INCORPORATING DENSITY ESTIMATION INTO OTHER EXPLORATORY TOOLS

David W. Scott, Rice University
Department of Statistics, MS 138, POB 1892, Houston, TX 77251-1892

Key Words: Density Estimation, Exploratory Data Analysis, Grand Tour, Scatter Diagrams, Parallel Coordinates, Averaged Shifted Histogram

Abstract

Preliminary understanding of a new data set is routinely accomplished with graphical tools, such as those popularized originally by EDA. A number of more recent ideas for multivariate data analysis have emerged and some are available in software packages or shareware such as XGobi. In this talk, we illustrate how many of the point-oriented techniques can be supplemented by incorporating nonparametric density estimates. Examples from the grand tour to parallel coordinates to clustering will be presented. Potential advantages include visual simplicity, recognition of unusual structure, and handling an additional dimension.

1. Introduction

Research in statistical graphics is driven about equally by abstract ideas and by hardware innovations. This paper focuses on somewhat abstract ideas for which the latter is becoming relevant. Computers, software systems, and practicing statisticians are being challenged and stressed in the new data environment. Here, we discuss software solutions that will improve the efficiency of the human client and offer new capabilities.

The evolution of the central components of a graphics workstation has not been even. Screen resolution most rapidly "saturated," reaching the quite useful 1280 by 1024 pixel mapping available on most machines today. Higher resolutions are available, but there does not seem to be much pressure in the marketplace for such improvements, compared to the pressure when low resolution screens dominated the PC market. CPU speed has had an impressive but fairly constant rate of improvement.

After lagging behind for years, there has been substantial and rapid progress in the size of computer hard disks and CPU memory available. A Unix workstation can reasonably be outfitted with 10GB disk, 512MB CPU, and the serial power of a Cray. The existence of this class of workstation has encouraged the growth of on-line and very large data sets. Initial experience by workers trying to understand these *Massive Data Sets* using graphical tools has been surprisingly negative. The hardware bottleneck is caused primarily by the slow graphical I/O, as well as by the slow disk I/O.

Thus we are in the situation of having sufficient data to warrant extensive graphical exploration, but hardware and software which are not (yet) up to the task. Graphical I/O and advanced visualization have not been top priorities for most manufacturers, and progress has only been steady, not rapid. The challenges of Massive Data Sets require simultaneous improvements in graphics I/O and screen resolution.

It is doubtful that computer monitors will ever have significantly more than several millions of pixels, and direct exploration of a billion data points will require compromise and innovation. One compromise is the selection of subsets of data, either randomly or by choosing chunks of time. One innovation is to build and design nonparametric tools into the software packages that currently support EDA. This paper describes how *Multivariate Density Estimation* (Scott, 1992) is well-suited for this task. Some items on our wish list of new tools cannot be accomplished with current generation hardware. Some will require multiple processors with tightly synchronized output to frame buffers for smooth animation viewing.

2. Existing Tools

Interactive tools for exploring multivariate data are widely available, but none is more attractive or complete than XGobi (Swayne, Cook, and Buja, 1991). This one well-designed program incorporates a number of features and techniques: jittered dot plots for univariate data; pairwise scatter plots for bivariate data; rotating scatter plots for trivariate data; and the grand tour (Asimov, 1985; Buja and Asimov, 1986) for higher dimensional data. Under the grand tour option, automated search for 2-D structure in the data can be accomplished by projection pursuit (Friedman and Tukey, 1974; Cook, Buja, and Cabrera, 1993). Exploration is aided by identification tools such as brushing, grouping, subsetting, and identification. Subsets of points can be highlighted through color, size, or shape of glyph (Cleveland, 1993).

Of course, it is easy to suggest enhancements to the basic XGobi features. For example, the univariate dot plots are only available one at a time — having side-by-side dot plots would be a nice addition. Also missing is the scatterplot matrix option. A scatterplot matrix of selected variables might also allow the user to select a single pairwise scatter diagram by pointing and clicking with the mouse (an easier interaction sequence than the current method of selecting variables manually or stepping through all pairs). More challenging enhancements might include trivariate projection pursuit options to feed into the rotation tool.

XGobi has recently added the ability to draw lines in addition to points. Thus elementary graphs or maps can be included. One alternative to scatter diagrams that uses lines is the parallel coordinate plot (Wegman, 1990)—a tool that might usefully be included.

A very advanced feature allows multiple XGobi windows to communicate with each other, and synchronize their actions. This feature facilitates general types of "linking" such as brushing in one window while viewing other variables in a second window (or even in three windows).

