Multivariate Density Estimation and Visualization

David W. Scott[1]

Rice University, Department of Statistics, MS-138, Houston, TX 77005-1892 USA
scottdw@rice.edu, http://www.stat.rice.edu/∼scottdw

1 Introduction

This chapter examines the use of flexible methods to approximate an unknown density function, and techniques appropriate for visualization of densities in up to four dimensions. The statistical analysis of data is a multilayered endeavor. Data must be carefully examined and cleaned to avoid spurious findings. A preliminary examination of data by graphical means is useful for this purpose. Graphical exploration of data was popularized by **?** in his book on exploratory data analysis (EDA). Modern data mining packages also include an array of graphical tools such as the histogram, which is the simplest example of a density estimator. Exploring data is particularly challenging when the sample size is massive or if the number of variables exceeds a handful. In either situation, the use of nonparametric density estimation can aid in the fundamental goal of understanding the important features hidden in the data. In the following sections, the algorithms and theory of nonparametric density estimation will be described, as well as descriptions of the visualization of multivariate data and density estimates. For simplicity, the discussion will assume the data and functions are continuous. Extensions to discrete and mixed data are straightforward.

Statistical modeling of data has two general purposes: (1) understanding the shape and features of data through the density function, $f(\mathbf{x})$, and (2) prediction of y through the joint density function, $f(\mathbf{x}, y)$. When the experimental setting is well-known, parametric models may be formulated. For example, if the data are multivariate normal, $N(\mu, \Sigma)$, then the features of the density may be extracted from the maximum likelihood estimates of the parameters μ and Σ. In particular, such data have one feature, which is a single mode located at μ. The shape of the data cloud is elliptical, and the eigenvalues and eigenvectors of the covariance matrix, Σ, indicate the orientation of the data and the spread in those directions. If the experimental setting is not well-known, or if the data do not appear to follow a parsimonious parametric form, then nonparametric density estimation is indicated. The major features of the density may be found by counting and locating the sample modes. The shape of the density cannot easily be determined algebraically, but visualization methodology can assist in this task. Similar remarks apply in the regression setting.

When should parametric methods be used and when should nonparametric methods be used? A parametric model enjoys the advantages of well-known properties and parameters which may be interpreted. However, using parametric methods to explore data makes little sense. The features and shape of a normal fit will always be the same no matter how far from normal the data may be. Nonparametric approaches can fit an almost limitless number of density functional forms. However, at the model, parametric methods are always more statistically accurate than the corresponding nonparametric estimates. This statement can be made more precise by noting that parametric estimators tend to have lower variance, but are susceptible to substantial bias when the wrong parametric form is invoked. Nonparametric methods are not unbiased, but the bias asymptotically vanishes for any continuous target function. Nonparametric algorithms generally have greater variance than a parametric algorithm. Construction of optimal nonparametric estimates requires a data-based approach in order to balance the variance and the bias, and the resulting mean squared error generally converges at a rate slower than the parametric rate of $O(n^{-1})$. In summary, nonparametric approaches are always appropriate for exploratory purposes, and should be used if the data do not follow a simple parametric form.

2 Visualization

2.1 Data Visualization

Visualization of data is a fundamental task in modern statistical practice. The most common figure for this purpose is the bivariate scatter diagram. Figure 1(a) displays the levels of blood fats in a sample of men with heart disease. The data have been transformed to a logarithm base 10 scale to minimize the effects of skewness. At a first glance, the data appear to follow a bivariate normal distribution. The sample correlation is only 0.22. One might examine each of the two variables separately as a univariate scatter diagram, which is commonly referred to as a "dot plot," but such figures are rarely presented. Tukey advocated the histogram-like stem-and-leaf plot or the box-and-whiskers plot, which displays simple summaries including the median and quartiles. Figure 1(b) displays box-and-whisker plots for these variables. Clearly triglyceride values vary more than cholesterol and may still be right-skewed.

As shown later in Section 3.3, there may be rather subtle clusters within these data. The eye can readily detect clusters which are well-separated, but the eye is not reliable when the clusters are not well-separated, nor when the sample size is so large that the scatter diagram is too crowded. For example,

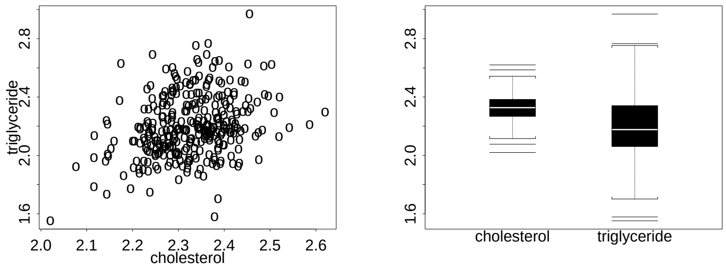

Fig. 1. Cholesterol and triglyceride blood levels for 320 males with heart disease.

consider the Old Faithful Geyser data (**?**), (x_t, y_t), where x_t measures the waiting time between successive eruptions of the geyser, and y_t measures the duration of the subsequent eruption. The data were blurred by adding uniform noise to the nearest minute for x_t and to the nearest second for y_t. Figure 2 displays histograms of these two variables. Interestingly, neither appears to follow the normal distribution. The common feature of interest is the appearance of two modes. One group of eruptions is only 2 minutes in duration, while the other averages over 4 minutes in duration. Likewise, the waiting time between eruptions clusters into two groups, one less than an hour and the other greater than one hour. The distribution of eruption durations appears to be a mixture of two normal densities, but the distribution of the waiting times appears more complicated.

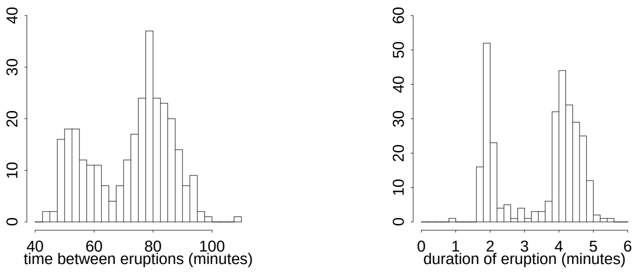

Fig. 2. Waiting time and duration of 299 consecutive eruptions of the Old Faithful Geyser.

