Presentation on Using SiZer

Lindsey Turner
Department of Statistics
University of British Columbia

April 16, 2003

Based on a paper by: Chaudhuri P. and Marron S (1999), ”SiZer for Exploration of Structures of Curves”, JASA, 94, 807-823.

What is SiZer?

- Graphical device for the display of significant features with respect to location and bandwidth.
- Assessment of **SI**nificant **ZER**o crossing of derivatives results in a **SiZer** map.
- A zero crossing is ”significant” when the derivative estimate is significantly different from zero on both sides.
- SiZer can be used for density or regression estimation.

Notation

- Kernel Density Estimation

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{1=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

- Look at family of smooths $\{\hat{f}_h(x) : h \in [h_{min}, h_{max}]\}$
  - ⋆ Note: Unless specified in SiZer, h_{min} =2*binwidth and h_{max} =range of data
- Based on Confidence Limits for $\hat{f}'_h(x)$:

$$\hat{f}'_h(x) \pm q.\hat{sd}(\hat{f}'_h(x))$$

q:An Appropriate Quantile

- Motivation: If u_1 and u_2 are sufficiently far apart then $\hat{f}_h(u_1)$ and $\hat{f}_h(u_2)$ are independent $\Rightarrow$ $\hat{f}'_h(u_1)$ and $\hat{f}'_h(u_2)$ are independent.
- Then the simultaneous Confidence Limit problem can be approximated by m independepent confidence intervals, i.e. m independent blocks.
- Estimate m through Estimated Sample Size (ESS):

$$ESS(x,h) = \frac{\frac{1}{h}\sum_{1=1}^{n} K\left(\frac{X_i - x}{h}\right)}{\frac{1}{h}K(0)}$$

q:An Appropriate Quantile (cont...)

- Choose m to be the number of independent blocks of average size available from the data set of size n. To avoid problems with small ESS, calculation of block size is given by:

$$m(h) = \frac{n}{avg_{x \in D_h} ESS(x,h)}$$

where $D_h = \{x : ESS(x,h) \geq 5\}$

- So, assuming independence of these $m(h)$ blocks of data, the approximate simultaneous quantile is given by:

$$q(h) = \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right)$$

Standard Deviation

Main idea behind the calculation of $\hat{sd}(\hat{f}'_h(x))$ in this context, is that the derivative estimator $\hat{f}'_h(x)$ is an average (of the derivative kernel functions)

$$\begin{aligned} \hat{var}(\hat{f}'_h(x)) &= \hat{var}\left(\frac{1}{n}\sum_{1=1}^{n}\frac{1}{h^2}K'\left(\frac{X_i - x}{h}\right)\right) \\ &= \frac{1}{n}s^2\left(\frac{1}{h^2}K'\left(\frac{X_1 - x}{h}\right), \ldots, \frac{1}{h^2}K'\left(\frac{X_n - x}{h}\right)\right) \end{aligned}$$

SiZer Map

Colour Scheme:

- **Blue**: zero above confidence limits
  $\Rightarrow$ Curve is significantly increasing

- **Red**: zero below confidence limits
  $\Rightarrow$ Curve is significantly decreasing

- **Purple**: zero within confidence limits
  $\Rightarrow$ Curve is not significantly increasing or decreasing

- **Grey**: Indicates regions where the data is too sparse to make statements about significance
  $\Rightarrow$ When the effective sample size (ESS) is less than 5.

Using SiZer

- Download SiZer code from Steve Marron's site:
  http://www.stat.unc.edu/faculty/marron/marron_software.html

