feature: R package

Tarn Duong

August 15, 2006

1 Introduction

Feature significance is an extension of kernel density estimation, which is used to establish the statistical significance of features (like local modes). See Chaudhuri & Marron (1999) for 1-dimensional data, Godtliebsen, Marron & Chaudhuri (2002) for 2-dimensional data and Duong, Cowling, Koch & Wand (2006) 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. We cover the options which a newcomer would most likely require. For the full range of options (e.g. varying plotting colours), see `?featureSignif`.

2 Examples

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. Figure 1 contains the kernel density estimate with bandwidth 0.1 (in orange). Superimposed in green are the sections of this density estimate which as significant gradient (i.e. significantly different from zero). The rug plot is the log –depth measurements.

Figure 2 contains the same kernel density estimate and significant gradient region plot along with the SiZer plot (cf. Chaudhuri & 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)

Loading required package: KernSmooth
KernSmooth 2.22 installed
Copyright M. P. Wand 1997
Loading required package: misc3d
Loading required package: rgl
feature 1.1-3 1.0-0 (2006)

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

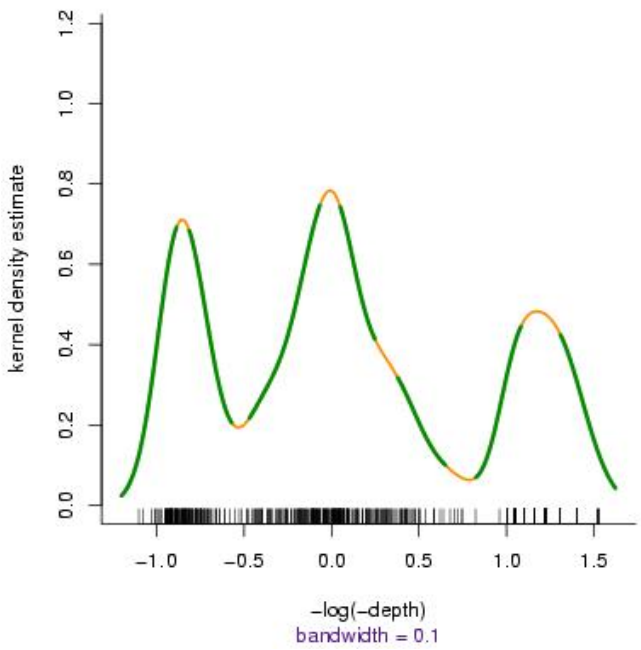

Figure 1: Univariate data: significant gradient region with kernel density estimate

```
> layout(matrix(1:2, nrow = 2))
> featureSignif(eq3, addSignifGradRegion = TRUE, xlab = "-log(-depth)",
+     bw = 0.1)
> xlim <- par()$usr[1:2]
> featureSignif(eq3, plotSiZer = TRUE, xlab = "-log(-depth)", xlim = xlim)
> lines(c(-2, 2), c(0.1, 0.1))
> layout(1)
```

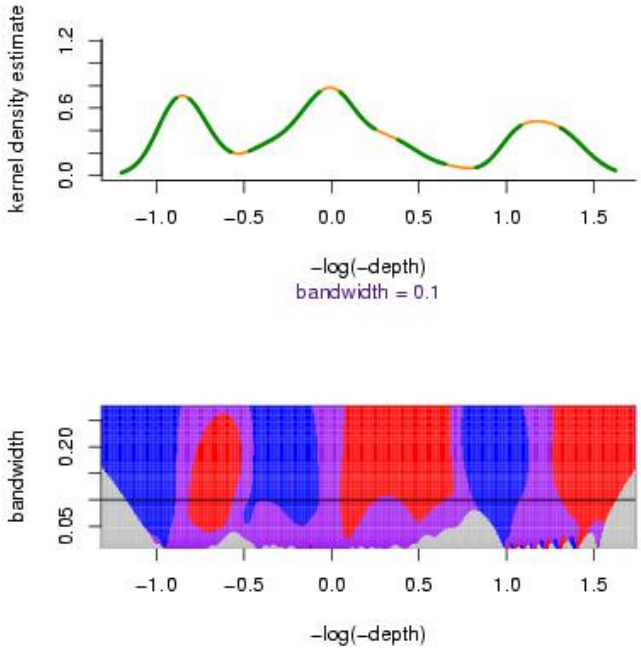

Figure 2: Univariate data: significant gradient regions and SiZer plot

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. Figure 3 has a kernel density estimate with bandwidth (4.5, 0.37), with the significant gradient regions in green and the significant curvature regions in blue superimposed.