Finally, in Figure 3 we examine the scatter diagrams of both (x_t, y_t) as well as the lagged values of eruption duration, (y_{t-1}, y_t). The common feature in these two densities is the presence of three modes. As mentioned earlier, the eye is well-suited to discerning clusters that are well-separated. From Figure 3(a), short waiting periods are associated with long eruption durations. From Figure 3(b), all eruptions of short duration are followed by eruptions of long duration. Missing from Figure 3(b) are any examples of eruptions of short duration following eruptions of short duration, which should be a plus for the disappointed tourist. The observant reader may notice an odd clustering of points at integer values of the eruption duration. A quick count shows that 23, 2, and 53 of the original 299 values occur exactly at $y = 2$, 3, and 4 minutes, respectively. Examining the original time sequence suggests that these measurements occur in clumps; perhaps accurate measurements were not taken after dark. Exploration of these data has revealed not only interesting features but also suggest possible data collection anomalies.

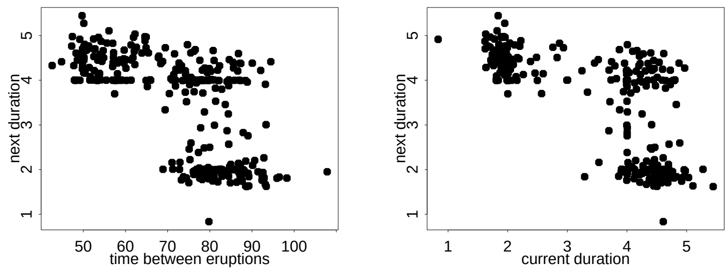

Fig. 3. Two scatter diagrams of the Old Faithful Geyser data.

Massive datasets present different challenges. For example, the Landsat IV remote sensing dataset described by **?** contains information on 22,932 pixels of a scene imaged in 1977 from North Dakota. The variables displayed in Figure 4 are the time of peak greenness of the crop in each pixel and the derived value of the maximum greenness, scaled to values 0-255 and blurred with uniform noise. Overplotting is apparent. Each successive figure drills down into the boxed region shown. Only 5.6% of the points are eliminated going to the second frame; 35.5% eliminated between the second and third frames; and 38.1% between the third and final frames, still leaving 8,624 points. Overplotting is still apparent in the final frame. Generally, gleaning detailed density information from scatter diagrams is difficult at best. Nonparametric density estimation solves this problem.

To see the difficulty of gleaning density information from the graphs in Figure 4, compare the bivariate histogram displayed in Figure 5 for the data in frame (b) from Figure 4. Using only the scatter diagram, there is no way to know the relative frequency of data in the two largest clusters except through the histogram.

The bivariate histogram uses rectangular-shaped bins. An interesting hybrid solution is to use hexagonal-shaped bins and to use a glyph to represent the bin count. **?** compared the statistical power of using squares, hexagons, and equilateral triangles as shapes for bins of bivariate histograms and concluded that hexagons were the best choice. **?** examined the use of drawing a glyph in each bivariate bin rather than the perspective view. For graphical

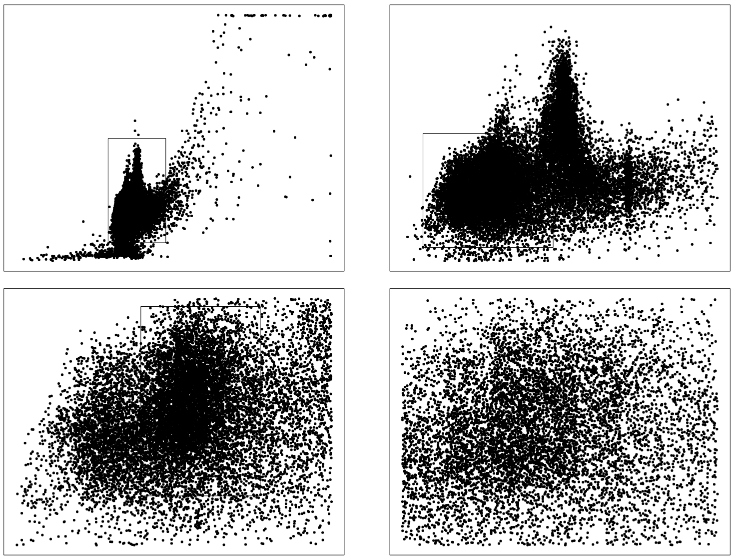

Fig. 4. Drilling into the Landsat IV data with $n = 22{,}932$.

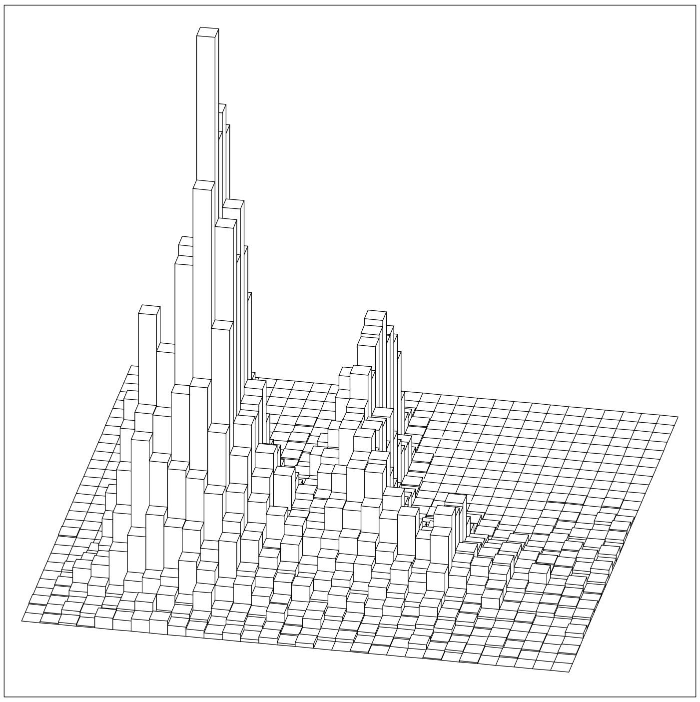

Fig. 5. Histogram of data in Figure 4(b)

reasons, Carr found hexagonal bins were more effective. The bin count is represented by a hexagonal glyph whose area is proportional to the bin count. Figure 6 displays the hexagonal mosaic map of the same data as in Figure 5. This representation gives a quite accurate summary of the density information. No bin counts are obscured as in the perspective view of the bivariate histogram.

