Nonparametric Function Estimation
Stat 550[1] Chapter 9
Special Topics

David W Scott[2]

Rice University

November 7
Fall 2023
Rice University

[1] A course based upon the 2nd edition of *Multivariate Density Estimation; Theory, Practice, and Visualization*, *John Wiley & Sons, 2015*

[2] www.stat.rice.edu/∼scottdw/

Chapter IX: Other Applications

- An abbreviated sampling of advanced applications
- Classification, Discrimination, and Likelihood Ratios
- The lipid data re-visited
- Risk analysis of plasma lipid dataset.

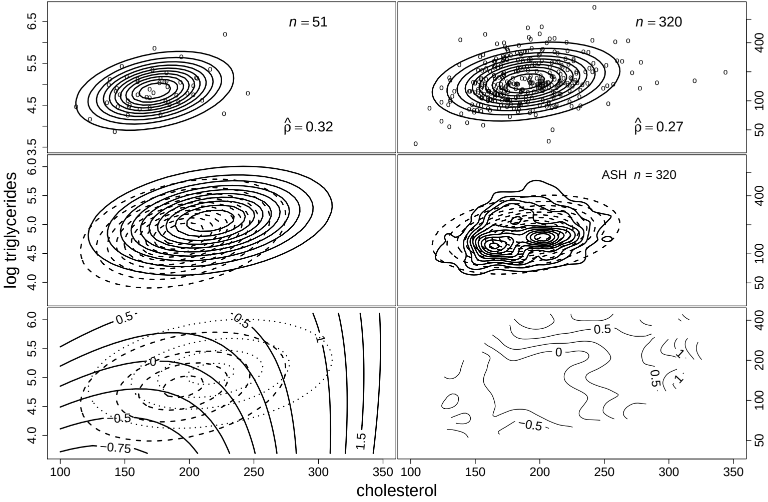


Figure: (Upper left) Normal fit "−" group. (Upper right) Normal fit "+" group. (Middle left) Overlay of 2 normal fits. (Middle right) Overlap of ASH of f_+ [biweight kernel $h = (21.7, 0.33)$] and normal fit to "+" group. (Lower left) Contours of parametric $\log_{10}(\text{LR})$. (Lower right) $\log_{10}(\text{LR})$ nonparametric estimate.

Likelihood Ratio Views Lipid Data: Parametric vs Nonparametric

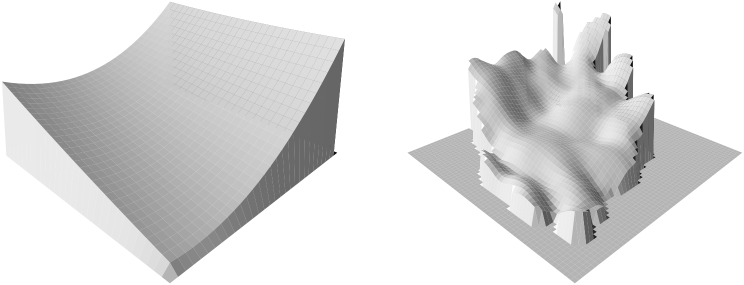


Figure: Perspective plots of the $\log_{10}$ likelihood ratio surfaces in previous Figure. The range of the vertical axes is $(-0.91, 0.91)$ in both frames, corresponding to a range of odds in favor of disease from 0.15:1 to 8.1:1.

LANDSAT Classification

Table: Classification Cross-Tabulations Based on Trivariate Gaussian and ASH Fits to the Landsat Data[a]

	PRED	Sunflwr	Wheat	Barley	% Right	Smoothed
NORM	Sunflwr	1,191	9	0	99.3%	100.0%
	Wheat	10	665	335	65.8%	80.3%
TRUTH:	Barley	10	314	1,066	76.7%	93.7%
ASH	Sunflwr	1,194	5	0	99.5%	100.0%
	Wheat	7	773	230	76.5%	93.7%
TRUTH:	Barley	3	361	1,026	73.8%	89.9%

[a] The first 3 columns summarize the predictions of the classifier using the training data. (book)(Classification!majority prediction) (book)(Classification!prediction) The last column summarizes the rates using a classification rule based on a majority rule of a pixel and its 8 neighbors.

