

## **INFORMATION TO USERS**

**This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.**

**The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.**

**In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.**

**Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.**

**Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.**

# **U·M·I**

University Microfilms International  
A Bell & Howell Information Company  
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA  
313/761-4700 800/521-0600



**Order Number 9408649**

**A test of mode existence with applications to multimodality**

**Minnotte, Michael C., Ph.D.**

**Rice University, 1993**

**U·M·I**  
300 N. Zeeb Rd.  
Ann Arbor, MI 48106



RICE UNIVERSITY

**A Test of Mode Existence  
with Applications to Multimodality**

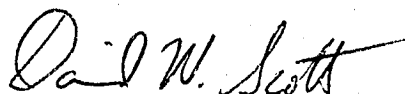
by

**Michael C. Minnotte**

A THESIS SUBMITTED  
IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE

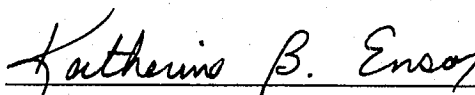
**Doctor of Philosophy**

APPROVED, THESIS COMMITTEE:



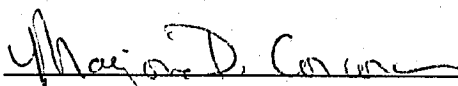
---

David W. Scott, Chairman  
Professor of Statistics



---

Katherine B. Ensor  
Assistant Professor of Statistics



---

Marjorie D. Corcoran  
Professor of Physics

Houston, Texas

July, 1992

# **A Test of Mode Existence with Applications to Multimodality**

Michael C. Minnotte

## **Abstract**

Modes, or local maxima, are often among the most interesting features of a probability density function. Given a set of data drawn from an unknown density, it is frequently desirable to know whether or not the density is multimodal, and various procedures have been suggested for investigating the question of multimodality in the context of hypothesis testing. Available tests, however, suffer from the encumbrance of testing the entire density at once, frequently through the use of nonparametric density estimates using a single bandwidth parameter. Such a procedure puts the investigator examining a density with several modes of varying sizes at a disadvantage. A new test is proposed involving testing the reality of individual observed modes, rather than directly testing the number of modes of the density as a whole. 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 combined with the graphical notion of a "mode tree," in which estimated mode locations are plotted over a range of kernel bandwidths. In this way, one can obtain a procedure for examining, in an adaptive fashion, not only the reality of individual modes, but also the overall number of modes of the density. A proof of consistency of the test statistic is offered, simulation results are presented, and applications to real data are illustrated.

## Acknowledgments

I would like to thank the members of my committee, Professors David W. Scott, Katherine B. Ensor, and Marjorie D. Corcoran for their advice, assistance, and moral support. I owe especially heartfelt thanks to my advisor, David Scott, whose expertise, guidance, and idea of the mode tree have helped immeasurably in the production of this research. Matthew P. Wand was also of great assistance. Finally, I would like to thank my wife, Letitia, for keeping me sane and for being understanding all those nights when I had to do “just one more simulation!” before coming home.

This research was supported in part by the Office of Naval Research under contract N00014-90-J-1176 and by Department of Defense NDSEG Fellowship grant number DAAL03-90-G-0169.

# Contents

Abstract	ii
Acknowledgments	iii
List of Illustrations	vii
List of Tables	ix
<b>1 Introduction</b>	<b>1</b>
1.1 The Problem of Multimodality . . . . .	1
1.2 Kernel Density Estimation . . . . .	2
1.3 Outline of Results . . . . .	7
<b>2 Survey of Current Techniques</b>	<b>8</b>
2.1 Bin-Difference Investigations . . . . .	8
2.2 The Maximum Penalized-Likelihood Method . . . . .	9
2.3 The Critical Kernel-Bandwidth Test . . . . .	11
2.4 The Dip Test . . . . .	15
2.5 The Excess Difference Test . . . . .	18
2.6 Nearest-Neighbor Procedures . . . . .	19
2.7 Density Confidence Sets . . . . .	21
2.8 Some Theoretical Results with Implications for Multimodality Tests .	21
2.9 A Multimodal Mode Estimation Procedure . . . . .	22
2.10 Final Thoughts . . . . .	23
<b>3 A Pair of Useful Tools</b>	<b>25</b>
3.1 A Univariate Point-Matching Algorithm . . . . .	25



3.2	The Mode Tree . . . . .	27
<b>4</b>	<b>The Mode-Existence Test</b>	<b>35</b>
4.1	Testing Individual Modes . . . . .	36
4.2	Calculating $p$ -values . . . . .	37
4.3	The Use of Sequential Monte Carlo $p$ -values . . . . .	41
4.4	Examining Multimodality of the Overall Data Set . . . . .	42
4.5	Implementation . . . . .	46
<b>5</b>	<b>Theoretical Investigations</b>	<b>49</b>
5.1	The Relationship Between $M_i$ and $h$ . . . . .	49
5.2	Consistency and Rate of Convergence of $M_i$ . . . . .	51
<b>6</b>	<b>Simulation Studies</b>	<b>63</b>
6.1	Statistical Significance . . . . .	63
6.2	Power . . . . .	64
6.3	The Adaptive Nature of the Test . . . . .	70
<b>7</b>	<b>Case Studies</b>	<b>74</b>
7.1	Chondrite Data . . . . .	74
7.2	Stamp Data . . . . .	76
7.3	Snowfall Data . . . . .	79
7.4	Galaxy Data . . . . .	81
7.5	LRL Data . . . . .	83
7.6	Rayleigh's Nitrogen Data . . . . .	85
<b>8</b>	<b>Future Directions</b>	<b>88</b>
8.1	Combined $p$ -values . . . . .	88
8.2	Adjusted $p$ -values — Greater Power . . . . .	88

8.3	Investigating Bumps . . . . .	89
8.4	Multivariate Data . . . . .	89
8.5	Regression and Spectral Density Estimation . . . . .	90
8.6	Adaptive Density Estimation . . . . .	90
<b>References</b>		<b>92</b>

## Illustrations

1.1	Normal kernel density estimate of the chondrite data . . . . .	4
1.2	Twenty Normal kernel density estimates of the chondrite data . . . . .	5
1.3	Perspective plot of $\psi(x, h) = \hat{f}_h(x)$ for the chondrite data . . . . .	6
2.1	An example of a troublesome density . . . . .	16
2.2	A kernel estimate of the above density, based on a sample of 100 points	16
3.1	Example of the point-matching algorithm . . . . .	26
3.2	Normal kernel mode tree for the chondrite data . . . . .	29
3.3	Biweight kernel mode tree for the chondrite data . . . . .	30
3.4	Enhanced Normal mode tree for the chondrite data . . . . .	32
3.5	$M_1$ , $M_3$ , and $M_5$ are equal to the shaded areas for the chondrite data when $h = 1.0$ . . . . .	33
3.6	Enhanced biweight mode tree for the chondrite data . . . . .	34
4.1	Examples of the density-choosing process . . . . .	40
4.2	Quantiles of sequential estimates of $p$ . . . . .	43
4.3	Test mode tree for the chondrite data . . . . .	45
6.1	Test density #1: standard Normal density . . . . .	65
6.2	Test density #2: Uniform density . . . . .	66
6.3	Test Density #3: $\frac{1}{2}N(-1.5, 1) + \frac{1}{2}N(1.5, 1)$ . . . . .	68

---

6.4	Test Density #4: $\frac{1}{2}N(-2, 1) + \frac{1}{2}N(2, 1)$ . . . . .	69
6.5	Test Density #5: $\frac{3}{4}N(0, 1) + \frac{1}{4}N(2, \frac{1}{3})$ . . . . .	71
6.6	Test Density #6: $\frac{1}{5}N(4, 1) + \frac{1}{5}N(8, 1) + \frac{3}{5}N(20, 5)$ . . . . .	72
7.1	Test mode tree for the chondrite data . . . . .	75
7.2	Kernel density estimates for the condrite data; $h = 0.75$ . . . . .	75
7.3	Test mode tree for the Hidalgo stamp data . . . . .	78
7.4	Kernel density estimates for the Hidalgo stamp data; $h = 0.0005$ . . . . .	78
7.5	Kernel density estimates for the Buffalo snowfall data; $h = 4.0$ . . . . .	80
7.6	Test mode tree for the Buffalo snowfall data . . . . .	80
7.7	Test mode tree for the galaxy velocity data . . . . .	82
7.8	Kernel density estimates for the galaxy velocity data; $h = 500$ . . . . .	82
7.9	Kernel density estimates for the LRL data; $h = 20$ . . . . .	84
7.10	Test mode tree for the LRL data . . . . .	84
7.11	Kernel density estimates for the Raleigh “nitrogen” data; $h = 0.001$ . . . . .	86
7.12	Test mode tree for the Rayleigh “nitrogen” data . . . . .	86

## Tables

1.1	Popular Kernels . . . . .	3
4.1	Estimated quantiles for $p$ -values calculated using sequential Monte Carlo techniques . . . . .	44
6.1	Results of testing 20 samples of size 100 from Test Density #6 . . . .	73
7.1	Modes of chondrite mode tree with $p$ -values less than 0.15 . . . . .	76
7.2	Modes of Hidalgo stamp mode tree with $p$ -values less than 0.15 . . .	77
7.3	Modes of Buffalo snowfall mode tree with $p$ -values less than 0.15 . . .	79
7.4	Modes of galaxy mode tree with $p$ -values less than 0.15 . . . . .	81
7.5	Modes of LRL mode tree with $p$ -values less than 0.15 . . . . .	85
7.6	Modes of Rayleigh nitrogen mode tree with $p$ -values less than 0.15 . .	87

# Chapter 1

## Introduction

### 1.1 The Problem of Multimodality

The identification of modes and their generalizations, “bumps,” 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). While by far the most common interpretation of multimodality is that of a mixture distribution containing several subpopulations, or as an indication of clustering, there are other possibilities as well. For example, Comparini and Gori (1986) discuss a family of densities related to nonlinear diffusion processes which satisfy the relation  $f'(x)/f(x) = -g(x)/v(x)$ , with  $g(x)$  and  $v(x)$  polynomials in  $x$  of given orders. In any case, identifying multimodality is desirable when it exists, but we do not wish to give too much importance to apparent modes caused merely by random fluctuations in the data. The resulting dilemma is a problem which has received substantial attention in recent years.

We begin with some definitions. A *mode* of a probability distribution is defined simply as a local maximum in the associated probability density function, or sometimes as a point  $m$  such that there exist points  $a$  and  $b$ ,  $a < m < b$ , where the distribution function is convex in  $(a, m]$  and concave in  $[m, b)$ . Clearly, these two definitions are equivalent except for the case of a density function with constant values at its peak. In this latter case, all of the points on this one constant peak shall be considered a single mode.

A *bump*, on the other hand is defined by Good and Gaskins (1980) (in a manner which specifically excludes the possibility of linear stretches in the density) as the

concave segment of a density function lying between two inflection points. Clearly a mode and some surrounding region would be an example of a bump, but the converse is not necessarily true. Bumps are of interest for much the same reason as modes, primarily because they can indicate mixtures. While the focus of this discussion is firmly on the problem of identifying multiple modes, the extension to bumps is relatively straightforward.

Different techniques in the field of multimodality testing have aimed at different goals. The key distinction for our purposes lies between *local* and *global* tests. 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. See, for example, the tests and procedures proposed by Silverman (1981), Hartigan and Hartigan (1985), Wong (1985), Roeder (1990), and Müller and Sawitzki (1991).

The ideas propounded in this paper, on the other hand, follow the more local direction 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.

## 1.2 Kernel Density Estimation

Questions of multimodality generally occur within an exploratory setting. In such a setting, the number and location of features such as modes and bumps is not known a priori. The parametric form of the density is generally poorly understood and any tentative choice may be too restrictive. After all, if the parametric form of the density is known, the number of modes (or at least an upper bound) will usually be known as well. Under such uncertain circumstances, a nonparametric density estimation technique can be highly valuable.

Kernel density estimation is a popular example of nonparametric density estimation (Rosenblatt, 1956; Parzen, 1962), and is central to several multimodality tests, including the one developed in this study. Given a sample  $\{X_1, \dots, X_n\}$  of size  $n$ , the kernel density estimate at  $x$  is computed as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i), \quad (1.1)$$

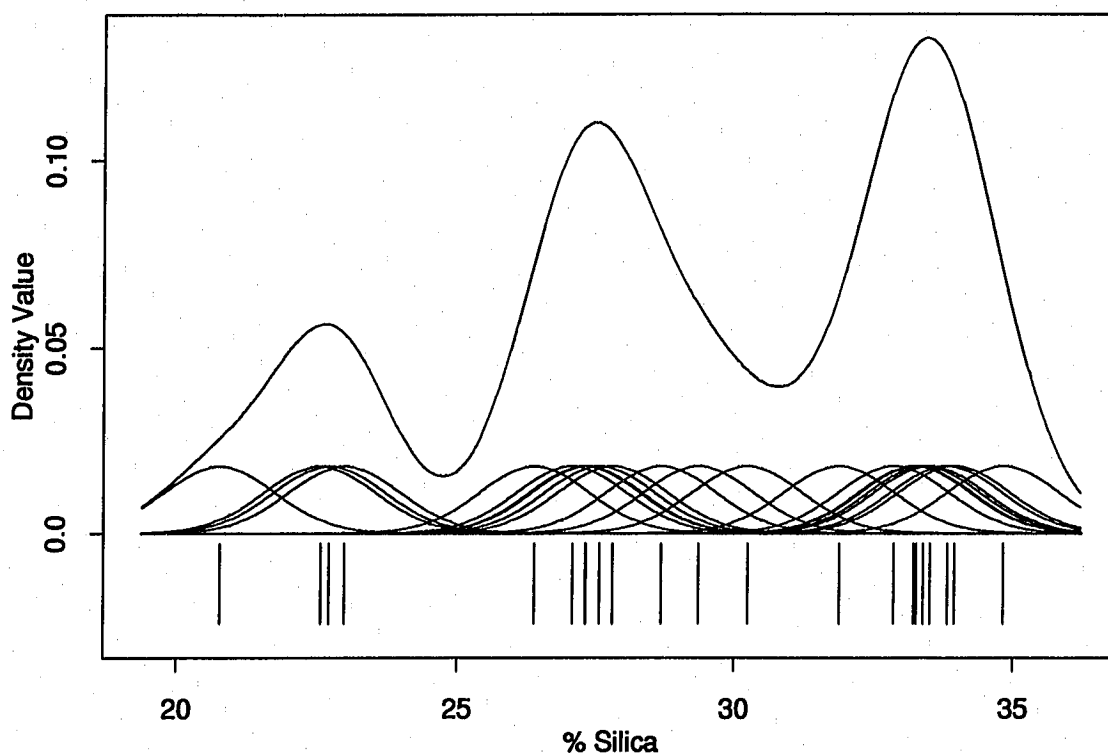
where  $K_h(t) = K(t/h)/h$  and  $K$  is known as the kernel. The kernel estimator is equivalent to a mixture density with the function  $K$  equally weighted and centered at each  $X_i$ . If the kernel  $K$  itself is assumed to be a density function, then  $\hat{f}_h(\cdot)$  is nonnegative and integrates to one.  $K$  is generally taken to be symmetric with mean 0 and positive variance. Popular examples are given in Table 1.1.

Kernel	Value	Nonzero Domain
Uniform	$\frac{1}{2}$	$[-1, 1]$
Triangle	$1 -  x $	$[-1, 1]$
Epanechnikov	$\frac{3}{4}(1 - x^2)$	$[-1, 1]$
Biweight	$\frac{15}{16}(1 - x^2)^2$	$[-1, 1]$
Normal	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$	$(-\infty, \infty)$

**Table 1.1** Popular Kernels

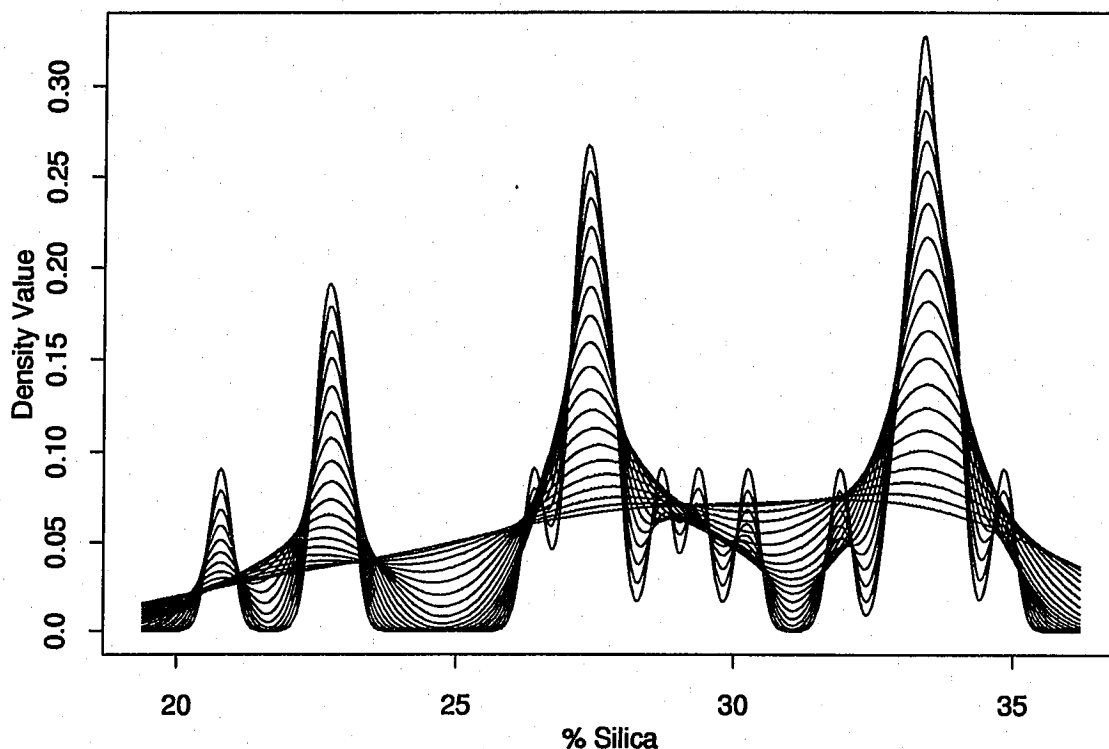
In Figure 1.1, a kernel density estimate and its component kernels are displayed for the chondrite meteorite data (Ahrens, 1965). These data, the percentages of silica in 22 chondrite meteors, were first discussed in the bump-hunting context by Good and Gaskins (1980). The vertical lines below the  $x$ -axis represent the values of the individual data points. The Normal kernel is used in this estimate, which appears trimodal. We shall return to the example of the chondrite data several times.





**Figure 1.1** Normal kernel density estimate of the chondrite data with  $h = 1.0$ , along with the data points (at bottom) and the individual kernel masses which make up the estimate.

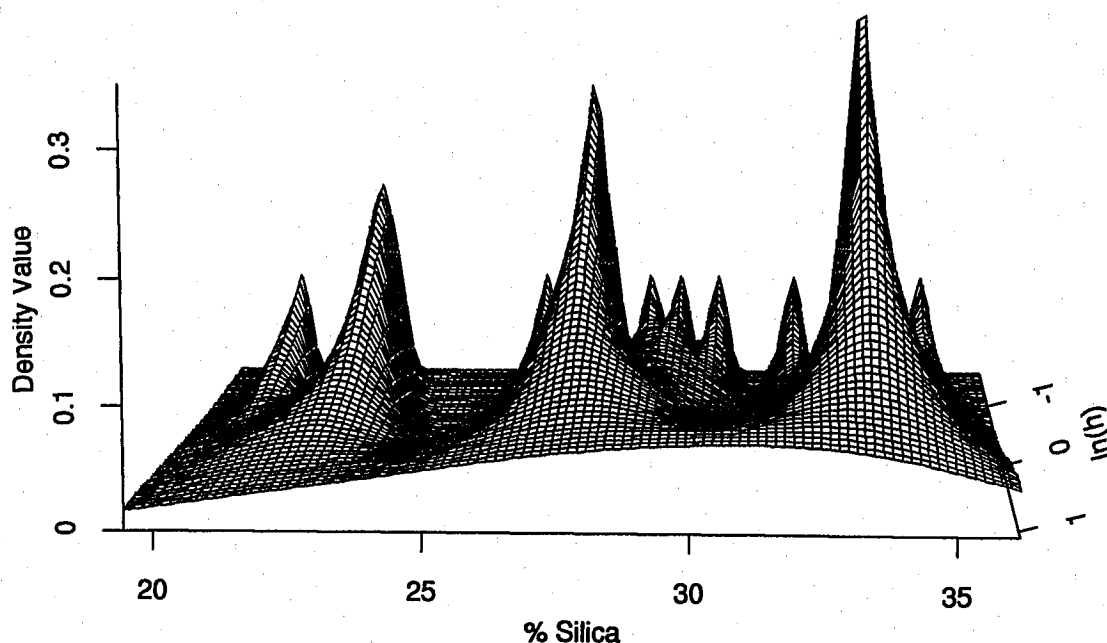
After  $K$  has been chosen, the remaining element to be specified is the bandwidth parameter  $h$ . The bandwidth is a rescaling factor which determines the extent of the region over which the probability mass for point  $X_i$  is spread. The choice of this parameter is quite crucial to the final density estimate both locally and globally. In Figure 1.2, we display twenty different Normal kernel density estimates of the chondrite data, with bandwidths ranging from 0.2 to 3.0. The amount of information about the number and location of potential bumps and modes is quite large in this figure, to the point of being overwhelming. The number of modes and bumps increases rapidly as  $h \rightarrow 0$  (though it can never become greater than  $n$ ).



**Figure 1.2** Twenty Normal kernel density estimates of the chondrite data. Bandwidths are equally spaced on a logarithmic scale, and range from 0.2 to 3.0.

Another view of the same information can be seen in Figure 1.3. Here, a sequence of fifty density estimates over the same range of bandwidths from 0.2 to 3.0 is viewed using a perspective plot of the function  $\psi(x, h) = \hat{f}_h(x)$ , again showing the vastly different estimates produced by density estimates using various bandwidths.

Unfortunately,  $h$  must be selected by the user, and no completely satisfactory method of doing so has been found. [Scott (1992) presents a survey of available techniques that focus on minimizing  $L_2$  error, a criterion which is only loosely related to bumps and modes.] Even assuming the best global choice for  $h$ , the fact remains that no single value of  $h$  will perform well for all points  $x$  (Terrell and Scott, 1992), as we will demonstrate in the Chapters 6 and 7. Jones (1990) and Terrell and Scott



**Figure 1.3** Perspective plot of  $\psi(x, h) = \hat{f}_h(x)$  for the chondrite data. Fifty density estimates are plotted, with bandwidths ranging on a logarithmic scale from 0.2 (at the rear of the plot) to 3.0 (at the front).

(1992) have investigated the theoretical and practical advantages of an adaptive kernel estimate introduced by Breiman, et al. (1977) and Abramson (1982):

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right) = \frac{1}{n} \sum_{i=1}^n K_{h_i}(x - X_i), \quad (1.2)$$

where  $h$  varies along  $x$ . Unfortunately, the original problem of choosing one smoothing parameter  $h$  in Equation 1.1 has been replaced by the necessity of choosing many smoothing parameters in Equation 1.2. In a related approach, Wand, Marron, and Ruppert (1991) have examined the use of data transformation families in Estimator 1.1 to address the problem. Here, the difficulty of choice of bandwidth remains, plus there is the additional problem of choice of transformation, requiring at least another pa-

parameter or two. No method seems to be sufficiently general to work for all densities or data sets.

### 1.3 Outline of Results

In this thesis, when we focus our attention on a particular mode, we desire an adaptive estimator. We compromise by using a fixed, but different, bandwidth for each potential mode. By testing the reality of each individual mode at a bandwidth appropriate to that mode, we can accomplish the goal of finding a usefully adaptive procedure, without the problems inherent in attempting to produce an adaptive density estimate. This is assisted by the use of a graphical tool known as a “mode tree,” which will be used both to choose the bandwidths at which to test, and to display the results after they are computed. Use of the mode tree will also allow us, in the end, to go back and examine the multimodality of the density as a whole.

In Chapter 2, we examine some key procedures which have been developed for investigating multimodality. We describe in detail the mode tree and a useful point-matching algorithm in Chapter 3, followed by the description of the new test in Chapter 4. Chapter 5 contains some key theorems relating to the mode-existence test, and Chapters 6 and 7 investigate the use of the procedure on simulated and real data, respectively. We close in Chapter 8 with some thoughts on directions which future investigations related to those of this thesis might take.

## Chapter 2

### Survey of Current Techniques

Before venturing into a discussion of the new procedure, it is worth discussing the alternatives which currently exist for the analyst interested in multimodality. In addition, we will mention several results which are not multimodality algorithms themselves, but which have important theoretical implications to such procedures.

#### 2.1 Bin-Difference Investigations

Cox (1966) provides an early example of investigating multimodality. Cox is concerned almost solely with detecting bimodality, though his suggestions are not limited to unimodal or bimodal densities in general. His first suggestion is to follow Haldane (1952) and examine test statistics on equality of pairs of adjacent population bin frequencies and on linearity of triplets of adjacent population bin frequencies. If  $\nu_i$  is the observed frequency in the  $i$ th bin, the relevant test statistics are

$$m_i = \frac{|\nu_{i+1} - \nu_i| - 1}{\sqrt{\nu_{i+1} + \nu_i}}$$

and

$$t_i = \frac{|\nu_{i+1} - 2\nu_i + \nu_{i-1}| - (3/2)}{\sqrt{2(\nu_{i+1} + \nu_i + \nu_{i-1})}}.$$

Binomial arguments indicate that if the equality and linearity conditions hold respectively, then the two statistics will be distributed approximately as standard normal random variables. Examination of all of the  $m_i$ 's and  $t_i$ 's can give an informal idea of existence and location of modes, at least in regions where the sample counts are sufficiently high.

