

INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.
2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.
3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.
4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.
5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

**University
Microfilms
International**
300 N. Zeeb Road
Ann Arbor, MI 48106

8416502

Boswell, Steven Blake

NONPARAMETRIC MODE ESTIMATION FOR HIGHER DIMENSIONAL
DENSITIES

Rice University

PH.D. 1984

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy.
Problems encountered with this document have been identified here with a check mark ✓.

1. Glossy photographs or pages _____
2. Colored illustrations, paper or print _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Pages with black marks, not original copy _____
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages ✓
8. Print exceeds margin requirements _____
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) _____ lacking when material received, and not available from school or author.
12. Page(s) _____ seem to be missing in numbering only as text follows.
13. Two pages numbered _____. Text follows.
14. Curling and wrinkled pages _____
15. Other _____

University
Microfilms
International

RICE UNIVERSITY

NONPARAMETRIC MODE ESTIMATION FOR HIGHER
DIMENSIONAL DENSITIES

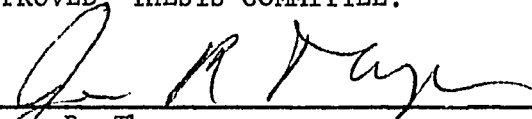
by

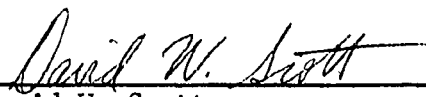
STEVEN BLAKE BOSWELL

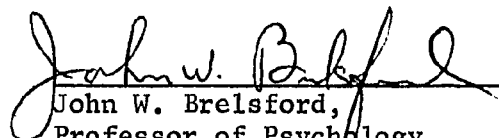
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

DOCTOR OF PHILOSOPHY

APPROVED, THESIS COMMITTEE:


James R. Thompson,
Professor of Mathematical Sciences
Chairman


David W. Scott,
Associate Professor of Mathematical Sciences


John W. Brelsford,
Professor of Psychology

HOUSTON, TEXAS

NOVEMBER, 1983

ABSTRACT

NONPARAMETRIC MODE ESTIMATION FOR HIGHER DIMENSIONAL DENSITIES

by

Steven B. Boswell

In this study a family of estimators is developed for local maxima, or modes, of a multivariate probability density function. The mode estimators are computationally feasible iterative optimization procedures utilizing nonparametric techniques of probability density estimation which generalize easily to sample spaces of arbitrary dimension. The estimators are proven to be strongly consistent for any distribution possessing mild continuity properties.

Three specific mode estimators are evaluated by extensive Monte Carlo testing upon samples from both classical unimodal and nonstandard unimodal and bimodal distributions. Detection of the presence of multiple modes is a matter of special concern in many investigations. Thus a global strategy is developed and tested to demonstrate the potential of the estimators for complete characterization of sample modality.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Professor James R. Thompson, for his guidance throughout my graduate career, and especially during the completion of this dissertation. In general I am indebted to the faculty of the Department of Mathematical Sciences for its intellectual and financial support. I relied heavily on earlier course work and on the excellent computing facilities assembled in recent years. In particular I would like to recognize the influence of my committee member, Dr. David W. Scott.

Finally, I want to thank my wife, Dr. Anthea J. Coster, for her love and her forbearance of my unique schedule, and my parents, Mr. and Mrs. George A. Boswell, for their unvarying and unselfish support in all circumstances.

This research was partially supported by the U.S. Army Research Office Grant DAAG-29-82-K-0014. Thanks are due also to the School of Mathematics of the Georgia Institute of Technology for arranging computer support for much of the final algorithmic development.

TABLE OF CONTENTS

	Page
I. INTRODUCTION AND SUMMARY	1
1.1. Introduction	1
1.2. Summary of Later Chapters.	4
II. KERNEL DENSITY ESTIMATION AND DERIVED ESTIMATORS OF THE MODE: STATISTICAL ANALYSIS	9
2.1. Review of Mode Estimators (One- Dimensional)	9
2.2. Kernel Estimators.	16
2.2.1. Introduction	16
2.2.2. Consistency in Mean Squared Error for Kernels with Infinite Support.	20
2.2.3. Mean Squared Error; Kernel Design	28
2.2.4. Variable Kernel Estimators	31
2.3. Consistency of Mode Estimators	35
III. THE MEAN UPDATE ALGORITHMS	49
3.1. Description of the Mean Update Algorithm and Discussion	49
3.2. Structure and Implementation of the Monte Carlo Trials	64
3.3. Presentation of Results with the Mean Update Algorithm	78
IV. MINIMIZATION OF THE ESTIMATED DENSITY BY NEWTON'S METHOD.	105
4.1. Introduction	105
4.2. Description of the Algorithm and Consis- tency.	108
4.3. Implementation Notes	111
4.4. Empirical Results.	120
V. WEIGHTED MEAN UPDATE	144
5.1. Derivation of the Procedure and Consis- tency Results.	144
5.2. Empirical Results.	154

	Page
VI. MULTISTART ALGORITHM AND SUGGESTIONS FOR FURTHER WORK	173
6.1. Consideration of Previous Chapters and Introduction	173
6.2. Multistart Algorithm	175
6.3. Further Work	193
APPENDIX: Generation of Equivalent Two-Component Gaussian Mixtures.	195
REFERENCES	205

I. INTRODUCTION AND SUMMARY

1.1. Introduction

Many techniques of statistical analysis rely ultimately upon the estimation of a probability distribution from available data. When it is reasonable to assume that the unknown distribution belongs to an easily summarized functional family, the estimation problem reduces to the estimation of a small number of parameters which identify individual members of the family. A parametric model grants the analyst more efficient estimators and more powerful inferential techniques, but only if the model that is used is appropriate. If the model fails to account for important features of the stochastic behavior being studied, the results of the parametric estimation may be seriously misleading, the more so as the distortion escapes the practitioner and propagates through subsequent analyses. In a less negative vein, an analyst may consciously seek to detect nonstandard distributional characteristics because of the information they carry about the phenomenon he is investigating. An example of such endeavor is the investigation of bumps and dips in mass spectra in scattering experiments [Good and Gaskins, 1980].

Among the most damaging distributional features, when undetected, and the most informative, when detected, is the presence of multimodality in the population density. This

is particularly true as dimension increases, since experience of workers in the field indicates that multimodal data is encountered very frequently in high dimensions, while at the same time recognition of data structure becomes a difficult problem.

For the reasons outlined above, the past two decades have witnessed the development of procedures for estimating probability density functions which employ assumptions only on the regularity of the function. Kernel type estimators, employed throughout the study, require for consistency only that the density be continuous, though finite sample applications depend upon an additional supposition that the estimated density be "not too rough." As a practical matter, an estimate which is too rough is identified by the presence of high frequency oscillation, and thus of many local minima and maxima. Good density estimates avoid such rapid oscillation but are usually just on the verge of acquiring it. Thus nonparametric density estimation and the investigation of multimodality exist in a special symbiotic relationship with one another. A principal motivation for density estimation is to evaluate the presence or absence of multiple modes; yet the density estimate is evaluated to a large extent by the modality pattern it yields.

The object of this research has been to devise and implement a nonparametric procedure for the detection of

modes in high dimensional data. Our effort has focused on the accurate identification of a local maximum (any local maximum). The task of locating the global mode, or of cataloging all local maxima, has been considered more briefly. By high dimensional we mean dimension greater than two, though we should point out that our estimators are valid and tested in low dimension as well. Nevertheless, the main interest has been in dimensions large enough that visual display of data structure is intractable.

Motivation for this study derived from difficulties encountered in utilizing nonparametric density estimation techniques in such dimensions. The difficulties stem from the loss of representational economy which necessarily accompanies nonparametric estimates. A part of the problem is computational. To characterize the density over a region, point estimates must be made on a lattice which covers the region with adequate precision. With the spacing held constant, the number of mesh points grows exponentially. A mesh with ten points along each coordinate is likely too sparse to detect pattern over the full support of the density; yet even such a rough grid will require 100,000 point evaluations in dimension five. The computational burden may be lessened by first locating regions of high concentration of the density, so as to focus computation where it is needed.

Perhaps more crucial is the difficulty of interpreting

density estimates in a high dimensional space. Constructing and displaying level sets, for example, is probably infeasible beyond dimension three or four. The presence and location of modes yield one of the few characteristics of a multivariate density, aside from low-order moments and principal component vectors, which can be easily visualized. Density estimation and mode estimation remain closely related activities, but as dimension increases, the detection of modes becomes an increasingly independent problem, more practical than full-scale density estimation, and more important in its own right.

1.2. Summary of Later Chapters

In this dissertation three mode estimators are studied. The first, known as the mean update estimator, is given by the following algorithm:

```

Let  $m_1$  be an initial iterate,
     $k$  be a fixed parameter;
 $i = 1$ ;
Repeat until  $m_{i+1} = m_i$ ;
    begin
        Find the sample points  $\{x_1, \dots, x_k\}$ 
            which are nearest to  $m_i$ ;
        These are called the neighbor set.
         $m_{i+1} = (1/k) \sum_{j=1}^k x_j$ ;
         $i = i+1$ ;
    end.
```

The mean update was studied by Fwu, Tapia, and Thompson [1980], who reported that the procedure occasionally appeared to cycle (failed to converge), and otherwise had a tendency to stop before a bona fide local maximum had been well approximated. We show that the mean update must converge in a finite number of iterations, and thus that the cycling problem does not arise. Extensive tests were conducted to evaluate the severity of the second fault, using different values of k , different sample sizes, and data of dimension as high as 100. In univariate and bivariate data the mean update is untrustworthy on small and moderate samples, but its ability to identify modes generally increases as dimension increases, especially up through dimension 10, and remains stable over the complete range tested. We were especially interested in the estimator's performance with k giving small percentages of the sample size, since limits on k which could be used would imply limitations on use of the algorithm for exploring multimodality. With univariate or bivariate data and moderate sample sizes, say 100 to 500, the rapidly accelerated variability of the estimators that occurred with truncation below 20% of the sample size made use of such small neighbor sets untenable. In high dimension, however, the use of neighbor sets even half this size gave reasonable results. The mean update procedure is discussed in Chapter III.

Chapter III also describes the design and implementation of the Monte Carlo testing used throughout the study, including the methods used for generating random variates and the means by which performance was measured and reported.

The main failing of the mean update is its tendency to fall short of the mode. To improve its "hill-climbing" ability we examined a two-stage procedure which followed the mean update with a Newton's method algorithm. This algorithm is discussed in Chapter IV. Asymptotically unbiased kernel estimators for the derivatives of a density function were developed. Modifications to the stopping criteria and line search phase were necessary to adapt Newton procedures to the mode estimation problem. Because estimating derivatives is intrinsically more difficult than estimation of the density itself, expectations for the Newton method second stage were modest. Nevertheless, on low-dimensional data in cases where the mean update stalled well short of a local maximum, and thus where a pronounced gradient should be evident, the second stage usually "unstuck" the mode estimate. As dimensionality increased, however, the effectiveness of the second stage diminished rapidly. With further refinement of the Newton's method implementation, particularly in the manner it searches for better iterates, its performance might be substantially improved. The fundamental limitation of Newton's method for our purposes is its reliance at each iteration upon a

one-dimensional line search. When gradient information is erratic, the direction chosen for the line search is unreliable, and there is inadequate provision in the procedure for recovering from a poor choice.

By examining expressions for critical points of the kernel estimate of a probability density function, a weighted mean update was devised which addresses the deficiencies of the Newton procedure. The weighted mean algorithm makes more complete use of the local information in the sample, and is less sensitive to the scale of the measurements. It has an iterative formulation similar to the previously described mean update:

Let m_1 be an initial iterate,

k and h be fixed parameters;

$i = 1$;

Repeat until a stopping criterion is met;

begin

Find the sample points $\{x_1, \dots, x_k\}$

which are nearest to m_i ;

$m_{i+1} = \sum_{j=1}^k w_j(m_i; h) x_j$, where

$w_j(x; h) = \pi_j(x; h) / \sum_{j=1}^k \pi_j(x; h)$ and

$\pi_j(x; h) = \exp\{-\frac{1}{2} (\|x^{(j)} - x\|/h)^2\}$;

$i = i+1$;

end.

The smoothing or scale parameter h may remain fixed, as described above, or may be tagged to local sample characteristics, such as $\|m_i - x_k\|$, the distance from the current iterate to its k -th neighbor. We note that the weighted mean update is similar to one of the computational forms of an M-estimate for multivariate location [Huber, 1981]. The weighted mean is the method of choice. Its properties and performance are discussed in Chapter V.

Chapter II contains an historical review of research in univariate mode estimation, and also conducts statistical analyses of kernel density estimation and related optimization procedures. These analyses underlie subsequent demonstrations of the consistency of our mode estimators.

Chapter VI presents a simple global strategy for organizing the iterative mode estimators with the goal of fully cataloging the modes of a probability density function. Performance of the estimators in multimodal environments is discussed, and topics are suggested for further research.

II. KERNEL DENSITY ESTIMATION AND DERIVED ESTIMATORS OF THE MODE: STATISTICAL ANALYSIS

2.1. Review of Mode Estimators (One-Dimensional)

If X is a continuous random variable taking values in p -dimensional Euclidean space, and $f(x)$ is its density with respect to Lebesgue measure, the mode θ of f is typically defined to be the (unique) point for which $f(\theta) \geq f(x)$, $x \in \mathbb{R}^p$. A more general definition will take θ as a mode of f if there exists an open convex set N containing θ , and θ is a (unique) local maximizer of f over N . In most cases we will consider modes locally without explicit comment. When it is necessary to identify the mode of the first definition, it will be called the global mode. A density possessing more than one local maximum is said to be multimodal; otherwise it is unimodal.

The problem of estimating modes received little attention prior to Parzen's article on nonparametric density estimation (1962). Since then a number of estimators designed specifically for the mode have been proposed. The three most heavily studied were in existence as early as 1965, when they were discussed by Dalenius, who conducted a brief Monte Carlo investigation of their performance on small samples from a chi-square distribution. The first of these estimators is the midpoint of the interval of prescribed length containing the maximum number of observations. It was proposed by Chernoff [1964]. For bias to vanish

asymptotically, it is necessary that the width of the interval, $h(n)$, go to zero as $n \rightarrow \infty$. Wegman [1971] established strong consistency for the Chernoff estimator if $h(n)$ converges to zero more slowly than $[\log(\log n)/n]^{1/2}$. The Chernoff estimate is equivalent to the mode of a kernel-type density estimate with uniform kernel.

A second and interesting estimator, due to Grenander [1965], is

$$M_S^* = B/A, \quad (2.1.1)$$

where

$$B = \frac{1}{2} \sum_{v=1}^{n-k} (x_{v+k} + x_v) (x_{v+k} - x_v)^{-s},$$

$$A = \sum_{v=1}^{n-k} (x_{v+k} - x_v)^{-s},$$

and

$$x_1 < x_2 < \dots < x_n$$

are the order statistics of a sample of size n . Grenander's estimate is based upon the observation that, under suitable regularity conditions, the quantity

$$\begin{aligned} M_{S+1} &= \int_{-\infty}^{\infty} x f^{S+1}(x) dx / \int_{-\infty}^{\infty} f^{S+1}(x) dx \\ &= \int_{-\infty}^{\infty} x K_S(x) dx, \quad K_S(x) = \frac{f^{S+1}(x)}{\int_{-\infty}^{\infty} f^{S+1}(x) dx} \quad (2.1.2) \\ &= E\left\{X \frac{f^{(s)}(X)}{\int_{-\infty}^{\infty} f^{(s+1)}(x) dx}\right\} \end{aligned}$$

will closely approximate the mode for s sufficiently large. To establish this point, we may assume a simplified version of conditions utilized by Venter [1967] and Sager [1975], which indicate primarily the sharpness of the curvature about the mode. Letting θ be the mode of f , define

$$\alpha(\delta, R) = \frac{\inf\{f(x) : |x-\theta| < \delta/R\}}{\sup\{f(x) : |x-\theta| \geq \delta\}}, \quad (2.1.3)$$

and suppose there are constants $\rho > 0$, $R > 1$, $c > 0$ such that, for all small δ , $\alpha(\delta, R) \geq 1 + \rho\delta^c$. Then, for $|x-\theta| > \delta$,

$$\begin{aligned} f^{s+1}(x) &\leq [\sup\{f(x) : |x-\theta| \geq \delta\}]^{s+1} \\ &\leq \left[\frac{\inf\{f(x) : |x-\theta| < \delta/R\}}{[1 + \rho(\delta/R)^c]} \right]^{s+1}. \end{aligned}$$

Also,

$$\begin{aligned} \int f^{s+1}(x) dx &\geq \int_{|x-\theta| < \delta/R} f^{s+1}(x) dx \\ &\geq [\inf\{f(x) : |x-\theta| < \delta/R\}]^{s+1} \cdot 2(\delta/R). \end{aligned}$$

Thus, again for $|x-\theta| > \delta$,

$$K_s(x) \leq \frac{2(\delta/R)}{[1 + \rho(\delta/R)^c]^s}.$$

Since the denominator is strictly greater than one, with ρ , R , c constant, then for all δ sufficiently small,

$$\sup_{|x-\theta| > \delta} K_s(x) \rightarrow 0 \text{ as } s \rightarrow \infty.$$

Redefining $\tilde{K}_s(x) = \tilde{K}_s(x+\theta)$, \tilde{K}_s satisfies

$$(i) \quad \tilde{K}_s(x) \geq 0$$

- (ii) $\int K_s = 1$ for all s
- (iii) On any subset of the real line bounded away from zero, $K_s \rightarrow 0$ uniformly as $s \rightarrow \infty$.

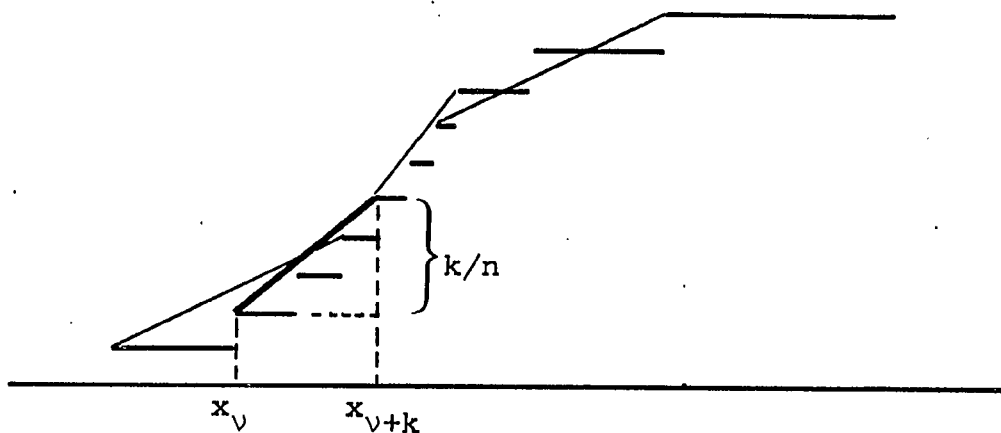
Thus K_s is an approximate identity and

$$M_{s+1} = \int x K_s(x-\theta) dx$$

converges to θ as $s \rightarrow \infty$ [Baggett and Fulks, p. 19].

Grenander estimates M_{s+1} via linear interpolation of the empirical cumulative distribution at jump points spread k order statistics apart (see Figure 2.1.1).

Figure 2.1.1.



This yields

$$\hat{f}_n(x) = \begin{cases} (k/n) [x_{v+k} - x_v]^{-1}, & x_v \leq x \leq x_{v+k} \\ 0 & , \quad x < x_1 \text{ or } x > x_n, \end{cases}$$

and substituting in (2.1.2) gives expression (2.1.1) for M_s^* .

It is worthwhile to note that M_s^* is a quasi-linear,

convex combination of the order statistics, i.e.,

$$M_S^* = \sum_{v=1}^k w_v x_v + \sum_{v=n-k+1}^n w_v x_v + \sum_{v=1+k}^{n-k} (w_{v-k} + w_v) x_v, \quad (2.1.4)$$

where

$$w_v = \frac{[x_{v+k} - x_v]^{-s}}{2 \sum_{j=1}^{n-1} [x_{v+k} - x_j]^{-s}} > 0,$$

and

$$\sum_{v=1}^n w_v < 1.$$

The formulas for w_v are non-linear functions of the order statistics themselves (more specifically, their spacings).

The rationale for the Grenander estimate depends upon the ability to take s large. However, Grenander finds that, for consistency, it is necessary to take $k > s$. For small sample sizes s must be kept small.

The third method, suggested by Dalenius [1965], estimates the mode by a point in the smallest interval (x_J, x_{J+k}) containing k points, where k is a prescribed parameter. We will call this the "nearest neighbor" mode estimator.

Usually the point estimate is chosen as the midpoint of the interval, giving $M_\infty^* = \frac{1}{2} x_J + \frac{1}{2} x_{J+k}$. Following a remark by Ekblom [1972], we note that as $s \rightarrow \infty$ in (2.1.4), the relative contribution of all w_v vanish except for $v = J$; hence M_∞^* is the limiting case of Grenander's M_S^* as $s \rightarrow \infty$. From this observation, we might suspect the consistency of the

nearest neighbor mode estimator. However, Venter [1967] proves strong consistency if the parameter k is chosen as a function of sample size satisfying

$$\begin{aligned} & \text{(i) } k(n)/n \rightarrow 0 \text{ as } n \rightarrow \infty, \\ & \text{and (ii) for all } \lambda \in (0,1), \sum_{n=1}^{\infty} n \lambda^{k(n)} < \infty. \end{aligned} \quad (2.1.5)$$

Sager [1975] extends Venter's analysis, relaxing conditions on the distribution of the random variable, and improving Venter's assessment of optimal rate of convergence. If $\alpha(\delta, R) \geq 1 + \rho \delta^c$ (see (2.1.3)), Sager suggests choosing

$$k(n) = A n^{2c/(1+2c)} \quad \text{for some } A > 0,$$

producing a mode estimate which converges as

$$|\hat{\theta}(n) - \theta| = O(n^{-(1/(1+2c))} (\log n)^{1/c}). \quad (2.1.6)$$

As both he and Venter remark, estimating the mode is a task of greatly varying difficulty. If c is large in (2.1.6), the convergence can be very fast indeed. For c small, the opposite is true.

Monte Carlo studies of the performance of the estimators have been reported by Dalenius [1965], Ekblom [1972], Robertson and Cryer [1974], and Andriano, Gentle, and Sposito [1978]. In all cases, the sample sizes available to the estimators were moderate or more typically, very small. No large sample testing has been reported. For the sample sizes considered, however, several conclusions are available.

First, all of the estimators are strongly affected by the shape of the density, in particular the local definition of the mode. For example, Andriano, et al. reported mean squared error on samples generated according to three different distributions -- a beta distribution with (3,50) degrees of freedom, chi-square with 8 degrees of freedom, and F-distribution with (25,25) degrees of freedom. The errors for the F-distribution with a sharp peak at .852, were on the average 100 times smaller than for the chi-square, which has a rather flat mode at 6. The errors observed with the beta distribution were further reduced two orders of magnitude.

Secondly, Grenander's estimate is also influenced by the shape of the density away from the mode, and this makes it more biased on skew data than its competitors. On the other hand, the Grenander estimate, since it averages contributions from every observation in the sample, is less variable than either the nearest neighbor or the Chernoff-type mode estimate. This increase in stability compensates for the greater bias, and the Grenander estimate overall appears the most successful of the three estimators, particularly as larger sample sizes will allow the parameter s to increase, with a corresponding mitigation of the bias problem.

A final remark to make about the common mode estimators is that they abstract specific features of well-known non-parametric density estimators, either kernel or nearest neighbor type. The Chernoff and nearest neighbor mode

estimates simply locate the maximal sets of the associated density estimates. The Grenander estimate also relies upon a nearest neighbor density estimate, and weights the density estimate's maximal value most heavily, though it averages it with the midpoints of other spacings in the sample.

The implication is that we will look to generalize the univariate mode estimators to high dimensional spaces via generalization of the associated density estimators. As we want to rule out exhaustive search through the sample space for reasons of economy, we will look for iterative optimization procedures based upon the estimated density to locate the mode or modes. The Grenander estimate, because of its essential dependence upon the ordering of the real line, cannot be extended verbatim beyond one dimension. However, its formulation (2.1.4) as a weighted average of the sample points certainly extends, although different weighting functions may be desirable. For example, the weighting function may be truncated so that local clusters of sample points, indicating local modes, may be identified. These considerations lead us to a study of kernel and variable kernel density estimates, the latter being an amalgamation of kernel and nearest neighbor techniques.

2.2. Kernel Estimators

2.2.1. Introduction

Consider a sample of n observations of a p -dimensional random vector X . Superscripts will denote individual

observations, i.e., $x^{(j)} = (x_1^{(j)}, \dots, x_p^{(j)})^T$ is the j -th observation in the sample. Suppose that f is the density of X with respect to Lebesgue measure on \mathbb{R}^p . We will be interested in designing kernel estimators for f , and, for use as components in optimization algorithms, in estimators for ∇f and $\nabla^2 f$ as well. Estimators will be distinguished from true values by a subscript indicating sample size. For example, $f_n(x)$ estimates $f(x)$. Constructive papers for such estimators are due to Parzen [1962], Epanechnikov [1969], and Singh [1976].

Parzen constructed kernel estimators for univariate densities, established consistency of the estimators, and calculated optimal asymptotic rates of convergence. He also established consistency and asymptotic normality of estimates of the mode obtained from the kernel estimator. Parzen's work was extended to multidimensional densities by Cacoullos [1966] and further considered by Epanechnikov. Epanechnikov's kernels are constructed as a product of univariate kernels applied to each coordinate projection of $x - x^{(j)}$; that is,

$$f_n(x) = \frac{1}{n} \sum_{j=1}^n \prod_{i=1}^p \frac{1}{h_i(n)} K_i\left(\frac{x_i - x_i^{(j)}}{h_i(n)}\right), \quad (2.2.1)$$

where each $K_i: \mathbb{R}^1 \rightarrow \mathbb{R}^1$ satisfies

- a) $0 \leq K_i(y) < c < \infty, i = 1, \dots, p$ (2.2.2)
- b) $K_i(y) = K_i(-y)$ (implying $\int y K(y) dy = 0$)
- c) $\int K_i(y) dy = 1$

$$d) \int y^2 K_i(y) dy = 1$$

$$e) \int y^m K_i(y) dy < \infty \text{ for } 0 \leq m < \infty.$$

Epanechnikov examined the mean squared error of the kernel estimate of f at a point $x \in \mathbb{R}^p$, and by expanding f in a Taylor series about x showed that asymptotically,

$$\text{Bias}(f_n(x)) \sim \frac{1}{2} \sum_{i=1}^p \frac{\partial^2 f(x_1, \dots, x_m)}{\partial^2 x_i} h_i^2(n), \quad (2.2.3)$$

and

$$\text{Var}(f_n(x)) \sim \frac{1}{n} f(x) \sum_{i=1}^p \left[\frac{1}{h_i(n)} \int_{-\infty}^{\infty} K_i^2(y) dy \right], \quad (2.2.4)$$

which gives consistency in mean square if $h_i(n) \rightarrow 0$ and

$$n \sum_{i=1}^p h_i(n) \rightarrow \infty \text{ as } n \rightarrow \infty.$$

By minimizing the expression for global integrated mean squared error with respect to $h(n)$, taking the same kernel and same $h(n)$ in each coordinate direction, he determined that the optimal rate of convergence of the kernel estimator is of order $n^{-4/(4+p)}$, achieved by taking $h(n)$ proportional to $n^{-1/(p+4)}$.

Finally, Epanechnikov discusses the efficiency of various kernel functions. Under his assumption (d), the kernel function enters the expression for mean squared error only through the integral $\int K^2(y) dy$. An optimal kernel will be one which minimizes $\int K^2(y) dy$ subject to the constraints (2.2.2). Using a simple variational argument, details of

which are reproduced in Section 2.2.3, he arrives at the quadratic

$$K_*(y) = (3/4\sqrt{5}) (1 - \frac{y^2}{5}) I_{[-\sqrt{5}, \sqrt{5}]}(y). \quad (2.2.5)$$

Taking as a measure of efficiency for K the ratio

$$r = \int K^2(y) dy / \int K_*^2(y) dy,$$

Epanechnikov shows that no intuitively appealing kernel function loses more than a few percentage points of efficiency, and thus that kernel functions may be safely chosen for computational ease or desired smoothness properties.

Singh [1976] addressed the estimation of mixed partials of a probability density function. Writing

$$f^{(\underline{r})}(x) = \frac{\partial^{r_1} \dots \partial^{r_p}}{\partial^{r_1} x_1 \dots \partial^{r_p} x_p} f(x),$$

where $\underline{r} = (r_1, r_2, \dots, r_p)^T \in \mathbb{R}^p$, his estimator is

$$f_n^{(\underline{r})}(x) = \frac{1}{n} \sum_{j=1}^n \prod_{i=1}^p \frac{1}{h_i(n)^{p_i+1}} K_i\left(\frac{x_i^{(j)} - x_i}{h_i(n)}\right). \quad (2.2.6)$$

Again by considering the Taylor formula for f about x , he finds that to avoid asymptotic bias we must have

$$\int y^s K_i(y) dy = 0 \quad \text{if } s < p_i, \quad (2.2.7)$$

and

$$\int y^{p_i} K_i(y) dy = p_i!. \quad (2.2.8)$$

Singh establishes the pointwise and uniform convergence of these estimators given simple assumptions on f and its derivatives, and appropriate sequence of smoothing parameters $h_1(n)$. He does not provide any insight regarding rate of convergence or kernel choice.

2.2.2. Consistency in Mean Squared Error for Kernels with Infinite Support

The estimators discussed by Epanechnikov and Singh both employ kernels formed as a product of univariate kernels having finite support. The product formulation is useful but not necessary for either analytical or computational purposes. The restriction to finite support is less arbitrary, as the optimality criteria of Epanechnikov yield finitely supported kernels, which may also have computational advantages. On the other hand, arguments by Silverman [1981] suggest the usefulness of Gaussian kernels for investigating multimodality. In addition, a Gaussian kernel underlies the weighted mean update, which is recommended as a mode estimator in Chapter V. To prove the consistency of the mode estimator, it is expedient to establish consistency of kernel density estimates using kernels with infinite support.

Approximating Kernels

Our asymptotic analysis follows roughly the original development by Parzen, and depends upon the following multi-dimensional analog of a result of Bochner [1955]:

Lemma 2.2.1. (A) Suppose $g: \mathbb{R}^p \rightarrow \mathbb{R}^1 \in L^1(\mathbb{R}^p)$, there exists $M > 0$ such that $|g(x)| \leq M$ for all x , and $K: \mathbb{R}^p \rightarrow \mathbb{R}^1$ satisfies:

$$\begin{aligned} \text{(i)} \quad & \int_{\mathbb{R}^p} K(y) dy < \infty \\ \text{(ii)} \quad & \sup_{y \in \mathbb{R}^p} |K(y)| < \infty. \end{aligned} \tag{2.2.9}$$

Let $\{H_n\}$ be a sequence of nonsingular matrices in $\mathbb{R}^{p \times p}$ such that, using a vector norm $\|\cdot\|$, the sequence of induced norms on $\{H_n\}$, which we will write $\{\nu(H_n)\}$, converges to zero. For each n , define

$$g_n(x) = |H_n^{-1}| \int_{\mathbb{R}^p} K(H_n^{-1}y) g(x-y) dy. \tag{2.2.10}$$

Then

$$g_n(x) \rightarrow g(x) \int_{\mathbb{R}^p} K(y) dy \text{ as } n \rightarrow \infty \tag{2.2.11}$$

at all continuity points of g .

(B) If g is uniformly continuous the convergence in (2.2.11) is uniform.

Proof. $|g_n(x) - g(x) \int_{\mathbb{R}^p} K(y) dy| =$

$$\begin{aligned} &= \left| \int_{\mathbb{R}^p} [g(x-y) - g(x)] |H_n^{-1}| K(H_n^{-1}y) dy \right| \\ &\leq \sup_{\|y\| \leq \delta} |g(x-y) - g(x)| \int_{\|y\| \leq \delta} |K(H_n^{-1}y)| |H_n^{-1}| dy \\ &\quad + \int_{\|y\| \geq \delta} |g(x-y) - g(x)| |H_n^{-1}| |K(H_n^{-1}y)| dy \end{aligned}$$

$$\begin{aligned}
&\leq \sup_{\|y\| \leq \delta} |g(x-y) - g(x)| \int_{\mathbb{R}^p} K(z) dz \\
&\quad + \int_{\|H_n z\| \geq \delta} |g(x-H_n z) - g(x)| |K(z)| dz. \quad (2.2.12)
\end{aligned}$$

Since by definition of the induced norm, $\|H_n z\| \leq v(H_n) \times \|z\|$, $\|H_n z\| \geq \delta \Rightarrow \|z\| \geq \delta/v(H_n)$.

Therefore, using the boundedness of g ,

$$\begin{aligned}
|g_n(x) - g(x)| \int_{\mathbb{R}^p} K(y) dy &\leq \sup_{\|y\| \leq \delta} |g(x-y) - g(x)| \int_{\mathbb{R}^p} K(z) dz \\
&\quad + 2M \int_{\|z\| \geq \delta/v(H_n)} |K(z)| dz. \quad (2.2.13)
\end{aligned}$$

For any $\delta > 0$, since $v(H_n) \rightarrow 0$ as $n \rightarrow \infty$, and K is integrable, the second term vanishes as $n \rightarrow \infty$. Thus, given $\epsilon > 0$, if g is continuous at x , there exists $\delta_\epsilon > 0$ such that

$$\sup_{\|y\| \leq \delta_\epsilon} |g(x-y) - g(x)| \int_{\mathbb{R}^p} K(z) dz < \epsilon/2,$$

and there is an n such that

$$2M \int_{\|z\| \geq \delta_\epsilon/v(H_n)} |K(z)| dz < \epsilon/2.$$

If g is uniformly continuous, δ_ϵ applies for all x . Since the above holds for arbitrary ϵ , the lemma is proved.

If g may be unbounded, additional conditions must be imposed upon the kernel function and the set of smoothing parameters.

Lemma 2.2.2. All definitions and assumptions of Lemma 2.2.1 apply except that g may be unbounded. Assume that $\|z\|^P K(z) \rightarrow 0$ as $\|z\| \rightarrow \infty$, that the H_i are diagonal, and that for some $c > 0$, and for all n ,

$$\min_i \{|h_i(n)|\} / \max_i \{|h_i(n)|\} \geq c.$$

Then again, as $n \rightarrow \infty$,

$$g_n(x) \rightarrow g(x) \quad \int_{\mathbb{R}^P} K(y) dy$$

at all continuity points of g , and the convergence is uniform if g is uniformly continuous.

Proof. The second term in (2.3.12) becomes

$$\begin{aligned} & \int_{\|z\| \geq \delta/v(H_n)} |g(x - H_n z) - g(x)| |K(z)| dz \\ & \leq \int_{\|z\| \geq \delta/v(H_n)} |g(x - H_n z)| |K(z)| (\|z\|^P / \|z\|^P) dz \\ & \quad + g(x) \int_{\|z\| \geq \delta/v(H_n)} |K(z)| dz. \end{aligned} \tag{2.2.14}$$

The second integral vanishes as $n \rightarrow \infty$ by the integrability of K . The first is bounded by

$$\sup_{\|z\| \geq \delta/v(H_n)} (\|z\| |K(z)|) \delta^{-P} \frac{v(H_n)^P}{\det(H_n)} \int_{\mathbb{R}^P} g(u) du.$$

With the ratio of the h_i bounded away from zero,

$v(H_n)^P / \det(H_n)$ remains bounded as $v(H_n) \rightarrow 0$. Thus, both

integrals in (2.2.14) vanish as $n \rightarrow \infty$, and the lemma is proved.

Kernel Estimates of a Density and Its Derivatives

We present an estimator for

$$f^{(\underline{r})}(x) = \frac{\partial^{r_1} \dots \partial^{r_p}}{\partial x_1^{r_1} \dots \partial x_p^{r_p}} f(x),$$

assuming the context given below:

- (i) There exists a neighborhood of x , $N(x)$, on which $f \in C^{|\underline{r}|}(N(x))$.
- (ii) Let $\{H_n\} = \{\text{diag}(h_1(n), \dots, h_p(n))\}$ be a sequence of diagonal scaling matrices such that $\max_i h_i(n) \rightarrow 0$ as $n \rightarrow \infty$, and for some $c > 0$, for each n , $\min_i(h_i(n))/\max_i(h_i(n)) > c$.
- (iii) $K(y): \mathbb{R}^p \rightarrow \mathbb{R}^1$ satisfies

$$|K(y)| \leq M \quad \forall y \in \mathbb{R}^p$$

$$\int y^{\underline{k}} K(y) dy = 0 \quad \forall \underline{k} \text{ with } |\underline{k}| < |\underline{r}|$$

$$\text{or with } |\underline{k}| = |\underline{r}| \text{ but } \underline{k} \neq \underline{r}.$$

$$\int y^{\underline{r}} K(y) dy = \underline{r}!$$

$$\int \|y\|^{|\underline{r}|} K(y) dy < \infty$$

$$\|y\|^{p+|\underline{r}|} K(y) \rightarrow 0 \text{ as } \|y\| \rightarrow \infty. \quad (2.2.15)$$

Note that a multi-index notation is being used in which

$$\underline{r} = (r_1, r_2, \dots, r_p)^T \in \mathbb{R}^p,$$

$$|\underline{r}| = \sum_{i=1}^p r_i, \quad \underline{r}! = \prod_{i=1}^p r_i!,$$

$$\binom{s}{\underline{r}} = \frac{s!}{(s-\underline{r})! \underline{r}!}, \quad s > |\underline{r}|,$$

and, for $x \in \mathbb{R}^p$,

$$x^{\underline{r}} = \prod_{i=1}^p x_i^{r_i}.$$

The kernel estimator is defined as

$$f_n^{(\underline{r})}(x) = \frac{1}{n} \frac{1}{\prod_{i=1}^p h_i(n)^{r_i+1}} \sum_{j=1}^n K(H_n^{-1}(x^{(j)} - x)). \quad (2.2.16)$$

Asymptotic Unbiasedness

$$\begin{aligned} Ef_n^{(\underline{r})}(x) &= \left(\prod_{i=1}^p h_i(n)^{-r_i} \right) |H_n|^{-1} \int_{\mathbb{R}^p} K(H_n^{-1}(y-x)) f(y) dy \\ &= \left(\prod_{i=1}^p h_i(n)^{-r_i} \right) \int_{\mathbb{R}^p} K(y) f(x+H_n y) dy. \end{aligned} \quad (2.2.17)$$

Suppose that the ball of radius ρ , centered at x , is contained in $N(x)$ (see 2.2.15), and split \mathbb{R}^p into the domains

$$\{y: \|H_n y\| \leq \rho\} \text{ and } \{y: \|H_n y\| \geq \rho\} \subset \{y: \|y\| \geq \rho/v(H_n)\}.$$

Under the assumptions of (2.2.15.ii), as described in the proof of Lemma 2.2.2, the contribution of the "tail" component of the domain vanishes asymptotically.

$$Ef_n^{(\underline{r})}(x) = \left(\prod_{i=1}^p h_i(n)^{-r_i} \right) \int_{\|H_n y\| \leq \rho} K(y) f(x+H_n y) dy + o(n).$$

By assumption, f in the integrand may be given by Taylor's formula with p -th order remainder term, yielding

$$\begin{aligned}
E f_n^{(\underline{r})}(x) &= \left(\prod_{i=1}^p h_i(n)^{-r_i} \right) \int_{\|H_n Y\| \leq \rho} K(y) [f(x) + \\
&\quad + \sum_{s=1}^{|\underline{r}|} \frac{1}{s!} \sum_{|\underline{q}|=s} \binom{s}{\underline{q}} f^{(\underline{q})}(x) (H_n Y)^{\underline{q}}] dy + o(n) \\
&\qquad\qquad\qquad (2.2.18) \\
&= \left(\prod_{i=1}^p h_i(n)^{-r_i} \right) \int_{\mathbb{R}^p} K(y) [f(x) + \\
&\quad + \sum_{s=1}^{|\underline{r}|} \frac{1}{s!} \sum_{|\underline{q}|=s} \binom{s}{\underline{q}} f^{(\underline{q})}(x) (H_n Y)^{\underline{q}}] dy + o(n).
\end{aligned}$$

Moment conditions on the kernel remove all terms except those with $\underline{q} = \underline{r}$. Therefore,

$$E f_n^{(\underline{r})} = \frac{1}{r!} \left(\prod_{i=1}^p h_i(n)^{-r_i} \right) f^{(\underline{r})}(x) \int_{\mathbb{R}^p} (H_n Y)^{\underline{r}} K(y) dy.$$

But

$$\begin{aligned}
(H_n Y)^{\underline{r}} &= \prod_{i=1}^p (h_i(n) y_i)^{r_i} \\
&= \left(\prod_{i=1}^p h_i(n)^{r_i} \right) y^{\underline{r}}.
\end{aligned}$$

Applying Lemmas 2.2.1 and 2.2.2, $f_n^{(\underline{r})}$ is asymptotically unbiased if $\max_i h_i(n) \rightarrow 0$ as $n \rightarrow \infty$, uniformly so if $f_n^{(\underline{r})}$ is uniformly continuous.

Asymptotic Variance

Since the $X^{(j)}$ are assumed independent and identically distributed,

$$\begin{aligned}
\text{Var } f_n^{(\underline{r})}(x) &= \frac{1}{n} \left(\prod_{i=1}^p h_i(n)^{-2r_i} \right) |H_n|^{-2} \text{Var } K(H_n^{-1}(x-X)) \\
&= \frac{1}{n} \left(\prod_{i=1}^p h_i(n)^{-2r_i} \right) |H_n|^{-2} \{ \int K^2(H_n^{-1}(x-y) f(y)) dy \\
&\quad - [\int K(H_n^{-1}(x-y)) f(y) dy]^2 \}.
\end{aligned}$$

Applying the Lemmas 2.2.1 and 2.2.2 to both integrals gives that

$$\begin{aligned}
\text{Var } f_n^{(\underline{r})}(x) &\rightarrow \frac{1}{n \prod_{i=1}^p h_i(n)^{2r_i}} \{ |H_n|^{-1} f(x) \int K^2(y) dy \\
&\quad - [f(x) \int K(y) dy]^2 \}
\end{aligned}$$

as $n \rightarrow \infty$. Now $v(H_n) \rightarrow 0 \Rightarrow \max_{i,j} |(h_{ij})_n| \rightarrow 0 \Rightarrow |H_n|^{-1} \rightarrow +\infty$ as $n \rightarrow \infty$, so that asymptotically the second term becomes negligible. Therefore,

$$\text{Var } f_n^{(\underline{r})}(x) \approx \frac{1}{n \left(\prod_{i=1}^p h_i(n)^{2r_i} \right) |H_n|} f(x) \int K^2(y) dy, \quad (2.2.19)$$

and for the variance of $f_n^{(\underline{r})}(x)$ to vanish as $n \rightarrow \infty$ it is necessary and sufficient to have

$$n \left(\prod_{i=1}^p h_i(n)^{2r_i} \right) |H_n| \rightarrow \infty.$$

The convergence is uniform if f is bounded, which is assured if $f_n^{(\underline{r})}$ is uniformly continuous.

2.2.3. Mean Squared Error; Kernel Design

One of the principal methods for directing the choice of smoothing parameters and kernel shape has been the analysis of the mean squared error of the estimator, either pointwise,

$$E(f_n^{(\underline{r})}(x) - f(x))^2,$$

or integrated over the support of the density. To perform such an analysis requires expanding Singh's moment conditions on the kernel function (2.2.7 and 2.2.8), so that, assuming product kernels, for some $t_i > r_i$ the one-dimensional kernels satisfy

$$\begin{aligned} \int y^s K_i(y) dy &= 0 \quad r_i < s < t_i \\ \int y^{t_i} K_i(y) dy &< \infty. \end{aligned}$$

In estimating the density itself we are held to only one additional zero moment by the desire to keep K_i non-negative. In dealing with derivatives of f there is no such restriction and we could take t_i to be any order higher than p_i , but there is little practical advantage gained by taking $t_i > p_i + 2$.

