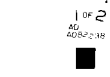
AD-A082 238 MASSACHUSETTS INST OF TECH CAMBRIDGE DEPT OF MATHEMATICS F/G 12/1
ITERATIVE ELLIPSOIDAL TRIMMING.(U)
FEB 80 L S GILLICK N00014-75-C-0555
UNCLASSIFIED TR-15 NL

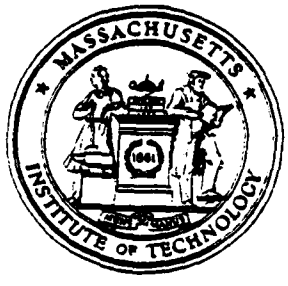


ITERATIVE ELLIPSOIDAL TRIMMING

BY

LAURENCE S. GILLICK
DEPARTMENT OF MATHEMATICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

DEPARTMENT OF MATHEMATICS
NORTHEASTERN UNIVERSITY



TECHNICAL REPORT NO. 15
FEBRUARY 11, 1980

PREPARED UNDER CONTRACT
N00014-75-C-0555 (NR-042-331)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF MATHEMATICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS

(C) ITERATIVE ELLIPSOIDAL TRIMMING,

by

(14) TR 121

(10) Laurence S. Gillick
Department of Mathematics
Massachusetts Institute of Technology

Department of Mathematics
Northeastern University

(15) N

ITERATIVE ELLIPSOIDAL TRIMMING

by

Laurence S. Gillick

Department of Mathematics
Massachusetts Institute of Technology

Department of Mathematics
Northeastern University

ABSTRACT

The iterative ellipsoidal trimming algorithm is introduced as both a clustering method and an estimator of location and shape. Its power as a data analytic tool is investigated and the asymptotic distribution of its stationary point is derived. In addition, several scale estimators are proposed and studied.

Some key words: asymptotics; cluster analysis; ellipscidal trimming; scale estimation

AMS 1970 subject classification: Primary 62E20; secondary 62H30

Acknowledgements

I would like to thank Professor Herman Chernoff for his encouragement, his suggestions, and his criticism during the preparation of this dissertation, and Ms Phyllis Ruby for a careful job of typing.

Table of Contents

Chapter 1

Introduction

It is not uncommon in scientific or technological work for an investigator to be confronted with a collection of entities and for him then to wonder whether that collection is, in some sense, homogeneous or whether, instead, it is made up of several distinct subgroups. That branch of statistics known as cluster analysis is concerned with providing a body of techniques which will be generally useful in discovering subpopulations. It is to this subject that this dissertation seeks to make a contribution.

The central aim of this work is to introduce what we shall refer to as the iterative ellipsoidal trimming algorithm (for brevity's sake, IET) as a method for discovering clusters and to study some of its properties. We define IET as follows. Suppose that we have n observations in R^k : $X_1,\ldots,X_n$. To start the algorithm, initial estimates (starting values) of the mean and covariance of the cluster being sought must be provided: $(\tilde{\mu}_0, \tilde{Z}_0)$. Sometimes, when we are in complete ignorance of the distribution of the X's, it is appropriate to let $\tilde{\mu}_0 = \bar{X}$; in other situations $\tilde{\mu}_0$ may be derived from previous analysis or it may be an arbitrary point in a certain region of R^k. Usually, $\tilde{Z}_0$ is taken to be I_k,

the $k \times k$ identity matrix. To perform an iteration, one specifies a p, which represents the proportion of the observations to be included in the computation of the next estimates, $(\tilde{\mu}, \tilde{\Sigma})$, and calculates the Mahalanobis distance of each X_i from $\tilde{\mu}_0$: $D_i^2 = (X_i - \tilde{\mu}_0)'\tilde{\Sigma}_0^{-1}(X_i - \tilde{\mu}_0)$. Then,

$$\tilde{\mu} = [np]^{-1} \sum_{i \in L} X_i$$

and

$$\tilde{\Sigma} = [np]^{-1} \sum_{i \in L} (X_i - \tilde{\mu})(X_i - \tilde{\mu})'$$

where $L = \{i : D_i^2 \leq D^2_{([np])}\}$, $D^2_{(r)}$ is the r^{th} order statistic of the D_i^2's, and $[t]$ is the greatest integer $\leq t$. Of course, one performs the next iteration by again choosing a p and then treating $(\tilde{\mu}, \tilde{\Sigma})$ as the new $(\tilde{\mu}_0, \tilde{\Sigma}_0)$. We will say that IET has converged (for fixed p) if on two successive iterations we find that L, the set of indices, does not change. Equivalently, we will say that IET has converged if $(\tilde{\mu}, \tilde{\Sigma})$ stays the same on two successive iterations. It is appealing to call this final estimate a stationary point (of the sample). Sometimes, it is too time consuming to wait for IET to converge; then it is reasonable to simply continue until successive changes in the estimates $(\tilde{\mu}, \tilde{\Sigma})$ are sufficiently small.

In successive iterations of the algorithm, p may be allowed to change or it may be kept constant; often this decision may fruitfully be made interactively, that is to say, after looking at the last $(\tilde{\mu}, \tilde{\Sigma})$. The choice of a sequence of p's will depend on the goal of the analysis. We will have two separate but closely related intentions in mind. First, we wish to find large clusters and second, having found a large cluster, we wish to obtain robust estimates of its mean and covariance, (μ, Σ), with the idea of using them in the search for smaller clusters in the tails of the large one. We hope to discuss the problem of finding such "hidden" clusters in a subsequent paper. It is pertinent to remark now, however, that we are especially interested in IET because it appears to yield a plausible estimator when there is asymmetric contamination (for instance, a small cluster) in the tails of the sample.

In chapter 2 we deal with the problem of using IET to discover clusters. There we also examine a certain lack of robustness characteristic of "k-means" algorithms, and how IET avoids this difficulty. Chapter 3 contains a discussion of the asymptotic properties of IET, always supposing that there are no outliers (isolated points far away from the bulk of the observations) or small hidden clusters. Instead, the data will be assumed to be purely

from some spherically symmetric or ellipsoidal distribution. The fourth chapter presents Monte Carlo results concerning the performance of IET in the presence of outliers and small clusters.

Finally, in chapter 5 we take up the question of scale estimation. When data that are far from $(\tilde{\mu}, \tilde{\Sigma})$ are trimmed, inevitably the next estimate of Σ is biased towards 0; basically, $\tilde{\Sigma}$ is an estimate of $c\Sigma$ for some c such that $0 < c < 1$. Now this is a matter of no consequence when we are only interested in finding the location and shape of a cluster. Furthermore, IET is unaffected when $\tilde{\Sigma}$ is multiplied by a constant, since the ordering of the Mahalanobis distances is unchanged. However, when we want to compare the D^2 of an X to two different clusters, it is imperative that we "scale up" our estimates of their Σ's. We must, in essence, estimate and divide out the "c" referred to above.

Iterative ellipsoidal trimming has been investigated before by other statisticians, most notably by Gnanadesikan and his coworkers Kettenring and Devlin [5, 7, 8]. The focus in the past, however, has been on Monte Carlo studies of the utility of this algorithm in the robust estimation of covariance matrices. It has been shown to be reasonably reliable for this purpose [5].

Chapter 2

Finding Clusters

In this chapter we shall discuss and illustrate the use of IET in discovering clusters in a data set, and carry out a comparison with a natural competitor, the k-means algorithm. It is expected that IET will be a useful tool when one wishes to find ellipsoidal clusters in a high dimensional Euclidean space (where by high we mean "greater than two"). The basic rationale behind its use is that it will tend to climb up density gradients, stopping when it reaches an ellipsoidal region containing [np] observations whose sample mean is just its center, and whose sample covariance will generate its "shape".

There are several k-means algorithms, the different versions differing as to whether or not a covariance matrix is estimated for each cluster and as to how many observations are reclassified before the k means (and possibly, covariances) are updated. We will consider two versions of this general method; both of them update the means and covariances after reclassifying all of the data. The following algorithm will be referred to as k-means 2. Suppose $X_1,\ldots,X_n$ are our observations and $(\tilde{\mu}_1,\tilde{Z}_1),\ldots,(\tilde{\mu}_k,\tilde{Z}_k)$ are our current estimates of the means and covariances of k clusters. Let

$D_{ij}^2 = (X_i - \tilde{\mu}_j)'\tilde{\Sigma}_j^{-1}(X_i - \tilde{\mu}_j)$ and classify X_i into the j_0^{th} cluster if

$$D_{ij_0}^2 = \min_{1 \le j \le k} D_{ij}^2.$$

Let $L_j = \{i : X_i$ is classified in the j^{th} cluster$\}$. Then, the next estimates are

$$\tilde{\mu}_j^{(1)} = |L_j|^{-1} \sum_{i \in L_j} X_i$$

and

$$\tilde{\Sigma}_j^{(1)} = |L_j|^{-1} \sum_{i \in L_j} (X_i - \tilde{\mu}_j^{(1)})(X_i - \tilde{\mu}_j^{(1)})',$$

(where $|S|$, for a set S, is the number of elements S contains), or, to put it another way, the sample means and covariances of the new clusters. If we fix $\tilde{\Sigma}_j = I$, the identity of the appropriate dimension, for all j and update only the means after reclassifying by Euclidean distance, then the above algorithm reduces to what we shall call k-means 1. For further discussion of k-means algorithms, one might refer to either Hartigan's book on clustering algorithms [9] or the volume by Duda and Hart on pattern recognition [6]. Chernoff appears to have been the first to suggest that

it would be helpful to estimate the covariances of the different clusters [3], though Rohlf [11] made a similar recommendation in the context of hierarchical clustering.

The basic idea which underlies both IET and k-means is that if one has an approximation to the mean of a cluster and one computes a new estimate of the mean based only on points currently thought to belong to the cluster (based on the approximation), then the new estimate will be better than the old one. In one dimension, in the case of IET, one imagines a situation like that in Figure 2.1. There, the new estimate is the mean of all the observations contained between the brackets and will clearly be closer to the true mean than was the starting value.

Now there are three important ways in which k-means and IET differ. IET defines the current cluster in a rather conservative fashion, namely as those points within some D^2 of the current $(\tilde{\mu}, \tilde{Z})$. On the other hand, k-means defines the j^{th} cluster as everything that is closer to $(\tilde{\mu}_j, \tilde{Z}_j)$ than to any of the other cluster centers. Hence, a k-means cluster may have unbounded volume. The second basic difference is that when using IET, the statistician specifies the proportion of data points to be included within the ellipsoidal "window" whose location, shape, and orientation the algorithm computes. Of course,

k-means lets the proportions vary according to the results of its classification scheme. Finally, and by definition, IET only searches for one cluster at a time, while k-means tries to locate k clusters simultaneously.

Several important consequences follow from these remarks. Since a k-means cluster can be unbounded, even if the starting value for a given cluster is very good, one iteration can actually carry the cluster center far away from the true center, if some outlying observation happens to be newly classified into that cluster. An outlier can, in essence, take hold of a perfectly good cluster and, in some cases, ultimately have it all to itself. In the case of k-means 2, a more subtle and, in some respects, a more disturbing occurrence is possible. If an outlier or an observation from another population is misclassified into a cluster, it will tend to increase the estimate of its scale, perhaps by a considerable amount. Since $\tilde{Z}$ will then be quite "large", Mahalanobis distances to that cluster will tend to be small; as a result, this cluster will tend to absorb all of the other clusters. This latter point brings to mind our earlier comment that k-means does not allow the statistician to control the number of data points in a cluster.

Since IET only looks at observations within a bounded ellipsoidal region that is forced to contain a certain number of data points, it is relatively immune to these sorts of robustness problems. It may be objected that to ask the statistician to specify p (essentially, the size of the region) is to ask too much, as he may have no prior information about the distribution of the sample. But it is our view that IET is an exploratory tool (in the sense that Tukey uses this expression [12]) and it must be used in an exploratory spirit. One specifies a sequence of p's and sees what the algorithm does, that is, where the sequence of estimates, $(\tilde{\mu}, \tilde{\Sigma})$, goes; one chooses a new starting value and a new sequence of p's and observes again. If there is an ellipsoidal cluster to be found and p is taken to be sufficiently small (no bigger than the proportion of points in the cluster), then our experience suggests that IET is likely to find it. The basic rule of thumb is that if several different starting points and different sequences of p's lead to convergence to approximately the same place, then one should suspect that a cluster has been found. One should then try increasing p, using the found cluster as a starting point, to get a feel for how big the cluster might be - if one increases p from 0.4 to 0.6 without changing the estimate of location, $\tilde{\mu}$, very much, then this finding

both reassures us that the cluster is real and provides a bound on its size. Of course it will then be important to make plots and perhaps use other devices to make sure that the sample really does have a mode where we think it does. One situation to beware of can arise when p is taken to be bigger than any cluster in the sample and is best described by the diagram in Figure 2.2. Here, each of the two clusters, one might imagine, contains 50% of the observations and p was set at 70%. Of course, given that one was seeking a cluster with 70% of the data, this result (the distribution in the box) isn't so bad.

A further important caveat concerns the situation when [np] is very small, for thenwe are likely to obtain very unreliable results. IET may then have points of convergence at many locations of no interest, just because of the granularity of the distribution at that "window" size (where by "window" size we just mean the number of observations inside the ellipsoid). One wants the ellipsoidal window to be big enough so that when we look through it, we can distinguish the trend, or signal, from the noise.

Next we will consider several examples so as to illustrate a number of the above generalities. First we review an example on page 195 of Duda and Hart [6]. A sample, consisting of 8 N(-2,1) and 17 N(2,1) observations

is to be clustered. We find that k-means 1 converges from a variety of starting values to two clusters with means -2.18 and 1.68 (we are assuming that we know that there are two clusters). This happens to be a perfect clustering, a result made possible by the fact that the data from the two distributions happen to be nonoverlapping. We find that k-means 2 also converges to these two clusters from "good" starting values like -2,2, but not from "bad" starting valueslike -2,7. Here, however, we will not focus on the relevance of good starting values, but instead on the question of robustness against outliers.

Suppose now that one additional observation, at x = 25, is added to the sample. We set the starting values at -2,2 and apply k-means 1, obtaining the results in Table 2.1. Obviously, what has happened is that the outlier at x = 25 now has the second cluster all to itself.

Next we apply k-means 2 to the same data. Since we are working in one dimension we will write $\tilde{\sigma}_j^2$ for $\tilde{\Sigma}_j$. Taking the starting values to be the population parameters, we obtain the successive estimates in Table 2.2. There, when the variance of cluster 2 becomes large, most points become "close" to it and "far" from cluster 1, whose variance gets ever smaller. What we are observing here

is an inherent instability in the k-means 2 algorithm.

The point of the preceding example is that one outlier can completely throw off our search for large clusters. In general, data with many outliers will lead to a k-means clustering which assigns the bulk of the centrally located data to one cluster and the rest to however many other clusters there are. Next we examine a more interesting example and compare the performance of IET and k-means.

We generate a sample of 300 $N(\mu_1, I_2)$, 300 $N(\mu_2, I_2)$, 300 $N(\mu_3, I_2)$, and 100 $N(\mu_1, 400I_2)$ observations where $\mu_1 = (0,0)'$, $\mu_2 = (6, 0)'$, and $\mu_3 = (12, 0)'$, and attempt to cluster it. When k-means 1 is applied to these data using starting values (1, 0)', (5, 0)', (14, 0)', it converges to three clusters whose means are (2.11, 0.30)', (5.80, 0.25)', (13.13, 0.31)', which is a relatively satisfying result. Unfortunately, if the three starting values are (-3, 0)', (2, 0)', (7, 0)', k-mean 1 converges to three clusters whose means are (-23.48, 3.55), (2.28, 0.19), (12.21, 0.21) and the first of the clusters contains only 27 points.

When k-means 2 is applied to the same data, using the population parameters for the starting values of the three clusters, the algorithm yields, on the fourth

iteration, these means and covariances:

$$\begin{pmatrix} 3.13 \\ 0.37 \end{pmatrix}, \begin{pmatrix} 6.87 & 4.22 \\ 4.22 & 6.03 \end{pmatrix}$$

$$\begin{pmatrix} 5.97 \\ -0.11 \end{pmatrix}, \begin{pmatrix} 10^{-3} & 2 \times 10^{-3} \\ 2 \times 10^{-3} & 8 \times 10^{-3} \end{pmatrix}$$

$$\begin{pmatrix} 11.98 \\ 0.07 \end{pmatrix}, \begin{pmatrix} 0.74 & -0.17 \\ -0.17 & 0.96 \end{pmatrix}$$

where the three clusters contain 723, 4, and 273 observations, respectively. The third cluster is quite good but the first cluster has absorbed practically all of the second one, as well as almost all of the outliers. On the fifth iteration, the rest of cluster 2 is absorbed by cluster 1. If the algorithm is allowed to continue, using only two clusters now, cluster 3 will also be gradually absorbed by cluster 1. What happens is that the outer edges of cluster 3 are nibbled away bit by bit; at each stage its variances are thereby reduced making all of its remaining points closer to cluster 1.

Suppose now that we undertake a search for these three clusters, using IET. It would be natural to take the sample mean and covariance of all of the data as a

starting value. Then, 3 iterations of IET, with p set equal to 0.3, lead to the estimates

$$\begin{pmatrix} 5.94 \\ -0.05 \end{pmatrix}, \begin{pmatrix} 1.08 & -0.07 \\ -0.07 & 0.92 \end{pmatrix}$$

which are excellent. If we take $\tilde{\mu} = (1, 2)'$ and $\tilde{\Sigma} = I_2$ as starting values and set $p = 0.3$, then IET computes the sequence of means: $(0.31, 0.15)'$, $(0.1, 0.04)'$, $(0.05, -0.01)'$ and stops with the final covariance estimate being

$$\begin{pmatrix} 0.9 & 0.02 \\ 0.02 & 0.94 \end{pmatrix}.$$

When, again, $p = 0.3$ and the starting values are $\tilde{\mu} = (10, 10)'$, $\tilde{\Sigma} = I_2$, the sequence of sample means is $(9.92, 1.38)'$, $(10.95, 0.34)'$, $(11.47, 0.2)'$, $11.67, 0.12)'$, $(11.79, 0.09)'$, $(11.87, 0.13)'$, $(11.90, 0.13)'$, $(11.92, 0.13)'$ and the final covariance estimate is

$$\begin{pmatrix} 0.98 & -0.06 \\ -0.06 & 1.10 \end{pmatrix}.$$

In examining the pattern of convergence in the last two

trials, one is struck by the fact that it is similar to that of geometric convergence, although the rate appears to gradually change as the stopping point is approached. In the next chapter an asymptotic analysis of IET will reveal why this phenomenon occurs.

So far p has been set to the "right" value (each cluster containing 30% of the entire sample with an additional 10% contamination by outliers). Therefore, one might worry that our success so far is rather artificial. So set $p = 0.2$ and take $\tilde{\mu} = (4, 4)'$, $\tilde{\Sigma} = I_2$ as starting values. Then the sequence of means is $(4.54, 0.69)'$, $(5.32, 0.45)'$, $(5.82, 0.3)'$, $(5.89, 0.22)'$, $(5.90, 0.13)'$, $(5.91, 0.10)'$, $(5.92, 0.10)'$, $(5.93, 0.09)'$, $(5.93, 0.08)'$. In general, setting p at too small a value will not be damaging, and we see in the preceding trial that the rate of convergence was not even substantially affected by the fact that p was only 2/3 of the true value for the cluster being sought.

On the other hand, when p is bigger than the proportion of points in the cluster, it is possible to encounter difficulties. For instance, if we again work with the same data, set $p = 0.5$ and use the sample mean and covariance of all the observations as starting values, after 11 iterations IET converges to a cluster with mean

and covariance:

$$\begin{pmatrix} 3.00 \\ 0.09 \end{pmatrix}, \begin{pmatrix} 9.98 & -0.29 \\ -0.20 & 0.48 \end{pmatrix}.$$

Furthermore, upon examining the 500 points included in this final cluster, we learn that 251 are from the $N((0, 0)', I_2)$ distribution, 246 are from the $N((6, 0)', I_2)$ distribution, and the other 3 are outliers (from the $N((0, 0)', 400I_2)$ distribution). This cluster is described in Figure 2.3.

We do not claim that the remarks we have made about k-means are novel and indeed it may be objected that some k-means algorithms currently in use have already been immunized against at least some of our criticisms. For instance, by allowing the introduction of a new cluster when an observation is "too far" from all of the current clusters, we can prevent outliers from taking hold of large clusters and pulling them far away from the bulk of the data. It may even be possible, through the use of other ad hoc addenda to the k-means algorithm, to eliminate the instability in k-means 2. Maronna and Jacovkis [10] compare several different variable metric clustering methods, (including k-means 2 with one observation

reclassified per iteration and conclude, after noting the instability we have pointed out, that the only such method they would recommend is that same k-means 2 but with the covariances normalized by the requirement $|\tilde{Z}_j| = 1$. This method has the weakness that it does not really allow for the possibility that different clusters may have quite different scales. Another approach would be to estimate the covariance using only the central portion of a cluster. (One might then use the methods of chapter 5 to "scale up" those estimates.) Our purpose in discussing k-means has been to highlight the basic differences between it and IET, which is a natural competitor. No doubt experienced users of either algorithm will be able to successfully cluster a sample with genuine ellipsoidal clusters.

