THE ESTIMATION OF MODES AND ITS APPLICATION TO PATTERN CLASSIFICATION

by

Michael William Blasgen

Memorandum No. ERL-M254

23 October 1968

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

ACKNOWLEDGEMENTS

The author is indebted to his research advisor, Professor Eugene Wong, for the guidance which led to this thesis, and for his assistance in its preparation.

The author also acknowledges the financial support of the National Science Foundation, and the Electronics Research Laboratory at the University of California through funds provided by AROD Contract DAHC04-67-C-0046.

TABLE OF CONTENTS

CHAPTER V : COMPARISON AND CONCLUSION

APPENDIX

ABSTRACT

This thesis deals with the problem of designing pattern classifiers with the aid of uncategorized samples. It is assumed that the samples have local concentrations, or clusters, and to aid in the design of the classifiers, two algorithms are introduced which estimate the location of these concentrations. More precisely, the algorithms estimate *modes* of essentially unknown probability distributions, given only a sequence of samples from the distribution. The bulk of this research is concerned with proving the convergence of these estimates as the number of samples becomes large for the unimodal case.

The algorithms discussed here are shown to be specific cases of general stochastic approximation methods, and the convergence is a consequence of this fact. Since the methods are particularly convenient computationally, requiring little storage and reasonable computations, the algorithms have been implemented on a computer, and the results of this effort are given for the multimodal case as well as the unimodal case.

CHAPTER I

INTRODUCTION

The work reported in this thesis is concerned with the design of classification procedures, when such design must be based on unclassified samples. This is a realistic problem, and examples of where the design must be of this type will be given in later paragraphs. The techniques that we shall explore make use of the relative concentrations of samples in the measurement space. For this reason, we refer to this body of techniques as cluster analysis.

One principal motivation is due to problems in pattern recognition. Most of the problems in pattern *recognition* that have been analyzed are really only problems in pattern *classification*. It seems to us that genuine recognition should involve the discovery of the existence of classes, as well as classification. It is to this problem of discovery that we address ourselves in this thesis.

The most familiar example in which the existence of classes has to be discovered is taxonomy. The entire hierarchy of classes must be discovered by examining the distribution of characteristics formed by the samples that have been collected. In problems of this type classification is a means rather than an end. Classification in this case serves the purpose of summarizing the detailed measurements of a sample by its membership in a class. For classification to be useful in this type of application, "variation" within a class must be relatively unimportant in comparison with "variation" among classes. This principle underlies much of cluster analysis.

Even in well specified classification problems such as automatic recognition of typewritten characters, cluster analysis may be useful in order to make a preliminary classification into various groups before final classification takes place. For example, the researcher may have correctly categorized samples available (i.e., samples of A's, samples of B's, etc.) but he may have difficulty achieving a high recognition rate. This often occurs when a single category, say the A category, contains capital A's, and small a's, and possibly different type fonts. The reason the recognition rate is low is that the names attached to a class have little in common with the natural groupings of the data, and the natural classes must be discovered. An appropriate procedure in this case is to "rename" the groups. Thus cluster analysis can be employed to make intermediate classifications, breaking the A category into several categories, say A_1, A_2, and so on.

Cluster analysis itself may be seen as a problem in pattern classification. A simplified model of a pattern classifier is shown below:

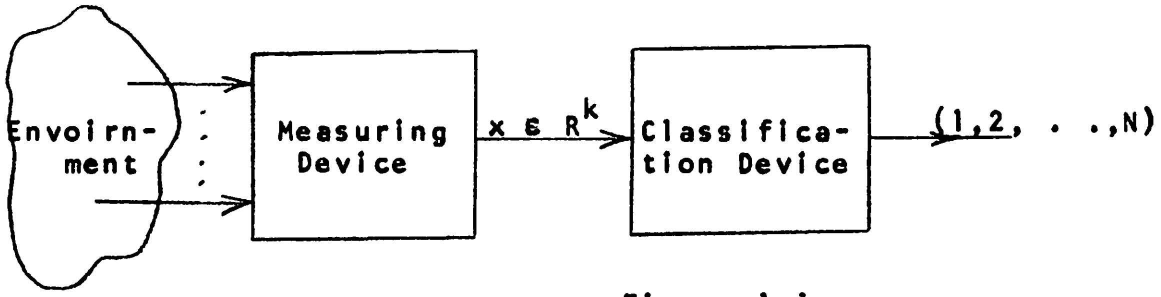


Figure 1.1

The measuring device converts, for purposes of classification, real objects into a set of numerically valued measurements. The classifier then maps these measurements into one of N classes, numbered 1 through N. For example, we may wish to distinguish football players (class 1) from jockeys (class 2). Suppose we measure the height and weight of the men (thus $k = 2$). Because of individual differences between men of the same category, the measurements within one class are distributed over a region in the measurement space. To handle such a distribution theoretically, we assume that the measurements are random variables with law $P(x \mid i)$ for class i. Thus a set of measurements is really a set of random variables with a probability distribution parameterized by the class index i. In the case above, the distribution of height and weight for jockeys is quite different from the distribution for football players. The classifier must detect this difference and classify accordingly.

The central problem of pattern classification is the design of these two blocks. Although the problem of choosing good measurements is a vitally important area for research, most research has been directed toward classifier design. To a certain extent, this thesis perpetuates this bias. We will thus concern ourselves with designing a good classifier, given an appropriate class of patterns to recognize.

It is clear that specifying the set of patterns to be recognized plays a major role in designing the classifier. In certain cases, for example, if the distribution of the samples (sample is a word used interchangeably with pattern) is completely known for each of the N classes, the "best" classifier can be specified without examining a single sample. On the other hand, if either the number of classes or the distribution is unknown (or partially unknown), or if the underlying probability law is represented only by samples which are correctly categorized, then the samples should be examined to determine the best classifier. For example, if only certain parameters of the underlying distributions are unknown, then the classifier should be designed using sample <u>estimates</u> of these parameters.

Various techniques for implementing estimation precedures and other classification schemes have been devised by various investigators. For a survey of some of this literature, see Nilsson [12]. Although this literature is fairly extensive, there has been little research done in the area of pattern classification where the least information is available--where the number of classes is unknown, where there are no correctly classified samples, and where the underlying probability distribution is essentially unknown. In such a situation pattern classification can be identified with cluster analysis.

The basic assumption in cluster analysis is that the data has local concentrations, or clusters. Thus if the patterns are represented in the measurement space as follows:

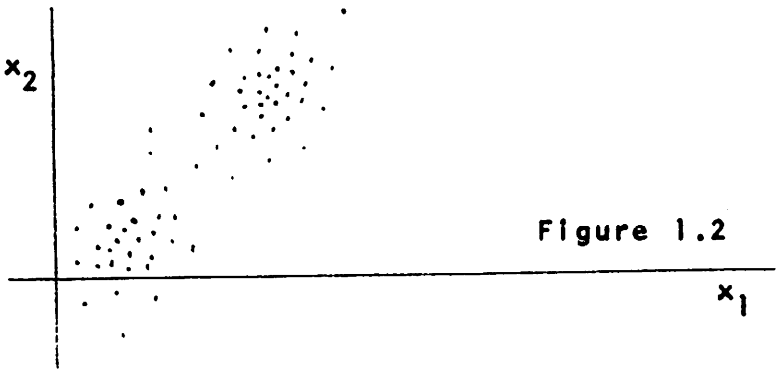


Figure 1.2

then the data is *clustered*, in this case into two clusters, and cluster analysis will attempt to sort these patterns into the two corresponding classes. In this way cluster analysis classifies data without prior specification of classes. As before, the classification will depend on the particular *unclassified* samples available, and cluster analysis has been called *learning without a teacher*, since there are no correctly classified samples available. For a summary of much of the literature in cluster analysis, see Ball [1].

The cluster analysis problem is to discover the clusters and classify data into classes corresponding to the natural clusters. There are relatively few different approaches to such a problem. One approach exploits the idea that "variation" within a cluster is small compared with "variation" between clusters. An attempt is thus made to find those sets S_i which make the sum of the variances on each set S_i a minimum. The collection $\{S_i\}$ is called a minimum variance partition, and such minimum variance procudures have received some attention in the literature. See, for example, Cox[8], who demonstrates the minimum variance partition for the normal case, MacQueen [11], who proves the convergence of a simple minimum variance algorithm, and Ball and Hall [2], who use an iterative procedure on a finite number of samples to obtain an approximation to the minimum variance partition. There are, however, certain theoretical difficulties with these algorithms, either in determining the correct number of clusters, or in assuring convergence.

Another approach in cluster analysis which is applicable only to the problem of classifying a finite number of samples is the similarity matrix method. A measure of "closeness" a_{ij} is defined between the i^{th} and j^{th} samples, and the similarity matrix is then defined as $A = \{a_{ij}\}$. An attempt is then made to group the samples in such a way that a_{ij} is small for i and j in the group, and large if i is in the group and j is outside. Bonner [4] uses such a formulation. There are, of course, computational difficulties with such a procedure if the number of samples is large. This thesis instead concentrates on methods which are easily computed, have well understood characteristics, and which still contribute significantly to the cluster analysis problem. The methods to be discussed here involve, as a third approach, estimation of <u>modes</u> of the underlying probability distribution. The following paragraphs explain why knowledge of the location of the modes is both useful and natural for cluster analysis.

Any classification device is determined by its partition of the measurement space. If the classification device is to be estimated, it is reasonable to try to estimate the partition. Generally speaking it is most convenient to estimate parameters, yet it is impossible to parameterize in any convenient way all possible partitions of the measurement space. It is appropriate, therefore, to examine techniques which estimate a reasonable number of parameters, choosing those parameters which allow good classification decisions to be made.

If the underlying distribution is parametric, then by definition there is a finite collection of parameters

$\{\alpha_1, \ldots \alpha_n\}$ such that the overall probability distribution P can be written as

$$P(x) = f(\alpha_1, \ldots \alpha_n, x)$$

where $f(\cdot)$ is a known function. In this case it is possible to estimate these parameters based on samples from the distribution P, and thus determine an estimate of P. Once P is known, a good classifier can be designed. Unfortunately, cluster analysis problems are non-parametric in nature (although some possibly relevant parametric work has been done--see Fralick [10]), and there are no obvious parameters to estimate. It is, of course, possible to estimate moments, and in this manner to estimate the distribution, but there are many problems with this idea. High order moments are required to approximate a multimodal distribution, and the estimators for high order moments are quite bad. The distribution can be estimated directly by using the sample distribution or the "potential function" method (see Chapter IV, Section 1), or by using the simpler sample histogram based on cells. If such a histogram procedure is used in cluster analysis, one does not normally want to cover the entire measurement space with small cells for computational reasons, and for this reason some work has been done on the problem of optimum placement of the cells. See Sebestyn [17]. However, the procedure is still computationally complex and has unknown theoretical properties. Rather than estimate the entire distribution function, it seems more profitable to attempt to estimate relevant parameters.

The one set of parameters that does have meaning for a large family of distributions and which would be appropriate for cluster analysis is the set of modes. A mode is, of course, a local maximum in the underlying density function.

The modes are perhaps the most natural measurement of concentration. It is certainly true that the best choice of a cluster analysis technique will depend on how the data is structured--that is, the meaning of "clustered." Although a precise definition of cluster has intentionally been avoided, almost any reasonable definition would either implicitly or explicitly use the idea that a cluster is a subset of the data such that the underlying probability is unimodal over this set. A bimodal cluster is a contradiction in terms. Thus knowing the number of modes is really equivalent to knowing the number of clusters.

Modes have other important properties. Not only are they natural in the above sense, but, at least at present, they are the only parameters which represent clustered data and which can be conveniently estimated. In the next chapter, an algorithm is presented which, based on experience in using the method, can estimate all the modes of a distribution over n-dimensional Euclidean space. This is a significant step toward solving the cluster analysis problem, since reasonable classification can usually be made with no more information than the number and positions of the modes.

To see one possible way modes can be used to categorize data, consider the following procedure, called "nearest neighbor classification." Suppose that each class is one or more clusters, and suppose that data is grouped in this fashion, that is, assume that the data is drawn from a distribution which is a super-position of n unimodal densities, each class corresponding to one or more densities. Assume in addition that, although nothing is known about the distribution, an infinite sequence of <u>classified</u> samples is available, denoted (x_1, i_1) , (x_2, i_2), ... where i_j indicates the category or class of the j^{th} sample x_j. Classify a sample y of unknown category as follows: Find the x_i which is closest to y, say x_m, and classify y as i_m. This procedure has much to recommend it in terms of low probability of error (See Cover and Hart [7]). However, two facts prevent it from being applied: (1) Every classified sample must be stored--thus an infinite memory is required, and (2) an infinite number of distances must be calculated in order to find the smallest distance. Even if only a finite number of categorized samples are used, the storage and time problems are considerable.

To cure these problems, and still preserve the basic idea of nearest neighbor classification, a modified procedure involving only "typical" or "representative" samples has been suggested. In this procedure, a sample is classified into category i if it is closer to a representative of category i than any other category. This is illustrated in Figure 1.3 . Note that in this rather idealized case, classification of an unknown vector will be done almost as accurately with representatives as with the whole data set.

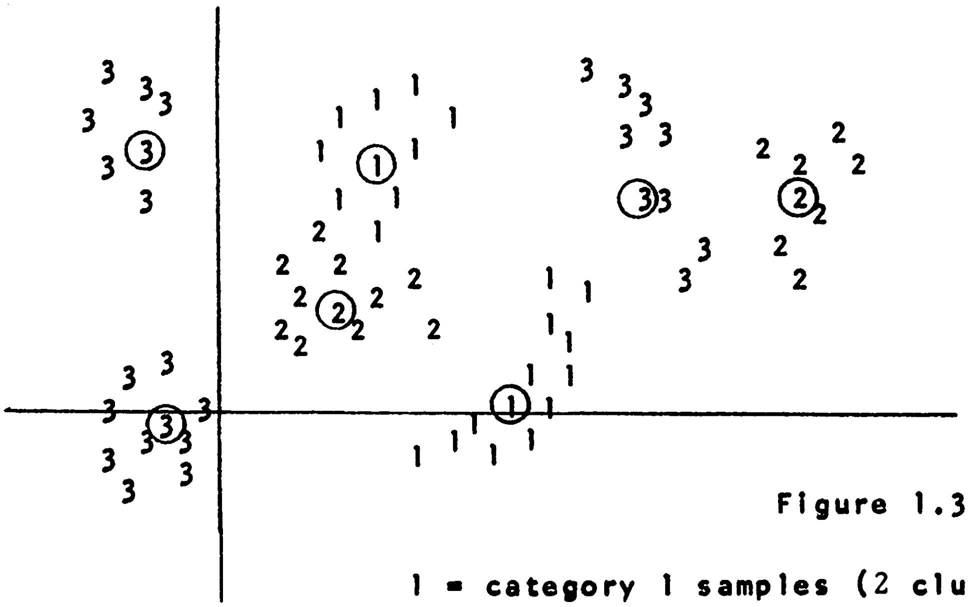


Figure 1.3

1 = category 1 samples (2 clusters)
2 = category 2 samples (2 clusters)
3 = category 3 samples (3 clusters)

"Typical" sample circled.

This is an eminently practical algorithm, since little storage or computation is required. However, difficulties arise when one attempts to choose the representatives. Here cluster analysis is useful again. It is clear that there should be at least one representative for each cluster, and if only one is chosen, the one to choose is the cluster "center." This suggests two possibilities--the conditional mean of the samples in the cluster, or the mode of the cluster. Both of these have much to recommend them, however the conditional means are relatively difficult to find since the correct sets to condition on are hard to find. Thus a mode estimation procedure seems most useful for classification in the nearest neighbor procedure.

It was mentioned above that good conditional means are fairly hard to estimate. Interestingly enough, even if it were possible to find a good set of conditional means, in most cases the classification can be done as well with the modes. An informative example is the case where each class is a normal distribution, and the overall distribution is a superposition of n equally weighted normals with identity covariance matrices and differing means. In this simplified case, the optimum decision rule is to choose class i if the sample is closest in Euclidean distance to μ_i, the mean of the i^{th} distribution. If the appropriate set of conditional means $\{\mu_i^o\}$ were found (by appropriate I mean that μ_i^o is the average over the set of points which are closest to μ_i) then classification could be done in a similar way, classifying a sample in class i if it is closest to μ_i^o. This gives an identical classification rule. However, this property is shared by the modes, since assuming that the means μ_i are not too close to each other, the same decision rule is defined by the modes. Thus it can be seen that for classification purposes means, modes and conditional means are equivalent in this case.