3. Smoothing Scatter Diagrams

XGobi provides a useful model for an exploratory tool. The idea of incorporating density smoothing is not new. In fact, Hurley and Buja (1990) demonstrate the ability to display the averaged shifted histogram (Scott, 1985) in XGobi. The ASH bins the data and applies a local convolution smoothing. The output is a matrix of the form $(t_i, f(t_i))$ where i ranges over the number of bins and $\{t_i\}$ are points at the bin centers. The ASH points (or lines) can be displayed. This feature is not implemented in the current XGobi release.

A different use of density smoothing was demonstrated by Stuetzle (1987) who suggested linking to a histogram, in which every data point is a "building block" stacked in bins that can be brushed. For our purposes, the smoothness of the ASH is relevant, as we propose to view sequences of ASHs based on grand tour projections. The ASH provides an essentially continuous view of the data as they are being rotated. As noted by Hurley and Buja, even small changes in the 1-D projection angle can result in distracting jumps in the histogram.

The term "scatterplot smoothing" can be applied equally to regression data (Cleveland, 1993) and non-regression data (Scott, 1991). The data we have in mind do not have a response variable, and hence the latter (but less common) meaning is assumed.

We have argued that the scatter diagram points to the density plot, especially as $n \to \infty$ (Scott and Thompson, 1983). Scott (1992) has discussed in detail the use of the ASH for visualizing data in 1, 2, 3, and more dimensions. We have found the bivariate ASH in particular an extremely effective tool for visualizing data coming out of the grand tour. The computational horsepower of today's machines is pushed to the limit while attempting to display an evolving bivariate ASH in real time. However, we are convinced that the contour surfaces of the 3-D ASH will eventually provide an even more compelling view of data coming out of the grand tour. However, the Silicon Graphics (SGI) computer we used for this work (Model 310GTX with hardware transparency) can barely handle one frame at the bin resolution we desire. Thus our 3-D examples are all oversmoothed to permit some semblance of the animation. We call such a sequence the *density grand tour*.

4. Visualizing Multivariate Densities

Let $\hat{f}(\mathbf{x})$ denote the ASH density estimate, which is really an array of values over a mesh. Depending on the dimension, we can view $\hat{f}(\mathbf{x})$ directly (graph or perspective plot) or indirectly through its contours (2 or 3 dimensions). We identify contours of $\hat{f}(\mathbf{x})$ as α-shells denoted by S_α; specifically, S_α is the set of points $\mathbf{x} \in \Re^d$ satisfying

$$S_\alpha = \{\mathbf{x} \in \Re^d \mid \hat{f}(\mathbf{x}) = \alpha \hat{f}(\mathbf{x}_m)\}$$

where $\alpha \in (0,1)$ is a real number and $\mathbf{x}_m$ is the global mode

$$\mathbf{x}_m = \arg\max_{\mathbf{x}} \hat{f}(\mathbf{x}).$$

For example, if $\hat{f}$ is a trivariate normal density, then an α-shell is an ellipsoidal solid (surface). As α varies, only the size and not the shape of the ellipsoid changes.

For $\mathbf{x} \in \Re^1$ or $\Re^2$, we graph the function $\hat{f}(\mathbf{x})$ itself, rather than its contours. However, in the bivariate case, if the ASH function surface is color-encoded by height, then viewing the density directly from above reveals the α-contours; see Scott and Salch (1995).

5. Examples of the Density Grand Tour

We implemented a prototype of the DGT in Splus, using the Geomview software on an SGI for visualization. We re-coded the ordinary grand tour (GT) so that we could do scatter diagrams and ASHs on the same projection views. Our GT follows the full-dimensional implementation of Wegman, Carr and Luo (1993), so that we could examine 3-D ASHs as well. Note that Wegman can display all the grand tour variables with parallel coordinates.

5.1. One-Dimensional Density Grand Tour: Iris Data

Figure 1 displays a subset of the 1-D DGT of the iris data at three different "times." The data have been subjected to a one-time jittering in the *vertical* direction, so that points move only left or right during the GT. XGobi displays dot plots vertically, but we choose a horizontal display to parallel the ASH x-axis. The corresponding DGT ASHs are shown to the right of the data point frames. Discerning the number of modes and clusters in the jittered dot plot is facilitated by examination of the corresponding ASH. The number of groups in the data is not always as clear as these three frames suggest. Clearly the jittered dot plot and ASH should (and can) be displayed simultaneously if n is not too large. For large n, only the ASH presentation is feasible.