Mixture: MCLUST Library (lipid example)

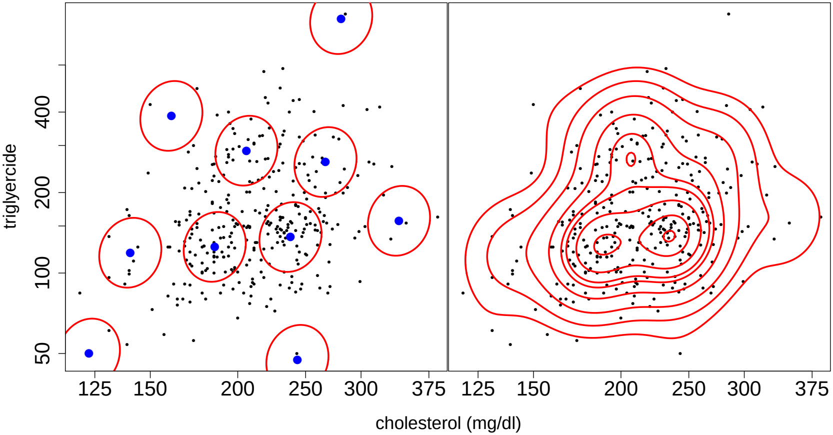


Figure: Mclust (2005 version) applied to log-lipid dataset ($n = 320$).

frametitleClustering: The Mode Tree (Geyser Data)

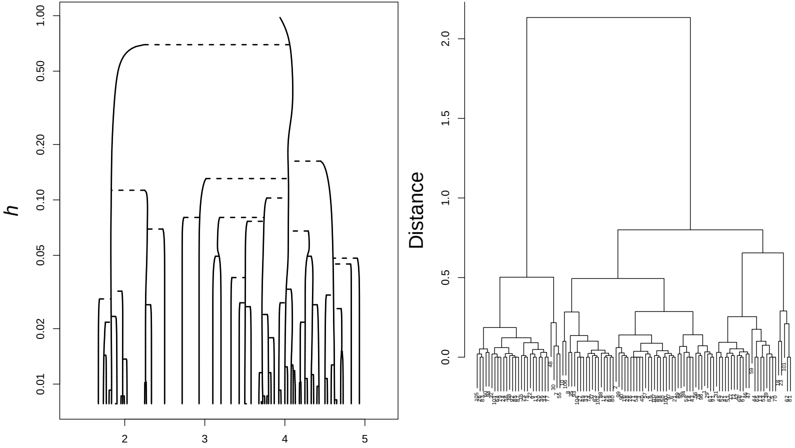


Figure: A mode tree and dendrogram of the geyser eruption times. The dendrogram is the hierarchical clustering tree based on average linkage.

Clustering: The 2-D Mode Tree (Lagged Geyser Data)

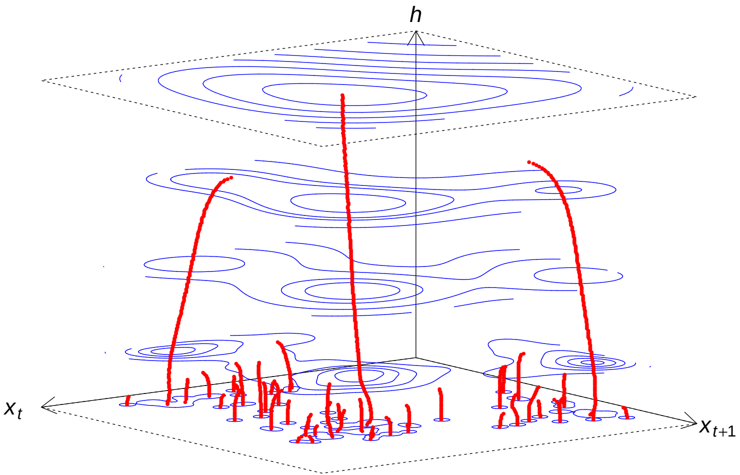


Figure: Bivariate mode tree of the lagged geyser dataset. Contours of the

Surprise: Number of Modes Not Monotone With h

A simple example where an equal mixture of 3 bivariate normal kernels at the corners of an equilateral triangle can have 1, 3, or 4 modes!