Of course IET does not need to be "patched up"; its simplest, most fundamental form is already, in a certain sense, robust. An important dividend is that the form of IET that we propose to use in practice (and have illustrated in this chapter) is sufficiently simple for it to admit fruitful asymptotic analysis, a task which we undertake in the next chapter.

Chapter 3 - Asymptotics

In this chapter we study the asymptotic behavior of IET when there is only one cluster to be found and there are no outliers. The word "asymptotic" is used here in two different senses: we imagine that (1) the sample size is large or infinite and/or (2) that certain parameters describing the algorithm approach limiting values. The first part of the chapter will be concerned with zero order asymptotics, which is to say, how the expectations of certain quantities behave (or, alternatively, how infinite samples behave). Later we shall obtain a first order asymptotic result: the large sample distribution of what we shall call the stationary point of IET.

It will be helpful to introduce some notation at this point. We suppose that $f(x)$ denotes a spherically symmetric density about $x = 0$ in k dimensions, that is, $f(x) = c_k^{-1} g(|x|^2)$, where c_k is the normalizing constant and $|x|^2 = \sum_{i=1}^{k} x_i^2$, and $F(x)$ is the corresponding cumulative distribution function (cdf). Furthermore, $T = X'X$ has the generalized chi-squared distribution with k degrees of freedom with density f_k, given in (A1.1), and cdf F_k. We shall also speak of the ellipsoidal distributions

generated from f : if $X \sim f$, then $Y = \mu + A^{1/2}X$ has an ellipsoidal distribution if A is symmetric and positive definite. We assume that g has the property that the necessary moments exist. Hence $E(Y) = \mu$, and the covariance of Y, as shown in Appendix 1, is $\Sigma = (2\pi c_k)^{-1} c_{k+2} A$. Two examples we shall refer to are the spherical normal with density $f(x) = c_k^{-1} g(x'x)$, where $c_k = (2\pi)^{k/2}$ and $g(t) = \exp(-t/2)$, and the multivariate normal, which is the corresponding ellipsoidal distribution. (In this case $\Sigma = A$). Numerous properties of symmetric and normal densities are collected in Appendices 1 and 2.

We shall denote the sphere of radius a centered at δ by $S_a(\delta) = \{x:(x-\delta)'(x-\delta) \leq a^2\}$. If we define

$$(3.1) \qquad P_a(\delta) = \int_{S_a(\delta)} f(x)dx,$$

and

$$(3.2) \qquad e_a(\delta) = \int_{S_a(\delta)} xf(x)dx,$$

then the conditional mean of X is

$$(3.3) \qquad \mu_a(\delta) = \frac{e_a(\delta)}{P_a(\delta)} = E(X|X \in S_a(\delta)).$$

When the sample size is large and $k = 1$, $\mu_a(\delta)$ is the approximate result of one iteration of IET when δ is the starting $\tilde{\mu}$.

The first problem we shall consider is the following: under what conditions will one iteration of IET improve the estimate of the mean of the distribution no matter what the starting value is, when $k = 1$?

Theorem 3.1: In the univariate case, when $a > 0$ and $|\delta| > 0$,

$$(3.4) \quad |\mu_a(\delta)| \leq |\delta| \quad \text{iff} \quad 0 \leq \frac{\bar{F}_c - \bar{F}_d}{\delta} \leq 2\bar{f}_c$$

where

$$\bar{F}_d = \frac{1}{2}(F(\delta+a) + F(\delta-a)),$$

$$\bar{F}_c = \frac{1}{2a}\int_{\delta-a}^{\delta+a} F(x)dx ,$$

and

$$\bar{f}_c = \frac{1}{2a}\int_{\delta-a}^{\delta+a} f(x)dx.$$

Proof: Observe that

$$\int_s^t xf(x)dx = tF(t) - sF(s) - \int_s^t F(x)dx$$

and

$$\int_s^t f(x)dx = F(t) - F(s).$$

Therefore, putting $s = \delta-a$, $t = \delta+a$, we have

$$\mu_a(\delta) = \frac{(\delta+a)F(\delta+a) - (\delta-a)F(\delta-a) - \int_{\delta-a}^{\delta+a} F(x)dx}{F(\delta+a) - F(\delta-a)}$$

which implies

$$\mu_a(\delta) = \delta - \frac{\bar{F}_c - \bar{F}_d}{\bar{f}_c}.$$

If $\delta > 0$, then $|\mu_a(\delta)| \leq \delta$ iff $0 \leq \bar{F}_c - \bar{F}_d \leq 2\delta\bar{f}_c$. Similarly, if $\delta < 0$, then $|\mu_a(\delta)| \leq -\delta$ iff $0 \geq \bar{F}_c - \bar{F}_d \geq 2\delta\bar{f}_c$. But these results, taken together, are equivalent to (3.4). ∎

The proof of this theorem did not rely on the existence of a density f: we could have replaced f by dF and $\bar{f}_c$ by $(2a)^{-1}\int_{\delta-a}^{\delta+a} dF(x)$. The distribution of F need not be symmetric about 0, although we stated Theorem 3.1 that way

because of our special concern wtih such distributions in this chapter. Finally, we do not even require that F have mean 0. So the theorem can be applied to a finite sample (with empirical distribution function F_n).

Next we will investigate the behavior of $\mu_a(\delta)$ as $\delta \to 0$ while a stays fixed when, again, $k = 1$. In essence we wish to answer the question of how, in one dimension, IET behaves when one iteration is performed and we are near the true mean.

Theorem 3.2: *In the univariate case, if f is continuous in a neighborhood of $x = a$, then if $F(a) > 1/2$,*

$$(3.5) \qquad \mu_a(\delta) = \frac{af(a)}{F(a) - 1/2}\delta + o(\delta) \quad \text{as} \quad \delta \to 0.$$

Furthermore, if f is $n-1$ times differentiable in a neighborhood of a, then $\mu_a^{(n)}(0)$ exists, and vanishes when n is even.

Proof: Since $e_a(0) = 0$, it follows that $\mu_a(0) = 0$. Observe that $e_a'(0) = af(a) + af(-a) = 2af(a)$ and $P_a(0) = 2(F(a) - 1/2)$. Hence,

$$\mu_a'(0) = \frac{e_a'(0)}{P_a(0)} - \frac{e_a(0)P_a'(0)}{P_a^2(0)} = \frac{af(a)}{F(a) - 1/2}$$

which implies (3.5). In general, the n^{th} derivatives of P_a and e_a are

$$(3.6) \qquad P_a^{(n)}(\delta) = f^{(n-1)}(\delta+a) - f^{(n-1)}(\delta-a)$$

and

$$(3.7) \qquad e_a^{(n)}(\delta) = (n-1)[f^{(n-2)}(\delta+a) - f^{(n-2)}(\delta-a)] + [(\delta+a)f^{(n-1)}(\delta+a) - (\delta-a)f^{(n-1)}(\delta-a)].$$

Since f is symmetric about $x = 0$, $f^{(n)}(x) = (-1)^n f^{(n)}(-x)$. Hence, (3.6) and (3.7) imply that

$$(3.8) \qquad p^{(n)} \equiv P_a^{(n)}(0) = \begin{cases} 0 & n \text{ odd} \\ 2f^{(n-1)}(a) & n \text{ even} \end{cases}$$

and

$$(3.9) \qquad e^{(n)} \equiv e_a^{(n)}(0) = \begin{cases} 2(n-1)f^{(n-2)}(a)+2af^{(n-1)}(a) & n \text{ odd} \\ 0 & n \text{ even.} \end{cases}$$

Let $\mu^{(n)} = \mu_a^{(n)}(0)$. Then, since $e_a = P_a \mu_a$,

$$(3.10) \qquad e^{(n)} = \sum_{k=0}^{n} \binom{n}{k} p^{(k)} \mu^{(n-k)}.$$

Now we use an induction argument to prove $\mu^{(2n)} = 0$. Of

course $\mu^{(0)} = 0$. Suppose $\mu^{(0)} = \mu^{(2)} = \ldots = \mu^{2(n-1)} = 0$ and consider the $(2n+1)^{st}$ instance of (3.10). Then,

$$e^{(2n)} = \sum_{k=0}^{2n} \binom{2n}{k} P^{(k)} \mu^{(2n-k)}.$$

Observe that all terms with k odd vanish because of (3.8). Similarly, for $k > 0$ and k even, the corresponding terms vanish by the induction hypothesis. Hence, $e^{(2n)} = P^{(0)} \mu^{(2n)}$. But $e^{(2n)} = 0$ by (3.9) and $P^{(0)} > 0$; hence $\mu^{(2n)} = 0$. ∎

If we put

$$b(a) = \frac{af(a)}{F(a) - 1/2},$$

then the preceding theorem asserts that $\mu_a(\delta) = b(a)\delta + o(\delta)$, or that for δ small, $b(a)$ represents the proportional reduction in bias achieved by performing one iteration of IET. It is interesting that $b(a)$ has a simple geometrical interpretation as the ratio of the density f at the boundary of $[-a, a]$ to the average density of f in $[-a, a]$,

$$(3.11) \qquad b(a) = \frac{f(a)}{\frac{1}{2a}(2(F(a) - 1/2))}.$$

It turns out that $b(a)$ is the $k=1$ version of $b_k(a)$,

the k-dimensional bias reducing function that appears in the next theorem (that is, $b(a) = b_1(a)$). We define

$$(3.12) \qquad b_k(a) = \frac{c_k^{-1} g(a^2)}{F_k(a^2)/V_k(a)}$$

where

$$V_k(a) = \frac{\pi^{k/2}}{\Gamma(\frac{k}{2} + 1)} a^k$$

is the volume of a k-dimensional sphere of radius a, given in (A1.5). The function $b_k(a)$ is the ratio of the density of X on the boundary of $S_a(0)$ to the average density of X inside $S_a(0)$. (Of course, $F_k(a^2) = \Pr(X'X \leq a^2)$.)

In the k-dimensional case, the approximate result of one iteration of IET is not quite given by $\mu_a(\delta)$ because it depends not only on δ but also on the initial estimate of the covariance matrix. Nevertheless the behavior of $\mu_a(\delta)$ expressed in Theorem 3.3 is relevant to our understanding of IET.

Theorem 3.3: If g is differentiable in a neighborhood of a^2, then

$$(3.13) \qquad \mu_a(\delta) = b_k(a)\delta + o(\delta) \quad \text{as} \quad \delta \to 0.$$

Proof: By (3.2), $e_a(\delta) = \int_{S_a(\delta)} xc_k^{-1} g(x'x)dx$. We change variables by setting $y = x-\delta$ and obtain

$$(3.14) \quad e_a(\delta) = \int_{S_a(0)} (y+\delta)c_k^{-1} g(y'y + 2y'\delta + \delta'\delta)dy.$$

Expanding $e_a(\delta)$ we find, to first order, that

$$(3.15) \quad e_a(\delta) = [\int_{S_a(0)} c_k^{-1} g(y'y)dy]\delta + [2\int_{S_a(0)} yy'c_k^{-1} g'(y'y)dy]\delta + o(\delta)$$

by using the spherical symmetry of $g(y'y)$. But the first term of (3.15) is just $F_k(a^2)\delta$ and the second term is $2M_3\delta$ by (A1.13) and spherical symmetry. Substituting (A1.19) for M_3, we find that

$$e_a(\delta) = (\frac{1}{2\pi} \frac{c_{k+2}}{c_k})(2f_{k+2}(a^2))\delta.$$

Furthermore, $P_a(\delta) = P_a(0) + o(1)$; hence,

$$(3.16) \quad \mu_a(\delta) = (\frac{1}{2\pi} \frac{c_{k+2}}{c_k}) \frac{2f_{k+2}(a^2)}{P_a(0)}\delta.$$

Using (A1.1) and (A1.5) it is easily verified that (3.16) is equivalent to (3.13). ∎

Two useful alternative formulas for $b_k(a)$ are given in (A1.3) and (A1.4) and some numerical values are presented in Table A2.1 when $X \sim N(0,I_k)$. In that table we write $p = F_k(a^2)$ for the probability of lying in $S_a(0)$; alternatively, $a = (F_k^{-1}(p))^{1/2}$. It is interesting to study $b_k((F_k^{-1}(p))^{1/2})$ for fixed p as k increases. For fixed p, $F_k^{-1}(p)$ increases with k; therefore, by Lemma A2.1, b_k decreases with k. For k large, in fact, by (A2.14), $b_k = O(k^{-1/2})$ as $k \to \infty$ when p is fixed. To summarize, for a given p, in high dimensional space, IET will converge faster than in a low dimensional space. This fact is, at first glance at least, rather surprising and indeed it depends on the sample size being sufficiently large, where what constitutes "large" may itself grow with k.

Later in this chapter we shall prove a generalization of the preceding theorem, Theorem 3.6, which will imply that b_k is still the bias reduction factor when one computes the expectation of X, given that it lies in an off center ellipsoid.

It is natural, now that we have studied the zero order asymptotics as $\delta \to 0$, to undertake the analogous investigation as $\delta \to \infty$. Two formulations of this problem are relevant. We can, first of all, let the radius a be fixed and send $\delta \to \infty$, thereby forcing the probability in

the spheres, $P_a(\delta)$ to go to 0. Alternatively, we can fix $P_a(\delta) = p$; then, as $\delta \to \infty$, a must also approach ∞. Theorem 3.4 deals with the first alternative while Theorem 3.5 deals with the second. Neither result is as general as it could be, but both provide some insight into the operation of IET.

Theorem 3.4: Suppose $X \sim N(0, I_k)$. Then,

$$\lim_{\delta \to \infty} |\mu_a(\delta) - d| = 0 \tag{3.17}$$

where d is the point in $S_a(\delta)$ closest to 0.

In order to prove this result we must make use of a rather technical lemma, the proof of which will follow later. Note that $d = d(\delta)$ and let

$$T_\varepsilon(\delta) = \{x : x \in S_a(\delta) \text{ and } |x| < |d(\delta)| + \varepsilon\} \tag{3.18}$$

and

$$U_\varepsilon(\delta) = S_a(\delta) - T_\varepsilon(\delta). \tag{3.19}$$

We shall also abbreviate $S_a(\delta)$ as S_δ.

Lemma 3.1: Suppose $X \sim N(0, I_k)$. Then,

$$(3.20) \qquad \frac{P(T_\varepsilon(\delta))}{P(S_\delta)} \to 1 \text{ as } \delta \to \infty \quad \forall\, \varepsilon > 0.$$

and

$$(3.21) \qquad \sup_{x \in T_\varepsilon(\delta)} |x-d| \to 0 \text{ as } \varepsilon \to 0 \text{ uniformly in } \delta .$$

Theorem 3.4 asserts that the conditional expectation of $X \sim N(0, I_k)$ given that $X \in S_\delta$ approaches the point of maximum density in the sphere, namely d, as the S_δ's move out to ∞. The two pieces of information in Lemma 3.1 are that as $\delta \to \infty$, more and more of the probability in S_δ is actually contained in a small subset, $T_\varepsilon(\delta)$, of S_δ, and that every point in $T_\varepsilon(\delta)$ is very close to d when ε is small.

Proof of Theorem 3.4: We may write

$$(3.22) \qquad \mu_a(\delta) = \frac{\int_{T_\varepsilon(\delta)} xf(x)dx + \int_{U_\varepsilon(\delta)} xf(x)dx}{P(T_\varepsilon(\delta)) + P(U_\varepsilon(\delta))} .$$

If we set $c_\delta = P(U_\varepsilon(\delta))/P(T_\varepsilon(\delta)) = \left(1 - \frac{P(T_\varepsilon(\delta)}{P(S_\delta)}\right)\frac{P(S_\delta)}{P(T_\varepsilon(\delta))}$,

then by Lemma 3.1, $c_\delta \to 0$ as $\delta \to \infty$. It is possible to rewrite (3.22) as

$$(3.23) \qquad \mu_a(\delta) = \frac{1}{1+c_\delta} E(X | X \in T_\varepsilon(\delta)) + \frac{c_\delta}{1+c_\delta} E(X | X \in U_\varepsilon(\delta))$$

which implies that

$$(3.24) \qquad \mu_a(\delta) - E(X|X \in T_\varepsilon(\delta)) =$$

$$\frac{c_\delta}{1+c_\delta}(E(X|X \in U_\varepsilon(\delta)) - E(X|X \in T_\varepsilon(\delta))).$$

Since both $E(X|X \in U_\varepsilon(\delta))$ and $E(X|X \in T_\varepsilon(\delta))$ lie in S_δ (a consequence of the convexity of S_δ), it follows that they can be no farther apart than 2a. Hence (3.24)) implies that

$$(3.25) \quad |\mu_a(\delta) - E(X|X \in T_\varepsilon(\delta))| \leq \frac{c_\delta}{1+c_\delta}(2a) \to 0 \quad \text{as} \quad \delta \to \infty$$

Now $E(X|X \in T_\varepsilon(\delta)) \in T_\varepsilon(\delta)$ since $T_\varepsilon(\delta)$ is convex. Fix $\eta > 0$ and choose $\varepsilon > 0$ so small that for all δ $\sup_{x \in T_\varepsilon(\delta)} |x-d| < \eta/2$, which we can do as a result of Lemma 3.1. Then it follows that

$$(3.26) \qquad |E(X|X \in T_\varepsilon(\delta)) - d| < \frac{\eta}{2}$$

Next choose A such that $|\delta| > A$ implies that

$$(3.27) \quad |\mu_a(\delta) - E(X|X \in T_\varepsilon(\delta))| < \frac{\eta}{2},$$

which we can do as a result of (3.25). Combining (3.26) and (3.27) we obtain the result that $|\delta| > A$ implies $|\mu_a(\delta)-d| < \eta$, which is what we desired to prove. ∎

Proof of Lemma 3.1: Without loss of generality we may assume that δ lies on the positive x_1 axis and that $|\delta| > 2a$. Then we may write $d = \delta-a$, using the convention that δ (or d) can mean either the vector or its length (equivalently, its x_1 component), depending on the context. Equation (3.21) is geometrically obvious, so we shall only prove (3.20). Let $V(\delta,\varepsilon)$ be the volume of $T_\varepsilon(\delta)$ and observe that

$$(3.28)\quad P(T_\varepsilon(\delta)) \geq P(T_{\varepsilon/2}(\delta)) \geq (2\pi)^{-k/2}\exp\left(-\tfrac{1}{2}(d+\varepsilon/2)^2\right)V(\delta,\tfrac{\varepsilon}{2})$$

and

$$(3.29)\quad P(U_\varepsilon(\delta)) \leq (2\pi)^{-k/2}\exp\left(-\tfrac{1}{2}(d+\varepsilon)^2\right)V_k(a)$$

where $V_k(a)$, the volume of S_δ, is given in (A1.5). It is easy to show that a sphere with radius $\varepsilon/4$ centered at $d+\varepsilon/2$ is contained in $T_\varepsilon(\delta)$. It then follows that

(3.30) $$V(\delta,\varepsilon) \geq V_k(\varepsilon/4) = O(\varepsilon^k)$$

Using (3.28), (3.29), and (3.30) we may now prove (3.20):

$$0 < \frac{P(U_\varepsilon(\delta))}{P(T_\varepsilon(\delta))} \leq \frac{\exp\ [-\frac{1}{2}(d+\varepsilon)^2]V_k(a)}{\exp\ [-\frac{1}{2}(d+\varepsilon/2)^2]V(\delta,\varepsilon/2)}$$

$$\leq \exp\ [-\frac{1}{2}(d\varepsilon + \frac{3}{4}\varepsilon^2)]\frac{V_k(a)}{V_k(\varepsilon/8)} \to 0 \quad \text{as} \quad \delta \to \infty$$

since $d \to \infty$ as $\delta \to \infty$. ∎

The next theorem shows what happens in the one dimensional case when the $S_a(\delta)$'s go out to infinity in such a way that the probability they contain goes to a limit which is non-zero. We shall write $\phi(x)$ and $\Phi(x)$ for the $N(0,1)$ density and cdf, respectively.

