Color versions of B&W figures in MDE, 2nd edition, 2015.

Click on figure number below to jump to that color graphic.

- Figures 1.4 1.6 1.21 1.22 1.23 1.24
- Figure 2.4
- Figures 3.4 3.23 3.27 3.28
- Figures 4.6 4.9
- Figures 6.18 6.19 6.24 6.25 6.26 6.28 6.29 6.30 6.31
- Figures 7.6 7.7 7.9 7.15 7.16 7.17
- Figures 8.6 8.11
- Figures 9.1 9.4 9.5 9.6 9.8 9.9 9.11 9.13 9.17

Version 1.0 March 14, 2015

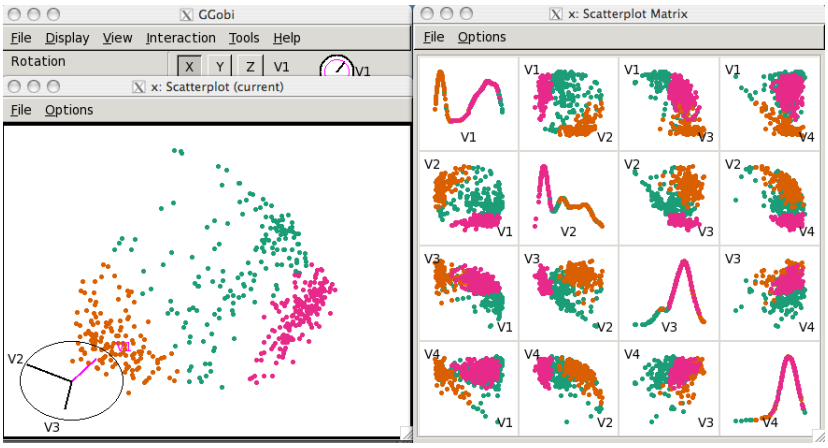

Figure 1.4 Pairwise scatterplots of the transformed PRIM4s data using the ggobi visualization system. Two clumps of points are highlighted by brushing.

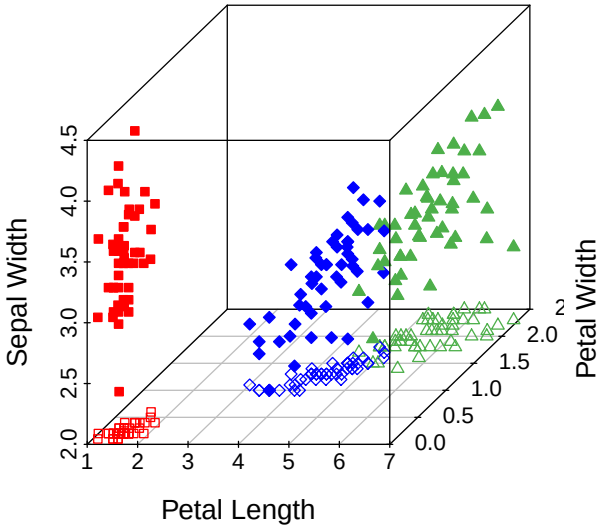

Figure 1.6 A three-dimensional scatter diagram of the Fisher-Anderson *Iris* data, omitting the sepal length variable. From left to right, the 50 points for each of the three varieties of *setosa*, *versicolor*, and *virginica* are distinguished by symbol type and color, (square, diamond, triangle) and (red, blue, green), respectively. The symbol/coloring is required to indicate the presence of three clusters rather than only two. The same basic picture results from any choice of three variables from the full set of four variables.

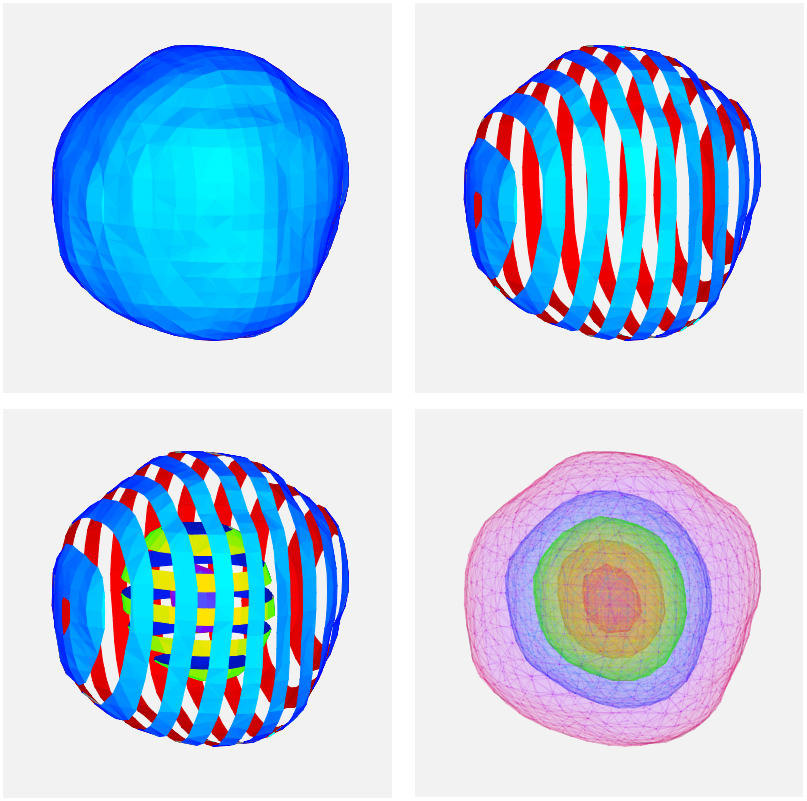

Figure 1.21 Trivariate normal examples.

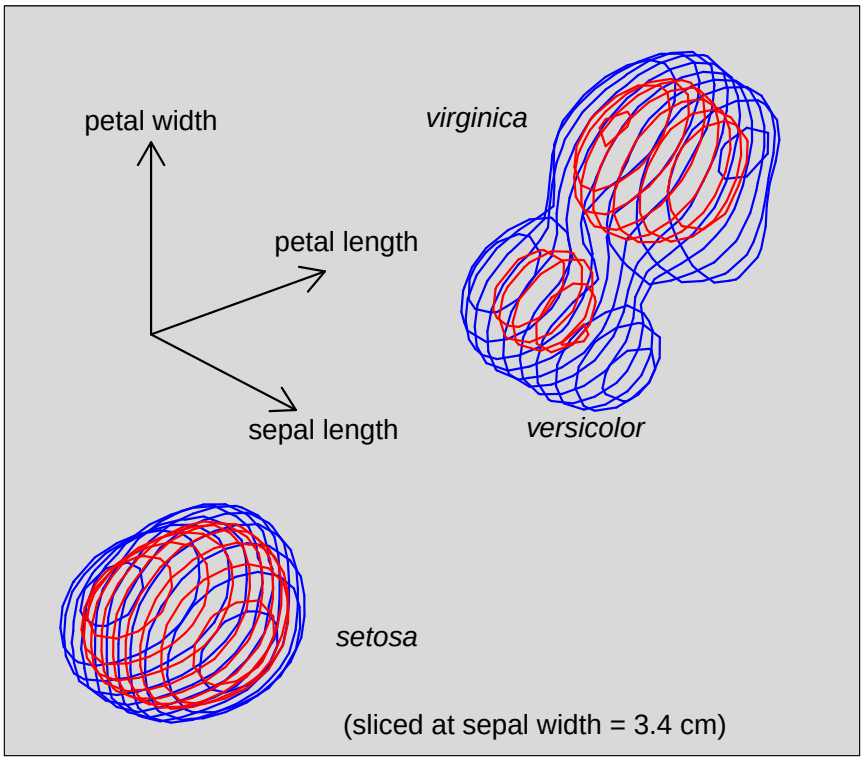

Figure 1.22 Two α-level contour surfaces from a slice of a five-dimensional averaged shifted histogram estimate, based on all 150 *Iris* data points. The displayed variables x, y, and z are sepal length, petal length and width, respectively, with the sepal width variable sliced at $t = 3.4$ cm. The blue $\alpha = 4\%$ contour reveals only two clusters, while the red $\alpha = 10\%$ contour reveals the three clusters.

