1 Multidimensional Smoothing and Visualization

DAVID W. SCOTT
Department of Statistics
Rice University
Houston, Texas

1.1 INTRODUCTION

Visualization of data and visualization of functionals of data are often treated separately. Exploratory visualization of data is often an end unto itself. Sometimes exploratory data may not meet the criteria for stochastic modeling, and no move towards estimation is contemplated, nor is further functional visualization appropriate. Even when exploratory visualization is undertaken as a data preparation step for parametric estimation, multivariate functional visualization is rare. However, statistical data points to statistical functions, and the opportunities for functional visualization are growing rapidly. This chapter describes some techniques and provides examples of functional visualization. For simplicity, the discussion will be limited to continuous functions. Whenever possible, data and functional visualization will be presented in parallel. Data visualization without functional visualization, or vice versa, represents an incomplete picture and missed opportunity.

Parametric modeling in classical statistics relies on tabular display of model coefficients, error bands, and p-values, rather than graphical display. A major component of model building is the process of reducing the complexity of the statistical model to its most parsimonious form. The very simplicity of parametric models aids in data understanding, even in high dimensional settings. Graphical display plays a secondary role, beginning with simple one and two dimensional data presentation through histograms and scatter plots, and ending with a diagnostic role using residual plots. Parametric models are preferred for their low variance, but are susceptible to high bias due to model misspecification.

Nonparametric approaches to density and regression estimation do not have such a parsimonious representation, so that graphical display is the most efficient means of understanding a nonparametric estimate. Nonparametric estimates are neither unbiased nor minimum variance, but are not as susceptible to high model bias, at least until the curse of dimensionality takes effect. As the number of variables increases, the number of visualization options quickly decreases, and the statistical efficiency of nonparametric methods also decreases rapidly. However, the more common availability of massive data sets and sophisticated graphical engines represents a clear opportunity for modern computational statistics.

A glance at most statistical journals and textbooks reveals a surprising lack of sophistication in common data and model presentation. Almost all figures are flat. Almost all are either line graphs (including contours) or point graphs. Color is rarely employed and often without sufficient impact or importance to support or justify its use. The tendency is to present results of the form $f(x)$ and include graphical presentation, but switch to tabular form with additional variables. When a bivariate function, $f(x, y)$, is displayed, contour and perspective presentations are about equally frequent, as these capabilities are now widely available. In many sciences, display of trivariate functions such as $f(x, y, z)$ has become common, as the data represent real spatial objects. The more general nature of statistical data and the common use of parametric models has made the use of volumetric visualization relatively rare among statisticians.

Several authors have written generally about statistical graphics. Cleveland (1985) describes the details that make a scatter diagram faithly convey the structure in the data. Tufte (1983) and Wainer (1997) have written more generally about visualization and presentation. Both focus on charts and maps as much as function visualization. These books do not include volumetric representation. Many collections do, including Wolff and Yaeger (1993). However, these tend to be computer graphics efforts at representing "real" things as graphical objects that can be manipulated. Statistical functions, such as a regression plane or a density surface, do not represent real objects, but rather mathematical abstractions. However, the same tools apply and statisticians may benefit from a more hands on exploratory view of nonparametric functionals. Some of the newer technologies in virtual reality, including the data cave (Cook et al., 1997) and the virtual workbench, offer unexplored opportunities for both data and function exploration.

1.2 DATA POINT VISUALIZATION

Data point visualization is a well-explored endeavor, from one to two to three and more dimensions. The most common figure by far is the ordinary (x_i, y_i) bivariate scatter diagram. For univariate data, a scatter diagram (commonly referred to as a *dot plot*) is seldom presented; rather, a histogram is the more common choice. A hybrid display is Tukey's (1977) stem-and-leaf plot, which displays individual data in such a manner that the frequency of the data in bins is apparent. For higher dimensional data, these ideas can be assembled for all pairs of variables into a scatterplot matrix.

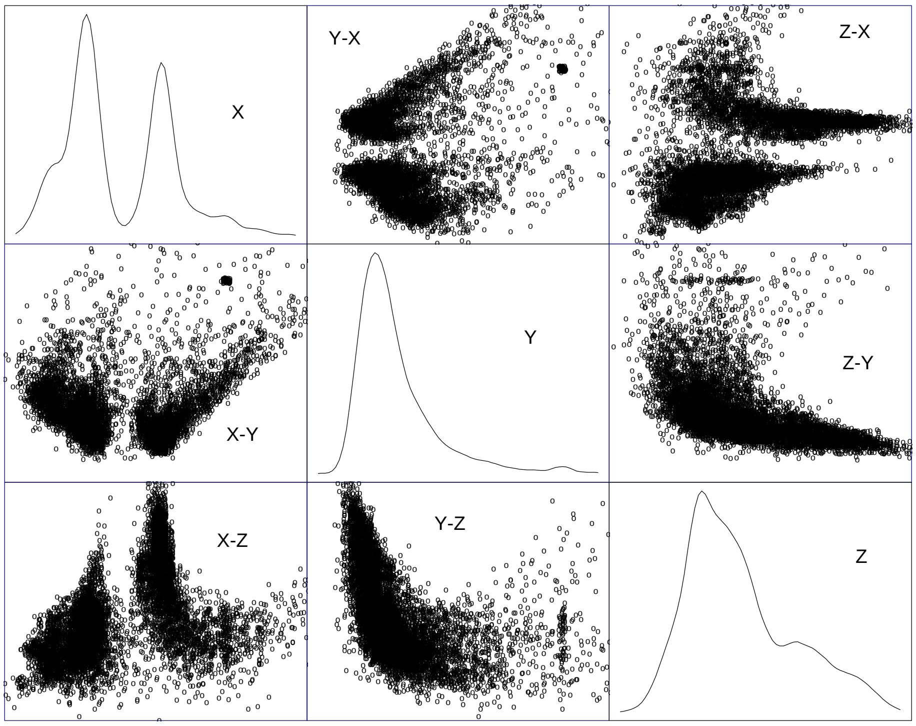

Fig. 1.1 Scatterplot matrix of remote sensing data (n = 8,088). Smoothed histograms (ASH) of the variables are shown in the diagonal frames.

In Figure 1, we display an enhanced 3×3 scatterplot matrix of $n = 8{,}397$ trivariate points from a remote sensing satellite data set, which will be described below. (Actually, only 8,088 points are displayed, as 3.7% of the data were truncated outside the viewing panel.) In the diagonal frames, smoothed histograms of the three variables are displayed. The off-diagonal frames contain the pairwise scatterplots. Each pair of variables appears twice. A first impression is that there are two dominate clusters, with possibly two other smaller bumps at least suggested in the X histogram. In fact, these data were extracted from Landsat images of size 117×196 pixels of land in agricultural production. Gray-scale and false-color images from these data may be found in Scott (1986, 1992), and the data are available at ftp.stat.rice.edu. Of the 22,932 pixels, we selected a subset for the crops sunflower, spring wheat, and spring barley for use in Figure 1. We will examine these data further below. The scatter diagrams in Figure 1 illustrate the common problem of overplotting when the sample size exceeds one thousand. The outliers are over-emphasized in a scatter plot. Carr, et al. (1992) have suggested a hybrid figure called a glyph histogram, which provides better balance among the two views. Figure 2 displays a hexagonal histogram glyph of the X-Z scatter diagram from Figure 1. This representation largely confirms that there are no hidden clusters in the high-density region of data.

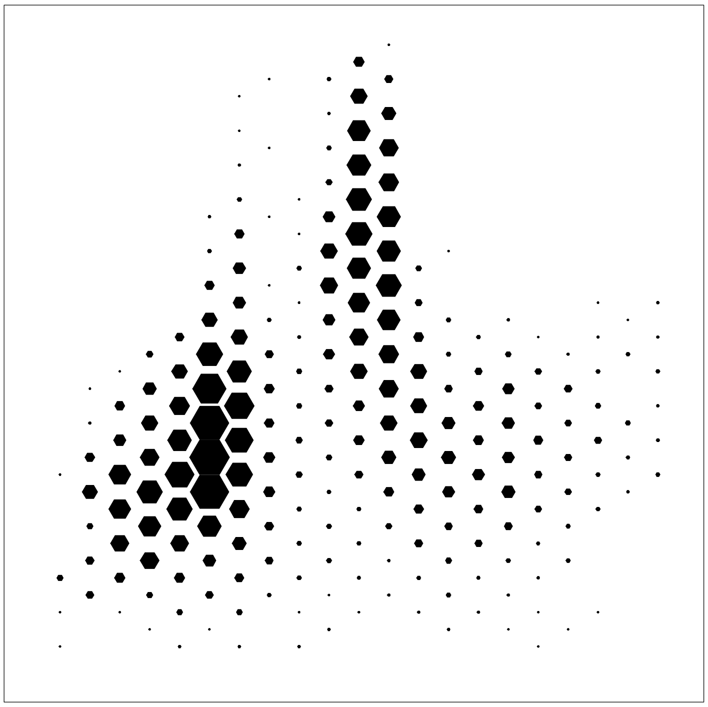

Fig. 1.2 Hexagonal histogram glyph of the X-Z scatter diagram of the remote sensing data. The size of each hexagonal glyph is approximately proportional to the number of data points in the hexagonal bin.

