ROBUST ESTIMATION; LIMIT THEOREMS AND THEIR APPLICATIONS

Brenton Ross Clarke

A thesis submitted to the
Australian National University
for the degree of
Doctor of Philosophy

July 1980

STATEMENT OF ORIGINALITY

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any University; and to the best of my knowledge and belief it does not contain any material previously published or written by any other person except where due reference is made in the text.

Signed Brenton Ross Clarke

Brenton Ross Clarke

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Professor C.R. Heathcote and Dr P. Hall, both of whom have supervised my work in the last three years.

I am also grateful for having had the opportunity to study at the Australian National University. To all the staff of the Statistics Department, S.G.S., I wish to express thanks for providing me with opportunities that made working and teaching here a satisfying and rewarding experience.

I thank Mrs H. Patrikka for her excellent typing.

Finally, I am thankful for the assistance of a Commonwealth post-graduate research award during the latter half of this work.

TABLE OF CONTENTS

Summary

This thesis is concerned with the asymptotic theory of general M-estimators and some minimal distance estimators. Particular attention is paid to uniform convergence theory which is used to prove limit theorems for statistics that are usually implicitly defined as solutions of estimating equations.

The thesis is divided into eight chapters and into three main sections. In Section A the theory of convergence is studied as a prelude to validating the use of the particular M-estimators given in Section B and C. Section B initially covers the view of robustness of Hampel (1968) but places more emphasis on the application of the notions of differentiability of functionals and on M-estimators of a general parameter that are robust against "tail" contamination. Sections A and B establish a base for a comparison of robustness and application aspects of minimal distance estimators, particularly with regard to their application to estimating mixtures of normal distributions. An important application of this is illustrated for the analysis of seismic data. This constitutes Section C.

Chapter 1 is devoted to the study of uniform convergence theorems over classes of functions and sets allowing also the possibility that the underlying probability mechanism may be from a specified family. A new Glivenko-Cantelli type theorem is proved which has applications later to weakening differentiability requirements for the convergence of loss functions used in this thesis.

For implicitly defined estimators it is important to clearly identify the estimator. By uniform convergence, asymptotic uniqueness in regions of the parameter space of solutions to estimating equations can be established. This then justifies the selection of solutions

through appropriate statistics, thus defining estimators uniquely for all samples. This comes under the discussion of existence and consistency in Chapter 2. Chapter 3 includes central limit theorems and the law of the iterated logarithm for the general M-estimator, established under various conditions, both on the loss function and on the underlying distribution. Uniform convergence plays a central role in showing the validity of approximating expansions. Results are shown for both univariate and multivariate parameters. Arguments for the univariate parameter are often simpler or require weaker conditions.

Our study of robustness is both of a theoretical and quantitative nature. Weak continuity and also Fréchet differentiability with respect to Prokhorov, Lévy and Kolmogorov distance functions are established for multivariate M-functionals under similar but necessarily stronger conditions than those required for asymptotic normality. Relationships between the conditions imposed on the class of loss functions in order to attain Fréchet differentiability and those necessary and sufficient conditions placed on classes of functions for which uniform convergence of measures hold can be shown. Much weaker conditions exist for almost sure uniform convergence and this goes part way to explaining the restrictive nature of this functional derivative approach to showing asymptotic normality.

In Chapter 5 the notion of a set of null influence is emphasized. This can be used to construct M-functionals robust (in terms of asymptotic bias and variance) against contamination in the "tails" of a distribution. This set can depend on the parameter being estimated and in this sense the resulting estimator is adaptive. Its construction is illustrated in Chapter 6 for the estimation of scale. Robustness against "tail" contamination is illustrated by numerical comparison with other

M-estimators. Particular applications are given to inference in the joint estimation of location and scale where it is important to identify the root to the M-estimating equations. Techniques justified by uniform convergence are used here. Uniform convergence also lends itself to the use of a graphical method of plotting "expectation curves". It can be used for either identifying the M-estimator from multiple solutions of the defining equations or in large samples (e.g. > 50) as a visual indication of whether the fitted model is a good approximation for the underlying mechanism. Theorems based on uniform convergence are given that show a domain of convergence (numerical analysis interpretation) for the Newton-Raphson iteration method applied to M-estimating equations for the location parameter when redescending loss functions are used.

The M-estimator theory provides a common framework whereby some minimal distance methods can be compared. Two established L_2 minimal distance estimators are shown to be general M-estimators. In particular a Cramér-Von Mises type distance estimator is shown to be qualitatively robust and have good small sample properties. Its applicability to some new mixture data from geological recordings, which clearly requires robust methods of analysis is demonstrated in Chapters 7 and 8.

SECTION A: CONVERGENCE AND LIMIT THEORY

CHAPTER 1

THE THEORY OF UNIFORM CONVERGENCE

§1.1 Basic Definitions

Terminology and results basic to later discussion are given in this section. The "observation space" is denoted by R, and it is assumed to be a separable metric space. By $\mathcal{B}$ we mean that smallest σ-field containing the class of open sets on R generated by the metric on R. It is called the Borel σ-field on R. If R is Euclidean k-space, E^k, the sets $\mathcal{B}$ are called k-dimensional Borel sets. A distribution on R is a non-negative and countably additive set function, μ, on $\mathcal{B}$, for which $\mu(R) = 1$, and it is well known that on E there corresponds a unique right continuous function F whose limits are 0 and 1 at $-\infty$, $+\infty$, defined by $F(x) = \mu\{(-\infty,x]\}$.

As usual (Ω,A,P) denotes an abstract probability space, i.e. A a σ-field of subsets of Ω, with P a probability measure on A. Ω is thought of as the sample space and elements of Ω, denoted by ω, are the outcomes. Then a sequence of random variables on Ω is defined via

$$\underset{\sim}{X}(\omega) = X_1(\omega),X_2(\omega),.....,X_n(\omega),..... \quad , \tag{1.1}$$

taking values in the infinite product space $(R^\infty,\mathcal{B}^\infty)$. The observed sample of size n is then written

$$(X_1(\omega),.....,X_n(\omega)) = \pi^{(n)} \circ \underset{\sim}{X}(\omega) \quad ,$$

while the n'th random variable is given by $X_n(\omega) = \pi_n \circ X(\omega)$. Both $\pi^{(n)}$ and π_n are then measurable maps with respect to $\mathcal{B}^\infty$. They induce

distributions $G^{(n)}$ and G_n respectively on $(R^n,\mathcal{B}^n)$ and $(R,\mathcal{B})$. Theorems concerning equivalent representations of infinite sequences of random variables and probability measures are found in Chung (1968, P.54-58).

We use the symbol $\mathcal{G}$ to denote the space of distributions on $(R,\mathcal{B})$. The sequence $\underset{\sim}{X}$ is independent and identically distributed (i.i.d.) if there exists a $G \in \mathcal{G}$ such that $G_n = G$, $n = 1,2,.....$, and for every $A^{(n)} = A_1 \times ..... \times A_n \in \mathcal{B}^n$

$$\int_{A^{(n)}} dG^{(n)} = \int_{A_1} dG \ ..... \int_{A_n} dG \quad .$$

The two modes of convergence of a sequence $\{X_n\}$ of random variables on $(\Omega,\mathcal{A},P)$ to a random variable X on that same probability space, convergence in probability and convergence with probability one, are defined in the usual way. Convergence with probability one, or almost sure convergence, apriori implies convergence in probability. But for clarity we give the following formulation that helps to relate statements about events to almost sure convergence.

DEFINITION 1.1: Sequences of statements $A_1(X_1),A_2(X_1,X_2),.....$ are said to hold for all sufficiently large n (f.a.s.l.n.) if

$$P\{\omega \mid \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n(X_1(\omega),.....,X_n(\omega))\} = 1 \ ,$$

and a sequence $\{T_n(X_1,...,X_n)\}$ of measurable maps $T_n : R^n \to M$, where M is some metric space, converges almost surely to T if and only if for every $\eta > 0$ the sequence of statements

$$d(T_n(X_1,...,X_n),\ T) < \eta \qquad n = 1,2,..... \ ,$$

d the metric on M, holds f.a.s.l.n..

For an account of this definition see Foutz and Srivastava (1979). From the definition we have an immediate lemma concerning sequences of statements occurring in conjunction

LEMMA 1.1: Let sequence of statements

$$A_1(X_1),\ A_2(X_1,X_2),\ldots\ ,\ \text{and}$$

$$B_1(X_1),\ B_2(X_1,X_2),\ldots$$

both hold f.a.s.l.n.. Then the sequence of statements $A_1(X_1)\cap B_1(X_1)$, $A_2(X_1,X_2)\cap B_2(X_1,X_2),\ldots$ hold f.a.s.l.n..

PROOF: Let $A = \bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n$, $B = \bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} B_n$. Then $\int_{A\cap B} dP(\omega) = 1$. The result follows from the identity

$$A\cap B = \left\{\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n\right\}\cap\left\{\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} B_n\right\} = \bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty}(A_n\cap B_n)\ .$$

Expectation of a real valued random variable on (Ω,A,P) is written $EX = \int_{\Omega} X(\omega)dP(\omega) = \int_{-\infty}^{+\infty} xdG(x)$, where G is the induced distribution function on the real line. The variance is denoted by $\text{var } X = E[(X-EX)^2]$. A fundamental result that is used frequently is the strong law of large numbers (S.L.L.N.), a classical expression of which can be found in Loève (1955, P.239). General extensions to the S.L.L.N. to normed linear, Fréchet, and Hilbert spaces for sequences of uncorrelated random variables are given in Padgett and Taylor (1973). Nagaev (1972) examines necessary and sufficient conditions for the S.L.L.N.. Application of the S.L.L.N. is exemplified by taking f to be any real valued measurable map with domain R, and $\underset{\sim}{X}$ an i.i.d. sequence of random variables on (Ω,A,P) taking values in R. Then since $f(X_1),f(X_2),\ldots$, forms an i.i.d. sequence of random variables on E, if $E|f(X)| < \infty$

$$\frac{f(X_1) + \ldots\ldots + f(X_n)}{n} - Ef(X) \xrightarrow{a.s.} 0 \ . \qquad (1.2)$$

This "ergodic property" is known to hold for much more general sequences $\underset{\sim}{X}$. Hannan (1970, P.202) presents the mixing condition as sufficient for ergodicity of a stationary sequence. Breiman (1968, P.105) notes that measurable transformations of a strictly stationary process are strictly stationary. So for any measurable f with $E|f(X_1)| < \infty$, and strictly stationary mixing sequence $\underset{\sim}{X}$ the ergodicity (1.2) is retained. Loève (1955, P.423) states that a stationary process $\underset{\sim}{X}$ is ergodic if and only if it is indecomposable, that is, if its invariant σ-field consists of ϕ and Ω only, up to an equivalence.

The other major convergence is that of convergence of measures on the general space $(R,\mathcal{B})$. We consider the mode of weak convergence described in Billingsley (1956). In particular we use the following characterization of weak convergence

PROPOSITION 1.1: The following statements are equivalent:

(1) $\mu_n \Rightarrow \mu$,

(2) $\lim_{n\to\infty} \mu_n(A) = \mu(A)$ for each continuity set A of μ, and

(3) for each bounded and uniformly continuous function $g(x)$ on R,

$$\lim_{n\to\infty} \int g \, d\mu_n = \int g \, d\mu \ .$$

This can be found in Alexandroff (1943) or Billingsley (1956).

The *empirical distribution function* will be the distribution (random) that assigns atomic mass $1/n$ to each point of the sample $\underset{\sim}{X}^{(n)} = (X_1,\ldots\ldots,X_n)$. We label it $F_n(x,\omega)$, and abbreviate it to $F_n(\cdot)$ for most purposes. A result of Varadarajan (1958) states that for an ergodic sequence $\underset{\sim}{X}$, with marginal distribution G,

$$P\{\omega \mid F_n(\cdot,\omega) \Rightarrow G\} = 1 \ . \qquad (1.3)$$

Therefore if we have estimators which are functions of the sample, $T_n(X_1,.....,X_n)$, that can be written as functionals $T[F_n]$, they will converge whenever the functional satisfies a continuity property in the topology of weak convergence. Invoking (1.3) and using the "deterministic" approach of weak convergence can be sufficient for showing convergence of functionals $T[F_n]$. But it is not always necessary to assume this weak continuity.

We let estimators and/or functionals take values in subsets of Euclidean r-space, E^r. For a vector $z = (z_1,...,z_r) \in E^r$, we denote the usual Euclidean norm of $\underset{\sim}{z}$ by $(z_1^2 + ... + z_r^2)^{\frac{1}{2}} = \|\underset{\sim}{z}\|$. For an arbitrary $n \times m$ matrix A we write, $\|A\| = \{\text{trace}(A'A)\}^{\frac{1}{2}}$.

§1.2 Uniform Convergence on Classes of Sets and Functions

Frequently we have a sequence of points $\{T_n\}$ that converge to T in some sense, but our interest lies in the convergence of

$$\sup_{f \in \mathcal{a}^*} |f(T_n) - f(T)| ,$$

for $\mathcal{a}^*$ a class of real valued functions on R, or more generally the almost sure limits of

$$\sup_{f \in \mathcal{a}} \left| \int f \, dF_n - \int f \, dG \right| ,$$

where G is the underlying marginal probability distribution. This convergence is of particular interest in statistical inference where P is a family of probability measures $\{P_\theta | \theta \in \Theta\}$, so that $\Theta \subset E^r$, and we wish to find consistent asymptotically normal estimates of some underlying $\theta_o \in \Theta$, when samples $\underset{\sim}{X}^{(n)}$ are generated from P_{θ_o}.

The classical theorem of uniform convergence of measures on sets is that of Glivenko and Cantelli which asserts that

$$P\{\omega | \lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x,\omega) - G(x)| = 0\} = 1 , \qquad (1.4)$$

for $\underset{\sim}{X}$ a strictly stationary ergodic sequence. The proof stems from the S.L.L.N. applied to the sequence of indicator function values $\{I_{(-\infty,x]}(X_i)\}_{i=1}^{\infty}$. Extensions of the Glivenko-Cantelli theorem to E^k space and more general sets have been carried out by Wolfowitz (1954), Ranga Rao (1962), Topsoe (1970) and Elker, Pollard, and Stute (1979). All of these authors consider the theorem for closed half spaces E^k. Rao uses Varadarajan's result (1.3), and examines classes for which weak convergence implies uniform convergence. Billingsley and Topsoe (1970) follow up this approach and investigate the necessary and sufficient condition for a class $\mathcal{a}$ of functions to be a P-uniformity class; that is a class $\mathcal{a}$ for which if $\{P_n\}$ are such that $P_n \Rightarrow P$, then

$$\lim_{n\to\infty} \sup_{f\in\mathcal{a}} \left| \int f\, dP_n - \int f\, dP \right| = 0 .$$

To describe their result let $U_\delta(x)$ be the open sphere with center x and radius δ. We adopt this common notation throughout. Then denote the oscillation of a function f on a set B by $w_f(B)$, and $W_{\mathcal{a}}(B)$ is the oscillation of the family of functions, i.e.

$$w_{\mathcal{a}}(B) = \sup\{w_f(B) \mid f \in \mathcal{a}\} .$$

We call $\partial_{\delta,\varepsilon}(f)$ the δ,ε-boundary of the function f, the set of points x in R such that $w_f(U_\delta(x)) > \varepsilon$. The necessary and sufficient condition is stated in Topsoe.

PROPOSITION 1.2 (Billingsley and Topsoe): $\mathcal{a}$ is a P-uniformity class if and only if

$$w_{\mathcal{a}}(R) < \infty ,$$

and

$$\lim_{\delta\to 0} \sup_{f\in\mathcal{a}} P(\partial_{\delta,\varepsilon}(f)) = 0 ,$$

holds for all $\varepsilon > 0$.

If $\mathcal{A}$ denotes the class of indicator functions of a class $\mathcal{U} \subset \mathcal{B}$ $\mathcal{U}$ is called a uniformity class. A uniformity class of sets with respect to a distribution G will be automatically a Glivenko Cantelli class; a class $\mathcal{D}$ for which

$$\sup_{D \in \mathcal{D}} \left| \int_D dF_n(x,\omega) - \int_D dG(x) \right| \xrightarrow{a.s.} 0 \ ;$$

but the converse need not be true (Example 2 of Elker, Pollard, and Stute).

So it is important to recognise the delineation between probability, or almost sure results, and those implied by weak convergence. The latter provides a convenient avenue for arriving at results of the former, but is not necessary. Topsoe does point out though that if one can show the class $\mathcal{U}$ to be a uniformity class with respect to the continuous part of the probability measure G, it is automatically then a Glivenko Cantelli class. Particular classes of sets that have been investigated are the convex measurable subsets on E^k.

PROPOSITION 1.3 (Fabian 1970): Let E^k be the Borel sets of E^k, C the set of all convex measurable subsets of E^k. Let $\{\mu_n\}$ be a sequence of measures converging pointwise on C to a measure μ on E^k. Suppose $\mu(\bar{C} - C^o) = 0$, for any $C \in C$ with $\bar{C}$ and C^o denoting closure and interior of C. Then $\{\mu_n\}$ converges to μ uniformly on C.

The result was proved as an if and only if result assuming $\mu_n \Rightarrow \mu$ by Rao (1962, Theorem 4.2).

Generalizations of the Glivenko Cantelli theorem to dependent sequences have been carried out by Tucker (1959). He considers the case where the stationary sequence may not be indecomposable. His result is stated

$$P\{\omega \mid \sup_{-\infty<x<\infty} |F_n(x,\omega) - G(x|T)| \xrightarrow[n\to\infty]{} 0\} = 1 \ ,$$

where $G(x|T)$ denotes the conditional distribution function of X_1 given T, the invariant σ-field of events of $\underset{\sim}{X}$. Rao (1962) considered the extension to the collection of half spaces in E^k.

Also Rao illustrates that uniform convergence over classes of functions almost surely need not require the class to be a uniformity class. To describe his result in its generality we say a sequence of random measures $\{\lambda_n(A,\omega), \lambda(A,\omega); n = 1,2,\ldots\}$ on R, possesses the "ergodic property" if for each real valued function $g(x)$ on R, for which

$$E\int |g(x)|\lambda(dx,\omega) < \infty ,$$

then

$$\lim_{n\to\infty} \int g(x)d\lambda_n = \int g(x)d\lambda \quad \text{a.e.} \tag{1.5}$$

PROPOSITION 1.4 (Rao 1962, Theorem 6.4): Let $\underset{\sim}{X}$ be a strictly stationary sequence and denote $G(dx,\omega)$ the random measure associated with the conditional expectation (as in that of Tucker). Set $\mathcal{a}$ to be an equicontinuous class of continuous functions and $g(x)$ a continuous function on R such that $|f(x)| \leq g(x)$ for each $f \in \mathcal{a}$ and $x \in R$. Suppose $E|g(X_1)|^{1+\alpha} < \infty$ for some $\alpha \geq 0$, then

(a) $P\{\lim_{n\to\infty} \eta_n = 0\} = 0$

(b) $\lim_{n\to\infty} E\eta_n^{1+\alpha} = 0$,

where

$$\eta_n = \sup_{f\in\mathcal{a}} \left| \frac{f(X_1)+\ldots+f(X_n)}{n} - \int f(x)G(dx,\omega) \right| .$$

The proof of this proposition relies on the more general expression of the S.L.L.N., in Birkhoff's ergodic theorem (c.f. Loève 1955, P.421). A result that clarifies the nature of the random measure $G(dx,\omega)$ is found in Blackwell (1956). For an indecomposable sequence

$G(dx,\omega)$ is the constant measure $dG(x)$.

Since functions f are permitted to be unbounded clearly $\mathcal{a}$ is not required to be a uniformity class with respect to any measure. To impress on the reader the action of the S.L.L.N. in this result we give the proof for an ergodic sequence $\underset{\sim}{X}$ of real valued random variables. This also serves the purpose of familiarizing us with a useful technique to be used later.

PROOF: Denote $C = (-c,c]$ and C' its complement. Since g is integrable ($g \in L_1(G)$ in symbols), given $\varepsilon > 0$ there exists $0 < c < \infty$ such that $\int_{C'} g\,dG < \varepsilon/2$. Since $\mathcal{a}$ is equicontinuous on $[-c,c]$, given arbitrary $\eta > 0$ there exists a finite partition

$$-c = a_0 < a_1 < \ldots\ldots < a_m = c\ ,$$

for which $a_{i-1} \le x < y \le a_i$ implies $|f(x) - f(y)| < \eta$ for every $f \in \mathcal{a}$, and $i = 1,\ldots,m$. Consider the possibly improper distribution G^* attributing weight $G(a_i) - G(a_{i-1})$ to the points $x_i = (a_i + a_{i-1})/2$, $i = 1,\ldots,m$. Then clearly

$$\left|\int_C f\,dG^* - \int_C f\,dG\right| \le \sum_{i=1}^{m} \int_{(a_{i-1},a_i]} |f(x_i) - f(x)|\,dG(x)$$

$$< \eta(G(c) - G(-c)) \le \eta\ . \tag{1.6}$$

Let $\alpha(c) = \sup_C |g(x)| < \infty$ and F_n^* be analogously constructed from F_n. Then

$$\sup_{f\in\mathcal{a}} \left|\int_C f\,dF_n^* - \int_C f\,dG^*\right| \le \alpha(c) \sum_{i=1}^{m} \int_{(a_{i-1},a_i]} d(F_n - G)|$$

$$< \eta \quad \text{f.a.s.l.n.}, \tag{1.7}$$

since $F_n(x) - G(x) \xrightarrow{\text{a.s.}} 0$ for every $x \in E$ by the S.L.L.N.. We deduce from (1.6) and (1.7) that

$$\sup_{f\in\mathcal{a}} \left|\int_C f\,dF_n - \int_C f\,dG\right| < 3\eta \qquad \text{f.a.s.l.n.}\ . \tag{1.8}$$

Finally

$$\sup_{f\in\mathcal{a}}\left|\int f\,dF_n - \int f\,dG\right| \le \sup_{f\in\mathcal{a}}\left|\int_{C'} f\,dF_n - \int_{C'} f\,dG\right|$$

$$+ \sup_{f\in\mathcal{a}}\left|\int_{C} f\,dF_n - \int_{C} f\,dG\right|$$

$$< \int_{C'} g\,dF_n + \int_{C'} g\,dG + \sup_{f\in\mathcal{a}}\left|\int_{C} f\,dF_n - \int_{C} f\,dG\right|$$

$$< \varepsilon + 3\eta \quad \text{f.a.s.1.n.} .$$

This is true from (1.8) and since $\int_{C'} g\,dF_n \xrightarrow{a.s.} \int_{C'} g\,dG$ by the S.L.L.N.

Only the S.L.L.N. were necessary in this proof of the proposition, not a Glivenko Cantelli result. Hence there exists an obvious corollary to independent but not identically distributed (i.n.i.d.) sequences $\underset{\sim}{X}$.

COROLLARY 1.1: Let $\mathcal{a}$ be an equicontinuous class of functions with domain E, and assume $\underset{\sim}{X}$ to be an i.n.i.d. sequence of real valued random variables with distributions $G_1, G_2, \ldots$. Further let g be a continuous function on the real line such that (a) $|f(x)| \le g(x)$ for each $f \in \mathcal{a}$ and $x \in E$; (b) for every $\varepsilon > 0$ there exists a $C = (-c,c]$, $c > 0$ such that $\int_{C'} g\,d\bar{G}_n < \varepsilon$ for $n \ge n_o(c)$; and (c) the S.L.L.N. holds for the sequence $\{g(X_i)I_{C'}(X_i)\}$. Here $\bar{G}_n = n^{-1}\Sigma_{i=1}^{n} G_i$. Then

$$P\{\lim_{n\to\infty} \eta_n = 0\} = 1 ,$$

where

$$\eta_n = \sup_{f\in\mathcal{a}}\left|\int f(x)\,d F_n(x) - \int f(x)\,d \bar{G}_n(x)\right| .$$

PROOF: From the previous proof we see that it is only necessary that the S.L.L.N. hold on the bounded sequence $\{I_{(-\infty,x]}(X_i)\}$ for all $x \in E$, for then $|F_n(x) - \bar{G}_n(x)| \xrightarrow{a.s.} 0$ for every $x \in E$. This is true by using Theorem 3.1.2 of Padgett and Taylor (1973).

It is difficult to see an extension of this method of proof to functions with domain E^k, even if easily verifiable conditions for the S.L.L.N. are satisfied, although we can resort to the weak convergence arguments. Sequences that are not identically distributed are of no particular interest for the parametric estimation investigated here and we do not pursue this line of discussion.

An important result that combines the notions of uniform convergence over both classes of functions and sets, or equivalently extends the classes of functions considered in Proposition 1.4, can be shown for univariate sequences on the real line. We consider a stationary ergodic sequence with marginal distribution G, and in notational abbreviation write $G(x^-) = \lim_{h \downarrow 0} G(x-h)$. The following are preparatory lemmas.

LEMMA 1.2: Let G be any distribution function for which $G(x) - G(x^-) < \eta/4$ for $x \in (a,b)$, where $a < b$ real, and $\eta > 0$ are given. If $G(b^-) - G(a) > \eta$, then there exists a finite partition

$$a = x_o < x_1 < \ldots\ldots < x_{k'} = b \ ,$$

so that

$$G(x_j^-) - G(x_{j-1}) < \eta \ , \quad j = 1,\ldots,k' \ .$$

PROOF: Define $G^{-1}(z) = \inf\{x | G(x) \geq z, \ z \in [a,b]\}$. Since G is right continuous $G(G^{-1}(z)) \geq z$. Choose

$$y_j = G^{-1}\{G(a) + \frac{j}{k} (G(b^-) - G(a))\} \ ,$$

where $k \geq 1$ is chosen so that

$$\frac{G(b^-) - G(a)}{\eta} < k < \frac{2(G(b^-) - G(a))}{\eta} \ .$$

Then

$$G(y_j) - G(y_{j-1}) \geq G(a) + \frac{j}{k}\,(G(b^-) - G(a)) - G(y_{j-1})$$

$$\geq G(a) + \frac{j}{k}\,(G(b^-) - G(a))$$

$$- \{G(a) + \frac{j-1}{k}\,(G(b^-) - G(a)) + \eta/4\}$$

$$= \frac{1}{k}\,(G(b^-) - G(a)) - \eta/4$$

$$\geq \eta/4 \ .$$

If $y_j \in (a,b)$, $j = 1,\ldots,k$, then $y_j > y_{j-1}$. For if $y_j = y_{j-1}$, then

$$G(y_j) - G(y_j^-) = G(y_j) - G(y_{j-1}^-) \geq G(y_j) - G(y_{j-1})$$

$$\geq \eta/4 \ .$$

But this contradicts the initial assumption. Now since

$$G(y_j^-) \leq G(a) + \frac{j}{k}\,(G(b^-) - G(a)) \ , \quad j = 1,\ldots,k \ ,$$

then

$$G(y_j^-) - G(y_{j-1}) \leq \frac{1}{k}\,(G(b^-) - G(a)) < \eta \ .$$

Note that $y_o = a$, $y_1 > a$, and if $y_k < b$, then $G(b^-) - G(y_k) = 0$.
Let $a = x_o < x_1 < \ldots\ldots < x_{k'} = b$, be formed from $\{y_j\}_{j=1}^{k} \cup \{b\}$.

LEMMA 1.3 (c.f. Proposition 1.4): Let $\underset{\sim}{X}$ be a stationary ergodic sequence of real valued random variables with marginal distribution G. Assume $\mathcal{A}$ to be a family of real valued functions on E such that (i) $\mathcal{A}$ is equicontinuous; and (ii) there exists a continuous function g such that $|f(x)| \leq g(x)$ for every $x \in E$, $f \in \mathcal{A}$. Then given any $c > 0$, setting $C = (-c,c]$

$$\sup_{x \in C} \sup_{f \in \mathcal{A}} \left| \int_{(-c,x]} f(y)dF_n(y) - \int_{(-c,x]} f(y)dG(y) \right| \xrightarrow{a.s.} 0 \ .$$

PROOF: Given $\eta > 0$, let $\{d_i\}_{i=1}^{\ell}$ be the at most finite set of points in C such that $G(d_i) - G(d_i^-) \geq \eta/(4.\alpha(c))$, where $\alpha(c) = \sup\{g(y) \,|\, y \in C\}$,

if they exist. Since the family $\mathcal{A}$ is equicontinuous and $\bar{C}$ is compact, we may choose a decomposition

$$-c = a_o < a_1 < \ldots\ldots < a_m = c$$

so that $a_{i-1} \leq x < y \leq a_i$ implies $|f(x) - f(y)| < \eta$, for every $f \in \mathcal{A}$, and $i = 1,\ldots,m$. Let $\{a_i^*\}_{i=o}^{k}$ be the further decomposition obtained be combining the points $\{a_i\}_{i=o}^{m}$ and $\{d_i\}_{i=1}^{\ell}$, so that $a_{i-1}^* < a_i^*$, $i = 1,\ldots,k$. From Lemma 1.2 if $G(a_i^{*-}) - G(a_{i-1}^*) > \eta/\alpha(c)$, then there exists a finite decomposition $\{x_{ij}\}_{j=o}^{n_i}$ so that

$$a_{i-1}^* = x_{io} < x_{i1} < \ldots\ldots < x_{in_i} = a_i^*$$

for which

$$G(x_{ij}^-) - G(x_{i(j-1)}) < \eta/\alpha(c) \ , \quad j = 1,\ldots,n_i \ . \tag{1.9}$$

If $G(a_i^{*-}) - G(a_{i-1}^*) \leq \eta/\alpha(c)$, set $n_i = 1$, $x_{io} = a_{i-1}^*$ and $x_{i1} = a_i^*$. That is, no further partitioning is necessary. Let $\{b_i\}_{i=o}^{n'}$ be the set of points that partition $(-c,c]$ formed from combining the points $\{x_{ij}\}_{j=o}^{n_i}$, $i = 1,\ldots,k$. Denote F^* the possibility improper distribution that attributes weight $G(b_i) - G(b_i^-)$ to the points b_i, and weight $G(b_i^-) - G(b_{i-1})$ to the points $p_i = \frac{1}{2}(b_i + b_{i+1})$, $i = 1,\ldots,n'$. Let $x \in C$. Then either; (1) there exists an $0 \leq i_x \leq n'-1$, such that $b_{i_x} < x < b_{i_x+1}$; or (2) there exists an $1 \leq i_x \leq n'$ for which $x = b_{i_x}$.

(1) If $b_{i_x} < x < b_{i_x+1}$ and $f \in \mathcal{A}$, then

$$\left| \int_{(-c,x]} f\, dF^* - \int_{(-c,x]} f\, dG \right| \leq \sum_{j=1}^{i_x} \int_{(b_{j-1},b_j)} |f(p_j) - f(x)|\, dG(x)$$

$$+ \left| \int_{(b_{i_x}, x]} f(x)\, d(F^*-G)(x) \right|$$

$$\leq \eta(G(c) - G(-c)) + 2\alpha(c)(G(b_{i_x+1}^-) - G(b_{i_x}))$$

$$< \eta + 2\eta = 3\eta \,. \tag{1.10}$$

(2) If $x = b_{i_x}$ for some $1 \leq i_x \leq n'$ and $f \in \mathcal{A}$, then

$$\left| \int_{(-c,x]} f\, dF^* - \int_{(-c,x]} f\, dG \right| \leq \sum_{j=1}^{i_x} \int_{(b_{j-1}, b_j)} |f(p_j) - f(y)|\, dG(y)$$

$$< \eta(G(c) - G(-c)) \leq \eta \,.$$

Hence

$$\sup_{x \in C} \sup_{f \in \mathcal{A}} \left| \int_{(-c,x]} f\, dF^* - \int_{(-c,x]} f\, dG \right| < 3\eta \,. \tag{1.11}$$

This is true for any distribution satisfying (1.9). In particular, since by the S.L.L.N.

$F_n(x_{ij}^-) - F_n(x_{i,j-1}) < \eta/\alpha(c)$, $j = 1,\ldots,n_i$, holds f.a.s.1.n., if we let F_n^* be the corresponding measure constructed from F_n, then

$$\sup_{x \in C} \sup_{f \in \mathcal{A}} \left| \int_{(-c,x]} f\, dF_n^* - \int_{(-c,x]} f\, dF_n \right| < 3\eta \,, \tag{1.12}$$

holds f.a.s.1.n.. Now consider case (2)

(2) For $x = b_{i_x}$, $1 \leq i_x \leq n'$,

$$\sup_{f \in \mathcal{A}} \left| \int_{(-c,x]} f(y)\, dF_n^*(y) - \int_{(-c,x]} f(y)\, dF^*(y) \right|$$

$$\leq \alpha(c) \left\{ \sum_{j=1}^{i_x} \left| \int_{(b_{j-1}, b_j)} d(F_n - G) \right| + \sum_{j=1}^{i_x} \left| \int_{\{b_j\}} d(F_n - G) \right| \right\}$$

$$\leq \alpha(c) \left\{ \sum_{j=1}^{n'} \left| \int_{(b_{i-1}, b_i)} d(F_n - G) \right| + \sum_{j=1}^{n'} \left| \int_{\{b_j\}} d(F_n - G) \right| \right\} \tag{*}$$

$$< \eta \quad \text{f.a.s.1.n.} \quad \text{by the S.L.L.N.} \tag{1.13}$$

For case (1) we can split it further into two further cases

(1)(a) For $b_{i_x} < x < p_{i_x}$ for some $0 \le i_x \le n'-1$

$$\sup_{f \in \mathcal{a}} \left| \int_{(-c,x]} f(y)dF_n^*(y) - \int_{(-c,x]} f(y)dF^*(y) \right|$$

$$\le (*) + \sup_{f \in \mathcal{a}} \left| \int_{(b_{i_x},x]} f\, d(F_n^* - F^*) \right|$$

$$= (*) < \eta \quad \text{f.a.s.l.n.} .$$

(1)(b) If $p_{i_x} \le x \le b_{i_x+1} \qquad 0 \le i_x \le n'-1$,

$$\sup_{f \in \mathcal{a}} \left| \int_{(-c,x]} f(y)dF_n^*(y) - \int_{(-c,x]} f(y)dF^*(y) \right|$$

$$\le (*) + \alpha(c) \left| F_n(b_{i_x+1}^-) - F_n(b_{i_x}) - G(b_{i_x+1}^-) - G(b_{i_x}) \right|$$

$$< 2\eta \quad \text{f.a.s.l.n. by the S.L.L.N. and (1.9)..}$$

So for both cases (1) and (2)

$$\sup_{x \in C} \sup_{f \in \mathcal{a}} \left| \int_{(-c,x]} f\, dF_n^* - \int_{(-c,x]} f\, dF^* \right| < 2\eta \quad \text{f.a.s.l.n.} \quad (1.14)$$

Then from (1.10) and (1.12) this shows that

$$\sup_{x \in C} \sup_{f \in \mathcal{a}} \left| \int_{(-c,x]} f\, dF_n - \int_{(-c,x]} f\, dG \right| < 8\eta \quad \text{f.a.s.l.n.}, \quad (1.15)$$

proving the lemma since $\eta > 0$ is arbitrary.

THEOREM 1.1:

Let $\underset{\sim}{X}$ be a stationary ergodic sequence of real valued random variables with marginal distribution G. Assume $g \in L_1(G)$ is a continuous function that bounds the equicontinuous family of real valued functions $f \in \mathcal{a}$. Then

$$\sup_{x \in E \cup \{+\infty\}} \sup_{f \in \mathcal{a}} \left| \int_{(-\infty,x]} f\, dF_n - \int_{(-\infty,x]} f\, dG \right| \xrightarrow{\text{a.s.}} 0 .$$

PROOF: Let $\varepsilon > 0$ be arbitrary and choose $c > 0$ so that $\int_{C'} g\,dG < \varepsilon/2$. Set

$$H_n(\mathcal{A},G,x) = \sup_{f\in\mathcal{A}}\left|\int_{(-\infty,x]} f\,dF_n - \int_{(-\infty,x]} f\,dG\right| .$$

If $x \in (-\infty,-c]$, then

$$H_n(\mathcal{A},G,x) \le \int_{(-\infty,x]} g\,dF_n + \int_{(-\infty,x]} g\,dG$$

$$\le \int_{(-\infty,-c]} g\,dF_n + \int_{(-\infty,c]} g\,dG$$

$$< \varepsilon \quad \text{f.a.s.l.n. uniformly in } x \in (-\infty,-c] .$$

This is by the S.L.L.N. on $\{g(X_i)I_{(-\infty,-c]}(X_i)\}_{i=1}^{\infty}$.

If $x \in (-c,c]$

$$H_n(\mathcal{A},G,x) \le \int_{(-\infty,-c]} g\,dF_n + \int_{(-\infty,-c]} g\,dG$$

$$+ \sup_{x\in C}\sup_{f\in\mathcal{A}}\left|\int_{(-c,x]} f\,dF_n - \int_{(-c,x]} f\,dG\right|$$

$$< 3\varepsilon/2 \quad \text{f.a.s.l.n. uniformly in } x \in (-c,c].$$

This follows from Lemma 1.3 and the S.L.L.N. on $\{g(X_i)I_{(-\infty,-c]}(X_i)\}_{i=1}^{\infty}$.

If $x \in (c,\infty]$ then similarly

$$H_n(\mathcal{A},G,x) \le \int_{C'} g\,dF_n + \int_{C'} g\,dG$$

$$+ \sup_{x\in C}\sup_{f\in\mathcal{A}}\left|\int f\,dF_n - \int f\,dG\right|$$

$$< 3\varepsilon/2 \quad \text{f.a.s.l.n. uniformly in } x \in (c,\infty] .$$

Combining the three possibilities, $x \in (-\infty,-c]$, $x \in (-c,c]$, and $x \in (c,\infty]$ we get

$$\sup_{x\in E\cup\{+\infty\}} H_n(\mathcal{A},G,x) < 3\varepsilon/2 \quad \text{f.a.s.l.n.} .$$

Since $\varepsilon > 0$ is arbitrary the theorem is proved.

This generalizes both the Glivenko-Cantelli result and Proposition 1.4 for univariate stationary ergodic sequences. The classical Glivenko-Cantelli lemma is obtained by taking $\mathcal{A} = \{1\}$. The result is proved for classes that need not be P-uniformity classes. Neither the oscillation of the family $\mathcal{A}$ is required to be finite, nor is it required that

$$\lim_{\delta \to 0} P(\partial_{\delta,\varepsilon}(fI_x)) = 0 \ ,$$

since not all points x need be continuity points of the distribution G. Particular properties of the real line were utilized in Lemma 1.2, and the Theorem does not appear to have a natural extension to k-space. However restricting G to be absolutely continuous, it is possible to speculate that a proof utilizing weak convergence and Proposition 1.3 could be constructed.

§1.3 Uniform Convergence for Parameterized Functions

The results of §1.2 are instrumental in the study of parametric loss functions. Typically we deal with classes $\mathcal{A}$ of either real or vector valued functions of two variables; one taking values in the observation space, and the other in the parameter space. The equicontinuity requirement is then examined in the observation space variable. It need only be shown at each individual point of the observation space and we have some simple Lemmas that provide criteria for recognizing it.

LEMMA 1.4: Let $\mathcal{A} = \{\psi(\cdot,\theta) | \theta \in \Theta\}$ be a family of real valued functions defined on E, that are continuously differentiable in $x \in E$ for each $\theta \in \Theta$. For each $x \in E$ write

$$S_x = \{(y,\theta) | |x-y| < 1, \ \theta \in \Theta\} \ .$$

Suppose there exist constants A_x, independent of θ, such that

$$\sup_{S_x} |(\partial/\partial y)\psi(y,\theta)| < A_x < \infty .$$

Then the family $\mathcal{A}$ is equicontinuous.

PROOF: By the mean value theorem,

$$\sup_{\theta \in \Theta} |\psi(x,\theta) - \psi(y,\theta)| < A_x|x-y| .$$

That is the family $\mathcal{A}$ is equicontinuous in x.

REMARK 1.1: It is easy to see that conditions of Lemma 1.4 can be relaxed to letting $\psi(x,\theta)$ be continuous in x and piecewise continuously differentiable in x for each $\theta \in \Theta$.

LEMMA 1.5: Let $\mathcal{A} = \{\psi(\cdot,\theta) | \theta \in \Theta\}$ be a family of real valued functions defined on E^k. Assume $\psi(x,\theta)$ is twice continuously differentiable in x for each $\theta \in \Theta$. For each $x \in E^k$ write

$$S_x = \{(y,\theta) | \|x-y\| < 1 , \theta \in \Theta\} .$$

If for each $x \in E^k$, there exist constants A_x, $B_x \geq 0$ such that

$$\sup_{S_x} \|(\partial/\partial y)\psi(y,\theta)\| < A_x < \infty ,$$

and

$$\sup_{S_x} \|(\partial^2/\partial y^2)\psi(y,\theta)\| < B_x < \infty ,$$

then the family $\mathcal{A}$ is equicontinuous in x.

PROOF: Consider the usual Taylor expansion

$$\psi(x,\theta) = \psi(y,\theta) + (x-y)'(\partial/\partial y)\psi(y,\theta) + \tfrac{1}{2}(x-y)'(\partial^2/\partial z^2)\psi(z,\theta)(x-y)\Big|_{z=\xi},$$

with ξ so that $\|y-\xi\| \leq \|x-y\|$. Take Euclidean norms and observe that for the $k\times k$ matrix A and $k\times 1$ vector x

$$\|Ax\| \leq \|A\|\|x\| .$$

Then

$$\sup_{S_x} |\psi(x,\theta) - \psi(y,\theta)| \leq \|x-y\|A_x + \tfrac{1}{2}\|x-y\|^2 B_x$$

$$\leq A_x + \tfrac{1}{2}B_x$$

$$< \infty ,$$

which implies that the family $\mathcal{A}$ is equicontinuous.

Differentiability, while being convenient for establishing equicontinuity, is by no means necessary. If the parameter space is a compact subset of E^r, the result may be established in a manner similar to Graves (1946, P.20, Th.23). Since many limit theorem arguments are local in the parameter space we need only consider a compact subset $D \subset \Theta$, restricting the family $\mathcal{A}$ accordingly.

LEMMA 1.6: Let $\mathcal{A} = \{\psi(\cdot,\theta) | \theta \in D\}$ be a family of real valued functions defined on E^k, and suppose $D \subset \Theta \subset E^r$ is compact. Assume ψ is a continuous function in x and θ. Then $\mathcal{A}$ forms an equicontinuous family of functions.

PROOF: For each fixed b, continuity implies

$$\lim_{\substack{h\to 0 \\ x\to b}} \psi(x,\theta+h) = g(\theta) \quad \text{on} \quad \Theta$$

where $g(\theta)$ is finite. Taking $h = 0$ we have $\lim_{x\to b} \psi(x,\theta) = g(\theta)$ for each $\theta \in D$, and further $g(\theta)$ is continuous on D. That is $\psi(x,\theta)$ is continuous on the closed set for which $\theta \in D$ and $x = b$ and, since D is compact, is uniformly continuous on that set. That is,

$$\lim_{\substack{h\to 0 \\ x\to b}} \psi(x,\theta+h) = g(\theta) \quad \text{uniformly on} \quad D.$$

This completes the proof.

We may combine any of these results with Proposition 1.4 or Theorem 1.1 to obtain uniform convergence over the parameter space. Given that limit theorems often rely on local arguments in the parameter

space it could be expected that uniform convergence theorems exist for some functions ψ under very weak conditions. Roussas (1969) illustrated one such result in a lemma, the original version of which was by Le Cam (1953). His assumptions on ψ are adopted for current presentation.

Assumptions (Roussas)

(R1) Θ is an open subset of E^r.

(R2) The process $\underset{\sim}{X}$ is stationary and metrically transitive (indecomposable).

(R3) For each $y \in E^k$, the function $\psi(y,\theta)$ is continuous in θ.

(R4) For each $\theta \in \Theta$ there exists a neighbourhood of it $U_{\rho_o(\theta)}(\theta) = U_o(\theta)$, (which lies in Θ) and a (measurable) function $H_G(y)$ such that

$$\|\psi(y,t)\| \leq H(y), \quad E_G H(X_1) < \infty \quad \text{for all} \quad t \quad \text{in} \quad U_o(\theta) .$$

(R5) For each $\theta \in \Theta$ both $\sup\{\psi(y,t) \mid t \in C\}$ and the $\inf\{\psi(y,t) \mid t \in C\}$ are $\mathcal{B}$ measurable for all compact subsets C of Θ.

Under these conditions Roussas showed that

$$\sup\{|W_n(t)| \mid t \in \bar{U}_{\frac{1}{3}\rho_o(\theta)}(\theta)\} \xrightarrow{a.s.} 0 \quad \text{as} \quad n \to \infty ,$$

where

$$W_n(t) = \frac{1}{n} \sum_{j=1}^{n} (\psi(X_j,t) - E_G\psi(X_j,t)) .$$

Both Theorem 1.1 and this result show that the assumption of continuity in both variables may be relaxed. However, global consistency arguments often require uniform convergence over the whole parameter space and not only compacts. Thus it is preferable to use Lemmas 1.4 and 1.5 if possible.

§1.4 Uniform Convergence in the Underlying Distribution

Uniform convergence results presented up to now have been established assuming a fixed underlying distribution. These are used in construction of proofs of consistency and asymptotic normality. But a justification for inference applications and point estimation suggested by Wald (1941, 1943), was the uniform convergence of these limit theorems with respect to the underlying probability distributions. Such results often appear difficult to establish but we give one possible avenue of approach here.

