The Annals of Statistics
1997, Vol. 25, No. 4, 1646–1660

NONPARAMETRIC TESTING OF THE EXISTENCE OF MODES

BY MICHAEL C. MINNOTTE[1]

Utah State University

Given a set of data drawn from an unknown density, it is frequently desirable to estimate the number and location of modes of the density. A test is proposed for the weight of evidence of individual observed modes. The test statistic used is a measure of the size of the mode, the absolute integrated difference between the estimated density and the same density with the mode in question excised at the level of the higher of its two surrounding antimodes. Samples are simulated from a conservative member of the composite null hypothesis to estimate p-values within a Monte Carlo setting. Such a test can be used with the graphical "mode tree" of Minnotte and Scott to examine, in a locally adaptive fashion, not only the reality of individual modes, but also (roughly) the overall number of modes of the density. A proof of consistency of the test statistic is offered and simulation results are presented.

1. Introduction. A *mode* of a probability distribution is defined as a local maximum in the associated probability density function. In the case of a density function with constant values at a peak, all of the points on this peak shall be considered a single mode. The identification of modes has applications in many fields of study, from high-energy physics [Good and Gaskins (1980)] and astronomy [Roeder (1990)], to philately [Izenman and Sommer (1988)]. The most common interpretation of multimodality is that of a mixture distribution containing several subpopulations, or as an indication of clustering. It is desirable to identify multimodality when it exists, but we do not wish to give too much importance to apparent modes caused merely by random fluctuations in the data.

Different techniques in the field of multimodality testing have aimed at different goals. The vast majority of techniques available to date have been global, testing for the unimodality, bimodality or multimodality of a data set as a whole.

For example, Silverman (1981) provides us with the most commonly used and studied test, based on "critical bandwidths," the infimum of those smoothing parameters h for which the kernel density estimate

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

Received November 1994; revised May 1996.

[1] Research supported by the Office of Naval Research under contract N00014-90-J-1176, the Department of Defense NDSEG Fellowship Grant DAAL03-90-G-0169 and Utah State University Faculty Research Grant SM-11128.

AMS 1991 *subject classifications*. Primary 62G10, 62G07; secondary 62G09, 62G20.

Key words and phrases. Bump hunting, kernel density estimation, mode estimation, multimodality.

using normal kernel $K = \phi$ is at most k-modal. The normal kernel was selected for the useful property that the number of modes in a normal kernel density estimate is nonincreasing as h increases. This property is not, in general, shared with other kernels; see Babaud, Witkin, Boudin and Duda (1986) and Minnotte and Scott (1993). The critical bandwidth test has been investigated theoretically in Silverman (1983), Mammen, Marron and Fisher (1992) and Hall and Wood (1993) and empirically in Matthews (1983) and Izenman and Sommer (1988). Müller and Sawitzki (1991) suggest a test based on use of the empirical cumulative distribution function

$$F_n(x) = \frac{\#\{x_i \le x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,\, x]}(x_i) \tag{2}$$

to estimate the amount of increase in probability mass above some level λ when $k+1$ disjoint connected sets are counted as compared to k. Other global univariate tests of multimodality can be found in Hartigan and Hartigan (1985) and Wong (1985). Tests of multivariate multimodality are investigated in Hartigan (1988), Hartigan and Mohanty (1992), Rozál and Hartigan (1994) and Polonik (1993). See Minnotte (1992) for a survey and comparison of such techniques.

As opposed to the global approach of the above authors, the ideas propounded in this paper follow the more local procedure of Good and Gaskins (1980). Instead of trying to perform a test on the entire data set, individual suspected modes are examined, which has several advantages. Frequently, merely knowing the number of modes is insufficient; knowing which modes are "real" is of greater value. Local tests also allow location information to be used more effectively, and simplify the possibility of adaptive procedures. Adaptivity can be important when modes occur on peaks of varying sizes, as the peaks can be poorly estimated using a nonadaptive density estimation technique such as fixed-bandwidth kernel density estimation.

In the mode-existence test, we achieve adaptivity by using a fixed, but different, bandwidth for each potential mode. By testing the reality of each mode at an appropriate bandwidth, we can accomplish the goal of finding a usefully adaptive procedure, without the problems inherent in attempting to produce an adaptive density estimate. Therefore, the null hypothesis which we wish to test is that "the mode seen at location x of our density estimate is an artifact of the sample" against the alternative "the mode seen at location x is a true feature of the population." In this way, one can gain the benefits of locality and increase power by taking advantage of mode location and size information which tests such as Silverman's ignore.

This is assisted by the use of a graphical tool known as a "mode tree" [Minnotte and Scott (1993)]. In its purest form, the mode tree plots a range of bandwidths h on the y-axis versus the locations of all modes seen in the kernel density estimates with those bandwidths. The result is a series of (roughly) vertical lines called *mode traces*. As h decreases, new modes appear. A new mode trace is connected to an existing one by a horizontal line indicating which peak contained the shoulder which has now become a mode (determined by the location of the new antimode). See Figure 2 for an example of a mode tree for the Ahrens (1965) chondrite data ($n = 22$).

2. Testing modes. We begin with a sample $\{X_1, \dots, X_n\}$ of size n from a population with density $f(x)$. We compute a normal kernel density estimate $\hat{f}(x)$ with kernel bandwidth h. If h is sufficiently large that there is only a single mode, then there is nothing to test, since all densities are assumed to have a minimum of one mode. Therefore, in the following discussion, it is assumed that there are $k \geq 2$ modes $u_1, u_3, \dots, u_{2k-1}$ and $(k-1)$ antimodes $u_2, u_4, \dots, u_{2k-2}$ in $\hat{f}(x)$. Note that $u_i < u_{i+1}$ for all i. For the purposes of the test statistic, the extreme x-values $-\infty$ and ∞ (or, more practically, the lowest and highest points x for which the estimate $\hat{f}(x)$ is calculated) are also considered antimodes and are denoted u_0 and u_{2k}, respectively.