For trivariate data in particular, rotating scatterplots have been popular since the mid-1980's with MacSpin (Donoho, et al., 1988), and are available in most packages today. Animation control is quite sophisticated these days. Alternatively, stereo representation provides a static 3-D view of trivariate data (Carr and Nicholson, 1988). Early crude red-green stereo techniques and cardboard glasses have given way to advanced LCD glasses. With virtual reality and data caves, true stereo representation can be merged with animation. High-resolution stereo representation seems to be more effective than rotation for displaying larger data sets. Given the number of individuals with various vision deficiencies such as color blindness or depth perception limitations, the newer LCD technologies are almost foolproof. Personal experience in a data cave (Loftin, 1997) suggests that total immersion can be disorienting and tiring. However, the capabilities are absolutely unique and powerful.

For higher dimensional data, an effective exploration tool is the grand tour (Asimov, 1985). Rather than look at a collection of random projections, a smooth set of projections is viewed in such a manner that eventually all two-dimensional projections are produced. Huber (1985) has provided estimates of how long that might take as the dimension increases, and the increase is rapid. The marriage of the grand tour and "optimal" projection selection via the grand tour (Friedman and Tukey, 1974) has proven highly successful. The grand tour can provide not only a set of different starting views for the nonlinear optimization inherent with projection pursuit, but can also provide search directions. A particularly effective implementation of the grand tour and projection pursuit may be found in the larger XGobi program (Swayne, et al., 1991). The idea of displaying frequency curves rather than data points has been investigated by Scott (1995), who described the resulting representation as the *density grand tour*. An implementation in C using *geomview* as the display engine has been described by Salch and Scott (1997). Geomview is distributed by The Geometry Center (www.geom.umn.edu). The density grand tour is available at ftp.stat.rice.edu.

With the exception of some work by Nason (1995), most projection pursuit indices are two dimensional. There is some discussion of the advantage, if any, of viewing three full coordinates in a grand tour and distinguishing the extra dimension from the movement through the data space. However, with the display of trivariate density estimates (described below) in the density grand tour, the advantage is clear. The same advantage should be apparent in total immersive environments such as the data cave.

Other glyph representations are available for high dimensional data. Perhaps the best known is the Chernoff face (Chernoff, 1973), in which variables are associated with facial features and determine the size or other attribute of those features. Since each data point is represented individually, sample size is a limiting factor. This representation attempts to display all variables simultaneously, although the assignment of variables to particular features is largely subjective.

A relatively simple device that also displays all variables and in a more symmetric manner is the parallel coordinate plot (Inselberg, 1985; Wegman, 1990). Axes are drawn in parallel equally spaced. The coordinate values of a data point are marked and connected by line segments. The proposers note a number of interesting dualities between points and lines in this representation. Parallel coordinate plots cannot handle as many points as a scatter diagram, but Wegman has partially addressed this limitation by plotting the intensity (or density) of ink on the diagram. This will reveal the patterns in these diagrams for much larger data sets.

1.3 FUNCTIONAL VISUALIZATION IN ONE AND TWO VARIABLES

Suppose $f(\cdot)$ represents a generic statistical functional, whether estimated parametrically or nonparametrically, where the dependence on the data $\{x_1, x_2, \ldots, x_n\}$ has been suppressed. High quality graphical display of $f(x)$ and $f(x,y)$ in postscript is routinely available. In the bivariate setting, as opposed to the univariate setting, the user has a number of choices to make—contour versus perspective representation, number of contour levels versus surface viewing angle, use of contour labels versus choice of wireframe or shaded surface, and so on.

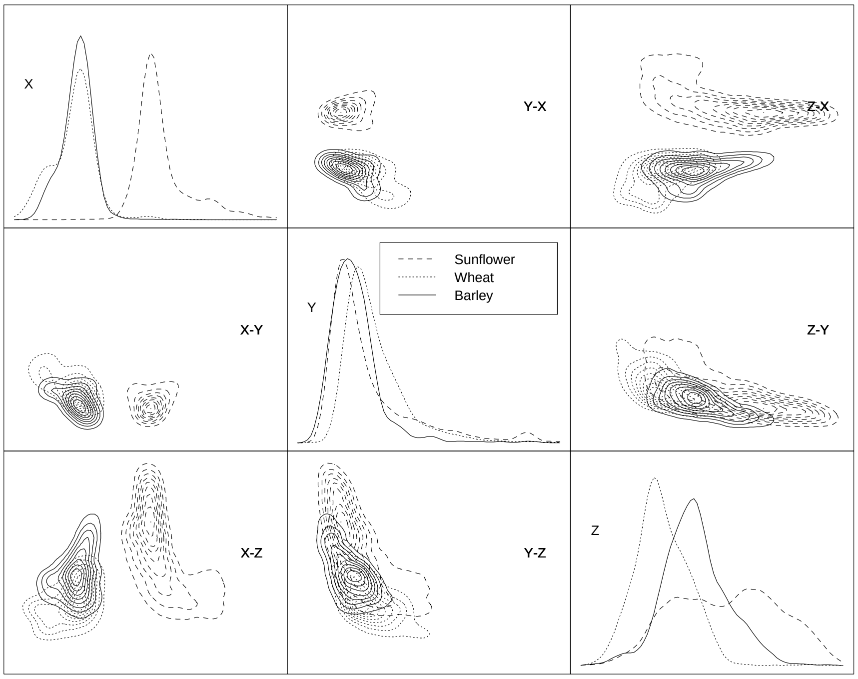

Fig. 1.3 Contourplot matrix of remote sensing data as in Figure 1. Contours of bivariate smoothed histograms (ASH) of for each of three crops (sunflower, spring wheat, and spring barley) are shown for each pair of variables. Smoothed histograms (ASH) of the variables for each crop are shown in the diagonal frames.

In Figure 3, we continue our examination of the remote sensing data. Extending the idea of a scatterplot matrix, a contourplot matrix is displayed. Furthermore, since the actual classification ("ground truth" in remote sensing jargon) is known for these data, each frame includes three density estimates, one for each crop. The original image contained 3,694 pure pixels where sunflowers were growing, 3,811 of spring wheat, and 892 of spring barley. Comparing Figures 1 and 3, the mixing of the three univariate densities into one becomes apparent (diagonal frames). The two clusters observed before — a well-separated sunflower cluster and one spring grain cluster — are really three classes. Spring wheat and spring barley are not sufficiently separated to manifest two modes.

Examining the off-diagonal frames, the earlier remark that scatterplots of large data sets tend to overemphasize low-density regions is reinforced. To be more specific, a definition of an individual contour of a multivariable function $f(\mathbf{x})$ for $\mathbf{x} \in \Re^d$ is required:

$$C_\alpha = \left\{ \mathbf{x} \, : \, f(\mathbf{x}) = \alpha \cdot \max_{\mathbf{x}} f(\mathbf{x}) \right\}, \quad \text{where } 0 \leq \alpha \leq 1.$$

We refer to C_α as an "α-contour." Thus C_1 gives the location of the (largest) mode, while C_0 would delineate the support of the function f. In Figure 3, nine equally spaced contours are displayed for each density estimate; that is, C_{α_k} are drawn in different line types for $\alpha_k = k/10$, for $k = 1, \ldots, 10$.

The α-contour concept may be applied to any (preferably continuous) function, and not just to density functions. A regression example is considered below. However, contours of a density function are usually closed, which aids and facilitates interpretation. In such cases, the equivalent concept of a β-contour, $\mathcal{C}_\beta$, may be considered, where again $0 \leq \beta \leq 1$. The fraction of the probability density contained in $\mathcal{C}_\beta$ defines the constant β, and $\mathcal{C}_\beta$ is the (unique) contour among C_α satisfying

$$\int_{\mathcal{C}_\beta} f(\mathbf{x}) \, d\mathbf{x} = \beta.$$

Here, $\mathcal{C}_0$ represents the largest mode and $\mathcal{C}_1$ represents the support of f. Thus, α and β are negatively correlated. Note that in the bivariate setting, most practical experience with contours comes from reading topographical maps, which are of the C_α category. In practice, β is usually estimated for a given α-contour by computing the density at each data point and counting the number of density values greater than or equal to α. If a particular value of β is required, then a simple line search over α may be used.

