ADAPTIVE DENSITY ESTIMATION BASED
ON THE MODE EXISTENCE TEST

by

Nizar Sami Jawhar

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

UTAH STATE UNIVERSITY
Logan, Utah
1996
ADAPTIVE DENSITY ESTIMATION BASED

ON THE MODE EXISTENCE TEST

by

Nizar Sami Jawhar

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Michael Minnotte
Major Professor

Daniel Coster
Committee Member

Ronald Canfield
Committee Member

James P. Shaver
Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah
1996
ABSTRACT

Adaptive Density Estimation
Based On The Mode
Existence Test

by

Nizar Sami Jawhar, Master of Science
Utah State University, 1996

Major Professor: Dr. Michael Minnotte
Department: Mathematics and Statistics

The kernel persists as the most useful tool for density estimation. Although, in general, fixed kernel estimates have proven superior to results of available variable kernel estimators, Minnotte's mode tree and mode existence test give us newfound hope of producing a useful adaptive kernel estimator that triumphs when the fixed kernel methods fail. It improves on the fixed kernel in multimodal distributions where the size of modes is unequal, and where the degree of separation of modes varies. When these latter conditions exist, they present a serious challenge to the best of fixed kernel density estimators. Capitalizing on the work of Minnotte in detecting multimodality adaptively, we found it possible to determine the bandwidth $h$ adaptively in a most original fashion and to estimate the mixture normals adaptively, using the normal kernel with encouraging results.

(70 pages)
ACKNOWLEDGMENTS

I would like to thank Dr. Michael Minnotte for his great ideas and bright intellect, and what that meant in achieving this work. His wonderfully helpful personality benefited me and gave me an experience few dare wish for. I would also like to thank each of my committee members, Dr. Ronald Canfield and Dr. Daniel Coster, my teachers and mentors, who embody the nobility of the intellect and the sacredness of the teaching occupation. I would like to thank Dr. Richard Cutler for his personality and all he taught me, and all my other teachers for their seldom matched competence. I warmly thank my father, Sami, and dedicate this work to him.

Nizar Sami Jawhar
# CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td></td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td></td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td></td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td></td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>DENSITY ESTIMATION</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>Parametric Methods</td>
<td>4</td>
</tr>
<tr>
<td>2.2</td>
<td>Nonparametric Density Estimation</td>
<td>5</td>
</tr>
<tr>
<td>2.2.1</td>
<td>The Histogram</td>
<td>6</td>
</tr>
<tr>
<td>2.2.2</td>
<td>The Naive Estimator</td>
<td>7</td>
</tr>
<tr>
<td>2.3</td>
<td>The Fixed Kernel Estimator</td>
<td>9</td>
</tr>
<tr>
<td>2.4</td>
<td>The Adaptive Kernel</td>
<td>9</td>
</tr>
<tr>
<td>2.4.1</td>
<td>The Balloon Estimators</td>
<td>10</td>
</tr>
<tr>
<td>2.4.2</td>
<td>The Sample Smoothing Estimators</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>THE MODE EXISTENCE TEST AND THE MODE TREE</td>
<td>13</td>
</tr>
<tr>
<td>3.1</td>
<td>The Mode-Existence Test</td>
<td>13</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Detecting Multimodal Distributions</td>
<td>13</td>
</tr>
<tr>
<td>3.1.2</td>
<td>The Mode Existence Test</td>
<td>14</td>
</tr>
<tr>
<td>3.2</td>
<td>The Mode Tree</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>A NEW VARIABLE KERNEL ESTIMATOR</td>
<td>17</td>
</tr>
<tr>
<td>4.1</td>
<td>Choosing “True” Modes</td>
<td>18</td>
</tr>
<tr>
<td>4.1.1</td>
<td>An Example in Mode Searching</td>
<td>19</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Modes by $h$</td>
<td>20</td>
</tr>
<tr>
<td>4.2</td>
<td>The Bandwidth Function $w(x)$</td>
<td>22</td>
</tr>
<tr>
<td>4.3</td>
<td>The Minnotte-Jawhar Variable Kernel</td>
<td>25</td>
</tr>
<tr>
<td>4.4</td>
<td>The Balloon Estimator</td>
<td>25</td>
</tr>
<tr>
<td>4.5</td>
<td>The Sample Smoothing Estimator</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>SIMULATIONS</td>
<td>28</td>
</tr>
<tr>
<td>5.1</td>
<td>Points of Focus</td>
<td>28</td>
</tr>
<tr>
<td>5.1.1</td>
<td>The Optimal Bandwidth Fixed Kernel</td>
<td>28</td>
</tr>
<tr>
<td>5.1.2</td>
<td>AVMISE</td>
<td>28</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Efficiency</td>
<td>29</td>
</tr>
</tbody>
</table>
5.2 Distributions Simulated ........................................... 30
5.3 The Sample Size .................................................... 33
5.4 The Mode Existence Test $\alpha$ Level ............................. 33
5.5 The $\alpha - n$ Connection .......................................... 34

6 RESULTS ........................................................................ 35
  6.1 AVMISE Report ......................................................... 35
  6.2 Distribution Effects .................................................. 35
  6.3 Results on the Sample Size Effects ............................... 36
  6.4 Results on the $\alpha$ Effect .......................................... 38
  6.5 Pointwise Confidence Intervals ...................................... 43

7 CONCLUSIONS .............................................................. 57
  7.1 Observations ........................................................... 57
  7.2 Future Directions ...................................................... 59

REFERENCES ................................................................. 60
LIST OF TABLES

6.1 AVMISE Results. .................................................. 36

6.2 Descriptions for Distributions in Table 6.1....................... 36
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The number of peaks increases as $h$ decreases.</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>The number of vertical lines increases with the fall of $h$.</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>The histogram.</td>
<td>6</td>
</tr>
<tr>
<td>2.2</td>
<td>The naive estimator.</td>
<td>8</td>
</tr>
<tr>
<td>2.3</td>
<td>The balloon estimator.</td>
<td>11</td>
</tr>
<tr>
<td>2.4</td>
<td>Sample smoothing estimator.</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>The mode tree for data sample from the trimodal distribution, $\alpha = 0.05$.</td>
<td>15</td>
</tr>
<tr>
<td>3.2</td>
<td>The mode tree for data sample from the trimodal distribution, $\alpha = 0.25$.</td>
<td>16</td>
</tr>
<tr>
<td>4.1</td>
<td>Mode-seeking example.</td>
<td>20</td>
</tr>
<tr>
<td>4.2</td>
<td>Modes by $p$ vs. modes by $h$.</td>
<td>21</td>
</tr>
<tr>
<td>4.3</td>
<td>Bandwidth assignment.</td>
<td>23</td>
</tr>
<tr>
<td>4.4</td>
<td>$h(x)$ vs. $h_{geo}(x)$.</td>
<td>24</td>
</tr>
<tr>
<td>4.5</td>
<td>Balloon estimates for the sample of size 40 from the trimodal density.</td>
<td>26</td>
</tr>
<tr>
<td>4.6</td>
<td>The sample smoothing estimate.</td>
<td>27</td>
</tr>
<tr>
<td>5.1</td>
<td>The $N(0,1)$ distribution.</td>
<td>30</td>
</tr>
<tr>
<td>5.2</td>
<td>Bimodal distributions under investigation. $\frac{1}{2}N(0,1) + \frac{1}{2}N(3,1)$.</td>
<td>30</td>
</tr>
<tr>
<td>5.3</td>
<td>Bimodal distributions under investigation. $\frac{1}{2}N(0,1) + \frac{1}{2}N(4,1)$.</td>
<td>31</td>
</tr>
<tr>
<td>5.4</td>
<td>Bimodal distributions under investigation. $\frac{3}{10}N(0,1) + \frac{7}{10}N(5,2)$</td>
<td>31</td>
</tr>
<tr>
<td>5.5</td>
<td>The trimodal distribution.</td>
<td>32</td>
</tr>
<tr>
<td>5.6</td>
<td>Asymmetric claw.</td>
<td>32</td>
</tr>
</tbody>
</table>
6.1 AVMISE decreases as \( n \) gets large. .............................. 37
6.2 M-J estimator’s success rate in correct modality detection. ......... 39
6.3 The efficiency and the sample size. ................................. 40
6.4 \( \alpha \) a smoothing parameter. .................................... 42
6.5 M-J estimator’s behavior with changes in \( \alpha \). ...................... 42
6.6 95% data-based pointwise confidence interval for \( N(0, 1) \). ... ... 45
6.7 95% data-based pointwise confidence interval for \( \frac{1}{2} N(0, 1) + \frac{1}{2} N(3, 1) \). ... 46
6.8 95% data-based pointwise confidence interval for \( \frac{1}{2} N(0, 1) + \frac{1}{2} N(4, 1) \). ... 47
6.9 95% data-based pointwise confidence interval for \( \frac{3}{10} N(0, 1) + \frac{7}{10} N(5, 2) \). ... 48
6.10 95% data-based pointwise confidence interval for the trimodal distribution. 49
6.11 95% data-based pointwise confidence interval for the asymmetric claw. ... 50
6.12 Pointwise 50% confidence interval (IQR) for \( N(0, 1) \). .............. 51
6.13 Pointwise 50% confidence interval (IQR) for \( \frac{1}{2} N(0, 1) + \frac{1}{2} N(3, 1) \). ... 52
6.14 Pointwise 50% confidence interval (IQR) for \( \frac{1}{2} N(0, 1) + \frac{1}{2} N(4, 1) \). ... 53
6.15 Pointwise 50% confidence interval (IQR) for \( \frac{3}{10} N(0, 1) + \frac{7}{10} N(5, 2) \). ... 54
6.16 Pointwise 50% confidence interval (IQR) for the trimodal distribution. ... 55
6.17 Pointwise 50% confidence interval (IQR) for asymmetric claw. ........ 56
CHAPTER 1