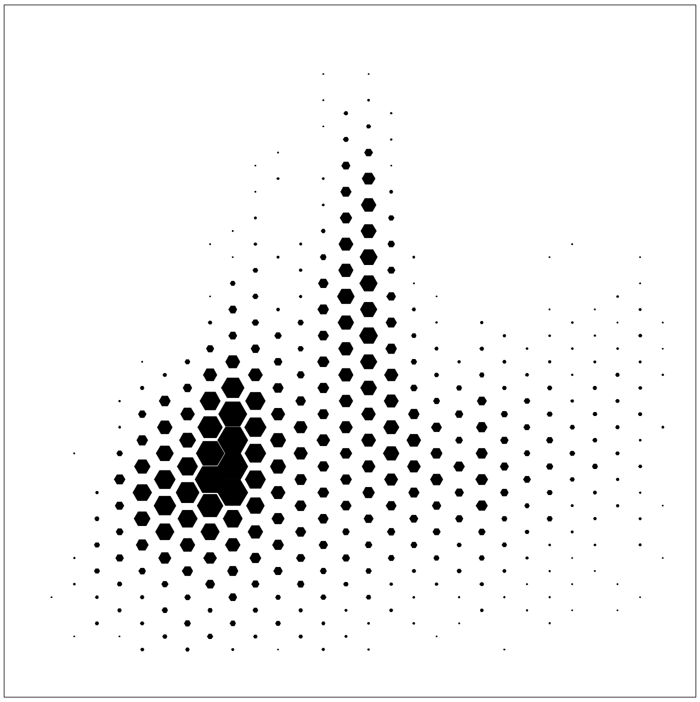

Fig. 6. Hexagonal bin glyph of the data in Figure 4(b)

In the next section, some of the algorithms for nonparametric density estimation and their theoretical properties are discussed. We then return to the visualization of data in higher dimensions.

3 Density Estimation Algorithms and Theory

This section includes enough algorithms and results to obtain a basic understanding of the opportunities and issues. Fortunately, there have been a number of readable monographs available for the reader interested in pursuing this subject in depth. In rough chronological order, excluding books primarily dealing with nonparametric regression, the list includes **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, and **?**.

The purpose of the next section is to provide a survey of important results without delving into the theoretical underpinnings and details. The references cited above are well-suited for that purpose.

3.1 A High-Level View of Density Theory

Smoothing Parameters

Every algorithm for nonparametric density estimation has one or more design parameters which are called the smoothing parameter(s) or bandwidth(s) of the procedure. The smoothing parameter controls the final appearance of the estimate. For an equally-spaced histogram, the bin width plays the primary role of a smoothing parameter. Of course, the bins may be shifted and the location of the bin edges is controlled by the bin origin, which plays the role of a secondary smoothing parameter. For a kernel estimator, the scale or width of the kernel serves as the smoothing parameter. For an orthogonal series estimator, the number of basis functions serves as the smoothing parameter. The smoothing parameters of a spline estimator also include the location of the knots. Similarly, a histogram with completely flexible bin widths has many more than two smoothing parameters.

No Unbiased Density Estimators

As a point estimator of $f(x)$, **?** proved that every nonparametric density estimator, $\hat{f}(x)$ is biased. However, it is usually true that the integral of all of the pointwise biases is 0. Thus mean squared error (MSE) and integrated mean squared error (IMSE) are the appropriate criteria to optimize the tradeoff between pointwise/integrated variance and squared bias.

Nonparametric density estimators always underestimate peaks and overestimate valleys in the true density function. Intuitively, the bias is driven by the degree of curvature in the true density. However, since the bias function is continuous and integrates to 0, there must be a few points where the bias does vanish. In fact, letting the smoothing parameter vary pointwise, there are entire intervals where the bias vanishes, including the difficult-to-estimate tail region. This fact has been studied by **?** and **?**. Since the bias of a kernel estimator does not depend on the sample size, these zero-bias or bias-annihilating estimates have more than a theoretical interest. However, there is much more work required for practical application. Alternatively, in higher dimensions away from peaks and valleys, one can annihilate pointwise bias by balancing directions of positive curvature against directions of negative curvature; see **?**. An even more intriguing idea literally adjusts the raw data points towards peaks and away from valleys to reduce bias; see **?**.

Rates of Convergence

The rate of convergence of a nonparametric density estimator to the true density is much slower than in the parametric setting, assuming in the latter case that the correct parametric model is known. If the correct parametric model is not known, then the parametric estimates will converge but the bias will not vanish. The convergence is slower still in high dimensions, a fact which is often referred to as the *curse of dimensionality.* Estimating the derivative of a density function is even harder than coping with an additional dimension of data.

If the k-th derivative of a density is known to be smooth, then it is theoretically possible to construct an order-k nonparametric density estimation algorithm. The pointwise bias is driven by the k-th derivative at x, $f^{(k)}(x)$. However, if $k > 2$, then the density estimate will take on negative values for some points, x. It is possible to define higher-order algorithms which are non-negative, but these estimates do not integrate to 1; see **?**. Thus higher-order density estimation algorithms violate one of the two conditions for a true density: $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Of course, there are cases where the first condition is violated for lower-order estimators. Two such cases include orthogonal series estimators (**??**) and boundary-corrected kernel estimators (**?**). Note that positive kernel estimators correspond to $k = 2$. **?** studied the efficacy of higher-order procedures and suggested $k = 3$ often provided superior estimates. **?** also studied this question and found some improvement when $k = 3$, which must be traded off against the disadvantages of negative estimates.

Choosing Bandwidths in Practice

Picking the best smoothing parameter from data is an important task in practice. If the smoothing parameter is too small, the estimate is too noisy, exhibiting high various and extraneous wiggles. If the smoothing parameter is too large, then the estimate may miss key features due to oversmoothing, washing out small details. Such estimates have low variance but high bias in many regions. In practice, bandwidths that do not differ by more than 10-15% from the optimal bandwidth are usually satisfactory.

A statistician experienced in EDA is likely to find all estimates informative for bandwidths ranging from undersmoothed to oversmoothed. With a complicated density function, no single choice for the bandwidth may properly represent the density for all values of x. Thus the same bandwidth may result in undersmoothing for some intervals of x, oversmoothing in another interval, and yet near optimal smoothing elsewhere. However, the practical difficulty of constructing locally adaptive estimators makes the single-bandwidth case of most importance. Simple transformations of the data scale can often be an effective strategy (**?**). This strategy was used with the lipid data in Figure 1, which were transformed to a $\log_{10}$ scale.

Consider the 21,640 x points shown in frame (b) of Figure 4. Histograms of these data with various numbers of bins are shown in Figure 7. With so much data, the oversmoothed histogram Figure 7(a) captures the major features, but seems biased downwards at the peaks. The final frame shows a histogram that is more useful for finding data anomalies than as a good density estimate.