The choice between a contour plot and a perspective representation of $f(x, y)$ is somewhat subjective. Personal experience suggests that the mathematical notion of equivalence between the two representations does not translate into practice. Indeed, trying to imagine a surface perspective view by studying a contour plot is rarely very successful. A perspective plot provides a more complete view of a function. A contour plot provides only a fraction, perhaps half, of the extra dimension. Of course, a portion of a perspective plot is likely to be hidden behind a peak, or the height of the function may vary so widely that small features cannot be simultaneously depicted as in a contour plot. Nevertheless, the essential difference in the use of continuity between contour and surface representations favors the latter. While individual contours are continuous, the jump from contour to contour is discontinuous. Furthermore, surfaces that undulate require special care reading in contour space, as few contour representations indicate whether nested levels are increasing or decreasing in value.

In Color Plate I, the bivariate X-Z variable density estimates for the three crops are shown, and should be compared to the lower left frame in Figure 3, rotated about 180°. The software program *geomview* was used for this display. The density estimates are based on smoothed histograms and have finite support. The greatest of the three densities is displayed at each point. If a nonparametric Bayesian classifier with equal priors was of interest, the highest color displayed would give nonparametric decision regions for classification. Geomview supports real time visualization, including rotation, shading, and transparency, so that a few manipulations eliminate any hidden regions of the surface. The perspective plot more clearly indicates the relative height of the three densities (compared to the contour picture), and provides much greater detail in the tails and regions of low density.

1.4 AVERAGED SHIFTED HISTOGRAMS

1.4.1 Univariate Data

Plotting routines require values of the function on a mesh — $\{f_i\}$, $i = 1, \ldots, m_1$ for a single-variable function, and $\{f_{ij}\}$, $i = 1, \ldots, m_1$ and $j = 1, \ldots, m_2$ for a two-variable function. Similarly, $\{f_{ijk}\}$ and $\{f_{ijk\ell}\}$ are required for three- and four-variable function visualization. Depending on the sophistication of the software, these mesh values may be interpolated by linear functions or even specialized spline approximations.

When f is a density function, the most natural estimator that provides a grid of values is the classical histogram with equally-spaced bins. Generally, the bin widths in a histogram are selected to minimize bias and variance, and not to match the desired graphical resolution of a plotting routine. Optimal bin widths are certain to be much wider than any desired graphical resolution. However, the *averaged shifted histogram* (ASH) of Scott (1985) is well-suited to both tasks. The ASH retains much of the computational advantage of the histogram for massive data sets, while computing precisely the array of estimates required for visualization.

The construction of the averaged shifted histogram is easily described in one variable and generalized to d variables. Consider not one, but m equally-spaced histograms, $\hat{p}_1$, $\hat{p}_2, \ldots$, $\hat{p}_m$, of a data set. Each histogram employs the same bin width, h, but the meshes are shifted by a fixed amount, h/m, from the preceding mesh. Define δ to be this shift, that is, $\delta = h/m$. The ordinary ASH, $\hat{f}(x)$, is formed by simply averaging the m different shifted histogram levels covering the point x:

$$\hat{f}(x) = \frac{1}{m} \sum_{k=1}^{m} \hat{p}_k(x).$$

Since each histogram, $\hat{p}_k(\cdot)$, is piecewise constant, then so is the ASH estimator, $\hat{f}(\cdot)$. However, the "apparent bin width" of the ASH is not h, but rather h/m, or δ. Thus h may be selected appropriately to minimize statistical error, while m (and hence δ) selected to control the mesh size. In one dimension, 100–400 mesh points might be desirable, but in three dimensions, 100^3 mesh points is beyond the comfortable range for most computers and 50^3 might be acceptable.

An explicit formula for the ASH can be derived. Focusing on the narrow bin intervals of width δ, let ν_k denote the bin counts. Without loss of generality, consider the value of the ASH over the bin $[0, h)$, which we label bin 0. Then the m histograms covering this interval are given by:

$$\begin{aligned}
\hat{p}_m &= (\nu_0 + \nu_1 + \cdots + \nu_{m-1})/nh \\
\hat{p}_{m-1} &= (\nu_{-1} + \nu_0 + \cdots + \nu_{m-2})/nh \\
&\vdots \\
\hat{p}_1 &= (\nu_{1-m} + \nu_{2-m} + \cdots + \nu_0)/nh.
\end{aligned}$$

Thus the ASH is a weighted average of these bin counts, namely,

$$\hat{f}(x) = \frac{\nu_{1-m} + 2\nu_{2-m} + \cdots + m\nu_0 + \cdots + 2\nu_{m-2} + \nu_{m-1}}{m \cdot nh}$$

for $0 \leq x < \delta$. In general, for x in bin B_ℓ,

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

The weight function in the ordinary ASH is sampled from the isosceles triangle, which is called a kernel function, K. Here, $K(t) = 1 - |t|$ for $-1 \leq t \leq 1$. The ASH may be generalized to allow for unequal weighting of the shifted histograms. Let this new set of weights be denoted by $\{w_m(k)\}$; then

$$\hat{f}(x) = \frac{1}{nh} \sum_{k=1-m}^{m-1} w_m(k)\ \nu_{\ell+k}, \quad \text{for } x \in B_\ell.$$

Given a kernel, K, defined on the interval $[-1, 1]$, the weights may be obtained by the formula

$$w_m(k) = m \times \frac{K(k/m)}{\sum_{i=1-m}^{m-1} K(i/m)}, \quad \text{for } k = 1-m, \ldots, m-1.$$

Other popular choices for the kernel are the Epanechnikov function, $3/4(1-t^2)$, and the biweight function, $15/16(1-t^2)^2$. The normal function, $\exp(-\pi x^2)$, does not have finite support and cannot be used without further approximation.

As the number of shifted histograms grows, the ASH converges to the well-known kernel estimator,

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

Thus the ASH may be thought of as a statistical approximation to a finite-support kernel estimator. With this understanding, the choice of smoothing parameter h need not be an integer multiple of δ in the ASH, as the normalization in the definition of the weight function works generally.

Alternatively, a numerical approximation to the kernel estimator may be considered. Let t_k denote the center of bin B_k with bin counts ν_k as above. Then consider rounding all data values to bin centers. The kernel estimate becomes

$$\hat{f}(x) = \frac{1}{nh} \sum_{k=-\infty}^{\infty} \nu_k \; K\left(\frac{x - t_k}{h}\right).$$

The desired grid of density values, $\{f_j\}$, is obtained by

$$f_j = \hat{f}(t_j) = \frac{1}{nh} \sum_{k=-\infty}^{\infty} \nu_k \; K\left(\frac{t_j - t_k}{h}\right).$$

Silverman (1982) has shown how the Fast Fourier Transform may be used to compute these grid values efficiently when the kernel is the normal density. To avoid aliasing problems, a large number of empty bin counts must be appended onto the end of the bin count vector. Thus in several variables, if twice as many bins are required in each dimension for padding, considerable extra storage and computation may be required.

1.4.2 Multivariate Data

The extension of the averaged shifted histogram to bivariate and multivariate data is straightforward. In the bivariate setting, a collection of shifted histograms is produced. Suppose the histogram bin is a rectangle of size $h_1 \times h_2$. Then construct $m_1 m_2$ histograms by performing m_1 equal shifts along the

x-axis, and m_2 equal shifts along the y-axis for each, and average:

$$\hat{f}(x,y) = \frac{1}{m_1 m_2} \sum_{k_1=1}^{m_1} \sum_{k_2=1}^{m_2} \hat{p}_{k_1 k_2}(x,y), \quad \text{for } (x,y) \in B_{k_1 k_2}.$$

The resulting estimate with general weights can be shown to be a simple product of two univariate weight functions as defined above, applied to the bivariate bin counts:

$$\hat{f}(x,y) = \frac{1}{nh_1h_2} \sum_{k_1=1-m_1}^{m_1-1} \sum_{k_2=1-m_2}^{m_2-1} w_{m_1}(k_1)\ w_{m_2}(k_2)\ \nu_{\ell_1+k_1,\ell_2+k_2},$$

for $(x,y) \in B_{\ell_1,\ell_2}$. Further generalization to trivariate data is immediate.

Software to compute these estimates in Splus is available at ftp.stat.rice.edu. Three separate functions with interfaces to Fortran have been written: *ash1*, *ash2*, and *ashn*. The latter has been written to return a trivariate slice of up to ten dimensions of data. As mentioned above, mesh size grows exponentially, so that in core meshes beyond four or five dimensions are not generally feasible.