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

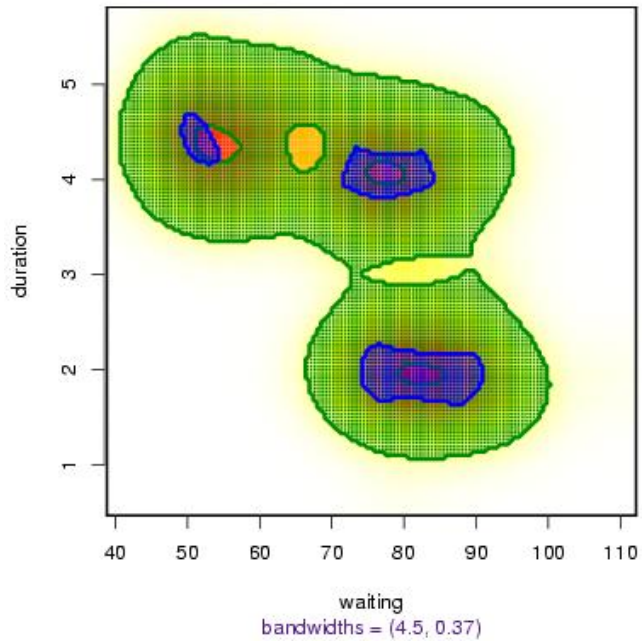

Figure 3: Bivariate data: significant gradient and curvature regions

A variation on plotting the significant regions is to plot the data points which fall inside these regions in Figure 4: significant gradient data points are in green, significant curvature data points are in blue.

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

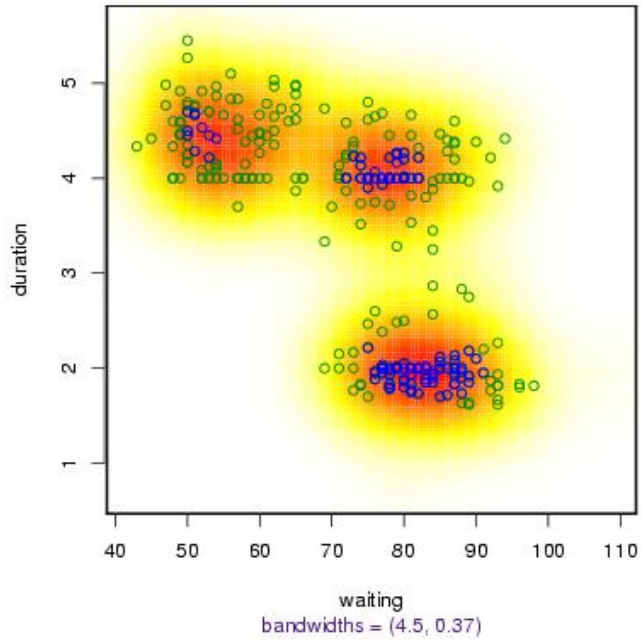

Figure 4: Bivariate data: significant gradient and curvature data points

Usually `featureSignif` returns invisbly to R but in this example, we assigned it to the variable `fs`.

```
> names(fs)

[1] "x"    "bw"   "fhat" "grad" "curv"
```

where x is the data, bw is the bandwidth, fhat$est is the kernel density estimate on the grid fhat$x.grid, grad is the matrix indicating signficant gradient and curv is the matrix indicating signficant curvature.

```
> fs$x[1:5, ]

  waiting duration
1      80 4.016667
2      71 2.150000
3      57 4.000000
4      80 4.000000
5      75 4.000000

> fs$bw

[1] 4.50 0.37

> fs$fhat$x.grid[[1]][30:35]

[1] 51.42667 51.95000 52.47333 52.99667 53.52000 54.04333

> fs$fhat$x.grid[[2]][101:106]

[1] 4.096111 4.134289 4.172467 4.210644 4.248822 4.287000

> fs$fhat$est[30:35, 101:106]

           [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
[1,] 0.01140390 0.01185379 0.01225223 0.01259516 0.01287930 0.01310208
[2,] 0.01169080 0.01214089 0.01253763 0.01287711 0.01315621 0.01337253
[3,] 0.01191348 0.01236082 0.01275322 0.01308690 0.01335892 0.01356710
[4,] 0.01207150 0.01251349 0.01289923 0.01322512 0.01348841 0.01368714
[5,] 0.01216557 0.01259999 0.01297718 0.01329373 0.01354706 0.01373545
[6,] 0.01219754 0.01262263 0.01298985 0.01329595 0.01353859 0.01371622

> fs$grad[30:35, 101:106]

     [,1] [,2] [,3]  [,4]  [,5]  [,6]
[1,] TRUE TRUE TRUE  TRUE  TRUE  TRUE
[2,] TRUE TRUE TRUE  TRUE  TRUE FALSE
[3,] TRUE TRUE TRUE  TRUE FALSE FALSE
[4,] TRUE TRUE TRUE  TRUE FALSE FALSE
[5,] TRUE TRUE TRUE FALSE FALSE FALSE
[6,] TRUE TRUE TRUE FALSE FALSE FALSE

> fs$curv[30:35, 101:106]
```

```
      [,1]  [,2]  [,3] [,4] [,5] [,6]
[1,] FALSE FALSE FALSE TRUE TRUE TRUE
[2,] FALSE FALSE FALSE TRUE TRUE TRUE
[3,] FALSE FALSE FALSE TRUE TRUE TRUE
[4,] FALSE FALSE  TRUE TRUE TRUE TRUE
[5,] FALSE FALSE FALSE TRUE TRUE TRUE
[6,] FALSE FALSE FALSE TRUE TRUE TRUE
```

`feature` includes feature significance for 3- and 4-dimensional data. However the displays in these dimensions rely on the rgl (Adler & Murdoch 2006) engine 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 `feature` in its non-interactive mode 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 nside this vignette, see `?featureSignif`.

References

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

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. (2006), 'Feature significance for multivariate kernel density estimation'. Submitted.

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