In testing the existence of mode u_i, the first step is to determine the bandwidth $h_{\text{test}, i}$ at which to test. We perform the test at the smallest bandwidth at which the mode still remains a single object; that is, at the bandwidth slightly greater than that at which the mode tree indicates the mode in question is splitting. Thus $h_{\text{test}, i}$ will be Silverman's critical bandwidth $h_{\text{crit}, k}$ for some k.

There are several reasons for this choice of test bandwidth. The first is simply that this is an objective choice, determined by the data, rather than the analyst. A more important reason is that this choice will give the greatest power (see Section 4, Theorem 1). This choice also allows us to use the theory already developed for Silverman's critical bandwidths.

A final feature of this selection is that if a given mode never splits, it will never be tested. This result is useful in the case of single points. An isolated point in the tail of a sample will not be tested, even if it produces a mode at fairly large values of h. In our view, this is desirable, and perhaps comparable to Hall and Wood's (1993) suggestion to truncate the data in the tails for Silverman's test or their modification of it. Unfortunately, if the data are binned, a mode consisting of a single bin will also never be tested, even if it contains many points. Minor modification of the procedure will allow us to test such modes. We will suggest a possible solution in the binned case in Section 3.

Having determined the bandwidth $h_{\text{test}, i}$ of our test [and having kernel density estimate $\hat{f}(\cdot)$], we can now calculate the test statistic

$$M_i = \int_{u_{i-1}}^{u_{i+1}} \left[\hat{f}(x) - \max\left(\hat{f}(u_{i-1}), \hat{f}(u_{i+1})\right)\right]_+ dx. \tag{3}$$

We note that M_i is the minimal L_1 distance from the density estimate to the set of continuous functions without a local maximum between the observed antimodes in the density function. M_i can be thought of as the area or probability mass of the mode above the higher of the two surrounding antimodes. In this light, the decision to use the smallest possible bandwidth makes sense; the smaller the bandwidth, the higher the modes and lower the antimodes, and the greater the probability mass in the region above the higher antimode.

M_i is also in a sense the single-mode equivalent of Müller and Sawitzki's (1991) excess mass functional. It differs from their statistic both in being local and in being computed from a specific density estimate, rather than from the empirical cumulative distribution function. It also shares some similarities with Hartigan's (1988) FILL parameter, which compares a continuous, but multimodal, density, to one which is unimodal, but possibly discontinuous, through the addition of "bridges" connecting the modes. FILL finds the minimal area under the bridges and above the valleys the bridges cross.

To estimate a p-value from M_i, we follow Silverman (1981) in the use of Monte Carlo methods. Standard Monte Carlo methods for obtaining p-values assume a simple null hypothesis from which to draw the new samples for comparison. Unfortunately, this is certainly not the case; the set of densities containing no modes in the observed region is infinite. Therefore, we settle for choosing a representative density of the null hypothesis which is both conservative and consistent with the observed data.

In order to keep our estimate consistent with the data in every way but that in which the hypothesis is concerned, we impose some constraints. We insist (with one exception, considered shortly) that the new density $\tilde{f}_i$ equal $\hat{f}$ everywhere outside the adjacent modes u_{i-2} and u_{i+2}. Of course, if the mode being tested is the left- or rightmost, this constraint will only apply to the density on the far side of the (sole) adjacent mode.

Within the region bounded by the indicated modes, we then impose additional constraints. In the region $u_{i-2} < x < u_i$, $\tilde{f}_i(x)$ cannot be greater than $\hat{f}(u_{i-2})$ unless $\hat{f}(x) > \hat{f}(u_{i-2})$. Likewise, $\tilde{f}_i(x)$ in the region $u_i < x < u_{i+2}$ is bounded by $\max\{\hat{f}(x), \hat{f}(u_{i+2})\}$. Finally, of course, in keeping with the null hypothesis, there can be no mode between u_{i-2} and u_{i+2}, resulting in an overall maximum for the entire region of $\max\{\hat{f}(u_{i-2}), \hat{f}(u_{i+2})\}$.

Working within this somewhat convoluted set of constraints, we find the admissible function $\tilde{f}_i$ which is closest in L_1 distance to the original $\hat{f}$. This will involve putting some of the probability mass of the mode into at least one of the antimode valleys on either side. The L_1 difference between the two functions will simply be twice the mass so moved, so this criterion will favor leaving as much mass as possible under the mode in question. The result will frequently be that all of the moved mass will go to one side or the other. Which side is filled will be determined by which choice will leave the region of the original mode the highest. If the side with the higher mode is filled to the level of that mode, but there is still mass to be accounted for which was removed from the mode down to that level, then the second side will fill, to a maximum of *its* mode. If this is done, and the two filled regions together still do not equal the excised region, then the entire density is rescaled to make up the difference (so that the whole integrates to one; this is the exception to the equality constraint mentioned above). If the mode being investigated is the left- or rightmost mode of $\hat{f}$, the density will be rescaled after filling only the one available valley. Some examples of the entire density-choosing process are shown in Figure 1.