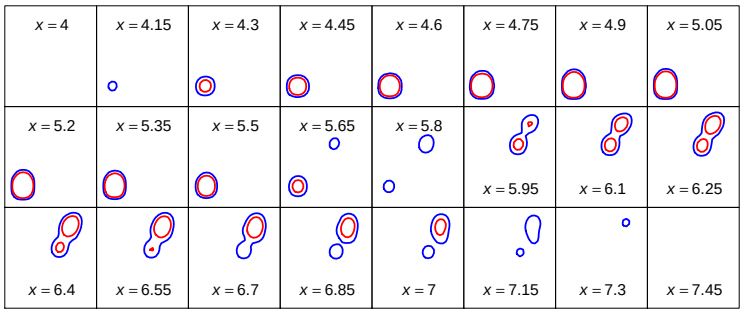

Figure 1.23 A detailed breakdown of the 3-D contours shown in Figure 1.22 taken from the ASH estimate $\hat{f}(x, y, z, t = 3.4)$ as the sepal length, x, ranges from 4.00 to 7.45 cm.

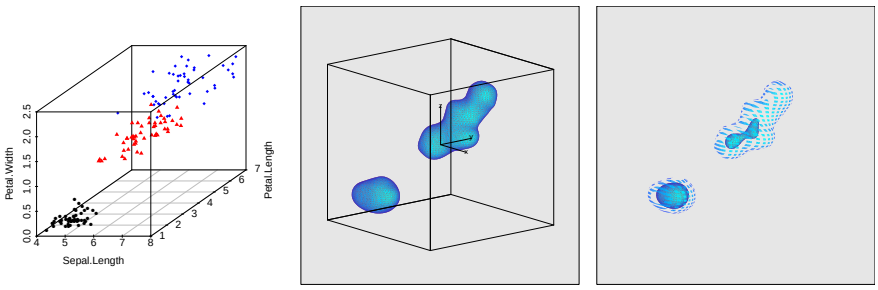

Figure 1.24 Analysis of 3 of the 4 *Iris* variables, omitting sepal width entirely, which should be compared to the slice shown in Figure 1.22. The middle contour ($\alpha = 0.17$) is superimposed upon the contour ($\alpha = 0.44$) in the right frame to help locate the shells.

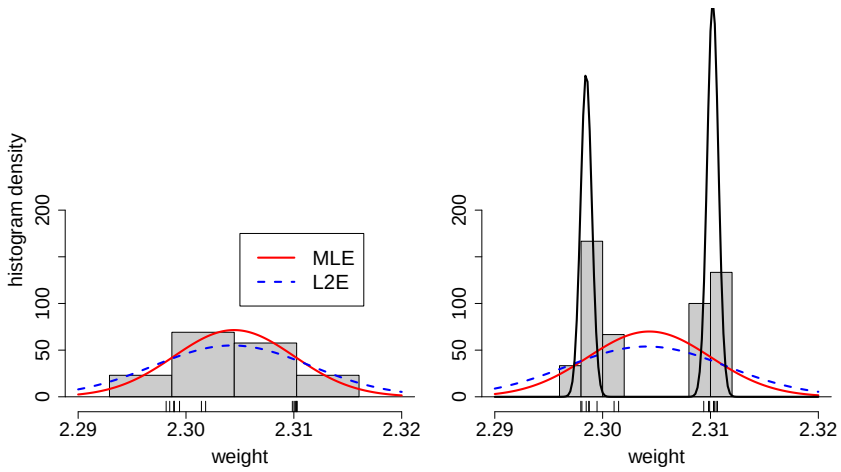

Figure 2.4 (Left frame) Histogram with MLE and L_2E normal fits to the Rayleigh data. (Right frame) L_2E normal mixture fit to blurred Rayleigh with common variance.

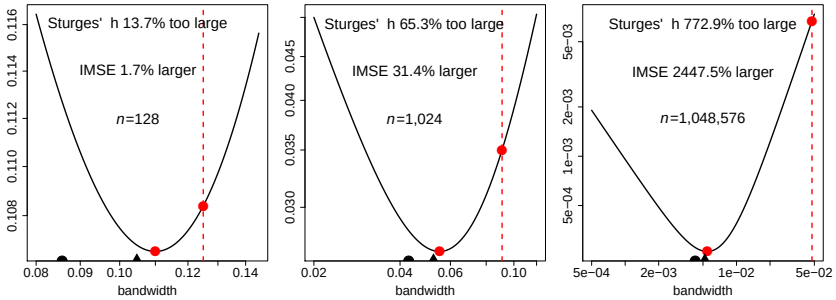

Figure 3.4 AMISE versus bandwidth for a $B(5,5)$ density. The best and Sturges' bandwidths are indicated by points on the curves. The Freedman-Diaconis and Scott reference bandwidths are shown as semicircular and triangular points, respectively, along the x-axis.

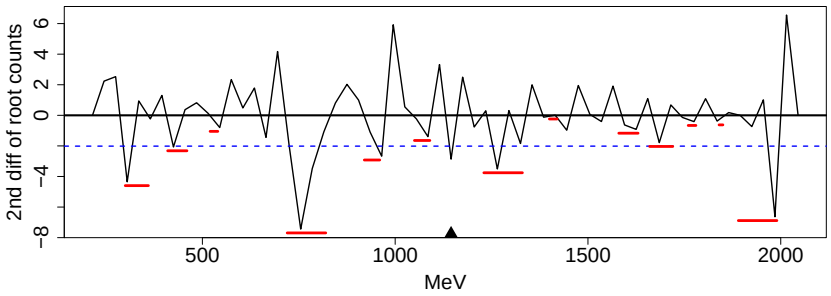

Figure 3.23 Second differences of square root of LRL histogram dataset with bin width of 30 MeV. The 13 bumps found by Good and Gaskins are indicated as before. The dashed line indicates the approximate 5% cutoff level for a bump to be significant.

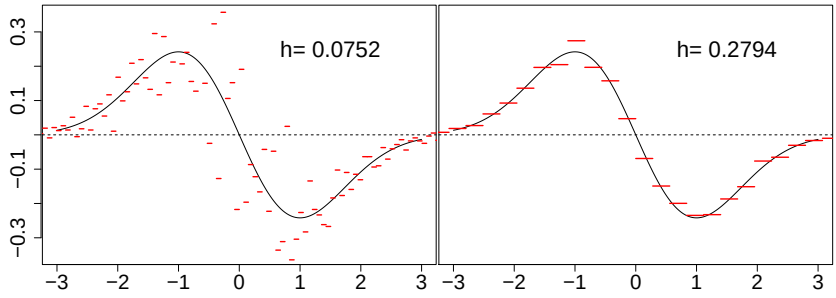

Figure 3.27 For a standard normal sample $n = 10^5$ points, comparison of the histogram "derivative" estimates using the optimal density and derivative bandwidths.