1.4.3 WARPing: ASH's For Regression

A multivariate averaged shifted histogram can be used to obtain the corresponding regression function by formally computing

$$\hat{r}(\mathbf{x}) = \int_y y\hat{f}(\mathbf{x},y)\,dy \Big/ \int_y \hat{f}(\mathbf{x},y)\,dy.$$

The well-known Nadarya-Watson regression estimator,

$$\hat{r}(x) = \frac{\sum_{i=1}^n y_i\, K((x-x_i)/h)}{\sum_{i=1}^n K((x-x_i)/h)}$$

falls from this expression when $\hat{f}$ is a product kernel estimator.

Whittaker and Scott (1994) and Scott and Whittaker (1996) discuss the derivation of the ASH regression estimator and its application to mapping and spatial smoothing of complex survey data. The additional information required is the bin response averages. In the case of bivariate data, these would be $\bar{y}_i$, and of the form $\bar{z}_{ij}$ for trivariate data. The two ASH regression estimates for these two cases are

$$\begin{aligned} \hat{r}_i &= \frac{\sum_a \bar{y}_{i+a} w_a \nu_{i+a}}{\sum_a w_a \nu_{i+a}} \\ \hat{r}_{ij} &= \frac{\sum_a \sum_b \bar{z}_{i+a,j+b} w_a w_b \nu_{i+a,j+b}}{\sum_a \sum_b w_a w_b \nu_{i+a,j+b}}. \end{aligned}$$

For the mapping applications above, the ordinary bin counts were replaced by the sampling frame weights. Taking account of such complicated features of real data turned out to be a trivial modification in the ASH setting, but not for other estimators.

The extension of the binning technology fundamental to the ASH to regression and average derivative estimation was described by Härdle and Scott (1992). They coined the acronym WARPing, which stands for weighted average of rounded points. Improvements to approximation by using linear binning counts rather than raw bin counts have been described by Hall and Wand (1996).

1.5 FUNCTIONAL VISUALIZATION IN THREE VARIABLES AND BEYOND

We return to our discussion of functional visualization. The averaged shifted histogram as described above is the basis of our color figures. An option in the *ashn* software outputs a full object-oriented file describing the contours, which may be piped into the *geomview* program for visualization. On a high-performance workstation, such as the Silicon Graphics High Impact Indigo 2 Model used to produce these figures, real-time rotation, animation, and contour transparency are supported. Geomview provides postscript output for our figures, but a screen capture program was used to capture the figures using transparency, as postscript does not support that capability.

1.5.1 Trivariate Functionals

Estimating and visualizing trivariate densities, $f(x, y, z)$, has progressed rapidly since I wrote Scott (1992). Widely available software packages such as Mathematica and Matlab provide such capabilities, for example. The two options of contouring and perspective plotting available for bivariate densities are now reduced to the contour representation. I prefer to use the *ashn* software for the level of control and design parameters (described below).

Three-dimensional contours corresponding to level sets of a four-dimension surface may not be familiar to everyone. For a trivariate normal density, the contours are egg-shaped surfaces or hyperellipses. The shape of the ellipses (as determined by the eigenvalues and eigenvectors of the covariance matrix) does not change as α or β vary, only the size of the ellipse. Using transparency techniques, several contours can be visualized simultaneously. Trivariate examples may be found in Scott (1992); a quadravariate example is given below.

The *ashn* software uses the marching cubes algorithm (Lorensen and Cline, 1987). The values at the eight vertices of a cube (or hyper-rectangle) are used to construct a number of triangles if the contour intersects the cube. This computation can be performed without reference to neighboring information and hence is fast. For many applications, further smoothing by Gouraud shading may be applied. We prefer to leave the underlying binning and interpolation information as part of the visualization, since the surfaces represent estimates and not noiseless functions. The details of bivariate and trivariate contouring are interesting in their own right and are covered in Scott (1992, Appendix A.1), among others.

Returning to the Landsat remote sensing example, trivariate ASH's for each of the three crops were computed using all three variables. A number of α-contours were computed to find the one containing half the estimated probability density. Thus a β-contour at the level $\beta = 0.5$ was computed for each crop density. These three contours, $\mathcal{C}_{0.5}$, are displayed in Color Plate II. The β-contours for wheat and barley are nearly disjoint. The sequence of figures of the Landsat data beginning with one variable at a time, and progressing to all three simultaneously emphasizes the value of multivariate analysis over univariate analysis. Adding variables to an analysis will hopefully provide more separation among the classes than additional noise. This feature space is not perfect for these classes, but clearly the trivariate feature space will outperform any one- or two-dimensional subspace of features.

The $\beta = 0.5$-contour has special appeal. This object is one candidate for a generalization of the median to more than one dimension. The β-contours may be obtained directly in the *ashn* software through option 18. When a value of β is entered, a simple line search is performed over α-contours to find the one with the correct probability mass (to four significant digits).

The image from which the remote sensing data were extracted is more complex. There are at least six significant crops identified in the ground truth. (Approximately half of the acre-sized pixels are identified as "mixed" and represent boundary pixels.) The median contour, $\mathcal{C}_{0.5}$, is displayed for all six averaged shifted histograms. The impression of the quality of the feature space is changed radically. It is interesting to see that spring oats are located in the middle of the spring wheat and spring barley portion of the feature space. The peanut contour in the bottom middle — well — really looks appropriate. The lower half of the sunflower contour is hidden by the β-contour of sugar beets. Of special interest is the disjointed nature of the β-contour for soybean crops. Given a random sample of a relatively homogeneous crop such as soybean, the non-connectedness of the contour is rather unexpected. A standard use of such "training data" is to fit a discriminate function, such as Fisher's linear discriminant function. Implicit in that parametric model is that the covariance matrix for all the groups is the same. Color Plate III makes a clear impression that not only does the size of the covariance matrices vary from class to class, but also the orientation of the major axes of the covariance matrices. Nonparametric analysis and visualization plays a diagnostic function, as well as a basis for a nonparametric classification algorithm.

1.5.2 Time Series Data

The study of univariate time series leads naturally to multivariate analysis as a means to understand the autocorrelation structure in the data. Univariate data, x_t, become the bivariate data, (x_t, x_{t+1}), and the trivariate data, (x_t, x_{t+1}, x_{t+2}). Nonparametric density estimation algorithms may be applied to correlated data, although care must be taken with regard to selection of the smoothing parameter.

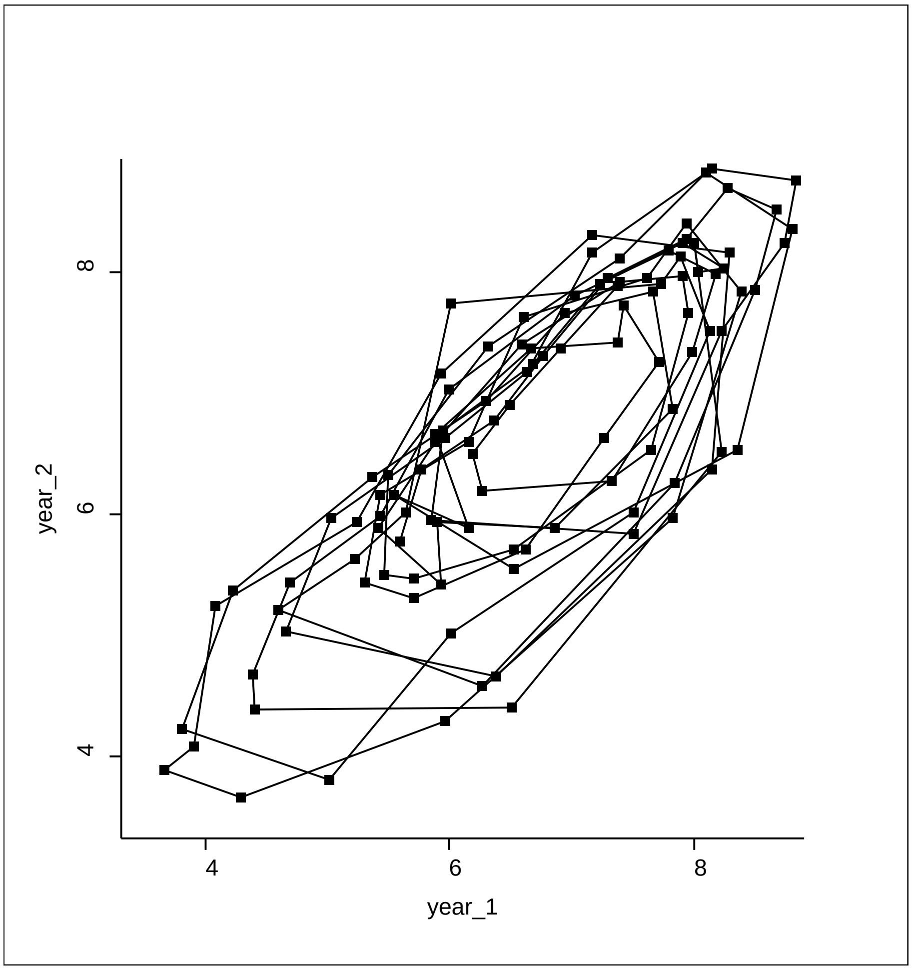

Fig. 1.4 Lagged values, (x_t, x_{t+1}), of the log-lynx time series from 1821 to 1934. Successive points are connected by line segments.

