Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies

KATHRYN ROEDER*

A method is presented for forming both a point estimate and a confidence set of semiparametric densities. The final product is a three-dimensional figure that displays a selection of density estimates for a plausible range of smoothing parameters. The boundaries of the smoothing parameter are determined by a nonparametric goodness-of-fit test that is based on the sample spacings. For each value of the smoothing parameter our estimator is selected by choosing the normal mixture that maximizes a function of the sample spacings. A point estimate is selected from this confidence set by using the method of cross-validation. An algorithm to find the mixing distribution that maximizes the spacings functional is presented. These methods are illustrated with a data set from the astronomy literature. The measurements are velocities at which galaxies in the Corona Borealis region are moving away from our galaxy. If the galaxies are clustered, the velocity density will be multimodal, with clusters corresponding to modes. Natural candidates for examining the distribution of the data are finite normal mixtures and histograms. The shortcomings of these methods become apparent from the analysis of these data. By finding a confidence set of densities a set of estimates is obtained, ranging from smooth to rough; the number of modes ranges from three to seven. The confidence set of densities is further substantiated by performing nonparametric tests for the number of modes.

KEY WORDS: Cross-validation; Normal mixtures; Spacings; Vertex exchange method.

1. INTRODUCTION

This article presents a novel approach to density estimation and illustrates this method using data from the astronomy literature. Since the data have some interesting features that lend themselves to a physical interpretation, I will first present a rudimentary sketch of the problem. After the Big Bang, it is believed that matter expanded at a tremendous rate. Because of the local attraction of matter, the galaxies formed. Astronomers predicted that gravitational pull would lead to some clustering of galaxies; however, there are data to suggest the presence of superclusters of galaxies, surrounded by large voids (Lapparent, Geller, and Huchra 1986), the so-called string-and-filament pattern. The forces causing this large-scale clustering are not yet understood. Historically, astronomers have mapped galaxies by measuring declination and right ascension—the latitude and longitude with respect to the earth. A third component of position, the distance from our galaxy to others, has recently become available. This distance is estimated using the red shift in the light spectrum in fashion analogous to the way the Doppler effect measures changes in speed via changes in sound. Given the expansion scenario of the universe, points furthest from our galaxy must be moving at greater velocities. Distance, then, is proportional to and can be estimated from velocity.

If the galaxies are clumped, the distribution of velocities would be multimodal, each mode representing a cluster as it moves away at its own speed. Conversely, if there is no cluster effect, the distribution would be determined by the sampling scheme. From our galaxy we sample a conic section of space, but we have a declining ability to detect galaxies at greater distances; thus the velocity density should increase initially and gradually tail off.

In an unfilled survey of the Corona Borealis region, velocities of 82 galaxies from 6 well-separated conic sections of space were measured (Table 1; Postman, Huchra, and Geller 1986). The error is estimated to be less than 50 km per second. Unfilled surveys differ from filled surveys in that the latter cover large continuous regions of the sky, but the sampling is shallow. Unfilled surveys cover less of the sky, but are deeper. A nonparametric density estimate of these velocities, using the method of least squares cross-validation (LSCV) with a normal kernel, is presented in Figure 1. The multimodality of the estimate seems to support the supercluster hypothesis; however, this is only a point estimate of the density function (hereafter point estimate means a single density estimate out of the class of all density functions). Different choices of the smoothing parameter will lead to quite different estimates. Although LSCV leads to an optimal choice of the smoothing parameter asymptotically (Hall 1983), for reasonably sized samples the method can perform quite poorly (Hall and Marron 1987a,b). Obviously, these data were obtained at great effort and expense. Can we extract more information from the data than is available from the kernel estimate? Yes, what is needed is a method of estimation that provides both a density point estimate and a confidence set of plausible densities. In this article such a method is developed, based on inverting a distribution-free goodness-of-fit test. The final product is a three-dimensional density estimate—the third dimension being a range of plausible smoothing parameters (Fig. 2). The

* Kathryn Roeder is Assistant Professor of Statistics, Yale University, Box 2179 Yale Station, New Haven, CT 06520. A portion of this work was presented in the author's doctoral dissertation under Bruce Lindsay at the Pennsylvania State University. This work was partially supported by the National Science Foundation under Grant DMS 9001421. The author thanks the editor and the referees for many comments that improved the article. The author also thanks Eric Feigelson for a stimulating discussion about astronomy.

Table 1. Data for an Unfilled Survey of the Corona Borealis Region

Velocity (km per second)		
9,172	9,350	9,483
9,558	9,775	10,227
10,406	16,084	16,170
18,419	18,552	18,600
18,927	19,052	19,070
19,330	19,343	19,349
19,440	19,473	19,529
19,541	19,547	19,663
19,846	19,856	19,863
19,914	19,918	19,973
19,989	20,166	20,175
20,179	20,196	20,215
20,221	20,415	20,629
20,795	20,821	20,846
20,875	20,986	21,137
21,492	21,701	21,814
21,921	21,960	22,185
22,209	22,242	22,249
22,314	22,374	22,495
22,746	22,747	22,888
22,914	23,206	23,241
23,263	23,484	23,538
23,542	23,666	23,706
23,711	24,129	24,285
24,289	24,366	24,717
24,990	25,633	26,960
26,995	32,065	32,789
34,279		

boundaries of the smoothing parameter are determined by a nonparametric test of fit. Clearly, this presentation is more informative than a single density estimate. Moreover, it is more informative than a point estimate with confidence bands. This example will be pursued further in Section 5.

2. BACKGROUND

There is a vast literature in the area of nonparametric density estimation. For reviews of this literature see Tapia and Thompson (1978), Wegman (1982), and Silverman (1986). Regardless of which method of density estimation is used, the key issue is choosing the smoothing parameter. A data-based selection procedure that has produced good results is LSCV (Bowman 1984; Rudemo 1982). Nonparametric density estimators that use cross-validation (CV) to choose the smoothing parameter have the advantage of objectivity. Though it is well known that the pointwise variance of density estimators is large—$O(n^{-2/3})$ for histograms and $O(n^{-4/5})$ for kernel estimators—LSCV yields no indication of what other densities are plausible alternatives for a given data set.