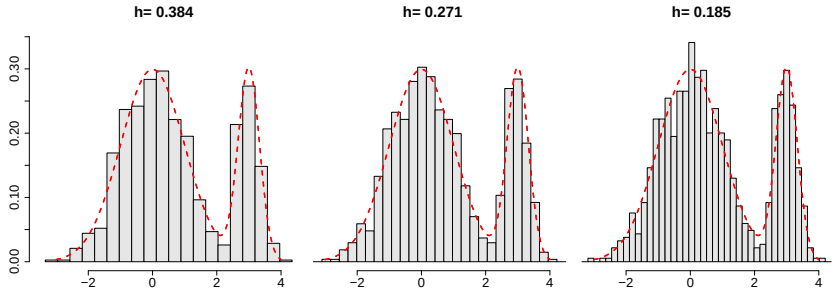

Figure 3.28 Three histograms of 1,000 points from the two-component mixture. The bandwidths (from left to right) are optimal for the left component, the mixture, and the right component, respectively.

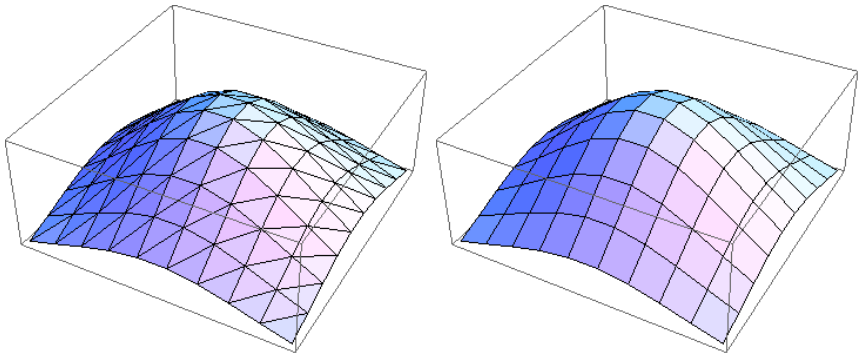

Figure 4.6 An example of the construction of a bivariate frequency polygon using triangular meshes (left) and linear blend elements(right).

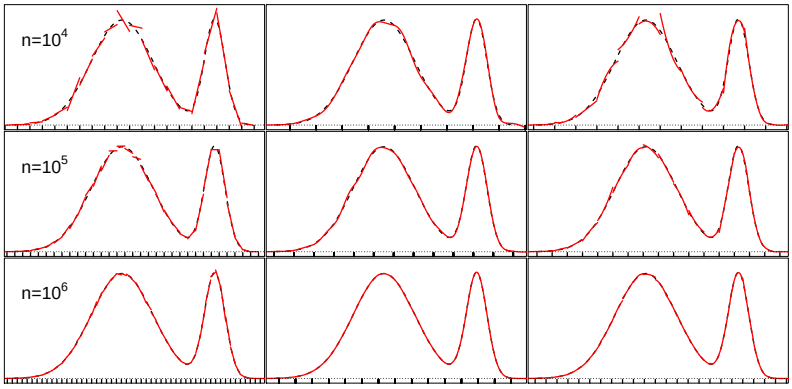

Figure 4.9 For three sample sizes from the mixture density, examples of the piecewise LPH, the continuous LPH, and the piecewise QPH estimates.

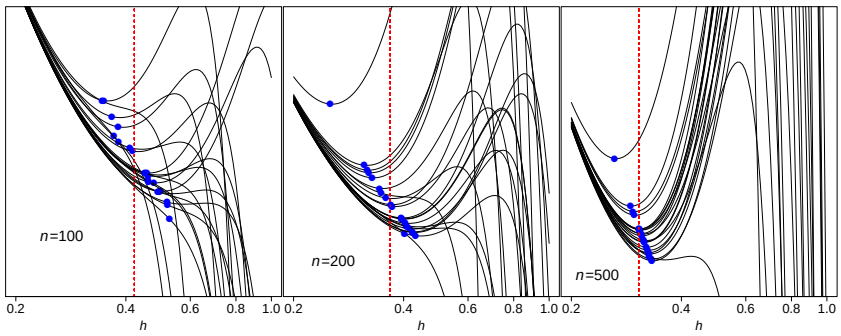

Figure 6.18 Twenty-one examples of the AMISE approximation of the plug-in rule with $N(0,1)$ data and a normal kernel. The plug-in bandwidth for each simulation is shown by the blue dot on the risk curve. The vertical dotted line indicates the normal reference rule (with $\sigma = 1$). Note that the horizontal axis is the same for each sample size, but the vertical scale (not labeled) zooms in on the relevant area.

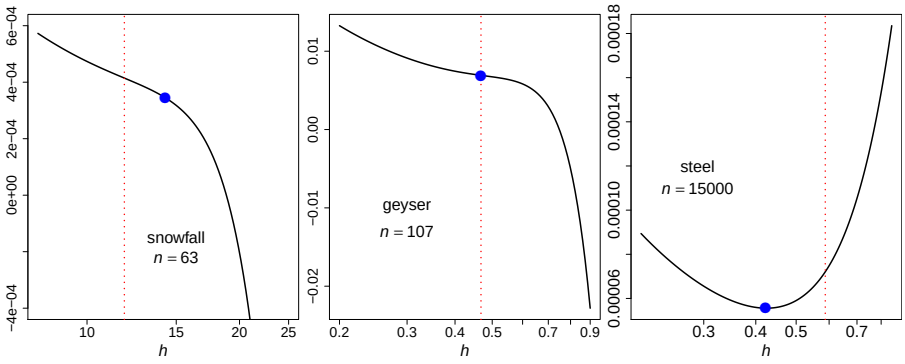

Figure 6.19 Plug-in cross-validation curves for the snowfall data ($n = 63$), the geyser dataset ($n = 107$) and the steel surface data ($n = 15{,}000$) for the normal kernel. The plug-in bandwidth obtained by formula (6.80) is indicated by the blue dot, and the oversmoothed bandwidth by the dashed red line.

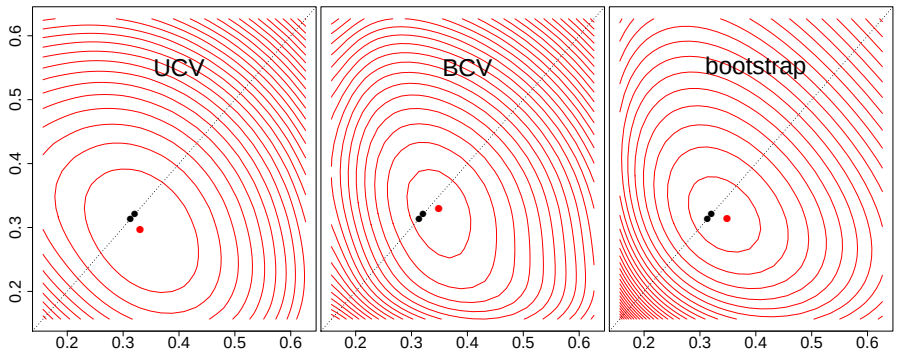

Figure 6.24 Estimated MISE(h_x, h_y) using UCV, BCV, and the bootstrap algorithms on 1,500 $N(\mathbf{0}_2, I_2)$ points. The two dots on each diagonal are h^* and the oversmoothed bandwidths. The dot locating the minimizer of each criterion is below the diagonal.