For mean square analysis, we make these modifications to (2.2.15):

- (i') There exists a neighborhood of x , $N(x)$, on which $f \in C^{|\underline{r}|+2}(N(x))$.

$$\begin{aligned}
\text{(iii)} \quad K(y) &= (-1)^{|\underline{r}|} K(-y), \quad y \in \mathbb{R}^p \\
\int_{\mathbb{R}^p} y^{\underline{k}} K(y) dy &= \underline{r}! \quad \text{for all } \underline{k} = \underline{r} + 2\mathbf{e}_i, \\
i &= 1, \dots, p, \text{ where } \mathbf{e}_i \text{ is the } i\text{-th Kronecker} \\
&\text{basis element.}
\end{aligned} \tag{2.2.20}$$

The expression (2.2.19) for asymptotic variance is unaffected by the revised assumptions. To evaluate the bias term, again f is expanded in a Taylor series about x . The moment conditions in (2.2.20) are designed to strip as many terms as possible from the expectation of the estimator, leaving, much as in (2.2.18),

$$\begin{aligned}
Ef_n^{(\underline{r})}(x) - f^{(\underline{r})}(x) &= \left\{ \sum_{j=1}^p \frac{1}{(\underline{r}+2\mathbf{e}_j)!} \left(\prod_{i=1}^p h_i(n)^{-r_i} \right) f^{(\underline{r}+2\mathbf{e}_j)}(x) \times \right. \\
&\quad \left. \int_{\mathbb{R}^p} (H_n y)^{\underline{r}+2\mathbf{e}_j} K(y) dy \right\} + o(n).
\end{aligned} \tag{2.2.21}$$

Taking $H_n = h(n)I$, I the $p \times p$ identity matrix, the expression simplifies to

$$Ef_n^{(\underline{r})}(x) - f^{(\underline{r})}(x) = \frac{1}{2} h(n)^2 \sum_{j=1}^p f^{(\underline{r}+2\mathbf{e}_j)}(x) + o(n). \tag{2.2.22}$$

Combining (2.2.22) and (2.2.19), we get

$$MSE(f_n^{(\underline{r})}(x)) = \frac{f(x) \int_{\mathbb{R}^p} K^2(y) dy}{nh(n)^{p+2} |\underline{r}|} + \frac{h(n)^4}{4} \left(\sum_{j=1}^p f^{(\underline{r}+2\mathbf{e}_j)}(x) \right)^2 + o(n). \tag{2.2.23}$$

Solving for a value of $h(n)$ which will minimize (2.2.23)

gives

$$h(n) = c_1 n^{-1/(p+4+2|\underline{r}|)}, \tag{2.2.24}$$

where c_1 is a constant which depends upon the higher derivatives of f in a complicated way, and an optimal rate of convergence for $f_n^{(r)}(x)$,

$$\text{MSE}(f_n^{(r)}(x)) = c_2 n^{-4/(p+4+2|r|)} . \quad (2.2.25)$$

Optimal Kernels

The form of the kernel function used for estimating $f^{(r)}(x)$ appears in the expression for mean squared error (2.2.23) only through the integral $\int K^2(y) dy$. The kernel which is optimal for reducing mean squared error is that function which minimizes $\int K^2(y) dy$, subject to the constraints (2.2.15) and (2.2.20). If the kernel used has a product formulation, each one-dimensional kernel must solve

$$\text{minimize } \int K_i^2(t) dt , \quad K_i \in L^2(-\infty, \infty) \quad (2.2.26)$$

S/T

$$\int t^k K_i(t) dt = \begin{cases} 0 & 0 \leq k < r_i \text{ or } k = r_i + 1 \\ r_i! , & k = r_i \text{ or } k = r_i + 2. \end{cases} \quad (2.2.27)$$

The constraints (2.2.27) are linear functionals applied to K_i and thus are their own derivatives. Their representants are the monomials $\Pi_j(x) = x^j$. Lagrange multiplier theory provides that K_i is an (r_i+2) -degree polynomial restricted to a compact interval; that is, for some $a > 0$, there exist scalars, $\lambda_0, \lambda_1, \dots, \lambda_{r_i+2}$, such that

$$K_i(t) = (\lambda_0 + \lambda_1 t + \dots + \lambda_{r_i+2} t^{r_i+2}) I_{[-a,a]}(t). \quad (2.2.28)$$

The coefficients are determined by solving the system of linear equations generated by substituting (2.2.20) into (2.2.27). The symmetry conditions immediately require $\lambda_k = 0$ for all odd or even k , depending on whether r_i is even or odd, respectively.

The optimal kernel for estimating the density itself (2.2.5) was given by Epanechnikov [1969]. Since we will be investigating Newton methods for finding local maxima of the estimated density, it is also necessary to design kernels for estimating the first and second partial derivatives of f . The functions obtained by the procedure described above are,

$$K_1(t) = \frac{15}{4} \left(\frac{42}{79}\right)^{3/2} [t - \frac{42}{49} t^3], \quad |t| \leq \left(\frac{79}{42}\right)^{1/2}, \quad (2.2.29)$$

for estimating first partials, and

$$K_2(t) = \frac{35}{2} \left(\frac{2}{3}\right)^{1/2} [-\frac{1}{4} + t^2 - \frac{5}{9} t^4], \quad |t| \leq \left(\frac{3}{2}\right)^{1/2}, \quad (2.2.30)$$

used for second partial derivatives. The kernels are pictured in Figure 4.1.1.

2.2.4. Variable Kernel Estimators

The choice of smoothing matrix H_n , or bandwidth, is the crucial issue in kernel estimation, both in asymptotic analysis and in designing estimates for practical applica-

tions on finite samples. Many methods have been developed for guiding the choice of the $h_i(n)$, based upon asymptotic expressions such as (2.2.24) and the observed derivatives of the estimated density [Scott, 1976; Tapia and Thompson, 1978; Factor, 1979; Nezames, 1980], the modality of f_n [Silverman, 1978], the likelihood of the sample under f_n [Duin, 1976], or cross-validation techniques [Wahba, 1977]. Typically the best values of $h_i(n)$ just exceed a threshold beneath which the density estimates rapidly evince local oscillation and spurious modes. For this reason, the character of estimates in the "good" range is sensitive to relatively minor changes in the bandwidth, and though the best of the procedures mentioned above are generally quite effective, they must still be confirmed by visual display and user interaction.

Throughout our development it has been assumed that the bandwidth is prescribed and held constant for all locations x . One adaptive technique that has received close attention in multiple dimensions is to make the kernel width used for the estimate $f_n(x)$ be some function of the distance from x to its closest neighbors in the sample. The most common function is simply the distance to the k -th neighbor, where k is a fixed integer, $1 \leq k \leq n$. Estimators with location dependent smoothing terms will be called "variable kernel" or nearest neighbor estimators. To be more specific, the class of variable kernel estimators is defined as follows: given the random sample $\{x_1, x_2, \dots, x_n\}$ and a point x , let

$\{r_1(x), r_2(x), \dots, r_n(x)\}$ be the distances $\{\|x - x_i\|, i = 1, \dots, n\}$, measured in some norm. Let $\{r_{[1:n]}(x), r_{[2:n]}(x), \dots, r_{[n:n]}(x)\}$ be the same distances arranged in ascending order; these are the order statistics of the random variable, $R = \|x - X\|$. Finally, let $\{x_{[1:n]}, x_{[2:n]}, \dots, x_{[n:n]}\}$ be the original sample points arranged by the order relation of R . Then, with $B(x; \rho)$ denoting the closed ball of radius ρ about x , and $V(x; \rho)$ denoting its volume, the nearest neighbor estimate of f is

$$\bar{f}_n(x; k) = \frac{1}{nV(x; r_{[k:n]}(x))} \sum_{i=1}^n K\left(\frac{x - x_i}{r_{[k:n]}(x)}\right).$$

Also, the parameter k is regarded as a function of n , written $k(n)$.

The appeal of the nearest neighbor estimator is twofold. First, k -th neighbor adaptation corresponds roughly to psychological perceptions of the role of bandwidth. In order to achieve stability in our estimates, we know it is necessary to average the effects over a containing region to obtain the estimate at a point. Where the density is high the clustering of points should be dense, and the smoothing interval can be relatively small; where the density is small, the smoothing interval ought to expand enough to capture some of the now sparsely distributed sample. The neighbor distances promise the ability of order statistics to sense the natural scale of the problem. Secondly, $k(n)$, or equivalently a fraction of the sample, $k(n)/n$, is a more convenient

control parameter for most users than the more abstract choice of bandwidth. It is less scale dependent. For example, the choice of $k(n)$ should be unaffected by the common practice of rescaling each coordinate to have unit variance; for $h(n)$ this is clearly not the case. In addition, the variable kernel estimators appear to be more robust with respect to the choice of $k(n)$ than fixed bandwidth estimators are with respect to $h(n)$. Such robustness was noted by Breiman, Meisel, and Purcell [1977] in testing of two-dimensional density estimators. Since as a rough rule of thumb, equivalent smoothing in \mathbb{R}^p should be obtained with $(k(n)/n)$ proportional to $h(n)^{1/p}$, this effect should become more pronounced as the dimension increases.

The variable kernel technique presents some theoretical problems which the fixed estimator bandwidth does not share. Comments by Moore and Yackel [1977] demonstrate that, if the support of the kernel is infinite, nearest neighbor distances break down as analogues of bandwidth in certain circumstances. Their remarks are echoed by Mack and Rosenblatt [1977], who study asymptotic convergence of the variable kernel estimator in mean squared error, and warn against heavy bias in the tails. However, these difficulties are confined to estimation over regions where the density may be zero or infinitely small. When the density in the region of interest may be bounded away from zero, as is certainly reasonable in the neighborhood of sample or population

modes, the variable and fixed bandwidth estimators have been shown [Moore and Yackel, 1977] to have equivalent convergence properties.

2.3. Consistency of Mode Estimators

As stated in Section 1 of this chapter, common univariate estimators of the mode are based on nonparametric estimators of a probability density function. The same will be true of the multivariate mode estimators we study. The following proposition provides general conditions, first given by Parzen [1962], under which local maximizers of the estimated density are consistent for local modes of the population density.

Proposition 2.3.1. Suppose $f: \mathbb{R}^p \rightarrow \mathbb{R}$ is uniformly continuous in a closed set D , and that f has a unique maximum in D at θ . Let f_n be an estimator of f based on a sample of size n , and let θ_n be the maximizer of f_n in D . Then

$$f_n \xrightarrow{P} f \text{ uniformly in } D \Rightarrow \theta_n \xrightarrow{P} \theta, \text{ and}$$

$$f_n \xrightarrow{\text{w.p.1}} f \text{ uniformly in } D \Rightarrow \theta_n \xrightarrow{\text{w.p.1}} \theta.$$

Proof. By the uniqueness of θ in D , given $\varepsilon > 0 \exists \eta > 0$ such that $x \in D$ and $\|x - \theta\| > \varepsilon \Rightarrow |f(x) - f(\theta)| > \eta$. To see this, assume the converse. Then, for some $\varepsilon > 0$, there is a sequence $\{x_n\} \subset D$ such that $\inf_n \|x_n - \theta\| \geq \varepsilon$ and $|f(x_n) - f(\theta)| \leq 1/n$. Therefore, $\sup_{x \in D \cap B(\theta, \varepsilon)} f(x) = f(\theta)$.

Since f is a probability density function and

$$\lim_{\|x\| \rightarrow \infty} f(x) = 0,$$

there exists an $R > 0$ such that also $f(\theta) = \sup f(x)$ over $x \in D \cap B(\theta, \varepsilon)^c \cap B(\theta, R)$. But the latter is a compact set, implying that f assumes its maximum in $D \cap B(\theta, \varepsilon)^c$, thus contradicting the uniqueness of θ .

With ε, η as above, $P[\|\theta_n - \theta\| > \varepsilon] \leq P[|f(\theta_n) - f(\theta)| > \eta]$. Therefore, if for every $\eta > 0$, $P[|f(\theta_n) - f(\theta)| > \eta] \rightarrow 0$ as $n \rightarrow \infty$, then for every $\varepsilon > 0$, $P[\|\theta_n - \theta\| > \varepsilon] \rightarrow 0$ as $n \rightarrow \infty$. That is $f(\theta_n) \xrightarrow{P} f(\theta) \Rightarrow \theta_n \xrightarrow{P} \theta$. Similarly, $f(\theta_n) \xrightarrow{w.p.1} f(\theta) \Rightarrow \theta_n \xrightarrow{w.p.1} \theta$. Now $|f(\theta_n) - f(\theta)| \leq |f(\theta_n) - f_n(\theta_n)| + |f_n(\theta_n) - f(\theta)|$. The first term on the right clearly inherits any uniform convergence properties of $f_n(\cdot)$. For the second term, suppose $f_n(\theta_n) > f(\theta)$. Then $f_n(\theta_n) - f(\theta) < f_n(\theta_n) - f(\theta_n) \leq \sup_{x \in D} |f_n(x) - f(x)|$. Similarly, $f(\theta) > f_n(\theta_n)$ implies that $f(\theta_n) - f_n(\theta_n) \leq \sup_{x \in D} |f_n(x) - f(x)|$. Thus $|f(\theta_n) - f(\theta)| \leq 2 \sup_{x \in D} |f_n(x) - f(x)|$. Therefore,

$$f_n \xrightarrow{P} f \text{ uniformly in } D$$

$$\Leftrightarrow \lim_{n \rightarrow \infty} P\{\sup_{x \in D} |f_n(x) - f(x)| > \eta\} = 0 \text{ for all } \eta$$

$$\Rightarrow \lim_{n \rightarrow \infty} P\{\sup_{x \in D} |f(\theta_n) - f(\theta)| > \eta\} = 0 \text{ for all } \eta$$

$$\Rightarrow \theta_n \xrightarrow{P} \theta,$$

and

$f_n \xrightarrow{w.p.1} f$ uniformly in D

$$\Leftrightarrow P\{\lim_{n \rightarrow \infty} (\sup_{x \in D} |f_n(x) - f(x)|) = 0\} = 1$$

$$\Rightarrow P\{\lim_{n \rightarrow \infty} |f(\theta_n) - f(\theta)| = 0\} = 1$$

$$\Rightarrow \theta_n \xrightarrow{w.p.1} \theta.$$

This proves the proposition.

Unfortunately, uniform consistency of kernel estimators is not guaranteed by our previous analyses of mean square convergence. From the study of asymptotic bias we have easily that $\sup_x |Ef_n(x) - f(x)| \rightarrow 0$ as $n \rightarrow \infty$ if f is uniformly continuous. However, the fluctuation of the estimate is given only in terms of expectations. We have from (2.2.19) that if

$$n \prod_{i=1}^p h_i(n) \rightarrow \infty$$

as $n \rightarrow \infty$ and f is uniformly continuous that $\sup_x \text{Var}(f_n(x)) \rightarrow 0$; however, this is not sufficient for the conclusion that $\sup_x |f_n(x) - Ef_n(x)| \rightarrow 0$.

Strong uniform consistency was proved by Van Ryzin [1969], under rather more restrictive conditions on the sequence of smoothing parameters. Van Ryzin's result is as follows:

Proposition 2.3.2. Let K be as described in (2.2.15), with $r = 0$. Let

$$k(t) = \int_{\mathbb{R}^p} e^{i\langle t, u \rangle} K(u) du,$$

where $\langle x, y \rangle$ is the standard Euclidean inner product, and assume that $k(\cdot)$ is absolutely integrable, and that $g(c) = \int |k(ct) - k(t)| dt$ is locally Lipschitz of order 1 at $c = 1$. Let $\{h(n)\}$ be a sequence such that

$$\begin{aligned} & \text{(i)} \quad h(n) \rightarrow 0 \text{ as } n \rightarrow \infty \\ & \text{(ii)} \quad nh(n)^{2p} \rightarrow \infty \text{ as } n \rightarrow \infty \\ & \text{(iii)} \quad \sum_{n=1}^{\infty} (nh(n)^p)^{-2} < \infty \end{aligned} \tag{2.3.1}$$

and

$$\text{(iv)} \quad \sum_{n=1}^{\infty} \frac{1}{nh(n)^{2p-1}} \left| \frac{1}{h(n+1)} - \frac{1}{h(n)} \right| < \infty.$$

Then with $H_n = h(n)I$, if $f(x)$ is uniformly continuous on \mathbb{R}^p ,

$$\sup_x |f_n(x) - f(x)| \rightarrow 0 \text{ w.p.1 as } n \rightarrow \infty.$$

Moore and Yackel [1977] extended Van Ryzin's result to apply also to a uniform kernel (which does not have an absolutely integrable characteristic function), and they further establish the equivalence of convergence properties of kernel and variable kernel density estimators over any set on which the value of the true density function is bounded away from zero. For the result we need, suppose that $r(n) \rightarrow 0$ and $nr(n)^p \rightarrow \infty$ as $n \rightarrow \infty$. Writing $\hat{f}_n(x; \alpha)$ for the fixed bandwidth kernel estimate of $f(x)$, using

$h(n) = \alpha r(n)$, and kernel K , $\hat{g}_n(x; \alpha)$ for the special case of a uniform kernel, and writing $f_n(x; \beta)$ for the variable kernel estimator with $k(n) = \beta n r(n)^p$, Moore and Yackel prove the following result:

Proposition 2.3.3. Let $\epsilon > 0$ be given and choose $0 < \delta < \epsilon/12$. For any $\beta > 0$ and $x \in \{z: f(z) \geq 2\delta\}$, $|f_n(x; \beta) - f(x)| > \epsilon$ implies that $|\hat{f}_n(x; \alpha) - f(x)| \geq \delta/2$ or $|\hat{g}_n(x; \alpha) - f(x)| \geq \delta/2$ for at least one of a finite set of values of α not depending on n , x , or the sample point w .

They impose one condition on the kernel function K that, though quite natural, is not strictly necessary for the consistency of the fixed bandwidth estimator, namely that $K(cu) \geq K(u)$ for all u and for any c , $0 \leq c \leq 1$.

The preceding three propositions allow us to prove that maximizers of either the kernel or variable kernel estimator over specified regions are consistent estimators of the corresponding local maxima of the population density.

Theorem 2.3.1. Suppose $\{h(n)\}$ is a sequence satisfying the conditions 2.3.1 (i)-(iv) of Proposition 2.3.2, and $K(n)$ is the sequence of integers $[nh(n)^p]$, where $[x]$ denotes the greatest integer not exceeding x . Let K be a function which is either uniform on the unit ball in some metric, or meets the conditions of Proposition 2.3.2. Assume that $f: \mathbb{R}^p \rightarrow \mathbb{R}$ is a uniformly continuous probability density function. Finally, suppose that for some $\eta > 0$, $D = \{z:$

$f(z) > \eta\}$ is not empty, and θ is the unique local maximizer of f over D . Then, for any $\alpha > 0$, the fixed bandwidth kernel estimate of f , $\hat{f}_n(x; \alpha)$, is uniformly strongly consistent; for any $\beta > 0$ the variable kernel estimator of f , $f_n(x; \beta)$ is uniformly strongly consistent; and in either case, if θ_n maximizes the estimated density over D , $\theta_n \rightarrow \theta$ with probability one as $n \rightarrow \infty$.

Proof. The consistency of $\hat{f}_n(x; \alpha)$ is given by Proposition 2.3.2. Now suppose that $\sup_{x \in D} |f_n(x; \beta) - f(x)|$ does not converge to zero with probability one. Then, for every $\varepsilon > 0$, for every n there is a larger integer $N(\varepsilon, n)$ and some $x \in D$, such that $|f_{N(\varepsilon, n)}(x; \beta) - f(x)| > \varepsilon$. Let $\delta = \min(\eta, \varepsilon/12)/2$. By Proposition 2.3.3, this implies that for one of a finite number of values of α , either

$$(i) \quad |\hat{f}_{N(\varepsilon, n)}(x; \alpha) - f(x)| > \min(\eta/2, \varepsilon/25), \text{ or}$$

$$(ii) \quad |\hat{g}_{N(\varepsilon, n)}(x; \alpha) - f(x)| > \min(\eta/2, \varepsilon/25).$$

Since the inequalities above hold for an infinite number of $N(\varepsilon, n)$, there is a single value α and one of the inequalities holding for an infinite number of values of n , $\{n_1 < n_2 < n_3 < \dots\}$. Suppose (i) holds. Then for small $\varepsilon > 0$, for every n , there is an $n_i > n$ and a point $x \in D$ such that $|\hat{f}_{n_i}(x; \alpha) - f(x)| > \varepsilon/25$. This contradicts the assumption that $\sup_x |\hat{f}_n(x; \alpha) - f(x)| \rightarrow 0$ with probability one as $n \rightarrow \infty$. Therefore, it must be that $f_n(x; \beta)$ is uniformly strongly consis-

tent over D . By Proposition 2.3.1, the local maximizer of either $\hat{f}_n(x; \alpha)$, $\hat{g}_n(x; \alpha)$, or $f_n(x; \beta)$ converges to θ with probability one as $n \rightarrow \infty$. This completes the theorem.

Note that, though the conditions (2.3.1) on $h(n)$ and by extension the conditions on $k(n)$ in Theorem 2.3.1 are fairly complicated, they are easily satisfied, for example, by $h(n) = \alpha n^{(s-1)/p}$ and $k(n) = \beta n^s$, for any $\alpha > 0$, $\beta > 0$, and $1/2 < s < 1$.

Iterative Procedures

The mode-seeking procedures discussed in Chapters III through V may all be related to variable kernel density estimates for which, in each case, Theorem 2.3.1 applies. In each case as well, the mode estimate is a local optimum for the corresponding density estimate. Thus, Theorem 2.3.1 gives some evidence that the procedures discussed in this dissertation are conceptually sound. However, the mode estimation is conducted by means of iterative algorithms which make no attempt at complete search of any region beyond the sequence of neighbor sets which they identify (cf. Algorithm 3.1.1), and for consistency of the associated density estimate, the neighbor sets will have zero Lebesgue measure in the limit as n approaches infinity. It is conceivable that, even if it starts, remains, and converges in a region D having a unique interior local maximizer of the estimated density, any one of our iterative procedures may fail to attain that value. For this reason,

Theorem 2.3.1 does not guarantee the consistency of the mode estimators we will present, even though it does guarantee consistency of modes of the associated kernel estimators.

Consistency results may be established for our iterative algorithms using the fact, which will be established for the various algorithms individually, that each update step produces an increase in the density estimate with which the algorithm is paired, and converges to a local maximizer of that density estimate. The arguments will require a preliminary result establishing consistency of variable kernel estimators of the gradient of the population density. With these facts in hand, it will be shown that asymptotically the mode estimators must converge to a local maximum or a saddlepoint of the population density.

Proposition 2.3.4. Let K_0 be a kernel function satisfying the provisions of (2.3.15) for estimating $\partial/\partial x_m f(\cdot)$. Let e_i be the i -th standard basis element for \mathbb{R}^p . Suppose $K_0(x) = c_1[K_1(x+we_m) - K_1(x-we_m)] - c_2K_2(x)$ for some constant w , where K_1 and K_2 are symmetric non-negative densities satisfying (2.2.15) for estimating f , and further satisfying the monotonicity requirement of Moore and Yackel, that

$$\text{for all } u \in \mathbb{R}^p \text{ and } t, 0 \leq t \leq 1, K_i(tu) \geq K_i(u).$$

(2.3.2)

Suppose that for some $\alpha > 0$ and $s, 1 > s > \max\{1/2, 2/(p+2)\}$, $h(n) = \alpha n^{(s-1)/p}$. Then with $k(n) = \beta n h(n)^p = \beta n^s$, for any

$\beta > 0$, the variable kernel estimator

$$\hat{f}_n(x; \beta) = \frac{1}{nr_k(x)^{p+1}} \sum_{i=1}^n K_0\left(\frac{X^{(i)} - x}{r_k(x)}\right) \quad (2.3.3)$$

consistently estimates $\partial/\partial x_m f(x)$ over any region in which $f(x)$ is bounded away from zero.

Proof. By the asymptotic results of Section 2.2, since $nh(n)^{p+2} \rightarrow \infty$, writing the fixed bandwidth kernel estimator as

$$\hat{f}_n(x) = \frac{1}{nh(n)^{p+1}} \sum_{i=1}^n K_0\left(\frac{X^{(i)} - x}{h(n)}\right), \quad (2.3.4)$$

then $\hat{f}_n(x) \rightarrow \frac{\partial}{\partial x_m} f(x)$ as $n \rightarrow \infty$, with convergence in the sense of quadratic mean. Similarly, writing

$$\hat{\zeta}_n(x) = \frac{1}{nh(n)^{p+1}} \sum_{i=1}^n K_1\left(\frac{X^{(i)} - x}{h(n)}\right)$$

and

$$\hat{\xi}_n(x) = \frac{1}{nh(n)^{p+1}} \sum_{i=1}^n K_2\left(\frac{X^{(i)} - x}{h(n)}\right),$$

both $\hat{\zeta}_n(x) \rightarrow f(x)$ and $\hat{\xi}_n(x) \rightarrow f(x)$ as $n \rightarrow \infty$.

Define

$$\zeta_n(x; \beta) = \frac{1}{nr_k(n)^{p+1}} \sum_{i=1}^n K_1\left(\frac{X^{(i)} - x}{r_k(x)}\right)$$

and

$$\xi_n(x; \beta) = \frac{1}{nr_k(n)^{p+1}} \sum_{i=1}^n K_2\left(\frac{X^{(i)} - x}{r_k(x)}\right).$$

Now

$$\left| f_n(x; \beta) - \frac{\partial}{\partial x_m} f(x) \right| \leq \left| f_n(x; \beta) - \hat{f}_n(x) \right| + \left| \hat{f}_n(x) - \frac{\partial}{\partial x_m} f(x) \right|,$$

and the results of Section 2.2 establish that $|\hat{f}_n(x) - \frac{\partial}{\partial x_m} f(x)| \rightarrow 0$ in quadratic mean or in probability as $n \rightarrow \infty$. In addition, with some straightforward algebra, it is possible to show that

$$\begin{aligned} \left| f_n(x; \beta) - \hat{f}_n(x) \right| &\leq |c_1| \{ |\zeta_n(x + wr_k(x)e_m; \beta) - \zeta_n(x; \beta)| \\ &\quad + |\zeta_n(x; \beta) - \hat{\zeta}_n(x)| \\ &\quad + |\hat{\zeta}_n(x) - \hat{\zeta}_n(x + wr_k(x)e_m)| \\ &\quad + |\zeta_n(x - wr_k(x)e_m; \beta) - \zeta_n(x; \beta)| \\ &\quad + |\zeta_n(x; \beta) - \hat{\zeta}_n(x)| \\ &\quad + |\hat{\zeta}_n(x) - \hat{\zeta}_n(x - wr_k(x)e_m)| \} \\ &\quad + |c_2| \{ |\xi_n(x; \beta) - \hat{\xi}_n(x)| \}. \end{aligned}$$

The arguments sustaining Proposition 2.3.3 and the first part of the proof of Theorem 2.3.1 carry through with the factor $r_k(n)^{p+1}$ replacing $r_k(n)^p$ in the denominator of $\zeta_n, \hat{\zeta}_n, \xi_n$, and $\hat{\xi}_n$, and they provide, since $f(x) > 0$, that both $|\zeta_n(x; \beta) - \hat{\zeta}_n(x)| \rightarrow 0$ and $|\xi_n(x; \beta) - \hat{\xi}_n(x)| \rightarrow 0$ with probability one as $n \rightarrow \infty$. As for the remaining terms, the

choice of $k(n)$ is such that $r_k(x) \rightarrow 0$ w.p.1 as $n \rightarrow \infty$. Since f is continuous at x and $\zeta_n, \hat{\zeta}_n, \xi_n$, and $\hat{\xi}_n$ are consistent for $f(x)$, it must be that $|f_n(x; \beta) - \hat{f}_n(x)| \rightarrow 0$ w.p.1 as $n \rightarrow \infty$, and thus that $|f_n(x; \beta) - \frac{\partial}{\partial x_m} f(x)| \rightarrow 0$ in probability.

Proposition 2.3.5. Let f be a uniformly continuous, continuously differentiable probability density function in \mathbb{R}^p , and $\{X^{(1)}, \dots, X^{(n)}\}$ be a sample of n independent observations drawn from f . Suppose $f_n(\cdot)$ is an estimator of f which is uniformly strongly consistent over any domain in which f is bounded away from zero. Let x^0 be a starting point, independent of n , for which $f(x^0) > 0$, and let $\{x_n^1, x_n^2, \dots\}$ be the sequence of iterates produced by an algorithm which guarantees that $f_n(x_n^{i+1}) > f_n(x_n^i)$ for all i , and suppose that the x_n^i converge to a point x_n^* at which $\nabla f_n(x_n^*) = 0$. Further, define $L_\eta = \{x \in \mathbb{R}^p : f(x) \geq \eta\}$ and suppose there is a real number η , $f(x_0) > \eta > 0$, such that for all $x \in L_\eta$, $\nabla f_n(x) \rightarrow \nabla f(x)$ in quadratic mean as $n \rightarrow \infty$. Then $\nabla f(x_n^*) \rightarrow 0$ almost surely as $n \rightarrow \infty$.

Proof. By the consistency of f_n and the monotonicity of the iterates $\{f(x_n^i), i = 1, 2, \dots\}$, there is an integer M such that for all $n > M$, $f_n(x^0) > \eta$ and $f_n(x_n^i) > \eta$ for all i . Thus, the sequence of mode estimates $\{x_n^*, n > M\}$ must lie in the compact set L_η . We will suppose that, for some $\epsilon > 0$, there is an infinite subsequence of $\{x_n^*, n > M\}$ such

that $\|\nabla f(x_n^*)\| > \varepsilon$ for every n belonging to this subsequence, and show that this leads to a contradiction. We will hereafter assume that $\{x_n^*\}$ is composed only of members of that subsequence, and not introduce new notation for it. Owing to its containment in a compact set, $\{x_n^*\}$ must have an accumulation point, call it x^* . Under the contrary assumption, since ∇f is continuous, it must be that $\|\nabla f(x^*)\| \geq \varepsilon$. On the other hand, $\nabla f_n(x)$ is consistent in quadratic mean everywhere in L_n , and for each n , $\nabla f_n(x_n^*) = 0$. Therefore, for arbitrary $\delta > 0$, there is some index N_δ such that for $i > N_\delta$, x_i^* lies within a distance δ of x^* , and $P\{\|\nabla f(x_i^*)\| < \delta\} > 1 - \delta$. Again, as ∇f is continuous, $P\{\|\nabla f(x^*)\| \leq \delta\} > 1 - \delta$. Since δ is arbitrary, then $P\{\|\nabla f(x^*)\| = 0\} = 1$. Thus, with probability 1 we obtain a contradiction, and the theorem is proved.

The mean update and weighted mean updates of Chapters III and V, respectively, both are gradient-based, steepest ascent type methods for which the conditions of Proposition 2.3.4 and Proposition 2.3.5 can be verified. Thus, the following theorem, which is the main result of this section, will be used to establish a quasi-consistency result for both update algorithms as estimators of local maxima of a probability density function. What keeps it from being a complete consistency result is that it does not eliminate the possibility of convergence to a saddlepoint. We believe it to be true that, asymptotically, convergence to a saddle-

point is an event of probability zero, and that in fact, the consistency of the update mode estimators need not be qualified. However, we do not have a rigorous proof of the full conjecture at this time.

Theorem 2.3.2. Let $f: \mathbb{R}^p \rightarrow \mathbb{R}$ be a uniformly continuous, continuously differentiable probability density function, and $\{x^{(1)}, \dots, x^{(n)}\}$ be a sample of n independent observations drawn from f . Let e_i be the i -th standard basis element in \mathbb{R}^p . Let K_0 be a kernel function satisfying (2.3.15) for estimating f , and suppose that for any m , $1 \leq m \leq p$, $\frac{\partial}{\partial x_m} K_0(x)$ is of the form $K'_0(x) = c_1 [K_1(x+we_m) - K_1(x-we_m)] - c_2 K_2(x)$ for some constant w , where also K_1 and K_2 satisfy (2.3.15) for estimating f , and for all $u \in \mathbb{R}^p$ and $t \in [0,1]$, $K_i(tu) \geq K_i(u)$, $i = 0,1,2$. Suppose that for some $\alpha > 0$ and s such that $1 > s > \max\{1/2, 2/p+2\}$, $h(n) = \alpha n^{(s-1)/p}$, and for some $\beta > 0$, $k(n) = \beta n h(n)^p = \beta n^s$. Let

$$\hat{f}_n(x) = \frac{1}{n r_k(x)^p} \sum_{i=1}^n K_0\left(\frac{x^{(i)} - x}{r_k(x)}\right)$$

and

$$\hat{\hat{f}}_n(x) = \frac{1}{n h(n)^p} \sum_{i=1}^n K_0\left(\frac{x^{(i)} - x}{h(n)}\right)$$

be the variable and fixed bandwidth kernel estimators of f , respectively. Let x^0 be a starting point, independent of n for which $f(x^0) > 0$, and let $\{x_n^1, x_n^2, \dots\}$ be the sequence of

iterates produced by an algorithm which guarantees that $g_n(x^{i+1}) > g_n(x^i)$ for all i , and that the x_n^i converge to a point x_n^* at which $\nabla g_n(x_n^*) = 0$, for either $g_n = f_n$ or $g_n = \hat{f}_n$. Then, with probability one, as $n \rightarrow \infty$, $\nabla f(x_n^*) \rightarrow 0$.

Proof. By Proposition 2.3.2 and the first part of the proof of Theorem 2.3.1, both f_n and \hat{f}_n are strongly uniformly consistent in the region $\{x \in \mathbb{R}^p : f(x) \geq f(x_i)/2\}$. The remainder of the proof follows immediately from Proposition 2.3.4 and Proposition 2.3.5.

III. THE MEAN UPDATE ALGORITHM

3.1. Description of the Mean Update Algorithm and Discussion

Our initial empirical investigations of the problem of locating modes in data of arbitrary dimension were conducted using a simple nearest neighbor procedure, which we refer to as the mean update. The mean update algorithm has a user-controlled parameter, k , or equivalently a specified fraction of the sample size, where k indicates the number of nearby sample points to include in each mean calculation. With k given and an initial guess, x_c , the mean update algorithm is:

Algorithm 3.1.1. Find the k observations in the sample, $\{x^{(1)}, \dots, x^{(k)}\}$ which are nearest to x_c ; these are called the neighbor set or k -neighbor set of x_c .

Update: $x_+ = \frac{1}{k} \sum_{i=1}^k x^{(i)}$; find the k -neighbor set of x_+ ;

If the neighbor set remains unchanged, then $x_+ = x_c$

and this is the mode estimate;

return;

else

$x_c = x_+$;

go to update;

end if.

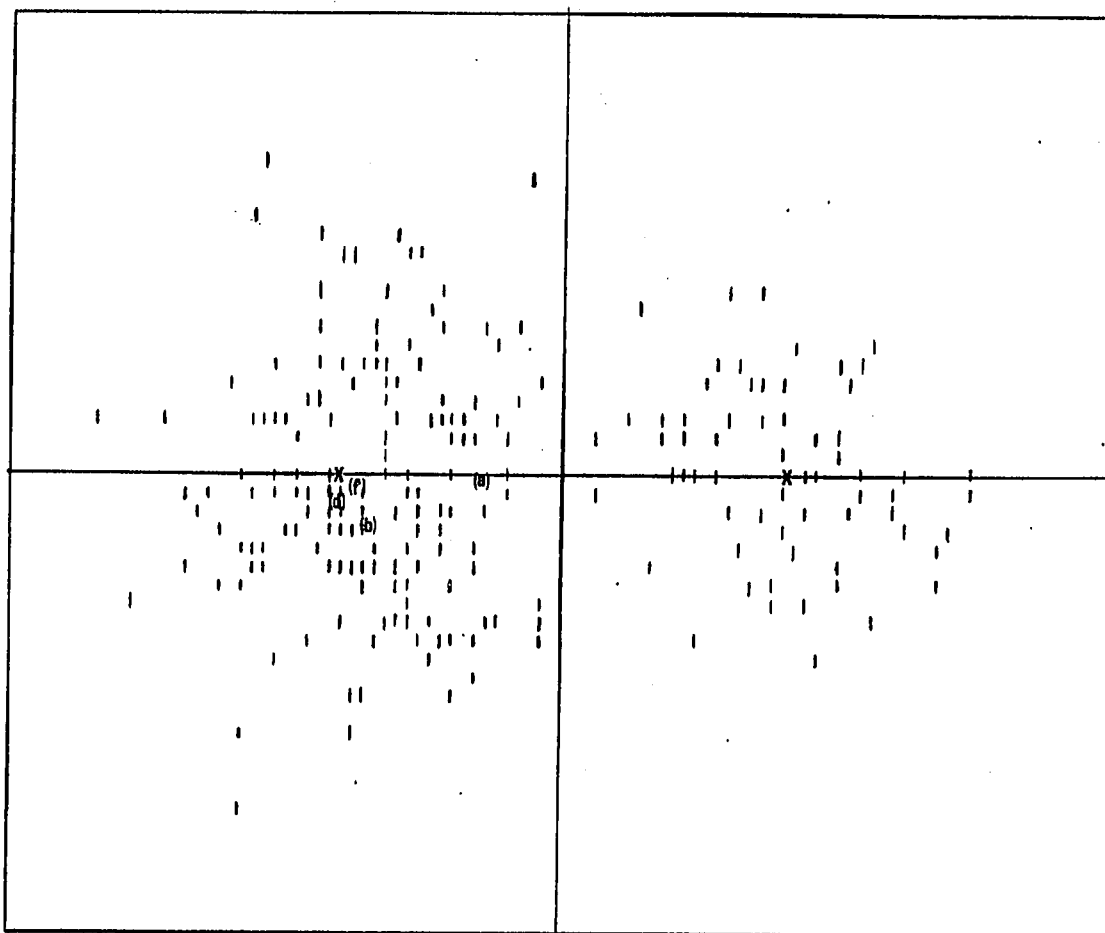
An example of the performance of the mean update is

given in Figure 3.1.1. The scatter plot is of a sample of 300 observations drawn from a two-component Gaussian mixture with means at -2.25 and $+2.25$, mixing proportions of $3/4$ and $1/4$, respectively, and unit covariance matrices. The starting point is the sample mean (location (a)), and initially $k = 50$. The mean update proceeds along the path indicated, and after 13 iterations stops at the location (b), at the periphery of a relatively tight cluster of points. With k decreased to 25, and, upon convergence with that value, to 12, the update gravitates to the center of the cluster, at location (d). With $k = 100$, the update essentially locates the sample mean of the dominant Gaussian component (location (f)). A complete trace of the sequence of iterates is given in Table 3.1.1. With each value of k indicated there, the mean update is tracked until it converges, and the location of each iterate, the radius of the associated neighbor set, and the Euclidean distance from the true norm are given.

The rationale for the mean update algorithm is that modes in the density function will be accompanied by denser clustering of the sample observations, that the local mean is an obvious measure of location for such a cluster, and that successive updates should be drawn by the local distribution of points in a direction of ascent of the density function.

In fact, the update step is exactly the Newton and

Figure 3.1.1. Partial Trace of the Mean Update on Two-Dimensional Gaussian Mixture



KEY

- X - local mode
- (a) - starting point for the update procedure
- (b), (d), (f) - stopping points with different parameter values

TABLE 3.1.1.

	<u>x-Coordinate</u>	<u>y-Coordinate</u>	<u>Radius of the k-Neighbor Set</u>	<u>L₂ Error</u>
k = 50 (a) :	-0.873	-0.0752	1.477	
	-1.210	-0.0907	.958	
	-1.381	-0.0716	.742	
	-1.452	-0.0681	.790	
	-1.529	-0.0619	.832	
	-1.653	-0.1075	.696	
	-1.744	-0.1176	.668	
	-1.828	-0.1558	.678	
	-1.897	-0.2466	.721	
	-1.971	-0.3496	.680	
	-2.017	-0.3764	.647	
	-2.037	-0.3918	.676	
	-2.046	-0.4079	.660	0.456
k = 25 (b) :	-2.046	-0.4079	.249	
	-2.100	-0.3926	.239	
	-2.165	-0.4089	.247	
	-2.223	-0.3868	.249	
	-2.289	-0.3524	.259	
	-2.320	-0.3717	.246	
	-2.346	-0.3522	.242	
	-2.378	-0.3561	.241	
(c) :	-2.401	-0.3386	.236	0.371
k = 12 (c) :	-2.401	-0.3386	.0763	
(d) :	-2.387	-0.3164	.0721	0.345

TABLE 3.1.1. Continued.

	<u>x-Coordinate</u>	<u>y-Coordinate</u>	Radius of the k-Neighbor Set	<u>L₂Error</u>
k = 50 (d) :	-2.387	-0.3164	.681	
	-2.352	-0.3056	.649	
	-2.328	-0.3105	.644	
	-2.302	-0.3293	.635	
(e) :	-2.270	-0.3281	.613	0.329
k = 100 (e) :	-2.270	-0.3281	1.677	
	-2.227	-0.2824	1.589	
	-2.173	-0.2464	1.563	
	-2.144	-0.2118	1.535	
(f) :	-2.120	-0.1877	1.472	0.228
k = 50 (f) :	-2.120	-0.1877	.686	
	-2.192	-0.2822	.698	
	-2.187	-0.3137	.666	0.331
(g) :	-2.191	-0.3255	.682	

gradient step of a variable kernel density estimate with quadratic kernel,

$$K(x) = c(1 - \frac{1}{2} \|x\|_2^2), \quad \|x\|_2 \leq 1,$$

and c is chosen so that K integrates to 1. Then, writing $r_k(x)$ for the radius of the k -neighbor set of x_c ,

$$f_n(x; k) = \frac{c'}{nr_k(x_c)^p} \sum_{i=1}^k [1 - \frac{1}{2} \frac{\|X^{(i)} - x\|^2}{r_k(x_c)^2}] , \quad (3.1.1)$$

$$\frac{\partial}{\partial x_m} f_n(x; k) = \frac{c'}{nr_k(x_c)^{p+2}} \sum_{i=1}^k (X_m^{(i)} - x_m) ,$$

and

$$\nabla^2 f_n(x; k) = - \frac{c'k}{nr_k(x_c)^{p+2}} I ,$$

where I is the identity matrix of appropriate dimension. Note that, if the smoothing parameter, once initialized as $h = r_k(x_c)$, is held fixed, then $\nabla^2 f_n(x; k)$ is negative definite and constant for all x . Again, regardless of the value of h , as long as the neighbor set is held fixed, the unique value satisfying $\nabla f_n(x; k) = 0$ is given by the mean of the neighbor set. Thus, the update $m_+ = (1/k) \sum_{i=1}^k X^{(i)}$ is the optimizer of the local quadratic model of the density function obtained from $f_n(x; k)$ and its derivatives with the neighbor set held fixed.

For use as a density estimator the kernel described above has some drawbacks. For one thing, the estimate it

yields is discontinuous at any point where the neighbor set changes. The form of the estimate over any region where the neighbor set is constant is a negative definite quadratic. The imposition of strict, locally uniform concavity in the estimate of the density is likely to introduce spurious local modes that might not arise with a kernel function, such as the Gaussian, whose concavity decreases with distance from the origin. As an example, consider the following one-dimensional sample and the estimate (3.1.1) drawn from it:

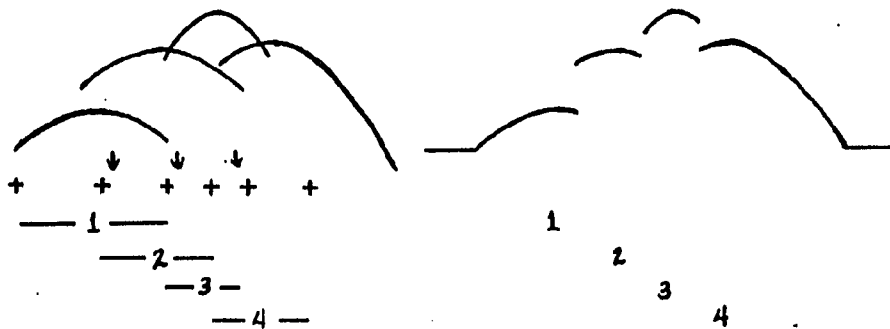


Figure 3.1.2.

The sample points are indicated by small crosses. Moving from left to right there are four successive neighbor sets of size three. These are indicated below the sample in sketch (a), with numerals marking the maximizers of the corresponding quadratic models. Vertical arrows mark the

locations where the neighbor set changes. In Figure (b) the resulting density estimate is pictured. The estimate retains the optima of each of its component quadratics, though overall, the estimated density has a unimodal character. Any optimum that remains is a potential stopping point for the mean update algorithm. Generally speaking, many of the component modes will be lost beneath overlapping quadratic segments, but the susceptibility to many technical modes remains.

Though the density estimate associated with the mean update immediately suggests a point of departure for improving the algorithm (i.e., replacing the quadratic kernel with one having variable concavity), it also allows us to observe properties of the mean update that guarantee its consistency. We would like to demonstrate that each step taken by the mean update increases the value of the associated variable kernel density estimate, since then Proposition 2.3.5 would apply immediately, but we have been unable to prove or disprove this assertion. However, it is possible to associate the mean update with a modified version of (3.1.1) for which each update step does yield an increased function value.