Theorem 3.5: Suppose $X \sim N(0,1)$. Then if $P_a(\delta) \to p$ as $\delta \to \infty$ for some p such that $0 < p < 1$, it follows that

(3.31) $$P_a(\delta) \to \frac{\phi(\Phi^{-1}(1-p))}{p}$$

Proof: Since $\delta \to \infty$, $\Phi(\delta+a) \to 1$. Hence, $\Phi(\delta-a) \to 1-p$, which implies that $\delta-a \to \Phi^{-1}(1-p)$. Of course, $\delta + a \to \infty$. Therefore, as $\delta \to \infty$

$$(3.32) \qquad \mu_a(\delta) = \frac{e_a(\delta)}{P_a(\delta)} \to \frac{\int_{\phi^{-1}(1-p)}^{\infty} x\phi(x)\,dx}{p}$$

But since $\int_c^{\infty} x\phi(x)dx = \phi(c)$, we conclude that

$\mu_a(\delta) \to p^{-1}\phi(\phi^{-1}(1-p))$. ∎

We may approximate the limiting conditional expectation of X in (3.31) in a simple fashion. In Woodroofe [13, p. 97] it is shown that if $x > 0$,

$$(3.33) \qquad 1 - \Phi(x) \le \frac{\phi(x)}{x} \le (1 + x^{-2})(1 - \Phi(x)).$$

By setting $x = \phi^{-1}(1-p)$ and applying (3.33) it is easy to derive that if $p < 1/2$,

$$(3.34) \qquad 0 \le \frac{\phi(\phi^{-1}(1-p))}{p} - \phi^{-1}(1-p) \le \frac{1}{\phi^{-1}(1-p)}.$$

Hence, as $p \to 0$, $p^{-1}\phi(\phi^{-1}(1-p)) - \phi^{-1}(1-p) \to 0$. Some numerical results are exhibited in Table 3.1. Observe that when $p = 0.8$, one iteration of IET using the worst possible starting value will still lead to an estimate with expectation 0.35. Of course this comment is based on the assumption that the observations are from a univariate normal distribution.

The rest of this chapter is devoted to an analysis of the first order asymptotics of IET; more particularly, we shall be interested in obtaining the asymptotic distribution of what we shall call the stationary point of this algorithm. Let $\phi = (b,\nu,B)$ be the parameter specifying the ellipsoid.

$$(3.35) \qquad E^*(\phi) = \{y : (y-\nu)'B^{-1}(y-\nu) \leq b^2\}$$

where b is a scalar, ν is a k-vector, and B is a symmetric positive definite $k \times k$ matrix with $\operatorname{tr} B = k$. We can represent the operation of IET as a sequence $\{\tilde{\phi}^{(m)}\}$ with $E_m^* = E^*(\tilde{\phi}^{(m)})$ and $\tilde{\phi}^{(m+1)} = T_n(\tilde{\phi}^{(m)})$ where the operator T_n is a function of the sample (of size n) and is defined by

$$(3.36) \qquad \tilde{\nu}^{(m+1)} = \frac{n}{[np]} \int_{E_m^*} y \, dH_n ,$$

$$(3.37) \qquad \tilde{B}^{(m+1)} = \frac{\int_{E_m^*} (y-\tilde{\nu}^{(m+1)})(y-\tilde{\nu}^{(m+1)})' dH_n}{k^{-1}\operatorname{tr} \int_{E_m^*} (y-\tilde{\nu}^{(m+1)})(y-\tilde{\nu}^{(m+1)}) dH_n}$$

and $\tilde{b}^{(m+1)}$ is the smallest value of b for which

$$(3.38) \qquad \frac{[np]}{n} = \int_{E_{m+1}^*} dH_n$$

where H_n is the empirical cdf of $Y = \mu + A^{1/2}X$, which has an ellipsoidal distribution with mean μ and covariance $\Sigma = (2\pi c_k)^{-1} c_{k+2} A$. We define the "true ϕ" to be $\phi_0 = (b_0, \nu_0, B_0)$ where $\nu_0 = \mu$, $B_0 = k\Sigma/(\text{tr } \Sigma)$, and b_0 is determined implicitly by $\Pr(Y \in E^*(\phi_0)) = p$.

Definition 3.1: A local stationary point is a sequence of random points $\tilde{\phi}_n = (\tilde{b}_n, \tilde{\nu}_n, \tilde{B}_n)$ such that $\tilde{\phi}_n - \phi_0 = o_p(n^{-1/2})$ and satisfying

$$\tilde{\phi}_n = T_n(\tilde{\phi}_n) + o_p(n^{-1/2}). \tag{3.39}$$

We shall say that there is an essentially unique local stationary point if whenever $\tilde{\phi}_n$ and $\tilde{\phi}_n'$ are local stationary points, $\tilde{\phi}_n - \tilde{\phi}_n' = o_p(n^{-1/2})$.

Because of the invariance of IET it will only be necessary to consider the special case where the distribution is spherically symmetric (i.e. $\mu = 0$ and $A = I_k$). It will be convenient to have special notation available for this situation. Let $\vartheta = (\lambda, \delta, \varepsilon)$ be the parameter specifying the ellipsoid

$$E(\vartheta) = \{x : (x-\delta)'(I_k+\varepsilon)^{-1}(x-\delta) \leq (a_0+\lambda)^2\} \tag{3.40}$$

where λ is a scalar, δ is a k-vector, and ε is a symmetric $k \times k$ matrix with $\text{tr}\ \varepsilon = 0$. Here the "true ϑ" is just $\theta_0 = (0,0,0)$ and $a_0^2 = F_k^{-1}(p)$. In this case, the random variable in question is X and f, F, and F_n are respectively its density, cdf, and empirical cdf.

Our next theorem, a generalization of Theorem 3.3, will play a crucial role in the derivation of the asymptotic distribution of what will turn out to be the essentially unique local stationary point. We define

$$(3.41) \quad |\theta|^2 = \lambda^2 + \sum_{i=1}^{k} \delta_i^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} \varepsilon_{ij}^2 .$$

Theorem 3.6: If g is differentiable in a neighborhood of a_0^2, then as $|\theta| \to 0$,

$$(3.42) \quad \int_{E(\vartheta)} f(x)dx = F_k(a_0^2) + 2a_0 f_k(a_0^2)\lambda + o(|\theta|)$$

$$(3.43) \quad \int_{E(\vartheta)} xf(x)dx = \left(\frac{c_{k+2}}{2\pi c_k}\right)(2f_{k+2}(a_0^2))\delta + o(|\theta|)$$

and

$$(3.44)\quad \int_{E(\theta)} xx'f(x)dx = \left(\frac{c_{k+2}}{2\pi c_k}\right) F_{k+2}(a_0^2) I_k + \left(\frac{c_{k+2}}{2\pi c_k}\right)(2a_0 f_{k+2}(a_0^2) I_k)\lambda + \left(\frac{c_{k+4}}{(2\pi)^2 c_k}\right)(2f_{k+4}(a_0^2))\varepsilon + o(|\theta|).$$

We shall find it useful to derive a lemma before embarking upon the proof of this result. Set $a = a_0 + \lambda$ and suppose that h is a vector valued function defined on R^k. We shall write $Dh(x)$ for the derivative matrix of h.

Lemma 3.2: If h is differentiable, as $|(\delta,\varepsilon)| \to 0$,

$$(3.45)\quad \int_{E(\theta)} h(x)f(x)dx = A + B + C + o(|(\delta,\varepsilon)|)$$

where

$$(3.46)\quad A = \int_{x'x \le a^2} h(x)c_k^{-1}g(x'x)dx$$

$$(3.47)\quad B = \int_{x'x \le a^2} [Dh(x)](\varepsilon x/2 + \delta)c_k^{-1}g(x'x)dx$$

and

$$(3.48)\quad C = \int_{x'x \le a^2} h(x)(x'\varepsilon x + 2x'\delta)c_k^{-1}g'(x'x)dx.$$

Proof of Lemma 3.2: We change variables by setting $y = (I_k + \varepsilon)^{-1/2}(x - \delta)$, where $(I_k + \varepsilon)^{-1/2}$ is chosen to be symmetric. Then, the range of integration in (3.45) is the sphere $\{y : y'y \leq a^2\}$. Since $tr\ (\varepsilon) = 0$, the Jacobian determinant is

$$(3.49) \qquad \left|\frac{\partial x}{\partial y}\right| = 1 + o(|\varepsilon|)$$

Furthermore,

$$(3.50) \quad x = (I_k + \varepsilon/2)y + \delta + o(|\varepsilon|)$$

and

$$(3.51) \quad x'x = y'y + y'\varepsilon y + 2y'\delta + o(|(\delta,\varepsilon)|).$$

Now, by recalling that $f(x) = c_k^{-1}g(x'x)$; substituting (3.49), (3.50), and (3.51) in the left hand side of (3.45); expanding h and g about y and $y'y$, respectively; and keeping only those terms which are zero or first order in (δ,ε), we obtain the lemma. ∎

Proof of Theorem 3.6: To derive (3.42), (3.43) and (3.44) we simply apply Lemma 3.2 with $h(x)$ set equal to 1, x, and $x_i x_j$, respectively. The final step is to expand the

results thereby obtained (as functions of a) around a_0. In all cases we shall make heavy use of spherical symmetry and Lemma A1.1. For convenience we shall write S for the region of integration $\{x : x'x \leq a^2\}$. When $h = 1$, we find that $A = F_k(a^2)$ and since $Dh = 0$, $B = 0$. Observe that the formula for C may be simplified to

$$C = (\int_S x_1^2 c_k^{-1} g'(x'x))(\Sigma\varepsilon_{ii}),$$

which implies that $C = 0$ since $tr\ (\varepsilon) = 0$. Equation (3.42) then follows from

$$\int_{E(\theta)} f(x)dx = F_k((a_0+\lambda)^2) + o(|(\delta,\varepsilon)|).$$

When $h(x) = x$, it is obvious that $A = 0$. Since $Dh(x) = I_k$, we obtain immediately that $B = F_k(a^2)\delta$. The equation for C is

$$C = \int_S (xx'\varepsilon x + 2xx'\delta)c_k^{-1} g'(x'x)dx,$$

but, by spherical symmetry, the integral of $xx'\varepsilon x c_k^{-1} g'(x'x)$ vanishes and the integral of the remaining term reduces to

$$C = [\int_S x_1^2 c_k^{-1} g'(x'x)dx](2\delta). \tag{3.52}$$

As a result of (A1.13) and (A1.19), we may write (3.52) as

$$C = \left[\left(\frac{c_{k+2}}{2\pi c_k}\right)(2f_{k+2}(a^2)) - F_k(a^2)\right]\delta.$$

Therefore, adding A, B, and C we find

$$\int_{E(\theta)} xf(x)dx = \left(\frac{c_{k+2}}{2\pi c_k}\right)(2f_{k+2}(a^2))\delta + o(|(\delta,\varepsilon)|)$$

from which equation (3.43) may be derived by expanding around a_0. Finally we shall obtain (3.44) by considering two cases: $h(x) = x_i x_j$ for $i < j$ and for $i = j$. First, when $i < j$, it is easy to see that $A = 0$. We find that when we drop terms that vanish because of spherical symmetry and the form of Dh (Dh has x_j as its i^{th} coordinate and x_i as its j^{th} coordinate, with the other coordinates equal to 0),

$$(3.53) \quad B = \int_S (1/2)(x_j^2\varepsilon_{ij} + x_i^2\varepsilon_{ji})c_k^{-1}g(x'x)dx .$$

By the symmetry of ε and (A1.8) and (A1.17), equation (3.53) leads to

$$(3.54) \qquad B = (2\pi c_k)^{-1}c_{k+2}F_{k+2}(a^2)\varepsilon_{ij}.$$

Again keeping only nonvanishing terms, we find

$$C = \int_S x_i^2 x_j^2 (\varepsilon_{ij} + \varepsilon_{ji}) c_k^{-1} g'(x'x)dx,$$

which simplifies to

$$(3.55) \quad C = 2[(\frac{c_{k+4}}{(2\pi)^2 c_k}) f_{k+4}(a^2) - (\frac{c_{k+2}}{2\pi c_k}) \frac{F_{k+2}(a^2)}{2}] \varepsilon_{ij}$$

upon the application of (A1.16) and (A1.20). But then, using (3.54) and (3.55) we derive

$$(3.56) \quad \int_{E(\theta)} x_i x_j f(x)dx = (\frac{c_{k+4}}{(2\pi)^2 c_k})(2f_{k+4}(a^2))\varepsilon_{ij} \text{ for } i < j.$$

Taking the next case, $i = j$, or equivalently $h(x) = x_i^2$, we find

$$A = (\frac{c_{k+2}}{2\pi c_k}) F_{k+2}(a^2)$$

and

$$B = (\frac{c_{k+2}}{2\pi c_k}) F_{k+2}(a^2) \varepsilon_{ii}$$

by making use of (A1.8) and (A1.17). Dropping terms that vanish we may write

$$C = \int_S x_i^2 (\sum_{\ell=1}^{k} x_\ell^2 \varepsilon_{\ell\ell}) c_k^{-1} g'(x'x)dx.$$

Then, using M_4, defined in (A1.20), and equations (A1.15) and (A1.16),

$$C = (3M_4)\varepsilon_{ii} + M_4\Big(\sum_{\ell\neq i}\varepsilon_{\ell\ell}\Big).$$

But since $\operatorname{tr}\varepsilon = 0$, it follows that $\sum_{\ell\neq i}\varepsilon_{\ell\ell} = -\varepsilon_{ii}$; hence,

$$C = 2\Big[\Big(\frac{c_{k+4}}{(2\pi)^2 c_k}\Big) f_{k+4}(a^2) - \Big(\frac{c_{k+2}}{2\pi c_k}\Big)\frac{F_{k+2}(a^2)}{2}\Big]\varepsilon_{ii}.$$

Therefore, summing A, B, and C we find

$$(3.57)\qquad \int_{E(\theta)} x_i^2 f(x)\,dx = \Big(\frac{c_{k+2}}{2\pi c_k}\Big) F_{k+2}(a^2) + \Big(\frac{c_{k+4}}{(2\pi)^2 c_k}\Big)(2f_{k+4}(a^2))\varepsilon_{ii}.$$

Combining (3.56) and (3.57) we obtain

$$\int_{E(\vartheta)} xx' f(x)\,dx = \Big(\frac{c_{k+2}}{2\pi c_k}\Big) F_{k+2}(a^2) I_k + \Big(\frac{c_{k+4}}{(2\pi)^2 c_k}\Big)(2f_{k+4}(a^2))\varepsilon$$

from which (3.44) follows upon expanding around a_0. ∎

It is easy to see that as a result of (3.42) and (3.43) and the formula for $b_k(a)$ given in (A1.4),

$$(3.58)\qquad E(X'X \in E(\vartheta)) = b_k(a_0)\delta + o(|\vartheta|).$$

Using (3.42) and (3.44), we obtain in a similar way that

$$E((X-\delta)(X-\delta)'|X \in E(\theta)) = (\frac{c_{k+2}}{2\pi c_k})\frac{F_{k+2}(a_0^2)}{F_k(a_0^2)}I_k + O(|\vartheta|)$$

which we may rewrite as

$$(3.59)\quad E[(X-\delta)(X-\delta)'|X \in E(\theta)] = c(k,p)\frac{c_{k+2}}{2\pi c_k}I_k + O(|\vartheta|)$$

where $c(k,p) = F_{k+2}(F_k^{-1}(p))/p$. (Note that $\frac{c_{k+2}}{2\pi c_k}I_k$ is the covariance matrix of X.) It is a simple matter to extend (3.58) and (3.59) to their ellipsoidal generalizations:

$$(3.60)\quad E[Y|Y \in E^*(\phi)] = \mu + b_k(a_0)(\nu - \mu) + O(|\phi - \phi_0|)$$

and

$$(3.61)\quad E[(Y-\nu)(Y-\nu)'|Y \in E^*(\phi)] = c(k,p)\Sigma + O(|\phi - \phi_0|).$$

At the end of Appendix 1 it is shown that $c(k,p)\Sigma$ is the covariance of the truncated ellipsoidal distribution, so (3.61) reassures us that when ϕ is near ϕ_0, the expected value of $\tilde{\Sigma}$ generated by IET will be close to the true truncated covariance. According to (3.60), the bias reduction factor plays the same role in the ellipsoidal case that it played in Theorems 3.2 and 3.3, which is another reassuring result. It is straightforward to write

down the first order terms in the expansion of the left hand side of (3.61). There is a term in $b-b_0$ and a term in $B-B_0$; the corresponding coefficients play the same sort of role that $b_k(a_0)$ does in the location case in telling us how one iteration of IET will reduce the bias in the covariance estimate, on the average.

Our next result concerns the uniformity of convergence of linear functionals of the empirical distribution function F_n on all ellipsoids $E(\theta)$ such that θ is sufficiently close to 0. Note that $|B|$, for a matrix $B = (b_{ij})$, will denote $(\sum_i \sum_j b_{ij}^2)^{1/2}$. Recall that $\theta = (\lambda,\delta,\varepsilon)$ where $\mathrm{tr}\,(\varepsilon) = 0$.

Lemma 3.3 - (Uniformity): Suppose F is a cdf on R^k with a bounded density and F_n is the corresponding empirical cdf. Let

$$V_n = \{\theta : |\theta| \leq Kn^{-1/2+b}\} \tag{3.62}$$

and

$$D_n(\theta) = \int_{E(\theta)} h(x)(dF_n - dF) - \int_{E(0)} h(x)(dF_n - dF). \tag{3.63}$$

Then there exists a $b > 0$ such that for any $K > 0$ and any bounded scalar, vector, or matrix valued h,

$$(3.64) \qquad \sup_{\theta \in V_n} |D_n(\theta)| = o_p(n^{-1/2})$$

Proof: It will suffice to show (3.64) for scalar valued h. First we cover V_n with balls of radius $\rho = n^{-1/2 - a}$, so that $V_n \subset \bigcup_{i=1}^{m} S_\rho(\theta_i)$, where $S_\rho(\theta_i) = \{\theta : |\theta - \theta_i| \leq \rho\}$. We will need m balls where

$$m \simeq K_1 \left(\frac{n^{-1/2 + b}}{n^{-1/2 - a}}\right)^{k^*} = K_1 n^{k^*(a+b)}$$

and k^* is the dimension of the θ space, that is

$$k^* = 1 + k + k(k+1)/2 - 1 = k(k+3)/2$$

If $\theta \in S_\rho(\theta_i)$, then

$$E(\theta) \Delta E(\theta_i) \subset R(\theta_i, \rho)$$

where $R(\theta_i, \rho)$ is a spherical shell in x-space with thickness bounded by $K_2 \rho$. Hence, if $\theta \in S_\rho(\theta_i)$,

$$|D_n(\theta) - D_n(\theta_i)| \leq \int_{R(\theta_i, \rho)} |h(x)| dF_n + \int_{R(\theta_i, \rho)} |h(x)| dF.$$

But $\int_{R(\theta_i,\rho)} |h(x)| dF \leq K_3 \int_{R(\theta_i,\rho)} dF \leq K_4\rho$ and

$$\int_{R(\theta_i,\rho)} h(x)\ dF_n \leq K_3 \int_{R(\theta_i,\rho)} dF_n \leq K_4\rho + K_5(\rho/n)^{1/2}Z$$

where Z is a random variable with mean 0 and variance 1. Hence,

$$\sup_{\theta\epsilon S_\rho(\theta_i)} |D_n(\theta) - D_n(\theta_i)| \leq K_6(\rho + (\rho/n)^{1/2}|Z|).$$

Since $\Pr(|Z| \geq \eta^{-1}) \leq \eta^2$, by Chebyshev's inequality we may conclude that

$$(3.65)\quad \Pr\left(\sup_{\theta\epsilon S_\rho(\theta_i)} |D_n(\theta) - D_n(\theta_i)| > K_6(\rho + (\rho/n)^{1/2}\eta^{-1})\right) \leq \eta^2.$$

Since $D_n(\theta_i)$ has expectation 0,

Var $D_n(\theta_i) \leq n^{-1} \int_{E(\theta_i)\Delta E(0)} h^2(x)dF \leq K_7 n^{-1}|\theta_i|$. Hence,

Var $D_n(\theta_i) \leq K_8 n^{-1}n^{-1/2+b} = K_8 n^{-3/2+b}$ for

$i = 1,\ldots,m$. But then, by Chebychev's inequality,

$$(3.66)\quad \Pr(|D_n(\theta_i)| > Kn^{-1/2-a}) \leq K_9 n^{-1/2+2a+b}$$

where $a > 0$. We are interested in showing that for some $a > 0$,

$$(3.67)\quad \Pr\left(\sup_{\theta\in V_n} |D_n(\theta)| > Kn^{-1/2-a}\right) = o(1).$$

Now the left hand side of (3.67) may be bounded by

$\sum_{i=1}^{m} \Pr\left(\sup_{\theta\in S_\rho(\theta_i)} |D_n(\theta)| > Kn^{-1/2-a}\right)$. In addition,

$$(3.68)\quad \Pr\left(\sup_{\theta\in S_\rho(\theta_i)} |D_n(\theta)| > Kn^{-1/2-a}\right) \le P_{1i} + P_{2i}$$

where P_{1i} is the left hand side of (3.66) and

$$(3.69)\quad P_{2i} = \Pr\left(\sup_{\theta\in S_\rho(\theta_i)} |D_n(\theta) - D_n(\theta_i)| > Kn^{-1/2-a}\right)$$

But if we choose η such that

$$K_6(\rho + \rho^{1/2}n^{-1/2}\eta^{-1}) \le Kn^{-1/2-a},$$

i.e. if we take $\eta = K_{10}n^{(a/2)-(1/4)}$, then equation (3.65) implies

$$(3.70)\quad P_{2i} \le K_{11}n^{a-(1/2)}.$$

But now, combining (3.65), (3.66), and (3.70) we obtain

$$(3.71)\quad \sum_{i=1}^{m} (P_{1i} + P_{2i}) \le K_{12}n^{k^*(a+b)}\left(n^{-(1/2)+b+2a} + n^{a-(1/2)}\right)$$

If we take $a = b = (4(k^* + 2))^{-1}$, then the right hand side of (3.71) goes to 0 as $n \to \infty$, which implies (3.67), which is equivalent to (3.64), the desired result. ∎

It is interesting to note that the a and b we needed in the preceding proof were quite small, that is

$$a = b = \frac{1}{2(k(k+3))}.$$

We are now ready to derive the asymptotic distribution of the local stationary point $\tilde{\phi}_n$ when Y has an ellipsoidal density.