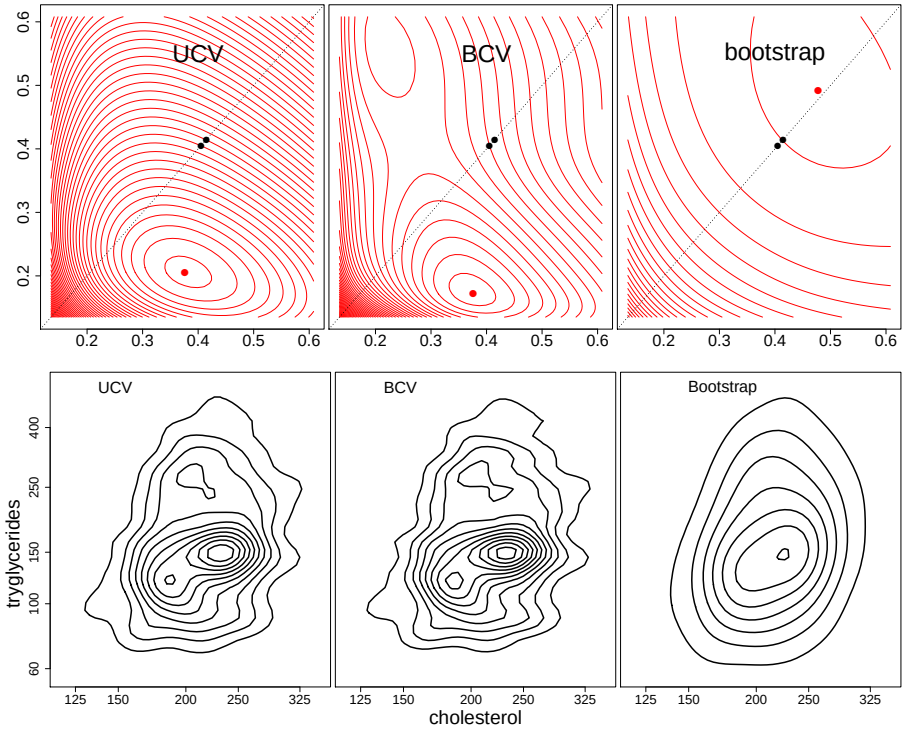

Figure 6.25 Same criterion as in Figure 6.24 for the standardized log lipid dataset ($n = 320$), together with the corresponding kernel estimates.

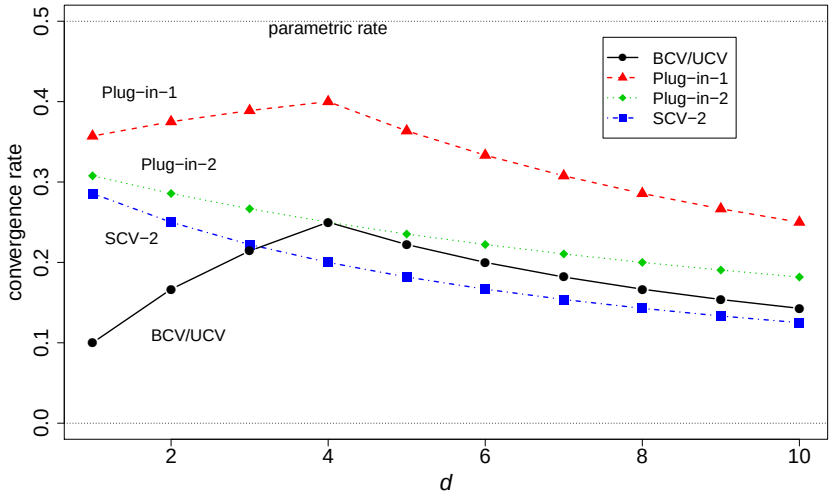

Figure 6.26 Magnitude of convergence rate exponents of several cross-validation algorithms. The best rate of $O(n^{-1/2})$ for parametric models would appear as $1/2$ on this graph.

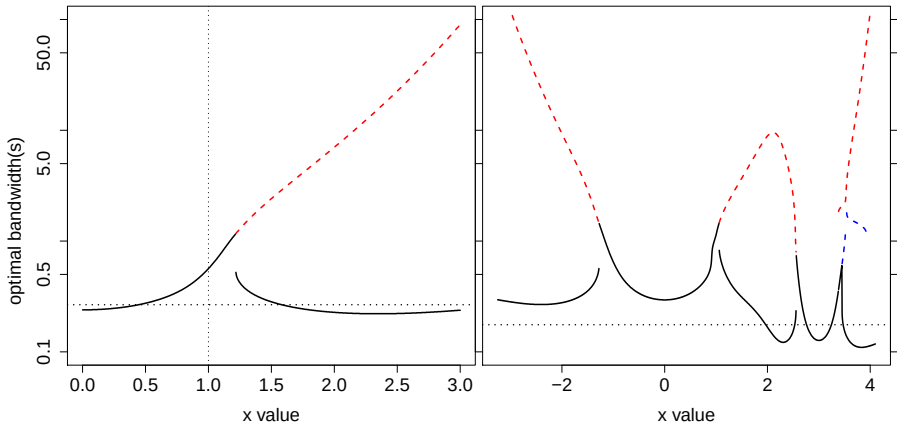

Figure 6.28 (Left) Asymptotically optimal bandwidths and zero-bias bandwidths for normal sample of size $n = 1000$. There are two optimal bandwidths when $x > 1.218$. The dashed line shows bandwidths close to zero-bias ones. (Right) Same for normal mixture but $n = 500$.

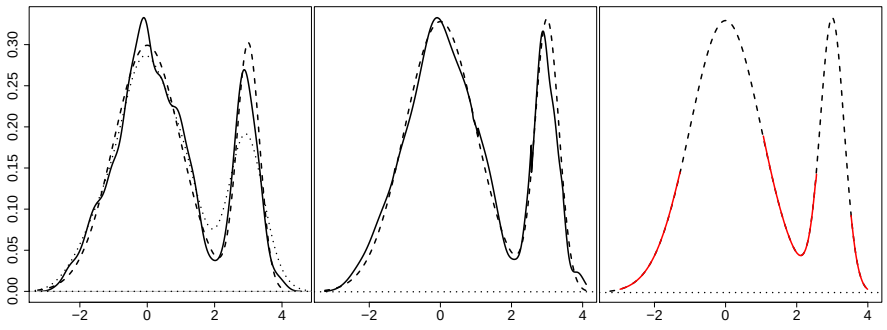

Figure 6.29 Several density estimates (solid lines) of a sample of size $n = 500$ from the mixture density in Equation (3.78), which is shown as a dashed line in each frame. (Left) A fixed kernel estimate with $h^* = 0.176$ and the normal reference rule $h = .473$ (dotted line). (Middle) Kernel estimate using $h^*(x)$, which is the smallest bandwidth in the right frame of Figure 6.28. (Right) Kernel estimate using $h_0^*(x)$ where it exists.