Lemma 3.1.1. Let $\{x_m\}$ be the sequence of iterates generated by the mean update procedure, and let $\{r_m\}$ be the associated k -neighbor distances. Let $\tilde{f}(x;h)$ be analogous to the kernel estimator (3.1.1) with smoothing parameter h replacing $r_k(x)$,

that is,

$$\tilde{f}(x;h) = \frac{\alpha}{nh^p} \sum_{i=1}^n \left[1 - \frac{1}{2} \frac{\|X^{(i)} - x\|^2}{h^2} \right],$$

where $n(x) = \{X^{(1)}, \dots, X^{(k)}\}$ in the k -neighbor set of x and

$$r = \max_{1 \leq i \leq k} \{ \|X^{(i)} - x\| \}.$$

Then

- (i) the mean update procedure terminates after a finite number of steps, and
- (ii) for any h , and for all m , $\tilde{f}(x_{m+1};h) > \tilde{f}(x_m;h)$.

Proof. Let $n(x_m) = \{X_m^{(1)}, \dots, X_m^{(k)}\}$ be the k -neighbor set of x_m . For arbitrary h let $\tilde{f}_{(m)}(\cdot;h)$ be the fixed bandwidth sample function based upon $n(x_m)$. That is,

$$\tilde{f}_{(m)}(x;h) = \frac{\alpha}{nh^p} \sum_{i=1}^k \left[1 - \frac{1}{2} \frac{\|X_m^{(i)} - x\|^2}{h^2} \right], \quad (3.1.2)$$

thus $\tilde{f}(x_m;h) = \tilde{f}_{(m)}(x_m;h)$ (and $f_n(x_m;k) = \tilde{f}(x_m;r_m)$).

Define

$$D_m(x) = \sum_{i=1}^k \|X_m^{(i)} - x\|^2.$$

Clearly for any two points z_1 and z_2 , $\tilde{f}_{(m)}(z_1;h) < \tilde{f}_{(m)}(z_2;h)$ if and only if $D_m(z_1) > D_m(z_2)$; also for any two iterates x_i and x_j , $\tilde{f}(x_i;h) < \tilde{f}(x_j;h)$ if and only if $D_i(x_i) > D_j(x_j)$. Therefore $\tilde{f}_{(m)}(z_1;h) < \tilde{f}_{(m)}(z_2;h)$ if and only if $\tilde{f}_{(m)}(z_1;h') < \tilde{f}_{(m)}(z_2;h')$ for any $h' > 0$. Similarly

$\tilde{f}(x_i;h) < \tilde{f}(x_j;h)$ if and only if $\tilde{f}(x_i;h') < \tilde{f}(x_j;h')$ for every $h' > 0$. Since at each iteration x_{m+1} is chosen to maximize $\tilde{f}_{(m)}(x;h)$.

$$D_m(x_{m+1}) \leq D_m(x_m), \quad (3.1.3)$$

with equality only if $x_m = x_{m+1}$. Now the neighbor set $n(x_{m+1})$ can be arranged and written in the form,

$$n(x_{m+1}) = \{x_m^{(1)}, \dots, x_m^{(k-s)}, x_{m+1}^{(k-s+1)}, \dots, x_{m+1}^{(k)}\},$$

where $\{x_m^{(1)}, \dots, x_m^{(k-s)}\} \subset n(x_m)$ and $\{x_{m+1}^{(k-s+1)}, \dots, x_{m+1}^{(k)}\} \subset n(x_m)^c$. Then

$$\begin{aligned} D_{m+1}(x_{m+1}) - D_m(x_{m+1}) &= \sum_{j=1}^s [\|x_{m+1}^{(k-s+j)} - x_{m+1}\|^2 \\ &\quad - \|x_m^{(k-s+j)} - x_{m+1}\|^2] \quad (3.1.4) \\ &\leq 0. \end{aligned}$$

Combining (3.1.3) and (3.1.4) yields

$$D_{m+1}(x_{m+1}) < D_m(x_m), \quad (3.1.5)$$

and accordingly,

$$\tilde{f}(x_{m+1};h) > \tilde{f}(x_m;h) \quad (3.1.6)$$

for any h , unless $x_{m+1} = x_m$, in which case, the mean update terminates with that value. Thus, the second statement of the lemma is proved. To prove the first statement, we recognize that the number of possible neighbor sets, and

hence the set of possible iterates x_m is finite, and the strict monotonicity of (3.1.6) insures that none of these iterates may be visited twice without causing termination. \square

The monotonicity of estimated function values $\{f_n(x_n^1), f_n(x_n^2), \dots\}$ generated by an iterative mode-finding procedure is required in Proposition 2.3.5 only as a means of assuring that the set of mode estimates obtained from a particular starting point belongs to a compact set contained in the domain of positivity of f . This compact containment can be established under the slightly weaker conditions we have just established.

Lemma 3.1.2. Let $x_1 \in \mathbb{R}^p$ be a starting point independent of n , and let $k(n) = O(n^s)$, $1 > s > \max\{1/2, 2/p+2\}$. Define $L_\eta = \{x: f(x) \geq \eta\}$. Let $\{x_1 = x_n^1, x_n^2, x_n^3, \dots\}$ be the sequence of iterates generated by the mean update process with sample of size n . Then for any η , $0 < \eta < f(x_1)$, there exists $N > 0$ such that $f(x_n^j) \in L_{\eta/2^{(p/2+1)}}$ for all j and for all $n > N$.

Proof. By Theorem 2.3.1, for any $\eta > 0$, the variable kernel estimator, $f_n(x; k)$ is (uniformly strongly) consistent over L_η . By the work of Van Ryzin [1969] and Moore and Yackel [1977; cf. Proposition 2.3.3], the fixed bandwidth estimator $\hat{f}_n(x; ch)$ is also uniformly strongly consistent everywhere, taking c as any positive constant and $h = r_{[k(n):n]}(x_1) = r_1$. Referring to the definitions of Lemma 3.1.1 and Theorem

2.3.2, observe that $\tilde{f}(x_1; r_1) = f_n(x_1; k) = \hat{f}_n(x_1, r_1)$, assuming of course the quadratic kernel (3.1.1). By the consistency of $f_n(x; k)$, if $\eta < f(x_1)$, then for n large enough, $f_n(x_1; k) > \eta + \delta$, for some $\delta > 0$.

From the preceding lemma, we have that $f_n(x_1; k) < \tilde{f}(x_n^m; r_1)$, where x_n^m is the m -th iterate of the mean update based upon a sample of size n . In what follows, $x_m^{(1)}, \dots, x_m^{(k)}$ are the nearest neighbors of x_n^m , listed in ascending order of distance $\|x_m^{(i)} - x_n^m\|$. Note that if $\|x_m^{(i)} - x_n^m\|^2 > 2h^2$, then $x_m^{(i)}$ makes a net negative contribution to the sample function $\tilde{f}(x_n^m; h)$. Let k' be the number of neighbors for which $\|x_m^{(i)} - x_n^m\|^2 < 2r_1^2$. Then,

$$\begin{aligned}
 f_n(x_1; k) &< \tilde{f}(x_n^m; r_1) \\
 &= \frac{\alpha}{nr_1^p} \sum_{i=1}^k \left[1 - \frac{1}{2} \frac{\|x_m^{(i)} - x_n^m\|^2}{r_1^2} \right] \\
 &< \frac{\alpha}{nr_1^p} \sum_{i=1}^{k'} \left[1 - \frac{1}{2} \frac{\|x_m^{(i)} - x_n^m\|^2}{r_1^2} \right] \quad (3.1.7) \\
 &< 2^{p/2} \frac{\alpha}{n(\sqrt{2} r_1)^p} \sum_{i=1}^{k'} \left[1 - \frac{1}{2} \frac{\|x_m^{(i)} - x_n^m\|^2}{2r_1^2} \right] \\
 &\leq 2^{p/2+1} \hat{f}_n(x_n^m; \sqrt{2} r_1).
 \end{aligned}$$

Therefore, for n large enough, $\hat{f}_n(x_n^m; \sqrt{2} r_1) > 2^{-p/2}(\eta + \delta)$, a constant bound which does not depend upon n or m . Since $\hat{f}_n(x_n^m; \sqrt{2} r_1)$ is consistent, again for n large enough, and for all m , $f(x_n^m) > 2^{-(p/2+1)}\eta$, and the lemma is proved.

We note in passing that the conclusions of this lemma follow directly from Lemma 3.1.1 without (3.1.7) and without the factor $2^{-p/2}$ if the bandwidth in the mean update procedure is held fixed. For a fixed bandwidth procedure, consistency can be established even if, say, an extra dimension is added with every new observation. It is likely that tighter analysis can remove the dimensional dependency of the constant $2^{-p/2}$, and thus eliminate the role of dimensionality from the consistency arguments and the design of the variable kernel mean update; however, such a goal is of secondary importance at this stage of our inquiry.

Lemma 3.1.2 allows us access to the consistency results of Section 2.3. An additional requirement for those results is that for arbitrary coordinate direction m , the partial derivative of the quadratic kernel (3.1.1) may be expressed in the form

$$\frac{\partial}{\partial x_m} K(x) = c_1 [K_1(x + we_m) - K_1(x - we_m)] - c_2 K_2(x), \quad (3.1.9)$$

where w , c_1 , and c_2 are real constants, e_m is the unit m -th coordinate vector, and K_1 and K_2 are probability density functions satisfying the moment conditions (2.3.15) for estimating f . For the quadratic kernel,

$$\frac{\partial}{\partial x_m} K(x) = \begin{cases} -x_m, & \text{if } \|x\|_2 \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

In the univariate case, the decomposition is simple and natural; we take

$$K_1(\tau) = 1, \quad \text{for } |\tau| \leq 1/2,$$

$$K_2(\tau) = 1 - |\tau|, \quad \text{for } |\tau| \leq 1,$$

and $c_1 = c_2 = 1$ and $w = -1$. In the multivariate case, the decomposition is complicated by the spherical support of K and its derivatives. Figure 3.1.3 depicts $\frac{\partial}{\partial x_m} K$ for bivariate random variables, as well as the component K_1 of the decomposition. To generalize the dimension p , we take $C = \{(x_1, \dots, x_p); x_m \geq 0 \text{ and } \|x\|_2 \leq 1\}$, let u be a random variable distributed uniformly over C , μ be the mean of u , V be the volume of C , and

$$K_1(\tau) = \begin{cases} 1/V, & \text{if } \tau - \mu \in C \\ 0, & \text{otherwise,} \end{cases}$$

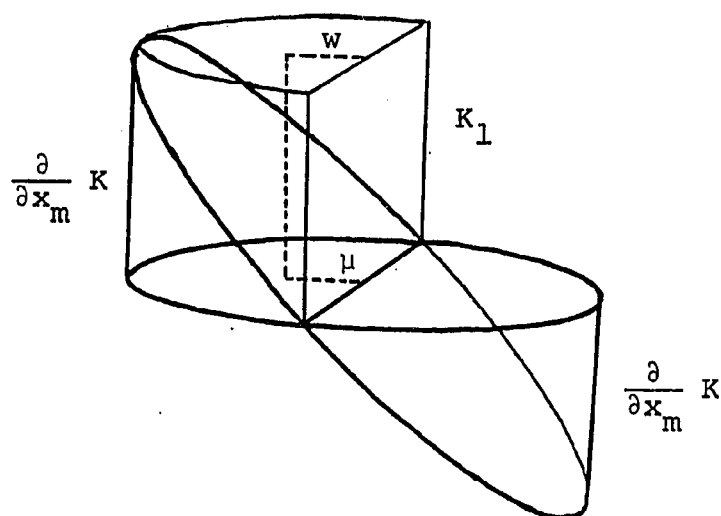
$$K_2(\tau) = \begin{cases} 1/V (1 - |\tau|), & \text{if } \|\tau\|_2 \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\frac{\partial}{\partial x_m} K(x) = K_1(x - \mu) - K_1(x + \mu) - K_2(x).$$

Because it is not symmetric with respect to the m -th coordinate, K_1 is not an appealing kernel function; nevertheless, it meets the criteria of Theorem 2.3.2, and thus a consistency result for the mean update follows immediately.

Figure 3.1.3. Partial Derivative of the Quadratic Kernel Corresponding to the Mean Update



Theorem 3.1.1. Let $f: \mathbb{R}^p \rightarrow \mathbb{R}^1$ be a uniformly continuous, continuously differentiable probability density function, and $\{x^{(1)}, \dots, x^{(n)}\}$ be a sample of n independent observations drawn from f . Suppose that for some $\beta > 0$ and some s satisfying $1 > s > \max\{1/2, 2/p+2\}$, $k(n) = \beta n^s$. Then the mean update, Algorithm 3.1.1, will converge to a mode estimate x_n^* in a finite number of steps, and with probability one, $\nabla f(x_n^*) \rightarrow 0$ as $n \rightarrow \infty$.

3.2. Structure and Implementation of the Monte Carlo Trials

The trace of the mean update in Table 3.1.1 gives limited indication of the character of the algorithm and the significance of its control parameter k . In the example nothing unexpected or extreme occurs; accuracy increases with the size of the neighbor set, while the drop to $k = 12$, or 4% of the sample size, has no serious consequences. However, had the algorithm started at the sample mean with $k = 12$, it would have halted very quickly, and far from either mode. In addition, with slightly less separation between the modes, even a neighbor set as low as $k = 100$, or 33% of the sample, could conceivably draw enough from both components that it would remain poised between the modes, behaving essentially as does the sample mean.

Mode estimation has two distinct but complementary uses. One is an estimation procedure for a single location parameter that will be resistant to data contamination. Under the assumption that, with minor surgery involving the removal of outlying points in a sample, the distribution

being observed can be "brought to symmetry," it seems desirable to have neighbor sets that are large, perhaps 80% or 90% of the sample size. Frequently in high dimensional settings that assumption is unfounded. The second use of mode estimation, and the one which motivated this study, is for investigating the presence of secondary modes. For this purpose, it is necessary for the estimation algorithm to be able to concentrate locally in regions near individual modes, and this implies that neighbor sets should be confined to a small fraction of the sample.

It is important to know, and difficult to predict analytically, over what range of values of k the mean update will be effective. We need to know how performance characteristics change with varying dimension and sample size, and how they are affected by the shape of the population density. The same issues apply to the quasi-Newton and weighted mean algorithms of Chapters IV and V. To get empirical answers to these questions an extensive battery of Monte Carlo tests was conducted. Each test requires the simulation of a particular family of multivariate probability distributions. Holding the sample size constant, the mode-seeking algorithm (or algorithms) being examined are applied to data of varying dimensions, and on each trial are run using several values of the control parameter k . The structure of the tests remained basically unchanged throughout the course of this study, as given here:

Algorithm 3.2.1

```

for I = 1 to #TRIALS
  for dimension P = 1,2,3,4,5,10,15,20(,50,100)

    Generate a sample of N independent observa-
    tions of a p-dimensional random vector
    simulated according to a specified prob-
    ability law.

    for a sequence of neighbor set sizes k,
      (either k = .1N, .2N, .3N, .4N
        or k = .75N, .90N, .95N, .99N)
      for each of the mode-seeking procedures
        in the test
        start at  $x_{init} = (-100.0, -100.0, \dots, -100.0)^T$ 
        run the current mode-seeking procedure
          with parameter k, and
          accumulate descriptive performance
            statistics

      end for
    end for
  end for
end for

```

Early testing of the mean update included sample spaces of dimension 50 and 100. Though the results were favorable enough to justify further exploration in such high dimen-

sions, it was decided, for reasons of economy and interest, to limit the dimensionality for further testing to a maximum of 20. In all tests, the number of trials was 25, thus all performance statistics quoted in this study will be based upon 25 independent executions of the algorithm under a fixed set of circumstances.

Measurement of Performance

A number of statistics were collected from each test to measure the accuracy of the mode-seeking procedures. Definitions of the five measures that figure in this report are given below along with acronyms by which we will often refer to them. In the definitions, it is assumed that the sample observations are p -dimensional, the true mode is x^* , and the mode estimate \hat{x} :

$$MSE = \frac{1}{p} \|\hat{x} - x^*\|_2^2 = \frac{1}{p} \sum_{i=1}^p (\hat{x}_i - x_i^*)^2;$$

$$SUP = \|\hat{x} - x^*\|_\infty = \max_i \{|\hat{x}_i - x_i^*|\};$$

$$MAD = \frac{1}{p} \|\hat{x} - x^*\|_1 = \frac{1}{p} \sum_{i=1}^p |\hat{x}_i - x_i^*|;$$

$$MED = \text{median} \{(\hat{x}_i - x_i^*), i = 1, \dots, p\};$$

$$L2 = p * MSE.$$

The MSE, of course, is the common quadratic loss, averaged over the variates to expedite the comparison of

results between different dimensional spaces. As its average over a series of trials incorporates errors arising from both variability and systematic bias, it is the most complete measure of accuracy, and will be the one used most often in this study to report and compare performance. Unaveraged quadratic loss, L_2 , will often be used with skew data because of the particular way in which that data was generated. The SUP measure is included out of concern for extreme errors that might occur, especially with small neighbor sets. MAD is used on occasion when the range of values of the MSE over different dimensions or between different procedures make them difficult to display on a linear scale. The MED measure has significance only due to the geometry of the test configuration. Unimodal data sets were generated so that the mode of the simulated density was located at the origin, and the mode-seeking procedures were started from $(-100.0, \dots, -100.0)$. Thus, the predominance and magnitude of negative values in the mode estimate gave evidence of the extent to which errors in the estimate stemmed from failure of the algorithm to complete its ascent operation rather than from sampling variability.

Distributions Used in the Tests

Tests of the mode-seeking algorithm were conducted with data sets simulated from several distributions. Testing began with uncorrelated Gaussian data. Subsequent data sets were chosen to expose the procedures to distributional

patterns that threatened to be problematic, and which we expected to occur frequently in practice. First among these is the possibility that the data lie in a manifold of smaller dimension than that in which the observations were recorded. To produce such sub-dimensionality, a highly correlated Gaussian distribution was used, having all variances equal to unity, and all covariances equal to 0.9. The determinants of the resulting covariance matrices in dimensions 5, 10, 15, and 20, are 0.00046 , 9.1×10^{-9} , 1.36×10^{-13} , and 1.81×10^{-18} , respectively, indicating that in the higher dimensions, the generated data will be nearly univariate in character.

A second distributional characteristic which was studied is erratic tail behavior or heavy-tailedness. The study was conducted primarily by the generation of multivariate-Cauchy data, uncorrelated, and centered at the origin.

The third shape characteristic we investigated is asymmetry about the mode or skewness. The generation of multivariate skew distributions is rather a difficult problem and little work has been done in the area [Kennedy and Gentle, 1980]. The situation is further complicated by the desire to have the severity of the skewness, or shape of the distribution, especially in the vicinity of the mode, be comparable across dimensions. A strategy was developed for generating two-component Gaussian mixtures which are

comparable from the standpoint of parametric classification.

The justification for the strategy is rather lengthy and hence is deferred to an appendix. The main idea is that for a two-component mixture density,

$$p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x),$$

where

$$p_i(x) = (2\pi)^{-n/2} [\Sigma_i]^{-1/2} \exp\left\{-\frac{1}{2} (x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\right\},$$

the modality of the mixture is completely summarized by the modality of the one-dimensional slice passing through the two mean vectors, so long as Σ_1 and Σ_2 share the eigenvector $\mu_1 - \mu_2$. It is then possible to generate mixtures whose two components are "equivalently separated" in arbitrary dimension by constraining the conditional density along the slice between the means to match a one-dimensional prototype. In the appendix, it is shown that the prototype matching technique produces mixtures which yield dimensionally constant error rates for an optimal Bayes classifier, or linear discriminant function.

The prototype we used for investigating algorithm performance in the presence of skewness was

$$p(x) = \frac{1}{\sqrt{2\pi}} \{0.3e^{(-1/2)(x+1.25)^2} + 0.7e^{(-1/2)(x-1.25)^2}\}, \quad (3.2.1)$$

but shifted to the left by an amount 1.1974 to place the mode at the origin. The p-dimensional density matched to

it was

$$p(x) = (2\pi)^{-p/2} s^{-p} \{ 0.3 e^{-\frac{1}{2s^2} \left\| x + \frac{\mu + 1.1974s}{\sqrt{p}} \underline{1} \right\|^2} + 0.7 e^{-\frac{1}{2s^2} \left\| x - \frac{\mu - 1.1974s}{\sqrt{p}} \underline{1} \right\|^2} \},$$

with $s = 1.2$, $\mu = 1.5$, and $\underline{1}$ the p -dimensional vector composed of all ones, $\underline{1} = (1, 1, \dots, 1)^T$. A graph of the prototype density is given in Figure 3.2.1, and Figure 3.2.2 contains a plot of 100 observations drawn from the bivariate mixture (3.2.2). In the scatterplot, an "X" marks the origin and the location of the single mode. Scatterplots of the first two components of equivalent three- and fifteen-dimensional samples are given in Figure 3.2.3.

As the "principal axis" of the skewness in the mixture density is aligned directly with the starting point in the tests, $(-100.0, \dots, -100.0)^T$, the iterative algorithms must contend with the "shoulder" created by the subordinate component, which is on the verge of generating a local optimum. Thus, the two-component mixture provides a slightly more stringent test of the hill-climbing abilities of the algorithms than either of the symmetric distributions.

We also wanted to focus on the behavior of the algorithms on skewed data in the vicinity of a mode, the

Figure 3.2.1. Unimodal Mixture G2SKEW, Univariate Prototype

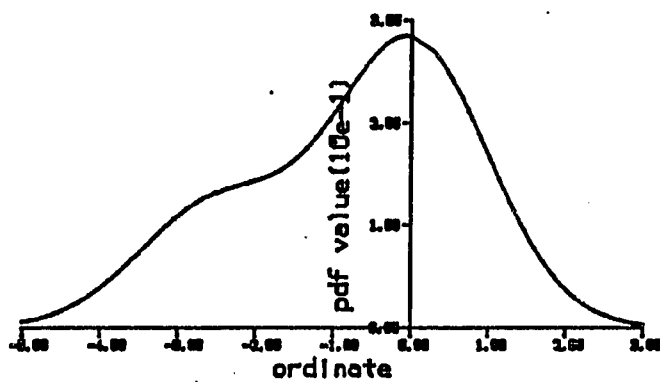


Figure 3.2.2. Scatter Plot of G2SKEW

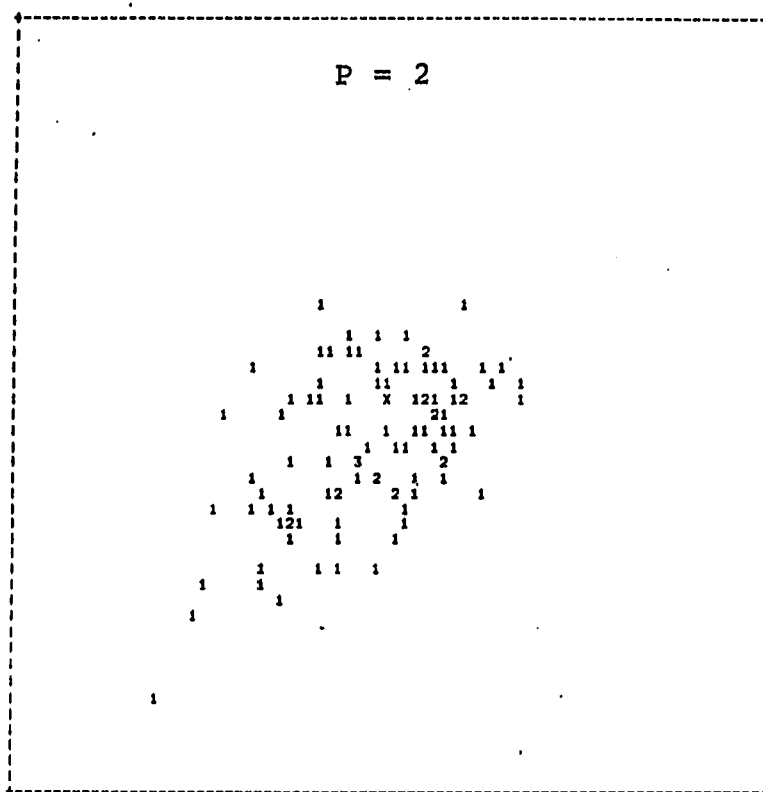
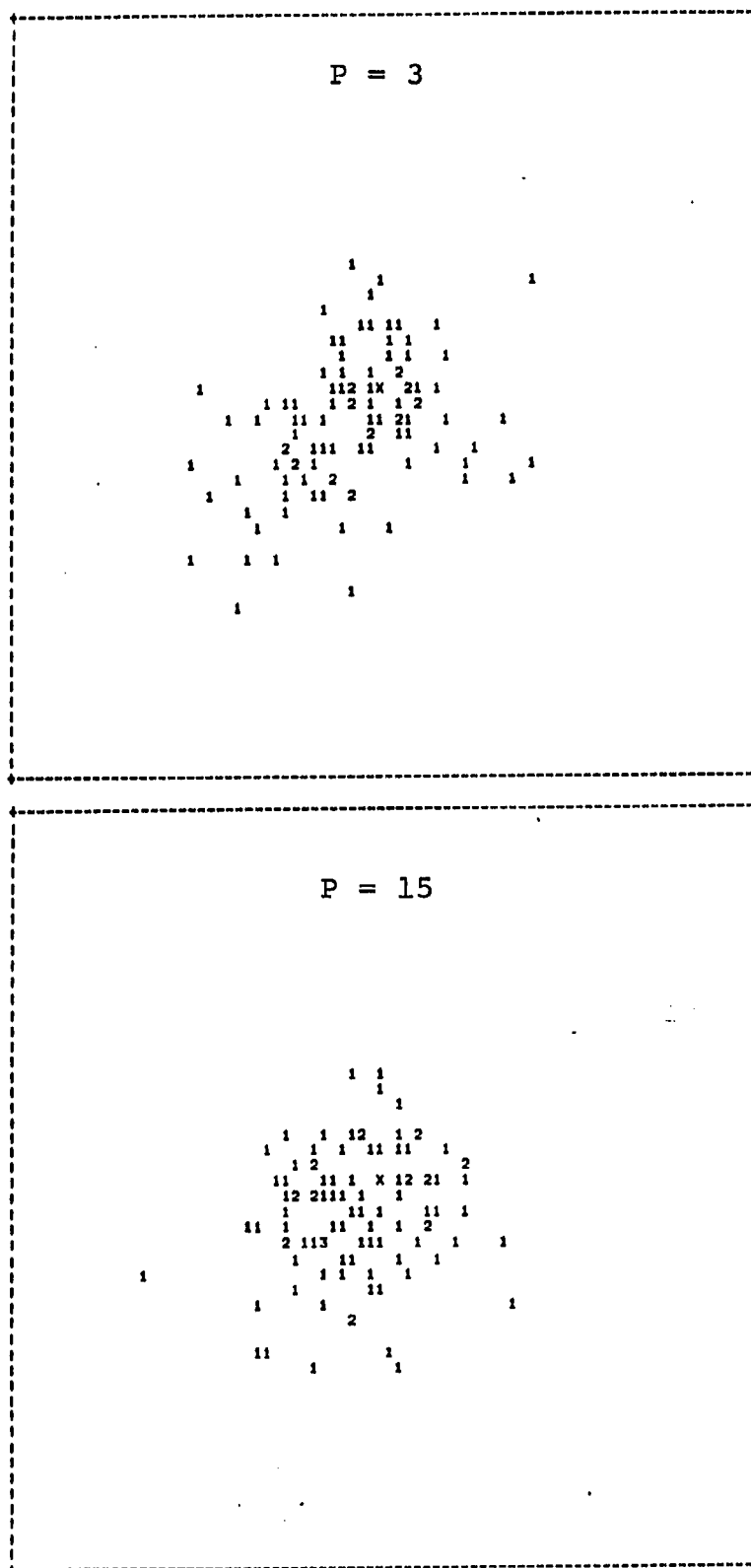


Figure 3.2.3. Scatter Plots of G2SKEW



main purpose being to assess the rate at which a biasing effect grows with increase in the size of the neighbor set. To effect this, the previous "least favorable" orientation was replaced with the "most favorable," by reversing the weighting of the two mixture components, or simply replacing x by $-x$ in (3.2.1) and (3.2.2).

Note that the factor \sqrt{p} in (3.3.2) implies that the Euclidean distance between the mean vectors is constant regardless of dimension. The difference in individual coordinates of the component means, however, decreases as $p^{-1/2}$, and coordinate average measures such as MAD and MSE reflect this. Therefore, accuracy on the skew data is typically reported in terms of the Euclidean metric, L_2 , which in this case is dimensionally stable.

The coalescence of the coordinates of the two mean vectors also indicates that the nature of the mixture densities does change with dimension. The existence of two component means introduces with the line passing through them a particular one-dimensional orientation in p -dimensional space. The "equivalent separation" of the component mixtures produced by the method of prototype matching is defined in terms of this one-dimensional projection. In high dimensional settings this is a very special orientation, and the separation of the mixture components will not be apparent from most vantage points. Though as Figures 3.2.2 and 3.2.3a suggest, the prototype matching technique

is reasonably effective in producing the desired skew effect in low and moderate dimensions, in high dimension most lower dimensional conditional densities will involve a mixture of two nearly identical Gaussians, and most views of the data (e.g., Figure 3.2.3b) will present a nearly spherical cloud. Nevertheless, the mixture generated by the prototype matching technique in either low or high dimensions present a mode-seeking algorithm with a stringent test. In low dimensions, the algorithm, working on G2SKEWLT, must bypass the near mode at a Euclidean distance of 2.94 units from the true mode. The skewing effect of the subordinate mixture component declines with increasing dimension, but as it does the mixture exhibits a departure from normality which is correspondingly more difficult for a nonparametric procedure to detect.

Finally, for generating bimodal distributions, Gaussian mixtures akin to the skew distributions were used, but with the component-wise separation held constant independent of dimension. This procedure can be justified as holding stable across dimension the expected value of the ratio SSB/SSW , commonly encountered in the analysis of variance in clustering procedures, where SSB is the factor or between group sum of squares, $SSB = \|\bar{x}_1 - \bar{x}_2\|^2 / (1/n_1 + 1/n_2)$, and SSW is the pooled within group sum of squares.

To recapitulate, a summary of the distributions generated in this study is given in Table 3.2.1.

TABLE 3.2.1. Test Data Sets Used in the Study.

Label	Functional Description	Key Features
GAUSS	$\phi(x; O, I) *$	well-conditioned test set
R9GAUSS	$\phi(x; O, \Sigma),$ $\Sigma_{ij} = \begin{cases} 1, & i = j \\ 0.9, & i \neq j \end{cases}$	correlation; subdimensionality
CAUCHY	$\frac{\Gamma(\frac{p+1}{2})}{\pi^{\frac{p+1}{2}}} \frac{1}{(1 + \ x\ _2^2)^{p+1/2}}$	heavy tails; gross error contamination
G2SKEWLT	$0.3\phi(x; \frac{\mu_1}{\sqrt{p}} \underline{1}, s^2 I) + 0.7\phi(x; \frac{\mu_2}{\sqrt{p}} \underline{1}, s^2 I),$	asymmetry, hill-climbing, heavy tails
where	$\mu_i = (-1)^i \mu - 1.1974s,$ $\mu = 1.25, s = 1.0,$ $\underline{1} = (1, \dots, 1)^T$	
G2SKEWRT	$0.3\phi(-x; \frac{\mu_1}{\sqrt{p}} \underline{1}, s^2 I) + 0.7\phi(-x; \frac{\mu_2}{\sqrt{p}} \underline{1}, s^2 I)$ $\mu_i = (-1)^i \mu + 1.1974s$	asymmetry in the vicinity of the mode; bias
G2SEPRT	$0.3\phi(x; O; I) + 0.7\phi(x; 2.5 \underline{1}; I)$	multimodality

* $\phi(x; \mu, \Sigma)$ is the multivariate Gaussian density with mean μ and variance-covariance matrix Σ .

Generation of Random Numbers

The simulation of variates from a given probability distribution usually begins with the simulation of a uniform random variable in the interval $(0,1)$. The most used of the uniform random number generators, the multiplicative congruential generator, has the disadvantage that successive p -tuples lie in sparse disjoint affine subsets of the unit hypercube, and can have very poor distributional patterns in high dimensions. A second class of random number generators, first suggested by Tausworthe [1965], draws integers from the string of bits generated by a feedback shift register (FSR) based upon a primitive polynomial over the field $GF(2)$. If the precision of the integers generated is L bits, the degree of the polynomial is r , L is relatively prime to $2^r - 1$, and $pL < r$, then Tausworthe showed that the FSR procedure achieves p -dimensional uniformity [Kennedy and Gentle, 1980, p. 150 ff]. We employed a generalized version of the FSR technique, due to Lewis and Payne [1973], with modifications to allow a variable seed value to initiate the extended shift register it maintains. If the degree of the generating polynomial is the exponent of a Mersenne prime (e.g., 89, 127, 521, 607), p -dimensional uniformity is guaranteed so long as $pL < r$. We chose $r = 127$, and generated 31 bit unsigned integers. Thus spatial uniformity (with 31-bit precision) was guaranteed up through dimension four, and uniformity of at least the high order

6 bits (and thus of a lattice of at least 64^D mesh points) was guaranteed throughout the dimensional range of the tests.

Gaussian random variables were simulated via the Box-Muller transformation of two independent variates U_1 and U_2 , distributed uniformly over the unit interval, defined

$$X_1 = [-2 \ln(U_1)]^{1/2} \cos(2\pi U_2)$$

$$X_2 = [-2 \ln(U_1)]^{1/2} \sin(2\pi U_2),$$

yielding the two independent $N(0,1)$ variates X_1 and X_2 . Cauchy data was taken from the ratio X_1/X_2 , safeguarded to avoid overflow.

3.3. Presentation of Results with the Mean Update Algorithm

To investigate the feasibility of the mean update algorithm Monte Carlo tests were first conducted using uncorrelated Gaussian data, sample spaces ranging from dimension 1 to 100, sample sizes of 100, 500, and 1000, and neighbor sets composed of 10, 20, 30, and 40 percent of the sample. Plots of average component squared error (MSE) and maximum component error (SUP), averaged across 25 trials, are given in Figures 3.3.1 and 3.3.2, for sample size $N = 100$; in Figures 3.3.3 and 3.3.4, for $N = 500$; and in Figures 3.3.5 and 3.3.6, for $N = 1000$. For a description of the test procedures and the conventions used in reporting test

results the reader should consult Section 3.2.

The study of the mean update was undertaken out of the conviction that modes of a density function are inherently as identifiable in high dimensions as low, more so if the distribution includes several independent components in the vicinity of the mode. The MSE and SUP plots in Figures 3.3.1 through 3.3.6 present clear evidence in support of this claim. The pronounced features of the plots are the rapid improvement of the mean update algorithm with dimension increasing above one and two, especially for the smaller neighbor set fractions, and the retention of stability in very high dimensions. The regularity of the plots in the high dimensional range suggests that the results in dimension twenty and beyond essentially characterize infinite-dimensional behavior, and that there is no intrinsic dimensional bound upon the application of the mean update procedure.

Limitations which do appear concern applications in dimension one or two, of little concern to us, and more significantly, the use of small neighbor sets. The impact of too small neighbor set selection is indicated in Figure 3.3.7, which contains histograms of the 25 MSE values obtained from 100 observations in dimension 1, 3, and 10. The left hand graph gives observed MSE using a neighbor set of size $K = 10 = .1N$; the right hand graph corresponds to $K = 40$. Vertical and horizontal scales are comparable in

Figure 3.3.1. GAUSS Distribution, $N = 100$, Average of 25 Trials

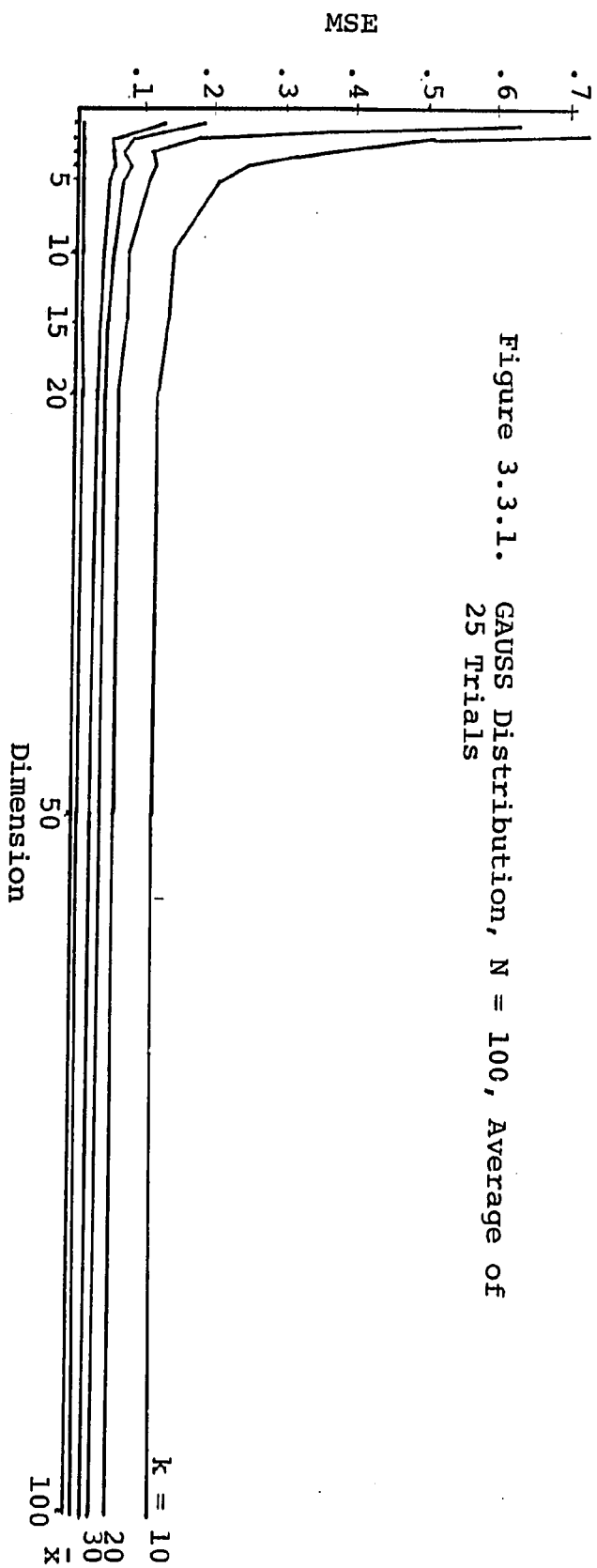


Figure 3.3.2.

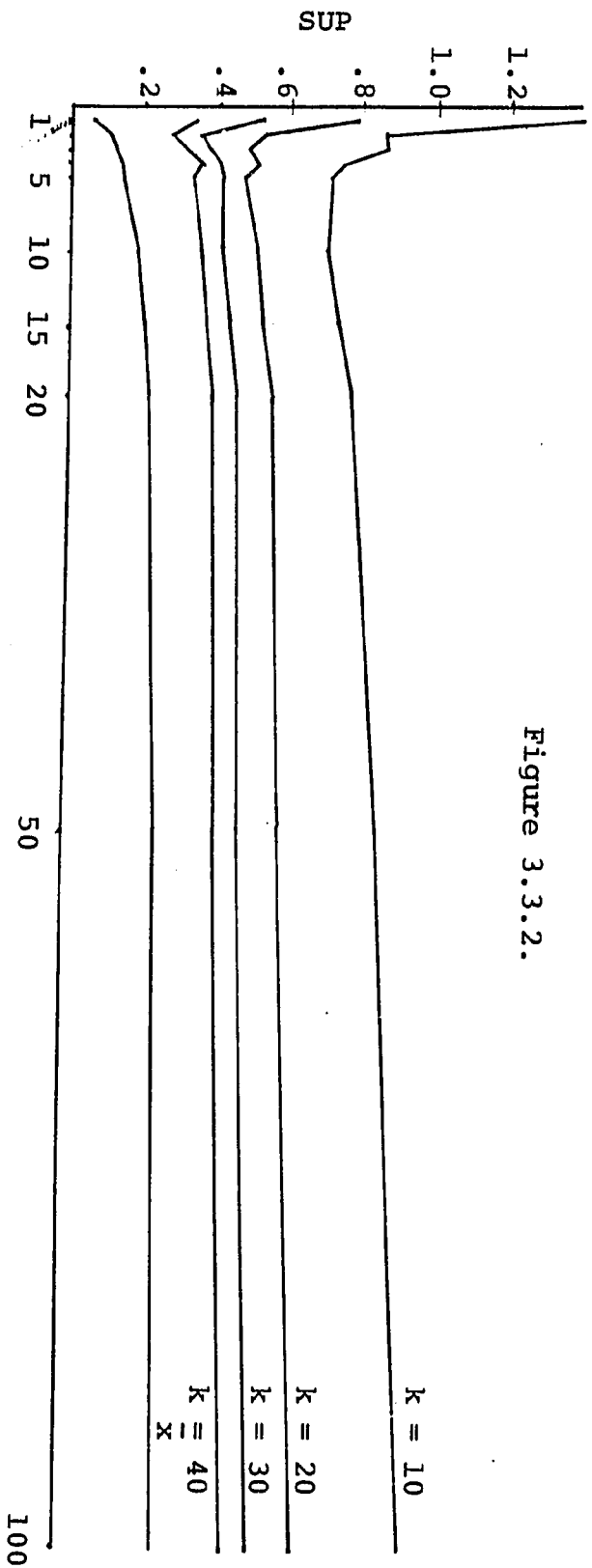


Figure 3.3.3. GAUSS Distribution, $N = 500$, Average of 25 Trials

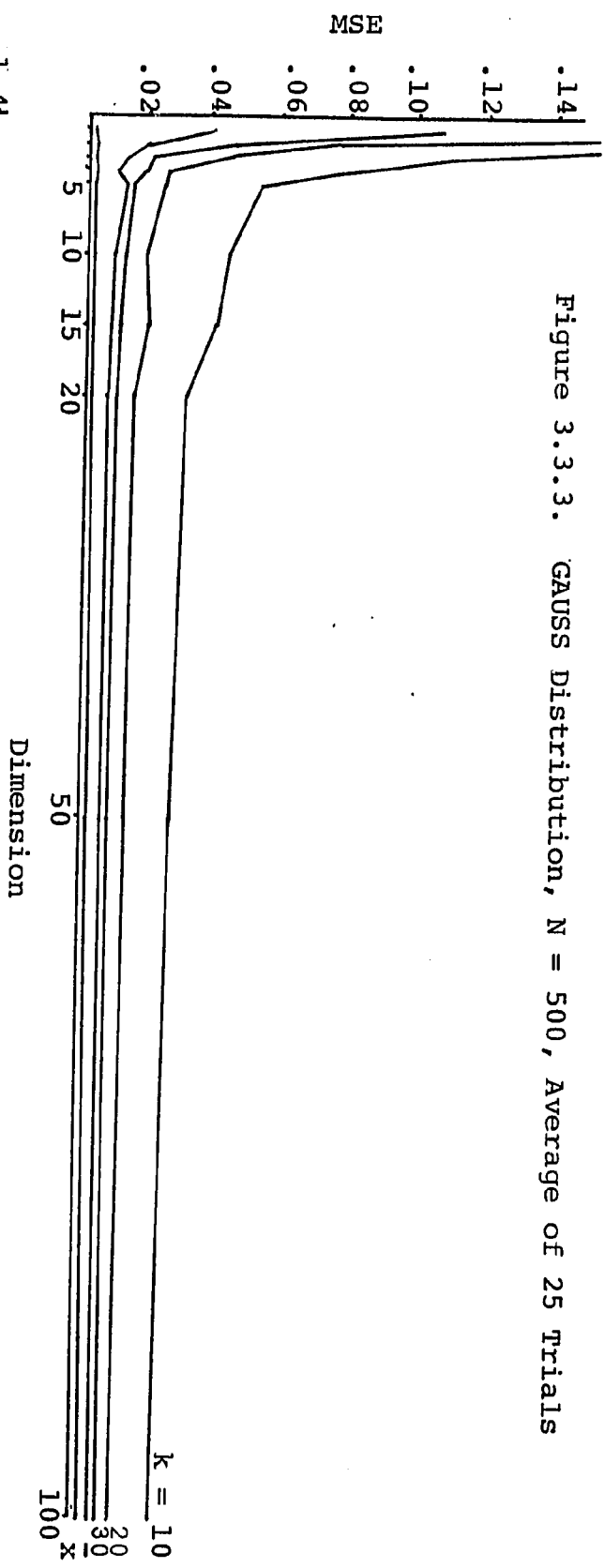


Figure 3.3.4.