Let $\mathcal{P}$ be the range of uncertainty in the underlying probability distribution. For instance $\mathcal{P}$ can represent the parametric family of probability laws. Some natural convergence criteria are:
For every $\varepsilon > 0$

$$P\{\omega \mid \|T_n - T\| > \varepsilon\} \to 0 \quad \text{uniformly in} \quad P \in \mathcal{P}, \tag{1.16}$$

and in the case of the central limit theorem (C.L.T.)

$$P\{\omega \mid \sqrt{n}(T_n - T) \leq x\} \to N(0,\sigma^2(P);x) \ , \tag{1.17}$$

uniformly in x and $P \in \mathcal{P}$. (The inequality $T \leq x$ is interpreted componentwise in Euclidean r-space and $N(0,\sigma^2(P);x)$ is the multivariate normal distribution with variance covariance matrix $\sigma^2(P)$). Parzen (1955) carried through a thorough investigation of this type of convergence in one dimension assuming $\mathcal{P}$ to be the parametric family. Results applied with proper interpretation of the symbols (see Loève 1950, P.84) to the r-dimensional case. No obvious extension is apparent for more abstract spaces.

We write $\mathcal{L}(X;x) = P\{X < x\}$, and $S_n = \sum_{i=1}^{n} X_i$, and consider a sequence $\underset{\sim}{X}$ that is i.i.d. and univariate. Assume that the X_n have common distribution, mean and variance given by F_θ, $m(\theta)$, and $\sigma^2(\theta)$ respectively. Then conditions for the uniform C.L.T. can be expressed

in terms of the mean and variance.

PROPOSITION 1.5 (Parzen P.38): Uniform C.L.T.: Suppose there exists constants K_1 and K_2 such that for all θ, $0 < K_1 \leq \sigma(\theta) \leq K_2 < \infty$. Then uniformly in θ and x

$$\mathcal{L}_\theta \left(\frac{S_n - m(\theta)}{\sqrt{n}} \; ; \; x \right) \to \Phi(x/\sigma(\theta)) \; ,$$

if and only if the variances are convergent uniformly in θ; that is,

$$\lim_{M \to \infty} \int_{|x|>M} (x-m(\theta))^2 dF_\theta(x) = 0 \quad \text{uniformly in} \quad \theta.$$

Our mode of investigation of many results is through the S.L.L.N.. Parzen reformulates the usual notion of almost sure convergence. The equivalent statement to that of

$$P\{\omega \,|\, (S_n - ES_n)/n \to 0\} = 1 \quad \text{is} \; ,$$

for every $\varepsilon > 0$ there is an N such that

$$P\{\omega \,|\, |S_n - ES_n| > \varepsilon n, \text{ for some } n > N\} < \varepsilon \; .$$

By the uniform S.L.L.N. is meant the statement: for every $\varepsilon > 0$ there is an N independent of θ such that, for every θ,

$$P_\theta\{\omega \,|\, |\frac{S_n - E_\theta S_n}{n}| > \varepsilon \quad \text{for some} \quad n > N\} < \varepsilon \; . \tag{1.18}$$

The sequence is said to converge almost surely p-uniformly in θ.

The following proposition follows as a consequence of the Kolmogorov inequality.

PROPOSITION 1.6 (Parzen, Th. 16A): If (i) $b_n \leq b_n(\theta) \leq B_n < \infty$, where $b_n \uparrow \infty$, (ii) for each θ, $\underset{\sim}{X}(\theta)$ is a sequence of independent random variables, (iii) $\Sigma_n\{\text{var}[X_n(\theta)]/b_n^2(\theta)\}$ is convergent and bounded by a constant K, uniformly in θ, then

$$\frac{1}{b_n(\theta)} \sum_{k=1}^{n} \{X_k(\theta) - EX_k(\theta)\} \xrightarrow{a.s.} 0 \quad \text{p-uniformly in } \theta.$$

It is now possible to combine the uniform convergence theorems over functions and sets with uniform convergence in the underlying probability measure.

THEOREM 1.2:

Let $\mathcal{A}$ be an equicontinuous class of continuous real valued functions on E, and $g(x)$ be a continuous function on E such that $|f(x)| \leq g(x)$ for each $f \in \mathcal{A}$ and $x \in E$. Suppose there exists some set $D \subset \Theta$ so that for every $\varepsilon > 0$ there exists a constant $c = c(\varepsilon)$ for which $C = (-c,c]$ satisfies $\int_{C'} g \, dF_{\theta_o} < \varepsilon$ for all $\theta_o \in D$, and further that $\text{var}[g(X(\theta_o))] < \infty$. Then for i.i.d. sequence $\underset{\sim}{X}(\theta_o)$ where $X_1(\theta_o)$ is distributed as F_{θ_o},

$$\sup_{f \in \mathcal{A}} \left| \int f \, dF_n^{\theta_o} - \int f \, dF_{\theta_o} \right| \xrightarrow{a.s.} 0 \,,$$

p-uniformly in $\theta_o \in D$.

OUTLINE OF PROOF: The proof follows that given for Proposition 1.4. Given $\varepsilon > 0$ we choose C so that $\int_{C'} g \, dF_{\theta_o} = \int_{C} g \, I_{C'} dF_{\theta_o} < \varepsilon$, where $I_{C'}$ is the indicator function of C'. Since

$$\text{var}_{\theta_o}[g(X(\theta_o))I_{C'}(X(\theta_o))] \leq \text{var}_{\theta_o}[g(X(\theta_o))] + 2E^2[g(X(\theta_o))]$$

$$< \infty \,,$$

it follows by Proposition 1.6 that,

$$\left| \int_{C'} g \, dF_n^{\theta_o} - \int_{C'} g \, dF_{\theta_o} \right| \xrightarrow{a.s.} 0 \quad \text{p-uniformly in } \theta_o \in D \,,$$

and hence

$$\int_{C'} g \, dF_n^{\theta_o} < \varepsilon \quad \text{holds almost surely p-uniformly in } \theta_o \in D. \quad (1.19)$$

Let $F_n^{\theta_o *}$ and $F_{\theta_o}^*$ be the improper distributions formed from $F_n^{\theta_o}$ and F_{θ_o} in the manner described in the proof of Proposition 1.4. Now

$$F_n^{\theta_o}(x) \xrightarrow{a.s.} F_{\theta_o}(x) \quad \text{p-uniformly in } \theta_o \in D$$

holds for every $x \in E$. This follows by Proposition 1.6 and considering bounded sequences $\{I_x(X_i(\theta_o)) - E[I_x(X(\theta_o))]\}$. Then the equivalent statement to that of (1.7) follows, namely

$$\sup_{f \in \mathcal{a}} \left| \int_C f \, dF_n^{\theta_o *} - \int_C f \, dF_{\theta_o}^* \right| < \eta$$

holds almost surely p-uniformly in $\theta_o \in D$. Combining this with (1.9) in the manner following (1.8), we obtain

$$\sup_{f \in \mathcal{a}} \left| \int f \, dF_n^{\theta_o} - \int f \, dF_{\theta_o} \right| < \varepsilon + 3\eta$$

almost surely p-uniformly in $\theta_o \in D$. Since ε and η are arbitrary this proves the theorem.

Similarly if the family of distributions $\{F_{\theta_o} | \theta_o \in D\}$ is such that atoms of each F_{θ_o}, if they exist, are independent of θ_o and there are at most a finite number of them in any compact set, then Theorem 1.1 may be extended in the same manner. The proof depends only on the S.L.L.N. .

A feature common to Propositions 1.5 and 1.6, and theorems of Parzen that concern sums S_n for more general sequences $\underset{\sim}{X}(\theta)$, is that they hold whenever the components of $\underset{\sim}{X}(\theta)$ are bounded in absolute value by a constant independent of θ. Let us also note that statements concerning uniform convergence in the probability laws of class P need not be restricted to a parametric class of probability laws. It is the recognition of these two facts that justifies the use of asymptotics to compare robust estimators. To assume a more general family P, so that

$\mathcal{F} = \{F \mid F$ is induced by $P \in \mathcal{P}\}$ extends $\{F_{\theta_o} \mid \theta_o \in D\}$, can severely delimit the range of $\mathcal{a}$ because of the assumption of Theorem 1.2: for every $\varepsilon > 0$ there exists some constant $c = c(\varepsilon)$ for which $\int_{C'} g\, dF < \varepsilon$ for all $F \in \mathcal{F}$. It may be necessary that there be some compact set C^* for which $g(x) = 0$ for all $x \in E - C^*$. That is all of the functions of $\mathcal{a}$ would be required to redescend to zero within the compact set. Examples where they do can be found in §6.

CHAPTER 2

MEASURABILITY, EXISTENCE, UNIQUENESS, AND CONSISTENCY

§2.1 Measurability of Implicitly Defined Estimators

We will mainly be concerned with the probability laws of maps $T_n[X_1,\ldots,X_n]$ from $(R^n,\mathcal{B}^n)$ to the parameter space $(\Theta,\mathcal{A})$ that are defined implicitly as a particular solution of equations

$$g_n(\underset{\sim}{X}^{(n)},T_n) = 0 \ , \tag{2.1}$$

given solutions exist f.a.s.l.n. . The function g_n is assumed to be Borel measurable. That means for any Borel set B of the real line $g_n^{-1}(B) \in \mathcal{B}^n \times \mathcal{A}$. Before any probability statements can be made about T_n it is first necessary to show its measurability. Rather than considering only the existence of a consistent measurable map we seek to define the maps uniquely for all samples $(X_1,\ldots,X_n)$, $n \in N$, or at least uniquely on a set of probability one for $n \geq n_o$. For instance defining the map as a solution of (2.1) would not define the map uniquely if there existed multiple solutions. The selection of the root for the map T_n can have implications for both the resulting probability law and practical application of the statistical procedure.

In the literature many estimators are given as a solution satisfying

$$f_n(\underset{\sim}{X}^{(n)},T_n[\underset{\sim}{X}^{(n)}]) = \inf_{\theta\in\Theta} f_n(\underset{\sim}{X}^{(n)},\theta) \ . \tag{2.2}$$

Note that the estimator is not uniquely defined should $\{\tau | f_n(\underset{\sim}{x}^{(n)},\tau) = \inf_{\theta\in\Theta} f_n(\underset{\sim}{x}^{(n)},\theta)\}$ consist of more than one point for some $\underset{\sim}{x}^{(n)} \in R^n$, $n \in N$. What is then established is the existence of a measurable map $T_n[\cdot]$ which will have the usual desired asymptotic properties when the $\{f_n\}$ are suitably regular. Strict convexity of

$f_n(\underset{\sim}{x}^{(n)},\theta)$ in θ resolves the difficulty in identifying the estimators uniquely, but this is not always desirable or possible.

Limit theorems for these statistics are often arrived at by investigating local arguments on the minimizing equations. Then the choice is $g_n(\underset{\sim}{x}^{(n)},\theta) = \nabla' f_n(\underset{\sim}{x}^{(n)},\theta)$ where ∇ represents the $1\times r$ operator $(\partial/\partial\theta_1,\ldots,\partial/\partial\theta_r)$. In fact it is common practice to search for the estimator (2.2) by examining the corresponding equation (2.1) for the set of solutions, which we label

$$H_n(\underset{\sim}{X}^{(n)}) = \{\theta \in \Theta \mid g_n(\underset{\sim}{X}^{(n)},\theta) = 0\} .$$

Then $f_n(\underset{\sim}{X}^{(n)},\cdot)$ is used as a "*selection statistic*" to determine the estimator so that

$$f_n(\underset{\sim}{X}^{(n)},T_n[\underset{\sim}{X}^{(n)}]) = \inf_{\theta\in H_n(\underset{\sim}{X}^{(n)})} f_n(\underset{\sim}{X}^{(n)},\theta) .$$

For this estimator to correspond to that of (2.2), assuming that is unique, it is necessary for f_n to attain its infimum value at some point in the interior of the parameter space, and the observation space must be independent of the parameter.

This approach is feasible for the construction of estimators more generally, and indeed this has been suggested for application to determine some robust estimators. For instance Hampel (1974) suggests that the M-estimator for location when a three part redescending influence function is used should be chosen as that root of the M-estimating equations

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0 ,$$

which is closest to the median. This is equivalent to choosing

$$g_n(\underset{\sim}{X}^{(n)},\theta) = \sum_{i=1}^{n} \psi(X_i - \theta), \text{ and } f_n(\underset{\sim}{X}^{(n)},\theta) = |\theta - \text{med}(X_1,\ldots,X_n)| .$$

Clearly the equations are not derived from the selection statistic.

Emphasis on the technique of using the selection statistic independent of the estimating equations is new. We consider specifically the situation where multiple solutions to the equations (2.1) exist even asymptotically, as can be the case with the redescending influence function of Tukey's biweight (c.f. §6.1). This situation motivates the examination of consistency arguments that rely chiefly on asymptotic behaviour of the roots of equations (2.1) and where the estimator is identified specifically from the multiple roots. This contrasts with the approach of Huber (1967) where asymptotically there exists a unique root to the M-functional equation but only the existence of a consistent sequence of roots is shown. Particular comparison of conditions is given at the end of the chapter. Other authors, e.g. Collins (1976) or Foutz (1977), show only the consistency of a root without identifying the root for all n.

The notion of a selection statistic or functional further allows the separation of robustness considerations into local and global arguments. Moreover limiting distributions of statistics are invariably displayed having assumed the underlying distribution, usually an F_{θ_o}. There is no loss of generality in examining first the asymptotic distribution of a measurable sequence $\{T_{n1}\}$ defined by the selection statistic $\|\theta - \theta_o\|$ and then showing that for any appropriate selection statistic $f_n(\underset{\sim}{X}^{(n)}, \theta)$ that defines a measurable sequence $\{T_{n2}\}$ it is true that the statement $T_{n1} = T_{n2}$ holds f.a.s.l.n., i.e. the limiting distributions of T_{n1} and T_{n2} are the same. The selection statistic $\|\theta - \theta_o\|$ is independent of the data and selects the root of the equations that is closest to the true parameter, while f_n is dependent on θ_o only through the sample $\underset{\sim}{X}^{(n)}$ which is supposedly generated from F_{θ_o}. A property of the selection statistic that is dependent on the sample is that the root that is selected converges almost surely to the true underlying parameter when the underlying probability law is within the

parametric family. This is often stated under the guise of Fisher consistency. It is then possible to take the efficacious approach of first constructing the equations (2.1) so that the limiting distribution of statistics determined by $\|\theta - \theta_o\|$ is only slightly perturbed on neighbourhoods of the distribution F_{θ_o}. Should the resulting equations not necessarily correspond to any minimizing equations of some distance, an appropriate selection statistic is then chosen. Even if there is a correspondence with a distance it may prove advantageous for robustness reasons to resort to an alternative selection statistic. That is, it is not necessary that $g_n(\underset{\sim}{x}^{(n)},\theta) = \nabla' f_n(\underset{\sim}{x}^{(n)},\theta)$.

We let the set of global minima amongst the solutions of (2.1) be

$$A_n(\underset{\sim}{X}^{(n)}) = \{\tau \in \Theta \mid f_n(\underset{\sim}{X}^{(n)},\tau) = \inf_{\theta \in H_n(\underset{\sim}{X}^{(n)})} f_n(\underset{\sim}{X}^{(n)},\theta)\} .$$

An ideal selection statistic would ensure $A_n(\underset{\sim}{x}^{(n)})$ is at most a single point set for all $\underset{\sim}{x}^{(n)} \in R^n$, $n \in N$. Then the estimator $T_n : R^n \to \Theta \cup \{+\infty\}$, where $\Theta \cup \{+\infty\}$ is the one point compactification of Θ described in Kelley (1967), is uniquely defined:-

$$T_n[\underset{\sim}{X}^{(n)}] = A_n(\underset{\sim}{X}^{(n)}) \quad \text{if} \quad A_n(\underset{\sim}{X}^{(n)}) \neq \phi$$

$$+\infty \quad \text{if} \quad A_n(\underset{\sim}{X}^{(n)}) = \phi .$$

But if this is not the case, then at least we would prefer that the set of $\underset{\sim}{x}^{(n)} \in R^n$ such that $A_n(\underset{\sim}{x}^{(n)})$ consists of more than one point be a null set with respect to the laws $F_\theta^{(n)}$ induced on $\underset{\sim}{X}^{(n)}$ by the probability measures P_θ, for all $\theta \in \Theta$. If for each $\theta \in \Theta$, P_θ is absolutely continuous with respect to P^+, then this is achieved if this set is null with respect to the law $G^{(n)+}$ induced by P^+ on $\underset{\sim}{X}^{(n)}$. If $\Theta \subset E$, we could for completeness, set

$$T_n[\underset{\sim}{X}^{(n)}] = \inf A_n(X^{(n)}) \quad \text{if} \quad A_n(\underset{\sim}{X}^{(n)}) = \phi$$
$$+\infty \quad \text{if} \quad A_n(\underset{\sim}{X}^{(n)}) \neq \phi \,. \tag{2.3}$$

Borel measurability of these maps follows from the theory of Brown and Purves (1973). Some notation is necessary.

If 0 is a set of ordered pairs, the projection of 0, or proj (0), is the set of all first co-ordinates of members of 0.

If $C \subset U \times V$ where U, V are metric spaces, S will be said to be a *Borel selection* of C whenever

(i) S is a Borel set;

(ii) $S \subset C$;

(iii) For $u \in U$, the section $S_u = \{v \in V | (u,v) \in S\}$ contains at most one point;

(iv) proj (S) = proj (C).

Corresponding to each selection S is the function ρ_s, which assigns to each $u \in$ proj (C) the second co-ordinate of the unique member of S with first co-ordinate u. Thus $(u, \rho_s(u)) \in C$, for all $u \in$ proj (C).

PROPOSITION 2.1 (Brown and Purves, Theorem 1): Let U, V be complete separable metric spaces and $C \subset U \times V$ be a Borel set. If for each $u \in U$, the section C_u is σ-compact there is a Borel selection S of C. Further proj (C) is a Borel set and ρ_s is a Borel measurable function defined on proj (C).

Letting D^* be the domain of the real valued function, f, of two variables, the infimum of the sets of reals, $\{f(x,\theta) | \theta \in D^*_x\}$, is abbreviated inf f_x. A function $f(x,\theta)$ is said to be *lower semicontinuous* in θ if

$$\inf_{\theta' \in W} f(x,\theta') \to f(x,\theta) \ ,$$

as the neighbourhood W shrinks to $\{\theta\}$.

COROLLARY 2.1 (Brown and Purves): Let V, $\bar{\Theta}$ be complete separable metric spaces, where $\bar{\Theta}$ is the closure of $\Theta \subseteq E^r$. Assume f to be a real valued Borel measurable function defined on a Borel subset D^* of $R \times \Theta$. (Assume $\Theta \subset E^r$ is Borel). Suppose that for each $x \in \text{proj } (D^*)$, the section D^*_x is σ-compact and $f(x,\cdot)$ is lower semicontinuous with respect to the topology on D^*_x. Then:

(i) the sets

$$J = \text{proj } (D^*)$$

$$Q = \{x \in D^* \mid \text{for some } \theta \in D^*_x,\ f(x,\theta) = \inf f_x\} \ ,$$

are Borel;

(ii) there is a Borel measurable function T from V to the extended real line satisfying for $x \in V$,

$$f(x,T(x)) = \inf f_x \quad \text{if} \quad x \in Q \ .$$

and

$$T(x) = +\infty \qquad \text{if} \quad x \notin Q \ .$$

In the proof of the Corollary, the functional T is defined as a map

$$T : x \to \rho(x, \inf f_x) \ , \quad x \in Q \ ,$$

where ρ is defined by a Borel selection of the set

$$B = \{((x,v),\theta) \in (V \times E) \times \bar{\Theta} \mid (x,\theta) \in D^*, f(x,\theta) \leq v\} \ ,$$

whence Proposition 2.1 is utilized. For $x \in Q$, $f(x,T[x]) = \inf f_x$. Since Q is shown to be Borel and T is the composition of Borel measurable maps ρ and $x \to (x, \inf f_x)$, it follows that T is Borel measurable.

Now the measurability of an estimator that is defined by equation (2.1) and some selection statistic will follow from Corollary 2.1. Suppose $g_n(\underset{\sim}{x}^{(n)},\theta)$ is lower semicontinuous in θ, and Θ is σ-compact. Then

$$H_n = \{(\underset{\sim}{x}^{(n)},\theta) \,|\, g_n(\underset{\sim}{x}^{(n)},\theta) = 0,\ \underset{\sim}{x}^{(n)} \in R^n,\ \theta \in \Theta\}$$

is Borel, and further $H_n(\underset{\sim}{x}^{(n)})$ is σ-compact. Measurability follows by the following construction

Construction (†): Set $D^* = H_n$ and $f(x,\theta) = f_n(\underset{\sim}{x}^{(n)},\theta)$ in the Corollary.

This does not conclude discussion. For in the proof of the Corollary Proposition 2.1 was used to produce a map ρ, and therefore a map T. It is possible for there to be several maps ρ from the set

$$A = \{(x,v) \in V \times E \,|\, \text{for some } \theta \in \Theta,\ ((x,v),\theta) \in B\}\ ,$$

into B. Hence the above construction (†) could reveal several possible maps $T[\underset{\sim}{x}^{(n)}]$.

There exist several possibilities for a sequence of estimators $\{T_n\}$ constructed under (†) using sequences of functions $\{g_n\}$, $\{f_n\}$. It is assumed that these functions satisfy

(i) g_n, f_n have domain $R^n \times \bar{\Theta}$ and are measurable for every $n \in N$; and

(ii) g_n, f_n are lower semicontinuous in θ for every $n \in N$.

Resulting estimators can be distinguished according to whether neither, one, or both of the subsequent criteria are true:

(1) $R^n - \text{proj}\ (H_n)$ is a null set with respect to $G^{(n)+}$

(2) There exists a Borel set $Q_n^* \subset \text{proj}\ (H_n)$ such that $\text{proj}\ (H_n) - Q_n^*$ is a null set with respect to $G_n^{(+)}$, where for every $\underset{\sim}{x}^{(n)} \in Q_n^*$ there

exists a unique $\theta \in \Theta$ such that $(\underset{\sim}{x}^{(n)}, \theta) \in H_n$ and

$$f_n(\underset{\sim}{x}^{(n)}, \theta) = \inf_{\theta \in H_n(\underset{\sim}{x}^{(n)})} f_n(\underset{\sim}{x}^{(n)}, \theta) \ .$$

On the set R^n - proj (H_n) the estimator takes the value $+\infty$ whilst on $Q_n - Q_n^*$ the estimator need not be uniquely determined. (When construction (†) is made Q_n corresponds to the set Q of the Corollary). Ideally both (1) and (2) are preferable, but often we tentatively forgo either or both of these to gain elegant asymptotic results. Some constructions g_n may have a set of decreasing probability in n in which solutions to equations (2.1) do not exist, in which case R^n - proj (H_n) is not a null set with respect to $G^{(n)+}$. Letting

$$Q_n^{**} = \{\underset{\sim}{x}^{(n)} \mid \text{there exists a unique } \theta \in \Theta \text{ such that } (\underset{\sim}{x}^{(n)}, \theta) \in H_n,$$

$$\text{and } f_n(\underset{\sim}{x}^{(n)}, \theta) = \inf_{\theta \in H_n(x^{(n)})} f_n(\underset{\sim}{x}^{(n)}, \theta)\} \ ,$$

the identification problem is apparent when proj $(H_n) - Q_n^{**}$ is not contained in a null set. But if $\underset{\sim}{X}^{(n)}$, the sample, is an element of this latter set even in the pathological case of it being a null set the identification becomes vital. By defining the set

$$H_n^+ = \{(\underset{\sim}{x}^{(n)}, \theta) \mid (\underset{\sim}{x}^{(n)}, \theta) \in H_n,\ f_n(\underset{\sim}{x}^{(n)}, \theta) = \inf_{\theta \in H_n(x^{(n)})} f_n(\underset{\sim}{x}^{(n)}, \theta)\},$$

which is a Borel set since the map $x \to \inf f_x$, $x \in$ proj D^* is measureable, we can introduce a refining selection statistic f_{n1} and operate with the construction (†) using $D^* = H_n^+$ and $f = f_{n1}$.

An important illustration of the Corollary is the following.

EXAMPLE 2.1: Let Θ be a compact subset of E^r. Then for any $\underset{\sim}{x}^{(n)} \in$ proj (H_n) the set $H_n(\underset{\sim}{x}^{(n)})$ is compact by lower semicontinuity of g_n. Similarly lower semicontinuity of f_n implies proj $(H_n) = Q_n$.

A suitable value θ for which

$$f_n(\underset{\sim}{x}^{(n)},\theta) = \inf_{\theta \in H_n(x^{(n)})} f_n(\underset{\sim}{x}^{(n)},\theta) \tag{2.4}$$

can be described explicitly in the manner sketched below.

For $r = 1$, Borel measurability of the estimator T_n of (2.3) is as a consequence of:

(*) Let H be a Borel set in $V \times E$, where V is a complete separable metric space. The set proj (H) is Borel. Then the function $x \to \inf H_x$, $x \in$ proj $(H) \subset V$ is Borel measurable.

This is an instance of Corollary 2.1 with f chosen to be zero on H and one on the complement of H. That A_n is Borel follows because (*) implies proj (H_n) is Borel and also from the proof of Corollary 2.1 (or otherwise) $\underset{\sim}{x}^{(n)} \to \inf_{\theta \in H_n(x^{(n)})} f_n(\underset{\sim}{x}^{(n)},\theta)$, $\underset{\sim}{x}^{(n)} \in$ proj (H_n) is Borel measurable. Since each $\underset{\sim}{x}^{(n)}$ section of A_n is σ-compact (*) implies that $T_n : \underset{\sim}{x}^{(n)} \to \inf A_n(\underset{\sim}{x}^{(n)}) = H_n^+(\underset{\sim}{x}^{(n)})$, $\underset{\sim}{x}^{(n)} \in$ proj $(A_n) =$ proj (H_n) is Borel measurable. The effective selection statistic in this instance is $\theta - \inf \Theta$.

The case where $r > 1$ is examined in Remark 3 of Brown and Purves.

Quite often it is useful for the purposes of limit theorems that are given only at the underlying distribution P_{θ_o} to define the statistic uniquely through the selection statistic $\|\theta - \theta_o\|$. If $\Theta \subset E$ the refining selection statistic might be $\theta - \inf \Theta$ for example. When both equations and selection statistic can be written as functions of the empirical distribution function we write $T_n[X^{(n)}] = T[F_n]$, and refer to the selection functional $f.(\theta)$. The selection statistic is then $f_{F_n}(\theta)$.

Given that maps $\{T_n\}$ are Borel measurable it follows that

$$C = \bigcap_{\ell=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{\omega \,|\, \|T_n[\underset{\sim}{X}^{(n)}(\omega)] - T\| < 1/\ell\} \in A,$$

and we see from Definition 1.1 that $T_n \xrightarrow{a.s.} T$ if and only if $P(C) = 1$. This is strong consistency. A sequence $\{T_n\}$ that is defined by $\{g_n\}$, $\{f_n\}$ via (t) will be asymptotically well defined provided the almost sure limit T is unique, whatever the representations are that T_n can take.

Pfanzagl (1969) establishes measurability of "minimum contrast" estimators, which he defines as follows:

DEFINITION 2.1: A strict estimate for the sample size n is a $\mathcal{B}^n$ measurable map $T_n : R^n \to \Theta$, which depends on $\underset{\sim}{x}^{(n)}$ only. A m.c. (minimum contrast) estimate for the sample size n is a strict estimate for which

$$\frac{1}{n} \sum_{i=1}^{n} \ell(x_i, T_n[\underset{\sim}{x}^{(n)}]) = \inf \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, t) \,|\, t \in \Theta \right\}.$$

For instance the M.L.E. is obtained by setting $\ell(x,\theta) = -\log f_\theta(x)$. The restriction of the mappings to Θ can be a strong one when Θ is a proper subset of E^r. Examples where this estimate does not exist because the infimum is not attained on Θ are common. This is seen for the M.L.E. of the mixture parameter in §8.1, or the M.L.E. of the parameters of a mixture of two normal distributions when both dispersions are unknown, discussed in section C.

PROPOSITION 2.2 (Pfanzagl, P.252): Set (W,U) to be a locally compact Hausdorff space with countable base, and $\sigma(U)$ the σ-algebra over W generated by U. Let $\ell(\cdot,t) : R \to [-\infty,\infty]$, $t \in W$, be such that

(0) for all $n \in N$ and all $(x_1,\ldots,x_n) \in R^n$,

$$\inf \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, t) \,|\, t \in W \right\} \text{ is attained in } W ;$$

(1) $t \to \ell(\cdot,t)$ is lower semicontinuous for all $x \in R$;

(2) $\inf \ell_D \subset B$ for all compact sets $D \subset W$.

Then for any $n \in N$ there exists a B^n, $\sigma(U)$-measurable function $T_n : R^n \to W$, such that

$$\frac{1}{n} \sum_{i=1}^{n} \ell(\underset{\sim}{x}^{(n)}, T_n[\underset{\sim}{x}^{(n)}]) = \inf \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i,t) \,|\, t \in W \right\}.$$

(Condition (2) is always fulfilled if instead of (1) the stronger condition

$$\lim_{s \to t} \ell(\cdot,s) = \ell(\cdot,t) \quad \text{for all} \quad t \in W \quad \text{holds.})$$

Reiss (1978), recognizing that the minimum contrast estimator need not exist, investigated the consistency of a more general class of estimators; asymptotic minimum contrast estimators.

§2.2 Existence; Relation to the Minimal Distance Approach

Assumption of a parametric family P inducing the family of marginal distributions $F = \{F_\theta \,|\, \theta \in \Theta\}$ on (R,B) is motivated by the idea that the underlying probability measure of a "sample" $X_1,\ldots,X_n$, is a $P \in P$; that is there is some $\theta_o \in \Theta$ so that F_{θ_o} is the induced marginal distribution from P. A natural requirement for point estimation to be unambiguous is that F be identifiable which is so whenever $\theta_1,\theta_2 \in \Theta$, $\theta_1 \neq \theta_2$ implies $F_{\theta_1} \neq F_{\theta_2}$. Presumably the parameter θ describes physical characteristics of "nature", and assuming the latter follows some sort of continuity in θ we attempt to estimate the true θ_o on the evidence presented, which is the sample.

In the literature, for i.i.d. sequences $\underset{\sim}{X}$ there exist many strongly consistent estimators proposed to be solutions of a set of $r \times 1$ (possibly nonlinear) equations which we label

$$K_n(\theta, \underset{\sim}{X}^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} \underset{\sim}{\psi}(X_i, \theta) = 0 \ , \tag{2.5}$$

where $\underset{\sim}{\psi}$ is some measurable mapping from $(R \times \Theta, \mathcal{B} \times \mathcal{E}^r(\Theta))$ to $(E^r, \mathcal{E}^r)$. Here $\mathcal{E}^r$ are the Borel sets on E^r, and $\mathcal{E}^r(\Theta)$ is the relative topology for Θ. *Fisher consistency* requires that

$$E_\theta[\psi(X, \theta)] = 0 \ , \tag{2.6}$$

the expectation being taken componentwise with respect to F_θ. Huber (1964, 1967) studied M-estimators through these equations, restricting ψ for reasons of identification or asymptotic identification of the estimator. Chanda (1954) and Foutz (1977) discussed aspects of consistency of a solution of (2.5) when ψ was assumed to be the efficient score.

Hampel (1968) took the approach of defining $\mathcal{F}_n = \{G_n\} \subset \mathcal{G}$, the set of probability measures whose atoms have probabilities equal to $1/n$ or a multiple of $1/n$ and regarding solutions $T_n[\underset{\sim}{X}^{(n)}]$ of (2.5) as functionals defined on $\mathcal{F}_n$. Writing $T_n[\underset{\sim}{X}^{(n)}] = T[\psi, F_n]$, consistency then follows from weak continuity of functionals $T[\psi, \cdot]$. The σ-algebra on $\mathcal{F}_n$ is that generated by the Prokhorov distance which we give here.

DEFINITION 2.2: Let d be the metric on the space R generating Borel sets $\mathcal{B}$. Denote A^δ to be the closed δ-neighbourhood of A. That is the closure of

$$\{x \mid \text{there exists a } y \in A, \text{ for which } d(x,y) \leq \delta\} \ .$$

Then the Prokhorov distance between two distributions F, G on $(R, \mathcal{B})$ is defined,

$$d_p(F,G) = \inf\{\varepsilon \mid F\{A\} \leq G\{A^\varepsilon\} + \varepsilon, \ G\{A\} \leq F\{A^\varepsilon\} + \varepsilon, \text{ for every } A \in \mathcal{B}\}.$$

This approach to consistency does not lead to proofs of consistency for many established estimators. Just as for uniform convergence theory

(§1.2), the same is true here. Results derived deterministically using weak convergence are not as broad as those obtainable through probabilistic arguments.

The term "*influence function*" will refer to the equivalence class of functions defined by the relation: $\psi_1 \sim \psi_2$ if and only if for each ψ_1, ψ_2 there exists some constant nonsingular $r \times r$ matrix A so that $\psi_1 = A\psi_2$. It is distinct from the specific curve introduced by Hampel (1968) called the influence curve (§4.4).

We stress that for many influence functions there can exist more than one solution to equations (2.5). If a selection statistic, that can be written $f_{F_n}(\theta)$, is required, the corresponding estimating functional may be written $T[\psi,f,\cdot]$. Both ψ and f characterize the functional.

The contemporary approach of authors Huber (1967), Pfanzagl (1969), Landers (1972) and Reiss (1978) has been related to m.c. estimators. They in general adopt an asymptotic identification criterion:

$$\int \ell(x,T[\ell,G])dG(x) < \int \ell(x,\theta)dG(x) \quad \text{for all} \quad \theta \neq T[\ell,G],\ G \in \mathcal{G}\ ,$$

and (2.7)

$$T[\ell,F_{\theta_o}] = \theta_o \quad \text{for all} \quad \theta_o \in \Theta\ .$$

This restriction may be regarded as a generalization of the Jensen inequality, used in Wald's (1949) argument for consistency of the maximum likelihood estimator. If Θ is an open interval and ℓ satisfies appropriate differentiability conditions, the estimators are based on (2.5), with $\psi(x,\theta) = \nabla\ell(x,\theta)$. The selection statistic is then $f_{F_n}(\theta) = \int \ell(x,\theta)dF_n(x)$. Questions of basic importance concern existence and asymptotic uniqueness of consistent solutions to the equations (2.5). *We will mean by an* M-*estimator, any estimator arrived at through equations* 2.5 (*viz construction* (†)). *They need not be minimum contrast estimators.*

Asymptotics of M-estimators are investigated as a consequence of the uniform convergence theory, although techniques equally apply to m.c. and minimal distance estimators. Wolfowitz (1952, 1954, 1957) was an initial exponent of the general methods of the latter, while Sahler (1970), Bolthausen (1977), and Pollard (1980) discuss further asymptotic properties in a general framework. More applied works that include asymptotics of minimal distance methods are by Blackman (1955), Choi and Bulgren (1968), Paulson, Holcomb and Leitch (1975), and Quandt and Ramsey (1978). Efficiency and robustness motivate discussion of certain L_2-norms in Heathcote (1977), and Beran (1977). An L_2 distance takes the form

$$L_n(\theta) = \int |h_n(t) - h(t,\theta)|^2 dw(t) ,$$

where h_n is characterized by the empirical distribution (e.g. characteristic function, moment generating function, kernel density estimate), and $\omega(t)$ is some weight function possibly dependent on θ. The asymptotic identification of the estimator under the parametric model is usually guaranteed assuming at least

$$L_{F_{\theta_o}}(\theta) = \int |h(t,\theta_o) - h(t,\theta)|^2 dw(t) > 0, \ \theta \neq \theta_o, \ \theta \in \Theta,$$

and arguing from the convergence of $h_n(t) \to h(t,\theta_o)$. By the convergence $L_n(\theta) \to L_{F_{\theta_o}}(\theta)$ uniformly in $\theta \in \Theta$ it is shown that the minimizing statistic $\hat{\theta}_n$ of $L_n(\theta)$ converges to θ_o. Slight perturbations from F_{θ_o} can possibly perturb the convergence greatly from the value θ_o, depending on the weight function $w(t)$.

But for some L_2-distances the minimizing equations correspond to M-estimating equations. This together with the intuitively simpler to work with construction of the M-estimator make it the central theme of

our robustness discussion. Asymptotics are discussed without necessarily assuming the underlying distribution $G_o \in F$. The emphasis on studying convergence of functionals in this case stems from robustness theory, where we look for stability in "neighbourhoods" of a given F_{θ_o}. This is merely a formalization of our uncertainty about the choice of model family which *we* put forward to explain a given sample.

A neighbourhood of a distribution G, $n(\cdot,G)$, is only required to be a subset of $\mathcal{G}$, containing G, and satisfying the ordering property,

$$n(\varepsilon_1,G) \subset n(\varepsilon_2,G) \quad \text{whenever} \quad 0 < \varepsilon_1 \leq \varepsilon_2 .$$

It may be determined by a metric, or otherwise. With this generality we can determine asymptotic limits, either of the M-functional or the M-estimator, by varying the neighbourhood. We consider existence of functionals on neighbourhoods about some specified distribution G_o under varying conditions on ψ. Functionals $T[\psi,\cdot]$ evaluated at a distribution G are a solution of

$$K_G(\theta) = \int \psi(x,\theta)dG(x) = 0 \tag{2.8}$$

if a solution exists. If not $T[\psi,G]$ is set equal to $+\infty$.

Properties of minimal distance estimators and or solutions to either (2.5) or (2.8) are dependent largely on Euclidean geometry. In the sequel we let $D \subset \Theta$ be some nondegenerate compact set that contains the parameter θ_o^* in its interior. We distinguish between two situations:

B1: The parameter θ_o^* is a zero of an $r\times 1$ continuously differentiable vector function $K_{G_o}(\theta)$, so that $\nabla K_{G_o}(\theta_o^*)$ is nonsingular.

B2: The parameter θ_o^* is an isolated local minima of twice continuously differentiable $Q_{G_o}(\theta)$. (This implies $\nabla\nabla' Q_{G_o}(\theta_o^*)$ is positive definite). This will be regarded as being some distance quantity.

Under restrictive conditions on ψ it is possible to see the correspondence between proofs of existence of minimal distance estimates and existence of solutions to equations (2.8). For instance we can set

$$Q_G(\theta) = \|K_G(\theta)\|^2 , \tag{2.9}$$

and then B1 implies B2. It can also happen that $K_G(\theta) = \nabla' Q_G(\theta)$, whence B2 is equivalent to B1. We use the condition

C1: For every $\varepsilon > 0$ there exists a $\delta > 0$, so that $G \in n(\delta, G_o)$ implies

$$\sup_{\theta \in D} \| \int \psi(x,\theta) dG(x) - \int \psi(x,\theta) dG_o(x) \| < \varepsilon .$$

This assumption implies in the situation (2.9) that given $\varepsilon > 0$ there exists a $\delta > 0$, so that $G \in n(\delta, G_o)$ implies

$$\sup_{\theta \in D} |Q_G(\theta) - Q_{G_o}(\theta)| < \varepsilon . \tag{2.10}$$

We denote by ∂U the boundary of a set U.

LEMMA 2.1: Let θ_o^*, G_o, and Q be so that B2 and (2.10) hold. Then there exists an $\varepsilon_1 > 0$ so that $U_{\varepsilon_1}(\theta_o^*) \subset D$, and given $\varepsilon > 0$ arbitrary there is a $\delta > 0$ so that $G \in n(\delta, G_o)$ implies

$$\inf_{\theta \in U_{\varepsilon^*}(\theta_o^*)} Q_G(\theta) < \inf_{\theta \in \partial U_{\varepsilon^*}(\theta_o^*)} Q_G(\theta) ,$$

where $\varepsilon^* = \min(\varepsilon, \varepsilon_1)$. Hence if $Q_G(\theta)$ is continuous on D there exists at least one local minima of Q in $U_{\varepsilon^*}(\theta_o^*)$.

PROOF: Let $\{d_k\}_{k=1}^r$ be the eigenvalues of the Hessian matrix $\nabla\nabla' Q_{G_o}(\theta)$. Write $d^-(\theta) = \min_{1 \le k \le r} d_k(\theta)$. Given arbitrary unit vector $\underset{\sim}{x}$, $\|\underset{\sim}{x}\| = 1$, we may express $\underset{\sim}{x}$ in terms of the orthonormal basis

$$\underset{\sim}{x} = a_1 \gamma_1(\theta) + \ldots\ldots + a_r \gamma_r(\theta) , \tag{2.11}$$

where the $\gamma_i(\theta)$'s are the orthonormal eigenvectors of $\nabla\nabla' Q_{G_o}(\theta)$. Observe,

$$\begin{aligned} \underset{\sim}{x}'\nabla\nabla' Q_{G_o}(\theta)\underset{\sim}{x} &= \underset{\sim}{x}'\Gamma(\theta)D(\theta)\Gamma'(\theta)\underset{\sim}{x} \\ &= \sum_{k=1}^{r} d_k(\theta)(\gamma_k(\theta)'\underset{\sim}{x})^2 \\ &\geq d^-(\theta) \sum_{k=1}^{r} a_k^2 = d^-(\theta) \ . \end{aligned}$$

By continuity choose ε_1 so that $\theta \in \bar{U}_{\varepsilon_1}(\theta_o^*)$ implies

$$\|\nabla\nabla' Q_{G_o}(\theta) - \nabla\nabla' Q_{G_o}(\theta_o^*)\| < d^-(\theta_o^*)/2 \ .$$

Then from the Taylor expansion, for any $\theta \in \partial U_{\varepsilon^*}(\theta_o^*)$, the boundary of $U_{\varepsilon^*}(\theta_o^*)$,

$$\begin{aligned} Q_{G_o}(\theta) &= Q_{G_o}(\theta_o^*) + \nabla' Q_{G_o}(\theta_o^*)(\theta-\theta_o^*) + \tfrac{1}{2}(\theta-\theta_o^*)\{\nabla\nabla' Q_{G_o}(\xi)\}(\theta-\theta_o^*) \\ &= Q_{G_o}(\theta_o^*) + 0 + \tfrac{1}{2}(\theta-\theta_o^*)'\{\nabla\nabla' Q_{G_o}(\theta_o^*)\}(\theta-\theta_o^*) \\ &\qquad + \tfrac{1}{2}(\theta-\theta_o^*)'\{\nabla\nabla' Q_{G_o}(\xi) - \nabla\nabla' Q_{G_o}(\theta_o^*)\}(\theta-\theta_o^*) \\ &> Q_{G_o}(\theta_o^*) + \tfrac{1}{2}\varepsilon^{*2} d^-(\theta_o^*) - \tfrac{1}{4}\varepsilon^{*2} d^-(\theta_o^*) \\ &= Q_{G_o}(\theta_o^*) + \tfrac{1}{4}\varepsilon^{*2} d^-(\theta_o^*) \ . \end{aligned}$$

Thus choose δ so that $G \in n(\delta, G_o)$ implies

$$\sup_{\theta \in D} |Q_G(\theta) - Q_{G_o}(\theta)| < \frac{1}{8}\varepsilon^{*2} d^-(\theta_o^*) \ ,$$

whence $\theta \in \partial U_{\varepsilon^*}(\theta_o^*)$ implies

$$\begin{aligned} Q_G(\theta) &> Q_{G_o}(\theta) - \frac{1}{8}\varepsilon^{*2} d^-(\theta_o^*) \\ &> Q_{G_o}(\theta_o^*) + \frac{1}{8}\varepsilon^{*2} d^-(\theta_o^*) \\ &> Q_G(\theta_o^*) \ , \end{aligned}$$

and the lemma is proved.

The assumption of the existence of two continuous partial derivatives of Q is required only at G_o. What has been shown is the existence of a nondegenerate compact region of arbitrarily small size about θ_o^* for which there exists a neighbourhood of G_o so that if the minimum of $Q_G(\theta)$ is attained on that region for any distribution G of the neighbourhood, the minimizing parameter is found in the interior of that region. If $Q_G(\theta)$ is continuously differentiable in θ, the minimum exists and is found by solving $\nabla' Q_G(\theta) = 0$, or equations (2.8) if Q is a m.c. distance or a suitable L_2 distance. But differentiability in θ of $Q.(\theta)$ on $n(\varepsilon, G_o)$ is not always possible.

If $Q_{G_o}(\theta)$ is given by (2.9) and $Q_{G_o}(\theta_o^*) = 0$, we need only assume existence of one continuous partial derivative of $K_{G_o}(\theta)$ to observe the existence of a minimum.

LEMMA 2.2: Let θ_o^*, G_o, and ψ be so that B1 and C1 hold. Then there exists an $\varepsilon_1 > 0$ so that $U_{\varepsilon_1}(\theta_o^*) \subset D$ and given $\varepsilon > 0$ arbitrary, there is a $\delta > 0$ for which $G \in n(\delta, G_o)$ implies

$$\inf_{\theta \in U_{\varepsilon^*}(\theta_o^*)} \|K_G(\theta)\| < \inf_{\theta \in U_{\varepsilon^*}(\theta_o^*)} \|K_G(\theta)\| ,$$

where $\varepsilon^* = \min(\varepsilon, \varepsilon_1)$. A minimum attained on $\bar{U}_{\varepsilon^*}(\theta_o^*)$, is attained in the interior of that set (which is the case if $K_G(\theta)$ is continuous on D).