The other method he discusses is that of fitting the parameters of a specified functional form (in his case, a mixture of two normal distributions) followed by comparison to a goodness-of-fit statistic, such as a chi-squared statistic. Cox acknowledges that the precision on the estimates of the mixing proportions are very low, even when all parameters of the two mixed distributions are known. He suggests a graphical method based on the assumption that the tails will be composed almost exclusively of data from the nearer of the two normal distributions. Plots on normal probability paper (or equivalently scaled by computer) of  $x$  versus the number of points less than  $x$  divided by the sample size and the probability of a data point being from the lower-meaned component should give a straight line for  $x$  sufficiently small. A similar result holds for the upper tail and the probability associated with the upper component.

## 2.2 The Maximum Penalized-Likelihood Method

Good and Gaskins (1980) proposal was one of the earliest of the recent group of bump hunting and multimodality testing procedures, and it contains several features which have not appeared since. In addition to being the only local procedure, it is also the only method for assessing bumps as well as modes. They refer to their method as Maximum Penalized Likelihood, or MPL.

The MPL description applies essentially to the density-estimation technique they use en route to the bump hunting procedure, so it will be described only briefly. In essence, they find the function  $f$  (built from either Fourier or Hermite polynomials), which maximizes a score  $\omega = \omega(f(X))$  given the data set  $\{X\}$ . The score  $\omega$  is defined as the log-likelihood of  $f$ , minus  $\beta \int [\gamma''(x)]^2 dx$ , where  $\gamma(x) = \sqrt{f(x)}$  and  $\beta$  is a positive parameter. We call  $\int [\gamma''(x)]^2 dx$  the roughness or, together with  $\beta$ , the roughness penalty.

Good and Gaskins use several sample-based estimates of the optimal  $\beta$ , involving chi-squared tests, sign tests, and the Kolmogorov-Smirnov statistic. They also men-

tion that cross-validation, such as in Wahba (1981) is another possibility, but do not use it themselves.

Good and Gaskins then use a process of “surgery” on those bumps which appear in their final density estimate, to calculate odds-ratios as to their existence. The surgery is aided for their primary data set example (of 25,752 events from a Lawrence Radiation Laboratory scattering experiment; see section 7.5) by the fact that their raw data are already binned, rather than being all 25,752 “true” measurements. As Scott (1980) points out in the discussion, their method also uses the property of series estimators that bumps tend to be underestimated.

Their method of surgery is to set the counts of the bins between the inflection points defining the bump in question to the values predicted by the estimate. The numbers of counts in all bins are then rescaled so that the total number of counts is again 25,752, and the density is again estimated from this modified data set. This procedure is then repeated until convergence occurs, or until some large number of iterations has been performed (Good and Gaskins use 14). If convergence has not occurred, the bump in question can be assured of existence; otherwise, the log-odds ratio  $\omega(f_1(X)) - \omega(f_2(X))$  can be calculated, where  $f_1(X)$  is the original estimate of the density,  $f_2(X)$  is the estimate with the bump in question excised, and both log-likelihoods are calculated with the entire original data set. High ratios clearly correspond to strong likelihoods of the existence of bumps.

The key problem with this proposal is the lack of any concept of power or significance. Likelihood ratios can be useful, but more useful still would be p-values. Other minor drawbacks include heavy computational costs and an essentially Bayesian viewpoint for its density estimation technique. Finally, and far more importantly, the examination occurs only at a single value of the smoothing parameter. Even if chosen “optimally,” some features may be hidden, while other, spurious, features may appear. A single clear mode may be obscured by the use of too small a bandwidth at which to test.

## 2.3 The Critical Kernel-Bandwidth Test

Silverman (1981) suggests a method of testing the null hypothesis that a density has at most  $k$  modes. While (as presented) the method does not locate the modes, doing so is a relatively straightforward extension. Silverman's test has some advantages over those of many others in ease of comprehension and computation. It has also been more widely investigated than other tests and procedures involving multimodality. The test utilizes almost exclusively kernel density estimates of the data (see Section 1.2)

Silverman's test is based on the intuitive idea that in a non-adaptive fixed-bandwidth kernel density estimate (in which  $h$  is constant throughout the domain of interest), the number of modes increases as the bandwidth  $h$  decreases. With very large  $h$ , the values are so smoothed as to have only a single mode; as  $h$  goes to 0, the resulting density estimate becomes  $n$ -modal, with a mode at each data point. Silverman (1981) shows that for a normal kernel, the number of maxima of  $\hat{f}$  and all derivatives is a right continuous decreasing function of  $h$ . Therefore, Silverman proposes as a test statistic,

$$h_{crit,k} = \inf\{h; \hat{f}(\cdot, h) \text{ has at most } k \text{ modes}\},$$

where  $\hat{f}(\cdot, h)$  is the normal density kernel estimate of  $f$  with bandwidth  $h$ . Calculation of  $h_{crit,k}$  can be done through a simple binary search procedure.

To calculate a  $p$ -value for this test, Silverman suggests using the  $k$ -modal estimate most consistent with the data, the density estimated by the kernel with smoothing parameter  $h_{crit,k}$ , as the representative of the null hypothesis. Sampling from this distribution is a simple matter; simply draw (with replacement)  $n$  values from the sample data, and add to each a  $\text{Normal}(0, h_{crit,k}^2)$  random variable. Since the variance of a kernel estimate (with  $h^2 = \int x^2 K(x) dx$  and sample variance  $\hat{\sigma}^2$ ) is  $\hat{\sigma}^2 + h^2$ , Silverman follows the proposal of Efron (1979) in recommending that the bootstrapped points be rescaled toward the mean by a factor of  $(\hat{\sigma}^2 / (\hat{\sigma}^2 + h_{crit,k}^2))^{1/2}$ . For each smoothed bootstrap sample, the test calculates  $h_{crit,k}^i$  (or, more easily and equivalently, checks



to see if there are more than  $k$  modes when  $h = h_{crit,k}$ ). Dividing the number with  $h_{crit,k}^i \leq h_{crit,k}$  or more than  $k$  modes by the total number of bootstrap samples gives an approximate  $p$ -value.

Silverman (1983) demonstrates the consistency of such a test under a few conditions. The requirements are that  $f$  be twice continuously differentiable on  $[a, b]$ , that  $f'(a+) > 0$  and  $f'(b-) < 0$ , and that

$$\min_{z: f'(z)=0} \frac{f''(z)^2}{f(z)} > 0.$$

Given these conditions, as  $n \rightarrow \infty$ ,  $h_{crit,k}$  converges stochastically to 0 if the density truly has no more than  $k$  modes, but  $h_{crit,k}$  remains bounded away from 0 if the density has more than  $k$ . More precisely, in the former case where the true number of modes is less than  $k$ , he shows that

$$\text{pliminf } n^{-1} h_{crit,k}^{-5} \log h_{crit,k}^{-1} \geq \frac{2}{3} \pi \sqrt{2} \min_{z: f'(z)=0} \frac{f''(z)^2}{f(z)}$$

and

$$\text{plimsup } n^{-1} h_{crit,k}^{-5} \log h_{crit,k}^{-1} < \infty.$$

Mammen, Marron and Fisher (1990) extend Silverman's asymptotic work and calculate the expected number of modes  $EN(h)$  for a normal kernel density estimator. Under the same conditions as above, and assuming that  $f$  has  $j$  local maxima  $z_1 < z_3 < \dots < z_{2j-1}$  and  $(j-1)$  local minima  $z_2 < \dots < z_{2j-2}$ , they arrive at the following result. If  $0 < \liminf_{n \rightarrow \infty} n^{1/5} h < \limsup_{n \rightarrow \infty} n^{1/5} h < \infty$ , then

$$EN(h) = j + \sum_{p=0}^{2j-2} H \left( \frac{\sqrt{nh^5} |f''(z_p)|}{\|\phi''\| \sqrt{f(z_p)}} \right) + o(1)$$

where  $\|\phi''\| = [f(\phi'')^2]^{1/2} = (3/8)^{1/2} (\pi)^{-1/4}$  and  $H(x) = \phi(x)/x + \Phi(x) - 1$ . ( $\phi$  is the standard normal density and  $\Phi$  is the standard normal distribution function.) Mammen (1991) expands this to a more general class of kernels  $K$  by replacing  $\|\phi''\|$  with  $\|K''\|$ .

Mammen, et al. (1990) also show that under the null hypothesis ( $k \geq j$ ),  $h_{crit,k}$  is of order  $n^{-1/5}$ , the same order as optimal bandwidths for global density estimations under most criteria. Finally, they provide evidence, though not proof, that a bootstrap test at the  $(1 - \alpha)$  level has no more than  $\alpha$  probability of type-I error (as desired).

Matthews (1983) provides some simulation results for Silverman's test. Matthews uses the decision rule to accept a density as  $k$ -modal for the smallest  $k$  in which the estimated  $p$ -value is greater than 0.60. [Actually, he follows an error in Silverman (1981) in which the stated  $p$ -values are actually  $1 -$  (the actual  $p$ -values). Thus his stated rule is to accept the smallest  $k$  for which the significance level is less than 0.40.] The mere fact that such a large value of the significance level must be used to have any real hope of identifying multimodal densities is an indication of both the difficulty of the problem and the low power of Silverman's test, at least at more common significance levels such as 0.05.

Due to computational limitations, the number of tests run for any given number of parameters is small (10 in most cases) and only 50 bootstrap samples are used for each  $p$ -value. For each test, he uses a mixture of two normal distributions with unit variance,

$$f(x) = p\phi(x) + (1 - p)\phi(x - \mu).$$

In the initial simulation ( $p = .5$  and  $\mu = 3$ ), Silverman's test appears to perform fairly well, identifying as bimodal 13 out of 15 samples of size  $n = 256$ . It's performance goes rapidly downhill, however. Using sample sizes of 40, 200, and 1000; and with parameters ( $p = .5$ ;  $\mu = 2.2, 2.7, 3.2$ ) and ( $p = .1, .3, .5$ ;  $\mu = 3.5$ ), Silverman's test and Matthew's decision rule never concluded bimodality for more than 7 out of 10 samples, and only decided correctly in 6 or 7 out of 10 samples in 3 of the 18 cases ( $p = .5, \mu = 3.2, n = 1000$ ; and  $p = .5, \mu = 3.5, n = 200, 1000$ ). In light of these results, it seems likely that the first simulation of 15 runs happened to be an

unusually “lucky” fluke, made more likely by the relatively few trials, than that the test has as much power as indicated for  $p = .5$  and  $\mu = 3$ .

Matthews concludes his simulation study by looking at the convolution of a normal density with a geometric density,

$$f(x) = \sum_{j=0}^{\infty} p(1-p)^j \phi\left(\frac{x-j\alpha}{\sigma}\right).$$

Using parameters  $p = .25$ ,  $\alpha = 2$ , and  $\sigma = .316$ , the density has infinitely many modes. Not surprisingly, Silverman’s test shows great instability in this case as Matthews investigates the first 5 critical bandwidths.

Fisher, Mammen and Marron (1991) suggest an alternative to the rescaling suggested by Silverman (1981) in resampling. Instead of rescaling the data (which they point out has drawbacks in terms of moving modes relative to one another and is useless for directional data), they suggest investigating the resampled data not at  $h_{crit,k}$ , but at some  $h_{crit,k}^*$  chosen to be on the same “smoothing scale,” as measured by the “smoothing ratio”

$$\frac{\int \text{var}(\hat{f}_h)}{\int \text{bias}^2(\hat{f}_h)}.$$

They accomplish this by using a pilot estimate of  $f$  and getting an estimate of

$$\left( \frac{\int (f'')^2}{\int (\hat{f}_{h_{crit,k}}'')^2} \right)^{1/5}.$$

Preliminary tests indicate that this variant tends to have a somewhat more trustworthy significance level at the cost of some power in marginal cases.

Also, Izenman and Sommer (1988) suggest the possibility of replacing the fixed-bandwidth kernel estimate and  $h_{crit,k}$  with an adaptive estimate and a statistic based on an average, rather than overall, bandwidth. This suggestion is motivated by the observation that two large, close, narrow modes may separate from each other at much smaller  $h$  than one at which small, possibly spurious, modes appear in the tail. This particular idea appears to be yet unfulfilled, although the suggestions of this thesis follow from similar thoughts.

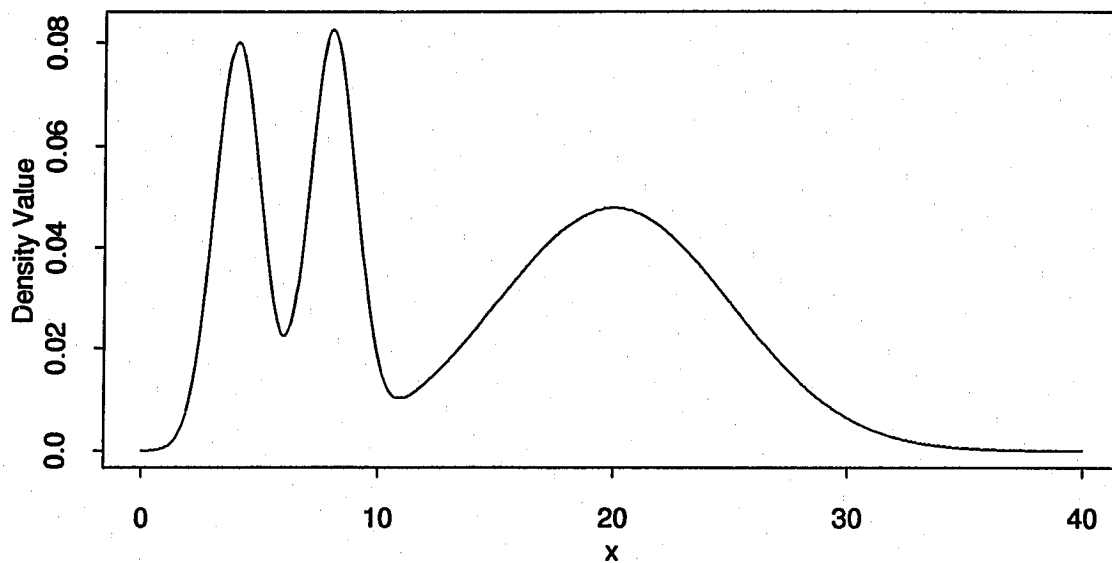
Silverman's test examines the density as a whole, rather than looking at the individual suspected modes. Therefore, it does not, in itself, locate the modes. However, by looking at the density corresponding to  $h_{crit,k}$ , one can easily identify the  $k$  modes established by the procedure.

But the location-independence of Silverman's test is not simply a problem in locating modes. It also results in lower power for the test under some circumstances than might otherwise be achieved. The test only counts the number of modes, ignoring factors such as location and size. Thus, small modes in the tail, perhaps the result of only a few points, are weighted as much as large, clear modes consisting of hundreds of points. In Figure 2.1, we show an example of a density which is vulnerable to such effects. A kernel density estimate of a sample from this density of size 100 can be seen in Figure 2.2. Note that the two narrow modes still appear to be a single mode, while the wide mode appears to be two. Also note that such behavior is not limited to such extreme densities. A single outlier sufficiently far out in a tail can produce such an effect on any density with two or more modes. This is among the problems addressed by the procedure of Chapter 4.

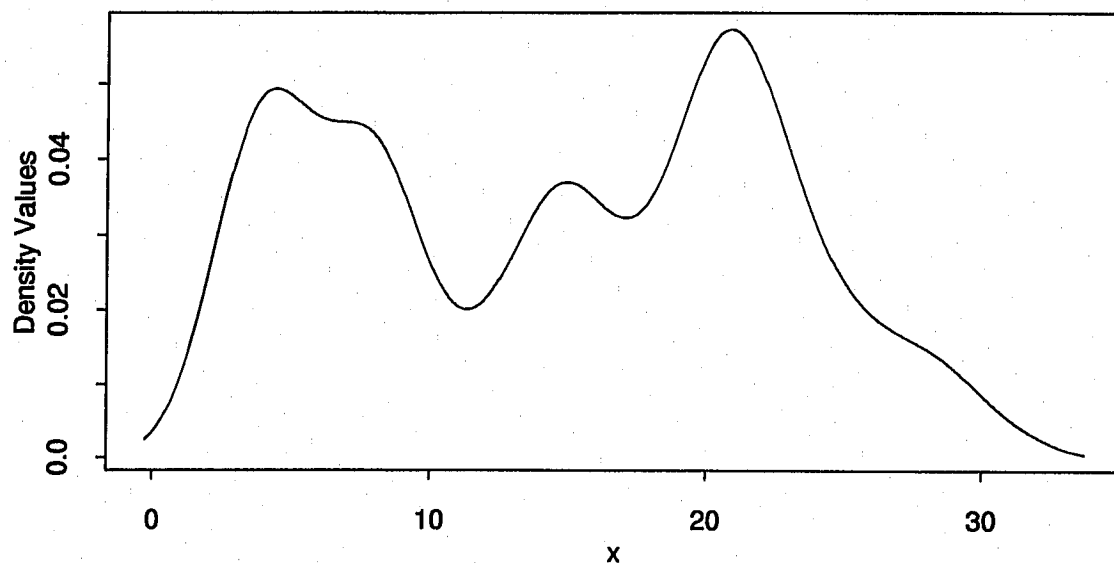
## 2.4 The Dip Test

Hartigan and Hartigan (1985) propose what they refer to as the *dip test of unimodality*. Although the dip test is formulated solely as a test of unimodality versus multimodality, it could (at least in principle) be extended to test for any number of modes. This test does not, however, have an easy generalization to testing for bumps, nor does it locate the modes. The test statistic is based on the empirical distribution function  $\hat{F}$ , rather than on any estimate of the density.

Let  $\rho(F, G) = \sup_x |F(x) - G(x)|$  for bounded functions  $F$  and  $G$ , and define  $\mathcal{U}$  as the class of unimodal distribution functions. Then  $D(F) = \inf_{G \in \mathcal{U}} \rho(F, G)$  is the *dip* of  $F$ . In other words, the dip of  $F$  is the largest difference from the "closest" unimodal distribution. Clearly  $D(F) = 0$  for  $F \in \mathcal{U}$ , and  $D(F) > 0$  for



**Figure 2.1** An example of a troublesome density.



**Figure 2.2** A kernel estimate of the above density, based on a sample of 100 points;  $h = h_{crit,3} = 1.497$ . Note the location of the third mode.

$F \notin \mathcal{U}$ . By The Glivenko-Cantelli Theorem, we have  $\rho(\hat{F}, F) \rightarrow 0$  almost surely, so  $D(\hat{F}) \rightarrow D(F)$  a.s., and a test based on the dip statistic will asymptotically distinguish any multimodal from any unimodal distribution.

Hartigan and Hartigan provide an algorithm for computing  $D(\hat{F}) = d$  based on finding a nondecreasing function  $G$  subject to several conditions. The requirements include that for some  $x_L \leq x_U$ ,  $G$  is the greatest concave minorant of  $\hat{F} + d$  and the least convex majorant in  $\hat{F} - d$  over  $(-\infty, x_L)$  and  $(x_U, \infty)$  respectively, that  $G$  has constant maximum slope in  $(x_L, x_U)$ , and that

$$d = \sup_{x \notin (x_L, x_U)} |\hat{F}(x) - G(x)| \geq \sup_{x \in (x_L, x_U)} |\hat{F}(x) - G(x)|.$$

They suggest comparing  $D(\hat{F})$  to critical values computed from simulations on the uniform distribution (which has an asymptotically higher dip statistic than any unimodal distribution with exponentially decreasing tails and, Hartigan and Hartigan speculate, all other unimodal distributions as well). They provide a small table of critical values for such a purpose.

Another possibility they mention is to follow Silverman (1981). We can draw bootstrap samples from the “closest” unimodal distribution, in this case  $G$  as defined above (which will be a mixture of uniforms), and use these simulations to calculate  $p$ -values. This will result in a more powerful test (though far more computer-intensive), as it essentially conditions on those observations not in the tails. Unless, for some reason, computer time is at a premium, this latter idea seems far preferable to the former.

The concept (though not necessarily the algorithm) is easily extended to tests of bimodality versus trimodality and higher comparisons. The class of unimodal distributions in the definition of the dip is simply replaced by the class of distributions appropriate to the null hypothesis. Finding and testing the statistic is unlikely to be easy, however, and the appropriate choice of null distribution is not clear.

## 2.5 The Excess Difference Test

Müller and Sawitzki (1991) propose a test based on mass found in connected sets (contours) above a given level. A major element in their proposal lies in viewing a mode, not as a local maximum in a density, but as a region where “an excess of probability mass is concentrated.” Their claim is that this is a more relevant definition statistically, and that one would prefer to view a mode as a location associated with a region of high probability.

While it is certainly true that modes associated with relatively low-probability regions will be difficult to detect at moderate sample sizes, this fact does not seem to us like a valid reason to exclude such small modes from our desired search objectives. Such low-probability modes may, in fact, be important features in the context of the data, despite being associated with relatively few points out of any given sample.

The primary element under study in Müller and Sawitzki’s method is called the *excess mass functional* at level  $\lambda$ , defined as

$$E(\lambda) = \int [f(x) - \lambda]_+ dx, \quad (2.1)$$

where the “+” subscript has its usual meaning of the “positive part;”  $[y]_+$  is equal to  $y$  if  $y$  is nonnegative and to 0 if  $y$  is negative. Thus the functional measures the amount of probability mass which lies above the contour level  $\lambda$  in density  $f$ . The authors define

$$E_m(\lambda) = \sup \sum_{j=1}^m \int_{C_j(\lambda)} (f(x) - \lambda) dx = \sup \sum_{j=1}^m H_\lambda(C_j),$$

where the supremum is taken over all families  $\{C_j : j = 1, \dots, m\}$  of pairwise disjoint connected sets. Also,  $H_\lambda = F - \lambda \cdot \text{Leb}$  where  $F$  is the distribution function associated with  $f$  and  $\text{Leb}$  stands for Lebesgue measure. Note that if  $f$  has no more than  $m$  modes,  $E_m(\lambda) = E(\lambda)$ .

To arrive at an excess mass estimate, Müller and Sawitzki simply replace the distribution function  $F$  with the empirical distribution function  $F_n$ , finally arriving

at

$$E_{n,M}(\lambda) = \sup \sum_{j=1}^M H_{n,\lambda}(C_j),$$

where  $M$  is the assumed maximum number of modes. The test statistic they recommend is

$$\Delta_{n,M} = \max_{\lambda} [E_{n,M}(\lambda) - E_{n,1}(\lambda)],$$

though it seems that for general  $M$ , it would be more appropriate to subtract  $E_{n,(M-1)}(\lambda)$  than  $E_{n,1}(\lambda)$ .

Müller and Sawitzki follow Hartigan and Hartigan (1985) in suggesting the use of critical values calculated from the uniform distribution. They provide a table of percentage points for several sample sizes. Again, the use of the uniform as a null hypothesis may cost the test some power.

## 2.6 Nearest-Neighbor Procedures

Wong (1985) presents a variant on Silverman's test using a  $k$ th nearest-neighbor ( $k$ NN) approach to density estimation, as opposed to the kernel estimator approach suggested by Silverman (1981). The  $k$ NN estimate of the density is calculated as

$$f_n(x, k) = k/(nV_k(x)),$$

where  $V_k(x)$  is the volume of the minimal sphere centered on  $x$  containing at least  $k$  sample observations. Here  $k$  plays much the same role of a smoothing parameter as the bandwidth  $h$  does in a kernel estimate, and, in a similar fashion, as  $k$  gets smaller,  $f_n(x, k)$  gets rougher and shows more modes.

To test the hypothesis that  $f$  has at most  $M$  modes, Wong uses as his test statistic  $k_{crit}(M)$ , the smallest value of  $k$  such that  $f_n(x, k)$  has at most  $M$  modes (in a manner exactly analogous to the use of  $h_{crit}$  in Silverman's test). He estimates significance in much the same way as Silverman as well, drawing samples from  $f_n(x, k_{crit}(M))$  (drawing random elements of  $\{X_1, \dots, X_n\}$  with replacement and adjusting each  $x_i$



by a uniform random variable with limits  $\pm d_{k_{crit}(M)}(x_i)$ , where  $d_k(x_i)$  is the distance to the  $k$ th nearest neighbor from  $x_i$ ). He then tests the number of modes in the density estimated from this new sample when  $k_{crit}(M)$  is used. A  $p$ -value is calculated in the same manner as in Silverman's test.

Wong's examples show his method working acceptably for two well-separated modes of equal size. It does not, however, do very well for a trimodal example in which the central mode had twice the mass of the two outer ones. In general, since nearest-neighbor estimates tend to produce very rough and variable estimates in regions of high sample concentration, this method seems especially vulnerable to spurious modes appearing on large modes before smaller true modes appear at all.

Wong and Schaack (1985) also propose another statistic for assessing multimodality. The procedure is based on the same nearest neighbor density estimates as Wong's test, but uses as a statistic  $S(m)$ , which they define to be the number of values of  $k$  for which the  $k$ th nearest neighbor estimate is  $m$ -modal. The authors make the claim that  $S(m)$  can be viewed as the number of observations in the smallest modal cluster, and that if a data set is strongly  $m$ -modal,  $S(m)/N$  will be significantly higher than 0.

In Monte Carlo investigations, Wong and Schaack demonstrate that well-balanced, widely separated bimodal and trimodal distributions do have notably larger values of  $S(2)$  and  $S(3)$ , respectively, than other, non bimodal or trimodal, densities do on average. Two problems, however, keep this from being an overly useful tool. The first is that it is merely a diagnostic procedure, with no method suggested for estimating significance. The other is that the procedure did not perform at all well on unbalanced densities, where the smaller number of observations in the secondary mode led to greatly reduced values of  $S(2)$  (for unbalanced bimodal densities). These factors, in conjunction with the problems inherent in the nearest neighbor estimate discussed above, severely limit the statistic's utility.

## 2.7 Density Confidence Sets

Roeder (1990) presents the idea of a confidence set of normal mixture density estimates based on a range of smoothing parameters. For each member of a large grid of values of  $h$ , she finds the fit  $\hat{F}$  of mixtures of normal distributions each with standard deviation  $h$  such that  $\hat{F}$  achieves the largest value of a function  $\phi(\hat{F})$  of the sample spacings, the intervals between consecutive ordered sample values. The function  $\phi$  is chosen to be asymptotically normal for sample spacings from the uniform distribution.