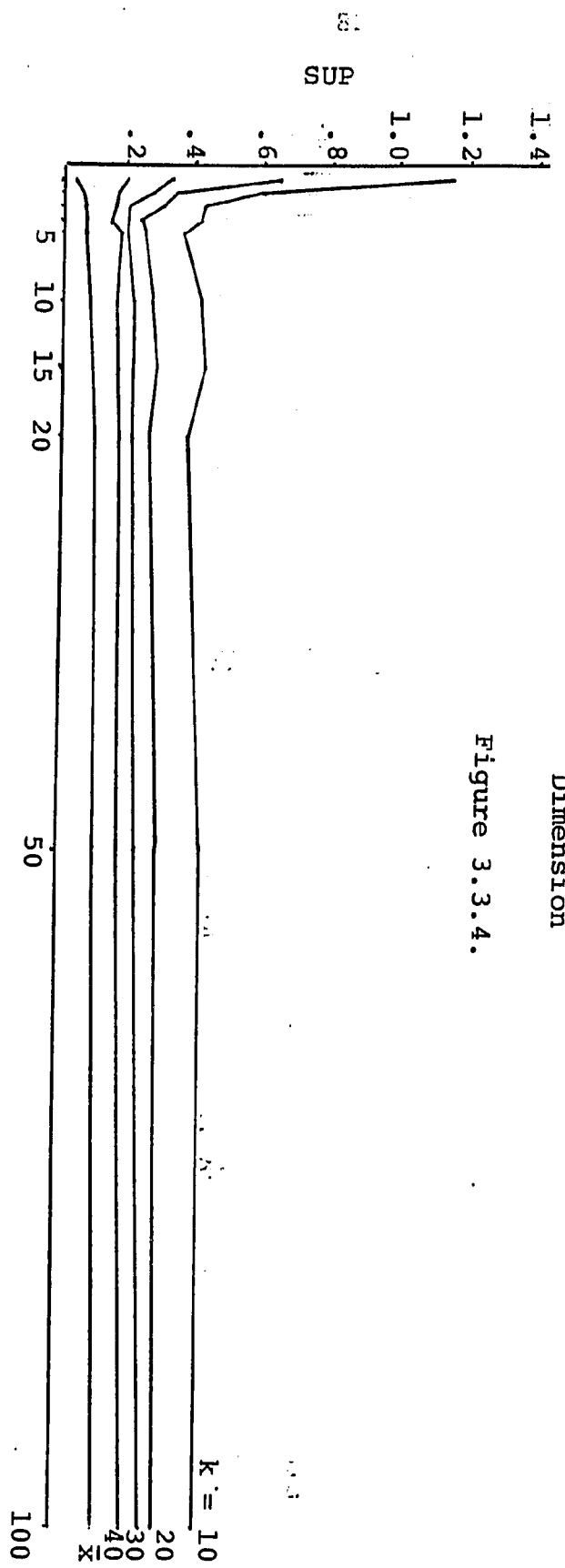


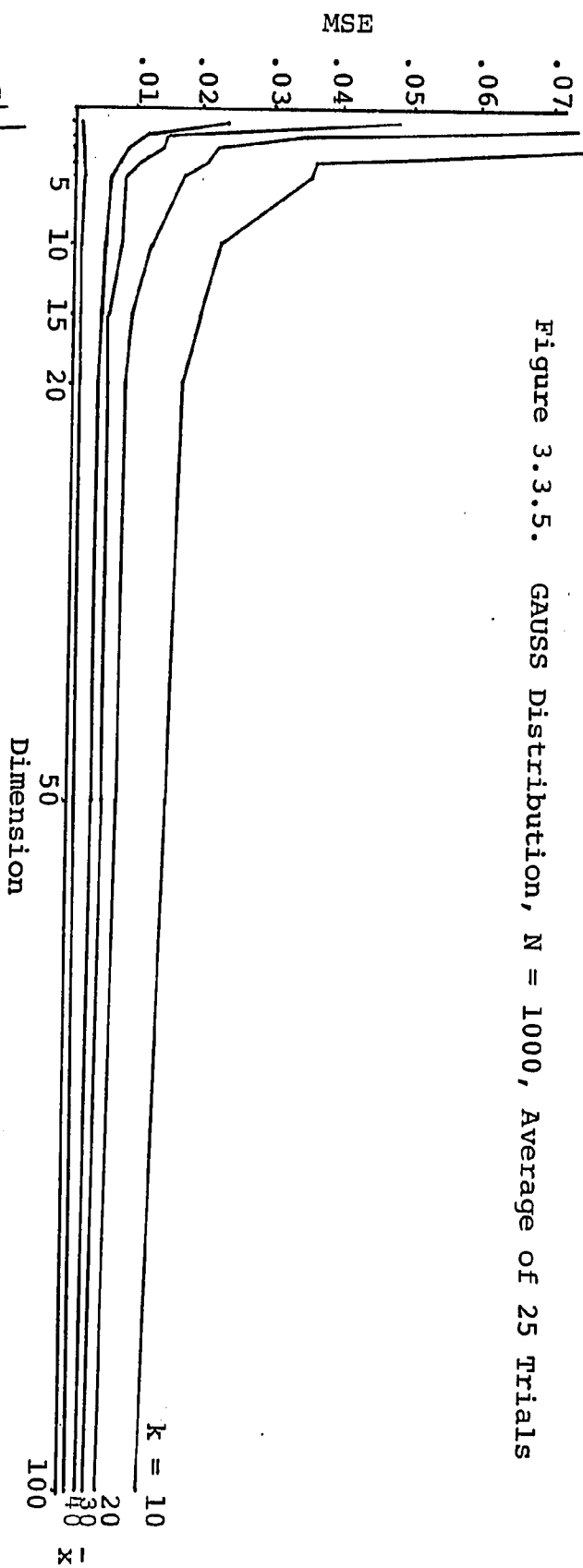
Figure 3.3.5. GAUSS Distribution, $N = 1000$, Average of 25 Trials

Figure 3.3.6

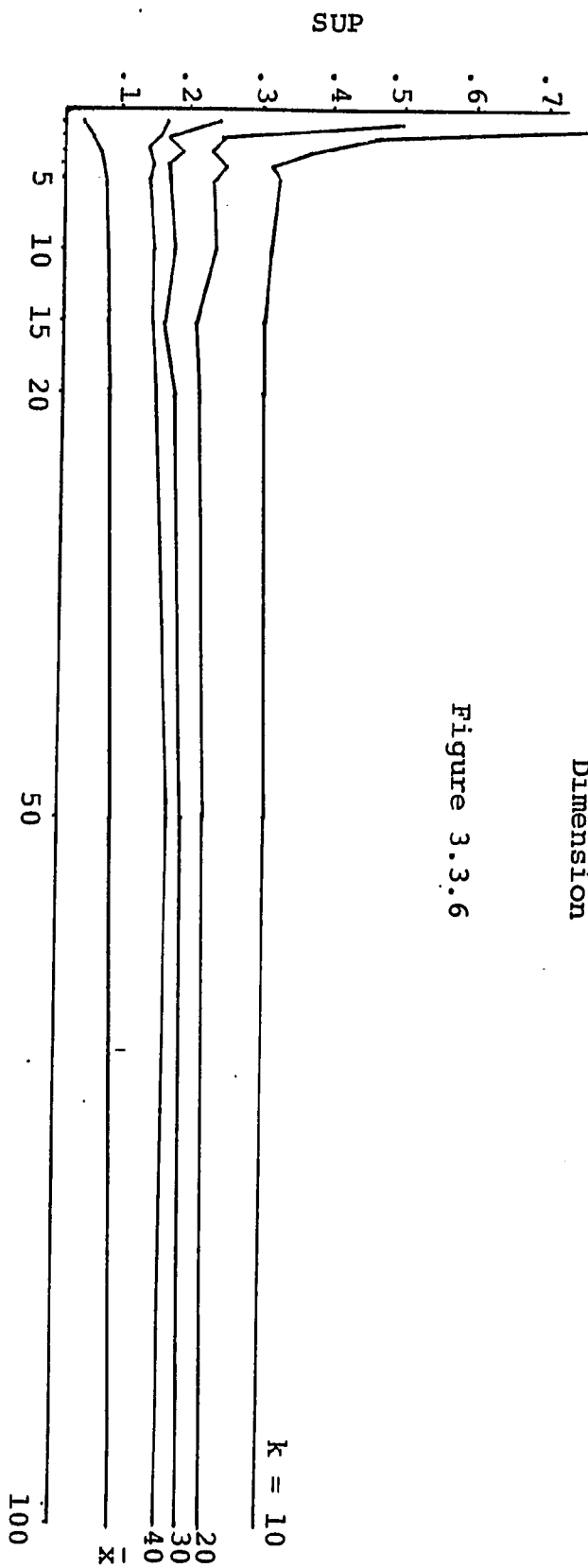
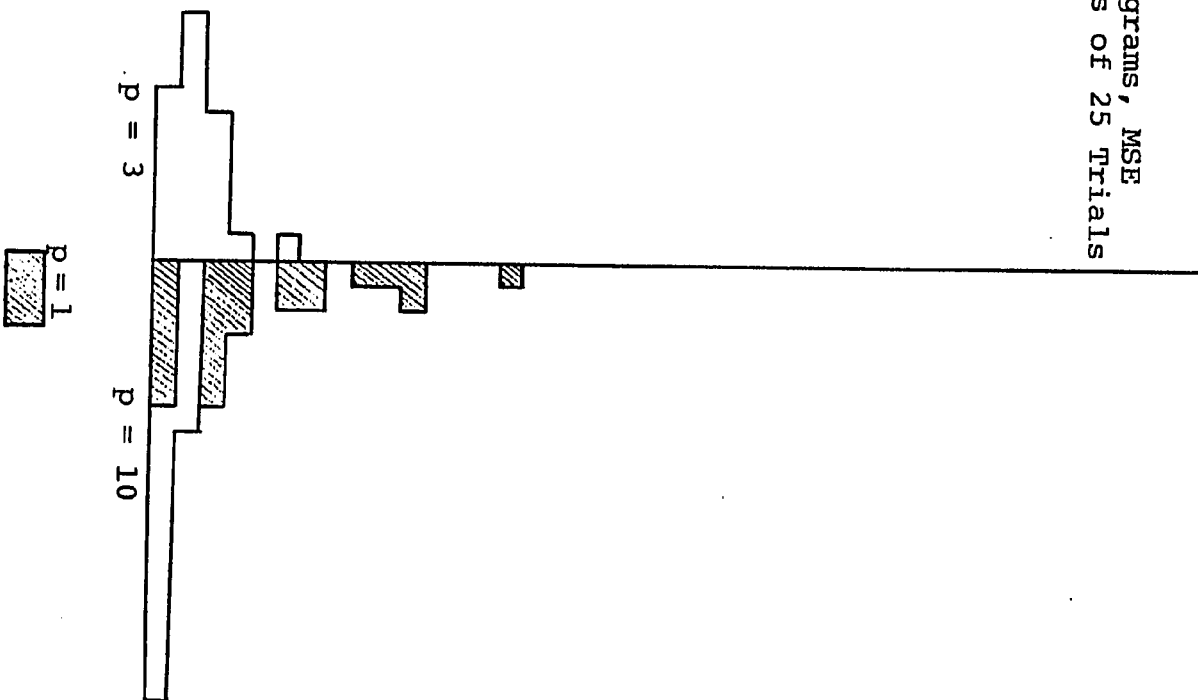
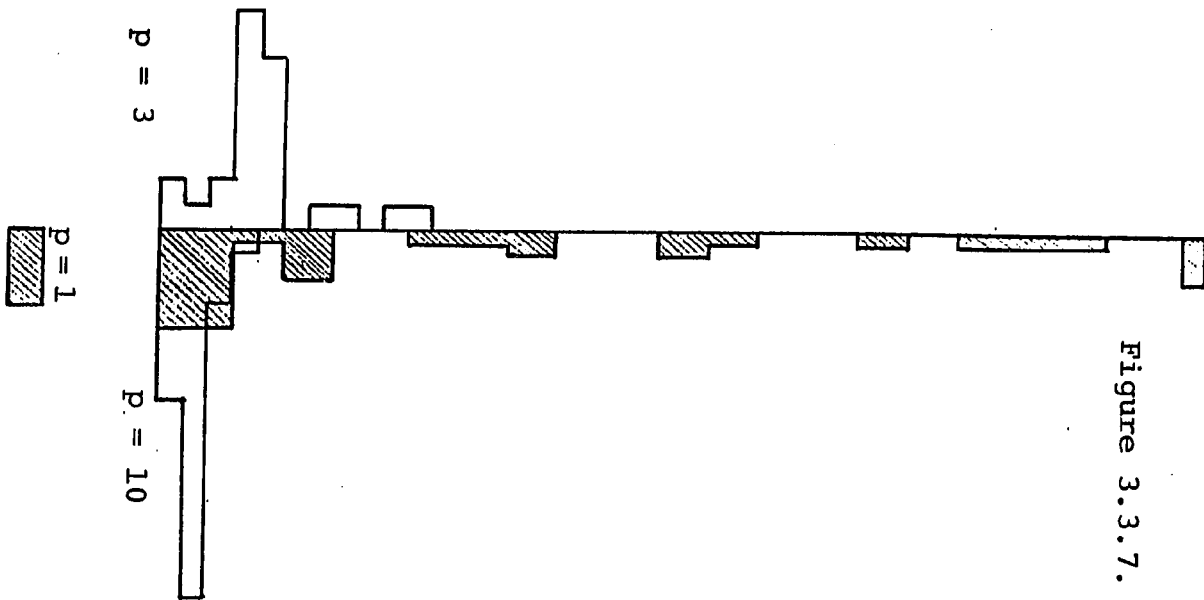


Figure 3.3.7. Histograms, MSE
Values of 25 Trials



the two graphs. The histograms clearly depict the "tailiness" of the univariate mean update as well as the diminution of extreme error which accompanies the introduction of additional independent variates.

The source of the frequent large errors in low dimension is suggested by Table 3.3.1. The measure MED used in the table is the median over all components of component-wise error in the mode estimate, i.e., median $\{(x_i - x_i^*), i = 1, \dots, p\}$. Table 3.3.1 gives the average MED recorded over 25 trials. As explained in Section 3.2, MED values indicate the extent to which the iterative algorithm is able to detach itself from the starting point of the search. Consistently negative values of MED indicate that the algorithm is easily trapped by insignificant gaps or clusters in the spatial distribution of the sample points, and thus unable to complete a path of ascent to a genuine mode of the population density. Table 3.3.1 thus suggests the extent to which increased sample size, neighbor set, or dimension improve the hill-climbing ability of the mean update. The figures therein clearly indicate that in low dimensions, the mean update tends to stall prematurely along a path connecting the initial guess and the population mode, but that with increasing dimension the stalling problem is steadily and effectively reduced.

On the uncorrelated Gaussian data, for large dimensions, as long as k/n is not too small, the accuracy of the

TABLE 3.3.1. MED (Ave. Over 25 Trials).

N =	100	K = .1N	.2N	.3N	.4N
P =	1	-1.402	-.7901	-.5240	-.3066
	2	-.6369	-.3558	-.1659	-.09529
	3	-.4974	-.1615	-.05835	-.03797
	4	-.3560	-.1502	-.08002	-.05674
	5	-.3087	-.1537	-.03932	-.001098
	10	-.1106	-.04134	-.03142	-.02521
	15	-.0757	-.05907	-.03515	-.02629
	20	-.0710	-.03143	-.02575	-.01239
	50	-.0228	-.01500	-.008324	-.007152
	100	.0216	-.007647	-.001475	-.002391
N =	500				
P =	1	-1.113	-.6240	-.2988	-.1179
	2	-.4513	-.1997	-.1058	-.4533e ⁻¹
	3	-.2587	-.7079e ⁻¹	-.3231e ⁻¹	-.2191e ⁻¹
	4	-.1318	-.1781e ⁻¹	-.2395e ⁻¹	-.1130e ⁻¹
	5	-.1131	-.3673e ⁻¹	-.1994e ⁻¹	-.1821e ⁻¹
	10	-.3400e ⁻¹	-.1152e ⁻¹	-.4451e ⁻²	-.4378e ⁻²
	15	-.2819e ⁻¹	-.1262e ⁻¹	-.9842e ⁻²	-.4294e ⁻²
	20	-.1416e ⁻¹	-.5292e ⁻²	-.4377e ⁻²	-.4281e ⁻²
	50	-.5697e ⁻²	-.6858e ⁻²	-.2507e ⁻³	-.3903e ⁻²
	100	-.3630e ⁻²	-.2816e ⁻²	-.6750e ⁻³	-.5813e ⁻²

TABLE 3.3.1. Continued.

N =	100	K = .1N	.2N	.3N	.4N
P =	1	- .9721	-.4699	-.2074	-.7179e ⁻¹
	2	- .3166	-.8313e ⁻¹	-.3690e ⁻¹	-.3365e ⁻¹
	3	- .1560	-.2317e ⁻¹	-.1423e ⁻¹	-.1476e ⁻¹
	4	- .3819e ⁻¹	-.1854e ⁻¹	-.9446e ⁻²	-.6702e ⁻²
	5	- .7613e ⁻¹	-.2941e ⁻¹	-.2283e ⁻¹	-.3142e ⁻²
	10	- .1775e ⁻¹	-.5670e ⁻²	-.3071e ⁻²	+.2921e ⁻²
	15	- .1894e ⁻¹	-.3409e ⁻²	-.1200e ⁻²	+.2957e ⁻²
	20	- .6299e ⁻²	-.4556e ⁻²	+.4169e ⁻²	-.3142e ⁻²
	50	- .4729e ⁻²	-.3573e ⁻³	-.1831e ⁻²	-.1612e ⁻²
	100	- .7554e ⁻³	-.5963e ⁻³	-.1483e ⁻²	-.6754e ⁻³

mean update depends upon the size of the neighbor set almost independently of the number of observations in the complete sample as an illustration, Figure 3.3.8 depicts the average MED values for $k = 200$ with $N = 500$ ("o") and $k = 200$ with $N = 1000$ ("0"), in dimensions 10 and above.

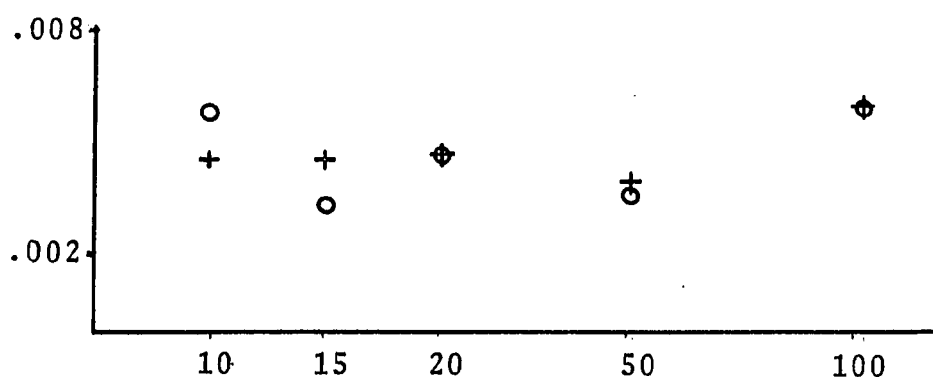


Figure 3.3.8

The elemental role of neighbor set size in high dimension is reinforced by the examination of MSE results. Figures 3.3.1, 3.3.3, and 3.3.5 clearly reveal that reduction of MSE is directly proportional to $1/k$ in dimension 20 or above; in other words, doubling the size of the neighbor set halves the observed MSE. Moreover, as indicated by Table 3.3.2, which gives results using neighbor sets of size 100, the average MSE of the mean update in high dimensions is determined almost entirely by k , without regard to total sample size.

In high dimensions, the rate of decrease observed in MSE with increasing neighbor set size equals the rate of decrease expected for the variance of an estimated mean as

TABLE 3.3.2. Average MSE with Neighbor Sets of Size $k = 100$.

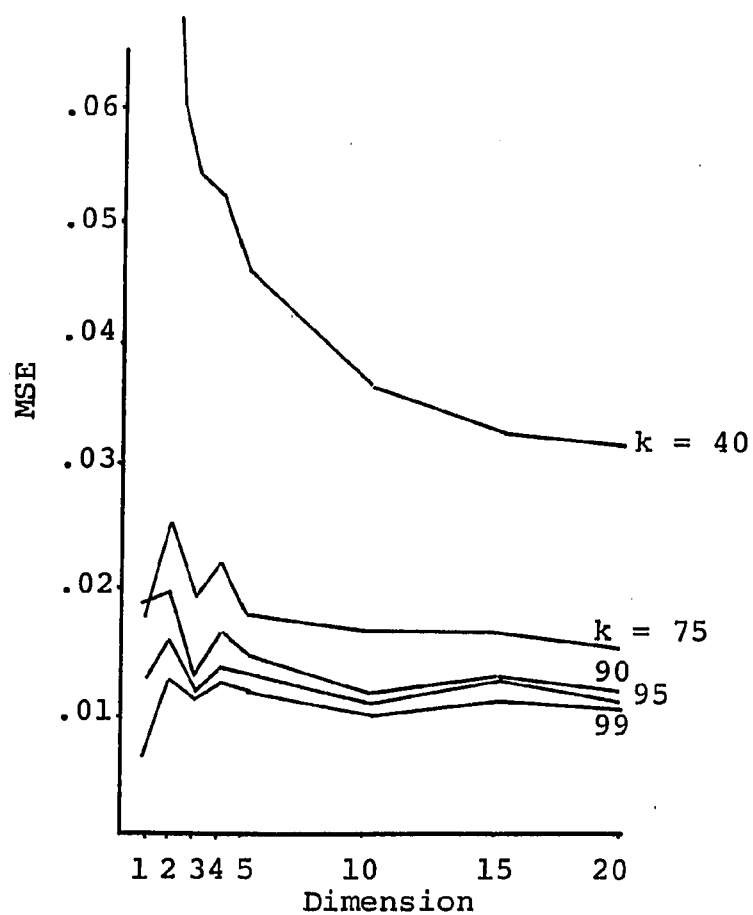
	P =	1	2	3	4	5
N =	100	.006	.008	.009	.009	.009
	500	.390	.071	.041	.023	.022
	1000	.945	.127	.060	.035	.034

	P =	10	15	20	50	100
N =	100	.010	.010	.010	.010	.010
	500	.017	.018	.014	.013	.012
	1000	.021	.018	.016	.014	.011

sample size increases. The implication is that with increasing dimension, the bias component of the MSE becomes negligible. Since with a symmetric distribution there is no inherent bias in the mean update, that proportion of the MSE which is not explainable by sampling variability in the final iterations reflects premature interruption of the ascent of the mean update. The disappearance of this additional error thus corroborates the assertion that in high dimensions the ascent operations of the mean update are completely realized.

Thus, with independent normal data, our empirical results indicate that the stalling problem of the mean update diminishes as dimension increases. In sufficiently high dimension the mean update yields an unbiased mode estimate whose final neighbor set contains the population mode within its convex hull. Moreover, the sampling variability of the mode estimate in these high dimensions essentially derives from the variability of the final mean operation, and though not specifically valued as such for our purposes, the mean update functions on symmetric distributions effectively as a trimmed mean, despite the absence of a unique order relation in the sample space. Moreover, as indicated by Figure 3.3.9, which compares MSE results for low and high neighbor set sizes, the amount of trimming which can be tolerated before the estimation procedure deteriorates increases appreciably with additional independent dimension.

Figure 3.3.9. GAUSS Distribution, $N = 100$, Average MSE, 25 Trials



"Problem" Data Sets

For testing the mean update on the "problem" distributions introduced in Section 3.2, the maximum dimension of the simulations was limited to twenty, and with occasional exceptions, the sample size used was 100. These two restrictions remained in place throughout the subsequent duration of the study, including tests of the Newton and weighted mean procedures discussed in later chapters. The dimensional restriction was desired to reduce data volume, and, as has already been discussed, every indication is that dimension twenty can function as proxy for yet higher dimensional behavior.

An additional change made in the testing format was the occasional introduction of a high range for neighbor set values. In such cases, tests of each algorithm on each distributional type were conducted in pairs using neighbor sets $k = .1N, .2N, .3N, .4N$, and $k = .75N, .90N, .95N, .99N$. The two k -ranges were employed in separate simulations with randomly generated seed values for the random number generation, and thus with different generated data. Comparison within a range (.2N vs. .3N, for example) will refer to exactly the same data. Comparisons between ranges will not, and small discrepancies between quantities comparable across the ranges, in particular, the full sample mean, will naturally occur.

The first of the alternate tests of the mean update was

performed using data generated from the multivariate Cauchy distribution. Cauchy data was employed to investigate two questions. First was to determine to what extent the ascent interruption and resulting gross errors characteristic of small neighbor sets are tail phenomena; that is, whether they occur primarily where the population density is nearly flat and numerically small. Second was to investigate limitations on the use of neighbor sets containing a large proportion of the sample.

Plots of measures MAD and SUP for the Cauchy simulation are given in Figure 3.3.11 and Figure 3.3.12. The well-known instability of the sample mean with Cauchy data is quite evident in the plots. The small and moderate neighbor set mean update is relatively impervious to the greatly exaggerated tails of the population density. In fact, in low dimensions the performance of the mean update with $k = 20$ to $k = 40$ betters that of the Gaussian trials, and for $k = 10$ the improvement with increasing dimension is quite rapid. The tails of the Cauchy distribution do not have a decided impact until the neighbor set exceeds 40% of the sample size. Beyond dimension two the performance of the mean update is virtually unchanged over the range $k = 20$ to $k = 75$.

The next test of the mean update utilized highly correlated Gaussian data, designed so that the data have a lower dimensional character than the dimension in which

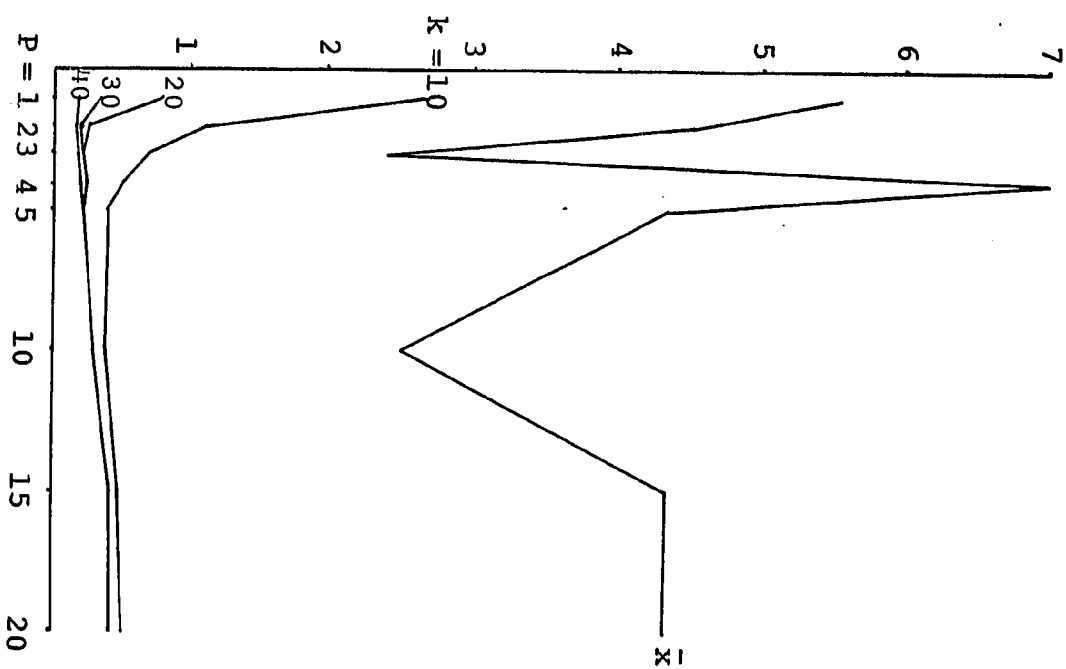
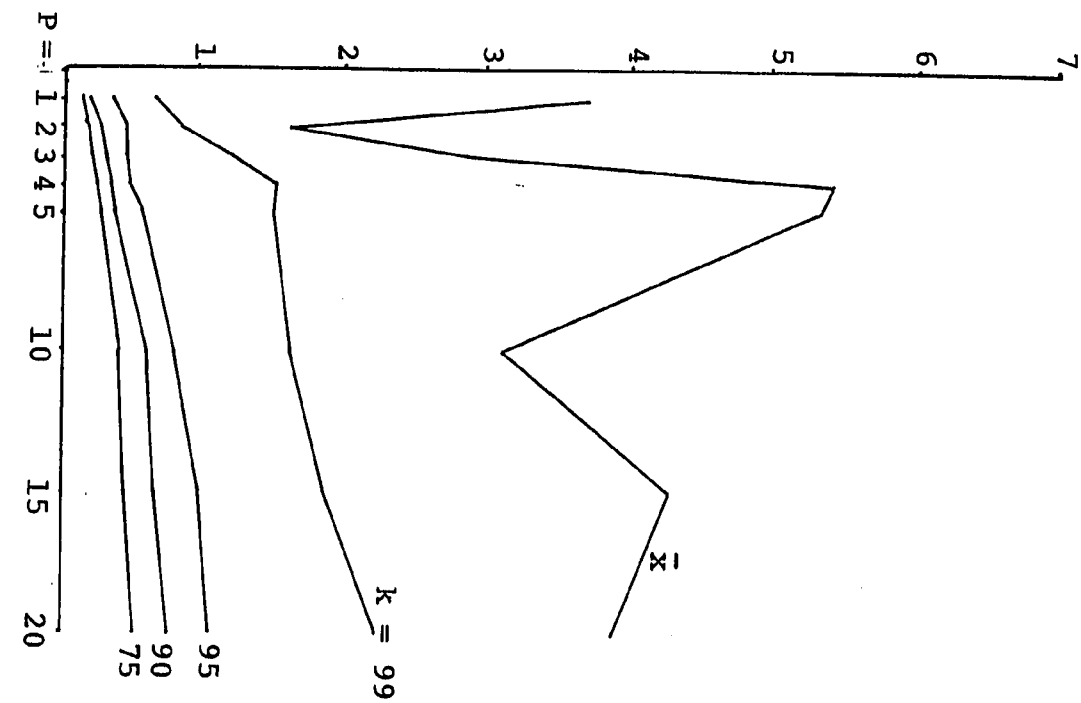
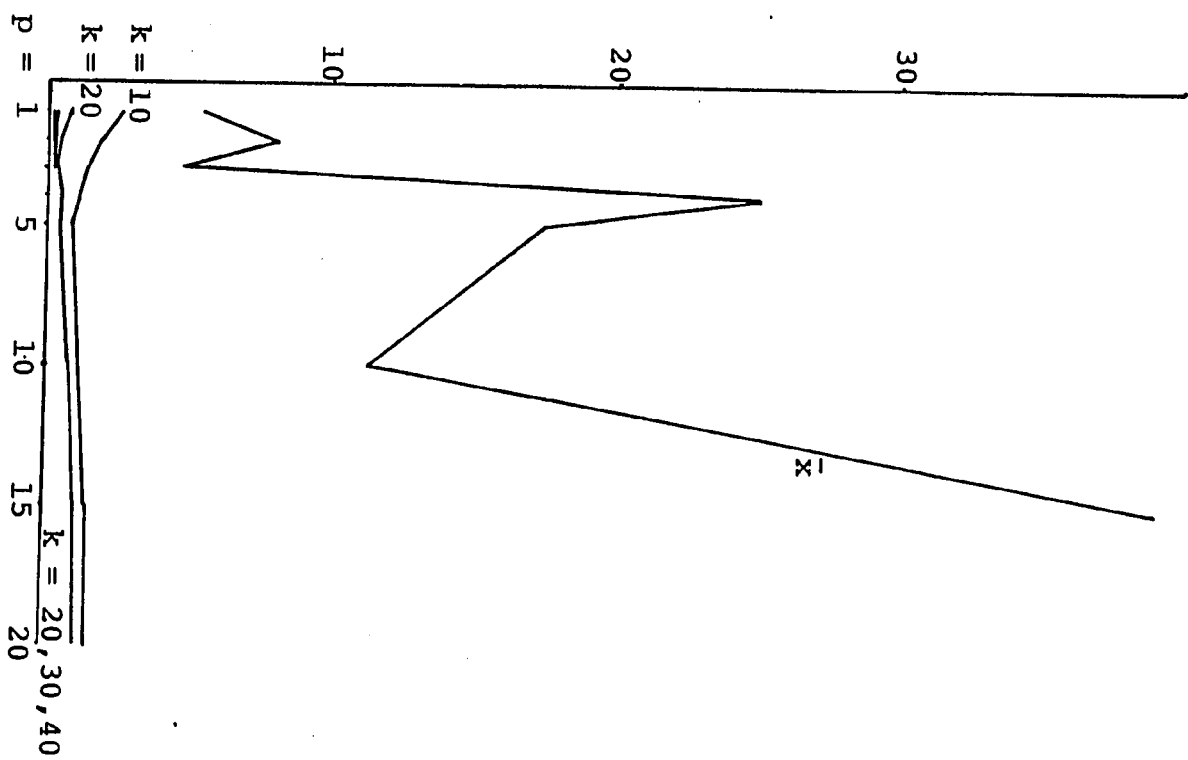
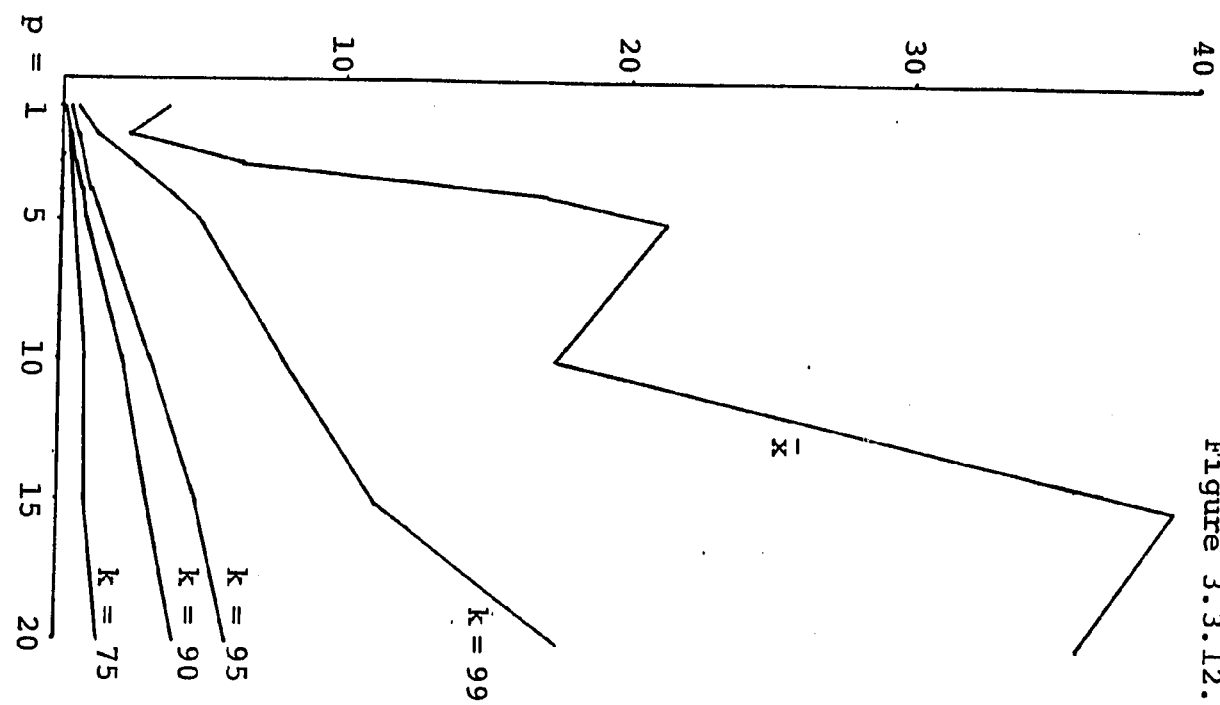
Figure 3.3.11. CAUCHY, $N = 100$, MAD

Figure 3.3.12. CAUCHY, $N = 100$, SUP

measurements are recorded. The correlated Gaussian distribution, denoted R9GAUSS, had all variances equal to 1.0, and all covariances 0.9. This produced increasingly ill-conditioned covariance matrices and increasingly extreme distributional patterns as dimension grew. Results with the correlated Gaussian data are presented in Figures 3.3.13 and 3.3.14. Both figures give MSE values for the correlated data on the left, compared with analogous results for uncorrelated data on the right. It is clear that, so far as it presents itself to the mean update algorithm, the correlated data is effectively one-dimensional over the range of dimensions tested; also the mean update "sees it" as one-dimensional regardless of the neighbor set used. It appears that subdimensionality is a problematic feature that the mean update may in severe cases be unable to circumvent. However, we should remark that the data sets we generated in the higher dimensions are quite extreme. Even data that is highly correlated in dimension 10, say, is likely to have four or five effectively independent coordinates or linear combinations of coordinates. In this case, by analogy with the standard Gaussian results, we would expect the mean update to show a several-fold improvement over the extreme correlation results of Figure 3.3.13.

Average L2 results for the skew two-component Gaussian mixtures are given in Figures 3.3.15, 3.3.16, and 3.3.17. The mixture densities are described in Section 3.2. Figures

Figure 3.3.13. Effect of Correlation on Mean Update

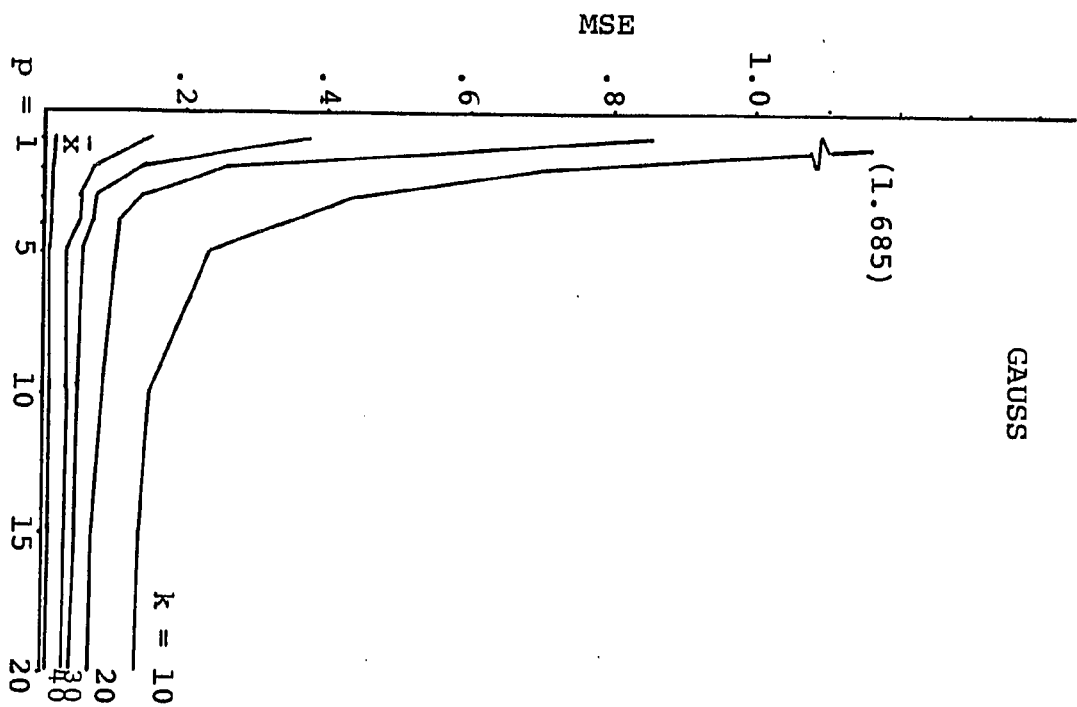
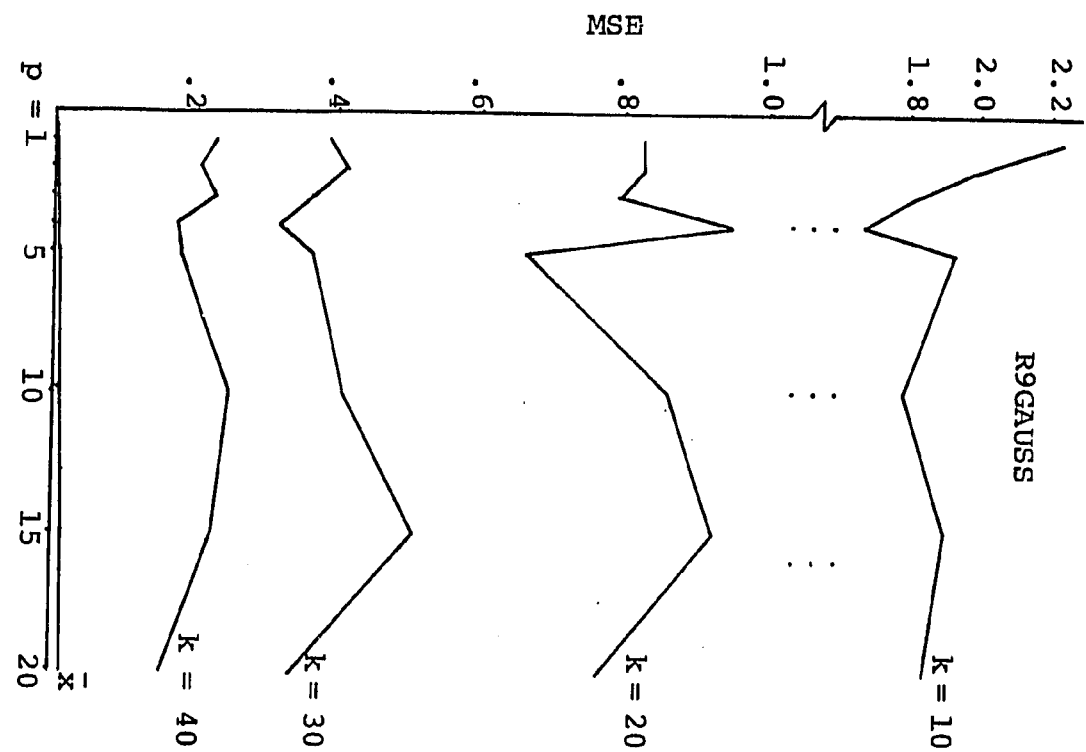