For a short example, consider the annual Canadian lynx trapping data from

1821 to 1934 (Elton and Nicholson, 1942), which is reproduced in Splus. A time series plot (not shown) shows a fairly regular sinusoidal pattern. The lagged data transformed to a logarithmic scale, (x_t, x_{t+1}, x_{t+2}), were fed into XGobi. A scatter plot of the first two dimensions is shown in Figure 4. Without the line segments connecting successive points, a fairly standard normal density is apparent. However, the flow of data is not random, but proceeds in clockwise fashion with a 9–12 year period to complete a cycle. A number of density estimates were examined informally, and none conformed to the donut shape apparent.

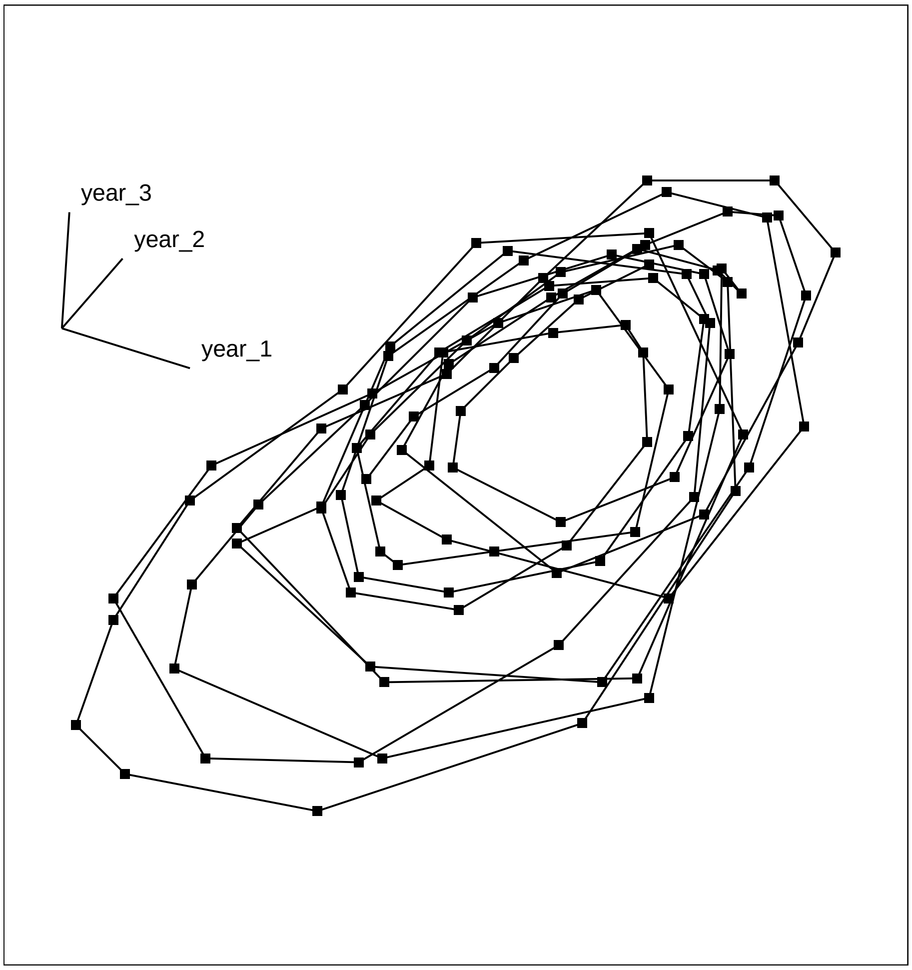

Fig. 1.5 Lagged values, (x_t, x_{t+1}, x_{t+2}), of the log-lynx time series from 1821 to 1934. Successive points are connected by line segments.

Viewing all three dimensions in XGobi and rotating gives the view shown in Figure 5. Rotating this point cloud reveals that the data have little variation in the orthogonal direction to the plot. The "hole" in the data is also more prominent. Indeed, the *ashn* view of these data (see Color Plate IV) shows the hole quite clearly now. It is instructive to look at the difference in the diameters of this $\alpha = 0.25$ contour at several cross-sections. The wider cross-section at around 10:30 is the result of more (not less data) concentration. (That is, the density function gets much higher inside this region.) The lowest density cross-section occurs at about 4:30. The orientation of the views in Figure 5 and Color Plate IV are nearly identical, so a comparison is possible. The extra separate little mass at 5:00 can be traced to the four points that are quite close there, resulting in a little local peak in the estimate. Notice that the density is not to taken as zero outside the red surface. Rather than density estimate goes smoothly to zero. There is one strong mode, located in the middle of the mass at around 1:00.

1.5.3 Quadravariate Functionals And Beyond

Trivariate contouring is a powerful tool, but not radical. Biochemists have used this tool to represent "real" pictures of molecules and bonding forces. Furthermore, trivariate contours live in our natural dimension of space. However, with four variables and beyond, we must leave the comfort of our real world and try to do the best we can at understanding the structure in higher dimensions.

In fact, we have already proposed going beyond three dimension earlier by drawing several trivariate contours at different levels and making the outer surface transparent in order to visualize all contours simultaneously. Just as the set of nine bivariate contours took us part way to the full three-dimensional function surface, so does the set of three to five trivariate contours take us part-way to the full four-dimensional function surface of $\hat{f}(x, y, z)$. These surfaces must be distinguished, and color provides a simple (discontinuous) manner to do so. By using lighting models, we can have gradual shading of a single color over a single contour, but different colors are used to indicate different α or β values.

Since the only other major cue left is motion (or animation), we propose to use that in a major role for visualization the quadravariate density $\hat{f}(x, y, z, t)$. By varying the fourth variable, t, we obtain a smoothly varying density surface, and hence, smoothly varying density contours as well. The grand tour, which is comprised of many discontinuous points, takes advantage of the continuity of the trajectory of each individual point during the grand tour. Earlier slicing algorithms (in MacSpin, for example) were less effective as points jumped in and then out of the display. Continuity is the most important link for viewing beyond three dimensions.

Thus we propose to make a movie of $\hat{f}(x, y, z, t)$ by choosing three to five *fixed* levels of α for contours. Each frame of the movie will have a slightly different value for the fourth variable, t, and the movie will show continuous and smoothly deforming contours.

As a specific example, consider the 500 epicenters of earthquakes preceding the 1982 eruption of Mount St. Helens (Weaver, et al., 1983). These data do follow a space-time format, with the depth variable, z, measured in log-units, and the time variable spanning five weeks before the eruption. In Color Plate V, a sample of nine of thirty frames in the movie are display from left to right and then down. The z-axis or depth is properly oriented. Examining the bottom set of frames first, three clusters of earthquakes are apparent. Given the logarithmic depth transformation (see Scott, 1992, for more details), the highest cluster is near the surface, while the middle cluster is about 1-2 km deep, and the bottom cluster approximately 6–8 km deep. Note that the same absolute α levels are maintained across the animation. Thus it is not until the fourth frame that the red contour level is pierced, and then the sixth frame for the green level. The initial three frames show low intensity, but an increasing intensity, of earthquake activity. The deepest earthquakes appear first, and a close examination of the progression from frame four to the final frame nine reveals a surge in activity at the middle then top levels, with the deepest activity tailing off.

It is important to note that the animation is *not* of the conditional density, $\hat{f}(x, y, z|t)$, but rather its unnormalized form, $\hat{f}(x, y, z, t = t_k)$. With the normalization implicit in conditioning, the overall level of the function is missing, and the animation less effective. In particular, the animation of Color Plate V would completely miss the dramatic conclusion. If the upper left frame is normalized by conditioning, then it will contain contours at the red and green levels. However, there is almost no data and the estimates would be useless. In fact, the blue contours might reasonably approximate the green level if conditioning were applied.

Thus with the use of color for the three different contour levels, and animation associated with a fourth variable, we are substantially along the way to visualizing a five-dimensional function surface. The data cave holds the promise of introducing other senses into the mix, including smell, sound, and touch (changes in air pressure and wind). Each cue is less precise than those employed here, but the goal of higher dimensional function visualization is not precision but rather overall understanding of the structure and its position in the full dimensional space.