Theorem 3.7: There is an essentially unique local stationary point $\tilde{\phi}_n = (\tilde{b}_n, \tilde{\nu}_n, \tilde{B}_n)$. As $n \to \infty$, $n^{1/2}(\tilde{\phi}_n - \phi_0)$ is asymptotically normal. In particular,

$$(3.72) \quad L(n^{1/2}(\tilde{\nu}_n - \mu)) \to N(0, \frac{c(k,p)}{p(1-b_k(a_0))^2} Z)$$

Proof: It is enough to prove the theorem in the spherically symmetric case where the true density is f(x). The invariance of IET leads to the more general result. Note that in the spherically symmetric situation we shall study in this proof,

$\tilde{\xi}_n = (\tilde{b}_n, \tilde{\nu}_n, \tilde{B}_n)$ is simply related to $\tilde{\epsilon}_n = (\tilde{\epsilon}_n, \tilde{\delta}_n, \tilde{E}_n)$ by $\tilde{b}_n^2 = (a_0 + \tilde{\epsilon}_n)^2$, $\tilde{\nu}_n = \tilde{\delta}_n$, and $\tilde{B}_n = I_k + \tilde{E}_n$, when we require $E^*(\tilde{\xi}_n) = E(\tilde{\epsilon}_n)$.

Then, as a result of (3.39) and the uniformity lemma, (Lemma 3.3),

$$(3.73)\quad \tilde{\delta}_n = p^{-1}\int_{E(0)} x dF_n + p^{-1}\int_{E(\tilde{\theta}_n)} x dF - p^{-1}\int_{E(0)} x dF + o_p(n^{-1/2}).$$

If we write $1(E)$ for the indicator function of the event E, and let

$$(3.74)\quad W_1 = n^{-1}\Sigma X_i 1(X_i \in E(0)) = \int_{E(0)} x dF_n$$

then upon making use of equations (3.43) of Theorem 3.6 and (A1.4) we find that (3.73) may be rewritten as

$$(3.75)\quad \tilde{\delta}_n = \frac{W_1}{p(1 - b_k(a_0))} + o_p(n^{-1/2}).$$

Now, equation (3.75) gives an explicit formula for $\tilde{\delta}_n$. We may apply the central limit theorem to W_1 and obtain, as a consequence, the asymptotic distribution of $\tilde{\delta}_n$. Because of (A1.8) and (A1.17)

$$(3.76)\quad L(n^{1/2}W_1) \to N(0, \ (\frac{c_{k+2}}{2\pi c_k})F_{k+2}(a_0^2)I_k).$$

Using the definition of $c(k,p)$ in (A1.23), and the transformation $Y = u + A^{1/2}X$, (3.76) immediately leads

to (3.72). In a similar fashion we may obtain explicit formulas for $\tilde{\lambda}_n$ and $\tilde{\varepsilon}_n$, analogous to (3.75), by making use of uniformity and (3.42) and (3.44) together with (3.39). So if we let $W_2 = n^{-1}\Sigma 1(X_i \in E(0))$, then

$$\tilde{\lambda}_n = \frac{p - W_2}{2a_0 f_k(a_0^2)} + o_p(n^{-1/2}). \tag{3.77}$$

Since $L(n^{1/2}(W_2 - p)) \to N(0, p(1-p))$, it follows that

$$L(n^{1/2}\tilde{\lambda}_n) \to N(0, \frac{p(1-p)}{(2a_0 f_k(a_0^2))^2}).$$

We can also obtain an explicit formula for $\tilde{\varepsilon}_n$ as a function of $W_3 = n^{-1}\Sigma X_i X_i' 1(X_i \in E(0))$ and $\tilde{\lambda}_n$:

$$I_k + \tilde{\varepsilon}_n = \frac{D}{k^{-1}\mathrm{tr}\,(D)} + o_p(n^{-1/2}) \tag{3.78}$$

where

$$D = W_3 + (2\tau c_k)^{-1} c_{k+2}(2a_0 f_k(a_0^2)) I_k \tilde{\lambda}_n$$

$$+ ((2\tau)^2 c_k)^{-1} c_{k+4}(2f_{k+4}(a_0^2))\tilde{\varepsilon}_n.$$

By computing $\mathrm{tr}(D)$, which would depend on $\tilde{\lambda}_n$ and W_3 but not on $\tilde{\varepsilon}_n$, we could use (3.78) to derive an explicit

formula for $\tilde{\varepsilon}_n$. Again, $n^{1/2}\tilde{\varepsilon}_n$ will have an asymptotically normal distribution but with a very messy formula for its covariance. (In Appendix 1, Lemma A1.1 provides the formulas which would be needed to actually compute that covariance matrix.) From what we have shown so far, it follows that $\tilde{\theta}_n$ itself, and hence $\tilde{\phi}_n$, is asymptotically normal. We have seen that if $\tilde{\phi}_n$ is a local stationary point, then $\tilde{\theta}_n$ must satisfy (3.75), (3.77), and (3.78). But since those three equations themselves yield an explicit formula for $\tilde{\theta}_n$ such that the corresponding $\tilde{\phi}_n$ satisfies (3.39), we have proved the theorem. ∎

We conjecture that if IET is allowed to iterate until it reaches a stopping point, that is, until for some m_0, $\tilde{\phi}^{(m_0)} = T_n(\tilde{\phi}^{(m_0)})$, then <u>first</u>, there will be such an m_0 (with probability approaching one) and second, this stopping point will be in a $o_p(n^{-1/2})$ neighborhood of the local stationary point whose asymptotic distribution we have just derived. If our conjecture is true, then we have just obtained the asymptotic distribution of the stopping point of IET. Were we to prove that IET does halt with probability approaching 1, it would then be enough to show that there can be no stationary point $\tilde{\phi}_n$ such that $\tilde{\phi}_n - \phi_0$ is bigger in order of magnitude than $o_p(n^{-1/2})$, since a stopping point is

certainly a stationary point.

It is interesting to note that if $Y \sim N(\mu,Z)$ then the result in Theorem 3.7 becomes

$$L(n^{1/2}(\tilde{\nu}_n-\mu)) \to N(0, \frac{1}{p(1-b_k(a_0)}Z)$$

since $c(k,p) = 1 - b_k(a_0)$ for this distribution.

Chapter 4

Monte Carlo Analysis

In this chapter we will be concerned with the question of how the presence of outliers and/or a small (infrequent) subpopulation influences the performance of IET when it is used to estimate the mean of the larger (more common) population present in the sample. More specifically, we shall, in each case considered here, generate a random sample composed of three kinds of data:

n_1 $N(\mu_1, \Sigma_1)$ observations

n_2 $N(\mu_2, \Sigma_2)$ observations

and

n_3 $N(\mu_1, \sigma^2 I_k)$ observations

where $n_2, n_3 << n_1$ and $\sigma^2 >> 1$. Then we shall use IET, as defined in Chapter 1 to estimate μ_1.

Given this setup, how shall we study the performance of IET? There are a considerable number of parameters which must be specified, some involving the algorithm and the others involving the simulated data, before the above process is well defined. To specify IET, one must select starting values for the mean and covariance as well

as a sequence of proportions or p's, while, on the other hand, before generating the data one must choose μ_1, μ_2, Σ_1, Σ_2, σ^2, n_1, n_2, n_3, and, implicitly, the dimensionality k, of the problem. Presumably our study should consist of setting these parameters at a sufficiently large number of values to encompass the range of possibilities likely to be encountered in practice. Unfortunately, there are too many parameters for such a comprehensive investigation to be feasible. Therefore we shall content ourselves with the examination of a relatively small number of cases, in the hope that we shall still get a feeling for the *important characteristics* of the behavior of IET.

To begin our study we now assume, without loss of generality, that $\mu_1 = 0$ and that μ_2 always lies on the positive x_1 axis. Furthermore we set $\Sigma_1 = \Sigma_2 = I_k$ and always take the starting values for IET to be the observed mean and covariance of the entire sample, which consists of $n = n_1 + n_2 + n_3$ observations. Then, to proceed we need only choose a single positive number for μ_2, the dimensionality k, σ^2, n_1, n_2, n_3 and the sequence of p's.

First we consider an example. Set $k = 3$, $\mu_2 = 5$, $n_1 = 100$, $n_2 = 20$, $n_3 = 0$ and let the sequence of proportions be 0.5, 0.6, 0.7. The overall mean of the random sample generated according to these specifications

is $(0.88, 0.05, -0.06)'$. The three successive estimates of μ_1 are

$$(0.06, 0.06, -0.16)'$$
$$(-0.07, 0.05, -0.08)'$$
$$(-0.06, 0.06, -0.07)'.$$

Incidentally, the final correlation matrix is

$$\begin{pmatrix} 1 & -0.05 & 0.04 \\ & 1 & 0.09 \\ & & 1 \end{pmatrix}.$$

Though the general behavior of IET in this example is typical of its behavior in many other examples with similar parameter settings, it would be nice, nevertheless, to have some quantitative information about its performance. The rest of this chapter is devoted to the presentation and interpretation of this kind of information.

Our procedure is as follows. Generate a random sample according to the appropriate parameters. Apply IET, using the sequence of proportions $p_1,\ldots,p_m$ to obtain the sequence of estimates $\tilde{\mu}^{(0)}, \tilde{\mu}^{(1)}, \ldots, \tilde{\mu}^{(m)}$, where $\tilde{\mu}^{(0)}$ is the overall mean, $\bar{x}$. Compute and save

$e_i^{(1)} = k^{-1/2}|\tilde{\mu}^{(i)}|$, where the superscript 1 indicates that this is the first Monte Carlo trial. Repeat this process for each of s samples with the same parameter settings. Report the means and standard errors of the $e_i^{(j)}$'s in the form

$$\begin{array}{ccccc} \bar{e}_0 & & \bar{e}_1 & \cdots & \bar{e}_m \\ (\hat{\sigma}_{\bar{e}_0}) & & (\hat{\sigma}_{\bar{e}_1}) & \cdots & (\hat{\sigma}_{\bar{e}_m}) \end{array} .$$

Of course,

$$\bar{e}_i = s^{-1} \sum_{j=1}^{s} e_i^{(j)}$$

and

$$\hat{\sigma}^2_{\bar{e}_i} = s^{-1}\hat{\sigma}^2_{e_i} = s^{-1}(s-1)^{-1} \sum_{j=1}^{s} (e_i^{(j)} - \bar{e}_i)^2.$$

Now suppose a sample has no contamination, i.e. $n_2 = n_3 = 0$. Then $\bar{X} \sim N(0, n^{-1}I_k)$. Hence $|\bar{X}| \sim n^{-1/2}\chi_{(k)}$. Now let

$$(4.1) \qquad m_k = (2/k)^{1/2} \frac{\Gamma((k+1)/2)}{\Gamma(k/2)}$$

Then a simple calculation reveals that

$E(\chi_{(k)}) = k^{1/2}m_k$ and $Var(\chi_{(k)}) = k(1-m_k^2)$. It then follows that, in this case,

$$E(e_0) = m_k n^{-1/2} \tag{4.2}$$

and

$$Var(e_0) = (1-m_k^2)n^{-1} \tag{4.3}$$

since $e_0 = k^{-1/2}|\bar{X}| \sim (nk)^{-1/2}\chi_{(k)}$.
We list some numerical values for m_k and $1-m_k^2$ in Table 4.1. Incidentally, as $k \to \infty$, $m_k \to 1$ and $1-m_k^2 \sim (2k)^{-1}$. These last two facts may be derived by observing that $\mathcal{L}(k^{1/2}(k^{-1}\chi^2_{(k)}-1)) \to N(0,2)$ as $k \to \infty$ and applying the δ-method.

The point of the preceding paragraph was to provide a benchmark for determining when $\bar{e}_i$ is big. The best one can hope to do is to have an average error $\bar{e}_i$ approximately equal to $m_k n_1^{-1/2}$ if there are n_1 observations of the major population (as long as the minor population isn't too close to the major one). For example, if $n_1 = 100$ and $k = 2$, then $m_k n^{-1/2} \tilde{=} 0.09$. Similarly we will expect the estimate of the standard error $\hat{\sigma}_{\bar{e}_i}$ to be approximately

$(0.215 \times 10^{-4})^{1/2} \tilde{=} 5 \times 10^{-3}$ if $s = 100$ and if p_i is such that about 100 observations were used in computing the i^{th} estimate $\tilde{\mu}(i)$ (and if not too many outliers are included).

Now we are ready to report some results. We shall describe and interpret 5 simulations, to be referred to as simulations A, B, C, D, E, whose results are recorded in Tables 4.2-4.6 at the end of this paper. In simulations A, B, and C we will set $k = 2$, $n_1 = 100$, $n_2 = 10$, $n_3 = 0$, and $s = 100$ (= number of samples). These three simulations differ only in regard to the sequence of proportions used. In A, $p_1 = p_2 = p_3 = p_4 = 0.5$; in B, $p_1 = 0.5$, $p_2 = 0.6$, $p_3 = 0.7$, $p_4 = 0.8$; in C, $p_1 = p_2 = p_3 = p_4 = 0.8$.

Simulations A, B, and C are intended to illustrate the effect (on IET estimates) of the position of the subpopulation when there are no outliers ($n_3 = 0$). Unfortunately, the standard errors, in parentheses, are often comparable in size to the differences between various $\bar{e}_i$'s. (To reduce the standard errors to about 0.001 would require at least 25 times the number of samples used here: 2500 samples of 110 observations!) Nevertheless we may derive some general conclusions of interest.

In simulation A, the presence of a subpopulation causes one to do somewhat worse than one would otherwise

do. Simulations B and C suggest that the estimation problem is hardest when the subpopulation is 1 or 2 standard deviations away. The use of an increasing sequence of proportions seems to slow down convergence (as compared to using the biggest P repeatedly), but not to affect $\bar{e}_4$ very much.

In simulation D we study the effects of having both outliers and a small subpopulation in the sample. To simplify the presentation, each run in Table 4.5 will be identified by a parameter vector: $(\mu_2, \sigma^2, P_1, P_2, P_3, P_4)$. For all of the runs, $n_1 = 100$, $n_2 = n_3 = 10$, $s = 100$, and $k = 2$.

The final average error, $\bar{e}_4$, in runs 2 through 7 is close to 0.1 and is based on $0.8(120) = 96$ observations. So we are doing about as well as we could hope to do. The presence of outliers, whether with variance 100 or 400, has little effect; ultimately IET screens out most of these points. Similarly, the final error is not influenced much by where, exactly, the subpopulation is (as long as it is at least 3 standard deviations away) nor by whether or not an increasing sequence of proportions is used. In practice, naturally, we do not know that the right proportion is, for example, 0.8. This is one reason why an increasing sequence might be useful - one imagines gradually raising P

until IET begins wandering off, a sign that P has become too big.

Our last simulation, simulation E in Table 4.6, involves four dimensional data. Each run is again identified by a parameter vector (μ_2, n_2, n_3), but in all cases $n_1 = 100$, $\sigma^2 = 400$ if $n_3 > 0$, and the sequence of p's is 0.5, 0.6, 0.7, 0.8. The results here are similar to those that have gone before, but note that the final average error in run #4 is 0.147, a rather large value, presumably a result of the fact that it is based on 0.8(130) = 104 observations, some of which must, of necessity, be outliers or belong to the subpopulation.

These Monte Carlo results, though they are neither terribly comprehensive nor terribly accurate (in the sense of having small standard errors) do reassure us that IET behaves basically in the way that intuition would expect it to. We have found that IET is insensitive to outliers and small subpopulations that are sufficiently far away, and that it does hone in on the major cluster we are seeking. These results are not directly related to the theoretical result in Chapter 3 concerning the distribution of the stationary point, since we did not let IET iterate to a limit but rather had it iterate a predecided number of times. It would be interesting, however, to carry out

another Monte Carlo study in order to get an idea of what, on the one hand, the small sample distribution of the stationary point looks like for spherically symmetric observations and how, on the other, it depends on the presence of various kinds of contamination.

The reader may wonder why we chose to report "absolute" errors rather than squared errors. The reason is that when we began doing simulations, we found that rather large values of $e_i^{(j)}$ would occur with rather surprising frequency. We were afraid that these occasional large values, if squared, would have an undue amount of influence. Reporting absolute errors was a way of downweighting that influence.

Chapter 5 - Scale Estimation

As we pointed out in Chapter 1, the estimate of covariance yielded by IET, the final $\tilde{Z}$, is biased towards 0. It turns out, in fact, that $\tilde{Z}$ is an estimate of cZ where c is an unknown constant between 0 and 1. In essence, then, IET provides an estimate of both the shape (the relative dimensions of the axes) and the orientation but not of the scale of the ellipsoid corresponding to Z. It is for this reason that we devote this chapter to scale estimation. First we shall describe the basic device with which we intend to obtain an unbiased estimate of Z. Then we shall discuss some properties of this method and investigate its efficiency. In addition, a few alternative scale estimators will be considered and compared.

Let us suppose that the conjecture at the end of Chapter 3 is correct: namely, that the stopping point of IET is guaranteed to be within $O_p(n^{-1/2})$ of the stationary point whose distribution we have derived. Then it follows that the limiting $\tilde{Z}$ produced by IET will, as $n \to \infty$, converge to the population covariance of the truncated ellipsoidal distribution given in equation (A1.22), $c(k,p)Z$, where $c(k,p) = F_{k+2}(a^2)/F_k(a^2)$ and $a^2 = F_k^{-1}(p)$. It would then be plausible to obtain an estimate of Z by

dividing $\tilde{Z}$ by $c(k,p)$

$$\tilde{Z}_U = c(k,p)^{-1}\tilde{Z} \tag{5.1}$$

(The "U" in $\tilde{Z}_U$ is supposed to indicate that it is an asymptotically unbiased estimate.)

The parameter p, which appears in $c(k,p)$, represents the proportion of points, from the cluster we have converged to included in successive IET estimates, and not the proportion of points from the entire sample, which was the meaning of p in Chapter 2. In Chapter 3, the two proportions coincided because there was only one cluster, but typically in our applications there will always be several clusters. So it is necessary to estimate p in order to compute $\tilde{Z}_U$.

One might object that if it is necessary to "know" approximately how many observations are in the cluster before one can scale up $\tilde{Z}$ to get an unbiased estimate, then there is little point to this estimation procedure, for one could simply use all of the points thought to belong to the cluster and thereby obtain an unbiased estimate directly. Though there is some justice to this criticism, it must be remarked that the latter estimate would surely be much less robust, as the observations

with the largest influence on it are precisely those points about whose correct classification we are least certain. Earlier we suggested one way to approximate the number of observations in a cluster: after reaching a stopping point, gradually increase the number of points included in the ellipsoid until IET begins to move away. Here the basic idea is to base $\tilde{Z}$ only on points whose membership in the cluster is reasonably certain; then, obtain the total number of points in the cluster by throwing in points which are likely to belong but about whose classification we nevertheless entertain some doubt. Another possibility, which really embodies the same idea is to plot a histogram of the D_i^2's. If we are lucky, there will be a relatively clean cutoff as there is in Figure 5.1. Unfortunately such cutoffs are rather uncommon.

One consolation is that it is easy to compute the bias introduced by using the wrong p in the equation for $\tilde{Z}_U$. More precisely, if one uses p' instead of p, then asymptotically, $\tilde{Z}_U$ will be off by the factor $c(k,p')/c(k,p)$. Indeed, for many ellipsoidal distributions of interest, $c(k,p)$ is increasing in p (in particular for the normal distribution: see Lemma A2.1 and equation (A2.9)). Hence if one felt that p lay in a certain interval, then $c(k,p)$ would lie in a corresponding

interval and one could study the variation in $\tilde{\bar{Y}}_U$ with p.

Later in this chapter we will introduce an alternative estimator, called $\tilde{v}_3$, which does not require an estimate of p. Unfortunately we shall find that it is much less efficient than a close relative of $\tilde{\bar{Y}}_U$ called $\tilde{v}_1$, at least when the data are normally distributed.