Figure 3.3.14. Effect of Correlation on Mean Update

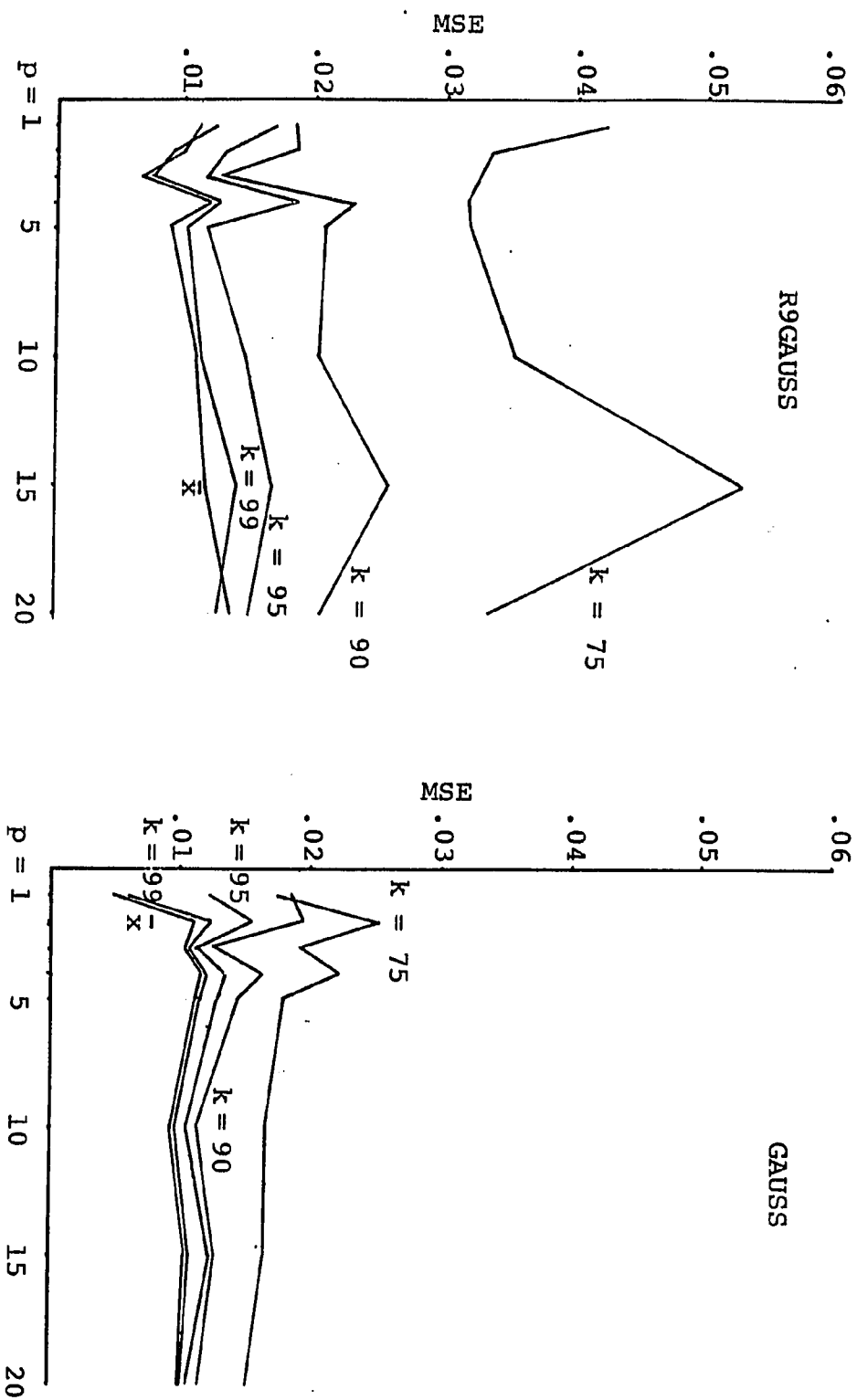


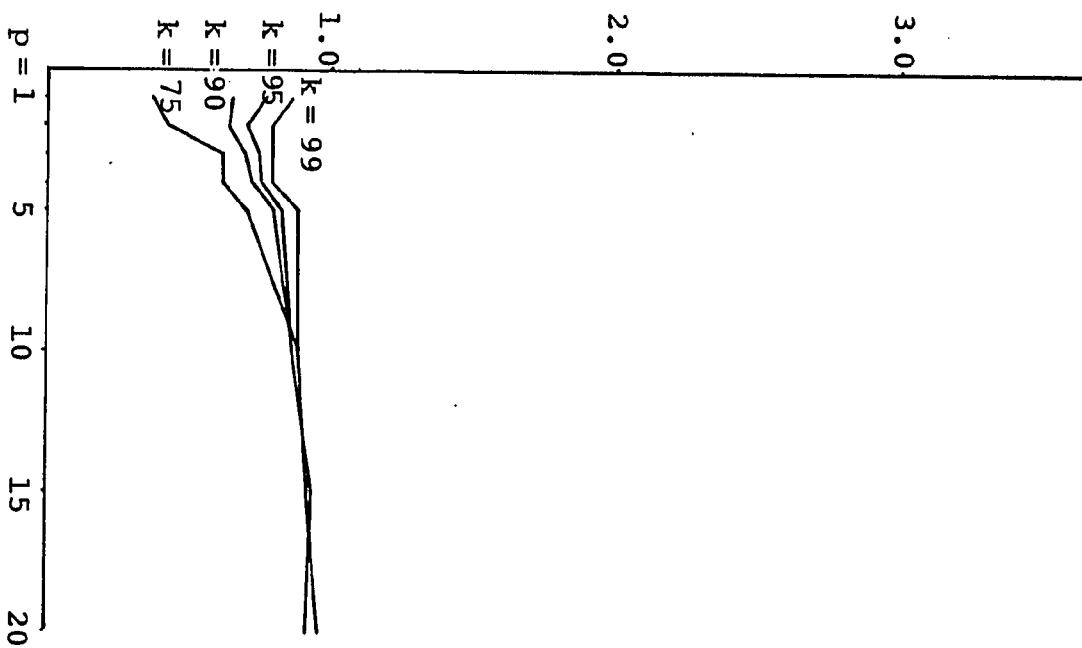
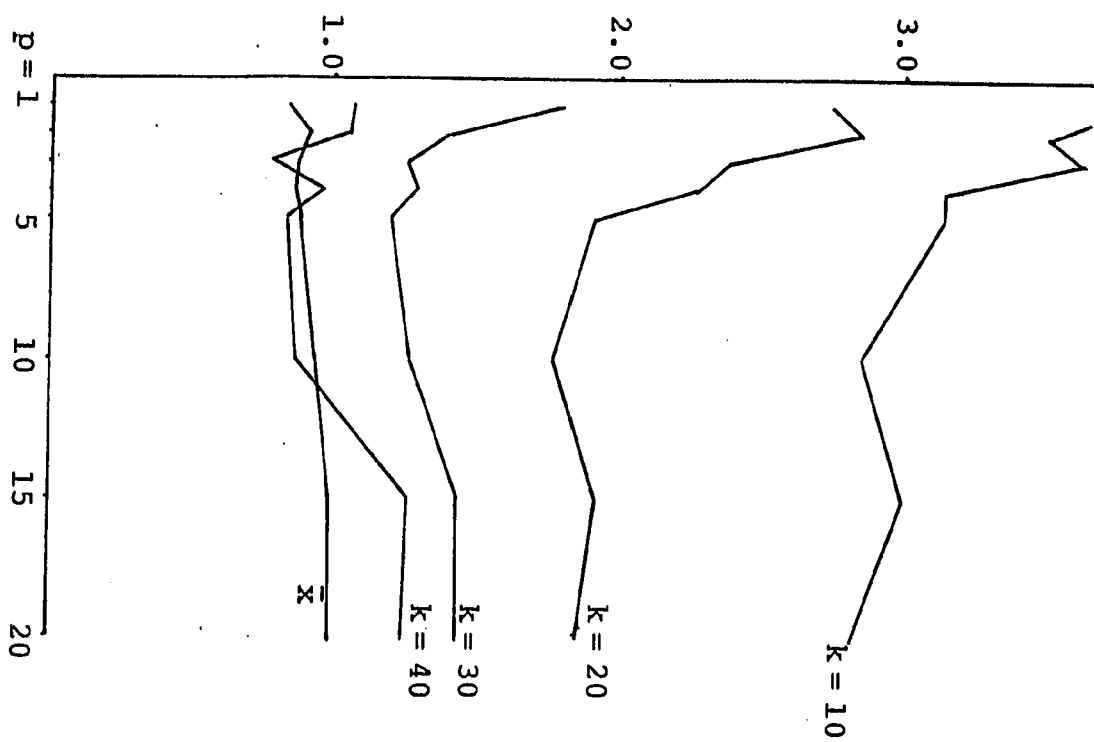
Figure 3.3.15. G2SKEWIT, $N = 100$, Euclidean Norm

Figure 3.3.16. G2SKEWRT, $N = 100$, Euclidean Norm

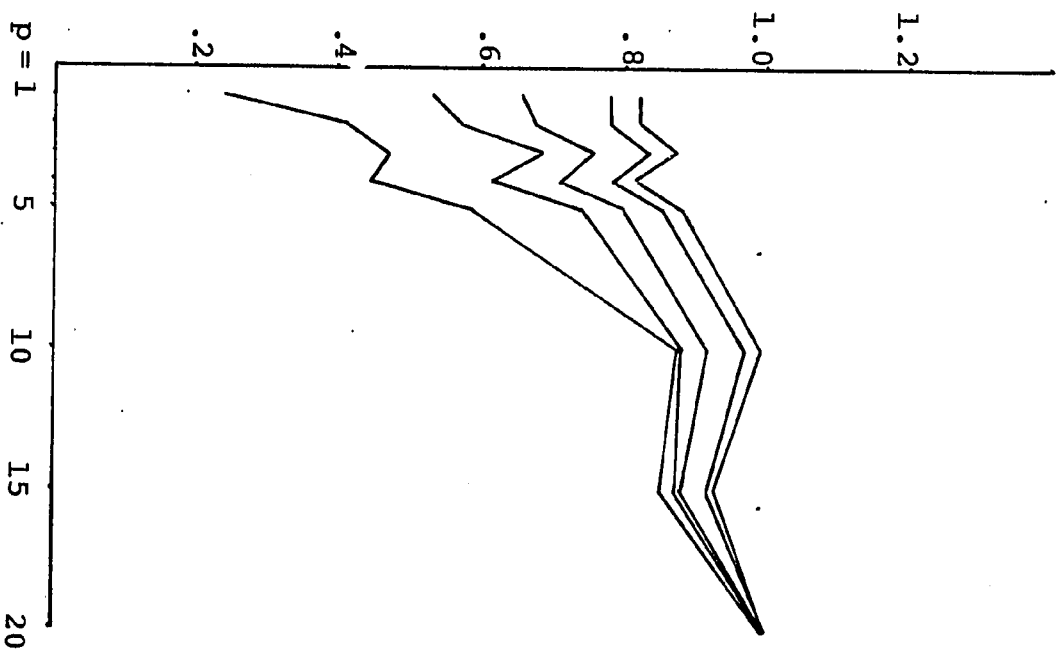
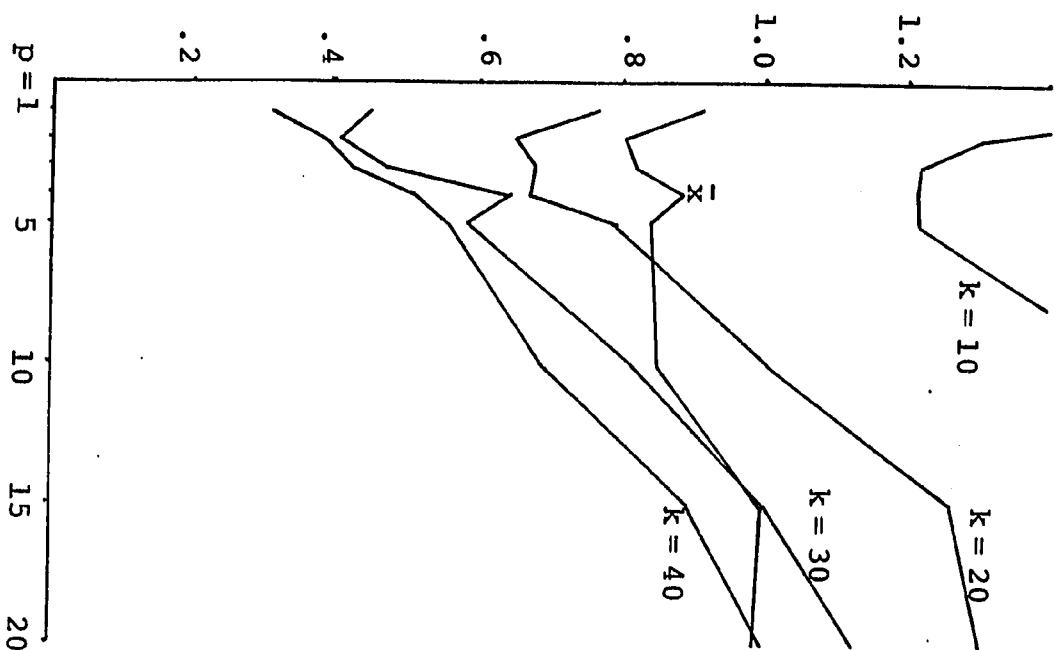
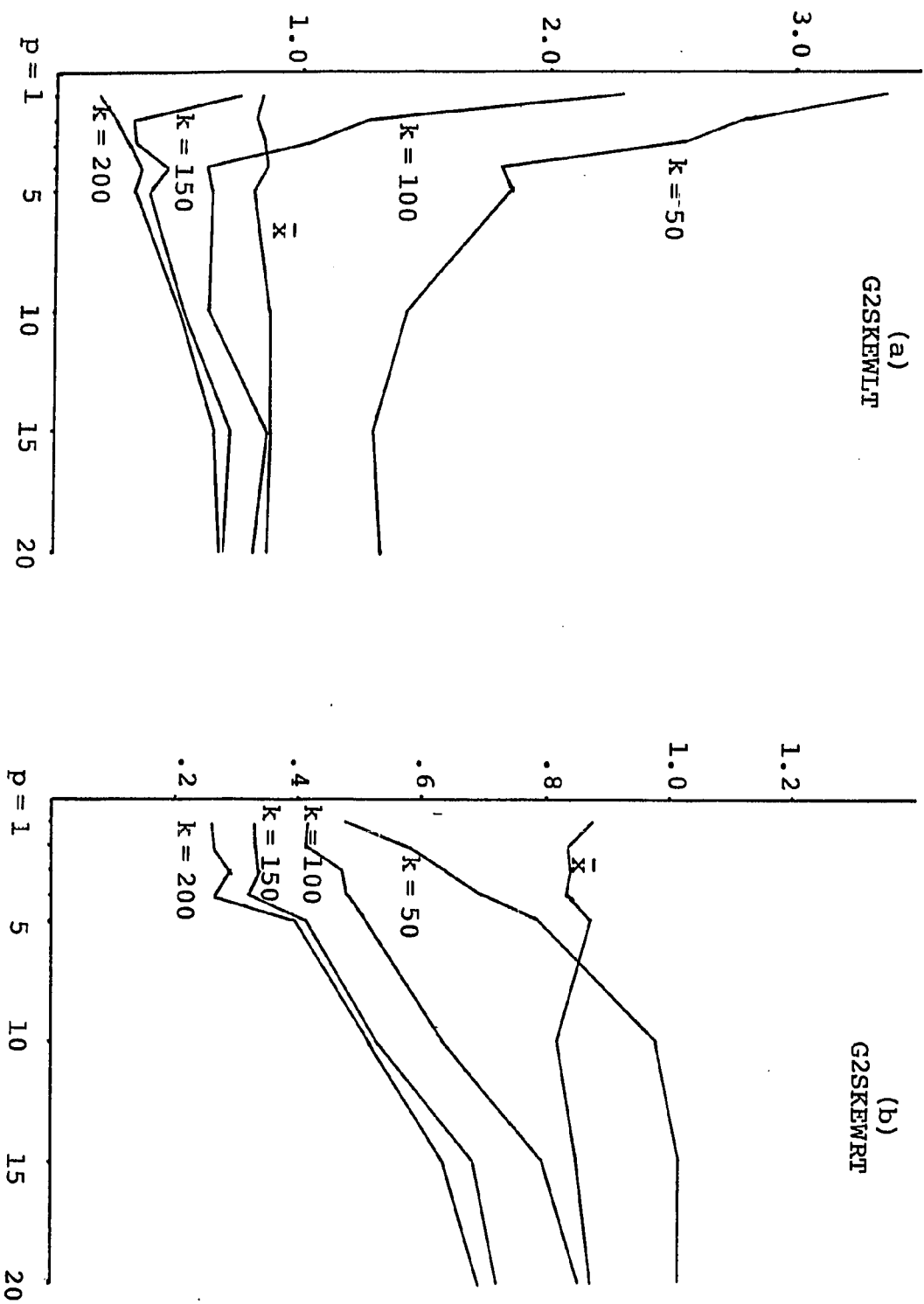


Figure 3.3.17. G2SKEW, N = 500, Euclidean Norm



3.3.15 and 3.3.16 are based on sample sizes of 100, and present results both for low and high ranges of the size of the neighbor set. Figure 3.3.17 is based upon samples of 500 observations, and considers only the smaller neighbor sets. The consistency of the mean update algorithm is guaranteed by Theorem 3.1.1 under conditions which include the provision that $k/N \rightarrow 0$ as $N \rightarrow \infty$. The ability of the mean update to respond in practice to significant asymmetries in the population density hinges upon its ability to perform effectively with small neighbor sets on practicable sample sizes. With an insufficient degree of local concentration, opportunity for the algorithm to detect fine structure in the data is lost. Figures 3.3.15b and 3.3.16b depict the mean update losing its grasp on the mode as k is increased from 75% of the sample size, converging eventually of course to the sample mean, which serves as a limiting example of non-adaptability.

Figures 3.3.15a and 3.3.16a suggest that the mean update with sample size $N = 100$ will have difficulty responding to relatively subtle distributional features such as those posed by the multivariate skew mixtures. The disparity between results obtained with the least favorable orientation of G2SKEWLT and the most favorable of G2SKEWRT indicate that the small-sample performance of the mean update will depend heavily upon the location of the initial guess and the shape of the population density. For G2SKEWLT

there does exist a range of neighbor set sizes (approximately $50 \leq k \leq 75$) for which the mean update is able to bypass the near mode in low dimensions and also distinguish the population mode from the population mean. Unfortunately, the range is small, and deviation from it, particularly on the low end, rapidly degrades performance. It should be remarked that the orientation dependence of the small neighbor set mean update, or any other mode estimator, may possibly be overcome by an iterated multistart algorithm to be discussed in Chapter VI; nevertheless, the "single execution" performance of the mean update with $N = 100$ is typically either more sporadic than desirable, or else too insensitive.

Fortunately, tests conducted with sample sizes of 500, given in Figure 3.3.17, demonstrate that the small sample handicaps of the mean update disappear fairly rapidly with additional observations. For G2SKEWLT in dimensions 2 through 10 the average error of the mean update drops by a factor of 2 or 3 for almost all of the neighbor set-dimension pairs. For G2SKEWRT the mean update already performs well in low dimensions with sample sizes of 100; however, the increase to $N = 500$ prolongs the advantage over the sample mean, except for $k = .2N$, throughout the range of dimensions tested. More significantly, with minimal exceptions, confined to dimension 1 or 2 or to very low k , the performance of the mean update is almost identical for

both G2SKEWLT and G2SKEWRT, indicating that a genuine and consistent estimation procedure is obtained irrespective of the arbitrary selection of a starting point for the algorithm.

Conclusions

The mean update will be effective as a hill-climbing local mode finder on moderately sized data sets, particularly in dimensions three and above; moreover, there is no apparent dimensional limit to its application. In multivariate settings, it provides reasonable efficiency with standard normal data, and it is able to adapt to all the types of irregularities devised for testing, though naturally the differentiation of some features requires larger sample sizes than the differentiation of others. Nevertheless, we would like to improve upon the performance of the mean update with small neighbor sets. For the sample sizes and distributional features we have considered the mean update is most effective utilizing 30% to 40% of the sample in its neighbor sets. Experience with bimodal data, to be presented in Chapter VI, has indicated that even with such a degree of local concentration, the mean update can smooth over and ignore important secondary modes.

Two approaches were considered to improve upon the mean update. First was the more explicit implementation of a Newton-Raphson algorithm, utilizing the analyses of Section 2.2 to derive optimal estimates of the population density

and its first and second derivatives. The results of this exercise, which enjoyed reasonable success in low dimensions, are presented in the next chapter. The second approach, motivated by observations made in Section 3.1, was to improve the adaptiveness of the mean update by working from a kernel density estimate with more flexible curvature. This approach led to the weighted mean and truncated weighted mean updates, discussed in Chapter V.

IV. MINIMIZATION OF THE ESTIMATED DENSITY BY NEWTON'S METHOD

4.1. Introduction

The update step of the mean update algorithm was introduced as corresponding to a Newton step performed upon the variable kernel density estimate (3.1.1). That estimate presents some conceptual difficulties because it is only piecewise differentiable, and is not consistent as an estimator of second order derivatives. A natural strategy for improving the adaptability and the statistical efficiency of the mean update is to replace (3.1.1) with a density estimate that is consistent and optimally efficient for derivative information, and design a Newton-like optimization procedure for the revised estimate. Statistical analysis of kernel density estimators is presented in Chapter II. The fixed bandwidth kernel estimator f_n (2.2.16) inherits the differentiability properties of the kernel function K . Thus if K is properly chosen, a Newton's algorithm applied to f_n will be well-defined, and convergence to a maximizer of f_n is guaranteed by standard results. Optimal kernels for estimation of f and its derivatives, based upon analysis of the asymptotic rate of convergence of the estimates in mean squared error, are given in equations (2.2.5), (2.2.29), and (2.2.30). As these kernel functions are not differentiable at the boundary of their supports, smoothed versions were developed, yielding product

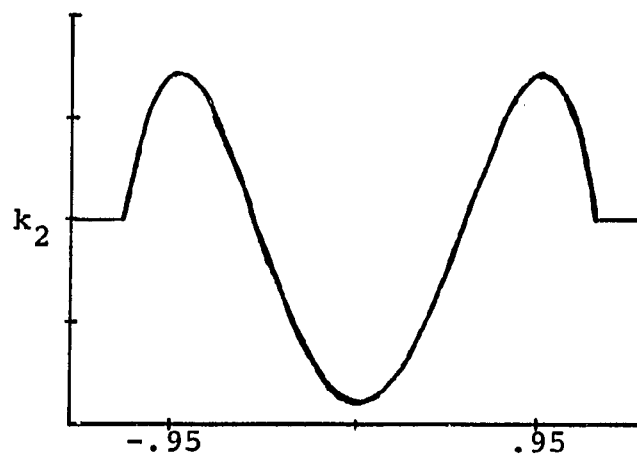
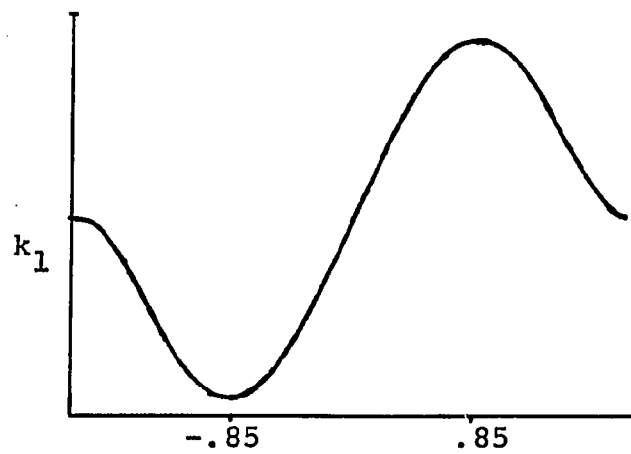
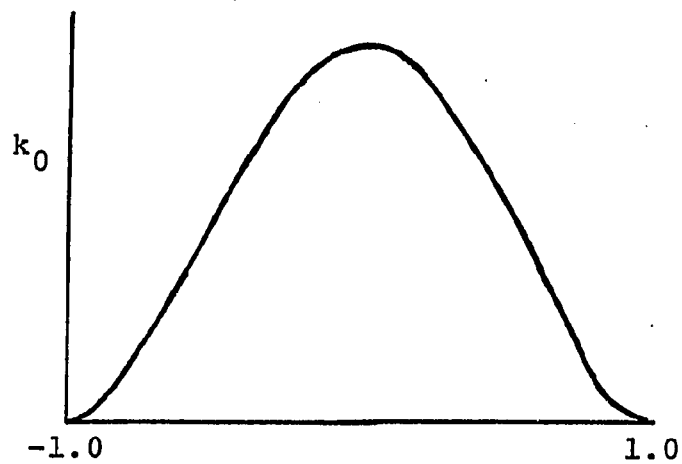
kernels with components

$$\begin{aligned}
 K_0(x) &= \frac{15}{16} (1-y^2)^2, \quad y^2 \leq 1 \\
 K_1(x) &= \frac{35}{48} \sqrt{3} y (1 - \frac{y^2}{3})^2, \quad y^2 \leq 3 \\
 K_2(x) &= \frac{35}{2} \sqrt{2/3} (-\frac{1}{4} + y^2 - \frac{5}{9} y^4), \quad y^2 \leq 3/2.
 \end{aligned} \tag{4.1.1}$$

The kernel functions are depicted in Figure 4.1.1.

As Newton-like methods for optimization depend upon the accuracy of their gradient information, and kernel estimates of the gradient may be expected to have appreciable sampling variability, Newton's method appears to be mismatched with the statistical problem of estimating modes, particularly in high dimensions. However, high accuracy in the gradient is essential only in the vicinity of the mode, where gradient values near zero must be compared. The algorithm retains its hill-climbing ability under much rougher conditions than permit accurate placement of the modes. In particular, a procedure incorporating Newton's method techniques should be able to detect the presence of multimodality even when the locations of the modes it estimates are only approximate.

Simulation results will be presented which indicate that a Newton's method phase, using the mean update to provide an initial guess, often does correct the overt stalling problems of the mean update, and supplies better sensitivity about the mode as well. The effectiveness of the Newton's

Figure 4.1.1. Kernels for f and its Partial Derivatives

procedure is greatest in low dimensions. In high dimension, it has a negligible effect in the presence of well-scaled, uncorrelated dispersion. However, in such cases it "does no harm", and on data which is highly correlated and has effectively less dimensionality, it retains its usefulness, particularly with smaller neighbor sets.

4.2. Description of the Algorithm and Consistency

The general construction of a Newton or quasi-Newton procedure is given in Algorithm 4.2.1.

Algorithm 4.2.1

1. Compute $\nabla f(\mathbf{x}_k)$ and terminate if stopping criteria are met.
 2. Compute $H_k = \nabla^2 f(\mathbf{x}_k)$ or an approximation to it.
 3. Factor H_k and perturb it if necessary to correct ill-conditioning or insure positive definiteness.
 4. Solve for the Newton step $H_k \mathbf{s}_k = -\nabla f(\mathbf{x}_k)$.
 5. Decide whether to take the full Newton step, $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$, or backtrack to choose a shorter step in the Newton direction.
- Go to Step 1.

Note that the algorithm presented is for function minimization; to apply it for finding modes of the estimated density, f_n , one minimizes $-f_n$. A complete introduction to Newton and quasi-Newton methods for unconstrained minimization may be found in Dennis and Schnabel [1983], from which

much of the following discussion is drawn.

Important features of the algorithm from the standpoint of assuring consistency are steps 3 and 5. These steps "globalize" the Newton's method procedure; that is, they insure that gradient values generated by the algorithm converge to zero. The proof of this fact is contained in two results of P. Wolfe [1969, 1971]. The linesearch routine, invoked from a location x_c , attempts first to take a full Newton step and then backtracks if necessary until a point x_{k+1} is found which satisfies the conditions

$$(i) \quad f(x_{k+1}) - f(x_k) \leq \alpha \frac{\|x_{k+1} - x_k\|}{\|s_k\|} \nabla f(x_k)^T s_k \quad (4.2.1)$$

$$(ii) \quad (\nabla f(x_{k+1}) - \beta \nabla f(x_k))^T (x_{k+1} - x_k) \geq 0,$$

for $0 < \alpha < \beta < 1$. The first result of Wolfe simply states that the conditions (4.2.1) can be satisfied simultaneously if $\nabla f(x_k)^T s_k < 0$. The second result is as follows:

Proposition 4.2.1. Let $f: \mathbb{R}^p \rightarrow \mathbb{R}$ be continuously differentiable on \mathbb{R}^p and assume there exists $\gamma \geq 0$ such that $\|\nabla f(z) - \nabla f(x)\| \leq \gamma \|z - x\|$ for every $x, z \in \mathbb{R}^p$. Then, given any $x_0 \in \mathbb{R}^p$, if f is bounded below, there exists a sequence $\{x_k\}$, $k = 0, 1, \dots$ obeying (4.2.1), and either

$$\nabla f(x_k)^T d_k < 0$$

or

$$\nabla f(x_k) = 0 \quad \text{and} \quad d_k = 0$$

for each $k > 0$, where

$$d_k = x_{k+1} - x_k.$$

Furthermore, for any such sequence, either

$$\begin{aligned} & \text{(i)} \quad \nabla f(x_k) = 0 \quad \text{for some } k \geq 0, \quad \text{or} \\ & \text{(ii)} \quad \lim_{k \rightarrow \infty} \frac{\nabla f(x_k)^T d_k}{\|d_k\|_2} = 0 \end{aligned} \tag{4.2.2}$$

Step 3 in Algorithm 4.2.1, by requiring positive definiteness of the matrix H , insures that each step direction is a descent direction; in addition the prevention of ill-conditioning in the Hessian approximation prevents the step s_k from approaching orthogonality with the gradient.

It is now possible to give the main theoretical result of this chapter, which states general conditions under which the maximization of kernel density estimates by Algorithm 4.2.1 yields consistent estimates of the modes of a probability density function.

Theorem 4.2.1. Let f be a bounded, continuously differentiable probability density function in \mathbb{R}^p , and let $\{x^{(1)}, \dots, x^{(n)}\}$ be a sequence of n independent observations drawn from f . Let $K: \mathbb{R}^p \rightarrow \mathbb{R}^1$ be continuously differentiable and ∇K be Lipschitz continuous on \mathbb{R}^p with constant γ . Suppose that K satisfies the conditions (2.2.15) for estimating f , and that for each i , $i = 1, \dots, p$, $\partial/\partial x_i K(\cdot)$ is a kernel function satisfying conditions (2.2.15) for estimating

$\partial/\partial x_i f$. In addition, suppose that the assumptions of Proposition 2.3.2 hold regarding the characteristic function of K . Finally let $\{h(n)\}$ be a sequence of real numbers satisfying $h(n) = \beta n^{(s-1)/p}$ for some $\beta > 0$ and $s \in (2/3, 1)$. Let x_n^* be the result of applying Algorithm 4.1.1 to the kernel density estimate,

$$f_n(x) = (nh(n)^p)^{-1} \sum_{j=1}^n K(h(n)^{-1}(x^{(j)} - x)).$$

Then $\nabla f(x_n^*) \rightarrow 0$ in probability as $n \rightarrow \infty$.

Proof. Since f_n is the constant multiple of a finite sum of evaluations of K , f_n inherits the differentiability of K and ∇f_n the Lipschitz continuity of ∇K . Due to the globalizing measures in Algorithm 4.1.1, $f_n(x_{k+1}) > f_n(x_k)$ and conditions (4.2.1) are satisfied for all k ; thus by Proposition 4.2.1, the quasi-Newton algorithm produces a point x_n^* at which $\nabla f_n(x_n^*) = 0$. The rate of decrease in the sequence $\{h(n)\}$ satisfies both $h(n) = O(n^{(s-1)/p})$, $1/2 < s < 1$, and $nh(n)^{3p} \rightarrow \infty$. By the first condition, Proposition 2.3.2 holds, establishing uniform strong consistency of f_n . The second condition guarantees consistency in quadratic mean of ∇f_n as an estimator of ∇f (cf., equation (2.2.19)). Thus Proposition 2.3.5 holds and provides that $\nabla f(x_n^*) \rightarrow 0$ in probability as $n \rightarrow \infty$.

4.3. Implementation Notes

The stochastic nature of the function evaluations and

the difficulty of assuring matched scales for function and derivative estimates led to some special considerations in the implementation of Algorithm 4.1.1, and these are described briefly here.

First, the magnitude of the density estimate in the neighborhood of the mode is unpredictable a priori, but in high dimensions will be very small unless a considerable amount of correlation is present. For mode-finding the estimates of f need only be consistently scaled, not true in magnitude. The easiest way to change the scale of f_n or ∇f_n is to multiply the kernel function by a constant factor. Since the mean update was used to provide an initial guess for the Newton procedure, the procedure begins with access to an easily evaluated preliminary density estimate, and can scale to give this estimate a numerically desirable value, such as 1.0.

Also, the smoothing parameters $h_i(n)$ in the estimator f_n (2.2.16) may be adapted from the radius of the final neighbor set of the mean update. The choice of smoothing parameters is extremely important but difficult to quantify. Reasons for making the choice based upon nearest neighbor distances such as those obtained from the mean update are presented in Section 2.2.4.

Theoretical principles for coordinating the smoothing parameters for function and derivative estimates are not complete. Analysis of mean squared error culminating in

expression (2.2.24) suggests that $h(n) = O(n^{-1/p+4})$ for estimating f , $h(n) = O(n^{-1/p+6})$ for gradient estimates, and $h(n) = O(n^{-1/p+8})$ for estimates of second partials, but in practice these asymptotic rules are not very instructive. Certain practical guidelines can be obtained by observing the analogy between kernel functions K_1 and K_2 of expression (4.1.1) and central finite difference approximations of derivatives. If a function f is twice continuously differentiable and $\nabla^2 f$ is Lipschitz continuous with constant γ , and if finite difference quotients for first and second partial derivatives are

$$g_i(x) = [f(x+\delta e_i) - f(x-\delta e_i)]/2\delta \quad (4.2.3)$$

and

$$a_{ij}(x) = [f(x+\delta(e_i+e_j)) - f(x+\delta(e_i-e_j)) - f(x-\delta(e_i-e_j)) + f(x-\delta(e_i+e_j))]/4\delta^2,$$

then it is straightforward to show that in exact arithmetic,

$$|g_i(x) - \frac{\partial}{\partial x_i} f(x)| \leq \gamma(2\delta)^2/24$$

and

$$|a_{ij} - \nabla^2 f(x)_{ij}| \leq \gamma(2\delta)(5/3).$$

If ϵ represents machine precision and η is an upper bound on the relative error in evaluation of f , then a bound on the total error in g_i is,

$$e(\delta) = \frac{2(\eta+\epsilon)f(x)}{2\delta} + \frac{\gamma(2\delta)^2}{24},$$

minimized by

$$(2\delta_1)^3 = 24\left(\frac{f(x)}{\gamma}\right)(\eta+\epsilon). \quad (4.2.4)$$

Similarly, a bound on the total error in a_{ij} is

$$e(\delta) = \frac{4(\eta+\epsilon)f(x)}{(2\delta)^2} + (5/3)\gamma(2\delta),$$

minimized when

$$(2\delta_2)^3 = \left(\frac{24}{5}\right)\left(\frac{f(x)}{\gamma}\right)(\eta+\epsilon).$$

Now $2\delta_2/2\delta_1 = 5^{-1/3} = 0.58$; that is, the distance between function evaluations in the Hessian finite difference quotients should be about .58 times the separation employed in first-order difference quotients. Taking the distance from peak to valley in K_1 and K_2 to measure the distance between function evaluations, the ratio of these distances for K_2 versus K_1 is almost exactly .58. From the point of view of finite difference approximation, then, K_1 and K_2 stand in optimal relation to one another as is; thus, it is reasonable to use the same smoothing parameter for both gradient and Hessian estimates.

Relating smoothing parameters for estimation of f and ∇f is a bit more difficult. Substituting expression (2.2.24) for optimal $h(n)$ into (2.2.5) and collecting terms, the mean squared error in $f_n(x)$ can be expressed as

$$\text{MSE}(f_n(x)) = (1 + \frac{1}{p}) [\text{trace}(\nabla^2 f(x))]^2 h(n)^4.$$

If a typical absolute error is of the order $h(n)^3$, and this is hypothetical but not unreasonable if $h(n)$ is near unity or smaller, then the relative error may be reasonably taken as $\eta = \omega h(n)^3$, where

$$\omega = (1 + \frac{1}{p}) \left[\frac{\text{trace}(\nabla^2 f(x))}{f(x)^{1/2}} \right]^2.$$

In this case, by (4.2.4), assuming $\eta \gg \epsilon$,

$$2\delta_1 = \omega' h(n), \quad (4.2.5)$$

where

$$\omega' = 2 \cdot 3^{1/3} \text{trace}(\nabla^2 f(x))^{2/3} \gamma^{-1/3}.$$

In other words, (4.2.4) suggests that the distance between the two function evaluations be chosen proportional to the smoothing parameter used for estimates of the density function. Since ω' cannot be estimated, the constant of proportionality is an open choice, and in our implementation it was taken as unity.

Though an effort has been made to choose smoothing factors with good theoretical properties, it is clear that optimal choices depend upon characteristics of the population density which will be inestimable in practice. Poorly matched smoothing factors for first and second order derivative information are likely to distort the magnitude of the

Newton step s_k , even if the direction of the step is appropriate. It has already been mentioned that a k -nearest neighbor density assessment provides information about the magnitude of f as well as offering smoothing parameters for the kernel estimators. It also proved advantageous to employ the k -neighbor set as a "trust region" for the length of step to be taken. Letting r_k be the distance to the k -th neighbor, and for two constants $0 < c_1 < c_2$, the Newton step is expanded or contracted if necessary to satisfy $c_1 r_k \leq \|s_k\| \leq c_2 r_k$. The upper bound $c_2 r_k$ constrains the algorithm from extrapolating too far a model based only upon the local sample contained within the radius r_k . The lower bound requires the algorithm to first attempt a fairly long Newton step. This requirement is motivated partially by the fact that standard optimization implementations usually benefit by taking a full Newton step whenever possible. In addition, early experience with the Newton implementation revealed that the impact of local oscillation in kernel (and other) density estimators is mitigated by taking the longest step possible consistent with condition (4.2.1).

Backtracking was conducted by estimating the density at a lattice of points distributed at regular intervals along the step direction. Density estimates on any pre-specified grid may be performed in parallel with one pass through the data. When the data set may be stored in high-

speed memory, this parallelism is not especially significant; however, when the data must be accessed from secondary store, so that input/output dominates time requirements, the savings from reducing data access are considerable.

Some testing was conducted with an alternate method of linesearch intended to improve the accuracy and robustness of the local quadratic model, at least along the one-dimensional path chosen for the step direction. The method was to fit a quadratic to the lattice of function evaluations in the backtrack grid, observing that consecutive error terms in the regression model are likely to be significantly correlated. The model has the form

$$y_i = f_n(\lambda_i) = q_2 \lambda_i^2 + q_1 \lambda_i + q_0 + \epsilon_i,$$

where λ_i is the displacement of the i -th evaluation along the step direction. If the error terms have variance-covariance matrix Σ , or correlation matrix R , the least square solution for the vector of quadratic coefficients is

$$q = (\Lambda^T R^{-1} \Lambda)^{-1} (\Lambda^T R^{-1}) y,$$

where

$$y = (y_1, \dots, y_m)^T,$$

$$\Lambda = \begin{pmatrix} \lambda_1^2 & \lambda_1 & 1 \\ \vdots & \vdots & \vdots \\ \lambda_m^2 & \lambda_m & 1 \end{pmatrix}$$

and m is the number of evaluations. An estimate of R was obtained by modeling the sequence of function values as a first order autoregressive process, in which $\hat{R}_{ij} = \hat{\rho}^{|i-j|}$, where $\hat{\rho}$ is the estimated correlation coefficient of lag 1. Once the coefficients q were obtained the next iterate would be taken as the minimizer of the associated quadratic, subject to the upper step length bound obtained from the local neighbor set.

An additional modification that was investigated was to adjust the smoothing parameter h to the minimal value for which the sequence of function values $f_n(\lambda_i)$ would be monotone or monotone on either side of a contained minimum, after a median smoothing. The rationale for this step was the observation of Chapter II that good smoothing parameters typically yield density estimates just on the verge of displaying pronounced local oscillation. The smoothing parameter chosen as above is the least (and f_n the least biased estimate) consistent with identifying a single mode in the vicinity of the current neighbor set. The notion is similar to one discussed by Silverman [1981] in exploring multimodality with Gaussian kernels.

The method of quadratic fits showed some promise in resisting the noisiness of erratic density estimates, limiting the influence of any compact clump of observations from becoming too great. Fits using the autoregressive model for estimating correlation in most cases conformed

well to the shape of the population density, reducing the effect of points at the boundaries of the sequence $\{\lambda_i\}$, which often distorted standard least square fits. Nevertheless, the fitted quadratic continued to be strongly influenced by the width of the window containing the lattice of function evaluations, the location of the minimizer being particularly effected. After some exploration of techniques for selecting a window width for the lattice of evaluations, it was concluded that the method of quadratic fits is too difficult to control in practice, while the backtracking method is effective and more stable. Further exploration of linesearch and smoothing strategies was rejected because it became apparent that the fundamental limitation of the Newton's method procedure was not in the conduct of the linesearch but in the restriction of the search for a better iterate to a specific one-dimensional path.

Finally, because of the uncertain scale in gradient values and the problem of small scale oscillation in the density estimate, the use of gradient values as a stopping criterion proved impractical, and instead the stopping criterion for the algorithm was based on the convergence of successive iterates. A queue was maintained holding the most recent six iterates, and from these iterates a "running mean" was calculated. At the same time, a cumulative mean was maintained of all iterates, beginning with the initial guess. The algorithm was said to have converged if the

relative change in either of these means fell below a specified tolerance, with the relative change from the current mean μ^c to the updated mean μ^+ defined as

$$\Delta = \max_{1 \leq i \leq p} \{ |\mu_i^+ - \mu_i^c| / \max(|\mu_i^c|, 1.0) \}.$$

In addition, the algorithm halted if the linesearch routine was unable to locate a point with lower function value than the current iterate. In fact this latter condition was the most frequent cause of algorithm termination, reflecting perhaps the fact that kernel K_0 is very nearly quadratic except at the periphery of its support, so that jumps directly to a local minimizer are common, but also suggesting again that reliance on a single step direction handicaps the algorithm in searching for a decrease in function value.

4.4 Empirical Results

The Newton's method second phase was tested on data simulated from three unimodal distributions, an uncorrelated Gaussian (GAUSS), a Gaussian with all correlations equal to 0.9 (R9GAUSS), and a multivariate Cauchy (CAUCHY). The distributions and their simulation are discussed in Section 3.2.

Three methods of handling second-derivative information were employed. The first was secant approximation of the Hessian by the method (BFGS) of rank 2 updates. A good reference on the secant method is Dennis and Schnabel [1983].

The second method utilized kernel estimators for second-order partial derivatives, as described earlier in this chapter. The third method simply took the Hessian as the identity matrix, in effect producing a steepest descent technique. The reason for considering three methods was simply to gain an empirical assessment of the quality of the second-derivative information available to the Newton's procedure. Secant approximations of the Hessian are generally preferable to finite difference approximations when the function evaluations are noisy, and thus there was some question as to whether kernel estimation of second derivatives should be avoided. The steepest descent technique was included as a touchstone to determine whether either Hessian approximation yielded any gain in the performance of the algorithm.

Summaries of average MSE results obtained from 25 trials of the mean update and Newton algorithms are given in Tables 4.4.1, 4.4.2, and 4.4.3. Each table includes results of tests in dimensions $P = 1, 2, 3, 4, 5, 10, 15$, and 20 , utilizing four neighbor set sizes. The sample size in all cases was $N = 100$. Table 4.4.1 corresponds to the uncorrelated Gaussian, Table 4.4.2 to independent Cauchy variates, and Table 4.4.3 to the highly correlated Gaussian. Each table gives the average MSE for the mean update phase in absolute form, and then expresses average values for the Newton phase as a fraction of the mean update values.

TABLE 4.4.1. GAUSS Average MSE, 25 Trials.

	Mean Update	Ratio to Mean Update		
		Secant	Kernel Hessian	Steepest Descent
P = 1				
K = 10	2.217	.526	.599	.530
20	.9732	.525	.470	.547
30	.3899	.623	.666	.614
40	.1792	.704	.753	1.376
P = 1				
K = 10	.8268	.693	.422	.720
20	.3383	.455	.384	.461
30	.1546	.644	.628	.730
40	.08756	.660	.660	.756
P = 3				
K = 10	.4727	.677	.710	.698
20	.1720	.895	.767	.769
30	.1024	.785	.760	.743
40	.06319	.648	.647	.695
P = 4				
K = 10	.2811	.758	.688	.718
20	.1364	.776	.738	.628
30	.06776	1.106	.811	.843
40	.05737	.868	.875	.843
P = 5				
K = 10	.2476	.855	.792	.686
20	.1140	.970	.830	.781
30	.07004	.909	.949	.908
40	.04531	.901	.979	.910
P = 10				
K = 10	.1437	.657	1.000	.971
20	.08638	.558	1.000	.967
30	.05477	.746	1.000	.997
40	.03850	.948	1.000	.999
P = 15				
K = 10	.1608	1.000	1.000	1.000
20	.07246	1.000	1.000	1.000
30	.04393	1.000	1.000	1.000
40	.03302	1.000	1.000	1.000

TABLE 4.4.1. Continued.

	Mean Update	Ratio to Mean Update		
		Secant	Kernel Hessian	Steepest Descent
P = 20				
K = 10	.1371	1.000	1.000	1.000
20	.07116	1.000	1.000	1.000
30	.04390	1.000	1.000	1.000
40	.03302	1.000	1.000	1.000

TABLE 4.4.2. CAUCHY Average MSE, 25 Trials.

	Mean Update	Ratio to Mean Update		
		Secant	Kernel Hessian	Steepest Descent
P = 1				
K = 10	13.43	.904	.0054	.931
20	.9858	.839	.312	.844
30	.1093	.559	.592	.486
40	.0471	.717	1.156	4.576
P = 2				
K = 10	1.559	.806	.169	.871
20	.0914	1.141	1.595	.876
30	.0663	.937	1.185	.859
40	.0545	.994	1.028	.994
P = 3				
K = 10	.6860	.703	.387	.785
20	.0644	1.491	1.100	1.032
30	.0601	1.090	.975	.999
40	.0625	.988	.940	.988
P = 4				
K = 10	1.054	.966	.693	1.010
20	.0815	1.021	.980	.991
30	.0779	.939	.941	.999
40	.0670	.967	1.000	1.000
P = 5				
K = 10	.4864	.797	.599	.995
20	.1027	1.043	1.017	.996
30	.0660	1.004	1.000	1.000
40	.0520	1.000	1.000	1.000
P = 10				
K = 10	.2574	1.000	1.000	1.000
20	.1526	1.000	1.000	1.000
30	.1253	1.000	1.000	1.000
40	.1316	1.000	1.000	1.000
P = 15				
K = 10	.3059	1.000	1.000	1.000
20	.2029	1.000	1.000	1.000
30	.1764	1.000	1.000	1.000
40	.1663	1.000	1.000	1.000

TABLE 4.4.2. Continued.

	Mean Update	Ratio to Mean Update		
		Secant	Kernel Hessian	Steepest Descent
P = 20				
K = 10	.3822	1.000	1.000	1.000
20	.2500	1.000	1.000	1.000
30	.2070	1.000	1.000	1.000
40	.2229	1.000	1.000	1.000

TABLE 4.4.3. R9GAUSS Average MSE, 25 Trials.