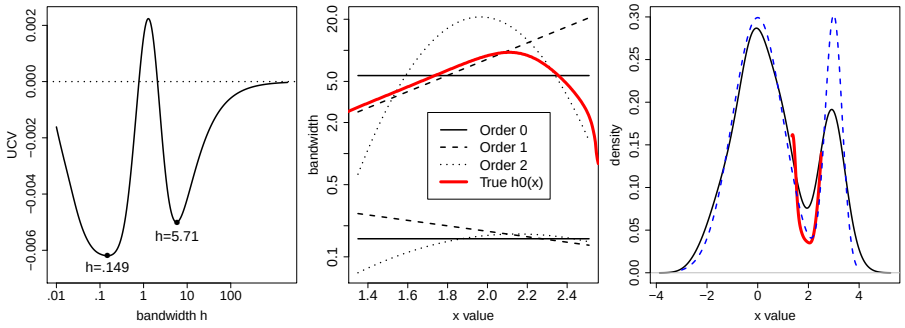

Figure 6.30 (Left) The UCV function over the interval $(a, b) = (1.35, 2.51)$ for the normal mixture dataset. (Middle) Log-polynomial fits to $h_0(x)$ that minimize UCV over (a, b). (Right) The zero-bias estimate over (a, b) together with the true mixture density and normal reference rule kernel estimate.

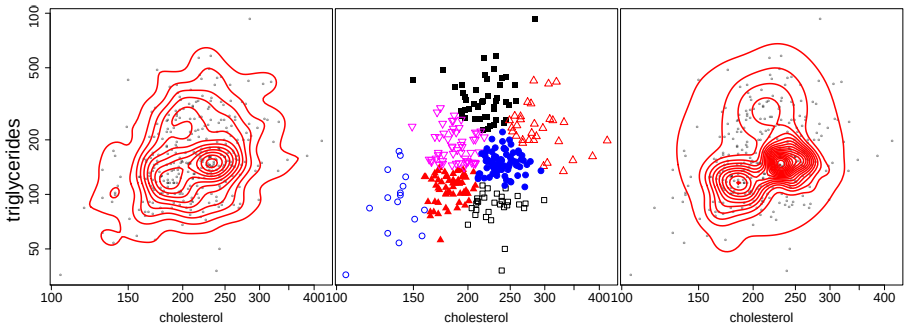

Figure 6.31 (Left) Twelve contours of the UCV-calibrated ($\hat{h} = 0.276$) bivariate Gaussian fixed-kernel estimate of the standardized log cholesterol and triglyceride data. (Middle) Seven clusters from k-means. (Right) The adaptive kernel estimator. The 7 bandwidths range from 0.174 to 2.36. The mode is 54% greater than in the left frame. The 19 contour levels are the same as in the left frame plus 7 more at higher levels.

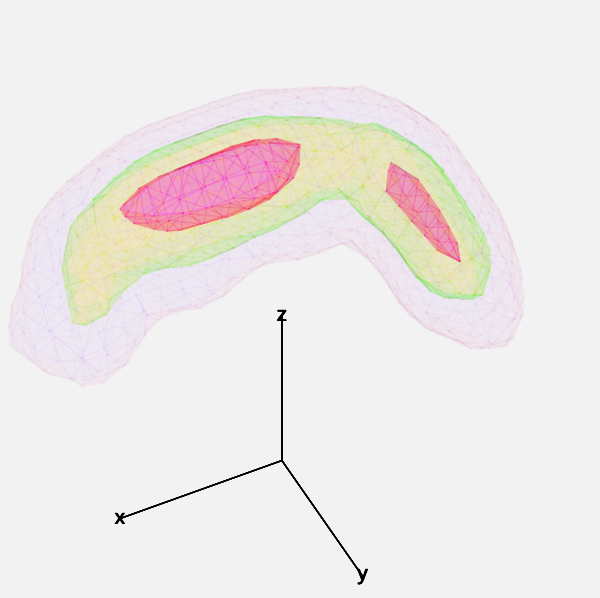

Figure 7.6 Three contour shells (α = 10%, 30%, and 60%) of a slice of the averaged shifted histogram of the four-dimensional PRIM4 data set with 500 points. These variables are heavily skewed, and the resulting density estimation problem more difficult.

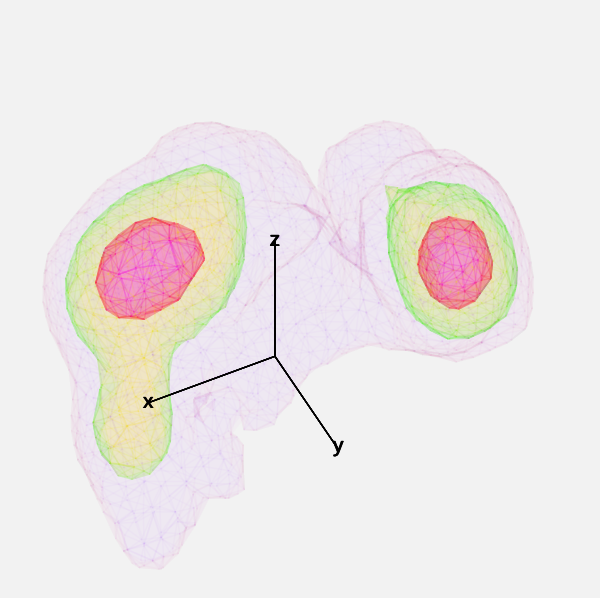

Figure 7.7 Three contour shell levels as in Figure 7.6 based on an ASH of the transformed PRIM4 data. The transformation was chosen to reduce skewness in each marginal variable. Such marginal transformations can greatly improve the quality of density estimation in multiple dimensions.

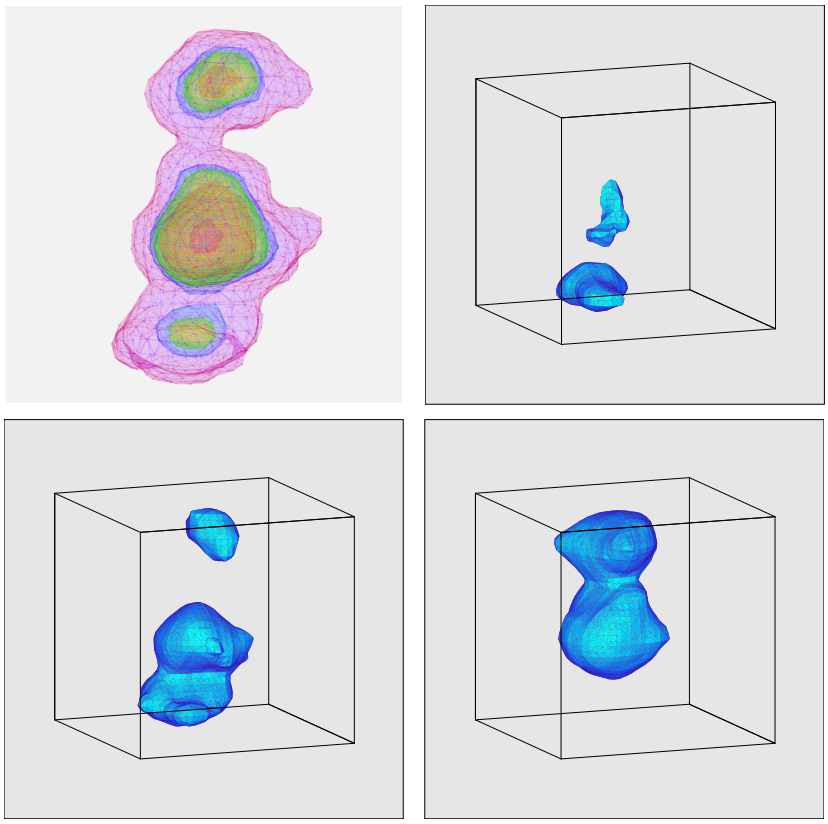

Figure 7.9 (Top left) ASH of the location of 510 earthquake epicenters and the transformed depths. The next three frames show the space-time ASH at approximately one week intervals leading up to the eruption (all at the same α level.)

