

Package ‘feature’

February 14, 2012

Version 1.2.8

Date 2011/08/09

Title Feature significance for multivariate kernel density estimation

Author Tarn Duong <tarn.duong@gmail.com> & Matt Wand <>wand@uow.edu.au>

Maintainer Tarn Duong <tarn.duong@gmail.com>

Depends R (>= 1.4.0), KernSmooth, ks (>= 1.8.0), tcltk

Suggests MASS

Description Feature significance for multivariate kernel density estimation

License GPL (>= 2)

URL <http://www.mvstat.net/tduong>

Repository CRAN

Date/Publication 2011-09-10 06:17:24

R topics documented:

earthquake	2
feature	2
featureSignif	3
featureSignifGUI	5
plot.fs	6
SiZer,siCon	8
Index	10

 earthquake

Mt St Helens earthquake data

Description

This data set is a reduced version of the full data set in Scott (1992). It contains the first three variables.

Usage

```
data(earthquake)
```

Format

A matrix with 3 columns and 510 rows. Each row corresponds to the measurements of an earthquake beneath the Mt St Helens volcano. The first column is the longitude (in degrees, where a negative number indicates west of the International Date Line), the second column is the latitude (in degrees, where a positive number indicates north of the Equator) and the third column is the depth (in km, where a negative number indicates below the Earth's surface).

Source

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley & Sons Inc., New York.

 feature

feature

Description

Package for feature significance for multivariate kernel density estimation.

Details

The **feature** package contains functions to display and compute kernel density estimates, significant gradient and significant curvature regions. Significant gradient and/or curvature regions often correspond to significant features (e.g. local modes).

There are two main functions in this package. `featureSignifGUI` is the interactive function where the user can select bandwidths from a pre-defined range. This mode is useful for initial exploratory data analysis. `featureSignif` is the non-interactive function. This is useful when the user has a more definite idea of suitable values for the bandwidths. For a more detailed example for 1-d and 2-d data, see `vignette("feature")`.

Author(s)

Tarn Duong <tarn.duong@gmail.com> & Matt Wand <>wand@uow.edu.au>

See Also

ks, sm, KernSmooth

featureSignif	<i>Feature significance for kernel density estimation</i>
---------------	---

Description

Identify significant features of kernel density estimates of 1- to 4-dimensional data.

Usage

```
featureSignif(x, bw, gridsize, scaleData=FALSE, addSignifGrad=TRUE,
              addSignifCurv=TRUE, signifLevel=0.05)
```

Arguments

x	data matrix
bw	vector of bandwidth(s)
gridsize	vector of estimation grid sizes
scaleData	flag for scaling the data i.e. transforming to unit variance for each dimension.
addSignifGrad	flag for computing significant gradient regions
addSignifCurv	flag for computing significant curvature regions
signifLevel	significance level

Details

Feature significance is based on significance testing of the gradient (first derivative) and curvature (second derivative) of a kernel density estimate. This was developed for 1-d data by Chaudhuri & Marron (1995), for 2-d data by Godtliebsen, Marron & Chaudhuri (1999), and for 3-d and 4-d data by Duong, Cowling, Koch & Wand (2007).

The test statistic for gradient testing is at a point \mathbf{x} is

$$W(\mathbf{x}) = \|\widehat{\nabla}f(\mathbf{x}; \mathbf{H})\|^2$$

where $\widehat{\nabla}f(\mathbf{x}; \mathbf{H})$ is kernel estimate of the gradient of $f(\mathbf{x})$ with bandwidth \mathbf{H} , and $\|\cdot\|$ is the Euclidean norm. $W(\mathbf{x})$ is approximately chi-squared distributed with d degrees of freedom where d is the dimension of the data.

The analogous test statistic for curvature is

$$W^{(2)}(\mathbf{x}) = \|\text{vech}\widehat{\nabla}^{(2)}f(\mathbf{x}; \mathbf{H})\|^2$$

where $\widehat{\nabla}^{(2)}f(\mathbf{x}; \mathbf{H})$ is the kernel estimate of the curvature of $f(\mathbf{x})$, and vech is the vector-half operator. $W^{(2)}(\mathbf{x})$ is approximately chi-squared distributed with $d(d+1)/2$ degrees of freedom.

Since this is a situation with many dependent hypothesis tests, we use a multiple comparison or simultaneous test to control the overall level of significance. We use a Hochberg-type procedure. See Hochberg (1988) and Duong, Cowling, Koch & Wand (2007).