	Mean Update	Ratio to Mean Update		
		Secant	Kernel Hessian	Steepest Descent
P = 1				
K = 10	1.973	.599	.531	.662
20	.873	.504	.544	.549
30	.440	.484	.791	.655
40	.206	.603	.636	1.719
P = 2				
K = 10	2.015	.541	.476	.739
20	.823	.633	.562	.642
30	.391	.809	.799	.810
40	.222	.753	.775	.670
P = 3				
K = 10	1.913	.557	.872	1.085
20	.702	.540	.464	.499
30	.259	.680	.739	.626
40	.134	.680	1.246	.636
P = 4				
K = 10	1.975	.692	.658	.674
20	.680	.528	.699	.574
30	.321	.736	.917	.592
40	.135	.697	.947	.672
P = 5				
K = 10	1.897	.722	1.017	.785
20	.731	.702	.877	.687
30	.366	.731	.983	.772
40	.249	1.399	.992	.877
P = 10				
K = 10	1.940	.691	1.000	.804
20	.893	.695	1.000	.817
30	.435	.815	1.000	1.077
40	.288	1.490	1.000	1.052
P = 15				
K = 10	1.758	.742	1.000	1.099
20	.844	.790	1.000	.868
30	.348	.743	1.000	1.175
40	.201	.940	1.000	.981

TABLE 4.4.3. Continued.

	Mean Update	Ratio to Mean Update		
		Secant	Kernel Hessian	Steepest Descent
P = 20				
K = 10	1.942	.835	1.000	1.097
20	.843	.894	1.000	1.066
30	.389	.791	1.000	1.074
40	.134	.996	1.000	1.140

A subset of the tabular information is also presented in graphical form in Figures 4.4.1 through 4.4.4. Figure 4.4.1 displays the MSE results for the GAUSS distribution, for neighbor sets of size $K = 10$ and $K = 40$ only. Figure 4.4.2 presents the analogous information based upon the CAUCHY data. Figures 4.4.3 and 4.4.4 correspond to the distribution R9GAUSS and together give the full range of neighbor set sizes. The figures have an orientation that is transposed from that of previous graphical displays; dimension increases with vertical movement down the plot, while the error measure increases with movement from left to right across a line.

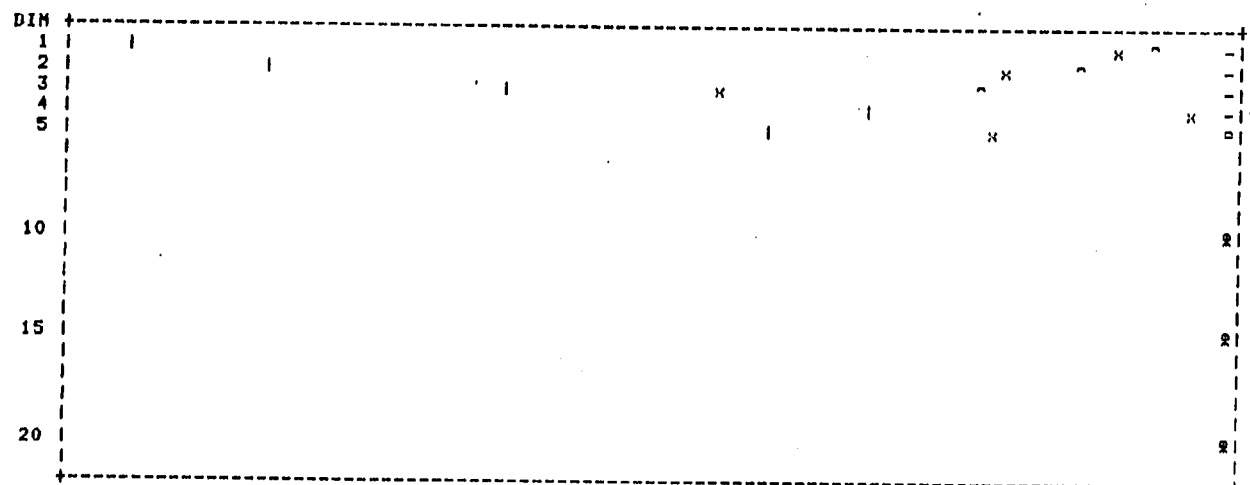
Some basic conclusions can be drawn from the summary information. The Newton's method second phase has a clearly beneficial impact on the performance of the mode-seeking procedure. The impact is strongest in low dimensions and in cases where the neighbor set size is small. As regards the efficacy of second-order information, the results are mixed. In dimensions 5 and below there is no strongly prevailing pattern. The use of kernel estimated Hessians is extremely successful on Cauchy data with $K = 10$ and overall appears the most successful of the three procedures in dimensions 1 and 2. In dimensions 3 through 5, particularly with the correlated Gaussian data the secant method often gives the best results. But there is nothing very conclusive in these dimensions; in fact, taken as a whole neither

FIGURE 4.4.2

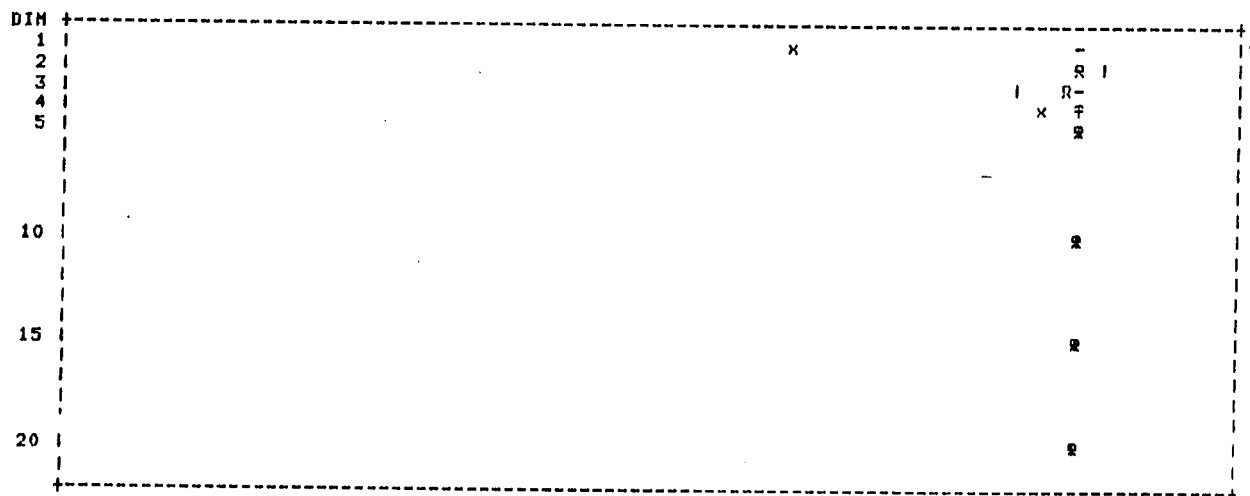
RATIO, MSE OF NEWTON ALGORITHMS TO MEAN UPDATE

CAUCHY

Neighbor set size = 10



Neighbor set size = 40



Key to methods used:

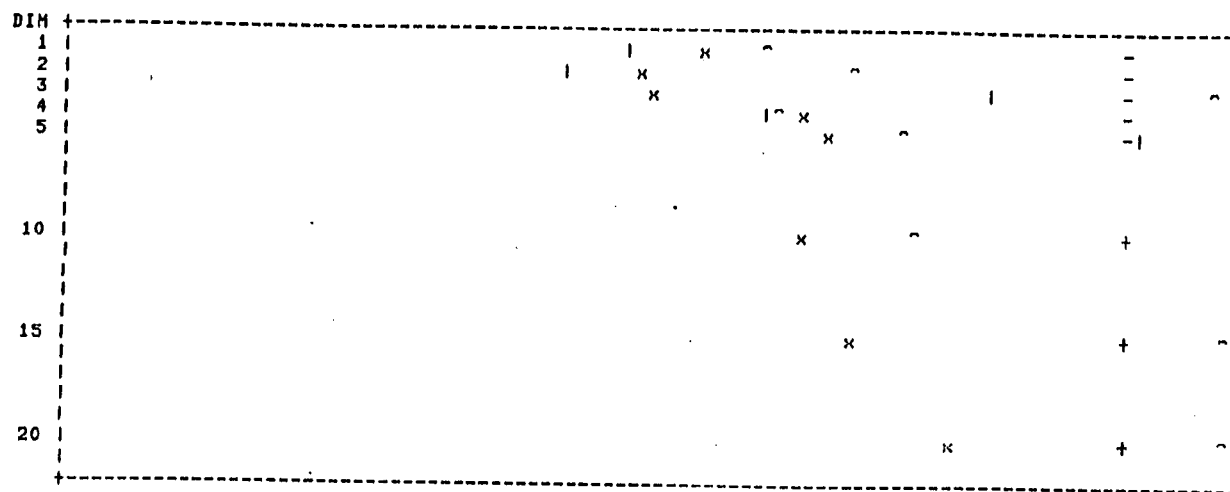
(-)	mean update	
(x)	newton-like, nnfixh(no reeval)	=secant=
(l)	newton-like, nnfixh(no reeval)	=Hessian=
(~)	newton-like, nnfixh(no reeval)	=steep=

FIGURE 4.4.3

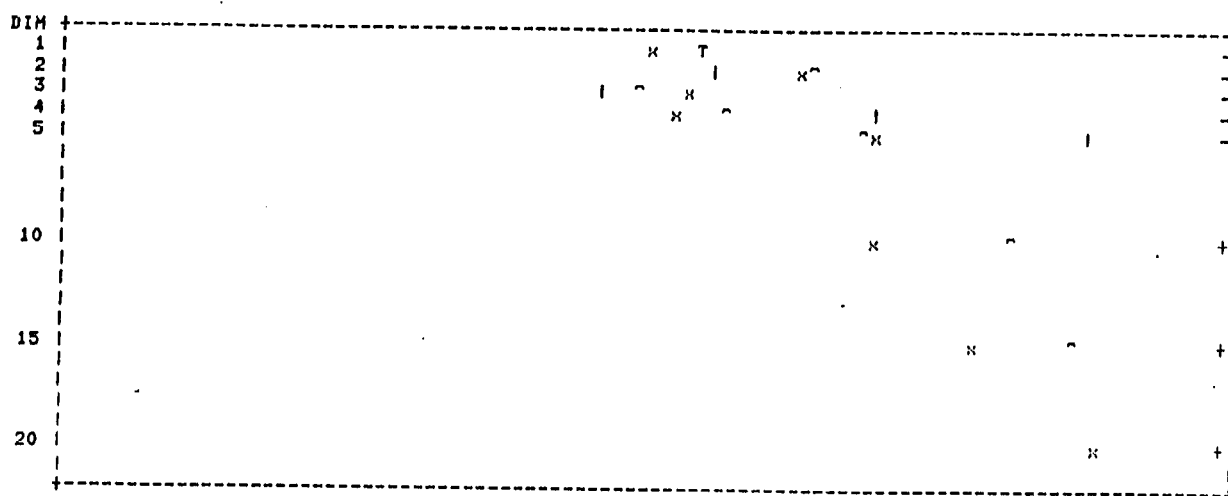
RATIO, MSE OF NEWTON ALGORITHMS TO MEAN UPDATE

R7GAUSS

Neighbor set size = 10



Neighbor set size = 20



Key to methods used:

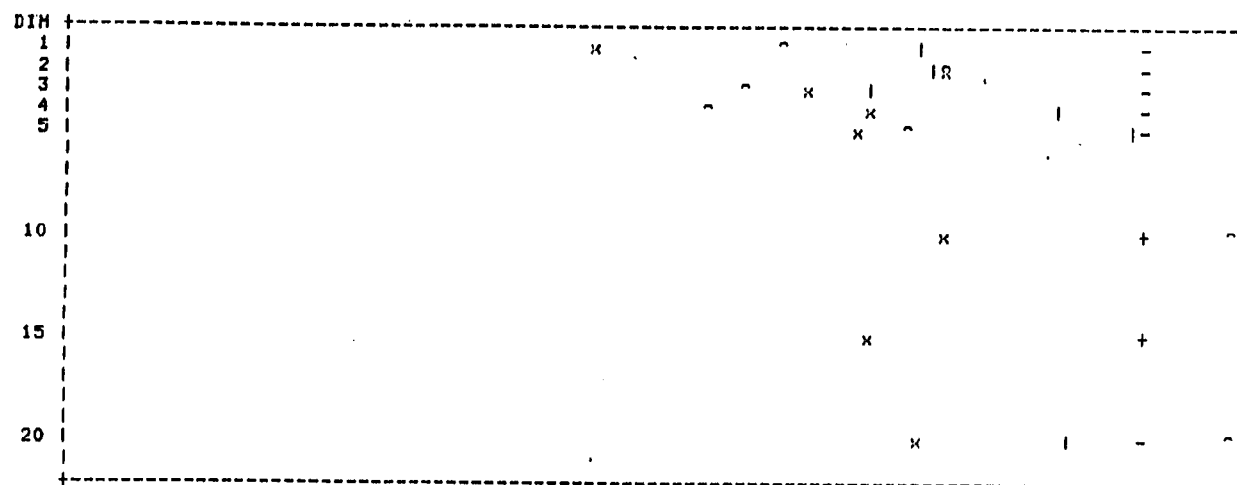
(-) mean update
 (x) newton-like, nnfixh(no reeval) =secant=
 (I) newton-like, nnfixh(no reeval) =Hessian=
 (T) newton-like, nnfixh(no reeval) =steep=

FIGURE 4.4.4

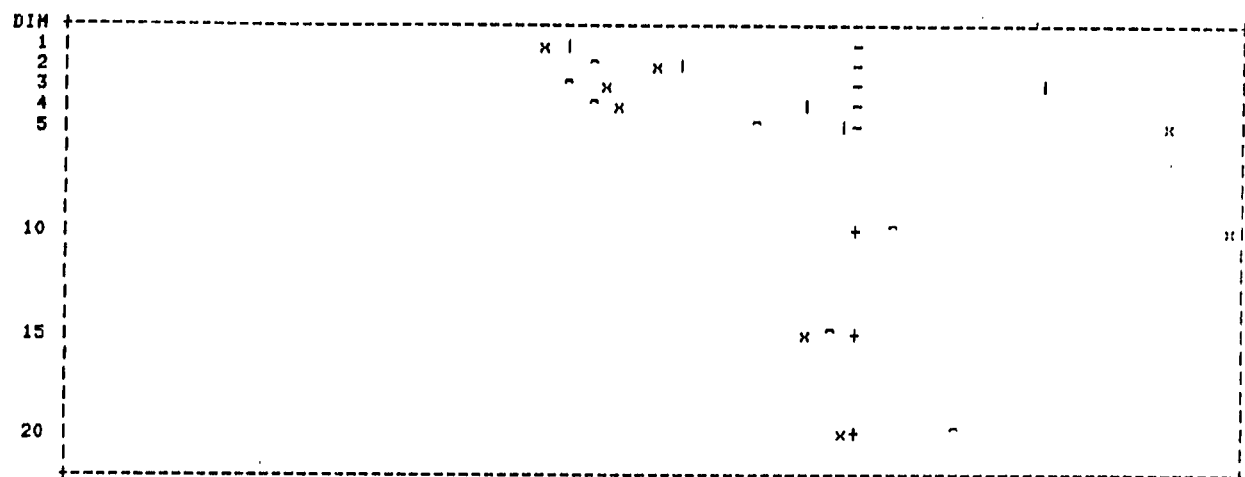
RATIO, MSE OF NEWTON ALGORITHMS TO MEAN UPDATE

R9GAUSS

Neighbor set size = 30



Neighbor set size = 40



Key to methods used:

(-)	mean update	
(x)	newton-like, nnfixh(no reeval)	=secant=
(l)	newton-like, nnfixh(no reeval)	=Hessian=
(^)	newton-like, nnfixh(no reeval)	=steep=

of the two Hessian approximations has a meaningful advantage over the steepest descent. The two small tables below give the number of (dimension, neighbor set) pairs for which the method labelling a row had lower average error than the method heading a column. In the tables "S" indicates the secant method, "K" indicates kernel estimated Hessians, and "D" stands for steepest descent.

TABLE 4.4.4.

GAUSS			
	S	K	D
S		7	9
K	12		10
D	10	10	

R9GAUSS			
	S	K	D
S		13	9
K	7		9
D	10	10	

For both GAUSS and R9GAUSS distributions, the steepest descent method outperforms the secant method ten times, and is outperformed nine. All the other pairwise comparisons show even splits as well. Thus, in the lower dimensional range, at least with the current methods of obtaining second-order information, and sample size $N = 100$, the quadratic estimates have an unpredictable impact on the performance of the search algorithm, and in most cases that impact is only marginally salutary.

With increasing dimension, however, the secant approximations emerge as being decidedly superior to the other techniques. This is particularly true in the case of the

highly correlated data, where the secant method continues to provide a noticeable, albeit decreasing improvement on the mean update through dimension 20.

To get a more detailed understanding of the operation of the Newton's algorithm we must look at the trials individually. Figure 4.4.5 gives a variant of stem-and-leaf plots of MSE values of 25 trials of the mean update and secant method second phase, applied to independent Gaussian data. The plots for the two methods are paired and placed back-to-back, with univariate results on the left and results in dimension 10 on the right. Neighbor sets of size 20 were used in both cases. The individual trials may be identified from their alphabetic symbols; the first trial is marked "A", the twenty-fifth "Y".

The univariate results indicate how much more precise and rapid a Newton's method procedure can be compared to the mean update. Trials A, M, C, H, N, K, and B, all with high errors, are brought to an accuracy matching or exceeding the best results of the mean update. Trial A, for example, drops from a squared error of more than 1.5 to one of less than 0.1 after relocation by the Newton phase.

At the same time, a number of poorly positioned points (G, F, D, R, V) are untouched by the secant method. No positions, however, are significantly worsened by the second phase. This is not surprising in one dimension, since first derivative estimates need only the correct algebraic sign to

(a)		GAUSS MSE 25 TRIALS		(b)	
P = 1		K = 20		P = 10	
G	G				
				J	
				O N	
M					
C					
A					J
F D	D F			E	
H				Q	Q
N K					
R B	R			K	
V	V			T S R	
				X	E N
Y O	O		Y W V U F A		
W P L J I	I J		H G	G K O	
	C P		P M	R U	
X U			D C	C W	
T S Q E	E M Q S W		L I B	A B T V X	
				G M P Y	
	Y			D H I L S	
	A B H K L N T U X				
Mean Update		Secant Method		Mean Update	
				Secant Method	

Figure 4.4.5.

initiate movement in the right direction, and stopping points for the univariate mean update are typically amid clusters in the tail portions of the data set, so that search away from the mode is likely only to bring a quick halt. Unfortunately, the same "conservatism" cannot always be relied upon in multiple dimensions.

The ten-dimensional results show a very favorable pattern of performance for the secant method and indicate considerable promise for a Newton-like algorithm in high dimension, especially if previously discussed limitations of our procedure can be removed. Individual relocations with $P = 10$ are not as dramatic as with $P = 1$, but we should keep in mind that the scales of the two plots in Figure 4.4.5 are not comparable; rather, average MSE values with $P = 10$ fall below the univariate average by almost an order of magnitude. Thus, the Newton's phase begins with quality initial guesses, and it provides an improvement in the accuracy of the initial guess consistently throughout the trials. Only trials Q, C, and B fail to benefit from the application of the second phase. The improvement in relatively poor initial estimates (O, N) and relatively good estimates (D, L, I) are alike emphatic.

A more mixed performance is exhibited in Figure 4.4.6, which again gives back-to-back plots for the mean update and secant method, now observed on the distribution R9GAUSS, in dimension 10 and with neighbor sets of size 20.

R9GAUSS

(a)		(b)	
P = 10	K = 20	P = 10	K = 20
D	D	D	D
	L		L
R		R	
W	R W	W	R W
U O J I	U	U O J I	U
M		M	
L	I J O	L	I J O
X T G C		X T G C	
P	G X	P	G X
V N K		V N K	
F	K N P	F	K N P
Q H		Q H	
Y S E	Q T V Y	Y S E	Q T V Y
A		A	
	A		A
	C E H M S		C E H M S
B	B F	B	B F
Significant Change		4 Iterations or more	

Figure 4.4.6.

In the figure a) the trials which undergo significant change as a result of the Newton procedure are boxed. Clearly the impact of the second phase is not as uniform as in Figure 4.4.5b. In particular, significant change occurs only for the mean update estimates which are already good or moderately good, leaving the positions which most need correction essentially as is.

As mentioned before, the Newton's procedure usually terminated not by the designed stopping criterion, which tested for accumulation of the sequence of iterates, but because the linesearch procedure, backtracking along the specified step direction, failed to locate a better function value. In many cases, including most of the poor initial estimates in Figure 4.4.6, this would happen on the first or second iterate. In Figure 4.4.6b the same plot as 4.4.6a is shown but with those trials boxed which completed at least four iterations of the Newton's procedure before terminating. Of the ten trials so identified, six showed marked improvement as pictured in Figure 4.4.6a, the accuracy of the remaining four either improved slightly or remained unchanged. Conversely, of the eight trials showing significant improvement during the Newton phase, only two (V, E) obtained the improvement in three iterations or less. The obvious conclusion is that the algorithm must be kept active casting out searches. The more it searches (more iterations), the better the expected results.

This conclusion seems to be contradicted by the results in Figure 4.4.7, which is identical in design to Figure 4.4.6, except that the algorithm was run with neighbor sets of size $K = 40$. There again the trials experiencing significant change in accuracy during the Newton phase, and those having four or more iterations within that phase, essentially coincide. Unfortunately, in this case all but one of the significant location changes are detrimental. Thus, it appears that the algorithm must guard against too much movement, and of course this is true in principle; however, the necessary controls have already been provided by conditions (4.2.1). In fact, the degraded accuracy illustrated in Figure 4.4.7 as well as the "missed opportunities" of Figure 4.4.6 can be explained by immobility, or lack of search activity, at the heart of Algorithm 4.2.1.

The situation is similar to the problem of the "resolution ridge" considered by Wilde [1964] and Brent [1973] in minimization without derivative values. A schematic illustration of the problem is given in Figure 4.4.8, which depicts typical level contours for a quadratic (or Gaussian) surface. Proceeding from the point (a), movement along the path 1, which would be indicated by accurate gradient information, leads to the optimizer of the function. However, with relatively minor deviation from that search direction, as in path 2, the search fails to detect any function value better than $f(a)$. One can imagine the

R9GAUSS

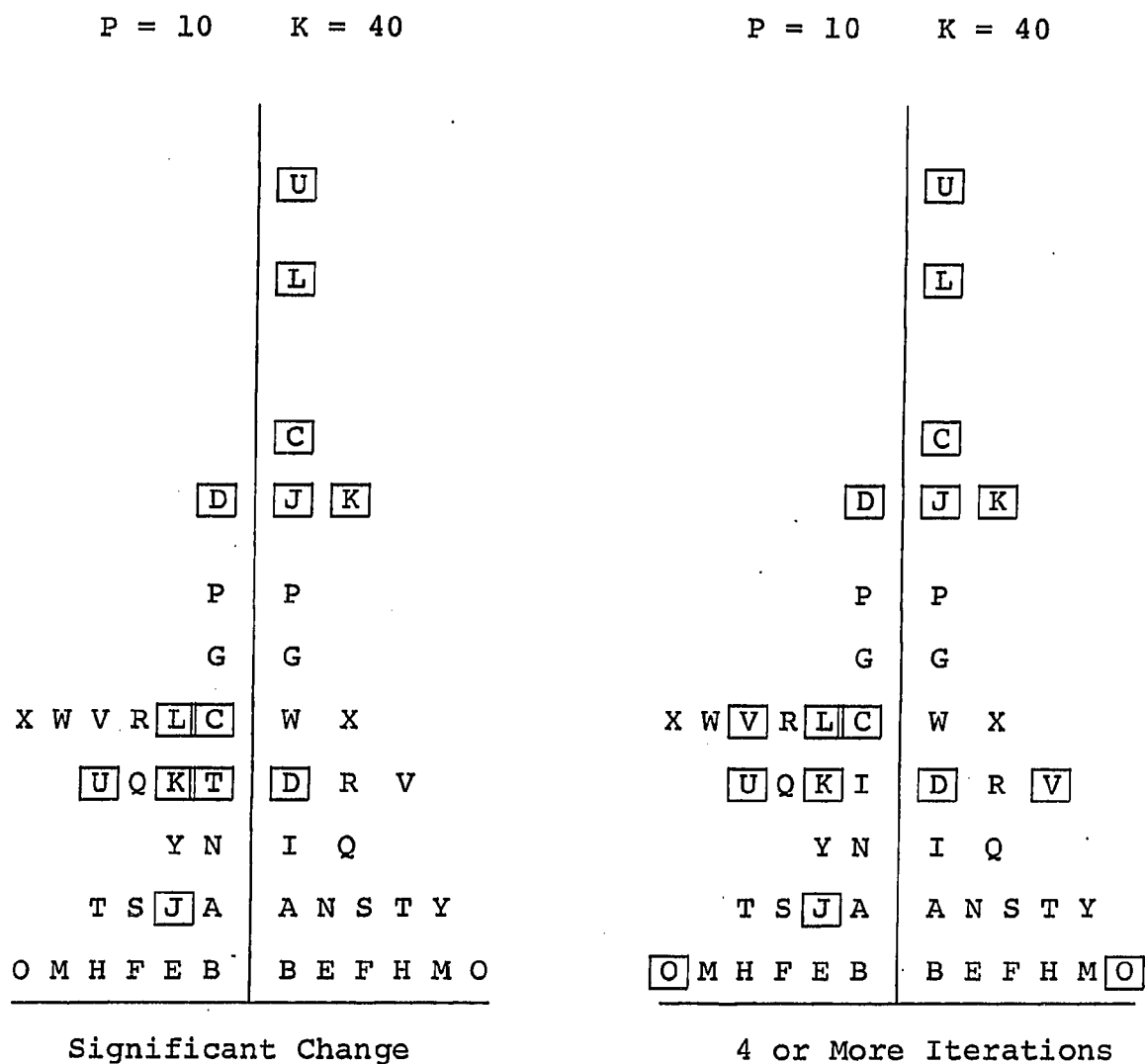
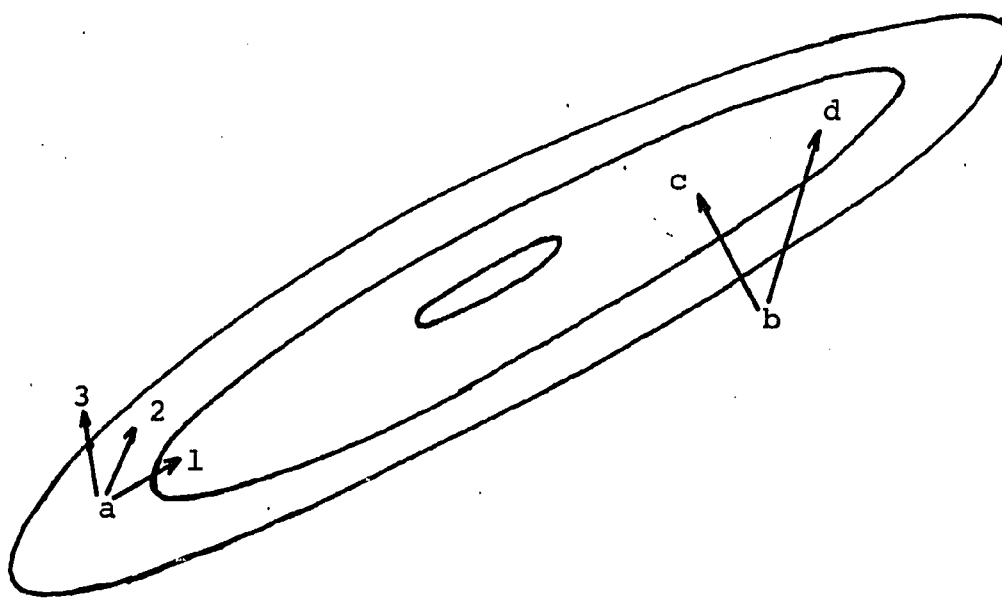


Figure 4.4.7.

Figure 4.4.8.



cone of ascent directions becoming more and more acute in higher dimensions. A point (b) with reasonably accurate gradient information will move quickly to the ridge (e.g., location (c)). Degradation can occur with rough gradient estimates, where the algorithm takes a step away from the mode ((b) to (d)), and then is unable to identify an ascent direction to recover.

Concluding Remarks

We have paid due to intrinsic difficulties in the application of Newton-like optimization techniques to the problem of finding modes of a probability density function. Nevertheless, the widespread and occasionally outstanding improvements that we have obtained upon the mean update by making largely cosmetic modifications to a conceptually deterministic technique are enough to suggest to us that further work on joining nonparametric density estimation and Newton-like optimization is justified. The first effort at repairing existing flaws must go toward replacing the linesearch step with a search strategy that will perhaps favor an indicated direction but will resort to full spatial sweep when necessary. The ability of kernel estimation procedures to evaluate a preselected lattice of points in parallel will probably feature in such a strategy.

When we say "Newton-like" we mean that the algorithm will make use of local quadratic (conceivably, other parametric) models of the density. A second topic for investi-

gation is the construction of models considering the stochastic nature of function and gradient evaluations upon which they must be based. Kernel-estimated Hessians, of course, are one such device, but beyond dimension two, they have not proved useful. Secant approximations, while their influence has been limited in most of our simulations, have had some success in high dimensions. It is likely that if the search strategy is improved so that more observations are made available to feed the approximations, the influence of the secant method will likewise increase. Other constrained or adaptive fitting techniques may also be worth considering.

V. WEIGHTED MEAN UPDATE

5.1. Derivation of the Procedure and Consistency Results

A multivariate kernel density estimate with product kernel has the form

$$\hat{f}_n(x;h) = \frac{c}{nh^p} \sum_{i=1}^n \left[\prod_{j=1}^p K\left(\frac{x_j^{(i)} - x_j}{h}\right) \right]. \quad (5.1.1)$$

A necessary condition for \hat{f}_n to have a mode at x , of course, is that $\nabla \hat{f}_n(x) = 0$. The k -th partial derivative of (5.1.1) is given by

$$(\partial/\partial x_k) \hat{f}_n(x) = - \frac{c}{nh^{p+1}} \sum_{i=1}^n \left[K'\left(\frac{x_k^{(i)} - x_k}{h}\right) \prod_{\substack{j=1 \\ j \neq k}}^p K\left(\frac{x_j^{(i)} - x_j}{h}\right) \right] \quad (5.1.2a)$$

In general (5.1.2) is difficult to summarize analytically. However, if we use a univariate Gaussian kernel,

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} t^2},$$

Then since $K'(t) = -tK(t)$,

$$(\partial/\partial x_k) \hat{f}_n(x) = \frac{c}{nh^{p+2}} \sum_{i=1}^n \left[(x_k^{(i)} - x_k) \prod_{j=1}^p K\left(\frac{x_j^{(i)} - x_j}{h}\right) \right]. \quad (5.1.2b)$$

Now the kernel product

$$\pi_i(x;h) = \frac{c}{nh^p} \prod_{j=1}^p K\left(\frac{x_j^{(i)} - x_j}{h}\right)$$

depends upon the location x , the smoothing parameter h , and upon the i -th sample point, but it is independent of the

coordinate with respect to which differentiation is performed. Therefore, critical points of \hat{f}_n are those \tilde{x} for which

$$\tilde{x} = \frac{\sum_{i=1}^n \pi_i(\tilde{x};h) x^{(i)}}{\sum_{i=1}^n \pi_i(\tilde{x};h)} \quad (5.1.3)$$

or

$$\tilde{x} = \sum_{i=1}^n w_i(\tilde{x};h) x^{(i)}, \quad (5.1.4)$$

with

$$w_i(x;h) = \pi_i(x;h) / \left(\sum_{i=1}^n \pi_i(x;h) \right) = \pi_i(x;h) / \hat{f}_n(x;h)$$

being the relative contribution the sample point $x^{(i)}$ makes to the density estimate at x . Expressed in matrix form, (5.1.4) is

$$\tilde{x} = Xw(\tilde{x};h) \quad (5.1.5)$$

where

$$X = (x^{(1)}, \dots, x^{(n)})$$

and

$$w(x;h) = (w_1(x;h), \dots, w_n(x;h))^T.$$

Thus, critical points of \hat{f}_n are obtained as solutions to the non-linear system of equations,

$$\phi(x;h) = x - Xw(x;h). \quad (5.1.6)$$

In one sense, the formulation above differs little from the Newton-like approach of Chapter IV, merely reexpressing the optimization problem as a search for roots of the gradient, and of course solutions of (5.1.6) may be obtained from Newton or quasi-Newton equations solvers. However, (5.1.6) presents the optimization of (5.1.1) with Gaussian kernels in a format that reveals special structure which may be exploitable computationally, and which will inform the diagnosis and interpretation of computational results.

Specifically, (5.1.6) expresses any mode estimate as a weighted average of other members of the sample, and also demonstrates that the nonlinearity in the system to be solved is contained entirely in the evaluation of these weights. Were (5.1.6) instead

$$\phi(x) = y - Xw(x)$$

with y a fixed vector, then (5.1.6) could be rewritten

$$\phi(w) = y - Xw; \quad (5.1.7)$$

that is, as an underdetermined linear regression problem with the weights w_i being functions of the data matrix. It is likely that solution techniques for similar problems (e.g., iteratively re-weighted least squares) may be adapted for (5.1.6) with considerable computational savings over the methods of Chapter IV, particularly with large sample sizes. For example, X may be once decomposed into

triangular or orthogonal factors and stored in factored form for use in repeated least squares operations.

Generalized inverses for rank deficient least squares problems commonly choose w as the minimum norm element of the row space of X which satisfies $X^T X w = X^T y$; such a strategy is particularly appropriate for mode estimation in that weighting of sample observations is thereby constrained to remain as nearly uniform as possible. The norm of w at mode estimates is relevant to the interpretation of the estimates, since locations which weight a small fraction of observations heavily are more likely to be spurious, while weights which are extremely uniform may indicate over-smoothing. Thus, the weighting vector at solutions of (5.1.6) may be useful in deciding which mode estimates to accept and reject, or how many modes are present in all.

Rather than become involved in the implementation details of solution methods such as those discussed above, we employed an algorithmically simpler fixed point method based upon (5.1.4). A fixed bandwidth version of the procedure is given below:

Algorithm 5.1.1a

Assume as input an initial guess x_c and a smoothing parameter h ;

Repeat until (stopping criteria are met)

$$x_c = \sum_{i=1}^n w_i(x_c; h) x^{(i)} \quad (5.1.8)$$

end repeat
Return x_c ;

The weights are calculated as

$$w_i(x;h) = \pi_i(x;h) / \sum_{i=1}^n \pi_i(x;h) \quad (5.1.9)$$

$$\pi_i(x;h) = \exp\{-\frac{1}{2}(\|x^{(i)} - x\|/h)^2\}.$$

The stopping criteria for the iteration were the same as those described in the implementation notes of Section 4.2.

Motivated by Algorithm 3.1.1, and by the reformulation of (5.1.1) as a variable kernel density estimate, the algorithm as implemented and used for testing followed the update step (5.1.8) with calculations of a k-neighbor set. The k-th neighbor distance was then used as the smoothing parameter in the next evaluation of (5.1.8).

Though Monte Carlo simulations to be presented indicate that the algorithm based on Gaussian kernels is superior in most cases to the methods of previous chapters, the infinite support of the kernel does counteract the local concentration needed for mode estimation. The balance that the tails provide may be largely responsible for the relative success of the method. Nevertheless, one may wish to control the kernel support to resist having secondary modes swamped by the aggregate influence of a dominant mode.

Let r_T be a truncation radius, and k_T be the number of sample observations falling within the distance r_T of the current

iterate x . Let the data matrix be permuted so that $X = (x^{(1)}, \dots, x^{(n)})$ where $\|x^{(1)} - x\| \leq \|x^{(2)} - x\| \leq \dots \leq \|x^{(n)} - x\|$. Then truncation of the kernel can be approached in three equivalent ways.

- a) Set $w_{k_T+1} = w_{k_T+2} = \dots = w_n = 0$, and rescale so that $\|w\|_1 = 1$ in (5.1.8);
- b) Calculate the weights w_i (or estimate $f_n(x)$) using

$$K(z) = c' e^{-\frac{1}{2} \|z\|^2}, \quad \|z\| \leq r_T$$

$$K(z) = 0, \quad \text{otherwise;}$$

- c) Isolate the k_T -neighbor set of x and ignore the remaining observations.

Methods a) and b) eliminate the need for explicitly forming the k_T -neighbor set. At the same time, since the weights w_i are monotonically decreasing functions of the distances $\|x^{(i)} - x\|$, the weight vector w can identify the neighbor sets of every size, and thus method a) can be used to adapt the truncation according to the local distribution of neighbor distances.

Incorporating the options of (1) nearest-neighbor smoothing parameters, and (2) truncation of the kernel support, the complete weighted mean procedure is given as algorithm 5.1.1b.

Algorithm 5.1.1b

Assume as input an initial guess x_c

```

and a smoothing parameter h    /*.Not. option 1*/
and a neighbor set size k ;    /*option 1*/
Repeat until (stopping criteria are met)
  if (option 1) then    /*variable kernel*/
    determine the distance  $r_k$  of the k-th
    nearest neighbor to  $x_c$ ;
     $h = r_k$ ;
  end if
  Calculate the weight vector  $(w_i(x_c;h), i = 1, \dots, n)$ ;
  if (option 2) then    /*truncate support*/
    determine a minimum weight  $w_0$ ;
    set all weights less than  $w_0$  to zero;
    rescale so that  $\|w\|_1 = 1$ ;
  end if
  
$$x_c = \sum_{i=1}^n w_i(x_c;h) x^{(i)} ;$$

End repeat
Return  $x_c$ ;

```

In this study option 1 is always invoked; option 2 is in force only in a few instances in Chapter VI.

Lemma 5.1.1. Let x_c and x_+ be successive iterates of Algorithm 3.1.1a, and let $\hat{f}_n(x_c;h)$ and $\hat{f}_n(x_+;h)$ be evaluations of (5.1.1). Then either $x_+ = x_c$ or $\hat{f}_n(x_+;h) > \hat{f}_n(x_c;h)$. The proof of Lemma 3.1.1 relied on the fact that the local model obtained from the associated density estimate at x_c was quadratic. Here, of course, K is not even

concave. The technique of proof is similar to one used by Dutter [1975] in justifying the solution of multiparameter location or regression problems by modified weights [Huber, 1981, Chapter 7]. The idea is to construct a negative definite function which minorizes \hat{f}_n everywhere but coincides with it at x_c , and which is minimized by x_+ .

Proof of Lemma 5.1.1. Assume $x_+ = x_c$. The above conditions suggest that the minorant be tangent to \hat{f}_n at x_c . Holding h fixed,

$$\begin{aligned} \nabla \hat{f}_n(x;h) &= \frac{c'}{h^{p+2}} \sum_{i=1}^n (x^{(i)} - x) K\left(\frac{x^{(i)} - x}{h}\right) \\ &= \frac{c'}{h^2} \sum_{i=1}^n (x^{(i)} - x) K_i(x;h). \end{aligned}$$

Define

$$u_i(x;h) = v_i\left(\frac{x^{(i)} - x}{h}\right),$$

where

$$\begin{aligned} v_i(\tau) &= -\frac{1}{2} K_i(x_c;h) \|\tau\|^2 + K_i(x_c;h) \\ &\quad + \frac{1}{2} K_i(x_c;h) \left\| \frac{x^{(i)} - x_c}{h} \right\|^2 \\ &= K_i(x_c;h) \left\{ 1 - \frac{1}{2} \|\tau\|^2 + \frac{1}{2h^2} \|x^{(i)} - x_c\|^2 \right\}, \end{aligned} \tag{5.1.10}$$

and define

$$Q_c(x;h) = c' \sum_{i=1}^n u_i(x;h). \tag{5.1.11}$$

Then

$$\nabla Q_C(x;h) = \frac{c'}{h^2} \sum_{i=1}^n (x^{(i)} - x) K_i(x_C;h),$$

and clearly,

$$Q_C(x_C;h) = \hat{f}_n(x_C;h),$$

$$\nabla Q_C(x_C;h) = \nabla \hat{f}_n(x_C;h)$$

and

$$\nabla Q_C(x_+;h) = 0. \quad (5.1.12)$$

In addition, each $u_i(x;h)$ is a quadratic in x , and by expanding and summing over i one sees $Q_C(x;h)$ is a quadratic in x with

$$\nabla^2 Q_C(x;h) = -\frac{c'}{h^2} \hat{f}(x_C;h) I,$$

I being the appropriate identity matrix. Thus x_+ maximizes $Q_C(\cdot;h)$. In particular, $Q_C(x_+;h) > Q_C(x_C;h) = \hat{f}_n(x_C;h)$.

Finally we show that $\hat{f}_n(x_+;h) > Q_C(x_+;h)$; in fact, that the inequality holds everywhere.

$$\frac{1}{c} [\hat{f}_n(x;h) - Q_C(x;h)] = \sum_{i=1}^n [K_i(x;h) - u_i(x;h)],$$

and

$$\begin{aligned} K_i(x;h) - u_i(x;h) &= K_i(x_C;h) \{K_i(x;h)/K_i(x_C;h) \\ &\quad - 1 - (\frac{1}{2h^2}) (\|x^{(i)} - x_C\|^2 - \|x^{(i)} - x\|^2)\} \\ &= K_i(x_C;h) \{e^\tau - (1+\tau)\}, \end{aligned} \quad (5.1.13)$$

with $2h^2\tau = \|x^{(i)} - x_c\|^2 - \|x^{(i)} - x\|^2$.

The term in brackets above is the remainder from the first-order Taylor series expansion of the exponential; thus

$$e^\tau - (1+\tau) = e^\omega \frac{\tau^2}{2} > 0,$$

for some ω such that $|\omega| < \tau^2/2$. Therefore $K_i(x;h) - u_i(x;h) > 0$, implying $\hat{f}_n(x;h) > Q_c(x;h)$ for all x , hence $\hat{f}_n(x_+;h) > \hat{f}_n(x_c;h)$, and the lemma is proved.

The monotonicity of the estimated function values permits a straightforward consistency argument.

Theorem 5.1.1. Let f be a uniformly continuous, continuously differentiable probability density function in \mathbb{R}^p , and let $\{x^{(i)}, \dots, x^{(n)}\}$ be a sample of n independent observations drawn from f . Let x^0 be a starting point, independent of n , for which $f(x^0) > 0$. Then Algorithm 3.1.1a will converge in a finite number of steps to a mode estimate, x_n^* , and if h is chosen as a function of n so that $h \rightarrow 0$ and $nh^{p+2} \rightarrow \infty$ as $n \rightarrow \infty$, then $\nabla f(x_n^*) \rightarrow 0$ as $n \rightarrow \infty$.

Proof. From Lemma 5.1.1, $\hat{f}_n(x_+;h) > \hat{f}_n(x_c;h)$ unless $x_+ = x_c$, and since $\hat{f}_n(\cdot;h)$ is bounded, the sequence of function values must converge to a finite limit. Moreover,

$$\begin{aligned}
\hat{f}_n(x_+;h) - \hat{f}_n(x_c;h) &> Q_c(x_+;h) - Q_c(x_c;h) \\
&= \int_0^1 \nabla Q_c(x_c + \alpha(x_+ - x_c);h)^T (x_+ - x_c) d\alpha \\
&= \frac{c'}{h^2} \int_0^1 \sum_{i=1}^n K_i(x_c;h) [X^{(i)} - (x_c + \alpha(x_+ - x_c))]^T (x_+ - x_c) d\alpha \\
&= \frac{c'}{h^2} \int_0^1 \sum_{i=1}^n K_i(x_c;h) [X^{(i)} - \alpha x_+ - (1-\alpha)x_c]^T (x_+ - x_c) d\alpha \\
&= \frac{c'}{h^2} \int_0^1 (1-\alpha) (x_+ - x_c)^T (x_+ - x_c) \sum_{i=1}^n K_i(x_c;h) d\alpha \\
&= \frac{\|x_+ - x_c\|^2}{2h^2} \hat{f}_n(x_c;h)
\end{aligned} \tag{5.1.14}$$

Thus $\|x_+ - x_c\|^2 < (2h^2/\hat{f}_n(x^0;h)) |\hat{f}_n(x_+;h) - \hat{f}_n(x_c;h)|$ and since the sequence of function values are converging the iterates must also. Let x_n^* be the limit of the sequence of iterates. Then x_n^* is a fixed point of (5.1.8) and solution of (5.1.6) and $\nabla f_n(x_n^*;h) = 0$. With h chosen as specified, $\nabla f_n(x;h)$ is consistent everywhere in quadratic mean (cf., equation 2.2.19). Thus, all the postulates of Proposition 2.3.5 apply, and $\nabla f(x_n^*) \rightarrow 0$ w.p.1.