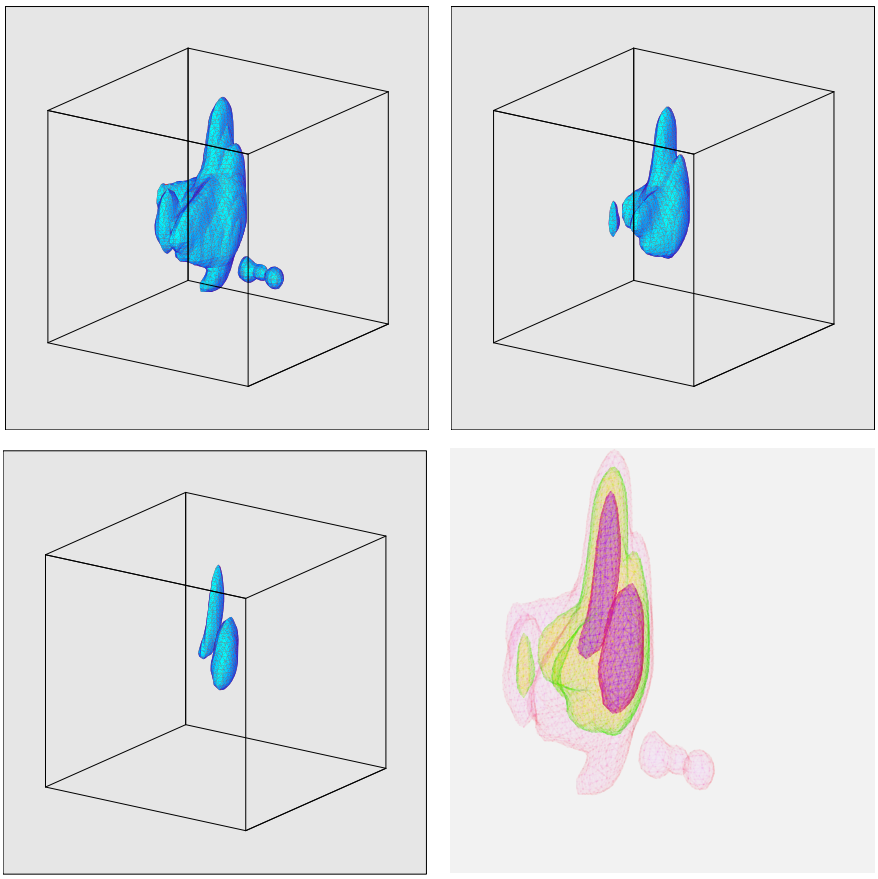

Figure 7.15 The $\alpha = 1\%$, 3.5%, and 15% level contours of a trivariate averaged shifted histogram of the Landsat data set of 22,932 points. The small disjoint second shell in the bottom right corner represents some of the outliers in the data set. The outliers resulted from singularities in the model-based data transformation algorithm from the original 24-dimensional Landsat data and were recorded at the minimum or maximum values. The final frame represents a composite of the first three frames using transparency. An examination of the crops being grown in this region reveals that the the tall cluster in the middle represents sunflowers; and the largest cluster on the right represents small grains including wheat, and the small small cluster on the far left represents sugar beets; see Section 1.4.3.

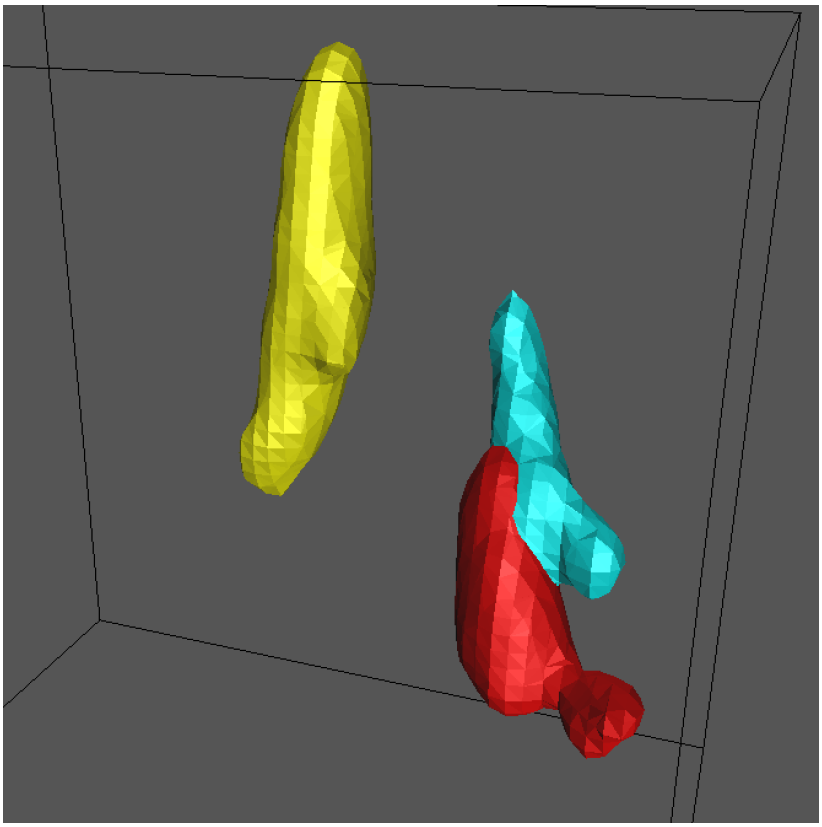

Figure 7.16 Median trivariate contours for sunflower ($n = 3694$, yellow), spring wheat ($n = 3811$, red), and barley ($n = 892$, cyan). The median contour contains 50% of the labelled data.

Figure 7.17 Median contours shown in Figure 7.16 with spring oats ($n = 459$, white), peanuts ($n = 304$, purple), soybeans ($n = 731$, magneta), and sugar beets ($n = 506$, green).

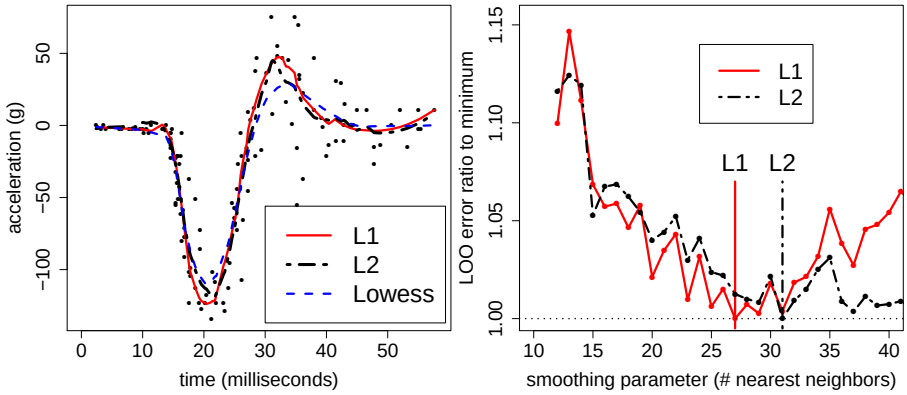

Figure 8.6 (Left) Local L_1 and L_2 quadratic fits to the motorcycle data, together with a LOWESS fit with $f = 0.25$. (Right) Normalized Leave-One-Out (LOO) cross-validation criteria, namely, the mean absolute error and the standard deviation for the L_1 and L_2 fits, respectively. The raw values range from $(16.2, 18.6)$ and $(18.5, 20.8)$, respectively. The best L_1 fit occurred with 27 points in each local neighborhood.

