The Annals of Statistics
1985, Vol. 13, No. 4, 1630–1638

TWO BOOKS ON DENSITY ESTIMATION

B. L. S. PRAKASA RAO, *Nonparametric Functional Estimation*. Academic Press, Orlando, 1983, xiv + 522 pages, $70.00.

LUC DEVROYE AND LÁSZLÓ GYÖRFI, *Nonparametric Density Estimation: The L_1 View*. John Wiley and Sons, New York, 1985, xi + 356 pages, $37.95.

REVIEW BY B. W. SILVERMAN

University of Bath

The basic problem of nonparametric probability density estimation is easily stated. Given identically distributed random variables $X_1, \ldots, X_n$ drawn from a density f, the aim is to construct an estimate of f without making parametric assumptions about the form of f. A typical approach to density estimation is the kernel method, where the estimate $\hat{f}$ is given, in the univariate case, by

$$\hat{f}(t) = n^{-1}h^{-1}\sum_{i=1}^{n} K\{(t - X_i)/h\};$$

here K is a kernel function (usually a symmetric probability density function) and h is a smoothing parameter or bandwidth, the value of which determines how much the data are smoothed to produce the density estimate.

The first published paper specifically on density estimation was Rosenblatt (1956); however, density estimates were suggested several years earlier by Fix and Hodges (1951) in a technical report. It is regrettable that this report was never published, since it contains a great deal of interesting discussion and insight, much of which is still pertinent today. In these days of computer graphics and "exploratory data analysis," it is often supposed that the primary purpose for which density estimation was ordained was as a method for producing pretty (perhaps too pretty) pictures from data. However, a glance at Fix and Hodges (1951) shows that this is not at all the case. Their interest in density estimation stemmed from the discrimination or classification problem of allocating an observation Z to one of two populations A and B. If the distributions underlying the populations A and B are not known but have to be estimated from data, and if the statistician is unwilling to make parametric assumptions of the kind tacitly present in Fisher's linear discriminant method, then a natural approach is to construct estimates of the densities f_A and f_B, and then to base a discriminant rule on the ratio of these two density estimates.

Since these early papers, there has been a large amount of research into density estimation and related subjects. The main emphasis of much of this work has been to investigate the theoretical properties, and particularly the asymptotic properties, of various methods of density estimation. In certain quarters,

Received June 1985; revised July 1985.
AMS 1980 *subject classifications*. Primary, 62G05, 62-02.
Key words and phrases. Density estimation, smoothing, nonparametric.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*. www.jstor.org

density estimation has something of a bad name: When, at a conference a few years ago, I presented a paper entitled "Density estimation: are theoretical results useful in practice?," a colleague proposed the answer "Yes, for writing papers." Similar remarks are often made about almost the whole of mathematical statistics and so it is unfair to level this indictment at density estimation in particular.

The two books under review are very different in their aims, and I shall deal first with the book by Prakasa Rao. The aim of this book is "to bring together the large amount of literature in the area scattered over various journals." The author covers an enormous amount of ground in great detail; his bibliography alone contains about 750 items, and the book is an impressive work of scholarship and a valuable reference text. The author's approach is to provide a discussion of various topics giving details drawn from what he considers the more important papers on the subject. Each chapter contains detailed bibliographical notes giving the provenance of the various theorems discussed. These are followed by "Problems," which consist essentially of statements, without proof, of further theorems, with appropriate references to the literature. It would be a very intrepid student indeed who regarded these problems as exercises and they are clearly not intended to be used as such. Overall this general approach is a good way of putting across the large amount of theoretical material that the author has aimed to describe.

The first half of the book consists of a detailed survey of the theory of univariate and multivariate density estimation. The available methods are described and their published properties surveyed in detail. In the second half of the book, the author goes on to deal with related topics such as the estimation of density derivatives and modes, sequential and recursive estimation of densities, estimation under order restrictions, and the nonparametric estimation of a distribution function. There is a brief theoretical discussion of nonparametric discrimination and of the estimation of mixtures. There is little discussion of practical matters however; two graphs of density estimates are reproduced and no other practical examples are included. In summary, Prakasa Rao's book is clearly intended to be a comprehensive account, rather than a critical view, of the theoretical aspects of the subject, and as such is an important contribution to the mathematical statistics literature.