INTRODUCTION

Given a random sample $X_1 \ldots X_n$ from an unknown distribution, it is often of interest to investigate the underlying density function. Estimating the density function could be pursued following two classes of methods.

The class of parametric methods works well if the assumptions about the distribution being from a parametric family are true. In fact, parametric methods are superior to the other class of methods if the parametric assumptions are met.

In the case where no prior knowledge exists about the distribution the random sample came from, nonparametric methods become useful and quite appropriate. Nonparametric methods need minimal assumptions about the underlying distribution, and we can always count on the empirical density to converge to truth given a large enough sample size.

A superior nonparametric method is kernel density estimation. For the random sample $X_1 \ldots X_n$, the kernel density estimate of the density function at $x$ is

\begin{equation}
\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),
\end{equation}

where $K$ is a density function centered about 0, and $K_h(y) = \frac{1}{h} K\left(\frac{y}{h}\right)$. As the formula suggests, the kernel estimate is a sum of kernels centered at each observation. Choosing the optimal bandwidth $h$, also called the smoothing parameter, persists as the most critical issue in kernel density estimation. A nonadaptive kernel estimator would keep the bandwidth $h$ constant across all points of estimation. Adaptive kernel estimation varies the bandwidth with the point of estimation or with the data points. The bandwidth controls
Figure 1.1: The number of peaks increases as $h$ decreases. From top to bottom, $h = 2$, $h = .88$, $h = .4$

the number of peaks (possible modes) that appear in the estimate.

In Figure 1.1, we plot three fixed kernel density estimates of the density function from a sample of size 40, at three different values of $h$. The sample came from a normal mixture $\frac{1}{5} N(4, 1) + \frac{1}{5} N(8, 1) + \frac{3}{5} N(20, 25)$. As the plot indicates, the number of modes in the estimates gets larger as $h$ gets smaller. That last fact is beautifully evident in Minnotte and Scott’s (1993) mode tree. Estimating the density function for numerous values of $h$, the mode locations are plotted against the bandwidth. It is obvious in Figure 1.2 how the number of modes detected increases with the decrease in the value of $h$. Also Minnotte
Figure 1.2: The number of vertical lines in creases with the fall of $h$. Filled circles indicate significant modes $\alpha = 0.25$.

devised the mode existence test, in which he tested apparent modes at a critical bandwidth $h$, using a new statistic (see Section 3.1), and recorded the results of the test on the mode tree, indicating significance with a filled circle. In the continuing search for improvements on the fixed kernel density estimate, through varying the bandwidth over the points of estimation, we found that the results of the mode existence test and the mode tree can be effectively and successfully used to choose a selection of bandwidths $h$ to produce an acceptable adaptive kernel density estimate $\hat{k}$ that is global, smooth, and continuously differentiable, and which integrates to 1. This new adaptive kernel demonstrated itself to be very useful, beating the fixed kernel calculated using the optimal bandwidth in some tricky distributions, and doing very well in most cases. Extensive simulation studies on mixture normals indicate the birth of a useful variable kernel density estimator. The averaged integrated mean squared error (AVIMSE) was the criterion we used in deciding the worth of this new estimator, with encouraging results (see Chapters 5 and 6).
CHAPTER 2

DENSITY ESTIMATION

The probability density function $f$ lies at the heart of the mathematics of statistics. It is often the case that an estimate of the density function $f$ is essential to the researcher, such as the investigator of multimodality or any other feature of interest. Density estimation can be a goal by itself or it could be a helpful tool to help the investigator go further. Given a random sample $X_1, X_2, \ldots, X_n$ from an unknown distribution, an estimate of the probability density function $\hat{f}(x)$ could be constructed from the data, making only a few assumptions. This is density estimation. The methods of estimations form two classes:

1. The class of parametric methods
2. The class of nonparametric methods.

2.1 Parametric Methods

Parametric methods are so named because the investigator assumes the random sample at hand comes from a particular parametric family. In this case the problem of estimating the probability function $f(x)$ is reduced to estimating the population parameters from the data set $X_1 \ldots X_n$ by looking for estimators that are hopefully unbiased, estimable, and of minimal variance. In case we suspect the presence of outliers or contamination, we would like our estimator to be robust. We usually look for sufficient or complete statistics or a linear combination of sufficient statistics, such as the sample average, the trimmed mean and the median. For example, suppose we can assume that a random sample $X_1 \ldots X_n$ is coming from a normal distribution. To estimate the density function we would estimate
\(\mu\) by

\[\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i\]

and we would estimate \(\sigma^2\) by

\[\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.\]

Then our density estimate would be

\[f(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}}.\]  

\[\hat{F}_n = F.\]

It has been said that "every multivariate density estimator that is in any reasonable sense nonparametric may be written in the form

\[\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} K_n(x_i, y),\]

where \(K_n\) is asymptotically a Dirac evaluation functional at \(y\)" (Terrel and Scott 1992, p.1238). We will see that all the following estimators conform to the above statement.
Figure 2.1: The histogram. Plot is for the sample of size 40 discussed in chapter 1, with bin width 5 and 3.