PROOF: By continuity of the matrix of partial derivatives of $\nabla K_{G_o}(\theta)$, and the nonsingularity of $\nabla K_{G_o}(\theta_o^*)$ let ε_1 be given as in the inverse function theorem (Appendix 1) so that

$$\sup_{\theta \in \bar{U}_{\varepsilon_1}(\theta_o^*)} \|\nabla K_{G_o}(\theta) - \nabla K_{G_o}(\theta_o^*)\| < \kappa ,$$

where $\kappa = (2\|\nabla K_{G_o}(\theta_o^*)^{-1}\|)^{-1}$. Then by (a) of that theorem $\theta \in \partial U_{\varepsilon^*}(\theta_o^*)$ it is true that

$$\varepsilon^* = \|\theta-\theta_o^*\| \leq 2\|\nabla K_{G_o}(\theta_o^*)^{-1}\|\|K_{G_o}(\theta)-K_{G_o}(\theta_o^*)\|$$

since $K_{G_o}(\theta_o^*) = 0$ for every $\theta \in \partial U_{\varepsilon^*}(\theta_o^*)$

$$\|K_{G_o}(\theta)\| \geq \kappa\varepsilon^* .$$

By C1 let $\delta > 0$ be so that $G \in n(\delta, G_o)$ implies

$$\sup_{\theta \in D}\|K_G(\theta) - K_{G_o}(\theta)\| < \varepsilon_o < \kappa\varepsilon^*/2 .$$

Then for $\theta \in \partial U_{\varepsilon^*}(\theta_o^*)$

$$\|K_G(\theta)\| > \kappa\varepsilon^* - \varepsilon_o > \varepsilon_o > \|K_G(\theta_o^*)\| ,$$

and the lemma is proved.

The importance of the formulations using neighbourhoods can be demonstrated with a simple choice of neighbourhood,

$$n_o(\delta, G_o) = \{G \in \mathcal{G} \mid \sup_{\theta \in D} \|K_G(\theta) - K_{G_o}(\theta)\| < \delta\} .$$

For this neighbourhood C1 automatically holds with $\delta = \varepsilon$.

EXAMPLE 2.2: Let $G_o = \Phi$, the standard normal distribution, and set

$$\psi(x) = x \qquad 0 < |x| \leq c$$

$$0 \qquad \text{otherwise}$$

Then $K_\Phi(\theta) = \int \psi(x-\theta)d\Phi(x)$ is continuously differentiable, $K_\Phi'(0) \neq 0$, and $K_\Phi(0) = 0$. Let $d > 0$ and n_o be defined for the compact set $D = [-d,d]$. By Lemma 2.2, for every $\varepsilon > 0$ there exists a $\delta > 0$ so that $G \in n_o(\delta, \Phi)$ implies that there exists a local minima of $K_n^2(\theta)$ in $(-\varepsilon,\varepsilon)$. But by Theorem 1.1 observe for any stationary ergodic sequence with marginal Φ,

$$\int \psi(x-\theta)dF_n(x) = \int_{[-c-\theta,c-\theta]} x\,dF_n(x) \xrightarrow{a.s.} \int \psi(x-\theta)d\Phi(x)$$

uniformly in $\theta \in D$.

That is "$F_n \in n_o(\delta,\Phi)$" holds f.a.s.l.n., and so there exists a consistent sequence, $\{\hat{\theta}_n\}$, of minima of $K_n^2(\theta)$, consistent to $\theta_o^* = 0$. It is important to note here that ψ is discontinuous in θ.

§2.3 Consistency and Uniqueness of the Multivariate M-functional

Considering the M-functional specifically as a solution $T[\psi,\cdot]$ of equations (2.8) we make the further assumptions

C2 $\psi(x,\theta)$ is a continuous function on $R \times D$.

C3 There exists a continuous function $g \in L_1(G_o)$ so that

$$\|\psi(x,\theta)\| < g(x) \quad \text{for all} \quad (x,\theta) \in (R \times D) .$$

Then assumption C2 leads to an extension of Lemma 2.2. Addition of C3 ensures a corollary giving existence of roots f.a.s.l.n. of the estimating equations.

LEMMA 2.3: Assume θ_o^*, G_o, and ψ are so that B1, C1, and C2 hold. Then there is an $\varepsilon_1 > 0$ so that $U_{\varepsilon_1}(\theta_o^*) \subset D$, and given $\varepsilon > 0$ arbitrary, there is a $\delta > 0$ so that $G \in n(\delta,G_o)$ implies there exists a solution, $t[G] \in U_{\varepsilon^*}(\theta_o^*)$, of equations (2.8) where $\varepsilon^* = \min(\varepsilon,\varepsilon_1)$.

PROOF: By continuity of the partial derivative let $\varepsilon_1 > 0$ be so that $\nabla K_{G_o}(\theta)$ is nonsingular on $\bar{U}_{\varepsilon_1}(\theta_o^*)$. Then consider the open ball $K_{G_o}(U_{\varepsilon^*}(\theta_o^*))$ containing $\underset{\sim}{0}$ and let $\bar{B}_r$ be a ball of positive radius r contained in $K_{G_o}(U_{\varepsilon^*}(\theta_o^*))$. Since K_{G_o} is a continuous map $K_{G_o}^{-1}(\bar{B}_r)$ is closed, as is then $W = K_{G_o}^{-1}(\bar{B}_r) \cap \bar{U}_{\varepsilon^*}(\theta_o^*)$. Consider the 1-1

homeomorphism $K_{G_o}(\theta)\big|_{\bar{B}_r}$, which maps W onto $\bar{B}_r$. If $\sup_{\theta \in W} \|g_G(\theta)\| \leq r$, we can transform the equation

$$K_{G_o}(\theta) + g_G(\theta) = 0 \quad \text{for} \quad \theta \in W$$

into

$$t + g_G(K_{G_o}^{-1}(t)) = 0 \quad \text{for} \quad \|t\| \leq r \ .$$

The map $t \to -g_G(K_{G_o}^{-1}(t))$ is continuous and maps the ball $\bar{B}_r$ into itself. So Brouwer's fixed point theorem (Appendix 1) guarantees a solution. Then letting δ be so that $G \in n(\delta, G_o)$ implies

$$\sup_{\theta \in D} \|K_G(\theta) - K_{G_o}(\theta)\| \leq r \ ,$$

the theorem is proved with $g_G(\theta) = K_G(\theta) - K_{G_o}(\theta)$.

Reeds (1976) used the Brouwer's fixed point theorem in a like manner, while another version was used in the consistency argument of Aitchison and Silvey (1958).

COROLLARY 2.2: Let θ_o^*, G_o, and ψ be so that B1, C2 and C3 hold, and G_o is the marginal of a stationary ergodic sequence $\underset{\sim}{X}$. Then given $\varepsilon > 0$ there exists a root $\hat{\theta}_n(\underset{\sim}{X}^{(n)})$ of equations (2.5), within $U_\varepsilon(\theta_o^*)$ f.a.s.l.n. .

PROOF: By Lemma 1.6 and Proposition 1.4, for fixed $\delta > 0$, see that C2 and C3 imply

$$F_n \in n_o(\delta, G_o) \quad \text{f.a.s.l.n.} \ , \tag{2.12}$$

and the result follows from the Lemma.

The assumption that $\psi(x,\theta)$ is continuous in θ is utilized here to ensure existence of a root to equations (2.5). But while Corollary 2.2 appears to describe the existence of a consistent root to θ_o^*, it is

far from complete in its description of the identity of the root. For instance the above Corollary states: there is a ball U_{ε_1} about θ_o^* so that given any $0 < \varepsilon < \varepsilon_1$ arbitrarily small, there exists a root of equations (2.5) in U_ε f.a.s.l.n.. There may be more than one root. But also there is the possibility that there exists roots in $U_{\varepsilon_1} - U_\varepsilon$. Similar statements can be made about the local minima of empirical distances $Q_{F_n}(\theta)$. This raises a natural question; "Do all the local minimizing parameters of $Q_{F_n}(\theta)$, or the zeros of $K_n(\theta)$, that are known to lie in the set $U_{\varepsilon_1}(\theta_o^*)$ converge to θ_o^* f.a.s.l.n.?" An answer for the latter, which is also an answer for the former should $Q_{F_n}(\theta)$ be continuously differentiable, is provided by the following Theorem:

THEOREM 2.1:

Let θ_o^*, G_o, and ψ be so that B1, C2, and C3 hold. By continuity let ε_1 be so that

$$\sup_{\theta \in U_{\varepsilon_1}(\theta_o^*)} \|\nabla K_{G_o}(\theta) - \nabla K_{G_o}(\theta_o^*)\| < (2\|\nabla K_{G_o}(\theta_o^*)^{-1}\|)^{-1}.$$

Let $\underset{\sim}{X}$ be an ergodic sequence with marginal G_o. Then, for any sequence $\{\hat{\theta}_n(\underset{\sim}{X}^{(n)})\}_{n=1}^{\infty}$ of roots of (2.5), such that "$\hat{\theta}_n(\underset{\sim}{X}^{(n)}) \in U_{\varepsilon_1}(\theta_o^*)$" holds f.a.s.l.n.,

$$\|\hat{\theta}_n - \theta_o\| \xrightarrow{a.s.} 0 .$$

PROOF: From (a) of the Inverse Function Theorem, since $\hat{\theta}_n$, $\theta_o^* \in U_{\varepsilon_1}(\theta_o^*)$

$$\begin{aligned}\|\hat{\theta}_n - \theta_o^*\| &\leq 2\|\nabla K_{G_o}(\theta_o^*)^{-1}\|\|K_{G_o}(\hat{\theta}_n) - K_{G_o}(\theta_o^*)\| \\ &= 2\|\nabla K_{G_o}(\theta_o^*)^{-1}\|\|K_{G_o}(\hat{\theta}_n)\| \\ &= 2\|\nabla K_{G_o}(\theta_o^*)^{-1}\|.\|K_{G_o}(\hat{\theta}_n) - K_n(\hat{\theta}_n)\| \qquad (2.13)\end{aligned}$$

$\xrightarrow{a.s.} 0$ by (2.12).

The mode of proof may be applied to any set of nonlinear equations for which there is uniform convergence over the parameter space. It is possible to show with further regularity conditions on ψ the existence of an asymptotically unique consistent root of the equations (2.5). In fact there is a small region about θ_o^* in which there will exist a unique root f.a.s.l.n.. Clearly this has no significance by itself but in the context of the parametric model where $G_o = F_{\theta_o}$, and ψ is suitably regular and satisfying (2.6), then these results apply to the parameter θ_o, as well as to other roots of $E_{\theta_o} \psi(X,\theta) = 0$. There exists a sequence of roots $\hat{\theta}_n$ consistent to θ_o, which are unique in a region about θ_o f.a.s.l.n.. But also, for suitable functions ψ, there exist other neighbourhoods of the distribution function F_{θ_o}, on which there exists a unique root of (2.8) within a region of the parameter space. This is fundamental to the theory of robustness where the indeterminacy assumed about the generating marginal of $\{F_n\}$ spans not only F, but also neighbourhoods of elements of F. It is also important in that many functional limits of the M-functional presume the existence and uniqueness of the M-functional on neighbourhoods of a parametric distribution F_{θ_o}. Under specific regularity conditions on ψ and the neighbourhood, we show existence and uniqueness of the M-functional simultaneously by a method that adopts the mode of proof in Foutz (1977). He showed existence and uniqueness of a consistent solution of the maximum likelihood equations. The results here extend to dependent sequences and contain the consistency argument for estimators in Markov chains detailed by Foutz and Srivastava (1979). Uniform convergence theory that is of interest in this thesis, gives practical conditions that cover the assumptions of Foutz.

For completeness we list the conditions at the risk of some repetition. The first condition appears slightly esoteric. But suppose for the moment that the distribution F_{θ_o} is known. We wish to illuminate the behaviour of the root of (2.8) that is closest to θ_o in Euclidean norm, on neighbourhoods of F_{θ_o}. This provides a temporary identification of the M-functional. We can later employ a more statistically applicable selection functional that does not presume the knowledge of θ_o but one that will be equivalent in its selection on small enough neighbourhoods of F_{θ_o}, to $\|\theta - \theta_o\|$.

CONDITIONS A

A0 $T[\psi,G]$ is a root of (2.8), if one exists, chosen by the selection functional $f_G(\theta) = \|\theta - \theta_o\|$ (independent of G). $T[\psi,G] = +\infty$ otherwise. $T[\psi,F_{\theta_o}] = \theta_o$.

A1 The $r\times 1$ vector function $\psi(x,\theta)$ is differentiable in θ, and has partial derivatives which are continuous on $R \times D$.

A2 The families of vector functions $\{\psi(\cdot,\theta)\,|\,\theta \in D\}$ and matrix functions $\{\nabla\psi(\cdot,\theta)\,|\,\theta \in D\}$ are bounded above in Euclidean norm by some function g on R, where $g \in L_1(G)$ for all $G \in n(\varepsilon_o,F_{\theta_o})$ and some $\varepsilon_o > 0$.

As a result of A0-1 resulting M-estimates are measurable. From A1-2 observe that vectors $K_G(\theta)$ and matrices

$$M(\theta,G) = \int \nabla\psi(x,\theta)dG(x)$$

are continuous in $\theta \in D$ for each $G \in n(\varepsilon_o,F_{\theta_o})$. Then the function $K_G(\theta)$ has partial derivative $M(\theta,G)$, since interchange of differentiation and integration is availed by A2 and dominated convergence. We will set $M(\theta) = M(\theta,F_{\theta_o})$, which is also continuous in $\theta \in D$.

A3 $M(\theta_o)$ is nonsingular.

A4 Given $\varepsilon > 0$ there exists a $\delta > 0$ so that $G \in n(\delta, F_{\theta_o})$ implies

$$\sup_{\theta \in D} \| \int \psi(x,\theta) dG(x) - \int \psi(x,\theta) dF_{\theta_o}(x) \| < \varepsilon ,$$

and

$$\sup_{\theta \in D} \| \int \nabla\psi(x,\theta) dG(x) - \int \nabla\psi(x,\theta) dG_o(x) \| < \varepsilon .$$

Assumption A3 is often replaced by the stronger property of positive definiteness of $M(\theta_o)$ when (2.5) represents the minimizing equations of some distance $Q_n(\theta)$, but nonsingularity is sufficient for the arguments here. A1 proves necessary for application of the inverse function theorem employed in the uniqueness argument. Assumption A4 can place some restriction on the function ψ, depending on the neighbourhood that is considered. Investigation of the influence curve of Hampel (1974) involves examining the M-functional in starlike neighbourhoods

$$n_x(\varepsilon, F_{\theta_o}) = \{(1-\delta)F_{\theta_o} + \delta\delta_x \mid 0 \leq \delta \leq \varepsilon, \text{ and } \delta_x \text{ is the d.f. determined by the point mass one at the given point } x \in R\} . \qquad (2.14)$$

For neighbourhoods n_x, assumption A4 follows as a result of A1-2. Neighbourhoods

$$n^*(\varepsilon, G_o) = \{G \mid G \in G,\ \sup_{\theta \in D} \| \int \psi(x,\theta) dG(x) - \int \psi(x,\theta) dG_o(x) \| < \varepsilon ,$$
$$\sup_{\theta \in D} \| \int \nabla\psi(x,\theta) dG(x) - \int \nabla\psi(x,\theta) dG_o(x) \| < \varepsilon \},$$

satisfy A4 immediately with δ chosen equal to ε and $G_o = F_{\theta_o}$. Again by Lemma 1.6 and Proposition 1.4 it follows that assumptions A1-2 imply that the empirical distribution function satisfies

$$F_n \in n^*(\varepsilon, G_o) \quad \text{f.a.s.l.n.} . \qquad (2.15)$$

Analogous statements hold for the maximum likelihood estimation of parameters in a Markov process $\ldots\ldots Y_{-1},Y_o,Y_1,\ldots\ldots$, by writing $X_i = (Y_{i-1},Y_i)$ and $\psi(x,\theta) = \nabla p_\theta(x)/p_\theta(x)$. Here $p_\theta(x)$ is the conditional density described in Roussas (1969, P.63).

EXAMPLE 2.3: To illustrate the practicality of conditions A we show A0-3 for the maximum likelihood estimator of location scale of an i.i.d. sequence from a univariate normal population. This is even though the uniqueness and asymptotics may be clearly demonstrated in this instance by other means. Here $\psi(x;\mu,\sigma) = \left(\frac{x-\mu}{\sigma}\ ,\ -1 + \{\frac{x-\mu}{\sigma}\}^2\right)'$ satisfies the restriction placed on ψ by A0. Let (μ_o,σ_o) be the parameter in question, where σ_o may take any positive number. Set $D = \{(\mu,\sigma) \mid \|(\mu,\sigma)' - (\mu_o,\sigma_o)'\| \leq \sigma_o/2\}$. Clearly ψ has continuous partial derivatives on D. Since uniformly on D it is true that

$$\left|\frac{x-\mu}{\sigma}\right| < (2/\sigma_o)(|x - \mu_o| + \sigma_o/2) ,$$

the vector function $\psi(x;\mu,\sigma)$ and matrix function of partial derivatives of ψ are bounded in Euclidean norm by

$$g(x) = \{1 + 4.(2/\sigma_o)(|x - \mu_o| + \sigma_o/2)\}^2 \max(1.,\sigma_o/2) .$$

Then A2 is satisfied if g is integrable with respect to each $G \in n(\varepsilon_o,F_{\theta_o})$. This is true for neighbourhoods n_x, $x \in E$. Since $\det\{M(\mu_o,\sigma_o)\} = 2\sigma_o^{-2} > 0$, A3 holds.

Preliminary results follow.

LEMMA 2.4: Let conditions A hold. Then there is a $\delta_1 > 0$ and an $\varepsilon_1 > 0$ so that for all $\theta \in U_{\delta_1}(\theta_o)$, $G \in n(\varepsilon_1,F_{\theta_o})$ implies the matrix $M(\theta,G)$ is nonsingular.

PROOF: By continuity of the determinant as a function of the elements of a matrix we may choose $\eta > 0$ such that for any matrix A with

$\|A - M(\theta_o)\| < \eta$, $|\det\{A\}| > \frac{1}{2}|\det\{M(\theta_o)\}|$. From the continuity of $M(\theta)$ in θ we may choose $\delta_1 > 0$ so that $\|M(\theta) - M(\theta_o)\| < \eta/2$ whenever $\theta \in U_{\delta_1}(\theta_o)$. Assume $U_{\delta_1} \subset D$. By A4 we can choose ε_1 such that $\|M(\theta,G) - M(\theta)\| < \eta/2$ holds for all $G \in n(\varepsilon_1, F_{\theta_o})$. The lemma is proved by the triangle inequality of norms.

The next result is vital to many limit arguments concerning the implicitly defined M-functional.

LEMMA 2.5: Let conditions A hold. Then given $\kappa > 0$ there exists an $\varepsilon > 0$ such that $G \in n(\varepsilon, F_{\theta_o})$ implies existence of $T[\psi,G] \in U_\kappa(\theta_o)$. Also there exists $\kappa^* > 0$ for which $T[\psi,G]$ is the unique zero of $K_G(\theta)$ in $U_{\kappa^*}(\theta_o)$, and $M(\theta,G)$ is nonsingular on $U_{\kappa^*}(\theta_o)$, for all $G \in n(\varepsilon, F_{\theta_o})$. For any positive null sequence $\{\varepsilon_k\}$ we may take an arbitrary sequence $\{G_k\}$, where $G_k \in n(\varepsilon_k, F_{\theta_o})$, and then

$$\lim_{k\to\infty} T[\psi,G_k] = T[\psi,F_{\theta_o}] \ . \tag{2.16}$$

If $\tilde{T}[\psi,\cdot]$ is any other functional satisfying (2.8) and (2.16) then

$$\tilde{T}[\psi,G_k] = T[\psi,G_k] \quad \text{for all} \quad k \geq k_o \ ,$$

where k_o is independent of the choice of sequence $\{G_k\}$.

PROOF: Write $\lambda = 1/(4.\|M(\theta_o)^{-1}\|)$. By continuity of $M(\theta)$ we may choose an open ball of radius $0 < \kappa^* \leq \min(\delta_1,\kappa)$ so that $\theta \in U_{\kappa^*}(\theta_o)$ implies $\|M(\theta) - M(\theta_o)\| < \lambda/2$. Let ε_1 be given by Lemma 2.4. For $G \in n(\varepsilon_1, F_{\theta_o})$, define $\lambda(G) = 1/(4.\|M(\theta_o,G)^{-1}\|)$. Choose $0 < \varepsilon^* \leq \varepsilon_1$ so that

$$\begin{aligned}\|M(\theta,G) - M(\theta_o,G)\| &\leq \|M(\theta,G) - M(\theta)\| + \|M(\theta_o,G) - M(\theta_o)\| \\ &\quad + \|M(\theta) - M(\theta_o)\| \\ &\leq \lambda < 2.\lambda(G) \ ,\end{aligned}$$

holds uniformly in $\theta \in U_{\kappa^*}(\theta_o)$ for all $G \in n(\varepsilon^*, F_{\theta_o})$. Properties (a) and (b) of the Inverse Function Theorem of Appendix 1 ensure $K_G(\cdot)$ is a one-to-one function from U_{κ^*} onto $K_G(U_{\kappa^*})$, and that the image set contains the open ball of radius $\lambda\kappa^*/2$ about $K_G(\theta_o)$. Choose $0 < \varepsilon' \leq \varepsilon^*$ such that

$$\|K_G(\theta_o) - 0\| < \lambda\kappa^*/2 .$$

Then it is clear that $0 \in K_G(U_{\kappa^*}(\theta_o))$ for all $G \in n(\varepsilon^*, F_{\theta_o})$, and that the image set contains the open ball of radius $\lambda\kappa^*/2$ about $K_G(\theta_o)$. Consider the inverse function

$$K_G^{-1} : K_G(U_{\kappa^*}(\theta_o)) \to U_{\kappa^*}(\theta_o), \text{ for } G \in n(\varepsilon', F_{\theta_o}) .$$

It is well defined whenever $K_G(\theta)$ is one-to-one. Since $0 \in K_G(U_{\kappa^*})$ for $G \in n(\varepsilon', F_{\theta_o})$ we may conclude that with $\varepsilon' = \varepsilon$ there exists a unique root $T[\psi, G]$ of (2.8) in $U_{\kappa^*}(\theta_o)$ whenever $G \in n(\varepsilon, F_{\theta_o})$. Now letting $\{\kappa_i^*\}_{i=1}^{\infty}$ be a positive null sequence for which $\kappa_i^* \leq \kappa^*$, there exists a corresponding sequence of $\{\varepsilon_i'\}$. Since $\{\varepsilon_n\}$ is null there is some $j(i)$ for which $\varepsilon_{j(i)} \leq \varepsilon_i'$, whence $G_{j(i)} \in n(\varepsilon_i', F_{\theta_o})$. Letting

$$T[\psi, G_k] = K_{G_k}^{-1}(0) \cap U_{\kappa^*}(\theta_o) ,$$

we see that

$$\lim_{k\to\infty} T[\psi, G_k] = T[\psi, F_{\theta_o}] .$$

The functional value is the unique root of (2.8) on $U_{\kappa^*}(\theta_o)$ for $k \geq j(1) = k_o$.

The versatility of Lemma 2.5 is immediately apparent. For if $\underset{\sim}{X}$ is a stationary ergodic sequence with marginal distribution F_{θ_o}, it

follows from (2.15), where $G_o = F_{\theta_o}$, that there exists an asymptotically unique consistent root of equations (2.5) to θ_o. This is Theorem 2 of Foutz. For the starlike neighbourhoods of (2.14) we see that contamination of a single point x of the observation space, provided that it is sufficiently small, leads to only small perturbations of the M-functional which is also unique in a region about θ_o.

Employing continuity properties on the asymptotics of the M-functional, for instance in terms of Prokhorov, Kolmogorov, or Lévy distance, does not necessarily make those properties relevant to the M-estimator. The next theorem alleviates this doubt.

THEOREM 2.2:

Let conditions A hold. Denote F_n (random) to be the empirical distribution function from a stationary ergodic sequence with marginal G_o where $G_o \in n(\varepsilon, F_{\theta_o})$, the value ε being given by Lemma 2.5. Then there exists a root $\theta(\psi, F_n)$ of equations (2.5) which has the property

$$\theta(\psi, F_n) \xrightarrow{a.s.} T[\psi, G_o] \ . \tag{2.17}$$

If $\tilde{\theta}(\psi, F_n)$ is any other functional satisfying (2.5) and (2.17)

$$\tilde{\theta}(\psi, F_n) = \theta(\psi, F_n) \quad \text{f.a.s.l.n.} \ .$$

PROOF: By Lemma 2.5, $T[\psi, G_o]$ exists and lies in $U_{\kappa^*}(\theta_o)$, and also $M(T[\psi, G_o], G_o)$ is nonsingular. Analogous statements to those of Lemma 2.5 are true taking neighbourhoods n^* centered at G_o and denoting the selection functional that determines $\theta(\psi, \cdot)$ equal to $\|\theta - T[\psi, G_o]\|$. That is given $\kappa_o > 0$ there exists $0 < \kappa_o^* < \kappa_o$, $0 < \varepsilon_o$, such that $G \in n^*(\varepsilon_o, G_o)$ implies existence of a functional value $\theta(\psi, G) \in U_{\kappa_o^*}(T[\psi, G_o])$ which is unique in that set. Given a null sequence $\{\varepsilon_{ok}\}$, for any sequence $\{G_k\}$, $G_k \in n^*(\varepsilon_{ok}, G_o)$, we have

$$\lim_{k\to\infty} \theta(\psi,G_k) = T[\psi,G_o] \ .$$

The theorem follows from (2.15).

REMARK 2.1: If ψ satisfies conditions A for Prokhorov, Kolmogorov, or Lévy neighbourhoods (determined by the respective metrics), and $T[\psi,G_o]$ is the unique functional such that $\|T[\psi,G_o] - \theta_o\| < \kappa^*$ whenever $G_o \in n(\varepsilon,F_{\theta_o})$, then since $d(F_n,G_o) \xrightarrow{a.s.} 0$ for each metric it follows that $F_n \in n(\varepsilon,F_{\theta_o})$ f.a.s.l.n.. So $\theta(\psi,F_n) = T[\psi,F_n]$ f.a.s.l.n. and a unique solution of (2.8) in $U_{\kappa^*}(\theta_o)$, even though generated by G_o.

The conditions introduced are not difficult to check. Under them the existence of a region about θ_o in the parameter space has been demonstrated for which the M-functional will exist and be unique in a small enough neighbourhood about F_{θ_o}. Moreover given any distribution in that neighbourhood generating the process (1.1) there will exist an asymptotically unique consistent root to the M-functional. If the neighbourhood is determined by a suitable metric the M-estimator will also be unique in the region about θ_o subject to conditions A being satisfied.

Theorems to this point have been local in nature. Indeed uniform convergence need only apply in a local region about some parameter θ_o^*, in order that a consistency result be attained. But what is sought for are theorems of a global nature which, in practice are more useful.

LEMMA 2.6: Let $f_n(\underset{\sim}{X}^{(n)}, \theta) \xrightarrow{a.s.} f_{G_o}(\theta)$ uniformly in $\theta \in \Theta$. Suppose there exists a θ_o^* so that for every neighbourhood N of θ_o^*

$$\inf_{\theta \notin N} f_{G_o}(\theta) - f_{G_o}(\theta_o^*) > 0 \ .$$

Define $H_n(\psi) = \{\theta \in \Theta \,|\, K_n(\theta) = 0\}$. Suppose further there exists a region

$U_{\kappa^*}(\theta_o^*)$, $\kappa^* > 0$, so that $H_n(\psi) \cap U_{\kappa^*}(\theta_o^*)$ consists of a single point f.a.s.l.n.. Assume there exists a sequence $\{\theta_n \in H_n(\psi)\}$ so that $\theta_n \xrightarrow{a.s.} \theta_o^*$. Then,

$$\inf_{\theta \in H_n(\psi)} f_n(\underset{\sim}{X}^{(n)}, \theta) = f_n(\underset{\sim}{X}^{(n)}, \theta_n) \quad \text{f.a.s.l.n..}$$

PROOF: Let $\varepsilon(\kappa^*) = \inf_{\theta \notin U_{\kappa^*}(\theta_o^*)} f_{G_o}(\theta) - f_{G_o}(\theta_o^*)$. Choose $0 < \delta_1 < \kappa^*$ so that

$$|f_{G_o}(\theta) - f_{G_o}(\theta_o^*)| < \varepsilon(\kappa^*)/2 \qquad \theta \in U_{\delta_1}(\theta_o^*) .$$

Since $\theta_n \in U_{\delta_1}(\theta_o^*)$ and $|f_n(\underset{\sim}{X}^{(n)}, \theta) - f_{G_o}(\theta)| < \varepsilon(\kappa^*)/4$ uniformly in $\theta \in \Theta$ f.a.s.l.n.,

$$\begin{aligned} f_n(\underset{\sim}{X}^{(n)}, \theta_n) &< f_{G_o}(\theta_n) + \varepsilon(\kappa^*)/4 \\ &< f_{G_o}(\theta_o^*) + (3/4)\varepsilon(\kappa^*) \\ &< f_{G_o}(\theta) - \varepsilon(\kappa^*)/4 \quad \text{uniformly in } \theta \notin N = U_{\kappa^*}(\theta_o^*) \\ &< f_n(\underset{\sim}{X}^{(n)}, \theta) \qquad \text{uniformly in } \theta \notin N . \end{aligned}$$

Hence

$$\inf_{\theta \in H_n(\psi)} f_n(\underset{\sim}{X}^{(n)}, \theta) = f_n(\underset{\sim}{X}^{(n)}, \theta_n) \quad \text{f.a.s.l.n.} .$$

If $f_n(\underset{\sim}{X}^{(n)}, \theta) = f_{F_n}(\theta)$, the functional $T[\psi, f, \cdot]$ leads to a measurable estimator that is consistent to $T[\psi, f, G_o] = \theta_o^*$. The statement of consistency is with respect to an underlying probability measure P which induces G_o. When $G_o = F_{\theta_o}$ and $\theta_o^* = \theta_o$ we may assert the estimation procedure is "globally consistent". With identifiability of the parametric family F it is desirable to have this property for all $\theta_o \in \Theta$.

The separation requirement of the Lemma can be shown to be satisfied

for a number of the L_2 norms of the Kolmogorov-Smirnov and Cramer Von Mises type. Pollard (1980) studies statistics of the type $\|F_n - F_\theta\|$, where distributions F_θ take values in $\mathfrak{X}$ where $(\mathfrak{X}, \|\cdot\|)$ is a normed linear space. Moreover the maps $\theta \to F_\theta$ are assumed continuous. Here we have $\mathfrak{X} = E$. Since $F_{\theta_1} \neq F_{\theta_2}$ whenever $\theta_1 \neq \theta_2$, it suffices to show there exists *at least one* compact neighbourhood N_o of θ_o for which

$$\inf_{\theta \notin N_o} \|F_\theta - F_{\theta_o}\| > 0 .$$

For then, given an N, the continuous function $\theta \mapsto \|F_\theta - F_{\theta_o}\|$ must be bounded away from zero on the compact set $N_o - \text{int } N$, and hence also on the set $(\Theta - N_o) \cup (N_o - \text{int } N) \supseteq \Theta - N$. These investigations are related to the separation arguments of Wolfowitz (1957) who considered strong consistency of minimal distance estimators. Other selection statistics yielding globally consistent estimators are based on (2.7).

§2.4 Global Consistency and Uniqueness of the Univariate M-functional

When the parameter space is a subset of the real line, arguments peculiar to the real line can be employed to prove existence of an asymptotically unique root of equations (2.5). Continuous differentiability of ψ can be relaxed. Moreover a selection statistic based on the expected slope of the influence function is possible. This obviates the need for searching for the function $\ell(x,\theta)$ of (2.7), which may not even exist if ψ is constructed for its local properties.

Small violations in differentiability of an influence function ψ can be overcome by considering $\{(\partial/\partial\theta)^- \psi(x,\theta) \mid \theta \in \Theta \subset E\}$, where $(\partial/\partial\theta)^-$ denotes left differentiation, which is often well defined on $R \times \Theta$. Individual investigation frequently shows

$$(d/d\theta)^{-}K_n(\theta) \xrightarrow{a.s.} \nabla K_{G_o}(\theta) \quad \text{uniformly in} \quad \theta \in D \ , \tag{2.18}$$

where $K_{G_o}(\theta)$ is continuously differentiable and

$$\nabla K_{G_o}(\theta) = \int (\partial/\partial\theta)^{-}\psi(x,\theta)dG_o(x) \ .$$

LEMMA 2.7: Let θ_o^*, G_o, and ψ be so that B1, C2 and C3 hold, and that (2.18) is also satisfied. Then there exists an open ball $U_\delta(\theta_o^*)$, $\delta > 0$, such that

(a) there exists a sequence $\{\theta_n(F_n)\}$ of zeros of $\{K_n(\theta)\}$ within $U_\delta(\theta_o^*)$, and

(b) for any other sequence $\{\theta(F_n)\}$ of zeros of $\{K_n(\theta)\}$ consistent for θ_o^*, $\tilde{\theta}_n(F_n) = \theta_n(F_n)$ f.a.s.l.n..

PROOF: Abbreviate $\lambda_o = \nabla K_{G_o}(\theta_o^*)$. By continuity choose the ball $U_\delta(\theta_o^*) \subset D$ so that $\nabla K_{G_o}(\theta) > \lambda_o/2$ for $\theta \in U_\delta(\theta_o^*)$. By (2.18)

$$(d/d\theta)^{-}K_n(\theta) > \lambda_o/4 \quad \text{uniformly in} \quad \theta \in \bar{U}_\delta(\theta_o^*), \text{ f.a.s.l.n.} \tag{2.19}$$

By the uniform convergence indicated by (2.12), and differentiability of $K_{G_o}(\theta)$,

$$K_n(\theta_o^*-\delta) < K_{G_o}(\theta_o^*-\delta) + \lambda_o\delta/2 < 0 < K_{G_o}(\theta_o^*+\delta) - \lambda_o\delta/2 < K_n(\theta_o^*+\delta) \quad \text{f.a.s.l.n.} \tag{2.20}$$

By continuity of $K_n(\theta)$ there exists $\theta_n \in U_\delta(\theta_o^*)$ f.a.s.l.n., and by strict monotonicity, (2.19), it is unique within U_δ f.a.s.l.n. . Fixing $\delta > 0$ and taking $0 < \delta_1 < \delta$ arbitrarily small we see that $\theta(F_n) \in U_{\delta_1}$ and is unique in U_δ f.a.s.l.n.. That is the functional values $\theta_n(F_n)$, defined by selection function $|\theta - \theta_o^*|$ are asymptotically unique and satisfy $\theta(F_n) \xrightarrow{a.s.} \theta(G)$. This completes the proof.

Then to each nonstationary zero θ_o^* of $K_{G_o}(\theta)$ there exists an asymptotically unique consistent sequence of roots of equations (2.5). They are identified as those closest to θ_o^*, the minimum if there occur two equidistant from θ_o^*. Not knowing G_o this belies the statistical estimation procedure. But with some mild conditions on ψ the estimation can be resolved when $G_o = F_{\theta_o}$. Let

$$H(\psi,F_{\theta_o}) = \{\theta = \theta(\psi,F_{\theta_o}) \,|\, K_{F_{\theta_o}}(\theta) = 0,\ \nabla K_{F_{\theta_o}}(\theta) \neq 0\ ,\ \theta \in \Theta\}\ .$$

A requirement for estimation is that $\theta_o \in H(\psi,F_{\theta_o})$ for all $\theta_o \in \Theta$. For a global consistency argument we consider the following assumptions

(a) $H(\psi,F_\theta)$ is finite $(= \theta_o,.....,\theta_N$ say).

(b) $\theta \in H(\psi,F_{\theta_o})$, $\theta \neq \theta_o$ implies $|\nabla K_{F_{\theta_o}}(\theta) - \nabla K_{F_\theta}(\theta)| > \xi(\theta_o) > 0$, for some $\xi(\theta_o)$.

(c) $|\nabla K_{F_{\theta_o}}(\theta_o)| \geq \ell$ for some $\ell > 0$.

By continuity there exists balls $U_{\delta_i}(\theta_o)$, $\delta_i > 0$ where for each $\theta \in U_{\delta_i}(\theta_i)$ it is true that

$$|\nabla K_{F_{\theta_o}}(\theta)| > \tfrac{1}{2}|\nabla K_{F_{\theta_o}}(\theta_i)|\ .$$

Assume further that,

(d) there exists $\varepsilon = \varepsilon(\ell)$ such that if

$$\theta \in \tau(\varepsilon,\psi,F_{\theta_o}) = \{\theta \,|\, |K_{F_{\theta_o}}(\theta)| < \varepsilon \quad \theta \in \Theta - \bigcup_{i=1}^{N} U_{\delta_i}(\theta_i)\}$$

implies $\nabla|K_{F_{\theta_o}}(\theta)| < s.\ell$ for some fixed $0 < s < \ell$, and

(e) $\nabla K_{F_\theta}(\theta)$ is a continuous function in $\theta \in \Theta$.

The motivation for assumption (d) is to allow for possible zeros of $K_{F_{\theta_o}}(\theta)$ which may be stationary points of that function. Or more importantly if $\lim_{\theta\to\infty} K_{F_{\theta_o}}(\theta) = 0$ it often occurs that in finite samples and for large θ there exists roots $\theta_n(F_n)$ of equations (2.5), yet the $\{\theta_n(F_n)\}$ need not correspond to an estimator sequence consistent to an element of $H(\psi,F_{\theta_o})$.

THEOREM 2.3:

Let $\theta_o^* = \theta_o$, $G_o = F_{\theta_o}$, and ψ be so that B1, C2, and C3 hold with $D = \Theta$. Assume (a)-(e) hold and (2.18) holds when $D = \Theta$. Abbreviate $(d/d\theta)^- K_n(\theta) = K_n'(\theta)$ and let

$$H_n(\psi,s.\ell) = \{\theta | K_n(\theta) = 0,\ |K_n'(\theta)| \geq s.\ell\ ,\ \theta \in \Theta\}\ ,$$

where ℓ is a known constant of (c) and $0 \leq s < 1$ is some constant. Define $\{T[\psi,F_n]\}$ to be that sequence of zeros of $\{K_n(\theta)\}$ if they exist (if not define $T[\psi,F_n] = +\infty$) which satisfy

$$\min_{\theta\in H_n(\psi,s.\ell)} |K_n'(\theta)-K_{F_\theta}'(\theta)| = |K_n'(T[\psi,F_n])-K_{F_{T[\psi,F_n]}}'(T[\psi,F_n])| \tag{2.21}$$

(let $T[\psi,F_n]$ be the least if there exists more than one solution.) Then

$$T[\psi,F_n] \xrightarrow{a.s.} \theta_o\ .$$

PROOF: Firstly observe from (2.12) and (2.18), with $D = \Theta$ and $G_o = F_{\theta_o}$,

$$H_n(\psi,s.\ell) \cap \{\Theta - \bigcup_{i=1}^{N} U_{\delta_i}(\theta_i)\} = \phi \quad \text{f.a.s.1.n.}\ .$$

By Lemma 2.7 there exist asymptotically unique sequences $\{\theta_{ni}(F_n)\}$

such that $\theta_{ni}(F_n) = H_n(\psi,F_n) \cap U_{\delta_i}(\theta_i) = \theta_{ni}(F_n) \xrightarrow{a.s.} \theta_i$, $i = 0,\ldots,N$.

Then by (2.12), continuity of $K_{F_{\theta_o}}(\theta)$ and (e) respectively

$$|K_n'(\theta_{ni}(F_n)) - K_{F_{\theta_o}}'(\theta_{ni}(F_n))|, \quad |K_{F_{\theta_o}}'(\theta_{ni}(F_n)) - K_{F_{\theta_o}}'(\theta_i)|,$$

$$|K_{F_{\theta_i}}'(\theta_i) - K_{F_{\theta_{ni}(F_n)}}'(\theta_{ni}(F_n))|,$$

are all less than $\xi(\theta_o)/6$ f.a.s.l.n., $i = 0,\ldots,N$, which by (b) implies

$$|K_n'(\theta_{ni}(F_n)) - K_{F_{\theta_{ni}(F_n)}}'(\theta_{ni}(F_n))| > \xi(\theta_o)/6 \quad \text{f.a.s.l.n.}$$

$$i = 1,\ldots,N.$$

Finally

$$|K_n'(\theta_{no}(F_n))| \geq s.\ell \quad \text{f.a.s.l.n., for any } 0 < s < 1 ,$$

whereby the lemma is proved since

$$|K_n'(\theta_{no}(F_n)) - K_{F_{\theta_{no}(F_n)}}'(\theta_{no}(F_n))| < \xi(\theta_o)/6 \quad \text{f.a.s.l.n.} .$$

If a lower bound, ℓ, is not known to exist uniformly in $\theta_o \in \Theta$ it may be true that there is an $\varepsilon > 0$ for which $\tau(\varepsilon,\psi,F_{\theta_o}) = \phi$. In that case Theorem 2.3 remains valid replacing $H_n(\psi,s.\ell)$ by $H(\psi,F_n)$ in (2.21). If a lower bound were not to exist and yet $\tau(\varepsilon,\psi,F_{\theta_o}) \neq \phi$ for all $\varepsilon > 0$ then a natural assumption in place of (d) is:

(f) There is an $\varepsilon > 0$ so that $\theta \in \tau(\varepsilon,\psi,F_{\theta_o})$ implies

$$|K_{F_{\theta_o}}'(\theta) - K_{F_\theta}'(\theta)| > \xi(\theta_o)/2 .$$

Then any solution $\hat{\theta} \in \tau(\varepsilon,\psi,F_{\theta_o})$ will be so that

$$|K_n'(\hat{\theta}) - K_{F_{\hat{\theta}}}'(\hat{\theta})| > \xi(\theta_o)/6 \quad \text{f.a.s.l.n.} .$$

Hence the root $\theta_{no}(F_n)$ will minimize the selection statistic

$f_n(\underset{\sim}{X}^{(n)},\theta) = |K_n'(\theta) - K_{F_\theta}'(\theta)|$ over all such roots. The particular assumptions imposed then will depend on the individual family F and influence function ψ. Note that uniform convergence over the whole of the parameter space is necessary for the Theorem.

Many authors prove existence of a consistent sequence of solutions of (2.5) but few also deal with uniqueness of the estimator. Very weak conditions were given in the consistency case B of Huber's (1967) monograph, that were sufficient for showing almost sure convergence of estimators $T_n : R^n \to \Theta$ that satisfied

$$\frac{1}{n} \sum_{i=1}^{n} \psi(X_i, T_n) \xrightarrow{a.s.} 0 \ . \tag{2.22}$$

For clarity we make a comparison.

Assumptions (Huber 1967)

(B-1) For fixed $\theta \in \Theta$, $\psi(x,\theta)$ is $\mathcal{B}$ measurable, and $\psi(x,\theta)$ is separable (see Huber 1967, P.222, (A-1)).

(B-2) The function ψ is a.s. continuous in θ:

$$\lim_{\theta' \to \theta} |\psi(x,\theta') - \psi(x,\theta)| = 0 \quad \text{a.s.}$$

(B-3) The expected value $K_{G_o}(\theta) = E_{G_o}\psi(X,\theta)$ exists for all $\theta \in \Theta$, and has a unique zero at $\theta = \theta_o$.

(B-4) There exists a continuous function which is bounded away from zero, $b(\theta) \geq b_o > 0$, such that

(i) $\sup_\theta \dfrac{|\psi(x,\theta)|}{b(\theta)}$ is integrable

(ii) $\liminf_{\theta \to \infty} \dfrac{|K_{G_o}(\theta)|}{b(\theta)} \geq 1$

(iii) $E_{G_o}\{\limsup_{\theta \to \infty} \dfrac{|\psi(x,\theta) - K_{G_o}(\theta)|}{b(\theta)} < 1\}$.

Relinquishing the requirement that T_n be a root of equations (2.5) allows the relaxation of the continuity assumption slightly in (B-2). Theorem 2.3 relaxes condition (B-3) and in particular (B-4)(ii). The assumption (B-4)(ii) bounds the curve $K_{G_o}(\theta)$ away from zero outside some compact set. At the model F_{θ_o} this would ensure $\tau(\varepsilon,\psi,F_{\theta_o})$ was empty for some $\varepsilon > 0$. But for instance consider the common case of estimating the location of a symmetric distribution, G_o say. Assume symmetry about the origin for instance. For those continuous odd psi-functions that are zero outside the compact set $[-c,c]$, $\psi(x-\theta)$ is zero whenever x lies in $(-\infty, -c+\theta] \cup [c+\theta, \infty)$. Clearly

$$\lim_{\theta\to\infty} K_{G_o}(\theta) = \lim_{\theta\to\infty} \int_{-c+\theta}^{c+\theta} \psi(x-\theta)dG_o(x) = 0 ,$$

which violates (B-4)(ii). But either assumption (d) or (f) is quite plausible since there will exist some set $U_\delta(0)$ so that $|K'_{G_o}(\theta)| < |K'_{G_o}(0)|$ uniformly in $\theta \in E - U_\delta(0)$.

Huber proves existence of a consistent root. We give a proof that identifies the estimator uniquely for all n.

CHAPTER 3

LIMIT THEOREMS FOR M-ESTIMATORS

§3.1 Asymptotic Normality of the Univariate M-Estimator

In this chapter we use uniform convergence theory to obtain limit theorems for suitably normed M-estimators. We consider i.i.d. sequences and limit theorems for dependent sequences will be regarded as peripheral to this thesis. The only change required in the proofs is to apply the appropriate classical law to the sums of random variables.