The book by Devroye and Györfi is again theoretical in its emphasis. However, there the similarity between the books ends. Devroye and Györfi present a very personal view of the subject based on the general theme that the natural space in which to consider probability densities is the space L_1 of integrable functions. Until now, when considering the limiting behaviour of density estimates, most authors have considered error measures based on the integrated square error $\int(\hat{f} - f)^2$ of the estimators. This emphasis has probably been as much due to mathematical tractability as to any honest belief that integrated square error is of any greater real interest than any other measure of error. (Compare the place of least-squares theory and the normal distribution in statistics generally.) Devroye and Györfi argue from a different point of view; to them, the most interesting and natural measure of error is the L_1 error $\int|\hat{f} - f|$. They have

another aim, and that is to produce theorems whose statements are uncluttered by unnecessary conditions. Inevitably these aims lead to a complicated and difficult development. As the authors say, rather tongue-in-cheek, in their preface: "Although we hope that this book is entertaining in places, most of it, in fact, is rather dull except perhaps to the odd technical fanatic. Thus, we do not recommend it for class notes or for reading during TV commercials."

The book contains large amounts of original and hitherto unpublished material, and at times feels like a collection of linked papers rather than a text as such. One irritating aspect of this (more the fault of the publishers than the authors) is the lack of a single bibliography; instead there are 12 separate chapter bibliographies, together with an inaccurate author index, with the unfortunate result that the book's value as a reference text is impaired.

Some indication of the remarkable generality of the results available in L_1 is given by the theorem that gives consistency of the kernel estimate. Let $J_n(f)$ be defined by

$$J_n(f) = \int |\hat{f} - f|$$

where $\hat{f}$ is constructed by the kernel method from an independent sample of size n drawn from f. Suppose that the kernel K is a nonnegative Borel function integrating to 1. Then, if $h \to 0$ and $nh \to \infty$, $J_n(f)$ will converge to zero in a stronger sense than almost surely for *all* densities f; otherwise $J_n(f)$ will not converge to zero in probability for a single f. The amazing property of this theorem is the total lack of conditions on the unknown density f. However, the authors are unfortunately not able to give an exact asymptotic form for the expected value $E(J_n)$; they take the majority of a chapter to prove that, for given f and K, the limiting value of $n^{2/5}E(J_n)$ can be bounded above and below by quantities differing by a factor of about $1 \cdot 34$. This is a little disappointing when compared with the exact asymptotic results easily available under suitable conditions for the perhaps more unnatural mean integrated square error, and indicates the magnitude of the task that Devroye and Györfi have set for themselves.

The majority of the book is concerned with consistency, rates of convergence and similar questions for the kernel estimator and other density estimation methods. Though there are no practical examples of any kind in the book and the general mathematical level is far too high for the vast majority of applied statisticians, there is an interesting chapter on the use of density estimation in simulation, the basic message of which is as follows. Suppose it is of interest to simulate from a density f. If f itself is not known, but a sample from f is available, then a natural approach is to construct an estimate $\hat{f}$ from the given data and then to simulate from this estimate. The authors point out that, in order for moderately large simulated samples from $\hat{f}$ to be, for all practical purposes, indistinguishable from samples generated from f, an astronomically large data set must be used to construct the estimate $\hat{f}$. Perhaps, on reflection, this conclusion is not as surprising as all that, but nevertheless it is enlightening for it to be quantified in the way that the authors have done. In addition to the

theoretical results, the chapter on simulation contains some useful practical hints for the construction of fast algorithms for simulating from $\hat{f}$.

To sum up, the book by Devroye and Györfi is, for the most part, a technical tour-de-force that will be of great interest to a rather small number of specialists. If I may make the distinction, its appeal is really to mathematicians rather than to statisticians, but within its own terms it is a remarkable achievement.