2.2.1 The Histogram

One of the oldest nonparametric methods of estimating $f$ is the histogram. We choose a bin width and make the bin height proportional to the number of observations that fall within that bin. Assuming the density function $f$ exists, the histogram estimate of any density function is

\[
\hat{f}(x) = \frac{1}{nh} \times \text{(No of data points in same bin as } x)
\]

A histogram representing that estimate for a sample of size 40 of the distribution described in Chapter 1 is in Figure 2.1. Notice how the estimate to the right, which uses a bin width of 3, is much rougher and shows a lot more detail than the histogram to the left at the larger bin width of 5. The choice of bin width $h$ is left to the researcher. Choosing the optimal $h$ is the key to a histogram that is close to the truth. The bin width $h$ is called the smoothing parameter. If $h$ is chosen too small, the number of peaks, which are usually
construed as modes, increases in the histogram, which will be larger than the number of modes characteristic of the distribution the sample came from. With large values of \( h \), the features of the histogram are oversmoothed. If too large of a value is chosen for \( h \), most features of the density will be obscured. If \( h \) is large enough, the estimate will then have one "fat" mode. Bin widths do not all have to be equal. We might have reasons to partition the range of the data into bins of different widths. We usually do that when we suspect a different amount of smoothing is required in different subsets of the range of estimation. In this case we can write in more general terms

\[
\hat{f}(x) = \frac{1}{nh_x} \times \text{(No of data points in same bin as } x) \tag{2.4}
\]

where \( h_x \) is the bin width of the bin \( x \) falls into. Allowing the bin width to vary within the same histogram is an example of variable bandwidth density estimation. The histogram as a nonparametric estimator could be written in the form

\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{hist}(x, X_i) \tag{2.5}
\]

by letting

\[
K_{hist}(x, t) = \begin{cases} 
\frac{1}{h_x} & \text{if } t \text{ in same bin as } x \\
0 & \text{otherwise}
\end{cases}
\]

2.2.2 The Naive Estimator

For any given \( h \), to estimate the probability that \( (t - h) \leq X \leq (t + h) \), Silverman defined

\[
w(t) = \begin{cases} 
\frac{1}{2} & \text{if } |x| < 1 \\
0 & \text{otherwise}
\end{cases}
\]

and wrote the naive estimator as

\[
\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h}\right)
\]
(Silverman 1986). This estimator works by placing a box of width $2h$ and height $(2nh)^{-1}$ on each observation and summing. In Figure 2.2 we plotted the naive estimator for the sample in Figure 2.1. The bandwidth value is 1 and 0.5. Note the stepwise nature of the plots. The naive estimator, then, works like a histogram, but is superior in the sense that the naive estimator is freed from the choice of bin origin by considering every point the center of a bin. Both methods are very useful, yet not satisfactory. Both estimates are discontinuous, due to their stepwise nature. That leads us to generalize the naive estimator and consider the kernel estimator.

### 2.3 The Fixed Kernel Estimator

To overcome the discontinuities of the naive estimator, consider replacing $w$ by a kernel function $K$ such that

$$\int_{-\infty}^{\infty} K(x) dx = 1.$$
Usually $K$ will be symmetric and centered around 0. Then the fixed kernel estimator is written as

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

The smoothing parameter $h$ is called the bandwidth. The kernel estimate of $f$ is continuous and infinitely differentiable if $K$ is, which renders the kernel estimate far superior to the naive estimator and the histogram. The lack of adaptivity of the fixed kernel, where the same bandwidth is used for all points of estimation, makes the fixed kernel undesirable in cases where the modes differ in size and degree of separation. In the case where the sample comes from a population with a distribution like the one featured in Figure 1.1, we want to use a smaller bandwidth to detect the “valley” between the two first taller modes than the bandwidth we need to detect or estimate the density around the “fatter” smoother third mode. By the time the bandwidth is small enough to detect the various modes, spurious modes erupt in the tails and at locations where there is minor random clustering of the data. Figure 1.1 illustrates that fact clearly.

### 2.4 The Adaptive Kernel

While the amount of smoothing controlled by $h$ is the same across all points of estimation in the fixed kernel estimate, a global bandwidth $h$ might not be optimal for all points of estimation. The variable bandwidth (adaptive) kernel varies $h$ in hopes of improving on the fixed kernel. Varying $h$ with the point of estimation was called by Terrell and Scott (1992) the *balloon estimator*, while they called the kernel that varies $h$ with the data points, the *sample smoothing estimator*. In Figures 2.3 and 2.4, we plotted two adaptive estimates, via two different methods, by using the bandwidth value 0.5 for points below
11 and the bandwidth value 2 for points above 11. It is obvious the estimates in Figures 2.3 and 2.4 are a lot more accurate than the fixed kernel estimate at any of the global bandwidths in Figure 1.1.

2.4.1 The Balloon Estimators

A balloon estimator of the probability function at $x$ is written as

$$
\hat{f}(x) = \frac{1}{nh_x} \sum_{i=1}^{n} K\left( \frac{X_i - x}{h_x} \right). 
$$

(2.7)

Loftsgaarden-Quesnberry-style estimators are examples of a balloon estimator. They are the generalized $k$th nearest neighbor estimators, which are defined by

$$
\hat{f}(x) = \frac{1}{nd_k(x)} \sum_{i=1}^{n} K\left( \frac{x - X_i}{d_k(x)} \right)
$$

(2.8)

where $d_k(x)$ is the distance to the nearest data point to $x$. The advantage of this estimator is a straightforward asymptotic analysis (Mack and Rosenblatt 1979). However, when used in a global fashion, the estimate does not integrate to one even if $K$ does, and thus is not in itself a density function. Terrell and Scott (1992) investigated this class of estimators, and found them to offer very modest improvement on the fixed kernel methods. Their criterion for comparison was taken to be the asymptotic mean square error (AMSE) at a single point of estimation. They also found these estimators to do poorly both asymptotically and with reasonably small sample sizes, in the univariate case. In the multivariate case, however, these estimators were found to be advantageous. We restrict our research to the univariate case. Notice how the first derivative of the balloon estimate in Figure 2.3 is not continuous and when summed does not sum up to 1.
2.4.2 The Sample Smoothing Estimators

The sample smoothing estimator is written as

\[ \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i(X_i)} K(\frac{X_i - x}{h(X_i)}). \]  

An example of a sample smoothing estimator is the adaptive kernel estimate of Breiman, Meisel, and Purcell (1997) defined as

\[ \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K(\frac{x_i - y}{h_i}), \]

where \( h_i \) is the Euclidean distance from \( x_i \) to the \( k^{th} \) nearest other data point. It is a mixture of individually scaled, yet identical, kernels centered at each observation. It has an advantage of being a density function any time \( K \) is. The main disadvantage is that points that are far away from the point of estimation may contribute significantly to the estimate. Figure 2.4 accentuates the difference in smoothness and continuity from Figure 2.3. Simulations conducted by Terrell and Scott (1992) showed good behavior for small to
Figure 2.4: The sample smoothing estimator. We used the $h$ values of figure 2.3

moderate sizes, and a deteriorating performance of the estimator in comparison to fixed kernel estimates as the sample size increased. The search for a useful adaptive density estimate is still on, and is exactly what this project is all about.
CHAPTER 3

THE MODE EXISTENCE TEST AND THE MODE TREE

3.1 The Mode-Existence Test

3.1.1 Detecting Multimodal Distributions

A lot of work has been done to investigate the possibility that a data set at hand comes from a multimodal distribution. The number of modes, the reality of each mode, and their locations and sizes are pieces of information of great interest to researchers. Silverman (1983) shows that for a normal kernel, the number of maxima of \( \hat{f} \) and all derivatives is a decreasing right continuous function of the bandwidth \( h \). Basically we can say that the number of modes increase as \( h \) decreases, a phenomenon familiar to any user of the histogram or any other method requiring a choice of window-width. Silverman's "Critical Kernel-Bandwidth Test" tests the hypothesis that a density has at most \( k \) modes, and computes p-values for his statistic nonparameterically, using bootstrap samples. The statistic used is

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

where \( \hat{f}(\cdot, h) \) is the normal density kernel estimate of \( f \) with bandwidth \( h \). Silverman (1983) further claims and demonstrates the consistency of this test, when a few conditions are met. The problem with the test is its lack of adaptivity, which renders it weak in finding modes of varying sizes and degrees of separation, as the alpha-level has to be close to 0.4. Moreover, it does not find the locations or sizes of the modes, just the number. The Maximum Penalized-Likelihood method of Good and Gaskins (1980) hunts for bumps
and modes in a local fashion. This test proved to have low power also. By combining the above-mentioned two procedures in their relevant aspects, Minnotte devised a test called the mode existence test (Minnotte 1992), which tests individual modes at location \( x \), which is adaptive and more powerful under less than optimal conditions.