Roeder shows that this largest value of  $\phi(\hat{F})$  is a decreasing function of  $h$ , and chooses for her  $(1 - \alpha)100\%$  confidence set those estimates for which the maximum value of  $\phi(\hat{F})$  lies between  $-Z_{\alpha/2}$  and  $+Z_{\alpha/2}$ , where  $Z_{\alpha}$  is the  $(1 - \alpha)$  percentile of the standard normal distribution.

Her example of galaxy velocities data gives a range of three to seven modes (with five at the “point estimate” density chosen by least squares cross-validation). Although Roeder’s emphasis is on the density estimates, rather than on the modes, this clearly has some potential in the search for multimodality. It is not obvious, however, what her upper bound signifies in light of Donoho’s notes about one versus two-sided inference (as will be discussed in the next section).

## 2.8 Some Theoretical Results with Implications for Multimodality Tests

For a conservative initial view of features such as multimodality, Terrell (1990) suggests using the “maximal smoothing principle” of Terrell and Scott (1985). Given sample size  $n$  and standard deviation  $s$ , using (for a kernel  $K$  with variance 1) a bandwidth of

$$h = \frac{3s}{(35)^{1/5}} \left( \int K^2 \right)^{1/5} n^{-1/5}, \quad (2.2)$$

will result in a conservative, oversmoothed density estimate. As Terrell (1990) points out, in such an environment, features such as multimodality which remain after such

maximal smoothing have great evidence in their favor, although such evidence may be hard to quantify and the single-bandwidth problem persists. While not a test of multimodality as such, viewing maximally smoothed plots can often give insights as to whether or not there is reason to investigate the possibility of multimodality further.

Donoho (1988) shows that one can produce confidence intervals on functionals, including the number of modes, of an unknown density, but that these confidence regions must be one-sided. We could say, for example, that we know with 95% confidence that this density has *at least* three modes. On the other hand, as there could be any number of sufficiently small modes (at any given sample size) that we would have no way of detecting, we could not give any sort of upper confidence limit or two-sided interval. The essential point is that near any given density, there are densities which, at any given sample size, would be empirically indistinguishable, yet would have arbitrarily large values of the functional in question (in our case, the number of modes). While not a multimodality test itself, this result clearly has important implications for our purposes and should be kept in mind when examining the results of such tests.

## 2.9 A Multimodal Mode Estimation Procedure

Finally, Comparini and Gori (1986) provide a method of estimating the numerical values of both modes and antimodes for univariate multimodal densities. Rather than basing their technique on a kernel or other density estimate, followed by reading off the local maxima and minima, they work on the idea that a mode will be found near the greatest clustering in a region, while an antimode will be found near the point of least clustering in a region.

To find the modes they offer the following algorithm. Given an ordered sample  $\{X_1, \dots, X_n\}$  and  $r_n < n/2$  such that  $\lim_{n \rightarrow \infty} (r_n/n) = 0$  and  $\sum_n n\lambda^{r_n} < \infty$  [for  $\lambda \in (0, 1)$ ; they suggest  $0.1n^{4/5}$ ], calculate  $V_i = X_{i+r_n} - X_{i-r_n}$  for  $i = r_n + 1, \dots, n - r_n$ .

Choose  $\sigma < d/2$  where  $d$  is the assumed minimum Euclidean distance between any pair of modes or antimodes, and let  $t = r_n + 1$ . If

$$V_t = \min_{(j: X_{t-\sigma} \leq X_j \leq X_{t+\sigma})} (V_j),$$

estimate a mode at one of  $X_t$ ,  $\text{mean}(X_j \in [X_{t-r_n}, X_{t+r_n}])$ , or  $.5(X_{t-r_n} + X_{t+r_n})$ ; otherwise increase  $t$  by 1. Antimodes are found in a similar fashion based on the maximum  $V_j$  with  $X_j$  between two adjacent modes. The three estimators have the same asymptotic properties and behave similarly for estimating modes, but the midpoint is a superior estimator for antimodes at moderate sample sizes than the mean or the median.

The primary drawback to this method lies in the necessity for having at least an idea of  $d$  in order to choose  $\sigma$ . If  $\sigma$  is chosen too large, modes will be lost, while spurious modes will appear when  $\sigma$  is too small, in a manner analogous to the effect of bandwidth choice on kernel estimation. If  $r_n$  and  $\sigma$  are chosen appropriately, however, the method provides a strongly consistent estimate of each mode and antimode of the density. Clearly this method works most successfully in conjunction with some other method of estimating the number and approximate locations of modes and antimodes (such as Silverman's test, or the test proposed in Chapter 4).

## 2.10 Final Thoughts

Of the tests and procedures discussed above, Silverman's critical-smoothing test is by far the most widely studied. It also has the advantages of being relatively easy to program; of providing  $p$ -values from a data-based choice of null hypothesis, rather than from a default of the uniform distribution; and of using the relatively stable kernel density estimate, rather than the far more unstable nearest-neighbor estimate of Wong's procedure. The only real drawback of Silverman's test is the lack of adaptivity, which results in very low power for finding modes of varying sizes and degrees of separation. This drawback which is shared with the other proposals above (with

the exception of Wong's nearest-neighbor procedures), but which the test proposed in Chapter 4 attempts to rectify.

## Chapter 3

### A Pair of Useful Tools

Before describing the test which is the focus of this work, we find it necessary to discuss in some detail two tools which, while central to the procedure, are also quite useful in broader settings. The first is a point-matching algorithm which will be used in several places. The other is a graphical tool which is useful for exploratory data analysis and for investigating questions of multimodality.

#### 3.1 A Univariate Point-Matching Algorithm

The matching procedure will be described for ordered sets of points  $\mathbf{a} = (a_1 < a_2 < \dots < a_{k_a})$  and  $\mathbf{b} = (b_1 < b_2 < \dots < b_{k_b})$ . Let  $\alpha_1 = (\alpha_{11}, \alpha_{12}, \dots, \alpha_{1k_a})$ , where

$$\alpha_{1i} = \arg \min_{\ell} \{|a_i - b_{\ell}| \leq |a_i - b_j| \text{ for all } j = 1, \dots, k_b\}.$$

Thus  $\alpha_{1i}$  is the index of the member of  $\mathbf{b}$  which lies closest to  $a_i$ . Let  $\alpha_2 = (\alpha_{21}, \alpha_{22}, \dots, \alpha_{2k_a})$ , where if  $a_i < b_1$  or  $a_i > b_{k_b}$ ,  $\alpha_{2i}$  equals 0, otherwise

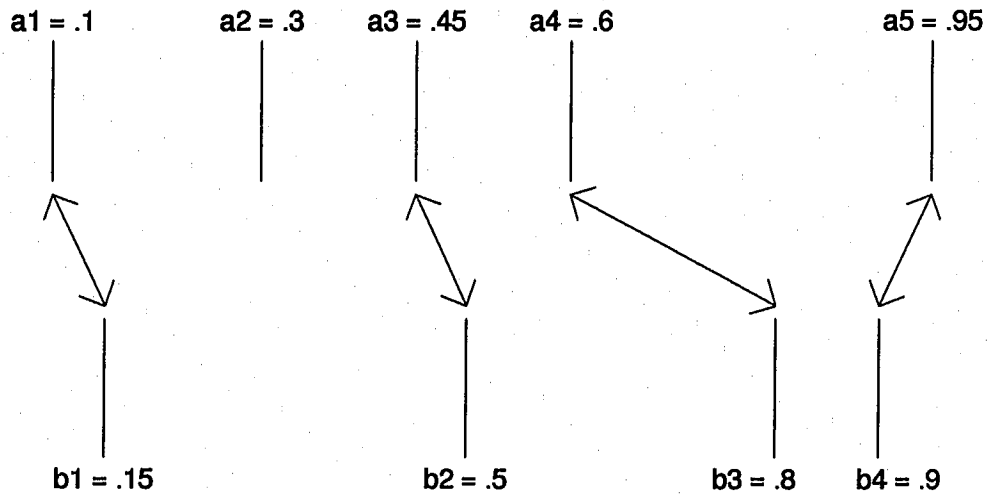
$$\alpha_{2i} = \begin{cases} \arg \min_{\ell} \{b_{\ell} > a_i\}, & \text{if } a_i > b_{\alpha_{1i}} \\ \arg \max_{\ell} \{b_{\ell} < a_i\}, & \text{if } a_i < b_{\alpha_{1i}} \\ \arg \min_{\ell} \{|a_i - b_{\ell}| \leq |a_i - b_j| \text{ for all } j \neq \alpha_{1i}\}, & \text{if } a_i = b_{\alpha_{1i}}. \end{cases}$$

Here  $\alpha_{2i}$  is assigned the index of the closest member of  $\mathbf{b}$  to  $a_i$  which lies on the other side of  $a_i$  from  $b_{\alpha_{1i}}$ . The vectors  $\beta_1$  and  $\beta_2$  are found by reversing  $\mathbf{a}$  and  $\mathbf{b}$  in the above formulas.

With the above four lists of indices,  $\alpha_1$ ,  $\alpha_2$ ,  $\beta_1$ , and  $\beta_2$ , we begin the actual process of matching. For each value  $i = 1, \dots, k_a$ , we see if  $\beta_{1\alpha_{1i}} = i$ , which would indicate that  $a_i$  and  $b_{\alpha_{1i}}$  are each the closest element in the opposite list to each other.

If this is true,  $a_i$  and  $b_{\alpha_i}$  are considered matched to each other and removed from further consideration. If  $\mathbf{a}$  and  $\mathbf{b}$  are closely related, most of the points should be matched in this step. After all such pairs are removed, the remainder are investigated further. For any  $i$  such that  $a_i$  was not matched in the first cycle, if  $b_{\alpha_i}$  also remains unmatched and  $\beta_{2\alpha_i} = i$ ,  $a_i$  and  $b_{\alpha_i}$  are matched and removed. The process is repeated on the remainder two more times, the first time matching still unmatched pairs if  $\beta_{1\alpha_{2i}} = i$  and the second matching any remaining pairs if  $\beta_{2\alpha_{2i}} = i$ . In each of these two cases,  $a_i$  is matched with  $b_{2\alpha_{2i}}$ .

For the example in Figure 3.1, consider  $\mathbf{a} = (0.1, 0.3, 0.45, 0.6, 0.95)$  and  $\mathbf{b} = (0.15, 0.5, 0.8, 0.9)$ . Then  $\alpha_1 = (1, 1, 2, 2, 4)$ ,  $\alpha_2 = (0, 2, 1, 3, 0)$ ,  $\beta_1 = (1, 3, 5, 5)$ , and  $\beta_2 = (2,$



**Figure 3.1** Example of the point-matching algorithm used in the mode tree and mode-existence test. The arrows indicate points among the two sets which are matched.

2, 4, 4). The first round matches  $a_1$  with  $b_1$ ,  $a_3$  with  $b_2$ , and  $a_5$  with  $b_4$ . (For example,  $\alpha_{13} = 2$  and  $\beta_{12} = 3$ .) Removing these pairs leaves  $a_2$ ,  $a_4$ , and  $b_3$ . No pairs are matched in the second and third rounds, but since  $\alpha_{24} = 3$  and  $\beta_{23} = 4$ ,  $a_4$  and  $b_3$  are matched in the final cycle. The point  $a_2$  remains unmatched.

By using a matching algorithm such as this, we can quickly determine which members of each set correspond to each of the other, as well as which elements should remain unmatched.

## 3.2 The Mode Tree

The difficulties inherent in the choice of a specific  $h$  in kernel density estimation suggest that another method of viewing the data might be appropriate, one which summarizes information from density estimates calculated for a large variety of values of  $h$ . One possibility is the overlaid density plot shown in Figure 1.2. While such a figure certainly has value, it is difficult to extract information on any single estimate. Roeder (1990) used the option of Figure 1.3, which treats  $h$  as a second independent variable in a perspective plot of the function  $\psi(x, h) = \hat{f}_h(x)$ . While we find this plot more informative than Figure 1.2, its use also invokes all of the disadvantages of perspective plots, most notably hidden features and difficulties in interpreting the complex surface. The figure seems to overemphasize estimates with small bandwidths. The eye is drawn to the high, sharp peaks at the back of the plot, to the neglect of the lower and broader, but potentially more realistic, features found at wider bandwidths.

### 3.2.1 The Basic Mode Tree

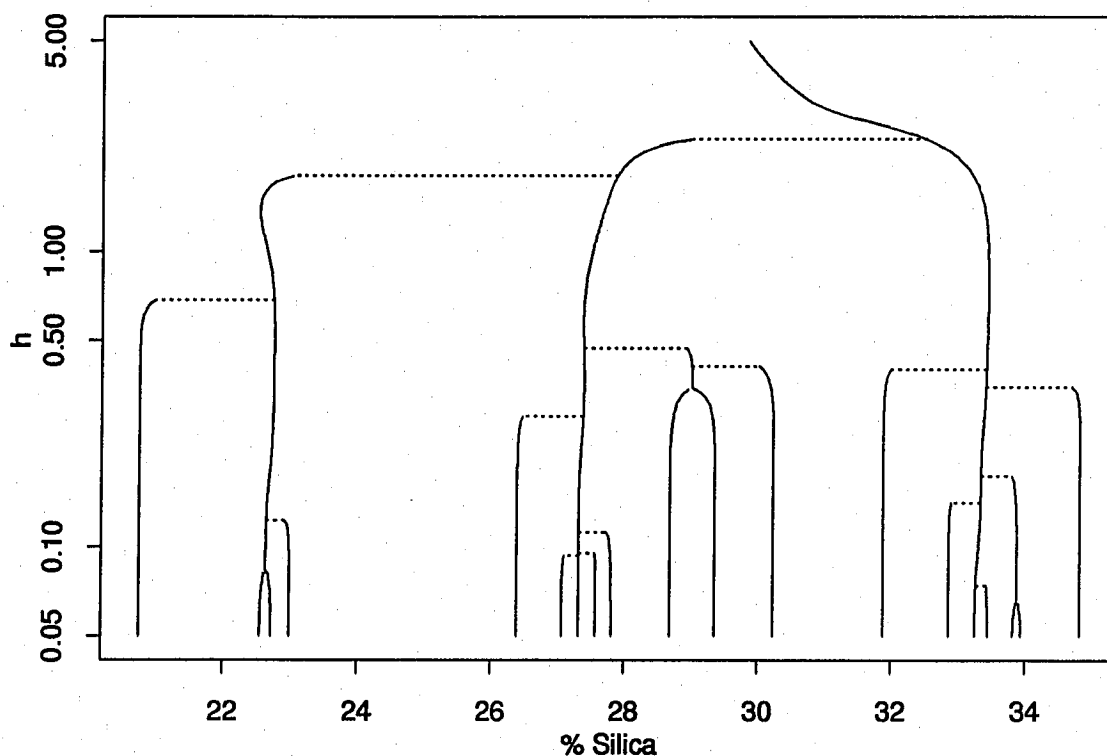
If the key information about the structure in an estimate at a given bandwidth can be condensed to a 1-dimensional plot, then the second dimension can be allocated to  $h$ , without the problems of perspective plots. As the number, location, and size of modes (i.e., the “wiggleness” of the density) are among the most easily visible changes in a density estimate due to varying  $h$ , these features provide a good 1-dimensional



surrogate for the density estimate as a whole. In addition, when there is strong reason to believe that multiple modes are present in the true density, they are often among the most important and interesting features. Thus, in lieu of presenting the entire collection of density estimates, information about their modes is a valuable contribution.

The basic *mode tree* is very simple to define. The mode locations are plotted against the bandwidth at which the density estimate with those modes is calculated. In Figure 3.2, the solid vertical lines represent the modes corresponding to those in the density estimates in Figure 1.2 for the chondrite data. A larger number of values of  $h$  (here, 200) is used in producing the mode tree than in the superimposed plot. Notice the choice of the logarithmic scale for the vertical axis. The values of  $h$  should be chosen to be equally spaced on a logarithmic scale, as large changes at high values of  $h$  have less of an effect on the density estimate than smaller changes at lower values of  $h$ . If the mode locations are plotted for all value of  $h$ , then a set of lines, or *mode traces*, will result. The matching algorithm from Section 3.1 can be used to determine the mode traces derived from a *finite* set of mode points so that they can be connected between levels.

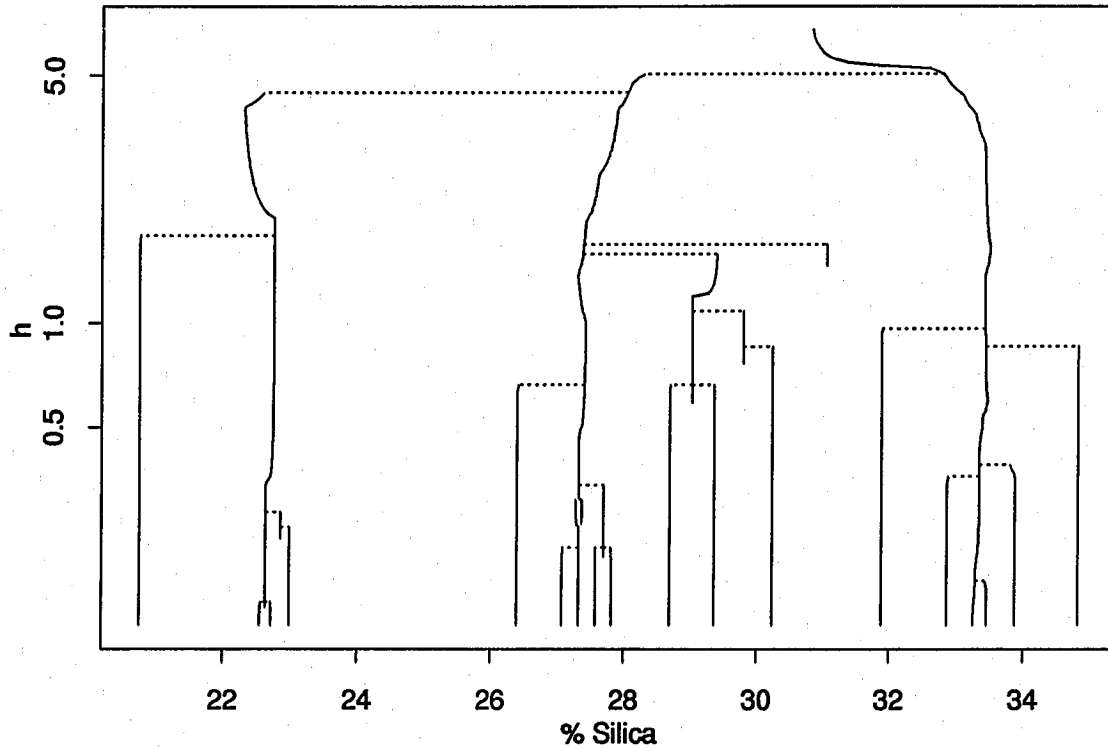
It is useful to think of modes as “splitting” as  $h$  decreases. Between any pair of adjacent modes is a valley containing exactly one antimode. As  $h$  decreases, a critical or saddle point appears between two adjacent modes and thereafter a new mode and antimode appear. This event will be apparent from the increased number of mode locations. The new mode is the one at the smaller value of  $h$  not matched to any mode from the previous, slightly greater, value of  $h$ . Depending on the relative location of the new mode/antimode pair, either the left or the right mode may be thought of as splitting. If the new mode appears to the right of the new antimode, it is the left mode which has split; otherwise, it is the right. Thus the mode tree may be made more informative by adding the horizontal connections seen as dashed lines in Figure 3.2. These connections show the relationships between new modes and the



**Figure 3.2** Normal kernel mode tree for the chondrite data. Solid lines represent mode locations at each bandwidth. The horizontal dashed lines indicate the “splitting” of a mode.

old modes from which they split, and they give this plot the structure which justifies the “tree” label.

The mode tree in Figure 3.2 was calculated using a Normal kernel. This choice was quite deliberate. As mentioned in Section 2.3, Silverman (1981) showed that for a Normal kernel, the number of zeroes in all derivatives of  $\hat{f}_h$  is monotone decreasing in  $h$ . These zeroes include all modes, antimodes, and inflection points. Silverman’s proof hinges on the facts that the Normal density is totally positive and that the convolution of a Normal kernel density estimate with bandwidth  $h_1$  and a Normal density with standard deviation  $h_2$  is also a Normal kernel density estimate with



**Figure 3.3** Biweight kernel mode tree for the chondrite data. Note the modes which appear briefly and then vanish, as well as the modes which continue after the point at which they split.

bandwidth  $h_3 = \sqrt{h_1^2 + h_2^2}$ . The desirable behavior in  $\hat{f}_h$  can be clearly seen in Figure 3.2. All modes found at a given level of  $h$  remain as  $h$  decreases.

Surprisingly, this result need not be true for other kernels. Figure 3.3 shows the tree generated by the same chondrite data set but using the biweight kernel. Sample modes appear and disappear in an irregular fashion. The appearance of such distracting features in practical situations is not well recognized. These spurious peaks are generally quite small and are more accidents of the estimation method than true features of the data. Similar results have been observed in estimates from other members of the Beta family (of the form  $c_m(1 - x^2)^m$ , where  $m$  is an integer and  $c_m$  is the constant which makes the kernel integrate to 1 in the range  $[-1, 1]$ ), even though

the family approaches a Normal kernel as  $m \rightarrow \infty$ . Therefore, we conclude that the Normal kernel should be used for applications in which modes take a key role.

### 3.2.2 Enhancements to the Mode Tree

As we have shown, the basic mode tree, in and of itself, can be quite informative and useful. Yet the possibility exists to significantly increase the information presented, thereby further improving our understanding of the data.

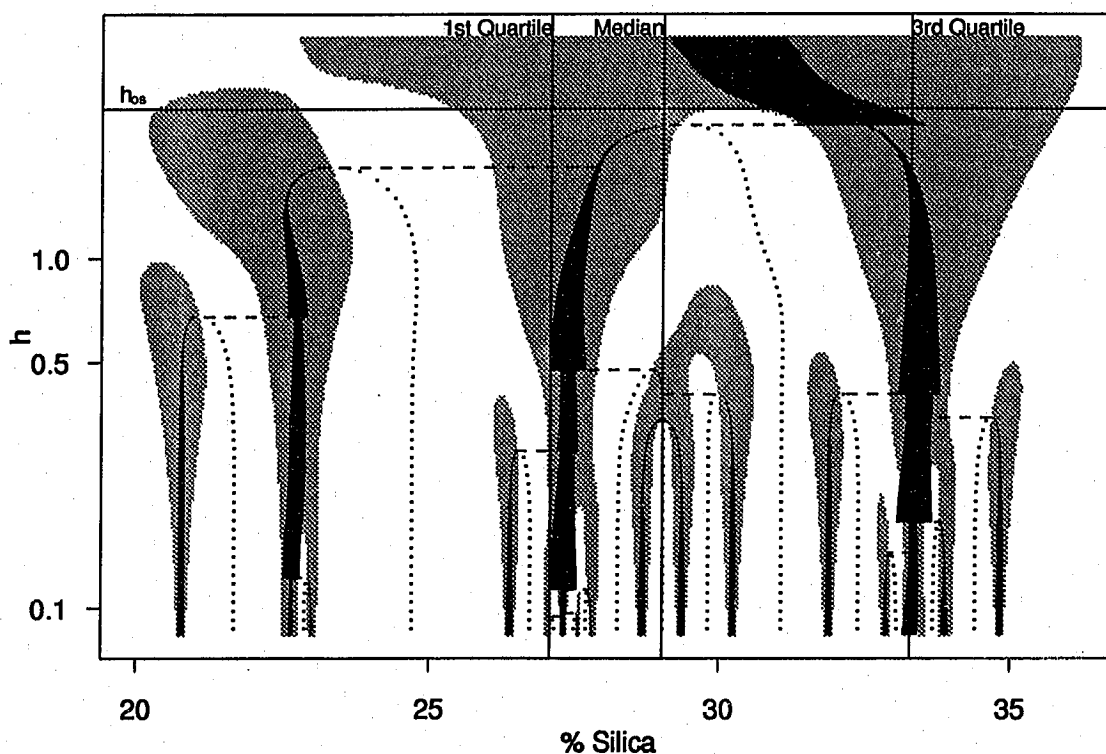
Our first enhancement is to add a dimension of information through the widths of the mode traces in the mode tree. Typically, that information relates to some feature of those modes. In Figure 3.4 we plot an enhanced mode tree of the chondrite data. The lines of Figure 3.2 have been replaced by the black regions centered on each mode location and whose horizontal width at each level of  $h$  is proportional to the quantity

$$M_i = \int_{u_{i-1}(h)}^{u_{i+1}(h)} \left[ \hat{f}(x, h) - \max \left( \hat{f}(u_{i-1}(h), h), \hat{f}(u_{i+1}(h), h) \right) \right]_+ dx. \quad (3.1)$$

The values  $u_{i-1}(h)$  and  $u_{i+1}(h)$  are respectively the left and right antimodes surrounding mode  $u_i(h)$  in  $\hat{f}(x, h)$  (and which may be  $-\infty$  and  $+\infty$ , respectively). The “+” denotes the positive part of the argument, as in Equation 2.1. In Figure 3.5, the values of  $M_1$ ,  $M_3$ , and  $M_5$  when  $h$  is 1 are equal to the indicated shaded areas. The value  $M_i$  is representative of the “size” of the mode, and in fact will be used as our statistic to test the reality of the mode in question in Chapter 4 (when the bandwidth is chosen appropriately).

A second enhancement is an indication of the locations of antimodes. The enhanced mode tree for the chondrite data shown Figure 3.4 uses dotted lines for the antimode traces. Observe that the additional information does not appreciably clutter the diagram as the mode and antimode traces are approximately parallel and do not cross.