Consider estimating the unknown density by a normal mixture density

$$f_{Q,h}(x) = \int h^{-1}K\left(\frac{x-\theta}{h}\right) dQ(\theta), \qquad (1)$$

where $K(\cdot)$ is the standard normal density, h is the smoothing parameter, and Q is an arbitrary probability measure called the mixing distribution. This is an extremely rich class of distributions. For h and Q unrestricted, any density can be closely approximated by a normal mixture. Precisely because this class is so large, the maximum likelihood estimator of Q and h fails to produce a meaningful result; the likelihood approaches infinity as $f(\cdot; Q, h)$ approaches a discrete distribution with spikes at the data points (e.g., Geman and Hwang 1982). Clearly, we must restrict either Q or h in some way to achieve a sensible estimator.

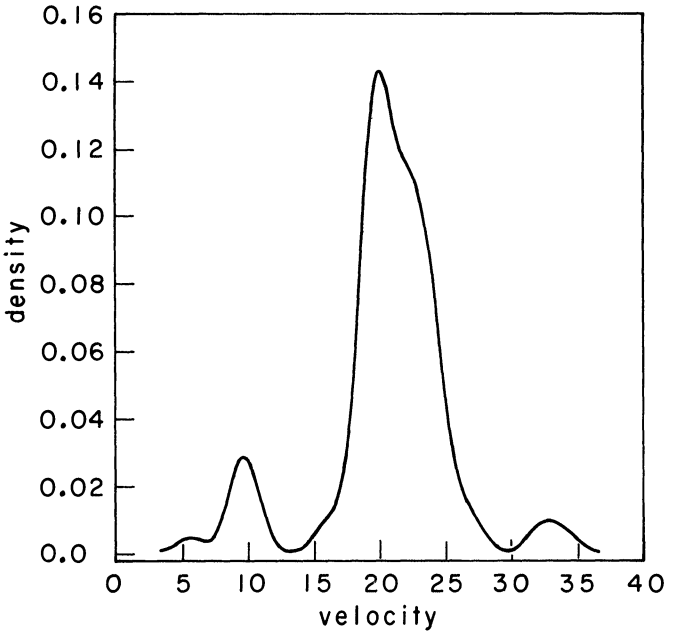

Figure 1. Least Squares Cross-validated Kernel Estimate of the Data in Table 1. Velocity is in 1,000s of km/second.

A diverse class of parametric models is obtained by considering finite mixture models. Consider the class of all mixing distributions $\{Q_v\}$ with positive probability on v points $\{\theta_1, \ldots, \theta_v\}$, $P_{Q_v}(\Theta = \theta_i) = \pi_i$ $(\pi_i > 0)$ and $\sum \pi_i = 1$ $(i = 1, 2, \ldots, v)$:

$$f_{Q_v,h}(x) = h^{-1}\sum_{i=1}^{v} \pi_i K\left(\frac{x-\theta_i}{h}\right). \qquad (2)$$

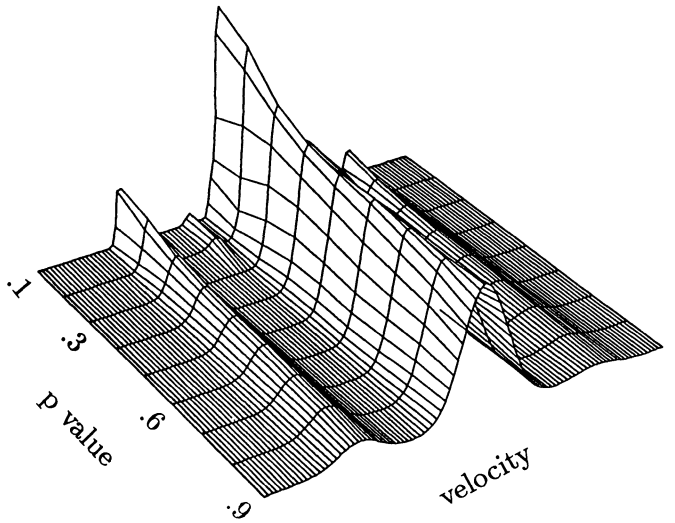

Figure 2. Confidence Set of Density Estimates. The foreground is the estimate associated with a p value of .10, and the background is the estimate associated with a p value of .90.

Maximizing the likelihood over all v-point mixtures is a parametric estimation problem [see Everitt and Hand (1981), McLachlan and Basford (1988), and Titterington, Smith, and Makov (1985) for a discussion of the finite mixture problem]. Unfortunately, there is not necessarily a unique mixing distribution $Q \in \{Q_v\}$ that maximizes the likelihood. Moreover, there is no clear statistical procedure for choosing the number of support points. Because of boundary problems, the usual asymptotic arguments do not apply (Aitkin and Rubin 1985; Ghosh and Sen 1985; J. A. Hartigan 1985; Quinn, McLachlan, and Hjort 1987). Alternatively, consider restricting h. For a given h, Lindsay (1983a) showed that there is a unique probability measure $Q(h)$ that maximizes $\sum \log f_{Q,h}(x_i)$—call this the nonparametric maximum likelihood estimate. By applying this method, one can obtain consistent estimates provided that h_n approaches 0 at an appropriate rate (Geman and Hwang 1982). Unfortunately, it is unclear how one might choose the arbitrary constant.

In this article I present the following: a method of estimation that selects a confidence set of densities from the class of normal mixtures, based on the spacings between ordered observations (Sec. 3); a data-driven method that can be employed to choose a point estimate of the underlying density (Sec. 4); further analysis of the astronomy data set (Sec. 5); and an algorithm that can be used to find $\hat{Q}(h)$ (Appendix).