Value

Returns an object of class `fs` which is a list with the following fields

`x` data matrix

`names` name labels used for plotting

`bw` vector of bandwidths

`fhat` kernel density estimate on a grid

`grad` logical grid for significant gradient

`curv` logical grid for significant curvature

`gradData` logical vector for significant gradient data points

`gradDataPoints` significant gradient data points

`curvData` logical vector for significant curvature data points

`curvDataPoints` significant curvature data points

References

Chaudhuri, P. & Marron, J.S. (1999) SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807-823.

Duong, T., Cowling, A., Koch, I. & Wand, M.P. (2008) Feature significance for multivariate kernel density estimation. *Computational Statistics and Data Analysis*, **52**, 4225-4242.

Hochberg, Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800-802.

Godtliebsen, F., Marron, J.S. & Chaudhuri, P. (2002) Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics*, **11**, 1-22.

Wand, M.P. & Jones, M.C. (1995) *Kernel Smoothing*. Chapman & Hall/CRC, London.

See Also

[featureSignifGUI](#), [plot.fs](#)

Examples

```
## Univariate example
data(earthquake)
eq3 <- -log10(-earthquake[,3])
fs <- featureSignif(eq3, bw=0.1)
plot(fs, addSignifGradRegion=TRUE)
```

```
## Bivariate example
library(MASS)
data(geyser)
fs <- featureSignif(geyser)
plot(fs, addSignifCurvRegion=TRUE)
```

```
## Trivariate example
data(earthquake)
```

```

earthquake[,3] <- -log10(-earthquake[,3])
fs <- featureSignif(earthquake, scaleData=TRUE, bw=c(0.06, 0.06, 0.05))
plot(fs, addKDE=TRUE)
plot(fs, addKDE=FALSE, addSignifCurvRegion=TRUE)

```

featureSignifGUI

GUI for feature significance for kernel density estimation

Description

GUI for feature significance for kernel density estimation.

Usage

```
featureSignifGUI(x, scaleData=FALSE)
```

Arguments

x	data matrix
scaleData	flag for scaling the data to the unit interval in each dimension

Details

In the first column are the sliders for selecting the bandwidths (one for each dimension). Move the slider buttons to change the value of the bandwidths. The text field is for the grid size which specifies the number of points in each dimension of the kernel estimation binning grid. Press the 'Compute significant features' button to begin the computation. This creates a plot of the kernel density estimate (KDE) from the data with the specified bandwidths by calling [featureSignif](#). Once this complete, a pop-up window will appear.

In the second column are the axis limits and labels. The last text field is for the (maximum) number of data points used in the display. Press the 'Reset plot (except KDE)' button to clear the plot of all added features except for the KDE itself.

In the third column are 5 buttons which can be used to add to the KDE plot such as the data points, significant gradient points/regions and significant curvature points/regions. For 1-d data, the button in the third column is 'Compute SiZer map'. Press this button to compute a gradient SiZer plot using the [SiZer](#) function. Once this complete, a pop-up window will appear. For 2- and 3-d data, the button in the third column is 'Reset plot'. This will clear the plot of all features as well as the KDE. This is useful for showing only the significant features when the KDE may interfere with their display.

For 3-d data, there is an extra fourth column of options: these are sliders for the transparency values for the features. Move the slider button along to the desired value (between 0 and 1) and then press the 'Add ...' button to the left. Repeatedly pressing the 'Add ...' button will cause the transparency of the features to decrease. In this case, press the one of the 'Reset plot' buttons to clear the plot window, and replot the significant feature with the desired transparency.

Examples

```
## Not run:
library(MASS)
data(geyser)
duration <- geyser$duration
featureSignifGUI(duration) ## univariate example
featureSignifGUI(geyser)  ## bivariate example

data(earthquake)          ## trivariate example
earthquake$depth <- -log10(-earthquake$depth)
featureSignifGUI(earthquake, scaleData=TRUE)
## End(Not run)
```

plot.fs

*Feature significance plot for 1- to 3-dimensional data***Description**

Feature significance plot for 1- to 3-dimensional data.