This choice of $\tilde{f}_i$ satisfies all of the desirable properties mentioned earlier. It is a member of the null hypothesis. It keeps the probability mass as close to that of the original data-produced density estimate as permissible under the

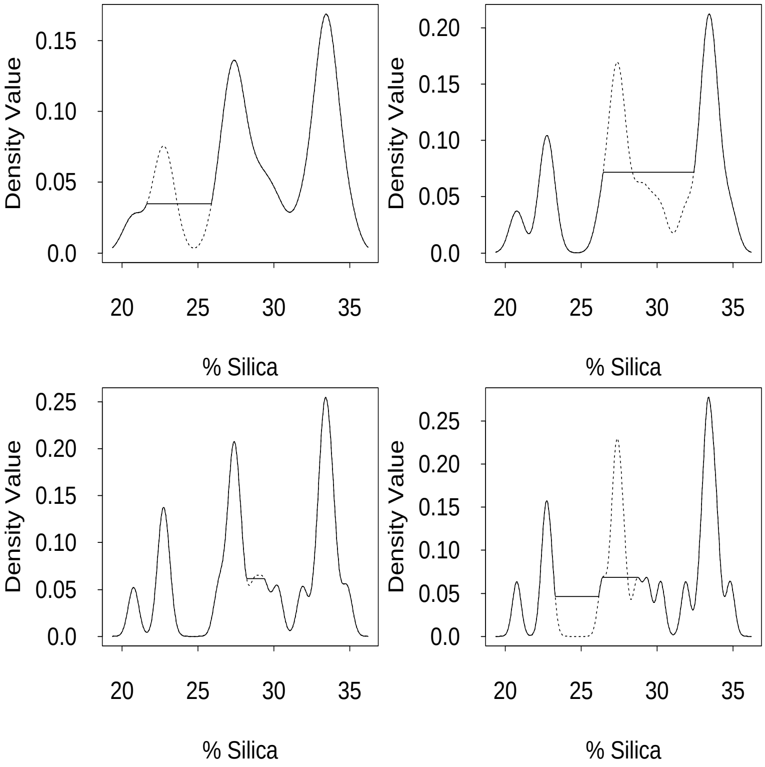

FIG. 1. *Examples of the density-choosing process using the chondrite data. The dotted line indicates the original density estimate, while the solid line represents the final choice of density (without the mode in question). Examples include* (a) $h = 0.700$, $i = 1$, (b) $h = 0.487$, $i = 3$, (c) $h = 0.347$, $i = 4$ *and* (d) $h = 0.286$, $i = 3$.

constraints (due to the L_1 requirement). Finally, it is conservative in general, as the algorithm will generally result in a density with a flattened bump where the mode used to be (stuck on the side, as it were, of one of the adjacent modes), and $\tilde{f}_i$ will be on the boundary of the null hypothesis, since it "almost" has a mode.

Given the null representative density $\tilde{f}_i$, new samples, each of size n, are then drawn from $\tilde{f}_i$. After a sample is drawn, a new density estimate $\hat{f}_i^j$ is calculated using the same bandwidth $h_{\text{test},i}$. The modes of $\hat{f}_i^j$ and $\hat{f}$ are matched, using a matching algorithm described for use in the mode tree in Minnotte and Scott (1993). The mode of $\hat{f}_i^j$ matching u_i might be used directly to estimate the p-value, but it is more conservative and appropriate to select the largest mode of $\hat{f}_i^j$ in the region of interest. This region is bounded by the matches of u_{i-2} and u_{i+2} of $\hat{f}$ or, lacking one or both of these, the mode locations themselves. Each mode in this region is measured with a test statistic exactly equivalent to M_i. The largest is then allowed to "experience" decreasing h until just before it splits (to give a true equivalent to our choice of $h_{\text{test}, i}$). The value of the final test statistic can be denoted M_i^j.

Estimation of a p-value could be performed as a standard Monte Carlo procedure, fixing the number of resamplings N, but it is more efficient to follow Besag and Clifford (1991) in using a sequential method, counting the number of samples, $\tilde{N}$, and the number with simulated test statistics at least as great as the observed value, $\tilde{L}$. We stop when $\tilde{N} = N$ [specifying p-value $(\tilde{L}+1)/(N+1)$] or when $\tilde{L} = L < N$ (assigning p-value $L/\tilde{N}$). In simulation studies, values of 16 for L and 399 for N seemed quite successful, though both might be increased for greater accuracy.

If we conduct such a test at each $h_{\text{crit}, k}$ for a given data set, we may plot the results on the same mode tree that already provided us with the critical bandwidths for our tests. For example, Figure 2 is the mode tree for the chondrite data, with filled circles indicating the location and test bandwidths of modes found significant at the $\alpha = 0.15$ level. Tests resulting in p-values greater than 0.15 are indicated by open circles. Of course, any choice of cutoff α-level could be selected. Using levels of higher than 5% has been suggested for other tests of multimodality [see Matthews (1983) and Izenman and Sommer (1988)]. The test might be conducted on a fixed number of modes (starting at the top of the mode tree) or on all splits until each point or bin forms its own mode. The author's implementation takes all splits down to the bandwidth equal to 0.005 times the range of the data.