3. SPACINGS, GOODNESS-OF-FIT TESTS, AND CONFIDENCE SETS

Suppose that we have a random sample from a continuous parametric family, $\{F_\theta : \theta \in \Omega\}$, and we want to estimate θ. A method of estimation called maximum product spacings was recently proposed by Cheng and Amin (1983) and Ranneby (1984). Consider an iid sample, $y_1, y_2, \ldots, y_n$, from a continuous distribution F_0. Let $x_1 < x_2 < \cdots < x_n$ be the ordered values of the sample. The log–product–spacings function is defined as $\mathrm{LPS}(F) = \sum \log(F[I(k)])$, where $I(k) = [x_k, x_{x+1}]$ is the interval between ordered sample values and F is a member of a family of continuous distribution functions. For notational convenience we identify both the distribution function and the probability measure by F. Call $\{F[I(k)]\}_{k=1}^{n+1}$ the spacings. The objective is to generalize this method of spacings and apply it to density estimation.

Let $F_{Q,h}$ denote the distribution corresponding to $f_{Q,h}$. The selection of LPS as an objective function is motivated by the following: the probability measure $\hat{Q}(h)$ that maximizes $\mathrm{LPS}(F_{Q,h})$ is asymptotically equivalent to the nonparametric maximum likelihood estimator (Roeder 1988), and the method of spacings naturally yields a goodness-of-fit statistic that has a distribution that is independent of the null hypothesis. In this respect, the method of spacings differs from the likelihood method, and this turns out to be the key element in preventing overfitting.

Suppose that the data are a random sample from F_0. Because the probability integral transform has a uniform distribution, $\{F_0[I(k)]\}_{i=1}^{n+1}$ has the same distribution as a set of uniform spacings. The distribution of functions of uniform spacings has been widely studied (e.g., Cressie 1976, 1979; Darling 1953; Hall 1986; Pyke 1965). Let $\phi(F_0)$ denote a function of the sample spacings that is asymptotically normally distributed. The following probability statement holds for n large: $\Pr[|\phi(F_0)| < Z_{\alpha/2}] = 1 - \alpha$. If a test is constructed that rejects F if $|\phi(F)| > Z_{\alpha/2}$, the inverse of this test yields a $(1 - \alpha)$ 100% confidence set of distribution functions: $\mathcal{C}(\alpha) = \{F : |\phi(F)| < Z_{\alpha/2}, F \text{ a continuous distribution function}\}$.

Since this test is defined only up to the vector of spacings, arbitrarily rough distributions are contained in the set. As an aid in describing this set of distributions, consider the graphical presentation of a smooth subset, for example, the best fitting normal mixture distributions that fall within $\mathcal{C}(\alpha)$, $\hat{\mathcal{C}}(\alpha) = \{F_{\hat{Q},h} : |\phi(F_{\hat{Q},h})| < Z_{\alpha/2}, \sup_Q \phi(F_{Q,h}) = \phi(F_{\hat{Q},h})\}$. If one can a priori presume that the density came from a class of normal mixtures with some positive variance σ^2, then the probability of coverage is at least $1 - \alpha$. In addition, the confidence set generally represents a set of consistent estimates of the density, provided that the underlying density is a normal mixture (Roeder 1988).

For notational convenience, because ϕ is now only a function of Q and h, let $\phi(Q, h)$ denote $\phi(F_{Q,h})$. Let $\mathcal{F}_h = \{f(\cdot; Q, h) : Q \text{ is a probability measure}, f(\cdot; Q, h) \text{ is a normal mixture}\}$. Choose $h_2 > h_1$. Any density in $\mathcal{F}_{h_2}$, say $f(x; Q_2, h_2)$, can be shown to be in $\mathcal{F}_{h_1}$ by the following argument: let Q^* be a convolution of a normal $(0, h_2^2 - h_1^2)$ with Q_2, from which it follows that $f(x; Q^*, h_1) = f(x; Q_2, h_2)$. Hence $\mathcal{F}_{h_2} \subset \mathcal{F}_{h_1}$. It follows that $\phi(\hat{Q}, h)$ is nonincreasing as a function of h. Take advantage of this property to select the desired confidence set; that is, find $G(\alpha) = \{h : |\phi(\hat{Q}, h)| < Z_{\alpha/2}\}$ to determine the acceptable range of smoothing parameters (Fig. 3).

The function ϕ that minimizes the length of $G(\alpha)$ is unknown. A number of authors (e.g., Cressie 1979; Hall 1986) have examined the equivalent problem in testing.

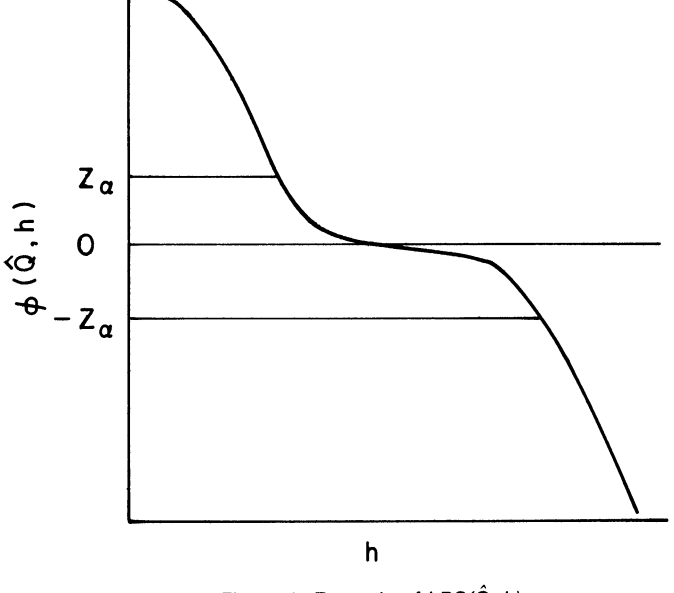

Figure 3. Example of LPS($\hat{Q}$, h).