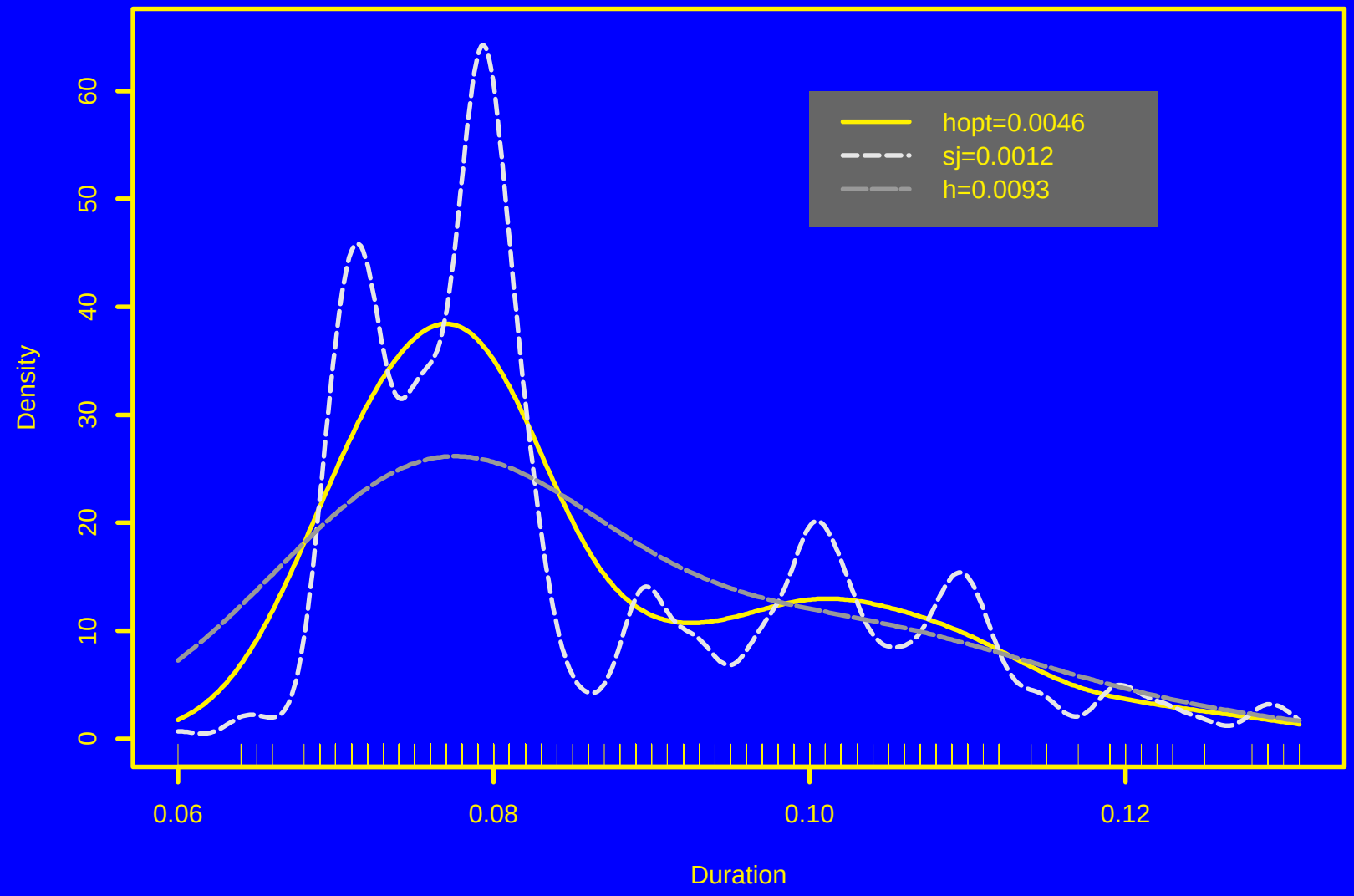
- Need function *norminv* to be able to run SiZer on our computers, can download it from:
  http://www.octave.sourceforge.net/afunclist.html

- Sizer Maps Code in Matlab;

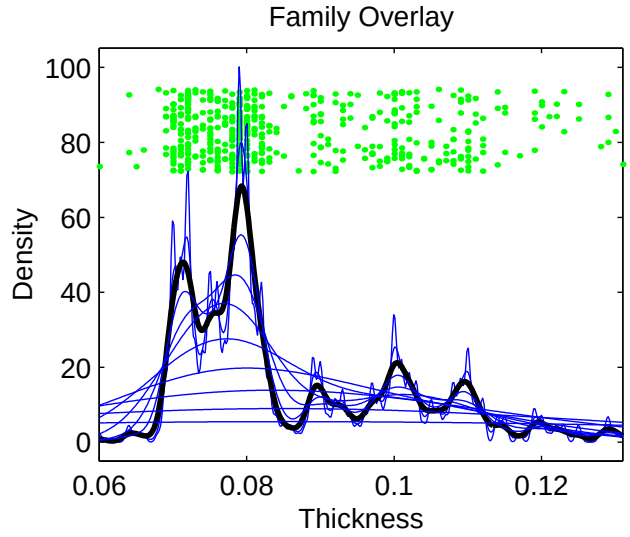
  - **SizerSM(data,struct('iout',3,'imovie',0,'hhighlight',h))**
    - *iout* specifies which plots to display
    - *imovie* specifies whether to make a movie version
    - *hhighlight* specifies whether to highlight a specific bandwidth
      - 0 don't highlight any h
      - -1 highlight h closest to data driven
      - h$>$0 highlight h closest to this value