The result may be useful as a rough exploratory tool for examining the multimodality of the overall data set. Of course, the dangers inherent in multiple testing are very real here, and we should approach the results with caution (see the final example in Section 5). A Bonferroni approach might be appropri-

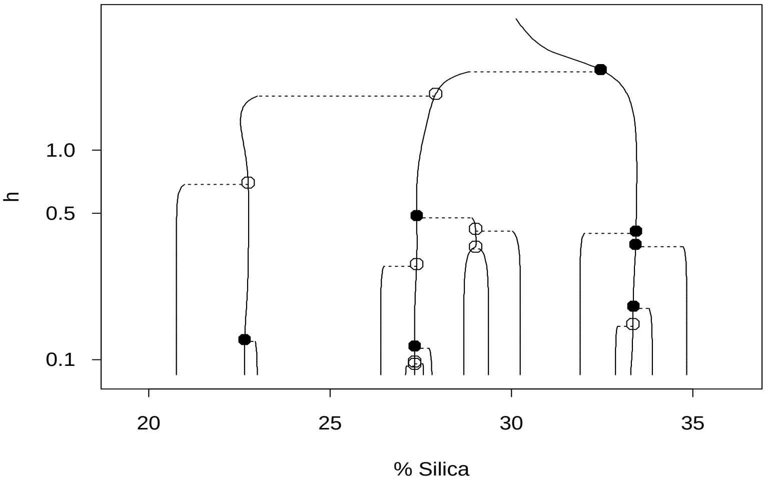

FIG. 2. *Test mode tree for the chondrite data. Filled circles are significant modes at the $\alpha = 0.15$ level; open circles are modes which are not significant at this level.*

ate, but appears likely to be too conservative to be worthwhile. Nonetheless, such a plot may be useful for pointing out modes which may be worthy of further study, including some which might be missed by nonadaptive tests such as Silverman's.

3. Implementation. The implementation of the above test used by the author follows Good and Gaskins (1980) in requiring binned data. While it is possible that an unbinned version of the test is feasible and perhaps slightly more accurate, the advantages in ease and speed of most of the required computations made the binned implementation attractive. Our choice of 500 bins resulted only in very small amounts of rounding, and likely resulted in very little difference from a nonbinned version of the procedure. The most important consideration here is that the bin widths be substantially smaller than anticipated modes, to ensure sufficient resolution that the modes can be clearly identified and to ensure that the modes of interest will eventually split and be tested.

For the initial (and simulated) density estimates, an implementation of Scott's (1985a) average shifted histogram (ASH) estimator was used as an approximation to the normal kernel density estimate. Although approximating the normal kernel is less efficient than a finite-support kernel, it can still be much quicker than an exact kernel calulation [see, e.g., Fan and Marron (1994)].

By binning the data and then using the ASH to calculate $\hat{f}(x)$ at the bin centers, it is simple to simulate from the modified distributions $\tilde{f}_j$ by approximating $\tilde{f}_j$ as a multinomial distribution. Since only the bin counts, rather than the actual data are required for the estimate, little or no accuracy is lost in this approximation.

The only remaining problem is that of prior binning. Generally, data are measured only to a certain level of accuracy. Although the mode-estimation procedure requires binning, it can produce problems if the binning is much cruder than we wish to use and, more importantly, if numerous bins contain a large number of points. If the original, rough binning is used, it can result in large modes never being tested, due to never splitting (see Section 2). If the unmodified points are placed in narrower bins, it can produce unreasonably low p-values for single bins. This effect is due to the large number of points in some bins, while several surrounding bins have zero points.

One possible solution is standard blurring. This is accomplished by adding a uniform random variable ranging from $-d/2$ to $d/2$, where d is the bin width, to each data point. While this is satisfactory if all interesting modes are wider than the individual bins, it unfairly dilutes modes as narrow as the original bins.

We found a useful compromise between standard blurring and the original data to be random blurring based on the frequency polygon of the data (*FP-blurring*). In the frequency polygon, histogram bin centers are connected with straight lines, leading to a smoother [and asymptotically superior; see Scott (1985b)] density estimate than the histogram itself.

The FP-blurring algorithm takes the frequency polygon estimate from each bin (which will include that bin center and half of each of the two linear regions connecting to adjacent bins), and rescales it into a density in its own right. A number of points equal to the original bin count is then drawn from the resulting density (by use of an inverse cdf method). Thus, when a bin is far larger than either of its neighbors, most of the points will be drawn near the center, where they will still provide evidence for a possible mode. On the other hand, a bin with a far smaller count than its neighbors will find most of its (few) points near the bin edges. This might be viewed as a rough inverse analogue to the linear binning examined in Jones (1989).

This compromise appears to work well. The points are distributed throughout the domain of the data, but in a manner consistent with the data's indication of likely mode and antimode locations.

4. Theoretical investigations. Theoretical investigations into the mode-existence test have proven fruitful in two major directions. The first is an indication of the utility of the mode tree in investigations of this sort, as Theorem 1 below shows that for a given mode, M_i is decreasing in h. This is a strong argument for using the appropriate h_{crit}, the bandwidth at which the mode is about to split, as the bandwidth for the test, as this ensures that the test statistic will be as large as possible and will give the greatest possible power for the results. Because of the focus on the bandwidth h in this theorem, we explicitly note the dependence on h (usually left unstated) of the density estimate $\hat{f}(x)$ and its derived statistics.

THEOREM 1. *For fixed data set $\{X_1, \dots, X_n\}$, let $h_1 < h_2$ be such that the normal kernel density estimates with bandwidths h_1 and h_2, $\hat{f}(x, h_1)$ and $\hat{f}(x, h_2)$, have the same number of modes. Define M_i, u_{i-1}, u_i, and u_{i+1} as in* (3). *Then $M_i(h_1) \geq M_i(h_2)$.*

The proof of the theorem requires the use of the following two related propositions, easily verified by differentiation, that are of substantial interest themselves.