But before going on to develop our ideas further, perhaps it will be helpful to examine some numerical values of $c(k,p)$ (for the normal case), which may be found in Table 5.1. When $k = 1$ and we include 25% of the observations, $\tilde{\sigma}^2$ is only 3.3% of σ^2, on the average, but if $k = 7$ and 25% of the data is included, $\tilde{\bar{Y}}$ is 42.5% of what it should be, on the average. Further values of $c(k,p)$ may be obtained from Table A2.1, after making use of the result in (A2.9), that $c(k,p) = 1 - b_k((F_k^{-1}(p))^{1/2})$. Note that (A2.9) is not true for all ellipsoidal families; it does hold, however, for the normal one. It is a remarkable coincidence that the bias reducing factor in chapter 3, $b_k(a)$, is related in this way to the covariance of the truncated normal.

Lemma A2.2 implies that

$$(5.2) \quad c(k,p) \simeq 1 - (2/k)^{1/2} \frac{\phi(\phi^{-1}(p))}{p} \quad \text{for large } k$$

and

$$(5.3)\quad c(k,p) \sim \frac{\Gamma((k/2)+1)^{2/k}}{(k/2)+1} p^{2/k} \quad \text{for small} \quad p$$

Hence, as $k \to \infty$, $c(k,p) \to 1$ as long as $p > 0$. Furthermore, $c(1,p) = O(p^2)$, $c(2,p) = O(p)$, and $c(7,p) = O(p^{2/7})$ as $p \to 0$, which is to say that $c(k,p) \to 0$ as $p \to 0$ more and more slowly as k increases. This fact accounts for the results in the above discussion when $p = 0.25$.

We would like to investigate the efficiency of the estimator $\tilde{\bar{X}}_U$. The most satisfying approach to this question would be to define a loss function, $L(\bar{X},\tilde{\bar{X}}_U)$, and compute its expected value as $n \to \infty$ when the data belongs to some family of ellipsoidal distributions. Then one would compare the asymptotic risk to the corresponding value for the optimal procedure. Unfortunately this program is a difficult one to carry out. Therefore, we shall make a variety of simplifying assumptions so as to make the problem more tractable. Later, we will return to comment upon how much we have lost in making these assumptions.

Henceforth we shall assume that there are n independent observations: $X_1,\ldots,X_n \sim N(0,vB)$, where B is a known $k \times k$ matrix such that $|B| = 1$. We shall be interested

in estimating v, which we shall sometimes refer to as the scale of the distribution. Note that in this formulation there is no contamination and, as a result, no ambiguity associated with "p". It is convenient to introduce some further notation at this point. Let $Z_i = X_i' B^{-1} X_i$ and note that $Z_i \sim v\chi^2_{(k)}$; then define $J_i = 1(Z_i \leq t^2)$, where $1(E)$ for an event E is the indicator function of that event, and observe that J_i is a Bernoulli random variable with parameter $p = F_k(t^2/v)$. (As before, $a^2 = F_k^{-1}(p)$.) We will often refer to the statistics $\bar{J} = n^{-1}\Sigma J_i$, $\overline{JZ} = n^{-1}\Sigma J_i Z_i$, and $\bar{Z} = n^{-1}\Sigma Z_i$. It will be helpful to write $\tilde{p} = \bar{J}$, since $\bar{J}$ is an estimate of p, and $\tilde{a}^2 = F_k^{-1}(\tilde{p})$. Finally we introduce $S = (n\bar{J})^{-1}\Sigma J_i X_i X_i'$.

Now we may present our first estimator of scale, $\tilde{v}_1$:

$$\tilde{v}_1 = \frac{\overline{JZ}/\bar{J}}{kc(k,\tilde{p})}. \tag{5.4}$$

The first natural question to ask is: why is this a plausible estimate of v? The numerator of (5.4) is essentially v multiplied by the sample mean of $n\tilde{p}$ observations of a $\chi^2_{(k)}$ truncated at a^2. On the other hand, the denominator is the expectation of a $\chi^2_{(k)}$ truncated at $\tilde{a}^2$. This latter observation follows from

(A1.7), (A1.17), (A2.3) and (A1.23). Of course we would prefer to use the expectation of $\chi^2_{(k)}$ truncated at a^2, but since $a^2 = t^2/v$ depends on v, we use the estimate $\tilde{a}^2$ instead. Nevertheless, $\tilde{v}_1$ is a consistent estimator of v, a fact which can easily be shown by applying the weak law of large numbers to $\bar{J}$ and $\overline{JZ}$ and using the continuity of F_k and b_k.

Since $\tilde{\Sigma}_U$ is computed by truncating at a random point and we are interested in studying simple estimators that are similar to it, one might well ask why we truncated at a fixed point in computing $\tilde{v}_1$. Indeed, we next introduce $\tilde{v}_1'$, an estimator which is identical to $\tilde{v}_1$ except that the truncation point is now random and such that exactly $[np]$ observations are small enough to be included. It will turn out however (see Theorem 5.6) that $\tilde{v}_1$ and $\tilde{v}_1'$ are asymptotically equivalent; as a result we can study whichever one is more convenient and that one is $\tilde{v}_1$. Before giving the formula for $\tilde{v}_1'$, it will be convenient to set $J_i' = 1(Z_i \leq Z_{([np])})$. Then,

$$\tilde{v}_1' = \frac{\overline{J'Z}/\overline{J'}}{kc(k,p)} = \frac{\Sigma J_i' Z_i}{[np]kc(k,p)} \tag{5.5}$$

Formula (5.5) is simply (5.4) with J_i replaced by J_i'

(with the change that since p is now "fixed" by the statistician, $a^2 = F_k^{-1}(p)$ is now known). The duality associated with either fixing t to get $\tilde{v}_1$, or fixing p to get $\tilde{v}_1'$ appears to be very general. For example, the "t-fixed" estimators $\tilde{v}_2$, $\tilde{v}_3$, $\tilde{v}_4$ to be studied shortly all have "p-fixed" counterparts $\tilde{v}_2'$, $\tilde{v}_3'$, $\tilde{v}_4'$. We believe that the proof of Theorem 5.6, which asserts the asymptotic equivalence of $\tilde{v}_1$ and $\tilde{v}_1'$ as well as of $\tilde{v}_2$ and $\tilde{v}_2'$, will provide some insight into this phenomenon. However, we conjecture that there is a theorem, or perhaps a metatheorem, which is yet to be formulated precisely, that would make clear the general nature of this duality. Such a theorem, we believe, would for example obviate the need for working with order statistics and their asymptotic distributions in many cases, as in the derivation of the asymptotic distribution of the median or the trimmed mean.

By now the reader must be wondering why we are so interested in $\tilde{v}_1$ and $\tilde{v}_1'$. Let $S' = (n\bar{J}')^{-1}\Sigma J_i' X_i X_i'$ and observe that another formula for $\tilde{v}_1'$ is

$$\tilde{v}_1' = \frac{k^{-1}\mathrm{tr}\,(S'B^{-1})}{c(k,p)}$$

Now suppose that we were to regard $\tilde{u}$ and

$\tilde{\Sigma}/|\tilde{\Sigma}|^{1/k}$ (obtained from IET) as "correct". Then, we could subtract off $\tilde{\mu}$ from each of the n observations, so that their population mean would be zero, and set $B = \tilde{\Sigma}/|\tilde{\Sigma}|^{1/k}$. It would follow that, since $S' = \tilde{\Sigma}$, we can write $\tilde{v}_1' = c(k,p)^{-1}|\tilde{\Sigma}|^{1/k}$. But then

$$\tilde{v}_1' B = \tilde{\Sigma}_U$$

Essentially we have factored $\tilde{\Sigma}_U$ into two parts and have chosen to study the variability of only the first, or scale, part.

Before going on to compute the asymptotic distribution of $\tilde{v}_1$, we shall introduce several alternative scale estimators. It is appropriate to consider first the MLE of v based on all of the data (i.e. $X_1,\ldots,X_n$), which we shall call $\tilde{v}_0$. Using the fact that the log-likelihood of one observation is $\ell(v) = C - (k/2)\log v - (2v)^{-1}x'B^{-1}x$, a simple calculation shows that the Fisher information associated with $\tilde{v}_0$ is $I_0(v) = k/2v^2$. It follows that

$$(5.6) \quad \mathcal{L}(n^{1/2}(\tilde{v}_0 - v)) \to N(0, V_0) \quad \text{as} \quad n \to \infty$$

where

(5.7) $$V_0 = 2v^2/k.$$

The formula for $I_0(v)$ is quite reasonable as it asserts that the information in k 1-dimensional observations is the same as that in one k-dimensional observation.

The next estimator, $\tilde{v}_2$, we will refer to as a "Bernoulli-type" estimator because it only depends on the J_i's:

(5.8) $$\tilde{v}_2 = t^2/F_k^{-1}(\tilde{p}).$$

Actually, $\tilde{v}_2$ is a maximum likelihood estimator, as we shall see later. The "p-fixed" version of $\tilde{v}_2$ is

(5.9) $$\tilde{v}_2' = Z_{([np])}/F_k^{-1}(p).$$

We discussed earlier how in any application of IET, we would not know the proportion of points, p, from a given cluster included in the final ellipsoid. The next estimator, $\tilde{v}_3$, offers the possibility of surmounting this problem in a formal way (as opposed to the ad-hoc ways we mentioned before). We define $\tilde{v}_3$ to be the MLE of v using the truncated normal density, which is

$$(5.10)\quad p_t(x;v)\begin{cases} = (2\pi v)^{-k/2}(F_k(t^2/v))^{-1}\exp\left(-\frac{1}{2v}x'B^{-1}x\right) \\ \qquad\qquad \text{for } x'B^{-1}x \le t^2 \\ = 0 \qquad\qquad \text{otherwise} \end{cases}$$

based of course on the M observations X_i such that $J_i = 1$. This estimator does not use any information concerning how many observations had $J_i = 0$. Of course, $\tilde{v}_3'$ is the MLE of v based on the m smallest Z_i's, where now m is not random. We conjecture but shall not prove that the limiting distribution of $\tilde{v}_3'$ is the same as that of $\tilde{v}_3$. Incidentally, Cohen extensively investigated estimation problems involving the truncated normal distribution in the 1950's, but never, as far as we can tell, studied the ellipsoidally truncated normal. See for example [4].

Finally, $\tilde{v}_4$ is the MLE for v based on observations of the censored normal distribution, which is to say that we "see" X_i if $J_i = 1$ but learn only that $J_i = 0$ otherwise. Therefore, this estimator is based on knowing how many observations are "missing", in contrast to $\tilde{v}_3$, which was based only on the X_i's with $J_i = 1$. Actually, $\tilde{v}_4$, then, makes use of exactly the same information as did $\tilde{v}_1$: $\bar{J}$ and $\overline{JZ}$; and it is therefore particularly

appropriate to compare then. Again we conjecture but shall not prove that $\tilde{v}_4$ has the same asymptotic distribution as $\tilde{v}_4'$, which is the MLE of v based on observing the [np] smallest Z_i's and knowing that there were n observations altogether. Incidentally, the likelihood for one observation corresponding to $\tilde{v}_4$ is

$$(5.11)\qquad \begin{cases} (2\pi v)^{-k/2}\exp\left(-\frac{1}{2}\frac{X'B^{-1}X}{v}\right) & \text{if } X'B^{-1}X \le t^2 \\ 1 - F_k(t^2/v) & \text{if } X'B^{-1}X > t^2. \end{cases}$$

Now we are ready to derive the asymptotic distributions of $\tilde{v}_1$ to $\tilde{v}_4$.

Theorem 5.1: As $n \to \infty$,

$$(5.12)\qquad L(n^{1/2}(\tilde{v}_1 - v)) \to N(0, V_1)$$

where

$$(5.13)\qquad V_1 = \frac{v^2}{F_{k+2}^2(a^2)}\left[F_k(a^2)(1-F_k(a^2))\frac{a^4}{k^2}\right.$$

$$\left. -2(1-F_k(a^2))F_{k+2}(a^2)\frac{a^2}{k} + \left(\frac{k+2}{k}\right)F_{k+4}(a^2) - F_{k+2}^2(a^2)\right]$$

Proof: $\tilde{v}_1$ is a function only of the statistics $(\bar{J},\overline{JZ})$. We will derive the joint asymptotic distribution of these two random quantities and obtain the asymptotic distribution of $\tilde{v}_1$ by applying the delta method. By the central limit theorem,

$$\mathcal{L}(n^{1/2}((\bar{J},\overline{JZ})' - (\mu_1,\mu_2)')) \to N(0,C)$$

where

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{pmatrix}$$

and $(\mu_1,\mu_2)'$, and C are, respectively, the mean and covariance of $(J_i,J_iZ_i)'$. Since J_i is Bernoulli, $\mu_1 = F_k(a^2)$, and by (A1.8), $\mu_2 = vkF_{k+2}(a^2)$. The variance of J_i is $c_{11} = F_k(a^2)(1 - F_k(a^2))$ and by (A1.8), $c_{12} = vk(1 - F_k(a^2))F_{k+2}(a^2)$. Finally, by (A1.8) and (A1.9), $c_{22} = v^2[k(k+2)F_{k+4}(a^2) - k^2F^2_{k+2}(a^2)]$. Using (A1.23), we may rewrite (5.4) as

$$\tilde{v}_1 = \frac{\overline{JZ}}{kF_{k+2}(F_k^{-1}(\bar{J}))}.$$

If we define the function

$$h(u_1,u_2) = \frac{u_2}{kF_{k+2}(F_k^{-1}(u_1))} ,$$

then $\tilde{v}_1 = h(\bar{J},\overline{JZ})$ and $v = h(\mu_1,\mu_2)$. Applying the so called delta method we have the result that as $n \to \infty$, $L(\sqrt{n}(\tilde{v}_1-v)) \to N(0,V_1)$ where

$$V_1 = \left(\frac{dh}{du}\right)' C \left(\frac{dh}{du}\right) \tag{5.14}$$

and the derivative above is evaluated at $u^{(0)} = (\mu_1,\mu_2)$. It is easy to compute that

$$\left.\frac{\partial h}{\partial u_1}\right|_{u^{(0)}} = - v \frac{f_{k+2}(a^2)}{F_{k+2}(a^2) f_k(a^2)}$$

and

$$\left.\frac{\partial h}{\partial u_2}\right|_{u^{(0)}} = \frac{1}{kF_{k+2}(a^2)} .$$

Then, using the fact that

$$\frac{f_{k+2}(a^2)}{f_k(a^2)} = \frac{a^2}{k} ,$$

a straightforward evaluation of (5.14) reveals that (5.13) holds. ∎

Before deriving the asymptotic distributions of the remaining estimators, it will be convenient to record a simple result concerning the effect of a reparameterization on the Fisher information.

Lemma 5.1: Let $I_1(\theta)$ and $I_2(\phi)$ be the Fisher informations associated with $f(x;\theta)$ and $f(x;g(\phi))$, respectively. If θ and ϕ are scalars and g is a differentiable strictly monotone transformation, then

$$I_2(\phi) = (g'(\phi))^2 I_1(g(\phi)). \tag{5.15}$$

Proof: The proof is a simple calculation. ∎

Next we derive the asymptotic distribution of $\tilde{v}_2$.

Theorem 5.2: As $n \to \infty$,

$$\mathcal{L}(n^{1/2}(\tilde{v}_2 - v)) \to N(0, V_2) \tag{5.16}$$

where

$$V_2 = v^2 \frac{F_k(a^2)(1-F_k(a^2))}{a^4 f_k^2(a^2)}. \tag{5.17}$$

Proof: Note that $\tilde{v}_2$, as defined in (5.9), is the MLE of v when we observe $J_1,\ldots,J_n$, which are i.i.d. Bernoulli random variables with parameter $p = F_k(t^2/v)$. Using the notation of Lemma 5.1, $I_1(p) = [p(1-p)]^{-1}$. Furthermore, since $\frac{dp}{dv} = -v^{-1}a^2 f_k(a^2)$,

$$I_2(v) = \frac{1}{v^2}(a^2 f_k(a^2))^2 [F_k(a^2)(1-F_k(a^2))]^{-1}$$

which is equivalent to (5.17) since $V_2 = I_2(v)^{-1}$. ∎

If there are m observations of X such that $X'B^{-1}X \leq t^2$, then using (5.10), the log-likelihood expressed as a function of $a = tv^{-1/2}$ is $\ell(a)$ where

$$(5.18)\quad m^{-1}\ell(a) = C + k \log a - \frac{1}{2}\left(\frac{tr\ (SB^{-1})}{t^2}\right)a^2 - \log F_k(a^2).$$

Then, the likelihood equation, $\frac{d\ell}{da} = 0$, is

$$(5.19)\qquad k^{-1}\ \frac{tr\ (SB^{-1})}{t^2}a^2 + b_k(a) = 1.$$

If we let $\tilde{a}_3$ be the solution to (5.19) and observe that by definition, $\tilde{v}_3 = t^2/\tilde{a}_3^2$, then we may write

$$\tilde{v}_3 = \frac{k^{-1}tr\ (SB^{-1})}{1 - b_k(\tilde{a}_3)} = \frac{k^{-1}tr\ (SB^{-1})}{c(k,F_k(\tilde{a}_3^2))}$$

an equation which strongly resembles (5.5). The next theorem shows that $\tilde{v}_3 = \infty$ with non-zero probability for a finite sample size.

Theorem 5.3:

$$(5.20) \qquad \frac{tr\ (SB^{-1})}{t^2} \geq \frac{k}{k+2} \Rightarrow \tilde{v}_3 = \infty$$

$$(5.21) \qquad \frac{tr\ (SB^{-1})}{t^2} < \frac{k}{k+2} \Rightarrow \tilde{v}_3 = \frac{t^2}{\tilde{a}_3^2},$$

where $\tilde{a}_3$ is the unique finite positive solution to the likelihood equation (5.19).

To prove this theorem we shall need a lemma, which is proved in Appendix 3.

Lemma 5.2: The function $s(a) = ka^{-2}(1 - b_k(a))$ is monotone decreasing for $a > 0$. As $a \to 0$, $s(a) \to k/(k+2)$ and as $a \to \infty$, $s(a) \to 0$.

Proof of Theorem 5.3: We may expand $F_k(a^2)$ in powers of a as in (A2.12) to obtain

$$F_k(a^2) = 2r_k \frac{a^k}{k}\left(1 - \frac{a^2}{2}\left(\frac{k}{k+2}\right) + \frac{a^4}{8}\left(\frac{k}{k+4}\right) + o(a^4)\right)$$

Then,

$$\log F_k(a^2) = \text{const.} + k \log a - \frac{a^2}{2}\left(\frac{k}{k+2}\right) + \frac{a^4}{2}\frac{k}{(k+4)(k+2)^2} + o(a^4)$$

and as a result of (5.18)

$$(5.22)\quad m^{-1}\ell(a) = \text{const.} + \frac{a^2}{2}\left(\frac{k}{k+2} - \frac{\text{tr}\,(SB^{-1})}{t^2}\right) - \frac{a^4}{2}\frac{k}{(k+4)(k+2)^2} + o(a^4).$$

Now let $c^2 = t^{-2}\,\text{tr}\,(SB^{-1})$ and observe that $(am)^{-1}\frac{d\ell}{da} = ka^{-2}(1 - b_k(a)) - c^2 = s(a) - c^2$. If $c^2 \geq k/k+2$, then by Lemma 5.2, $s(a) - c^2 < 0 \quad \forall a > 0$ and therefore $\frac{d\ell}{da} < 0 \quad \forall a > 0$. Hence, the maximum value of $\ell(a)$ is achieved at $a = 0$. On the other hand, if $c^2 < k/k+2$, then there is a unique a_c such that $s(a_c) = c^2$ by Lemma 5.2. Furthermore, a_c is the unique $a > 0$ such that $\frac{d\ell}{da} = 0$. But, by (5.22), $\ell(a)$ is increasing for a near 0; since $\ell(a) \to -\infty$ as $a \to \infty$, there is at least one local maximum at an $a > 0$. We may now conclude that the global maximum occurs at a_c. Of course, $\tilde{a}_3 = a_c$. ∎

Why is it that when $t^{-2}\text{tr}\,(SB^{-1}) \geq k/k+2$, $\tilde{v}_3 = \infty$? Suppose that X has a uniform distribution in

the ellipsoid $\{x : x'B^{-1}x \leq t^2\}$ and let $Y = B^{-1/2}X$. Then $Y'Y = X'B^{-1}X$ and Y has a uniform distribution in $S_t(0)$, the sphere of radius t. By (A1.5), the volume of this sphere is $d_k t^k$, where $d_k = \pi^{k/2}/\Gamma((k/2)+1)$, and its surface area is $kd_k t^{k-1}$. Hence

$$E(Y'Y) = (d_k t^k)^{-1} \int_0^t r^2 k d_k r^{k-1} dr = (k/(k+2))t^2$$

and $t^{-2}E(X'B^{-1}X) = k/k+2$. Therefore, when $t^{-2}tr\,(SB^{-1}) \geq k/k+2$, the sample looks, at best, as if it is from a uniform distribution. Of course, as $v \to \infty$, the truncated normal approaches the uniform distribution.