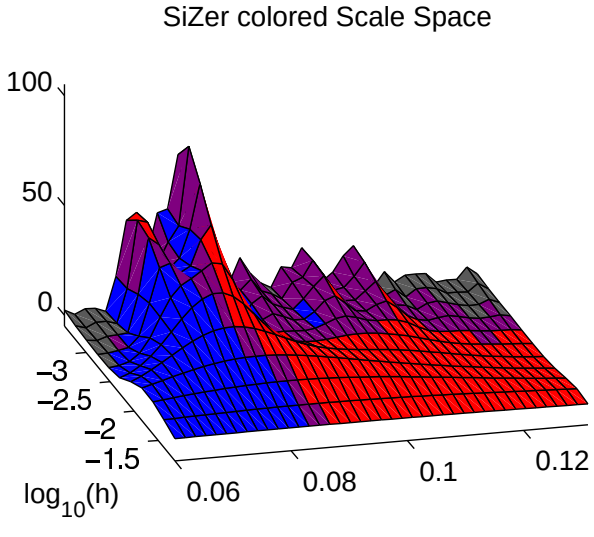
  Note: bandwidth is not highlighted when family plot is not displayed and is data driven when imovie is set to 1.

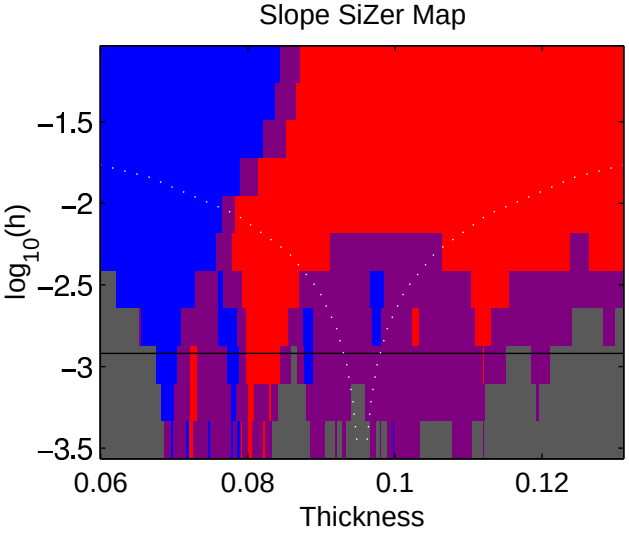
Hidalgo Stamp Data