Even more importantly, the following chapters will indicate that modes can easily be estimated, whereas means and good conditional means are much harder to find without apriori assumptions about the form of the underlying distribution.

Another property of modes is demonstrated by classifying data in the following way: Define a procedure based on the points $\{\xi_i\}_{i=1}^{n}$ by letting $\varepsilon > 0$ be small and deciding a sample x is in class i if $d(x,\xi_i) < \varepsilon$ for some metric d, and otherwise make no decision. Then simply by the definition of mode, as ε becomes small the use of this procedure based on the modes will classify a greater proportion of the samples than any other n points. That is

$$\Pr\{\text{making a decision} \mid \text{modes}\} \geq \Pr\{\text{making a decision} \mid \text{any } n \text{ points}\}$$

In addition, if the samples close to a mode are all from a single class (this is a reasonable assumption common to much of cluster analysis) then for ε small <u>every</u> classification decision made by the above procedure using the modes will be correct, and this will not in general be true for any other n points.

Modes have another property which makes them useful as descriptors of clustered data. If a certain random variable $x \in R^k$ has a density $p(\cdot)$ which has n modes at locations m_1 , m_2, . . m_n , then for any linear transformation A from R^k onto R^k , the random variable Ax has a density which has n modes at locations Am_1, . . . , Am_n. Thus it can be seen that, using the modes as the essential property of clusters, a data set and any linear transformation of the same data set will have the same clusters. This makes a classification scheme based on modes relatively independent of the units in which the measurements are made, and independent of the particular quantities measured, so long as the "proper" set can be derived from them by an onto linear transformation.

These last few paragraphs have attempted to show that the set of modes is the set of parameters most relevant to cluster analysis. They are very useful in the modified nearest neighbor classification technique, and they induce the same classification rule as means in the simplest normal case. The modes are, by definition, in the regions of highest probability,and finally, modes are invariant under linear transformation. This is not the whole argument, which in general emphasizes the "naturalness" of modes. The modes are the most intuitively satisfying indicators of where the clusters are centered, since, as was pointed out earlier, clusters are by nature unimodal. Because of this, the modes are a reasonable measure of the location of the cluster. Thus if modes can be conveniently estimated, the cluster analysis problem will be considerably simplified.

The next chapter introduces two algorithms which estimate modes in an extremely convenient manner. In one case, the algorithm estimates points which are not quite modes in the strict sense (i.e., local maxima) but which are "averaged" modes and may actually be <u>more</u> useful for cluster analysis. Both algorithms are shown to be convergent in the unimodal case, and are expected to converge in the multimodal case, although no proof of this has been found.

In Chapter III, the important generalization to n-space is made, and it is shown that all the one dimensional results carry over to R^k. Chapter IV presents actual results of these algorithms as they operate on computer generated data, both unimodal and bimodal. Chapter IV also contains the rather limited multimodal results obtained in this research. Finally, the last chapter briefly compares these techniques with both conventional mode estimation procedures and with other mode estimation procedures based on stochastic approximation. Some extensions and generalizations are also suggested in the last chapter.

CHAPTER II

MODE SEEKING ALGORITHMS

Section 1. Introduction.

In the first chapter it was seen that a significant problem in cluster analysis would be solved if a technique could be devised for locating modes in multimodal populations. In this chapter, we demonstrate such a technique. In particular, a method is given which, when provided a sequence of samples from an unknown unimodal distribution, extracts an estimate of the mode which converges to the actual value of the mode. The method has been carefully devised so that it will continue to work in the multimodal situation.

To provide some intuition regarding the particular algorithm to be discussed here, consider the unimodal density below:

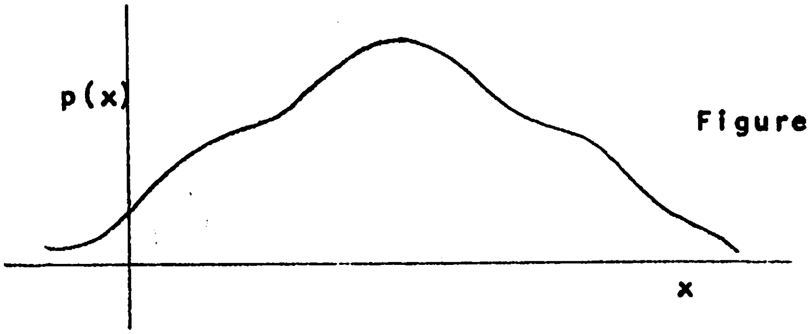


Figure 2.1

In this case, the mean and mode almost coincide, so that the mode can be estimated by estimating the mean. However, if the method is to work for multimodal as well as unimodal distributions, estimating the mean is unsatisfactory. If we can restrict our attention to a subset of the data whose conditional density looks like Figure 2.1, then the mode can be estimated by the conditional mean-- $\mu_n = \frac{1}{n}\sum^n x_i$ where x_i is a sample from the distribution in Figure 2.1 . This procedure estimates the mode quite well when the distribution a symmetric about the mode. However, if the distribution is skewed, the conditional mean is somewhat less satisfactory as an estimate of the mode.

There is an easy cure for this. Consider the density in Figure 2.2.

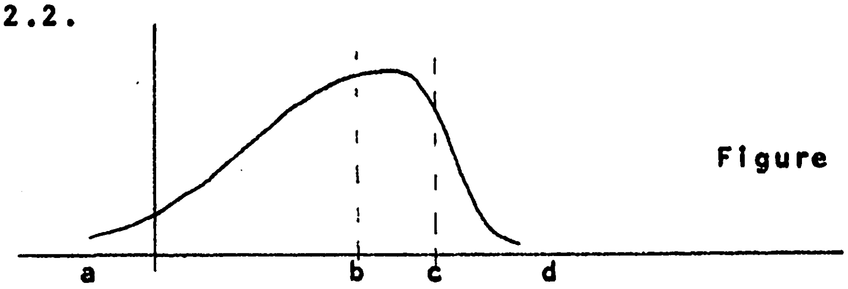


Figure 2.2

If a smaller subset than the interval (a,d) is chosen, and if it is located properly, like the interval (b,c), then the error in estimating the mode by estimating the conditional mean on (b,c) is considerable smaller.

This discussion may seem a bit academic, for if the mode location is unknown apriori, these subsets cannot be placed appropriately. That is, if the best location for these intervals is known, the mode location is also known, without estimating anything.

Fortunately there is a useful procedure based on the above discussion. We will attempt to estimate a good subset of the data at the same time as the mode is estimated. Considering only the unimodal case, one possible simultaneous estimation scheme is as follows:

Let $x_1, x_2, \ldots$ be a sequence of independent and identically distributed random variables with a common unimodal density $p(\cdot)$. Center an interval of length 2L, with $L>0$ and specified, about x_1, and call this center μ_1. Suppose we call the next sample which falls in this interval y_2. The density of the random variable y_2 given that $\mu_1=\xi$ is, as one might guess, the density $p(\cdot)$ conditioned on being in the interval, and thus

$$p_{y_2|\mu_1=\xi}(x) = \frac{p(x)}{\int_{\xi-L}^{\xi+L} p(z)\,dz} \qquad \text{for } |x-\xi|\leq L$$

$$= 0 \qquad \text{elsewhere.}$$

Since $p(\cdot)$ is unimodal, the expected value of y_2 is closer than μ_1 to the mode, and if we define $\mu_2 = (\mu_1+y_2)/2$ then μ_2 is closer, on the average, to the mode than μ_1 is.

A sequence $\mu_1, \mu_2, \ldots$ can now be generated by extending this process. Thus if $\mu_n=\xi$ is the present center we define y_{n+1} as the next sample which falls in the interval $(\xi+L,\xi+L)$. The new interval center μ_{n+1} is then obtained by averaging:

$$\mu_{n+1} = \frac{1}{n+1} (n\mu_n + y_{n+1}) \quad .$$

If p(•) is unimodal, then for each n, μ_{n+1} is closer to the mode than μ_n is, and in this case convergence might be expected. In fact convergence does obtain in such a situation. The assumption is made in the following form: If y_{n+1} is closer, on the average, to the mode than μ_n is, we will say the distribution is unimodal on the average. This is a natural condition, and one which many unimodal distributions satisfy.

The algorithm has been devised to enable μ_n to converge to a mode in the multimodal situation as well. (However, no completely satisfactory proof of convergence in this general situation has been found.) Thus, depending on the starting point μ_1, the algorithm will converge to one of the modes, and if it is restarted at a new μ_1 it will converge again, this time to a possible different mode. This, combined with a method for distinguishing the different values to which μ_n is converging, provides a simple and effective algorithm for finding all the modes of a multimodal distribution.

An important consideration here is the ease with which the algorithm can be applied, since so little has to be stored or computed. The samples x_1, x_2, . . . are used sequentially and then discarded and the only parameter which must be stored and updated is μ_n. The computations are also straightforward.

In the next section, the algorithm is precisely stated, and its convergence is shown under the "unimodal on the average" assumption. Actually a slightly generalized version of the algorithm is studied; whereas we have considered in this section

$$\mu_{n+1} = \frac{1}{n+1}(n\mu_n + y_{n+1})$$

$$= \mu_n - \frac{1}{n+1}(\mu_n - y_{n+1}) \quad ,$$

in the next section we consider

$$\mu_n = \mu_n - a_n(\mu_n - y_{n+1}).$$

In Section 3 the unimodality assumption is more carefully studied, and there it is shown that most unimodal densities satisfy it. Section 4 introduces a new, though similar, algorithm, and Section 5 is concerned with rates of convergence.

Section 2. A mode seeking algorithm, and its convergence proof.

<u>THEOREM 1</u> Let $x_1, x_2, . .$ be a sequence of independent identically distributed real random variables, each with a fixed probability density $p(\cdot)$, and assume $Ex^2 < \infty$. Let L be a fixed positive number and derive a new sequence $y_1, y_2, . . .$ and $\mu_1, \mu_2, . . .$ as follows:

For $n = 1,2, . .$ define

$$y_n = x_{j_n} \qquad \text{where} \qquad j_1 = 1 \text{ and for } m = 1,2, . .$$

$$j_{m+1} = \min_{j_m < i} \{ i \mid |\mu_m - x_i| < L \}$$

and define $\mu_1 = y_1$

$$\mu_{n+1} = \mu_n - a_n(\mu_n - y_n)$$

for a sequence a_n of positive numbers.

Let the density $p(\cdot)$ be such that the function

$$f(\xi) = \xi - \frac{1}{\int_{\xi-L}^{\xi+L} p(x)} \int_{\xi-L}^{\xi+L} xp(x)\, dx \quad , \xi \text{ in the support of } p$$

satisfies

$$(\xi - \theta) f(\xi) > 0 \qquad \text{for } \xi \neq \theta \quad . \tag{2.1}$$

Let a_n satisfy

$$\sum a_n = \infty \qquad \sum a_n^2 < \infty \quad . \tag{2.2}$$

Then $E(\mu_n - \theta)^2 \to 0$ and $\mu_n \to \theta$ with prob. one as $n\to\infty$.

REMARK Condition (2.1) is a type of unimodality requirement, denoted in the sequel as unimodal on the average about θ. Because p(•) is judged only by its characteristics averaged over an interval of length 2L, p(•) may differ slightly from a true unimodal distribution. In fact, if L is made very large, many densities become unimodal on the average, and in the limiting case of L = ∞ , any density is unimodal with θ taken as the mean. Our primary interest is in smaller values of L.

Proof of THEOREM 1: Condition (2.1) is a requirement on the expected value of the random variable y_n(given that $\mu_{n-1}=\xi$). y_n has a density given by

$$p_{y_n|\mu_{n-1}=\xi}(x) = \frac{p(x)}{\int_{\xi-L}^{\xi+L} p(y)} \qquad \text{for } |x-\xi| \leq L$$

$$= 0 \qquad \text{elsewhere}$$

as can be seen by considering, for any set $A \subset (\xi-L,\xi+L)$, the probability that y_2 is in the set A, given that $\mu_1 = \xi$. That is,

$$\begin{aligned}\text{Prob}\{y_2 \varepsilon A|\mu_1=\xi\} = \Pr\{&x_2 \varepsilon A \text{ or} \\ &x_2 \not\varepsilon (\xi-L,\xi+L) \text{ and } x_3 \varepsilon A \text{ or} \\ &x_2,x_3 \not\varepsilon (\xi-L,\xi+L) \text{ and } x_4 \varepsilon A \text{ or} \\ &\ldots\ldots \} \quad .\end{aligned}$$

The events on each line are disjoint, so we can write

$$= \Pr\{x_2 \varepsilon A\} + \sum_{i=3}^{n} \Pr\{x_i \varepsilon A \text{ and}$$

$$x_j \not\varepsilon (\xi+L,\xi+L) \; 2\leq j<i\}.$$

Since the x_i's are independent and identically distributed, this can be written as

$$= \Pr\{x_2 \varepsilon A\}\sum_{i=1}^{\infty}(1-\Pr\{x\varepsilon(\xi-L,\xi+L)\})^i$$

$$= \Pr\{x \varepsilon A\} / \Pr\{(\xi-L,\xi+L)\}$$

as desired.

Thus the function $f(\cdot)$ appearing in (2.1) can be written as

$$f(\xi) = \xi - E\,Y_\xi \tag{2.3}$$

where the random variable $Y_\xi = (y_n \text{ given } \mu_{n-1}=\xi)$. The subscript n can be dropped since the distribution of y_n given $\mu_{n-1}=\xi$ is independent of n.

Using this result, the theorem follows as a consequence of the following stochastic approximation theorem. See Dvoretsky [9].

THEOREM 2 Let Z_ξ be a one parameter family of real random variables, $\xi \in R$, and define $g(\xi) = EZ_\xi$. Assume $g(\xi)$ exists for every ξ. Define a sequence x_n as x_1 arbitrary and

$$x_{n+1} = x_n - a_n z_n$$

where z_n is an observation of the random variable Z_{x_n}, and where a_n is a positive sequence such that

$$\sum a_n = \infty \qquad \sum a_n^2 < \infty$$

Let Z_ξ have uniformly bounded variances, and let $g(\cdot)$ satisfy

$$|g(\xi)| < A|\xi| + B \quad \text{some } A,B \quad , \tag{2.4}$$

$$g(\xi)\ (\xi - \Theta) > 0 \quad \text{for } \xi \neq \Theta \quad \text{for some } \Theta, \tag{2.5}$$

$$\inf_{t_1<|\xi-\Theta|<t_2} |g(\xi)| > 0 \quad \text{all } 0<t_1<t_2<\infty. \tag{2.6}$$

Under these circumstances, $x_n \to \Theta$ in quadratic mean and with probability one as $n\to\infty$.

To apply this result to the mode seeking algorithm we define a random variable $Z_\xi = \xi - Y_\xi$ where Y_ξ is the previosly defined r.v. with density

$$p_{Y_\xi}(x) = \frac{p(x)}{\int_{\xi-L}^{\xi+L} p(y)} \qquad \text{for } |x-\xi| \leq L$$

$$= 0 \qquad \text{elsewhere.}$$

Defining, as before, the function $f(\xi) = E\,Z_\xi$, we identify this $f(\cdot)$ with the $g(\cdot)$ appearing in THEOREM 2 and observe that since $f(\cdot)$ is continous, (2.5) implies (2.6). By (2.1), (2.5) holds. To verify (2.4) we note that by the nature of the random variable Y_ξ, $|\xi-Y_\xi| \leq L$ with prob. one, and thus $|f(\xi)| \leq L$. In addition, $E\,Z_\xi^2 \leq 4L^2$. These facts, combined with condition (2.2) of THEOREM 1, guarantee that $\mu_n \to \theta$.

Section 3. Unimodal on the average examined.

In the last section a simple algorithm was introduced which observes samples through a window, and centers this window on the average of the observed samples, and it was shown that the estimates μ_n converge to a value Θ if the underlying density satisfies a condition (2.1) which we called unimodal on the average. It is of considerable interest to examine the class of densities which have this property, and to study the meaning of the value Θ .