PROPOSITION 1. *For a normal kernel estimator $\hat{f}(x, h)$,*

$$\frac{\partial}{\partial h}\hat{f}(x, h) = h\frac{\partial^2}{\partial x^2}\hat{f}(x, h).$$

PROPOSITION 2. *Let $x(h)$ be a point chosen in a manner dependent on h, for example, a mode of a particular kernel density estimate. For a normal kernel estimator $\hat{f}(x(h), h)$,*

$$\frac{\partial}{\partial h}\hat{f}(x(h), h) = h\frac{\partial^2}{\partial x^2}\hat{f}(x(h), h) + \frac{\partial}{\partial h}x(h)\frac{\partial}{\partial x}\hat{f}(x(h), h).$$

The second key theoretical result is the consistency and rate of convergence of the test statistic M_i. Because of the local nature of kernel density estimates as n goes to ∞, this can be reduced to two primary cases, a unimodal density with two (false) estimated modes, and a bimodal density with two (presumably true) estimated modes. Theorems 2 and 3, respectively, explore the asymptotic behavior of M_i under these two conditions.

For the purposes of these two theorems, we will follow Mammen, Marron and Fisher (1992) in making the following assumptions on f.

ASSUMPTIONS.

(A1) f is a bounded density with bounded support $[a, b]$.
(A2) f is twice continuously differentiable on (a, b).
(A3) $f'(a+) > 0$, $f'(b-) < 0$.
(A4) $f''(x) \neq 0$ and $f(x) > 0$ for all x with $f'(x) = 0$.

We also require one further assumption.

(A5) $|f''(x)| < \infty$ for all x with $f'(x) = 0$.

THEOREM 2. *Let f be a density satisfying assumptions* (A1)–(A5) *with single mode z_1. Let M_1 and M_3 be the values of the test statistic M_i for the two modes, u_1 and u_3, of $\hat{f}_n$ observed for $h_{\mathrm{crit},1} \geq h > h_{\mathrm{crit},2}$.* (*Recall, u_2 is the antimode. One of M_1 and M_3 will be evaluated at $h_{\mathrm{crit},2}$, the other at $h_{\mathrm{crit},k}$ for some $k > 2$.*) *Then $M_i = O_P(n^{-3/5}(\log n)^{3/4})$ for $i = 1, 3$.*

THEOREM 3. *Let f be a bimodal density satisfying assumptions* (A1)–(A5) *with modes z_1 and z_3, and antimode z_2. Let*

$$\mathscr{M}_1 = \int_{-\infty}^{z_2} [f(x) - f(z_2)]_+ \, dx$$

and

$$\mathscr{M}_3 = \int_{z_2}^{\infty} [f(x) - f(z_2)]_+ \, dx.$$

Let M_1 and M_3 be the values of the test statistic M_i for the two modes of $\hat{f}_n$ observed for $h_{\mathrm{crit},1} \geq h > h_{\mathrm{crit},2}$. Then $|M_i - \mathscr{M}_i| = O_P(n^{-2/5}(\log n)^{1/2})$ for $i = 1, 3$.

These theorems imply that the test statistics M_i converge in probability to 0 when the modes are spurious, and to $\mathscr{M}_i > 0$ for real modes.

5. Simulation studies. In order to investigate the properties of the mode-existence test, a number of simulation studies were conducted. The first aspect we examined was the reported p-values provided by the test. Since these represent the likelihood of seeing so extreme a result from a member of the null hypothesis, we tested samples of various sizes drawn from unimodal distributions. For each sample, the modes existing at $h_{\mathrm{crit},2}$ were noted and tested at the appropriate bandwidths. Although Theorem 2 suggests that both test statistics should be converging in probability to 0, at the relatively small sample sizes investigated one of the two will probably contain the true mode and thus has the potential to still have large M_i. As we are concerned primarily with the probability of declaring a second mode, we focus on the larger (less significant) of the two p-values generated from each sample.

For each tested density, 1000 samples were generated for sample sizes 40, 100 and 250. For each sample, the first two modes to appear were tested at the appropriate bandwidths. If the reported p-values are to be accurate (or conservative), the probability of getting a p-value of α from the null hypothesis should be no greater than α. This does indeed turn out to be the case for the standard normal and uniform densities, as can be seen in Table 1 for $\alpha = 0.05$ and 0.15.

Visual examination of the quantiles of the larger p-values (not shown) confirms the test's conservative nature at all α levels. The uniform results are somewhat less conservative than those of the normal, giving credence to the suggestion of authors such as Hartigan and Hartigan (1985) and Müller and Sawitzki (1991) that the uniform be used in calculating "worst-case" p-values in tests of unimodality.

As the mode-existence test appears to be acting appropriately for unimodal densities, we now turn our attention to the alternative (bimodal) case. We repeated the procedure used above for examining the significance levels in an attempt to examine the power of the test. Again, for each density and sample size, we tested the first two modes of each of 1000 samples. Because we are interested in the likelihood of finding both modes significant, we again examine the larger p-value (presumably from the smaller of the two modes). Here, of course, it is desirable to see large probabilities of small p-values.