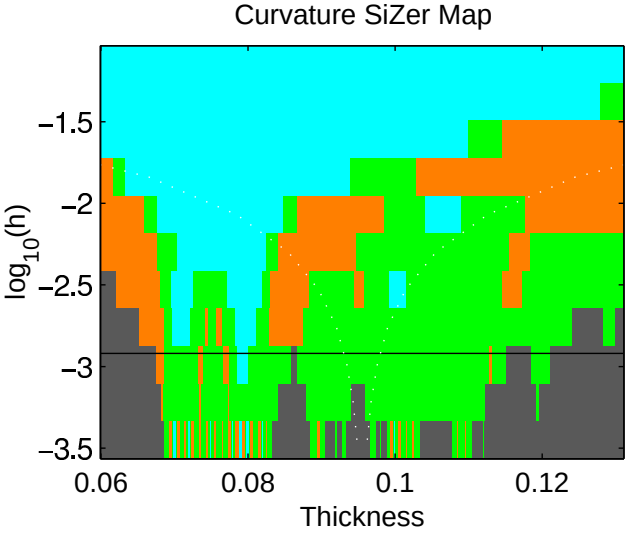


Sizer Map: SJ Bandwidth

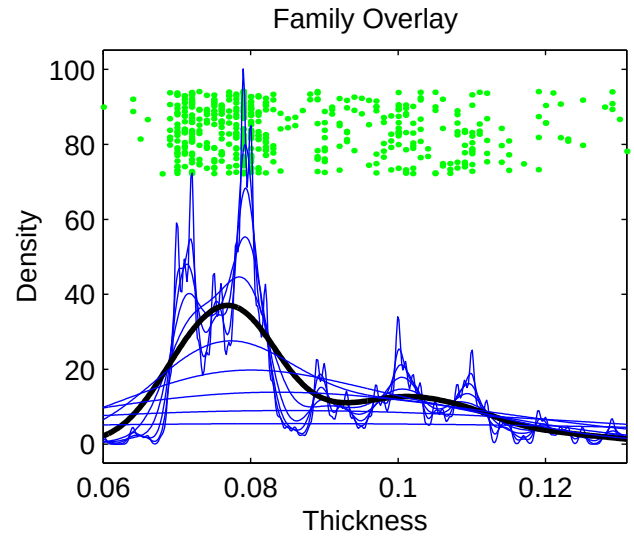


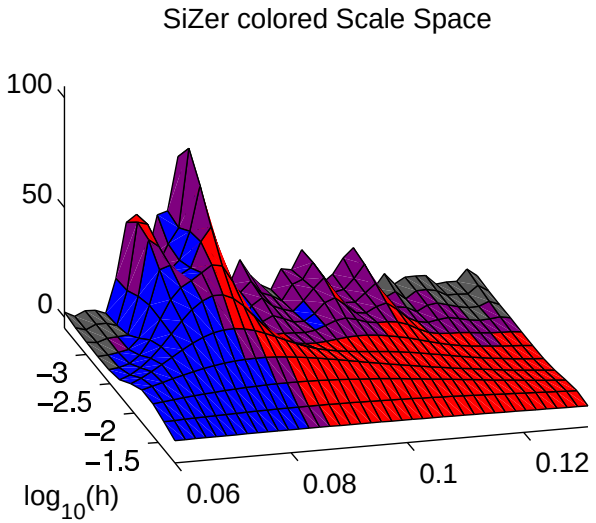


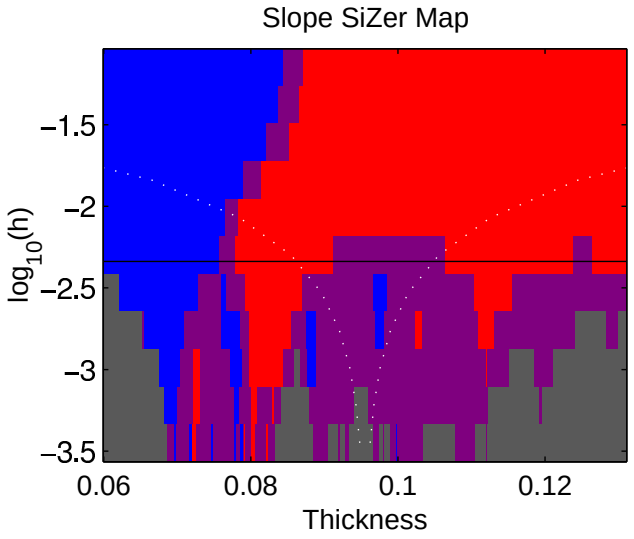


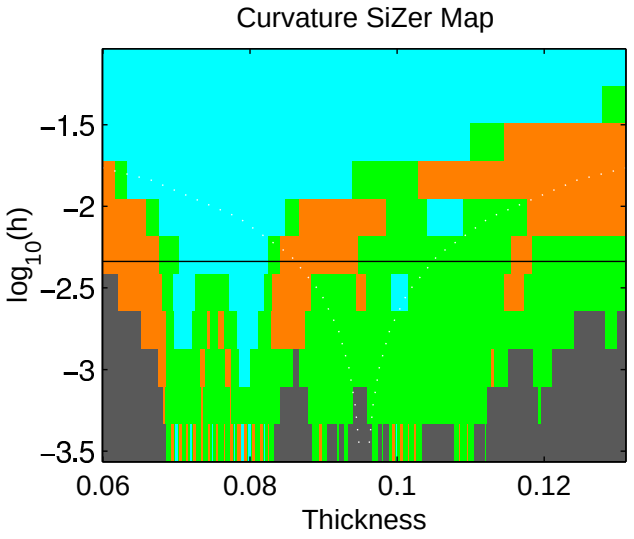