Before proceeding to the computation of the asymptotic distribution of $\tilde{v}_3$, we investigate numerically the dependence of $\tilde{v}_3$ on S in the one dimensional case. We set $B = 1$ and $s^2 = S$ and note that $\tilde{v}_3 = \infty$ when $s^2 \geq t^2/3$. Table 5.2 contains some numerical values of s^2/t^2 and the corresponding $\tilde{v}_3/s^2$, the factor by which we must multiply s^2 to get our estimate.

Theorem 5.4: As $n \to \infty$

$$(5.23) \qquad \mathcal{L}(n^{1/2}(\tilde{v}_3 - v)) \to N(0, V_3)$$

where

$$(5.24)\quad V_3 = \frac{4v^2}{k}\{F_k(a^2)\,[2-b_k(a)\,(a^2+2-k+kb_k(a))]\}^{-1}$$

Proof: The derivative of the log of the density in (5.10), written as a function of a, is

$$\frac{d \log \phi_t}{da} = \frac{k}{a} - \frac{x'B^{-1}x}{t^2}a - \frac{k}{a}b_k(a)$$

and the second derivative is

$$\frac{d^2 \log \phi_t}{da^2} = -\frac{k}{a^2} - \frac{x'B^{-1}x}{t^2} + \frac{k}{a^2}b_k(a) - \frac{k^2}{a^2}b_k(a)\left(1 - \frac{a^2}{k} - b_k(a)\right).$$

Using the fact that

$$E\left(\frac{X'B^{-1}X}{t^2}\right) = \frac{k}{a^2}(1 - b_k(a)),$$

where the expectation is taken with respect to ϕ_t, the density of the truncated distribution, and recalling (A1.7), (A2.7), and $t^2 = a^2v$, we find that the Fisher information for the parameter a is

$$(5.25)\quad I_1(a) = 2\frac{k}{a^2}(1 - b_k(a)) + \frac{k^2}{a^2}b_k(a)\left(1 - \frac{a^2}{k} - b_k(a)\right).$$

Since $\frac{da}{dv} = -\frac{1}{2}tv^{-3/2}$, it follows by Lemma 5.1 that the

information associated with v is

$$(5.26) \qquad I_2(v) = \frac{a^2}{4v^2} I_1(a).$$

As $m \to \infty$,

$$(5.27) \qquad \mathcal{L}(m^{1/2}(\tilde{v}_3 - v)) \to N(0, I_2(v)^{-1}).$$

But as $n \to \infty$, $m/n \to p = F_k(a^2)$. Therefore applying (5.25), (5.26), and (5.27), we have the result in (5.23) with $V_3 = (F_k(a^2) I_2(v))^{-1}$. ∎

Theorem 5.5: As $n \to \infty$

$$(5.28) \qquad \mathcal{L}(n^{1/2}(\tilde{v}_4 - v)) \to N(0, V_4)$$

where

$$(5.29) \qquad V_4^{-1} = V_2^{-1} + V_3^{-1}.$$

Proof: The proof is a calculation similar to those done in the proofs of Theorems 5.2 and 5.4. ∎

The fact that $V_4^{-1} = V_2^{-1} + V_3^{-1}$ in the preceding theorem is an instance of a very general theorem suggested

by P. K. Bhattacharya. We will not give a formal proof but will simply outline the argument for a simple version of it. Suppose X has a density $f_\theta(x)$ and suppose further that the space in which X takes values, $\mathcal{X}$, is partitioned into two parts: A and $\bar{A}$; so $\mathcal{X} = A \cup \bar{A}$. Let $P_\theta = \int_A f_\theta(x)dx$ and define three new random variables, 2 truncated versions of X and an indicator based on X:

$$X_1 = X \cdot 1[X \in A], \tag{5.30}$$

$$X_2 = X \cdot 1[X \in \bar{A}], \tag{5.31}$$

and

$$X_3 = 1[X \in A]. \tag{5.32}$$

Then X_1 has density

$$f_\theta^{(1)}(x) = P_\theta^{-1} f_\theta(x) 1[x \in A]$$

and X_2 has density

$$f_\theta^{(2)}(x) = (1-P_\theta)^{-1} f_\theta(x) 1[x \in \bar{A}].$$

Of course X_3 is a Bernoulli random variable: $\Pr(X_3 = 1) = P_\theta$. Now let $I_1(\theta)$, $I_2(\theta)$, $I_3(\theta)$ be the Fisher informations associated with X_1, X_2, X_3 and $I(\theta)$ the information associated with X. Then if we suppress some algebra and assume that all necessary formal manipulations are valid, we may compute

$$(5.33) \qquad I_1(\theta) = \int_A \left[\frac{d}{d\theta}\log \frac{f_\theta}{P_\theta}\right]^2 \frac{f_\theta}{P_\theta}dx$$

$$= \int_A \left[\frac{d}{d\theta}\log f_\theta\right]^2 f_\theta dx - \frac{(P_\theta')^2}{P_\theta^2} ,$$

$$(5.34) \qquad I_2(\theta) = \int_{\bar{A}} \left[\frac{d}{d\theta}\log \frac{f_\theta}{1-P_\theta}\right]^2 \frac{f_\theta}{1-P_\theta}dx$$

$$= \int_{\bar{A}} \left[\frac{d}{d\theta}\log f_\theta\right]^2 f_\theta dx - \frac{(P_\theta')^2}{(1-P_\theta)^2} ,$$

and

$$(5.35) \qquad I_3(\theta) = \frac{(P_\theta')^2}{P_\theta(1-P_\theta)}$$

where (5.35) follows from Lemma 5.1. But using (5.33), (5.34), and (5.35), we find that

$$(5.36) \qquad I(\theta) = P_\theta I_1(\theta) + (1-P_\theta)I_2(\theta) + I_3(\theta).$$

The factors P_θ and $(1-P_\theta)$ appear in (5.36) because X_1 and X_2 are only non-zero with probabilities P_θ and $1-P_\theta$.

What we have shown is that the information in X can be partitioned into three easily interpreted parts: information from 2 complementary truncated random variables and information from a Bernoulli random variable which says which truncated variable is observed. No doubt a theorem stating that such a partitioning is possible can be proved in considerable generality.

We have now studied five estimators of v: $\tilde{v}_0$ through $\tilde{v}_4$ and have obtained their asymptotic variances: V_0 through V_4 given in (5.7), (5.13), (5.17), (5.24), and (5.29). We define the efficiency of $\tilde{v}_i$ to be $E_i = V_0/V_i$ for $1 \leq i \leq 4$. Of course by Theorem 5.5, $E_4 = E_2 + E_3$. We also expect that $E_4 \geq E_1$, since $\tilde{v}_4$ and $\tilde{v}_1$ use the same data and $\tilde{v}_4$ is the MLE. In Table 5.2 we give numerical values for E_1 through E_4 for $k = 1,\ldots,7$ and $p = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.999$. Note that all of the efficiencies depend only on $a = tv^{-1/2}$ or, alternatively, on $p = F_k(a^2)$.

Several interesting remarks may be made about the results in Table 5.3. Note first that $\tilde{v}_3$ is extremely

inefficient: for example, when $k = 1$ and $p = 0.7$, $E_3 = 0.032$. This means that if one bases $\tilde{v}_3$ on 100r observations, one will do about as well as one would do using $\tilde{v}_0$ when there were 3r observations. It is important to remember, of course, that only about 70r observations will be used in computing $\tilde{v}_3$; the others will be truncated. Still, however, the result is striking. A crucial point is that using $\tilde{v}_3$ is the best we can do if we decide that we cannot guess the proportion of points that are not truncated (included in the estimate).

When $k = 1$ and $p = 0.7$, $E_2 = 0.556$; hence, not knowing any of the actual values of the observations results in a loss of only about 50% of the information. This remark is less surprising when one considers the similarity between the mode of estimation used in $\tilde{v}_2'$ and, for instance, the use of the median to estimate the mean of a normal (see equation (5.9)). Our next theorem will demonstrate that $\tilde{v}_2$ and $\tilde{v}_2'$ are asymptotically equivalent. Finally, we observe that, in general, E_1/E_4 is about 0.8; hence, the loss of information due to not using the MLE is not severe. It seems appropriate to conclude that $\tilde{Z}_U$ is likely to be a reasonably efficient estimate.

Theorem 5.6: As $n \to \infty$

(5.37) $\mathcal{L}(n^{1/2}(\tilde{v}_1' - v)) \to N(0,V_1)$

and

(5.38) $\mathcal{L}(n^{1/2}(\tilde{v}_2' - v)) \to N(0,V_2)$.

Proof: We will prove the theorem by demonstrating that

(5.39) $\tilde{v}_1' = \tilde{v}_1 + o_p(n^{-1/2})$

and

(5.40) $\tilde{v}_2' = \tilde{v}_2 + o_p(n^{-1/2})$.

First we derive (5.40). We will write F_{kn} for the empirical distribution corresponding to F_k. It is possible to rewrite (5.8) and (5.9) as

(5.41) $$\tilde{v}_2 = v\,\frac{F_k^{-1}(p)}{F_k^{-1}(\tilde{p})}$$

and

(5.42) $$\tilde{v}_2' = v\,\frac{F_{kn}^{-1}(p)}{F_k^{-1}(p)}$$

By a Taylor expansion, since $\tilde{p} - p = o_p(n^{-1/2})$,

$$(5.43) \qquad F_k^{-1}(\tilde{p}) = F_k^{-1}(p) + \frac{\tilde{p} - p}{f_k(F_k^{-1}(p))} + o_p(n^{-1/2})$$

and using the Bahadur representation [2],

$$(5.44) \qquad F_{kn}^{-1}(p) = F_k^{-1}(p) + \frac{p - \tilde{p}}{f_k(F_k^{-1}(p))} + o_p(n^{-1/2}).$$

Hence, by (5.43)

$$(5.45) \qquad \frac{F_k^{-1}(p)}{F_k^{-1}(\tilde{p})} = 1 - \frac{\tilde{p} - p}{F_k^{-1}(p) f_k(F_k^{-1}(p))} + o_p(n^{-1/2})$$

and from (5.44) we derive

$$(5.46) \qquad \frac{F_{kn}^{-1}(p)}{F_k^{-1}(p)} = 1 - \frac{\tilde{p} - p}{F_k^{-1}(p) f_k(F_k^{-1}(p))} + o_p(n^{-1/2}).$$

But (5.45) and (5.46) imply (5.40). The argument for (5.39) is very similar. First observe that we may rewrite (5.4) and (5.5) as follows:

$$(5.47) \qquad \tilde{v}_1 = v\left[\frac{\int_0^{F_k^{-1}(p)} x dF_{kn}}{\int_0^{F_k^{-1}(\tilde{p})} x dF_k}\right] \frac{n\tilde{p}}{[n\tilde{p}]}$$

and

$$\tilde{v}_1' = v\left[\frac{\int_0^{F_{kn}^{-1}(p)} x\,dF_{kn}}{\int_0^{F_k^{-1}(p)} x\,dF_k}\right] \frac{np}{[np]} \tag{5.48}$$

Using a Taylor expansion we may write

$$\int_0^{F_k^{-1}(\tilde{p})} x\,dF_k = \int_0^{F_k^{-1}(p)} x\,dF_k + F_k^{-1}(p)(\tilde{p}-p) + o_p(n^{-1/2}), \tag{5.49}$$

and by the uniformity lemma of chapter 3 (Lemma 3.3), since (5.44) holds and $\tilde{p}-p = o_p(n^{-1/2})$,

$$\begin{aligned}\int_0^{F_{kn}^{-1}(p)} x\,dF_{kn} &= \int_0^{F_k^{-1}(p)} x\,dF_{kn} + \int_0^{F_{kn}^{-1}(p)} x\,dF_k \\ &\quad - \int_0^{F_k^{-1}(p)} x\,dF_k + o_p(n^{-1/2}) \\ &= \int_0^{F_k^{-1}(p)} x\,dF_{kn} \\ &\quad + F_k^{-1}(p) f_k(F_k^{-1}(p))(F_{kn}^{-1}(p)-F_k^{-1}(p)) \\ &\quad + o_p(n^{-1/2}).\end{aligned} \tag{5.50}$$

Substituting (5.44) in (5.50) we obtain

$$(5.51)\quad \int_0^{F_{kn}^{-1}} x dF_{kn} = \int_0^{F_k^{-1}(p)} x dF_{kn} + F_k^{-1}(p)(p-\tilde{p}) + o_p(n^{-1/2}).$$

If we divide $\int_0^{F_k^{-1}} x dF_{kn}$ by (5.49) and expand, then we find

$$(5.52)\quad \frac{\int_0^{F_k^{-1}(p)} x dF_{kn}}{\int_0^{F_k^{-1}(\tilde{p})} x dF_{kn}} = \frac{\int_0^{F_k^{-1}(p)} x dF_{kn}}{\int_0^{F_k^{-1}(p)} x dF_k}\left(1 - \frac{F_k^{-1}(p)(\tilde{p}-p)}{\int_0^{F_k^{-1}(p)} x dF_k}\right) + o_p(n^{-1/2})$$

Since $\int_0^{F_k^{-1}(p)} x dF_{kn} = \int_0^{F_k^{-1}(p)} x dF_k + O_p(n^{-1/2})$, we may write (5.52), using (5.47), as

$$(5.53)\quad \tilde{v}_1/v = \frac{\int_0^{F_k^{-1}(p)} x dF_{kn}}{\int_0^{F_k^{-1}(p)} x dF_k} - \frac{F_k^{-1}(p)(\tilde{p}-p)}{\int_0^{F_k^{-1}(p)} x dF_k} + o_p(n^{-1/2})$$

But using (5.48), and dividing (5.51) by $\int_0^{F_k^{-1}(p)} x dF_k$, we may conclude

$$(5.54)\quad \tilde{v}_1'/v = \frac{\int_0^{F_k^{-1}(p)} x dF_{kn}}{\int_0^{F_k^{-1}(p)} x dF_k} - \frac{F_k^{-1}(p)(\tilde{p}-p)}{\int_0^{F_k^{-1}(p)} x dF_k} + o_p(n^{-1/2}),$$

which demonstrates (5.39). ∎

All of the analysis we have done in this chapter has been based on the assumption that B is known and that the mean of the normal distribution we are working with is 0, or, equivalently, that it is known, say equal to μ. What relevance do our results have when μ, B are themselves unknown, as is the case in most applications?

It is interesting that the information matrix $I(v,\mu,B)$ has a block diagonal structure, that is,

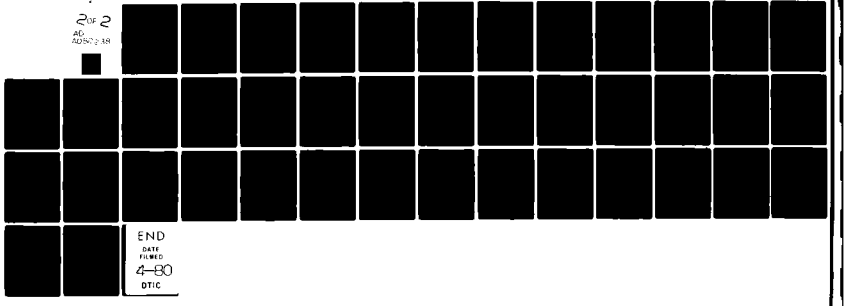
$$(5.55) \qquad I(v,\mu,B) = \begin{pmatrix} I_1(v) & 0 \\ 0 & I_2(\mu,B) \end{pmatrix}$$

when the density of X is

$$(5.56) \quad \phi(x;v,\mu,B) = (2\pi v)^{-k/2} \exp\left(-\frac{1}{2v}(x-\mu)'B^{-1}(x-\mu)\right).$$

To see that this assertion is true, we first reparameterize θ by a differentiable transformation $(\mu,B) = g(\vartheta)$, where ϑ ranges over some open subset of a Euclidean space. (Recall that B is subject to the restriction $|B| = 1$.) Then we may rewrite (5.56) as

AD-A082 238 MASSACHUSETTS INST OF TECH CAMBRIDGE DEPT OF MATHEMATICS F/G 12/1
ITERATIVE ELLIPSOIDAL TRIMMING.(U)
FEB 80 L S GILLICK N00014-75-C-0555
UNCLASSIFIED TR-15 NL



(5.57) $\log \phi(x;v,\theta) = -\frac{k}{2} \log (2\pi v) - \frac{1}{2v} h(x,\theta)$

where h is a differentiable function. By differentiating (5.57) with respect to θ and noting that $E(\frac{\partial}{\partial\theta} \log \phi) = 0$, we conclude that $E(\frac{\partial}{\partial\theta} h(X,\theta)) = 0$. But that implies that $E(\frac{\partial^2}{\partial v \partial\theta} \phi) = E((2v^2)^{-1} \frac{\partial}{\partial\theta} h(X,\theta)) = 0$, which means that (5.55) is true. Of course, the consequence of the fact that the information matrix is block diagonal is that the asymptotic distribution of the MLE of v, $\tilde{v}$, is the same whether (μ,B) is known or is simultaneously estimated with v. So, for instance, the asymptotic variance of $\tilde{v}_0$, $V_0 = 2v^2/k$, given in equation (5.7) is also the asymptotic variance of the MLE of v when μ and B are estimated at the same time.

The information matrix will also be of the form shown in (5.55) if we observe the truncated or censored version of (5.56), the densities for which are given in equations (5.10) and (5.11) (though in these formulas x must be replaced by $x-\mu$). But the block diagonal structure is dependent on having the truncation or censoring done with respect to the <u>correct</u> values of μ and B! Certainly, in our applications the truncation or censoring is done in a data dependent way and therefore does not

satisfy this requirement.

In practice, we expect that the estimates of μ and B to be used in the various scale estimates will be derived from ellipsoidal trimming. Conceivably, then, one could derive the asymptotic distributions of $\tilde{v}_2$, $\tilde{v}_3$, $\tilde{v}_4$ (when μ and B are estimated) using the known asymptotic distribution of the stationary point of the IET algorithm. Such a line of analysis appears to present considerable difficulties, although perhaps they are not insurmountable. The asymptotic distribution of $\tilde{\Sigma}_U$ (see (5.1)) follows as a direct consequence of the distribution of the stationary point; since $\tilde{v}_1' = |\tilde{\Sigma}_U|^{1/k}$, its distribution may be computed directly.

We shall not attempt to carry out any of the above program here and it is true, as a result, that our knowledge concerning the various estimates of scale remains fundamentally incomplete. It is our hope nevertheless that the various numerical and analytical results concerning $V_1,\ldots,V_4$ do provide a rough picture of these estimators. For instance, they provide, at the minimum, lower bounds to the true squared error.

Chapter 6

Conclusion

In this thesis we have studied the iterative ellipsoidal trimming algorithm from a number of different points of view. Based on our experience with IET as a data analytic tool and the plausibility arguments in Chapter 2 concerning its behavior, we would conclude that it can serve an important role as a clustering algorithm when the clusters being sought are approximately ellipsoidal. It will be especially useful when the statistician wishes to simultaneously find a cluster and estimate its location and shape (perhaps as a prelude to searching for smaller hidden clusters in its tails).

In Chapters 3 and 4 we obtained analytical and numerical results concerning the performance of IET in its capacity as an estimation tool. Certainly the most pressing work still to be done in that area is the proof of the conjecture we stated at the end of Chapter 3 concerning the distribution of the stopping point of IET.

The fifth chapter dealt with the problem of estimating the scale of a cluster after one has already obtained estimates of its location and shape. The status of the results presented there is somewhat unsatisfactory, as

the results we derived concerning asymptotic variances of the estimates were based on the assumption that the location and shape are known, rather than estimated.

Appendix 1

Spherically Symmetric Distributions

In this appendix we collect a variety of results concerning spherically symmetric distributions. Suppose that, as in chapter 3, $g(t)$ is a nonnegative function defined on the nonnegative reals whose behavior as $t \to \infty$ and degree of differentiability are suitable for our purposes, where "suitable" means that all of the formal manipulations involving g that we perform are, in fact, valid. We generate a spherically symmetric density for each dimension $k \geq 1$, $f(x) = c_k^{-1} g(|x|^2)$, where $|x|^2 = \sum_{i=1}^{k} x_i^2$. The density of the generalized $\chi^2_{(k)}$ variable, $T = |X|^2$, where $X \sim f(x)$, may easily be derived using (A1.6) and equals

$$(A1.1) \qquad f_k(t) = \frac{\pi^{k/2} t^{(k/2)-1} g(t)}{\Gamma(k/2) c_k}$$

and its cdf is $F_k(t)$.

From each $f(x)$, in turn, we may generate a multivariate location-scale family, $f(y; \mu, A)$, where $f(y; \mu, A)$ is the density of

$$Y = \mu + A^{1/2} X,$$

and A is symmetric and positive definite. It is reasonable to refer to these distributions as ellipsoidal distributions. Incidentally, while the expectation of Y is μ, the covariance of Y is not, in general, A. Since the covariance matrix for X is just $(2\pi c_k)^{-1} c_{k+2} I_k$ (a fact which follows from equation (A1.7) below), the covariance of Y is $\Sigma = (2\pi c_k)^{-1} c_{k+2} A$.