When investigating equal mixtures of normal densities with equal standard deviations, degree of separation is critical. When the modes are separated by only three standard deviations, the test does not distinguish the density from

TABLE 1

Percentage of samples (*out of* 1000) *from the indicated distribution and sample size in which the larger of the p-values for the first two modes was less than* 0.05 *or* 0.15. *Small numbers are desirable for unimodal densities, large ones for bimodal densities*

Sample size	40		100		250	
Significance level	0.05	0.15	0.05	0.15	0.05	0.15
Unimodal distributions						
$N(0,1)$	0.2	3.6	0.7	2.1	0.6	3.7
$U(0,1)$	2.6	9.1	2.5	6.8	2.2	6.8
Bimodal distributions						
$\frac{1}{2}N(-\frac{3}{2},1)+\frac{1}{2}N(\frac{3}{2},1)$	3.1	8.2	4.3	12.9	15.2	32.8
$\frac{1}{2}N(-2,1)+\frac{1}{2}N(2,1)$	27.6	46.6	63.3	82.4	96.0	98.7
$\frac{3}{4}N(0,1)+\frac{1}{4}N(2,(\frac{1}{3})^2)$	3.9	12.1	12.6	28.8	33.6	58.3

a unimodal one at sample sizes of 40 or 100. Even at 250 points, the test indentifies both modes only a small fraction of the time. With four standard deviations between the modes, however, the test finds both modes frequently even at 40 points, and most of the time at 100 or more.

A bimodal normal mixture with different weights and variances gave results between those of the first two bimodal examples. The density is the sum of 3/4 of a standard normal density and 1/4 of a normal density with mean 2 and standard deviation 1/3. As the two modes were separated by 1.5 times the sum of the standard deviations, it should not surprise us that the results were similar to those of the closer equal mixture.

To investigate the effectiveness of the adaptive nature of the test, a final density was chosen. The density is a trimodal mixture of three normal densities in a ratio of $1:1:3$. The first two have standard deviations of 1, and means of 4 and 8, while the third has standard deviation 5 and mean 20.

The density is easy to identify in that all three modes are large, clear and highly separated (by amounts comparable to the second bimodal test density). It is difficult in that the two left modes are so much narrower than the right mode, that highly different bandwidths are appropriate for density estimation of the two regions. Five hundred samples of size 100 were drawn from this density and tested both by the mode-existence test (on a complete mode tree) and by Silverman's critical-bandwidth test (testing the hypotheses of one through five modes for each sample). The results can be seen in Table 2.

As would be expected from the nature of the density, Silverman's test had severe problems detecting the separate natures of the two smaller modes. The test rejected unimodality for bimodality in all but 11 of the 500 tests performed, when tested at the $\alpha = 0.05$ level. Unfortunately, it only rejected bimodality in favor of trimodality in 25 samples, none of which were among the 158 samples in which the three true modes were the first three to appear. In none of the other 342 is there *any* chance to successfully identify the three true modes and no false ones using a global test such as Silverman's.

Even with an α of 0.4, almost three out of four samples still fail to reject bimodality (correctly or otherwise). One final indication of the problems a

TABLE 2
Percentages from testing 500 *samples of size* 100 *from* $\frac{1}{5}N(4,1) + \frac{1}{5}N(8,1) + \frac{3}{5}N(20,5^2)$ *using the mode-existence test and Silverman's critical bandwidth test with several choices of* α *level*

Modes found	**Mode-existence test**			**Critical-bandwidth test**			
Significance level	0.05	0.10	0.15	0.05	0.10	0.15	0.4
1	12.2	3.2	0.4	2.2	0.4	0.2	0.0
2	51.8	30.8	14.2	92.8	91.0	87.4	73.8
3 (correct)	26.8	32.8	28.2	0.0	0.2	0.6	1.8
3 (incorrect)	4.6	10.4	12.2	5.0	7.6	10.6	12.4
4	4.6	17.8	28.4	0.0	0.6	0.8	4.0
5	0.0	4.6	12.4	0.0	0.2	0.4	4.6
6 (or more)	0.0	0.4	4.2	0.0	0.0	0.0	3.4

global test can have with a density such as this is that for 223 of the samples (including 84 of the 158 with the proper first three modes), the p-value for rejecting bimodality in favor of trimodality is the largest p-value of the first five.

The mode-existence test, on the other hand, had considerably greater success with this density. At the $\alpha = 0.05$ level, 134 of the 500 densities indicated the correct three modes (and no others), as did 164 and 141 at the $\alpha = 0.10$ and 0.15 levels, respectively. It is clear, however, that the far larger number of tests here leads to a less conservative procedure overall; 46 of the samples found spurious modes at the $\alpha = 0.05$ level, and by $\alpha = 0.15$, 286 did. Even at the $\alpha = 0.4$ level, Silverman's test found spurious modes in only 122 of the samples. In a sense, this is analogous to a bias versus variance tradeoff. Silverman's test has (in this case) a very low variance, but is biased. The mode-existence test, on the other hand, is far less biased, but displays greater variability. Clearly, there is a choice to be made here between using a test which will have trouble finding modes of varying sizes, and using a procedure which, though conservative for individual tests, will more often find spurious modes when repeated for many apparent modes from a single data set.

6. Future directions.

A variety of potential lines of inquiry exist for improving, generalizing or making use of the mode existence test.

For example, it would be desirable to generate a p-value, not only on the existence of individual modes, but also on the number of modes itself, similar to the way that Silverman's test does. Whether there is some way of combining the information gained from the test into a single p-value on the number of modes is an open question, but if possible, it would be worthwhile.

The simulation results of Section 5 indicate that the test tends to be conservative, resulting in higher p-values than necessary. It might be possible to take results such as those for the uniform at a given sample size and make a monotone transformation so that the final values would be precise for the uniform density and less conservative for other unimodal densities. This would reduce the stated p-values, resulting in greater power, while keeping the p-values and α levels meaningful.

Another improvement might come in the manner of simulation. By drawing from a modified kernel density estimate, we open ourselves to variance inflation caused by the smoothing. If we could find a set of weights for the points which provided a weighted kernel density estimate consistent with the null hypothesis, we could use those weights to select weighted, nonsmoothed bootstrap samples from the data and avoid this spreading. The proper manner of choosing such weights appears to be the most difficult aspect of this proposal.