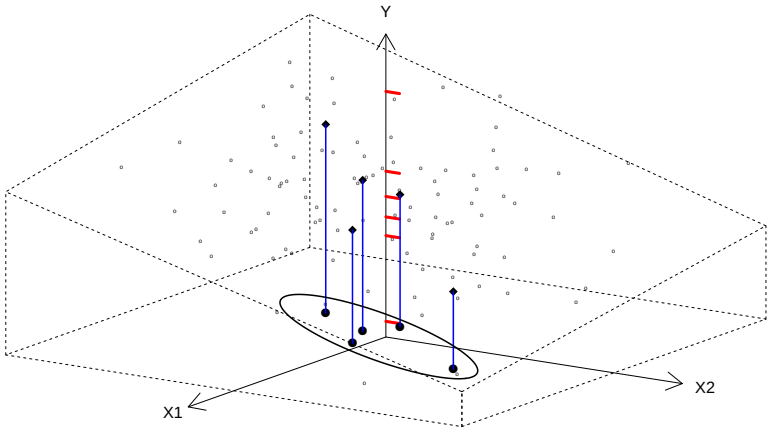

Figure 8.11 Example of the SIR dimension reduction technique.

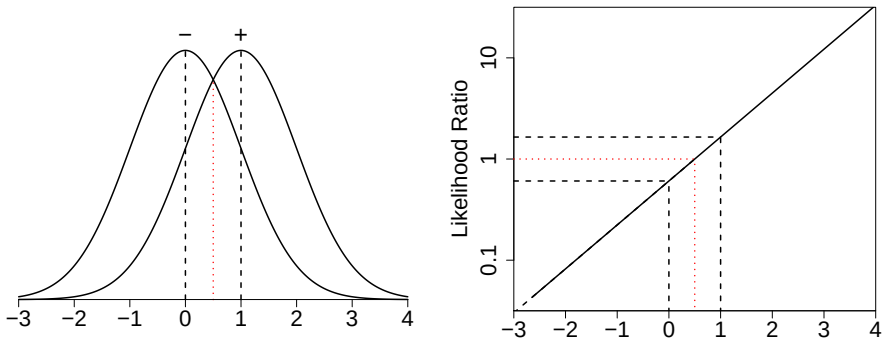

Figure 9.1 Shifted-normal model of two populations for a single risk factor or covariate. The log–likelihood ratio is linear in x.

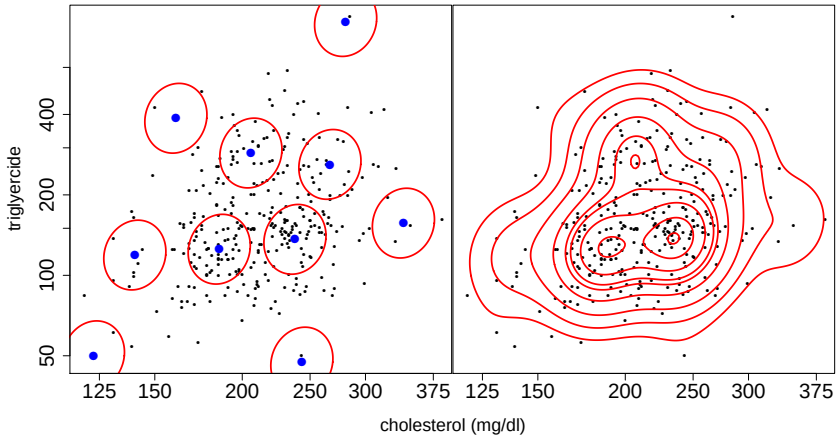

Figure 9.4 Mclust (2005 version) applied to log-lipid dataset ($n = 320$).

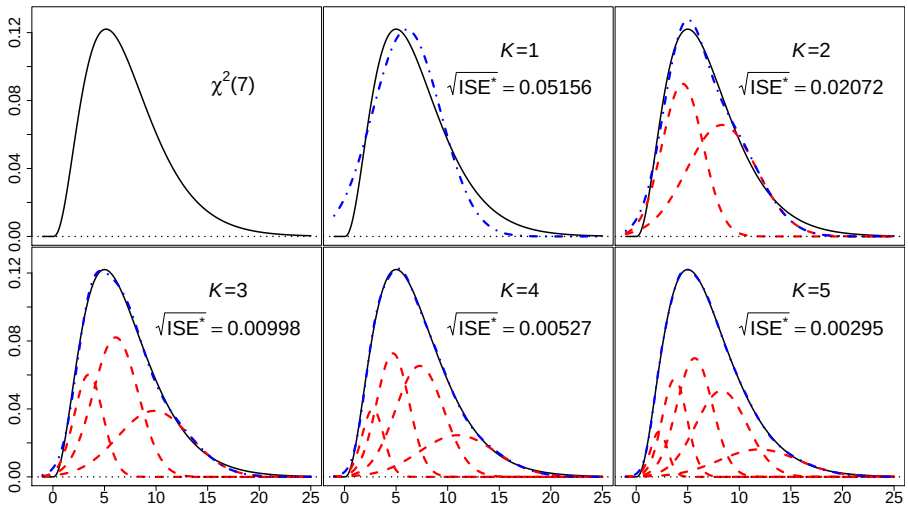

Figure 9.5 Numerically best Gaussian mixture approximations to a $\chi^2(7)$ density for $1 \leq K \leq 5$. The root optimal ISE is shown in each figure.

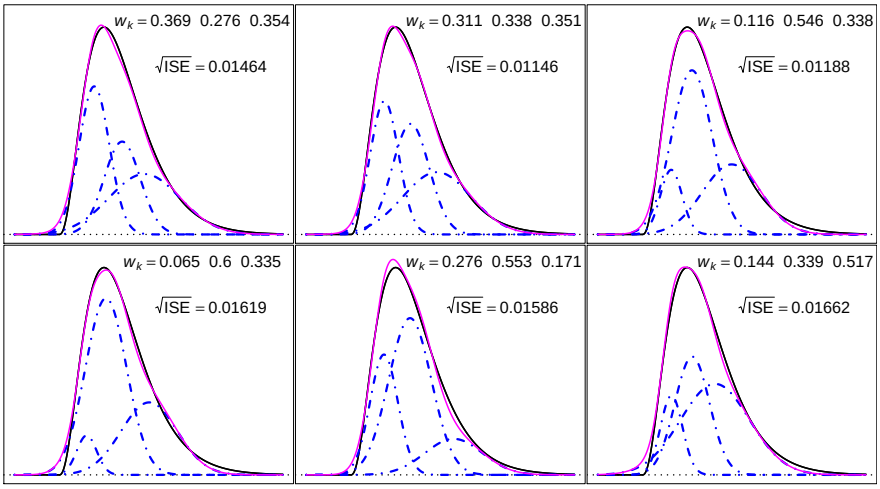

Figure 9.6 Six examples of nearby mixture solutions when $K = 3$ that are within a one-sigma confidence hyperellipse. The weights and criterion value are shown in each frame.

