hnm (Hnm), 15
Hnm.diag (Hnm), 15
Hns, 3, 4, 16
hns, 3
hns (Hns), 16
Hns.diag (Hns), 16
Hns.kcde (Hns), 16
hns.kcde (Hns), 16
Hpi, 3–5, 11, 12, 14, 17, 20, 27, 66
hpi, 3, 5
hpi (Hpi), 17
Hpi.diag, 3
Hpi.diag (Hpi), 17
Hpi.diag.kcde (kcde), 22
Hpi.diag.kfde (kfe), 42
Hpi.kcde, 5
Hpi.kcde (kcde), 22
hpi.kcde, 5
hpi.kcde (kcde), 22
Hpi.kfe, 4, 5, 38
hpi.kfe (kfe), 42
hpi.kfe, 4, 5
hpi.kfe (kfe), 42
hsct, 18
Hscv, 3, 4, 12, 14, 18, 19, 27, 66
hscv, 3
hscv (Hscv), 19
Hscv.diag, 3
Hscv.diag (Hscv), 19
Hucv (Hlscv), 13
hucv (Hlscv), 13
Hucv.diag (Hlscv), 13
invvec (vector), 70
invvech (vector), 70
ise.mixt, 20
kcde, 5, 16, 22, 25, 49, 54
kcopula, 5, 24
kcurv, 4
kcurv (kdde), 30
kda, 4, 26, 55, 56
kcde, 6, 28
kdde, 4–6, 16, 30, 40, 45, 46, 56
kd.de.truncate (kde.truncate), 39
kde, 3–6, 16, 23, 25, 27, 29–32, 32, 35, 40–42,
45, 47, 49, 50, 56, 58, 71
kde.balloons, 4
kde.balloon (vkde), 70
kde.boundary, 4, 25, 34
kde.local.test, 5, 36, 39, 60, 61
kde.sp, 4
kde.sp (vkde), 70
kde.test, 5, 37, 38, 44
kde.truncate, 3, 39
kdr, 5, 40
kfe, 42
kfs, 6, 44, 63
kms, 5, 46, 63
kms.part (plot.kde.part), 61
kroc, 5, 48, 64
ks (ks-package), 3
ks-package, 3
ksupp, 3, 50
mise.mixt (ise.mixt), 20
mixt, 51
mvnorm.mixt.mode (mixt), 51
mvnorm.mixt.part (plot.kde.part), 61
pkde (rkde), 67
plate (quake), 67
plot.histde, 13, 52
plot.kcde, 5, 23, 54
plot.kda, 4, 28, 55
INDEX

plot.kdde, 4, 56
plot.kde, 3, 4, 34, 53–57, 58, 60, 63–65, 71
plot.kde.loctest, 37, 60
plot.kde.part, 61
plot.kfs, 6, 46, 63
plot.kroc, 5, 64
plotmxt, 65
pre.scale, 17, 19, 43
pre.scale (pre.transform), 66
pre.sphere, 17, 19, 43
pre.sphere (pre.transform), 66
pre.transform, 66
predict.histde (histde), 12
predict.kcde (kde), 22
predict.kda (kda), 26
predict.kdde (kdde), 30
predict.kde (kde), 32
predict.kroc (kroc), 48
qkde (rkde), 67
quake, 67
rkde, 67
rmvnorm.mxt (mixt), 51
rmv.t.mxt (mixt), 51
rnorm.mxt (mixt), 51
summary.kms (kms), 46
summary.kroc (kroc), 48
tempb, 69
unicef, 69
vec (vector), 70
vech (vector), 70
vector, 70
vkde, 70
worldbank, 72