Usage

```
## S3 method for class 'fs'
plot(x, ..., xlab, ylab, zlab, xlim, ylim, zlim,
      add=FALSE, addData=FALSE, scaleData=FALSE, addDataNum=1000,
      addKDE=TRUE, jitterRug=TRUE,
      addSignifGradRegion=FALSE, addSignifGradData=FALSE,
      addSignifCurvRegion=FALSE, addSignifCurvData=FALSE,
      addAxes3d=TRUE, densCol, dataCol="black", gradCol="green4",
      curvCol="blue", axisCol="black", bgCol="white",
      dataAlpha=0.1, gradDataAlpha=0.3, gradRegionAlpha=0.2,
      curvDataAlpha=0.3, curvRegionAlpha=0.3)
```

Arguments

x	an object of class fs (output from featureSignif function)
xlim, ylim, zlim	x-, y-, z-axis limits
xlab, ylab, zlab	x-, y-, z-axis labels
scaleData	flag for scaling the data i.e. transforming to unit variance for each dimension
add	flag for adding to an existing plot
addData	flag for display of the data
addDataNum	maximum number of data points plotted in displays
addKDE	flag for display of kernel density estimates

jitterRug	flag for jittering of rug-plot for univariate data display
addSignifGradRegion, addSignifGradData	flag for display of significant gradient regions/data points
addSignifCurvRegion, addSignifCurvData	flag for display of significant curvature regions/data points
addAxes3d	flag for displaying axes in 3-d displays
densCol	colour of density estimate curve
dataCol	colour of data points
gradCol	colour of significant gradient regions/data points
curvCol	colour of significant curvature regions/data points
axisCol	colour of axes
bgCol	colour of background
dataAlpha	transparency of data points
gradRegionAlpha, gradDataAlpha	transparency of significant gradient regions/data points
curvRegionAlpha, curvDataAlpha	transparency of significant curvature regions/data points
...	other graphics parameters

Value

Plot of 1-d and 2-d kernel density estimates are sent to graphics window. Plot for 3-d is sent to RGL window.

See Also

[featureSignif](#)

Examples

```
library(MASS)
data(geyser)
fs <- featureSignif(geyser, bw=c(4.5, 0.37))
plot(fs, addKDE=FALSE, addData=TRUE) ## data only
plot(fs, addKDE=TRUE) ## KDE plot only
plot(fs, addSignifGradRegion=TRUE)
plot(fs, addKDE=FALSE, addSignifCurvRegion=TRUE)
plot(fs, addSignifCurvData=TRUE, curvCol="cyan")
```

 SiZer, siCon

SiZer and SiCon plots for 1-dimensional data

Description

SiZer (**S**ignificant **Z**ero crossings) and (**S**ignificant **C**onvexity) plots for 1-dimensional data.

Usage

```
SiZer(x, bw, gridsize, scaleData=FALSE, signifLevel=0.05,
      plotSiZer=TRUE, logbw=TRUE, xlim, xlab,
      addLegend=TRUE, posLegend="bottomright")
```

```
SiCon(x, bw, gridsize, scaleData=FALSE, signifLevel=0.05,
      plotSiCon=TRUE, logbw=TRUE, xlim, xlab,
      addLegend=TRUE, posLegend="bottomright")
```

Arguments

x	data vector
bw	vector of range of bandwidths
gridsize	number of x- and y-axis grid points
scaleData	flag for scaling the data i.e. transforming to unit variance for each dimension.
signifLevel	significance level
plotSiZer, plotSiCon	flag for displaying SiZer/SiCon map
logbw	flag for displaying log bandwidths on y-axis
xlim	x-axis limits
xlab	x-axis label
addLegend	flag for legend display
posLegend	legend position

Details

The gradient SiZer and curvature SiCon maps of Chaudhuri & Marron (1999) are implemented. The horizontal axis is the data axis, the vertical axis are the bandwidths. The colour scheme for the SiZer map is red: negative gradient, blue: positive gradient, purple: zero gradient and grey: sparse regions. For the SiCon map, orange: negative curvature (concave), blue: positive curvature (convex), green: zero curvature and grey: sparse regions.

Value

SiZer plot sent to graphics window.

References

Chaudhuri, P. & Marron, J.S. (1999) SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807-823.

See Also

[featureSignif](#)

Examples

```
data(earthquake)
eq3 <- -log10(-earthquake[,3])
SiZer(eq3)
SiCon(eq3)
```

Index

- *Topic **datasets**
 - earthquake, [2](#)
- *Topic **hplot**
 - plot.fs, [6](#)
 - SiZer, siCon, [8](#)
- *Topic **package**
 - feature, [2](#)
- *Topic **smooth**
 - featureSignif, [3](#)
 - featureSignifGUI, [5](#)

earthquake, [2](#)

feature, [2](#)

featureSignif, [2](#), [3](#), [5–7](#), [9](#)

featureSignifGUI, [2](#), [4](#), [5](#)

plot.fs, [4](#), [6](#)

SiCon (SiZer, siCon), [8](#)

SiZer, [5](#)

SiZer (SiZer, siCon), [8](#)

SiZer, siCon, [8](#)