Fertile ground for generalization of the mode existence test lies in several directions. Extending the test to a multivariate density estimation setting seems conceptually straightforward, but practically quite challenging. Within the univariate setting, the test could be extended to examine "bumps," regions of negative second derivative, much as it currently handles modes. It might also be fruitful to examine local maxima in settings other than density estimation, such as nonparametric regression or time series spectral density estimation.

Finally, Minnotte and Jawhar (1995) take the results of the mode-existence test and the mode tree, and apply them to produce a useful adaptive density estimate, which might show only those modes deemed significant at whatever level is desired.

APPENDIX

PROOF OF THEOREM 1. If the estimates are unimodal, the theorem follows trivially, as $u_0(h_1) = u_0(h_2) = -\infty$, $u_2(h_1) = u_2(h_2) = +\infty$ and $\hat{f}(u_0(h_1), h_1) = \hat{f}(u_2(h_1), h_1) = \hat{f}(u_0(h_2), h_2) = \hat{f}(u_2(h_2), h_2) = 0$. Clearly, in this case, $M_1(h_1) = M_1(h_2) = 1$.

If the estimates are not unimodal, then for some h satisfying $h_1 \le h \le h_2$, suppose $\hat{f}(u_{i-1}(h), h) \ge \hat{f}(u_{i+1}(h), h)$. Define $w(h)$ to be the unique solution to $\hat{f}(x, h) = \hat{f}(u_{i-1}(h), h)$ in the range $u_i(h) < x \le u_{i+1}(h)$. Then

$$\begin{aligned}\frac{\partial}{\partial h} M_i(h) = {}& \big[\hat{f}(w(h), h) - \hat{f}(u_{i-1}(h), h)\big]\frac{\partial}{\partial h} w(h)\\ &- \big[\hat{f}(u_{i-1}(h), h) - \hat{f}(u_{i-1}(h), h)\big]\frac{\partial}{\partial h} u_{i-1}(h)\\ &+ \int_{u_{i-1}(h)}^{w(h)} \frac{\partial}{\partial h}\big[\hat{f}(x, h) - \hat{f}(u_{i-1}(h), h)\big]\,dx.\end{aligned}$$

The difference factors in the first two terms are 0, so we can confine our investigation to the third term, using Propositions 1 and 2:

$$\begin{aligned}\frac{\partial}{\partial h} M_i(h) = {}& \int_{u_{i-1}(h)}^{w(h)} h \frac{\partial^2}{\partial x^2} \hat{f}(x, h)\,dx\\ &- \int_{u_{i-1}(h)}^{w(h)} \left[h \frac{\partial^2}{\partial x^2} \hat{f}(u_{i-1}(h), h) + \frac{\partial}{\partial x} \hat{f}(u_{i-1}(h), h) \frac{\partial}{\partial h} u_{i-1}(h)\right] dx.\end{aligned}$$

The second term of the second integral will be 0 since $u_{i-1}(h)$ is an antimode. Therefore, we arrive at

$$\begin{aligned}\frac{\partial}{\partial h} M_i(h) = {}& h\left[\frac{\partial}{\partial x} \hat{f}(w(h), h) - \frac{\partial}{\partial x} \hat{f}(u_{i-1}(h), h)\right]\\ &- h\big[w(h) - u_{i-1}(h)\big] \frac{\partial^2}{\partial x^2} \hat{f}(u_{i-1}(h), h).\end{aligned}$$

The bandwidth h and the other factors of the second term are positive. The derivatives with respect to x of $\hat{f}(u_{i-1}(h), h)$ and $\hat{f}(w(h), h)$ are 0 and nonpositive, respectively. Therefore, the derivative with respect to h of $M_i(h)$ is negative. A similar argument shows that this last conclusion also holds when $\hat{f}(u_{i-1}(h), h) \le \hat{f}(u_{i+1}(h), h)$. Since M_i is strictly decreasing in h, the theorem follows. □

PROOF OF THEOREM 2. We prove Theorem 2 for the case M_1 being tested at $h_{\mathrm{crit},\,2}$. The other cases have similar proofs. Let $u_1 < u_3$ be the modes of $\hat{f}_n$, and u_2 be the antimode. Also, let $w < u_1$ be such that $\hat{f}_n(w) = \hat{f}_n(u_2)$:

$$\begin{aligned} M_1 &= \int_w^{u_2} \left|\hat{f}_n(x) - \hat{f}_n(u_2)\right| dx \\ &\le (u_2 - w)\big(\hat{f}_n(u_1) - \hat{f}_n(u_2)\big) \\ &\le (|u_2 - z_1| + |z_1 - w|)\big(\left|\hat{f}_n(u_1) - f(z_1)\right| + \left|f(z_1) - \hat{f}_n(u_2)\right|\big). \end{aligned}$$

Mammen, Marron and Fisher (1992) show that the first element is $O_P(n^{-1/5})$. The same result, combined with Silverman's (1978) result that when h is of order $n^{-1/5}$, $\sup_x |\hat{f}_n(x) - f(x)|$ is $O_P(n^{-2/5}(\log n)^{1/2})$, can show that the elements of the second term share the latter rate, as does $|f(w) - f(z_1)|$. This and the fact that $f(w) - f(z_1) = (w - z_1)^2 f''(\xi)/2$ for some ξ between w and z_1 imply that $|z_1 - w|$ is $O_P(n^{-1/5}(\log n)^{1/4})$. The theorem follows. □

