PROBABILITY DENSITY ESTIMATION IN HIGHER DIMENSIONS

David W. Scott and James R. Thompson
Rice University, Houston, Texas

For the estimation of probability densities in dimensions past two, representational difficulties predominate. Experience indicates that we should investigate the locations of the modes and proceed to describe the unknown density using these as local origins. The scaling system to be employed should also be data determined. Using such a philosophy, density estimation has been successfully carried out in the three dimensional case. Color and motion can be used as enhancement devices so that estimation in dimensions past three becomes feasible.

1. INTRODUCTION

Much of the work in nonparametric probability density estimation is oriented to pointwise estimation. At any point in R_n we can do a good job of estimating the density function and its derivatives if only the sample size is sufficiently large. Unfortunately, almost all of the supporting numerical investigations have been done in R_1 where even a rusty nail works reasonably well. As we increase the dimension of the domain of the density past 2, we find that asymptotic pointwise rates of convergence are at best matters of secondary importance. Issues of L_1, L_2, or L_∞ convergence in density estimation are overwhelmed in importance by representational considerations. It is hard to perceive that this is so when the domain of the density is R_1 or R_2, for these lend themselves to two and three dimensional representations, respectively (recall that the density function itself takes up a dimension), and for these dimensions the *empty space phenomenon* is usually not a problem.

But as we proceed to densities with domains in spaces of dimension higher than 2, we observe that if we impose a Cartesian mesh on the domain which is of width in each dimension roughly appropriate for the one dimensional marginal, then most of the multivariate parallelepipeds contain no sample points. Moreover, we lose, in great measure, our geometrical intuitions, which are largely three dimensional. Therefore, we cannot simply rely on computing a density function estimate on a grid chosen by dead reckoning, trusting in our visual perceptions somehow to get us out of the soup. We are forced, somehow, to face the task of estimating the density where it is relatively large. To some extent, then, the problem of density estimation in higher dimensions involves first of all finding where the action is.

Let us attempt to illustrate this point by an example. In the case of a one dimensional density, we are given the choice of two information packets:

A: a random sample of size 100

or

B: exact knowledge of the density on a equispaced mesh of size 100 between the .5 and 99.5 percentiles.

Most of us, most of the time, will choose B. Indeed the main thrust in nonparametric density estimation has been to use information of the A type to obtain a degraded version of B.

Now, if the density in question has domain R^4, then chopping each axis into a mesh with three nodes will exhaust 81 of the 100 mesh points available in B. Information packet B is now rather unattractive. A random sample of size 100, however, is much more useful. Since the data points themselves tend to come from those parts of R^4 where the density is high, they can be used as a means of focusing our attention for regions where the density should be estimated.

It is tempting, then, to go with the data set itself, to abandon the density estimation problem altogether, and stick with the scattergram. There are several reasons why we have chosen not to take this approach. First of all, we take some issue with those who would like to treat the scattergram sui generis. The scattergram will obviously change as sample size changes. The density function is the fixed entity to which the scattergram merely points. If it is argued that most data generating mechanisms are, in fact, somewhat nonstationary even over the time frame where the data is being collected, we agree. But this nonstationarity is seldom so great as to negate the practical value of looking for a density function model.

The second reason for going beyond the scattergram is that an unprocessed collection of data points does not fully exploit the natural continuities generally present in a data generating process. It is not advisable, in our view, to insist on software which assumes that a human observer is always to be available to smooth by eye a screen output, as in the case of a scattergram display. Naturally, in the formative stages, we will rely a good deal on using our visual perceptions to assist us in software creation. But our goal should not be cyborg data analysis (CDA). We hope to reach a point where our software does much more than interactively provide graphical displays to a human observer. The density function appears to have much greater potential for full automation than does the scattergram and its variants.

2. THE AVERAGED SHIFTED HISTOGRAM

The number of techniques for nonparametric density estimation is large. They include series estimators [1], kernel estimators [2,3,4], maximum penalized likelihood estimators [5,6,7], isotonic estimators [8]-- to name only a few. Some are much better than others. Some are bound to the one dimensional case. However, in this paper, we shall focus on a modified version [9] of the oldest density estimator--the histogram [10].