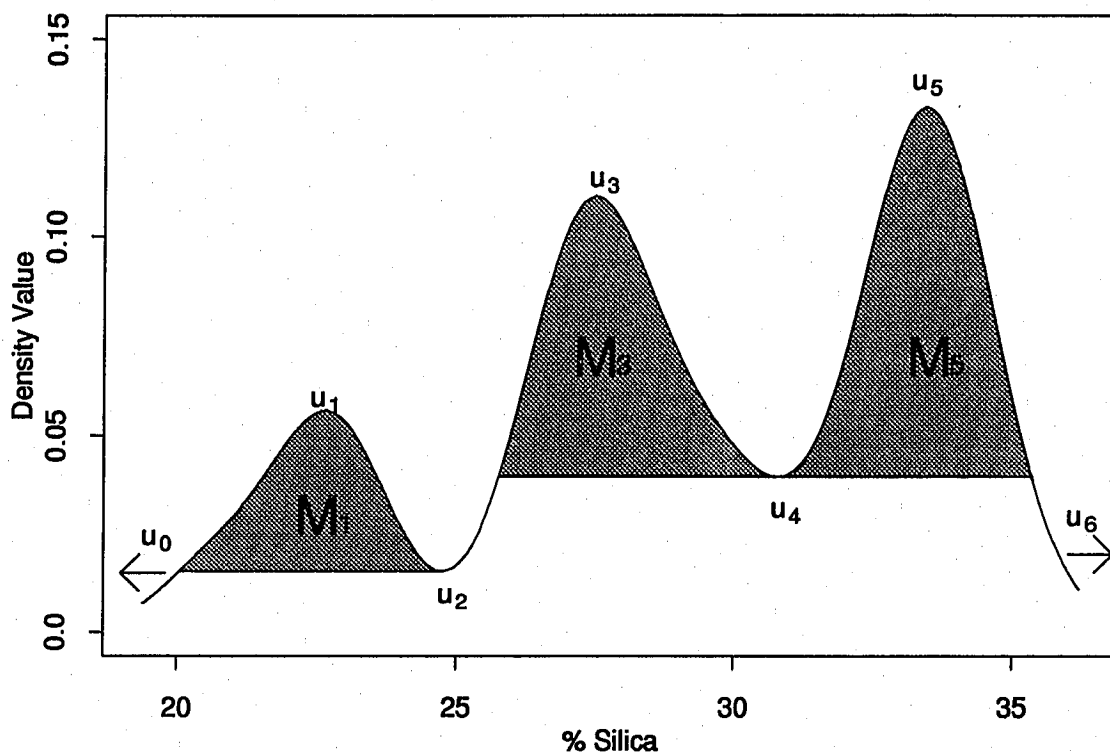
In addition to modes, one can easily add information about bumps to the mode tree. While inflection points could simply be plotted with one or two additional line



**Figure 3.4** Enhanced Normal mode tree for the chondrite data. The widths of the black regions are proportional to  $M_i$  for the modes at each level. The dots represent the locations of antinodes, and the gray regions are the bumps.

types, shading the entire area of each bump, as in the gray regions of Figure 3.4, is more visually striking and suggestive. We prefer to use color when available, rather than gray levels. Clearly, at any given  $h$ , each mode has a surrounding bump, but the reverse is not necessarily true. It is worth noting that a new mode will always start at the boundary of its shaded bump region. In hindsight, this is to be expected, but it is still interesting to view graphically.

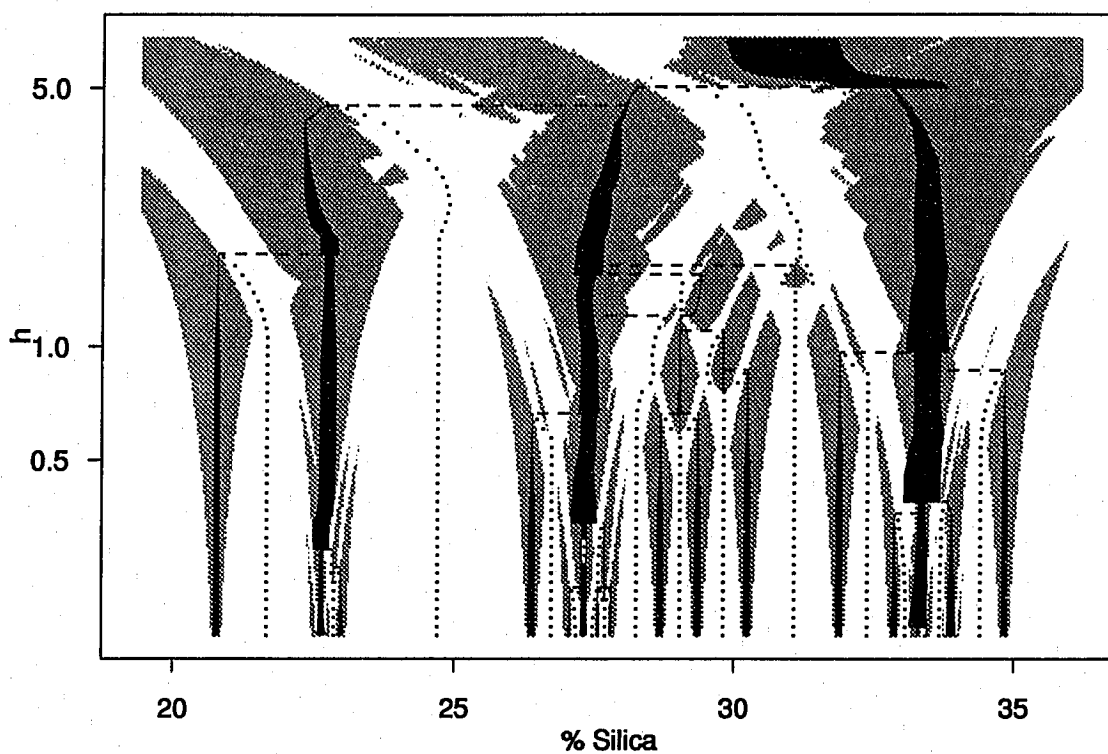
Finally, important values of either  $x$  or of  $h$  can be highlighted. Values of the sample mean, median, or quartiles might be shown. Appropriate special values of  $h$  include cross-validation bandwidths or Terrell and Scott's (1985) oversmoothed Normal bandwidth given by Equation 2.2, which for the Normal kernel equals  $3s(70\sqrt{\pi n})^{-1/5}$ .



**Figure 3.5**  $M_1$ ,  $M_3$ , and  $M_5$  are equal to the shaded areas for the chondrite data when  $h = 1.0$ .

The oversmoothed bandwidth  $h_{os}$  and the sample median and quartiles for the chondrite data are indicated in Figure 3.4.

It should be noted that the undesirable behavior observed for non-Normal kernels illustrated in Figure 3.3 is not limited to the mode. In Figure 3.6, a plot is shown of modes, bumps, and antimodes, from biweight kernel estimates of the chondrite data. Clearly, the same undesirable behavior observed in the modes is occurring in the bumps and antimodes as well. Note, for example, the behavior near  $x = 31$  and  $h = 1.5$ . An antinode briefly becomes a mode, with surrounding bump, before quickly becoming an antinode again. Such occurrences, along with the highly erratic behavior of modes, antimodes, and inflection points, all point to how strongly inappropriate the biweight kernel is for multimodality investigations and bump hunting, especially



**Figure 3.6** Enhanced biweight mode tree for the chondrite data. In comparison to Figure 3.4, note the erratic behavior of modes, antinodes, and bumps.

when using a concept such as Silverman's (1981) critical bandwidths. However, we note that the undesirable analytical behavior is limited to relatively flat regions near a mode or antinode and is difficult to discern in a plot of the density estimate.

## Chapter 4

### The Mode-Existence Test

In studying the tests and procedures currently available for evaluating multimodality, there are a number of features available in individual tests which would be advantageous to pass on to the next generation of tests. For example, the locality of Good and Gaskins' procedure is enviable, as is their ability to investigate bumps as well as modes. Silverman's test is easily understood and implemented, produces sensible  $p$ -values, and has a fair amount of asymptotic background supporting it. Unfortunately, neither test has the desirable properties just listed for the other. In addition, Silverman's test, while providing  $p$ -values, has somewhat low power, especially when dealing with several modes of varying peak widths. While this is to be expected in the nonparametric domain, especially in light of the results of Terrell and Scott (1985) and Donoho (1988), a more powerful test is desirable.

However, by combining certain aspects of the two procedures, along with some of the excess mass concepts of Müller and Sawitzki, a test related to (though not directly of) multimodality can be produced with the best features of both. By using the idea of "surgery" from Good and Gaskins, and the idea of sampling from the null hypothesis to obtain  $p$ -values from Silverman, one can obtain a test on the existence of individual modes, which will be adaptive and, we hope, more powerful than others under non-optimal conditions. 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 locality benefits of Good and Gaskins while maintaining the other benefits of Silverman and increasing power considerably



by taking advantage of mode location and size information which Silverman's test ignores.

## 4.1 Testing Individual Modes

We begin with a sample  $\{X_1, \dots, X_n\}$  of size  $n$  from a population with density  $f(x)$ . We compute a 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  $\hat{f}(x)$ . For the purposes of the test statistic, the extreme  $x$ -values  $-\infty$  and  $\infty$  (or, more practically, the lowest and highest points 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  $h_{test,i}$  at which to test. We perform the test at the lowest 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  $h_{test,i}$  will be Silverman's critical bandwidth  $h_{crit,k}$  for some  $k$  (Section 2.3).

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 largest value for the test statistic (see Theorem 5.1), and thus should give the greatest power.

Another crucial reason for this choice of  $h$  can also be found in Chapter 5. By choosing one of Silverman's critical bandwidths, we can use the theory already developed involving these estimates. In particular, we can use the result of Mammen, Marron, and Fisher (1990), that if  $f$  has  $j$  modes and  $k$  is at least as great as  $j$ , then  $h_{crit,k}$  is of order  $n^{-1/5}$ . This will prove invaluable in our consistency investigations.

A final feature of this selection should also be mentioned. 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 misleadingly tested, even if it produces a mode at fairly large values of  $h$ . Unfortunately, if the data are binned, a mode consisting of a single bin will also never be tested, even if it contains many points. We will return to this problem, and suggest a possible solution in Section 4.5.

Having determined the bandwidth  $h_{test,i}$  of our test [and having kernel density estimate  $\hat{f}(\cdot)$ ], we can now calculate the test statistic. Recall the definition of Equation 3.1,

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

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 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.

## 4.2 Calculating $p$ -values

To calculate 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, like Silverman, 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 highest inflection points (on the sides toward  $u_i$ ) of its adjacent modes  $u_{i-2}$  and  $u_{i+2}$ . This means, for example, that if at  $h_{test,i}$  there is one inflection point  $v_p$  in  $\hat{f}$  between mode  $u_{i-2}$  and antinode  $u_{i-1}$  and three inflection points  $v_{q-2} < v_{q-1} < v_q$  between antinode  $u_{i+1}$  and mode  $u_{i+2}$ , then in general  $\tilde{f}_i(x) = \hat{f}(x)$  for  $x < v_p$  and  $x > v_q$ .

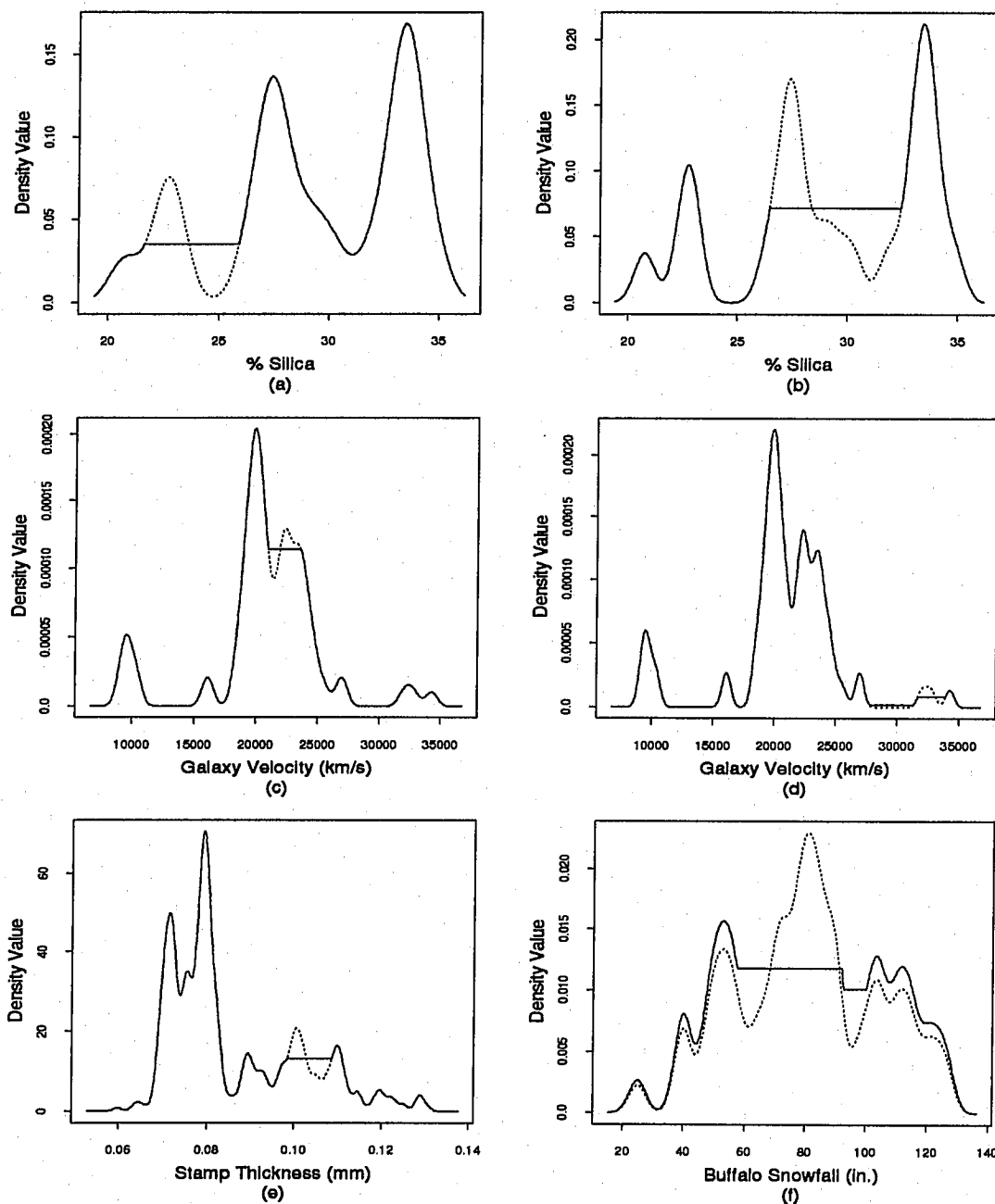
Within the region bounded by the indicated inflection points  $v_p$  and  $v_q$ , we then impose additional constraints. In the region  $v_p < x < u_i$ ,  $\tilde{f}_i(x)$  cannot be greater than  $\hat{f}(v_p)$  unless  $\hat{f}(x) > \hat{f}(v_p)$ . Likewise, the region  $u_i < x < v_q$  is bounded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

Working within this admittedly somewhat convoluted set of constraints, we find the admissible function 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 antinode 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 inflection point is filled to the level of that inflection point, 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* inflection point. 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 4.1.

This choice of  $\tilde{f}_i$  satisfies all of the desirable properties mentioned earlier. It will be a member of the null hypothesis. It will keep the probability mass as close to that of the original (data-produced) density estimate as permissible under the constraints (due to the  $L_1$  requirement). Finally, it will be 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 will thus be conservatively 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  $h_{test,i}$ . The modes of  $\hat{f}_i^j$  and  $\hat{f}$  are matched, using the same matching algorithm as described for use in the mode tree. 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  $u_{i+2}$  or, lacking one or both of these, the inflection points  $v_p$  and  $v_q$  of  $\hat{f}$ . 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  $h_{test,i}$ ). The value of the final test statistic can be denoted  $M_i^j$ .



**Figure 4.1** Examples of the density-choosing process. 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 (see Chapter 7): (a) chondrite data,  $h = .700$ ,  $i = 1$ , and (b)  $h = .487$ ,  $i = 3$ ; (c) galaxy data,  $h = 475$ ,  $i = 4$ , and (d)  $h = 363$ ,  $i = 7$ ; (e) stamp data,  $h = 0.00098$ ,  $i = 8$ ; and (f) Buffalo snowfall data,  $h = 2.83$ ,  $i = 4$ .

### 4.3 The Use of Sequential Monte Carlo $p$ -values

This procedure could be performed as a standard, fixed- $N$  Monte Carlo procedure, with estimated  $p$ -value

$$\frac{1}{N+1} \left( \sum_{j=1}^N I_{(M_i^j \geq M_i)} + 1 \right), \quad (4.1)$$

where  $I_A$  is the indicator function

$$I_A = \begin{cases} 1 & \text{if } A \\ 0 & \text{otherwise.} \end{cases}$$

The additions of one in (4.1) cause the results to round up; this ensures (among other things) no inappropriate  $p$ -values equal to zero.

It is more efficient, however, to follow Besag and Clifford (1991) in using a sequential method of Monte Carlo  $p$ -value estimation. In this technique, not only is  $N$  chosen, but a second (smaller) value  $L$  as well. Two counters,  $\tilde{N}$  and  $\tilde{L}$ , are initially set to zero. After each iteration  $\tilde{N}$  is incremented by 1, so that it remains a count of the number of iterations to date, and  $\tilde{L}$  is increased by 1 if  $M_i^j \geq M_i$ , so that it remains a count of simulated test statistics at least as great as the observed value. If  $\tilde{L} = L$ , then the  $p$ -value

$$\frac{L}{\tilde{N}}$$

is declared. If  $\tilde{L} < L$ , but  $\tilde{N} = N$ , then

$$\frac{\tilde{L} + 1}{N + 1}$$

is the estimated  $p$ -value.

This approach is valuable in that it invests the greatest time and effort into those tests where the  $p$ -value is low. When the  $p$ -value will be high, and thus less interesting, a relatively low value of  $L$  will ensure quick termination. The computing time thus saved allows the use of a higher value of  $N$  than would otherwise be practical. In essence, the sequential procedure trades lower accuracy for greater speed among the high  $p$ -values, and longer running times for greater accuracy among the lower  $p$ -values.

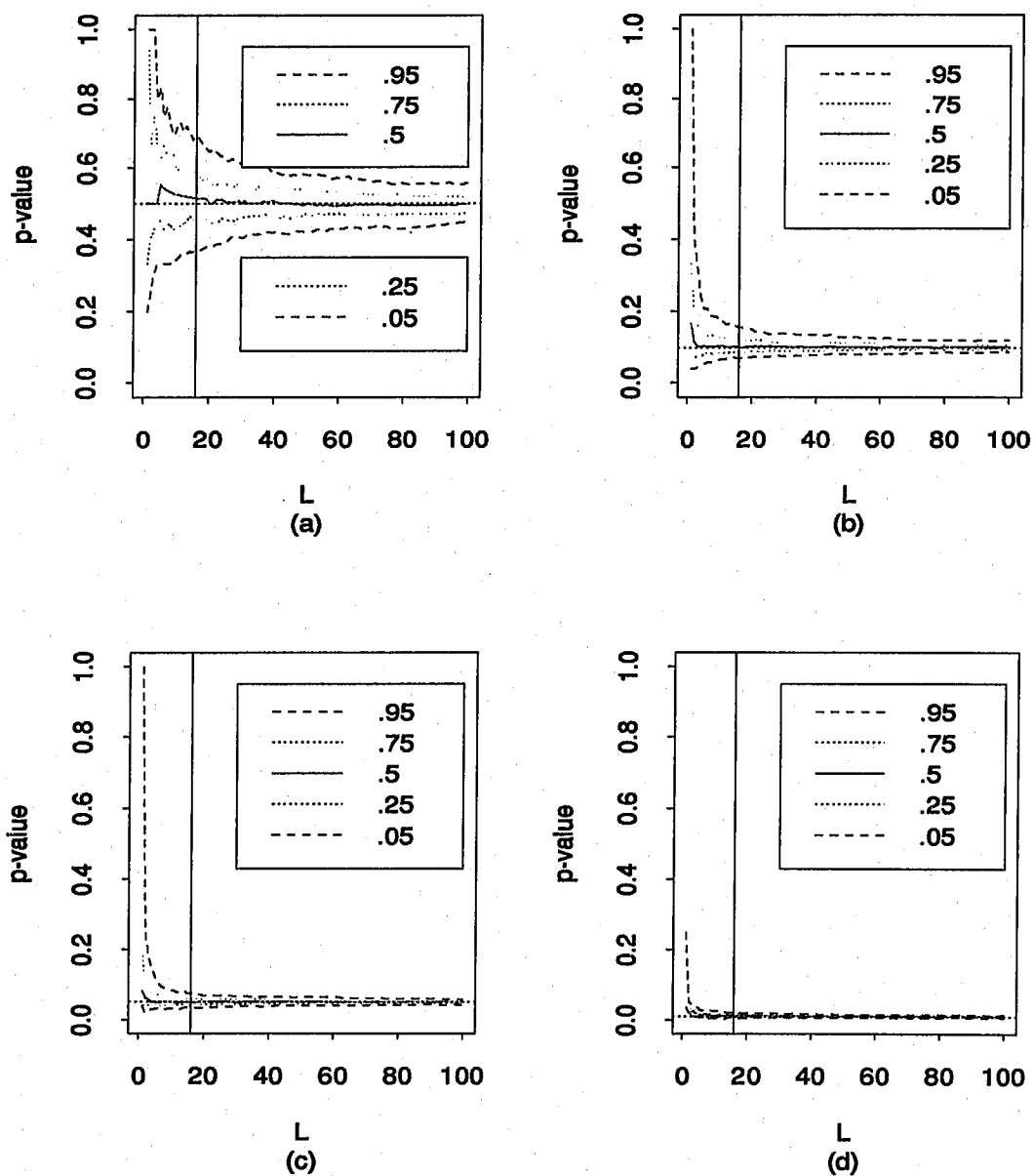
This is a very worthwhile tradeoff, as it is the low  $p$ -values where such accuracy is highly desirable. At the higher  $p$ -values, we are less concerned with such accuracy; we are generally content knowing that the value is high.

In simulation studies, values of 16 for  $L$  and 399 for  $N$  seemed quite successful, though both might be increased for still greater accuracy (at all levels) for analyses in which precision is more important than speed. As a check, the equivalent problem of estimating  $p$  sequentially in Bernoulli trials was simulated with  $p$  equal to .8, .5, .2, .1, .05, .02, and .01. At each value of  $p$ , 200 random sequences were generated to determine the estimate obtained when  $L$  equals the whole numbers 1 through 100. As the sequential nature of the algorithm provides the greatest precision at and just above the value  $L/(N + 1)$ , in all cases  $N$  was set equal to  $(25L - 1)$  to maximize accuracy near .05.

In Table 4.1, we show the 5th, 25th, 50th, 75th, and 95th percentiles for each estimate for the values of  $L$  of 5 (with  $N = 124$ ), 16 (with  $N = 399$ ), and 50 (with  $N = 1249$ ). Sixteen is our choice for use in the mode-existence test, while 50 requires a run time more than triple that of 16. In Figure 4.2, we show these same percentiles for all values of  $L$  for  $p$  equal to .5, .1, .05, and .01. While, as expected, the higher values of  $L$  do result in tighter bounds on the estimates, the effect is most pronounced at high values of  $p$ , where we have little interest beyond knowing that  $p$  is "large." In our opinion, the decreased computing time of a relatively small value of  $L$  (such as 16) greatly outweighs the slight increase in accuracy of the estimates at larger values, at least for our purposes.

#### 4.4 Examining Multimodality of the Overall Data Set

After all of the tests are completed, the results can be displayed on that same mode tree. For example, with the chondrite data, the results of the test are indicated just above each split in the unenhanced mode tree of Figure 4.3. Filled circles indicate that the null hypothesis has been rejected at the fairly generous  $\alpha = 0.15$  level (suggesting



**Figure 4.2** Quantiles of sequential estimates of  $p$  when  $p$  equals (a) 0.5, (b) 0.1, (c) 0.05, and (d) 0.01.  $N = 25L - 1$ . The vertical line indicates  $L = 16$ , and the dotted horizontal line indicates the true value of  $p$ .

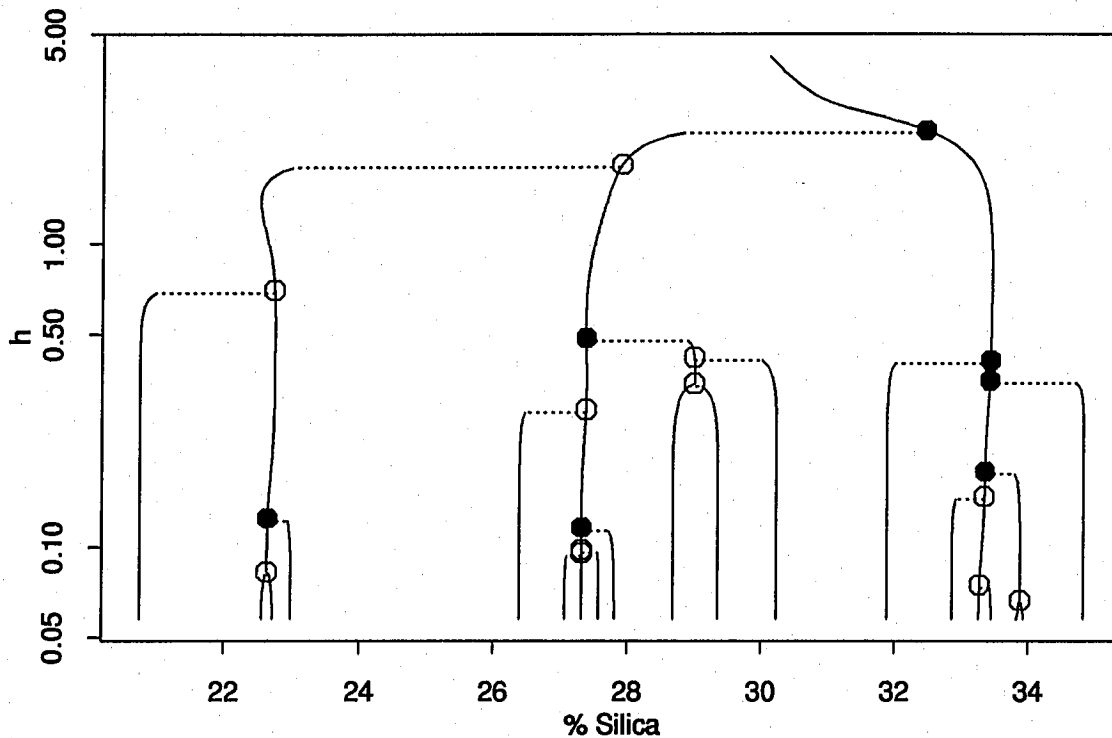


$p$	$L$	$Q_{.05}$	$Q_{.25}$	$Q_{.5}$	$Q_{.75}$	$Q_{.95}$
.8	5	.56	.71	.83	1.0	1.0
	16	.67	.76	.80	.89	.94
	50	.70	.77	.81	.83	.89
.5	5	.33	.45	.56	.63	.83
	16	.37	.46	.52	.59	.70
	50	.42	.47	.50	.54	.58
.2	5	.12	.17	.22	.29	.42
	16	.15	.18	.21	.24	.31
	50	.17	.19	.20	.22	.25
.1	5	.053	.081	.105	.139	.208
	16	.070	.088	.103	.119	.156
	50	.081	.092	.101	.110	.129
.05	5	.032	.040	.052	.071	.122
	16	.035	.041	.049	.058	.077
	50	.041	.046	.050	.055	.063
.02	5	.008	.016	.024	.032	.049
	16	.011	.018	.023	.028	.035
	50	.014	.018	.021	.023	.027
.01	5	.008	.008	.016	.024	.032
	16	.005	.008	.013	.015	.020
	50	.006	.009	.011	.013	.016

**Table 4.1** Estimated quantiles for  $p$ -values calculated using sequential Monte Carlo techniques.

that the mode is real), while open circles indicate a failure to reject the null hypothesis at that level. Choosing an  $\alpha$ -level greater than 5% has been commonly recommended for this problem; see Matthews (1983) and Izenman and Sommer (1988).