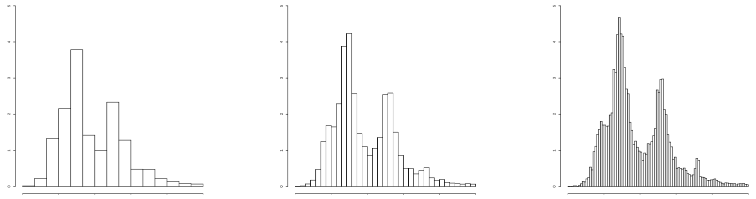

Fig. 7. Histograms of x variable in Figure 4(b) with 15, 35, and 100 bins.

The differences are more apparent with a smaller sample size. Consider the 320 $\log_{10}$-cholesterol levels shown in Figure 1. Three histograms are shown in Figure 8. The extra one or two modes are at least suggested in the middle panel, while the histogram in the first panel only suggests a rather unusual non-normal appearance. The third panel has many large spurious peaks. We conclude from these two figures that while an oversmoothed estimator may have a large error relative to the optimal estimator, the absolute error may still be reasonably small for very large data samples.

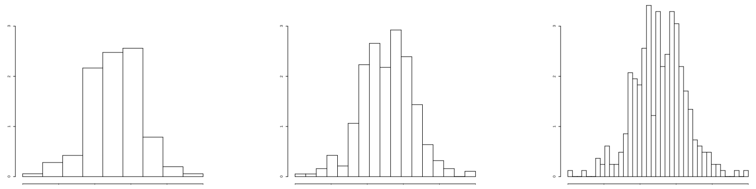

Fig. 8. Histograms of $\log_{10}$-cholesterol variable in Figure 1 with 9, 19, and 39 bins.

Oversmoothed Bandwidths

While there is no limit on how complicated a density may be (for which $\int f^{(k)}(x)^2\,dx$ may grow without bound), the converse is not true. **?** and **?** show that for a particular scale of a density (for example, the range, standard deviation, or interquartile range), there is in fact a lower bound among continuous densities for the roughness quantity

$$R(f^{(k)}) = \int_{-\infty}^{\infty} f^{(k)}(x)^2\,dx. \tag{1}$$

In a real sense, such densities are the smoothest possible and are the easiest to estimate. The optimal bandwidth for these "oversmoothed densities" serves as an upper bound. Specifically, any other density with the same scale will have more complicated structure and will require a smaller bandwidth to more accurately estimate those features. Since oversmoothed bandwidths (and reference bandwidths as well) only use the data to estimate the scale (variance, for example), these data-based estimates are quite stable. Obtaining similar highly stable data-based nearly optimal bandwidth estimators requires very sophisticated estimates of the roughness function given in Equation 1. One algorithm by **?** is often highly rated in practice (**?**). **?** noted the small-sample behavior of the algorithm seemed closely related to the oversmoothed bandwidths. These approaches all rely on asymptotic expansions of IMSE rather than an unbiased risk estimate, which underlies the least-squares or unbiased cross-validation algorithm introduced by **?** and **?**. However, the unbiased risk approach has numerous extensions; see **?** and **?**. Another algorithm that should be mentioned is the bootstrap bandwidth. For a Gaussian kernel, the bootstrap with an infinite number of repetitions has a closed form expression; see **?**. Multivariate extensions are discussed by **?**.

Many details of these ideas may be found in the literature and in the many textbooks available. The following section provides some indication of this research.

3.2 The Theory of Histograms

The basic results of density estimation are perhaps most easily understood with the ordinary histogram. Thus more time will be spent on the histogram with only an outline of results for more sophisticated and more modern algorithms.

Given an equally spaced mesh $\{t_k\}$ over the entire real line with $t_{k+1} - t_k = h$, the density histogram is given by

$$\hat{f}(x) = \frac{\nu_k}{nh} \quad \text{for } t_k < x < t_{k+1}, \tag{2}$$

where ν_k is the number of data points falling in the k-th bin. Clearly, $\sum_k \nu_k = n$ and ν_k is a Binomial random variable with mean $p_k = \int_{t_k}^{t_{k+1}} f(x)\,dx$; hence, $\mathrm{E}\,\nu_k = np_k$ and $\mathrm{Var}\,\nu_k = np_k(1-p_k)$. Thus the pointwise variance of the histogram (2) is $np_k(1-p_k)/(nh)^2$, which is constant for all x in the k-th bin. Thus, the integrated variance (IV) over $(-\infty, \infty)$ may be found by integrating the pointwise variance over the k-th bin (i.e., multiply by the bin width h), and summing over all bins:

$$\mathrm{IV} = \sum_{k=-\infty}^{\infty} \frac{np_k(1-p_k)}{(nh)^2} \times h = \sum_{k=-\infty}^{\infty} \frac{p_k(1-p_k)}{nh} = \frac{1}{nh} - \sum_k \frac{pk^2}{nh}\,, \tag{3}$$

since $\sum p_k = \int_{-\infty}^{\infty} f(x)\,dx = 1$. The final term may be shown to approximate $n^{-1} \int f(x)^2\,dx$, which is asymptotically negligible. Thus the global integrated variance of the histogram can be controlled by collecting more data or choosing a wider bin width.

Next consider the bias of $\hat{f}$ at a fixed point, x, which is located in the k-th bin. Note that $\mathrm{E}\,\hat{f}(x) = np_k/nh = p_k/h$. A useful approximation to the bin probability is

$$p_k = \int_{t_k}^{t_{k+1}} f(y)\,dy = h\,f(x) + h^2 \left(\frac{1}{2} - \frac{x-t_k}{h}\right) f'(x) + \cdots\,, \tag{4}$$

replacing the integrand $f(y)$ by $f(x) + (y-x)f'(x) + \cdots$. Thus the pointwise bias may be approximated by

$$\mathrm{Bias}\,\hat{f}(x) = \mathrm{E}\,\hat{f}(x) - f(x) = \frac{p_k}{h} - f(x) = h\left(\frac{1}{2} - \frac{x-t_k}{h}\right) f'(x) + \cdots\,. \tag{5}$$

Therefore, the bias is controlled by the first derivative of the unknown density at x. Since $t_k < x < t_{k+1}$, then the factor $(1/2 - (x-t_k)/h)$ in Equation 5 varies from $-1/2$ to $1/2$. Thus the bias is also directly proportional to the bandwidth, h. To control the bias of the histogram estimate, the bandwidth h should be small. Comparing Equations (3) and (5), the global consistency of the histogram can be guaranteed if, as $n \to \infty$, $h \to 0$ while ensuring that the product $nh \to \infty$ as well, for example, if $h = 1/\sqrt{n}$ (see **?**).