Typically proofs of the C.L.T. (central limit theorem) for implicitly defined estimators use Taylor expansions. For the M-estimator this requires that ψ be at least continuously differentiable in θ for each x in the observation space. It is often not possible to escape situations where apparently "slight" violations of this assumption occur. Portnoy (1977) gave a proof that showed "sharp corners" do not affect the asymptotic distribution of the M-estimator of location on the real line. The following theorem gives an extension of this result to the M-estimator of a general univariate parameter.

THEOREM 3.1:

Let G_o be the common distribution function of an i.i.d. sequence $\underset{\sim}{X}$, and suppose θ_o^*, G_o, and ψ are such that B1 of §2.2 and C2, C3 of §2.3 hold. Assume also that $\Theta \subset E$ and (2.18) holds. Denote

$$\lambda_o = K'_{G_o}(\theta_o^*) = \int (\partial/\partial\theta)^{-}\psi(x,\theta)dG_o(x)\Big|_{\theta_o^*} \neq 0 .$$

Let $\{\psi_m\}$ be a sequence of functions that are continuously differentiable in θ for all $x \in R$, and suppose there exists sets $\{\delta_m(\theta)\}$, $\delta_m(\theta) \subset R$, that belong to a Glivenko-Cantelli class of G_o, so that there exists a

neighbourhood $W \subset D$ of θ_o^* for which the following hold:

(i) $\psi_m(x,\theta) = \psi(x,\theta) \qquad x \in R - \delta_m(\theta),\ \theta \in \Theta$

(ii) $p_m = \sup_{\theta \in W} \int_{\delta_m(\theta)} dG_o(x) \to 0 \quad \text{as} \quad m \to \infty$

(iii) $\beta_m = \sup_{\theta \in W} \sup_{x \in R} |\psi_m(x,\theta) - \psi(x,\theta)| = o(m^{-\frac{1}{2}})$, and

(iv) $\beta'_m = \sup_{\theta \in W} \sup_{x \in R} |(\partial/\partial\theta)^{-}\psi(x,\theta) - (\partial/\partial\theta)\psi_m(x,\theta)| = O(1).$

Then if $\psi(X,\theta_o^*)$ has a finite second moment, there exists on asymptotically unique consistent sequence of zeros $\{\theta(F_n)\}$ of $\{K_n(\theta)\}$, the latter given in (2.5), such that $\sqrt{n}(\theta(F_n) - \theta_o^*) \xrightarrow{D} N\{0,\sigma^2(\psi,G_o,\theta_o^*)\}$, the normal random variable with zero mean and variance given by

$$\sigma^2(\psi,G,\theta) = \{\text{var}_G\ \psi(X,\theta)\}/\lambda_o^2(G) \ . \tag{3.1}$$

PROOF: Asymptotic uniqueness of $\{\theta(F_n)\}$ follows from Lemma 2.7. Let $K_n^{(m)}(\theta)$ and $K_{G_o}^{(m)}(\theta)$ be the quantities corresponding to K_n and K_{G_o} when ψ is replaced by ψ_m. If $m(n) = n$ then $|K_n^{(m)}(\theta) - K_n(\theta)| \leq \beta_n$, and so by (2.12)

$$K_n^{(m)}(\theta_o^*-\delta) < K_{G_o}(\theta_o^*-\delta) + \lambda_o\delta/2 < 0 < K_{G_o}(\theta_o^*+\delta) - \lambda_o\delta/2 < K_n^{(m)}(\theta_o^*+\delta) \ ,$$

f.a.s.l.n., when $m(n) = n$ and $U_\delta(\theta_o^*) \subset W$. Label zeros of $K_n^{(m)}(\theta)$ and $K_{G_o}^{(m)}(\theta)$ uniquely defined by the selection statistic $|\theta - \theta_o^*|$, as $\theta_m(F_n)$ and $\theta_m(G_o)$ respectively. They exist within any given neighbourhood of θ_o^* f.a.s.l.n.. Consider the expansion

$$\sqrt{n}(\theta(F_n)-\theta_o^*) = \sqrt{n}(\theta(F_n)-\theta_m(F_n)) + \sqrt{n}(\theta_m(F_n)-\theta_m(G_o)) + \sqrt{n}(\theta_m(G_o)-\theta_o^*) \ .$$

The required result is obtained by letting $m(n) = n$, and considering the three terms on the right separately. We show the first and third terms to be asymptotically negligible and the limiting distribution of

the second term, and thus that of $\sqrt{n}(\theta(F_n) - \theta_o^*)$, is normal $N\{0,\sigma^2(\psi,G_o,\theta_o^*)\}$. This would be expected if ψ were continuously differentiable.

Let $Y_{nm}(\theta) = n\int_{\delta_m(\theta)} dF_n(x)$. By the Glivenko-Cantelli assumption $\sup_m \sup_{\theta\in W} |n^{-1}Y_{nm}(\theta) - p_m(\theta)| \xrightarrow{a.s.} 0$. Then

$$\begin{aligned}|K'_{G_o}(\theta)-K_n^{(m)'}(\theta)| &\le |K'_{G_o}(\theta)-(d/d\theta)^{-}K_n(\theta)| \\ &\quad + \left|\frac{1}{n}\sum_{i=1}^{n}(\partial/\partial\theta)^{-}\{\psi(X_i,\theta)-\psi_m(X_i,\theta)\}\right| \\ &\le |K'_{G_o}(\theta)-(d/d\theta)^{-}K_n(\theta)| + n^{-1}Y_{nm}(\theta)\beta'_m \\ &\xrightarrow{a.s.} 0 \quad \text{uniformly in } \theta\in W \text{ as } m(n)=n\to\infty. \qquad (3.2)\end{aligned}$$

Let $U \subset W$ be an open ball containing θ_o^* with the property that for some $\xi > 0$, $K'_{G_o}(\theta) > \xi$ for all $\theta \in U$. Then (3.2) implies that

$$K_n^{(m)'}(\theta) > \xi/2 \quad \text{uniformly in } \theta\in U \quad \text{f.a.s.l.n.} \qquad (3.3)$$

Term 1 may be written,

$$\sqrt{n}\{\theta(F_n)-\theta_m(F_n)\} = \{K_n^{(m)'}(\hat\theta_{nm})\}^{-1}\sqrt{n}\,K_n^{(m)}(\theta(F_n))$$

where $\hat\theta_{nm}$ lies between $\theta(F_n)$ and $\theta_m(F_n)$. Now,

$$\begin{aligned}|\sqrt{n}\,K_n^{(m)}(\theta(F_n))| &= |\sqrt{n}\{K_n^{(m)}(\theta(F_n)) - K_n(\theta(F_n))\}| \\ &= \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\psi_m(X_i,\theta(F_n)) - \psi(X_i,\theta(F_n))\}\right| \\ &\le n^{-\frac{1}{2}}Y_{nm}(\theta(F_n))\beta_m \\ &\xrightarrow{a.s.} 0 .\end{aligned}$$

Also from (3.3), $K_n^{(m)'}(\hat\theta_{nm}) > \xi/2$ f.a.s.l.n., whence

$$\sqrt{n}(\theta(F_n) - \theta_m(F_n)) \xrightarrow{a.s.} 0 .$$

Term 3 may be written

$$\sqrt{n}(\theta_m(G_o)-\theta_o^*) = -\{K_{G_o}^{(m)'}(\tilde{\theta}_{nm})\}^{-1}\sqrt{n}\,K_{G_o}^{(m)}(\theta_o^*)$$

$$= -\{K_{G_o}^{(m)'}(\tilde{\theta}_{nm})\}^{-1}\{\int \sqrt{n}[\psi_m(x,\theta_o^*) - \psi(x,\theta_o^*)]dG_o(x)\} .$$

Since the sequence $\sqrt{n}[\psi_m(x,\theta_o^*) - \psi(x,\theta_o^*)]$ is dominated by an integrable variable and tends to zero a.s., and $K_G^{(m)'}(\theta) > \xi/2$ uniformly in $\theta \in U$ as $m(n) = n \to \infty$, then

$$\sqrt{n}(\theta_m(G_o)-\theta_o^*) = -\{K_n^{(m)'}(\theta_{nm}^*)\}^{-1}\sqrt{n}\,K_n^{(m)}(\theta_o^*) .$$

Here θ_{nm}^* lies between $\theta_m(F_n)$ and $\theta_m(G_o)$. Now as both $\theta_m(F_n)$, $\theta_m(G_o)$ converge a.s. to θ_o^*, from (3.2) and since a.s. convergence apriori implies convergence in probability

$$K_n^{(m)'}(\theta_{nm}^*) - K_{G_o}'(\theta_o^*) \xrightarrow{P} 0 . \tag{3.4}$$

Further,

$$\sqrt{n}\,K_n^{(m)}(\theta_m(G_o)) = \sqrt{n}\,K_n^{(m)}(\theta_o^*) + \sqrt{n}\,K_n^{(m)'}(\bar{\theta}_{nm})(\theta_m(G_o)-\theta_o^*)$$

$$= \sqrt{n}\,K_n^{(m)}(\theta_o^*) + o_p(1) . \tag{3.5}$$

Now

$$\sqrt{n}|K_n^{(m)}(\theta_o^*) - K_n(\theta_o^*)| = \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\psi_m(X_i,\theta_o^*) - \psi(X_i,\theta_o^*)\}\right|$$

$$\leq n^{-\frac{1}{2}}Y_{nm}(\theta_o^*)\beta_m$$

$$\xrightarrow{p} 0 \quad \text{as} \quad m(n) = n \to \infty .$$

Then since

$$\sqrt{n}\,K_n(\theta_o^*) \xrightarrow{D} N\{0,\mathrm{var}_{G_o}\,\psi(X,\theta_o^*)\} ,$$

this implies from (3.5) that

$$\sqrt{n}\,K_n^{(m)}(\theta_o^*) \xrightarrow{D} N\{0,\mathrm{var}_{G_o}\,\psi(X,\theta_o^*)\} .$$

From (3.4) and the above convergence, using Slutsky's theorem detailed in Cramér (1946, §20.6)

$$\sqrt{n}(\theta_m(F_n) - \theta_m(G_o)) \xrightarrow{D} N\{0,\sigma^2(\psi,G_o,\theta_o^*)\} ,$$

and the proof is complete.

The theorem avoids the necessity for continuous partial derivatives of ψ, and the local nature of the argument makes it applicable to non-monotonic (in the parameters) influence functions that have "sharp corners". Huber (1977) sketches a proof of Hampel (1973) giving asymptotic normality of a psuedo solution of the estimating equations under very weak conditions on a ψ that is monotonic in θ for each $x \in R$.

The classical basis for inference about a parameter estimate is the notion that $G_o \in F = \{F_\theta | \theta \in \Theta\}$. Moreover the convergence in the central limit theorem should be uniform in the underlying distribution. Even if the first should be true, to establish the latter it is initially required to show uniform consistency.

LEMMA 3.1: Let ψ be continuous on $R \times \Theta$, where $\Theta \subset E$. Write $K_n^{\theta_o}(\theta)$ to be $K_n(\theta)$ when the i.i.d. sequence $\underset{\sim}{X}(\cdot) = \underset{\sim}{X}(\theta_o,\cdot)$ is generated by F_{θ_o}. Assume $K_{F_{\theta_o}}(\theta_o) = 0$ for every $\theta_o \in \Theta$, and $K_{F_{\theta_o}}(\theta)$ is continuously differentiable in θ for each $\theta_o \in \Theta$. Then if

(1) $K_n^{\theta_o}(\theta) - K_{F_{\theta_o}}(\theta) \xrightarrow{a.s.} 0$ uniformly in $\theta \in \Theta$, p-uniformly in $\theta_o \in \Theta$; and

(2) there exists a $\delta > 0$, $\lambda > 0$ such that

$$\nabla K_{F_{\theta_o}}(\theta) > \lambda \quad \text{for} \quad \theta_o - \delta < \theta < \theta_o + \delta \quad \text{uniformly in} \quad \theta_o \in \Theta;$$

then there exists a measurable sequence $\{T_n[\underset{\sim}{X}^{(n)}(\theta_o)]\}$

such that $T_n[\underset{\sim}{X}^{(n)}(\theta_o)] - \theta_o \xrightarrow{a.s.} 0$ p-uniformly in $\theta_o \in \Theta$.

PROOF: Given arbitrary $\varepsilon > 0$, let $\delta^* = \min(\varepsilon, \delta)$. Since

$$K_n^{\theta_o}(\theta_o - \delta^*) < K_{F_{\theta_o}}(\theta_o - \delta^*) + \lambda\delta^* < 0 < K_{F_{\theta_o}}(\theta_o + \delta^*) - \lambda\delta^* < K_n^{\theta_o}(\theta_o + \delta^*)$$

holds p-uniformly in $\theta \in \Theta$, there exists a measurable root of $K_n^{\theta_o}(\theta) = 0$, within radius ε of θ_o. Define $T_n[\underset{\sim}{X}^{(n)}(\theta_o)]$ to be the root closest to θ_o, the least if two are equidistant. The proof is then complete.

Conditions on ψ sufficient for (1) are found by examining Theorem 1.2. It is not necessary to give the broadest possible conditions here. We merely point out that a basis for the result is the uniform S.L.L.N.. Asymptotic uniqueness of the root may be approached through p-uniform convergence of the partial derivatives. This assumption can also be used to establish the uniform C.L.T.

THEOREM 3.2:

Let $\psi(x,\theta)$ be continuously differentiable so that variances $\sigma^2(\psi, F_{\theta_o}, \theta_o)$ are convergent uniformly in $\theta \in \Theta$ and bounded above and below uniformly. Assume (1), (2) of Lemma 3.1, and

(3) there exists a $\delta > 0$ such that

$$|\nabla K_{F_{\theta_o}}(\theta_1) - \nabla K_{F_{\theta_o}}(\theta_o)| \leq M_\delta |\theta_1 - \theta_o|$$

for all $\theta_o - \delta < \theta_1 < \theta_o + \delta$ uniformly in $\theta_o \in \Theta$.

(4) $\nabla K_n^{\theta_o}(\theta) \xrightarrow{a.s.} \nabla K_{F_{\theta_o}}(\theta)$ uniformly in $\theta \in \Theta$ p-uniformly in $\theta_o \in \Theta$.

Then uniformly in $z \in E$ and $\theta_o \in \Theta$, there exists a sequence $\{T_n[\underset{\sim}{X}^{(n)}(\theta_o)]\}$ of roots of $K_n^{\theta_o}(\theta) = 0$ for which

$$P_{\theta_o}\{\omega \mid \sqrt{n}(T_n[\underset{\sim}{X}^{(n)}(\theta_o)] - \theta_o) \leq z\} \to \Phi(z/\sigma(\psi, F_{\theta_o}, \theta_o)) \ .$$

PROOF: Let $\{T_n[\underset{\sim}{X}^{(n)}(\theta_o)]\}$ be the consistent sequence of roots of $K_n^{\theta_o}(\theta) = 0$. Expanding using the mean value theorem

$$\sqrt{n}(T_n - \theta_o) = - \frac{\sqrt{n}\, K_n^{\theta_o}(\theta_o)}{\nabla K_n^{\theta_o}(\xi_n)} \ ,$$

where ξ_n lies between T_n and θ_o and so $\xi_n \xrightarrow{a.s.} \theta_o$ p-uniformly in $\theta_o \in \Theta$. This, together with (3) and (4) gives that

$$\nabla K_n^{\theta_o}(\xi_n) \xrightarrow{a.s.} \nabla K_{F_{\theta_o}}(\theta_o) \quad \text{p-uniformly in } \theta_o \in \Theta \ .$$

The numerator, $\sqrt{n}\, K_n^{\theta_o}(\theta_o)$, is a normed sum of i.i.d. random variables with zero means and variances $\sigma^2(\psi, F_{\theta_o}, \theta_o)$. So by Proposition 1.5 its distribution tends to $\Phi(z/\sqrt{\text{var}_{\theta_o}\, \psi(X, \theta_o)})$ uniformly in $\theta_o \in \Theta$. The conclusion follows from the uniform analogue of Slutsky's theorem (Parzen, 1955, P.48, Theorem 18D).

Restricting uncertainty to a parametric family F ignores the general trend of robustness theory where relatively wide departures from an underlying F_{θ_o} are considered. Uniform convergence in the underlying parameter with indeterminacy restricted to F would appear to be a necessity for justifying inferences made from the asymptotic distribution. But such a result can be illusory in practice since the underlying model is inevitably not in F. A remedy is to consider an influence function that redescends to zero inside some compact set, is zero on the complement of that set and is so that conditions A hold. Then by Theorem 1.2 the equivalent of (1) and (4) hold in neighbourhoods of an F_{θ_o}. Conditions similar to (2) and (3) may be shown in small enough

neighbourhoods of an F_{θ_o}. Whether these results are uniform over the whole parameter space or just on compacts of it generates further complications.

§3.2 The Law of the Iterated Logarithm

Limit theorems for estimators are not restricted to the central limit theorem. A law of the iterated logarithm for the M-estimator of location was shown by Boos (1977) and Boos and Serfling (1980). Under suitable regularity conditions this may be extended to the general M-estimator of a univariate parameter. We extend the conditions C.

C4 The partial derivatives of ψ exist and are continuous on $R \times D$.

C5 There exists a continuous function $g \in L_1(G_o)$, so that

$$\|\nabla\psi(x,\theta)\| < g(x) \quad \text{for all} \quad x \in R,\ \theta \in D\ .$$

THEOREM 3.3:

Let $\Theta \subset E$, $\underset{\sim}{X}$ be an i.i.d. sequence generated by G_o and assume θ_o^* and ψ are so that B1 of §2.2, C2 and C3 of §2.3 and C4, C5 hold. Further let $\psi(X,\theta_o^*) \in L_2(G_o)$. Then there exists an asymptotically unique sequence $\{\theta(F_n)\}$ of zeros of $\{K_n(\theta)\}$ defined by selection statistic $|\theta - \theta_o^*|$, such that

$$P\{\omega \mid \limsup \frac{\sqrt{n}\ \lambda_o\{\theta(F_n) - \theta_o^*\}}{\sqrt{2\sigma^2\ \ell n(\ell n\ n)}} = 1\} = 1 \qquad (3.6)$$

$$P\{\omega \mid \liminf \frac{\sqrt{n}\ \lambda_o\{\theta(F_n) - \theta_o^*\}}{\sqrt{2\sigma^2\ \ell n(\ell n\ n)}} = -1\} = 1\ ,$$

where $\lambda_o = \nabla K_{G_o}(\theta_o^*)$ and $\sigma^2 = \mathrm{var}_{G_o}\{\psi(X,\theta_o^*)\}$.

PROOF: By continuity, given $0 < \eta < \lambda_o/2$ choose an open ball $U_\delta(\theta_o^*)$, $\delta > 0$, on which $\lambda_o - \eta < \nabla K_{G_o}(\theta) < \lambda_o + \eta$. By Lemma 2.7 there exists an asymptotically unique sequence $\{\theta(F_n)\}$ of zeros of $\{K_n(\theta)\}$ consistent to θ_o^*. By Lemma 1.1

$$P\left\{\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^\eta\right\} = 1 , \tag{3.7}$$

where

$$A_n^\eta(\underset{\sim}{X}^{(n)}) = \{\omega \mid \theta(F_n) \in U_\delta , \ K_n(\theta(F_n)) = 0 ,$$

$$0 < \frac{1}{\lambda_o+2\eta} < \{\nabla K_n(\theta)\}^{-1} < \frac{1}{\lambda_o-2\eta} \text{ uniformly in } \theta \in U_\delta\}.$$

Let $\omega \in A_n^\eta$. By the mean value theorem

$$\theta(F_n) - \theta_o^* = -\{\nabla K_n(\xi_n)\}^{-1} K_n(\theta_o^*)$$

lies inside the interval with end points $-\frac{1}{\lambda_o+2\eta} K_n(\theta_o^*)$ and $-\frac{1}{\lambda_o-2\eta} K_n(\theta_o^*)$. For $a,b > 0$ we write $[a,b]c$ to be the closed interval $[ac,bc]$ if $c > 0$, and $[bc,ac]$ if $c < 0$. Now for each $m = 1,2,\ldots$ $\omega \in \bigcap_{n=m}^{\infty} A_n^\eta$ implies

$$b_m(\omega) = \sup_{n\geq m} \frac{\sqrt{n}\{\theta(F_n)-\theta_o^*\}}{\sqrt{2\sigma^2 \ell n(\ell n\ n)}} \in \left[\frac{1}{\lambda_o+2\eta} , \frac{1}{\lambda_o-2\eta}\right] \sup_{n\geq m} \frac{-\sqrt{n}\ K_n(\theta_o^*)}{\sqrt{2\sigma^2\ \ell n(\ell n\ n)}} . \tag{3.8}$$

Let

$$C = \{\omega \mid \limsup_{n} \sum_{t=1}^{n} - \frac{\psi(X_t,\theta_o^*)}{\sqrt{2\sigma^2\ n\ \ell n(\ell n\ n)}} = 1\} .$$

For an i.i.d. sequence where $\sigma^2 < 0$ the well known law of the iterated logarithm result given in Breiman (1968, P.64) gives that

$$P\{C\} = 1 \ . \tag{3.9}$$

Let $\omega \in \left\{ \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n^{\eta} \right\} \cap C$. Then there exists an m_o such that $m \geq m_o$ implies $\omega \in \bigcap_{n=m}^{\infty} A_n^{\eta}$, whence (3.8) holds. However, $b_1 \geq b_2 \geq \cdots\cdots \geq b_{m_o} \geq b_{m_o+1} \geq \cdots\cdots$, implies $\inf_m b_m = \inf_{m \geq m_o} b_m$. Then since $\lambda_o - 2\eta > 0$

$$\begin{aligned}
\inf_m b_m(\omega) &= \inf_{m \geq m_o} b_m(\omega) \\
&\in \left[\frac{1}{\lambda_o + 2\eta} \ , \ \frac{1}{\lambda_o - 2\eta} \right] \inf_{m \geq m_o} \sup_{n \geq m} \frac{-\sqrt{n}\, K_n(\theta_o^*)}{\sqrt{2\sigma^2 \ \ell n(\ell n \ n)}} \\
&= \left[\frac{1}{\lambda_o + 2\eta} \ , \ \frac{1}{\lambda_o - 2\eta} \right] \inf_m \sup_{n \geq m} \frac{-\sqrt{n}\, K_n(\theta_o^*)}{\sqrt{2\sigma^2 \ \ell n(\ell n \ n)}} \\
&= \left[\frac{1}{\lambda_o + 2\eta} \ , \ \frac{1}{\lambda_o - 2\eta} \right] \quad \text{by (3.9).}
\end{aligned}$$

Since η is arbitrary

$$P\{\omega \,|\, \lim \sup \frac{\sqrt{n}(\theta(F_n) - \theta_o^*)}{\sqrt{2\sigma^2 \ \ell n(\ell n \ n)}} = \frac{1}{\lambda_o}\} = 1 \ .$$

The associated result of (3.6) follows in an analogous manner. This completes the proof.

Following Theorem 3.1 the conditions of continuous differentiability can be relaxed slightly. An issue related to the law of the iterated logarithm for the estimator concerns the asymptotic expansion in the estimating equations.

LEMMA 3.2: Let ψ satisfy the assumptions of Theorem 3.3 and further be twice differentiable such that

$$\nabla^2 K_n(\theta) \xrightarrow{a.s.} \nabla^2 K_{G_o}(\theta) \quad \text{uniformly in} \quad \theta \in D \tag{3.10}$$

Then the asymptotically unique consistent sequence $\{\theta(F_n)\}$ satisfies

$$P\{\omega \mid \limsup\, n\{\ell n(\ell n\, n)\}^{-1} \left| K_n(\theta_o^*) + \lambda_o(\theta(F_n) - \theta_o^*) \right| \leq \frac{\alpha\sigma}{\lambda_o^2} + \frac{2\sqrt{\kappa\sigma}}{\lambda_o}\} \leq 1$$

where $\alpha = \alpha(\psi, G_o, \theta_o^*) = \left|\nabla^2 K_{G_o}(\theta_o^*)\right|$, and $\kappa = \kappa(\psi, G_o, \theta_o^*) = \mathrm{var}_{G_o} \nabla\psi(X, \theta_o^*)$.

PROOF: The usual Taylor expansion gives

$$0 = K_n\{\theta(F_n)\} = K_n(\theta_o^*) + \nabla K_n(\theta_o^*)\{\theta(F_n) - \theta_o^*\} + \tfrac{1}{2}\nabla^2 K_n(\xi_n)\{\theta(F_n) - \theta_o^*\}^2,$$

with ξ_n between $\theta(F_n)$ and θ_o^*, from which we write

$$K_n(\theta_o^*) + \nabla K_{G_o}(\theta_o^*)\{\theta(F_n) - \theta_o^*\} = \{\nabla K_{G_o}(\theta_o^*) - \nabla K_n(\theta_o^*)\}\{\theta(F_n) - \theta_o^*\}$$
$$- \tfrac{1}{2}\nabla^2 K_n(\xi_n)\{\theta(F_n) - \theta_o^*\}^2. \qquad (3.11)$$

Note that $\xi_n \xrightarrow{a.s.} \theta_o^*$ and (3.10) imply $\left|\nabla^2 K_n(\xi_n)\right| \xrightarrow{a.s.} \left|\nabla^2 K_{G_o}(\theta_o^*)\right| = 0$

Note that $\xi_n \xrightarrow{a.s.} \theta_o^*$ and (3.10) imply $\left|\nabla^2 K_n(\xi_n)\right| \xrightarrow{a.s.} \left|\nabla^2 K_{G_o}(\theta_o^*)\right| = \alpha$

Multiplying (3.11) by $n\{\ell n(\ell n\, n)\}^{-1}$ and taking absolute values we observe from the usual law of the iterated logarithm result

$$\limsup \sqrt{n\{\ell n(\ell n\, n)\}^{-1}} \left| \nabla K_{G_o}(\theta_o^*) - \nabla K_n(\theta_o^*) \right| < \sqrt{2\kappa} \quad \text{a.s.},$$

and from the result of Theorem 3.3

$$\limsup \sqrt{n\{\ell n(\ell n\, n)\}^{-1}} \left| \theta(F_n) - \theta_o^* \right| < \sqrt{2\sigma}/\lambda_o \quad \text{a.s.},$$

the Lemma holds true.

A particular version of the Lemma is to consider the M-estimator for location of a symmetric distribution on E, where $\psi(x,\theta) = \psi(x-\theta)$ for ψ an odd function, twice continuously differentiable. For G_o symmetric about zero there exists an M-functional value $\theta_o^* = 0$, and

$\alpha(\psi,G_o,\theta_o^*) = 0$. Then

$$P\{\omega \mid \limsup n\{\ell n(\ell n\, n)\}^{-1}|K_n(\theta_o^*) + \lambda_o\{\theta(F_n) - \theta_o^*\}| \leq c\} = 1 ,$$

where $c = \sqrt{\kappa\sigma}/\lambda_o$. This is a refinement of Carroll (1978a) who only claimed existence of a constant c. Conditions here appear stronger but extend to the more general M-estimator.

§3.3 The Multivariate M-estimator

More intricate arguments are necessary to extend limit theorems under equivalent conditions to the multiparameter models. But under suitable regularity conditions one can derive the asymptotic normality of the M-estimator when the underlying distribution lies in a neighbourhood of an $F_{\theta_o} \in F$.

THEOREM 3.4:

Let ψ satisfy conditions A of §2.3 and suppose $\underset{\sim}{X}$ is an i.i.d. sequence of r.v.'s generated by $G_o \in n(\varepsilon,F_{\theta_o})$, where ε is given by Lemma 2.5. Suppose $T[\psi,G]$ is the functional determined by selection statistic $\|\theta - \theta_o\|$. Assume $\psi(X,T[\psi,G_o])$ has finite second moments. Then there exists an asymptotically unique consistent sequence $\theta(\psi,F_n)$, consistent to $T[\psi,G_o]$, for which

$$\sqrt{n}(\theta(\psi,F_n) - T[\psi,G_o]) \xrightarrow{D} N\{0,\sigma^2(\psi,G_o,T[\psi,G_o])\} \tag{2.12}$$

where

$$\sigma^2(\psi,G,\theta) = M(\theta,G)^{-1}\Sigma(\psi,G,\theta)\{M(\theta,G)^{-1}\}' ,$$

and $\Sigma(\psi,G,\theta) = \mathrm{var}_G\psi(X,\theta)$. Here $\Theta \subset E^r$.

PROOF: By Theorem 2.2 $\{\theta(\psi,F_n)\}$ is an asymptotically unique consistent sequence to $T[\psi,G_o]$, defined by selection statistic $\|\theta - T[\psi,G_o]\|$. The two term Taylor expansion

$$0 = K_n(\theta[\psi,F_n]) = K_n(T[\psi,G_o]) + \nabla K_n(\tilde{\theta}_n)(\theta(\psi,F_n) - T[\psi,G_o]) ,$$

where $\nabla K_n(\theta)$ is evaluated at possibly different points $\tilde{\theta}_n$ on the diagonal between $\theta(\psi,F_n)$ and $T[\psi,G_o]$ (cf. Appendix 1), can be made. Since $\tilde{\theta}_n \xrightarrow{a.s.} T[\psi,G_o]$ and assumptions A1-2 give that

$$\|\nabla K_n(\theta) - M(\theta,G_o)\| \xrightarrow{a.s.} 0 \quad \text{uniformly in} \quad \theta \in D, \text{ then}$$

$$\|\nabla K_n(\tilde{\theta}_n) - M(T[\psi,G_o],G_o)\| \xrightarrow{a.s.} 0 .$$

So clearly the expansion implies $\sqrt{n}(\theta(\psi,F_n) - T[\psi,G_o]) = O_p(1)$. This follows since

$$\sqrt{n}\, K_n(T[\psi,G_o]) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i,T[\psi,G_o])$$

$$\xrightarrow{D} N\{0,\Sigma(\psi,G_o,T[\psi,G_o])\} ,$$

by the multivariate central limit theorem of Anderson (1958, P.74). The expansion can be validly re-expressed

$$\sqrt{n}(\theta(\psi,F_n) - T[\psi,G_o]) = -M(T[\psi,G_o],G_o)^{-1}\sqrt{n}\, K_n(T[\psi,G_o])$$

$$+ M(T[\psi,G_o],G_o)^{-1}\{M(T[\psi,G_o],G_o) - \nabla K_n(\tilde{\theta}_n)\}\sqrt{n}(\theta(\psi,F_n) - T[\psi,G_o]).$$

The latter term is now $o_p(1)$. This completes the proof.

REMARK: An alternative method of the multivariate C.L.T. proof is to make use of the Cramér-Wold (1936) device which shows that to study the asymptotic distribution of a vector ξ_n it is sufficient to study the one-dimensional asymptotic distribution of $Z_n = \underset{\sim}{z}'\xi_n$ for arbitrary unit $\underset{\sim}{z}$. (Crowder 1976).

By Theorems 2.2 and 3.4 we have introduced some new conditions which for suitably regular ψ functions are not difficult to check. The existence of a region about θ_o in the parameter space has been

demonstrated for which the M-functional will exist and be unique in a small enough neighbourhood about F_{θ_o}. Moreover, given any distribution in that neighbourhood the M-estimator will be an asymptotically unique consistent root to the M-functional, and as well be asymptotically normal. The arguments used are local and the choice of selection statistic is left to be resolved. But this will not affect the asymptotic distribution.

Asymptotic normality theorems at a specified distribution are common. A very weak set of conditions presented for showing asymptotic normality was given by Huber (1967). He considered a sequence $\{T_n[\underset{\sim}{X}^{(n)}]\}$ satisfying

$$(1/\sqrt{n}) \sum_{i=1}^{n} \psi(X_i, T_n[\underset{\sim}{X}^{(n)}]) \xrightarrow{P} 0 \ . \tag{3.13}$$

For instance if ψ is not continuous in θ there is no guarantee of a solution to the equations (2.5) as there otherwise would be by Corollary 2.2. But Lemma 2.2 does indicate the existence of consistent local minima of $\|K_n(\theta)\|^2$. It is interesting to compare Huber's assumptions with ours.

Assumptions: Huber (1967)

(N-1) For each fixed $\theta \in \Theta$, $\psi(x,\theta)$ is $\mathcal{B}$ measurable and $\psi(x,\theta)$ is separable in the sense of Doob: there is a P-null set N and a countable subset $\Theta' \subset \Theta$ such that for every open set $U \subset \Theta$ and every closed interval A, the sets $\{x|\psi(x,\theta) \in A, \forall\ \theta \in U\}$, $\{x|\psi(x,\theta) \in A, \forall\ \theta \in U \cap \Theta'\}$ differ by at most a subset of N.

Put $u(x,\theta,d) = \sup_{\|\tau-\theta\| \leq d} \|\psi(x,\tau) - \psi(x,\theta)\|$.

(N-2) There is a $T[G_o]$ such that $K_{G_o}(T[G_o]) = 0$.

(N-3) There are strictly positive numbers a,b,c,d_o such that

(i) $K_{G_o}(\theta) \geq a\|\theta - T[G_o]\|$ for $\|\theta - T[G_o]\| \leq d_o$.

(ii) $\int u(x,\theta,d)dG_o(x) \leq b.d$ for $\|\theta - T[G_o]\| + d \leq d_o \quad d \geq 0$

(iii) $\int u(x,\theta,d)^2 dG_o(x) \leq c.d$ for $\|\theta - T[G_o]\| + d \leq d_o \quad d \geq 0$.

(N-4) $\int \psi(x,T[G_o])\psi(x,T[G_o])'dG_o(x)$ is finite.

PROPOSITION 3.1 (Huber 1967, Theorem 3): Assume (N-1)-(N-4) hold and that T_n satisfies (3.13). If $P\{\|T_n - T[G_o]\| \leq d_o\} \to 1$, then

$$(1/\sqrt{n}) \sum_{i=1}^{n} \psi(X_i,T[G_o]) + \sqrt{n}\, K_{G_o}(T_n) \to 0$$

in probability.

COROLLARY 3.1 (Huber 1967, Corollary): Under the conditions of Proposition 3.1, assume K_{G_o} has a nonsingular derivative at $\theta = T[G_o]$. Then $\sqrt{n}(T_n - T[G_o])$ is asymptotically normal with mean zero and variance covariance matrix $\sigma^2(\psi,G_o,T[G_o])$.

The assumption (N-3)(ii) can be looked upon as a type of weakened Lipschitz condition on ψ over the parameter space, an essentially different condition to that of equicontinuity at each point in the observation space. While Huber's conditions are asserted to be very weak they appear difficult to check. For the influence functions associated with the symmetric location estimation a common condition imposed is that $|\psi(x) - \psi(y)| < M|x-y|$, for a constant $M < \infty$ and all $x,y \in E$. Clearly then $\{\psi(\cdot-\mu) | \mu \in E\}$ forms an equicontinuous family and condition (N-3)(ii) and (iii) are simultaneously satisfied.

The Lipschitz type condition to some extent emanates from the classical assumptions made by Cramér (1946, P.501) to prove asymptotic normality of the M.L.E. for i.i.d. random variables. They are equivalent to

CR1: There exists an $\alpha > 0$ such that $\psi(x,\theta)$ is twice differentiable with respect to θ, $\forall\ \theta \in S_\alpha = \{\theta \mid \|\hat{\theta} - T[G]\| \leq \alpha\}$ and the derivatives are for fixed θ, $\mathcal{B}$ measurable and almost surely continuous of $\theta \in S_\alpha$. Functions $(\partial^2/\partial\theta_i\partial\theta_j)\psi_k(x,\theta)$, $i,j,k = 1,\ldots,r$ are each $\mathcal{B}$ measurable for every $\theta \in S_\alpha$ for some $\alpha > 0$.

CR2: $\forall\ \theta \in S_\alpha$, $|\psi_k(X,\theta)| < F_1$, $|(\partial/\partial\theta_j)\psi_k(X,\theta)| < F_2$, $|(\partial^2/\partial\theta_i\partial\theta_j)\psi_k(X,\theta)| < H$, where F_1, F_2 are integrable over $(-\infty,\infty)$ and $\int H(x)dG_o(x) < M$.

CR3: $\int \psi(x,\theta)\psi(x,\theta)'dG_o(x) < \infty$.

CR2 is a type of weakened Lipschitz condition, on $\psi(x,\theta)$ in the parameter θ, in the sense that it is averaged over the x values by assuming an integrable third derivative. Our approach is a uniform restriction about each individual x-value over all $\theta \in D$, a compact subset of Θ. In general we cannot reconcile the two even though in specific cases such as the location parameter there appears to be some overlap.

§3.4 Relaxing Differentiability of the Multivariate Influence Function

The Cramér conditions assume the second partial derivatives of ψ exist and are almost surely continuous. Carroll (1978) makes similar assumptions to those of Cramér, but takes into account the possibility of a set $B(\theta)$ of Lebesgue measure zero in $R = E^k$, k-dimensional Euclidean space, at which the influence function does not have a continuous partial derivative at θ. A cornerstone to his proof is a lemma in which he shows

$$\|\hat{\theta}_n - \theta_o\| < Cn^{-\frac{1}{2}}\{\ell n(\ell n\ n)\}^{\frac{1}{2}} \quad \text{f.a.s.l.n.}.$$

This result is derived as a consequence of the uniform convergence

theory through Theorem 2.1 and results from the theory of empirical processes. Our approach has attempted to exhibit possible advantages of using a theory of uniform convergence. By it we also obtain easily checkable conditions on the multivariate ψ that do not require continuous differentiability.

LEMMA 3.3: Let $\underset{\sim}{X}$ be an i.i.d. sequence of real valued random variables with common distribution G_o that is continuous. Let θ_o^*, G_o, and ψ be such that B1 of §2.2 and C2, C3 of §2.3 all hold. Further suppose there exists a bounded integrable function $w(y)$ so that

$$W(x) = \int_{-\infty}^{x} w(y)dy \in L_2(G_o) \quad \text{and}$$

$$\sup_{\theta \in D} |d\psi_i(y,\theta)| < w(y)dy \qquad i = 1,\ldots,r \ .$$

Then there exists a constant C so that for any sequence $\{\hat{\theta}_n(\underset{\sim}{X}^{(n)})\}$ of zeros of $\{K_n(\theta)\}$ consistent to θ_o^*

$$\|\hat{\theta}_n - \theta_o^*\| < Cn^{-\frac{1}{2}}\{\ell n(\ell n\ n)\}^{\frac{1}{2}} \text{ holds f.a.s.1.n..} \qquad (3.14)$$

PROOF: By (2.13) it is sufficient to show

$$\sup_{\theta \in D} \|K_n(\theta) - K_{G_o}(\theta)\| = \sup_{\theta \in D} \left\| \int \psi(x,\theta)d[F_n(x) - G_o(x)] \right\|$$
$$< C_1 n^{-\frac{1}{2}} \{\ell n(\ell n\ n)\}^{\frac{1}{2}} \ .$$

Observe for the component functions ψ_i, integration by parts gives

$$\left| \int_{-\infty}^{+\infty} \psi_i(x,\theta)d[F_n(x) - G_o(x)] \right| = \left| - \int_{-\infty}^{+\infty} [F_n(x) - G_o(x)]d\psi_i(x,\theta) \right|$$
$$\leq \int_{-\infty}^{+\infty} |F_n(x) - G_o(x)||d\psi_i(x,\theta)|$$
$$\leq \int_{-\infty}^{+\infty} |F_n(x) - G_o(x)|w(x)dx \ . \qquad (3.15)$$

Let $\mathcal{B}(E)$ be the space of bounded real valued functions on the real line, with the sup norm. For $n \geq 3$, and $x \in E$ set

$$V_n(x) = \frac{n^{\frac{1}{2}}[F_n(x)-G_o(x)]}{\sqrt{2\ \ell n(\ell n\ n)}}\, w(x) \ .$$

By Corollary 1 of James (1975, P.771), there is a set of probability one on which the sequence $\{V_n\}_{n\geq 3}$ is relatively compact on $\mathcal{B}(E)$ with set of limit points

$$K^*_{w,G_o} = \{w(x)f(G_o(x)) \,|\, f \in K^*\} \ .$$

Here K^* is the set of absolutely continuous functions f in $\mathcal{B}([0,1])$ such that $f(0) = 0 = f(1)$ and $\int_o^1 [f'(t)]^2 dt \leq 1$. Then for a realization of a limit point of V_n we have from (3.15)

$$\left|\int_{-\infty}^{+\infty} \psi_i(x,\theta)\, d[F_n(x)-G_o(x)]\right| \leq n^{-\frac{1}{2}}\{2\ \ell n(\ell n\ n)\}^{\frac{1}{2}} \int_{-\infty}^{+\infty} w(x)\,|f(G_o(x))|\,dx$$

(by parts)
$$\leq n^{-\frac{1}{2}}\{2\ \ell n(\ell n\ n)\}^{\frac{1}{2}} \int_{-\infty}^{+\infty} W(x)\,|f'(G_o(x))|\,dG_o(x)$$

(Holder's inequality)
$$\leq n^{-\frac{1}{2}}\{2\ \ell n(\ell n\ n)\}^{\frac{1}{2}} \left(\int_{-\infty}^{+\infty} W(x)^2\, dG_o(x)\right)^{\frac{1}{2}}$$
$$\times \left(\int_{-\infty}^{+\infty} [f'(G_o(x))]^2\, dG_o(x)\right)$$
$$\leq n^{-\frac{1}{2}}\{2\ \ell n(\ell n\ n)\}^{\frac{1}{2}} \left(\int_{-\infty}^{+\infty} W(x)^2\, dG_o(x)\right)^{\frac{1}{2}}$$

and this proves the Lemma.

The result used from empirical processes in the Lemma is in fact a direct result of Finkelstein (1971). James allows the possibility of a much more general weight function w for the limit process of V_n to be attained. But for most purposes we need only that w be bounded. Results that appear to be for very restrictive classes $\mathcal{a}$ for which

$$\sup_{f\in \mathcal{a}} \left|\int_{-\infty}^{+\infty} f\, dF_n - \int_{-\infty}^{+\infty} f\, dG_o\right| << n^{-\frac{1}{2}}\{\ell n(\ell n\ n)\}^{\frac{1}{2}} \quad \text{f.a.s.l.n.}$$

can also be found in Kaufman and Phillip (1978).

It is emphasized that the component functions $\psi_i(x,\theta)$ need not be continuously differentiable in θ, but rather a continuity in the observation space variable is necessary. If $\Theta \subset E$, Theorem 3.1 can be applied. The proof of that theorem took a semi-deterministic approach which is also applicable to the law of the iterated logarithm and the uniform convergence to asymptotic normality. When $\Theta \subset E^r$, $r > 1$, a purely probabilistic approach through uniform convergence over classes of functions can avoid the necessity for continuous differentiability.