At an ASA data exposition in 1986, we attempted a visualization of a five-variable density, $\hat{f}(x, y, z, u, v)$ in the following manner. A 7×7 grid in the (u, v) space was constructed by looking at the bivariate ASH, $\hat{f}(u, v)$. Then 49 six-dimensional ASH surfaces were estimated using *ashn*, and figures depicting $\hat{f}(x, y, z, u = u_i, v = v_j)$ such as those in Color Plate V were placed at the appropriate position in (u, v) space. Actually, 49 photographs were made and pinned into a 7×7 array. The data had had a chunk removed by David Coleman, and points forming the word EUREKA inserted. The density estimate was able to detect the hole in the data (by an unusual nesting of contours—see Scott, 1992), but, of course, could not spell or detect the word inserted in the hole.

For six-variable data, this dual visualization frame and reference frame model will still work (especially in a data cave). Appropriate paths in the three-dimensional reference space are not obvious. Without such deliberate paths, one could resort to random paths, much as the grand tour does, and we then return to our earlier discussion of the density grand tour. But extreme detail cannot be the goal of visualization beyond four to six dimensions. Thus projection algorithms, including projection pursuit, to take us to three to five dimensional space should be explored more carefully.

In fact, the remote sensing Landsat example represents such a careful projection, in that case from 24 to 3 dimensions. The 24 original dimensions were built up by combining six Landsat IV satellite images taken during the summer of 1977. Each image was four dimensional, as the early Landsat passive sensors recorded measurements in four spectral bands. An agronomist (Badhwar, et al, 1982) spent two years constructing a three-parameter growth model that was applied to the 24-dimensional data pixel by pixel. While statistical techniques such as principal components and projection pursuit can often accomplish the same goal much faster, most scientists resist the use of purely "statistical" techniques in favor of models that incorporate the known science in the experiment. The original data are probably lost, but it would be interesting to see how modern statistical algorithms would compare to the work of one knowledgeable scientist.

1.6 VISUALIZATION OF REGRESSION FUNCTIONS

We conclude this chapter by briefly indicating how the visualization techniques which have primarily been applied to density functions can apply equally to regression functions. For simplicity, we do this in the setting of an experimental design in three variables. We consider a linear and quadratic response function fit to data collected over a cube according to some design.

If we restrict our visualization to the cube itself, and do not extrapolate outside the support of the data, we have some interesting results. Of course, with a linear model, the best response will occur in one corner of the cube. The level sets of a linear function are just planes. In Color Plate VI, we show three contours of an example of a linear response function. It perhaps serves to remind us of the shapes possible when intersecting a plane with a solid cube. (The hexagon always seems most surprising.)

A more interesting model to consider is a fully quadratic fit. The example shown in Color Plate VII turned out to have a definite rather than indefinite covariance, with an optimal setting at the middle of the green contour. In fact, all of the contours are spherical in this parametric model, but again the three contours shown were truncated to the cube in which data were measured.

Similar pictures may be constructed using nonparametric regression algorithms, including local polynomial algorithms. Such displays do not offer a certain precision in location that is an important component of regression predictions. That can be obtained by plotting low-dimensional slices of the regression function, much as proposed above for higher dimensional density visualization. At least one program, coplot in the trellis package found in Splus, presents a conditioning view of a response surface. Trellis is based on the ideas of Cleveland (1993).

1.7 SUMMARY

In summary, we have presented a number of examples and algorithms for visualizing data and functionals of data in one to five dimensions. Most nonparametric modeling takes place in one and two variables, and is important. The ability to attempt to extend that experience to the next several dimensions more easily is at hand, both with regards to software and inexpensive hardware. Presentation is an important part of science, and we can learn much about effective presentation in several variables by collaborating with our colleagues in the sciences, computer science, and psychology. All software for generating these examples is available from the author at scottdw@rice.edu.

Acknowledgments

This research supported in part by the National Science Foundation under grant DMS-9306658 and the Department of Defense under contract MDA 904-95-C-2203.

REFERENCES

1. Asimov, D., "The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM J. Sci. Statist. Comp.* **6**, 128–143, 1985.

2. Badhwar, G.D., J.G. Carnes, and W.W. Austin, "Use of Landsat-Derived Temporal Profiles for Corn-Soybean Feature Extraction and Classification," *Remote Sensing of Environment* **12**, 57–79, 1982.

3. Carr, D.W. and W.L. Nicholson, "Explor4: A Program for Exploring Four-Dimensional Data," in E.J. Wegman and D.J. Depriest, eds., *Statistical Image Processing and Graphics*, Marcel Dekker, New York, 1986, pp. 215–250.

4. Carr, D.B., A.R. Olsen, and D. White, "Hexagon Mosaic Maps for the Display of Univariate and Bivariate Geographical Data," *Cartograph. Geograph. Information Systems* **19**, 228–231, 1992.

5. Chernoff, H., "The Use of Faces to Represent Points in k-Dimensional Space Graphically," *J. Amer. Statist. Assoc.* **68**, 361–368, 1973.

6. Cleveland, W.S., *The Elements of Graphing Data*, Wadsworth, Monterey, CA, 1985.

7. Cleveland, W.S., *Visualizing Data*, Hobart Press, Summit, NJ, 1993.

8. Cook, D., C. Cruz-Neira, B.D. Kohlmeyer, U. Lechner, N. Lewin, L. Elson, A. Olsen, S. Pierson, and J. Symanzik, "Exploring Environmental Data in a Highly Immersive Virtual Reality Environment," *Inter. J. Environ. Monitoring and Assessment*, **51(1-2)**, 441–450, 1997.

9. Donoho, A.W., D.L. Donoho, and M.Gasko, "MacSpin: Dynamic Graphics on a Desktop Computer," in W.S. Cleveland and M.E. McGill, eds., *Dynamic Graphics for Statistics*, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1988, pp. 331–351.

10. Elton, C. and Nicholson, M., "The Ten-Year Cycle in Numbers of Lynx in Canada." *Journal of Animal Ecology* **11**, 215–244 (1942).

11. Friedman, J.H. and J.W. Tukey, "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions in Computers* **C-23**, 881–890, 1974.

12. Hall, P. and M.P. Wand, "On the Accuracy of Binned Kernel Density Estimators," *The Journal of Multivariate Analysis*, **56**, 165–184, 1996.

13. Härdle, W. and D.W. Scott, "Smoothing in Low and High Dimensions by Weighted Averaging Using Rounded Points." *Computational Statistics* **7**, 97–128, 1992.

14. Huber, P.J., "Projection Pursuit (with Discussion)," *Ann. Statist.* **13**, 435–525, 1985.

15. Inselberg, A., "The Plane with Parallel Coordinates," *The Visual Computer* **1**, 69–91, 1985.

16. Loftin, R.B., "NASA/UH Virtual Environment Technology Laboratory," Department of Computer Science, University of Houston, Brochure, July, 1997.

17. Lorensen, W.E. and H.E. Cline, "Marching Cubes: A High Resolution 3D Surface Construction Algorithm," *Computer Graphics* **21**, 163–169, 1987.

18. Nason, G., "Three-Dimensional Projection Pursuit," *Applied Statistics* **44**, 411–430, 1995.

19. Salch, J.D. and D.W. Scott, "Data Exploration with the Density Grand Tour," *Statistical Graphics and Computing Newsletter*, ASA, Alexandria, VA, **8**, 7–11, Summer, 1997.

20. Scott, D.W., "Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions," it Ann. Statist. **13**, 1024–1040, 1985.

21. Scott, D.W., "Data Analysis in 3 and 4 Dimensions With Nonparametric Density Estimation," in E.J. Wegman and D. DePriest, eds., *Statistical Image Processing and Graphics*, Marcel Dekker, New York, 1986, pp. 291-305.

22. Scott, D.W., "Multivariate Density Estimation: Theory, Practice, and Visualization," John Wiley, New York, 1992.

23. Scott, D.W., "Incorporating Density Estimation Into Other Exploratory Tools." *Proceedings of the Statistical Graphics Section*, ASA, Alexandria, VA, 1995, pp. 28–35.

24. Scott, D.W. and G. Whittaker, "Multivariate Applications of the ASH in Regression," *Communications in Statistics* **25**, 2521–2530, 1996.

25. Silverman, B.W., "Algorithm AS176. Kernel Density Estimation Using the Fast Fourier Transform, *Appl. Statist.* **31**, 93–99, 1982.

26. Swayne, D., D. Cook, and A. Buja, "XGobi: Interactive Dynamic Graphics in the X Window System with a Link to S," *ASA Proceedings of the Section on Statistical Graphics*, ASA, Alexandria, VA, 1991, pp. 1–8.