The number of significant modes estimated from the data is calculated in a recursive fashion. Each mode tested leads to two branches (which may have one or more tests themselves). If the mode is non-significant, the number of modes it passes up to its parent mode in the tree is simply the sum of the number of modes passed up by its two branches. If the mode is significant, it passes that sum or 1, whichever is greater. This means, for example, that the significant modes near  $x = 33.4$  in



**Figure 4.3** 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.

Figure 9 are all counted as a single “real” mode, while there are three such modes in the entire distribution. Non-significant modes at the same location but different bandwidths as those found significant are disregarded, being viewed as simply being tested at inappropriate bandwidths. Either  $h$  is too large, and the mode has not yet appeared sharply, or  $h$  is too small, and too much of the mode’s mass has been separated into other modes which have already split off. The mode tree is an integral part of summarizing this recursive procedure.

There is a danger to be found in this procedure, one which is shared (to a lesser extent) by the tests of Chapter 2. This is the danger of multiple testing. By conducting a large number of tests, we greatly increase the chance that we will select a false

positive, even if the  $p$ -values for each test are individually accurate or conservative. No clear solution to this has appeared, short of limiting the number of tests conducted and using an appropriately small value of  $\alpha$ .

## 4.5 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 non-binned 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. This estimator approximates a kernel density estimate with far fewer computations, especially for large data sets. The ASH takes binned data, and distributes the weight of each bin among the bins corresponding to the support of the kernel. As initially conceived, the ASH was a true average of  $m$  histograms with the same bin width but different bin edges, and approximated the triangle kernel. However, with appropriate choice of weight function, any kernel can be approximated. See Scott (1992) for further details on the ASH.

A Normal kernel approximation is a far less efficient use of the ASH than finite-domain kernels due to the infinite tails, thus requiring computations over the entire range, rather than over small subsets. Nonetheless, our experiences with the mode tree (Section 3.2), together with the results of Chapter 5, all indicated that the Normal kernel was the only reasonable approach.

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  $\hat{f}_j$  by approximating  $\hat{f}_j$  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 4.1). 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 a double uniform/rejection 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 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.

## Chapter 5

### 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 5.1 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.

The second key theoretical result is of 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 infinity, 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 5.2 and 5.3, respectively, explore the asymptotic behavior of  $M_i$  under these two conditions.

#### 5.1 The Relationship Between $M_i$ and $h$

Theorem 5.1 demonstrates the relationship between  $M_i$  and  $h$ . Because of the focus on  $h$  of this theorem, we explicitly note the dependence (usually left unstated) of the density estimate  $\hat{f}(x)$  and its derived statistics.

**Theorem 5.1** 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 Equation 3.1. 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 5.1** 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 5.2** For a Normal kernel estimator  $\hat{f}(x(h), h)$ ,

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

**Proof of Theorem 5.1** 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  $h_1 \leq h \leq h_2$ , suppose  $\hat{f}(u_{i-1}(h), h) \geq \hat{f}(u_{i+1}(h), h)$ . Define  $w(h)$  to be the unique solution to  $\hat{f}(x, h) = \hat{f}(u_{i-1}(h), h)$  in the range  $u_i(h) < x \leq u_{i+1}(h)$ . Then

$$M_i(h) = \int_{u_{i-1}(h)}^{w(h)} [\hat{f}(x, h) - \hat{f}(u_{i-1}(h), h)] dx,$$

without requiring  $[\cdot]_+$  and

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

By definition, the difference factors in the first two terms are zero, so we can confine our investigation to the third term:

$$\frac{\partial}{\partial h} M_i(h) = \int_{u_{i-1}(h)}^{w(h)} \frac{\partial}{\partial h} \hat{f}(x, h) dx - \int_{u_{i-1}(h)}^{w(h)} \frac{\partial}{\partial h} \hat{f}(u_{i-1}(h), h) dx$$

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

We can ignore the second term of the second integral since  $u_{i-1}(h)$  is an antimode and thus  $\frac{\partial}{\partial x} \hat{f}(u_{i-1}(h), h)$  is zero. Therefore, we arrive at

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

In the second term, all three factors are positive, the last since  $u_{i-1}(h)$  is an antimode. In the first term,  $h$  is also positive. As above, the element  $\frac{\partial}{\partial x} \hat{f}(u_{i-1}(h), h)$  is zero, and  $\frac{\partial}{\partial x} \hat{f}(w(h), h)$  is nonpositive by the requirement that there be only a single mode in the range  $u_{i-1}(h) < x \leq u_{i+1}(h)$ . Therefore,  $\frac{\partial}{\partial h} M_i(h)$  is negative. A similar argument shows that  $\frac{\partial}{\partial h} M_i(h)$  is negative when  $\hat{f}(u_{i-1}(h), h) \leq \hat{f}(u_{i+1}(h), h)$ . Since  $M_i$  is strictly decreasing in  $h$ , the theorem follows.  $\square$

## 5.2 Consistency and Rate of Convergence of $M_i$

In this section, we present proofs concerning the asymptotic behavior of the test statistic  $M_i$ . As indicated in the introduction to this chapter, the asymptotically local behavior of kernel density estimates allows us to concentrate on the two key cases of a unimodal distribution and a bimodal distribution. In each situation, we focus on the first two modes to appear.

For the purposes of these two theorems, we will follow Mammen, Marron, and Fisher (1990) 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 5.2** 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 of  $\hat{f}_n$  observed for  $h_{crit,1} \geq h > h_{crit,2}$ . (One of  $M_1$  and  $M_3$  will be evaluated at  $h_{crit,2}$ , the other at  $h_{crit,k}$  for some  $k > 2$ .) Then  $M_i = O_P(n^{-3/5}(\log n)^{3/4})$  for  $i = 1, 3$ .

**Theorem 5.3** Let  $f$  be a bimodal density satisfying assumptions (A1) - (A5) with modes  $z_1$  and  $z_3$ , and antimode  $z_2$ . Let

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

and

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

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  $h_{crit,1} \geq h > h_{crit,2}$ . Then  $|M_i - \mathcal{M}_i| = O_P(n^{-2/5}(\log n)^{1/2})$  for  $i = 1, 3$ .

These theorems imply that under the conditions of Theorems 5.2 and 5.3,  $M_i$  converges in probability to 0 and  $\mathcal{M}_i$ , respectively.

We will first prove Theorem 5.2 for the case  $M_1$  being tested at  $h_{crit,2}$ . 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  $\hat{f}_n(w) = \hat{f}_n(u_2)$ . We will use extensively the following lemma, which appears as Corollary 2.1 of Mammen, Marron, and Fisher (1990).

**Lemma 5.1** Assume  $f$  has  $j$  modes and  $k \geq j$ . Then for all  $\alpha_n \rightarrow 0$ ,  $\beta_n \rightarrow \infty$ ,

$$P(\alpha_n n^{-1/5} \leq h_{crit,k} \leq \beta_n n^{-1/5}) \rightarrow 1.$$

**Lemma 5.2** For  $f$  as in Theorem 5.2,  $|f(u_i) - f(z_1)| = O_P(n^{-2/5})$  for  $i = 1, 2, 3$ .

**Proof** By Taylor's Theorem and the fact that  $z_1$  is a mode, we have

$$|f(u_i) - f(z_1)| = |u_i - z_1|^2 \frac{|f''(\xi)|}{2}$$

for some  $\xi$  between  $u_i$  and  $z_1$ . Mammen, Marron, and Fisher (1990) show that  $|u_i - z_1| = O_P(n^{-1/5})$  by their Theorem 3. This implies that  $\xi \xrightarrow{P} z_1$  and  $f''(\xi) \xrightarrow{P} f''(z_1)$ , a finite constant by (A5). Therefore,

$$|f(u_i) - f(z_1)| = O_P(n^{-1/5})^2 O_P(1) = O_P(n^{-2/5}).$$

□

**Lemma 5.3** For  $f$  and  $\hat{f}_n$  as in Theorem 5.2,

$$|\hat{f}_n(u_i) - f(z_1)| = O_P(n^{-2/5}(\log n)^{1/2})$$

for  $i = 1, 2, 3$ .

**Proof**

$$\begin{aligned} \left| \hat{f}_n(u_i) - f(z_1) \right| &\leq \left| \hat{f}_n(u_i) - f(u_i) \right| + \left| f(u_i) - f(z_1) \right| \\ &\leq \sup_x \left| \hat{f}_n(x) - f(x) \right| + \left| f(u_i) - f(z_1) \right|. \end{aligned}$$

Silverman (1978) shows that when  $h$  is of order  $n^{-1/5}$  (as in our case by Lemma 5.1), the first term is  $O_P(n^{-2/5}(\log n)^{1/2})$ . The other element is  $O_P(n^{-2/5})$  by Lemma 5.2. The lemma follows.  $\square$

**Lemma 5.4** For  $f$  and  $\hat{f}_n$  as in Theorem 5.2,

$$|f(w) - f(z_1)| = O_P(n^{-2/5}(\log n)^{1/2}).$$

**Proof** Since  $\hat{f}_n(w) = \hat{f}_n(u_2)$ ,  $|\hat{f}_n(w) - f(z_1)| = O_P(n^{-2/5}(\log n)^{1/2})$  by Lemma 5.3.

$$\begin{aligned} \left| f(w) - f(z_1) \right| &\leq \left| f(w) - \hat{f}_n(w) \right| + \left| \hat{f}_n(w) - f(z_1) \right| \\ &\leq \sup_x \left| \hat{f}_n(x) - f(x) \right| + \left| \hat{f}_n(w) - f(z_1) \right| \\ &= O_P(n^{-2/5}(\log n)^{1/2}) + O_P(n^{-2/5}(\log n)^{1/2}) \\ &= O_P(n^{-2/5}(\log n)^{1/2}). \end{aligned}$$

$\square$

**Lemma 5.5** Given  $f$  as in Theorem 5.2, let  $y_j$  be the inflection points of  $f$ . Choose  $\lambda_1$  such that

$$\max_j \{f(y_j)\} < \lambda_1 < f(z_1),$$

and find  $v_1 < v_2$  such that  $f(v_1) = f(v_2) = \lambda_1$ . Then for all  $\varepsilon > 0$ , there exists  $N_\varepsilon$  such that

$$P(w \in [v_1, v_2]) \geq 1 - \varepsilon \quad \text{for all } n > N_\varepsilon.$$

**Proof** By Lemma 5.4, for all  $\varepsilon > 0$ , there exists  $L_\varepsilon$  and  $N_{1\varepsilon}$  such that

$$P\left(\left|f(w) - f(z_1)\right| > L_\varepsilon n^{-2/5}(\log n)^{1/2}\right) \leq \varepsilon \quad \text{for all } n > N_{1\varepsilon}$$

and therefore

$$P\left(f(z_1) - f(w) > L_\varepsilon n^{-2/5}(\log n)^{1/2}\right) \leq \varepsilon \quad \text{for all } n > N_{1\varepsilon}.$$

Let  $N_{2\varepsilon} = \arg \min_n \{L_\varepsilon n^{-2/5}(\log n)^{1/2} < f(z_1) - \lambda_1\}$ , and let  $N_\varepsilon = \max\{N_{1\varepsilon}, N_{2\varepsilon}\}$ .

Then

$$P\left(f(z_1) - f(w) > f(z_1) - \lambda_1\right) \leq \varepsilon \quad \text{for all } n > N_\varepsilon$$

and

$$P\left(f(w) < \lambda_1\right) \leq \varepsilon \quad \text{for all } n > N_\varepsilon.$$

Thus, since  $f(x) \geq \lambda_1$  only in the range  $[v_1, v_2]$ ,

$$P\left(w \in [v_1, v_2]\right) \geq 1 - \varepsilon \quad \text{for all } n > N_\varepsilon.$$

□

**Lemma 5.6** For  $f$  as in Theorem 5.2,  $|w - z_1| = O_P(n^{-1/5}(\log n)^{1/4})$ .

**Proof** Given  $\varepsilon$ , let  $\delta = \varepsilon/(1 + \varepsilon)$ . By Taylor's Theorem and the fact that  $z_1$  is a mode, we have

$$f(w) - f(z_1) = (w - z_1)^2 \frac{f''(\xi)}{2}$$

for some  $\xi$  between  $w$  and  $z_1$ . By Lemma 5.4, there exists  $L_\delta$ , and  $N_{1\delta}$  such that

$$P\left(\left|f(w) - f(z_1)\right| > L_\delta n^{-2/5}(\log n)^{1/2}\right) \leq \delta \quad \text{for all } n > N_{1\delta}.$$

Therefore,

$$P\left(\frac{1}{2}\left|f''(\xi)\right|\left|w - z_1\right|^2 > L_\delta n^{-2/5}(\log n)^{1/2}\right) \leq \delta \quad \text{for all } n > N_{1\delta},$$

and

$$P\left(\left|w - z_1\right|^2 > \frac{2L_\delta}{|f''(\xi)|} n^{-2/5} (\log n)^{1/2}\right) \leq \delta \quad \text{for all } n > N_{1\delta}.$$

Defining  $\lambda_1, v_1$ , and  $v_2$  as in Lemma 5.5, we have

$$\begin{aligned} & P\left(\left|w - z_1\right|^2 > \frac{2L_\delta}{|f''(\xi)|} n^{-2/5} (\log n)^{1/2}\right) \\ & \geq P\left(\left|w - z_1\right|^2 > \frac{2L_\delta}{\min_{x \in [v_1, v_2]} |f''(x)|} n^{-2/5} (\log n)^{1/2}\right) \times \\ & \quad P\left(|f''(\xi)| \geq \min_{x \in [v_1, v_2]} |f''(x)|\right) \\ & \geq P\left(\left|w - z_1\right|^2 > \frac{2L_\delta}{\min_{x \in [v_1, v_2]} |f''(x)|} n^{-2/5} (\log n)^{1/2}\right) \times \\ & \quad P\left(w \in [v_1, v_2]\right). \end{aligned}$$

Therefore,

$$P\left(\left|w - z_1\right|^2 > \frac{2L_\delta}{\min_{x \in [v_1, v_2]} |f''(x)|} n^{-2/5} (\log n)^{1/2}\right) \leq \frac{P\left(\left|w - z_1\right|^2 > \frac{2L_\delta}{|f''(\xi)|} n^{-2/5} (\log n)^{1/2}\right)}{P\left(w \in [v_1, v_2]\right)},$$

and since

$$P\left(\left|w - z_1\right|^2 > \frac{2L_\delta}{|f''(\xi)|} n^{-2/5} (\log n)^{1/2}\right) \leq \delta \quad \text{for all } n > N_{1\delta}$$

and (by Lemma 5.5), for some  $N_{2\delta}$ ,

$$P\left(w \in [v_1, v_2]\right) \geq 1 - \delta \quad \text{for all } n > N_{2\delta},$$

we obtain

$$P\left(\left|w - z_1\right|^2 > \frac{2L_\delta}{\min_{x \in [v_1, v_2]} |f''(x)|} n^{-2/5} (\log n)^{1/2}\right) \leq \frac{\delta}{1 - \delta}$$

for all  $n > \max\{N_{1\delta}, N_{2\delta}\}.$

Since all factors are positive and  $\delta/(1 - \delta) = \varepsilon$ ,

$$P(|w - z_1| > L_\varepsilon n^{-1/5}(\log n)^{1/4}) \leq \varepsilon \quad \text{for all } n > N_\varepsilon,$$

where

$$L_\varepsilon = \sqrt{\frac{2L_\delta}{\min_{x \in [v_1, v_2]} |f''(x)|}} \quad \text{and} \quad N_\varepsilon = \max\{N_{1\delta}, N_{2\delta}\},$$

and therefore  $|w - z_1| = O_P(n^{-1/5}(\log n)^{1/4})$ .  $\square$

### Proof of Theorem 5.2

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

Mammen, et al. (1990) gives us the first element as  $O_P(n^{-1/5})$ , and Lemma 5.6 gives us  $|z_1 - w| = O_P(n^{-1/5}(\log n)^{1/4})$ . Lemma 5.3 tells us both elements on the right are  $O_P(n^{-2/5}(\log n)^{1/2})$ , and the theorem follows.  $\square$

We now move to the proof of Theorem 5.3. Although the proof is in many ways similar to that of Theorem 5.2, there are sufficient differences to warrant inclusion of the additional proof. Again, we consider the case of  $M_1$  being tested at  $h_{crit,2}$ , with the other cases having 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  $\hat{f}_n(w) = \hat{f}_n(u_2)$  and  $t < z_1$  be such that  $f(t) = f(z_2)$ . We will continue to use extensively Lemma 5.1.

**Lemma 5.7** Given  $f$  and  $\hat{f}_n$  as in Theorem 5.3,  $|u_i - z_i| = O_P(n^{-1/5})$

for  $i = 1, 2, 3$ .

**Proof** Mammen, Marron, and Fisher (1990) show in their Theorem 3 that  $|u_i - z_j| = O_P(n^{-1/5})$  for some  $j$ . The fact that

$$\sup_x |\hat{f}_n(x) - f(x)| = O_P(n^{-2/5}(\log n)^{1/2})$$

and the mandatory orderings  $z_1 < z_2 < z_3$ ,  $u_1 < u_2 < u_3$ ,  $f(z_2) < f(z_1)$ ,  $f(z_2) < f(z_3)$ ,  $\hat{f}_n(u_2) < \hat{f}_n(u_1)$ , and  $\hat{f}_n(u_2) < \hat{f}_n(u_3)$ , combine to require that  $j = i$ .  $\square$

**Lemma 5.8** Given  $f$  as in Theorem 5.3,  $|f(u_i) - f(z_i)| = O_P(n^{-2/5})$  for  $i = 1, 2, 3$ .

**Proof** As in Lemma 5.2, by Taylor's Theorem and the fact that  $z_i$  has zero first derivative, we have

$$|f(u_i) - f(z_i)| = |u_i - z_i|^2 \frac{|f''(\xi)|}{2}$$

for some  $\xi$  between  $u_i$  and  $z_i$ . We have  $|u_i - z_i| = O_P(n^{-1/5})$  by Lemma 5.7. This implies that  $\xi \xrightarrow{P} z_1$  and  $f''(\xi) \xrightarrow{P} f''(z_i)$ , a finite constant by (A5). Therefore,

$$|f(u_i) - f(z_i)| = O_P(n^{-1/5})^2 O_P(1) = O_P(n^{-2/5}).$$

$\square$

**Lemma 5.9** Given  $f$  and  $\hat{f}_n$  as in Theorem 5.3,  $|\hat{f}_n(u_i) - f(z_i)| = O_P(n^{-2/5}(\log n)^{1/2})$  for  $i = 1, 2, 3$ .

**Proof**

$$\begin{aligned} |\hat{f}_n(u_i) - f(z_i)| &\leq |\hat{f}_n(u_i) - f(u_i)| + |f(u_i) - f(z_i)| \\ &\leq \sup_x |\hat{f}_n(x) - f(x)| + |f(u_i) - f(z_i)|. \end{aligned}$$

As in Lemma 5.3, the first term is  $O_P(n^{-2/5}(\log n)^{1/2})$  by Silverman (1978). The second term is  $O_P(n^{-2/5})$  by Lemma 5.8 and the lemma follows.  $\square$

**Lemma 5.10** Given  $f$  and  $\hat{f}_n$  as in Theorem 5.3,  $|f(w) - f(t)| = O_P(n^{-2/5}(\log n)^{1/2})$ .

**Proof** Since  $\hat{f}_n(w) = \hat{f}_n(u_2)$  and  $f(t) = f(z_2)$ ,  $|\hat{f}_n(w) - f(t)| = O_P(n^{-2/5}(\log n)^{1/2})$  by Lemma 5.7.

$$\begin{aligned}
|f(w) - f(t)| &\leq |f(w) - \hat{f}_n(w)| + |\hat{f}_n(w) - f(t)| \\
&\leq \sup_x |\hat{f}_n(x) - f(x)| + |\hat{f}_n(w) - f(t)| \\
&= O_P(n^{-2/5}(\log n)^{1/2}) + O_P(n^{-2/5}(\log n)^{1/2}) \\
&= O_P(n^{-2/5}(\log n)^{1/2})
\end{aligned}$$

□

**Lemma 5.11** Given  $f$  and  $\hat{f}_n$  as in Theorem 5.3, choose  $\lambda_2$  such that  $\lambda_2 < f(t)$  and  $f(t) + \lambda_2 < f(z_1)$ , and find  $\eta_1, \eta_2, \eta_3$  such that  $\eta_1 < \eta_2 < z_1 < \eta_3 < z_2$ ,  $f(\eta_1) = f(t) - \lambda_2$ , and  $f(\eta_2) = f(\eta_3) = f(t) + \lambda_2$ . Then for all  $\varepsilon > 0$ , there exists  $N_\varepsilon$  such that

$$P(w \in [\eta_1, \eta_2]) > 1 - \varepsilon \quad \text{for all } n > N_\varepsilon.$$

**Proof** Given  $\varepsilon > 0$ , let  $\delta = \varepsilon/2$ . By Lemma 5.10, there exists  $L_{1\delta}$  and  $N_{1\delta}$  such that

$$P(|f(w) - f(t)| > L_{1\delta} n^{-2/5}(\log n)^{1/2}) \leq \delta \quad \text{for all } n > N_{1\delta}.$$

Let  $N_{2\delta} = \arg \min_n \{L_{1\delta} n^{-2/5}(\log n)^{1/2} < \lambda_2\}$ . Then

$$P(|f(w) - f(t)| > \lambda_2) \leq \delta \quad \text{for all } n > \max\{N_{1\delta}, N_{2\delta}\}.$$

By Lemma 5.7, there exists  $L_{2\delta}$  and  $N_{3\delta}$  such that

$$P(|u_1 - z_1| > L_{2\delta} n^{-2/5}(\log n)^{1/2}) \leq \delta \quad \text{for all } n > N_{3\delta}$$

and therefore

$$P(u_1 - z_1 > L_{2\delta} n^{-2/5}(\log n)^{1/2}) \leq \delta \quad \text{for all } n > N_{3\delta}.$$



Find  $N_{4\delta} = \arg \min_n \{L_{2\delta}n^{-2/5}(\log n)^{1/2} < \eta_3 - z_1\}$ . Then

$$P(u_1 > \eta_3) \leq \delta \quad \text{for all } n > \max\{N_{3\delta}, N_{4\delta}\}$$

and, since  $w < u_1$ ,

$$P(w > \eta_3) \leq \delta \quad \text{for all } n > \max\{N_{3\delta}, N_{4\delta}\}.$$

Letting  $N_\varepsilon = \max\{N_{1\delta}, N_{2\delta}, N_{3\delta}, N_{4\delta}\}$ , we have

$$\begin{aligned} P(w \in [\eta_1, \eta_2]) &\geq 1 - P(|f(w) - f(t)| > \lambda_2) - P(w > \eta_3) \\ &\geq 1 - 2\delta \quad \text{for all } n > N_\varepsilon \\ &= 1 - \varepsilon \quad \text{for all } n > N_\varepsilon. \end{aligned}$$

□

**Lemma 5.12** For  $f$  and  $\hat{f}_n$  as in Theorem 5.3,  $|w-t| = O_P(n^{-2/5}(\log n)^{1/2})$ .

**Proof** Given  $\varepsilon$ , let  $\delta = \varepsilon/(1 + \varepsilon)$ . By Taylor's Theorem, we have

$$f(w) - f(t) = (w - t)f'(\xi)$$

for some  $\xi$  between  $w$  and  $t$ . By Lemma 5.10, there exists  $L_\delta$ , and  $N_{1\delta}$  such that

$$P\left(|f(w) - f(t)| > L_\delta n^{-2/5}(\log n)^{1/2}\right) \leq \delta \quad \text{for all } n > N_{1\delta}.$$

Therefore,

$$P\left(|f'(\xi)||w - t| > L_\delta n^{-2/5}(\log n)^{1/2}\right) \leq \delta \quad \text{for all } n > N_{1\delta},$$

and

$$P\left(|w - t| > \frac{L_\delta}{|f'(\xi)|} n^{-2/5}(\log n)^{1/2}\right) \leq \delta \quad \text{for all } n > N_{1\delta}.$$

Defining  $\lambda_2, \eta_1$ , and  $\eta_2$  as in Lemma 5.11, we have

$$\begin{aligned}
& P\left(|w-t| > \frac{L_\delta}{|f'(\xi)|} n^{-2/5} (\log n)^{1/2}\right) \\
& \geq P\left(|w-t| > \frac{L_\delta}{\min_{x \in [\eta_1, \eta_2]} |f'(x)|} n^{-2/5} (\log n)^{1/2}\right) \times \\
& \quad P\left(|f'(\xi)| \geq \min_{x \in [\eta_1, \eta_2]} |f'(x)|\right) \\
& \geq P\left(|w-t| > \frac{L_\delta}{\min_{x \in [\eta_1, \eta_2]} |f'(x)|} n^{-2/5} (\log n)^{1/2}\right) \times \\
& \quad P\left(w \in [\eta_1, \eta_2]\right).
\end{aligned}$$

Therefore,

$$P\left(|w-t| > \frac{L_\delta}{\min_{x \in [\eta_1, \eta_2]} |f'(x)|} n^{-2/5} (\log n)^{1/2}\right) \leq \frac{P\left(|w-t| > \frac{L_\delta}{|f'(\xi)|} n^{-2/5} (\log n)^{1/2}\right)}{P\left(w \in [\eta_1, \eta_2]\right)},$$

and since

$$P\left(|w-t| > \frac{L_\delta}{|f'(\xi)|} n^{-2/5} (\log n)^{1/2}\right) \leq \delta \quad \text{for all } n > N_{1\delta}$$

and (by Lemma 5.11), for some  $N_{2\delta}$ ,

$$P\left(w \in [\eta_1, \eta_2]\right) \geq 1 - \delta \quad \text{for all } n > N_{2\delta},$$

we get

$$P\left(|w-t| > \frac{L_\delta}{\min_{x \in [\eta_1, \eta_2]} |f'(x)|} n^{-2/5} (\log n)^{1/2}\right) \leq \frac{\delta}{1 - \delta} \quad \text{for all } n > \max\{N_{1\delta}, N_{2\delta}\}.$$

Since  $\delta/(1 - \delta) = \varepsilon$ ,

$$P\left(|w-t| > L_\varepsilon n^{-2/5} (\log n)^{1/2}\right) \leq \varepsilon \quad \text{for all } n > N_\varepsilon,$$

where

$$L_\varepsilon = \frac{L_\delta}{\min_{x \in [\eta_1, \eta_2]} |f'(x)|} \quad \text{and} \quad N_\varepsilon = \max\{N_{1\delta}, N_{2\delta}\}.$$

Thus  $|w-t| = O_P(n^{-2/5} (\log n)^{1/2})$ . □

### Proof of Theorem 5.3

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

Lemmas 5.7 and 5.12 tell us that the first and last elements of the first term are  $O_P(n^{-1/5})$  and  $O_P(n^{-2/5}(\log n)^{1/2})$  respectively, and the remaining element is a constant ( $O_P(1)$ ). Lemma 5.9 and Silverman (1978) tells us the elements of the second term are also  $O_P(n^{-2/5}(\log n)^{1/2})$ . The theorem follows.  $\square$

## Chapter 6

### Simulation Studies

In order to investigate the properties of the mode-existence test, a number of simulation studies were conducted. By running part or all of the procedure on samples drawn from known densities, it is possible to get an idea of how the test performs under various circumstances.