A one-dimensional scatter diagram does not hold much promise, and the expectation for the ASH view may not be much greater. However, there is a certain fascination watching the ASH undulate while simultaneously watching points sliding back and forth across the screen.

Of course, the search for multivariate *structure* requires multivariate views. But the one-dimensional views allow for discovery of very simple grouping structure. Unless the multivariate structure in each group is likely to be substantially similar, subsetting at this time is probably a prudent strategy at some stage of the exploration process.

5.2. Two-Dimensional Density Grand Tour: PRIM7 Data

The PRIM7 data have much structure, as indicated in Figure 2. These data were used by Friedman and Tukey (1974) to illustrate their work on projection pursuit. The view at $t = 64$ shows most of the data piled into one clump with two "arms." A second view at $t = 2$ very effectively shows three clusters falling roughly in a triangular shape. All four frames are at the same viewing angle, but the scaling for the points and ASHs is not exactly the same.

Friedman and Tukey found a Z-shaped structure in these data. Jee (1987) used a different PP criterion and obtained a more compelling view of a triangle with points clustered in the vertices even more pronounced than in Figure 3 at $t = 2$. Cook, Buja, and Cabrera (1993) found pairs of whiskers at each vertex.

The eye is very capable of discerning discrete structure or grouping, even if it is changing rapidly. The eye is much less successful as discriminating based on density or concentration of data. The height (or color) of the ASH is invaluable for this secondary goal; compare the ASH and scatterplots again at $t = 64$.

Notice that the ASH estimates tend to be somewhat oversmoothed in this figure. Users should have the smoothing parameter at their control, together with a history replay mechanism to dig deeper into recent views for golden nuggets of structure and information.

5.3. Three-Dimensional Density Grand Tour With Brushing: Iris Data

One frame of the data in a GT is shown in the upper left frame of Figure 3. By brushing on the species indicator variable, three ASH contours (all at the same α-level) are shown, one for each species; see the upper right frame of Figure 3. This view is in contrast to the corresponding view (same projection) without the labeling information; see the lower left frame in Figure 3. The first ASH shows the *Setosa* variety quite clearly, but at a higher α level, the exploratory view displays 4, not 3 modes; see the lower right frame in Figure 3. The weight of evidence in favor of the 4th mode can be determined to be slight, as views of the data in nearby projections show the presence of a long and narrow ridge in the bigger contour, that sometimes shows 1, 2, or 3 modes, but never at much "depth." By depth, we mean the height of the density halfway between modes. Not much depth is observed at any projection.

6. Controlling the Density Grand Tour

While a number of design issues go into the construction of an ASH, the basic controlling knob goes to the smoothing parameter $m = h/\delta$; here δ is the binning resolution and h is the usual kernel bandwidth. Our examples in the previous section clearly indicate that useful views of the data are provided even when h is far from "optimal." Nevertheless, a user should have the ability to control h in real time.

Color is a relatively crude marker (as anyone who has tried to use color to brush scatterplots with large n and several groups knows). We have found it useful in two different ways. The first is to differentiate contour levels (α-shells in the 3-D case) or density heights (in the 1-D and 2-D cases). This simple device makes it very easy to spot the number and location of high-density regions by focusing on one color.

The second use of color is for identification when grouping. This is most similar to the ordinary use

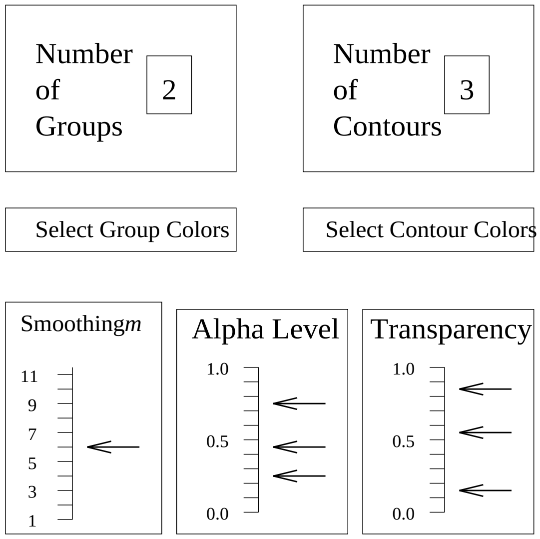

Figure 1: Sample Control Panel For the Density Grand Tour.

of color when brushing scatter plots. Here, each group's density estimate is colored uniquely for identification. The question of the best (or most effective) use of color when multiple α-shells and groups are desired is not fully answered. Clearly, things can get very busy on the screen if there are lots of groups and several α-shells displayed.