A more complete analysis of the bias (**?**) shows that the integrated squared bias is approximately $h^2\,\mathrm{R}(f')/12$, where $\mathrm{R}(f') = \int f'(x)^2\,dx$, so that the IMSE is given by

$$\mathrm{IMSE}\,[\hat{f}_k] = \frac{1}{nh} + \frac{1}{12}h^2\mathrm{R}(f') + O(n^{-1}). \tag{6}$$

From this equation, the optimal bandwidth is seen to be

$$h^* = \left[\frac{6}{n\mathrm{R}(f')}\right]^{1/3} \quad \text{and} \quad \mathrm{IMSE}^* = (\frac{9}{16})^{1/3}\,\mathrm{R}(f')^{1/3}\,n^{-2/3}\,. \tag{7}$$

Thus the optimal bandwidth approaches zero at the rate $O(n^{-1/3})$ and not the rate $O(n^{-1/2})$ as suggested by **?** nor the rate $O(\frac{1}{\log n})$ as suggested by **?**. With regards to IMSE, the best rate a histogram can achieve is of order $O(n^{-2/3})$, which falls well short of the parametric rate of $O(n^{-1})$. From Equation (7), the larger the value of the roughness $\mathrm{R}(f')$ of the true density, the smaller the optimal bandwidth and the larger the average error.

Finally, the smoothest density with variance σ^2 is

$$g(x) = \frac{15}{16\sqrt{7}\sigma}\left(1 - \frac{x^2}{7\sigma^2}\right)^2 \qquad -\sqrt{7}\sigma < x < \sqrt{7}\sigma \tag{8}$$

and zero elsewhere, for which $R(g') = 15\sqrt{7}/(343\sigma^3)$. Since $\mathrm{R}(f') \geq \mathrm{R}(g')$ for any other continuous density, f,

$$h^* = \left[\frac{6}{n\mathrm{R}(f')}\right]^{1/3} \leq \left[\frac{6}{n\mathrm{R}(g')}\right]^{1/3} = \left[\frac{686\,\sigma^3}{5\sqrt{7}n}\right]^{1/3} = 3.73\,\sigma\, n^{-1/3}\,, \tag{9}$$

which is the "oversmoothed bandwidth" rule. Consider the normal reference rule, $f = \phi = N(\mu, \sigma^2)$, for which $\mathrm{R}(\phi') = 1/(4\sqrt{\pi}\sigma^3)$, which when substituted into Equation (7) gives $h^* = 3.49\,\sigma\, n^{-1/3}$, a value that is only 6.4% narrower than the oversmoothed bandwidth.

The oversmoothing rule (9) may be inverted when the scale is the range of the density to obtain a lower bound of $\sqrt[3]{2n}$ on the number of bins in the optimal histogram. This formula should be compared to Sturges' rule of $1+\log_2 n$, which is in common use in many statistical packages (**?**). In fact, the histograms in the first frames of Figures 7 and 8 correspond to Sturges' rule, while the second frames of these figures correspond to the oversmoothed bandwidths. Presumably the optimal bandwidth would occur somewhere between the second and third frames of these figures. Clearly Sturges' rule results in oversmoothed graphs since the optimal number of bins is severely underestimated.

3.3 ASH and Kernel Estimators

The histogram is an order-one density estimator, since the bias is determined by the first derivative. The estimators used most in practice are order-two estimators. (Recall that order-three estimators are not non-negative.) Perhaps the most unexpected member of the order-two class is the frequency polygon (FP), which is the piecewise linear interpolant of the midpoints of a histogram. (**?**) showed that

$$\mathrm{IMSE}\,[\hat{f}_{\mathrm{FP}}] = \frac{2}{3nh} + \frac{49}{2880}h^4\mathrm{R}(f'') + O(n^{-1}). \tag{10}$$

Compared to Equation (6), the integrated variance of a FP is 33% smaller and the integrated squared bias is two orders of magnitude smaller. Clearly, $h^* = O(n^{-1/5})$ and $\mathrm{IMSE}^* = O(n^{-4/5})$. Thus the error converges at a faster rate than the histogram, by using bins which are wider and an estimator which is not discontinuous. Examples and other results such as oversmoothing may be found in **?**.

The use of wider bins means that the choice of the bin origin has a larger impact, at least subjectively. Given a set of m shifted histogram, $\hat{f}_1(x), \ldots, \hat{f}_m(x)$, one might use cross-validation to try to pick the best one. Alternatively, **?** suggested the averaged shifted histogram (ASH), which is literally defined:

$$\hat{f}_{\mathrm{ASH}}(x) = \frac{1}{m}\sum_{k=1}^{m}\hat{f}_k(x). \tag{11}$$

To be specific, suppose the collection of m histograms has meshes shifted by an amount $\delta = h/m$ from each other. Recompute the bin counts, ν_k, on the finer mesh, $t'_k = k\delta, -\infty < k < \infty$. Then a bin count for one of the histograms with bin width h may be computed by adding m of the bin counts on the finer mesh. For x in the ℓ-th (narrow) bin, there are m shifted histograms that include the (narrow) bin count, ν_ℓ. Adding these m shifted histograms together and averaging gives:

$$\frac{\nu_{\ell+1-m} + 2\nu_{\ell+2-m} + \cdots + m\,\nu_\ell + \cdots + 2\nu_{\ell+m-2} + \nu_{\ell+m-1}}{m \times nh}, \qquad (12)$$

or after re-arranging

$$\hat{f}_{\text{ASH}}(x) = \frac{1}{nh} \sum_{k=1-m}^{m-1} \left(1 - \frac{|k|}{m}\right) \nu_{\ell+k}\,. \qquad (13)$$

As the number of shifted histograms $m \to \infty$, the weights on the bin counts approaches the triangular kernel given by $K(t) = 1 - |t|$ for $|t| < 1$ and zero elsewhere. The ASH may be generalized to handle general weights by sampling from an arbitrary kernel function, $K(t)$, which is any symmetric probability density defined on the interval $[-1, 1]$. In this case,

$$\hat{f}_{\text{ASH}}(x) = \frac{1}{nh} \sum_{k=1-m}^{m-1} w_m(k)\,\nu_{\ell+k} \quad \text{where} \quad w_m(k) \propto K(k/m)\,. \qquad (14)$$

Like the FP, the ASH is an order-two algorithm, but more efficient in the statistical sense.