THEOREM 3.5:

Set $\underset{\sim}{X}$ to be an i.i.d. sequence generated by $G_o \in G$. Let θ_o^*, G_o, and ψ be as in Lemma 3.3, and $\psi(X,\theta_o^*) \in L_2(G_o)$. Suppose the partial derivatives of ψ exist and are continuous at θ for all $x \in R - B(\theta)$. Denote by B_n the set $\cup\{B(\theta) \,|\, \|\theta - \theta_o^*\| \leq \delta_n\}$, where $\delta_n = Cn^{-\frac{1}{2}}\{\ell n(\ell n\, n)\}^{\frac{1}{2}}$, and suppose $P_{G_o}(\bar{B}_n) = O(\delta_n)$. Assume there is an extension of $\nabla\psi(x,\theta)$ to points $x \in B(\theta)$, $\theta \in D$, so that

$$\nabla K_n(\theta) \xrightarrow{a.s.} \nabla K_{G_o}(\theta) \quad \text{uniformly in} \quad \theta \in D .$$

By Corollary 2.3 there exists a consistent sequence $\{\hat{\theta}_n\}$ of roots of (2.5) f.a.s.l.n., consistent to θ_o^*. Assume they satisfy (3.14). Finally let there exist a constant $H_1 < \infty$ so that for large enough n

$$\|\psi(x,\theta)-\psi(x,\theta_o^*)\| \leq H_1\|\theta - \theta_o^*\| \quad \text{uniformly in} \quad x \in B_n, \text{ and}$$

$$\|\nabla\psi(x,\theta)\| < H_1 \quad \text{uniformly in} \quad x \in B_n .$$

Then for the strongly consistent sequence $\{\hat{\theta}_n\}$,

$$\sqrt{n}(\hat{\theta}_n - \theta_o^*) \xrightarrow{D} N\{0,\sigma^2(\psi,G_o,\theta_o^*)\} .$$

PROOF: Set $S_n = \{i \,|\, X_i \in B_n,\ 1 \leq i \leq n\}$, and write

$$K_n^{**}(\theta) = \sum_{i \in S_n} \psi(X_i, \theta) .$$

Then let $K_n(\theta) = K_n^{*}(\theta) + K_n^{**}(\theta)$. For $\hat{\theta}_n$, a root of (2.5),

$$\begin{aligned} 0 = K_n(\hat{\theta}_n) &= K_n^{*}(\theta_o^{*}) + \nabla K_n^{*}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_o^{*}) + K_n^{**}(\hat{\theta}_n) \\ &= K_n(\theta_o^{*}) + \nabla K_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_o^{*}) - \nabla K_n^{**}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_o^{*}) \\ &\quad + K_n^{**}(\hat{\theta}_n) - K_n^{**}(\theta_o^{*}) . \end{aligned}$$

Hence

$$\begin{aligned} \| \sqrt{n}\, K_n(\theta_o^{*}) + \nabla K_n(\tilde{\theta}_n) \sqrt{n}(\hat{\theta}_n - \theta_o^{*}) \| &< n^{\frac{1}{2}} . 2 . H_1 \frac{\#X_t\text{'s} \in B_n}{n} \| \hat{\theta}_n - \theta_o^{*} \| \\ &= 2H_1 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{B_n}(X_i) \| \hat{\theta}_n - \theta_o^{*} \| . \end{aligned}$$

So for large n

$$P\{\omega \mid \| \sqrt{n}\, K_n(\theta_o^{*}) + \nabla K_n(\tilde{\theta}_n) \sqrt{n}(\hat{\theta}_n - \theta_o^{*}) \| > \varepsilon_n\}$$

$$\leq P\{\omega \mid \hat{\theta}_n \in B_n,\ 2H_1 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{B_n}(X_i) . \| \hat{\theta}_n - \theta_o^{*} \| > \varepsilon_n\} + P\{\omega \mid \hat{\theta}_n \notin B_n\}$$

$$\leq P\{\omega \mid 2 . H_1 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{B_n}(X_i) . \delta_n > \varepsilon_n\} + o_p(1)$$

$$= P\{\omega \mid \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{B_n}(X_i) > \frac{\varepsilon_n}{2 . H_1 . \delta_n}\} + o_p(1)$$

$$\leq \varepsilon_n^{-2} . 4 . H_1^2 . \delta_n^2\, P(B_n)((n-1)P(B_n) + 1) + o_p(1) \text{ (Chebyshev's Inequality).}$$

Letting $\varepsilon_n = \delta_n^{\frac{1}{2}}$ see that

$$\varepsilon_n^{-2} . 4 . H_1^2 . \delta_n^2 . (n-1) . P(B_n)^2 = O(n\delta_n^3) = o(1) \text{ , and}$$

$$\varepsilon_n^{-2} . 4 . H_1^2 \delta_n^2\, P(B_n) = O(\delta_n^2) = o(1). \text{ So,}$$

$$P\{\omega \mid \| \sqrt{n}\, K_n(\theta_o^{*}) + \nabla K_n(\tilde{\theta}_n) \sqrt{n}(\hat{\theta}_n - \theta_o^{*}) \| > \delta_n^{\frac{1}{2}}\} \to 0 \quad \text{as} \quad n \to \infty.$$

Finally, by the uniform convergence, and since $\tilde{\theta}_n \xrightarrow{a.s.} \theta_o^{*}$

$$\nabla K_n(\tilde{\theta}_n) \xrightarrow{a.s.} \nabla K_{G_o}(\theta_o^*) \ , \quad \text{and then}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_o^*) = - \{\nabla K_{G_o}(\theta_o^*)\}^{-1} \sqrt{n}\, K_n(\theta_o^*) + o_p(1) \ ,$$

which proves the result by the multivariate central limit theorem used in Theorem 3.4.

Uniform convergence is the essential ingredient in the proof of this theorem.

SECTION B: ROBUSTNESS - THEORY AND APPLICATION

CHAPTER 4

WEAK CONTINUITY AND FUNCTIONAL DERIVATIVES

§4.1 Background

In data analysis the P that describes the underlying process is most likely not in $\mathcal{P}$. This has two effects on the theory. Firstly, estimators T_n are consistent to some $\theta_1 \in \Theta$, but P_{θ_1} does not uniquely correspond to the generating P in any known way. Secondly, the predicted rate of convergence under the model (for instance through asymptotic variance) can be replaced by a radical departure even if P is "close" to a P_{θ_o} in the sense of some neighbourhood. Further criteria are necessary to take into account these latter considerations when choosing an estimator.

The development of robustness theory saw the structuring of such criteria. This began with the most important case; that of location for an i.i.d. sequence $\underset{\sim}{X}$ generated from a symmetric distribution G_o. Tukey (1960) gave an example with a set of mixtures of two normal distributions, the "contaminating" one having the same mean as the other, but a larger variance. In this situation there is no question as to what one is estimating, and the simple model provides a vivid illustration of the unsatisfactory behaviour of the sample mean and sample standard deviation under mild perturbation from strict normality. This compares with the acceptable behaviour of trimmed and Winsorized means. Huber (1964) arrived at an M-estimator for which the asymptotic variance was minimax amongst asymptotic variances of M-estimators consistent in

symmetric ε-contaminated neighbourhoods of the normal distribution. Again in 1972 he reviewed several methods that were stable under small symmetric departures from the underlying model. The lead away from the location parameter was taken by Hampel (1968). He argued heuristically in his thesis that sensitivity of estimators to observation values can be examined through the properties of the estimator functional at the model F. Robustness in that sense can be basically summed up: if the distance of the true G_o from F_{θ_o} is small enough, then the distance between the induced distribution laws on the estimators, $\mathcal{L}_{G_o}(T_n)$ and $\mathcal{L}_{F_{\theta_o}}(T_n)$ respectively, is also small. We are interested in the parametric procedure and not in a particular G_o, and hence consider neighbourhoods of an F_θ.

Erratic behaviour of an estimating functional $T[\cdot]$ on ε-contaminated neighbourhoods

$$n(\varepsilon,F_\theta) = \{(1-\delta)F_\theta + \delta H \mid 0 \leq \delta < \varepsilon\ ,\quad H \in G\}\ ,$$

will indicate the possible erratic behaviour of $T[F_n]$ when a single observation is permitted to greatly vary from the rest of the sample. On the other hand the Prokhorov neighbourhood allows for small round-off errors with large probability and gross errors with small probability within the sample. The former can be described as a modification of events A to A^δ, the latter as a probability error of size δ. Our model assigns $F_\theta\{A\}$ to A; in fact we observe A^δ plus gross errors with probability $G(A) = F_{\theta_o}(A^\delta) + \delta$. When distribution functions are close (within δ) in Prokhorov metric, then functionals T defined on the distribution space G which are weakly continuous are insensitive to such kinds of contamination.

DEFINITION 4.1: Let T be defined everywhere in $\mathcal{G}$. Then T is defined to be weakly continuous at $F \in \mathcal{G}$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $G \in \mathcal{G}$, $d_p(F,G) < \delta$ implies $\|T[F] - T[G]\| < \varepsilon$.

That weak continuity of the functional T should be equivalent to uniform continuity in n of the estimators $T_n[\underset{\sim}{X}^{(n)}] = T[F_n]$ is seen as follows.

PROPOSITION 4.1 (Hampel 1971): Let T be defined everywhere in $\mathcal{G}$ and put $T_n = T[F_n]$. We say that T_n is consistent at F if T_n tends to $T[F]$ in probability, where F is the true underlying distribution.

(i) If T is weakly continuous at all F, then T_n is consistent at all F, and $F \to \mathcal{L}(T_n)$ is weakly continuous uniformly in n.

(ii) If T_n is consistent and $F \to \mathcal{L}(T_n)$ is weakly continuous uniformly in n at all F, then T is weakly continuous.

Weak continuity alone may be satisfied by a broad class of estimators, even within the class of M-functionals. What we seek then are additional quantitative as well as qualitative results to help settle on a good robust statistical procedure. Quantitative results may relate to behaviour of asymptotic variance and bias within neighbourhoods of F_θ. "Breakdown points" or sizes of smallest neighbourhoods on which functionals exceed set values of asymptotic variance or bias, possibly infinite, may also be considered. In this thesis we emphasize M-estimators of a general parameter that have an influence function giving zero weight to "tail" regions of an F_θ. Hampel initially proposed the notion of these influence functions for symmetric location M-estimators, giving an example in his 1974 paper. This is a qualitative approach, although quantitative evidence may be found in support of it.

While robustness theory indicates the use of some general statistical procedures, specific application is often neglected. Invariably one requires algorithms for the search of solutions, or estimators, T_n. By examining asymptotics it is possible to reveal some relations which give a clearer picture to the identification of solutions of estimating equations. This is in addition to that already afforded in the manner of consistency arguments.

§4.2 Weak Continuity of M-Functionals

The relationship between weak continuity and consistency of Proposition 4.1 follows from the fact that if $\{\mu_n\}$ is any sequence of probability measures such that $\mu_n \Rightarrow \mu$ then this is equivalent to $d_p(\mu_n,\mu) \to 0$ (Prokhorov 1956). Consistency then follows from the result of Varadarajan (1958), $F_n \Rightarrow G_o$ almost surely. From this follows Fisher consistency. Also bearing in mind the intuitive interpretation of the Prokhorov distance a natural robustness or stability requirement is that T should be continuous with respect to the weak topology (at least at the model distribution, but if possible for G in some pencil of neighbourhoods of the model).

Weak continuity of the M-functional $T[\psi,\cdot]$ is closely related to condition B1 and C1 (of §2.2). Huber (1977) points out that for the M-functional for location to be weakly continuous, it is sufficient that ψ be bounded and continuous. Also there is an assumption of monotonicity of ψ, any corresponding loss function of which is convex. For more general parameters and influence function ψ, weak continuity will also depend on the nature of the selection statistic. Endeavouring to establish C1 for Prokhorov neighbourhoods

$$n_p(\varepsilon,G_o) = \{G \in \mathcal{G} \,|\, d_p(G,G_o) < \varepsilon\} \ ,$$

we take the now familiar approach via equicontinuity giving some preparatory results.

LEMMA 4.1: Given some distribution function G_o, a continuity set $A \in \mathcal{B}$ of G_o, and $\kappa > 0$, there exists $\delta > 0$ such that for any $G \in \mathcal{G}$:

$$d_p(G,G_o) < \delta \text{ implies } |G(A) - G_o(A)| < \kappa .$$

PROOF: Since A is a continuity set there is some $\delta_1 > 0$ for which $G_o(A^{\delta_1}) < G_o(A) + \kappa/2$. We may choose $\delta_1 < \kappa/2$. Then if $d_p(G,G_o) < \delta_1$,

$$G(A) < G_o(A^{\delta_1}) + \delta_1 < G_o(A) + \kappa .$$

Similarly, since $R - A$ shares the same boundary as A there is some $\delta_2 > 0$ for which $G_o(A^{-\delta_2}) = G_o(R - (R-A)^{\delta_2}) > G_o(A) - \kappa/2$. Choose $\delta_2 < \kappa/2$ and suppose $d_p(G,G_o) < \delta_2$. Then,

$$G(A) > G_o(A^{-\delta_2}) - \delta_2 > G_o(A) - \kappa .$$

Taking $\delta = \min(\delta_1,\delta_2)$ gives the lemma.

PROPOSITION 4.2 (Rao 1962): Let G_o be a probability measure on $(R,\mathcal{B})$ and $\mathcal{a}$ be an equicontinuous family of real valued functions on R. Then given arbitrary compact $C \subset R$, for each $\eta > 0$ there exists a finite number of sets $\{A_j\}_{j=1}^{n}$, where $n = n(\eta)$ such that

(1) $\bigcup_{j=1}^{n} A_j = C$; (2) $A_j \cap A_{j'} = \phi$, for $j \neq j'$; and (3) for each j, A_j is a continuity set for G_o; (4) for any $x,y \in A$ and $f \in \mathcal{a}$ $|f(x) - f(y)| < \eta$.

REMARK 4.1: Both assumptions of separability and metrizeability of R are utilized in the proof of this proposition.

THEOREM 4.1:

Let $\mathcal{A}$ be a class of continuous functions on R possessing the following two properties; (1) $\mathcal{A}$ is uniformly bounded, that is, there exists a constant H such that $|f(x)| \leq H < \infty$ for all $f \in \mathcal{A}$ and $x \in R$; and (2) $\mathcal{A}$ is equicontinuous. Suppose $G_o \in \mathcal{G}$. Then for every $\varepsilon > 0$ there exists a $\delta' > 0$ such that $d_p(G,G_o) < \delta$ implies

$$\sup_{f \in \mathcal{A}} \left| \int f \, dG - \int f \, dG_o \right| < \varepsilon .$$

PROOF: Since R is separable and complete there exists a compact set such that $G_o(R - C) < \varepsilon/(16.H)$. We can assume C to be a continuity set of G_o. Let $\eta = \varepsilon/4$ and $\{A_j\}$ subsequently be formed in the manner of Proposition 4.2. Choose $\{y_j\}_{j=1}^n$ arbitrarily in $\{A_j\}_{j=1}^n$ respectively and let G_o^* be the possibly improper measure attributing weight $G_o(A_j)$ to the point y_j, for each $j = 1,\ldots,n$. Then for each $f \in \mathcal{A}$

$$\left| \int_C f \, dG_o - \int_C f \, dG_o^* \right| \leq \sum_{j=1}^n \int_{A_j} |f(x) - f(y_j)| \, dG_o(x)$$

$$< \varepsilon/4 .$$

Hence,

$$\sup_{f \in \mathcal{A}} \left| \int_C f \, dG_o - \int_C f \, dG_o^* \right| \leq \varepsilon/4 .$$

Similarly, given $G \in \mathcal{G}$ we let G^* be that measure attributing weight $G(A_j)$ to y_j for each $j = 1,\ldots,n$. And so

$$\sup_{f \in \mathcal{A}} \left| \int_C f \, dG - \int_C f \, dG^* \right| \leq \varepsilon/4 .$$

Now

$$\left| \int_C f \, dG_o^* - \int_C f \, dG^* \right| \leq H. \sum_{j=1}^n |G_o(A_j) - G(A_j)| .$$

By Lemma 4.1 choose δ_j such that if $G \in \mathcal{G}$, $d_p(G,G_o) < \delta_j$ implies $|G_o(A_j) - G(A_j)| < \varepsilon/(4.n.H)$. Let δ_o be so that if $G \in \mathcal{G}$, $d_p(G,G_o) < \delta_o$ implies $|G_o(R-C) - G(R-C)| < \varepsilon/(16.H)$. Taking $\delta = \min(\delta_o,\delta_1,\ldots,\delta_n)$, if $G \in \mathcal{G}$ and $d_p(G,G_o) < \delta$ it follows that

$$\begin{aligned}
\sup_{f\in\mathcal{A}}\left|\int f\,dG_o - \int f\,dG\right| &\le \sup_{f\in\mathcal{A}}\left|\int_{R-C} f\,dG_o - \int_{R-C} f\,dG\right| \\
&\quad + \sup_{f\in\mathcal{A}}\left|\int_C f\,dG_o - \int_C f\,dG_o^*\right| + \sup_{f\in\mathcal{A}}\left|\int_C f\,dG - \int_C f\,dG^*\right| \\
&\quad + \sup_{f\in\mathcal{A}}\left|\int_C f\,dG_o^* - \int_C f\,dG^*\right| \\
&< H.(G_o(R-C) + G(R-C)) \\
&\qquad + \varepsilon/4 + \varepsilon/4 + \varepsilon/4 \\
&< \varepsilon\ .
\end{aligned}$$

The Theorem is proved.

REMARK 4.2: Clearly if $R = E$ and the decomposition of C is into the set $-c = a_o < a_1 < \ldots\ldots < a_n = c$, where the a_i are continuity points of G_o and $A_i = (a_{i-1},a_i]$, then Theorem 4.1 holds with d_p replaced by either:

(1) the Kolmogorov metric d_k, where

$$d_k(G,G_o) = \sup_{x\in E}|G(x) - G_o(x)|, \text{ or}$$

(2) the Lévy metric d_L, where

$$d_L(G,G_o) = \inf\{\varepsilon \mid G(x) \le G_o(x+\varepsilon) + \varepsilon,\ G_o(x) \le G(x+\varepsilon) + \varepsilon, \text{ for all } x \in E\}\ .$$

In a sense the assumptions of the theorem describe the most general class for which this result holds. Considering a weaker condition that $\sup_{f\in\mathcal{A}}\int |f|\,dG_o < +\infty$ and letting $\mathcal{A}$ be unbounded it is possible to choose sequences $\{f_n\} \in \mathcal{A}$ and $\{y_n\} \in R$ so that $|f_n(y_n)| \to +\infty$ as $n \to \infty$. But then given any $\delta > 0$ letting $G_n = (1-\delta)G_o + \delta\delta_{y_n}$, where

δ_{y_n} is that degenerate distribution attributing unit mass to the point $y = y_n$,

$$\sup_{f \in \mathcal{a}} \left| \int f \, dG_o - \int f \, dG_n \right| > \delta \left| \int f_n dG_o - f_n(y_n) \right|$$

$$> \delta(|f_n(y_n)| - \sup_{f \in \mathcal{a}} \int |f| dG_o)$$

$$\to +\infty \quad \text{as} \quad n \to \infty .$$

This is even though $d_p(G_o, G_n) \leq \delta$. Hence the result of the theorem could not hold if $\mathcal{a}$ is permitted to be unbounded.

When the family $\mathcal{a}$ is not equicontinuous there exists an $\varepsilon > 0$ and an $x \in R$ for which

$$\sup_{y \in N_x(\eta)} \sup_{f \in \mathcal{a}} |f(x) - f(y_n)| > \varepsilon .$$

So if $G_o = \delta_x$ and $G_n = \delta_{y_n}$, it is true that $d_p(G_o, G_n) \to 0$ but $\sup_{f \in \mathcal{a}} \left| \int f \, dG_o - \int f \, dG_n \right| > \varepsilon$. This again prevents the assertion of Theorem 4.1. This does not mean the conditions on $\mathcal{a}$ are necessary. Since in the latter example the distribution G_o was chosen in relation to the family $\mathcal{a}$. That is, if G_o were simply chosen to be a continuous distribution function, it is quite plausible that the theorem can hold for broader families $\mathcal{a}$ than those specified in Theorem 4.1.

We now observe that if $\{\psi(\cdot, \theta) | \theta \in D\}$ form an equicontinuous family of functions that are bounded uniformly by a constant, that condition C1 of §2.3 is satisfied for the Prokhorov metric. That is given $\varepsilon > 0$ there exists a $\delta > 0$ so that $G \in n_p(\delta, G_o)$ implies

$$\sup_{\theta \in D} \left\| \int \psi(x, \theta) dG_o(x) - \int \psi(x, \theta) dG(x) \right\| < \varepsilon .$$

Here n_p is the Prokhorov neighbourhood generated by the metric d_p. So as a result of Lemma 1.6 and Theorem 4.1, it follows from condition C2 of §2.3 and uniform boundedness of ψ on $R \times D$, that condition C1 is satisfied. The next result follows from Lemma 2.3.

LEMMA 4.3: Let B1, and C2 hold, and assume $\{\psi(\cdot,\theta)|\theta \in D\}$ is a family of functions uniformly bounded by a constant. Define the functional $T[\psi,\cdot]$ uniquely by the selection functional $\|\theta - \theta_o^*\|$. Then $T[\psi,\cdot]$ is weakly continuous at G_o.

PROOF: The assumptions of Lemma 2.3 are satisfied.

Similarly assumptions A can be verified with the application of Theorem 4.1. By Lemma 2.5 and Theorem 2.5 they are sufficient for uniqueness of solutions to the estimating functional equations, (2.8), in a set region of the parameter space. In particular note that if A1 is satisfied and families $\{\psi(\cdot,\theta)|\theta \in D\}$, $\{\nabla\psi(\cdot,\theta)|\theta \in D\}$ are uniformly bounded in Euclidean norm by some constant, then assumptions A2 and A4 immediately hold.

The conditions sufficient for the weak continuity of $T[\psi,\cdot]$ are simple. Moreover if conditions A hold we can assert that there exists a region about θ_o for which there is a unique solution to equations (2.8) on a neighbourhood of F_{θ_o}. Returning to Remark 2.1 following Theorem 2.2, there is also a unique solution to equations (2.5) in that region f.a.s.l.n. if the sequence $\underset{\sim}{X}$ is generated from a distribution G of that neighbourhood. Influence functions ψ possessing "monotonicity properties", ensuring at most a unique root to the equations (2.5) or (2.8) for whatever distribution $G \in \mathcal{G}$, then generate weakly continuous M-functionals. But if more than one root is allowed to exist further consideration must be given to the global arguments identifying the M-functional.

THEOREM 4.2:

Assume $f_G(\theta)$ is continuous in $\theta \in \Theta$ for all $G \in \mathcal{G}$, and given $F_{\theta_o} \in \mathcal{F}$ we have that for every neighbourhood N of $T[\psi,f,F_{\theta_o}] = \theta_o$

$$\inf_{\theta \notin N} f_{F_{\theta_o}}(\theta) - f_{F_{\theta_o}}(T[\psi,f,F_{\theta_o}]) > 0 .$$

Suppose for every $\eta > 0$ there exists $\varepsilon > 0$ such that $d_p(G,F_{\theta_o}) < \varepsilon$ implies

$$\sup_{\theta \in \Theta} |f_G(\theta) - f_{F_{\theta_o}}(\theta)| < \eta .$$

Then if conditions A of §2.3 are satisfied with respect to the Prokhorov neighbourhood n_p, $T[\psi,f,\cdot]$ is weakly continuous at F_{θ_o}.

PROOF: By Lemma 2.5, there is $\kappa^* > 0$ and $\varepsilon > 0$ such that $d_p(G,F_{\theta_o}) < \varepsilon$ implies that $H_o(\psi,G) \cap U_{\kappa^*}(\theta_o)$ consists of a single point, where

$$H_o(\psi,G) = \{\theta | \theta \in \Theta, K_G(\theta) = 0\} .$$

Denote $\delta(\kappa^*) = \inf_{\theta \in U_{\kappa^*}(\theta_o)} f_{F_{\theta_o}}(\theta) - f_{F_{\theta_o}}(\theta_o)$. Choose $0 < \kappa' < \kappa^*$ so that

$$|f_{F_{\theta_o}}(\theta) - f_{F_{\theta_o}}(\theta_o)| < \delta(\kappa^*)/2 \quad \text{for} \quad \theta \in U_{\kappa'}(\theta_o) .$$

For $\kappa' > 0$ choose $0 < \varepsilon' \leq \varepsilon$ so that $d_p(G,F_{\theta_o}) < \varepsilon'$ implies there exists a root $\theta(\psi,G)$ in $U_{\kappa'}(\theta_o)$, and

$$|f_G(\theta) - f_{F_{\theta_o}}(\theta)| < \delta(\kappa^*)/4 \quad \text{uniformly in} \quad \theta \in \Theta .$$

The root is unique in $U_{\kappa^*}(\theta_o) \supset U_{\kappa'}(\theta_o)$. Then

$$f_G(\theta(\psi,G)) < f_{F_{\theta_o}}(\theta(\psi,G)) + \delta(\kappa^*)/4$$

$$< f_{F_{\theta_o}}(\theta_o) + 3.\delta(\kappa^*)/4$$

$$< f_{F_{\theta_o}}(\theta) - \delta(\kappa^*)/4 \quad \text{uniformly in} \quad \theta \in \Theta - U_{\kappa^*}(\theta_o)$$

$$< f_G(\theta) \quad \text{uniformly in} \quad \theta \in \Theta - U_{\kappa^*}(\theta_o) \ .$$

So

$$\inf_{\theta \in H_o(\psi,G)} f_G(\theta) = f_G(\theta(\psi,G)) \ .$$

That is $T[\psi,f,G] = \theta(\psi,G)$, where $\theta(\psi,G)$ is determined by the selection statistic $\|\theta - \theta_o\|$. Hence

$$\|T[\psi,f,G] - T[\psi,f,F_{\theta_o}]\| < \kappa'$$

whenever $d_p(G,F_{\theta_o}) < \varepsilon'$. The functional T is weakly continuous.

Theorem 4.2 provides sufficient conditions for weak continuity of the M-functional when a selection statistic is required. It is emphasized that the conditions are by no means necessary. For instance the selection functional $f_G(\theta)$ can be weakly continuous at F_{θ_o} in the sense described in the Theorem, but it is equivalent in its action to the selection functional

$$f_G^*(\theta) = f_G(\theta) + \int_R \|x\| dG(x) \ ,$$

where R is a Fréchet space. We assume the latter term is finite (set it equal to zero if not). Clearly $f_G^*(\theta)$ does not satisfy the weak continuity property.

§4.3 Fréchet Differentiability

In a recent Ph.D. thesis Reeds (1976) examined the definition of von Mises functionals and their derivatives. Von Mises (1947) gave an initial framework of the functional derivative, and this was subsequently followed up by Fillipova (1962), and Kallianpur (1963), although domains of the functional varied and there existed some confusion as to the

nomenclature of the derivatives. The latter author discussed specifically the M.L.E. with emphasis on its asymptotic efficiency among a class of Von Mises functionals of second order. Relatively strong conditions that included Cramér's conditions for asymptotic normality were imposed to prove existence of derivatives.

Statistical application of the approach through functional derivatives requires establishing asymptotic normality and the examination of higher order properties of estimators. It brings into a common structure estimators of diverse origin.

If we consider functionals defined on the linear space of finite signed measures

$$M = \{aF + bG \mid a,b \text{ real}, F,G \in G\} ,$$

we can give simply three notions of derivative. M is a normed linear space with respect to $\|H\|^* = \sup_x |H(x) - H(-\infty)|$ where $R = E$, or $\|H\|^*$ = total variation of H, when $H \in M$. Let the functional $T : M \to E^r$. Define *derivatives* for T at $G_o \in M$ as follows: first, for any (4.1) continuous linear map $L : M \to E^r$, and $H \in M$ define associated *remainder*

$$R(G_o+tH) = \begin{cases} T[G_o+tH] - T[G_o] - L[tH] , & t \neq 0 \\ 0 , & t = 0 . \end{cases}$$

Suppose

$$\left\| \frac{R(G_o+tH)}{t} \right\| \to 0 \quad \text{as} \quad t \to 0 . \tag{4.2}$$

(i) T is said to be *Gâteaux differentiable* at G_o with derivative $T'_{G_o} = L$ if (4.2) holds for all $H \in M$;

(ii) if (4.2) holds *uniformly* for H lying in an arbitrary compact subset of M, T'_G is called a *compact derivative* of T at G_o;

(iii) if (4.2) holds uniformly for H lying in an arbitrary bounded subset of M, T'_G is called the *Fréchet derivative* of T at G_o.

These derivatives are successively stronger. Now T is Fréchet differentiable at G_o if and only if

$$\|T[G]-T[G_o] - T'_{G_o}(G-G_o)\| = O(\|G-G_o\|^*) \tag{4.3}$$

for some continuous linear functional T'_{G_o}. But many statistical functionals are not defined on M, and it is necessary to specify what is meant by the derivative. (Hampel 1968, P.39 implies that the extension of the functional T from G to the space of signed measures can be made in a "natural way", but the extension is not specified.) If T is a vector valued functional defined on a subset $G' \subset G$ of distribution functions including G_o and d is a metric on G, we say the statistical functional T is Fréchet differentiable at G_o with respect to (G',d) when it can be approximated by a linear functional T'_{G_o} such that for all $G \in G'$

$$\|T[G]-T[G_o] - T'_{G_o}(G-G_o)\| = O(d(G,G_o)) \ . \tag{4.4}$$

This definition was essentially used to define the Fréchet derivative in Kallianpur and Rao (1955). They used $G' = F$, the parametric family of univariate distributions for which $\Theta \subset E$. The metric was the Kolmogorov metric defined by

$$d_k(G,G_o) = \sup_{-\infty<x<\infty} |G(x) - G_o(x)| \ .$$

The two definitions of Fréchet differentiability (4.3) and (4.4) effectively depart when the metric distance between G and G_o cannot be guaged solely from the difference $G - G_o$. Metrics which can, are the Kolmogorov metric, the total variation metric, and the bounded Lipschitz

metric (Huber 1977). Boos (1979) adopts the expansion (4.3) although he restricts the domain of the derivative to be in line with the domain of the statistical functional T. Boos and Serfling (1980) take this approach to establish asymptotic normality of the M-estimator of location. Reeds (1976) on the other hand introduced the compact derivative in order to accommodate asymptotic normality arguments when statistical functionals were not necessarily Fréchet differentiable.

Kallianpur and Rao (1955) introduced Fréchet differentiability for a class of Fisher consistent estimators, showing that any statistic belonging to this class was consistent and asymptotically normally distributed with asymptotic variance greater than or equal to $[n\, I(\theta)]^{-1}$, where $I(\theta)$ is the Fisher information function. Rao (1957) was able to show that the M.L.E. for the multinomial distribution was a member of this class, and hence efficient with respect to this class. But Kallianpur (1963) reported that in general neither author could under any reasonable set of assumptions (on the density function in the continuous, and the probability function in the infinite discrete case) prove Fréchet differentiability of the M.L.E.. It was felt that Fréchet differentiability was too severe a restriction when dealing with the infinite dimensional (non multinomial) situation. This was the motivation for the latters article on "Volterra" derivatives. By results from §4.2 and §2.3 we can obtain some restrictive conditions under which Fréchet differentiability can be established for the M-functional. They are restrictive only in the sense that for a number of parametric families they will not be satisfied by the M.L.E., which can still be an efficient estimator in the first order sense of Rao (1963).

THEOREM 4.3:

Assume conditions A of §2.3 hold with respect to the neighbourhoods generated by a metric d on $\mathcal{G}$. Suppose further that for $G \in \mathcal{G}' \subset \mathcal{G}$

$$\int \psi(x,\theta_o)d(G-F_{\theta_o})(x) = O(d(G,F_{\theta_o})) \ . \tag{4.5}$$

Let $\mathcal{F} \subset \mathcal{G}'$. Then $T[\psi,\cdot]$ is Fréchet differentiable at F_{θ_o} with respect to $(\mathcal{G}',d)$, and has derivative

$$T'_{F_{\theta_o}}(G-F_{\theta_o}) = -M(\theta_o)^{-1}\int \psi(x,\theta_o)d(G-F_{\theta_o})(x)$$

where $M(\theta_o)$ is given in condition A3.

PROOF: Abbreviate $T[\psi,\cdot] = T[\cdot]$, and let κ^*, ε be given by Lemma 2.5. Let $\{\varepsilon_k\}$ be so that $\varepsilon_k \downarrow 0^+$ as $k \to \infty$, and let $\{G_{\varepsilon_k}\}$ be any sequence such that $G_{\varepsilon_k} \in n(\varepsilon_k,F_{\theta_o}) \cap \mathcal{G}'$. Here $n(\varepsilon_k,F_{\theta_o})$ is the neighbourhood of distributions within distance ε_k from F_{θ_o}. It is sufficient to show

$$\|T[G_{\varepsilon_k}] - T[F_{\theta_o}] - T'_{F_{\theta_o}}(G_{\varepsilon_k} - F_{\theta_o})\| = o(\varepsilon_k) \ .$$

By Lemma 2.5, $T[G_{\varepsilon_k}]$ exists and is unique in $U_{\kappa^*}(\theta_o)$. Also note that by assumption A4

$$\|M(\theta,G_{\varepsilon_k}) - M(\theta)\| \xrightarrow{k\to\infty} 0 \quad \text{uniformly in} \quad \theta \in D \ . \tag{4.6}$$

Examine the two term Taylor expansion,

$$0 = K_{G_{\varepsilon_k}}(T[G_{\varepsilon_k}]) = K_{G_{\varepsilon_k}}(\theta_o) + M(\tilde{\theta}_k,G_{\varepsilon_k})(T[G_{\varepsilon_k}] - \theta_o) \ ,$$

where $\tilde{\theta} \in U_{\kappa^*}(\theta_o)$ for $k \geq k_o$ and is evaluated at different points for each component function expansion. In particular by Lemma 2.5

$$\|\tilde{\theta}_k - \theta_o\| \leq \|T[G_{\varepsilon_k}] - \theta_o\| \to 0 \quad \text{as} \quad k \to \infty.$$

Then from the expansion and (4.6) see that

$$\|T[G_{\varepsilon_k}] - \theta_o\| = O(K_{G_{\varepsilon k}}) = O(\varepsilon_k) .$$

Consider the reformulation

$$T[G_{\varepsilon_k}]-\theta_o = -M(\theta_o)^{-1}K_{G_{\varepsilon_k}}(\theta_o)+M(\theta_o)^{-1}(M(\tilde{\theta}_k,G_{\varepsilon_k})-M(\theta_o))(T[G_{\varepsilon_k}]-\theta_o) . \tag{4.7}$$

Since by continuity of $M(\theta)$ and (4.6) $\|M(\tilde{\theta}_k,G_{\varepsilon_k})-M(\theta_o)\| = o(1)$

$$\|T[G_{\varepsilon_k}]-\theta_o-T'_{F_{\theta_o}}(G_{\varepsilon_k}-F_{\theta_o})\| = o(1)O(d(G_{\varepsilon_k},F_{\theta_o})) = o(\varepsilon_k).$$

The theorem is proved.

Now if $\psi(x,\theta_o)$ is a function of total bounded variation, and if for all $G \in \mathcal{G}$ integration by parts

$$\int \psi(x,\theta_o)d(G-F_{\theta_o}) = -\int (G-F_{\theta_o})(x)d\psi(x,\theta_o) , \tag{4.8}$$

is valid, then (4.5) is easily established for the Kolmogorov metric. Bearing in mind Remark 4.1 Fréchet differentiability with respect to the Kolmogorov metric may be established by this theorem for certain M-functionals of univariate distributions. If (4.8) were to hold for distributions on E^k, where $G(x) = G(x_1,\ldots,x_k)$ given by

$$G(x) = G(x_1,\ldots,x_k) = \int_{(-\infty,x_1]\times\ldots\times(-\infty,x_k]} dG(x) ,$$

it follows that (4.5) holds for the total variation metric. Conditions on $\psi(x,\theta_o)$ and $F_{\theta_o}(x)$ are not so clear that we can easily establish (4.5) for the Lévy or Prokhorov metrics. But for instance if F_{θ_o} were an absolutely continuous distribution function on the real line, possessing a bounded density in which case

$$\sup_{x\in E} F_{\theta_o}(x+\delta) - F_{\theta_o}(x) < c\delta \quad \text{uniformly in} \quad \delta > 0 ,$$

it is then not difficult to observe that if

$$F_{\theta_o}(x) \leq G(x+\delta) + \delta \ ,$$

and $G(x) \leq F_{\theta_o}(x+\delta) + \delta$ holds uniformly in $x \in E$, then

$$\sup_{x \in E} |G(x) - F_{\theta_o}(x)| < (c+1)\delta \ .$$

Hence

$$\sup_x |G(x) - F_{\theta_o}(x)| \leq d_L(G, F_{\theta_o})(c+1) \ ,$$

and (4.5) can be established immediately from (4.8) for the Lévy metric, and also the Prokhorov metric since $d_L \leq d_p$. The latter metric is defined on more general spaces and cannot always be compared with d_k. From the relations $d_L \leq d_k, d_p \leq d_T$, Fréchet differentiability with respect to the Lévy, Prokhorov, or Kolmogorov metrics implies differentiability with respect to the total variation metric. Beran (1977) advocated the Hellinger metric, d_H, for robust parametric estimation. The topology generated by d_H is equivalent to that generated by the total variation metric since $d_T \leq d_H \leq \sqrt{2}\, d_T$ (Stautde 1978).

If a selection functional is used then the same condition on the selection functional as that described in Theorem 4.2 ensures Fréchet differentiability of $T[\psi, f, \cdot]$ at F_{θ_o}. This is because differentiability is only a local argument.

If $T[\psi, \cdot]$ is Fréchet differentiable with respect to (G, d_k) at a continuous distribution F_{θ_o}, and if Ao of conditions A in §2.3 is satisfied, there exists an expansion

$$\sqrt{n}(T[\psi, F_n] - T[\psi, F_{\theta_o}]) = -M(\theta_o)^{-1}\sqrt{n} \int \psi(x, \theta_o) dF_n(x) + \sqrt{n}\, o(d_k(F_n, F_{\theta_o})). \tag{4.9}$$

If F_n is the empirical generated by F_{θ_o}, then for the i.i.d. sequence

$\underset{\sim}{X}$, a result of Kolmogorov gives that

$$\lim_{n\to\infty} P\{n^{\frac{1}{2}} \sup_x (F_n(x) - F_{\theta_o}(x)) < \lambda\} = 1 - e^{-2\lambda^2}$$

(Hájek and Šidák 1967, P.199). That is $\sqrt{n}\, o(d_k(F_n, F_{\theta_o}))$ is $o_p(1)$. Hence $\sqrt{n}(T[\psi, F_n] - T[\psi, F_{\theta_o}])$ converges in law asymptotically to a normal random variable with mean zero and variance covariance matrix $\sigma^2(\psi, F_{\theta_o}, \theta_o)$. More generally if F_n is generated by an i.i.d. univariate sequence taken from arbitrary $G \in \mathcal{G}$ it is shown in Appendix 2 that $d_k(F_n, G) = O_p(n^{-\frac{1}{2}})$.

The methods used to prove Fréchet differentiability of the M-functional follow similar lines to proving existence of consistent asymptotically normal roots of M-estimating equations when the underlying distribution lies in neighbourhoods of an F_{θ_o} (cf. Theorem 3.4). Clearly the asymptotic normality proof via direct expansion of the estimating equations affords greater generality as it includes cases where $\psi(x,\theta)$ is unbounded, whence condition A4 does not hold for metrics d_L, d_k, or d_p.

EXAMPLE: We illustrate by showing Fréchet differentiability of the M.L.E. of the multinomial parameter. Since observations are congregated on k points, representing k classes, there is no need for the Prokhorov metric to cover the possibility of rounding or gross errors. Apparently, Frechet differentiability with respect to $(\mathcal{G}', d_k)$, where $\mathcal{G}'$ is the subset of distributions of $\mathcal{G}$ whose support is contained within k points $y_1 < y_2 < \ldots\ldots < y_k$ say, is sufficient to show asymptotic normality. Rao (1957) considered a representation of k classes with hypothetical frequencies $\pi_1(\theta), \ldots\ldots, \pi_k(\theta)$, while the observed frequencies were written $p_1, \ldots, p_k$. So in a sample size n

the likelihood is proportional to $L_n(\pi) = \pi_1^{np_1} \ldots \pi_k^{np_k}$. The maximum likelihood equations are then written

$$\frac{p_1}{\pi_1} \frac{d\pi_1}{d\theta} + \ldots + \frac{p_k}{\pi_k} \frac{d\pi_k}{d\theta} = 0 .$$

For a multivariate parameter we can easily write these by considering

$$\psi(y,\theta) = \begin{cases} \frac{1}{\pi_1(\theta)} \nabla\pi_1(\theta) & y \leq y_1 \\ \frac{1}{\pi_j(\theta)} \nabla\pi_j(\theta) + \left(\frac{y-y_j}{y_{j+1}-y_j}\right) \frac{1}{\pi_{j+1}(\theta)} \nabla\pi_{j+1}(\theta) & y_j \leq y \leq y_{j+1},\ 1 \leq j \leq k-1 \\ \frac{1}{\pi_k(\theta)} \nabla\pi_k(\theta) & y \geq y_k \end{cases} \qquad (4.10)$$

and they are then written $\int \psi(x,\theta) dF_n(x) = 0$. We let F_θ be the distribution function attributing weight $\pi_j(\theta)$ to the point y_j, $1 \leq j \leq k$. Rao made assumptions

RA1: The expression

$$I(\theta_o,\theta) = - \sum_{i=1}^{k} \pi_i(\theta_o) \log \frac{\pi_i(\theta)}{\pi_i(\theta_o)} , \quad (I \geq 0)$$

which provides the average amount of discrimination between the multinomial distribution defined by θ and the true one defined by θ_o, is bounded away from zero $\|\theta - \theta_o\| > \delta$ for each $\delta > 0$,

RA2: $\pi_1(\theta),\ldots,\pi_k(\theta)$ have continuous partial derivatives of the second order in a neighbourhood of the true value θ_o,

RA3: $\pi_j(\theta_o) \neq 0$ for each j, and $(d\pi_j(\theta)/d\theta) \neq 0$ for at least one j. (As a consequence of this assumption $I(\theta_o) = (d/d\theta)K_{F_{\theta_o}}(\theta)\Big|_{\theta=\theta_o}$, which is Fisher's information at θ_o is $\neq 0$),

RA4: $\pi_i(\theta) = \pi_i(\xi)$ for all i implies $\theta = \xi$.

Under RA1-RA4 Rao showed existence of a neighbourhood of the true proportions $\pi(\theta_o)$, say $N(\pi(\theta_o))$, and a positive δ such that $p \in N(\pi(\theta_o))$ implies

(i) There exists one and only one root $\hat{\theta}$ of the likelihood equation which differs from the true value of θ_o by less than δ. This root, as a function of the relative frequencies, is continuous at $\pi(\theta_o)$ where it tends to θ_o and is Fisher consistent.

(ii) $\hat{\theta}$ is Fréchet differentiable (with respect to (F, d_k)).

(iii) $\hat{\theta}$ is the unique M.L. estimate and is therefore the M.L.E. estimate.

The proof is given by Rao for $\Theta \subset E$. Using M-estimation theory we extend the proof to $\Theta \subset E^r$ and Fréchet differentiability is given with respect to the wider class G'. We replace RA2 and RA3 by

RA2': $\pi_1(\theta), \ldots, \pi_k(\theta)$ have continuous partial derivatives of the second order in a neighbourhood of the true value θ_o and are uniformly bounded away from zero on Θ.

RA3': $\pi_j(\theta_o) \neq 0$ for each j, and

$$\nabla K_{F_{\theta_o}}(\theta_o) = -I(\theta_o) = \int \{\nabla \psi(x, \theta_o)\} dF_{\theta_o}(x) = -\sum_{j=1}^{k} \frac{\{\nabla' \pi_j(\theta_o)\}\{\nabla \pi_j(\theta_o)\}}{\pi_j(\theta_o)}$$

is nonsingular, and hence negative definite.

Let $\bar{U}_{\delta_1}(\theta_o) \subset \Theta$. By continuity of the partial derivatives and RA3', there exists a constant H so that

$$\sup_{\theta \in \bar{U}_{\delta_1}(\theta_o)} \sup_{1 \leq j \leq k} \left\| \frac{1}{\pi_j(\theta)} \nabla \pi_j(\theta) \right\| < H < \infty .$$

Then

$$\sup_{\theta \in \bar{U}_{\delta_1}(\theta_o)} \left\| \frac{\partial \psi}{\partial y}(y,\theta) \right\| \leq \max_{1 \leq j \leq k-1} \frac{1}{y_{j+1}-y_j} \cdot H ,$$

and so by Lemma 1.4 $\{\psi(\cdot,\theta) \mid \theta \in \bar{U}_{\delta_1}(\theta_o)\}$ is an equicontinuous family uniformly bounded by some constant. Similarly

$$\sup_{\theta \in \bar{U}_{\delta_1}(\theta_o)} \sup_{1 \leq j \leq k} \left\| -\frac{\{\nabla' \pi_j(\theta)\}\{\nabla \pi_j(\theta)\}}{\pi_j(\theta)^2} + \frac{1}{\pi_j(\theta)} \nabla' \nabla \pi_j(\theta) \right\| < H' < \infty ,$$

and so the family $\{\nabla\psi(\cdot,\theta) \mid \theta \in \bar{U}_{\delta_1}(\theta_o)\}$ is equicontinuous and bounded by a constant also. Conditions A are then satisfied with respect to the Kolmogorov metric and with $D = \bar{U}_{\delta_1}(\theta_o)$, $\delta_1 > 0$. Then (i) can be seen directly from the remarks following Lemma 2.5, while (ii) follows as a consequence of Theorem 4.3, since (4.5) is easily checked when $G \in \mathcal{G}'$. Hence the estimate $\hat{\theta}$ is asymptotically normal. Finally we can choose the selection statistic $f_n(\underset{\sim}{x}^{(n)},\theta) = -L_n(\pi(\theta))$, which satisfies

$$f_n(\underset{\sim}{x}^{(n)},\theta) = -\sum_{j=1}^{k} p_j \log \pi_j(\theta) \xrightarrow{a.s.} -\sum_{j=1}^{k} \pi_j(\theta_o) \log \pi_j(\theta_o) = f_{F_{\theta_o}}(\theta).$$

The convergence is uniform in $\Theta \subset E^r$ provided the $\pi_j(\theta)$ are uniformly bounded away from zero. Since it is true that by RA1 that

$$\inf_{\|\theta-\theta_o\|>\delta} f_{F_{\theta_o}}(\theta) - f_{F_{\theta_o}}(\theta_o) = \inf_{\|\theta-\theta_o\|} \sum_{j=1}^{k} \pi_j(\theta_o) \log \frac{\pi_j(\theta_o)}{\pi_j(\theta)} > 0,$$

part (iii) of the Theorem is confirmed by Lemma 2.6.

§4.4 The Influence Curve

Relationships between Fréchet differentiability and robustness follow from the fact that Fréchet differentiability of the M-functional implies boundedness of the influence curve, defined by Hampel (1974) as follows:

DEFINITION 4.2: Let T be a vector valued functional of real numbers, defined on some subset of the set of all distribution functions on R for which T is defined. Denote by δ_x the distribution function determined by the point mass one at any given point $x \in R$. Mixtures of G and some δ_x are written as $(1-\varepsilon)G + \varepsilon\delta_x$, for $0 < \varepsilon < 1$. Then the influence curve $IC_{TG}(\cdot)$ of ("the estimator") T at ("The underlying distribution function") G is defined pointwise by

$$IC_{TG}(x) = \lim_{\varepsilon \downarrow o} \{T[(1-\varepsilon)G+\varepsilon\delta_x] - T[G]\}/\varepsilon ,$$

provided that the limit is defined for every point $x \in R$.

In some situations we will consider it sufficient to show the existence of the limit at all bar a finite number of points.