### 3.1.2 The Mode Existence Test

Minnotte (1992) investigated multimodality in an adaptive fashion. He derived a mode existence test, with which he tested the hypothesis

\[ H_0 : \text{The mode detected at location } x \text{ is an artifact of the sample} \]

against the alternative

\[ H_a : \text{the mode detected at location } x \text{ is a true feature of the population} \]

Every mode was tested individually at the smallest bandwidth \( h \) for which the mode is still a single entity, using the statistic

\[
M_j(h) = \int_{a_j(h)}^{b_j(h)} \left[ \hat{f}_h(x) - \max(\hat{f}_h(a_j(h)), \hat{f}_h(b_j(h))) \right]_+ dx,
\]

where \( a_j(h) \) and \( b_j(h) \) are the left and right antimodes surrounding mode \( j \) in \( \hat{f}_h(x) \). \( M_j \) was recomputed from bootstrap resamples to obtain an empirical p-value (Minnotte 1992). This critical bandwidth coincides with Silverman's critical bandwidth. This coincidence is not by chance, as it allows Minnotte to make use of already established theory. There are other reasons for this choice of \( h \). One is that the choice of \( h_{crit} \) is determined by the data. Another is that this choice of \( h_{crit} \), gives the largest value of the test statistic, which makes the test more powerful. Based on results in Mammen, Marron, and Fisher (1992), Minnotte demonstrated that in the case of unimodal and bimodal densities the test statistic of the first two modes tested converge to zero and the true density, respectively.
3.2 The Mode Tree

Minnotte and Scott (1993) developed a most useful tool called the mode tree. After computing a fixed kernel density estimate for a large number of values (200) of the bandwidth \( h \), they plotted all mode locations against the values of \( h \), which were chosen to be equally separated on a logarithmic scale. In Figures 3.1 and 3.2 we display two mode-trees for the same sample we have been using, at the two \( \alpha \) levels of 0.05 and 0.25. The logarithmic scaling was deemed necessary, "as large changes at high values of \( h \) have less of an effect on the density estimate than smaller changes at lower values of \( h \)" (Minnotte, 1992, p.21). The solid vertical lines denote the modes. The graph obtained is desirably continuous, due to the use of the normal kernel. Other kernels do not share the continuity factor of the normal kernel. By looking at the tree, the poorest of observers would see how as the bandwidth \( h \) decreases, the number of modes increases. Each mode seems to split at a certain threshold value of \( h \). Minnotte conducted his mode existence test at that value of \( h \) at which the mode is still one entity, and recorded the test results on the mode.
Figure 3.2: The mode tree for data sample from the trimodal distribution, $\alpha = 0.25$.

tree (as a filled circle if $p < \alpha$).
CHAPTER 4

A NEW VARIABLE KERNEL ESTIMATOR

Recall that given a sample $X_1 \ldots X_n$, the balloon estimate at $x$ of $f(x)$ is

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

while the sample smoothing estimate at a point $x$ of $f(x)$ by

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

We have investigated two new variable kernel estimators, a sample smoothing estimator and a balloon estimator. As Minnotte has used the normal kernel in his test and mode tree production, we will use the normal kernel in our estimators. We favor the normal kernel for its continuity and nature as a density function.

Given a sample of size $n$, the mode tree supplies a wealth of information about modes and their behavior as $h$ varies. When using the normal kernel to produce the tree, the plot is continuous. The tree indicates for any value of $h$, the number of modes, their locations, and if at that value of $h$ a mode is splitting. Also the tree tests the mode for significance, the null hypothesis $H_0$ being that the mode is not a true feature of the distribution, and reports the p-value of the test. The task of deciding how to use this information in deciding which modes to "believe," and then choosing the bandwidth vector $h$ to adaptively estimate the density function in an efficient useful manner, is the main project in this thesis.

In the next section, we describe our methodology for using the mode tree to choose the modes that we will treat as "true." In Section 4.2, we define the function $w(x)$, which
assigns the bandwidth to the points of estimation. When that is done we can forge forward and declare our estimators, which will be investigated in later chapters.

4.1 Choosing “True” Modes

In fixed kernel density estimation, a large enough value of the bandwidth $h$ results in a unimodal estimate. From Figure 4.1 we can see how as $h$ decreases, the one “fat” mode begins to split, until it actually splits into two modes at a critical $h$. As $h$ continues to decrease, one and later both newborn modes split, each into two new modes and each at a different critical $h$, and so on. As $h$ goes to 0, the number of modes approaches $n$. Each mode that splits becomes a father mode and has therefore two branches, each of which in turn will split and become a father mode, which in turn has two branches and so on. So, at the top of the tree is the first ancestor mode, while at the bottom of the tree, where the value of $h$ is very small, we find the offspring modes. Moving up the mode tree, we look for modes that have been tested and have a $p$-value less than $\alpha$. As we climb up the mode tree, we abruptly stop when we come to a father mode that has at least one significant offspring in at least two branches. That means this mode has split into two significant modes. Here we explored two strategies.

1. *Modes by $h$:* Climbing back down along both branches of the father mode, we pick to be true, from each branch, the mode that has the largest value of $h_{\text{crit}}$ at which it was tested regardless of the result of the test. This method allows $h$ to be as large as possible for a smoother estimate.

2. *Modes by $p$:* In this approach and backing up along both branches of the father mode, we pick from each branch the first mode we encounter with a $p$-value less than $\alpha$. 
4.1.1 An Example in Mode Searching

We can make things clearer around these techniques by following an example. In Figure 4.1 we have a mode tree. We start from the bottom of the tree.

Look at the mode indicated by the tree around the coordinates (3,0.1) with an empty circle. This mode will be assigned a status of 0 because it has no significantly tested modes in its (rather short) branch extensions below it. The same is true for modes tested at the coordinates (4.2,0.24), (16.7,0.1), (16.7,0.16), (18,0.5), (8,0.4), (22.5,0.9), and (27,0.9).

Modes at locations (3.7,0.45), (8.7,0.46), and (18,0.75) get a status of 1. Even though they have no significant offsprings, or in other words, modes that have status greater than 0 in their branches below, they themselves are significant and get a status of 1 more than 0, 0 being the sum of significant modes in the branches.

The mode at location (18,0.99) gets a 1 also, as the significant mode below it is seen as the same mode tested significant at both bandwidths, yet there are no indications that it has split into two different modes because there are no significant modes in the branches splitting off this mode. The father mode at coordinates (4.5,1.97) gets a status of 2, the sum of the status of the modes in its branches. This is the mode we stop at, because it has a status bigger than 1, and back up along both branches.

In the case of the mode at (4.5,1.97) with a status of 2, the decision taken by both methods would be to consider the true modes at (3.7,0.42) and (8.7,0.46). But the decision is not the same, in the case of the mode that comes at location (18,1.86). That mode gets a status of 1 since it is related to significant modes along only one branch. So we keep going up the tree until we reach the ancester mode about location 14 at bandwidth of 5.8. That mode gets a status of 3, the sum of the status of the modes in both branches. This
Figure 4.1: Mode-seeking example. Modes not marked with a number are assigned the status 0.

is where we back up down the branch and choose the true mode differently using both methods. While method 1 (modes by $h$) will stop backing up when it reaches the tested mode at location (18,1.86), the first mode tested it encounters, not caring about the test result, method 2 (modes by $p$) continues descending down the bandwidth scale until it reaches the first significant mode at location (17.6,0.99). The end result is three modes at locations 3.7, 8.7, and 18 and bandwidth 0.42, 0.46, and 1.86, respectively, when using method 1 and three modes at locations 4.1, 8.7, and 18 and bandwidth 0.42, 0.46, and 0.99 when method 2 is used. The estimates from these two methods are in Figure 4.2.