27. Tufte, E.R., *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT, 1983.

28. Tukey, J.W., *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

29. Wainer, H., *Visual Revelations*, Springer-Verlag, New York, 1997.

30. Weaver, C.S., J.E. Zollweg, and S.D. Malone, "Deep Earthquakes Beneath Mount St. Helens: Evidence for Magmatic Gas Transport?" *Science* **221**, 1391–1394, 1983.

31. Wegman, E.J., "Hyperdimensional Data Analysis Using Parallel Coordinates," *J. Amer. Statist. Assoc.* **85**, 664–675, 1990.

32. Whittaker, G. and D.W. Scott, "Spatial Estimation and Presentation of Regression Surfaces in Several Variables Via the Averaged Shifted Histogram," *Computing Science and Statistics* **26**, 8–17, 1994.

33. Wolff, R.S. and L. Yaeger, *Visualization of Natural Phenomena*, Springer-Verlag, New York, 1993.

Appendix: Color Plate Captions

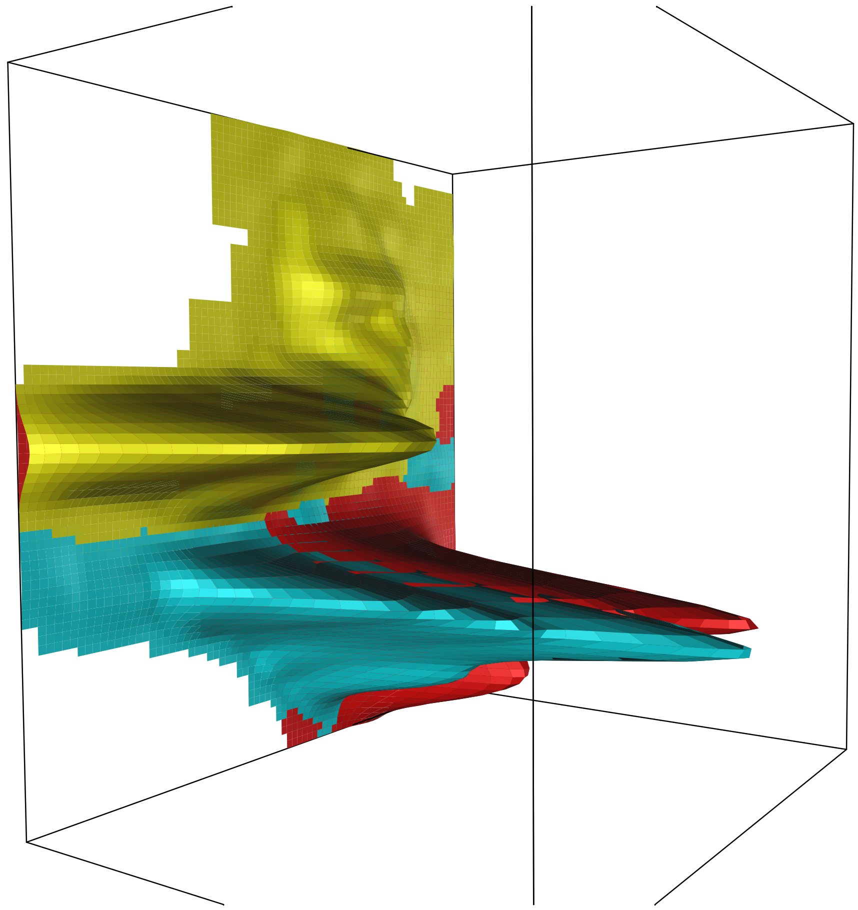

Fig. A.1 Color Plate I. Perspective plots for the X-Y variables of three estimated densities for each of the three crops in the remote sensing example (yellow=sunflower; red=wheat; blue=barley). At each point, (x, y), the greatest of the three density values is displayed.

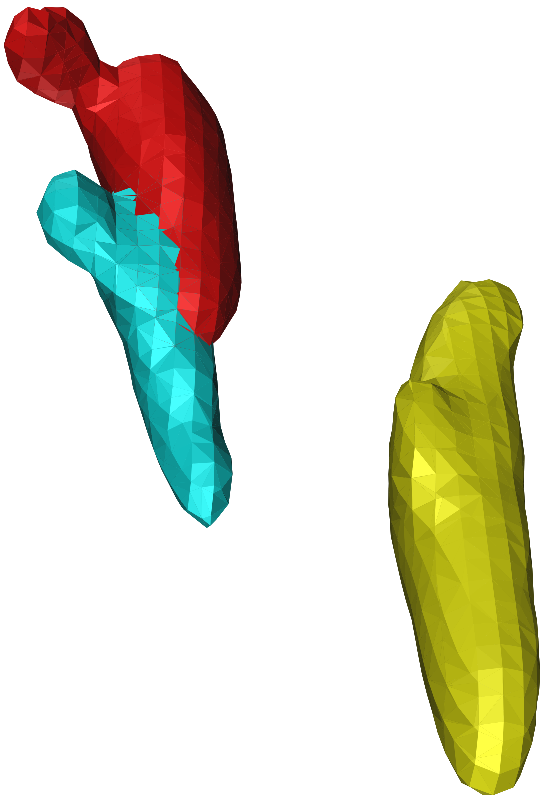

Fig. A.2 Color Plate II. For the remote sensing data, three β-contours with $\beta = 0.5$ of averaged shifted histograms of sunflower (yellow), wheat (red), and barley (blue) crops.

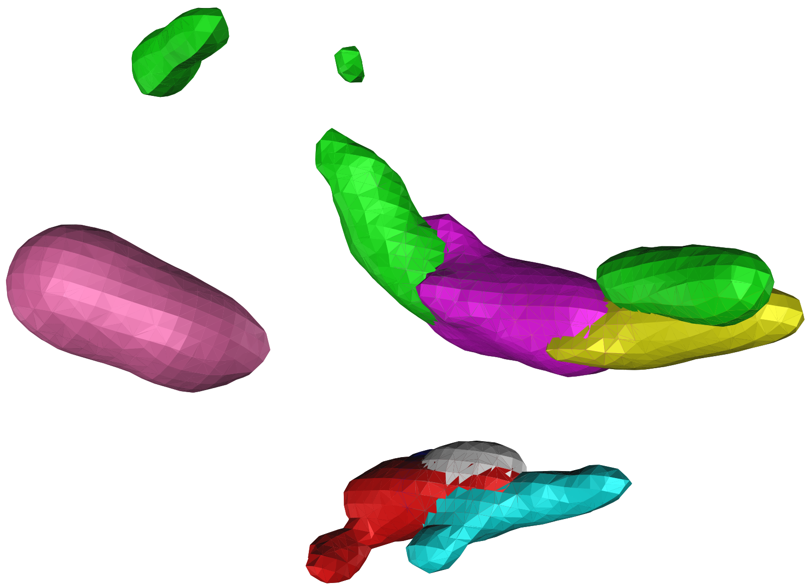

Fig. A.3 Color Plate III. For the remote sensing data, $\beta = 0.5$-contours of the six major crops in the image. In addition to sunflower (yellow), wheat (red), and barley (blue) crops, we have soybeans (green, $n = 731$), sugar beets(magenta, $n = 506$), spring oats (white, $n = 459$), and ¡peanuts (pink, $n = 306$).

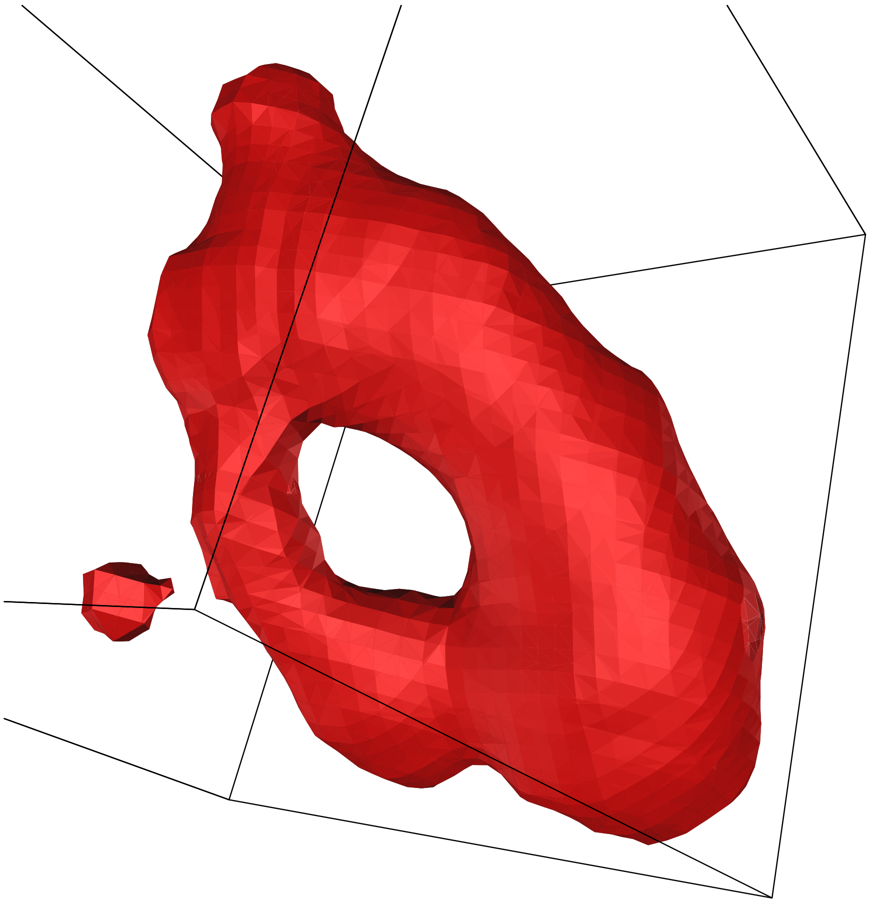

Fig. A.4 Color Plate IV. A trivariate contour of the lagged log-lynx data at the $\alpha = 0.25$ level.