They studied the Pitman asymptotic relative efficiency of tests that F is uniform versus a general alternative with density $f_n(x) = 1 + l(x)n^{-1/4}$, $\int_0^1 l(x)\,dx = 1$. (A probability integral transform converts any continuous null into this framework.) For tests based on the spacings, the efficiency is greater for sums of squared spacings than for sums of any other function of the spacings. Moreover, even greater improvements can be obtained for functions of higher-order gaps (i.e., let $I_m(k) = [x_k, x_{k+m}]$). This suggests that better results would be obtained for $\phi(Q, h) = \sum F^2[I_m(k)]$ $(m > 1)$ than for LPS(Q, h). Nonetheless, preliminary simulations show that these results do not apply in this semiparametric framework (semiparametric because h is real, whereas Q is infinite-dimensional). In simulations, squared spacings proved to be inferior in the selection of both Q and h (Roeder 1988). For parametric models, log-spacings are considerably more efficient than squared spacings for both estimation and testing. Normal mixtures, being smooth, mimic parametric models. Presumably, then, the semiparametric estimation procedure using LPS inherits some of the features of parametric likelihood estimators.

In simulations (Roeder 1988) I found that the second-order gaps $(I_2(k))$ performed as well as simple spacings. Clearly, second-order gaps are more robust to near ties. Thus I recommend a slight modification of the LPD method: maximize

$$\phi(Q, h) = N^{-1/2}\left(\sum_k \log F_{Q,h}[I_2(k)] + n\log(n+1) + \gamma - 1\right) \div (5\pi^2/6 - 3) \tag{3}$$

over $\mathfrak{H}_h$ (γ is Euler's constant). $\phi(F_0)$ is asymptotically standard normal (Cressie 1976). The function $\sum \log F[I_2(k)]$ can be interpreted as a composite-rank likelihood function (Lindsay 1988). That is, let R_i denote the rank of observation x_i; $F_Q[I_2(k)]$ equals the conditional likelihood that $R_i = k$ given the value of all of the observations except for x_i.

Compare this method of estimation to the normal kernel estimator:

$$\hat{f}_h(x) = \int h^{-1}K\left(\frac{x-y}{h}\right) dE_n(y), \tag{4}$$

where E_n is the empirical distribution function. For h sufficiently large, a rather surprising result emerges: $\hat{Q}$ has no more than $n/2$ support points (Roeder 1988). Hence $f_{\hat{Q},h}$ may differ substantially from the kernel estimator (which has n support points). For any fixed h, $f_{\hat{Q},h}$ achieves greater composite-rank likelihood than $\hat{f}_h$. When h is selected using some optimal method, this procedure should inherit optimality features much like Stein estimators. If the data are from a model similar to a finite normal mixture model and h is selected close to the mixture model standard deviation, then $f_{\hat{Q},h}$ should converge to f_0 faster than the usual kernel estimator. On the other hand, if the data arose from a model dissimilar to a finite normal mixture, then an optimal choice of h would yield a small h, resulting in an estimator similar to the kernel estimator.

4. POINT ESTIMATION

In a large number of nonparametric estimation problems for which a smoothing parameter must be selected, the method of CV has proved to be quite effective. For instance, Geman and Hwang (1982) obtained promising results using CV for an estimation scheme that is similar to this approach.

The method of LSCV requires that we find h to minimize

$$M_0(h) = \int f^2_{\hat{Q},h} - 2n^{-1}\sum_i f_{\hat{Q}-i,h}(x_i),$$

where $f_{\hat{Q},h}$ is the estimate derived from the full data set and $f_{\hat{Q}-i,h}$ is the density estimate constructed from all of the data points except x_i. This procedure is computationally intensive because for each value of h, $\hat{Q}_{-i}$ must be determined for $i = 1, \ldots, n$. This is not as computationally formidable a task as it initially appears, however, since once $\hat{Q}$ is obtained, $\hat{Q}_{-i}$ can be obtained with a few steps of the vertex exchange method (VEM) algorithm (see the Appendix). Nevertheless, for larger data sets consider a modification. Randomly partition the data into m groups of size l. (If l does not divide n evenly, then the mth group contains the remainder.) Let $f_{\hat{Q}-j,h}$ be the density estimate constructed from all data points except the jth group $(x_{j1}, \ldots, x_{jl})$. Then minimize

$$M_0^*(h) = \int f^2_{\hat{Q},h}(x) - 2n^{-1}\sum_{j=1}^{m}\sum_{k=1}^{l} f_{\hat{Q}-j,h}(x_{jk}).$$

The confidence set will help to determine a range of smoothing parameters (a grid) over which to calculate $M_0(h)$. The idea of randomly dividing the data into subgroups has some connections to Marron (1988). In the context of kernel density estimation, grouped LSCV provided improved rates of convergence. In kernel estimation the optimal smoothing parameter depends on n in a well-specified way, and hence more refined results are available. Here I appeal to the general results from the CV and jackknife literature, which suggest that for large samples "leave one out" stategies are essentially equivalent to "leave l out" stategies (e.g., Miller 1968).

5. SUPERCLUSTERS OF GALAXIES

In this section our density estimation procedures are applied to the astronomy data. Figure 4 presents a selection of histograms (note that velocities are in 1,000s of km per second). Figure 4a is nearly unimodal, and Figure 4b suggests roughly six to eight clusters. Figures 4c and 4d have the same interval width, but different starting values; notice that the latter has four modes and the former has only three. Clearly, the arbitrary nature of the histograms is problematic. Postman et al. (1986) displayed a histogram that is nearly identical to Figure 4b; incidentally, they