We have already seen that condition (2.1) is a requirement on the expected value of the random variable y_ξ which appears in the theorem. y_ξ is the random variable defined by considering for $L>0$ fixed only those values of a basic r.v. x (with density $p(\cdot)$) which fall in an interval $(\xi-L, \xi+L)$. Thus the density of y_ξ is

$$
\begin{aligned}
p_{y_\xi}(z) &= \frac{p(z)}{\int_{\xi+L}^{\xi+L} p(x)} && |z-\xi| < L \\
&= 0 && \text{elsewhere}
\end{aligned}
$$

and is seen to be parameterized by ξ , a real number which can vary over the support of $p(:)$. Condition (2.1) says that the expectation of y_ξ is always closer to Θ than ξ .

This is an intuitively satisfying requirement, and one would expect it to be satisfied by many unimodal distributions.

To understand which ones satisfy it, we first note that if $p(\cdot)$ is a symmetric unimodal density then $p(\cdot)$ is almost unimodal on the average. More precisely:

THEOREM 3 Let $p(\cdot)$ satisfy $p(x) \leqq p(y)$ for $|y-\Theta| \leqq |x-\Theta|$. Let $f(\xi)$ be defined for ξ in the support of p and L>0 by

$$f(\xi) = \xi - \frac{1}{\int_{\xi-L}^{\xi+L} p(x)} \int_{\xi-L}^{\xi+L} xp(x)\, dx\ .$$

Then, for every L>0,

$$f(\xi)(\xi-\theta) \leqq 0. \tag{2.7}$$

Proof: Let L>0 be fixed, and take $\theta = 0$. Note that

$$\xi - f(\xi) = \xi + \frac{1}{\int_{\xi-L}^{\xi+L} p(x)} \int_{-L}^{L} xp(x+\xi)\, dx$$

so (2.7) is satisfied for $\xi \geqq 0$ if

$$\int_{-L}^{L} xp(x+\xi)\, dx \leqq 0.$$

This requires verifying that

$$\int_{0}^{L} x[p(x+\xi) - p(-x+\xi)]\, dx \leqq 0\ ,$$

which is certainly satisfied if

$$p(x+\xi) \leqq p(-x+\xi) \qquad \text{for } x \text{ in } (0,L).$$

Considering $\xi>x$ it is true by assumption, and noting that $p(x+\xi) = p(-x-\xi)$ it is true for $\xi<x$. A similar argument holds for $\xi<0$. This completes the proof.

Thus any unimodal symmetric distribution is almost unimodal on the average. However, condition (2.7), a weaker version of (2.1), is not sufficient for convergence. This is demonstrated by taking $p(\cdot)$ as a uniform distribution of base 2a, and choosing $L<a$. Such a $p(\cdot)$ satisfies (2.7) and in this case μ_n will not converge, but will wander about in the region $(-a+L,a-L)$.

The condition we want -- unimodal on the average -- is obtained if $p(\cdot)$ is strictly unimodal as follows:

THEOREM 4 Let $p(\cdot)$ be symmetric about Θ and assume $p(x) < p(y)$ if $|x-\Theta| > |y-\Theta|$. Then, for every $L>0$, $p(\cdot)$ is unimodal on the average, that is:

$$(\xi-\Theta)f(\xi) > 0 \qquad \xi \neq \Theta$$

Proof: From the preceding proof we see that we must show that, for $\xi>0$,

$$\int_{-L}^{L} xp(x+\xi)\, dx < 0$$

which again is true if $p(x+\xi) < p(-x+\xi)$ for a set of positive measure of x in $(0,L)$. The strict inequality holds excluding only the point $x=\xi$ and points where $p(x+\xi) = p(-x+\xi) = 0$.

This result is easily applied to show that the normal and exponential distributions, among others, are unimodal on the average about the mode.

In fact, $p(\cdot)$ need not be strictly unimodal, as "flat spots" are permitted:

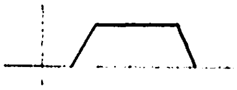
THEOREM 5 Let $L>0$ be fixed, and let $p(\cdot)$ be unimodal and symmetric about θ. Assume there is an $L^\circ < L$ such that for x, y in the support of p,

$$p(x) = p(y) \quad \text{implies } |x-y|<2L^\circ \quad \text{or } (x-\theta) = -(y-\theta)$$

Then $p(\cdot)$ is unimodal on the average about θ.

Proof: As before, it is sufficient to show that $p(x+\xi) < p(-x+\xi)$ for all x in some set $S \subset (0,L)$ which has positive measure. This will be satisfied if we take $S = (L^\circ,L)$, since for every x in this S,

$$|(x+\xi) - (-x+\xi)| > 2L^\circ .$$

This result shows that the uniform distribution and the trapezoidal distribution  are unimodal on the average provided that the intervals where the density is nonzero and constant are of length less than $2L$.

If $p(\cdot)$ is strictly unimodal but not necessarily symmetric about the mode Θ , then it is true that for every $L>0$, there exists a Θ' in $(\Theta-L,\Theta+L)$ such that $p(\cdot)$ is unimodal on the average about Θ'. More precisely:

THEOREM 6 Let $L>0$ be fixed, and let $p(\cdot)$ satisfy $p(x)< p(y)$ for $x<y<\Theta$, and $p(x)<p(y)$ for $x>y>\Theta$. Then there exists a Θ' in $(\Theta-L,\Theta+L)$ such that $f(\Theta') = 0$. If all such Θ' satisfy

$$\frac{L\ [\ p(\Theta'+L)+\ p(\Theta'-L)\ ]}{\int_{\Theta'-L}^{\Theta'+L} p(x)} < 1 \tag{2.8}$$

then Θ' is unique, and $p(\cdot)$ is unimodal on the average about Θ' for this L.

Proof: Construct an arbitrary density of this type as follows:

$$p(x) = \begin{cases} p_1(x) & x \geq 0 \\ p_2(x) & x<0 \end{cases}$$

where p_1 and p_2 are symmetric and strictly unimodal. Thus they are unimodal on the average for this L. Clearly

$$f(\mu) = \mu - \frac{1}{\int_{\mu-L}^{\mu+L} p(x)} \int_{-L}^{+L} xp(x)\ dx$$

has the correct properties (that is, satisfies 2.1) for $|\mu| \leqq L$. We need to check the case $|\mu| < L$. Define Θ' as a root of f. f has a root, since f is continuous and is positive for $\mu > L$ and negative for $\mu < -L$. (We take $\Theta = 0$ here). That this Θ' is unique, and that f has the correct property, can be most easily verified by noting that $f(\cdot)$ is differentiable and has a positive derivative at Θ' if and only if

$$\int_{\Theta'-L}^{\Theta'+L} p(x) \quad - L[p(\Theta+L) + p(\Theta-L)] \quad > \quad 0.$$

This is true by assumption (2.8). This proves that Θ' is unique, since if $f(\cdot)$ had more than one root, the derivative of f would change sign. But we have only a single sign. Thus only one root, and $f(\cdot)$ has the desired properties.

The four preceding theorems have shown that a large class of unimodal densities is unimodal on the average. This class is even larger, since the criteria for unimodality used in this chapter is relatively insensitive to slight modifications to the density function.

<u>THEOREM 7</u> Let $L>0$ be fixed. Let $p(\cdot)$ be unimodal on the average about zero and let $q(\cdot)$ be a symmetric density with bounded support satisfying

$$\frac{Lq(L)}{\int_{-L}^{+L} q(x)} \quad < \quad 1$$

Then there is a $\lambda^\circ < 1$ such that for all λ in $(\lambda^\circ, 1)$, the density $\lambda p + (1-\lambda)q$ is unimodal on the average about zero for this value of L.

Proof: We must show, for $\mu>0$, that

$$\int_{\mu-L}^{\mu+L} x[\lambda p(x) + (1-\lambda)q(x)] < \mu\int_{\mu-L}^{\mu+L} [\lambda p(x) + (1-\lambda)q(x)]$$

By assumption, $p(\cdot)$ satisfies

$$\int_{\mu-L}^{\mu+L} xp(x) = \mu\int_{\mu-L}^{\mu+L} p(x) + h(\mu) \qquad \text{where} \quad h(\mu) > 0.$$

Thus we must show

$$\int_{\mu-L}^{\mu+L} xq(x) < \mu\int_{\mu-L}^{\mu+L} q(x) + \frac{h(\mu)}{1-\lambda} .$$

A Taylor serives argument shows that for small μ, say $|\mu|<\varepsilon'$

$$\int_{\mu-L}^{\mu+L} xq(x) < \mu\int_{\mu-L}^{\mu+L} q(x) .$$

Thus we choose λ° as the smallest positive number such that

$$\sup_{\mu\in S} \int_{\mu-L}^{\mu+L} xq(x) - \mu\int_{\mu-L}^{\mu+L} q(x) - \frac{h(\mu)}{1-\lambda^\circ} < 0$$

where the set $S = \{x \mid x \text{ in support } p \text{ and } |x|<\varepsilon\}$. Such a λ° always exists, and satisfies the requirements of the theorem.

To illustrate this result we can take

$$q(x) = |\cos ax| \qquad |x| < M$$
$$= 0 \qquad \text{elsewhere.}$$

Then with proper normalization, and with any M, any a, and roughly half the possible values for L, $\lambda p + (1-\lambda)q$ is still unimodal on the average for λ close to 1, yet this density is clearly no longer unimodal.

More general results of this type can be obtained. For example if $q(\cdot)$ is neither symmetric nor locally unimodal, the theorem is still basically true, except that the "mode" may be moved slightly.

Incidentally, this proof suggests another fact.

FACT If p_1 and p_2 are unimodal on the average about Θ, then for any λ in $(0,1)$, $\lambda p_1 + (1-\lambda)p_2$ is also unimodal on the average about Θ.

The results presented in this section show that the condition--unimodal on the average--which is required for convergence of the mode seeking algorithm is satisfied for a large class of probability densities, both unimodal and "nearly" unimodal. This is not surprising, since the condition arises in a fairly natural way.

Section 4. A shrinking window algorithm.

One of our goals in this research has been to find modes in *multimodal* distributions. No completely satisfactory proof has been found that the algorithm discussed in section 2 converges in the multimodal situation, but convincing arguments and computer results seem to indicate that convergence does hold. The argument regarding convergence goes as follows: at step n, the algorithm only considers samples from the distribution within L units of μ_n. If μ_n is nearing a mode, and L is chosen small enough, then the algorithm "sees" only samples from the region about that mode, and does not see the other modes. This argument is at least valid for some special multimodal distributions--see Chapter IV, section 3.

The idea of putting in a window (μ_n-L,μ_n+L) can be generalized. Consider an algorithm which reduces the size of the window as it proceeds. This should improve the ability of the algorithm to converge in the multimodal situation, since with the width of the window decreasing toward zero, sooner or later the algorithm will concentrate on only *one* of the modes.

There is a second justification for such a "shrinking window" algorithm. Recall that the method discussed in Section 2--the "fixed window" algorithm--may not converge to the true mode. Intuitively speaking, it should be possible to make the shrinking window algorithm converge to the true mode of a unimodal distribution. This is indeed the case, and is shown below.

THEOREM 8 Let L_n be a sequence of positive numbers. For a fixed probability density $p(\cdot)$ with $Ex^2<\infty$ define a random variable Y_ξ^n by its density:

$$p_{Y_\xi^n}(x) = \frac{p(x)}{\int_{\xi-L_n}^{\xi+L_n} p(y)} \qquad |x-\xi|<L_n$$

$$= 0 \qquad \text{elsewhere.}$$

Let $p(\cdot)$ have derivative $p'(\cdot)$ defined a.e. which satisfies, for x in the support of p,

$$(x-\theta)\frac{p'(x)}{p(x)} < 0 \qquad \text{for } x\neq\theta \qquad (2.9)$$

$$\inf_{t_1<|x-\theta|<t_2} |p'(x)| > 0 \qquad \text{for } 0<t_1<t_2< \infty \qquad (2.10)$$

Define a sequence of r.v.s μ_n by $\mu_1=x$ (x a r.v. with density $p(\cdot)$) and

$$\mu_{n+1} = \mu_n - a_n(\mu_n-y_n)$$

where y_n is an observation on $Y_{\mu_n}^n$ and a_n is a sequence of positive numbers satisfying, along with L_n,

$$L_n \to 0$$

$$\sum a_n L_n^2 = \infty \qquad (2.11)$$

$$\sum a_n^2 L_n^2 < \infty.$$

Under these circumstances, $\mu_n \to \theta$ as $n \to \infty$ with probability one and in quadratic mean.

<u>REMARKS</u> Condition (2.9) and (2.10) are again unimodality conditions of a considerably stronger nature than the earlier unimodal on the average. We are no longer permitted the luxury of averaging over a set, but must have a very strict unimodality: $p(\cdot)$ differentiable and $p'(x)<0$ for $x>\theta$, and $p'(x)>0$ for $x<\theta$.

This algorithm has been presented in a slightly different form from THEOREM 1. An equivalent sequence μ_n and y_n can be derived from a sequence $x_1, x_2, \ldots$ of i.i.d. random variables with common density $p(\cdot)$ as follows:

$$\mu_1 = x_1$$

$$\mu_n = \mu_n - a_n(\mu_n - y_{n+1})$$

where

$$y_n = x_{j_n} \qquad \text{where } j_0 = 1 \text{ and for } m = 1, 2, \ldots$$

$$j_{m+1} = \min_{j_m < i} \{i \mid |\mu_{m+1} - x_i| < L_m\}$$

To prove the theorem, we need the following lemma:

<u>LEMMA</u> Let the density $p(\cdot)$ be differentiable a.e., and assume $(x-\Theta)p'(x)<0$ for $x\neq\Theta$. Then for every $L>0$,

$$(\xi-\Theta)\Big(\xi - \frac{1}{\int_{\xi-L}^{\xi+L} p(x)} \int_{\xi-L}^{\xi+L} xp(x)\, dx\Big) > 0 \qquad |\xi-\Theta|>L \qquad (2.12)$$

Proof of lemma: Write $p(\cdot)$ as

$$\begin{aligned} p(x) &= p_1(x) \qquad x \geq \Theta \\ &= p_2(x) \qquad x < \Theta \end{aligned}$$

where p_1 and p_2 are symmetric densities. By the nature of p, p_1 and p_2 are strictly unimodal. For $\xi-\Theta>L$, (2.12) involves only p_1, and is thus valid (by THEOREM 4), and similarly for $-\xi+\Theta>L$, (2.12) involves only p_2.

Proof of THEOREM 8: We take $\Theta=0$ without loss of generality. Define the function

$$f_n(\xi) = E[\xi - Y_\xi^n] \qquad \text{and note that}$$

$$|f_n(\xi)| < L_n \qquad \text{all } \xi, n. \qquad (2.13)$$

Also,

$$\mathrm{Var}[\xi - Y_\xi^n] < 4L_n^2 \quad \text{all } \xi, n \qquad (2.14)$$

and using the lemma

$$\xi f_n(\xi) > 0 \qquad \text{for } |\xi|>L_n. \qquad (2.15)$$

In addition, $f_n(\cdot)$ can be expanded as a series in the form

$$f_n(\xi) = f(\xi)L_n^2 + g(\xi,L_n^2) \tag{2.16}$$

where $\frac{g(\xi,L_n^2)}{L_n^2} \to 0$ as $L_n \to 0$. By writing $p(\cdot)$ as a Taylor series and evaluating $f_n(\cdot)$ we deduce that

$$f(\xi) = -\frac{p'(\xi)}{p(\xi)}$$

and thus by (2.9) and (2.10) we have

$$\xi f(\xi) > 0 \qquad \text{for } \xi \neq \theta \text{ , } \xi \text{ in support of } p \tag{2.17}$$

$$\inf_{t_1<|\xi-\theta|<t_2} |f(\xi)| > 0 \qquad 0<t_1<t_2<\infty \quad . \tag{2.18}$$

We will assume that

$$\frac{g(\xi,L_n^2)}{L_n^2} \to 0 \quad \text{as } L_n \to 0$$

uniformly for ξ in a compact set. This is true, for example, if $p'(\cdot)$ is continuous, in which case (2.9) implies (2.10).

