feature: an R package for feature significance for multivariate kernel density estimation

Tarn Duong

20 September 2010

1 Introduction

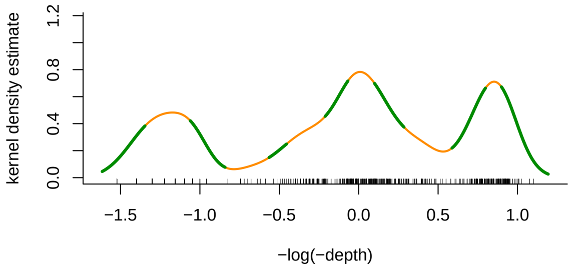
Feature significance is an extension of kernel density estimation which is used to establish the statistical significance of features (e.g. local modes). See Chaudhuri and Marron (1999) for 1-dimensional data, Godtliebsen et al. (2002) for 2-dimensional data and Duong et al. (2007) for 3- and 4-dimensional data. The `feature` package contains a range of options to display and compute kernel density estimates, significant gradient and significant curvature regions. Significant gradient and/or curvature regions often correspond to significant features. In this vignette we focus on 1- and 2-dimensional data.

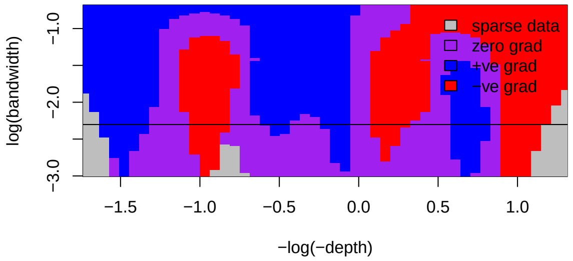
2 Univariate data example

The `earthquake` data set contains 510 observations, each consisting of measurements of an earthquake beneath the Mt St Helens volcano. The first is the longitude (in degrees, where a negative number indicates west of the International Date Line), second is the latitude (in degrees, where a positive number indicates north of the Equator) and the third is the depth (in km, where a negative number indicates below the Earth's surface). For the univariate example, we take the log(–depth) as our variable of interest. The kernel density estimate with bandwidth 0.1 is the orange curve. Superimposed in green are the sections of this density estimate which have significant gradient (i.e. significantly different from zero). The rug plot is the log(–depth) measurements.

Below this is the SiZer plot of Chaudhuri and Marron (1999). In the SiZer plot, blue indicates significantly increasing gradient, red is significantly decreasing gradient, purple is non-significant gradient and grey is data too sparse for reliable estimation. The horizontal black line is for the bandwidth 0.1.

```
> library(feature)
> data(earthquake)
> eq3 <- log10(-earthquake[, 3])
> eq3.fs <- featureSignif(eq3, bw = 0.1)
> plot(eq3.fs, xlab = "-log(-depth)")
> eq3.SiZer <- SiZer(eq3, bw = c(0.05, 0.5), xlab = "-log(-depth)")
```

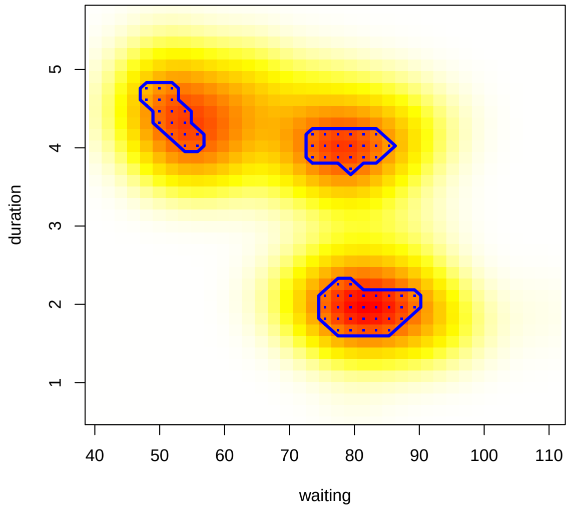




3 Bivariate data example

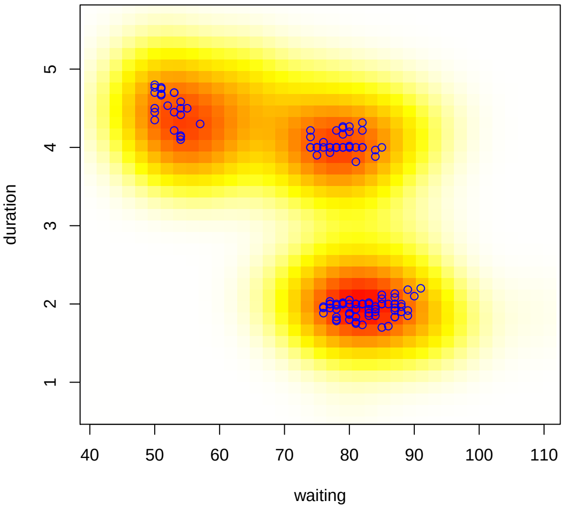
For bivariate data, we look at an Old Faithful geyser data set, in the `MASS` library. The horizontal axis is the waiting time (in minutes) between two eruptions, and the vertical axis is the duration time (in minutes) of an eruption. Below is a kernel density estimate with bandwidth (4.5, 0.37) with the significant curvature regions in blue superimposed.

```
> library(MASS)
> data(geyser)
> geyser.fs <- featureSignif(geyser, bw = c(4.5, 0.37))
> plot(geyser.fs, addSignifCurvRegion = TRUE)
```



A variation on plotting the significant regions is to plot the data points which fall inside these regions: significant curvature data points are in blue.

```
> plot(geyser.fs, addSignifCurvData = TRUE)
```



The result of `featureSignif` is an object of class `fs` which is a list with fields

```
> names(geyser.fs)
```

```
 [1] "x"                 "names"             "bw"                "fhat"
 [5] "grad"              "curv"              "gradData"          "gradDataPoints"
 [9] "curvData"          "curvDataPoints"
```

where `x` is the data, `names` are the name labels used for plotting, `bw` is the bandwidth, `fhat` is the kernel density estimate, `grad` is the logical matrix indicating signficant gradient on a grid, `curv` is the logical matrix indicating signficant curvature on a grid, `gradData` is the logical vector indicating signficant gradient data points, `gradDataPoints` are the signficant gradient data points, `curvData` is the logical vector indicating signficant curvature data points, and `curvDataPoints` are the signficant curvature data points.

4 Functionality not documented in this vignette

This package includes feature significance for 3-dimensional data. However these displays rely on an rgl engine which is not integrated with `Sweave` so we have excluded examples for the time being. See the example code in `?featureSignif`. The function `featureSignif` is non-interactive. Its interactive version is `featureSignifGUI`. We don't illustrate it in this vignette, see `?featureSignifGUI`.

References

Chaudhuri, P. and 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., and Wand, M. P. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, **52**, 4225–4242.

Godtliebsen, F., Marron, J. S., and Chaudhuri, P. (2002). Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics*, **11**, 1–21.