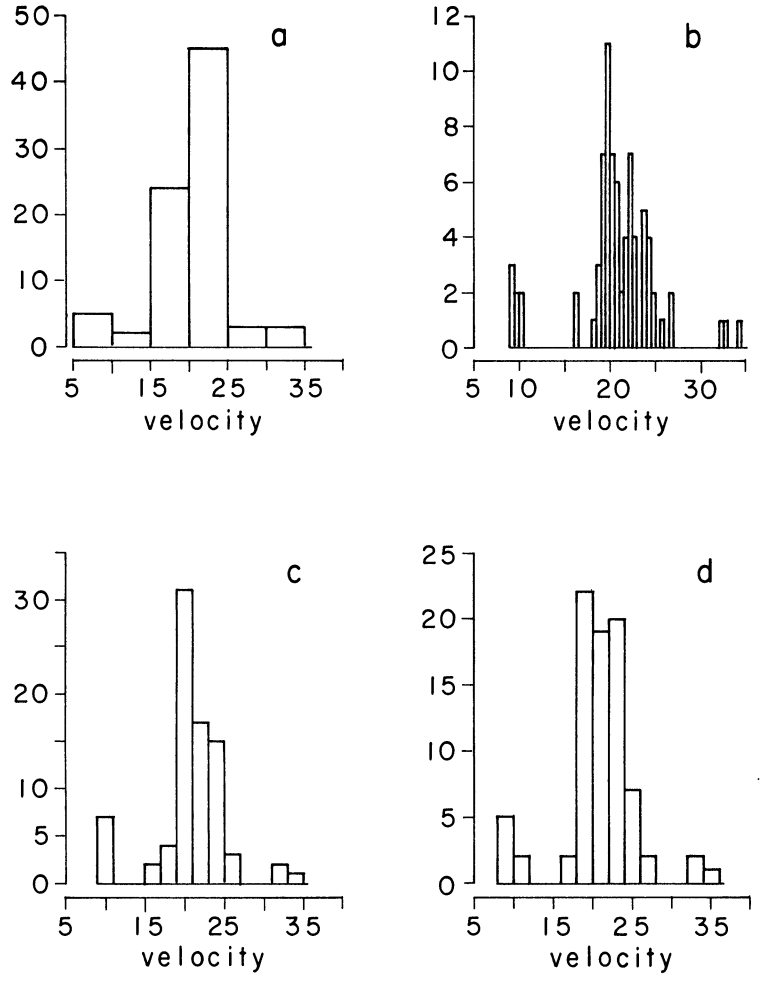

Figure 4. Histograms of Data in Table 1: (a) bin width = 5; (b) bin width = .5; (c) bin width = 2, first endpoint = 11; (d) bin width = 2, first endpoint = 10.

argued that the data demonstrate a deviation from the uniform expansion theory.

To illustrate the problems inherent in finite mixture estimation, the data are fit to finite normal mixtures using the EM algorithm (Aitkin and Tunnicliffe Wilson 1980; Dempster, Laird, and Rubin 1977). Two types of models were considered. Model I is a finite mixture of normals with equal variances (2). Model II is a mixture of normals with unequal means and unequal variances: $f_{Q,\mathbf{h}}(x) = \sum \pi_i h_i^{-1} K((x - \theta_i)/h_i)$. Figures 5a and 5b represent the fitted densities with four and five support points, respectively. Notice that the estimates differ fairly substantially depending on whether equal or unequal variances were used. Figures 5c and 5d present the fitted density with six support points for Models I and II, respectively. The latter estimate is a degenerate density (variance is approaching 0 for two of the components). This is one of the major problems with mixture estimation using Model II—the likelihood approaches infinity for estimates with spikes at observations; this occurs when $\theta_i = x_j$ and $h_i \to 0$.

For Model I, singularities in the likelihood do not occur; however, the likelihood is frequently multimodal. Different modes can result in dramatically different fits. Without experimenting with multiple starting values, it is difficult to find the global maximum. For Model I the four-, five-, and six-point mixtures achieved Akaike-corrected log-likelihoods (Bozdogan 1987) of -212.3, -210.4, and -203.3, respectively. For Model II the four-, five-, and six-point mixtures achieved Akaike-corrected log-likelihoods of -210.5, -202.0, and ∞, respectively. Ideally, one would like to conduct a simple likelihood ratio test to select a model; however, since the problem is nonregular, we cannot rely on the distribution of the likelihood ratio to be approximately chi squared.

Recall that there is a unique density that maximizes (3). To obtain a point estimate, LSCV was used. $M_0(h)$ is minimized at $h = .95$ [Fig. 6; $M_0^*(h)$ is also minimized at $h = .95$, $l = 5$], so the point estimate $f_{\hat{Q},0.95}$ was selected. The point estimate suggests five superclusters (Fig. 7). Also note that $\hat{Q}(.95)$ has more than five support points (Table 2). This is partly an artifact of the algorithm. If the grid were very fine and the convergence criteria were quite strict, then some of these points would coalesce, reducing the number of support points; however, this would make very little difference in our estimated densities.

In Figure 2, the best fitting normal mixtures for a range of smoothing values ($\hat{C}(.20)$) are presented. In the foreground is the roughest estimate in our set, which is ob-