With these facts, the theorem can now be proven as a consequence of Dvoretsky's [9] general stochastic approximation theorem, which appears in the appendix. We show that A - D are satisfied.

Rewrite

$$\mu_{n+1} = \mu_n - a_n f_n(\mu_n) - Z_n = T_n(\mu_n) - Z_n$$

where

$$Z_n = a_n[\mu_n - Y^n_{\mu_n} - f_n(\mu_n)].$$

By the definition of f_n, $EZ_n = 0$ and thus D is satisfied. (2.14) yields $EZ_n^2 < 4a_n^2L_n^2$ and by (2.11) $\sum EZ_n^2 < \infty$, thus C is satisfied, since we have assumed that $E\mu_1^2 = Ex^2 < \infty$.

To verify that A is satisfied for the transformation

$$T_n(\mu_n) = \mu_n - a_nf_n(\mu_n) ,$$

we let b_n be a sequence of positive numbers tending to zero such that

$$\sum a_nL_n^2b_n = \infty . \tag{2.19}$$

Define $\rho_n = L_n^2b_n$. We define the sequence η_n of positive numbers as

$$\inf_{\eta_n \leq |\xi| \leq 1} |f_n(\xi)| > \rho_n.$$

To show that η_n can be taken to go to zero, we remark that if any function $f(\cdot)$ satisfies (2.18) then if $\beta_n \to 0$, one can choose a sequence $\eta_n \to 0$ such that

$$\inf_{\eta_n \leq |\xi| \leq 1} |f(\xi)| > \beta_n .$$

To apply this to our situation define $\beta_n = \frac{\rho_n}{L_n^2} + \gamma_n$ where

$$\gamma_n = \sup_{\eta_n < |\xi| < 1} \frac{g(\xi, L_n^2)}{L_n^2}$$

Note that regardless of η_n, $\beta_n \to 0$ by the uniformity assumption. By construction $\rho_n/L_n^2 \to 0$ and thus $\beta_n \to 0$. Now using the above remark we define $\eta_n \to 0$ by

$$\inf_{\eta_n \le |\xi| \le 1} |f(\xi)| > \beta_n \quad .$$

This means

$$\inf_{\eta_n \le |\xi| \le 1} |f(\xi)| - \sup_{\eta_n \le |\xi| \le 1} |g(\xi,L_n^2)/L_n^2| > \rho_n/L_n^2$$

which implies

$$\inf_{\eta_n \le |\xi| \le 1} [\ |f(\xi) \ - \ g(\xi,L_n)/L_n^2|\ ] > \rho_n/L_n^2$$

which in turn implies

$$\inf_{\eta_n \le |\xi| \le 1} [\ |f(\xi)L_n^2 + g(\xi,L_n^2)|] > \rho_n$$

as desired. By using this sequence, and breaking up the range of the argument of $T_n(\cdot)$ we show that condition A is satisfied:

If $|r_n| \le L_n$ then

$$|T_n(r_n)| = |r_n - a_n f_n(r_n)| \le L_n + a_n L_n \ .$$

If $|r_n| \ge L_n$ then we know r_n and $f_n(r_n)$ have the same sign (2.15) and if $r_n < a_n f(r_n)$ then

$$|r_n - a_n f_n(r_n)| = a_n |f_n(r_n)| - |r_n| < a_n L_n \quad .$$

Alternatively, if $a_n f_n(r_n) < r_n$ we have

$$\begin{aligned} |r_n - a_n f_n(r_n)| &= |r_n| - a_n |f_n(r_n)| \\ &< |r_n| - a_n \rho_n && \text{for } 1 \geq |r_n| \geq \eta_n \\ &< \eta_n && \text{for } |r_n| < \eta_n \\ &< |r_n| - a_n \rho_n && \text{for } M \geq |r_n| \geq \eta_n \\ & && \text{and } n > n_M . \end{aligned}$$

In any event, for n large,

$$|T_n(r_n)| < \max \, [L_n + a_n L_n \; , \; \eta_n, \; |r_n| - a_n \rho_n]$$

and $L_n + a_n L_n \to 0$, $\eta_n \to 0$, and using (2.19), $\sum a_n \rho_n = \infty$. Thus A is satisfied for γ_n as in the extension. This completes the proof.

This section, plus section 2, contain all the basic convergence properties which have been found regarding these two mode seeking algorithms. To summarize, it has been shown that under very reasonable,conditions--unimodal on the average, which include most "unimodal" densities (THEOREMS 3 to 7)--the fixed window algorithm converges to a value Θ which is at most a known distance from the true mode (THEOREM 1) If more accuracy is desired, the shrinking window algorithm converges to the true mode for densities which are well behaved and strictly unimodal (THEOREM 8). Thus accuracy can be traded for generality of result by choosing one or the other of these algorithms.

Section 5. Rates of Convergence.

In this section the rates of convergence of the two mode seeking algorithms are discussed. In addition, some comments are made about the asymptotic distribution of the r.v. μ_n. Many of these results are based upon known characteristics of stochastic approximation methods.

The following theorem considers the original fixed window algorithm.

<u>THEOREM 9</u> Let $\mu_1, \mu_2, \ldots$ be the sequence of random variables derived from the fixed window mode seeking algorithm as defined in section 2. Let the underlying density $p(\cdot)$ be such that the function

$$f(\xi) = \xi - \frac{1}{\int_{\xi-L}^{\xi+L} p(x)} \int_{\xi-L}^{\xi+L} xp(x)\, dx$$

satisfies

$$(\xi-\theta) f(\xi) > 0 \qquad \text{if } \xi \neq \theta \tag{2.20}$$

$$f(\xi) = \beta_1(\xi-\theta) + o(\xi-\theta) \qquad \text{for } \beta_1 > 0 . \tag{2.21}$$

Let the weighting sequence a_n be of the form $a_n = a/n$ with $2a > 1/\beta_1$. In this case, $n^{.5}(\mu_n - \theta)$ is asymptotically normal with zero mean and variance $a^2\sigma^2/(2a\beta_1 - 1)$ where $\sigma^2 = \text{Var } Y_\Theta$.

Rðoof: This is an application of Sacks'[14] theorem on asymptotic distribution, which appears for reference in the appendix. Our $f(\cdot)$ is his $M(\cdot)$ and we must verify that $f(\xi) \leq K(\xi-\Theta)$ for some $K>0$. This is a conseqùence of (2.21) and $|f(\xi)| \leq L$. Finally $E(\xi-Y_\xi)^{2+\nu} < \infty$ since $|\xi-Y_\xi|<L$ a.s.

This result demonstrates that the random variable μ_n has a variance about the mode Θ which decreases as $1/n$, where n is the number of samples x_i which have been averaged into the current estimate μ_n. The more interesting question from a comptational standpoint is, of course, how does the variance behave as a fuction of $n°$, the number of samples either used <u>or</u> discarded. (Remember that the random variable Y_ξ is derived from x by discarding all samples which do not fall in the interval $(\xi-L,\xi+L)$). Fortunately $n°$ is related by a constant factor to n, at least asymptotically. It is easy to show that, for large values of n,

$$n° \int_{\Theta-L}^{\Theta+L} p(x)\, dx \cong n \quad .$$

Thus the variance of μ_n decreases as $1/n°$ as well.

The question of when (2.21) holds arises immediately. The following result is appropriate:

<u>THEOREM 10</u> Let $p(\cdot)$ satisfy

$$K_L = \frac{L[p(\Theta+L) + p(\Theta-L)]}{\int_{\Theta-L}^{\Theta+L} p(x)} < 1 \qquad (2.22)$$

Then (2.21) holds with $\beta_1 = 1 - K_L$.

Proof: A Taylor series expansion of $f(\cdot)$ can easily be found to be

$$f(\xi) = [\ 1 - \frac{L[p(\Theta+L) + p(\Theta-L)]}{\int_{\Theta-L}^{\Theta+L} p(x)}\ ](\xi-\Theta) + o(\xi-\Theta)$$

and the result follows.

The following result covers the rate of convergence of the shrinking window algorithm.

THEOREM 11 Let $\mu_1, \mu_2, \ldots$ be the sequence of random variables derived from the shrinking window mode seeking algorithm. Assume the function

$$f(\xi) = \xi - \frac{1}{\int_{\xi-L}^{\xi+L} p(x)} \int_{\xi-L}^{\xi+L} xp(x)\ dx$$

satisfies

$$AL^2 \leq f(\xi)/(\xi-\Theta) \quad \text{for some } A>0. \qquad (2.23)$$

Let $a_n = a/n^\alpha$ and let $a_n L_n^2 = b/n$ with $Ab>1$.

Under these circumstances,

$$E(\mu_n - \Theta)^2 \leq K/n \quad \text{for some } K<\infty. \tag{2.24}$$

Proof: This proof follows the lines of the proof of a simple Robbins Munro stochastic approximation theorem appearing in Sakrison[16]. By definition,

$$\mu_{n+1} = \mu_n - a_n(\mu_n - y_n)$$

Thus, taking $\Theta=0$,

$$E(\mu_{n+1}^2 | \mu_n=\xi) = \xi^2 - 2a_n\xi E(\mu_n - y_n | \mu_n=\xi)$$
$$+a_n^2 E[(\mu_n-y_n)^2 | \mu_n=\xi]$$

Using (2.23) plus $E[(\mu_n-y_n)^2|\mu_n=\xi] \leq 4L_n^2$ we obtain

$$E[\mu_{n+1}^2 | \mu_n=\xi] \leq (1 - 2Aa_nL_n^2)\xi^2 + 4a_n^2L_n^2 \quad .$$

Taking unconditional expectation and denoting $b_n = E\mu_n^2$ we have

$$b_{n+1} = b_n(1 - 2Aa_nL_n^2) + 4a_n^2L_n^2 \quad .$$

We may iterate this to obtain

$$b_n = b_1\beta_{1,n-1} + 4\sum_{k=1}^{n-1} a_k^2L_k^2\beta_{k,n-1} \tag{2.25}$$

where

$$\beta_{k,n} = \sum_{j=k+1}^{n} (1 - 2Aa_jL_j^2) \qquad 0\leq k\leq n$$
$$= 0 \qquad\qquad k\geq n \quad .$$

Taking logs of (2.26) and using the inequality

$$\log(1 - x) \le -x$$

we may show

$$\beta_{m,n} \le \exp\{-2A \sum_{j=m+1}^{n} a_j L_j^2\} \ ,$$

Directly approximating

$$\sum_{j=m+1}^{n} a_j L_j^2 = \sum_{j=m+1}^{n} b/j \qquad \text{with } Ab > 1$$

yields

$$\beta_{m,n} \le (m+1/n)^2 \ .$$

Similarly the second term in (2.25) can be bounded:

$$4 \sum_{k=1}^{n-1} a_k^2 L_k^2 \beta_{k,n-1} \le \frac{4ab}{n^2} \sum_{k=1}^{n-1} \frac{(k+1)^2}{k^{1+\alpha}}$$

$$\le C/n^{\alpha} \ .$$

Thus b_n goes to zero at least as fast as $1/n^{\alpha}$; and (2.24) follows.

To explain the relationship between the previously defined n and n°, for the shrinking window algorithm, the following theorem has been proven.

THEOREM 12 Let $a_n = a/n^{\alpha}$ and $L_n = L/n^{(1-\alpha)/2}$. Then THEOREM 11 states that $E(\mu_n-\Theta)^2 \to 0$ with a $1/n^{\alpha}$ rate. Let $n°$ be the number of samples either used or discarded before the nth sample is averaged in. Then, for large n,

$$n° \approx K(n+1)^{(3-\alpha)/2} \tag{2.28}$$

and therefore a $1/n^{\alpha}$ rate is a $1/n°^{2\alpha/(3-\alpha)}$ rate, and if $\alpha \in (0,1)$ then $2\alpha/(3-\alpha) \in (0,1)$.

Proof: The expected number of samples used between the ith and (i+1)st computations, given μ_i, is

$$1/P_i = 1/\int_{\mu_i-L_i}^{\mu_i+L_i} p(x)\ dx \quad .$$

Thus at the nth step in the algorithm we have used or discarded approximately

$$n° = \sum_{i=1}^{n} 1/P_i$$

samples. Since $\mu_n \to \theta$, it can be shown that if $L_n \to 0$, and $p(\cdot)$ continuous then

$$P_i/P_{\theta_i} = [\int_{\mu_i-L_i}^{\mu_i+L_i} p(x)\ ]/[\int_{\theta-L_i}^{\theta+L_i} p(x)\ ] \to 1 \ .$$

This result can be used to show that

$$[\ \sum_{i=1}^{n} 1/P_i\ ]/[\ \sum_{i=1}^{n} 1/P_{\Theta_i}\ ] \to 1$$

and thus

$$n^\circ \approx \sum_{i=1}^{n} 1/P_{\theta_i}$$

for large n. (This is the expected number of samples used or discarded if the window had been centered on θ , instead of μ_i.) This yields

$$n^\circ \approx \sum_{i=1}^{n} 1/ \; [\int_{\theta-L_i}^{\theta+L_i} p(x) \;] \approx \sum_{i=1}^{n} 2L_i p(\theta+L_i)$$

$$\approx \frac{1}{2p(\theta)} \sum_{i=1}^{n} 1/L_i \quad .$$

If $L_i = 1/i^{(1-\alpha)/2}$ then

$$n^\circ \approx K \sum_{i=1}^{n} i^{\alpha} \leq K'(n+1)^{\alpha+1}$$

which is (2.28). This completes the proof.

CHAPTER III

MODE SEEKING IN HIGHER DIMENSIONS

Section 1. Introduction.

Since the motivation for mode seeking algorithms is pattern classification, it is particularly important to extend the results of Chapter II to higher dimensions. Pattern classification is àlmost always done in higher dimensional spaces, since in practical situations many measurements are made on a single object, making up a vector in R^k with $k>1$. Thus any tool useful for pattern classification must operate in high dimensional spaces. This chapter extends most of the results of the previous section to the estimation of modes in probability distributions over R^k, and this extension is a most natural one. We assume R^k is an inner product space with inner product

$$<x,y> = x^T y = \sum x_i y_i \qquad \text{where } x = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} .$$

and norm

$$|x| = <x,x>^{1/2}$$

Section 2. The fixed window algorithm.

THEOREM 13 Let $x_1, x_2, \ldots$ be a sequence of independent identically distributed random variables taking values in R^k, each with a fixed probability density $p(\cdot)$ over R^k. Let S be a fixed bounded set, and derive two new sequences $y_1, y_2, \ldots$ and $\mu_1, \mu_2, \ldots$ as follows:

$$\mu_1 = x_1$$

For $n = 1,2,3 \ldots$ define

$$y_n = x_{j_n} \quad \text{where} \quad j_1 = 1 \quad \text{and for } m = 1,2, \ldots$$

$$j_{m+1} = \min_{j_m < i} \left\{ i \mid (\mu_m - x_i) \in S \right\}$$

and

$$\mu_{n+1} = \mu_n - a_n(\mu_n - y_n)$$

for a sequence a_n of positive numbers.

Let the density $p(\cdot)$ be such that the function

$$f(\xi) = \xi - \frac{1}{\int_{S+\xi} p(x)} \int_{S+\xi} xp(x)\, dx$$

satisfies

$$\langle \xi - \Theta, f(\xi) \rangle > 0 \qquad \text{if } \xi \neq \Theta,\ \xi \text{ in support } p \tag{3.1}$$

Let a_n satisfy

$$\sum a_n = \infty \qquad \sum a_n^2 < \infty \tag{3.2}$$

Then $|\mu_n - \Theta| \to 0$ as $n \to \infty$ with prob. one.

<u>REMARK</u> Condition 3.1 is an obvious analog of 2.1, and again can be denoted as <u>unimodal on the average</u>. The extension is a natural one, since if we take $k = 1$ and $S = [-L,+L]$ for $L>0$ then this theorem reduces to the previous one. In this sense S is the k dimensional generalization of the "window" $[-L,+L]$. S may be arbitrary, though the particular set $S = [-L,+L]^k$ will be seen several times here.

