

## Notes on Kernel Density Estimation

The R function `density` handles most of the estimation approaches we studied in class. R also has a library (`KernSmooth`) of kernel density functions, but it seemed a bit piecemeal and outdated.

We can load the `datasets` package in R, and estimated the density of both `eruptions` and `waiting` of the `faithful` data (that's Old Faithful) using default options in `density`. Note that the default title and x axis for the `density` command are not attractive and should always be relabelled. If you do not remember the name of the `faithful` data set, you can use the `ls` command to look at the names of all the data sets to refresh your memory. The `summary` command will provide variable names. An alternative (and more attractive option) would be to enter `data()`, find `volcano` on the list, and then type `help(volcano)`. The first approach is recreated below.

```
search()           # Find the package "datasets" position on the list
ls(pos=8)          # The list of "datasets" includes "faithful"
summary(faithful)
attach(faithful)
plot(density(eruptions),main="Density Estimate of Eruption Duration",
      xlab="Duration in Minutes")
plot(density(waiting),main="Density Estimate of Eruption Interval",
      xlab="Interval in Minutes")
```

The default bandwidth method is `bw="nrd0"`; this is Silverman's Rule of Thumb method ( $b = .9\min(s, IQR/1.34)m^{-1/5}$ ), and often oversmooths. The bandwidth used for the plot is printed at the bottom of the page and also included as part of the `density` command output. Feel free to adjust the bandwidth by specifying a constant as we did in class (E.g., `bw=.25`).

```
plot(density(eruptions,bw=.25),main="Density Estimate of Eruption
      Duration",xlab="Duration in Minutes")
```

### LSCV

The package `locfit` can produce LSCV plots using the `lscvplot` command, though the command is currently misbehaving on our system, perhaps because our version of R isn't quite up-to-date. In the interim, I have included code to generate a plot. The code requires numerical integration of the square of the

kernel density estimate. *CAUTION*: The `integrate` function in R passes the vector of all quadrature points to the function that evaluates the integrand. Make absolutely sure your function can evaluate a *vector* rather than just a scalar.

The function I've written (`lscv`) may need some fine-tuning of the range of the bandwidth, and the function `kern` can be changed from the default normal density kernel. Results from the output generally agree with the bandwidth generated by `density` using the `bw="ucv"` option; differences may occur since I evaluate an approximation of the LSCV function, while R uses exact methods. Other than that, here is the typical usage:

```
source{"z:/stat 740/lscv.txt"} # Skip this step if you're pasting
                                # the file directly into R
attach(faithful)
lscv(waiting)
```

The estimated bandwidth is based on  $s$ , so the procedure is easily “fooled” by bimodal data. In such cases, the default scaling limits (`s1=.5`, `su=2.5`) can be changed in the call to `lscv`. This is absolutely necessary for `eruptions`, for instance, which requires a very small bandwidth.