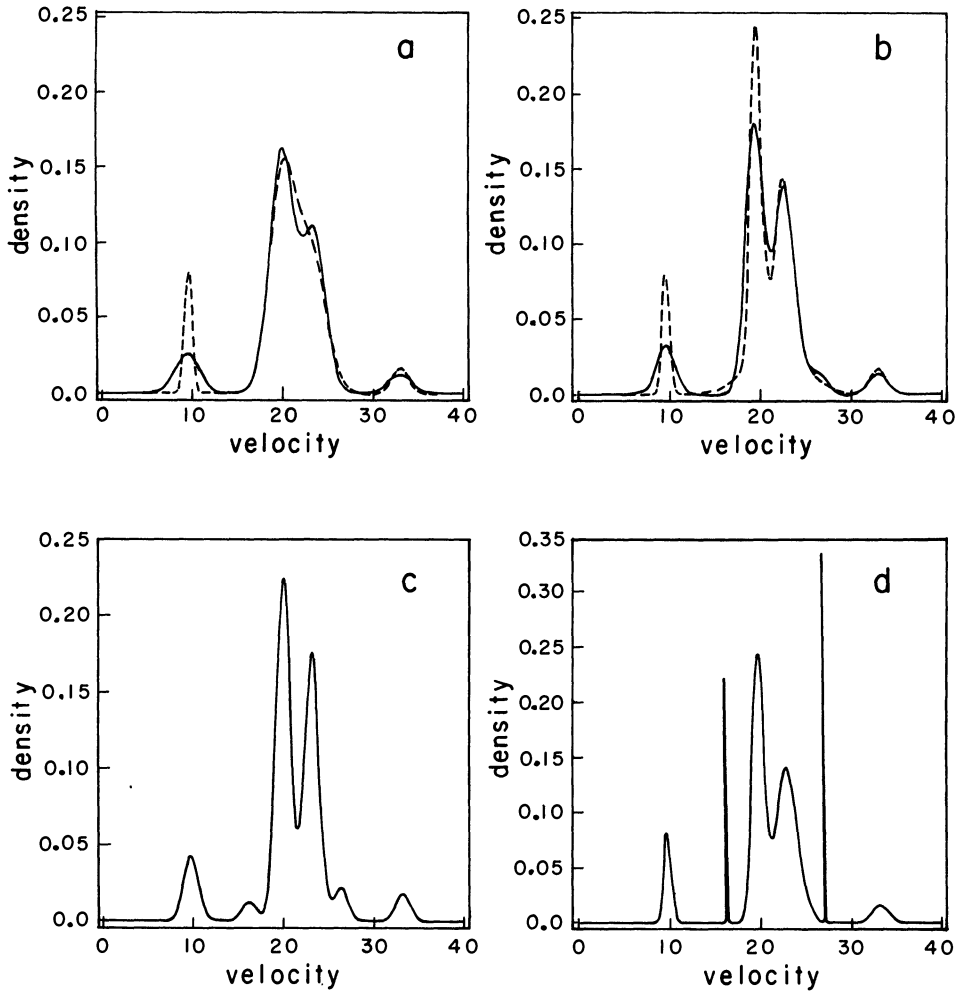

Figure 5. Normal Mixture Estimates of Data in Table 1: (a) mixture of four normals—the smooth curve represents the estimate for Model I (equal variance), and the dashed line represents the estimate for Model II (unequal variances); (b) mixture of five normals, Models I and II; (c) mixture of six normals, Model I; (d) mixture of six normals, Model II.

tained by inverting a size .10 test and selecting that distribution, $F_{\hat{Q},h}$, with the smallest h that is not rejected [$h = .28$, p value $= .10$; p value $= \Pr(Z > \phi(\hat{Q}, h))$]. In the background is the smoothest estimate in our set, which was selected by choosing that distribution with the largest h that is not rejected ($h = 1.50$, p value $= .90$). The point estimate lies approximately in the center of this figure ($h = .95$, p value $= .45$). By comparing the LSCV kernel estimate (Fig. 1; $h_{cv} = .93$) with this point estimate, we see that, though the smoothing parameter is nearly the same, the kernel estimator is much smoother. This is because $\hat{Q}$ accentuates clustering of the data. Nonetheless, the kernel estimate would clearly fall within the set of plausible estimates. Notice that many of the wrinkles are smoothed out as we move from the smoothest to the roughest distributions; however, at least three modes are present in every level. Contrary to the point estimate, where five modes were observed, we see that the confidence set contains at least three, but no more than seven, modes.

Nonparametric tests for multimodality also support the conjecture that the data have more than one mode. The dip test provides a nonparametric test of unimodality versus bimodality (Hartigan and Hartigan 1985; P. M. Hartigan 1985). This test rejects the hypothesis of unimodality ($p < .01$). Silverman (1981, 1983) derived a test based on the amount of smoothing required to force the normal kernel estimator (4) to have $\leq k$ modes. If considerable smoothing is required to remove the multimodality, this suggests that the observed modes are not merely random noise. The test statistic is $h_{\text{crit}} = \inf\{h : \hat{f}_h(\cdot)$ has at most k modes$\}$. Approximate p values are obtained by simulation [Table 3; see Izenman and Sommer (1988) for an excellent synopsis of this method and an application to philatelic data]. Silverman's test provides us with a lower bound of three modes. This test is known to be conservative, and hence it tends to underestimate the number of modes (Silverman 1983). One cannot construct a nonparametric upper bound on the number of modes (Donaho 1988).

APPENDIX: AN ALGORITHM FOR ESTIMATING $\hat{Q}$

In this section h is considered fixed. We are trying to find $\hat{Q} \in \mathfrak{M}$ such that $\phi(Q)$ is maximized for all $Q \in \mathfrak{M}$. $\phi : \mathfrak{M} \to R$ is a concave functional (Roeder 1988). The type of algorithm employed is the vertex direction method (VDM). This algorithm has been shown to converge to the global maximum (Federov

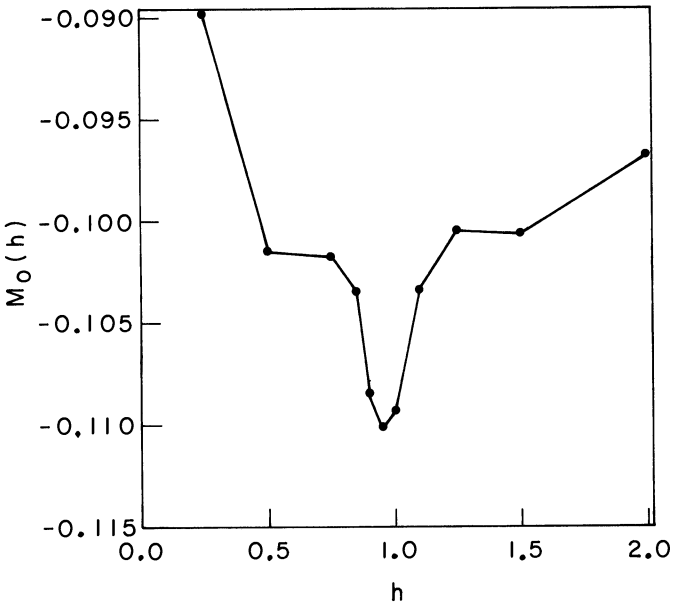

Figure 6. $M_0(h)$ Calculated for h Between .25 and 2.0.

1972; Lindsay 1983b; Wynn 1970). Various modifications to this simple method have been proposed in the design literature (e.g., Atwood 1976; Boehning 1985, 1986; Wu 1978a,b). We will focus on a modification dubbed the vertex exchange method (VEM) by Boehning.