Proof of THEOREM 13: As before, condition (3.1) is a requirement on the random variable $Y_\xi = y_n$ given $\mu_{n-1}=\xi$. Y_ξ has the distribution

$$p_{Y_\xi}(x) = \frac{p(x)}{\int_{S+\xi} p(y)} \qquad \text{for } x-\xi \in S$$

$$= 0 \qquad \text{elsewhere}$$

Thus the function $f:R^k \to R^k$ can be written as

$$f(\xi) = \xi - EY_\xi.$$

The proof of the above theorem follows, as before, from the following stochastic approximation result (See Sacks[14]).

<u>THEOREM 14</u> Let Z_μ be a one parameter family of random variables in R^k, with $\mu \in R^k$, and define $g(\mu) = EZ_\mu$. Assume $g(\mu)$ exists for every μ. Define a sequence x_n as x_1 arbitrary and $x_{n+1} = x_n - a_n z_n$ where z_n is an observation

on Z_{x_n} and where a_n is a sequence of positive numbers satisfying $\sum a_n = \infty$, $\sum a_n^2 < \infty$. Let Z_μ have uniformly bounded variances and assume $g(\cdot)$ satisfies

$$|g(\mu)| < K_1|\mu-\theta| \tag{3.3}$$

$$<g(\mu),\mu-\theta> > 0 \qquad \text{for } \mu\neq\theta \quad \text{for some } \theta, K_1>0. \tag{3.4}$$

$$\inf_{t_1\leq|\mu-\theta|\leq t_2} |g(\mu)| > 0 \qquad \text{for } 0<t_1<t_2<\infty \quad . \tag{3.5}$$

Then $x_n \to \theta$ with probability one.

To apply this result, we define $Z_\xi = \xi - Y_\xi$ and note $f(\xi) = EZ_\xi$. We then identify $f(\cdot)$ with the $g(\cdot)$ appearing in THEOREM 14 and conclude that (3.1) guarantees that (3.4) and (3.5) hold, since $f(\cdot)$ is continuous. We note also that since S is bounded, $|\xi-Y_\xi| \leq K$ w.p.1 for some $K>0$ and therefore $|f(\xi)| \leq K$. Thus to check (3.3) we need only show that for some $\varepsilon>0$,

$$|f(\xi)| < K_2|\xi-\theta| \qquad \text{for } |\xi| < \varepsilon \tag{3.6}$$

Since $f(\cdot)$ is differentiable, (3.6) holds if the Jacobian of f at θ has finite entries, that is

$$\left|\frac{\partial f^i(\xi_1, : : \xi_k)}{\partial \xi_j}\right|_{\xi=\theta} < \infty \quad .$$

This is easily verified. The Jacobian is explicitly exhibited in the proof of THEOREM 19 (Section 5 of this chapter).

Section 3. Unimodal on the average examined.

The significance of condition (3.1) is considered here. Generally, the results of Chapter II, Section 3 carry over. As before, if $p(\cdot)$ is a spherically symmetric unimodal density then $p(\cdot)$ is *almost* unimodal on the average.

THEOREM 15 Let $p(\cdot)$ satisfy $p(x) \leq p(y)$ for all $|x-\theta| \geq |y-\theta|$. Then for any symmetric set S, $f(\cdot)$ satisfies

$$\langle \xi-\theta, f(\xi) \rangle \geq 0 \ .$$

Proof. Consider $\theta=0$. We must show

$$\xi^T \left[\frac{1}{\int_{S+\xi} p(y)} \int_{S+\xi} xp(x) \ dx \right] \leq |\xi|^2 \quad .$$

A change of variables iddicates that we need only show

$$\xi^T \int_S xp(x+\xi) \ dx \leq 0.$$

Define the set $S_\xi^\circ = \{x \in S| \quad \xi^T x \geq 0\}$ and note that since S is symmetric,

$$\xi^T \int_S xp(x+\xi) \ dx = \int_{S_\xi^\circ} \xi^T x \ [p(x+\xi) - p(-x+\xi)] \ dx \ .$$

Since for all $x \in S_\xi^\circ$, $\xi^T x \leq 0$ and $|x+\xi| \geq |-x+\xi|$ the unimodality assumption yields

$$p(x+\xi) - p(-x+\xi) \leq 0$$ which completes the proof.

Unimodal on the average is obtained if $p(\cdot)$ is spherically symmetric and strictly unimodal as follows:

THEOREM 16 If $p(\cdot)$ satisfies $p(x) = p(y)$ whenever $|x-\Theta|=|y-\Theta|$ and $p(x) < p(y)$ whenever $|x-\Theta|>|y-\Theta|$ then for any symmetric set S, $f(\cdot)$ satisfies

$$<\xi-\Theta, f(\xi)> > 0 \qquad \xi\neq\Theta \;.$$

Proof: Following the last proof, we show that

$$\xi^T \int_S xp(x+\xi)\, dx > 0 \qquad \text{for } \xi\neq\Theta \;.$$

Define $S^{\circ}_{\xi} = \{x \in S|\ \xi^T x > 0\}$ and note that since S is symmetric

$$\xi^T \int_S xp(x+\xi)\, dx = \int_{S^{\circ}_{\xi}} \xi^T x\, [p(x+\xi) - p(-x+\xi)]\, dx \quad .$$

Since for all $x \in S^{\circ}_{\xi}$, $\xi^T x > 0$, and since $|x+\xi|>|-x+\xi|$ for all $x \in S^{\circ}_{\xi}$, the result follows.

Section 4. The shrinking window algorithm.

The shrinking window algorithm extends to R^k in an obvious manner as follows:

<u>THEOREM 17</u> Let L_n be a sequence of positive numbers. For a fixed probability density $p(\cdot)$ over R^k define a random variable Y_ξ^n , $\xi \varepsilon R^k$, taking on values in R^k by its density:

$$p_{Y_\xi^n}(x) = \frac{p(x)}{\int_{S_n+\xi} p(y)} \qquad x-\xi \ \varepsilon \ S_n$$

$$= 0 \qquad \text{elsewhere}$$

where $S_n = [-L_n,+L_n]^k$. Assume $p(\cdot)$ is differentiable with gradient $\nabla p(\cdot)$ defined a.e. which satisfies

$$<x-\Theta,\nabla p(x)/p(x)> < 0 \qquad \text{for } x \neq \Theta \text{ and } x \text{ in the support of } p. \qquad (3.7)$$

$$\inf_{t_1 \leq |x-\Theta| \leq t_2} |\nabla p(x)| > 0 \qquad 0<t_1<t_2<\infty \qquad (3.8)$$

$$|\nabla p(x)/p(x)| \leq K|x-\Theta| \qquad . \qquad (3.9)$$

Define a sequence of random variables μ_n in R^k by $\mu_1 = x$ (x a r.v. with density $p(\cdot)$) and

$$\mu_{n+1} = \mu_n - a_n(\mu_n - y_n)$$

where y_n is an observation on $Y_{\mu_n}^n$, and a_n is a sequence which satisfies, along with L_n,

$$\begin{aligned} &L_n \to 0 \\ &\sum a_n L_n^2 = \infty \\ &\sum a_n^2 L_n^2 < \infty \quad . \end{aligned} \tag{3.10}$$

Under these circumstances, $\mu_n \to \Theta$ as $n \to \infty$ with prob. one.

<u>REMARK</u> Again(3.7) and (3.8) are logical extensions of their one dimensional counterparts (2.9) and (2.10). All are clearly unimodality requirements.

Proof of THEOREM 17: We will follow the one dimension proof fairly closely, using the Sacks and Derman[15] extension of Dvoretsky's theorem. We again take $\Theta=0$.

Let b_n be the sequence of positive numbers tending to zero such that

$$\sum a_n L_n^2 b_n = \infty \tag{3.11}$$

and let $\rho_n = L_n^2 b_n$. Define $\eta_n \to 0$ as

$$\inf_{\eta_n \leq |\xi| \leq 1} \frac{<\xi, f_n(\xi)>}{|\xi|} > \rho_n (1+K' a_n^2 L_n^2)^{1/2} = \rho_n' \tag{3.12}$$

where $f_n(\cdot)$ is defined as

$$f_n(\xi) = \xi - \frac{1}{\int_{S_n+\xi} p(x)} \int_{S_n+\xi} xp(x)\ dx \tag{3.13}$$

that is,

$$f_n(\xi) = E(\xi - Y_\xi^n) \ . \tag{3.14}$$

We can expand $f_n(\cdot)$ as follows:

$$f_n(\xi) = h(\xi)L_n^2 + g(\xi,L_n^2) \tag{3.15}$$

where again we have

$$|g(\xi,L_n^2)/L_n^2| \to 0 \quad \text{as } L_n \to 0,$$

and we <u>assume</u> that such convergence is uniform on a compact set of ξ's. Again this will certainly be true if ∇p is continuous, in which case (3.7) implies (3.8). By expanding the density $p(\cdot)$ as a Taylor series and identifying $h(\cdot)$ as the coefficient of the L_n^2 term, we find

$$h(\xi) = - p'(\xi)/p(\xi) \tag{3.16}$$

and (3.7) and (3.8) yield

$$<\xi,h(\xi)> > 0 \quad \text{for} \quad \xi \neq 0 \tag{3.17}$$

$$\inf_{t_1 \leq |\xi| \leq t_2} |h(\xi)| > 0 \qquad 0<t_1<t_2<\infty \quad . \tag{3.18}$$

To show the η_n defined above can be taken to go to zero, we define $\beta_n = \rho_n'/L_n^2 + \gamma_n$ where

$$\gamma_n = \sup_{\eta_n \leq |\xi| \leq 1} \frac{|<\xi,g(\xi,L_n^2)>|}{|\xi|} \leq \sup_{\eta_n \leq |\xi| \leq 1} |g(\xi,L_n^2)| .$$

Such a $\gamma_n \to 0$ by the uniformity assumption. Thus $\beta_n \to 0$.

We now *define* $\eta_n \to 0$ as

$$\inf_{\eta_n \le |\xi| \le 1} \frac{\langle \xi, h(\xi)\rangle}{|\xi|} > \beta_n$$

which can always be done when h(•) satisfies (3.17) and (3.18). Rearranging we obtain

$$\inf_{\eta_n \le |\xi| \le 1} \frac{\langle \xi, h(\xi)L_n^2 + g(\xi, L_n)\rangle}{|\xi|} > \rho_n'$$

as desired. Finally note that (3.9) implies that we can find a sequence $k_n \to 0$ such that for $k_n \le |\xi|$,

$$\frac{|f_n(\xi)|}{L_n^2} \le K'|\xi| \ . \tag{3.19}$$

We can now use the preceding and

$$|f_n(\xi)| \le K_1 L_n \tag{3.20}$$

to show that A to D are satisfied. We define $T_n(\cdot)$ by

$$\mu_{n+1} = \mu_n - a_n f_n(\mu_n) - Z_n = T_n(\mu_n) - Z_n$$

where $Z_n = a_n[\mu_n - Y^n_{\mu_n} - f_n(\mu_n)]$. By (3.14), D is satisfied. C follows by noting that since $|\xi - Y^n_\xi| < K_1 L_n$ with prob. one, $EZ_n^2 < K_2 a_n^2 L_n^2$, and (3.10) yields $\sum EZ_n < \infty$. Finally we have

$$|T_n(r_n)|^2 = |r_n - a_n f_n(r_n)|^2$$

$$= |r_n|^2 - 2a_n<r_n,f_n(r_n)> + a_n^2|f_n(r_n)|^2 .$$

If $\max(k_n,\eta_n) \leq |r_n| \leq 1$ we have from (3.19) and (3.12)

$$|T_n|^2 \leq |r_n|^2(1+K'a_n^2L_n^2) - 2a_n\rho_n'|r_n| \quad .$$

This yields

$$|T_n| \leq |r_n|(1+K'a_n^2L_n^2)^{1/2} - a_n\rho_n$$

$$\leq |r_n|(1+K'a_n^2L_n^2) - a_n\rho_n \quad . \qquad (3.21)$$

Using (3.10) and (3.11) we see A is satisfied for this range of r_n. For $|r_n| < \max(k_n,\eta_n)$ we use (3.20) to obtain

$$|T_n| \leq \max(k_n,\eta_n) + K_1a_nL_n \ . \qquad (3.22)$$

A check of the proof shows that (3.21) holds if $\eta_n \leq |r_n| \leq M$ for M arbitrary, and by the extension, the result follows.

Section 5. Rates of Convergence.

This section discusses the rate at which $E|\mu_n-\Theta|^2$ goes to zero for both the fixed window and the shrinking window algorithms.

<u>THEOREM 18</u> Let $\mu_1,\mu_2,$ . . be the sequence of random variables derived from the fixed window procedure as defined in THEOREM 13. Let $p(\cdot)$ be such that the function

$$f(\xi) = \xi - \frac{1}{\int_{S+\xi} p(x)} \int_{S+\xi} xp(x)\ dx$$

satisfies

$$<\xi-\Theta, f(\xi)> \ > 0 \qquad \text{for } \xi\neq\Theta \tag{3.23}$$

$$f(\xi) = B(\xi-\Theta) + \delta(\xi-\Theta) \quad \text{for positive definite } B \text{ and } \frac{|\delta(\xi-\Theta)|}{|\xi-\Theta|} \to 0 \text{ as } |\xi-\Theta|\to 0. \tag{3.24}$$

Let the weighting sequence a_n be of the form $a_n = a/n$ with $2ab_k>1$, b_k being the smallest eigenvalue of B. Assume Cov $Y_\Theta = \pi$ for π non-negative definite. Under these circumstances, $n^{1/2}(\mu_n-\Theta)$ is asymptotically normal with zero mean and covariance Σ defined as follows:

Let b_1, . . . b_k be the eigenvalues of B in decreasing order. Write $B = PDP^{-1}$ where P is orthogonal and D is a diagonal matrix whose diagonal elements are b_1, . . b_k.

Define π°_{ij} as the i,jth element of $P^{-1}\pi P$. Then $\Sigma = PQP^{-1}$ where the i,jth element of Q is

$$a^2(ab_i + ab_j - 1)^{-1}\pi^{\circ}_{ij} \quad .$$

Proof: This follows from a theorem due to Sacks [14] which appears in the appendix. First we verify that

$$|f(\xi)| \leq K|\xi-\Theta|$$

which follows from (3.24) and $|f(\xi)| \leq K'$. Finally, since S is bounded, $|\xi-Y_\xi| < K''$, and thus

$$E|\xi-Y_\xi|^{2+\nu} < K'' \qquad \text{for } \nu \geq 0.$$

The question arises as to what densities satisfy (3.24). This requires examing the Jacobian of $f(\cdot)$, since $B = \{b_{ij}\}$ where

$$b_{ij} = \left.\frac{\partial f^i(\xi)}{\partial \xi_j}\right|_{\xi=\Theta} \quad .$$

If we assume for $S = [-L,+L]^k$ that $p(\cdot)$ has some symmetry, then B will be diagonal, and (2.24) is easily verified.

<u>THEOREM 19</u> Let $p(\cdot)$ satisfy, for some $L>0$, and all $i=1, \ldots k$,

$$p(x_1, \ldots, \underset{i^{th}}{\underset{\uparrow}{L}}, \ldots x_k) = p(x_1, \ldots, \underset{i^{th}}{\underset{\uparrow}{-L}}, \ldots x_k) \ .$$

Then for $S = [-L,+L]^k$, B is diagonal, and is positive definite if and only if

$$2L \int_{-L}^{+L} \cdots \int_{-L}^{+L} p(x_1, \ldots, \underset{i\text{th}}{\underset{\uparrow}{L}}, \ldots x_k) < \int_S p(x)\, dx \tag{3.25}$$

for all $i = 1, \ldots k$.