A full control panel can be imagined. There would be sliders for the number of α-shells; a slider for the values of each α level; menus for the color of an α-shell and group; and finally, a slider for transparency of each α-shell. An example control panel is shown in Figure 4.

7. Discussion

As with all graphical devices, experimentation is the true test. One can contemplate even more esoteric objects such as a 4-D DGT, which would display an 3×3 or 4×4 array of 3-D DGT slices going along. This idea can be implemented as with the 3-D DGT by having have 9 or 16 times as many single processors, each running an ASH and providing a single view to an output device.

Objects that allow brushing of bivariate histograms might be a natural extension of brushing scatter diagrams. Scott (1988) examined the theoretical benefits of using hexagonal bins, while Carr (1990) found hexagonal bins graphically appealing.

A more sophisticated idea would be to replace the brushing tool (typically a rectangle) by a "kernel," whose location is controlled by the mouse, and whose shape can be set by mouse interactions beforehand (as is the size and shape of the rectangular brush in XGobi currently). This would provide "density brushing" as opposed to simple grouping. A conditional density might be the output in a coupled window.

With many dimensions, the grand tour is exhausting before all interesting directions can be exhausted. Projection pursuit is attractive for low dimensions, but new technology such as Terrell's informative components analysis (Terrell, 1985) provides endless possibilities for subspaces to tour, as well as ending subspaces. We plan to examine such ideas and look forward to ever more powerful graphical exploration tools capable at the high end.

Acknowledgments

Paper presented at the Annual Meetings of the ASA, Orlando, Florida, August 15, 1995. The author would like to thank John D. Salch for assisting in the creation of the video tape and figures in this paper, and Keith Baggerly for his comments.

David W. Scott is Professor, Department of Statistics, Rice University, Houston, TX 77251. This research was supported in part by the National Science Foundation under grant DMS-9306658 and the National Security Agency under grant MOD 9086-93.

References

Asimov, D. (1985), "The Grand Tour: A Tool For Viewing Multidimensional Data," *SIAM J. Scient. Stat. Comp.*, **6**, 128-143.

Buja, A. and Asimov, D. (1986), "Grand Tour Methods: An Outline," *Comp. Sci. Stat: Proc 17th Symp. Interface*, pp. 63-67.

Carr, D.B. (1990), "Looking at Large Data Sets Using Binned Data Plots," PNL-7301, Battelle Labs, Richland, WA.

Cleveland, W.S. (1993), *Visualizing Data*, Hobart Press.

Cook, D., Buja, A. and Cabrera, J. (1993), "Projection Pursuit Indexes Based on Orthonormal Function Expansions," *J. Comp. Graphical Statist.*, **2**, 225-250.

Friedman, J.H. and Tukey, J.W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Trans. Comp.* C-**23**, 881-890.

Hurley, C. and Buja, A. (1990), "Analyzing High-dimensional Data With Motion Graphics," *SIAM J. Scient. Statist. Comp.*, **11**, 1193-1211.

Jee, J.R. (1987), "Exploratory Projection Pursuit Using Nonparametric Density Estimation," *ASA Proc. Statist. Comp. Sect.*, pp. 335-339.

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

Scott, D.W. (1988), "A Note on Choice of Bivariate Histogram Bin Shape," *J. Official Statistics* **4**, 47-51.

Scott, D.W. (1991), "On Estimation and Visualization of Higher Dimensional Surfaces," In *IMA Computing and Graphics in Statistics*, P. Tukey and A. Buja, Eds., pp. 187-205, Springer-Verlag, New York.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York, John Wiley.

Scott, D.W. and Salch, J.D. (1995), "Demonstration of the Density Grand Tour With Brushing Extensions," Video Tape, Department of Statistics, Rice University, Houston.

Scott, D.W. and Thompson, J.R. (1983), "Probability Density Estimation in Higher Dimensions," *Proc. 15th Interface*, J.E. Gentle, ed, pp. 173-179, North-Holland.

Stuetzle, W. (1987), "Plot Windows," *J. Amer. Statist. Assoc.*, **82**, 466-475.

Swayne, D.F., Cook, D., and Buja, A. (1991), "XGobi: Interactive Dynamic Graphics in the X Window System with a Link to S," In *ASA Proc. Sect. Statist. Graphics*, pp. 1-8.

Terrell, G.R. (1985), "Projection Pursuit Via Multivariate Histograms," TR 85-7, Dept. Math Sciences, Rice University, Houston.

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

Wegman, E.J., Carr, D.B., Luo, Q. (1993), "Visualizing Multivariate Data," In *Multivariate Analysis: Future Directions*, pp. 423-466.