The usual histogram in R^1 is constructed by dividing the interval of significant density mass [a,b] into M intervals of length $h=(b-a)/M$. We then count, of the total of the N data points, the number which fall into the jth interval--say N_{1j}. Then the histogram estimator at each point x in the jth interval is

$$\hat{f}_{1H}(x) = N_{1j}/(Nh).$$

To form the averaged shifted histogram we shift the left hand end point to $a + h/m$, where m is the number of shiftings we wish to perform. We then carry out the usual histogram procedure on the shifted support [a + h/m, b + h/m] again using intervals of length h. This leads us to the second histogram estimator $\hat{f}_{2H}$. We continue in this manner until we have m histogram estimators $\hat{f}_{1H}, \hat{f}_{2H}, \ldots, \hat{f}_{mH}$. The averaged shifted histogram estimator at each point x in $[a+h, b]$ is given by

$$\hat{f}_{ASH}(x) = \frac{1}{m} \sum_{i=1}^{m} \hat{f}_{iH}(x).$$

Experience indicates that a m of 5 produces density estimates which are visually nearly indistinguishable from the best of the usual nonparametric density estimates. In Figure 1, we show a sequence of average shifted histograms as m doubles for a Chondrite data set [11] with $n=22$ and $h=2$. The improvement in statistical efficiency is apparent. The limiting estimator is, in fact, a kernel estimator.

Extensions of the shifted histogram to higher dimensions are obvious. The main advantage of the technique in several dimensions is its speed--typically 50 times faster than that of kernel techniques. The speed makes possible real time displays for handling density estimation in five dimensions when using a minicomputer such as the VAX 780.

To indicate the graphical representation of a density in higher dimensions, we note first that the two dimensional case is straightforward. We simply use the density for the vertical axis and the variables for the horizontal and depth dimensions.

Let us go then to the first representationally difficult case--that of three dimensional data. A trivariate data set may be viewed directly by a rotating scattergram or clever use of color to permit visualization of the third dimension on a two dimensional graphics terminal. The underlying smooth density function may be viewed as a thunderstorm where the density f(x,y,z) at a spatial point (x,y,z) gives the relative intensity of the storm (or the density of moisture, etc.). In other words, the thunderstorm gives the 3-D representation of the 4-D density surface f(x,y,z). Any particular contour level $f(x,y,z)=c$ gives a surface or disjoint surfaces in R^3. These contour surfaces are displayed by intersecting them with a series of equally spaced planes perpendicular to the x-axis (and occasionally the y-axis). No hidden line algorithms are used in these plots.

For quadravariate data, we extend the thunderstorm analogy by arbitrarily assigning one variable to represent time. Hence the contours of the five-dimensional density surface f(x,y,z,t) may be observed by time-lapse photography of the thunderstorm's progression from calm beginning through the height of the storm to its inevitable conclusion. Clearly, in a reactive environment, each of the four variables may be sequentially assigned to time.

3. Trivariate Example

The data in this section were obtained by processing remote sensing satellite data measured over North Dakota during the summer growing season of 1977 and were furnished by Dr. Richard Heydorn of NASA/Houston. The sample contains approximately 24,000 points, each representing a 1.1 acre pixel, over a 5 by 6 nautical mile region. On each Landsat satellite overpass, the 4 spectral reflectance intensities on an individual pixel were combined into a "greenness" variable. The greenness of a pixel was then plotted as a function of time and Badhwar's [12] growth model was fitted to these points. From this model three variables were extracted: the time of maximum greenness (x); the ripening period of the crop (y); and the value of the greenness at time x (z). In the left column of Figure 2, we display each of the three variables positioned over the 5 by 6 nautical mile region. Dark pixels represent relatively large values of the pixels which were actually drawn in color. Field structure is easily seen in the color versions. In the second column are histograms of the three variables and an arbitrary composite of the variables.

It is difficult to present a satisfactory picture of a three-dimensional scatter diagram of these data shown in Figure 3. Color was used to represent the value of the component orthogonal to the viewing plane. Notice that the data have been projected; the original axes are shown projected and at the origin. Most of the data are hidden in the picture, only 3000 points are displayed on this 512 by 512 screen.

In Figure 4 we display the two disjoint contours of the averaged shifted histogram with $m=2$ at density level 1% of the mode. Comparing Figures 3 and 4 which are drawn on the same scale we see how many of the points observable in the scatter diagram are outside the 1% contours. In Figure 3 our eyes are drawn away from the large majority of the data. The small cylindrical 1% contour represents the line of points behind the large cloud and largely represents fields of sugar beets (which are harvested by plowing rather than cutting). In Figure 5 we zoom in on the large 1% contour shape and plot the 20% contour which is bimodal. Also drawn within the front 20% contour is the single 50% contour. Many modes are apparent in these data, perhaps corresponding to different types of vegetation. On a color graphics terminal we may simultaneously view these and other contours by using different colors to draw each contour level. The density estimate was computed over a 30 by 30 by 30 mesh as the data were sequentially read from disk in less than a minute of real time. In comparison a straightforward kernel estimate over this mesh was projected to required several hours of CPU time.