Proof: This simply requires evaluating the various partial derivatives of $f(\cdot)$. Thus: for $i \neq j$,

$$b_{ij} = 1 - \frac{1}{\Pr S} \underbrace{\int_{-L}^{+L} \cdots \int_{-L}^{+L} \quad \int_{-L}^{+L} \cdots \int_{-L}^{+L}}_{\substack{\uparrow \\ j\text{th} \\ \text{missing}}} g_{ij}(x)\, dx$$

where

$$g_{ij}(x) = x_i [p(x_1 \ldots, \underset{j\text{th}}{\underset{\uparrow}{L}}, \ldots x_k) - p(x_1 \ldots, \underset{j\text{th}}{\underset{\uparrow}{-L}}, \ldots x_k)],$$

$$b_{ii} = 1 - \frac{L}{\Pr S} \underbrace{\int_{-L}^{+L} \cdots \int_{-L}^{+L}}_{\substack{\uparrow \\ i\text{th} \\ \text{missing}}} h_i(x)\, dx$$

where

$$h_i(x) = p(x_1, \ldots, \underset{i\text{th}}{\underset{\uparrow}{L}}, \ldots x_k) + p(x_1, \ldots, \underset{i\text{th}}{\underset{\uparrow}{-L}}, \ldots x_k).$$

By the symmetry assumption, $b_{ij} = 0$, thus B is diagonal. (3.25) guarantees that $b_{ii} > 0$, which insures that B is positive definite. Note that b_{ij} are always finite, a result used in THEOREM 14.

Turning to the shrinking window algorithm, we have the following result on rate of convergence:

<u>THEOREM 20</u> Let μ_1, μ_2 be the sequence of random variables derived from the shrinking window algorithm. Let the function $f(\cdot)$ defined in THEOREM 18 with $S = [-L,+L]^k$ satisfy

$$AL^2 \leq \frac{<\xi-\theta, f(\xi)>}{|\xi-\theta|^2} \quad \text{for some } A>0. \tag{3.26}$$

Let the sequences a_n and L_n satisfy

$$a_n L_n^2 = K/n \qquad a_n = a/n^\alpha \quad .$$

If $AK > 1$, then

$$E|\mu_n - \theta|^2 < K_1/n^\alpha \quad .$$

Proof: The proof follows the one dimension proof of THEOREM 11 (Chapter II, Section 5) exactly. Assuming $\theta=0$ we write

$$|\mu_{n+1}|^2 = |\mu_n - a_n(\mu_n - y_n)|^2$$

$$= |\mu_n|^2 - 2a_n<\mu_n, \mu_n - y_n> + a_n|\mu_n - y_n|^2 \quad .$$

Taking conditional expectations we can bound the middle term as before, using (3.26), and the proof follows.

CHAPTER IV

COMPUTER SIMULATION

This chapter contains computational results derived from application of the mode seeking algorithms to computer generated data. In the first section, the variance of the estimates μ_n is computed for both of the algorithms discussed in previous chapters, and the results indicate that μ_n does converge in the bimodal situation as well as the unimodal. In the second section, a similar mode estimation procedure involving simultaneous estimation of all the modes is defined, and some computational results given. Finally, the third section presents some limited theoretical results on the multimodal problem.

Section 1. Program description and results.

The results of this section, summarized in Figures 4.1 to 4.4, show that the mode seeking algorithms do converge for both unimodal and bimodal data. The values used in these graphs are derived from a computer program which first generates a data set, and then implements either the fixed window or shrinking window algorithm and applies it to the data. The process is repeated several times to obtain estimates of the variance. Specifically:

1. The values A, NSH, BETA, L, N, and ITER are specified, and VAR is initially set to zero.
2. Samples are generated according to a density $p(x) = [p_1(x) + p_1(x - \mathrm{NSH})]/2$ where $p_1(\cdot)$ is a symmetric triangular density with base 10.
3. The samples are used sequentially in the mode estimation procedure $\mu_n = \mu_n - a_n(\mu_n - y_n)$ with $a_n = A/\ n^{1-\mathrm{BETA}/2}$ until N samples have been averaged into the estimate, using a window of initial length 2L, which is reduced to $2L/n^{\mathrm{BETA}}$ after n samples have been averaged in.
4. DIST is computed as the smaller of the distances of μ_N to 5, and μ_N to 5 + NSH, and the variance computed:

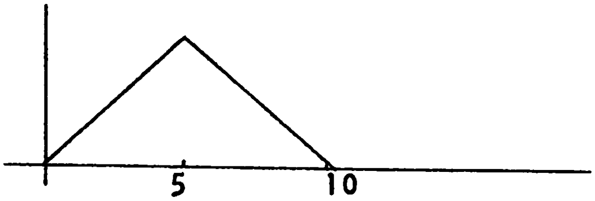
   $$\mathrm{VAR} = \mathrm{VAR} + (\mathrm{DIST})^2 \quad .$$
5. If the mode has been estimated ITER times, and thus VAR has ITER terms averaged in, then VAR/ITER, the sample variance, is printed. Otherwise go to 2.

The preceding is a multipurpose program. If BETA is zero, the so called fixed window algorithm is implemented, and if 0 < BETA < 1, then the shrinking window algorithm is used. Also, if NSH is zero, then the distribution is strictly unimodal, and if |NSH| > 5, then the density has two sharp peaks, one at 5, and the other at 5 + NSH.

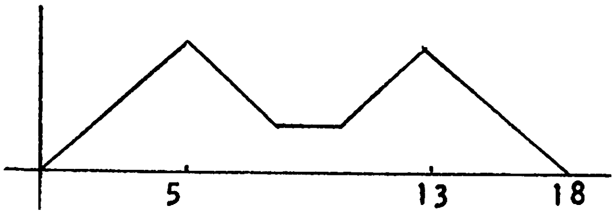
This general program has been run with all permutations of the basic options, as follows:

FIGURE	ALGORITHM TYPE	DATA TYPE
4.1	FIXED WINDOW	UNIMODAL
4.2	FIXED WINDOW	BIMODAL NSH = 8
4.3	SHRINKING WINDOW BETA = .2	UNIMODAL
4.4	SHRINKING WINDOW BETA = .2	BIMODAL NSH = 8

Thus the unimodal data has a density



and the bimodal data has density



For each of these four cases, the variance (VAR) is graphed on the following pages for values of N between 25 and 225. The results demonstrate two things: (1) the variance decreases roughly as predicted and (2) the estimates of the modes are in fact converging to the modes, since the variance is computed about the true modes.

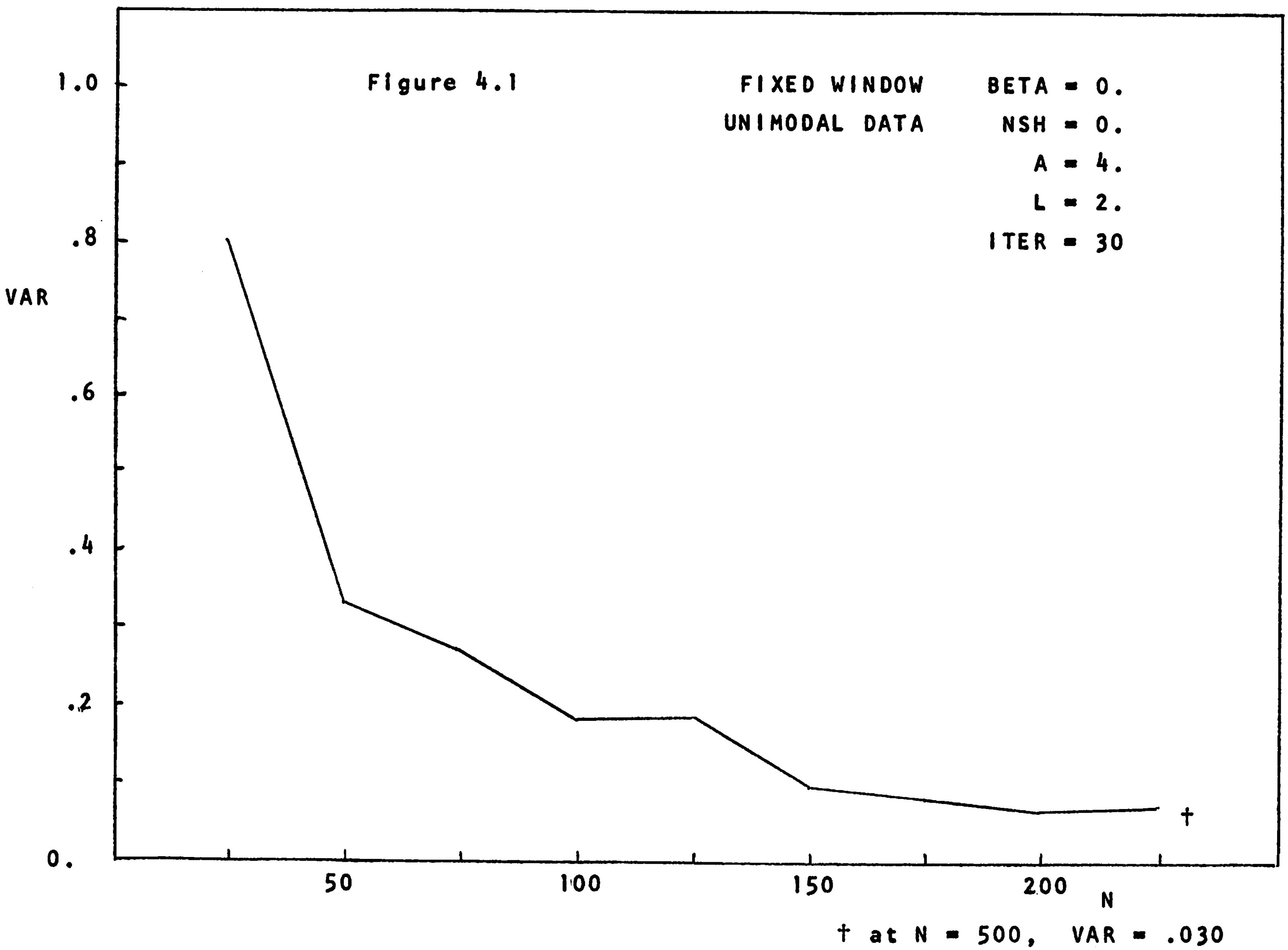


Figure 4.1

FIXED WINDOW
UNIMODAL DATA
BETA = 0.
NSH = 0.
A = 4.
L = 2.
ITER = 30
VAR
1.0
.8
.6
.4
.2
0.
50
100
150
200
N
†
† at N = 500, VAR = .030