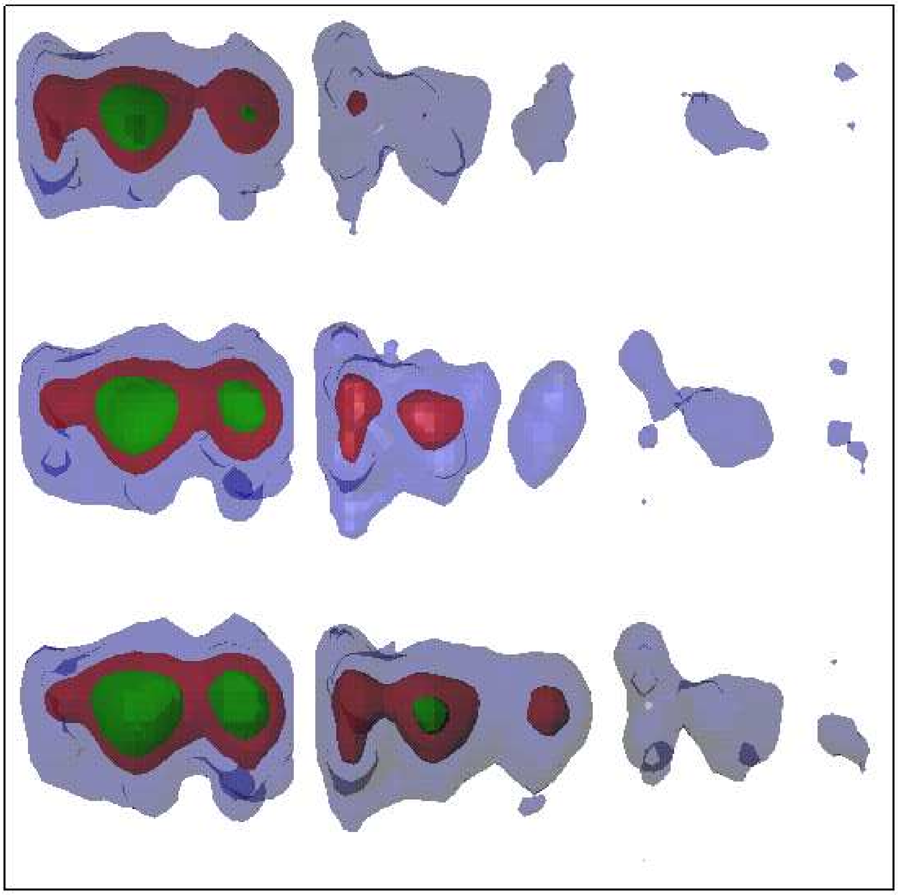

Fig. A.5 Color Plate V. Nine frames from a movie of the quadravariate density estimate, $\hat{f}(x, y, z, t)$, of the Mount St. Helens data. Time moves across and then down in this figure. Three α-contours are displayed at each time slice, with $\alpha = 0.02, 0.10, 0.25$.

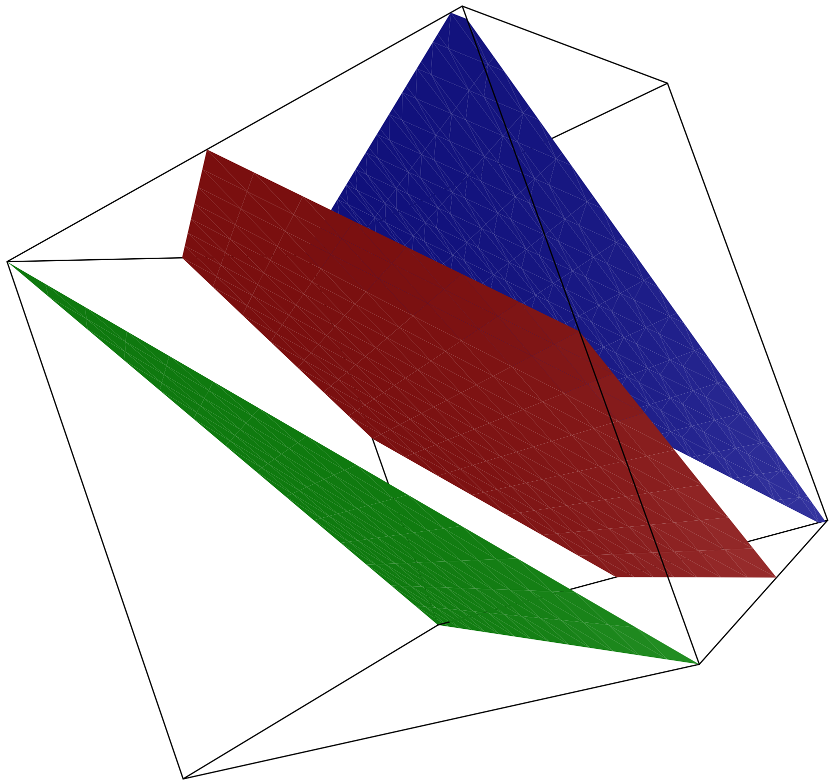

Fig. A.6 Color Plate VI. Three trivariate contours of a linear parametric model of experimental design data.

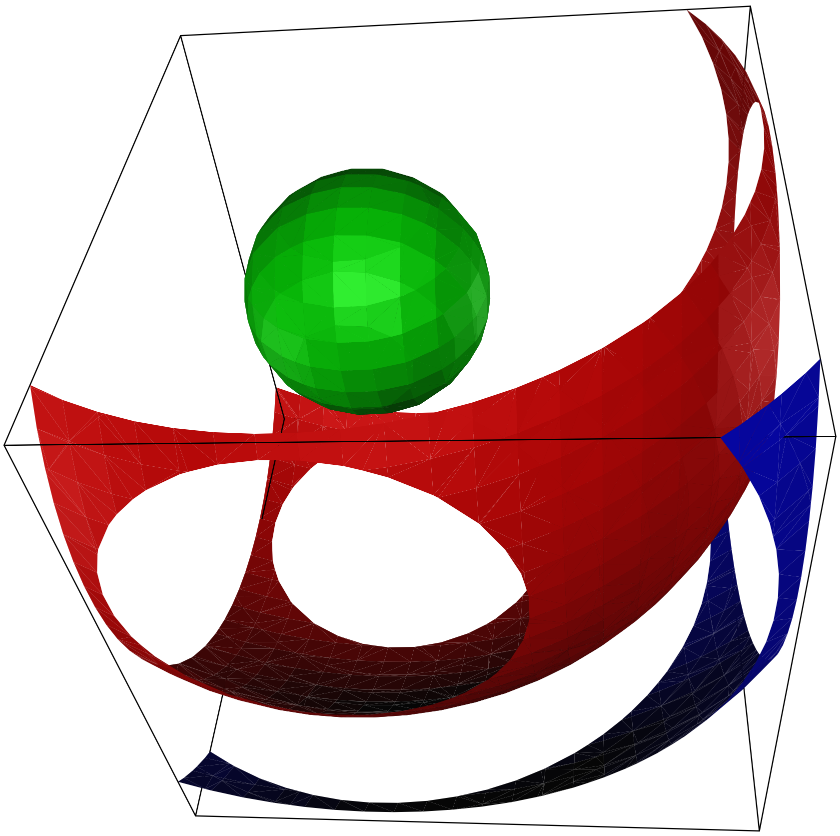

Fig. A.7 Color Plate VII. Three trivariate contours of a quadratic parametric model of experimental design data.