It is inevitable that a relatively well-developed subject cannot be adequately covered in two books, and what is missing from both of the books under current review is any strong feeling that density estimation is a practical, or even a potentially practical, technique. As I have already hinted, this epitomizes the situation in which mathematical statistics, generally, finds itself at present. The regrettable, but perhaps inevitable, decision to establish the journal *Statistical Science* is a symptom of this difficulty. To put it baldly, anyone whose only knowledge of statistics came from reading the *Annals of Statistics* would need some convincing that statistics *is* a practical subject and would find it hard to imagine that "mathematical statistics" originated as a study to reinforce and improve the practical capabilities of statistics. A particularly disappointing feature of the technical nature of much of the literature on density estimation is that it may even have had a negative effect, by scaring off potential users of the methods and by making it difficult for courses on the subject to be constructed.

In order to balance the view that might be obtained from reading *only* the two books under review and nothing else, it may be worth briefly discussing a few practical contexts in which density estimates arise. I have tried to provide a fuller description of these and other aspects of the subject, together with further bibliographic references, in Silverman (1986).

An important use of density estimates is for the *presentation* and *exploration* of data. An example where a density estimate displays structure not easily visible using other methods is given in Figures 1 and 2. The data consist of pairs of readings of plasma lipid concentrations taken on 320 diseased patients in a heart disease study [see Scott et al. (1978)]. I am very grateful to David Scott for making the data available to me. The clear bimodality visible in the density estimate is hard to see in the scatter plot, even after consideration of the density estimate. Even if the clustering were clearer from the scatter plot, the density estimate would still have the advantage of providing estimates of the positions of the two modes. Scott et al. (1978) used a density estimate similar to Figure 2 to divide the population into two groups, by drawing a line perpendicular to the join of the two modes and intersecting this join at the point of smallest estimated density. An interesting clinical difference was found between the two groups.

Since the original proposal by Fix and Hodges (1951) discussed above, *nonparametric discriminant analysis* using density estimates has been investigated in a practical context by several authors, and it is the subject of a monograph by Hand (1982). The computer package ALLOC80 [see Hermans et al. (1982)] is based on these ideas and is in quite wide practical use. The package has the capability of dealing with mixed and discrete data of various kinds and also provides a method of variable selection, a useful feature for high-dimensional