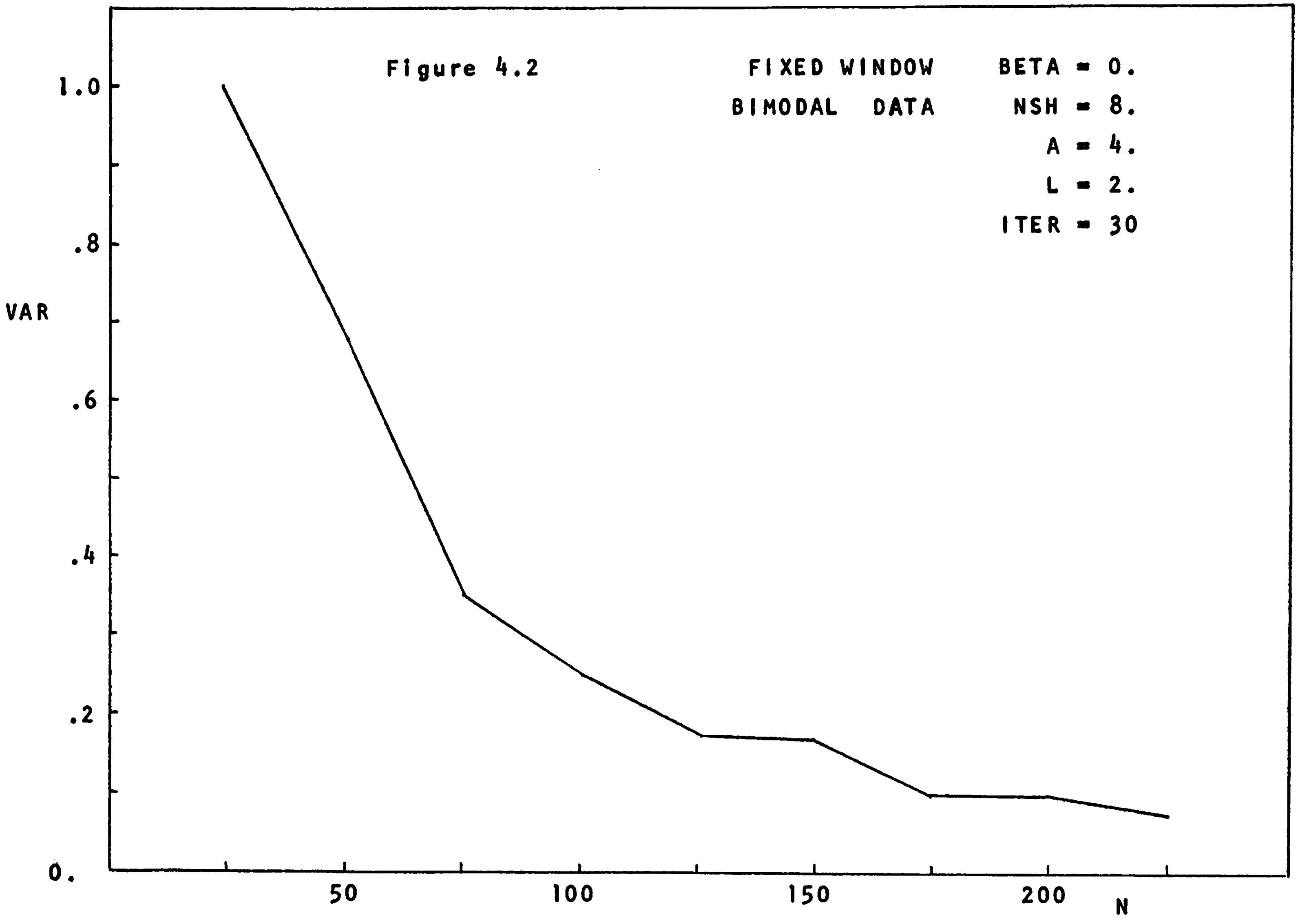


Figure 4.2

FIXED WINDOW
BIMODAL DATA
BETA = 0.
NSH = 8.
A = 4.
L = 2.
ITER = 30
VAR
1.0
.8
.6
.4
.2
0.
50
100
150
200
N

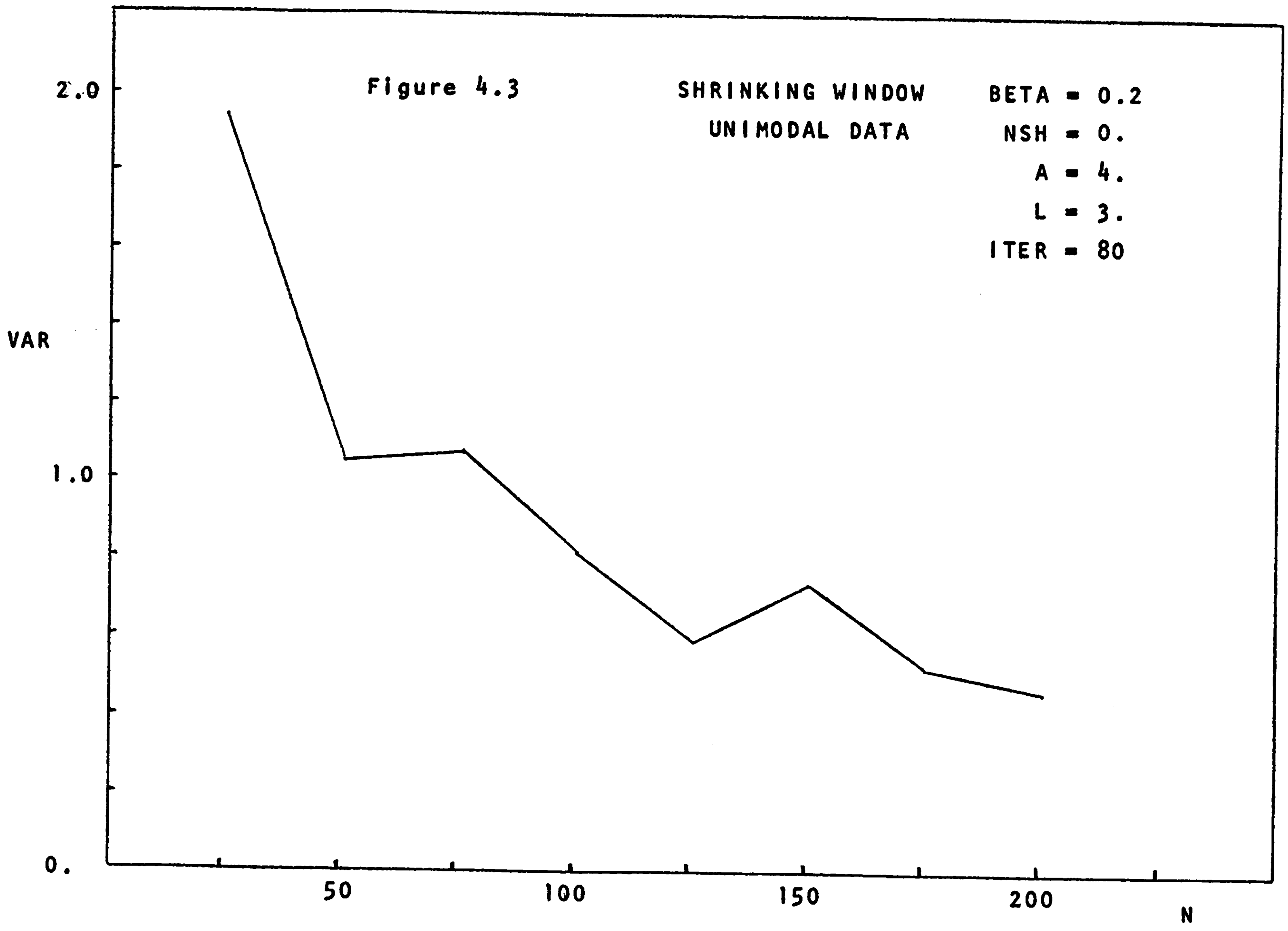


Figure 4.3

SHRINKING WINDOW
UNIMODAL DATA
BETA = 0.2
NSH = 0.
A = 4.
L = 3.
ITER = 80
VAR
2.0
1.0
0.
50
100
150
200
N

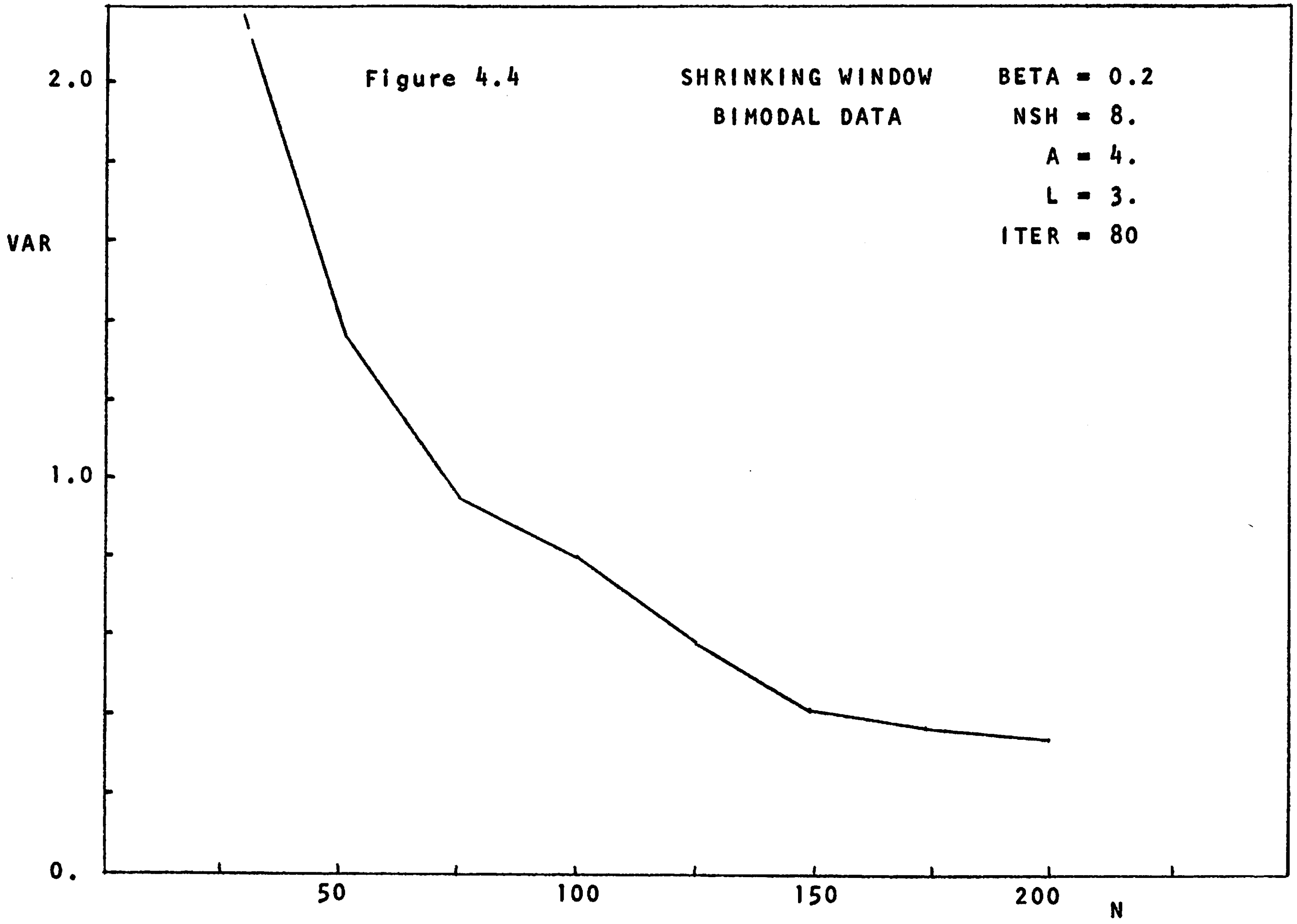


Figure 4.4

SHRINKING WINDOW
BIMODAL DATA
BETA = 0.2
NSH = 8.
A = 4.
L = 3.
ITER = 80
VAR
2.0
1.0
0.
N
50
100
150
200

Section 2. Simultaneous estimation of modes.

The algorithms as implemented in Section 1 of this chapter are rather wasteful of the data generated. Typically many more samples must be generated than are actually used in estimating the mode. Thus if a distribution is made up of two widely separated modes, it is unlikely that any samples from the region about one of the modes will ever be used. To make the program more "efficient" it seems that several modes should be estimated at once. Such a procedure has been implemented.

The procedure considered here uses a sample which fails to "hit" the intervals of interest to center a new interval as follows: Start the fixed width window algorithm as usual. When a sample does not fall in the window, use it to center a second window of the same size. Continue the procedure with two windows until another sample falls outside, and in that case center a third interval about this point. This process can be continued. Whenever one interval overlaps another, simple shorten the interval which has had fewer points averaged into it, doing the shortening in a symmetric fashion, so that the interval keeps the same center, and is still symmetric, although shorter by an amount sufficient to insure disjointness.

Experience has show that with a well chosen value for L, the (initial) window width, the results of the above technique represent quite well the various modes in the data. Figures 4.5 and 4.6 indicate the results of such a procedure. The data (500 and 1500 samples, respectively) is two dimensional, and is drawn from a superposition of two spherically symmetric triangular distributions (which perhaps should be called *conical*) with support set of radius five indicated by dotted lines. The procedure is implemented in its most obvious two dimensional form, with the set $S = \{ x | \quad |x| < L \}$. The number of samples averaged into each ball (nominally of radius $L = 3$, but in many cases reduced by the disjointness requirement) is indicated in each ball. The results are generally satisfactory, although several extraneous "modes" are estimated, as would be expected.

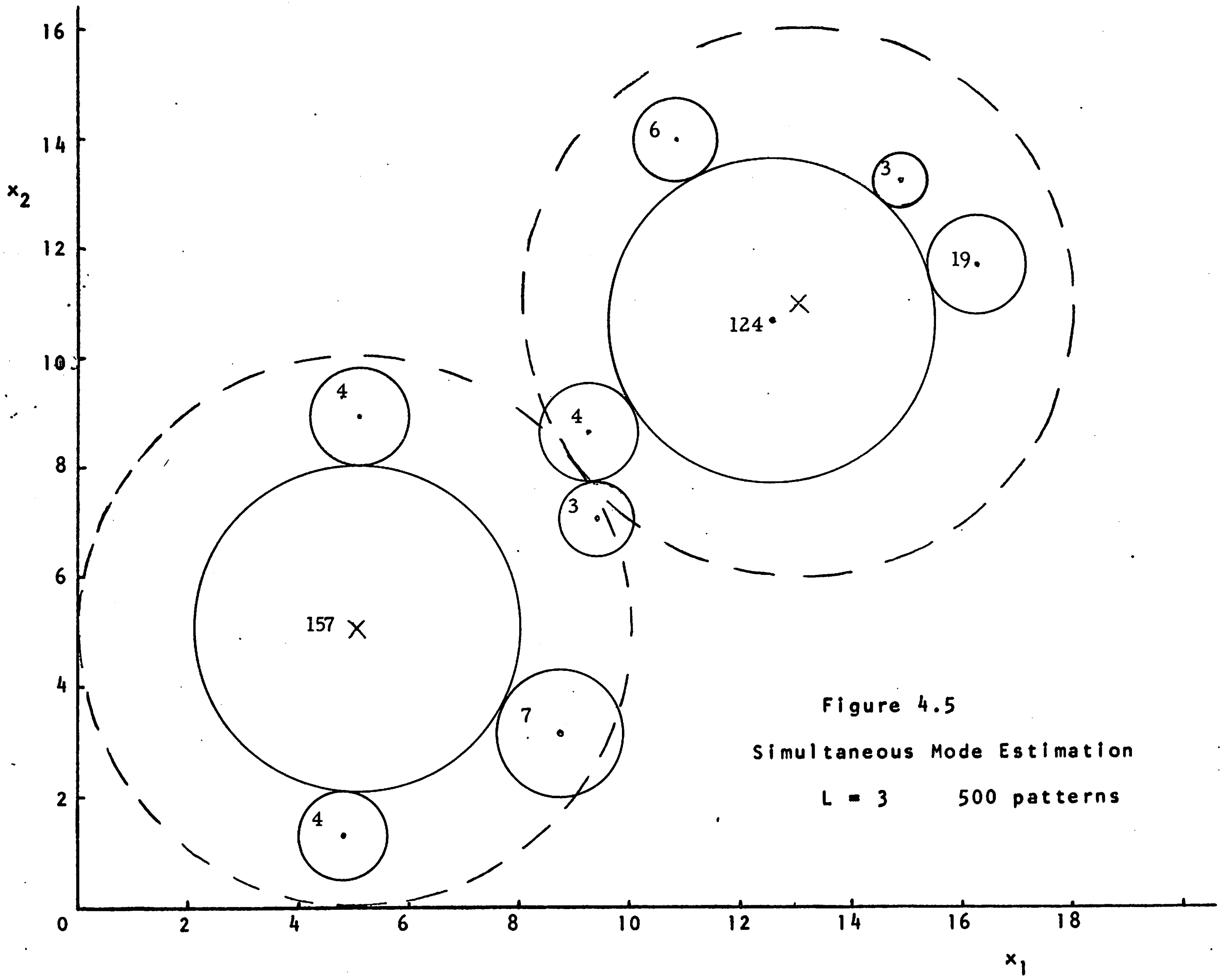


Figure 4.5

Simultaneous Mode Estimation

L = 3 500 patterns

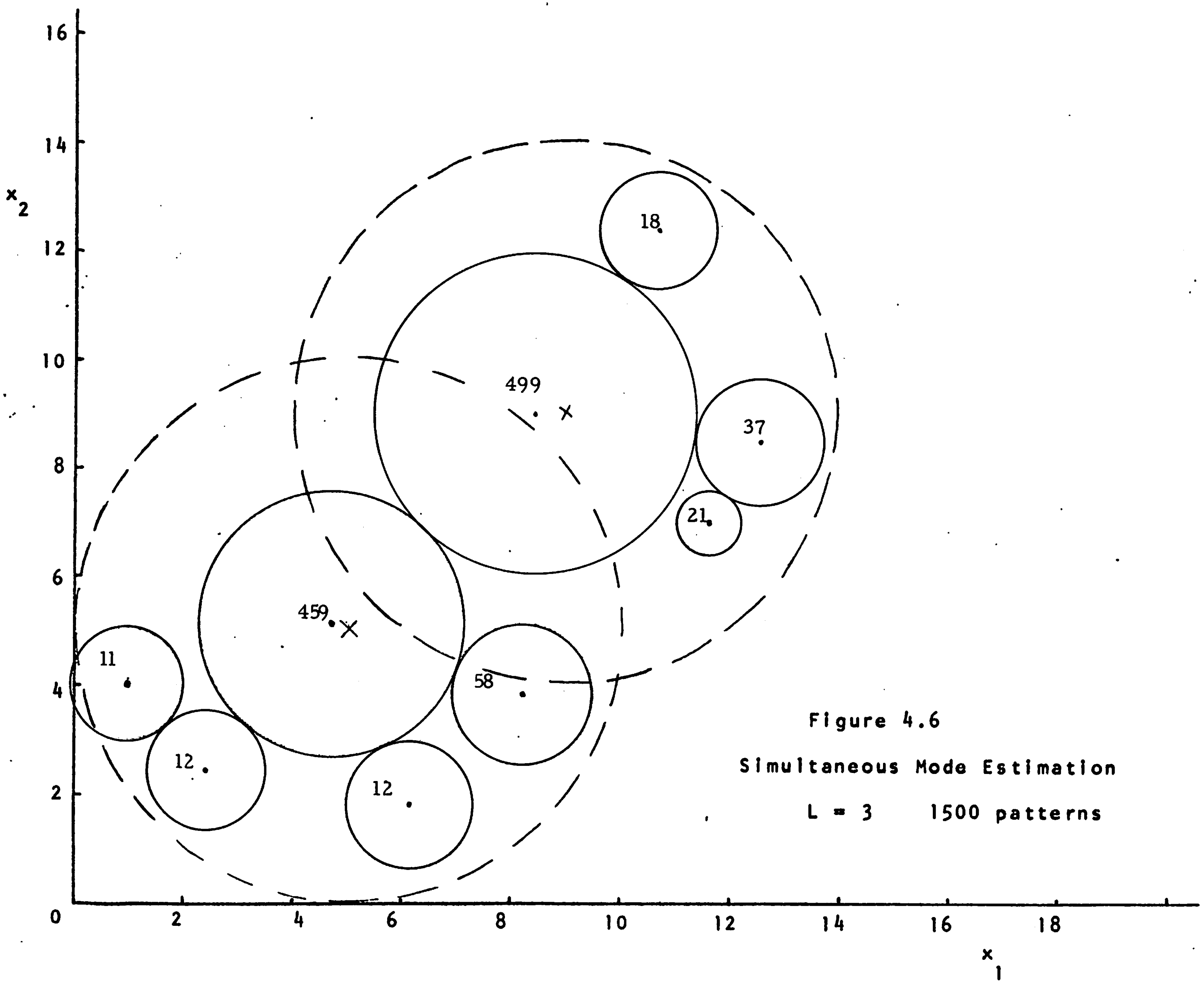


Figure 4.6

Simultaneous Mode Estimation

L = 3   1500 patterns

Section 3. Multimodal convergence.

The results of the computer simulation with bimodal data seem to indicate that the methods perform satisfactorily in multimodal situations. This agrees with the intuitive argument which points out that as the estimate μ_n moves toward a mode, only samples near that mode are averaged into the estimate, and if this is true, then the unimodal convergence result justifies convergence. Just such an argument will prove convergence for certain special distributions. We need an assumption that each mode is "isolated" so that μ_n cannot move from one mode to another.

<u>THEOREM 21</u> Let $L>0$ be fixed and let the density $p(\cdot)$ on the real line be of the form

$$p(x) = \sum_{i=1}^{n} \alpha_i p_i(x)$$

where $\alpha_i > 0$, $\sum\alpha_i = 1$, and $p_i(\cdot)$ unimodal on the average about Θ_i, $i = 1, \ldots n$. Define $L_i = [a_i, b_i]$ as the smallest interval containing the support of $p_i(\cdot)$. Assume, without loss of generality, that $a_1 \leq a_2 \leq \ldots \leq a_n$. Assume

$$a_{i+1} - b_i \geq L \qquad i = 1, \ldots n-1. \tag{4.1}$$

Then the random variable μ_n generated by the fixed window algorithm (THEOREM 1) converges (in quadratic mean and with probability one) to the simple random variable θ which takes on the value θ_i with probability α_i.

