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

Tarn Duong

4 November 2008

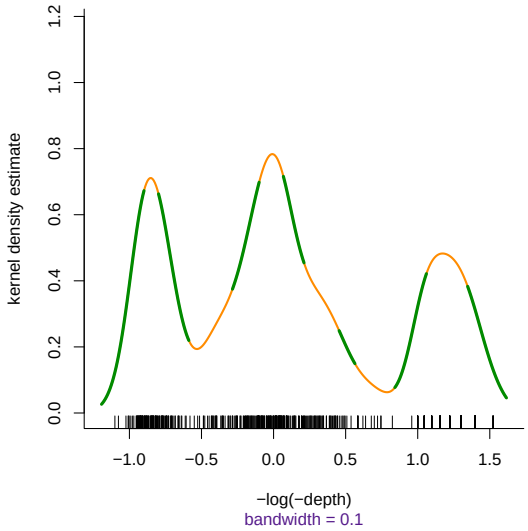
1 Introduction

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. `feature` is an R package for feature significance for 1- to 4-dimensional data. There is one main function in this package, `featureSignif`. It has a range of options which allow the user to compute and display 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

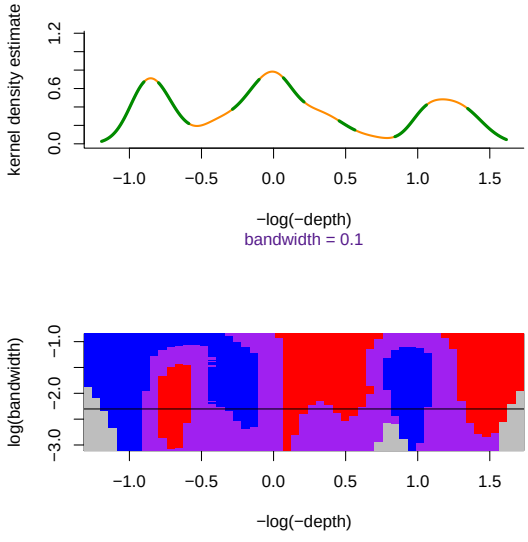
The `earthquake` data set is contained in `feature`. It contains 510 observations, each consisting of measurements of an earthquake beneath the Mt St Helens volcano. The first is the longitude (in degrees, where a negative number indicates west of the International Date Line), second is the latitude (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.

```
> library(feature)
> data(earthquake)
> eq3 <- -log10(-earthquake[, 3])
> featureSignif(eq3, addData = TRUE, addSignifGradRegion = TRUE,
+     xlab = "-log(-depth)", bw = 0.1)
```



Below is the same kernel density estimate and significant gradient region plot along with 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.

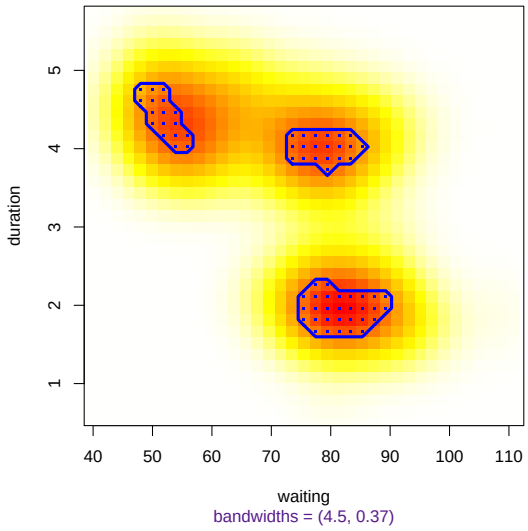
```
> featureSignif(eq3, addSignifGradRegion = TRUE, xlab = "-log(-depth)",
+     bw = 0.1)
> featureSignif(eq3, plotSiZer = TRUE, xlab = "-log(-depth)")
> lines(c(-2, 2), c(log(0.1), log(0.1)))
```



3 Bivariate data example

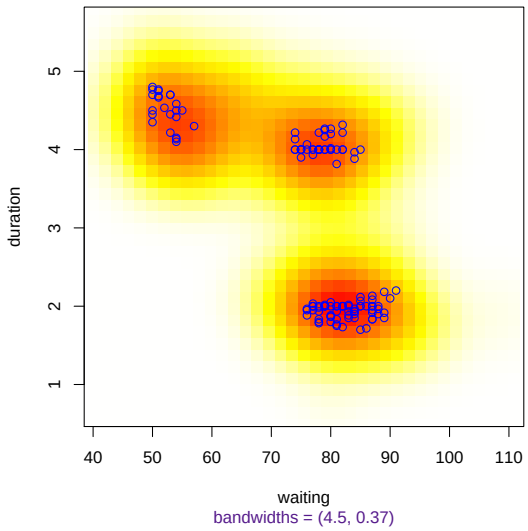
For bivariate data, we look at an Old Faithful geyser data set, in the `MASS` library. The horizontal axis is the waiting time (in minutes) between two eruptions, and the vertical axis is the duration time (in minutes) of an eruption. Below is a kernel density estimate with bandwidth (4.5, 0.37) with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

```
> library(MASS)
> data(geyser)
> featureSignif(geyser, addSignifCurvRegion = TRUE, bw = c(4.5,
+     0.37))
```



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.

```
> fs <- featureSignif(geyser, addSignifCurvData = TRUE, bw = c(4.5,
+     0.37))
```



Usually `featureSignif` returns invisibly to the command line but in this example, we assigned it to the variable `fs`.

```
> names(fs)
```

```
[1] "x"              "bw"             "fhat"           "curv"
[5] "curvData"       "curvDataPoints"
```

where `x` is the data, `bw` is the bandwidth, `fhat$est` is the kernel density estimate on the grid `fhat$x.grid`, `curv` is the logical matrix indicating signficant curvature on a grid, `curvData` is the logical vector indicating signficant curvature data points. So the points with significant curvature are

```
> fs$x[fs$curvData, ]
```

```
  waiting duration
1      80 4.016667
4      80 4.000000
5      75 4.000000
6      77 2.000000
9      77 2.033333
```

⋮

4 Functionality not documented in this vignette

`feature` includes feature significance for 3- and 4-dimensional data. However the displays in these dimensions rely on the rgl engine (Adler and Murdoch, 2006) which is not quite integrated with `Sweave` so we have excluded examples for the time being. See the example code in `?featureSignif`.

These examples have used `featureSignif` in its non-interactive mode i.e. where the user supplies a particular value of the bandwidth. In its interactive mode, the user is able to choose a bandwidth from a range of bandwidths and the significant features are displayed in real-time. Again it's not possible to illustrate this in this vignette, see `?featureSignif`.

References

Adler, D. and Murdoch, D. (2006). *rgl: 3D visualization device system (OpenGL).* R package version 0.67-2.

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.