4.1.2 Modes by $h$

In our final density estimator, we decide to follow method 1 (modes by $h$ approach), as the other method was less successful in estimating the density. Figure 4.2 is an example of the difference in behaviors of two estimates produced by each method, using the same
Figure 4.2: Modes by $p$ vs. modes by $h$. Two estimates produced following the two different methods.

sample. The estimate from the first method (to the right) is smoother than the one obtained from the second method (to the left), as one might have to go farther down the branch to locate a significant mode. Recall that $h$ decreases down the tree, which will make the critical bandwidth $h$ of the modes that are farther down the tree, smaller, causing the estimate to be rougher and almost always less accurate.

Let mode $j$ be the $j^{th}$ mode chosen to be true, and let $h_j$ be the critical bandwidth at which the $j^{th}$ mode was tested. Having chosen the modes, we now have available to us:

1. Mode locations.
2. Antimode locations.
3. The critical bandwidth $h_j$ at which they were tested.
The mode with the smallest location $x$ will have as the left antimode $-\infty$, and modes with the largest location $x$ will have $+\infty$ as the right antimode.

4.2 The Bandwidth Function $w(x)$

What is new about our estimator is its method of assigning bandwidths $h$ to each point of estimation (in the balloon case) or data point (in the sample smoothing case). Once the modes to be true are identified, we proceed with the bandwidth choice. This latter task involved many trials and different approaches. We started by reasoning that if a point of estimation or a data point, $x$, on the real line falls between the left antimode $a_j$ and the right antimode $b_j$ of a given mode $j$, then that point $x$ takes on the critical bandwidth value $h_j$, at which the mode containing $x$ was tested. Now two problems appear:

1. What to do with points not falling under any mode chosen.
2. What to do with points falling under more than one mode.

We tried a few ways to solve both problems. One way was to consider $x$ under the mode adjacent to it to the left. A second was considering it falling under the adjacent mode to the right. A third was assigning it a bandwidth that is the average of the two modes adjacent to it. The difference in estimates resulting from those choices was little. Finally we decided to assign $x$ the bandwidth $h_j$ of the father mode of the two modes sharing $x$ in problem 1, and the bandwidth $h_j$ of the father mode of the two modes surrounding $x$ in problem 2. The value of $h_j$ of the father mode of course is larger than the value of $h_j$ for both offspring modes. We define the function $h(x)$ as

$$h(x) = \begin{cases} h_j & \text{if } a_j < x < b_j \\ h_{father} & \text{if } a_j < x < b_j \\ & \text{for } j = k \text{ and } j = l \text{ with } k \neq l \\ & \text{or if } \not{j} j \\ & \text{such that } a_j < x < b_j \end{cases}$$
$h_{father}$ is the critical bandwidth at which the father of the two modes containing $x$, or the father of the two modes surrounding $x$ without containing it, was tested. In Figure 4.3 we have added to the tree in Figure 4.1 a horizontal dashed line at the modes chosen by the modes by $h$ method starting at the left antimode and ending at the right antimode. Points greater than 12.55, the left antimode of the mode at location 18 and tested at $h = 1.82$, will get assigned the bandwidth value of 1.82. The right antimode was extended to infinity as discussed at the end of Section 4.1. Points less than 4.2, the right antimode of the mode at location 3.7 tested at the $h$ value of .42, will be assigned a bandwidth of .42. Points between 5.9 and 12.2, the left and right antimodes of the mode at location 8.7 tested at bandwidth .46, will take on the $h$ value of .46. Now we can see that points between 4.2 and 5.9 are not covered by any mode. Then according to the methodology described above, we assign those points the bandwidth of the father mode of the two modes surrounding it, the mode marked by a status value of 2, which was tested at bandwidth 1.97.
Figure 4.4: $h(x)$ vs. $h_{geo}(x)$. Estimate with $h(x)$ is the solid line. The dashed line is the estimate with $h_{geo}$. Dotted line is the true density.

Estimates with this function $h(x)$ produced unwanted shoulders, that we found could be remedied by a slight change in the function $h(x)$. The new function $h_{geo}$ is defined as

$$h_{geo}(x) = \sqrt{h_{j}*h_{father}},$$

$h_{father}$ being the $h_{j}$ of the father of the mode considered containing $x$. Simply, $h_{geo}$ is the geometric mean of the critical bandwidth for the mode considered containing $x$, and the critical bandwidth of its father mode.

In the unimodal case, where we only detect one mode, we found that choosing

$$h_{os}(x) = 1.53 \left( \frac{1}{\sqrt{2\pi}} \right)^\frac{1}{2} n^{-\frac{1}{5}}$$

performed better as producing smoother estimates. This oversmoothed bandwidth $h_{os}$ was suggested by Scott (1992). Figure 4.3 illustrates the shoulder effect well of the $h(x)$ function, and demonstrates how $h_{geo}(x)$ solved the problem. In addition, $h_{geo}(x)$ proved to have a smaller AIMSE than $h(x)$.
We call the conclusion of our search, the bandwidth function \( w \) and we define it by

\[
(4.3) \quad w(y) = \begin{cases} 
    h_{geo}(y) & \text{if we detected multimodality} \\
    h_{so}(y) & \text{if only 1 mode is detected.}
\end{cases}
\]

4.3 The Minnotte-Jawhar Variable Kernel

Given a sample of size \( n \), \( X_1 \ldots X_n \), we produce a mode tree to detect the modes locations and record the significance level of the tests conducted on apparent modes at their respective critical bandwidths. Any distribution will have at least one mode and at most \( n \) modes. Making use of the mode tree information, following the methods described in the previous section, we can arrive at the number and locations of modes we deem to be "true modes." Also available to us are the locations of the left and right antimodes of those modes. We define and pronounce our balloon estimator at \( x \) of \( f(x) \) as

\[
(4.4) \quad f_b(x) = \frac{1}{nw(x)} \sum_{i=1}^{n} K \left( \frac{X_i - x}{w(x)} \right)
\]

and our new sample smoothing estimator

\[
(4.5) \quad f_s(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{w(X_i)} K \left( \frac{X_i - x}{w(X_i)} \right)
\]

where \( K \) is the normal kernel, and \( w(x) \) as described by Equation 4.3.

4.4 The Balloon Estimator

Recall that given a sample \( X_1 \ldots X_n \), the balloon estimate at \( x \) of \( f(x) \) is

\[
(4.4) \quad f_b(x) = \frac{1}{nw(x)} \sum_{i=1}^{n} K \left( \frac{X_i - x}{w(x)} \right)
\]

Terrell and Scott (1992) have looked closely at the family of density estimators that attempt to vary the value of \( h \) with the point of estimation and compared its asymptotic
AMSE to that of the fixed kernel. They concluded that the adaptive balloon estimator is at least as good as fixed kernel estimates asymptotically. But upon investigation of the available balloon estimators, they found that they appear to be of little to no use unless dealing with a multivariate density. We suspected our balloon estimator would conform to these findings. Even though the performance of the balloon estimator was acceptable by some measures, its unfavorable properties of not being continuously differentiable and not a density function in itself (as it does not integrate to 1) led us to quickly abandon this estimator and turn all our attention to the sample smoothing estimator. Examples of the performance of the balloon estimator, on which the samples of Figures 4.1 and 4.3 were based, are in Figure 4.5. The choice of bandwidth in Figure 4.5 was decided by our algorithm detailed in Section 4.2, and the value of the bandwidths assigned the various data points and points of estimation was as mentioned towards the end of that section.