Proof: The proof is obvious. If the first sample, x, is from $p_k(\cdot)$ (this happens with probability α_k) then all succeeding samples which are averaged into μ_n also come from $p_k(\cdot)$, because of (4.1). Since $p_k(\cdot)$ is unimodal on the average, μ_n given that x is from $p_k(\cdot)$ converges to θ_k by applying THEOREM 1.

A similar result is true for the shrinking window algorithm:

<u>THEOREM 22</u> Again assume the density $p(\cdot)$ is of the form $p(x) = \sum^{n} \alpha_i p(x)$, $p_i(\cdot)$ satisfying (2.9) and (2.10), i.e. differentiable and strictly unimodal about θ_i. Define a_i and b_i as before, and assume that for some $\varepsilon>0$,

$$a_{i+1} - b_i > \varepsilon \qquad i=1, \ . \ . \ ,n-1 \ .$$

Then the random variable μ_n generated by the shrinking window algorithm (THEOREM 8) will converge to a simple random variable taking on the values θ_i $i=1, . . ,n$, each with positive probability.

Proof: We use the fact that $L_n \to 0$. Assume that for $n > n_\varepsilon$, $L_n < \varepsilon$. Then if μ_{n_ε} is in the support of $p_k(\cdot)$

(this occurs with positive probability), then as in THEOREM 21 only samples from $p_k(\cdot)$ are averaged into μ_n for $n > n_\varepsilon$. THEOREM 8 yields the convergence result.

CHAPTER V

COMPARISON AND CONCLUSION

Section 1. Comparison.

This thesis has discussed two mode seeking methods. There are, of course, other mode estimation procedures. Most of them are of a very different nature from the methods suggested in this work, in that almost no assumptions are made about the density function whose mode is to be estimated. One class of procedures uses the samples to construct an estimate of the underlying distribution function (d.f.), and estimates the true mode (a unique maximum of the density is assumed to exist) by the mode of the sample d.f.

For example, Parzen [13] uses the sample d.f., defined for a sample $x_1, \ldots, x_n$ as

$$F_n(x) = n^{-1}\{\text{number of } x_i\text{'s which are less than } x\}$$

to construct estimates $\Theta_1, \Theta_2, \ldots$ by

$$\Theta_n = \max_{-\infty<x<\infty} \int_{-\infty}^{\infty} h^{-1} K\left(\frac{x-y}{h}\right)\, dF_n(y)$$

where K(·) is a "weighting"function. Parzen is able to prove under smoothness conditions on K(•) that provided $h(n) \to 0$ as $n \to \infty$, Θ_n is asymptotically normal and converges to Θ, the true mode. He shows $E(\Theta_n - \Theta)^2 < Cn^{-\beta}$ where $\beta < 1/2$.

Similarly, in a recent paper Venter [18] has suggested a slightly different technique, using ordered samples $Y_1, \ldots, Y_n$. A sequence of estimates Θ_n is defined as the midpoint of the interval formed by the first and last m (m fixed) consecutive Y_i's which are closest together. He shows $\Theta_n \to \Theta$, the true mode (assuming it exists), with convergence rate no better than $n^{-1/3}$.

The deficiencies of these techniques are numerous. Neither generalizes in an obvious way to R^k, and this is a very serious limitation. In addition, the computations required to implement either scheme are almost overwhelming. A sample d.f. is hard to compute, and obtaining ordered samples from samples is non-trivial, requiring the storage of <u>all</u> the samples. Similarly, computing the integral and finding the maximum in Parzen's scheme are quite difficult, and Venter's estimates are no easier to find.

Thus if one is willing to assume that the underlying distribution is unimodal (or, using the computational or theoretical results of Chapter IV, several distinct modes), then the mode seeking methods presented here are preferable to these "estimate the d.f." procedures.

A very different approach, which has not been applied to the problem of estimating modes but which is obviously related, is the approximation of the density p(•) by a function q(•), where

$$q(x) = \sum_{i=1}^{M} \alpha_i \phi_i(x)$$

where the $\phi_i(\cdot)$ are known functions. The "best" choice of the α_i's can be found using stochastic approximation, involving samples from the density p(•). See Blaydon [3]. The mode can then be estimated by finding the maximum value of q(x). Of course, unless p(•) is of a special form, no finite value of M and no choice of the ϕ_i will give an accurate estimate of the true mode, but in many cases the estimate based on q(•) will be sufficiently close. This may be a useful procedure. The method is based on the "potential function" approach to pattern recognition developed by Braverman [5].

Another approach to mode estimation which also uses stochastic approximation has been suggested by Burkholder [6]. He extends stochastic approximation to the problem of the estimation of the point of inflection of a regression function. (The methods suggested in this thesis simply require estimating the <u>root</u> of a regression function.) Since a mode is by definition a point of inflection of the distribution function, this procedure seems ideally suited to the estimation of modes. An apparent difficulty with Burkholder's presentation is that he requires samples from a family of random variables Y_γ such that $EY_\gamma = F(\gamma)$ where $F(\cdot)$ is the d.f. whose mode is to be estimated. However, we are given only the random variable x with d.f. $F(\cdot)$. In this case Y_γ can be derived from x as follows:

$$\begin{aligned} Y_\gamma(\omega) &= 1 \quad \text{if } x(\omega) \leq \gamma \\ &= 0 \quad \text{if } x(\omega) > \gamma . \end{aligned}$$

In this way the difficulty is avoided. With the family Y_γ is can be shown that the sequence of random variables x_n generated by

$$x_{n+1} = x_n - \frac{a_n}{c_n^2} [y_{3n-1} - (y_{3n-2}+y_{3n})/2]$$

converges with probability one to the mode Θ. y_{3n-2}, y_{3n-1}, and y_{3n} are observations on the random variables $Y_{x_n-c_n}$, Y_{x_n} , and $Y_{x_n+c_n}$, respectively, and the sequences a_n and c_n satisfy

$$c_n \to 0 \qquad \sum a_n = \infty \quad \sum a_n^2/c_n^4 < \infty.$$

$F(\cdot)$ must be strictly unimodal, with requirements on the density quite comparable to those placed on the density in the shrinking window algorithm discussed in this work. Burkholder is also able to prove, under even more restrictive conditions on the density, that x_n is asymptotically normal, with a variance which decreases as $n^{-\xi}$, with $\xi<1/2$.

This procedure possesses many of the properties of the methods described in this work, and thus may be quite useful in pattern classification and cluster analysis.

Section 2. Conclusion.

This work has established two basic facts: first, that mode estimation prodecures are quite useful for pattern recognition and cluster analysis, and second, that two particularly convenient mode seeking algorithms have several desirable properties.

This was accomplished by first introducting the concept of cluster analysis and discussing its relationship to pattern recognition. It was argued that any cluster analysis procedure would be more effective if the mode locations were known. For this reason, a mode estimation technique was described which could estimate modes in multimodal distributions and a proof of convergence was given in the unimodal case. A similar method, the shrinking window algorithm, was then introduced, and its convergence properties were found. Since it is of considerable interest to know about the rates of convergence and the extensions to higher dimensional spaces of both the algorithms, this was discussed. Finally, computational results were presented which confimed convergence in both unimodal and bimodal situations, and in the previous section, the mode estimation procedures defined in this work were favorably compared with other methods.

Several questions on this topic have been left unanswered by this work. The most significant result which was not obtained was the proof of convergence in multimodal populations. The procedures do converge under certain assumptions about disjoint support sets, but the computational results seem to indicate that a more general result may be true. This problem of multimodal convergence may be framed more generally by asking under what conditions does a stochastic approximation process converge to one of <u>several</u> roots. All present theorems consider only one root. A related problem is the development of a good technique for distinguishing the various modes to which the algorithms are converging.

Various outher topics may be worthy of further study. One important question is how to choose the value A in $a_n = A/n$ or $a_n L_n^2 = A/n$ in the mode seeking algorithms. Based on the rate of convergence results, there are optimum values for this parameter, yet without knowledge of the density $p(\cdot)$, the best choice of A cannot be made.

Several generalizations are apparent, though their usefulness is at times not clear. For example, the results of Chapter III in extending the procedures to R^k could probably be generalized to much more general inner product spaces. Or, different sets over which unimodality is judged might be considered. In the one dimensional versions we consider the set $[-L,+L]$, or in the shrinking window case, the set $[+L_n,+L_n]$. A most interesting generalization would be to allow L or L_n to be random variables. Similarly, another extension would be to consider the properties of the algoritms when the sequence μ_n is generated by, for some $K(\cdot)$,

$$\mu_{n+1} = \mu_n - a_n K(\mu_n - y_n).$$

This work deals only with the special case

$$K(x) = x \qquad \text{if } |x|<L$$
$$= 0 \qquad \text{elsewhere,}$$

or, in the case of higher dimensions

$$K(x) = x \qquad \text{if } x \varepsilon S$$
$$= 0 \qquad \text{elsewhere.}$$

Finally, nothing in this work is concerned with mode estimation with a finite number of samples. The two procedures described in this thesis are consistent and therefore asymptotically unbiased. However, they are inefficient in the use of samples, since only a fraction of any data set is actually used to calculate the estimate, and they do not have well understood properties for the finite case.

APPENDIX

Section 1. Stochastic Approximation.

<u>THEOREM A</u> (Dvoretsky [3]) Let α_n, β_n, and γ_n be positive sequences such that $a_n \to 0$, $\sum \beta_n < \infty$, and $\sum \gamma_n = \infty$. Let $T_n(\cdot)$ be measureable transformations satisfying

A. $|T_n(r_1, . . . r_n) - \theta| \leq \max\{\alpha_n, (1+\beta_n)|r_n - \theta| - \gamma_n\}$

Let x_1 and Y_n for $n=1,2,. . .$ be random variables and define

B. $x_{n+1} = T_n(x_1, . . . , x_n) + Y_n$

Assume

C. $Ex_1^2 < \infty$ and $\sum EY_n^2 < \infty$

D. $E[Y_n \mid x_1, . . . x_n] = 0$ w.p.1 .

Then

$$\lim_{n\to\infty} E(x_n - \theta) = 0 \qquad \text{and}$$

$$\Pr\{\lim_{n\to\infty} x_n = \theta\} = 1 \quad .$$

<u>EXTENSION</u>: γ_n may be replaced by a non-negative function $\gamma_n(r_1, . . . r_n)$ and the result holds provided

$\sum \gamma_n = \infty$ uniformly for all sequences $r_1, r_2, . . .$ for which $\sup_n |r_n| < M$ for some M.

The following extension to R^k is due to Sacks and Derman [7].

<u>THEOREM B</u> Let the same conditions as in THEOREM A be satisfied with these modifications: $\{x_n\}$, $\{Y_n\}$, and $\{T_n\}$ are k dimensional random vectors, C. should be interpreted as $E|Y_n|^2$, and the absolute value in A. should be read as norm. The conclusion is that $|x_n-\theta| \to 0$ w.p.1 .

<u>EXTENSION</u> The extension of the one dimensional result permitting random γ_n remains valid.

Section 2. Sacks' Results on Asymptotic Distribution.

These results are due to Sacks [14].

<u>THEOREM</u> Let $Y(x)$ be a real random variable for each x and define $M(x) = EY(x)$. Assume $M(x) = 0$ has a unique solution $x = \Theta$. Define

$$x_{n+1} = x_n - a_n Y(x_n)$$

where a_n satisfies $\sum a_n = \infty$, and $\sum a_n^2 < \infty$, and x_1 arbitrary. Assume

A1 $(x-\Theta)M(x) > 0$ for $x \neq \Theta$

A2 $|M(x)| < K_1|x-\Theta|$ and $\inf\limits_{t_1<|x-\Theta|<t_2} |M(x)| > 0$

$0<t_1<t_2<\infty$.

A3 $M(x) = K(x-\Theta) + \delta(x,\Theta)$

where $\delta(x,\Theta)/(x-\Theta) \to 0$ as $x \to \Theta$.

A4 $EY^2(x) < K_2$ $EY^2(\Theta) = \sigma^2$.

A5 $E|Y(x)|^{2+\nu}$ for some $\nu > 0$.

Let $a_n = A/n$ with $2AK > 1$. Then $n^{1/2}(x_n - \Theta)$ is asymptotically normal with zero mean and variance $A^2\sigma^2(2AK - 1)^{-1}$.

<u>THEOREM</u> (R^k) Let $Y(x)$ be a set of random variables in R^k, and define $M(\cdot)$ and x_n as before. Assume

A1 $\langle x-\Theta, M(x)\rangle > 0$ $x \neq \Theta$

A2' $|M(x)| < K_1|x-\Theta|$ and $\inf\limits_{t_1<|x-\Theta|<t_2} |M(x)| > 0$

for $0<t_1<t_2< \infty$.

A3' $M(x) = B(x-\Theta) + \delta(|x-\Theta|)$ where B is positive definite

A4' $E|Y(x)|^2 < K_3$ $E|Y(\Theta)|^2 = \pi$ for π non negative definite

A5' $E|Y(x)|^{2+\nu}$ some $\nu>0$.

Assume $a_n = A/n$ with $Ab_k > 1/2$. Then $n^{1/2}(x_n-\Theta)$ is asymptotically normal with zero mean and covariance Σ defined (along with b_k) by:

Let $b_1, \ldots b_k$ be the eigenvalues of B in decreasing order. Define P as $B = PDP^{-1}$ where D is diagonal with entries $b_1, \ldots b_k$. Define π°_{ij} as the i,j th element of $P^{-1}\pi P$. Then $\Sigma = PQP^{-1}$ where $Q = \{q_{ij}\}$ and

$$q_{ij} = A\,(Ab_i + Ab_j + 1)^{-1}\pi^\circ_{ij} \quad .$$

REFERENCES

1. G. Ball, "Data Analysis in the Social Sciences," Proc. Fall Joint Comp. Conf. Vol 27 Part 1 p 533 (1965)

2. G. Ball and D. Hall "Isodata, A Novel Method of Data Analysis and Pattern Classification," Stanford Research Institute, Menlo Park (1965).

3. C. Blaydon, "Recursive Algörithms for Pattern Classification," Technical Report #520 Harvard Univ. (1967).

4. R.E.Bonner "On Some Clustering Techniques," IBM J. Res. and Dev. Jan 1964.

5. E.M. Braverman "The Application of the Potential Function Method to Pattern Classification Without Supervision," Automation and Remote Control Vol 10 (Oct. 1966).

6. D.L. Burkholder, "On a Class of Stochastic Approximation Processes," Ann. Math. Stat. 27, 1044 (1956).

7. T. Cover and P. Hart, "Nearest Neighbor Classification," IEEE Trans. on Info. Theory (1966).

8. D.R. Cox, "Note on Grouping," J. of Amer. Stat. Assn. Vol 52 No. 280 (1957).

9. A. Dvoretsky, "On Stochastic Approximation," Proc. 3rd Berk. Symp. on Math., Stat., and Prob. p.39 (1956).

10. S. Fralick, "The Synthesis of Machines which Learn without a Teacher," Tech. Report 6103-8 Stanford (1964).

11. J. MacQueen, "On Convergence of k-means and Partitions with Minimum Average Variance," UCLA Report (Abstract Only in Ann. Math. Stat 36(3), 1084) (1965).

12. N. Nilsson, "Adaptive Pattern Reocoginition: A Survey," 1966 Bionics Symposium.

13. E. Parzen, "On Estimation of the Probability Density Function and Mode, " Ann. Math. Statis. 33,1065 (1962).

14. J. Sacks, "Asymptotic Distribution of Stochastic Approximation Procedures," Ann. Math. Statis. 29,373 (1958).

15. J. Sacks and C. Derman, "On Dvoretsky's Approximation Theorem," Ann. Math. Statis. 32,601 (1958).

16. D. Sakrison, "Stochastic Approximation," in Advances in Communication Systems, Vol 2, Academic Press, N.Y. (1966).

17. G. S. Sebestyen, "Pattern Recognition by an Adaptive Process of Sample Set Construction," IRE Trans. on Info. Theory, Vol IT-8 (1962).

18. J. Venter, "On Estimation of the Mode," Ann. Math. Statis. 38,1446 (1967).