The key element in VDM-type methods is the relationship between $\hat{Q}$ and the gradient function $D(Q, \theta)$. We start by defining the directional derivative of $\phi(F_Q)$ from F_Q toward F_{Q_1}:

$$D(Q, Q_1) = \lim_{\varepsilon \to 0} \varepsilon^{-1}\{\phi[(1 - \varepsilon)F_Q + \varepsilon F_{Q_1}] - \phi(F_Q)\}$$
$$= \sum_k (F_{\theta}[I(k)]/F_Q[I(k)] - 1).$$

Since $\int D(Q, \theta)dH(\theta) = D(Q, H)$, the gradient function $D(Q, \theta)$ determines the value of all directional derivatives. A probability measure $\hat{Q}$ maximizes $\phi(Q)$ iff (a) $D(\hat{Q}, \theta) \leq 0 \; \forall \; \theta$ and (b) $D(\hat{Q}, \theta) = 0$ *w.p.*1 under $\hat{Q}$ (Lindsay 1983b; Roeder 1988).

Select a finite probability measure, Q_1, as a starting mixture. Let $\text{supp}(Q_1) = \{\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{s_1}^{(1)}\}$ denote the support of Q_1. For the mth step, let Q_m be the current estimator. At each step follow this procedure:

1. Find $\theta_{\max}$ such that $\sup_\theta D(Q_m; \theta) = D(Q_m, \theta_{\max}), \theta \in \Omega$.
2. Find $\theta_{\min}$ such that $\inf_\theta D(Q_m; \theta) = D(Q_m; \theta_{\min}), \theta \in \text{supp}(Q_m)$. Let $\pi_{\min}$ denote the support at $\theta_{\min}$ for Q_m, subject to the condition that $\pi_{\min} > 0$.
3. Find β_m $(0 < \beta \leq 1)$ to maximize $\phi(Q_{m+1}[\beta])$, where $Q_{m+1}[\beta] = \sum \pi_i^{(m)}\theta_i^{(m)} + \beta\pi_{\min}(\theta_{\max} - \theta_{\min})$. β_m may be found by using the Newton–Raphson method, provided that one checks that the solution remains within (0, 1). For an algorithm that automatically falls within the boundary, see Boehning (1985).

In summary, the method either adds a support point ($\beta < 1$) or exchanges a point ($\beta = 1$) in the support of Q in such a way as to maximize the increase in $\phi(Q)$ obtainable by exchanging these two support points. The VEM converges monotonically (Boehning 1985). The solution has been reached when $D(Q_m; \theta) \leq 0 \; \forall \; \theta \in \Omega$. For practical purposes we will want to know when we are quite close to $\hat{Q}$. Let $\Delta_m = \phi(\hat{Q}) - \phi(Q_m)$ be the residual. The best upper bound for Δ, given only knowledge of $\delta = \sup D(Q; \theta) : \theta \in \Omega$, is $\Delta \leq (n + 1)\log[1 + \delta/(n + 1)]$ (Lindsay 1983b). Note that in practice it is difficult to find the $\sup_\theta D(Q, \theta)$ over a continuous interval, so we reduce Ω to

Table 2. Estimated Mixing Distribution When h = .95

π	θ
.097	10.0
.024	19.2
.026	19.5
.016	19.8
.400	20.0
.023	22.3
.300	23.0
.004	23.2
.045	24.4
.003	26.3
.004	26.6
.018	27.0
.004	27.2
.001	27.5
.001	27.8
.006	32.7
.029	33.0

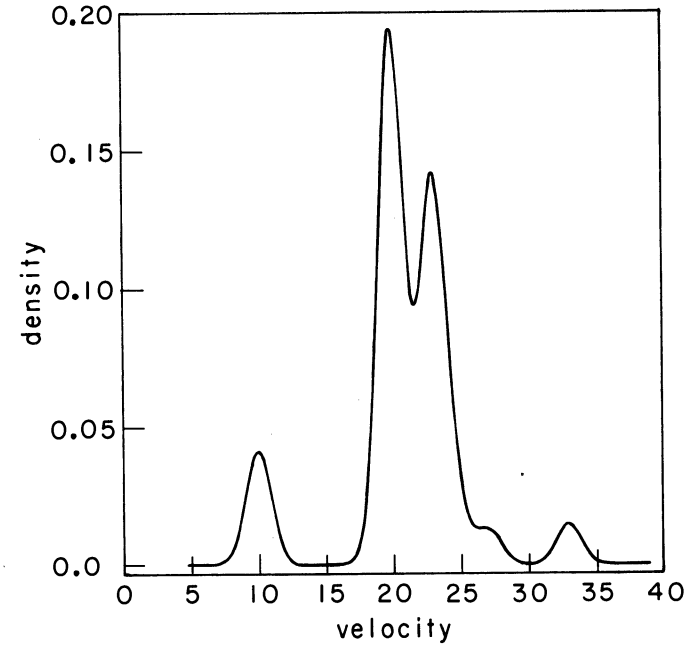

Figure 7. Normal Mixture Density Estimate, h = .95. This estimate corresponds to the value of h that minimizes $M_0(h)$.

Table 3. Critical Values and p Values for Siverman's Test for Multimodality

k	h	p Value
7	.45	.63
6	.67	.20
5	.73	.32
4	.88	.25
3	.94	.53
2	2.50	.00

NOTE: Three hundred bootstrap repetitions were performed to obtain the approximate null distribution.

a grid. This grid can be as fine as computational constraints will allow.

[*Received May 1989. Revised March 1990.*]

REFERENCES

Aitkin, M., and Rubin, D. B. (1985), "Estimation and Hypothesis Testing in Finite Mixture Models," *Journal of the Royal Statistical Society,* Ser. B, 47, 67–75.

Aitkin, M., and Tunnicliffe Wilson, G. (1980), "Mixture Models, Outliers, and the EM Algorithm," *Technometrics,* 22, 325–332.

Atwood, C. L. (1976), "Convergent Design Sequences for Sufficiently Regular Optimality Criteria," *The Annals of Statistics,* 4, 1124–1138.

Boehning, D. (1985), "Numerical Estimation of a Probability Measure," *Journal of Statistical Planning and Inference,* 11, 57–69.

——— (1986), "The Vertex-Exchange Method in *D*-Optimal Design Theory," *Metrika,* 33, 337–347.

Bowman, A. W. (1984), "An Alternative Method of Cross-validation for the Smoothing of Density Estimates," *Biometrika,* 71, 353–360.