4. Quadravariate Example

The 500 4-variable physics particle data have been analyzed by Friedman and Tukey using projection pursuit methods [13] and more recently by Tukey and Tukey using other multivariate graphical techniques [14]. Examination of a series of representations of scatter diagrams of three of the four variables suggests that the data lie predominately on two line segments in the shape of a "V" in four-space. For our thunderstorm analysis we first re-expressed the data variables x, y, z, and t in order to minimize skewness of the variables. The transformations were log(x), sqrt(log(1+y)), sqrt(log(1-z)), and sqrt(log(1-t)). After transformation the first variable was bimodal. For the thunderstorm analysis we let the fourth variable represent time and formed an averaged shifted histogram with $m = 5$ over a 30^4 mesh. In Figure 6 we display the 1%, 15%, 50%, and 80% contours (if any) at selected time points 9, 14, 19, 24, and 29 from the time lapse sequence. The axes are fixed in the same orientation as in Figures 3-5. The "V" shape is easily seen although the "left" segment is both longer and contains relatively more density mass. There appears to be relatively little data where the two segments join, at least in this re-expression. If we choose the bimodal first variable to represent time, we actually can transverse the segments individually. Other interesting shapes appear along these segments. Notice that many data points fall in the region between the segments. One feature not easily seen in this orientation without watching the figure being drawn is that the 1% contours are extremely flat like a pancake.

5. Higher Dimensions

We believe we can extend the thunderstorm representation at least one more dimension. For data sets with large numbers of variables we would first project the data onto a three or four dimensional space by a convenient method such as principal components or projection pursuit. The power of the averaged shifted histogram approach is that it removes many of the limitations of scatter diagrams which are useful only for data sets with sample sizes within a relatively narrow range. We have also nearly achieved the statistical efficiency of kernel methods without resorting to numerical approximation techniques. This speed makes density estimation in several dimensions feasible within the framework of a reactive computer environment of powerful individual workstations with extensive color capabilities. Data analysis has a powerful new tool.

REFERENCES

[1] Kronmal, R.A. and M.E. Tarter (1968), "The estimation of probability densities and cumulatives by Fourier series methods," *J. American Statistical Assoc.* 1:223-235.

[2] Parzen, E. (1962), "On estimation of a probability density function and mode," *Annals of Math. Statistics* 33:1065-1076.

[3] Rosenblatt, M. (1956), "Remarks on some nonparametric estimates of a density function," *Annals of Math. Statistics* 27:832-835.

[4] Scott, D.W., R.A. Tapia, and J.R. Thompson (1977), "Kernel density estimation revisited," *J. Nonlinear Analysis* 1:339-372.

[5] Good, I.J. and R.A. Gaskins (1972), "Global nonparametric estimation of probability densities," *Virginia J. Science* 23:171-193.

[6] Scott, D.W., R.A. Tapia, and J.R. Thompson (1980), "Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria," *Annals of Statistics* 8:820-832.

[7] Scott, D.W. (1976), NDMPLE, IMSL, Inc. Houston, Texas.

[8] Wegman, E.J. (1970), "Maximum likelihood estimation of a unimodal density function," *Annals Math. Statistics* 41:457-471.

[9] Scott, D.W. (1983), "Averaged Shifted Histograms for Multidimensional Density Estimation," Technical Report No. 83-101, Rice University, Houston, Texas.

[10] Graunt, J. (1662), *Natural and Political Observations on the Bills of Mortality*.

[11] Good, I.J. and R.A. Gaskins (1980), "Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data," with discussion, *J. Amer. Statist. Assoc.* 75:42-73.

[12] Badhwar, G.G., J.G. Carnes, and W.W. Austin (1982), "Use of Landsat- derived temporal profiles for corn-soybean feature extraction and classification," *Remote Sensing of Environment* 12:57-59.

[13] Friedman, J.H. and J.W. Tukey (1974), "A projection pursuit algorithm for exploratory data analysis," *IEEE Trans. Comp.* C-23, 881-890.

[14] Tukey, P.A. and J.W. Tukey (1981), "Preparation; prechosen sequences of views," in *Interpreting Multivariate Data*, V.Barnett, editor, John Wiley and Sons: New York, 189-213.