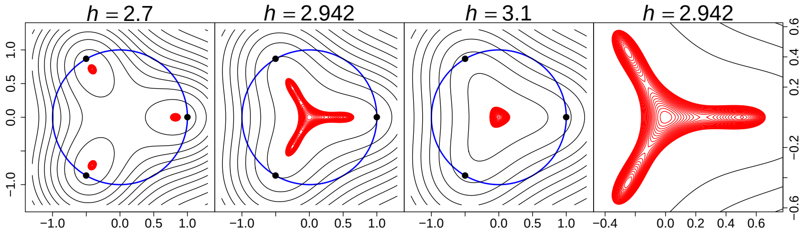


Figure: Contours of a bivariate Gaussian kernel density estimator with $n = 3$ points (black dots) on the unit circle forming an equilateral triangle. A highly nonlinear set of contour levels are displayed, so that the contours near the modes are emphasized.

Clusters in Images

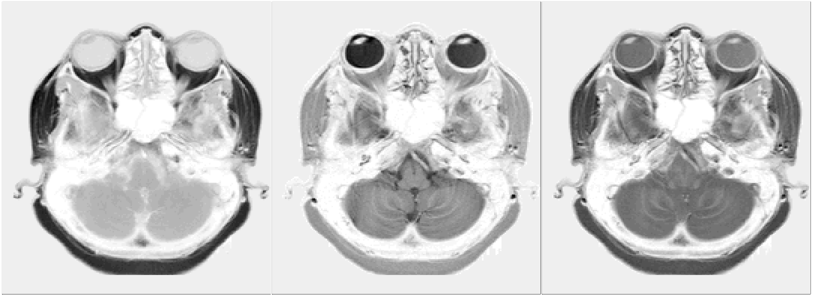


Figure: Gray scale images of the three MRI variables (t_1, t_2, sd).

Hill-Climbing to Local Modes

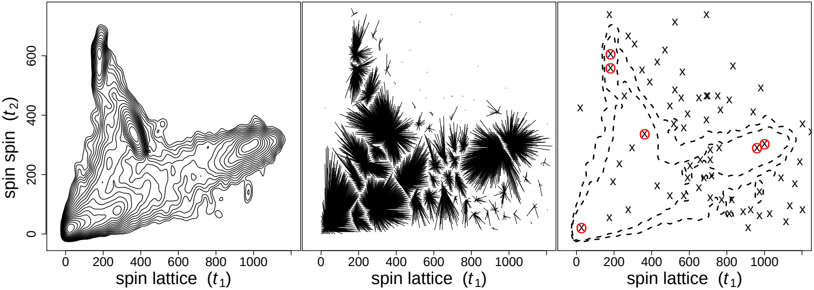


Figure: (Left) ASH contours of (t_1, t_2) of an MRI image with 24,476 pixels. (Center) Hill-climbing of individual pixel values to the nearest mode. (Right) The 70 modes found are superimposed on two contours.