4.5 The Sample Smoothing Estimator

We find the sample smoothing estimate at a point $x$ of $f(x)$ by

$$f_s(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{w(X_i)} K\left(\frac{X_i - x}{w(X_i)}\right)$$
where $K$ is the standard normal density function. In general these estimators have many favorable properties, such as continuity of the first derivative, which leads to a smoother estimate, and the fact that the estimator itself is a density function. The sample smoothing estimate for the same sample in Figure 4.5 is in Figure 4.6. We expected this estimate to do well, and we were not disappointed. In fact, the estimator performed up to the expectations we had and especially in the areas of concern to the fixed kernel. The estimates produced were continuous and numerically summed up to 1. Terrell and Scott (1992) investigated the usefulness of the sample smoothing estimator and found it to be ideal and superior to the fixed kernel in small sample sizes ($n < 40$), and where the modes are of unequal sizes and with different degrees of separation. But they found it not as efficient as the fixed kernel in large sample sizes.
CHAPTER 5
SIMULATIONS

5.1 Points of Focus

5.1.1 The Optimal Bandwidth Fixed Kernel

In their paper “Exact Mean Integrated Squared Error,” Marron and Wand (1992), when investigating the class of mixture normals, derived the optimal value of $h$, which minimizes $\text{AMISE}(h)$, given by the equation

$$\text{AMISE}(h) = n^{-\frac{1}{2}} h^{-1} \int K^2 + \frac{1}{4} h^4 \mu_2^2(K) \int [f'']^2$$

for $K$, a positive second-order kernel, and for $f$ having a continuous second derivative. In Equation 5.1, $\mu_2(K)$ is the second moment of the density function $K$. The $h$ that minimizes $\text{AMISE}(h)$ is

$$h_{\text{AMISE}} = \left[ \int K^2 \mu_2^2(K) \int [f'']^2 n \right]^{1/5}$$

$h_{\text{AMISE}}$ approximates $h_{\text{MISE}}$, which minimizes $\text{MISE}(h)$.

As all of our simulations are going to involve mixture normals, we compared our estimator to the $h_{\text{AMISE}}$ fixed kernel estimates. We refer in this thesis to $h_{\text{AMISE}}$ as the optimal fixed kernel bandwidth.

5.1.2 AVMISE

We calculated the averaged mean integrated square error (AVMISE) for our estimator and the optimal bandwidth fixed kernel. The AVMISE of an estimate $\hat{f}$ is given by
Equation 5.3

\[ \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{m} \sum_{j=1}^{m} (\hat{f}_{ij}(x) - f(x))^2 \right) \]

where \( \delta \) is the distance between any two points of estimation, \( m \) is the size of the vector estimated, and \( N \) is the number of simulation replications at the sample size \( n \). These averages are calculated for our estimator and for the optimal bandwidth fixed kernel, at each of the sample sizes and \( \alpha \) levels. Small values of the AVMISE for our estimates are highly desirable.

5.1.3 Efficiency

We will use a ratio of efficiency to compare our estimator to the fixed kernel. That ratio is \( OAVMISE/AAVMISE \), the ratio of the AVMISE of the optimal bandwidth kernel estimate to the AVMISE of our adaptive estimate. We will refer to this ratio as the efficiency of our estimator. Large values of the efficiency ratio indicate some success at outperforming the fixed kernel.

The expression in Equation 5.2 depends on the density function \( f \), which is usually unknown and thus is unattainable until a pilot search is conducted, and even then one can often be hopelessly guessing. When we are faced with a random sample from an unknown distribution, calculation of Equation 5.1 and Equation 5.2 is not within reach. Without any knowledge of the distribution underlying the sample, our estimator, on the average, reaches towards that smallest value that any fixed kernel can realize, and outperforms the fixed kernel at the optimal theoretical bandwidth \( h_{AMISE} \), especially when the sample size is small and where the distribution requires an adaptive technique. No prior knowledge of the distribution underlying the sample is needed.
5.2 Distributions Simulated

After implementing our strategy described extensively in the earlier chapter, we went ahead with an extensive plan of simulations. We chose our distributions based on various levels of difficulty of estimation. The distributions simulated are indicated in Figures 5.1-5.6.

Figure 5.1: The $N(0,1)$ distribution.

Figure 5.2: Bimodal distributions under investigation. $\frac{1}{2}N(0,1) + \frac{1}{2}N(3,1)$. 
Figure 5.3: Bimodal distributions under investigation. $\frac{1}{2}N(0,1) + \frac{1}{2}N(4,1)$.

Figure 5.4: Bimodal distributions under investigation. $\frac{3}{10}N(0,1) + \frac{7}{10}N(5,2)$.

The unimodal choice is obvious, as we want to detect unimodality as well as multimodality accurately. Detecting unimodality accurately and efficiently with respect to the fixed optimal bandwidth kernel indicates a "smart" kernel. A kernel that does not require a knowledge of the distribution and its parameters and rates of the mixture to decide based upon that a bandwidth of estimation, as the optimal bandwidth fixed kernel requires.
We investigated three bimodal densities that are of interest to our research. The first two bimodal densities have two equal modes, separated by $3\sigma$ in one and $4\sigma$ in the other. We expected to have relatively easy success detecting the right number of modes and their location in the $4\sigma$ case and were curious how well we would do in the $3\sigma$ case.

The third distribution was chosen with two modes of different size and scale, and a moderate degree of separation. The difference in the size and scale of the modes will create a problem in estimation for the fixed kernel, but we expected our adaptive estimator would overcome the shortcomings of the fixed kernel.
The trimodal distribution we were anxious to investigate confirmed the usefulness of our estimator. When using a global bandwidth fixed across all points of estimation, spurious modes appear in the right tail of this estimate and the third mode begins to disintegrate into many modes by the time the bandwidth is small enough to detect the valley between the two modes at locations 4 and 8. It would take an adaptive estimate to be accurate, as the amount of smoothing needed for the estimation interval left of the valley (local minimum) between the mode at location 8 and the mode at location 20, is definitely a lot less than the amount of smoothing needed to the right of that valley.

The last distribution we explored was the multimodal asymmetric claw, where the same phenomenon of varying size, scale, and degree of separation of modes was elaborately pronounced.

From each distribution, 40 samples were randomly drawn for each sample size.

5.3 The Sample Size

The sample size is an important factor in the accuracy of estimation. We like any estimator to be unbiased, and to converge to truth when the sample size grows indefinitely. Also an estimator that does well for small sample sizes is largely desired. We tested our estimator at three sample sizes. The sample sizes were 40, 100, and 250.

5.4 The Mode Existence Test \( \alpha \) Level

As the significance level \( \alpha \) for the mode existence test gets larger, one expects the number of modes detected to increase, as more modes will "pass" the test at higher \( \alpha \) levels. Also the significance level at which we are testing the individual modes is the proportion of time we make a Type I error and decide that an apparent mode is a true
feature of the density function, when in reality it is an artifact of the sample. Naturally with a larger Type I error, the proportion of the modes we inaccurately deem true is larger, leading to an estimate that is “wigglier” and less smooth than estimates at lower significance levels. But this should not prove to be an area of great concern, as large changes in the mode existence test significance levels produce small and sometimes no change in the accurate detection of the modality of the distribution. A sample that exhibits a certain modality at some $\alpha$ level, say 0.05, will still exhibit the same modality at the higher $\alpha$ level of 0.25, when there are no modes with p-values between 0.05 and 0.25. In addition, with the increase in the significance level, there is a limited increase in the number of modes detected. In other words, there is a cap on the modality detected by the test far smaller than $n$, no matter how large $\alpha$ gets. We tested our estimator at $\alpha$ levels 0.01, 0.05, and 0.15.