#### 6.1 Statistical Significance

The first aspect to be investigated 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 determined to test samples of various sizes drawn from unimodal distributions. For each sample, the modes existing at  $h_{crit,2}$  were noted and tested at the appropriate bandwidths. The results of Theorem 5.2 suggest that both test statistics should be converging in probability to 0. Nonetheless, 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 and sample size, 100 samples were generated; for each sample, the first two modes to appear were tested at the appropriate bandwidths. The 100 larger  $p$ -values are plotted on uniform quantile-quantile plots. If the reported  $p$ -values are to be accurate (or conservative), the quantiles of these larger  $p$ -values should lie on (or above) the line  $x = y$  (indicating that the probability of getting a

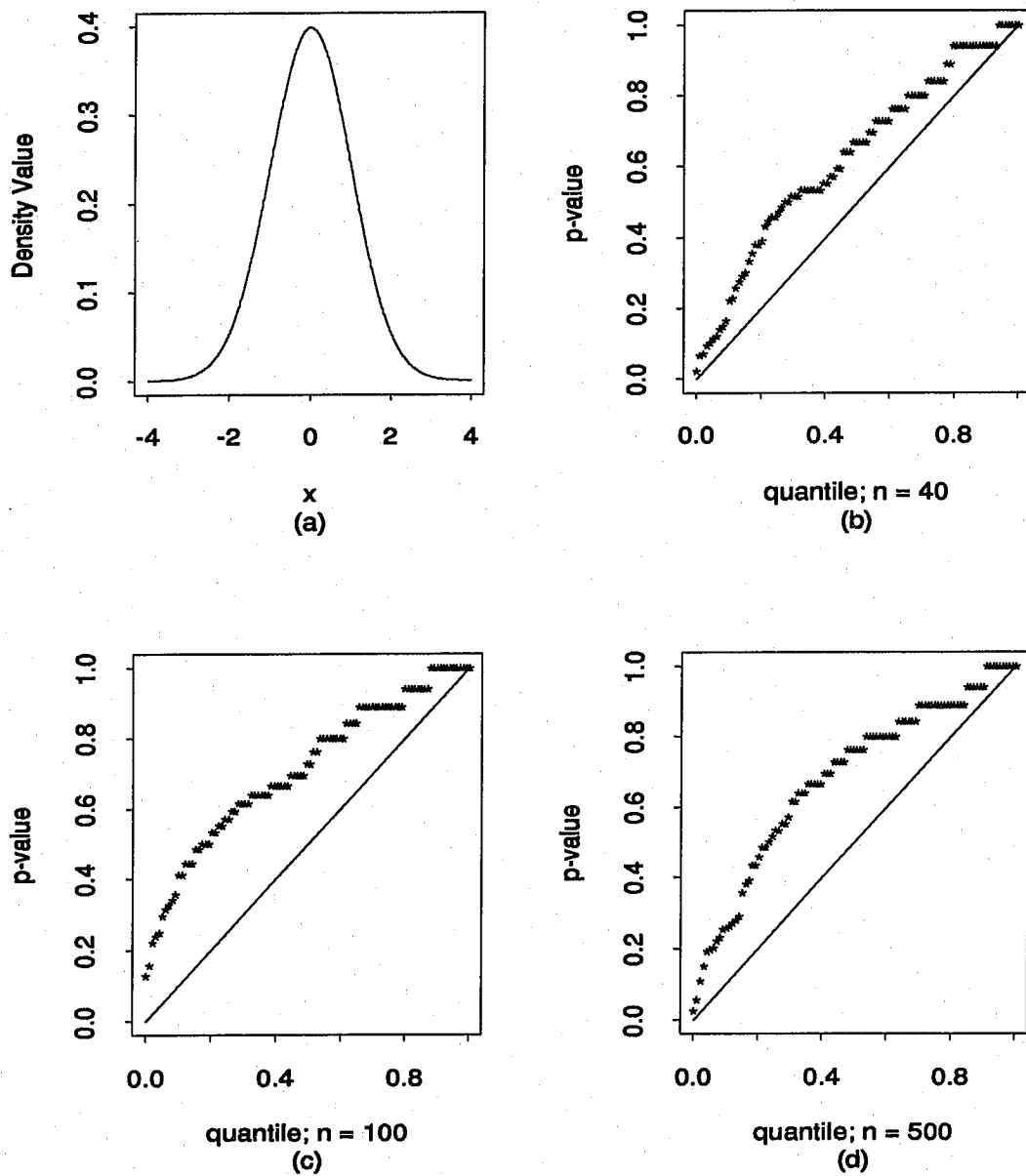
$p$ -value of  $\alpha$  from the null hypothesis is no greater than  $\alpha$ ). This does indeed turn out to be the case.

The first null density selected was the standard Normal density, as in Figure 6.1 (a). Figure 6.1 (b), (c), and (d) contain the q-q plots generated with sample sizes of 40, 100, and 500. The quantiles of the  $p$ -values (points) are all well above the line  $x = y$  (the solid line). This does not seem to be strongly influenced by sample size, with the 40-point results very similar to the 500-point results. Note that the visible discrete levels at the high end of the  $p$ -values are the result of the sequential Monte Carlo calculations, where 16 (our choice of  $L$ ) divided by integers (greater than or equal to 16) are the only possible values above 0.04.

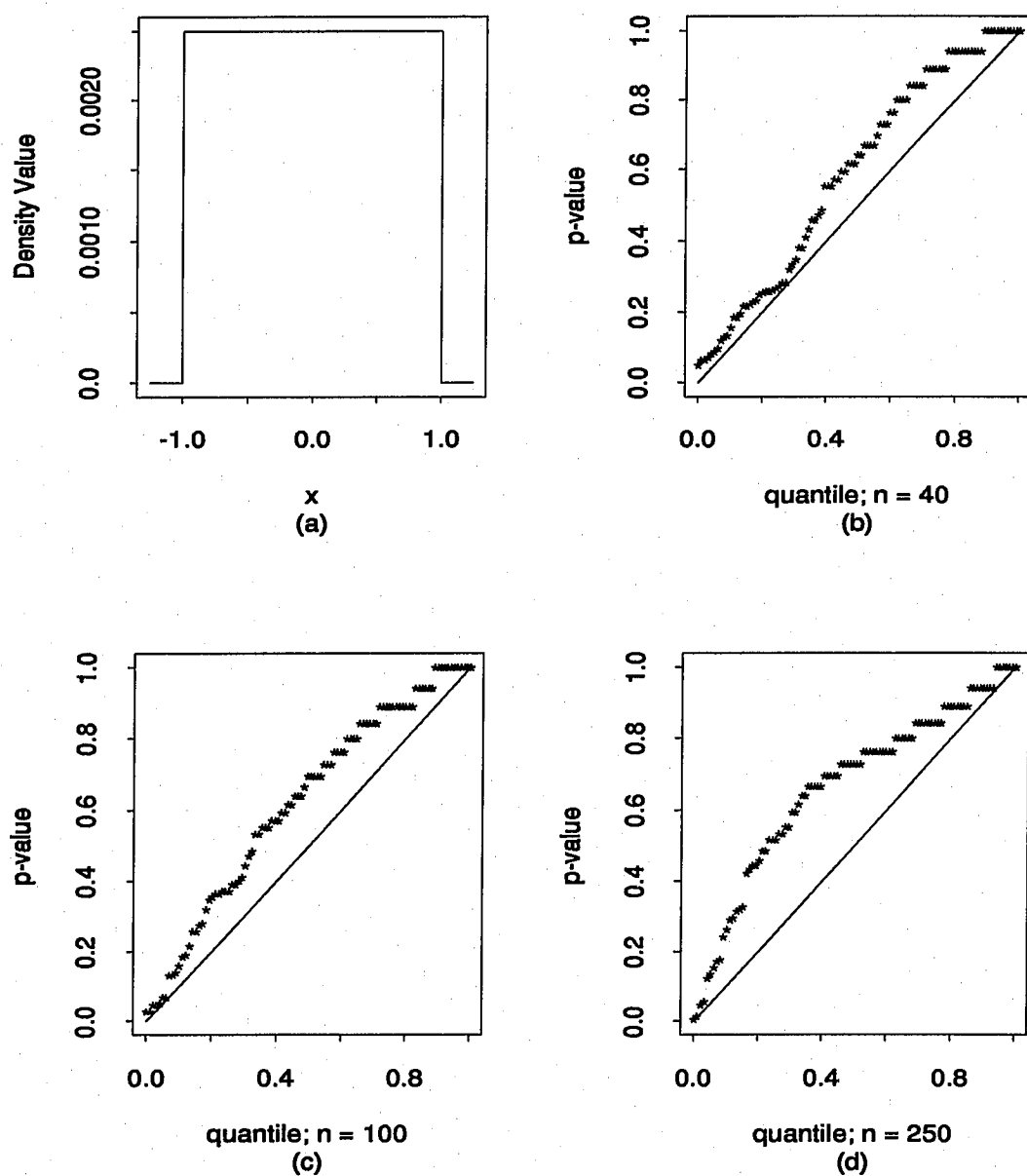
The second null density selected was the uniform  $[-1,1]$  density; see Figure 6.2. The uniform has been used by several authors (Hartigan and Hartigan, 1985; Müller and Sawitzki, 1990) as the standard for computing  $p$ -values, on the basis that it is “worse” than almost any other unimodal density. Figure 6.2 seems to support this conclusion, at least slightly. While the results for all of the sample sizes of 40, 100, and 250 are comfortably above the  $x = y$  line, the first two are somewhat closer to the line than the Normal results of Figure 6.1. In any event, these results also indicate the validity of the reported  $p$ -values.

## 6.2 Power

As the mode-existence test appears to be acting appropriately for unimodal densities, we now turn our attention to the alternative case, in which the density is bimodal. 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 100 samples. Because we are interested in the likelihood of finding both modes significant, we again confine our examinations to the sample mode with the larger  $p$ -value (presumably the smaller of the two). Because



**Figure 6.1** Test density #1: standard Normal density. Underlying density (a), and 100  $p$ -values generated from test density #1 using sample sizes (b) 40, (c) 100, and (d) 500.



**Figure 6.2** Test density #2: Uniform density. Underlying density (a), and 100  $p$ -values generated from test density #1 using sample sizes (b) 40, (c) 100, and (d) 250.

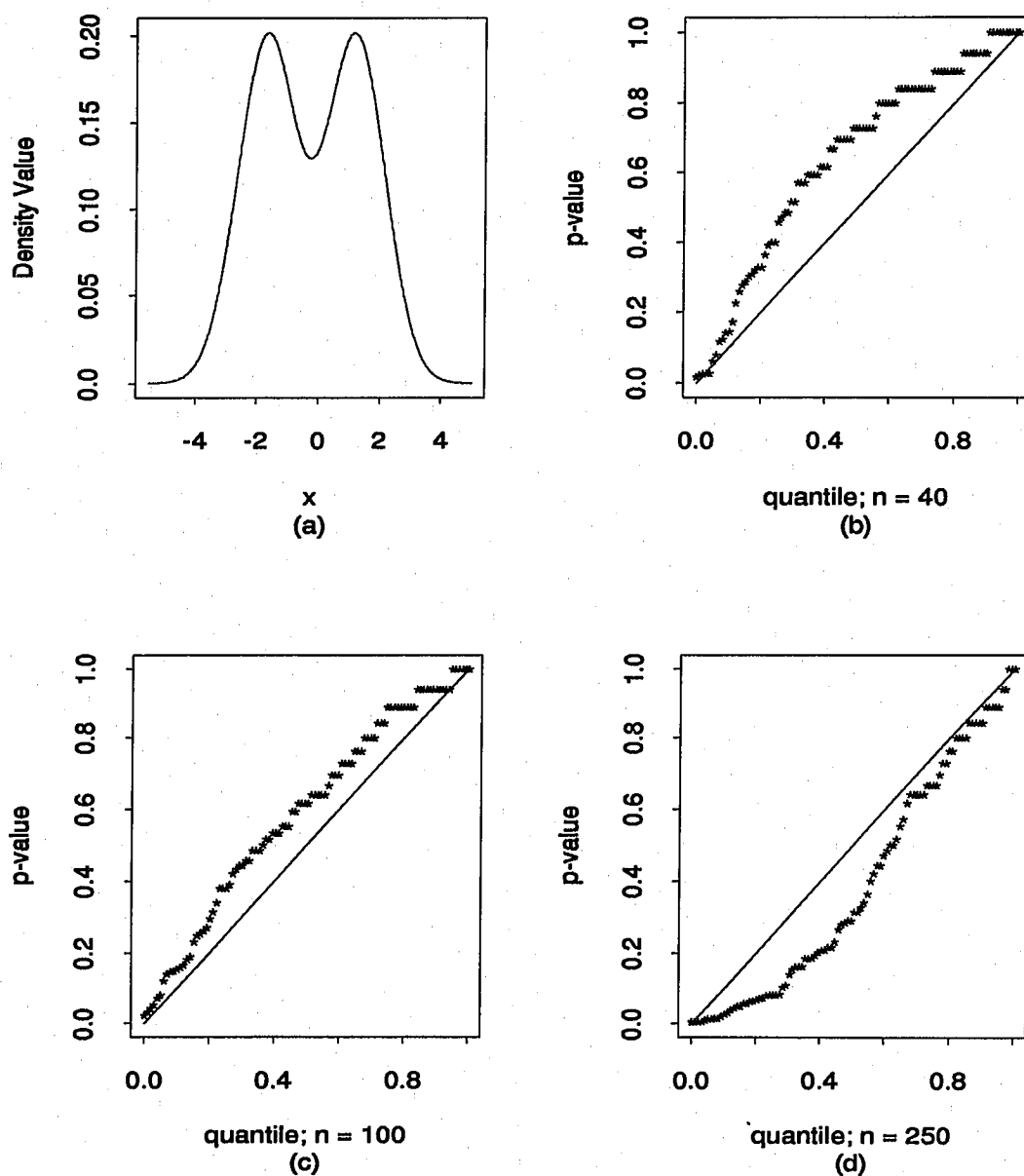
the density truly is bimodal, here it is desirable to see the quantiles of the  $p$ -values fall well below the  $x = y$  line.

The first density tested was an equal mixture of two Normal densities, centered at -1.5 and 1.5, each with standard deviation 1; see Figure 6.3 (a). Though clearly bimodal, the fraction of the mass actually in the modes above the antimode is relatively small, and the test had corresponding difficulties at small sample sizes. In Figure 6.3 (b), (c), and (d), we show the quantiles for the larger  $p$ -values generated from sample sizes of 40, 100, and 250. For the two smaller sample sizes, the density is indistinguishable from a unimodal density; in each, the collection of quantiles lies well above the  $x = y$  line. By the time the sample size reaches 250, things are somewhat better. Here at least we can see that the  $p$ -values are lower, on average, than they would be from a unimodal density. But even at this sample size, only 15% of the samples would reject the null hypothesis and accept the mode at the  $\alpha = 0.05$  level, and only 31% would reject even at the  $\alpha = 0.15$  level.

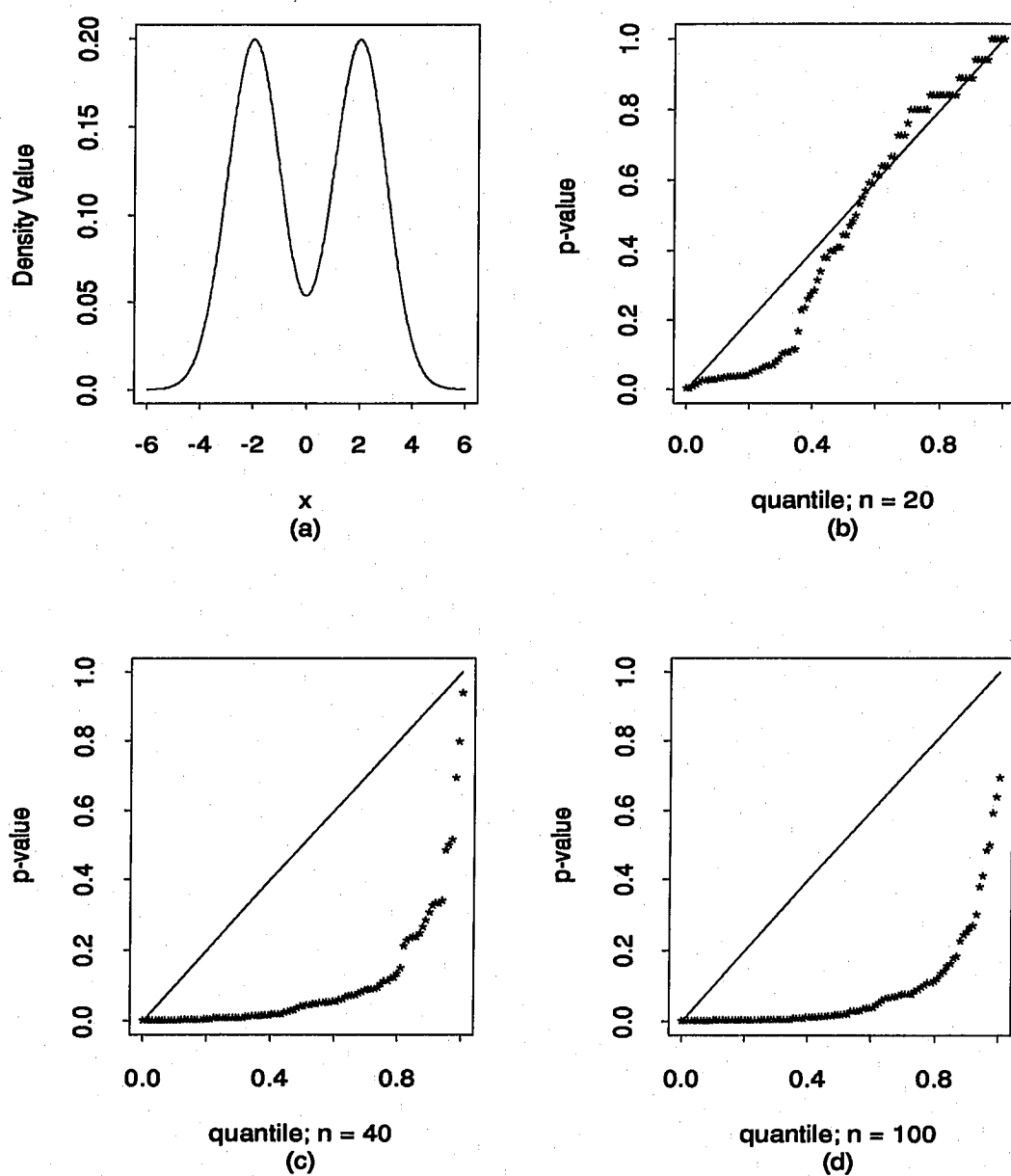
The second bimodal density investigated was another equal mixture of Normal densities with standard deviations of 1, but this time centered at -2 and 2. The extra standard deviation of distance between the modes results in a much deeper valley between the two peaks and, not surprisingly, makes the discrimination task far easier. Figure 6.4 shows the underlying density and the quantiles for sample sizes of 20, 40 and 100. By the time the sample size reaches 100, rejection of the null is very common; 61% reject at the  $\alpha = 0.05$ , and 83% reject at the  $\alpha = 0.15$  level. In fact, the test does nearly this well when the sample size equals 40; 54% reject when  $\alpha = 0.05$ , and 80% reject when  $\alpha = 0.15$ . Even at the very small sample size of 20, fully 21% reject at the 0.05 level, and 35% reject when  $\alpha = 0.15$ . Clearly, the large, well-separated modes of this example are easy for the test to distinguish.

The third bimodal density used was again a mixture of two Normal densities, but this time with different weights and variances. 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





**Figure 6.3** Test Density #3:  $\frac{1}{2}N(-1.5, 1) + \frac{1}{2}N(1.5, 1)$ . Underlying density (a), and 100  $p$ -values generated from test density #1 using sample sizes (b) 40, (c) 100, and (d) 250.



**Figure 6.4** Test Density #4:  $\frac{1}{2}N(2,1) + \frac{1}{2}N(-2,1)$ . Underlying density (a), and 100  $p$ -values generated from test density #1 using sample sizes (b) 20, (c) 40, and (d) 100.

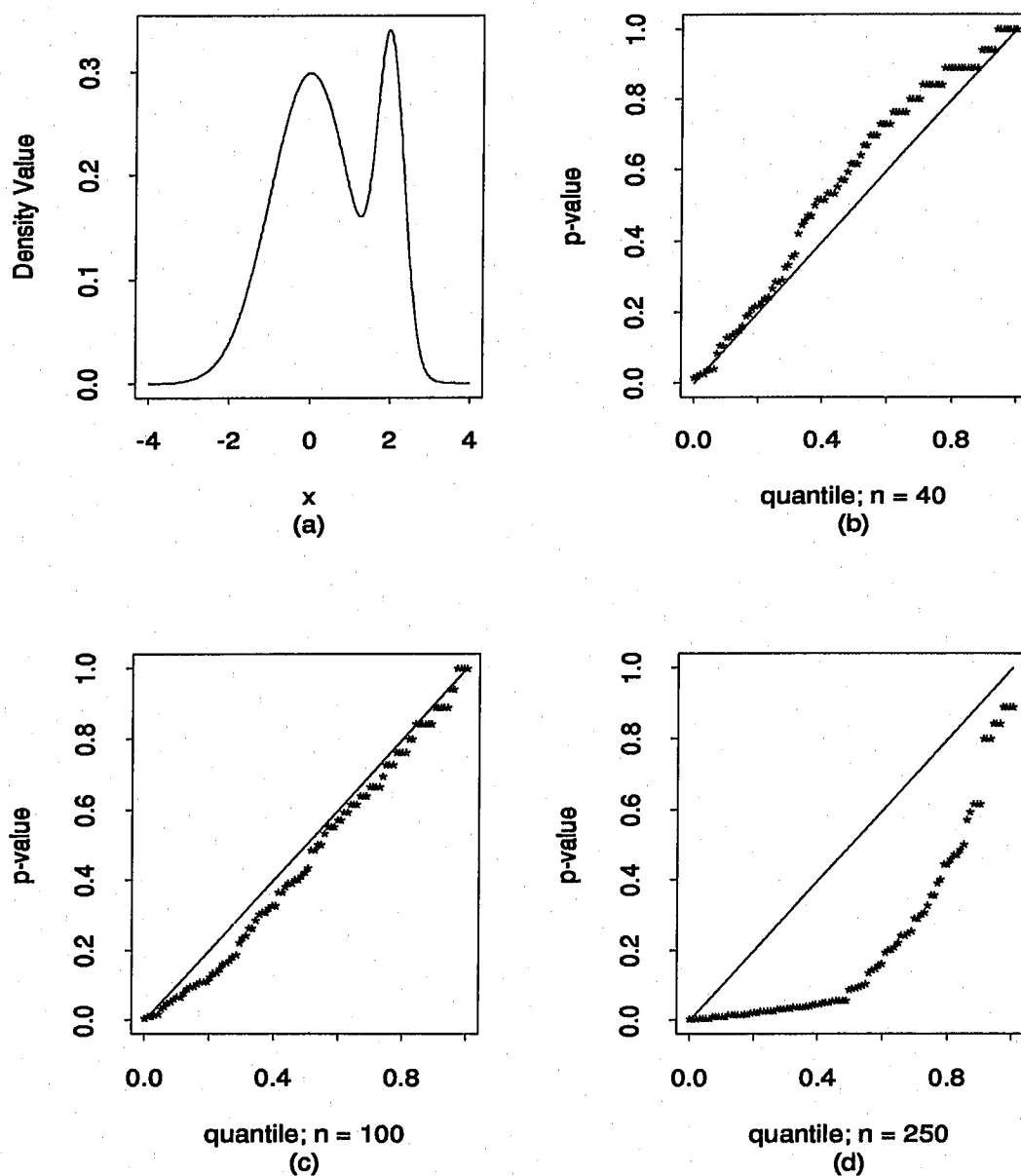
deviation  $1/3$ ; see Figure 6.5. The results were similar to those of the first bimodal example. With a sample size of 40, the results could have been produced by a unimodal density. When the sample size was 100, the  $p$ -value quantiles fell below the  $x = y$  line, but just barely. Only 8% would reject at the  $\alpha = 0.05$  level, and only 24% would reject at the  $\alpha = 0.15$  level. By sample size 250, though, the situation is greatly improved, as 41% reject at the 0.05 level, and 58% reject at the 0.15 level.