In Figure 9, two ASHs of the $\log_{10}$-cholesterol data are shown. The bin edge effects and discontinuities apparent in the ordinary histogram in Figure 8 are removed. The extra features in the distribution are hinted at.

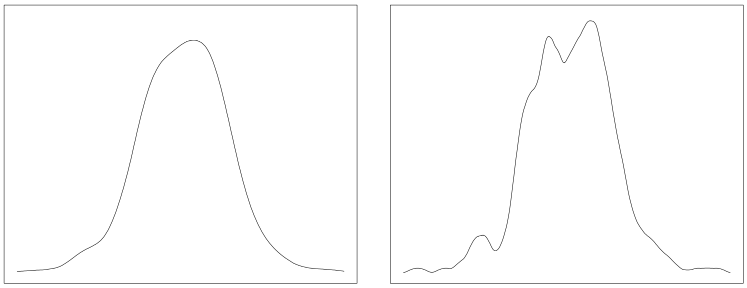

Fig. 9. Averaged shifted histograms of the $\log_{10}$-cholesterol data.

The extension of the ASH to bivariate (and multivariate) data is straightforward. A number of bivariate (multivariate) histograms are constructed with equal shifts along the coordinate axes and then averaged together. Figure 10 displays a bivariate ASH of the same lipid data displayed in Figure 1. The strong bimodal and weak trimodal features are evident. The third mode is perhaps more clearly represented in a perspective plot; see Figure 11. (Note that for convenience, the data were rescaled to the intervals $(0, 1)$ for these plots, unlike Figure 1.) The precise location of the third mode above (and between) the two primary modes results in the masking of the multiple modes when viewed along the cholesterol axis alone. This masking feature is commonplace and a primary reason for trying to extend the dimensions available for visualization of the density function.

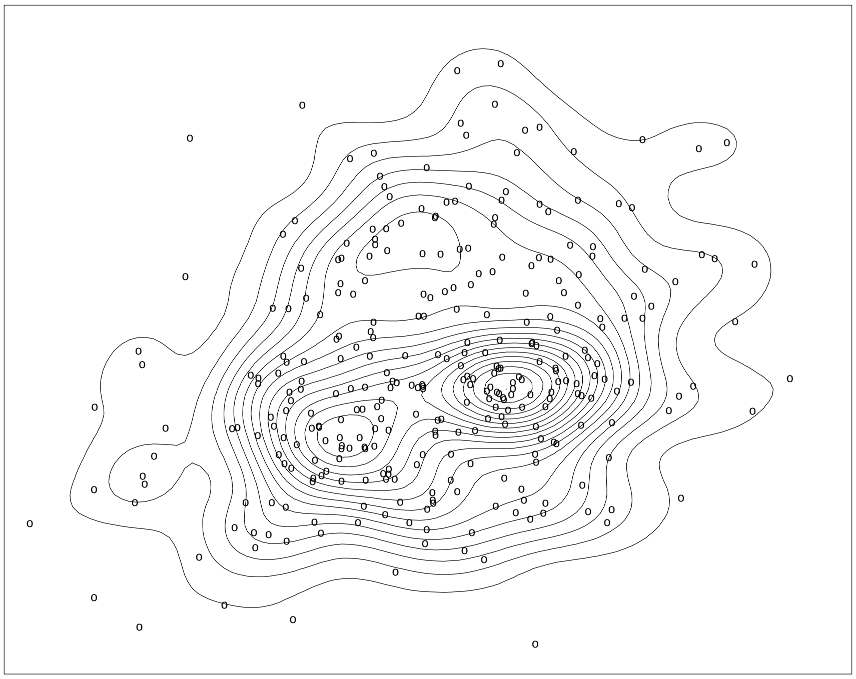

Fig. 10. Bivariate ASH of the $\log_{10}$-cholesterol and $\log_{10}$-triglyceride data.

3.4 Kernel and Other Estimators

The ASH is a discretized representation of a kernel estimator. Binned kernel estimators are of great interest to reduce the computational burden. An alternative to the ASH is the fast Fourier transform approach of **?**. Kernel methods were introduced by **?** and **?** with earlier work by Evelyn Fix and Joe Hodges completed by 1951 in San Antonio, Texas (see **?**).

Given a kernel function, $K(t)$, which is generally taken to be a symmetric probability density function, the kernel density estimate is defined by

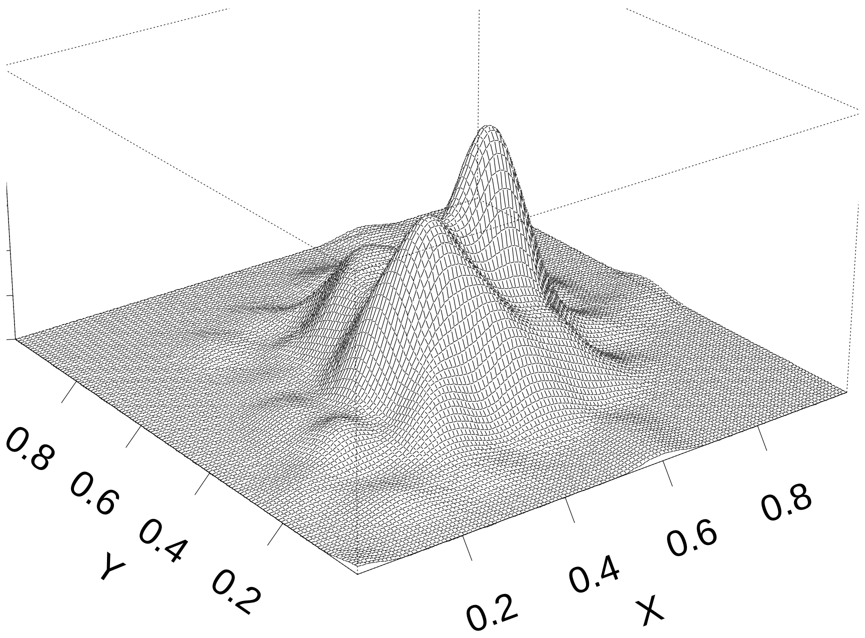

Fig. 11. Perspective view of the Bivariate ASH in Figure 10.

$$\hat{f}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)\,, \tag{15}$$

letting K_h denote the kernel density transformed by the scale factor, h; that is, $K_h(t) = K(t/h)/h$. Among kernels with finite support, Beta densities shifted to the interval $(-1, 1)$ are popular choices. Among kernels with infinite support, the normal density is by far the most common choice. An important paper by **?** showed that the normal kernel has the unique property that the number of modes in the kernel estimate monotonically decreases as the smoothing parameter increases. For many exploratory purposes, this property alone is reason to use only the normal kernel. **?** proposed graphing the locations of all modes at all bandwidths in the "mode tree." **?** proposed an extension of Silverman's bootstrap test (**?**) for the number of modes to test individual modes. Software for the ASH, kernel estimates, and the various mode tests may be found on the web; see statlib at www.stat.cmu.edu, for example.