5.5 The $\alpha - n$ Connection

It is logical and desirable for our estimator to require a smaller significance level for the mode existence test as the sample size increases. Also we suggest that as $n$ gets smaller, a larger significance level will produce more accurate estimates. How strong is this relationship and how does it behave? We attempt to explore those two questions in Chapter 6.
CHAPTER 6

RESULTS

In this chapter we report our findings and make observations concerning the specifics of the distributions estimated, the sample size effects and the $\alpha$ level effect on accuracy of estimation, the rate of success in detecting the correct number of modes detected and efficiency with respect to the optimal bandwidth fixed kernel.

6.1 AVMISE Report

In Table 6.1 below we report the AVMISE results for our estimator and the optimal bandwidth fixed kernel. The reported AVMISE in Table 6.1 is the actual AVMISE multiplied by $10^3$. That was done for convenience only.

6.2 Distribution Effects

In the unimodal case we came very close to the performance of the fixed kernel, which in this case is ideal. In the bimodal cases of equal sizes and scales, the fixed kernel had a smaller AVMISE than our adaptive estimator. The distribution we hoped would prove our estimator a success, did exactly that. For the trimodal distribution in Figure 5.3, the efficiency ratio was more than one at the smaller sample size of 40 and at the $\alpha$ level 0.01 and 0.05. The same is true for the asymmetric claw distribution where we beat the fixed kernel at the sample size 40. For the bimodal distribution with the two different size modes, efficiency was very high (98%) at the smaller sample sizes.
Table 6.1: AVMISE Results. Table 6.2 Describes the densities D

<table>
<thead>
<tr>
<th></th>
<th>AVMISE</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>0.01</td>
<td>0.05</td>
<td>0.15</td>
<td>Optimal</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>100</td>
<td>250</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>10.9</td>
<td>6.5</td>
<td>3</td>
<td>12.4</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>10.8</td>
<td>6.6</td>
<td>4.4</td>
<td>13.3</td>
<td>7.4</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>11.6</td>
<td>5.8</td>
<td>16.6</td>
<td>9.6</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>10.7</td>
<td>3.2</td>
<td>9.3</td>
<td>10.7</td>
</tr>
<tr>
<td>5</td>
<td>11.5</td>
<td>9.6</td>
<td>4.8</td>
<td>12.28</td>
<td>9.1</td>
</tr>
<tr>
<td>6</td>
<td>37.4</td>
<td>30</td>
<td>27</td>
<td>63</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 6.2: Descriptions for Distributions in Table 6.1.

<table>
<thead>
<tr>
<th>D</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N(0,1)</td>
</tr>
<tr>
<td>2</td>
<td>(\frac{1}{3}N(0,1) + \frac{1}{3}N(3,1))</td>
</tr>
<tr>
<td>3</td>
<td>(\frac{1}{2}N(0,1) + \frac{1}{2}N(4,1))</td>
</tr>
<tr>
<td>4</td>
<td>(\frac{1}{10}N(0,1) + \frac{2}{10}N(5,2))</td>
</tr>
<tr>
<td>5</td>
<td>(\frac{1}{2}N(4,1) + \frac{1}{3}N(8,1) + \frac{2}{3}N(20,5))</td>
</tr>
<tr>
<td>6</td>
<td>(\frac{1}{2}N(0,1) + \sum_{i=2}^{\infty}(2^{-i}/31)N(1 + \frac{1}{2}, (2^{-i}/10)^2))</td>
</tr>
</tbody>
</table>

6.3 Results on the Sample Size Effects

The averaged mean integrated squared error (AVMISE), given by Equation 5.3, of our estimator decreased consistently as the sample size grew, regardless of the distribution estimated. An exception to that was the bimodal distribution with different mode sizes. This is likely to be an artifact of the small number of replications. Figure 6.1 shows the decrease in AVMISE as \(n\) increases. The rate of success in detecting the correct number of modes, which almost always indicates the modes detected are in the immediate vicinity of the correct mode locations, increases with the sample size in general, as one would expect.
Figure 6.1: AVMISE decreases as n gets large. Top plot $\alpha = 0.01$, $\alpha = 0.05$ middle plot and $\alpha = 0.15$ bottom plot.
Figure 6.2 indicates that fact. The efficiency of the estimator went down as the sample size went up, indicating that the estimator is more desirable at the smaller sample sizes. Figure 6.3 illustrates the decrease in efficiency with the growth of the sample size. At the $\alpha = 0.01$ the decrease in efficiency with the increase of $n$ was faster and more pronounced and less so at the $\alpha = 0.05$ level or the $\alpha = 0.15$ level. The results concerning the sample size were not terribly surprising as in their analysis of the general class of sample smoothing kernels, Terrel and Scott (1992) concluded that the sample smoothing estimators will have an advantage at small sample sizes, and less so as the sample size grows. This leads us to conclude that the new estimator ought to be considered seriously when the sample size is small, and where the distribution is suspected to be multimodal with uneven modes and with the modes having different degrees of separation. The efficiency of our estimator for small sample sizes, renders it ideal in situations where the sample is small.

6.4 Results on the $\alpha$ Effect

The significance level chosen for the mode existence test behaved as a smoothing parameter. As we increase the significance level of our tests, more of the modes tested will be significant, thus increasing the number of modes in the estimate. This does not happen continuously, as we have explained in Section 5.4.

We found the $\alpha$ level of 0.01 to be best for estimation of the unimodal standard normal, and for minimizing the AVMISE of the bimodal distribution with equal modes and $3\sigma$ of separation. But in the bimodal distribution with $3\sigma$ mode separation, the 0.01 level produced unimodal estimates 100% of the time, which of course is not what we are looking for. We have found that a higher $\alpha$ level would be more suitable, because the narrow valley between the modes is hard to detect, and thus requires a higher $\alpha$ level.
Figure 6.2: M-J estimator’s success rate in correct modality detection. From top to bottom: $N(0, 1), \frac{1}{2}N(0, 1)+\frac{1}{2}N(3, 1), \frac{1}{2}N(0, 1)+\frac{1}{2}N(4, 1), \frac{3}{10}N(0, 1)+\frac{7}{10}N(5, 2)$, Trimodal, Asymmetric claw.
Figure 6.3: The efficiency and the sample size. The order of plots is as in efficiency plot in Figure 6.2.
Increasing the $\alpha$ level to 0.15 detected the right modality of the distribution 30% of the time.

In the case where the mode separation was $4\sigma$, a higher $\alpha$ of 0.05 produced a smaller AVMISE and detected the correct modality of the distribution 44% of the time. As the separation between the modes gets larger, a lower $\alpha$ becomes sufficient to detect the bimodality of the sample. It logically takes a higher $\alpha$ to detect the valley between the two modes in the $3\sigma$ case, than the $\alpha$ required to detect the wider valley in the $4\sigma$ case.

The $\alpha$ level had little effect on the estimates of the AVMISE of the trimodal distribution. We had the best efficiency at the sample size 40 where we beat the optimal bandwidth fixed kernel, but the change was very modest for the three values of $\alpha$.

In Figure 6.4 we see two estimates of the trimodal sample at the $\alpha$ level 0.15 and 0.45. Note the robustness of the estimate to this huge change in $\alpha$.

Figure 6.5 explores the behavior of our estimator as $\alpha$ changes for all the distributions investigated. One can plainly see the small to moderate difference in the AVMISE of the different estimates produced at the various $\alpha$ levels for a certain sample size.