Tumor Detection

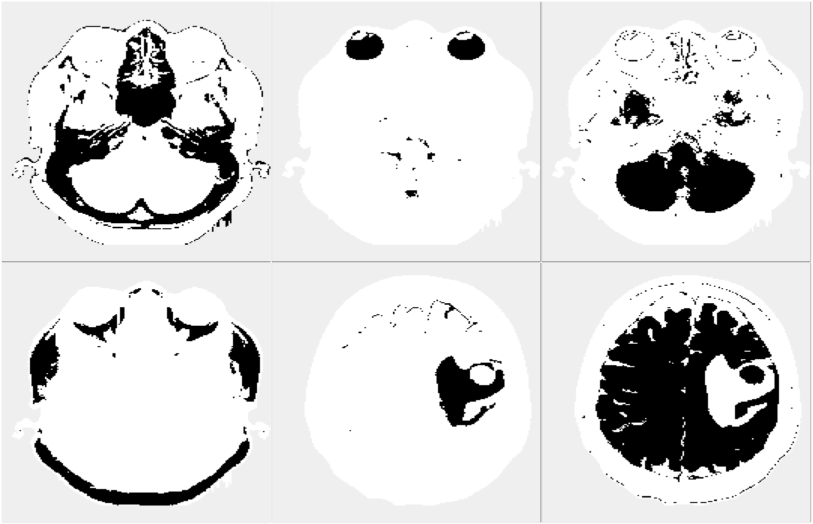


Figure: MRI images and subsets.

Data With Holes in $\Re^3$

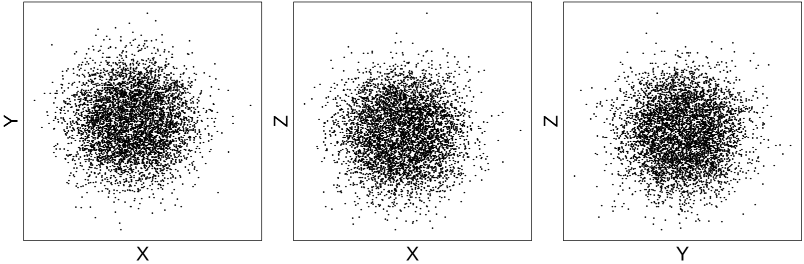
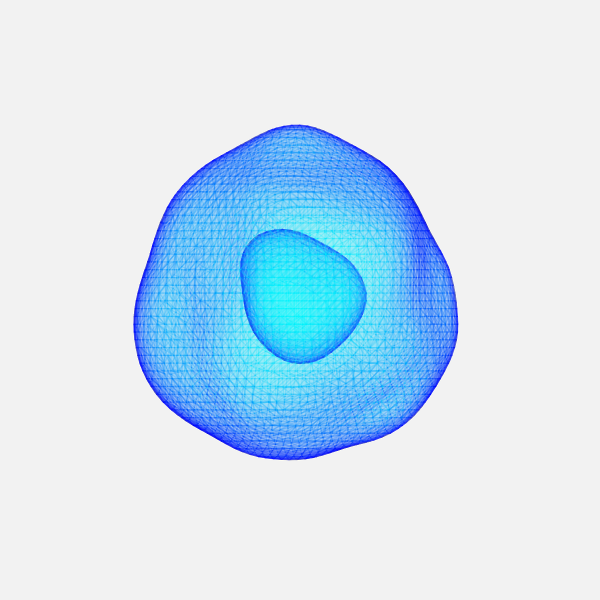
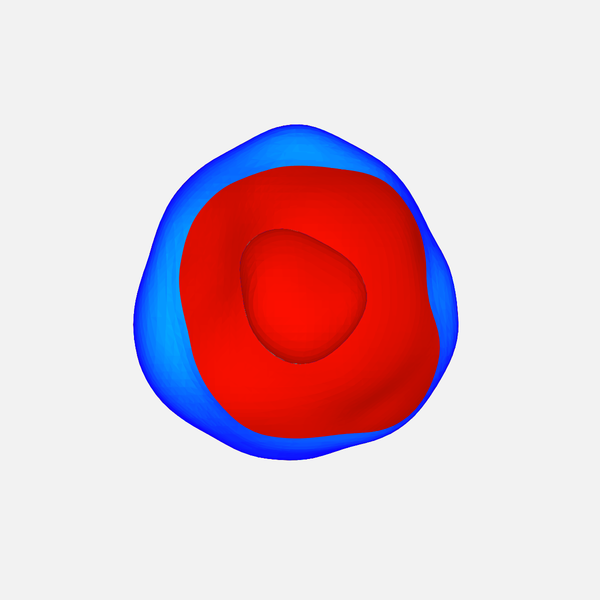
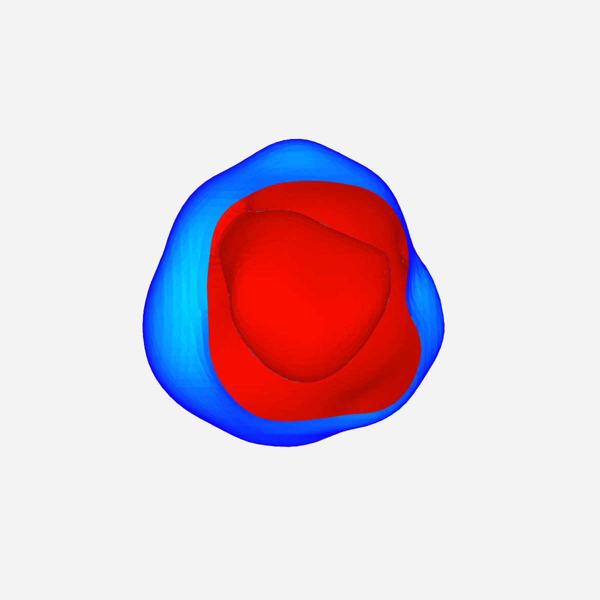
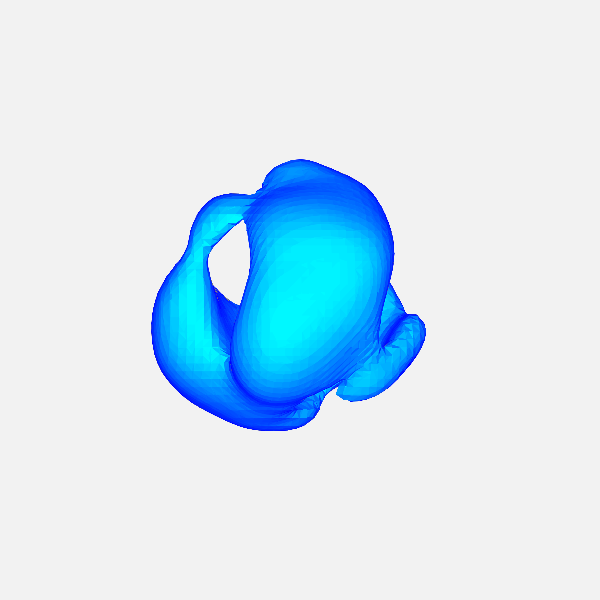