5.2. Empirical Results

Testing of the weighted mean update was conducted with a variety of distributions (symmetric, highly correlated Gaussian, skewed and bimodal Gaussian mixtures). Except for the bimodal mixtures, this is the same set of distributions

used for testing the mean update. The distributions, accuracy measures, and design of the simulations are described in Section 3.2.

A first concern was to determine how the weighted mean algorithm compared with the performance of the mean update, which is discussed at length in Section 3.3. The MSE measure is used as a basis for comparison. If x^* is a target mode of the density, and \hat{x} an estimate of that mode, then

$$MSE(\hat{x}) = \frac{1}{p} \sum_{i=1}^p (\hat{x}_i - x_i^*)^2,$$

where p is the dimension of the sample space. Twenty-five trials of both the weighted mean and mean updates were performed on the five standard test distributions, and average MSE values over the 25 trials obtained. Letting \overline{MSE}_w and \overline{MSE}_u be the average MSE's for the weighted mean and (unweighted) mean update, respectively, Table 5.2.1 gives the ratio $\overline{MSE}_w / \overline{MSE}_u$ for all the test distributions and all dimensions and neighbor set sizes in the trials. The trials are conducted with samples of size $N = 100$, and for this sample size it is clear that, comparing the two updates with same smoothing parameters head to head, the weighted mean outperforms the mean update almost without exception. The advantage is particularly strong with the GAUSS, R9GAUSS, and G2SKEWL data sets, where the hill-climbing ability of the algorithm is the dominant factor in its success (cf., Section 3.3). The relative advantage of the weighted mean

TABLE 5.2.1. Ratio of Average MSE, Weighted Mean vs.
Mean Update, 25 Trials, N = 100.

		<u>GAUSS</u>	<u>CAUCHY</u>	<u>R9GAUSS</u>	<u>G2SKEWL</u>	<u>G2SKEWR</u>
P = 1						
k = 10		.635	.160	.460	.608	.356
20		.329	.277	.387	.329	.346
30		.229	.252	.366	.142	.349
40		.356	.506	.331	.066	3.031
P = 2						
k = 10		.276	.507	.413	.497	.176
20		.269	.587	.240	.170	.528
30		.308	.514	.222	.273	.759
40		.392	.571	.348	.174	.732
P = 3						
k = 10		.303	.105	.153	.125	.292
20		.330	.761	.122	.046	.429
30		.377	.762	.190	.101	.730
40		.367	.740	.193	.245	.955
P = 4						
k = 10		.231	.173	.401	.111	.361
20		.334	.582	.152	.073	.698
30		.373	.727	.292	.163	.772
40		.407	.812	.434	.317	1.377
P = 5						
k = 10		.148	.294	.184	.044	.283
20		.154	.609	.258	.089	.507
30		.209	.565	.267	.236	.974
40		.319	.686	.384	.540	.954
P = 10						
k = 10		.114	.396	.212	.070	.237
20		.165	.665	.233	.169	.492
30		.244	.665	.289	.364	.807
40		.318	.704	.205	.820	1.122
P = 15						
k = 10		.097	.458	.230	.091	.296
20		.170	.566	.140	.213	.530
30		.244	.686	.153	.386	.817
40		.316	.742	.285	.512	1.068

TABLE 5.2.1. Continued.

	<u>GAUSS</u>	<u>CAUCHY</u>	<u>R9GAUSS</u>	<u>G2SKEWL</u>	<u>G2SKEWR</u>
P = 20					
k = 10	.094	.426	.181	.106	.231
20	.172	.571	.145	.241	.485
30	.257	.647	.241	.401	.662
40	.329	.756	.394	.550	.818

is also greatest with small neighbor sets, decreasing consistently throughout the table as the neighbor set is increased. On the symmetric distributions, this simply reflects the rapid improvement of the mean update as neighbor set is increased from a very sub-optimal level. On asymmetric data degradation of the weighted mean method is also a factor.

It may be misleading to judge the two updates by comparing only results obtained with matched smoothing parameters, since the effective bandwidth of the Gaussian kernel is larger than that of the quadratic (3.1.1). A more meaningful comparison, though one more difficult to obtain, would be to compare best obtainable results with the two methods. A partial development of such a comparison is given in Table 5.2.2. This table contains the ratio of average MSE values for both the mean update and weighted mean update, compared to average MSE values for the sample mean. The test distribution is G2SKEWL, and results are given based upon 25 trials with samples of size $N = 100$ and $N = 500$.

The most striking patterns in the table are:

- 1) the marked improvement of the mean update with moderate neighbor sets as sample size increases from 100 to 500 (hill-climbing difficulties are overcome);
- 2) the optimality of large neighbor sets for the mean update;

- 3) the optimality of small neighbor sets for the weighted mean, with optimal neighbor set fraction k/N becoming very small in high dimensions.

Except for certain instances of the weighted mean update, Table 5.2.2 only supplies the direction for seeking optimal parameter values; it does not directly indicate the best attainable MSE ratios. However, from Figure 3.3.15 we know that the optimal neighbor set for the mean update, with $N = 100$, typically lies between $k = 40$ and $k = 75$, and the ratios for the mean update in Table 5.2.2 with $k = 40$ and $N = 100$ are probably near the best that can be achieved. Since for consistency of the estimates, k/N approaches zero as N goes to infinity, we would expect best results from the mean update on G2SKEWL with $N = 500$ also for k/N below .75; moreover, ratios with $k/N = .40$ should be good indicators of best attainable results with the mean update.

For the weighted mean update no such bracketing is available. For $N = 100$ it appears that optimal smoothing occurs with $k \geq 40$ for $p = 1, 2$; $k = 30$ for $p = 3, 4$; $k = 20$ for $p = 5$; and $k < 10$ for $p > 5$. For $N = 500$, best k is below $N/10$ beginning with dimension 4. Performance reported for the weighted mean update with G2SKEWL in high dimensions is not the best obtainable, and it is likely that the measures of its high-dimensional performance given in Table 5.2.2 can be significantly improved.

TABLE 5.2.2. Average MSE, 25 Trials
(Ratio to Sample Mean).

		Mean Update	Weighted Mean	Mean Update	Weighted Mean
		G2SKEWL N = 100	G2SKEWL N = 100	G2SKEWL N = 500	G2SKEWL N = 500
P = 1					
K/N =	.10	20.58	12.51	16.007	.534
	.20	10.89	3.58	8.075	.166
	.30	5.326	.756	1.881	.056
	.40	2.426	.159	.076	.034
P = 2					
K/N =	.10	14.71	7.303	11.847	.281
	.20	10.11	1.714	3.555	.120
	.30	3.221	.880	.180	.090
	.40	2.042	.355	.126	.117
P = 3					
K/N =	.10	17.25	2.163	9.287	.145
	.20	8.452	.387	2.419	.121
	.30	2.949	.298	.189	.162
	.40	1.238	.304	.174	.227
P = 4					
K/N =	.10	13.25	1.476	5.646	.187
	.20	7.470	.547	.843	.218
	.30	3.008	.490	.491	.291
	.40	1.566	.497	.208	.369
P = 5					
K/N =	.10	12.71	.562	6.266	.178
	.20	5.346	.478	.980	.234
	.30	2.120	.500	.260	.306
	.40	.990	.534	.201	.415
P = 10					
K/N =	.10	9.518	.662	3.379	.508
	.20	4.055	.684	.579	.589
	.30	1.953	.711	.421	.641
	.40	.897	.735	.368	.680
P = 15					
K/N =	.10	9.301	.844	2.386	.694
	.20	4.039	.861	1.051	.736
	.30	2.274	.878	.721	.763
	.40	1.732	.887	.599	.786

TABLE 5.2.2. Continued

	Mean Update	Weighted Mean	Mean Update	Weighted Mean
	G2SKEWL N = 100	G2SKEWL N = 100	G2SKEWL N = 500	G2SKEWL N = 500
P = 20				
K/N = .10	8.286	.881	2.579	.737
.20	3.666	.883	.905	.770
.30	2.205	.885	.669	.795
.40	1.627	.894	.629	.815

Nevertheless, the high-dimensional behavior of the weighted mean update on G2SKEWL suggests that the influence of the tails of the Gaussian kernel is quite strong with the range of smoothing parameters reported in Table 5.2.2. This influence casts doubts on the ability of the weighted mean for exploring multimodality. To investigate the performance of the updates in the presence of multiple modes, a set of bimodal normal mixtures was generated. The bimodal mixtures are referred to as G2SEP, and defined by

$$f(x) = 0.3 \phi(x; -0.0526 \tilde{1}, I) + 0.7 \phi(x; 2.4474 \tilde{1}, I), \quad (5.2.1)$$

where $\phi(x; \mu, \Sigma)$ is the normal density with parameter (μ, Σ) , I is the identity matrix of appropriate dimension, and $\tilde{1}$ is the vector of all 1's. G2SEP is obtained from G2SKEWL by removing the factor $p^{-1/2}$ in determining the mean separation, and translating so that the secondary mode is at the origin. For $p = 1$, G2SEP and G2SKEWL have the same shape, but for $p \geq 2$, G2SEP is bimodal. The MSE value of the major mode is $(2.4474)^2 = 5.99$.

In tests with G2SEP the initial guess supplied to the updates is $x_0 = 0.5 \tilde{1}$, a location which would normally be identified with the nearer minor mode at the origin. The simulations were conducted with samples of size $N = 100$, once with the standard neighbor set range ($k = 10, 20, 30, 40$) and once with a smaller range ($k = 2, 5, 10, 15$). In

addition a truncated version of the weighted mean was added as one possibility for reducing tail influence. Truncation occurred at the k -th neighbor, while the smoothing parameter was set to $h = \frac{1}{2} r_k(x)$, $r_k(x)$ being the current k -th neighbor distance. This produced univariate kernels equivalent to a Gaussian truncated two standard deviations from the mean. Results with G2SEP are summarized in Tables 5.2.3 and 5.2.4.

The MSE values in these tables indicate the average distance of mode estimates from the minor mode. MSE values around 6.0 are associated with the dominant mode, and large and intermediate MSE values do not necessarily indicate ragged algorithmic performance. The results of Tables 5.2.3 and 5.2.4 are best considered in light of typical frequency distributions of the MSE values produced during the sequence of trials in the simulations. Histograms of the MSE values produced by each of the updates acting on G2SEP in dimension 5 are given in Figures 5.2.1 through 5.2.4. The neighbor set sizes reported in the histograms are $k = 5$, $k = 10$, $k = 20$, and $k = 40$, respectively. Sample size was $N = 100$.

The three updates are based upon kernel estimates with effective bandwidths that are greatest for the weighted mean, and least for the truncated weighted mean. To put it another way, the truncated weighted mean is the update most influenced by a small number of points around the

TABLE 5.2.3. G2SEP Average MSE, 25 Trials.

	Mean Update	Weighted Mean	Truncated Weighted Mean $\alpha = .5$
P = 1			
k = 2	.2510	.2466	.2490
5	.2318	.2774	.2404
10	.2702	.3946	.2646
15	.3618	.8504	.4092
P = 2			
K = 2	.2739	.2929	.2902
5	.2678	.5030	.4375
10	.2346	1.125	.3298
15	.2952	2.111	.5148
P = 3			
k = 2	.2155	.2098	.2958
5	.2022	.2549	.3398
10	.1366	.4730	.3504
15	.1177	2.229	.3002
P = 4			
k = 2	.2903	.3404	.3217
5	.2154	.5077	.2995
10	.1507	.5754	.2399
15	.1493	2.195	.2038
P = 5			
k = 2	.2349	.2364	.3352
5	.1844	.1077	.2983
10	.1063	.2726	.2137
15	.07343	1.494	.1766
P = 10			
k = 2	.2837	.2413	.4134
5	.1630	.06391	.3520
10	.1022	.5359	.2596
15	.08180	1.970	.2077
P = 15			
k = 2	.3328	.1053	.4571
5	.1644	.04992	.3899
10	.09711	.2718	.2641
15	.06432	1.225	.1942

TABLE 5.2.3. Continued.

		Mean Update	Weighted Mean	Truncated Weighted Mean = .5
		<hr/>	<hr/>	<hr/>
P = 20				
k =	2	.3596	.05778	.5956
	5	.1910	.04557	.4861
	10	.1203	.2794	.3310
	15	.07893	.2768	.2071
		<hr/>		

TABLE 5.2.4. G2SEP Average MSE, 25 Trials.

	Mean Update	Weighted Mean	Truncated Weighted Mean $\alpha = .5$
P = 1			
k = 10	.3637	.7611	.3138
20	.6371	2.316	.3902
30	1.360	4.736	.9151
40	3.257	5.139	2.131
P = 2			
k = 10	.2245	1.503	.2141
20	.3706	4.745	.4799
30	1.537	6.182	1.410
40	4.192	5.983	1.620
P = 3			
k = 10	.1964	.9155	.3275
20	.2856	3.643	.4054
30	.7158	5.592	.3179
40	2.496	6.007	.6051
P = 4			
k = 10	.1622	.5605	.3175
20	.07911	5.284	.1754
30	.1590	6.055	.07882
40	2.441	5.996	.04681
P = 5			
k = 10	.1133	.5124	.2422
20	.05822	3.353	.1509
30	.07354	5.750	.06541
40	1.338	5.969	.04171
P = 10			
k = 10	.1149	.04977	.2910
20	.06118	2.170	.1460
30	.05735	5.757	.06117
40	.9219	5.977	.04142
P = 15			
k = 10	.1091	.04317	.2928
20	.05200	2.956	.1302
30	.07142	5.825	.04990
40	.7748	6.043	.03227

TABLE 5.2.4. Continued.

	Mean Update	Weighted Mean	Truncated Weighted Mean $\alpha = .5$
P = 20			
k = 10	.1001	.2776	.3423
20	.05938	2.654	.1397
30	.1024	4.580	.06074
40	1.199	5.746	.04010

Figure 5.2.1.

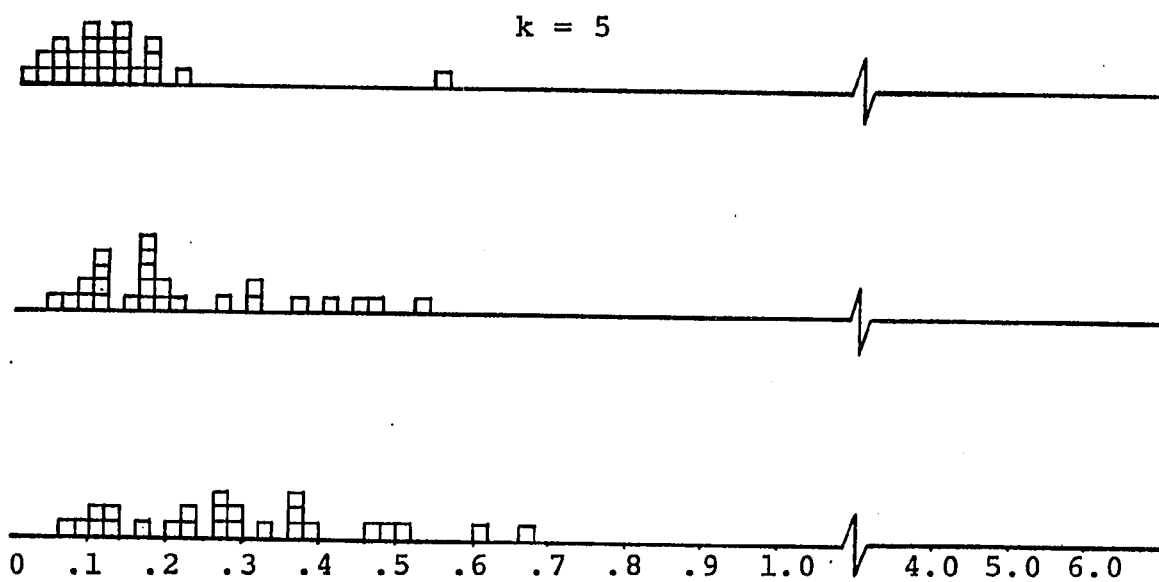
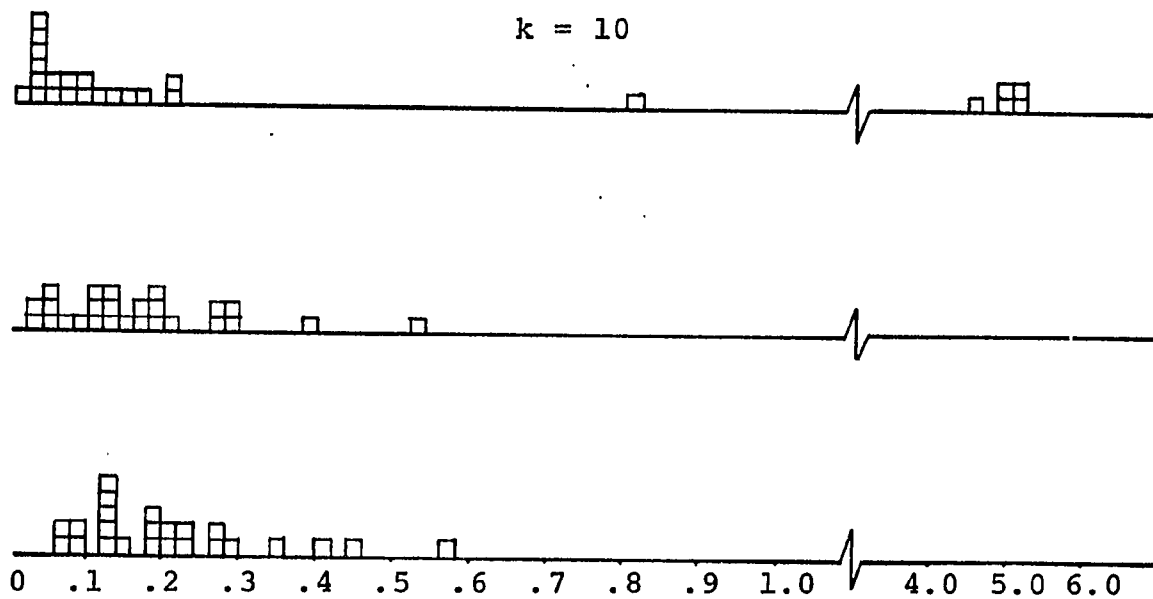


Figure 5.2.2.



Distribution G2SEP, Dimension = 5, MSE Values of 25 Trials

Figure 5.2.3.

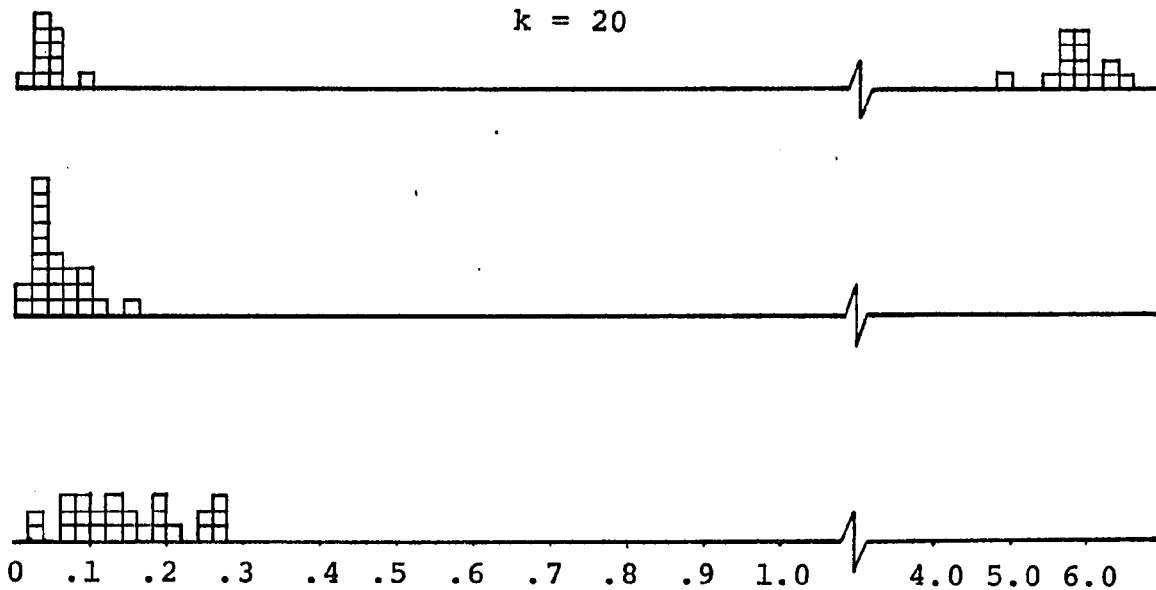
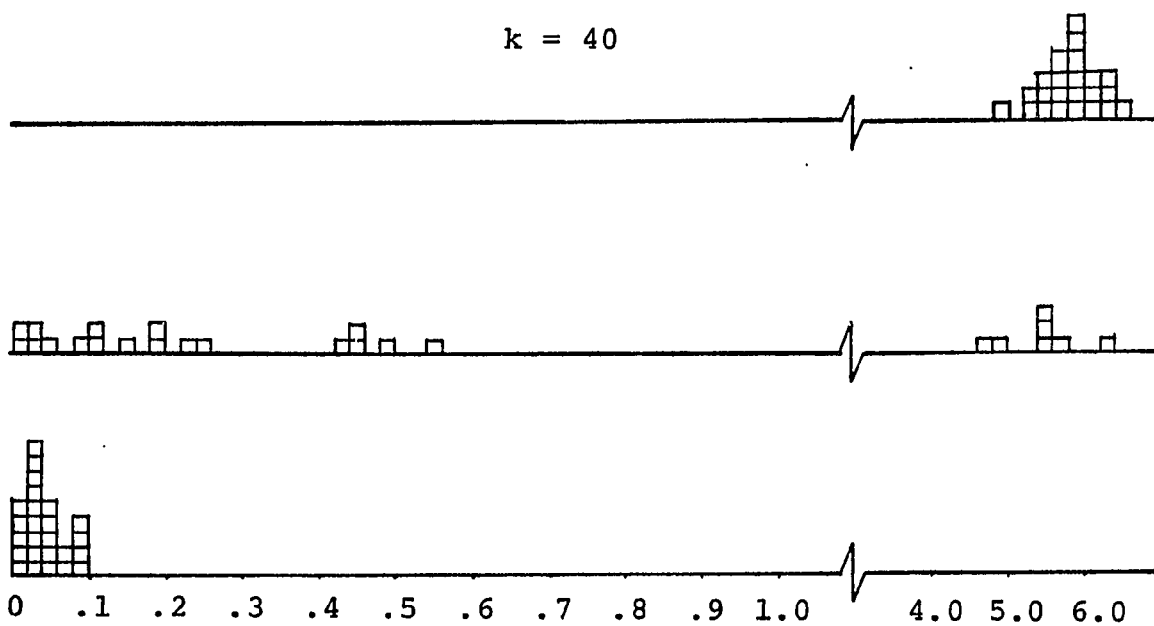
 $k = 20$ 

Figure 5.2.4.

 $k = 40$ 

current iterate. The sample mean assigns equal weight to all the members of the neighbor set, zero elsewhere. The full weighted mean distributes the weighting of its update step over the entire sample. Thus, it is natural that with small neighbor set sizes, as in Figure 5.2.1, the weighted mean update exhibits the least variance of the three methods, its truncated version the greatest.

The weighted mean is easily the most mobile of the update methods. The truncated version reported in these pages does not share these attributes, but the severity of the truncation can be varied by controlling the smoothing parameters h . As stated before, the tests reported here took h as one-half of the k -th neighbor distance. Other tests were made taking h as the full k -th neighbor distance, and in this case the truncated version behaved almost exactly as the nontruncated version. A continuum of intermediate responses is certainly available.

The full weighted mean provides the quickest and most focused identification of a mode. Even in the case pictured in Figure 5.2.2, where the weighted mean is attracted to both modes a certain proportion of the time, the set of mode estimates divides clearly into two groups, each clumped more tightly and more accurately than the sets of mode estimates obtained with the truncated updates. In Figures 5.2.2 and 5.2.3 one can observe a migration of estimates from the minor to the major mode. This transfer is complete

by Figure 5.2.4. Thus, the weighted mean appears to lose touch with the subordinate mode more rapidly than the other updates. Still, in all cases examined, it identifies clearly one or both of the population modes. With $k = 40$, for example, we still have evidence of a nearly unbiased mode estimate; the dominant mode has MSE of about 6.0, the sample mean an MSE of 4.2.

The mean update and truncated weighted mean are seen lagging behind the weighted mean, but going through essentially the same stages. The differences that appear are, first, that the truncated weighted mean is improving its focus on the minor mode throughout, and secondly, that the mean update is less definitive than the weighted mean. With $k = 30$ (not shown) the mean update begins to migrate toward the dominant mode. The migration is well-developed in Figure 5.2.4. With the mean update the transfer is more blurred than with the weighted mean. More trials terminate at intermediate locations divided between the two modes, and neither of the modal clusters is as tight.

Concluding Remarks

Based upon our observations thus far, then, the weighted mean should be preferred to the mean update. The weighted mean update is far more effective at locating the global mode of the distribution, in both unimodal and bimodal data. With tail influence controlled, either by the use of small smoothing parameters, or truncation of the

weights, the weighted mean is also able to identify secondary modes with precision certainly comparable, and generally superior, to that of the mean update.

By including the possibility of truncation combined with control of the smoothing parameter, the weighted mean acquires a degree of flexibility that is unattainable by the mean update. For example, if we fix a radius r_T about the current iterate, take $h = r_T$, and truncate at the same radius, the effect is equivalent to use of a Gaussian kernel truncated at one standard deviation, which in turn is virtually equivalent to the quadratic kernel underlying the mean update. Maintaining the same truncation, and taking h very large, we approach a uniform kernel. With h much less than r_T the kernel is again essentially equivalent to an unmodified Gaussian. Much more work remains to be done before it will be clear how best to utilize this flexibility. Nevertheless, the presence of adaptive capability and potential self-governance (through examination of the weights) is clearly an advantage. Without involving himself in the issue of truncation, our experimentation has indicated that the user can expect good results from the full weighted mean with most reasonable choices of the smoothing parameter, and by varying that parameter can concentrate on the global mode or press the search for additional local modes.

VI. MULTISTART ALGORITHM AND SUGGESTIONS FOR FURTHER WORK

6.1. Consideration of Previous Chapters and Introduction

The express goal of this investigation has been the complete enumeration of the modes of probability density functions, particularly high dimensional densities. This includes recognition of unimodality where appropriate, determination of the number of modes when not, and accurate estimates of location in either case. We have envisioned the task as requiring effort at two levels. First is the implementation of sensitive local stochastic optimization procedures. Secondly, a "global," or supervisory strategy is required for directing the application of the local procedures and interpreting their results.

Our research thus far has been devoted primarily to the development of good local optimization algorithms, and related to this, if implicitly, good density estimation procedures. Local performance is the natural initial concern, since identification of a local maximum of the density function is the primitive operation in the kind of exploration of data structure we are proposing. The capabilities and limitations of this primitive action define the potential for a comprehensive exploration of modality.

Extensive testing with simulations of various distributional characteristics directed us to seek refinements in the mean update algorithm with which we began, and led to

the consideration of three classes of optimization procedures. The methods are considered in detail in the preceding three chapters, and their essential characteristics and utility are discussed in the introductory and concluding remarks of the associated chapters.

Here we identify two significant patterns which evolved in the development of the local optimizers. The first trend was the increasing attention paid to skew and bimodal mixtures once feasibility of the optimization effort was demonstrated on standard unimodal distributions. Particular geometries were isolated to test the discriminatory power of the candidate mode estimators. For example, in tests with the family of bimodal normal mixtures G2SEP, reported in Chapter V, the mode estimators were constrained to begin at a location poised between the two modes; moreover this was a location associated more closely with the subordinate mixture component, which accounted for an average of only thirty percent of the probability mass. By comparison, the unimodal mixture G2SKEW included a subordinate component with the same thirty percent probability, and the ability of the mode estimators to distinguish this structure from the bimodal structure in G2SEP was examined. One of the orientations chosen (G2SKEWLT) was intended to make the distinction most difficult. Thus as testing proceeded it became increasingly a matter of abstracting critical configurations arising in the enumeration of multiple modes. By isolating these

geometries we felt that we could explore the behavior of the complete mode enumeration process, and do so in the most precise and economical way. Experience with a global implementation, to be presented in this chapter, supports this claim.

The second major development in our study was the movement toward adaptive local optimizers, as manifest in the weighted mean and truncated weighted mean algorithms of Chapter V. What is significant in these algorithms is the recognition of a regression model implicit in the formation of mode estimates, hence the identification of a problem which is dual to the location problem. That is the determination of the weights used to characterize the mode (equivalently, the coefficients of the solved regression model). The dual problem seems likely to us to be more invariant and more amenable to analysis than other approaches to the evaluation of mode estimates. Thus, it may be able to supply the linkage needed between the local and supervisory components of a complete assessment of sample modality.

6.2. Multistart Algorithm

To investigate the potential of our local optimizers for a comprehensive exploration of data structure, a basic multistart driver was implemented and applied to the skew and bimodal mixture distributions. The multistart method is probably the organization most frequently employed in multivariate searches for global optimizers (Rubinstein, 1981,

Chapter 7). The method is simply to run an iterative local method from a sequence of starting points, $(x_1, x_2, x_3, \dots, x_N)$, producing a set of termination points $T = (x_1^*, x_2^*, x_3^*, \dots, x_N^*)$. The global maximum is estimated as $\max_i \hat{f}(x_i^*)$, where \hat{f} is the exact or estimated objective function. For enumeration of all local modes it will be necessary to identify clustering of the elements of T . The multistart algorithm, despite its expense, is well suited to mode estimation because the sample observations immediately provide an appropriate set from which to choose starting points.

Our multistart algorithm started once from every observation in the sample, and recorded the sequence of termination points, which then formed a set of observations of the same cardinality and dimension as the original sample. The algorithm was run on distributions G2SKEW and G2SEP with samples of size $N = 100$, with samples spaces of dimension 1, 2, 3, 4, 5, 10, 15, and 20, and of course various values of the control parameters. Summary statistics of repeated simulations with G2SEP are presented later (in Tables 6.2.1.a and 6.2.1.b), but first some illustrative displays of the action of the algorithms in dimension five are presented.

The displays require some explanation. What was sought was an analogue of two-dimensional scatterplots, which would indicate visually the clustering of sample observations or mode estimates. This was accomplished by taking one coordinate for the plots to be the line connecting the means

of the component mixtures, and taking the other coordinate to represent the length of the orthogonal projection onto that line. A schematic diagram of the geometry of the plots is given in Figure 6.2.1. The rectangular box amounts

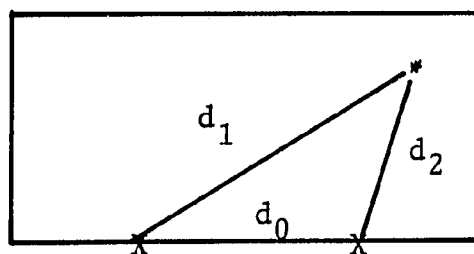


FIGURE 6.2.1. Schematic, Multivariate Scatterplots.

to a window on the sample space, with the bottom edge representing the line between the component means. The position of the means are indicated by "X's" overstruck on the border. A point z with z with $\|z - \mu_1\| = d_1$ and $\|z - \mu_2\| = d_2$ will be placed at the location of the asterisk in the figure. A sense of scale can be obtained by relation to the distance between the means, $d_0 = \|\mu_1 - \mu_2\|$.

Because many points may occupy the same print position in the plots, a count is maintained of the occupancy of each print position. If that count cannot be expressed with a single digit, the greatest integer multiple of 10 is used to index a character of the alphabet, and that character and the one-digit remainder are overstruck. For example, "F" overstruck with "3" marks the coincidence of 63 points at

Figure 6.2.2. Scatter of 100 Observations from G2SKEW, $p = 5$

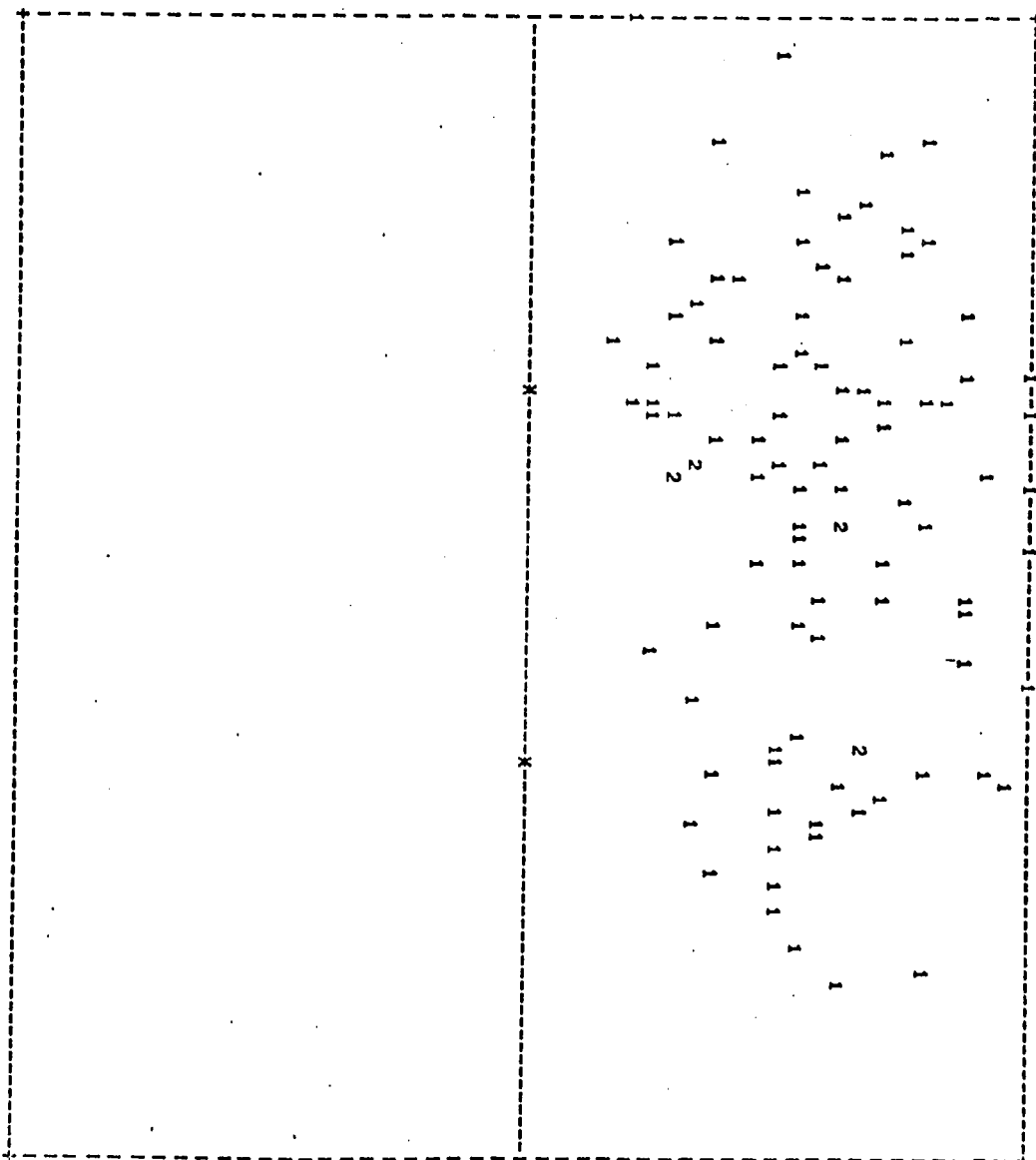


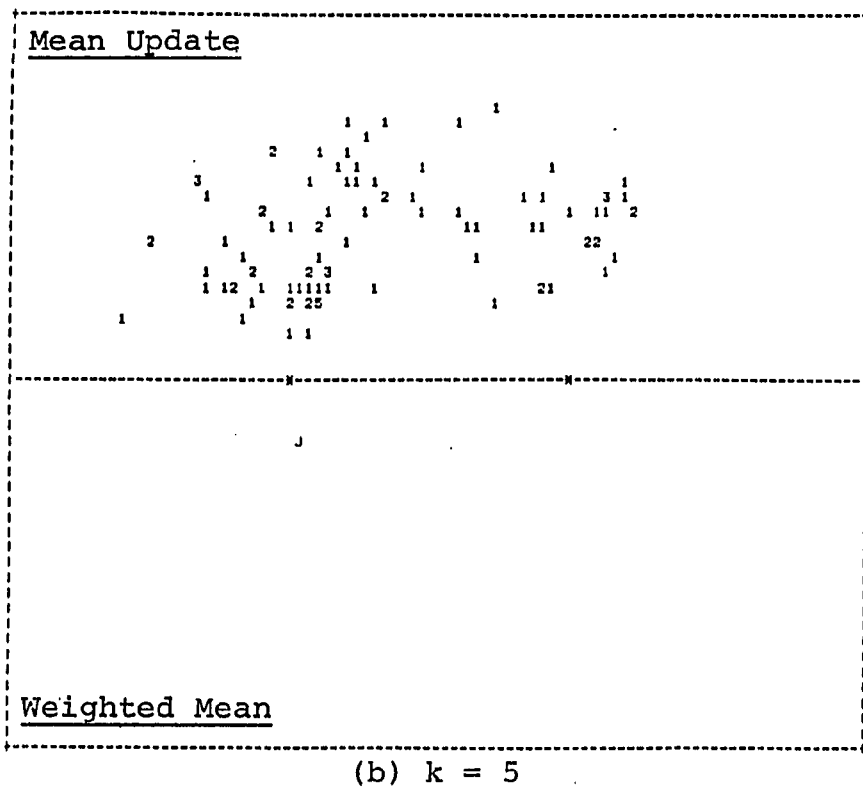
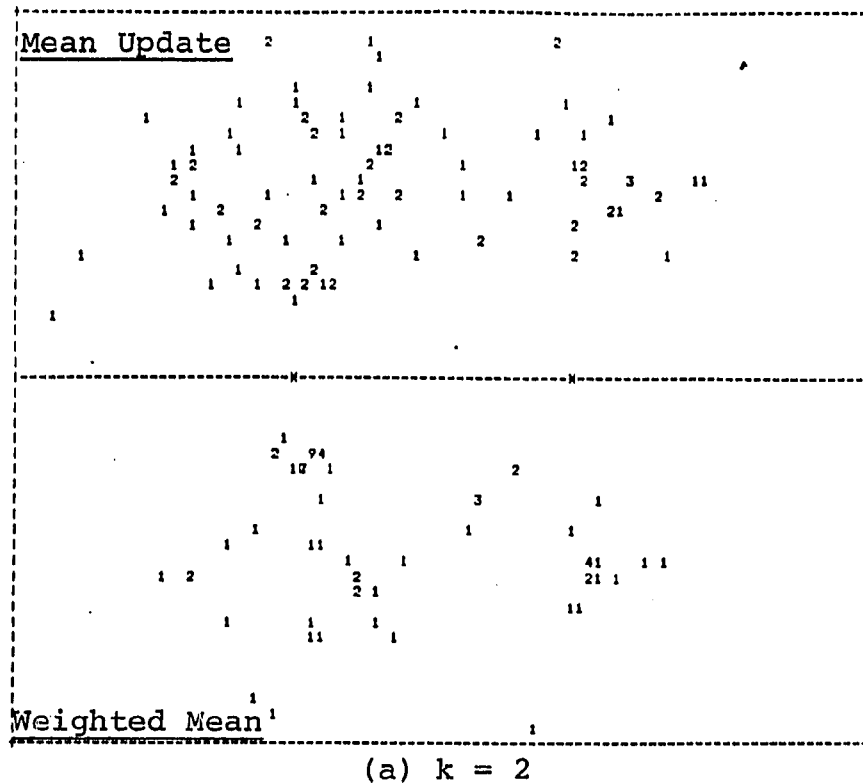
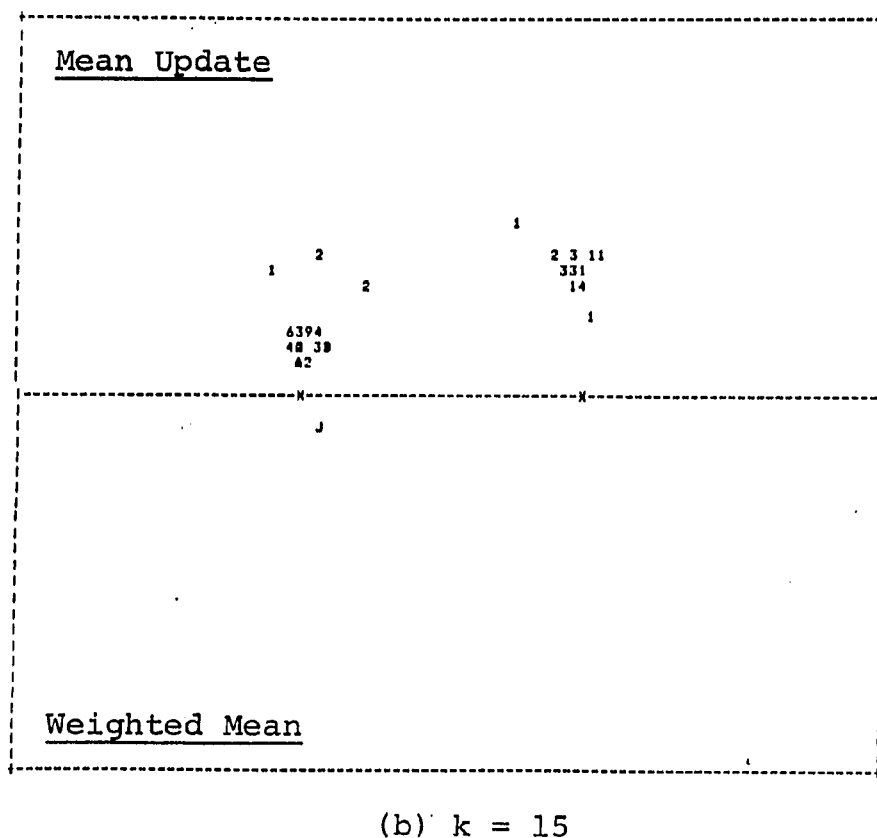
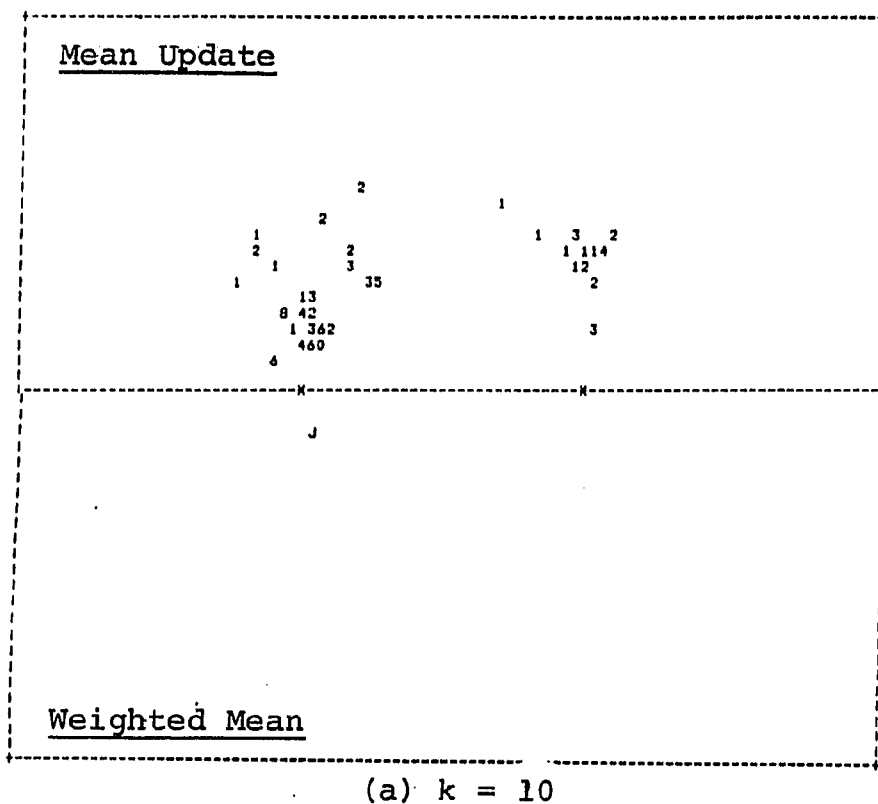
Figure 6.2.3. Transformed Scatter, G2SKEW, $p = 5$ 

Figure 6.2.4. Transformed Scatter, G2SKEW, $p = 5$ 

the indicated position. The single character "J" marks the shared location of all 100 points.

To compare two plots, as in comparing the application of two optimization methods to the same problem, one of the plots may be flipped about the horizontal axis so that the means appear on the upper edge of the window, and the two plots may then be positioned together sharing the same mean markers. In all cases where we compare algorithms using adjacent scatter plots the mean update appears in the upper window and the nontruncated weighted mean in the lower.