If $g(t) = e^{-t/2}$, then the preceding construction leads to the multivariate normal family of distributions, with $c_k = (2\pi)^{k/2}$. Another, more general, form for g that is a valuable source of examples is

$$g(t) = \exp(-rt^s). \tag{A1.2}$$

In this case, $c_k = \pi^{k/2} r^{-k/2s} s^{-1} \Gamma(k/2)^{-1} \Gamma(k/2s)$.

The bias reducing function introduced in Chapter 3, $b_k(a) = \dfrac{c_k^{-1} g(a^2)}{F_k(a^2)/V_k(a)}$, can be easily reexpressed in terms of f_k and F_k as

$$b_k(a) = \frac{2a^2 f_k(a^2)}{k F_k(a^2)} \tag{A1.3}$$

or as

$$(A1.4) \qquad b_k(a) = [(2\pi c_k)^{-1} c_{k+2}] \frac{2f_{k+2}(a^2)}{F_k(a^2)}$$

by making use of the formula for the volume of a k dimensional sphere of radius a,

$$(A1.5) \qquad V_k(a) = \frac{\pi^{k/2}}{\Gamma((k/2)+1)} a^k$$

given, for instance, in Apostol [1, p. 411]. It will be useful to record, in addition, that the surface area of such a sphere is

$$(A1.6) \qquad A_k(a) = \frac{2\pi^{k/2}}{\Gamma(k/2)} a^{k-1}.$$

The next lemma expresses a variety of integrals in terms of the generalized χ^2 density and distribution function. We will set $S = S_a(0)$, the ball of radius a centered at the origin.

Lemma A1.1:

$$(A1.7) \qquad \int_S (x'x) c_k^{-1} g(x'x)\,dx = kM_1$$

$$(A1.8) \qquad \int_S x_1^2 c_k^{-1} g(x'x)\,dx = M_1$$

$$(A1.9) \qquad \int_S (x'x)^2 c_k^{-1} g(x'x)\,dx = k(k+2)M_2$$

$$(A1.10)\quad \int_S x_1^4 c_k^{-1} g(x'x)\,dx = 3M_2$$

$$(A1.11)\quad \int_S x_1^2 x_2^2 c_k^{-1} g(x'x)\,dx = M_2$$

$$(A1.12)\quad \int_S (x'x) c_k^{-1} g'(x'x)\,dx = kM_3$$

$$(A1.13)\quad \int_S x_1^2 c_k^{-1} g'(x'x)\,dx = M_3$$

$$(A1.14)\quad \int_S (x'x)^2 c_k^{-1} g'(x'x)\,dx = k(k+2)M_4$$

$$(A1.15)\quad \int_S x_1^4 c_k^{-1} g'(x'x)\,dx = 3M_4$$

$$(A1.16)\quad \int_S x_1^2 x_2^2 c_k^{-1} g'(x'x)\,dx = M_4$$

where

$$(A1.17)\quad M_1 = (2\pi c_k)^{-1} c_{k+2} F_{k+2}(a^2)$$

$$(A1.18)\quad M_2 = (2\pi)^{-2} c_k^{-1} c_{k+4} F_{k+4}(a^2)$$

$$(A1.19)\quad M_3 = (2\pi c_k)^{-1} c_{k+2} f_{k+2}(a^2) - F_k(a^2)/2$$

$$(A1.20)\quad M_4 = (2\pi)^{-2} c_k^{-1} c_{k+4} f_{k+4}(a^2) - (2\pi c_k)^{-1} c_{k+2} F_{k+2}(a^2)/2.$$

Proof: We may demonstrate (A1.7) by putting $r^2 = x'x$, integrating with respect to $A_k(r)dr$, and then setting $t = r^2$. Of course (A1.8) is a trivial consequence of (A1.7). To derive (A1.9), use the same substitutions as for (A1.7). To obtain (A1.10), let $r = (x'x)^{1/2}$, and $y_i = x_i$ for $1 \leq i \leq k-1$; then the integral may be written as

$$2\int_0^a dr\, c_k^{-1} g(r^2) r \int_{\sum_{i<k} y_i^2 \leq r^2} (r^2 - \sum_{i<k} y_i^2)^{-1/2} y_1^4 dy.$$

But it is easy to show that

$$\int_{\sum_{1<i<k} y_i^2 \leq r^2 - y_1^2} (r^2 - y_1^2 - \sum_{1<i<k} y_i^2)^{-1/2} dy_2 \cdots dy_{k-1}$$

$$= (r^2 - y_1^2)^{(k-3)/2} \frac{\pi^{(k-1)/2}}{\Gamma((k-1)/2)}$$

from which it follows that

$$\int_{\sum_{i<k} y_i^2 \leq r^2} (r^2 - \sum_{i<k} y_i^2)^{-1/2} y_1^4 dy = \frac{3\pi^{k/2}}{k(k+2)\Gamma(k/2)} r^{k+2}$$

Then, finally, the integral in (A1.10) is straightforward

to obtain. Since $(x'x)^2$ is the sum of k terms of the form x_i^4, and $k(k-1)$ terms of the form $x_i^2 x_j^2$ $(i \neq j)$, by symmetry we conclude that

$$\int_S (x'x)^2 f(x)dx = k \int_S x_1^4 f(x)dx + k(k-1) \int_S x_1^2 x_2^2 f(x)dx$$

which implies (A1.11). The five integrals, (A1.12) - (A1.16), are entirely analogous and are obtained in a similar way, the only difference being that an integration by parts is necessary to get rid of the g' (hence, the presence of 2 terms in the expressions for M_3 and M_4). ∎

It follows from equation (A1.8) that the covariance of the truncated spherically symmetric distribution is just

$$\text{(A1.21)} \quad \text{Cov}(X | X'X \leq a^2) = (2\pi c_k)^{-1} c_{k+2} \frac{F_{k+2}(a^2)}{F_k(a^2)} I_k.$$

Similarly, the covariance of the truncated ellipsoidal distribution is

$$\text{(A1.22)} \quad \text{Cov}(Y | (Y-\mu)'A^{-1}(Y-\mu) \leq a^2) = (2\pi c_k)^{-1} c_{k+2} \frac{F_{k+2}(a^2)}{F_k(a^2)} A.$$

Recalling that the covariance of Y is $\Sigma = (2\pi c_k)^{-1} c_{k+2} A$, and defining

$$\text{(A1.23)} \quad c(k,p) = \frac{F_{k+2}(F_k^{-1}(p))}{p},$$

where $p = F_k(a^2)$, we may finally write the covariance of the truncated distribution as simply $c(k,p)\Sigma$.

Appendix 2

Multivariate Normal Distribution

Our two intentions in this appendix are to specialize the results of appendix 1 to the normal distribution and to give some detailed analytical and numerical information about the function $b_k(a)$ in this case.

First note that since $c_k = (2\pi)^{k/2}$ when $X \sim N(0,I_k)$, as we shall always assume in this appendix, it follows that $(2\pi c_k)^{-1}c_{k+2} = 1$ and $(2\pi)^{-2}c_k^{-1}c_{k+4} = 1$, which results in several simplifications in the formulas of appendix 1. For example, (A1.4) becomes

$$b_k(a) = \frac{2f_{k+2}(a^2)}{F_k(a^2)}. \tag{A2.1}$$

A well known property of the chi-squared distribution is that

$$2f_{k+2}(t) = F_k(t) - F_{k+2}(t), \tag{A2.2}$$

a fact which may be verified by differentiating both sides of the equation (and which is not true in the general spherically symmetric case). By making use of (A2.2), we find that equations (A1.17) - (A1.20) may be greatly simplified. Now,

(A2.3) $$M_1 = F_{k+2}(a^2)$$

(A2.4) $$M_2 = F_{k+4}(a^2)$$

(A2.5) $$M_3 = -\frac{1}{2}F_{k+2}(a^2)$$

(A2.6) $$M_4 = -\frac{1}{2}F_{k+4}(a^2).$$

Actually, (A2.5) and (A2.6) follow immediately from (A2.3) and (A2.4), since $g'(t) = -\frac{1}{2}g(t)$ for the normal case. Another important fact is that as a result of (A2.1) and (A2.2),

(A2.7) $$b_k(a) = 1 - \frac{F_{k+2}(a^2)}{F_k(a^2)}.$$

But as a consequence of (A2.7) we find that the covariance of the truncated multivariate normal (the specialization of (A1.22)) is:

(A2.8) $$\mathrm{Cov}(Y \mid (Y-\mu)'\Sigma^{-1}(Y-\mu) \leq a^2) = (1-b_k(a))\Sigma.$$

Another way of saying the same thing is that for a normal distribution, by (A1.23),

(A2.9) $$c(k,p) = 1 - b_k((F_k^{-1}(p))^{1/2}).$$

We collect a few facts about b_k in the next lemma.

Lemma A2.1: The function $b_k(a)$ is strictly decreasing on $[0,\infty)$. As a goes from 0 to ∞, b_k decreases from 1 to 0. Furthermore,

(A2.10) $$b_k'(a) = (k/a)b_k(a)(1 - a^2/k - b_k(a))$$

and

(A2.11) $$b_k(a) = 1 - a^2/(k+2) + o(a^2) \quad \text{as} \quad a \to 0.$$

Proof: Let $r_k = [2^{k/2}\Gamma(k/2)]^{-1}$. Then,

$$f_k(a^2) = r_k a^{k-2}(1 - a^2/2 + o(a^2))$$

and

(A2.12) $$F_k(a^2) = 2k^{-1}r_k a^k(1 - (k+2)^{-1}ka^2/2 + o(a^2)).$$

Therefore, using (A1.3),

$$b_k(a) = (1 - a^2/2)(1 + (k+2)^{-1}ka^2/2) + o(a^2)$$

which implies (A2.11), which in turn implies that $b_k \to 1$ as $a \to 0$. To prove that b_k is decreasing we shall show that

(A2.13) $\qquad b_k(a) > 1 - a^2/(k+2) \quad \forall\, a > 0,$

which, in conjunction with (A2.10) and the fact that $b_k > 0$, implies the result. Of course (A2.13) is trivially true when $a^2 \geq k+2$. So assume $a^2 < k+2$, let $t = a^2$, and call

$$d(t) = \frac{2tf_k(t)}{kF_k(t)} - (1 - t/(k+2)).$$

Then it will suffice to show $d(t) > 0$ for $0 < t < k+2$. Let $d_0(t) = d(t)F_k(t)/(1 - t/(k+2))$. Since d_0 has the same sign as d, it will be enough to prove that $d_0(t) > 0$, or since $d_0(0) = 0$, that $d_0'(t) > 0$ for $0 < t < k+2$. But

$$d_0'(t) = \frac{2f_k(t)}{k(1-t/(k+2))^2}\left(\frac{t}{k+2}\right)^2 > 0$$

for $0 < t < k+2$. ∎

Our next result describes the asymptotic behavior of b_k.

Lemma A2.2:

$$b_k((F_k^{-1}(p))^{1/2}) \sim (2/k)^{1/2} \frac{\phi(\Phi^{-1}(p))}{p} \quad \text{as } k \to \infty \tag{A2.14}$$

$$b_k((F_k^{-1}(p))^{1/2}) = 1 - \frac{\Gamma((k/2)+1)^{2/k}}{(k/2)+1} p^{2/k} + o(p^{2/k}) \quad \text{as } p \to 0 \tag{A2.15}$$

Proof: $\sqrt{2k}\, f_k(k + a\sqrt{2k}) \to \phi(a)$ as $k \to \infty$

and

$$\frac{F_k^{-1}(p)-k}{\sqrt{2k}} \to \Phi^{-1}(p) \quad \text{as} \quad k \to \infty.$$

Hence, $\sqrt{2k}\, f_k(F_k^{-1}(p)) \to \phi(\Phi^{-1}(p))$ as $k \to \infty$.

But then

$$(k/2)^{1/2} b_k((F_k^{-1}(p))^{1/2}) = (2/k)^{1/2} F_k^{-1}(p) f_k(F_k^{-1}(p))/p$$

$$= (2/k)^{1/2} \left(\frac{F_k^{-1}(p)-k}{\sqrt{2k}} + (k/2)^{1/2}\right) \frac{\sqrt{2k}\, f_k(F_k^{-1}(p))}{p}$$

$$\to \frac{\phi(\Phi^{-1}(p))}{p} \quad \text{as} \quad k \to \infty$$

which implies (A2.14).

By (A2.12),

$$p = F_k(a^2) = (2/k)\frac{a^k}{2^{k/2}\Gamma(k/2)} + o(a^k) \quad \text{as} \quad a \to 0.$$

Hence

$$a^2 = [2^{k/2}\Gamma((k/2)+1)p]^{2/k} + o(p^{2/k}) \quad \text{as} \quad p \to 0.$$

Using (A2.11) we conclude that (A2.15) holds. ■

We end this appendix by presenting some numerical values of b_k in Table A2.1.

Appendix 3 - Proof of Lemma 5.2

We prove Lemma 5.2.

Lemma 5.2: The function $s(a) = ka^{-2}(1 - b_k(a))$ is monotone decreasing for $a > 0$. As $a \to 0$, $s(a) \to k/(k+2)$ and as $a \to \infty$, $s(a) \to 0$.

Proof: By (A2.11) in Lemma A2.1, $b_k(a) = 1 - a^2/(k+2) + o(a^2)$. Therefore, $s(a) = k/(k+2) + o(1)$ as $a \to 0$. Since b_k is bounded, $s(a) \to 0$ as $a \to \infty$. After some algebra we find that

$$s'(a) = -ka^{-3}[2 + (k-2-a^2)b_k(a) - kb_k^2(a)].$$

Let

$$Q(x) = x^2 - (\frac{k-2-a^2}{k})x - 2/k.$$

Then, $s'(a) < 0$ iff $Q(b_k(a)) < 0$. The quadratic equation $Q(x) = 0$ always has both a positive and a negative solution; we define $h_k(a)$ to be the positive solution:

$$h_k(a) = [(\frac{a^2+2-k}{2k})^2 + 2/k]^{1/2} - \frac{a^2+2-k}{2k}.$$

Since as $x \to \pm \infty$, $Q(x) \to +\infty$, we may conclude that

$Q(b_k(a)) < 0$ iff $b_k(a) < h_k(a)$, since $b_k(a) > 0 \quad \forall a \geq 0$. Note that $h_k(0) = 1$ and, therefore, $h_k(0) - b_k(0) = 0$. We will show that $g_k(a^2) = (h_k(a) - b_k(a))F_k(a^2)/h_k(a)$ is positive for all $a > 0$ by showing that its derivative is always positive. Let $s = a^2$, $q = (s+2-k)/2k$, and $r = (q^2 + 2/k)^{1/2}$. Then, after a tedious computation we find that

$$g_k'(s) = \frac{f_k(s)}{r(r-q)^2}\left[\frac{(k+2)}{k^2}r - \frac{((k+2)^2 - s(k-2))}{2k^3}\right].$$

It will suffice to show that

$$\frac{k+2}{k^2}r > \left|\frac{(k+2)^2 - s(k-2)}{2k^3}\right|.$$

But upon squaring and expanding both sides of this inequality, and after more tedious algebraic manipulations, we find that it is equivalent to $(k+2)^2 > (k-2)^2$, which is, of course, true for $k > 0$. ■

References

[1] Apostol, Tom M. (1969). Calculus. Volume 2. Blaisdell Publishing Company, Waltham, Massachusetts.

[2] Bahadur, R. R. (1966). "A note on quantiles in large samples." Ann. Math. Stat. 37, 577-580.

[3] Chernoff, Herman (1970). "Metric considerations in cluster analysis." Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley and Los Angeles.

[4] Cohen, A. C., Jr. (1950). "Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples." Ann. Math. Stat. 21, 557-69.

[5] Devlin, S. J., Gnanadesikan R., and Kettenring, J. R. (1975). "Robust estimation and outlier detection with correlation coefficients." Biometrika. 62, 531-45.

[6] Duda, Richard, and Hart, Peter (1973). Pattern Classification and Scene Analysis. Wiley, New York.

[7] Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. Wiley, New York.

[8] Gnanadesikan, R. and Kettenring, J. R. (1972). "Robust estimates, residuals, and outlier detection with multiresponse data." Biometrics. 28, 81-124.

[9] Hartigan, J. A. (1975). Clustering Algorithms. Wiley, New York.

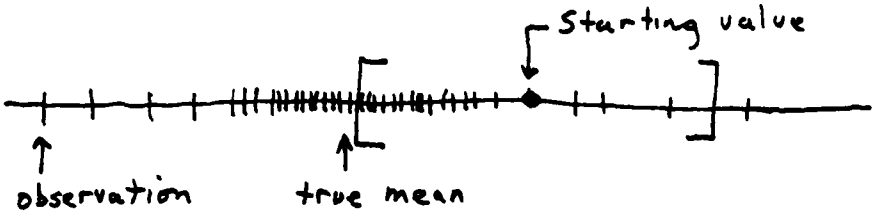
[10] Maronna, Ricardo and Jacovkis, Pablo M. (1974). "Multivariate clustering procedures with variable metrics." Biometrics. 30, 499-505.

[11] Rohlf, F. J. (1970). "Adaptive hierarchical clustering schemes." Syst. Zool. 19, 58-83.

[12] Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts.

[13] Woodroofe, Michael (1975). Probability with Applications. McGraw-Hill, New York.

Figure 2.1



Note: This figure represents one iteration of IET in one dimension.

Figure 2.2

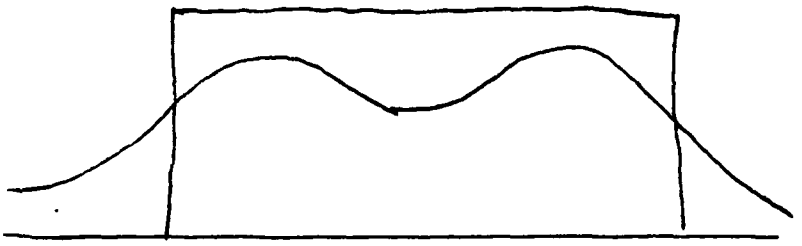
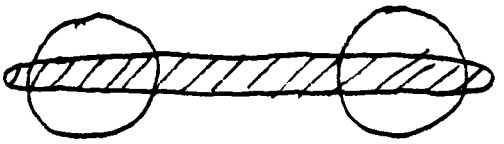


Figure 2.3



Note: The stopping "window" of IET may contain two clusters if p is too big. Figures 2.2 and 2.3 illustrate this phenomenon in 1 and 2 dimensions, respectively.

Table 2.1 - Example 1

$\tilde{\mu}_1$	$\tilde{\mu}_2$	$\lvert L_1 \rvert$	$\lvert L_2 \rvert$
-2	2	8	18
-2.17	2.98	11	15
-1.51	3.52	14	12
-1.05	4.25	18	8
-0.53	5.72	20	6
-0.23	6.80	23	3
0.17	10.75	25	1
0.45	25	25	1

Note: The successive estimates are produced by k-means 1 operating on 8 N(-2, 1) and 17 N(2, 1) observations and one outlier at x = 25.

Table 2.2 - Example 2

$\tilde{\mu}_1$	$\tilde{\sigma}_1^2$	$\tilde{\mu}_2$	$\tilde{\sigma}_2^2$	$\lvert L_1 \rvert$	$\lvert L_2 \rvert$
-2	1	2	1	8	18
-2.18	0.59	2.98	29.8	6	20
-2.21	0.15	2.47	29.3	3	23
-2.41	0.013	1.89	27.8	2	24
-2.49	10^{-4}	1.72	27.3	0	26

Note: The successive estimates are produced by k-means 2 operating on the same sample as was used in Example 1.

Table 3.1

p	$\Phi^{-1}(1-p)$	$p^{-1}\phi(\Phi^{-1}(1-p))$
0.8	-0.84	0.35
0.5	0	0.80
0.2	0.84	1.40
0.1	1.28	1.76
0.001	3.09	3.37

Table 4.1 - Values of m_k

k	m_k	$1-m_k^2$
1	0.798	0.363
2	0.886	0.215
3	0.921	0.151
4	0.940	0.116

Note: m_k is given in equation (4.1).

Table 4.2 - Simulation A

	$\bar{e}_0$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_3$	$\bar{e}_4$
$\mu_2 = 0$	0.092	0.109	0.124	0.133	0.138
	(0.004)	(0.006)	(0.007)	(0.008)	(0.008)
$\mu_2 = 1$	0.104	0.126	0.148	0.161	0.168
	(0.006)	(0.006)	(0.007)	(0.008)	(0.008)
$\mu_2 = 2$	0.154	0.146	0.155	0.165	0.170
	(0.006)	(0.007)	(0.007)	(0.008)	(0.008)
$\mu_2 = 3$	0.209	0.152	0.152	0.161	0.170
	(0.006)	(0.008)	(0.008)	(0.009)	(0.009)
$\mu_2 = 4$	0.267	0.159	0.151	0.155	0.160
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)

Note: In all runs, $k = 2$, $n_1 = 100$, $n_2 = 10$, $n_3 = 0$, $s = 100$, $p_i = 0.5$ for $1 \leq i \leq 4$, and $\mu_2 = j$ is to be interpreted as $\mu_2 = (j,0)$. Standard errors are in parenthesis. In all simulations, $\mu_1 = (0,0)$.