Multivariate extensions of the kernel approach generally rely upon the product kernel. For example, with bivariate data $\{(x_i, y_i),\ i = 1, \ldots, n\}$, the bivariate (product) kernel estimator is

$$\hat{f}_K(x, y) = \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x - x_i) K_{h_y}(y - y_i)\,. \tag{16}$$

A different smoothing parameter for each variable generally gives sufficient control. A full bivariate normal kernel may be used in special circumstances, effectively adding one additional smoothing parameter in the form of the correlation coefficient. However, an equivalent estimate may be obtained by rotating the data so that the correlation in the kernel vanishes, so that the product kernel may be used on the transformed data.

In higher dimensions, some care must be exercised to minimize the effects of the curse of dimensionality. First, marginal variable transformations should be explored to avoid a heavily skewed appearance or heavy tails. Second, a principal components analysis should be performed to determine if the data are of full rank. If so, the data should be projected into an appropriate subspace. No nonparametric procedure works well if the data are not of full rank. Finally, if the data do not have many significant digits, the data should be carefully blurred. Otherwise the data may have many repeated values, and cross-validation algorithms may believe the data are discrete and suggest using $h = 0$. Next several kernel or ASH estimates may be calculated and explored to gain an understanding of the data, as a preliminary step towards further analyses.

An extensive body of work also exists for orthogonal series density estimators. Originally, the Fourier basis was studied, but more modern choices for the basis functions include wavelets. These can be re-expressed as kernel estimators, so we do not pursue these further. In fact, a number of workers have shown how almost any nonparametric density algorithm can be put into the form of a kernel estimator; see **?** and **?**, for example. More recent work on local likelihood algorithms for density estimation further shows how closely related parametric and nonparametric thinking really is; see **?** for details and literature.

4 Visualization of Trivariate Functionals

The field of scientific visualization has greatly enhanced the set of tools available for the statistician interested in exploring the features of a density estimate in more than two dimensions. In this section, we demonstrate by example the exploration of trivariate data.

We continue our analysis of the data given by the duration of 299 consecutive eruptions of the Old Faithful geyser. A graph of the histogram of these data is displayed in Figure 2(b). We further modified the data as follows: the 105 values that were only recorded to the nearest minute were blurred by adding uniform noise of 30 seconds in duration. (The remaining data points were recorded to the nearest second). An easy way to generate high-dimensional data from a univariate time series is to group adjacent values. In Figure 12, ASH's of the univariate data $\{y_t\}$ and the lagged data $\{(y_{t-1}, y_t)\}$ are shown. The obvious question is whether knowledge of y_{t-1} is useful for predicting the value of y_t. Clearly, the answer is in the affirmative, but the structure would not be well-represented by an autoregressive model.

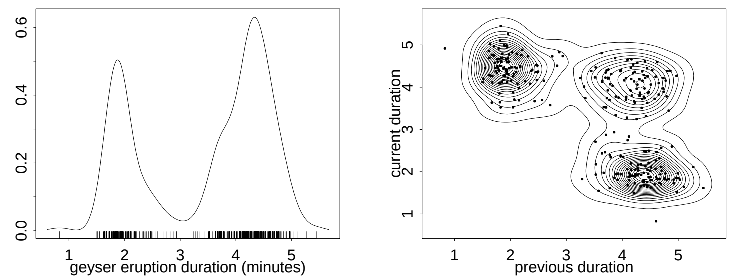

Fig. 12. Averaged shifted histograms of the Old Faithful geyser duration data.

Next, we computed the ASH for the trivariate lagged data $\{(y_{t-2}, y_{t-1}, y_t)\}$. The resulting estimate, $\hat{f}_{\text{ASH}}(y_{t-2}, y_{t-1}, y_t)$, may be explored in several fashions. The question is whether knowing y_{t-2} can be used to predict the joint behavior of (y_{t-1}, y_t). This may be accomplished, for example, by examining *slices* of the trivariate density. Since the (univariate) density has two modes at $x = 1.88$ and $x = 4.33$ minutes, we examine the slices $\hat{f}_{\text{ASH}}(1.88, y_{t-1}, y_t)$ and $\hat{f}_{\text{ASH}}(4.33, y_{t-1}, y_t)$; see Figure 13. The 297 data points were divided into two groups, depending on whether $y_{t-2} < 3.0$ or not. The first group of points was added to Figure 13(a), while the second group was added to Figure 13(b).

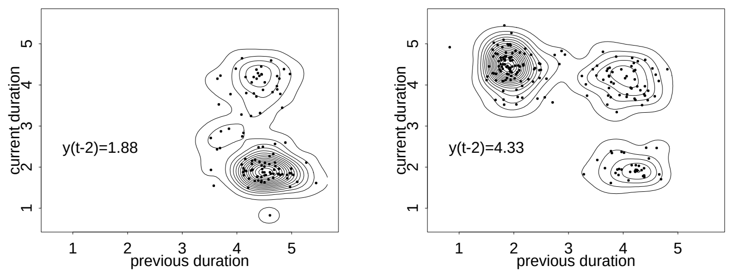

Fig. 13. Slices of the trivariate averaged shifted histogram of lagged values of the Old Faithful geyser duration data.

Since each axis was divided into 100 bins, there are 98 other views one might examine like Figure 13. (An animation is actually quite informative.) However, one may obtain a holistic view by examining level sets of the full trivariate density. A level set is the set of all points $\mathbf{x}$ such that $\hat{f}_{\text{ASH}}(\mathbf{x}) = \alpha \hat{f}_{\max}$, where $\hat{f}_{\max}$ is the maximum or modal value of the density estimate, and $\alpha \in (0, 1)$ is a constant that determines the contour level. Such contours are typically smooth surfaces in $\Re^3$. When $\alpha = 1$, then the "contour" is simply the modal location point. In Figure 14, the contour corresponding to $\alpha = 58\%$ is displayed. Clearly these data are multimodal, as five well-separated high-density regions are apparent. Each cluster corresponds to a different sequence of eruption durations, such as long-long-long. The five clusters are now also quite apparent in both frames of Figure 13. Of the eight possible sequences, three are not observed in this sequence of 299 eruptions.