Sizer Map: $\mathbf{h_{opt}}$ Bandwidth

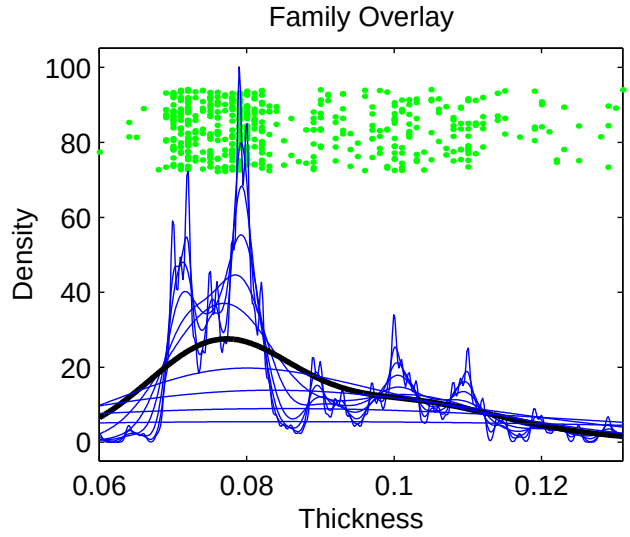


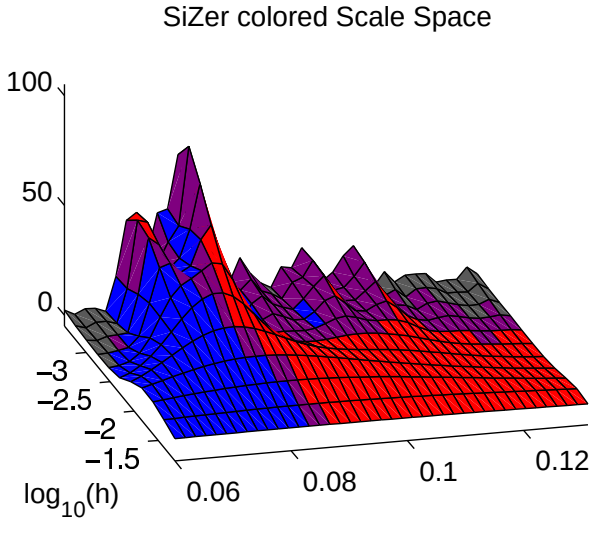


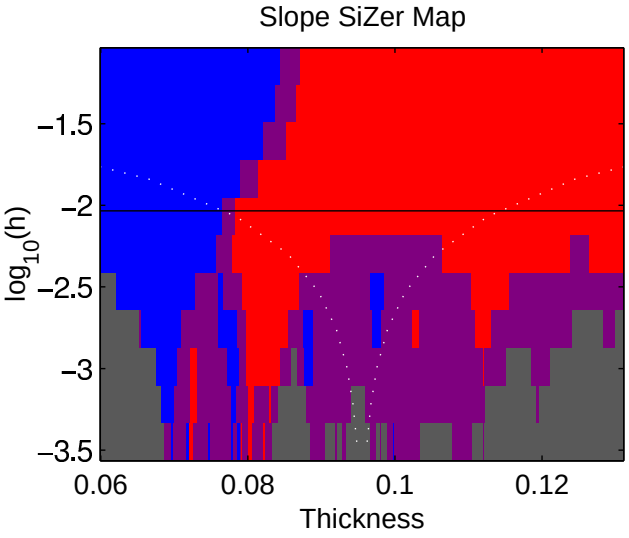


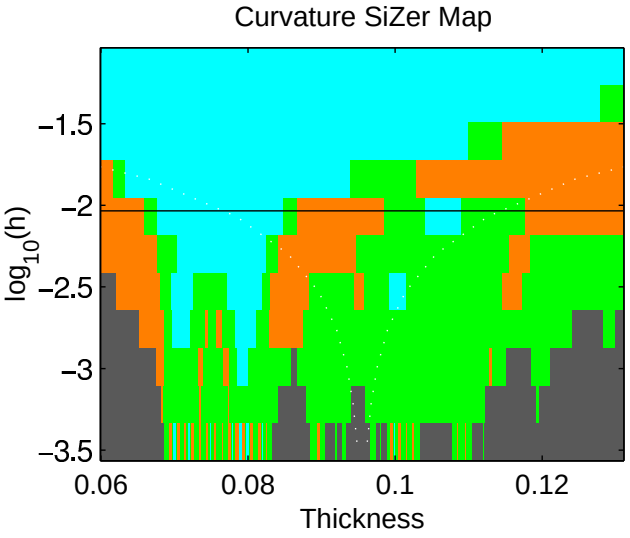


Sizer Map: Oversmoothed Bandwidth