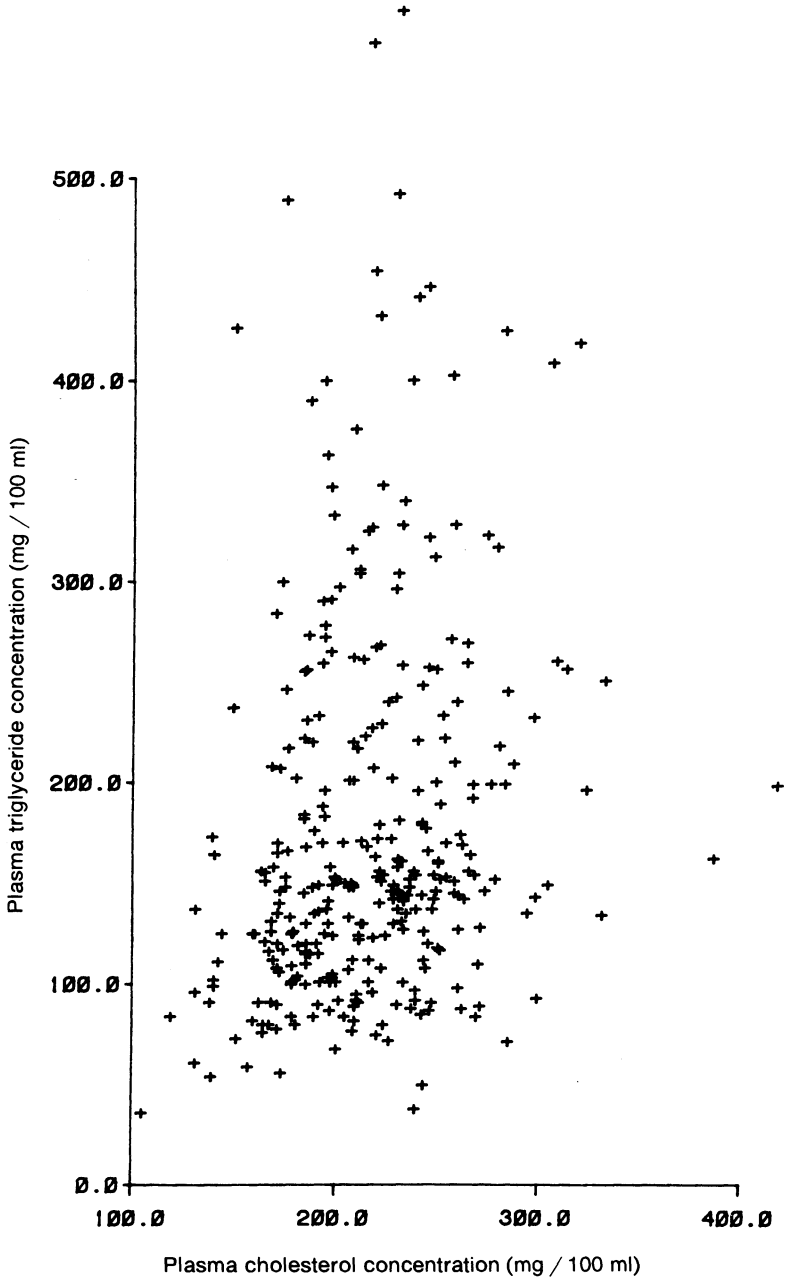

FIG. 1. *Scatter plot of plasma lipid data.*

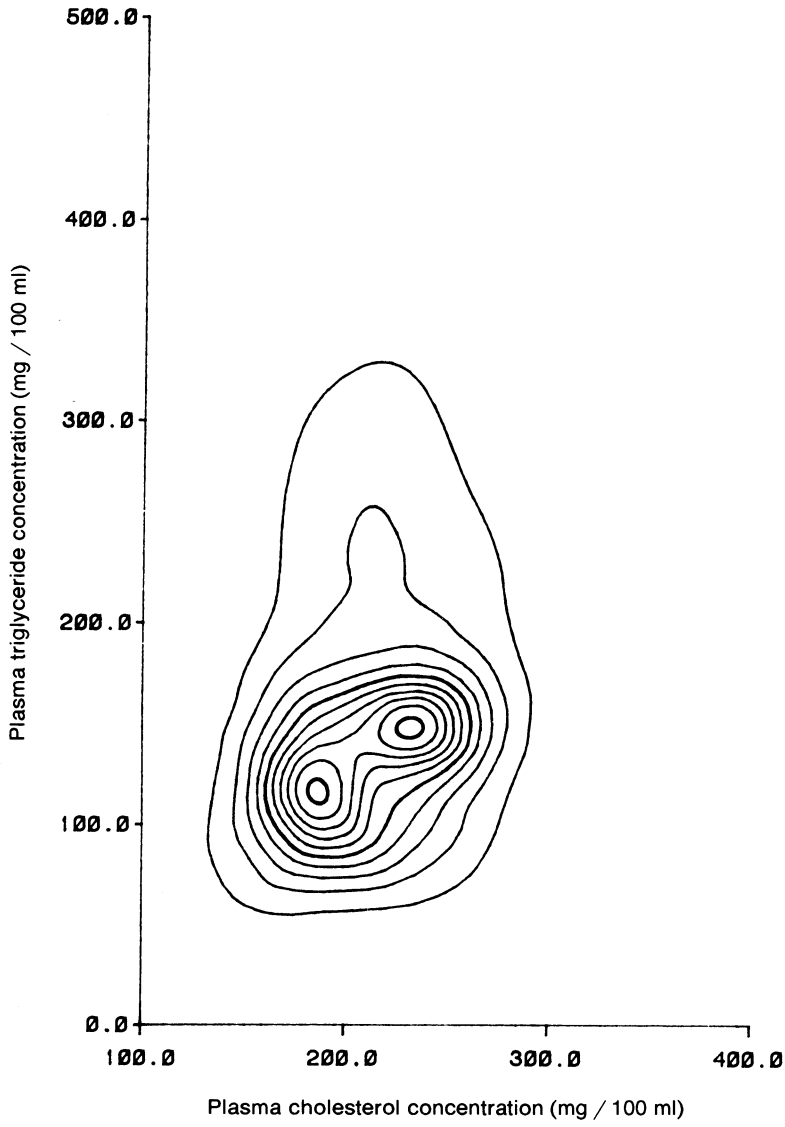

FIG. 2. *Density estimate for plasma lipid data shown in Figure 1.*

data. Developments in knowledge-based computer "expert systems" and the growth of discriminant problems involving the very large data sets collected by remote sensing and similar devices, will no doubt increase the demand for nonparametric methods of discriminant analysis in the future.

Cluster analysis using density estimates has already been alluded to in the example discussed above. Cluster analysis (also called classification by some authors) is a somewhat more contentious procedure than discriminant analysis, because it is often difficult or even impossible to formulate precisely the aim of the procedure. Nevertheless, cluster analysis is a popular statistical technique. As in the medical example of Figures 1 and 2, clusters in a data set can be considered as corresponding to modes in a density estimate constructed from the data. Various different ways of putting this idea into practice have been suggested [see, for example, Koontz, Narendra, and Fukunaga (1976), Fukunaga and Hostetler (1975), and Kittler (1976)]. These three papers are mostly concerned with the applications of cluster analysis in pattern recognition, an important and growing field.

An activity closely related to cluster analysis is *bump-hunting*. The distinction between bump-hunting and cluster analysis (if it exists at all), appears to be that in cluster analysis the aim is to separate the given *data* into groups, while in bump-hunting the data are only important because of the information they provide about the assumed underlying *model*; the object is to discern, for their own sake, modes or bumps (corresponding to extrema of the derivative) in a probability density function. The existence of multiple bumps has been discussed by Cox (1966) as a "descriptive feature likely to indicate mixing of components." Good and Gaskins (1980) discuss a problem arising in high-energy physics where bumps in the underlying density of the observed data give evidence concerning elementary particles.