Figure: Pairwise scatterplots of 5,000 trivariate simulated points with a hole. The hole is actually a region of lower density rather than a region around the origin with no data.









LANDSAT Images and Histogram Equalization

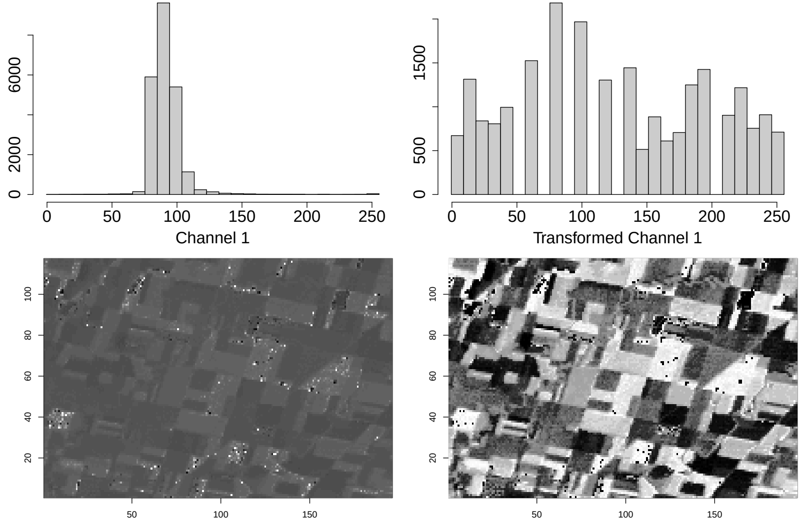


Figure: Histograms of raw data from Landsat scene and transformed data that are more nearly uniform. The increased dynamic range in the gray scale images may be observed.

THANK YOU!!

Any questions: scottdw@rice.edu