A five-dimensional sample of 100 observations from G2SKEW is depicted in Figure 6.2.2. The mode of the density is coincident with the left-most mean marker. The sample consists of 66 observations generated from the major component and 34 from the minor component of the mixture density. This sample is the input to the multistart algorithm whose output is pictured in Figures 6.2.3 and 6.2.4. In Figure 6.2.3(a) the optimizations are conducted with a neighbor set of size $k = 2$, and with neighbor sets of $k = 5$, 10, and 15 in Figures 6.2.3(b), 6.2.4(a), and 6.2.4(b) respectively. Again, individual points in these plots are termination points for either the mean update or weighted mean update starting at one of the sample points given in Figure 6.2.2, and conducted upon that sample.

The parameter value $k = 2$ is of course much too small for the mean update, whose effect in Figure 6.2.3(a) is

essentially to pair off nearest neighbors, leaving the scatter of the sample unchanged. By comparison, though considerable dispersion remains, the weighted mean performs a more distinctive transformation of the raw sample. Thirty-seven termination points are aligned together at one place, with nine points in the diagonally adjacent print position. Thus, there is already a tight cluster comprising about one-half of the sample which gives evidence of the location of a mode.

With k as low as five the coalescence of the mode estimates generated by the weighted mean is complete and striking. The weighted mean algorithm is emphatic in recognizing a single mode, whose location is estimated with good accuracy, increasing slightly as k is increased. Also as k is increased there is a more gradual coalescence of the termination points of the mean update, which admittedly is still not in its optimal range. Though the number of termination points classified with the minor component mean drops from 31, with $k = 2$, to 21, with $k = 15$, the mean update is clustering about both of the component means. Since the subordinate component represents an effect which would have important subject matter implications, it may not be undesirable to have attention drawn to it. However, for estimation of modes per se, the two clusters are misleading.

Figure 6.2.5 depicts the raw five-dimensional sample from the bimodal distribution G2SEPRT. For this distribution

the major mode is coincidental with the right-most mean marker, and the sample is composed of 40 observations from the minor component, 60 from the major. Clearly, the separation of the groups is greater here than in the unimodal mixture. Output from the mean and weighted mean updates are presented in the six successive panels of Figures 6.2.6 through 6.2.8 for parameter values $k = 5, 10, 15, 20, 30$, and 40 , respectively. We should note that the raw sample generated for $k = 30$ and $k = 40$ was different than that used for the smaller four parameter values. It included 26 observations from the subordinate component, 74 from the other.

Certain of the trends evidenced in the sequence of panels are predictable. Coalescence of the set of termination points increases with increase in the smoothing parameter. Up to some point this coalescence brings with it clearer identification of the modes. Beyond that point it represents loss of detail. The demarcation point is higher for the mean update than for the weighted mean, and higher still for the version of the truncated weighted mean which we investigated. The results of the truncated weighted mean are pictured in Figure 6.2.9, for $k = 10, 20, 30$, and 40 .

Were the differences in the performance of the algorithms solely differences in the range of desirable smoothing parameters there would be little reason to prefer one algorithm above the other. However, there are further characteristic differences which lead us to prefer the weighted mean.

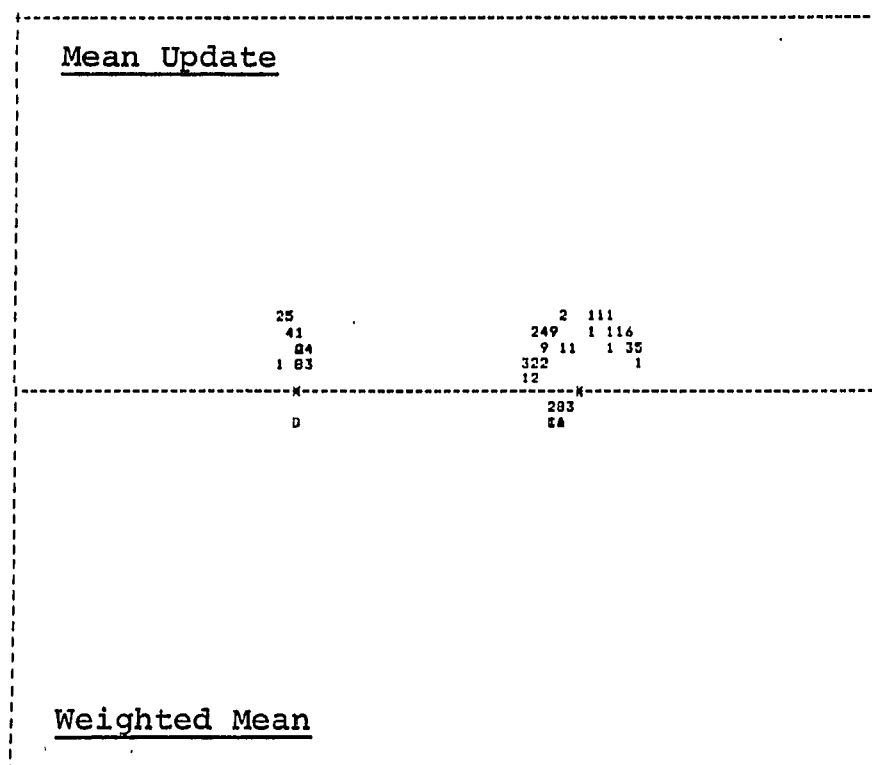
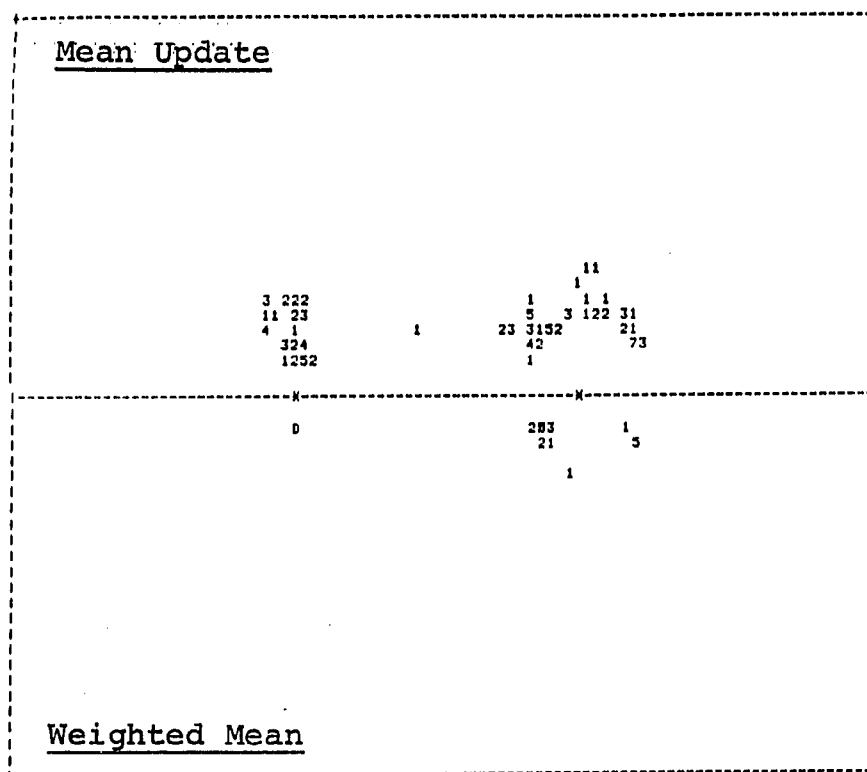
Figure 6.2.6. Transformed Scatter, G2SEP, $p = 5$ 

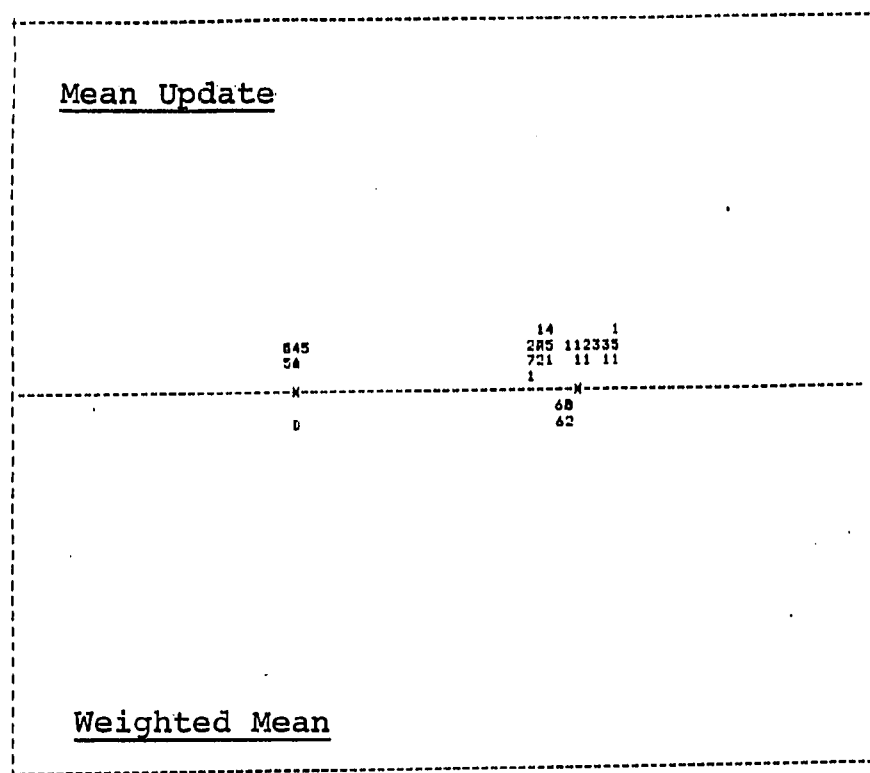
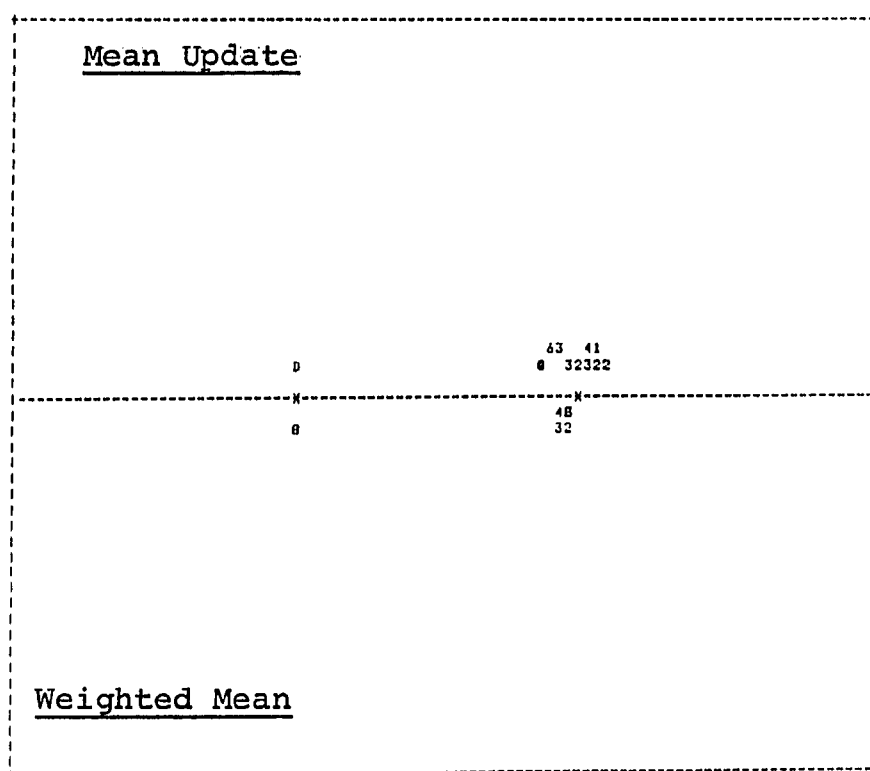
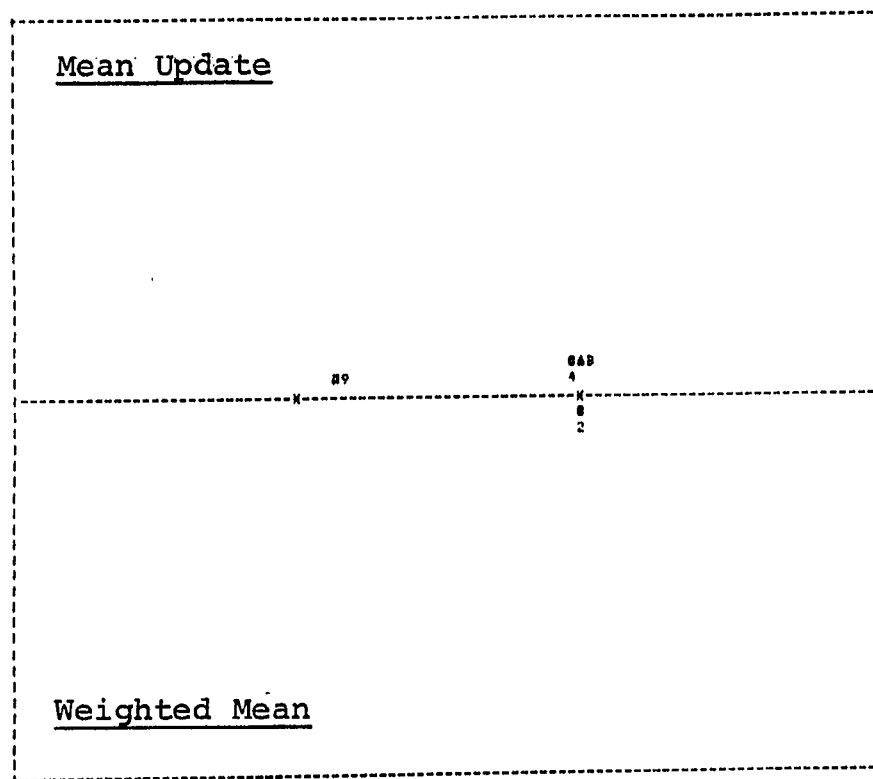
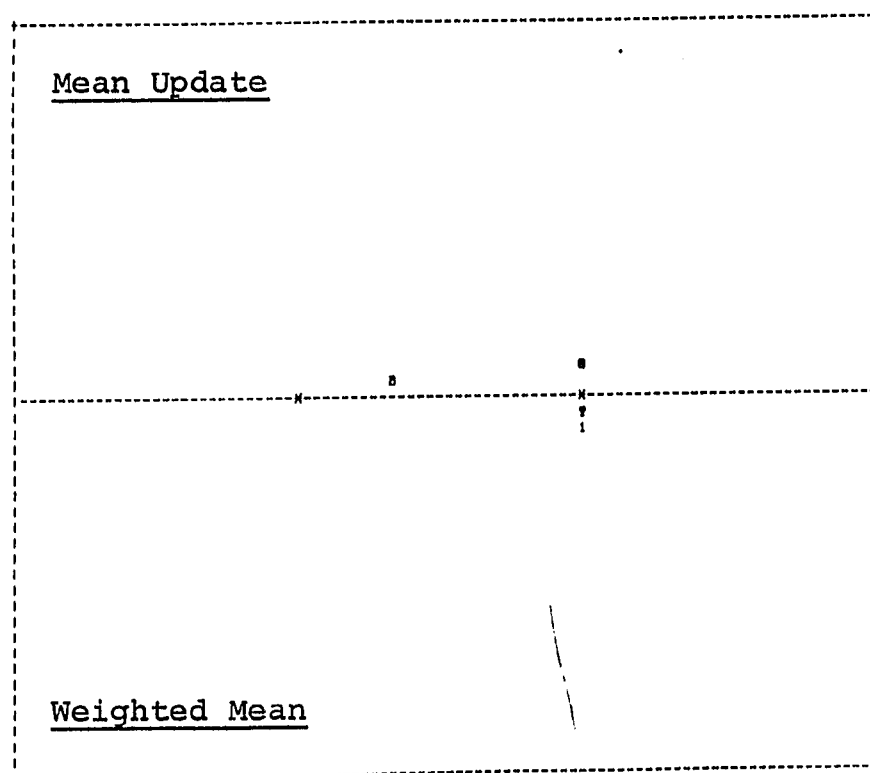
Figure 6.2.7. Transformed Scatter, G2SEP. $p = 5$ (a) $k = 15$ (b) $k = 20$

Figure 6.2.8. Transformed Scatter, G2SEP, $p = 5$ (a) $k = 30$ (b) $k = 40$

Though clustering about two locations is certainly well advanced by $k = 20$ (see Figure 6.2.7(b)), the mean update does not approach the degree of coalescence of the weighted mean until $k = 30$ (see Figure 6.2.8(a)), and by this time as the panel below with $k = 40$ makes clearer, the location of the second cluster has already begun to drift away from the minor mode. The weighted mean is able to achieve a high degree of coalescence without muddying the distinction between the modes or displacing them from their true locations. Patterns recognized by the weighted mean are recognized decisively, and transitions which occur as the smoothing parameter varies occur abruptly. Because of this the class of candidate structures presented to the user (unimodal, bimodal, etc.) are more clearly laid out, and the decisions which the user must make are better defined. In addition, the feedback mechanism provided by the final weighting vector supplies the user with diagnostic support for making those decisions. Thus we shall reiterate our preference, all factors considered, for the use of the weighted mean algorithm.

The one reservation to be made regarding the full weighted mean concerns its strong attraction to the dominant mode, which perhaps makes the possibility of ignoring secondary effects too great. To limit this attraction the method of truncation described in Section 5.1 was considered, and results of its performance displayed in Figure 6.2.9. With such truncation much larger smoothing parameters are required,

but with $k = 40$ the truncated weighted mean achieves nearly the same resolution of the modes as that attainable with the full weighted mean, and does so without biasing the estimates of location. There is not sufficient evidence yet to conclude whether a full or truncated version of the weighted mean algorithm is superior.

The scatter plots we have looked at thus far essentially represent a single trial of the various algorithms. An idea of the representativeness of these trials may be obtained from Tables 6.2.1 (a) and 6.2.1 (b), which report summary statistics for five such trials on the bimodal family of distributions G2SEPRT. The first statistic report is COUNT1, which is the number of termination points identified with the subordinate mode. The expected value of COUNT1 in the raw sample is 30. The second statistic measures primarily the degree of coalescence of the set of termination points. Entitled PVNE, or proportion of variance not explained, it is defined as:

$$PVNE = \frac{\sum_{i=1}^N (x_i^* - m^{(i)})^2}{\sum_{i=1}^N (x_i^* - \bar{x}^*)^2}$$

where \bar{x}^* is the mean of the set of termination points T , and $m^{(i)}$ is the (known) mode of the distribution nearest to x_i^* . Because known modes are used PVNE also penalizes modal clusters for drift from the true location.

In many cases, particularly in higher dimensions, and

TABLE 6.2.1.(a). Mean Statistics of 5 Trials of the Multistart Algorithm.

		Mean Update		Nontruncated Weighted Mean		Truncated Weighted Mean	
		Count 1	PVNE	Count 1	PVNE	Count 1	PVNE
P = 1.							
K =	5	33.6	.3199	30.8	.3231	32.6	.3384
	10	33.2	.3152	28.4	.4808	33.0	.3288
	15	32.4	.3519	18.8	.5120	33.0	.3774
	20	26.4	.4682	4.2	-	28.4	.4894
	30	33.0 ¹	.6597 ¹	7.0	-	15.6	-
	40	7.8	-	0.0	-	14.6	-
P = 2							
K =	5	31.4	.3194	28.2	.2319	30.2	.3565
	10	28.8	.2529	26.6	.1474	29.4	.2810
	15	28.4	.1818	26.0 ¹	.0764 ¹	28.8	.2106
	20	28.0	.1427	9.2	-	28.2	.1632
	30	21.8	.3180	0.0	-	20.8	.2789
	40	12.0	-	0.0	-	29.0 ¹	.0539 ¹
P = 3							
K =	5	32.8	.2879	31.4	.2130	33.6	.3567
	10	31.6	.1867	29.2	.1350	32.6	.2609
	15	31.4	.1294	29.0 ²	.0392 ²	32.0	.2245
	20	31.6	.1188	27.3 ¹	.0395 ¹	31.8	.1840
	30	20.8	.1523	0.0	-	23.3 ²	.1673 ²
	40	5.4	-	0.0	-	24.3 ¹	.1741 ¹
P = 4							
K =	5	29.2	.2885	28.8	.1794	29.8	.3837
	10	29.6	.1716	28.2	.1321	29.2	.2364
	15	29.4	.1339	22.8	.0905	29.4	.1743
	20	29.4	.1050	11.2	-	29.4	.1510
	30	27.3 ²	.0676 ²	0.0	-	27.2	.1028
	40	29.3 ¹	.1014 ¹	0.0	-	27.2	.0903

¹Average of 3 observations.²Average of 4 observations.

TABLE 6.2.1.(b). Mean Statistics of 5 Trials of the Multistart Algorithm.

		Mean Update		Nontruncated Weighted Mean		Truncated Weighted Mean	
		Count 1	PVNE	Count 1	PVNE	Count 1	PVNE
P = 5	K = 5	35.2	.2339	34.6	.1106	33.2	.3536
	10	33.2	.1337	34.6	.0547	35.2	.2191
	15	34.8	.0983	34.0	.0367	35.0	.1539
	20	35.0	.0757	30.8	.0363	35.0	.1204
	30	30.0	.0539	0.0	-	30.0	.0775
	40	29.8	.1534	0.0	-	30.0	.0606
	P = 10	K = 5	27.2	.2216	27.0	.0318	27.2
10		27.2	.1221	26.4	.0265	27.2	.2513
15		27.2	.0843	23.8	.0272	27.2	.1507
20		27.2	.0634	22.3 ¹	.0267 ¹	27.2	.1122
30		32.0	.0353	9.4	-	32.0	.0838
40		32.0	.1017	0.0	-	32.0	.0559
P = 15		K = 5	30.8	.1893	30.6	.0241	30.8
	10	30.8	.1007	29.6	.0220	30.8	.2539
	15	30.8	.0700	28.0	.0255	30.8	.1321
	20	30.8	.0537	29.3 ²	.0198 ²	30.8	.1011
	30	31.4	.0338	7.0	-	31.4	.0599
	40	31.4	.1277	0.0	-	31.4	.0413
	P = 20	K = 5	30.4	.1808	30.4	.0172	30.4
10		30.4	.0914	30.4	.0159	30.4	.2854
15		30.4	.0574	30.2	.0154	30.4	.1492
20		30.4	.0418	27.4	.0162	30.4	.0963
30		29.6	.0346	7.8	-	29.6	.0453
40		29.6	.1722	0.0	-	29.6	.0325

¹Average of 3 observations.²Average of 4 observations.

particularly with the full weighted mean, complete coalescence to a single mode would occur, in which case the measured PVNE would have no meaning. If this circumstance occurred only once or twice out of the five trials, statistics were reported on the basis of the remaining trials, and that fact duly recorded.

The figures in Table 6.2.1 confirm observations made earlier. The full weighted mean collapses on the major mode unless the neighbor set chosen is fairly small. However, it is effective over an acceptably wide range of parameter values, and in that range the results are more positive than any of the results obtained with the other algorithms. The truncated weighted mean is still improving at $k = 40$ and may match the performance of the full weighted mean with a larger smoothing parameter, though that is questionable.

6.3. Further Work

We have presented a working and analytically supported algorithm for the analysis of modes, and one that we believe is already successful in achieving reasonable goals set for it. Further work remains, however, before that algorithm is in a mature stage of development.

It is certainly desirable to seek improvement in the computational efficiency of the algorithm. We have already suggested the possible use of an imbedded regression model to accelerate the convergence of the local optimizer. The

elementary multistart algorithm we utilized wants refinement to eliminate unnecessary redundancy. A strategy for reducing the number of local optimizations is needed, especially for applications with large samples. Also parallelism in the operations may be exploited, for example, by performing the update step on all sample points in unison, or by employing two values of the smoothing parameter simultaneously. A clustering procedure is needed to produce a meaningful list of candidate mode estimates from the output of the multistart algorithm.

From a theoretical standpoint one would like to identify the general properties which a weighting function must satisfy to insure consistency of the update algorithm. This would justify the use of weightings with a lesser computational burden than the Gaussian-based weights used thus far, such as weights drawn from the quartic kernel, $K(x) = (1-x^2)^2$, $|x| \leq 1$. The most pressing need, though, is for a deeper understanding of the adaptive capabilities of the mode-seeking algorithm. Of particular interest is the distribution of the regressive weighting vector in the vicinity of a mode, which if known would greatly assist in the choice of appropriate smoothing parameters, and in assessing the significance of mode estimates.

APPENDIX

Generation of Equivalent Two-Component Gaussian Mixtures

For a Gaussian mixture with m components,

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

$$p_i(x) = (2\pi)^{-n/2} |\Sigma_i|^{-1/2} \exp\left\{-\frac{1}{2} (x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\right\},$$

we have

$$\nabla p(x) = - \sum_{i=1}^m \alpha_i p_i(x) \Sigma_i^{-1} (x-\mu_i)$$

and

$$\nabla^2 p(x) = \sum_{i=1}^m \alpha_i p_i(x) [\Sigma_i^{-1} (x-\mu_i) (x-\mu_i)^T \Sigma_i^{-1} - \Sigma_i^{-1}].$$

In a two-component mixture, critical points occur for x satisfying

$$0 = \alpha_1 p_1(x) \Sigma_1^{-1} (x-\mu_1) + \alpha_2 p_2(x) \Sigma_2^{-1} (x-\mu_2).$$

Thus, if x is a critical point of the mixture density, there exists a constant $c < 0$ such that

$$\Sigma_1^{-1} (x-\mu_1) = c \Sigma_2^{-1} (x-\mu_2).$$

If $(\mu_2 - \mu_1)$ is an eigenvector of both Σ_1 and Σ_2 , then any critical point must lie on the straight line passing through the component means.

To see this, assume x is a critical point of p and

decompose x as $x = \bar{d} + z$, where \bar{d} is the orthogonal projection of x onto the line determined by the mean vectors, and $z^T \bar{d} = 0$. For some t then, $\bar{d} = t\mu_1 + (1-t)\mu_2$. Thus,

$$\Sigma_1^{-1}(x - \mu_1) = c \Sigma_2^{-1}(x - \mu_2)$$

yields

$$\Sigma_1^{-1}[(1-t)(\mu_2 - \mu_1) + z] = c \Sigma_2^{-1}[(-t)(\mu_2 - \mu_1) + z]$$

or

$$(c\Sigma_2^{-1} - \Sigma_1^{-1})z = \left[\frac{1-t}{\lambda_1} + c \frac{t}{\lambda_2}\right](\mu_2 - \mu_1),$$

where λ_1 and λ_2 are eigenvalues of Σ_1 and Σ_2 respectively. By the construction of z and the properties of eigenspaces of symmetric, positive definite matrices, z lies in a subspace which is invariant under both Σ_1 and Σ_2 , or under their inverses, and $\mu_2 - \mu_1$ is orthogonal to this subspace. Therefore $z = 0$.

In other words, for many two-component mixtures the modality of the mixture is completely characterized by the modality of the one-dimensional conditional density of points lying on the line passing through the two mean vectors. We used this observation to motivate a procedure for generating mixtures in various dimensions which would be equivalent as far as separation and definition of the modes are concerned. The procedure begins with a prototype one-dimensional density, and for each dimension n assumes that

the covariance matrices $\Sigma_1(n)$ and $\Sigma_2(n)$ are fixed in advance, as well as the orientation of the mean vectors, which we will write as

$$d(n) = \frac{\mu_2(n) - \mu_1(n)}{\|\mu_2(n) - \mu_1(n)\|}.$$

The mixture density is written in the form

$$p(x) = \sum_{i=1}^2 \frac{\alpha_i(n)}{(2\pi)^{n/2} \sigma_i(n)^n |\Sigma_i(n)|^{1/2}} \exp\left\{-\frac{1}{2\sigma_i(n)^2} (x - \mu_i(n))^T \Sigma_i(n)^{-1} (x - \mu_i(n))\right\}, \quad (1)$$

and we restrict our attention to the line connecting the means, parameterized as

$$\phi(\rho) = \frac{\mu_1(n) + \mu_2(n)}{2} + \rho \delta \frac{[\mu_2(n) - \mu_1(n)]}{2},$$

where for the moment the precise specification of the scale factor δ is left open.

With these conventions,

$$\begin{aligned} (2\pi)^{n/2} p(\phi(\rho)) &= \\ &= \frac{\alpha_1(n)}{\sigma_1(n)^n |\Sigma_1(n)|^{1/2}} \exp\left\{-\frac{1}{2\sigma_1(n)^2} (\rho\delta+1)^2 \Delta(n)^T \Sigma_1(n)^{-1} \Delta(n)\right\} \\ &+ \frac{\alpha_2(n)}{\sigma_2(n)^n |\Sigma_2(n)|^{1/2}} \exp\left\{-\frac{1}{2\sigma_2(n)^2} (\rho\delta-1)^2 \Delta(n)^T \Sigma_2(n)^{-1} \Delta(n)\right\}, \end{aligned}$$

with

$$\Delta(n) = \frac{\mu_2(n) - \mu_1(n)}{2}$$

$$= Wd, \quad W = \|\mu_2(n) - \mu_1(n)\| / 2.$$

The mixture parameters over which we have control are $\alpha_1(n)$, $\alpha_2(n)$, $\sigma_1(n)$, $\sigma_2(n)$, and W . It is possible to formulate the above expression as an (unnormalized) one-dimensional density, namely

$$p(\phi(\rho)) = (2\pi)^{-1/2} \left\{ \frac{\alpha_1'}{\sigma_2'} e^{-\frac{1}{2} \left(\frac{\rho + \frac{1}{\delta}}{\sigma_1'} \right)^2} + \frac{\alpha_2'}{\sigma_2'} e^{-\frac{1}{2} \left(\frac{\rho - \frac{1}{\delta}}{\sigma_2'} \right)^2} \right\}, \quad (2)$$

where

$$\sigma_i' = \frac{\sigma_i(n)}{W\delta q_i(n)},$$

$$q_i(n) = [d^T \Sigma_i(n)^{-1} d]^{1/2},$$

$$\alpha_i' = \frac{\alpha_i(n)}{(2\pi)^{\frac{n-1}{2}} W\delta q_i(n) |\Sigma_i(n)|^{1/2} \sigma_i(n)^{n-1}}.$$

The number of parameters may be reduced to three by taking

$$\delta = \frac{1}{W} \left[\frac{\sigma_1(n)}{q_1(n)} \frac{\sigma_2(n)}{q_2(n)} \right]^{1/2}.$$

With this choice,

$$\sigma_1' = \left[\frac{\sigma_1(n)}{\sigma_2(n)} \frac{q_1(n)}{q_2(n)} \right]^{1/2} = 1/\sigma_2',$$

$$\alpha'_1 = \frac{\sigma'_1 \alpha_1(n)}{(2\pi)^{\frac{n-1}{2}} \sigma_1(n)^n |\Sigma_1(n)|^{1/2}}, \quad (3)$$

$$\alpha'_2 = \frac{(1/\sigma'_1) \alpha_2(n)}{(2\pi)^{\frac{n-1}{2}} \sigma_2(n)^n |\Sigma_2(n)|^{1/2}},$$

and $p(\phi(\rho))$ has a quasi-symmetrical form. Accordingly, we choose our prototype densities to have means evenly spaced about the origin and variance which are multiplicative inverses. Thus, a prototype will be of the form

$$\sqrt{2\pi} p^*(\rho) = \frac{\alpha_1}{s} e^{-\frac{1}{2} \left(\frac{\rho+\mu}{s}\right)^2} + \frac{\alpha_2}{1/s} e^{-\frac{1}{2} \left(\frac{\rho-\mu}{1/s}\right)^2}.$$

We imitate the prototype by matching equation (2) with equation (4).

To match the equations, we first require

$$s = \sigma'_1 = [(\sigma_1(n)/\sigma_2(n)) (q_2(n)/q_1(n))]^{1/2},$$

or

$$\frac{\sigma_1(n)}{\sigma_2(n)} = s^2 \frac{q_1(n)}{q_2(n)}.$$

Note that the ratio $q_1(n)/q_2(n)$ is fixed by the choices of d , $\Sigma_1(n)$, and $\Sigma_2(n)$. Secondly, we want

$$\alpha'_1/\alpha'_2 = \alpha_1/\alpha_2.$$

From the equations (3) then,

$$\frac{\alpha_1}{\alpha_2} = (\sigma_1')^2 \frac{\alpha_1(n)}{\alpha_2(n)} \left[\frac{\sigma_2(n)}{\sigma_1(n)} \right]^n \left[\frac{|\Sigma_2(n)|}{|\Sigma_1(n)|} \right]^{1/2} ;$$

hence

$$\frac{\alpha_2(n)}{\alpha_1(n)} = \frac{\alpha_2}{\alpha_1} s^{2(1-n)} \left[\frac{q_2(n)}{q_1(n)} \right]^n \left[\frac{|\Sigma_2(n)|}{|\Sigma_1(n)|} \right]^{1/2} .$$

Defining $\beta = \alpha_2(n)/\alpha_1(n)$, we take

$$\alpha_1(n) = 1/(1+\beta),$$

$$\alpha_2(n) = 1 - \alpha_1(n) = \beta/(1+\beta).$$

Thus the mixing proportions $\alpha_1(n)$ and $\alpha_2(n)$ are determined by

- (i) the mixing proportions and variance parameter of the prototype density,
- (ii) the variance-covariance matrices of the multivariate Gaussians, and
- (iii) the orientation of the difference between the means, d .

The mean separation W , as seems natural, is a function of the mean separation in the prototype mixture. To match the prototype we require, now, that

$$\delta = \frac{1}{W} \left[\frac{\sigma_1(n)\sigma_2(n)}{q_1(n)q_2(n)} \right] = \frac{1}{\mu} .$$

Substituting (5), we get

$$W = \mu s (\sigma_2(n)/q_2(n)) = (\mu/s) (\sigma_1(n)/q_1(n)) .$$

There is a free choice of the ratio $\sigma_2(n)/q_2(n)$. Suppose $\sigma_2(n)/q_2(n) = cs$. Then

$$W = \mu c; \quad \sigma_2(n) = cq_2(n)/s; \quad \sigma_1(n) = scq_1(n).$$

To recapitulate, the procedure for generating equivalent mixtures in arbitrary dimension is as follows:

- 1) choose a prototype two-component univariate Gaussian mixture, $p^*(\rho)$.
- 2) for each dimension n ,
 - a) select $\Sigma_1(n)$, $\Sigma_2(n)$, d , and calculate

$$q_i(n) = [d^T \Sigma_i(n)^{-1} d]^{1/2}, \quad i = 1, 2.$$

- b) fix a constant c and take

$$\sigma_2(n) = (c/s)q_2(n)$$

$$\sigma_1(n) = scq_1(n)$$

$$W = \|\mu_2(n) - \mu_1(n)\| / 2 = \mu c.$$

- c) calculate

$$\beta = \frac{\alpha_2}{\alpha_1} s^{2(n+1)} \left[\frac{q_2(n)}{q_1(n)} \right]^n \left[\frac{|\Sigma_2(n)|}{|\Sigma_1(n)|} \right]^{1/2}$$

and take

$$\alpha_1(n) = 1/(1+\beta), \quad \alpha_2(n) = 1-\alpha_1(n).$$

Classification Analogue

As a commentary on the above generation procedure, we may consider its relationship in specific cases to the two-category classification problem. If it is known that

observations may arise from either of two populations, each normally distributed with identical covariance matrices, then the optimal discriminant function is linear, and has the form $W^T(x-x_0)$, where

$$W = \Sigma^{-1}(\mu_1 - \mu_2) ,$$

$$x_0 = \frac{\mu_1 + \mu_2}{2} - \frac{\ln(P(\omega_1)/P(\omega_2))}{(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)} (\mu_1 - \mu_2) .$$

$P(\omega_1)$ and $P(\omega_2)$ are prior probabilities, analogous to our mixing proportions. The probability of classification error is

$$P_e = P\{W^T(x-x_0) < 0 \mid \omega_1\}P(\omega_1) + \\ + P\{W^T(x-x_0) > 0 \mid \omega_2\}P(\omega_2) .$$

Now, since $W^T x \sim \mathcal{N}(W^T \mu_1, W^T \Sigma W)$,

$$P\{W^T(x-x_0) < 0 \mid \omega_1\} = \frac{1}{\sqrt{2\pi} (W^T \Sigma W)^{1/2}} \int_{-\infty}^{W^T x_0} e^{-\frac{1}{2} \frac{(t-W^T \mu_1)^2}{W^T \Sigma W}} dt \\ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{W^T(x_0-\mu_1)}{(W^T W)^{1/2}}} e^{-u^2/2} du .$$

Substituting for W and x_0 , we get

$$P\{W^T(x-x_0) < 0 \mid \omega_1\} = P\{z < -\frac{1}{2} (Q + \frac{\ln(P(\omega_1)/P(\omega_2))}{Q})\} ,$$

where

$$Q = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) .$$

Similarly,

$$P\{W^T(x-x_0) > 0 | \omega_2\} = P\{z > \frac{1}{2} (Q - \frac{\ln(P(\omega_1)/P(\omega_2))}{Q})\} .$$

In both cases, z represents a standard normal random variable. Thus,

$$\begin{aligned} P_e = P(\omega_2) \{ & 1 - \Phi[\frac{1}{2}(Q - \frac{\ln(P(\omega_1)/P(\omega_2))}{Q})] \\ & + \frac{P(\omega_1)}{P(\omega_2)} \Phi[-\frac{1}{2}(Q + \frac{\ln(P(\omega_1)/P(\omega_2))}{Q})] \} . \end{aligned}$$

The probability of misclassification remains constant if Gaussian mixtures are generated so that $P(\omega_1)$ and $P(\omega_2)$ are constant, and Q remains constant.

The requirement of identical covariance matrices implies that $\sigma_1(n), \Sigma_1(n) = \sigma_2(n), \Sigma_2(n)$. For simplicity we may take $\Sigma_1(n) = \Sigma_2(n)$, in which case also $q_1(n) = q_2(n)$. Choosing the prototype to have equal variances, or taking $s = 1$, yields

$$p^*(\rho) = (2\pi)^{-1/2} [\alpha_1 e^{-(\rho+\mu)^2/2} + \alpha_2 e^{-(\rho-\mu)^2/2}] .$$

From equations (3), it follows that $\alpha_1(n)/\alpha_2(n) = \alpha_1/\alpha_2$. Also, observing $W = \mu c = \mu/q_1(n)$,

$$\begin{aligned}
Q &= (\mu_1(n) - \mu_2(n))^T \Sigma(n)^{-1} (\mu_1(n) - \mu_2(n)) \\
&= 4(Wd)^T \Sigma(n)^{-1} (Wd) \\
&= 4W^2 q_1(n)^2 \\
&= 4\mu^2.
\end{aligned}$$

Thus, both the mixing proportions and the quantity Q are, as desired, dimensionally invariant. Therefore, at least in the case where a linear discriminant function is efficient, and measuring the separation of the mixture components by the optimal classification error rate, the mixture densities generated by the procedure described in this appendix are comparable irrespective of dimension.

REFERENCES

- Andriano, K., Gentle, J., and Sposito, V. A. (1978). Comparisons of some estimators of the mode. Proceedings of the Social Statistics Section of the American Statistical Association, p. 760-764.
- Baggett, L. and Fulks, W. (1979). Fourier Analysis. Anjou Press, Boulder, Co.
- Behboodian, J. (1970). On the modes of a mixture of two normal distributions. Technometrics, Vol. 12, No. 1, p. 131-139.
- Bochner, S. (1955). Harmonic Analysis and the Theory of Probability. U. of California Press.
- Breiman, L., Meisel, W., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. Technometrics, V. 19, No. 2, p. 135-144.
- Brent, R. P. (1973). Algorithms for Minimization without Derivatives. Prentice-Hall, Englewood Cliffs, NJ.
- Cacoullos, T. (1966). Estimation of a multivariate density. Ann. Math. Stat., Tokyo, V. 18, p. 179.
- Chernoff, H. (1964). Estimation of the mode. Annals of the Institute of Statistical Mathematics 16, p. 31-41.
- Dalenius, T. (1964). The mode -- a neglected statistical parameter. J.R. Statist. Soc. Ser. A, V. 128, p. 110-117.
- Dennis, J. E. and Schnabel, R. (1983). Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, Englewood Cliffs, N.J.
- Duin, R. P. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Trans. Comp., C-125, p. 1175.
- Dutter, R. (1975). Robust regression: Different approaches to numerical solutions and algorithms. Res. Rep. No. 6, Fachgruppe fur Statistik, Eidgen. Technische Hochschule, Zurich.
- Eisenberger, I. (1969). Genesis of bimodal distributions. Technometrics, V. 6, No. 4, p. 357-363.

- Ekblom, H. (1972). A Monte Carlo investigation of mode estimators in small samples. Applied Statistics 21, p. 177-184.
- Epanechnikov, V. A. (1969). Nonparametric estimates of a multivariate probability density. Theor. Prob. Appl., V. 14, p. 153.
- Factor, L. E. (1979). Comparison of data-based methods for non-parametric density estimation. Masters dissertation at Rice University, Houston, Texas.
- Fwu, C., Tapia, R. A., and Thompson, J. R. (1980). The nonparametric estimation of probability densities in ballistics research. Proceedings of the Twenty-Sixth Conference on the Design of Experiments in Army Research and Testing.
- Good, I. J. and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. J.A.S.A., V. 75, No. 369, p. 42-73.
- Grenander, U. (1965). Some direct estimates of the mode. Ann. Math. Statist. 36, p. 131-138.
- Hartigan, J. A. (1975). Clustering Algorithms. John Wiley and Sons, NY.
- Huber, P. J. (1981). Robust Statistics. John Wiley and Sons, NY.
- Kennedy, W. J. and Gentle, J. E. (1980). Statistical Computing. Marcel Dekker, NY.
- Konstantellos, A. C. (1980). Unimodality conditions for Gaussian sums. IEEE Transactions on Automatic Control, AC-25, No. 4, p. 838.
- Lewis, T. G. and Payne, W. H. (1973). Generalized feedback shift register pseudorandom number algorithm. J. Ass. Computing Machinery, V. 20, No. 3, p. 456-468.
- Loftsgaarden, D. O. and Quesenberry, C. P. (1965). A non-parametric estimate of a multivariate density function. Ann. Math. Statist., V. 36, p. 1049-1051.
- Mack, Y. P. and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. J. Multivariate Analysis, V. 9, p. 1-15.

- Moore, D. S. and Yackel, J. W. (1977). Consistency properties of nearest neighbor density function estimates. Ann. Statist., V. 5, p. 143-154.
- Nezames, D. D. (1980). Some results for estimating bivariate densities using kernel, orthogonal series and penalized likelihood procedures. Doctoral dissertation, Rice University, Houston, Texas.
- Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist., V. 33, p. 1065-1076.
- Robertson, T. and Cryer, J. (1974). An iterative procedure for estimating the mode. J.A.S.A., V. 69, p. 1012-1016.
- Rubinstein, R. Y. (1981). Simulation and the Monte Carlo Method. John Wiley and Sons, NY.
- Sager, T. W. (1975). Consistency in nonparametric estimation of the mode. Ann. Statist., V. 3, p. 698-706.
- Scott, D. W. (1976). Nonparametric probability density estimation by optimization theoretic techniques. Doctoral dissertation, Rice University, Houston, Texas.
- Silverman, B. W. (1978). Choosing the window width when estimating a density. Biometrika, V. 65, p. 1-11.
- Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. J. R. Statist. Soc. Ser. B, V. 43, No. 1, p. 97-99.
- Singh, R. S. (1976). Nonparametric estimation of mixed partial derivatives of a multivariate density. J. Multivariate Analysis, V. 6, p. 111-122.
- Singh, R. S. (1979). On necessary and sufficient conditions for uniform strong consistency of estimators of a density and its derivatives. J. Multivariate Analysis, V. 9, p. 157-164.
- Tapia, R. A. and Thompson, J. R. (1978). Nonparametric Probability Density Estimation. The Johns Hopkins University Press, Baltimore.
- Van Ryzin, J. (1969). On strong consistency of densities estimates. Ann. Math. Statist., V. 40, p. 1765-1772.

- Wahba, G. (1977). Optimal smoothing of density estimates. Classification and Clustering, J. Van Ryzin, ed., Academic Press, NY.
- Wegman, E. J. (1971). A note on the estimation of the mode. Ann. Math. Statist., V. 42, p. 1909-1915.
- Wilde, D. J. (1964). Optimum Seeking Methods. Prentice-Hall, Englewood Cliffs, NY.
- Wolfe, P. (1969). Convergence conditions for ascent methods. SIAM Review 11, p. 226-235.
- Wolfe, P. (1971). Convergence conditions for ascent methods II: Some corrections. SIAM Review 13, p. 185-188.