Existence of the Fréchet derivative with respect to (G,d) at F_{θ_o} where d is either Kolmogorov or Prokhorov metric, implies that for the gross error model

$$n(\varepsilon,F_{\theta_o}) = \{G|G = (1-\delta)F_{\theta_o} + \delta H,\ H \in G,\ 0 \leq \delta < \varepsilon\}$$

$$T[G]-T[F_{\theta_o}] = \int IC_{TF_{\theta_o}}(x)dG(x) + o(\varepsilon)$$

$$= \varepsilon \int IC_{TF_{\theta_o}}(x)dH(x) + o(\varepsilon) .$$

In that case

$$b_1(\varepsilon) = \sup_{G \in n(\varepsilon,F_{\theta_o})} \|T[G] - T[F_{\theta_o}]\| = \varepsilon\gamma^* + o(\varepsilon) ,$$

with $\gamma^* = \sup_x \|IC_{TF_{\theta_o}}(x)\|$. We call this the gross error sensitivity. Hampel (1968) used the equivalent Euclidean norm, where $\|a\|$ is given as the $\sup|a|$, where $|a|$ is the vector of absolute values of components of a. Hence he defines gross error sensitivity $\gamma^* = \sup_x \sup|IC_{TF_{\theta_o}}(x)|$.

Huber admonishes us, pointing out that if only the weaker Gateaux derivative is available, then there exist examples where

(i) $\gamma^* < \infty$ but $b_1(\varepsilon) \equiv \infty$ for $\varepsilon > 0$

(ii) $\gamma^* = \infty$ but $\lim b(\varepsilon) = 0$ for $\varepsilon \to 0$

The influence curve exists for the M-functional under weaker conditions on ψ than those imposed to show Fréchet differentiability with respect to the metrics. But the mode of proof is the same.

THEOREM 4.4:

For $x \in R$ assume there exists a set $D = D_x$ such that conditions A of §2.3 hold for the starlike neighbourhoods $n_x(\varepsilon, F_{\theta_o})$. Then the influence curve $T[\psi,\cdot]$ at F_{θ_o} exists and is given by

$$C_{T[\psi,\cdot],\theta_o}(x) = -M(\theta_o)^{-1}\psi(x,\theta_o) .$$

PROOF: Letting $\{\varepsilon_k\}$ be so that $\varepsilon_k \downarrow 0$ then for $G_{\varepsilon_k} = (1-\varepsilon_k)F_{\theta_o} + \varepsilon_k\delta_x$

$$\|M(\theta, G_{\varepsilon_k}) - M(\theta)\| = \varepsilon_k\|M(\theta) - \nabla\psi(x,\theta)\| \to 0 . \qquad (4.11)$$

From (4.7)

$$T[G_{\varepsilon_k}]-\theta_o = -M(\theta_o)K_{G_{\varepsilon_k}}(\theta_o) + M(\theta_o)^{-1}(M(\tilde{\theta}_k, G_{\varepsilon_k}) - M(\theta_o))(T[G_{\varepsilon_k}]-\theta_o) ,$$

and since $K_{G_{\varepsilon_k}}(\theta_o) = \varepsilon_k\psi(x,\theta_o)$

$$T[G_{\varepsilon_k}]-\theta_o = -M(\theta_o)\varepsilon_k\psi(x,\theta_o) + o(\varepsilon_k) .$$

So

$$\lim_{k\to\infty} \{T[(1-\varepsilon_k)F_{\theta_o}+\varepsilon_k\delta_x]-\theta_o\}/\varepsilon_k = -M(\theta_o)\psi(x,\theta_o) .$$

Sometimes there can exist a few points $x \in R$ at which $\psi(x,\theta)$ has not a continuous partial derivative at θ_o, but the proof holds at

all other points of the observation space.

Those M-functionals at which $\gamma^* = +\infty$ are considered non-robust. Those that are Frechet differentiable necessarily have finite γ^* and the M-estimators are asymptotically normal with covariance matrix

$$\sigma^2(\psi, F_{\theta_o}, \theta_o) = \mathrm{var}_{F_{\theta_o}}(IC_{TF_{\theta_o}}(X)) \ .$$

CHAPTER 5

QUANTITATIVE AND QUALITATIVE CRITERIA

§5.1 Sensitivity and Breakdown Verses Efficiency

Weak continuity and Fréchet differentiability are useful notions in studying the robustness of the M-functional in infinitesimal departures from the model F_{θ_o}. Kallianpur and Rao (1955) show that if the M.L.E. is Frechet differentiable, then it is asymptotically efficient among the class of Fréchet differentiable functionals (with respect to (F,d_k)). It is then the natural choice of estimator. Under weak regularity conditions on ψ the M-estimator for the invariate parameter is asymptotically normal under a parametric family F. If F is absolutely continuous with respect to measure μ on R with a corresponding family of density functions, $\{f_\theta | \theta \in \Theta\}$ whose support is independent of θ, then $M(\theta)$ defined in conditions A is in fact given by

$$M(\theta) = -\text{cov}_{F_\theta} [\psi(X,\theta), \nabla \log f_\theta(X)] .$$

From this, or directly from the Cramer-Rao bound, we obtain the inequality

$$\sigma^2(\psi,F_\theta,\theta) \geq I(\theta)^{-1} \tag{5.1}$$

where

$$I(\theta) = \int \{\nabla \log f_\theta(x)\}^2 f_\theta(x) d\mu(x) < \infty ,$$

is the Fisher Information Matrix. The inequality (5.1) concerns the diagonal elements only in the multivariate case. For the univariate parameter the asymptotic efficiency is defined

$$e(\theta) = \frac{M(\theta)^2}{I(\theta)\text{var}_{F_\theta}\{\psi(X,\theta)\}} . \tag{5.2}$$

Then $0 \leq e(\theta) \leq 1$ for any M-estimator. The M.L.E. attains efficiency one. But for parameters of normal families, the influence curve of the M.L.E. which corresponds to the efficient score is unbounded. Then the M.L.E. is neither Fréchet differentiable, nor robust. Normal families are so often used for the ostensible reason that normality frequently occurs in nature. Realistically they present a convenient mathematical representation. It is not necessary to abandon normality, but perhaps to abandon the M.L.E., if circumstances warrant it.

Hampel (1968) introduced the measures of sensitivity and breakdown to compare M-estimators in departures from a model. In 1971 he formalized a definition of breakdown of an estimator functional T as follows:

DEFINITION 5.1: Let $\{T_n\}$ be a sequence of estimators. The breakdown point δ_H of $\{T_n\}$ at some probability measure G is defined

$$\delta_H = \delta_H(\{T_n\},G) = \sup\{\delta \leq 1 \mid \text{there exists a compact set } K = K(\delta) \text{ which is a proper subset of the parameter space so that } F \in n_p(\delta,G) \text{ implies } F\{T_n \in K\} \to 1 \text{ as } n \to \infty\}.$$

LEMMA 5.1: If ψ,f determine an M-functional let

$$\delta_B = \{\delta \leq 1 \mid \text{there exist a compact set } K = K(\delta) \text{ which is a proper subset of the parameter space such that } F \in n_p(\delta,G) \text{ implies } T[\psi,f,\cdot] \text{ is weakly continuous at } F, \text{ and } T[\psi,f,F] \in K^{int}\}.$$

Then $\delta_B \leq \delta_H(\{T[\psi,f,F_n]\},G)$.

PROOF: The proof follows from Proposition 4.1 and consistency.

Another definition of breakdown is found in Huber (1977). His criticism of previous breakdown point investigations was that the convergence $F\{T_n \in K\} \to 1$ need not be uniform in $F \in n_p$. Unfortunately

it is sometimes impractical to evaluate δ_B and δ_H and even more so this latter breakdown. If it were known that there existed at most a unique solution of the equations (2.8) then δ_B would be the smallest neighbourhood of G such that $H(\psi,F) = \phi$, for some F in the neighbourhood. But in the estimation of scale using Huber's Proposal 2 it was possible to observe that $\delta_B = 1$ even though there existed a sequence of distributions $\{G_n\}$ for which $T[\psi,G_n] \to \infty$ and $d_p(G_n,G) \to 1$. What is necessary is an indicator of behaviour of the M-functional on neighbourhoods of a distribution G that can be readily computed.

DEFINITION 5.2: Let $\{T_n\}$ be a sequence of estimators. The local breakdown point δ_L of $\{T_n\}$ with respect to the triple (G,K,n), where $G \in \mathcal{G}$, K is a subset of the parameter space, and n is a neighbourhood centered at G, is

$$\delta_L = \delta_L(\{T_n\},G,K,n) = \sup\{\delta \leq 1 \mid F \in n(\delta,G) \text{ implies } F[T_n \in K] \to 1 \text{ as } n \to \infty\}.$$

This can often be readily computed in the ε-contaminated neighbourhoods for the univariate M-functional (see §6.3).

§5.2 Redescending Influence Functions

At the root of understanding robustness is the question; "What is the significance of $T[G]$ when $G \notin \mathcal{F}$?" This depends to a large extent on the nature of the departure from $\mathcal{F}$, and on $\mathcal{F}$ itself. The most easily represented departure is ε-contamination

$$G_o = (1-\varepsilon)F_{\theta_o} + \varepsilon H \tag{5.3}$$

where $0 < \varepsilon < 1$ is small and H varies in $\mathcal{G}$. Setting $H = \delta_x$ and letting $x \to \infty$ simulates the behaviour of T in the presence of grossly erroneous observations. For the representation (5.3), $\|T[G] - T[F_{\theta_o}]\|$

denotes the "asymptotic bias". As it turns out breakdown bounds and asymptotic bias using either ε-contamination or Prokhorov neighbourhoods are determined at a particular θ_o. They do not naturally transfer to the whole family F as for the location family which is an exception.

This realization suggests that if we are to guard against a particular kind of contamination H we may need to proceed adaptively in $\theta \in \Theta$. Our object then is to gather information that contributes to F_{θ_o} but to screen out observations that fall in regions corresponding to "tails" of F_{θ_o} Since θ_o is not known we consider a construction for general θ. Then since the estimator is assumed close to θ_o continuity of F_θ in the parameter leads to approximately the right action on the "tail regions". For the moment we define the selection statistic (functional) to be $f(\theta) = \|\theta - \theta_o\|$ for a departure of the form (5.3). It is necessary to make a judicious choice of ψ for both asymptotic efficiency at the model and asymptotic bias away from the model. We link the observation space to the parameter space in the following manner.

DEFINITION 5.3: The set of null influence $N(\psi,\theta) \subset R$ associated with an influence function ψ is defined

$$N(\psi,\theta) = \{x \in R | \psi(x,\theta) = 0\} .$$

Observe that if G_o is given by (5.3) where $\int_{N(\psi,\theta_o)} dH = 1$, $\theta_o = T[\psi,F_{\theta_o}]$, and $M(\theta_o)$ is nonsingular, then it is true that $T[\psi,G_o] = \theta_o$ and $M(T[\psi,G_o],G_o) = (1-\varepsilon)M(\theta_o)$. For suitably regular ψ there exists an asymptotically unique consistent sequence of roots $\{T[\psi,F_n]\}$ to θ_o for which $\sqrt{n}(T[\psi,F_n] - \theta_o)$ is asymptotically normal with variance covariance matrix $\sigma^2(\psi,G_o,T[\psi,G_o]) = (1-\varepsilon)^{-1}\sigma^2(\psi,F_{\theta_o},\theta_o)$. Then the only departure in the usual convergence at the model is an increase in asymptotic variance of the order $(1-\varepsilon)^{-1}$.

The optimal choice for $N(\psi,\theta)$ is a tail region of F_θ, where a single observation is least likely to fall. It can be employed also according to the sample size. For instance it may be possible to choose $N_n(\psi,\theta)$ so that

$$1 - P_\theta\{X^{(n)} \in (R - N_n(\psi,\theta))^n\} = .05 \quad n = 1,2,\ldots \tag{5.4}$$

Finally, note that the influence curve is zero on $N(\psi,\theta)$.

It was remarked by Hogg (1977) that the notion of the redescending influence function and asymptotic efficiency one were not necessarily incompatible. His illustration was the M.L.E. for location of a Cauchy distribution, where

$$\psi(x-\theta) = \frac{2(x-\theta)}{1 + (x-\theta)^2}$$

which tends to zero as $|x-\theta| \to \infty$. The contrast is the M.L.E. for location of a normal distribution for which

$$\psi(x-\theta) = x-\theta \ .$$

To determine an optimality criterion that balances efficiency and sensitivity Hampel (1968) provided his Lemma 5. There is a simple extension of it that includes the notion of a set of null influence. Since we have only a single parameter we use the abbreviation $\dot{\psi}(x,\theta) = \nabla \log f_\theta(x)$.

LEMMA 5.2: Let $f_\theta(x)$ be from a model family of densities with respect to (σ-finite) measure μ (on R) and let $\theta \subset E$. Densities are assumed regular in the sense that each density is positive on (closed) support S (not depending on the parameter θ), and $\{\nabla f_\theta(x) | \theta \in \Theta\}$ are measureable on S, zero elsewhere, with $\int \nabla f_\theta(x) d\mu(x) = 0$ and $I(\theta) = \int \dot{\psi}(x,\theta)^2 f_\theta(x) d\mu(x) < \infty$. Set $N(\theta) \subset R$ to be a specified set of null influence for which $\int_{R-N(\theta)} f_\theta(x) d\mu(x) > 0$; let $b(\theta) > 0$ be some constant; define

$$\psi^*_\alpha(x,\theta) = \begin{cases} \min(|\dot\psi(x,\theta)-\alpha|, b(\theta))\operatorname{sign}(\dot\psi(x,\theta)-\alpha) & x \in R-N(\theta) \\ 0 & \text{otherwise.} \end{cases}$$

Then there is an $\alpha(\theta)$ so that $\int \psi^*_{\alpha(\theta)}(x,\theta) f_\theta(x)d\mu(x) = 0$. Define $\tilde\psi(x,\theta) = \psi^*_{\alpha(\theta)}(x,\theta)$ and $c(\theta) = \int \tilde\psi(x,\theta)\nabla f_\theta(x)d\mu(x)$. Assuming $c(\theta) \neq 0$, then $\tilde\psi(\cdot,\theta)$ minimizes $\sigma^2(\psi,F_\theta,\theta)$ among all ψ with $\int \psi(x,\theta)\nabla f_\theta(x)d\mu(x) \neq 0$, and with the same upper bound

$$k(\theta) = b(\theta)/\left|\int \tilde\psi(x,\theta)\nabla f_\theta(x)d\mu(x)\right|$$

for the sensitivity.

PROOF: For brevity we drop the function arguments where possible and let $\nabla f_\theta(x) = f'$. By the dominated convergence theorem $\int \psi^*_\alpha f d\mu$ is a continuous function of α, and as $\alpha \to \pm\infty$, this inegral tends to $\mp\, b\int_{R-N} f\,d\mu$; hence there is an $\alpha(\theta)$ with $\int \psi^*_{\alpha(\theta)} f\,d\mu = 0$. Without loss of generality we can assume $\int \psi f\,d\mu = 0$ and $\int \psi f'\,d\mu = c$ so that we have to consider $\int \psi^2 f\,d\mu$. Then

$$\begin{aligned}\int [(\dot\psi-\alpha(\theta))-\psi]^2 f\,d\mu &= \int (\dot\psi-\alpha(\theta))^2 f\,d\mu - 2\int \psi\dot\psi f\,d\mu + 2\int \alpha(\theta)\psi f\,d\mu \\ &\quad + \int \psi^2 f\,d\mu \\ &= \int (\dot\psi-\alpha(\theta))^2 f\,d\mu - 2\int \psi f'\,d\mu + 2\alpha(\theta)\int \psi f\,d\mu \\ &\quad + \int \psi^2 f\,d\mu \\ &= \int (\dot\psi-\alpha(\theta))^2 f\,d\mu - 2c + 0 + \int \psi^2 f\,d\mu\,.\end{aligned}$$

On the other hand

$$\int [(\dot\psi-\alpha(\theta))-\psi]^2 f d\mu = \int_N (\dot\psi-\alpha(\theta))^2 f d\mu + \int_{(R-N)\cap\{|\dot\psi-\alpha(\theta)|>b\}} (\dot\psi-\alpha(\theta)-\psi)^2 f d\mu$$

$$+ \int_{(R-N)\cap\{|\dot\psi-\alpha(\theta)|<b\}} (\dot\psi-\alpha(\theta)-\psi)^2 f d\mu$$

and since $|\psi| \le b$ this is minimized when $\psi = \tilde\psi$ whence the lemma follows.

COROLLARY 5.1: If $N(\theta) = \phi$ then $c(\theta) > 0$, and the resulting Lemma corresponds to Lemma 5 of Hampel (1968).

COROLLARY 5.2: If $b(\theta) = +\infty$, then

$$\tilde{\psi}(x,\theta) = \begin{cases} \dot{\psi}(x,\theta) - P(\theta) & x \in R - N(\theta) \\ 0 & x \in N(\theta) \ , \end{cases}$$

where $P(\theta) = \int_{R-N} \dot{\psi}(x,\theta) f_\theta(x) d\mu(x) \Big/ \int_{R-N} f_\theta(x) d\mu(x)$, and whenever

$$\int \tilde{\psi}(x,\theta) \nabla f_\theta(x) d\mu(x) \neq 0 \ .$$

REMARK 5.1: Care should be observed in applying the resulting $\tilde{\psi}$. Discontinuities may preclude use of presented asymptotic normality theorems. It is clear though that one can obtain a sufficiently smooth sequence of functions $\{\psi_j\}$ that satisfy $\psi_j(\cdot,\theta) \overset{j}{\to} \tilde{\psi}(\cdot,\theta)$ a.e. μ as $j \to \infty$. Then $\int_R \psi_j^2 f \, d\mu \to \int_{R-N} \tilde{\psi}^2 f \, d\mu$ and $\int \psi_j f' \, d\mu \to \int \tilde{\psi} f' \, d\mu$, so that $\sigma^2(\psi_j, F_\theta, \theta) \xrightarrow{j\to\infty} \sigma^2(\tilde{\psi}, F_\theta, \theta)$.

EXAMPLE 5.1: For the normal location family setting $N(\theta) = \phi$, and $b(\theta) = c$ reveals the solution corresponding to Huber's (1964) minimax psi-function, $\psi(x) = \min(\max(-c,x),c)$.

EXAMPLE 5.2: Setting $b(\theta) = +\infty$ and $N(\theta) = (-\infty,-c+\theta) \cup (c+\theta,\infty)$ for the normal location family gives the influence function of Example 2.2. There, consistency of $\hat{\theta}_n = \theta(F_n)$ was established. Also observe $K_n(\hat{\theta}_n) \xrightarrow{a.s.} K_\phi(0) = 0$. Without loss of generality suppose $\hat{\theta}_n < 0$. By expanding $\psi(x-\hat{\theta}_n) = \psi(x) + \psi'(\xi)\hat{\theta}_n$ in the region $[-c,\hat{\theta}_n+c]$, and noting $\psi'(\xi) = 1$, there exists the corresponding expansion

$$K_n(\hat{\theta}_n) = K_n(0) + n^{-1}\left\{\sum_{i=1}^{n} I_{[-c,\hat{\theta}_n+c]}(X_i)\right\}(\hat{\theta}_n-0)$$

$$+\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\theta}_n)I_{[\hat{\theta}_n-c,-c]}(X_i) - \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\theta}_n)I_{[\hat{\theta}_n+c,c]}(X_i). \quad (5.5)$$

For consistent $\hat{\theta}_n$ the two latter terms are $o_p(n^{-\frac{1}{2}})$. Supposing that $\sqrt{n}\,K_n(\hat{\theta}_n) \xrightarrow{a.s.} 0$, asymptotic normality follows from the C.L.T. for $\sqrt{n}\,K_n(0)$. Then $\sqrt{n}\,\hat{\theta}_n$ is asymptotically normal with the usual asymptotic variance,

$$\sigma^2(\psi,\Phi,0) = \int \psi^2(x)d\Phi(x)/\left(\int \psi'(x)d\Phi(x)\right)^2 .$$

The two latter terms of the expansion (5.5) are composed of those observations that contribute greatest towards the values of $K_n(\hat{\theta}_n)$ and $K_n(0)$. Observations X_i near the underlying location parameter have little weight. This suggests that contamination about cut off points $\pm c$ can be a vital concern for behaviour of the estimator. We choose the null set $N(\theta) \subset R$ to dampen the asymptotic bias given a particular kind of contamination. Choosing an upper bound $b(\theta)$ for the influence curve keeps asymptotic variances stable in small departures from the model. A combination of the two would reduce the sensitivity of the estimator to contaminations near $\partial N(\theta)$.

Estimators based on linear combinations of order statistics and rank statistics do not necessarily behave in the same manner as the M-estimator in departures from the model. Huber (1977, P.24) has noted that using the influence curve as a tool to identify robust estimators at the parametric family F may not be adequate. The influence curve should also be examined under departures from F. If the set of null influence, $N(\theta)$, for an L-statistic (linear combination of order statistics) is determined by its influence curve at F_θ it is most likely that it will not correspond to the set of null influence at G_o given by (5.3),

even if $\int_{N(\theta)} dH = 1$. That is when an observation falls into the null set it is still taken into account to determine the ordering of the other observations.

§5.3 Multiparameter Models

Construction of practical estimators that are both robust and efficient for multivariate models is governed by a number of complexities. Increasing the number of parameters can increase the possibility of zeros in the estimating equations. All zeros must be searched for numerically before a selection is made. This involves time consuming numerical searching algorithms. Robustness requires bounded influence curves but there is no natural choice for such, even in attempting to trade off efficiency with sensitivity.

Hampel (1978, P.436) published an incomplete result extending Corollary 5.1 to the multivariate parameter. To be realistic an analogous proof to that of Lemma 5.2 or its corollaries in the multivariate parameter case would necessarily assume $M(\theta)$ to be a constant matrix. This is generally not possible. But in estimation of location and scale where the influence function takes the form $\psi(x,\theta) = \psi(\frac{x-\mu}{\sigma})$, it is possible to write the matrix

$$M(\mu,\sigma) = \sigma^{-1}\begin{bmatrix} \int \psi_1'(x)dF_o(x) & 0 \\ 0 & \int x\psi_2'(x)dF_o(x) \end{bmatrix}.$$

Here $\psi = (\psi_1,\psi_2)'$ and F_o is the model distribution from which the family F is derived. Given a vector $\underset{\sim}{b}(\mu,\sigma)$ the asymptotic variance is minimized among all the estimators with the same upper bound for sensitivity

$$\underset{\sim}{k}(\mu,\sigma) = |M(\mu,\sigma)^{-1}b(\mu,\sigma)|,$$

where $|\cdot|$ represents the vector of absolute values of the elements, the

bound being elementwise. Huber's (1964) Proposal 2 for estimation of location and scale of the normal distribution given by

$$\psi_1(x) = \min(\max(-c,x),c) \quad \text{and} \quad \psi_2(x) = \psi_1^2(x)-\beta_c^2, \text{ where}$$

$$\beta_c = \int \psi_1^2(x)d\Phi(x)$$

minimizes the asymptotic variance of all M-estimators with upper bound $k(\mu,\sigma)$ given by

$$b(\mu,\sigma) = (c,(c^2-1)(2\Phi(c)-1) + 2c\phi(c)) \ .$$

Any other M-estimators whose influence functions are so that their corresponding value of $k(\mu,\sigma)$ is bounded above by that of Proposal 2 have asymptotic variance that is greater than or equal to that of Proposal 2. Alternatively if we set an upper bound on the sensitivity (Hampel's 1968 definition)

$$\gamma^*(\mu,\sigma) = \gamma.\sigma$$

for some constant γ, the Winsorizing constants of ψ_1 and ψ_2 that minimize the asymptotic variance, are different. They are given by the equations

$$\gamma = \frac{c_1}{2\Phi(c_1)-1} = \frac{(c_2^2-1)(2\Phi(c_2)-1) + 2c_2\phi(c_2)}{2.(2.\Phi(c_2)-1) - 4c_2\phi(c_2)} \ .$$

The choice of estimator is still left fairly open with no clear guidance for the choice of γ. This is without considering the possibility of a set of null influence.

With more complex parametric families the requirement of bounded influence curves is satisfied by some minimal distance estimates. It will become clear in Ch. 6 that efficiency at the model as a sole criterion must be abandoned, and that estimators with only average efficiency can be quite attractive under small departures from the model.

Then the main factor is the cost in employing the minimal distance estimate, particularly if the estimator is to be used on small samples where it may be necessary to account for small sample bias. Implementing bias reduction procedures, such as the jackknife of Quenoville (1956) may be costly due to the nature of implicitly defined M-estimates. Nevertheless Reeds (1978) demonstrates that asymptotics of the jackknifed estimates are equivalent to the M-estimates. But often it can be easier to correct for the small sample bias adopting the approach of Cox and Hinkley (1974, P.310). Taking the Taylor expansion approach the bias of the M-estimate is given by

$$E[\hat{\theta}_n - T[G]] = \left. \frac{-[2\int \psi(x,\theta)\nabla\psi(x,\theta)dG(x) + \int \psi^2(x,\theta)dG(x) \int \nabla^2\psi(x,\theta)dG(x)]}{2n[\int \nabla\psi(x,\theta)dG(x)]} \right|_{\theta=T[G]}$$

$$+ o(n^{-1})$$

$$= \frac{b(\theta,G)}{n} + o(n^{-1}) \ . \qquad (5.6)$$

The bias is estimated by $b(\hat{\theta}_n, F_{\hat{\theta}_n})/n$, whereby we adhere to the assumption of the model, at least approximately.

CHAPTER 6

APPLICATIONS TO LOCATION AND SCALE ESTIMATION

§6.1 Theory for Location M-estimates

The theory of M-estimates of location of symmetric distributions has been well developed since Huber's (1964) introduction of the minimax M-estimator. Specifically M-estimates of location are defined as solutions of

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0 , \tag{6.1}$$

where $\psi(x)$ is assumed to be an odd function in x.

Equation (6.1) is not scale invariant in the sense that the solution formed from taking a multiple of the sample can be different from that using the original sample. Assuming a standard normal distribution is unrealistic and a scale estimate is often used. A typical statistic for that purpose is the minimum absolute deviation estimator

$$d = \text{med}(|X_i - \text{med}(X_i)|)/.6745 .$$

The M-estimator then solves

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \theta}{d}\right) = 0 .$$

The value .6745 provides d as a consistent estimator when the underlying distribution is $\Phi(x/\sigma)$. With σ known $\psi(x) = \min(\max(-1.5,x),1.5)$ gives efficiency greater than 95%. If σ is unknown the asymptotic distribution of the M-estimator depends also on the statistic d. Then one can only speculate through behaviour of the strict location estimate with σ-known on the behaviour or optimality of a particular criterion (c.f. §5.3).

Other authors recognizing that observations that do not belong to the normal distribution should be neglected introduce redescending influence functions. Some examples are:

1. Hampel psi-function

$$\psi_{a,b,c}(x) = \begin{cases} x & 0 \le |x| \le a \\ a \text{ sign } (x) & a \le |x| \le b \\ \frac{c-|x|}{c-b} a \text{ sign } (x) & b \le |x| \le c \\ 0 & |x| \ge c \end{cases}$$

2. Wave of Andrews

$$\psi(x) = \begin{cases} \sin (x/c) & |x| < c\pi \\ 0 & |x| > c\pi \end{cases}$$

3. Biweight of Tukeys

$$\psi_{BS}(x) = \begin{cases} x(1-(x/c)^2)^2 & |x| \le c \\ 0 & |x| > c \ . \end{cases}$$

Examples 1. and 3. have set of null influence $(-\infty,-c] \cup [c,\infty) \cup \{0\}$, while that of example 2. is $(-\infty,-c\pi] \cup [c\pi,\infty) \cup \{0\}$. Hogg (1979) provides the following suggested values

1.) $a = 1.7$, $b = 3.4$, $c = 8.5$.

2.) $c = 1.5$ or 2.

3.) $c = 5.$ or 6.

For 3., $c = 4.685$ determines 95% efficiency with σ known.

If contamination were restricted to the set of null influence, ideally the choice of null set would exclude all observations that fell more than just over three standard deviations from the location. In a sample of size 30 generated from the normal distribution the probability that the largest observation is greater than 3.5 is $1 - \Phi(3.5)^{30} \doteqdot .007$. It is extremely unlikely to obtain an observation in $(-\infty,-3.5] \cup [3.5,\infty)$.

But all parameter values given above weight observations in that region. Admittedly the weight is exceedingly diminished from that which is attributed by the M.L.E. ($\psi(x) = x$). However it could be argued that it is preferable not to weight these aberrant observations at all. Large parameter values of c were born out of the Monte Carlo studies of Andrews et al (1972) where the apparent concern is for both asymptotic variance and small sample variance in symmetric departures from the model. With large samples asymptotic bias becomes the predominant concern and smaller values of c would be more in harmony with the notion of a model normal distribution.

Perhaps most criticism of redescending psi-functions has been that they allow multiple roots to the estimating equation. Hampel (1974) suggests iterating a few times from an initial consistent estimate of the location, proposing the median to be an appropriate starting point. This gives rise to a natural question; "Will such an iteration take us closer to the M-estimator or in fact take us farther away?" We answer the question partially with the following analysis.

Consider the classical Newton-Raphson iteration when solving the equation

$$f(x) = 0 \ . \tag{6.2}$$

Starting with x_0 as an initial estimate it takes successive estimates by setting

$$x_{\nu+1} = x_\nu - \frac{f(x_\nu)}{f'(x_\nu)} \qquad \nu = 0,1,2,\ldots \ . \tag{6.3}$$

For a particular root ξ of (6.2) we want to know of a region about ξ in which the method eventually converges to ξ for any choice of x_0 in the region. A rate of convergence is also of interest. Ostrowski (1960, P.44) describes an answer to both of these points in

his Theorem 7.2. The proof of that theorem can be carried through assuming f is continuously differentiable with piecewise continuous second partial derivative.

LEMMA 6.1: Let $f(\xi) = f''(\xi) = 0$, $f'(\xi) = \lambda_o > 0$. Suppose $\{f_n(x)\}$ is a sequence of functions satisfying

(1) $f_n(x) \to f(x)$, $f_n'(x) \to f'(x)$, $f_n''(x) \to f''(x)$, uniformly in $x \in E$ as $n \to \infty$. Then there exists an open neighbourhood $N(\xi)$, such that for sufficiently large n (f.s.l.n.) there is a root ξ_n in $N(\xi)$ of $f_n(x) = 0$. Further starting with any $x_o \in N(\xi)$, the sequence $x^{(n)}$ formed by the recurrence formula

$$x_{\nu+1}^{(n)} = x_\nu^{(n)} - \frac{f_n(x_\nu^{(n)})}{f_n'(x_\nu^{(n)})} \qquad \nu = 0,1,2,.... \tag{6.5}$$

all lie in $N(\xi)$ and we have

$$x_\nu^{(n)} \longrightarrow \xi_n \qquad (\nu \to \infty) .$$

Further given any $\varepsilon > 0$ there exists a neighbourhood $N(\xi,\varepsilon)$ with the property that f.s.l.n.

$$\text{(a)} \quad \frac{|x_{\nu+1}^{(n)} - x_\nu^{(n)}|}{|x_\nu^{(n)} - x_{\nu-1}^{(n)}|^2} \leq \varepsilon \qquad (\nu = 1,2,....) .$$

PROOF: By continuity choose $0 < \delta < \lambda_o/3$ so that

$$\text{(2)} \quad \frac{12}{13}\lambda_o < f'(x) < \frac{14}{13}\lambda_o \qquad x \in (\xi-\delta,\xi+\delta)$$

$$|f''(x)| < 1 \qquad x \in (\xi-2\delta,\xi+2\delta).$$

By (1) the inequalities (2) are true when f is replaced by f_n for all n greater than some $n(\delta)$. Let $0 < \kappa < 1/3$ and choose $n(\delta,\kappa) \geq n(\delta)$ so that $n > n(\delta,\kappa)$ implies $\xi-\kappa\delta < \xi_n < \xi+\kappa\delta$ (cf. proof of Lemma 2.7). Choose $x_o^{(n)}(\eta) = \xi_n - \eta$ for any $0 < |\eta| < (1-\kappa)\delta$ say.

Observe by the mean value theorem

$$\frac{12}{13}\lambda_o\eta < f_n(\chi_o^{(n)}(\eta)) < \frac{14}{13}\lambda_o\eta .$$

Then it is true that

$$\frac{6}{7}\eta < \frac{f_n(\chi_o^{(n)}(\eta))}{f_n'(\chi_o^{(n)}(\eta))} < \frac{7}{6}\eta \quad \text{uniformly in} \quad n \geq n(\delta,\kappa) . \quad 0 < |\eta| < (1-\kappa)\delta$$

We set $h_o^{(n)}(\eta) = f_n(\chi_o^{(n)}(\eta))/f_n'(\chi_o^{(n)}(\eta))$ and

$$J_o^{(n)}(\eta) = [\chi_o^{(n)}(\eta),\ \chi_o^{(n)}(\eta) + 2h_o^{(n)}(\eta)] \subset (-2\delta,2\delta) .$$

Hence

$$\sup_{x\in J_o^{(n)}(\eta)} |f_n''(x)| \leq \sup_{\xi-2\delta<x<\xi+2\delta} |f_n''(x)| \leq 1 .$$

So

$$2.h_o^{(n)}(\eta)\sup_{x\in J_o^{(n)}(\eta)} |f_n''(x)| < 2.\frac{7}{6}\eta.1 < \frac{7}{3}\delta < \frac{7}{9}\lambda_o < \frac{12}{13}\lambda_o < |f_n'(\chi_o^{(n)}(\eta))|$$

uniformly in $n \geq n(\delta,\kappa)$, $0 < |\eta| < (1-\kappa)\delta$. By Theorem 7.2 of Ostrowski we can set

$$N(\xi) = (\xi - (1-2\kappa)\delta,\ \xi + (1-2\kappa)\delta) ,$$

since observe that $(\xi-\kappa\delta,\xi+\kappa\delta) \subset N(\xi)$ and further given any $\xi_n \in (\xi-\kappa\delta,\xi+\kappa\delta)$ all points of $N(\xi)$ are contained in the interval $(\xi_n - (1-\kappa)\delta,\xi_n + (1-\kappa)\delta)$.

We now observe the last part of the Lemma. Since δ may be chosen arbitrarily small, assume

$$|f''(x)| < \lambda_o \frac{24}{13}.\varepsilon \quad \text{uniformly in} \quad (\xi-2\delta,\xi+2\delta) .$$

For $n \geq n(\delta,\kappa)$ assume the corresponding inequality holds with f replaced by f_n. Then

$$M_n(\eta) = \sup_{x \in J_o^{(n)}(\eta)} |f_n''(x)| < \sup_{\xi-2\delta<x<\xi+2\delta} |f_n''(x)| < \frac{\lambda_o .24}{13} \varepsilon$$

uniformly in $n \geq n(\delta,\kappa)$ and $0 < |\eta| < (1-\kappa)\delta$. Then letting $x_\nu^{(n)}(\eta)$ be the sequence generated by (6.5) with initial starting point $x_o^{(n)}(\eta)$ we have the result from (a) and (b) of Theorem 7.2 of Ostrowski by observing

$$\frac{M_n(\eta)}{2.|f_n'(x_\nu^{(n)}(\eta))|} < \frac{\lambda_o \frac{24}{13} \varepsilon}{2. \frac{12}{13} \lambda_o} = \varepsilon .$$

The Lemma is proved.

What has been shown is the existence of a region about the zero of f for which any initial estimate chosen in this region will lead to the Newton-Raphson iteration to converge to the zero of f_n for n large enough. A quadratic rate of convergence holds uniformly in n for large n. But the region in which convergence occurs can be explored further. The next Theorem pertains more closely to the M-estimating equations.

THEOREM 6.1:

Let $f(\xi) = 0$, $f'(\xi) = \lambda_o < 0$ and set $\{f_n\}$ to be a sequence of continuous functions with piecewise continuous derivatives satisfying

(3) $\quad f_n(x) \to f(x)$ uniformly in $x \in E$

$\quad f_n'(x) \to f'(x)$ uniformly in $x \in E$,

where $f_n'(x)$ is the left hand derivative assumed to exist for all $x \in E$. Given x_o we set $x_o^{(n)} = x_o$ for all n and define $\{x_\nu^{(n)}\}_{\nu=0}^{\infty}$ to be sequences defined by (6.5). Suppose $\delta > 0$ is such that f.s.l.n., $x_o \in U_\delta(\xi)$ implies the sequence $\{x_\nu^{(n)}\}_{\nu=0}^{\infty}$ converges to the root $\xi_n \in U_\delta(\xi)$ of $f_n(x) = 0$. For some $\varepsilon > 0$ suppose there exists an $x^* > \delta$ for which

$$2(x-\xi) - \frac{f(x)}{f'(x)} < -\varepsilon \qquad x \in [\xi-\chi^*,\xi-\delta] \tag{6.6}$$

$$2(x-\xi) - \frac{f(x)}{f'(x)} > \varepsilon \qquad x \in [\xi+\delta,\xi+\chi^*]$$

$$f'(x) < -\alpha(\chi^*) < 0 \qquad x \in [-\chi^*,\chi^*]$$

Then f.s.l.n. $\chi_o \in [\xi-\chi^*,\xi+\chi^*]$ implies the sequence $\{\chi_\nu^{(n)}\}_{\nu=0}^{\infty}$ converges to $\xi_n \in U_\delta(\xi)$ which is unique in $[\xi-\chi^*,\xi+\chi^*]$.

PROOF: F.s.l.n. it is true that $f_n'(x) < -\alpha/2$ for all $x \in [\xi-\chi^*,\xi+\chi^*]$. Since $f(\xi+\delta) < -\alpha\delta$, $f(\xi-\delta) > \alpha\delta$, then f.s.l.n. $f_n(\xi+\delta) < -\alpha\delta/2$, $f_n(\xi-\delta) > \alpha\delta/2$, and hence also

$$\frac{f_n(x)}{f_n'(x)} > \frac{\alpha\delta/2}{\alpha} = \delta/2 \qquad x \in [\xi+\delta,\xi+\chi^*] \tag{6.7}$$

$$\frac{f_n(x)}{f_n'(x)} < -\delta/2 \qquad x \in [\xi-\chi^*,\xi-\delta] .$$

Also f.s.l.n.

$$\left|\frac{f_n(x)}{f_n'(x)} - \frac{f(x)}{f'(x)}\right| < \varepsilon/2 . \tag{6.8}$$

Given n for which all the above inequalities hold, consider $\chi_\nu^{(n)} \in [\xi+\delta,\xi+\chi^*]$. By (6.7)

$$\chi_{\nu+1}^{(n)} = \chi_\nu^{(n)} - \frac{f_n(\chi_\nu^{(n)})}{f_n'(\chi_\nu^{(n)})} < \chi_\nu^{(n)} - \delta/2 .$$

By (6.6) and (6.8)

$$\chi_{\nu+1}^{(n)} = \chi_\nu^{(n)} - \frac{f_n(\chi_\nu^{(n)})}{f_n'(\chi_\nu^{(n)})} > 2\xi - \chi_\nu^{(n)} + \varepsilon/2 .$$

So

$$-(\chi_\nu^{(n)}-\xi) + \varepsilon/2 < (\chi_{\nu+1}^{(n)}-\xi) < (\chi_\nu^{(n)}-\xi) - \delta/2 .$$

That is

$$|x_{\nu+1}^{(n)}-\xi| < |x_{\nu}^{(n)}-\xi| + \min(\varepsilon,\delta)/2 \ . \tag{6.9}$$

The inequality (6.9) is also true when $x_{\nu}^{(n)} \in [\xi-x^*,\xi-\delta]$. Therefore there exists a $\nu_o = \nu_o(n)$ for which

$$x_{\nu_o}^{(n)} \in [\xi-x^*,\xi-\delta] \cup [\xi+\delta,\xi+x^*]$$

and

$$x_{\nu_o+1}^{(n)} \in (\xi-\delta,\xi+\delta) \ .$$

The theorem is proved.

EXAMPLE 6.1: Consider the domain of convergence of the Tukey Bi-weight M-estimating equations. Necessarily the statements "for sufficiently large n" are interpreted as probability statements "for all sufficiently large n" or "with probability going to one".

Observe that if the underlying distribution G is symmetric about some value ξ, letting

$$K_G(\theta) = \int \psi_{BS}(x-\theta)dG(x)$$

it is true that

$$K_G(\xi) = K_G''(\xi) = 0 \ , \ K_G'(\xi) = -\int \psi_{BS}'(x-\xi)dG(x) < 0 \ . \tag{6.10}$$

Let us write

$$\psi_{BS}(x-\theta) = \begin{cases} (x-\theta)(1-(x-\theta)^2/c^2)^2 & -c \leq x-\theta \leq c \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \left(-\theta + \frac{2\theta^3}{c} - \frac{\theta^5}{c^4}\right) + x\left(1 - \frac{6}{c^2}\theta^2 + \frac{5\theta^4}{c^4}\right) \\ + x^2\left(\frac{6\theta}{c^2} - \frac{10\theta^3}{c^4}\right) + x^3\left(-\frac{2}{c^2} + \frac{10\theta^2}{c^4}\right) \\ + x^4\left(\frac{-5\theta}{c^4}\right) + x^5\left(\frac{1}{c^4}\right) & \theta - c \le x \le \theta + c \\ 0 & \text{otherwise} . \end{cases}$$

Note that $\psi_{BS}(x)$ is continuously differentiable with piecewise continuous second partial derivatives. Hence so also is $K_n(\theta) = \int \psi_{BS}(y-\theta)dF_n(y)$. If we consider any closed interval $C = [a,b]$, it is not hard to construct uniformly continuous functions $\{h_j(x)\}_{j=o}^{5}$ so that

$$h_j(x) = x^j \qquad x \in [a-c, b+c] , \quad j = 0,\ldots,5$$

and

$$\sup_{x \in E} |h_j(x)| = \max_{x \in [a-c,b+c]} |x^j| .$$

The family $\mathcal{A} = \{h_j(x) \mid 0 \le j \le 5\}$ will be both bounded and equicontinuous. By Theorem 1.1

$$\sup_{x \in E} \sup_{0 \le j \le 5} \left| \int_{-\infty}^{x} h_j(y)dF_n(y) - \int_{-\infty}^{x} h_j(y)dG(y) \right| \xrightarrow{a.s.} 0 .$$

But this implies

$$\sup_{\theta \in [a,b]} \sup_{0 \le j \le 5} \left| \int_{\theta-k}^{\theta+k} y^j dF_n(y) - \int_{\theta-k}^{\theta+k} y^j dG(y) \right| \xrightarrow{a.s.} 0 .$$

More precisely, for any compact set $[a,b]$ it is true that

$$K_n(\theta) \xrightarrow{a.s.} K_G(\theta) \quad \text{uniformly in} \quad \theta \in [a,b] .$$

Similarly we can apply this approach to the derivatives $K_n'(\theta)$, $K_n''(\theta)$ so that

$$K_n(\theta) \xrightarrow{a.s.} K_G(\theta),\ K_n'(\theta) \xrightarrow{a.s.} K_G'(\theta),\ K_n''(\theta) \xrightarrow{a.s.} K_G''(\theta),\ \text{uniformly on } C\ . \tag{6.11}$$

From (6.10) and (6.11) we see that $K_n(\theta)$ satisfies the requirements of Lemma 6.1. That is there exists a domain of attraction for the Newton-Raphson iteration. Given an $x_o \in (\xi-\delta,\xi+\delta)$ the sequence $\{\chi_\nu^{(n)}\}$ will converge f.a.s.1.n. to the unique $\xi_n \in (\xi-\delta,\xi+\delta)$ for which $K_n(\xi_n) = 0$. To explore this convergence further we need only investigate the nature of the function

$$S_G(\theta) = 2(\theta-\xi) - \frac{K_G(\theta)}{K_G'(\theta)}$$

so as to apply Theorem 6.1. The function $S_G(\theta)$ is even about ξ. Letting $G = \Phi$, the function $S_\Phi(\theta)$ is explored for four different values of c. For each of these $S_\Phi(0) = 0$, $S_\Phi'(0) > 0$. Designating $\theta^* = \theta^*(c)$ the minimum positive value θ for which $S_\Phi(\theta) = 0$ the following values were obtained

c	θ^*	c	θ^*
3.5	1.218	5.	1.674
4.685	1.574	6.	1.999

For each value of c, $K_\Phi'(\theta) < -\alpha(c) < 0$ for all $\theta \in [-\theta^*(c),\theta^*(c)]$ and some $\alpha(c)$. Also $S_\Phi'(\theta^*) < 0$ for each c. Hence given $0 < \delta < \theta(c)$, by Lemma 6.1 any value $x^* \in [\delta,\theta(c))$ will satisfy the criterion of Theorem 6.1 with

$$\varepsilon = \tfrac{1}{2} \min_{\theta \in [\delta,x^*]} S_\Phi(\theta) > 0\ .$$

Then for any $0 \leq \theta' < \theta^*$, $[-\theta',\theta']$ is a domain of convergence for the Newton-Raphson iteration on the estimating equations (6.1).