An exception to that rule is the asymmetric claw, where the 0.15-level estimate has a much smaller AVMISE than the estimates at the other significance levels. That is an observation consistent with that distribution having five very narrowly separated modes, and requiring a larger $\alpha$ to detect the various modes. The asymmetric claw is an extremely hard distribution to estimate, and we expect to have to let the sample size grow to more than a thousand before we can hope for some serious, accurate estimation. Figure 6.5 is a superimposed plot of the three plots in Figure 6.1.

The indifference of the AVMISE to changes in $\alpha$ is not shared by the rate of successfully
Figure 6.4: \( \alpha \) a smoothing parameter. Left estimate \( \alpha = 0.15 \), right estimate \( \alpha = 0.45 \).

Figure 6.5: M-J estimator's behavior with changes in \( \alpha \). This plot superimposes the three plots of Figure 6.1.
detecting the correct modality of the distribution. In Figure 6.2, we can observe the change in the rate of our success in detecting the correct number of modes as \( \alpha \) changes. It is obvious that depending on the distribution's modality and degree of separation between the modes, the success peaks at a certain \( \alpha \) and then decreases as \( \alpha \) gets larger.

6.5 Pointwise Confidence Intervals

Estimation confidence bounds of 95% for all distributions at all three sample sizes and \( \alpha \) levels 0.01, 0.05, and 0.15 are in Figure 6.6 through Figure 6.11. Since the number of replications is 40 across the board, the 95% confidence interval for a point estimate would be the \( X_{(2)} \) and \( X_{(39)} \) ordered sample statistics. In all the confidence interval plots, first row \( \alpha = 0.01 \), second \( \alpha = 0.05 \), and third \( \alpha = 0.15 \). And first column \( n = 40 \), second column \( n = 100 \), and third column \( n = 250 \). We can tell from the plots of confidence intervals that we have contained truth in those intervals for all distributions at the \( \alpha \) level of 0.15, for all sample sizes. We had problems containing truth in our confidence interval at the 0.01 mode existence test significance level and sample size 40 for the various distributions, except of course the unimodal case. The confidence intervals confirm what we have already declared about the appropriateness of the \( \alpha \) levels for the different sample sizes and distributions. The confidence interval got narrower with the increase in sample size, and wider and contained truth as \( \alpha \) got larger. We also plotted the Inter-Quartile Range and median to get a closer look at the performance of our adaptive estimator away from the outlier effects. Those plots for all distributions, sample sizes, and \( \alpha \) levels are in Figures 6.12 through 6.17. The 50% confidence interval failed to contain truth most of the time, but came very close every time. Yet at the \( \alpha \) of 0.15, our interval contained truth almost always. These results are encouraging and show our estimator “homing” in
on the truth with great diligence.
Figure 6.6: 95% data-based pointwise confidence interval for $N(0,1)$. 
Figure 6.7: 95% data-based pointwise confidence interval for $\frac{1}{2}N(0, 1) + \frac{1}{2}N(3, 1)$. 
Figure 6.8: 95% data-based pointwise confidence interval for $\frac{1}{2}N(0, 1) + \frac{1}{2}N(4, 1)$. 
Figure 6.9: 95% data-based pointwise confidence interval for $\frac{3}{10} N(0, 1) + \frac{7}{10} N(5, 2)$. 
Figure 6.10: 95% data-based pointwise confidence interval for the trimodal distribution.
Figure 6.11: 95% data-based pointwise confidence interval for the asymmetric claw.
Figure 6.12: Pointwise 50% confidence interval (IQR) for $N(0, 1)$. 
Figure 6.13: Pointwise 50% confidence interval (IQR) for $\frac{1}{2}N(0, 1) + \frac{1}{2}N(3, 1)$. 
Figure 6.14: Pointwise 50% confidence interval (IQR) for $\frac{1}{2}N(0, 1) + \frac{1}{2}N(4, 1)$. 
Figure 6.15: Pointwise 50% confidence interval (IQR) for $\frac{3}{10}N(0,1) + \frac{7}{10}N(5,2)$. 
Figure 6.16: Pointwise 50% confidence interval (IQR) for the trimodal distribution.
Figure 6.17: Pointwise 50% confidence interval (IQR) for asymmetric claw.
CHAPTER 7

CONCLUSIONS

7.1 Observations

The performance of our estimator in the distributions investigated indicates that it is superior to the fixed kernel at the trimodal distribution we studied, and almost as good in the other cases. Further simulations on a variety of distributions are quite appropriate, especially distributions that have similar characteristics to the trimodal distribution we investigated.

We suspect the asymmetric claw density would prove our estimator to be very useful at a larger sample size. Even sample size 250 is too small for accurate estimation of such a difficult density function. But the confidence interval in Figure 6.11 clearly shows how, even at the relatively small sample size of 250, the estimator is, on the average, detecting the peaks at the right locations. A higher value of $\alpha$ up to 0.4 was not enough to improve the estimate.

Having explored the behavior of our estimator with the changes in the sample size and the $\alpha$ levels and the different distributions, we can safely make the following observations:

1. Our estimator performed best and outperformed the optimal bandwidth fixed kernel when the distribution in question was multimodal and with different size modes and varying degrees of mode separation, the exact result we would hope for from an adaptive estimator. The efficiency of the estimator with respect to the fixed kernel was almost one in the unimodal case. The fixed optimal bandwidth kernel is the
most efficient for estimating the standard normal, and our estimator proved to be just as efficient.

2. The alpha level should be a decreasing function of the sample size. As the sample size gets larger, a smaller $\alpha$ is required for more accurate results. Also, the $\alpha$ level suitable for successful modality detection at the sample sizes and distributions explored was 0.05, with a slight finetuning as the sample size grew larger or smaller, and as the degree of separation got larger or smaller.

3. When compared to the fixed kernel, our estimator fared better for smaller sample sizes among the sample sizes we used. It is possible that our estimator will “catch up” to the optimal bandwidth fixed kernel as we push the sample size towards 1000. We already can see that as the sample size gets larger, we produce a smaller AVMISE and need a smaller significance level to detect the modes correctly.

It is our opinion that we have developed a useful adaptive density estimator, which is data based and chooses the bandwidth $h$ for variable kernel density estimation in an automatic fashion. It is advisable according to the results to use an $\alpha$ level of 0.05 in cases where the separation of modes is thought to be small to moderate. A larger value of $\alpha$ should be used in cases where detecting the correct number of modes is more important than a small squared error. The probability of detecting the correct number of modes increases with the $\alpha$ level, in a multimodal setting. The performance of the estimator at the distributions we investigated in this thesis is quite desirable and considered to be successful.

Of course one might say that we substituted the smoothing parameter $h$ by another ($\alpha$). The $\alpha$ choice is easier to deal with than the $h$ choice because the smoothing parameter
\( \alpha \) is less sensitive to moderate variations, and strongly pronounced modes persist at very small values of \( \alpha \) as well as at large values of \( \alpha \). Also there is a cap on the number of modes we can detect that is much much less than \( n \) no matter how large \( \alpha \) is, which is not the case when dealing with the smoothing parameter \( h \).

We recommend this estimator in investigating normal mixtures with small sample sizes (less than 100). The smaller the sample size, the more the benefit of using this estimator. Another case where we would recommend this estimator is in the case where a multimodal distribution has modes with different sizes and degrees of separation, as the trimodal distribution investigated in this thesis.

### 7.2 Future Directions

It is worth the effort to test this estimator at larger sample sizes than the ones attempted here, to study its asymptotic behavior. Also this adaptive estimator has shown enough promise that it would be of great benefit to explore its advantages in distributions other than the ones investigated here, in particular where the density function has the characteristics of the trimodal example or the asymmetric claw example where adaptive techniques are required.

We could also study the behavior of the estimator as \( \alpha \) increases or decreases indefinitely, in more detail than was done here. We suspect that the estimator will show desirable robustness to variations in \( \alpha \), especially in large sample sizes.
REFERENCES