The use of density estimates in *simulation* has already been mentioned above in the discussion of the book by Devroye and Györfi. A very closely related technique is the *smoothed bootstrap* as defined by Efron (1982). Here the object of interest is a functional $\rho(f)$ of an unknown density f; typically $\rho(f)$ is the sampling standard deviation of some parameter estimate $\hat{\theta}(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ is a sample drawn from f. The essence of the smoothed bootstrap is to replace $\rho(f)$ by $\rho(\hat{f})$, where $\hat{f}$ is a kernel density estimate based on the given data, and then to construct a Monte Carlo estimate of $\rho(\hat{f})$. An example where the smoothed bootstrap does well is presented by Efron (1981). In this example, where the parameter θ is Fisher's variance-stabilized transformed correlation coefficient based on a bivariate sample of size 14, the smoothed bootstrap gives much better estimates of the standard error than the more popular standard bootstrap.

There are several contexts in which it is possible to use density estimates in order to obtain *estimates of functionals* of the density. An example is *projection pursuit* (Friedman and Tukey, 1974), a method for producing interesting low-dimensional projections of a high-dimensional data set, which has at its core the idea of evaluating the interest of a particular projected data set by an index depending, at least in part, on a density estimate constructed from the data. The basic idea, elaborated by Huber (1985) and Jones (1983), is that a density f with

a high value of, say $\int f^2$, is likely to contain features of interest, and hence a projected data set that yields a high value of $\int \hat{f}^2$ will likewise correspond to an interesting view of the original data.