Bozdogan, H. (1987), "Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions," *Psymtrka,* 52, 345–370.

Cheng, R. C. H., and Amin, N. A. K. (1983), "Estimating Parameters in Continuous Univariate Distributions With a Shifted Origin," *Journal of the Royal Statistical Society,* Ser. B, 45, 394–403.

Cressie, N. (1976), "On the Logarithms of High-Order Spacings," *Biometrika,* 63, 343–355.

——— (1979), "An Optimal Statistic Based on Higher Order Gaps," *Biometrika,* 66, 619–627.

Darling, D. A. (1953), "On a Class of Problems Relating to the Random Division of an Interval," *The Annals of Mathematical Statistics,* 24, 239–253.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data Via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 39, 1–38.

Donoho, D. L. (1988), "One-Sided Inference About Functionals of a Density," *The Annals of Statistics,* 16, 1390–1420.

Everitt, B. S., and Hand, D. J. (1981), *Finite Mixture Distributions,* New York: Chapman & Hall.

Federov, V. V. (1972), *Theory of Optimal Experiments,* New York: Academic Press.

Geman, S., and Hwang, C. R. (1982), "Nonparametric and Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics,* 10, 401–414.

Ghosh, J. K., and Sen, P. K. (1985), "On the Asymptotic Performance of the Log Likelihood Ratio Statistic for the Mixture Model and Related Results," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (Vol. 2), eds. L. M. LeCam and R. A. Olshen, Monterey, CA: Wadsworth, pp. 789–806.

Hall, P. (1983), "Large Sample Optimality of Least Squares Cross-validation in Density Estimation," *The Annals of Statistics,* 11, 1156–1174.

——— (1986), "On Powerful Distributional Tests Based on Sample Spacings," *Journal of Multivariate Analysis,* 19, 1156–1174.

Hall, P., and Marron, S. (1987a), "Extent to Which Least-Squares Cross-validation Minimizes Integrated Squared Error in Nonparametric Density Estimation," *Probability Theory,* 74, 567–581.

——— (1987b), "On the Amount of Noise Inherent in Bandwidth Selection for a Kernel Density Estimator," *The Annals of Statistics,* 15, 163–181.

Hartigan, J. A. (1985), "A Failure of Likelihood Asymptotics for Normal Mixtures," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (Vol. 2), eds. L. M. LeCam and R. A. Olshen, Monterey, CA: Wadsworth, pp. 807–810.

Hartigan, J. A., and Hartigan, P. M. (1985), "The Dip Test of Unimodality," *The Annals of Statistics,* 13, 70–84.

Hartigan, P. M. (1985), "AS217 Computation of the Dip Statistic to Test for Unimodality," *Applied Statistics,* 34, 320–325.

Izenman, A. J., and Sommer, C. (1988), "Philatelic Mixtures and Multimodal Densities," *Journal of the American Statistical Association,* 83, 941–953.

Lapparent, V. D., Geller, M., and Huchra, J. (1986), "A Slice of the Universe," *The Astrophysical Journal,* 302, L1–L5.

Lindsay, B. G. (1983a), "The Geometry of Mixture Likelihoods, Part II: The Exponential Family," *The Annals of Statistics,* 11, 783–792.

——— (1983b), "The Geometry of Mixture Likelihoods, Part I: A General Theory," *The Annals of Statistics,* 11, 86–94.

——— (1988), "Compositive Likelihood Methods," *Contemporary Mathematics,* 80, 221–239.

Marron, J. S. (1988), "Partitioned Cross-validation," *Econometric Reviews,* 6, 271–284.

McLachlan, G. J., and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering,* New York: Marcel Dekker.

Miller, R. G. (1968), "Jackknifing Variances," *The Annals of Statistics,* 39, 567–582.

Postman, M., Huchra, J. P., and Geller, M. J. (1986), "Probes of Large-Scale Structures in the Corona Borealis Region," *The Astronomical Journal,* 92, 1238–1247.

Pyke, R. (1965), "Spacings" (with discussion), *The Journal of the Royal Statistical Society,* Ser. B, 27, 395–449.

Quinn, B. G., McLachlan, G. J., and Hjort, N. L. (1987), "A Note on the Aitkin–Rubin Approach to Hypothesis Testing in Mixture Models," *Journal of the Royal Statistical Society,* Ser. B, 49, 311–314.

Ranneby, B. (1984), "The Maximum Spacings Method: An Estimation Method," *Scandinavian Journal of Statistics,* 11, 93–112.

Roeder, K. (1988), "Method of Spacings for Semiparametric Inference," unpublished Ph.D. dissertation, The Pennsylvania State University, Dept. of Statistics.

Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Estimators," *Scandinavian Journal of Statistics,* 9, 65–78.

Silverman, B. W. (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society,* Ser. B, 43, 97–99.

——— (1983), "Some Properties of a Test for Multimodality Based on Kernel Density Estimates," in *Probability, Analysis and Statistics,* (IMS Lecture Notes No. 79), eds. J. F. C. Kingman and G. E. H. Reuter, Cambridge, U.K.: Cambridge University Press, pp. 248–259.

——— (1986), *Density Estimation for Statistics and Data Analysis,* New York: Chapman & Hall.

Tapia, R. A., and Thompson, J. R. (1978), *Nonparametric Probability Density Estimation,* Baltimore, MD: Johns Hopkins University Press.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions,* New York: John Wiley.

Wegman, E. J. (1982), "Density Estimation," in *Encyclopedia of Statistical Sciences* (Vol. 2), eds. S. Kotz and N. Johnson, New York: John Wiley, pp. 309–315.

Wu, C.-F. (1978a), "Some Algorithmic Aspects of the Theory of Optimal Designs," *The Annals of Statistics,* 6, 1286–1301.

——— (1978b), "Some Iterative Procedures for Generating Nonsingular Optimal Designs," *Communications in Statistics,* 7, 1399–1412.

Wynn, H. P. (1970), "The Sequential Generation of *D*-Optimum Experimental Designs," *The Annals of Mathematical Statistics,* 41, 1655–1664.