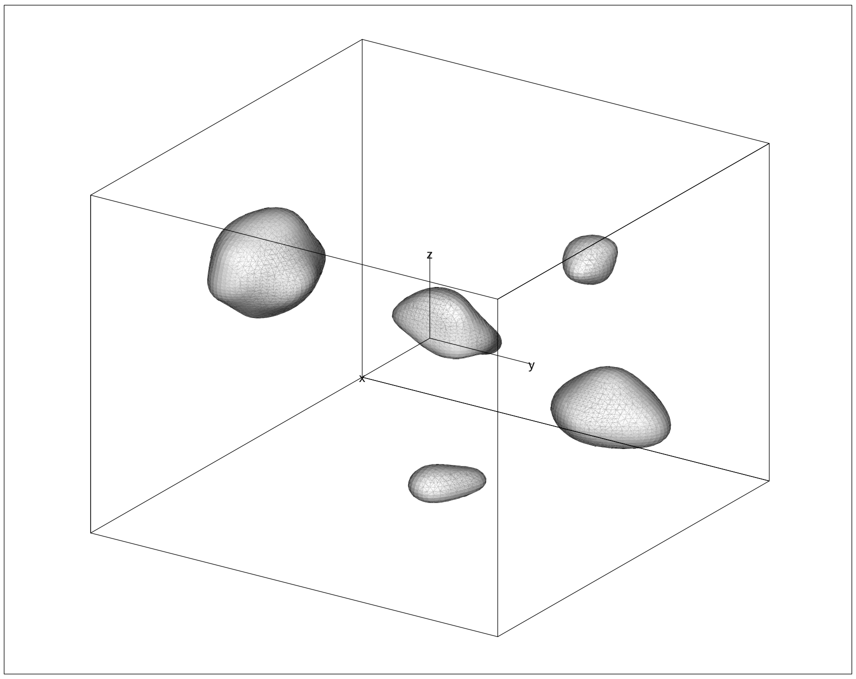

Fig. 14. Visualization of the $\alpha = 58\%$ contour of the trivariate ASH of the lagged geyser duration data.

A single contour does not convey as much information as several. Depending on the display device, one may reasonably view three to five contours, using transparency to see the higher density contours that are "inside" the lower density contours. Consider adding a second contour corresponding to $\alpha = 28\%$ to that in Figure 14. Rather than attempt to use transparency, we choose an alternative representation which emphasizes the underlying algorithms. The software which produced these figures is called *ashn* and is available at the author's website. ASH values are computed on a three-dimensional lattice. The surfaces are constructed using the marching cubes algorithm (**?**), which generates thousands of triangles that make up each surface. Here, we choose not to plot all of the triangles but only every other "row" along the second axis. The striped effect allows one to interpolate and complete the low-density contour, while allowing one to look inside and see the high-density contour. Since there are five clusters, this is repeated five times. A smaller sixth cluster is suggested as well.

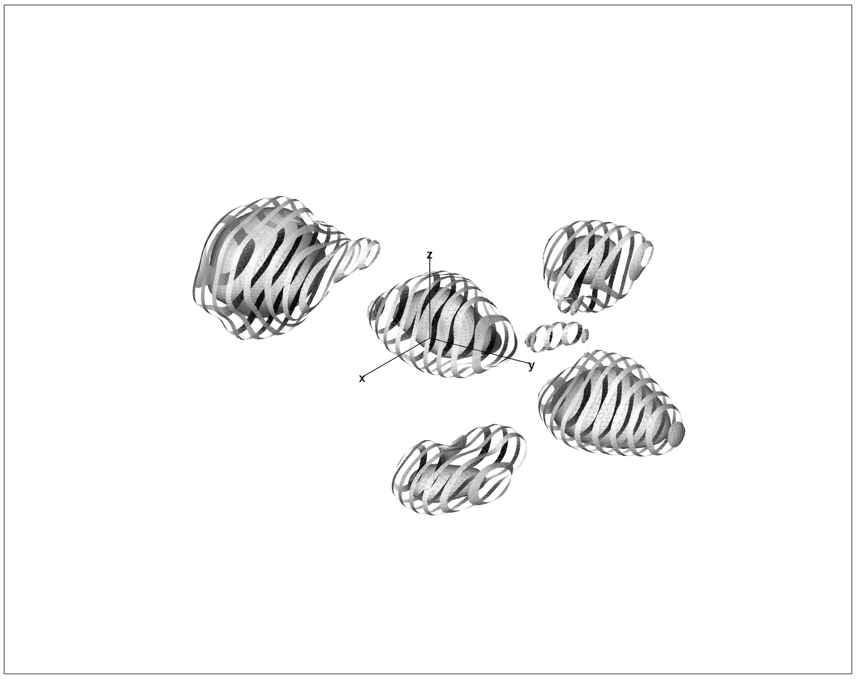

Fig. 15. Visualization of the $\alpha = 28\%$ and 58% contours of the trivariate ASH of the lagged geyser duration data.

5 Conclusions

Exploring data is an important part of successful statistical model building. General discussions of graphical tools may be found in **?**, **?**, **??**, **?** and **?**, for example. Advanced exploratory software may be found in many commercial packages, but of special note is the XGobi (**?**) system and successors. Immersive environments are also of growing interest (**?**). A general visualization overview may be found in **?**.

Especially when the data size grows, point-oriented methods should be supplemented by indirect visualization techniques based upon nonparametric density estimation or by parallel coordinates (**??**). Many density algorithms are available. The use of order-two algorithms is generally to be recommended. These should be calibrated by several techniques, starting with an oversmoothed bandwidth and the normal reference rule.

For data beyond three dimensions, density estimates may be computed and slices such as $\hat{f}(x, y, z, t = t_0)$ visualized. If advanced hardware is available, the surfaces can be animated as t varies continuously over an interval (t_0, t_1); see **??**. Obviously, this is most useful for data in four and five dimensions. In any case, multivariate density estimation and visualization are important modern tools for EDA.

6 Acknowledgments

This research was supported in part by the National Science Foundation grants NSF EIA-9983459 (digital government) and DMS 02-04723 (non-parametric methodology).

References