PROOF OF THEOREM 3. We consider Theorem 3 for the case M_1 being tested at $h_{\mathrm{crit},\,2}$, with the other cases having similar proofs. Let u_1, u_2, u_3 and w be as in the proof of Theorem 2. Also, let $t < z_1$ be such that $f(t) = f(z_2)$:

$$\begin{aligned} \left|M_1 - \mathscr{M}_1\right| &\le (\max\{z_2, u_2\} - \min\{t, w\})\big(\left|\hat{f}_n(u_2) - f(z_2)\right| + \sup_x\left|\hat{f}_n(x) - f(x)\right|\big) \\ &\le \big(\left|u_2 - z_2\right| + \left|z_2 - t\right| + \left|t - w\right|\big) \\ &\quad \times \big(\left|\hat{f}_n(u_2) - f(z_2)\right| + \sup_x\left|\hat{f}_n(x) - f(x)\right|\big). \end{aligned}$$

Again, using Silverman's (1978) result, the final element is $O_P(n^{-2/5}(\log n)^{1/2})$, as is the penultimate element when we bring in Mammen, Marron and Fisher's (1992) result. These also allow us to show that $|u_2 - z_2|$ is $O_P(n^{-1/5})$, and, with the fact that $f(w) - f(t) = (w - t)f'(\xi)$ for some ξ between w and t, imply that $|t - w|$ is $O_P(n^{-2/5}(\log n)^{1/2})$. As both of these elements are dominated by the constant $[O_P(1)]$ $|z_2 - t|$ term, the theorem follows. □

Acknowledgments. The author would like to thank his dissertation advisor, David W. Scott, for his invaluable aid in this research. He would also like to thank the Editors and reviewers for suggestions which have substantially improved this paper.

REFERENCES

AHRENS, L. H. (1965). Observations on the Fe–Si–Mg relationship in chondrites. *Geochim. Cosmochim. Acta* **29** 801–806.

BABAUD, J., WITKIN, A. P., BAUDIN, M. and DUDA, R. O. (1986). Uniqueness of the Gaussian Kernel for scale-space filtering. *IEEE Trans. Pattern Anal. Machine Intel.* **PAMI-8** 26–33.

BESAG, J. and CLIFFORD, P. (1991). Sequential Monte Carlo p-values. *Biometrika* **78** 301–304.

FAN, J. and MARRON, J. S. (1994). Fast implementations of nonparametric curve estimators. *J. Comput. Graph. Statist.* **3** 35–56.

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

HALL, P. and WOOD, A. T. A. (1993). Approximations to distributions of statistics used for testing hypotheses about the number of modes of a population. Technical Report CMA-SR14-93, Centre for Mathematics and its Applications, Australian National Univ.

HARTIGAN, J. A. (1988). The span test of multimodality. In *Classification and Related Methods of Data Analysis* (H. H. Bock, ed.) 229–236. North-Holland, Amsterdam.

HARTIGAN, J. A. and HARTIGAN, P. M. (1985). The dip test of unimodality. *Ann. Statist.* **13** 70–84.

HARTIGAN, J. A. and MOHANTY, S. (1992). The RUNT test for multimodality. *J. Classification* **9** 63–70.

IZENMAN, A. J. and SOMMER, C. (1988). Philatelic mixtures and multimodal densities. *J. Amer. Statist. Assoc.* **83** 941–953.

JONES, M. C. (1989). Discretized and interpolated kernel density estimates. *J. Amer. Statist. Assoc.* **84** 733–741.

MAMMEN, E., MARRON, J. S. and FISHER, N. I. (1992). Some asymptotics for multimodality tests based on kernel density estimates. *Probab. Theory Related Fields* **91** 115–132.

MATTHEWS, M. V. (1983). On Silverman's test for the number of modes in a univariate density function. B. A. honors thesis, Dept. Statistics, Harvard Univ.

MINNOTTE, M. C. (1992). A test of mode existence with applications to multimodality. Ph.D. thesis, Dept. Statistics, Rice Univ.

MINNOTTE, M. C. and JAWHAR, N. S. (1995). Adaptive kernel density estimation using mode information. *Comput. Sci. Statist.* **27** 545–549.

MINNOTTE, M. C. and SCOTT, D. W. (1993). The mode tree: a tool for visualization of nonparametric density features. *J. Comput. Graph. Statist.* **2** 51–68.

MÜLLER, D. W. and SAWITZKI, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738–746.

POLONIK, W. (1993). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. Technical report, Beitraege zur Statistik Nr. 7, Univ. Heidelberg.

ROEDER, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* **85** 617–624.

ROZÁL, G. P. M. and HARTIGAN, J. A. (1994). The MAP test for multimodality. *J. Classification* **11** 5–36.

SCOTT, D. W. (1985a). Average shifted histograms: effective nonparametric density estimators in several dimensions. *Ann. Statist.* **13** 1024–1040.

SCOTT, D. W. (1985b). Frequency polygons: theory and applicaton. *J. Amer. Statist. Assoc.* **80** 348–354.

SILVERMAN, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6** 177–184.

SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43** 97–99.

SILVERMAN, B. W. (1983). Some properties of a test for multimodality based on kernel density estimates. In *Probability, Analysis, and Statistics. IMS Lecture Notes* (J. F. C. Kingman and G. E. H. Reuter, eds.) **79** 248–259. Cambridge Univ. Press.

WONG, M. A. (1985). A bootstrap testing procedure for investigating the number of subpopulations. *J. Statist. Comput. Simulation* **22** 99–112.

DEPARTMENT OF MATHEMATICS AND STATISTICS
UTAH STATE UNIVERSITY
LOGAN, UTAH 84322-3900
E-MAIL: minnotte@math.usu.edu