Another example of the use of density estimates to give an estimate of a quantity that depends on the density is discussed by Diggle and Gratton (1984). Their concern is with models where the relationship between the parameters and the data is such that it is possible to simulate data for any given values of the parameters, but impossible to obtain tractable expressions for the likelihood function of the parameters. They suggest an approach where, given a set of observed data, density estimates are used to construct an *estimated likelihood* and hence to carry out inference for the unknown parameters. Models of the kind discussed arise in a wide variety of biostatistical contexts.

Thinking about density estimation in a practical way, naturally, raises all sorts of questions. A rather random short selection of interesting topics for future thought and research follows. Of course, some of these problems are partly solved already and others are, no doubt, insoluble.

1. What measure of error is appropriate when estimating a density for exploratory purposes? It may be something incorporating the error in the estimation of the derivatives of the density.

2. In some applications, though not necessarily for exploratory purposes, automatic methods for choosing the smoothing parameter are needed. Stone (1984) has shown that a method called *least-squares cross validation* (LSCV) is asymptotically optimal in terms of mean integrated square error. However, there are other methods that are asymptotically equivalent to LSCV but give different results in finite samples. Can LSCV be "tuned" to give finite sample optimality?

3. For any particular application, such as nonparametric discrimination, find methods that give the best choice of smoothing parameter for a particular data set.

4. Perform a careful comparative study of the relative merits of various methods of nonparametric discrimination, making use of real problems rather than simulated data.

5. What is the best way of presenting a three- or four-dimensional density estimate?

6. Develop hybrid methods that combine nonparametric smoothing in some variable directions with parametric fitting in others. When will these methods give good results?

7. Under what finite-sample circumstances is the smoothed bootstrap preferable to the standard bootstrap?

8. Adaptive methods, which adjust the amount of smoothing in the tails of the sample, have already been found to give better results in certain circumstances. What is the best way of implementing the idea of adaptivity, particularly in the low-dimensional multivariate case? How much improve-

ment can really be made, in applications, if adaptive methods are used? Can penalized likelihood approaches give a practical adaptive method within a firm philosophical framework?

9. Persuade the general statistical community that density estimation is by no means the answer to all the problems of statistics but is a technique that statisticians should know at least a little about.

The last of these topics is of course the hardest to implement, but it is the development most likely to lead to genuine practical and methodological advances!

REFERENCES

COX, D. R. (1966). Notes on the analysis of mixed frequency distributions. *British J. Math. Statist. Psych.* **19** 39–47.

DIGGLE, P. J. and GRATTON, R. J. (1984). Monte Carlo methods of inference for implicit statistical models (with discussion). *J. Roy. Statist. Soc. Ser. B* **46** 193–227.

EFRON, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* **68** 589–599.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: SIAM.

FIX, E. and HODGES, J. L. (1951). Discriminatory analysis, nonparametric estimation: consistency properties. Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–889.

FUKUNAGA, K. and HOSTETLER, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory* **IT-21** 32–40.

GOOD, I. J. and GASKINS, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75** 42–73.

HAND, D. J. (1982). *Kernel Discriminant Analysis.* Chichester: Research Studies Press.

HERMANS, J., HABBEMA, J. D. F., KASANMOENTALIB, T. K. and RAATGEVER, J. W. (1982). Manual for the ALLOC80 discriminant analysis program. Dept. of Medical Statistics, University of Leiden, The Netherlands.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–525.

JONES, M. C. (1983). The projection pursuit algorithm for exploratory data analysis. Ph.D. thesis, University of Bath.

KITTLER, J. (1976). A locally sensitive method for cluster analysis. *Pattern Recognition.* **8** 23–33.

KOONTZ, W. L. G., NARENDRA, P. M. and FUKUNAGA, K. (1976). A graph-theoretic approach to nonparametric cluster analysis. *IEEE Trans. Comput.* **C-25** 936–943.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

SCOTT, D. W., GOTTO, A. M., COLE, J. S. and GORRY, G. A. (1978). Plasma lipids as collateral risk factors in coronary heart disease—a study of 371 males with chest pain. *J. Chronic Diseases.* **31** 337–345.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

SCHOOL OF MATHEMATICS
UNIVERSITY OF BATH
BATH BA2 7AY
UNITED KINGDOM