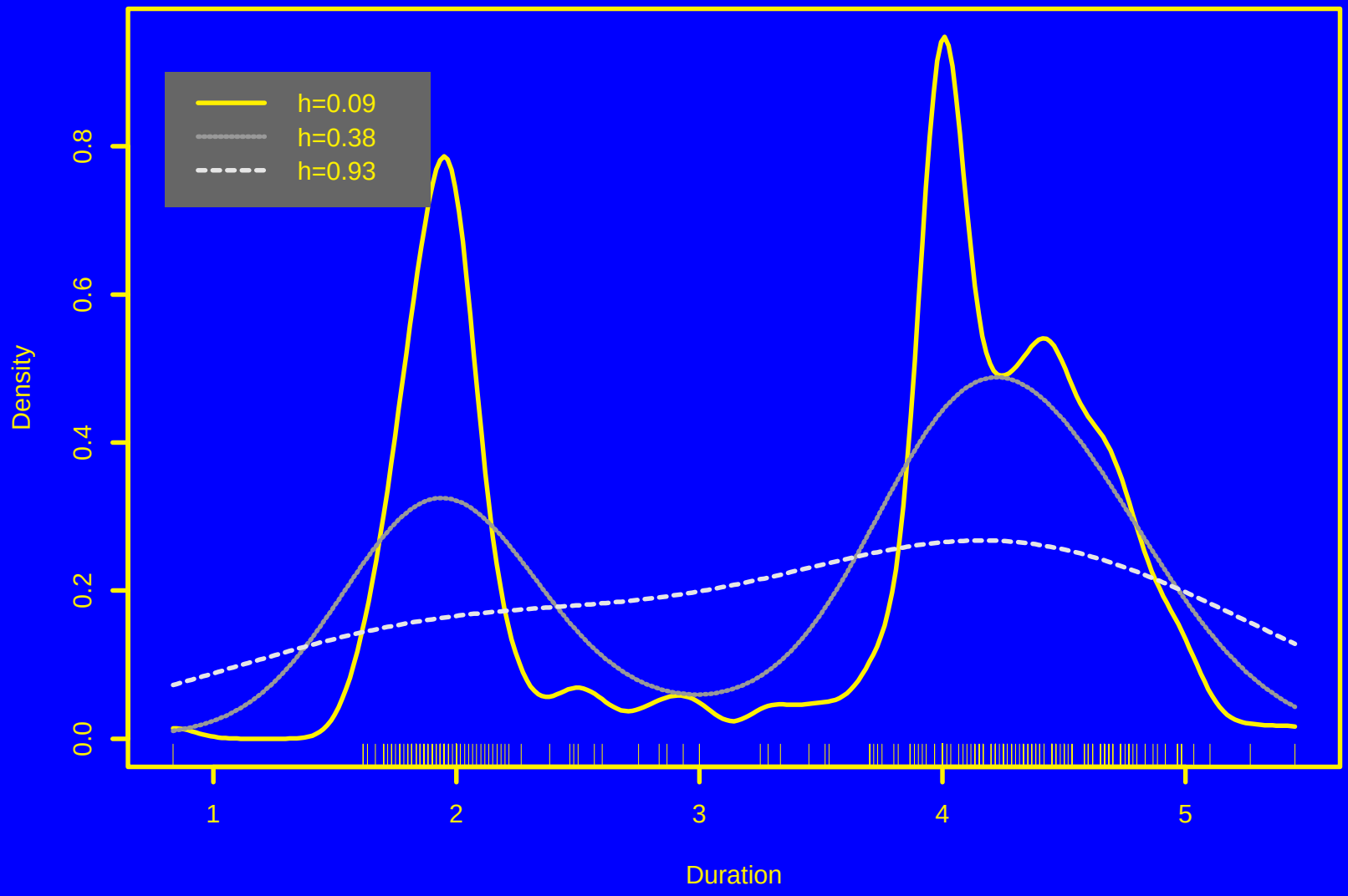




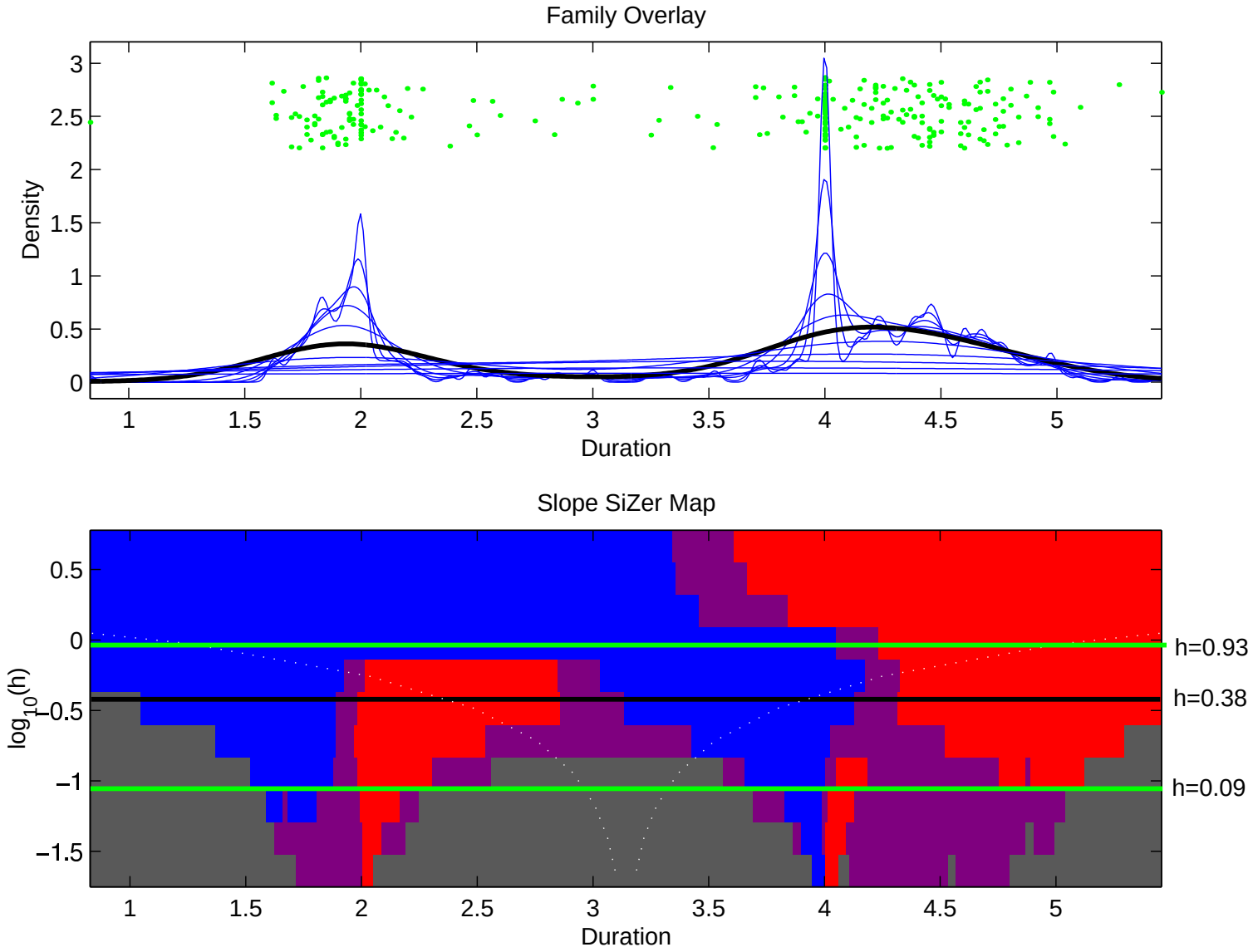




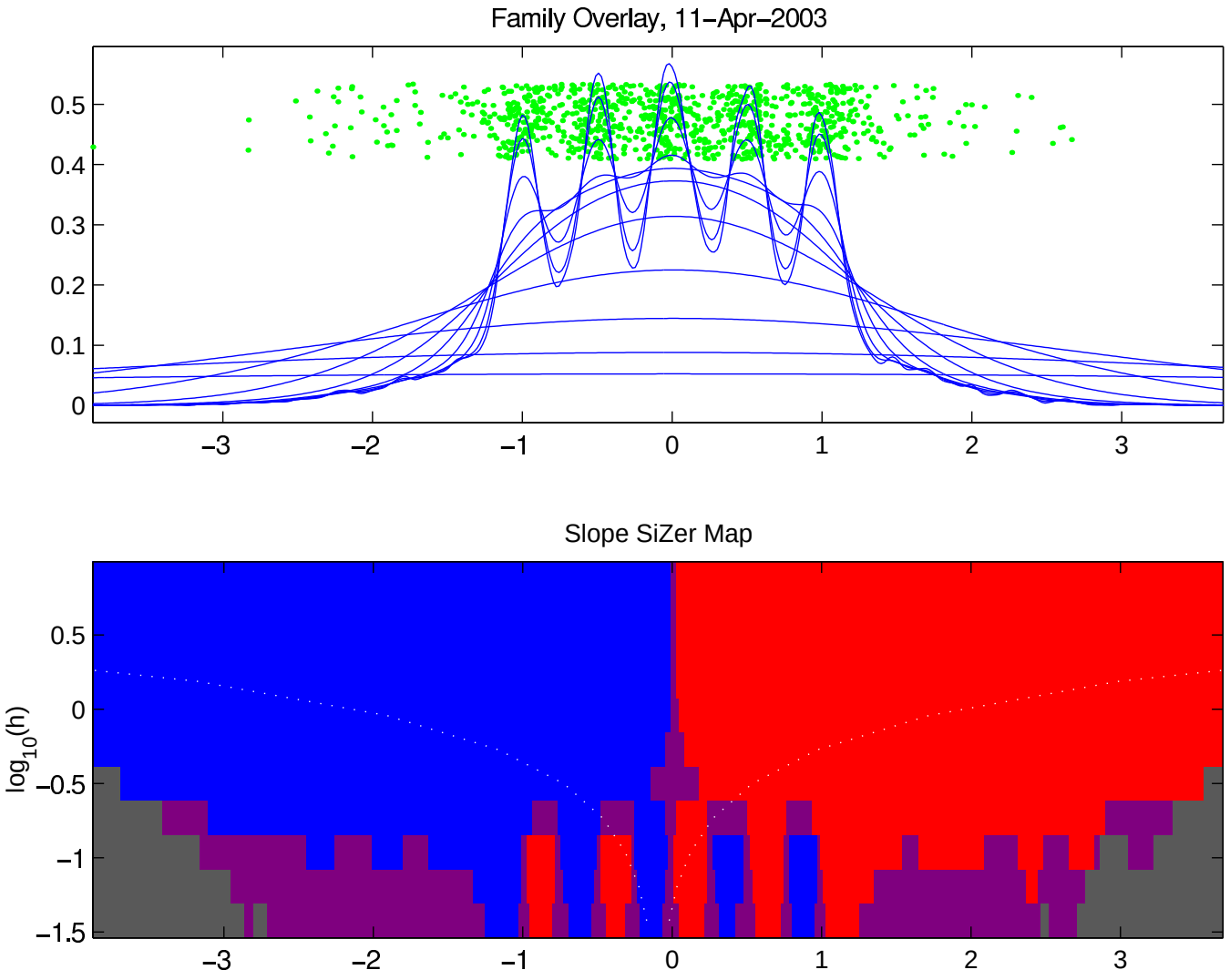
Geyser Data



Sizer Map for Geyser Data



Sizer Map for Generated Data



Summary

- SiZer can be thought of as adding statistical inference to a family of smooths.
- SiZer looks at a number of bandwidths and gives an indication as to which features in the smooth are ”really there”.
- The SiZer map counts the number of significant modes at different bandwidths and gives information about mode locations.
- There is a trade off though!!
  The SiZer map tends to be more conservative than mode tests that specifically target the number of modes.