Table 4.3 - Simulation B

	$\bar{e}_0$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_3$	$\bar{e}_4$
$\mu_2 = 0$	0.089	0.103	0.122	0.122	0.114
	(0.004)	(0.006)	(0.006)	(0.006)	(0.006)
$\mu_2 = 1$	0.096	0.111	0.127	0.131	0.127
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
$\mu_2 = 2$	0.157	0.142	0.141	0.143	0.134
	(0.006)	(0.006)	(0.008)	(0.007)	(0.008)
$\mu_2 = 3$	0.206	0.149	0.136	0.133	0.119
	(0.007)	(0.008)	(0.008)	(0.007)	(0.006)
$\mu_2 = 4$	0.269	0.169	0.142	0.125	0.109
	(0.007)	(0.008)	(0.008)	(0.008)	(0.006)

Note: Same parameters as in simulation A except for $p_1 = 0.5$, $p_2 = 0.6$, $p_3 = 0.7$, $p_4 = 0.8$.

Table 4.4 - Simulation C

	$\bar{e}_0$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_3$	$\bar{e}_4$
$\mu_2 = 0$	0.092	0.100	0.105	0.107	0.107
	(0.004)	(0.005)	(0.006)	(0.006)	(0.006)
$\mu_2 = 1$	0.107	0.114	0.124	0.128	0.129
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
$\mu_2 = 2$	0.161	0.136	0.136	0.135	0.135
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
$\mu_2 = 3$	0.199	0.116	0.112	0.112	0.111
	(0.006)	(0 .006)	(0.006)	(0.006)	(0.006)
$\mu_2 = 4$	0.273	0.117	0.120	0.120	0.120
	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)
$\mu_2 = 5$	0.329	0.103	0.107	0.111	0.112
	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)

Note: Same parameters as in simulation A except for $p_i = 0.8$ for $1 \leq i \leq 4$.

Table 4.5 - Simulation D

Run					
1.	(3, 100, 0.5, 0.5, 0.5, 0.5)				
	0.302	0.195	0.175	0.171	0.169
	(0.015)	(0.010)	(0.009)	(0.008)	(0.008)
2.	(3, 100, 0.5, 0.6, 0.7, 0.8)				
	0.280	0.186	0.143	0.121	0.106
	(0.013)	(0.010)	(0.008)	(0.007)	(0.007)
3.	(3, 100, 0.8, 0.8, 0.8, 0.8)				
	0.310	0.126	0.111	0.110	0.112
	(0.015)	(0.006)	(0.006)	(0.006)	(0.006)
4.	(4, 100, 0.5, 0.6, 0.7, 0.8)				
	0.328	0.202	0.145	0.111	0.096
	(0.016)	(0.010)	(0.007)	(0.006)	(0.005)
5.	(4, 400, 0.5, 0.6, 0.7, 0.8)				
	0.533	0.270	0.170	0.126	0.099
	(0.026)	(0.012)	(0.008)	(0.006)	(0.005)
6.	(4, 400, 0.8, 0.8, 0.8, 0.8)				
	0.511	0.138	0.109	0.106	0.105
	(0.027)	(0.008)	(0.006)	(0.006)	(0.006)
7.	(6, 400, 0.5, 0.6, 0.7, 0.8)				
	0.558	0.265	0.171	0.131	0.102
	(0.029)	(0.012)	(0.008)	(0.006)	(0.005)

Note: The format is

$(\mu_2, \sigma^2, p_1, p_2, p_3, p_4)$

$\bar{e}_0$ $\bar{e}_1$ $\bar{e}_2$ $\bar{e}_3$ $\bar{e}_4$

$(\hat{\sigma}_{\bar{e}_0})$ $(\hat{\sigma}_{\bar{e}_1})$ $(\hat{\sigma}_{\bar{e}_2})$ $(\hat{\sigma}_{\bar{e}_3})$ $(\hat{\sigma}_{\bar{e}_4})$.

For all runs, $n_1 = 100$, $n_2 = n_3 = 10$, $s = 100$, $k = 2$.

Table 4.6 - Simulation E

Run					
1.	(3, 10, 0)				
	0.167	0.139	0.139	0.131	0.121
	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)
2.	(6, 10, 0)				
	0.279	0.129	0.132	0.124	0.115
	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)
3.	(7, 10, 10)				
	0.596	0.193	0.137	0.118	0.100
	(0.020)	(0.008)	(0.005)	(0.004)	(0.004)
4.	(7, 10, 20)				
	0.707	0.219	0.136	0.103	0.147
	(0.024)	(0.006)	(0.005)	(0.004)	(0.004)

Note: The format is the same as in simulation D except that the parameter vector is (μ_2, n_2, n_3). For all runs $n_1 = 100$, $\sigma^2 = 400$, $p_1 = 0.5$, $p_2 = 0.6$, $p_3 = 0.7$, $p_4 = 0.8$, and $k = 4$.

Figure 5.1

Distribution of Mahalanobis Distances

within a Cluster

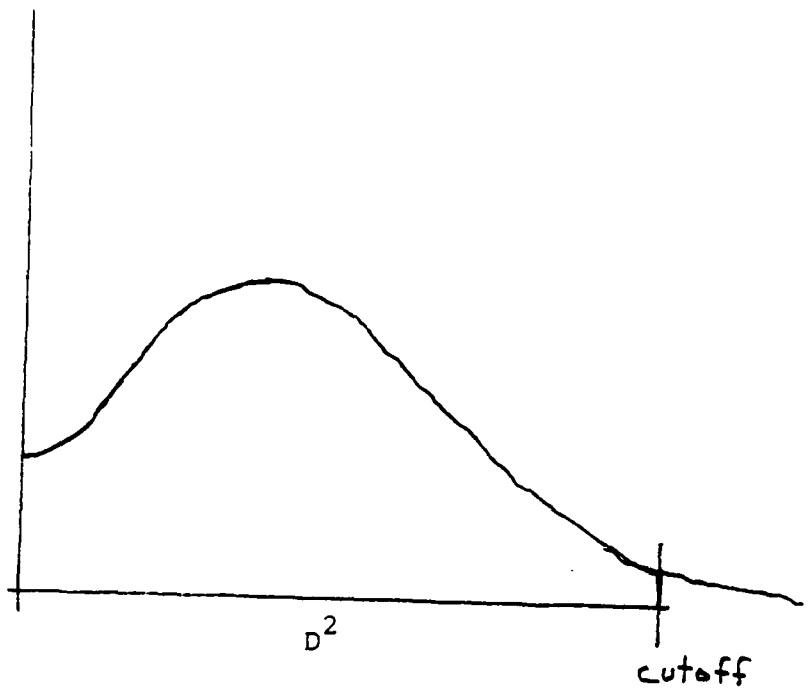


Table 5.1

Values for c(k,p) in the normal case

k/p	0.025	0.50	0.99
1	0.033	0.143	0.925
2	0.137	0.307	0.953
7	0.425	0.590	0.980

Table 5.2

Numerical values of $\tilde{v}_3$ when k = 1

s^2/t^2	$\tilde{v}_3/s^2$
0.081	1.006
0.108	1.03
0.146	1.10
0.193	1.29
0.214	1.42
0.235	1.66
0.255	2.00
0.274	2.53
0.290	3.44
0.324	12.40

Table 5.3 - Efficiencies of Scale Estimators

p	k	E_1	E_2	E_3	E_4
.1	1	0.012	0.055	0.000	0.055
	2	0.076	0.100	0.000	0.100
	3	0.098	0.131	0.000	0.132
	4	0.113	0.153	0.001	0.154
	5	0.125	0.170	0.001	0.171
	6	0.135	0.183	0.002	0.185
	7	0.143	0.194	0.002	0.196
.2	1	0.101	0.120	0.000	0.120
	2	0.155	0.199	0.001	0.200
	3	0.191	0.246	0.002	0.249
	4	0.216	0.278	0.004	0.282
	5	0.234	0.300	0.006	0.305
	6	0.248	0.316	0.007	0.324
	7	0.260	0.329	0.009	0.338
.3	1	0.158	0.194	0.000	0.194
	2	0.238	0.297	0.003	0.300
	3	0.283	0.351	0.007	0.358
	4	0.314	0.385	0.011	0.396
	5	0.335	0.407	0.015	0.422
	6	0.352	0.424	0.018	0.442
	7	0.366	0.437	0.021	0.458
.4	1	0.228	0.277	0.001	0.278
	2	0.324	0.391	0.009	0.400
	3	0.376	0.445	0.017	0.462
	4	0.409	0.477	0.024	0.501
	5	0.433	0.497	0.030	0.527
	6	0.450	0.512	0.035	0.547
	7	0.465	0.523	0.040	0.562
.5	1	0.307	0.368	0.004	0.372
	2	0.414	0.480	0.020	0.500
	3	0.469	0.528	0.034	0.561
	4	0.503	0.553	0.045	0.598
	5	0.527	0.569	0.054	0.623
	6	0.545	0.580	0.062	0.642
	7	0.559	0.588	0.068	0.656
.6	1	0.398	0.463	0.012	0.475
	2	0.510	0.560	0.040	0.600
	3	0.564	0.594	0.062	0.656
	4	0.597	0.610	0.079	0.690
	5	0.620	0.620	0.092	0.712
	6	0.637	0.625	0.103	0.728
	7	0.650	0.629	0.111	0.740

p	k	E_1	E_2	E_3	E_4
	1	0.503	0.556	0.032	0.588
	2	0.612	0.621	0.079	0.700
	3	0.662	0.637	0.111	0.748
.7	4	0.692	0.641	0.134	0.775
	5	0.712	0.642	0.152	0.794
	6	0.727	0.642	0.165	0.807
	7	0.738	0.640	0.176	0.817
	1	0.628	0.632	0.079	0.712
	2	0.723	0.648	0.152	0.800
	3	0.765	0.639	0.197	0.836
.8	4	0.789	0.630	0.226	0.856
	5	0.805	0.621	0.248	0.869
	6	0.817	0.614	0.265	0.879
	7	0.826	0.607	0.278	0.886
	1	0.781	0.640	0.207	0.847
	2	0.847	0.589	0.311	0.900
	3	0.874	0.555	0.365	0.920
.9	4	0.890	0.532	0.400	0.932
	5	0.900	0.515	0.424	0.939
	6	0.907	0.502	0.442	0.944
	7	0.913	0.491	0.457	0.948
	1	0.876	0.552	0.368	0.920
	2	0.917	0.472	0.478	0.950
	3	0.934	0.430	0.531	0.961
.95	4	0.943	0.404	0.563	0.967
	5	0.949	0.385	0.585	0.971
	6	0.953	0.372	0.602	0.973
	7	0.956	0.361	0.615	0.975
	1	0.970	0.280	0.703	0.983
	2	0.982	0.214	0.776	0.990
	3	0.986	0.185	0.807	0.992
.99	4	0.988	0.168	0.826	0.994
	5	0.990	0.157	0.838	0.995
	6	0.991	0.148	0.847	0.995
	7	0.991	0.142	0.854	0.996
	1	0.997	0.068	0.930	0.998
	2	0.998	0.048	0.951	0.999
	3	0.999	0.039	0.960	0.999
.999	4	0.999	0.035	0.965	0.999
	5	0.999	0.032	0.968	0.999
	6	0.999	0.030	0.970	1.000
	7	0.999	0.028	0.972	1.000

Table A2.1 - Numerical Values for b_k

p		k = 1	2	3	4	5	6	7
.1	a	0.126	0.459	0.764	1.031	1.269	1.485	1.683
	$b_k(a)$	0.995	0.948	0.887	0.831	0.782	0.741	0.705
.5	a	0.674	1.177	1.538	1.832	2.086	2.313	2.519
	$b_k(a)$	0.857	0.693	0.593	0.526	0.477	0.440	0.410
.8	a	1.282	1.794	2.154	2.447	2.700	2.925	3.131
	$b_k(a)$	0.562	0.402	0.326	0.281	0.249	0.226	0.208
.95	a	1.960	2.448	2.795	3.080	3.327	3.548	3.751
	$b_k(a)$	0.241	0.158	0.123	0.103	0.090	0.081	0.074
.99	a	2.576	3.035	3.368	3.644	3.884	4.100	4.298
	$b_k(a)$	0.075	0.047	0.035	0.029	0.025	0.022	0.020

Note: $a = [F_k^{-1}(p)]^{-1/2}$

OFFICE OF NAVAL RESEARCH
STATISTICS AND PROBABILITY PROGRAM

BASIC DISTRIBUTION LIST
FOR
UNCLASSIFIED TECHNICAL REPORTS

JANUARY 1980

	Copies
Statistics and Probability Program (Code 436) Office of Naval Research Arlington, VA 22217	3
Defense Technical Information Center Cameron Station Alexandria, VA 22314	12
Office of Naval Research New York Area Office 715 Broadway - 5th Floor New York, New York 10003	1
Commanding Officer Office of Naval Research Branch Office Attn: Director for Science 666 Summer Street Boston, MA 02210	1
Commanding Officer Office of Naval Research Branch Office Attn: Director for Science 536 South Clark Street Chicago, Illinois 60605	1
Commanding Officer Office of Naval Research Branch Office Attn: Dr. Richard Lau 1030 East Green Street Pasadena, CA 91101	1
Office of Naval Research San Francisco Area Office One Hallidie Plaza - Suite 601 San Francisco, CA 94102	1
Office of Naval Research Scientific Liaison Group Attn: Scientific Director American Embassy - Tokyo APO San Francisco 96503	1
Applied Mathematics Laboratory David Taylor Naval Ship Research and Development Center Attn: Mr. G.H. Gleissner Bethesda, Maryland 20084	1
Commandant of the Marine Corps (Code AX) Attn: Dr. A.L. Slafkosky Scientific Advisor Washington, DC 20380	1
Director National Security Agency Attn: Mr. Stanly and Dr. Maar (R51) Fort Meade, MD 20755	2
Navy Library National Space Technology Laboratory Attn: Navy Librarian Bay St. Louis, MS 39522	1

Copies

U.S. Army Research Office
P.O. Box 12211
Attn: Dr. J. Chandra
Research Triangle Park, NC 27706 1

OASD (I&L), Pentagon
Attn: Mr. Charles S. Smith
Washington, DC 20301 1

ARI Field Unit-USAREUR
Attn: Library
c/o ODCSPER
HQ USAEREUR & 7th Army
APO New York 09403 1

Naval Underwater Systems Center
Attn: Dr. Derrill J. Bordelon
Code 21
Newport, Rhode Island 02840 1

Library, Code 1424
Naval Postgraduate School
Monterey, California 93940 1

Technical Information Division
Naval Research Laboratory
Washington, DC 20375 1

Dr. Barbara Bailar
Associate Director, Statistical Standards
Bureau of Census
Washington, DC 20233 1

Director
AMSAA
Attn: DRXSY-MP, H. Cohen
Aberdeen Proving Ground, MD 21005 1

Dr. Bernard Heiche
Naval Air Systems Command
NAIR 03
Jefferson Plaza No. 1
Arlington, Virginia 20360 1

Copies

B. E. Clark
RR #2, Box 647-B
Graham, North Carolina 27253 1

ATAA-SL, Library
U.S. Army TRADOC Systems Analysis Activity
Department of the Army
White Sands Missile Range, NM 88002 1

OFFICE OF NAVAL RESEARCH
STATISTICS AND PROBABILITY PROGRAM

MODELING AND ESTIMATION DISTRIBUTION LIST
FOR
UNCLASSIFIED TECHNICAL REPORTS

JANUARY 1980

	Copies
Technical Library Naval Ordnance Station Indian Head, MD 20640	1
Bureau of Naval Personnel Department of the Navy Technical Library Washington, DC 20370	1
Library Naval Ocean Systems Center San Diego, CA 92152	1
Professor Robert Serfling Department of Mathematical Sciences The Johns Hopkins University Baltimore, Maryland 21218	1
Professor Ralph A. Bradley Department of Statistics Florida State University Tallahassee, FL 32306	1
Professor G. S. Watson Department of Statistics Princeton University Princeton, NJ 08540	1
Professor P. J. Bickel Department of Statistics University of California Berkeley, CA 94720	1
Professor F. J. Anscombe Department of Statistics Yale University Box 2179 - Yale Station New Haven, Connecticut 06520	1
Professor S. S. Gupta Department of Statistics Purdue University Lafayette, Indiana 47907	1
Professor R.E. Bechhofer Department of Operations Research Cornell University Ithaca, New York 14850	1
Professor D. B. Owen Department of Statistics Southern Methodist University Dallas, Texas 75275	1
Professor Herbert Solomon Department of Statistics Stanford University Stanford, CA 94305	1
Professor P.A.W. Lewis Department of Operations Research Naval Postgraduate School Monterey, CA 93940	
Dr. D. E. Smith Desmatics, Inc. P.O. Box 618 State College, PA 16801	1

	Copies
Professor R. L. Disney Dept. of Industrial Engineering and Operations Research Virginia Polytechnic Institute and State University Blacksburg, VA 24061	1
Professor H. Chernoff Department of Mathematics Massachusetts Institute of Technology Cambridge, MA 02139	1
Professor F. A. Tillman Department of Industrial Engineering Kansas State University Manhattan, Kansas 66506	1
Professor D. P. Gaver Department of Operations Research Naval Postgraduate School Monterey, CA 93940	1
Professor D. O. Siegmund Department of Statistics Stanford University Stanford, CA 94305	1
Professor M. L. Puri Department of Mathematics Indiana University Foundation P.O. Box F Bloomington, Indiana 47401	1
Dr. M. J. Fischer Defense Communications Agency Defense Communications Engineering Center 1860 Wiehle Avenue Reston, Virginia 22090	1
Defense Logistics Studies Information Exchange Army Logistics Management Center Attn: Mr. J. Dowling Fort Lee, Virginia 23801	1
Professor Grace Wahba Department of Statistics University of Wisconsin Madison, Wisconsin 53706	1
Mr. David S. Siegel Code 210T Office of Naval Research Arlington, VA 22217	1
Reliability Analysis Center (RAC) RADC/RBRAC Attn: I. L. Krulac Data Coordinator/ Government Programs Griffiss AFB, New York 13441	1
Mr. Jim Gates Code 9211 Fleet Material Support Office U.S. Navy Supply Center Mechanicsburg, PA 17055	1
Mr. Ted Tupper Code M-311C Military Sealift Command Department of the Navy Washington, DC 20390	1
Mr. Barnard H. Bissinger Mathematical Sciences Capitol Campus Pennsylvania State University Middletown, PA 17057	1
Professor Walter L. Smith Department of Statistics University of North Carolina Chapel Hill, NC 27514	1
Professor S. E. Fienberg Department of Applied Statistics University of Minnesota St. Paul, Minnesota 55108	1
Professor Gerald L. Sievers Department of Mathematics Western Michigan University Kalamazoo, Michigan 49008	1
Professor Richard L. Dykstra Department of Statistics University of Missouri Columbia, Missouri 65201	1

Copies

Professor Franklin A. Graybill
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523 1

Professor J. S. Rustagi
Department of Statistics
Ohio State University Research Foundation
Columbus, Ohio 43212 1

Mr. F. R. Del Priori
Code 224
Operational Test and Evaluation Force (OPTEVFOR)
Norfolk, Virginia 23511 1

Professor Joseph C. Gardiner
Department of Statistics
Michigan State University
East Lansing, MI 48824 1

Professor Peter J. Huber
Department of Statistics
Harvard University
Cambridge, MA 02318 1

Dr. H. Leon Harter
Department of Mathematics
Wright State University
Dayton, Ohio 45435 1

Professor F. T. Wright
Department of Mathematics
University of Missouri
Rolla, Missouri 65401 1

Professor Tim Robertson
Department of Statistics
University of Iowa
Iowa City, Iowa 52242 1

Professor K. Ruben Gabriel
Division of Biostatistics
Box 630
University of Rochester Medical Center
Rochester, NY 14642 1

Professor J. Neyman
Department of Statistics
University of California
Berkeley, CA 94720 1 Copy

Professor William R. Schucany
Department of Statistics
Southern Methodist University
Dallas, Texas 75275 1 Copy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 15	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): ITERATIVE ELLIPSOIDAL TRIMMING		5. TYPE OF REPORT & PERIOD COVERED: Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Laurence S. Gillick		8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0555
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Mathematics Massachusetts Institute of Technology Cambridge, Massachusetts 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-042-331)
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research Statistics & Probability Program Code 436 Arlington, Virginia 22217		12. REPORT DATE: February 11, 1980 13. NUMBER OF PAGES: 126
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

asymptotics; cluster analysis; ellipsoidal trimming; scale estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See reverse side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The iterative ellipsoidal trimming algorithm is introduced as both a clustering method and an estimator of location and shape. Its power as a data analytic tool is investigated and the asymptotic distribution of its stationary point is derived. In addition, several scale estimators are proposed and studied.