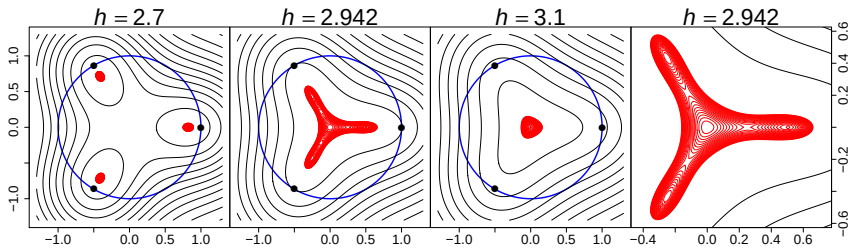

Figure 9.8 Contours of a bivariate Gaussian kernel density estimator with $n = 3$ points (black dots) on the unit circle forming an equilateral triangle. A highly nonlinear set of contour levels are displayed, so that the contours near the modes are emphasized.

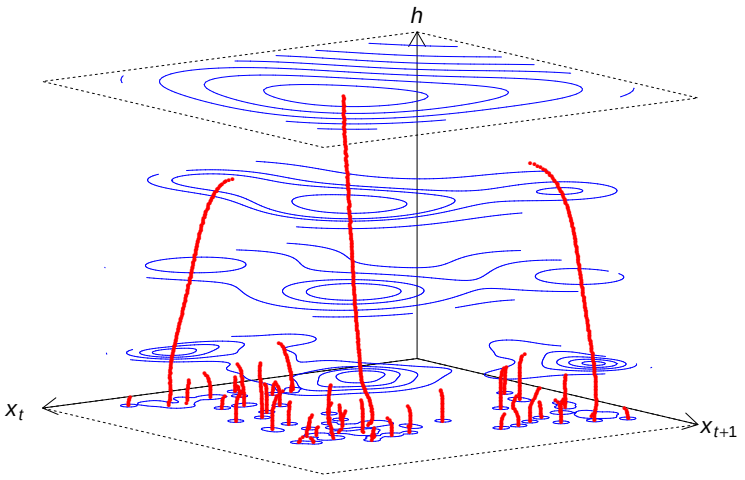

Figure 9.9 Bivariate mode tree of the lagged geyser dataset. Contours of the Gaussian kernel estimates are shown for $h \in (.05, .20, .45, .70, 1.00)$.

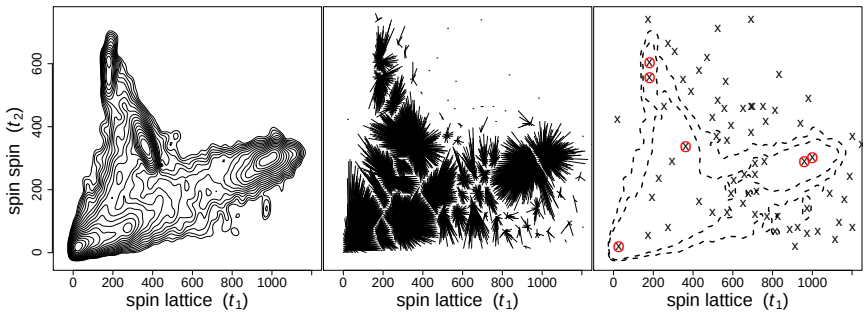

Figure 9.11 (Left) ASH contours of (t_1, t_2) of an MRI image with 24,476 pixels. (Center) Hill-climbing of individual pixel values to the nearest mode. (Right) The 70 modes found are superimposed on two contours.

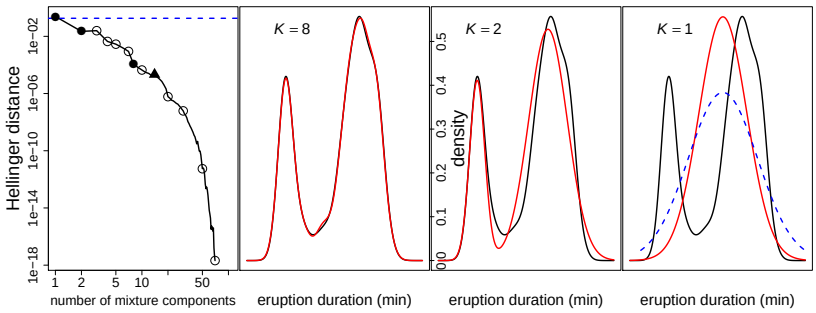

Figure 9.13 IPRA of the geyser duration dataset ($n = 107$). The smoothing parameter for the Gaussian kernel estimate is $h = 0.20$. The MLE normal fit and its Helling distance are shown as dashed lines in the fourth and first frames, respectively.

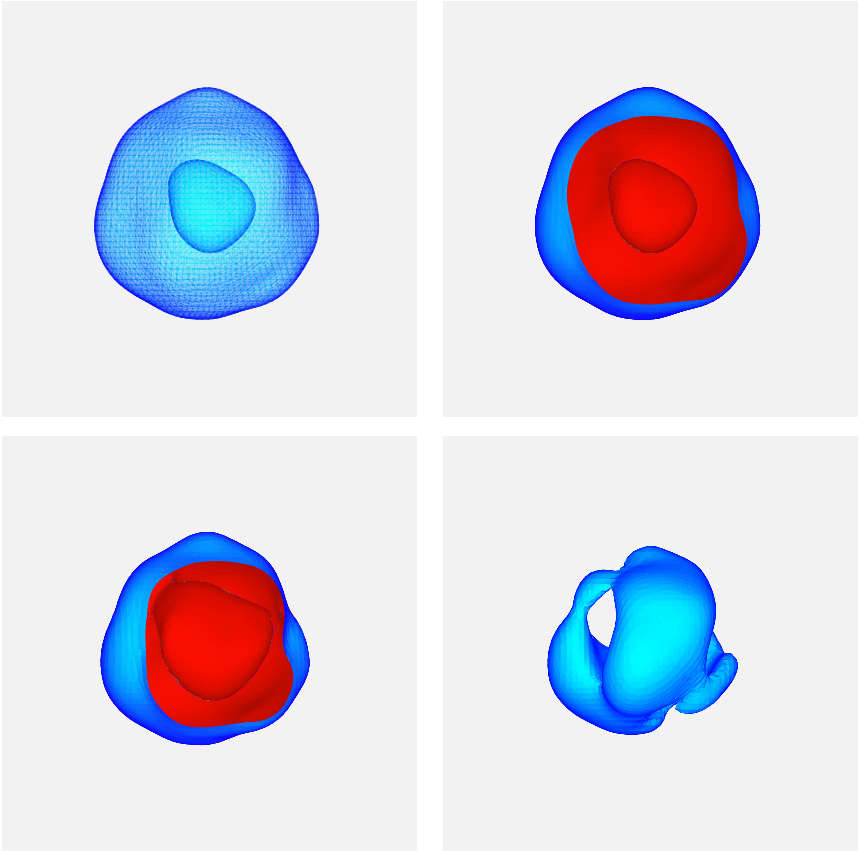

Figure 9.17 Contour shells derived from averaged shifted histogram estimate of a pseudo-random sample of 5,000 points from a trivariate density with a "hole" in the middle. (First frame) The single α-level contour displayed is a pair of nested and nearly spherical shells. At values of α lower than shown, the inner shell shrinks and then vanishes. (Second frame) Same as the first frame, with the outer shell pealed away. The contour surfaces colored red point to the higher density regions. In this case the higher region is between the nested shells. Theoretically, the mode is a sphere, although finite samples will not achieve this exactly. (Third frame) At a slightly higher α-level, the outer shell has shrunk and the inner shell has merged, and in fact, they have merged in the back. (Fourth frame) At an even higher α-level, the shells have broken apart, although theoretically they should be converging to a sphere (the mode).