### 6.3 The Adaptive Nature of the Test

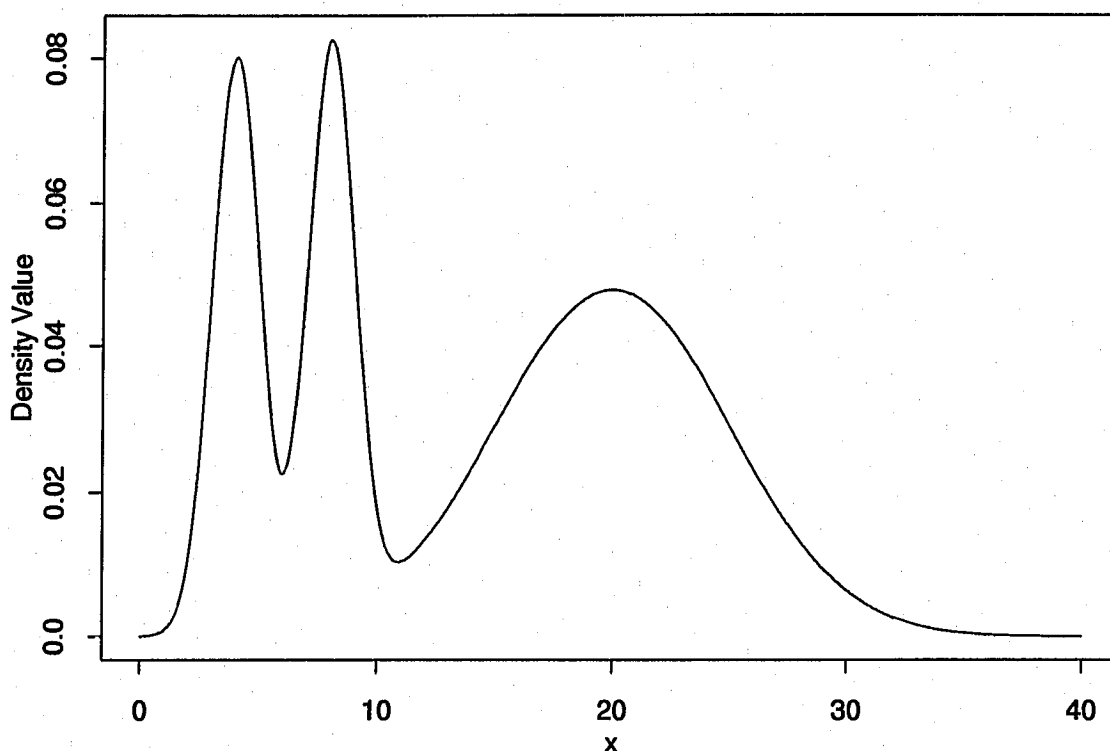
To investigate the effectiveness of the adaptive nature of the test, a test density was chosen that was designed to be both easy and difficult. The density can be seen in Figure 6.6, and is a trimodal mixture of three Normal densities in a ratio of 1 : 1 : 3. The first two have standard deviations of 1, and means of 4 and 8. The third is a much wider density, with standard deviation 5 and mean 20.

The density is easy in the sense that all three modes are large, clear, and highly separated (by amounts comparable to those of test density # 4). 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. Twenty 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 6.1.

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 19 of the 20 tests performed, when tested at the  $\alpha = 0.05$  level. Unfortunately, it did not reject bimodality for trimodality in any of those 19 cases. The one case in which unimodality was *not* rejected at the 5% level was also unique in that it was also the only sample which *did* reject bimodality for trimodality at the  $\alpha = 0.15$  level. By the  $\alpha = 0.4$  level, 8 of the 20 samples reject bimodality,



**Figure 6.5** Test Density #5:  $\frac{3}{4}N(0, 1) + \frac{1}{4}N(2, \frac{1}{3})$ . Underlying density (a), and 100  $p$ -values generated from test density #1 using sample sizes (b) 40, (c) 100, and (d) 250.



**Figure 6.6** Test Density #6:  $\frac{1}{5}N(4, 1) + \frac{1}{5}N(8, 1) + \frac{3}{5}N(20, 5)$

but only 3 samples stop there. In addition, the three modes at  $h_{crit,3}$  do not include all three of the true modes in 2 of the 3 cases where trimodality is indicated! The remaining 5 samples each indicate four or more modes for the density, and all 5 had higher  $p$ -values for rejecting bimodality than for rejecting trimodality. Thus none of the 5 would indicate trimodality for any choice of  $\alpha$  level. One further indication of the problems Silverman's test has with a density such as this is that for 11 of the 20 samples, 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 far greater success with this density. At the  $\alpha = 0.05$  level, 12 of the 20 densities indicated trimodality, though only 7 of the 12 actually found the correct three modes. It is clear, however, that

Modes Found	Mode-Existence Test		Critical Bandwidth Test		
	$\alpha$				
	0.05	0.15	0.05	0.15	0.4
1	2	0	1	0	0
2	2	2	19	19	12
3 (correct)	7	3	0	0	1
3 (incorrect)	5	3	0	1	2
4	4	6	0	0	2
5	0	4	0	0	1
6 (or more)	0	2	0	0	2

**Table 6.1** Results of testing 20 samples of size 100 from Test Density #6 using the mode-existence test and Silverman's critical bandwidth test with several choices of  $\alpha$  level.

the far larger number of tests of the mode-existence test leads to a less conservative procedure overall; 9 of the 20 samples found spurious modes at the  $\alpha = 0.05$  level, and 15 of the 20 did so at the  $\alpha = 0.15$  level. Even at the  $\alpha = 0.4$  level, Silverman's test found spurious modes in only 7 of the samples. In a sense, this is analogous to the bias versus variance tradeoff commonly found in density estimation. 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 the full procedure is followed.

## Chapter 7

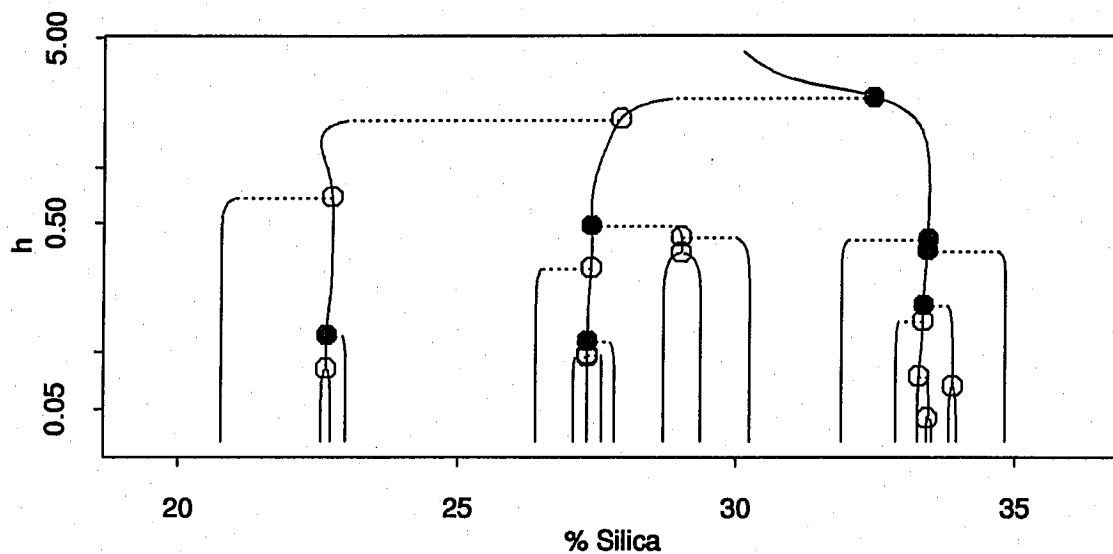
### Case Studies

We now examine the behavior of the mode-existence test in analyzing some of the “classic” data sets in the bump-hunting literature.

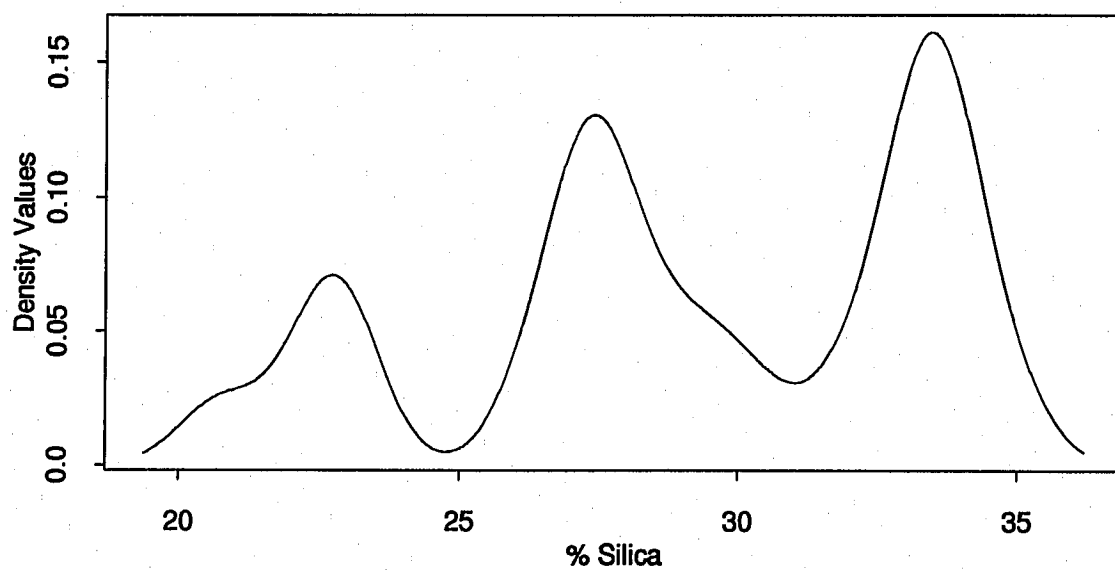
#### 7.1 Chondrite Data

The first data set we shall investigate is Ahrens’ (1965) chondrite meteorite data set which we have been using as an example throughout the course of this work. The data consist of the percent silica in 22 chondrite meteors. Though first used in the bump-hunting context by Good and Gaskins (1980), it has been a favorite test case by many authors in the area, including Silverman (1981), Wong (1985), and Müller and Sawitzki (1991). The reasons for its popularity no doubt are its strong apparent multimodality and its small sample size. Despite the fact that the data set has only 22 points, every author except Wong has decided that the chondrite density is trimodal. (Wong indicated that it was merely bimodal.)

Figure 7.1 is a repeat of Figure 4.3, showing the mode tree generated from the chondrite data set with significant modes (at the  $\alpha = 0.15$  level) indicated. The tree indicates agreement with the majority of authors who have analyzed this data set before; it seems to be clearly trimodal. Figure 7.2 is the kernel estimate of the data made with  $h = 0.75$ . The three large modes, with deep valleys between them and roughly equal widths, contribute heavily to the ability to identify this density as multimodal even from a sample size as small as 22. Table 7.1 lists the modes, test bandwidths, and  $p$ -values for the seven tests with  $p$ -values less than 0.15.



**Figure 7.1** 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.



**Figure 7.2** Kernel density estimates for the chondrite data;  $h = 0.75$ .



Modes	$h$	$p$ -value
Single Modes		
22.6	0.125	0.061
27.4	0.486	0.054
27.3	0.116	0.015
33.4	0.410	0.005
33.4	0.355	0.040
33.4	0.180	0.005
Composite Modes		
32.5	2.412	0.000

**Table 7.1** Modes of chondrite mode tree with  $p$ -values less than 0.15.

## 7.2 Stamp Data

We next examine the 1872 Hidalgo stamp thicknesses ( $n = 485$ ) measured by Walton van Winkle and analyzed in Wilson (1983) and Izenman and Sommer (1988). Wilson hand-smoothed a histogram on the 437 unwatermarked stamps without gum, and came to the conclusion that the data were bimodal, with modes near 0.077 mm and 0.105 mm. Izenman and Sommer applied the critical-bandwidth test described in Silverman (1981) to the full set of 485 measurements. Their results supported Wilson in rejecting unimodality, but indicated *seven* modes (located at 0.072, 0.080, 0.090, 0.100, 0.110, 0.120, and 0.130 mm), rather than Wilson's two. When Izenman and Sommer applied a square root transformation to the data in an effort to check for spurious modes in the right tail, Silverman's test supported *nine* modes, the same seven modes, plus an additional two in the *left* tail at 0.060 mm and 0.064 mm. We note that the small mode at 0.060 mm is based on only a single data point.

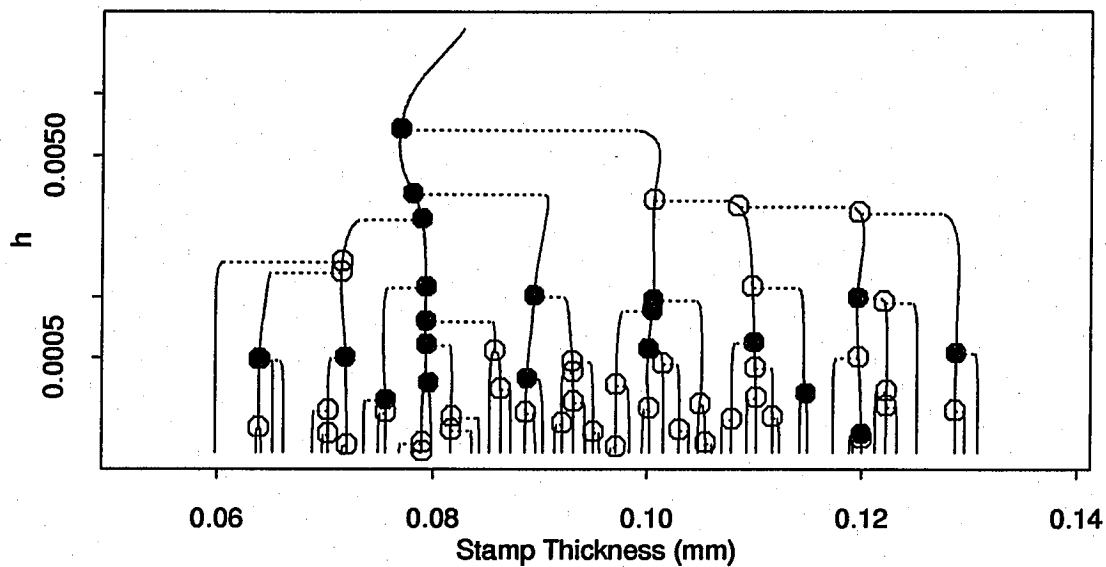
Table 7.2 and the mode tree of Figure 7.3 indicates that the transformation attempt may be misleading in two ways. All seven modes of the original test are indicated, as well as the larger of the two small left-tail modes found by Izenman and Sommer after transformation. In addition, however, we found two additional and

tantalizing modes at 0.075 mm and 0.115 mm. We note that the original data were effectively binned into 71 bins by the resolution of the measurements. The data were adjusted by the FP-blurring technique of Section 4.5 before analysis.

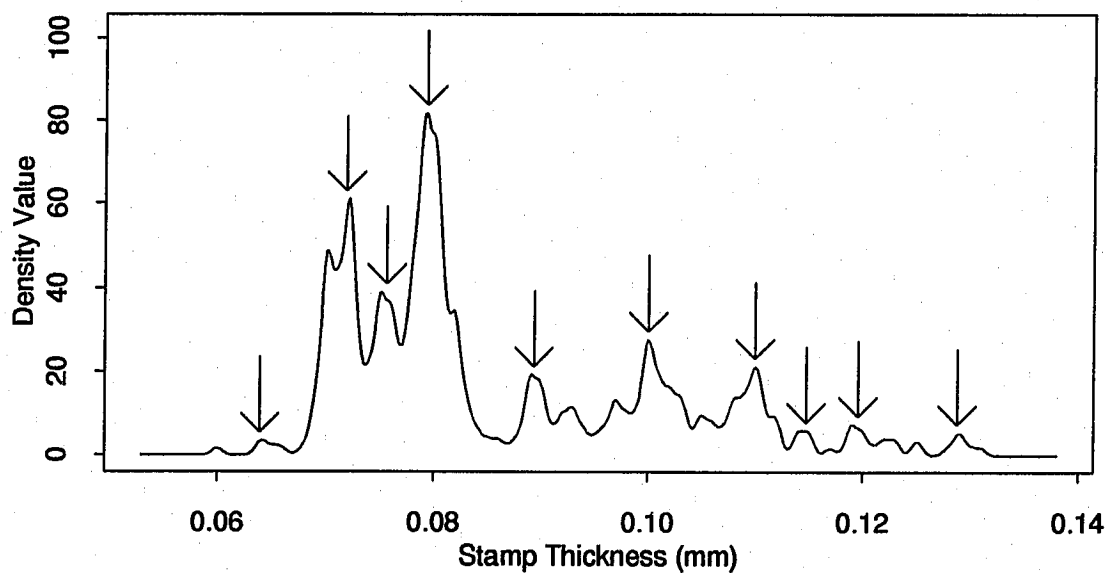
The more impressive of the new modes is at 0.075 mm, and may be clearly seen in Figure 7.4, as the medium-sized peak between the two large modes at 0.072 mm and 0.080 mm. It is not surprising that Silverman's test missed this mode and the one at 0.115 mm, for by the time the bandwidth is small enough to distinguish these modes from their large neighbors, numerous spurious modes have appeared in the right tail of the density. This is a prime example of the inability of a single  $h$  to work well for all  $x$ , mentioned in Section 1.2. Under such circumstances, any procedure

Modes	$h$	$p$ -value
Single Modes		
0.064	0.00049	0.148
0.071	0.00051	0.003
0.076	0.00031	0.109
0.079	0.00077	0.008
0.079	0.00059	0.003
0.080	0.00038	0.003
0.089	0.00102	0.119
0.089	0.00040	0.112
0.101	0.00098	0.013
0.100	0.00087	0.008
0.100	0.00056	0.059
0.110	0.00060	0.013
0.115	0.00034	0.120
0.120	0.00100	0.085
0.129	0.00053	0.023
Composite Modes		
0.077	0.00680	0.000
0.078	0.00328	0.013
0.078	0.00245	0.003
0.079	0.00113	0.018

**Table 7.2** Modes of Hidalgo stamp mode tree with  $p$ -values less than 0.15.



**Figure 7.3** Test mode tree for the Hidalgo stamp data. Filled circles are significant modes at the  $\alpha = 0.15$  level, open circles are modes which are not significant at this level.



**Figure 7.4** Kernel density estimates for the Hidalgo stamp data;  $h = 0.0005$ . Modes found significant at the  $\alpha = 0.15$  level are indicated by arrows.

testing multimodality globally using a single bandwidth is likely to fail to recognize such modes.

### 7.3 Snowfall Data

Next on the agenda are the measurements of yearly snowfall in Buffalo, New York, from 1910 to 1972. This data set is interesting because it appears to be a difficult choice, not between two adjacent possibilities such as unimodality and bimodality, but between unimodality and trimodality; see Figure 7.5. Parzen (1979) leans toward the unimodal interpretation, indicating that the trimodal possibility appears to be an overfitting of the data. He indicates, however, that Carmichael (1976) and Thaler (1974) view the trimodal estimate as more likely.

As Figure 7.6 indicates, we find ourselves agreeing with Parzen, that (even at the  $\alpha = 0.15$  level), we fail to reject the null of nonexistence for any but the large, central mode. Thus, we concur, that these 63 data points fail to provide conclusive evidence in favor of the possibility of trimodality. It is worth noting, however, that the  $p$ -values found include a value of 0.168 for the mode at 54.29 ( $h = 2.01$ ), and a value of 0.160 for the mode at 104.42 ( $h = 0.92$ ). Table 7.3 lists the key information about those tests which indicated significance.

Modes	$h$	$p$ -value
Single Modes		
80.3	7.45	0.000
80.8	2.82	0.005
89.8	0.57	0.082

**Table 7.3** Modes of Buffalo snowfall mode tree with  $p$ -values less than 0.15.

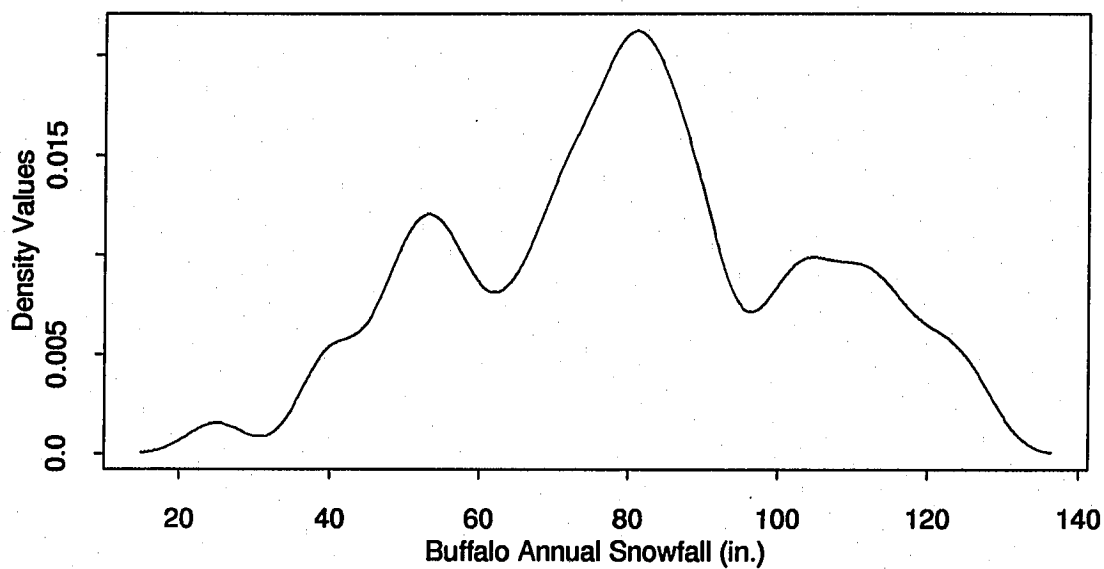


Figure 7.5 Kernel density estimates for the Buffalo snowfall data;  $h = 4.0$ .

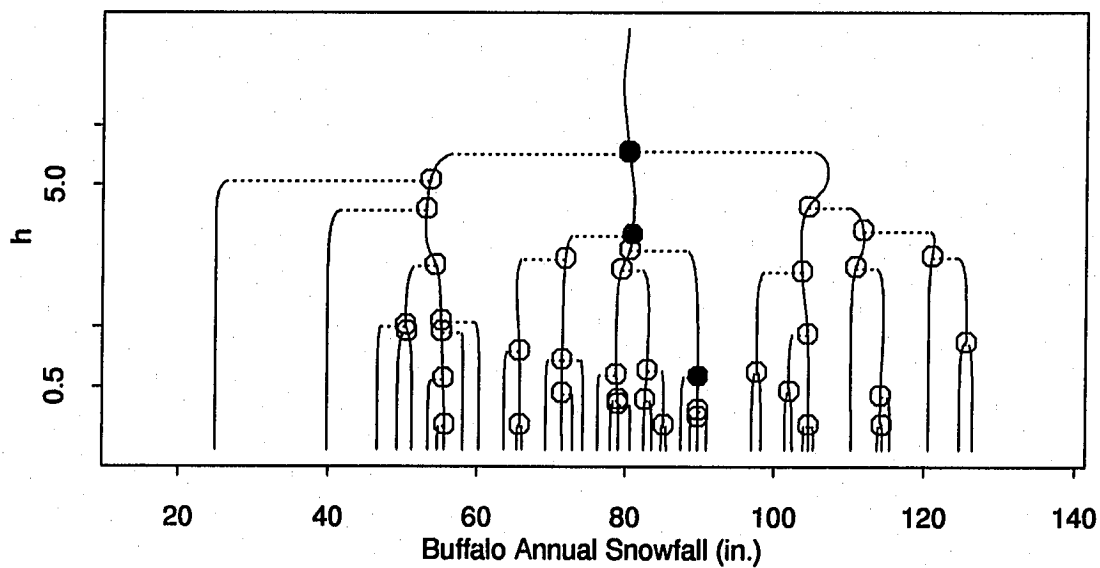


Figure 7.6 Test mode tree for the Buffalo snowfall data. Filled circles are significant modes at the  $\alpha = 0.15$  level, open circles are modes which are not significant at this level.

## 7.4 Galaxy Data

The data set which produced the mode tree of Figure 7.7 comes from Postman, Huchra, and Geller (1986), and was introduced to the bump-hunting community in Roeder (1990). It consists of the velocities (in km/s) of 82 galaxies, measured in an unfilled survey of the Corona Borealis region. An unfilled survey is one in which only small portions of the sky are investigated (here, 6 well-separated conic sections), but in which the sampling is very deep. Figure 7.8 is the kernel density estimate of this density with  $h$  equal to 500, and Table 7.4 lists the significant tests.

Two points are worthy of note here. The first is positive. The mode tree provides strong evidence of the leftmost mode in Figure 7.8 and the large central mode ( $p$ -values of 0.005 and 0.0025, respectively), yet the mass on the far right is not found to be a significant mode (with  $p$ -value 0.25). This mass is sufficiently well separated from the rest to result in at least one apparent mode in an estimate with any reasonable value of  $h$ . Nonetheless, it consists of only three points. It seems far more reasonable to indicate, as a  $p$ -value of 0.25 does, that this is a *potential* mode, one that with more data might be strongly evident, but one that does not yet contain enough points to ensure that it is not merely a tail fluctuation.