For large n any initial estimate within one standard deviation

of the true location will ensure convergence of the Newton-Raphson iteration to the unique M-estimator. Using a consistent estimator for the initial estimate the rate of convergence of the iteration will be made arbitrarily large with increasing n. This suggests for large samples, only a few iterations from a consistent estimate are sufficient to retain reasonable accuracy to the M-estimator. Asymmetric departures or even departures from normality in the underlying distribution can make these observations redundant. But boundedness of the influence function and its derivatives would also make the domain of convergence robust against small departures from normality. In particular for contamination $G = (1-\varepsilon)\Phi + \varepsilon H$, where H attributes zero weight to $[-c-\theta^*(c), c+\theta^*(c)]$ observe that $K_G(\theta) = (1-\varepsilon)K_\Phi(\theta)$ for all $-\theta^* \leq \theta \leq \theta^*$. So $K_G(0) = K_G''(0) = 0$ and $K_G'(0) = (1-\varepsilon)K_\Phi'(0) < 0$. The domain of convergence of the population (ε small) will remain the same since $S_G(\theta) = S_\Phi(\theta)$ on $[-\theta^*, \theta^*]$.

While ψ_{BS} has only piecewise continuous second partial derivatives it is clear that families $\{\psi_{BS}(\cdot-\theta) \mid \theta \in E\}$, and $\{\psi_{BS}'(\cdot-\theta) \mid \theta \in E\}$ are both equicontinuous and bounded above by a constant. Hence the uniform convergence of the resulting M-estimator in the underlying distribution is assured. This completes the example.

§6.2 Identification and Goodness of Fit; A Graphical Approach

Uniform convergence over the whole parameter space does not only lend itself to limit theorems and convergence criteria that identify the M-estimator. On estimation it is necessary to determine the adequacy of the model family F in explaining the data. Having guarded against slight contamination it is still important to know whether the underlying mechanism conforms or departs radically from the estimated distribution. In large samples the geometry of curves $K_n(\theta)$ approximates that of

$K_{G_o}(\theta)$. Since G_o is estimated by $F_{T[\psi,F_n]}$ a plot of $K_n(\theta)$ and $K_{F_{T[\psi,F_n]}}(\theta)$ can be advantageous as a graphical measure of goodness of fit. Examples $\psi_{a,b,c}$ and ψ_{BS} are known to have good efficiency within small symmetric perturbations from the normal distribution. But to make inferences based on asymptotics at the normal distribution the bulk of the population should be compared for its normal nature. Hampel (1978, P.427) supports this in his statement; "In careful and high-quality samples of measurement data from astronomy and geodesy, without any visible gross errors, typically have a higher kurtosis than the normal distribution . . . . .". The following example illustrates the graphical procedure and also shows how the M-estimator is identified in small samples using the global consistency argument of Theorem 2.3.

EXAMPLE 6.2: Consider the Hampel psi-function estimating equations

$$K_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \psi_{a,b,c}(X_t - \theta) = 0 \ . \tag{6.12}$$

Then

$$\begin{aligned}(d/d\theta)^{-}K_n(\theta) &= n^{-1}\left[\frac{a}{c-b}\{\#X_t's \in (\theta-c,\theta-b] \cup (\theta+b,\theta+c]\}\right. \\ &\qquad \left. --\{\#X_t's \in (\theta-a,\theta+a]\}\right] \\ &= \left[\frac{a}{c-b}\{(F_n(\theta-b) - F_n(\theta-c)) + (F_n(\theta+c) - F_n(\theta+b))\}\right. \\ &\qquad \left. - \{F_n(\theta+a) - F_n(\theta-a)\}\right] \\ &\xrightarrow{a.s.} K'_{G_o}(\theta) = \int (\partial/\partial\theta)^{-}\psi_{a,b,c}(x-\theta)dG_o(x)\end{aligned}$$

uniformly in $\theta \in E$, by the Glivenko-Cantelli theorem. The statistician who proposes the model $F = \{\Phi(x-\theta) \mid \theta \in E\}$, may choose

$$\ell = \int \psi'_{a,b,c}(x)d\Phi(x) \ ,$$

where ℓ is as in (d) of §2.4. Note $K_{\Phi'}(0) < 0$. For any sample of size n there exist roots of equations (6.12) for all $\theta \leq X_{(1)} - c$, $\theta \geq X_{(n)} - c$, where $X_{(i)}$ is the i'th order statistic. There exists one root between. The first two sets of roots are excluded from $H_n(\psi_{a,b,c}; s.\ell)$, $0 < s < 1$, since $|(d/d\theta)^{-}K_n(\theta)| < s.\ell$.

Two samples, one of size 10 and one of size 60 were generated from a double exponential distribution located at the origin. Solutions $\theta(F_n)$ were found for the parameters $(a,b,c) = (2.,3.,3.5)$. Assuming normality the value $\ell = -.9456$. For the sample of size ten, -3.61235, -1.39213, -1.18903, -.52096, -.15880, .02580, .084251, .44745, 6.10150, 7.95135, three solutions between $X_{(1)} - 3.5 = -7.11$ and $X_{(10)} + 3.5$ $X_{(10)} + 3.5 = 11.45$ were found. They were $\theta_1(F_{10}) = -.6719$, $\theta_2(F_{10}) = 3.5192$, and $\theta_3(F_{10}) = 7.0264$ respectively. Corresponding values of $(d/d\theta)^{-}K_n(\theta)$ evaluated at θ_1, θ_2, θ_3 were respectively -.7, 1.2, and -.2. According to Theorem 2.3 this determines $T[\psi,F_{10}] = -.6719$. Plots can also be used to indicate which root is the M-estimator, namely the root that yields the closest fit to $K_n(\theta)$ of

$$K_{\theta_i}(\theta) = \int \psi_{a,b,c}(x-\theta)d\Phi(x-\theta_i) .$$

The sample of size 60 yields only the one root, $T[\psi,F_{60}] = -.0447$ between $X_{(1)} - 3.5 = -7.84$ and $X_{(60)} + 3.5 = 11.45$, at which the derivative $(d/d\theta)^{-}K_n(\theta) = -.7667$. There is still a relatively large discrepancy between this value and $(d/d\theta)K_{\Phi}(\theta)\big|_{\theta=0} = -.9456$. This follows from the wrong selection of model. This is made apparent by diagram 6.2 where it is observed that the empirical curve follows the curve $K_{G_o}(\theta)$ more closely than the fitted curve $K_{F_{T[\psi,F_n]}}(\theta)$. We dub the curves "expectation curves".

DIAGRAM 6.1

Plots of "expectation curve", $K_n(\theta)$, $K_{G_o}(\theta)$, and $K_{\theta_i}(\theta)$, $i = 1,2,3$, where G_o is the double exponential distribution

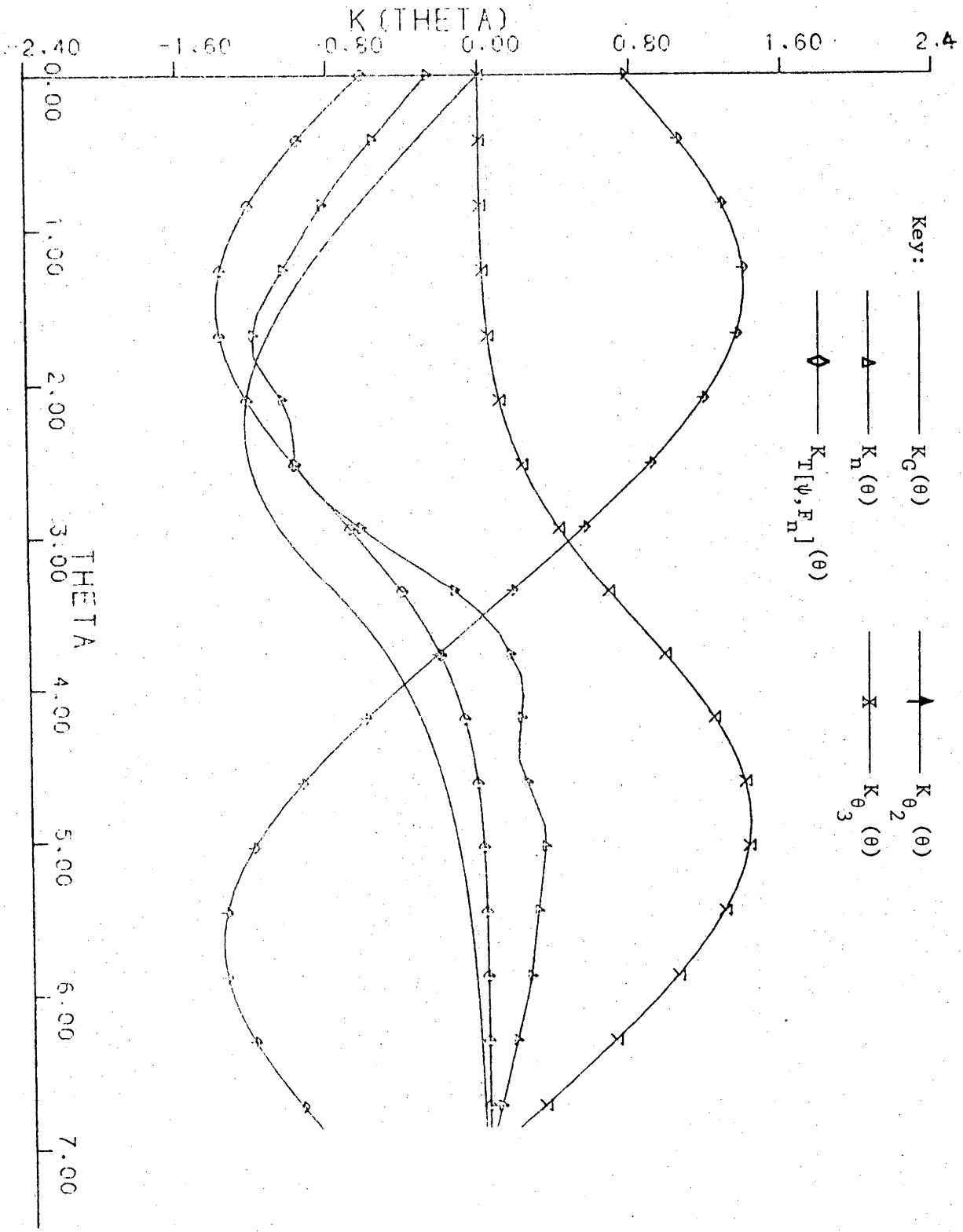
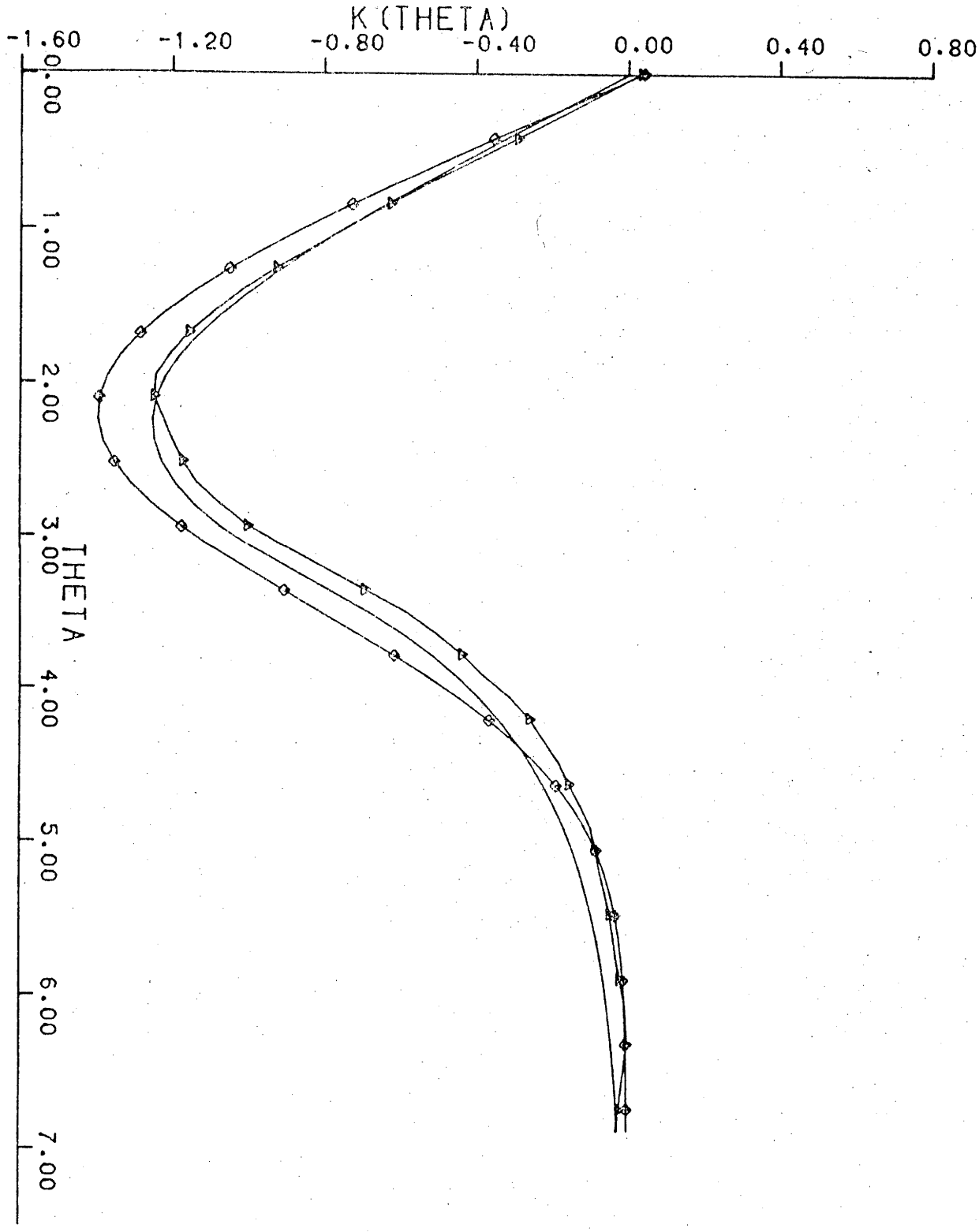


DIAGRAM 6.2

A graphical goodness of fit test.

Expectation curves of $K_n(\theta)$, $K_{F_{T[\psi,F_n]}}(\theta)$, and $K_{G_o}(\theta)$



§6.3 Robust Estimation of Scale

Estimation of scale has always proven difficult in that there is no natural symmetry in which to appeal to, as in the symmetric location situation. The scale parameter may be regarded as a measure of dispersion. However, in general, such measures are purely arbitrary. Indeed Bickel and Lehman (1976) proposed three alternatives to the standard deviation,

$$SD(G) = [\int (x-\mu)^2 dG(x)]^{\frac{1}{2}} ,$$

where μ is the expectation of the random variable X with distribution function G.

Huber (1972) considered estimation of scale by taking logarithms, reducing the problem to estimation of location. But again this results in highly asymmetric distributions not easily handled by the symmetric theory of location. Recently Thall (1979) extended this approach giving a minimax theorem for estimation of scale when departure is of the form $\{G | \sup_x |G(x) - F(x)| < \varepsilon\}$. The parametric family was $F = \{F(x/\sigma) | \sigma > 0\}$. The approach necessarily neglects any asymptotic bias incurred.

Influence functions of the form $\psi(x/\sigma)$ ensure that $T[\psi, F(\cdot/\sigma_o)] = \sigma_o T[\psi, F(\cdot)]$, for $\sigma_o > 0$. They allow for scale transformations of the data. Perhaps an obvious question to ask for symmetric distributions F defined on E is; "Does there exist a class of psi-functions, C, for which one retains consistency to the parameter σ_o whenever the uncertainty is allowed to vary in the form

$$G(x) = (1-\varepsilon)F(x/\sigma_o) + \varepsilon H(x) \quad \text{for all} \quad \sigma_o > 0 , \tag{6.13}$$

H being from the class of symmetric distributions about zero?" The answer is no. Since to retain consistency it is necessary that $K_G(\sigma_o) = 0$. However, this would mean ψ has to be odd in x, in which

case

$$(\partial/\partial\sigma)K_{F_{\sigma_o}}(\sigma)\Big|_{\sigma=\sigma_o} = \int -\frac{x}{\sigma^2}\,\dot{\psi}(\frac{x}{\sigma})dF(x/\sigma_o)\Big|_{\sigma=\sigma_o} = 0 .$$

Then we lose consistency under the model regardless. A true minimax theorem should weight both asymptotic bias and variance in its measure of loss. Experimental results lead us more quickly to insight into estimation of scale.

When $F = \Phi$, the M.L.E. of scale is $T[\psi,F_n] = \sum_{i=1}^{n} X_i^2/n$, and

$$\psi(x/\sigma) = \{(\partial/\partial\sigma)\phi(x;\sigma)\}^{-1}\phi(x;\sigma) ,$$

is the efficient score. Here $\phi(x;\sigma) = (\sqrt{2\pi}\ \sigma)^{-1}\exp(-x^2/2\sigma^2)$ is the normal density with standard deviation σ, and $\psi(x) = -1+x^2$. With the model of indeterminacy (6.13) it is necessary to dampen down values of $\psi(X_i/\sigma_o)$ when $\phi(X_i;\sigma_o)$ is too small relative to $\phi(x;\sigma_o)$ in a region judged to be where the bulk of the probability lies. Such a region is clearly related to the contours of the density function, and in that way is heuristically linked to the maximum likelihood method of estimation. Hence the Huber Proposal 2 is a natural construction. Alternatively, arguing that given a model normal distribution observations outside the region of approximately three times the scale occur with negligible probability we propose the following redescending influence function. Let

$$\psi_{a,b,c}(x) = \begin{cases} -1 + x^2 - P & 0 \le |x| \le a \\ -1 + a^2 - P & a \le |x| \le b \\ (-1 + a^2 - P)\,\dfrac{c-|x|}{c-b} & b \le |x| \le c \\ 0 & c \le |x| < \infty . \end{cases} \qquad (6.14)$$

The parameters $0 < a \leq b < c$ are given and the parameter P satisfies

$$\int \psi_{a,b,c}(x)\phi(x;1)dx = 0 .$$

The corresponding set of null influence is then $N(\psi_{a,b,c};\sigma) = \{\pm\sqrt{1+P} \cup (-\infty,-c\sigma] \cup [c\sigma,\infty)\}$. The effect of such a construction in small departures from the model is clearly demonstrated by the asymptotic values in Table 6.1 for symmetric contamination $H = \Phi(x/\alpha\sigma_o)$. Asymptotics of the M.L.E. are easily derived since $\psi(x) = -1+x^2$ gives $K_G(\sigma) = -1+\sigma^{-2}\{(1-\varepsilon)\sigma_o^2 + \varepsilon\alpha^2\sigma_o^2\}$, whence $T[\psi,G] = \sigma_o\{(1-\varepsilon)+\varepsilon\alpha^2\}^{\frac{1}{2}} = \sigma_1$, say. The asymptotic variance for the M.L.E. is then

$$\sigma^2(\psi,G,\sigma_1) = \frac{\sigma_o^4}{3\sigma_1^2}\left[3\{(1-\varepsilon)+\varepsilon\alpha^4\} - \{(1-\varepsilon)+\varepsilon\alpha^2\}^2\right] .$$

Since for each choice of ψ, $K_G(\sigma)$ is a function of σ/σ_o, if $\sigma_o = 1$ yields bias $\sigma^*(1)$ then for arbitrary $\sigma_o > 0$ the bias $\sigma^*(\sigma_o) = \sigma_o\sigma^*(1)$, (keeping ε, α, and ψ fixed).

Table 6.1 shows that minute contamination of .1% and .5% perturbs the asymptotic variance of the M.L.E. alarmingly and includes also an asymptotic bias. This throws into doubt the applicability of the M.L.E. to small samples on the basis of efficiency, the latter evaluated at the true parametric model. On the other hand Proposal 2 and the redescending type estimators (R.E.'s) remain relatively unperturbed.

When contamination of 10% is present, which is more the rule than the exception as noted by Hampel (1978, P.427), the M.L.E. behaves disastrously in both bias and variance. Redescending influence functions perform favourably in that asymptotic bias from the heavy tailed contamination is negligible and asymptotic variances remain near values at the true model; a property which is desirable for inference.

TABLE 6.1

Asymptotic bias and variance for maximum likelihood, Proposal 2, and redescending type estimators of scale with $G=(1-\varepsilon)\Phi(x)+\varepsilon\Phi(x/3)$

ε		Bias	Var	12	Bias	Var
0		0.	.5	Redescending	0.	1.02
.001	M.L.E.	.004	.55	a=1.645 b=2.4	0.	1.02
.005		.020	.75	c=3.0	.002	1.02
.1		.342	3.3		.049	1.01
0		0.	.64		0.	.79
.001	Proposal 2	.001	.64	Redescending	0.	.79
.005	c=1.645	.005	.65	a=1.645 b=2.	.002	.79
.1		.111	.79	c=3.3	.049	1.05
0		0.	.57		0.	.95
.001	Proposal 2	.001	.57	Redescending	0.	.95
.005	c=1.96	.006	.58	a=1.96 b=2.5	.002	.95
.1		.130	.73	c=3.0	.005	.93

Application of the Leibnitz rule on the partial derivative of $K_G(\sigma)$ when ψ is given by Huber's Proposal 2 gives

$$(\partial/\partial\sigma)K_G(\sigma) = -\sigma^3 \int_{-c\sigma}^{c\sigma} x^2 dG(x) < 0 . \qquad (6.15)$$

So there can exist at most a single root to $K_G(\sigma) = 0$. The greatest effect on the asymptotic variance expression $\sigma^2(\psi,G,\sigma)$, through contaminating H affecting the denominator $(\partial/\partial\sigma)K_G(\sigma) = M(\sigma,G)$, is when H attributes all of its weight to the region $(-\infty,-c\sigma] \cup [c\sigma,\infty)$. Then $(\partial/\partial\sigma)K_G(\sigma) = (1-\varepsilon)M(\sigma)$.

For redescending influence functions the rate at which the influence function redescends proves important in guarding against worst possible contamination. For symmetric contamination H

$$(\partial/\partial\sigma)K_G(\sigma) = \frac{2}{\sigma^2}\{(1-\varepsilon)\sigma_o[-\sigma_o/\sigma + \frac{-1+a^2-P}{c-b}(\phi(\frac{\sigma b}{\sigma_o};1)-\phi(\frac{\sigma c}{\sigma_o};1))]$$

$$+ \frac{\varepsilon}{\sigma}[-\int_{-\infty}^{+\infty} x^2 dH(x) + \sigma(\frac{-1+a^2-P}{c-b})\int_{b\sigma}^{c\sigma} x dH(x)]\} . \qquad (6.16)$$

Since $P < -1+a^2$, the major effect of an H on $(\partial/\partial\sigma)K_G(\sigma)$ is seen by maximizing the functional

$$f(b,c,d;H) = -\int_{-\infty}^{\infty} x^2 dH(x) + d\int_b^c x dH(x) ,$$

over $H \in G$, when parameters b,c,d are assumed fixed positive constants.

LEMMA 6.2: The functional $f(b,c,d;H)$ is maximized over all $H \in G$ in the regions $d < b$, $b < d \leq 2b$, $2b \leq d \leq 2c$, and $2c \leq d < \infty$ respectively by the Heaviside functions at 0, b, $d/2$, and c. In each case it is the unique $H \in G$ at which the maximum is achieved. The maximum values attained are respectively 0, $-b(b-d)$, $d^2/4$, and $-d(c-d)$. When $d = b$ distributions $(1-\eta)\delta_o + \eta\delta_b$, $0 \leq \eta \leq 1$ attain the maximum value of zero.

PROOF: For any $H \in G$ let $H = H_1 + H_2$ be the decomposition of H into components so that the support of H_1 is contained in $\{0\} \cup [b,c]$ and the support of H_2 is $E - \{\{0\} \cup [b,c]\}$. Then

$$f(b,c,d;H) = f(b,c,d;H_1) + f(b,c,d;H_2) .$$

For a distribution H_1 with support $\{0\} \cup [b,c]$, $f(b,c,d,H_1) = \int_b^c x(d-x)dH_1(x)$. If $d < b$ then $x(d-x) < 0$ uniformly in $x \in [b,c]$. So f is maximized by the Heaviside function δ_o for

which $f(b,c,d,\delta_o) = 0$. If $b < d \leq 2b$, then $b(d-b) > x(d-x)$ uniformly in $x \in \{0\} \cup (b,c]$. Then f is maximized uniquely by the distribution function δ_b, whence $f(b,c,d;\delta_b) = b(d-b)$. Similarly $x = d/2$ and $x = c$ maximize $x(d-x)$ on $\{0\} \cup [b,c]$ when $2b \leq d \leq 2c$ and $2c \leq d < \infty$ respectively. Distribution functions $\delta_{d/2}$ and δ_c maximize f on those regions. When $d = b$ the maximum of $x(d-x)$ is attained at $x = 0$ and $x = b$, implying distributions $(1-\eta)\delta_o + \eta\delta_b$ maximize f, and the maximum is zero.

Note that in each case the maximum value attained is ≥ 0. For any H, a positive measure of finite total variation with support $E - (\{0\} \cup [b,c])$, $f(b,c,d;H) < 0$. This completes the Lemma.

The maximum non-negative contribution to (6.15) from any contamination is then

$$\frac{2\varepsilon}{\sigma^3} \max_{H \in G} f(b\sigma, c\sigma, d\sigma; H) = \frac{2\varepsilon}{\sigma} \max_{H \in G} f(b,c,d;H), \text{ where}$$

$$d = \frac{-1+a^2-P}{c-b}.$$

This maximum attains its least value when $\frac{-1+a^2-P}{c-b} \leq b$, suggesting a reasonable separation of values b and c to be appropriate.

TABLE 6.2

Redescending Estimators

	a	b	c	$\max_{H \in G} f(b,c,d;H)$
1.	1.645	2.	3.3	0.
2.	1.645	2.4	3.	1.6
3.	1.96	2.5	3.	8.7
4.	2.	2.9	3.1	39.5

A premium is paid in loss of efficiency, although relatively small, in order to attain non-positive contribution from the functional f. This is compensated for by asymptotic bias in departure from the model if c is made too large in attempting to reduce d. Steep descents, as in 4., are particularly susceptible to increased asymptotic variance when contamination appears near tails of the distribution $\Phi(x/\sigma_o)$, that is near $c\sigma_o$.

Non positive contribution from f has the added advantage for $0 \leq \varepsilon < 1$ and all $H \in G$

$$(\partial/\partial\sigma)K_G(\sigma) \leq \frac{2}{\sigma^2}(1-\varepsilon)\sigma_o[-\sigma_o/\sigma + \frac{-1+a^2-P}{c-b}(\phi(\frac{\sigma b}{\sigma_o};1) - \phi(\frac{\sigma c}{\sigma_o};1))]$$

$$< 0 \quad \text{uniformly in} \quad \sigma \in (0,\infty),\ a = 1.645,\ b = 2.,\ c = 3.3$$

Then there exists at most one root $T[\psi,G]$.

Setting $\varepsilon = \varepsilon^*$ and $G = (1-\varepsilon^*)\Phi(x/\sigma_o) + \varepsilon^* H(x)$, $H \in G$, the M.L.E. fails completely in the presence of heavy tailed contamination. But a lower bound for $T[\psi,G]$ is attained when $H = \delta_o$. This indicates the behaviour of the M.L.E. for scale when accumulated "super efficient" observations near the true location parameter are present. For Proposal 2 monotonicity, (6.15), and the inequality

$$(1-\varepsilon^*)\int \psi(x/\sigma)\phi(x;\sigma_o)dx - \varepsilon^*\beta_c^2 \leq K_G(\sigma) \leq (1-\varepsilon^*)\int \psi(x/\sigma)\phi(x;\sigma_o)dx + \varepsilon^*(c^2-\beta_c^2)\ , \tag{6.17}$$

ensure that regardless of contamination H

$$\sigma_\ell(\varepsilon^*) \leq T[\psi,G] \leq \sigma_u(\varepsilon^*)\ , \tag{6.18}$$

where σ_ℓ and σ_u are the unique zeros of the lower and upper bounds on $K_G(\sigma)$. For a R.E.

$$(1-\varepsilon^*)\int \psi_{a,b,c}(x/\sigma)\phi(x;\sigma_o)dx - \varepsilon^*(1+P) \leq K_G(\sigma)$$

$$\leq (1-\varepsilon^*)\int \psi_{a,b,c}(x/\sigma)\phi(x;\sigma_o)dx + \varepsilon^* d \ . \qquad (6.19)$$

While K_G need not be monotonic in σ, numerical investigations give unique lower and upper bounds for zeros σ_ℓ, and σ_u. So any solution of $K_G(\sigma) = 0$ is bounded by them. The value of ε^* at which we cease to retain a zero in the region $\sigma/\sigma_o < 1.5$ to the upper bounds of (6.17) and (6.19) is denoted a local breakdown point. Proposal 2, as expected, compares favourably in regard to this pessimistic approach of worst possible contamination. The sensitivity, a measure of robustness in infinitesimal departures, reveals analogous results.

TABLE 6.3

Bounds on asymptotic bias in ε-contaminated neighbourhoods, breakdown points, and sensitivities

Estimator	$\varepsilon = .01$ σ_ℓ	σ_u	$\varepsilon = .1$ σ_ℓ	σ_u	Sensitivity γ^*	Breakdown $\delta^*(1.5)$
M.L.E.	.995	∞	.949	∞	∞	0.
Proposal 2, c=1.645	.993	1.02	.922	1.22	1.67	.176
Proposal 2, c=1.96	.994	1.02	.934	1.29	2.03	.138
Redescending 1.	.992	1.02	.910	1.24	1.96	.165
Redescending 2.	.991	1.02	.912	1.23	1.87	.169
Redescending 3.	.993	1.02	.926	1.32	2.29	.133
Redescending 4.	.993	1.02	.931	1.32	2.22	.132

The influence curve of the M.L.E. is a quadratic, indicating the scale parameter to be extremely sensitive to heavy tailed contamination. This is indicated clearly, for with $H = \Phi(x/\alpha\sigma_o)$, $T[\psi,G] = \sigma_o\sqrt{(1-\varepsilon)+\varepsilon\alpha^2}$. This is perturbed upward from σ_o for $\alpha > 1$. In contrast

$\sigma_\ell(\varepsilon^*) = \sigma_o\sqrt{1-\varepsilon^*}$. Contamination by super efficient observations is not a problem. But for the more robust estimators that guard against heavy tails, there is a greater sensitivity to observations near the origin. Nevertheless the values of $|\sigma_\ell - \sigma_o|$ remain well below $\sigma_u - \sigma_o$.

§6.4 M-Estimators of the Exponential Distribution Parameter

The most important application of the sole estimation of scale is in the case of the exponential distribution used to describe lifetime data. The density is regular with $R = E$, $S = E^+ \cup \{0\}$, $\Theta = E^+$, and μ taken to be Lebesgue measure so that $f_\theta(x) = \theta e^{-\theta x}$. The scale is given by θ^{-1}.

Setting $b(\theta) = -\theta^{-1}(1-x_o-e^{-x_o})$, for some $x_o > 1.593$, and $N(\theta) = \phi$ in Lemma 5.2, and observing $\dot{\psi}(x,\theta) = \theta(1-\theta x)$, $x > 0$, the resulting influence function is

$$\tilde{\psi}(x,\theta) = \theta^{-1}(1-\min(\theta x, x_o) - e^{-x_o}) \ . \tag{6.20}$$

This M-estimator behaves analogously to a Huber Proposal 2 under the normal distribution. Large observations are "brought in", although the Winsorizing is done adaptively, depending also on θ. But the proportion of Winsorized values at the model remains at e^{-x_o}. Sensitivity and asymptotic variance at f_θ are respectively

$$\gamma^*(\theta) = \theta(1-x_o-e^{-x_o})/(1-(1-x_o)e^{-x_o})$$

$$\sigma^2(\psi, f_\theta, \theta) = \theta^2(1-2x_oe^{-x_o}-e^{-2x_o})/(1+e^{-x_o}(1-x_o))^2 \ .$$

Smaller values of x_o than 1.593 give an influence function that dampens down the effect on small observations also. But this would effectively Winsorize at least 20% of the population.

TABLE 6.4

Efficiency, sensitivity, and local breakdown bounds for Winsorizing M-estimates of the exponential family

% Winsorizing	Efficiency	Sensitivity $\gamma^*(1)$	Breakdown Bound $\delta^*(.5)$
15	.7829	1.852	.2104 (*)
10	.8472	2.094	.2240
5	.9175	2.556	.1808
2.5	.9563	3.074	.1495

Efficiencies are uniform in $\theta \in E^+$ but sensitivity is proportional to θ. Breakdown values are the minimum proportion of contamination H so that $\theta_o/T[\tilde{\psi},G]$ is outside the region $[.5, 1.5]$. We can observe here that behaviour of the estimator for 10-15% Winsorizing in quantitative departures (Breakdown) cannot be inferred from the infinitesimal departure indicated by γ^*. The (*) indicates that breakdown is achieved by point contamination at the origin.

Thall's (1979) solution, that minimizes the maximum variance in the Kolmogorov neighbourhood $n_k(\varepsilon,F_o)$ where ε is specified and $F_o(x) = 1-e^{-x}$, is

$$\psi_o(x) = \begin{cases} k \tan((k/2)\log a) & \text{if } 0 < x < a \\ k \tan((k/2)\log x) & a \leq x \leq b \\ x - 1 & b \leq x \leq c \\ c - 1 & c \leq x < \infty . \end{cases}$$

This is continuous and has piecewise continuous bounded derivatives. It is a solution only for values $\varepsilon \leq \varepsilon_o = .0095$, which is clearly restrictive. Strictly speaking to make the estimator unbiased at the model $F_o(\theta x)$ the value of $\int \psi_o(x)dF_o(x)$ should be subtracted from ψ_o. Thall corrects after the estimation dividing through by $b(\varepsilon)$ which is

a solution of

$$\int_0^\infty \psi_0(x/b(\varepsilon))e^{-x}dx = 0 \ .$$

Efficiency at the model is poor, while the introduction of an asymptotic bias at the model is alarming considering small values of contamination that are allowed.

TABLE 6.5

ε	k	a	b	c	b(ε)	eff
.005	1.6005	.6085	3.4751	4.1506	.9896	.5323
.008	1.5935	.6115	3.6075	3.799	.9860	.5430
.0095	1.5902	.6123	3.6673	3.6732	.9835	.5482

If the bulk of the distribution is approximately exponential but there is the possibility of tail contamination a redescending influence function is appropriate. We are particularly concerned for those large samples where asymptotic bias and not variance are important. Let

$$\psi(x,\theta) = \begin{cases} 1-\theta x-P & 0 \le \theta x < a \\ 1-a-P & a \le \theta x < b \\ P^*(\theta x-c) & b \le \theta x < c \\ 0 & c < \theta x < \infty \ , \end{cases}$$

where $P^* = \frac{P+a-1}{c-b}$, and

$$P = \frac{ae^{-a} + (1-a)(e^{-a}-e^{-b}) + \frac{a-1}{c-b}\left((b+1-c)e^{-b} - e^{-c}\right)}{1 - e^{-b} - \frac{1}{c-b}\left((b+1-c)e^{-b} - e^{-c}\right)} \ .$$

Observations falling in the region $[c/\theta,\infty)$ are neglected in the estimating equation at θ, since this is a set of null influence. The

choice of null set can depend on n. For instance for a sample of size n if we set $P\{X_{(n)} > c/\theta\} = \varepsilon$, then $c = -\log(1-(1-\varepsilon)^{1/n})$. Some values are given below.

TABLE 6.6

ε \ n	10	15	20	50	100
.025	5.98	6.39	6.67	7.59	8.28
.05	5.28	5.68	5.97	6.88	7.58
.1	4.56	4.96	5.25	6.16	6.86

Asymptotically the proportion of values either Winsorized or trimmed tends to e^{-a}. Values $a = 2.303, 2.996$ correspond to proportions .1, .05 respectively. Several values of a,b,c were chosen and asymptotics given.

TABLE 6.7

a	b	c	eff.	$\gamma^*(1)$	$\delta^*(.5)$
2.303	3.5	4.56	.66	2.53	.194
2.303	3.5	5.98	.71	2.39	.203 (*)
2.303	3.5	6.67	.75	2.30	.208 (*)
2.303	4.5	6.67	.78	2.21	.216 (*)
2.996	3.5	5.25	.74	3.06	.169
2.996	4.5	6.67	.84	2.73	.177
2.996	4.5	8.28	.87	2.67	.178

As $b \to c$ efficiency is increased under the model and sensitivity

decreased, but this should be balanced against behaviour of variances and sensitivities away from the model (viz. the value of P^* tends to $+\infty$ as $b \to c$).

Asymptotics of the M-estimator are justified since the influence function is continuous with piecewise continuous derivatives. Suppose the true scale is θ_o^{-1}. As it represents the expectation of the distribution it is not unreasonable to search for the consistent solution of the M-estimating equations in a compact set $0 < M_1 < \theta < M_2 < \infty$. (In a Monte Carlo study it would be natural to let M_1, M_2 depend on θ_o.) Since $\psi(x,\theta)$ is continuous on $S \times [M_1,M_2]$, the family

$$\mathcal{A}(M_1,M_2) = \{\psi(\cdot,\theta) \,|\, M_1 \leq \theta \leq M_2\}$$

is equicontinuous. Clearly $\mathcal{A}(M_1,M_2)$ is uniformly bounded since

$$-1-P \leq \psi(x,\theta) \leq -1+a-P \qquad (x,\theta) \in S \times \Theta \ .$$

By Proposition 1.4

$$K_n(\theta) \xrightarrow{a.s.} K_G(\theta) \quad \text{uniformly in} \quad \theta \in [M_1,M_2] \tag{6.21}$$

for any $G \in \mathcal{G}$ given. Since ψ is only piecewise continuously differentiable with

$$(\partial/\partial\theta)^{-}\psi(x,\theta) = \begin{cases} -x & 0 \leq x \leq a/\theta \\ 0 & a/\theta \leq x \leq b/\theta \\ P^*x & b/\theta \leq x \leq c/\theta \\ 0 & c/\theta \leq x < \infty \end{cases}$$

it is amenable to consider functions

$$h(x,\theta) = \begin{cases} 0 & x < 0 \\ -x & 0 \le x \le a/\theta \\ -\frac{a}{\theta} + P^* \frac{b+a}{b-a}\left(x - \frac{a}{\theta}\right) & a/\theta \le x \le b/\theta \\ P^* x & b/\theta \le x \le c/\theta \\ P^* c/\theta & c/\theta \le x < \infty \end{cases} .$$

Then $\{h(\cdot,\theta) \mid M_1 \le \theta \le M_2\}$ is an equicontinuous and bounded family of functions. By Theorem 1.1

$$\int_{[0,x]} h(y,\theta)dF_n(y) \xrightarrow{a.s.} \int_{[0,x]} h(y,\theta)dG(y) \quad \text{uniformly in} \quad x \in S .$$

Observe then that

$$\begin{aligned} (\partial/\partial\theta)^{-}K_n(\theta) &= \int (\partial/\partial\theta)^{-}\psi(y,\theta)dF_n(y) \\ &= \int_{[0,a/\theta]} h(y,\theta)dF_n(y) + \int_{[0,c/\theta]} h(y,\theta)dF_n(y) \\ &\qquad - \int_{[0,b/\theta]} h(y,\theta)dF_n(y) \\ &\xrightarrow{a.s.} \int_{[0,a/\theta]} h(y,\theta)dG(y) + \int_{[0,c/\theta]} h(y,\theta)dG(y) \\ &\qquad - \int_{[0,b/\theta]} h(y,\theta)dG(y) \\ &= \int (\partial/\partial\theta)^{-}\psi(y,\theta)dG(y) \quad \text{uniformly in} \quad \theta \in [M_1,M_2] . \end{aligned}$$

So if $G(x) = F_{\theta_o}(x) = 1 - e^{-x\theta_o}$

$$(\partial/\partial\theta)^{-}K_n(\theta) \xrightarrow{a.s.} (\partial/\partial\theta)K_{F_{\theta_o}}(\theta) \quad \text{uniformly in} \quad \theta \in [M_1,M_2]. \quad (6.22)$$

See that $K_{F_{\theta_o}}(\theta)$ is continuously differentiable in θ, and

$$(\partial/\partial\theta)K_{F_{\theta_o}}(\theta) = \frac{1}{\theta_o}\, k(\theta_o/\theta) \ ,$$

where

$$k(\lambda) = 1 + e^{-\lambda a}(\lambda a - 1) + P^*(e^{-b\lambda}(\lambda b + 1) - e^{-\lambda c}(\lambda c + 1)) \ .$$

Also

$$(\partial/\partial\theta)K_{F_{\theta_o}}(\theta_o) = \frac{1}{\theta_o}\, k(1) = \frac{1}{\theta_o}\, \{1 + e^{-a}(a-1) + P^*(e^{-b}(b+1) - e^{-c}(c+1))\},$$

and $\lim_{\lambda \to +\infty} k(\lambda) = 1$, $\lim_{\lambda \to 0} k(\lambda) = 0$.

Values of $k(1)$ are tabulated for the parameters of Table 6.7, together with the region in which $k(\lambda) > 0$.

a	b	c	k(1)	$\{\lambda \mid k(\lambda) > 0\} = (\lambda^+, \infty)$	$\{\lambda \mid k(\lambda) - k(1) = 0,\ \lambda \neq 1\}$
2.303	3.5	4.56	1.23	(.308,∞)	.741
2.303	3.5	5.98	1.28	(.272,∞)	.544
2.303	3.5	6.67	1.29	(.259,∞)	.489
2.303	4.5	6.67	1.20	(.269,∞)	.546
2.996	3.5	5.25	1.31	(.316,∞)	.564
2.996	4.5	6.67	1.20	(.292,∞)	.477
2.996	4.5	8.28	1.22	(.255,∞)	.378

Since $(\partial/\partial\theta)K_{F_{\theta_o}}(\theta) = \frac{1}{\theta_o}\, k(\theta_o/\theta) > 0$ uniformly in $\theta < \theta_o/\lambda^+$, then there exists a unique consistent solution $\theta(\psi, F_n)$ in the region $M_1 \leq \theta \leq \min(\theta_o/\lambda^+, M_2)$ f.a.s.l.n.. This follows from the uniform convergence (6.21) and (6.22). Suppose there did exist a sequence of solutions $\{\hat{\theta}_n\}$ existing in the region $\theta_o/\lambda^+ \leq \theta \leq M_2$. Since

$$(\partial/\partial\theta)K_n(\theta)\big|_{\theta=\hat{\theta}_n} \xrightarrow{a.s.} (\partial/\partial\theta)K_{F_{\theta_o}}(\theta)\big|_{\theta=\hat{\theta}_n} = \frac{1}{\hat{\theta}_n}\, k(\theta_o/\hat{\theta}_n) \ ,$$

then

$$\left|(\partial/\partial\theta)K_n(\hat{\theta}_n) - \frac{1}{\hat{\theta}_n}k(1)\right| \xrightarrow{a.s.} \frac{1}{\hat{\theta}_n}\left|k(\theta_o/\hat{\theta}_n) - k(1)\right|$$

$$\geq \frac{1}{M_2}\inf_{\{\lambda\,|\,k(\lambda)\geq 0\}}\left|k(\lambda) - k(1)\right|$$

$$> 0 .$$

By Lemma 2.7 there exists a sequence $\theta(\psi,F_n) \xrightarrow{a.s.} \theta_o$. So

$$\left|(\partial/\partial\theta)K_n(\theta(\psi,F_n)) - \frac{1}{\theta(\psi,F_n)}k(1)\right| \xrightarrow{a.s.} 0 .$$

Then if we let $H(\psi,F_n) = \{\theta\,|\,K_n(\theta) = 0,\ M_1 \leq \theta \leq M_2\}$, the M-estimator is defined to be

$$\inf_{\theta\in H(\psi,F_n)}\left|(\partial/\partial\theta)K_n(\theta) - \frac{1}{\theta}k(1)\right| = \left|(\partial/\partial\theta)K_n(T[\psi,F_n]) - \frac{1}{T[\psi,F_n]}k(1)\right|.$$

This excludes all those solutions in $[\theta_o/\lambda^+,M_2]$ f.a.s.l.n.. The estimator is consistent and asymptotically normal. The latter observation follows from Theorem 3.1.

§6.5 Robust Estimation of Location and Scale

It is clearly of interest to estimate location and scale jointly Many authors, e.g. Collins (1976), Andrews et al (1972), advocate the simple inclusion of a nonparametric scale estimate in the equation for location. This is not conducive to the easy derivation of the asymptotic variance of the location estimator. But estimators satisfying equations

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\mu}{\sigma}\right) = 0 \qquad (6.23)$$

for suitable 2×1 vector functions $\psi = (\psi_1,\psi_2)'$ do have an

asymptotically normal distribution with variance covariance matrix $\sigma^2(\psi,G,T[\psi,G])$. They are obviously location and scale invariant. It is apparent from §5.3 and Hampel (1978) that optimality of a multivariate M-estimator is not fully resolved. Thus a numerical investigation is useful. We consider the following three estimators:

(i) M.L.E. with $\psi(x) = (x, -1+x^2)'$

(ii) Huber's Proposal 2 with $\psi_c(x) = (\max\{-c,\min(x,c)\},$
$\min(x^2,c^2) - \beta_c^2)'$, and

(iii) Redescending Estimator with

$$\psi_{a,b,c}(x) = \begin{cases} (x, -1+x^2-P)' & 0 \le |x| \le a \\ (a\,\mathrm{sign}(x), -1+a^2-P)' & a \le |x| \le b \\ \frac{c-|x|}{c-b}\,(\ \text{"}\ ,\ \text{"}\ )' & b \le |x| \le c \\ 0 & c \le |x| < \infty \end{cases}$$

The redescending estimator has a set of null influence $N(\psi_{a,b,c};\mu,\sigma) = (-\infty, \mu-c\sigma] \cup [\mu+c\sigma,\infty)$. Given a sample of size n generated from $\Phi(\frac{x-\mu}{\sigma})$ the probability that any observation falls in that set is $1 - (2\Phi(c)-1)^n$. At least one observation falls in N if either $X_{(1)} < \mu - c\sigma$, or $X_{(n)} > \mu + c\sigma$, where $X_{(i)}$ is the i'th order statistic. The value of c should be chosen so as to leave this value minimal, especially in smaller samples, so as not to lose information from medium size samples (e.g. $n < 50$) unnecessarily.