Modes	$h$	$p$ -value
Single Modes		
9471	278	0.005
9463	81	0.088
21184	2528	0.010
20012	935	0.003
19954	252	0.005
Composite Modes		
21255	3069	0.000

**Table 7.4** Modes of galaxy mode tree with  $p$ -values less than 0.15.

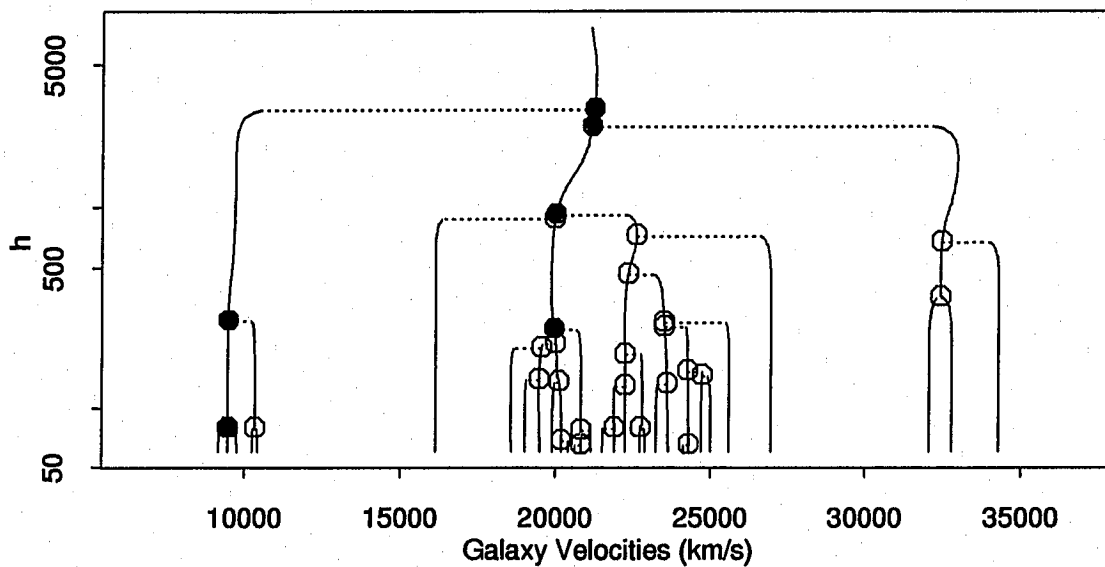


Figure 7.7 Test mode tree for the galaxy velocity data. Filled circles are significant modes at the  $\alpha = 0.15$  level, open circles are modes which are not significant at this level.

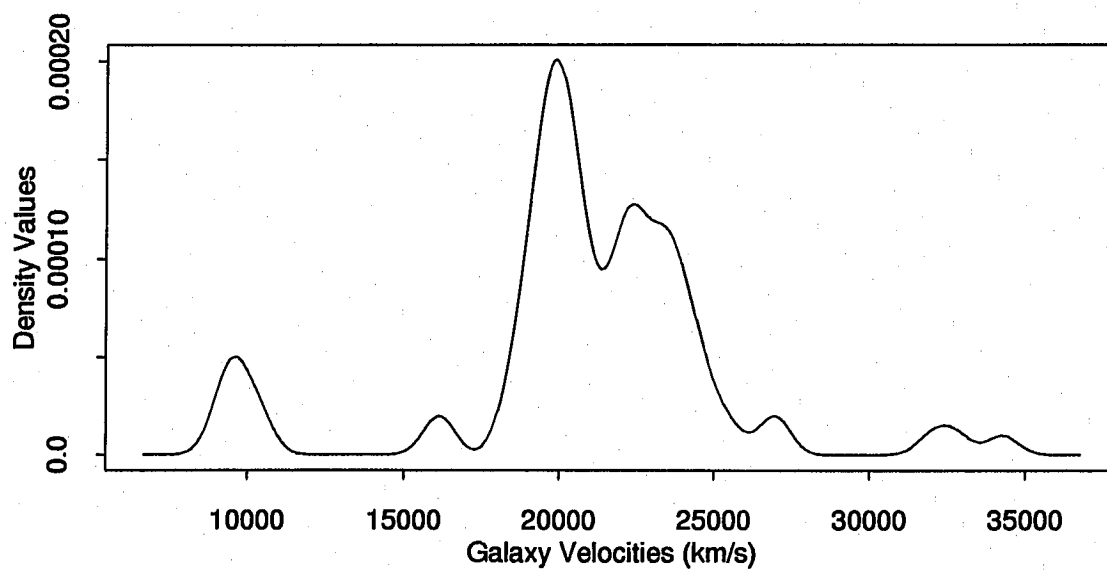


Figure 7.8 Kernel density estimates for the galaxy velocity data;  $h = 500$ .

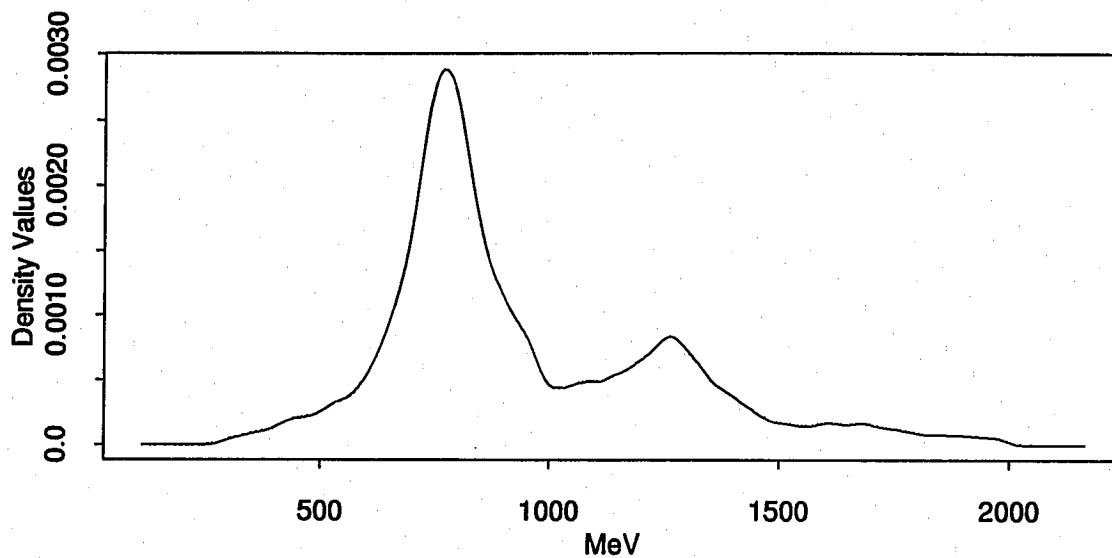
On the other hand, Figures 7.7 and 7.8 point out a potential small-sample problem with the proposed procedure. The right peak in the large central mass of Figure 7.8 is judged highly nonsignificant by the test (0.667), but is tested at the relatively high bandwidth of 474. While this mode is not significant at this level, and neither of the two smaller modes it splits into are significant at lower levels, it is possible that the entire region (now two modes) might be significant as a single mode if measured and tested as one. Perhaps some such variation would be useful for instances where an antinode is higher than the two on either side. How this would be implemented is not clear, and it would complicate an already complex procedure, but the thought bears consideration.

## 7.5 LRL Data

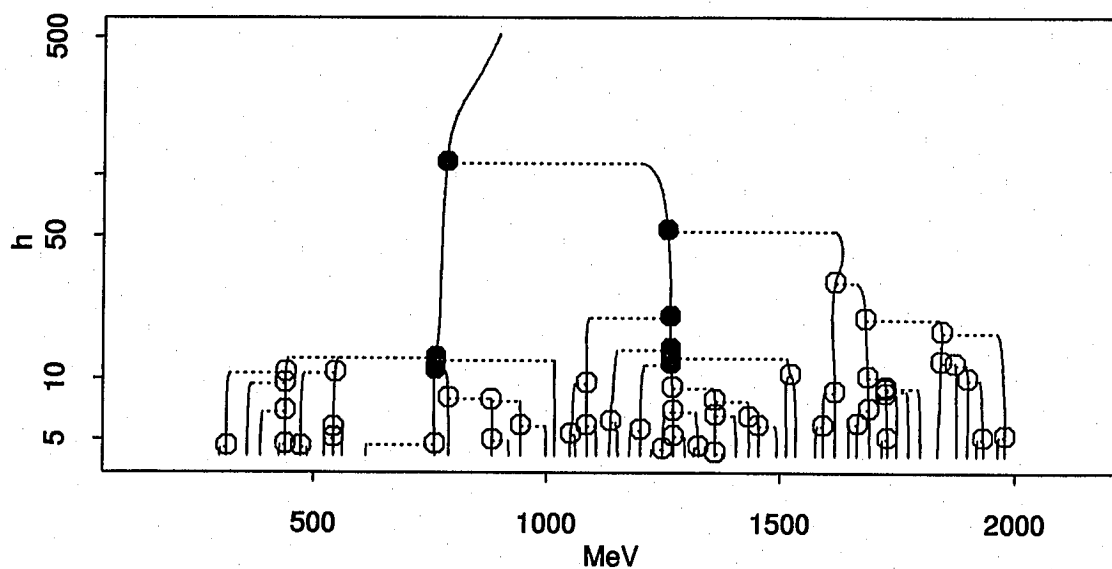
Next we investigate the data from 25,752 scattering events measured at the Lawrence Radiation Laboratory commonly referred to as the LRL data. This data set was published (and probably collected) as 172 bin counts. Therefore, we again used FP-blurring to adjust the data before proceeding. The Normal kernel density estimate generated from these data when  $h = 20$  can be seen in Figure 7.9. These data were first examined in the bump hunting context by Good and Gaskins (1980). In their analysis they found 13 bumps which they claimed to be “odds-on,” that is, having (Bayesian) probability of existence of greater than  $1/2$ . It is not clear from the article how many of these bumps were actually modes, but it appears from their Figure A that this was true of only four (two large modes and two small ones in the right tail).

In Figure 7.10, we show the mode tree for this data set. Figure 7.10 and Table 7.5 present the surprising result that only the two large modes show up as significant, even at the  $\alpha = 0.15$  level, despite the large sample size. (On the other hand, these two were *very* significant, both receiving the lowest possible  $p$ -value of 0.0025.) The only others which had  $p$ -values of less than 0.4 were a mode at 1687 MeV (tested at  $h = 10.3$  for a  $p$ -value of 0.21), and another at the edge of the data at 1979 MeV





**Figure 7.9** Kernel density estimates for the LRL data;  $h = 20$ .



**Figure 7.10** Test mode tree for the LRL data. Filled circles are significant modes at the  $\alpha = 0.15$  level, open circles are modes which are not significant at this level.

Modes	$h$	$p$ -value
Single Modes		
760	12.85	0.003
760	12.55	0.003
758	11.11	0.003
1256	53.80	0.003
1264	20.38	0.003
1265	14.17	0.061
1265	12.85	0.066
1266	11.95	0.070
Composite Modes		
781	116.94	0.000

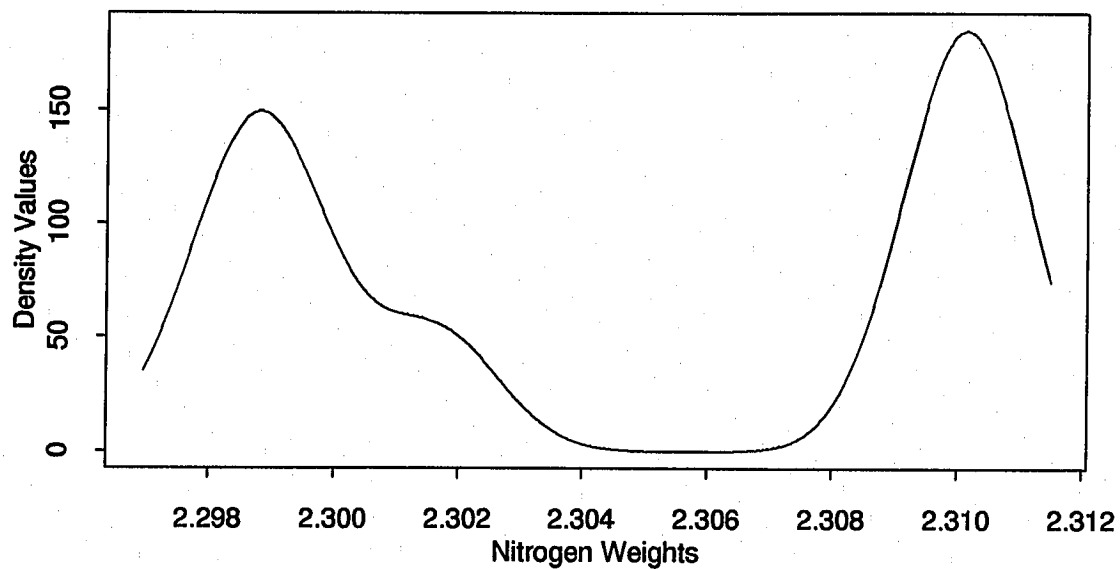
**Table 7.5** Modes of LRL mode tree with  $p$ -values less than 0.15.

(tested at  $h = 5.2$  for a  $p$ -value of 0.16). Considering the sheer number of tests performed, several false positives were a very real possibility. It is surprising that no other modes tested at less than 0.4, and there is little evidence to support the reality of these two.

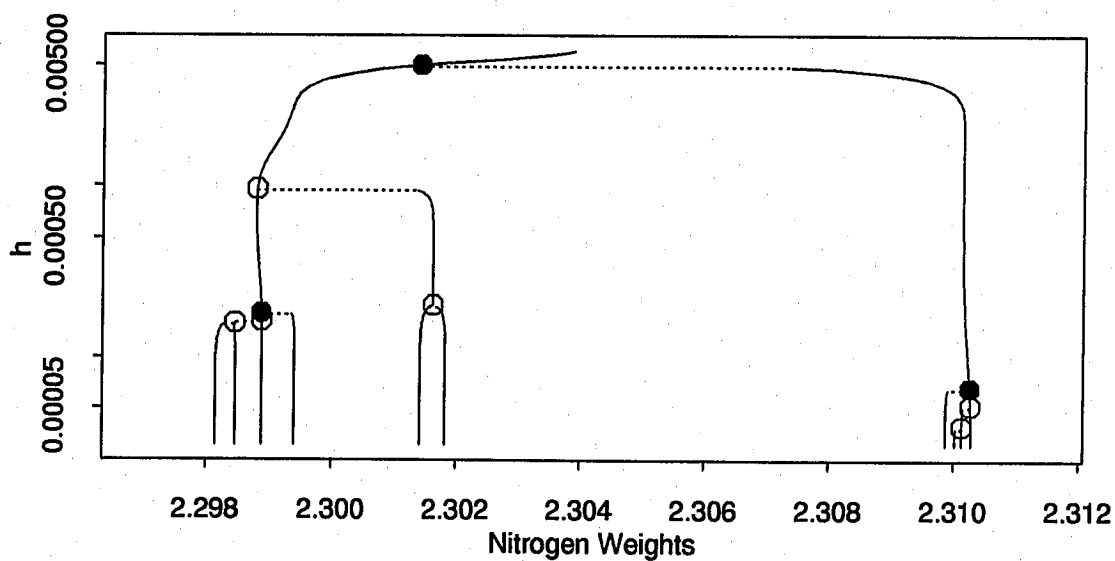
## 7.6 Rayleigh's Nitrogen Data

Our final example comes from Tukey (1977). On page 49, he lists the results of 15 experiments to measure the weight of Nitrogen conducted by Lord Rayleigh in 1893 and 1894. Seven of these experiments involved removing oxygen from air; the remaining eight involved decomposition of a nitrogen-containing compound. Each resulted in a value representing the weight (in grams) of a quantity of gas contained in a glass globe at a standard volume, temperature, and pressure.

In Figure 7.11, we show a kernel density estimate of the “nitrogen” data, under the assumption that all of the measurements come from a single distribution. Figure 7.12 contains the associated mode tree. Despite the small sample size, it indicates two highly significant modes which Table 7.6 verifies. This is comforting, given the ex-



**Figure 7.11** Kernel density estimates for the Raleigh “nitrogen” data;  $h = 0.001$ .



**Figure 7.12** Test mode tree for the Rayleigh “nitrogen” data. Filled circles are significant modes at the  $\alpha = 0.15$  level, open circles are modes which are not significant at this level.

Modes	$h$	$p$ -value
Single Modes		
2.299	0.000183	0.005
2.310	0.000067	0.003
Composite Modes		
2.301	0.004980	0.000

**Table 7.6** Modes of Rayleigh nitrogen mode tree with  $p$ -values less than 0.15.

treme nature of the data, and the fact that its obviously bimodal nature led Lord Rayleigh to the discovery of argon, for which he received the 1904 Nobel Prize in Physics!

This data set is so extreme that judging it bimodal is not terribly impressive. Nonetheless, it serves as a useful example of how even a very small sample (such as these 15 points) can produce very strong results ( $p$ -values of 0.0025 and 0.005), provided the evidence is sufficiently strong.

## Chapter 8

### Future Directions

The investigation so far into the potential of this new test for multimodality is only beginning. In this chapter, some of the areas which will be investigated in the near future are examined.

#### 8.1 Combined $p$ -values

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.

#### 8.2 Adjusted $p$ -values — Greater Power

The simulation results of Section 6.1 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 fall *on* the line  $x = y$ , rather than above it (given data drawn from a unimodal density). This would reduce the stated  $p$ -values, resulting in greater power, while keeping the  $p$ -values and  $\alpha$  levels meaningful.

### 8.3 Investigating Bumps

It appears to be relatively straightforward to modify the mode-existence test to investigate bumps as well as modes. The test statistic would be similar to  $M_i$ , but based on subtracting a linear value, rather than merely a constant one. The problem of how to most appropriately replace the mass for simulations could be difficult, and will likely be the aspect of this variant which proves most problematical. Assuming this difficulty can be overcome, the power of the resulting procedure is unlikely to be as great as for the modes, but the test would be useful nonetheless.

### 8.4 Multivariate Data

The extension of the test to the multivariate setting is conceptually simple, but fraught with technical difficulties. The idea behind the test translates into higher dimensions virtually unchanged, at least for modes; for bumps, even the multivariate definition is unclear.

Actual implementation of the test in multiple dimensions may prove difficult, however. The most basic problem under the current implementation is simply the question of binning. While 500 bins in one dimension seems quite satisfactory, the storage space and time required to manipulate the equivalent 250,000 bins in even two dimensions is likely to prove impractical, much less higher dimensions. The loss of information involved in reduction to 10 or so bins per dimension is likely to be intolerable, so, realistically, the test is likely to be restricted to at most two or three dimensions under the current setup. Other factors which will prove more difficult in more than one dimension will be the identification of modes and the determination of the level at which the modes may be cut off to be excised.

A related problem which might be worth investigation would be the investigation of ridges in two dimensions. As ridges are essentially long modes, the only difference

between a ridge test and the two dimensional mode test would be identification of the ridges and possibly the method of surgery.

## 8.5 Regression and Spectral Density Estimation

The basic ideas behind the mode-existence test could be extended to a number of other settings relatively easily. Heckman (1991) investigates the problem of bump hunting in the nonparametric regression setting and derives some asymptotic results for “modes” defined as  $k$  increasing estimation points followed by  $k$  decreasing estimation points.

It is clear that the nonparametric regression setting could be fertile ground for an extension of the mode-existence test. The same can be said for the area of time series spectral density estimation. In each case, a procedure based on the mode-existence test could be used to investigate the possibilities of features analogous to modes in density estimates. The primary differences between these tests and the current one will likely be their methods of resampling.

## 8.6 Adaptive Density Estimation

Finally, there is reason to hope that one might be able to 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.

The obvious choice seems to be to find the highest significant bandwidth at which each mode was tested, and use that bandwidth for all points between that mode’s adjacent antimodes (at that level). Unfortunately, this still leaves unclear what to do about points that do not fall in any such regions, or points in areas of “overlap” due to adjacent modes being tested at different bandwidths with the intermediate antimode moving toward the mode tested at the higher level as  $h$  approaches the lower mode’s test level.

Nonetheless, such difficulties do not seem insurmountable. The result could be an adaptive density estimate which would be extremely useful in showing the density as the mode-existence procedure indicates it, with just the significant modes at appropriate bandwidths for each. Such an estimate could be a valuable addition to the density estimation toolkit.



## References

- Abramson, I.S., (1982), "On Bandwidth Variation in Kernel Estimates—a Square Root Law," *The Annals of Statistics*, **10**, 1217–1223.
- Ahrens, L.H., (1965), "Observations on the Fe-Si-Mg Relationship in Chondrites," *Geochimica et Cosmochimica Acta*, **29**, 801–806.
- Besag, J., and Clifford, P., (1991), "Sequential Monte Carlo  $p$ -values," *Biometrika*, **78**, 301–304.
- Breiman, L., Meisel, W., and Purcell, E., (1977), "Variable Kernel Estimates of Multivariate Densities," *Technometrics*, **19**, 135–144.
- Carmichael, J.P., (1976), "The Autoregressive Method", unpublished Ph.D. thesis, State University of New York at Buffalo, Statistical Science Division.
- Comparini, A., and Gori, E., (1986), "Estimating Modes and Antimodes of Multimodal Densities," *Metron*, **44**, 307–332.
- Cox, D.R., (1966), "Notes on the Analysis of Mixed Frequency Distributions," *The British Journal of Mathematical and Statistical Psychology*, **19**, 39–47.
- Donoho, D.L., (1988), "One-Sided Inference about Functionals of a Density," *The Annals of Statistics*, **16**, 1390–1420.
- Efron, B., (1979), "Bootstrap Methods — Another Look at the Jack-knife," *The Annals of Statistics*, **7**, 1–26.
- Fisher, N.I., Mammen, E., and Marron, J.S., (1991), "Testing for Multimodality," Technical Report No. 629, Universität Heidelberg.
- 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), *Journal of the American Statistical Association*, **75**, 42–73.
- Haldane, J.B.S., (1952), "Simple Tests for Bimodality and Bitangentiality," *The Annals of Eugenics*, **14**, 309–318.
- Hartigan, J.A., and Hartigan, P.M., (1985), "The Dip Test of Unimodality," *The Annals of Statistics*, **13**, 70–84.

- Heckman, N.E. (1991), "Bump Hunting in Regression Analysis," in press.
- Izenman, A.J., and Sommer, C., (1988), "Philatelic Mixtures and Multimodal Densities," *Journal of the American Statistical Association*, **83**, 941-953.
- Jones, M.C., (1990), "Variable Kernel Density Estimates," *Australian Journal of Statistics*, **32**, 361-371.
- Mammen, E., (1991), "On Qualitative Smoothness of Kernel Density Estimates," Technical Report No. 614, Universität Heidelberg.
- Mammen, E., Marron, J.S. and Fisher, N.I., (1990), "Some Asymptotics for Multimodality Tests Based on Kernel Density Estimates," Technical Report No. 601, Universität Heidelberg.
- Matthews, M.V., (1983), "On Silverman's Test for the Number of Modes in a Univariate Density Function", unpublished B.A. Honors Thesis, Harvard University, Dept. of Statistics.
- Müller, D.W., and Sawitski, G., (1991), "Excess Mass Estimates and Tests for Multimodality," *Journal of the American Statistical Association*, **86**, 738-746.
- Parzen, E., (1962), "On Estimation of Probability Density Function and Mode," *Annals of Mathematical Statistics*, **33**, 1065-1076.
- Parzen, E., (1979), "Nonparametric Statistical Data Modeling," (with discussion), *Journal of the American Statistical Association*, **74**, 105-131.
- 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.
- Roeder, K., (1990), "Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies," *Journal of the American Statistical Association*, **85**, 617-624.
- Rosenblatt, M., (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics*, **27**, 832-837.
- Scott, D.W., (1980), Comment on Good, I.J., and Gaskins, R.A., "Density Estimation and Bump-Hunting by the Penalized Maximum Likelihood Method Exemplified by Scattering and Meteorite Data," *Journal of the American Statistical Association*, **75**, 42-73.
- Scott, D.W., (1985a), "Average Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions," *The Annals of Statistics*, **13**, 1024-1040.
- Scott, D.W., (1985b), "Frequency Polygons: Theory and Application," *Journal of the American Statistical Association*, **80**, 348-354.

- Scott, D.W., (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley.
- Silverman, B.W., (1978), "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives," *The Annals of Statistics*, **6**, 177-184.
- Silverman, B.W., (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society, 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 No. 79), eds. J.F.C. Kingman and G.E.H. Reuter, Cambridge, U.K.: Cambridge University Press, 248-259.
- Terrell, G.R., (1990), "The Maximal Smoothing Principle in Density Estimation," *Journal of the American Statistical Association*, **85**, 470-477.
- Terrell, G.R. and Scott, D.W., (1985), "Oversmoothed Nonparametric Density Estimates," *Journal of the American Statistical Association*, **80**, 209-214.
- Terrell, G.R. and Scott, D.W., (1992), "Variable Kernel Density Estimation," *The Annals of Statistics*, in press.
- Thaler, H., (1974), "Nonparametric Probability Density Estimation and the Empirical Characteristic Function", unpublished Ph.D. thesis, State University of New York at Buffalo, Statistics Department..
- Tukey, J.W., (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.
- Wahba, G., (1981), "Data-Based Optimal Smoothing of Orthogonal Series Density Estimates," *The Annals of Statistics*, **9**, 146-156.
- Wand, M.P., Marron, J.S., and Ruppert, D., (1991), "Transformations in Density Estimation," (with discussion), *Journal of the American Statistical Association*, **86**, 343-361.
- Wilson, I.G., (1983), "Add a New Dimension to Your Philately," *The American Philatelist*, **97**, 342-349.
- Wong, M.A., (1985), "A Bootstrap Testing Procedure for Investigating the Number of Subpopulations," *Journal of Statistical Computation and Simulation*, **22**, 99-112.
- Wong, M.A., and Schaack, C., (1985), "Diagnostics for Assessing Multimodality," in *Computer Science and Statistics: Proceedings of the Sixteenth Symposium on the Interface*, L. Billard, ed., 287-296.