Validity of asymptotics follows from uniform convergence theory. To establish asymptotic normality of M-estimators of (ii) and (iii) we use the results of §3.4. First observe by Lemma 1.4 that $\{\psi(\frac{\cdot-\mu}{\sigma}) \mid (\mu,\sigma) \in E \times (E^+ - [0,\eta]),\ \eta > 0\}$ is an equicontinuous family of functions uniformly bounded by $\sqrt{a^2 + (-1+a^2-P)^2}$. The Fisher consistency assumption at the model is satisfied; that is $\int \psi(x)d\Phi(x) = 0$. Suppose

Suppose $K_G(\mu,\sigma)$ has a zero (μ_o^*,σ_o^*) and is continuously differentiable at that point with non-singular derivative matrix. This is the case in the examples of Table 6.10. By Theorem 2.1 there exist consistent sequences of solutions of equations (6.23) to the zeros of $K_G(\mu,\sigma)$. Clearly from an analysis similar to that of §6.4 for any compact subset D of the parameter space bounding σ away from zero we observe

$$\nabla K_n(\mu,\sigma) \xrightarrow{a.s.} \nabla K_G(\mu,\sigma) \quad \text{uniformly in} \quad \theta \in D\ , \tag{6.24}$$

whenever F_n is generated from G. Points $B(\mu,\sigma)$ in the observation space at which the derivative of $\psi_{a,b,c}$ does not exist are $\pm a\sigma + \mu$, $\pm b\sigma + \mu$, and $\pm c\sigma + \mu$. Since each G has a bounded density, $B_n = \cup\{B(\mu,\sigma) \mid \sqrt{(\mu-\mu_o^*)^2 + (\sigma-\sigma_o^*)^2} \leq \delta_n\}$ is so that $P_G(\bar{B}_n) = O(\delta_n)$. Assumptions of Lemma 3.3 are satisfied, whence Theorem 3.5 asserts the asymptotic normality of the sequence of consistent roots.

On compacts D, $\{\psi(\frac{\cdot-\mu}{\sigma}) \mid (\mu,\sigma) \in D\}$ is a family of functions all of which redescend to zero within a compact subset of E. So by the comment comments proceeding Theorem 1.2 it can be expected that the asymptotic normality is attained uniformly in neighbourhoods of the normal distribution, justifying inference applications while slight contamination may be present.

To examine the correspondence between small sample behaviour and the asymptotic distribution a Monte Carlo experiment was performed. For the effective implementation of the redescending estimator it is necessary to identify it from multiple roots of (6.23). Rey (1977) identified this problem to be occurring in the regression estimation of Andrews (1974). It was overcome in the following manner. Starting with a grid of initial starting values $(\hat{\mu},\hat{\sigma})$ about the underlying value of (0,1), a selection statistic was used to identify the M-estimator.

Analogous to the single parameter estimation (§2.4), the selection statistic was based on comparison of the derivatives of $K_n(\mu,\sigma)$ and $K_{G_o}(\mu,\sigma)$. Since the distribution G_o is hypothesized to be some $\Phi\left(\frac{x-\mu_o}{\sigma_o}\right)$ and it is known there exists a sequence $\{(\hat{\mu}_n,\hat{\sigma}_n)\}$ of roots consistent to (μ_o,σ_o) then this is approximated by $\Phi\left(\frac{x-\hat{\mu}_n}{\hat{\sigma}_n}\right)$. That is the selection statistic employed was

$$f_n(\mu,\sigma) = \left\| \int \nabla\psi\left(\frac{x-\mu}{\sigma}\right) dF_n(x) - \int \nabla\psi\left(\frac{x-\mu}{\sigma}\right)(1/\hat{\sigma}_n)\phi\left(\frac{x-\hat{\mu}_n}{\hat{\sigma}_n}\right) dx \right\| , \qquad (6.25)$$

where ϕ is the standard normal density. This statistic is evaluated from the sample without prior knowledge of the underlying distribution. It is appropriate since the partial derivatives of ψ are bounded, and little weight is attributed to observations falling into the set of null influence. Investigations showed the resulting M-estimator most often to be the root closest to $(\bar{X},S)$ as one would expect under the model.

In notation we let v_{11}, v_{22} be the asymptotic variances of location and scale M-estimates respectively. Assuming a normal parametric family we can estimate v_{11} by

$$\hat{v}_{11} = \hat{\sigma}_n^2 \frac{\int \psi_1^2(x)\phi(x)dx}{\left(\int \psi_1'(x)\phi(x)dx\right)^2} . \qquad (6.26)$$

That is the scale is estimated by $\hat{\sigma}_n$. Then in practice the $100(1-\alpha)\%$ confidence interval for location would be

$$\hat{\mu}_n \pm Z_{\alpha/2}\sqrt{\hat{v}_{11}/n} ,$$

where $Z_\alpha = \Phi^{-1}(1-\alpha)$.

With 500 replications of each experiment, samples of size $n = 20$ and 100 were generated from a standard normal distribution. Corresponding

asymptotic variance compared favourably with resulting mean squared errors, labelled $m\nu_{11}$ and $m\nu_{22}$ respectively for location and scale estimates, and were even a good approximation with $n = 20$. Also, 90, 95% confidence intervals for the location estimator derived from the asymptotic distribution, compared favourably with corresponding empirical confidence intervals generated from the experiment.

TABLE 6.8

Comparison of asymptotic variance and mean squared error of location and scale M-estimators using 500 replications of samples generated from the standard normal distribution

Estimator	Sample size n	ν_{11}/n	ν_{22}/n	$m\nu_{11}$	$m\nu_{22}$	$(2\Phi(c)-1)^n$
	20	.5-1	.25-1	.53-1	.24-1	
M.L.E.	100	.1-1	.5-2	.96-2	.52-2	
Proposal 2	20	.51-1	.32-1	.51-1	.34-1	
c = 1.645	100	.1-1	.64-2	.1-1	.76-2	
R.E.	20	.54-1	.38-1	.63-1	.63-1	.9808
(a,b,c)=(1.645,2.,3.3)	100	.11-1	.75-2	.11-1	.92-2	.9078

TABLE 6.9

Comparison of the asymptotic confidence interval with the empirical confidence interval (E.C.I.) for location

Estimator	Sample size n	$\pm Z_{.025}\sqrt{\nu_{11}/n}$	95% E.C.I.		$\pm Z_{.025}\sqrt{\nu_{11}/n}$	90% E.C.I.	
M.L.E.	20	.438	-.413	.478	.368	-.349	.398
	100	.196	-.196	.195	.165	-.149	.158
Proposal 2	20	.444	-.476	.384	.373	-.402	.321
c = 1.645	100	.199	-.199	.195	.167	-.173	.164
R.E.2.	20	.456	-.517	.426	.382	-.438	.364
	100	.204	-.191	.209	.171	-.172	.185

We can conclude from Tables 6.8 and 6.9 that the confidence intervals derived from the asymptotic distribution are a satisfactory approximation to the confidence intervals for μ that exist in smaller samples. It is then reasonable to compare the behaviour of the asymptotic distribution in small departures from normality. This is particularly the case when intervals are estimated using (6.26). To give them approximate validity in small perturbations of the underlying distribution the asymptotic variance should not be perturbed greatly. Nor should the asymptotic distribution of $\hat{\sigma}_n$ be far from that generated under the normal model. Asymptotic variances, v_{11} and v_{22}, were found for various underlying distributions of the form

$$G(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(\tfrac{x+\Delta}{\alpha}) \ . \qquad (6.27)$$

The functional T was determined uniquely by the selection functional $\sqrt{\mu^2 + (\sigma-1)^2}$ in this case. Table (6.10) exhibits the perturbations from $T[\psi,\Phi] = (0,1)$ and the asymptotic variances for small values of ε. As emphasized in §4.1 we need not be interested in the interpretation of the functional values but rather their stability behaviour in small neighbourhoods of a model distribution.

Redescending estimators reduce asymptotic bias for heavy tailed contamination, e.g. when $G(x) = .95\Phi(x) + .05\Phi(\tfrac{x+1.5}{3.})$ the asymptotic bias is reduced to (.08, .03) which compares with (.08, .23) for the M.L.E. and (.1, .06) for Huber's Proposal 2. They can be further recommended for the estimation of confidence intervals since the asymptotic distribution, particularly v_{22}, is not greatly perturbed from that under the model. But sharply redescending influence functions can be susceptible to contamination near $\pm c$. Overall the estimators with bounded influence curves stabilize bias and variance near the model at little cost in asymptotic efficiency at the model. This supports the

TABLE 6.10

M-functional values and asymptotic variances of M-estimators of location and scale for distributions (6.27).

$H(x) = \Phi(\frac{x+\Delta}{\alpha\sigma})$			M.L.E.				Proposal 2 (c = 1.645)			
ε	Δ	α	$T[\psi,G]$		v_{11}	v_{22}	$T[\psi,G]$		v_{11}	v_{22}
	0.	1.	0.	1.	1.	.5	0.	1.	1.03	.64
.05	1.5	3.	-.08	1.23	1.51	3.06	-.10	1.06	1.19	.82
.05	3.5	.1	-.18	1.24	1.53	1.11	-.10	1.1	1.33	1.03
.1	0.	.1	0.	.95	.90	.52	0.	.93	.90	.52
.1	2.5	1.	-.25	1.25	1.56	1.04	-.20	1.18	1.54	1.19
			Redescending 1.				Redescending 2.			
	0.	1.	0.	1.	1.09	.78	0.	1.	1.08	.75
.05	1.5	3.	-.08	1.03	1.20	.91	-.08	1.03	1.19	.88
.05	3.5	.1	0.	1.	1.16	.84	0.	1.	1.14	.80
.1	0.	.1	0.	.91	.98	.96	0.	.91	.98	.92
.1	2.5	1.	-.14	1.13	1.78	.73	-.145	1.13	1.85	1.72

use of confidence intervals derived from the asymptotic distributions. Further tables and conclusions may be found in Appendix 3.

The fact that we do not fully understand what we are estimating in perturbations from the model, as exhibited in the sole estimation of scale or even location in asymmetric departures, should not deter us from using the parametric model. It is important for estimators to have robust asymptotic behaviour in small perturbations from the model.

Amongst the robust M-estimators it is clearly possible to choose contamination to show one estimator in a better light than another. But the philosophy (or reassurance) behind the choice of redescending estimator is the elimination of observations not belonging to the bulk of the sample according to the model to be fitted. Simultaneously an estimate is provided.

A final point concerns the small sample bias of the scale estimate. Jackknifing is precluded due to the numerical searching for roots of the equations. Carroll (1979, P.677) also warns that one step M-estimates from a consistent estimate of location and with nonparametric estimate for scale do not have good jackknifed estimates of variance. However from the Taylor expansion method of (5.6) it can be established for the joint estimation of location and scale of a symmetric distribution $F_o(\frac{x-\mu}{\sigma})$,

$$E[\hat{\sigma}_n - \sigma] = \frac{1}{n}\frac{\sigma}{E[X\,\psi_2']}\left\{ -\frac{E[\psi_2'\,\psi_1]}{E[\psi_1']} - \frac{E[X\,\psi_2'\psi_2]}{E[X\,\psi_2']} + \frac{E[\psi_1^2]E[\psi_2'']}{2E[\psi_1']^2} + \frac{E[\psi_2^2]E[X^2\,\psi_2'']}{2E[X\,\psi_2']^2}\right\} + o(n^{-1}).$$

Here $E[\cdot]$ is the expectation operator with respect to $F_o(x)$. When $F_o = \Phi$ and $\sigma = 1$ values of bias are given below.

a	b	c	Bias = $E[\hat{\sigma}-\sigma]$
1.645	∞	∞	-.7390
1.96	∞	∞	-.9444
1.645	2.4	3.	.7043
1.645	2.	3.3	1.1948
1.96	2.5	3.	.9291
2.	2.91	3.	1.6381

The effect of the redescending influence function is to change the sign of the bias. Again the sharply redescending influence function should be avoided; this time because it incurs a larger small sample bias.

SECTION C: APPLICATION TO ESTIMATION IN MIXTURES OF TWO NORMAL DISTRIBUTIONS

CHAPTER 7

MINIMAL DISTANCE ESTIMATES FOR MIXTURES

§7.1 Robustness and Relationships with Minimal Distance Methods

There are a variety of minimal distance methods for estimating the parameters in a finite mixture of normal distributions given by

$$F_\theta(x) = \varepsilon_1 \Phi\left(\frac{x-\mu_1}{\sigma_1}\right) + \varepsilon_2 \Phi\left(\frac{x-\mu_2}{\sigma_2}\right) + \ldots + \varepsilon_N \Phi\left(\frac{x-\mu_N}{\sigma_N}\right) ,$$

$$\sum_{i=1}^{N} \varepsilon_i = 1 , \quad \theta = (\varepsilon_1,\ldots,\varepsilon_N, \mu_1,\ldots,\mu_N, \sigma_1,\ldots,\sigma_N) .$$

The apparent failure of the M.L.E. in the mixture model to attain a global maximum of the log-likelihood as described in Odell and Basu (1976, P.1099) motivated the use of minimal distance methods some of which we briefly review in this section. These had been sufficiently developed since the initial articles of Wolfowitz in the 1950's so that specific distances were proposed by Bartlett and Macdonald (1968), Choi and Bulgren (1969) and Macdonald (1971). The latter authors investigated Cramer-von Mises type distances which have been recently used in the testing area when parameters of the null distribution must be estimated by Durbin, Knott, and Taylor (1975). Pollard (1980) and Silvapulle (1980) discuss them in a more general setting.

Related distances compare empirical and almost sure limits of characteristic functions and density estimators. Paulson, Holcomb, and Leitch (1975) and Heathcote (1977) investigated the Integrated Squared Error (I.S.E.) distance

$$I_n(\theta) = \int |\tilde{\phi}_n(t) - \phi(t,\theta)|^2 dW(t)$$

where $\tilde{\phi}_n(t) = n^{-1} \sum_{i=1}^{n} \exp\{itX_j\}$ is the empirical characteristic function with expectation $\phi(t;\theta)$. The I.S.E. estimator has direct application to estimating parameters in the stable laws where there is a ready representation in terms of the characteristic function (It is instructive to read the remarks of Hall (1980) concerning the correct forms of this representation.). The periodic nature of the characteristic function precludes any single observation from attributing undue weight to the distance $I_n(\theta)$ which lends it a robustness quality. In fact it can be a form of M-estimator in the sense defined in this thesis. For assuming interchange of integration and differentiation the minimizing equations correspond to the choice of

$$\psi(x,\theta) = \int [\{\cos(t.x)-u(t,\theta)\}\nabla u(t,\theta)+\{\sin(t.x)-v(t,\theta)\}\nabla v(t,\theta)]dW(t), \quad (7.1)$$

where $u(t,\theta)$ and $v(t,\theta)$ are the real and imaginary parts of the characteristic function of F_θ. This can be observed from the equations of Heathcote (1977, P.257). With a suitably regular weight function $W(t)$ this representation of the influence function can be simplified.

LEMMA 7.1: Assume W is an absolutely continuous function with respect to Lebesgue measure on E^k, having density $\tilde{g}(t) : E^k \to E$ which is the characteristic function of $g(t) \in L_2$, $\tilde{g}(t) = \int e^{it.x} g(t)dt$. Suppose further that g is real and symmetric about the origin; that is $g(\delta_1 t_1,\ldots,\delta_k t_k) = g(t_1,\ldots,t_k)$ where each δ_i can take values either ± 1.

Assume densities f_θ have partial derivatives that exist and are bounded above by an L_1 function. Then the I.S.E. estimator is determined by influence function

$$\psi(x,\theta) = (2\pi)^k \{ \int \nabla f_\theta(x-y)g(y)dy - \iint \nabla f_\theta(x-y)g(y)dy f_\theta(x)dx\}, \qquad (7.2)$$

and selection statistic $I_n(\theta)$. Each integration is over E^k.

PROOF: By the Fourier inversion on E^k, and noting $\int f_\theta(x-y)g(y)dy$ has Fourier transform $\phi(t,\theta)\tilde{g}(t)$

$$\int f_\theta(x-y)g(y)dy = \frac{1}{(2\pi)^k} \int e^{-it.x}\, \phi(t,\theta)\tilde{g}(t)dt \ .$$

Here $t.x = \sum_{i=1}^{k} t_i x_i$ is the inner product. So

$$\int \nabla f_\theta(x-y)g(y)dy = \frac{1}{(2\pi)^k} \int e^{-it.x}\, \nabla\phi(t,\theta)\tilde{g}(t)dt \ .$$

By the Parseval relation

$$\int f_\theta(x) \int \nabla f_\theta(x-y)g(y)dydx = \frac{1}{(2\pi)^k} \int \phi(t,\theta)\overline{\nabla\phi(t,\theta)\tilde{g}(t)}dt$$

$$= \frac{1}{(2\pi)^k} \int \phi(t,\theta)\overline{\nabla\phi(t,\theta)}\tilde{g}(t)dt$$

(as symmetry of g implies symmetry of $\tilde{g}$).

Hence the influence function is that of (7.1).

$$\psi(x,\theta) = \int e^{-it.X_i} \nabla\phi(t,\theta)\tilde{g}(t)dt - \int \phi(t,\theta)\overline{\nabla\phi(t,\theta)}\tilde{g}(t)dt$$

$$= \int [\{\cos(t.X_i) - u(t,\theta)\}\nabla u(t,\theta) + \{\sin(t.X_i)$$

$$- v(t,\theta)\}\nabla v(t,\theta)]\tilde{g}(t)dt \ .$$

Using this simpler form of the influence function it is possible to evaluate explicit expressions for the influence curve more easily. In both Paulson, Holcomb and Leitch (1975) and Thornton and Paulson (1977) the weight function $dW(t) = e^{-t^2}dt$ is used to estimate the location parameter of a univariate normal population. A short calculation reveals this to be an M-estimator with a bounded redescending influence curve

$$IC_{\mu}(x) = \frac{73}{24}(x-\mu)\exp\{-\frac{25}{48}(x-\mu)^2\} .$$

A robust estimator of a different type is that derived from the Hellinger distance advocated by Beran (1977)

$$\int (f_n^{\frac{1}{2}} - f_\theta^{\frac{1}{2}})^2 dx ,$$

where $f_n(x)$ is chosen as an appropriate density estimator. In a heuristic way the minimizing equations

$$\int f_n^{\frac{1}{2}}(x) f_\theta^{\frac{1}{2}}(x) \nabla f_\theta(x) dx = 0$$

have similar asymptotic bias properties to an M-estimator with influence function $\psi(x,\theta) = f_\theta^{-\frac{1}{2}}(x)\nabla f_\theta(x)$. For assuming g to be the density of the underlying distribution the resulting asymptotic equation resolves to

$$\int g^{\frac{1}{2}}(x) f_\theta^{\frac{1}{2}}(x) \nabla f_\theta(x) dx = 0 .$$

For a normal location family see that

$$f_\theta^{-\frac{1}{2}}(x)\nabla f_\theta(x) = -(2\pi)^{-\frac{1}{4}}(x-\theta)\exp\{-(x-\theta)^2/4\} ,$$

which is again a redescending function. Asymptotically the estimator is robust and efficient but the density must be estimated first. Small sample biases are most likely to be present in moderate sample sizes.

The redescending nature of these functions also allows the possibility of more than one root to the estimating equations. Assuming large numbers of parameters it may be a cumbersome problem to search out all roots of the equations.

The robustness and applicability of many L_2-distance estimators can therefore often be observed by studying the estimating equations. Simple illustrations in the single parameter case often provide insight into the behaviour of the distance estimator when larger numbers of parameters are present. This is certainly the case when estimating mixtures of normal distributions where with only two component distributions there are five parameters that must be estimated. The literature on the latter problem is extensive and we briefly set out some of the considerations involved before returning to the use of minimal distance estimates.

§7.2 Estimating Mixtures of Normal Distributions

The problem of "decomposing" a mixture of normal distributions by estimating the unknown parameters of the component distributions and mixing proportions is important in scientific and economic investigations. This is emphasized in the survey work of Macdonald (1975) and Odell and Basu (1976).

Several situations are possible. They depend on the parameters to be estimated, the number of component distributions, how much the component distributions overlap, and how the data is collected. The theory in estimation of mixtures assumes the mixture to be identifiable. A mixture of a family F is called "identifiable" if

$$G = \int F_\theta(x)dH^*(\theta) = \int F_\theta(x)dH(\theta)$$

implies $H^* = H$. Letting H be the class of mixing distributions and G^* the resulting class of mixtures we say G^* is identifiable if every $G \in G^*$ is identifiable. Chandra (1969) and Yakowitz (1969) proved the normal (multivariate and univariate) mixtures to be identifiable, while Teicher (1960, 1961, 1963) examined identifiability more generally.

Assuming a mixture with two completely specified component distributions, many estimators can be found for the proportion parameter. A class of estimators can be formed through

$$\hat{\varepsilon}_n(B) = \frac{F_n\{B\} - F_{\theta_2}\{B\}}{F_{\theta_1}\{B\} - F_{\theta_2}\{B\}} \quad , \quad \theta_1 \neq \theta_2$$

for any Borel sets B for which the two mixed distributions satisfy $F_{\theta_1}\{B\} \neq F_{\theta_2}\{B\}$. Each is unbiased, converges to the true mixing parameter with probability one, and has variance $O(1/n)$. Unfortunately the estimator can take values outside of the region $[0,1]$ although this is common amongst minimal distance estimators. Boes (1966) investigates estimators derived from those such as $\hat{\varepsilon}_n$, giving necessary and sufficient conditions for uniform attainment of the Cramér-Rao lower bound. This cannot be attained for the mixture of two normal distributions but can almost be attained if components are well separated. Should the data be grouped in the form of knowledge of the empirical distribution function evaluated at r, the M.L.E. is in this form with

$$\hat{\varepsilon}_n(r) = \frac{F_n(r) - \Phi\left(\frac{r-\mu_2}{\sigma_2}\right)}{\Phi\left(\frac{r-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{r-\mu_2}{\sigma_2}\right)} \tag{7.3}$$

James (1978) shows the efficiency in theory and application of this estimator. The use of such a simple statistic is supported by the

qualitative conclusion of Hill (1963) who claimed that large and often impractical sample sizes were required to obtain moderate precision in estimating the proportion should the mixing distributions be poorly separated. But with moderate separation this need not be the case and more frequently the component distributions must also be estimated.

The M.L.E. is known to work well under the model for those cases other than when both dispersion parameters are unknown and have to be estimated. In the latter case there exist singularities in the likelihood. Day (1969) showed the good behaviour of the M.L.E. for multivariate normal distributions assuming equal covariance matrices of the component distributions. Kiefer (1978) showed existence of a consistent root of the maximum likelihood equations when unequal variances are assumed but Fowlkes (1979, P.74) shows that even in a sample as large as $n = 200$ there can exist a number of local maxima of the log likelihood. Even assuming the model is true the correctness of the M.L.E. procedure has not yet been resolved. Apart from more recent work using minimal distance estimators the only procedure that has been discussed frequently as an alternative to the M.L.E. has been the method of moments introduced to the mixtures problem by Pearson (1894). Accounts of it being used for mixtures estimation are found in Cohen (1967) and Day (1969), while Bowmen and Shenton (1973) investigate regions for the method of moments solution to exist. Both the M.L.E. and the method of moments may be discounted on robustness grounds. This is clearly the case for the method of moments and can be observed to be so for the M.L.E. by observing the unboundedness of the influence function derived by partial differentiation of the log of the mixed normal density

$$\ell n\left\{\sum_{j=1}^{N} \varepsilon_j \phi\left(\frac{x-\mu_j}{\sigma_j}\right)\right\};\ \phi(x) = (\sqrt{2\pi})^{-1}\exp(-x^2/2),\ j = 1,\ldots,N\ .$$

Derivatives with respect to proportion parameters are bounded but not those with respect to location or scale parameters.

It has been a tendency in the literature to illustrate the applicability of the M.L.E. and other distance estimators with computer generated data assuming the model distribution. However, some estimators are very sensitive to departures from the distributional assumptions. Iterative computer algorithms and graphical methods are liable to work for no data other than the original author's example. Alternatives to the M.L.E. have been put forward simply under the criteria of estimability without thought for robustness.

To examine quantitative and qualitative robustness criteria specifically for mixtures is difficult since the notion of identifiability is easily lost. For instance for even small $\varepsilon > 0$ there may exist distributions F_{θ_1}, F_{θ_2}, $\theta_1 \neq \theta_2$, in the parametric family of mixed distributions for which $d(F_{\theta_1}, F_{\theta_2}) < \varepsilon$, but values of θ_1, θ_2 may be well separated in the parameter space. This is one more reason for examining each local minima of a distance to see the degree to which it may explain the data, rather than to proceed directly to the global minimum of the distance. Solutions to minimizing equations may have locally robust properties but these may not be reflected by the corresponding distance.

The implication of the existence of multiple local minima is that a search must be undertaken with a host of initial estimates. But to combine only two parameter values each of $(\varepsilon, \mu_1, \mu_2, \sigma_1, \sigma_2)$ in a grid of initial estimates would require execution of the equation solving or minimizing algorithm 2^5 times. Any Monte Carlo study that searches out all of the local minima becomes prohibitive in computing time. A study for the mixture of two normals was undertaken by Quandt and Ramsey (1978) by letting the initial estimate be the underlying parameter.

They minimized a distance

$$\sum_{j=1}^{k} \left(\int e^{s_j x} \, dF_n(x) - \int e^{s_j x} \, dF_\theta(x) \right)^2 .$$

For a univariate parameter and only a single weighting point $\{s_1\}$ the minimizing equations reduce to the M-estimating equations corresponding to

$$\psi(x,\theta) = e^{s_1 x} - \int e^{s_1 y} \, dF_\theta(y) .$$

This is assuming $\nabla \int e^{s_1 x} \, dF_\theta(x) \neq 0$ which is necessary for estimability. The unboundedness of ψ, and at a rate which is exponential in the tails, causes the resulting estimator to be nonrobust. This together with uncertainty in the choice of points $\{s_j\}_{j=1}^{k}$ makes it a poor choice of estimator. Discriminating between procedures is often more quickly done by examining the simple cases. Clarke and Heathcote (1978) compare the efficiency of this estimator with that of the Integrated Squared Error estimator for the location parameter again using a single weighting point. The latter is clearly superior. Kumar, Nicklin, and Paulson (1979) adopted the approach of plotting the sample and expected values of the moment generating function and characteristic functions in order to discriminate between the two types of general procedure. For both it is true that the convergence is uniform in compacts of the origin but the rate at which the convergence occurs appears faster for the characteristic function than for the moment generating function. The former has the advantage of being bounded.

We will now investigate the applicability of some specific distances based on robustness and small sample criteria to the estimation of parameters in a mixture of two normal distributions.

§7.3 The Minimal Mean Squared Error: A Robust Estimator

A known robust L_2-distance in estimating location was investigated by Knüsel (1969),

$$\omega_n^2(\theta) = \int_{-\infty}^{+\infty} (F_n(x) - F_\theta(x))^2 dx \quad . \tag{7.4}$$

Assuming a symmetric location parametric family the minimizing equations yield an M-estimator with bounded monotonic influence function

$$\psi(x) = F(x) - \tfrac{1}{2} \quad .$$

Heathcote and Silvapulle (1980) investigated this distance in the estimation of location and scale. They observed the solution $(\hat{\mu}_n, \hat{\sigma}_n)$ is unique due to the convexity of $\omega_n^2(\mu,\sigma)$. The attractive property of the location estimator is that the asymptotic variance of $\sqrt{n}(\hat{\mu}_n - \mu)$ is $(12)^{-1}\left\{\int_{-\infty}^{+\infty} f^2(x)dx\right\}^{-2}$ which corresponds to that of the Hodges-Lehmann (1963) estimator, the median of the pairwise averages $(X_i + X_j)/2$; $i,j = 1,\ldots,n$. At the normal model this corresponds to 96% efficiency. It is emphasized though that the estimators depart in behaviour away from the model.

More generally robustness of the estimator derived from (7.4) with a weighting function dW replacing Lebesgue measure, can be observed from the minimizing equations, written

$$\nabla'\omega_n^2(\theta) = \int_{-\infty}^{+\infty} (F_n(x) - F_\theta(x))\nabla' F_\theta(x)dW(x) = 0 \quad .$$

On the real line integration by parts of each component equation reveals the M-estimating equations corresponding to

$$\psi(x,\theta) = \int_{-\infty}^{x} \nabla' F_\theta(y)dW(y) - \int_{-\infty}^{+\infty}\int_{-\infty}^{x} \nabla' F_\theta(y)dW(y)dF_\theta(x) \quad .$$

For the mixture of two normal distributions we will write

$$F_\theta(x) = \varepsilon\Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) + (1-\varepsilon)\,\Phi\left(\frac{x-\mu}{\sigma_2}\right), \tag{7.5}$$

since this is a convenient form for the application considered in the next chapter. We will also take $dW(x) = dx$ and briefly describe the evaluation of a more explicit form for ψ. From this we can observe Fréchet differentiability of the resulting M-functional.

LEMMA 7.2: The influence function corresponding to the estimator that minimizes $\omega_n^2(\theta)$ given by $\nabla'\omega_n^2(\theta)$ where $\nabla = (\partial/\partial\varepsilon,\ \partial/\partial\mu,\ \partial/\partial\Delta,\ \partial/\partial\sigma_1,\ \partial/\partial\sigma_2)$, is given by

$$\psi_1(x,\theta) = 2\left[A(x;\mu,\Delta,\sigma_1,\sigma_2) - \Delta\Phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) - \sqrt{\sigma_1^2+\sigma_2^2}\,\phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) + \frac{\sigma_2}{\sqrt{\pi}}\right.$$

$$\left. - \varepsilon\left\{\Delta\left(1-2\Phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)\right) + \frac{\sigma_1+\sigma_2}{\sqrt{\pi}} - 2\sqrt{\sigma_1^2+\sigma_2^2}\,\phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)\right\}\right]$$

$$\psi_2(x,\theta) = 1 - 2\left\{\varepsilon\Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) + (1-\varepsilon)\,\Phi\left(\frac{x-\mu}{\sigma_2}\right)\right\}$$

$$\psi_3(x,\theta) = -2\,\varepsilon\,\phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) - \frac{\varepsilon^2}{\sqrt{\pi}} - 2\varepsilon(1-\varepsilon)\,\Phi\left(\frac{-\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$$

$$\psi_4(x,\theta) = 2\,\varepsilon\,\phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) - \frac{\varepsilon^2}{\sqrt{\pi}} - 2\varepsilon(1-\varepsilon)\,\frac{\sigma_1}{\sqrt{\sigma_1^2+\sigma_2^2}}\,\phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$$

$$\psi_5(x,\theta) = 2(1-\varepsilon)\,\phi\left(\frac{x-\mu}{\sigma_2}\right) - \frac{(1-\varepsilon)^2}{\sqrt{\pi}} - 2(1-\varepsilon)\varepsilon\,\frac{\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\,\phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right). \tag{7.6}$$

The term $A(x;\mu,\Delta,\sigma_1,\sigma_2) = A(x)$ say, in the first expression, is given by

$$A(x) = (x-\mu)\left\{\Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) - \Phi\left(\frac{x-\mu}{\sigma_2}\right)\right\} + \sigma_1\phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) - \sigma_2\phi\left(\frac{x-\mu}{\sigma_2}\right) + \Delta\Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right).$$

A brief outline of the proof follows. Beginning with the first expression we may write

$$\frac{1}{n}\sum_{i=1}^{n}\psi_1(X_i,\theta) = (\partial/\partial\varepsilon)\omega_n^2(\theta)$$

$$= \int_{-\infty}^{+\infty} -2\{F_n(x) - F_\theta(x)\}\left\{\Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) - \Phi\left(\frac{x-\mu}{\sigma_2}\right)\right\}dx . \quad (7.7)$$

Letting

$$a(x-\mu;\Delta,\sigma_1,\sigma_2) = \Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) - \Phi\left(\frac{x-\mu}{\sigma_2}\right)$$

it can be observed on integrating by parts that

$$A(x) = \int_{-\infty}^{x} a(y-\mu;\Delta,\sigma_1,\sigma_2)dy .$$

Now integration by parts also gives

$$\int_{-\infty}^{+\infty} F_n(x)a(x-\mu)dx = \Delta - \int_{-\infty}^{+\infty} A(x)dF_n(x) . \quad (7.8)$$

It is possible either by using calculus or repeated integrations by parts to establish the following identities

(1) $$\int_{-\infty}^{+\infty} \frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right)\Phi\left(\frac{x-\mu_2}{\sigma_2}\right)dx = \Phi\left(\frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$$

(2) $$\int_{-\infty}^{+\infty} a(x)^2dx = (\mu_1-\mu_2)\left(2\,\Phi\left(\frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) - 1\right) - \frac{\sigma_1+\sigma_2}{\sqrt{\pi}}\,\phi\left(\frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$$

(3) $$\int_{-\infty}^{+\infty} \Phi\left(\frac{x-\mu}{\sigma_2}\right)a(x-\mu)dx = \Delta - \Delta\Phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) - \sqrt{\sigma_1^2+\sigma_2^2}\,\phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) + \frac{\sigma_2}{\sqrt{\pi}} .$$

Returning to (7.7) we can write

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i(X_i,\theta) = \int_{-\infty}^{+\infty} -2\{F_n(x) - \varepsilon a(x-\mu) - \Phi\left(\frac{x-\mu}{\sigma_2}\right)\}a(x-\mu)dx$$

$$= -2\left\{\int_{-\infty}^{+\infty} F_n(x)a(x-\mu)dx - \varepsilon\int_{-\infty}^{+\infty} a(x-\mu)^2dx\right.$$

$$\left. - \int_{-\infty}^{+\infty}\Phi\left(\frac{x-\mu}{\sigma_2}\right)\, a(x-\mu)dx\right\}$$

from which we obtain the expression for $\psi_1(x,\theta)$ using (7.8) and (1), (2), and (3).

Expressions for $\psi_2 - \psi_5$ may be obtained in the same manner. A useful result in establishing (1), (2), and (3) is to set $\gamma = (\mu_1,\mu_2,\sigma_1,\sigma_2)$ and

$$\alpha(\gamma) = \frac{\mu_1\sigma_1^2 + \mu_2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\quad,\quad \beta(\gamma) = \frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\,,$$

and observe that

$$\phi\left(\frac{x-\mu_1}{\sigma_1}\right)\phi\left(\frac{x-\mu_2}{\sigma_2}\right) = \phi\left(\frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)\phi\left(\frac{x-\alpha(\gamma)}{\beta(\gamma)}\right)\,.$$

To see its applicability we derive $\psi_4(x,\theta)$. Now

$$\frac{1}{n}\sum_{i=1}^{n}\psi_4(X_i,\theta) = (\partial/\partial\sigma_1)\omega_n^2(\theta)$$

$$= \int_{-\infty}^{+\infty} 2\{F_n(x) - F_\theta(x)\}\varepsilon\,\frac{x-\mu+\Delta}{\sigma_1^2}\,\phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right)dx$$

by parts

$$= 2\varepsilon\int_{-\infty}^{+\infty}\phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right)dF_n(x)$$

$$- 2\varepsilon\int\phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right)\left\{\frac{\varepsilon}{\sigma_1}\,\phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) + \frac{1-\varepsilon}{\sigma_2}\,\phi\left(\frac{x-\mu}{\sigma_2}\right)\right\}dx$$

$$= \frac{2\varepsilon}{n}\sum_{i=1}^{n}\phi\left(\frac{X_i-\mu+\Delta}{\sigma_1}\right) - \sqrt{\frac{2}{\pi}}\,\frac{\varepsilon^2}{\sigma_1}\int\phi\left(\frac{\sqrt{2}\,x}{\sigma_1}\right)dx$$

$$- \frac{2\varepsilon(1-\varepsilon)}{\sigma_2} \phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) \int \phi\left(\frac{x-\alpha}{\beta}\right) dx \ ,$$

where

$$\alpha = \alpha(\theta) = \frac{-\Delta\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \ , \quad \beta = \beta(\theta) = \frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

$$= \frac{2}{n} \sum_{i=1}^{n} \phi\left(\frac{X_i-\mu+\Delta}{\sigma_1}\right) - \frac{\varepsilon^2}{\sqrt{\pi}} - 2\varepsilon(1-\varepsilon) \frac{\sigma_1}{\sqrt{\sigma_1^2 + \sigma_2^2}} \phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) \ ,$$

from which we easily obtain ψ_4.

A major reason for using the minimal mean squared error estimator for mixtures of two normal distributions is the following:

THEOREM 7.1:

Let $T[\psi,\cdot]$ be the M-functional defined by the selection functional $\|\theta - \theta_o\|$ and the M.M.S.E. influence function $\psi(x,\theta)$ given by (7.6). Let the parameter space be $\Theta = \{\theta | -\infty < \mu_1\Delta < \infty, \sigma_1 > 0, \sigma_2 > 0, 0 < \varepsilon < 1\}$ Then for each $\theta_o \in \Theta$ at which $M(\theta_o)$, defined in conditions A of §2.3, is positive definite, $T[\psi,\cdot]$ is Fréchet differentiable at F_{θ_o} with respect to $(\mathcal{G},d)$. The distance d can be either of d_k, d_L, or d_p.

PROOF: We first establish conditions A. Clearly A0 is satisfied from the expression (7.6). Suppose $\theta_o = (\varepsilon_o,\mu_o,\Delta_o,\sigma_{10},\sigma_{20})$. Then for any $0 < \delta < \frac{1}{2} \min(\varepsilon_o,1-\varepsilon_o,\sigma_{10},\sigma_{20})$ the function $\psi(x,\theta)$ is uniformly bounded above in norm by a constant for all $(x,\theta) \in E \times \bar{U}_\delta(\theta_o)$. Note that $(\partial/\partial x)\psi(x,\theta) = \nabla' F_\theta(x)$ is also uniformly bounded above by some constant on $E \times \bar{U}_\delta(\theta_o)$. From Lemma 1.4 the family $\{\psi(\cdot,\theta) | \theta \in \bar{U}_\delta(\theta_o)\}$ is equicontinuous. Similarly the family of matrix functions $\{\nabla\psi(\cdot,\theta) | \theta \in \bar{U}_\delta(\theta_o)\}$ involve terms which in the observation parameter, x, either decrease exponentially, or are uniformly bounded as $|x|$ tends to $+\infty$. Equicontinuity follows similarly since $(\partial/\partial x)\nabla\psi(x,\theta) = \nabla\nabla' F_\theta(x)$

which is uniformly bounded on $E \times \bar{U}_{\delta}(\theta_o)$. Then assumption A4 follows from Theorem 4.1 for the Prokhorov neighbourhood, and also the Kolmogorov and Lévy neighbourhoods by Remark 4.1. Finally since $\psi(x,\theta_o)$ is a function of total bounded variation so that (4.8) is valid,

$$\int \psi(x,\theta_o)d(G - F_{\theta_o})(x) = O(d_k(G,F_{\theta_o})) .$$

Since each F_{θ_o} is an absolutely continuous distribution function on the real line possessing a bounded density this is true also for d_L and d_p. Given that $M(\theta_o)$ is positive definite conditions A are satisfied and so are the assumptions of Theorem 4.3. This proves the theorem.

The actual M.M.S.E. selection functional is

$$f_G(\theta) = \int (G(x) - F_\theta(x))^2 dx .$$

Unfortunately this does not satisfy the weak continuity that is sufficient for robustness since for any given $\kappa > 0$ and any $\theta \in \Theta$, given $\varepsilon > 0$ and setting $G = (1-\kappa)F_\theta + \kappa\delta_y$ so that $d_p(G,F_{\theta_o}) \leq \kappa$ it is possible to make $f_G(\theta) > \varepsilon$ by letting $y \to \infty$. For mixtures of normal distributions the only statistic that does not fail in this regard is the Cramér-von Mises statistic. The I.S.E. is precluded since $\{\cos(t\,\cdot)|t \in E\}$ and $\{\sin(t\,\cdot)|t \in E\}$ are not equicontinuous. So we conclude that there exists a locally robust root of the minimizing equations of the M.M.S.E. but robust identification of this root is not guaranteed by the M.M.S.E. selection functional.

Should the true proportion fall on the perimeter of Θ, $\varepsilon_o = 0$ or 1, F_{θ_o} may be represented by any one of the parameters satisfying $0 \leq \varepsilon \leq 1$, $\Delta = 0$, $\sigma_1 = \sigma_2$, and so $M(\theta_o)$ will be singular. Values of $T[\psi,G]$ need not fall within the parameter space unless T should be

defined as such. For instance, the practicing statistician would simply truncate an estimate of $\hat{\varepsilon}_n < 0$ to the value 0, and use a single location and scale family for the estimation.

§7.4 The Minimal Mean Squared Error: Statistical Application

The expression for the asymptotic variance of the M.M.S.E. estimator is obtained directly from the expansion of $\nabla'\omega_n^2(\theta)$. Abbreviating F_θ to F it is observed under the model F

$$\nabla\nabla'\omega_n^2(\theta) \xrightarrow{a.s.} \left(\int_{-\infty}^{+\infty} \frac{\partial F}{\partial \theta_i}(x) \cdot \frac{\partial F}{\partial \theta_j}(x)dx \right) = M(\theta) .$$

Also

$$\sqrt{n}\, K_n(\theta) = \sqrt{n}\, \nabla'\omega_n^2(\theta) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int [I_j(x) - F(x)] \frac{\partial F}{\partial \theta_i}(x)dx \right) ,$$

which is a suitably normalized sum of i.i.d. random variables so that the asymptotic variance

$$\Sigma_\theta = \left(\operatorname{cov}\left[\int [I_j(x) - F(x)] \frac{\partial F}{\partial \theta_i}(x)dx, \int [I_j(y) - F(y)] \frac{\partial F}{\partial \theta_i}(y)dy \right] \right)$$

$$= \left(\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [F\{\min(x,y)\} - F(x)F(y)] \frac{\partial F}{\partial \theta_i}(x) \frac{\partial F}{\partial \theta_j}(y)dxdy \right) .$$

Splitting the integral into two half planes and using Fubini's theorem

$$\Sigma_\theta = \left(\int_{-\infty}^{+\infty} F(x) \frac{\partial F}{\partial \theta_i}(x) \int_x^{\infty} \frac{\partial F}{\partial \theta_j}(y)dydx + \int_{-\infty}^{+\infty} F(x) \frac{\partial F}{\partial \theta_j}(x) \int_x^{\infty} \frac{\partial F}{\partial \theta_i}(y)dydx \right.$$

$$\left. - \int_{-\infty}^{+\infty} F(x) \frac{\partial F}{\partial \theta_i}(x)dx \int_{-\infty}^{+\infty} F(x) \frac{\partial F}{\partial \theta_j}(x)dx \right) .$$

In particular, if $i = j$ and setting $\Sigma_\theta = (\sigma_{ij}(\theta))$ and

$$L_i(x) = \int_{-\infty}^{x} \frac{\partial F}{\partial \theta_i}(y)dy ,$$

the variance terms are given by

$$\sigma_{ii}(\theta) = 2L(\infty)\int_{-\infty}^{+\infty} F(x)\,\frac{\partial F}{\partial \theta_i}(x)dx - L^2(\infty) + \int_{-\infty}^{+\infty} L^2(x)dF(x) - \left\{\int_{-\infty}^{+\infty} F(x)\,\frac{\partial F}{\partial \theta}(x)dx\right\}^2 .$$

This expression allows simple application of numerical integration methods on the real line, or in the case of mixtures further decomposition can be made. In particular the estimating equation for proportion yields the explicit representation

$$\hat{\varepsilon}_n = \frac{\int_{-\infty}^{+\infty}\left\{F_n(x) - \Phi\left(\frac{x-\mu}{\sigma_2}\right)\right\} a(x-\mu)dx}{\int_{-\infty}^{+\infty} a(x)^2 dx} , \tag{7.9}$$

or if using $\frac{1}{n}\sum_{i=1}^{n} \psi_1(X_i,\theta) = 0$ can be written

$$\hat{\varepsilon}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{A(X_i) - \Delta\Phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) - \sqrt{\sigma_1^2+\sigma_2^2}\,\phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) + \frac{\sigma_2}{\sqrt{\pi}}}{\left\{\Delta\left(1-2\,\Phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)\right) + \frac{\sigma_1+\sigma_2}{\sqrt{\pi}} - 2\sqrt{\sigma_1^2+\sigma_2^2}\,\phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)\right\}} \tag{7.10}$$

$$= \frac{1}{n}\sum_{i=1}^{n} Z_i \quad \text{say} .$$

So $\hat{\varepsilon}_n$ is the sum of i.i.d. bounded random variables Z_i. Clearly $E_\theta[\hat{\varepsilon}_n] = E_\theta[Z_i]$ and by the S.L.L.N. $\hat{\varepsilon}_n \xrightarrow{a.s.} E_\theta[Z]$. The almost sure limits of $\hat{\varepsilon}_n$ can also be realized by observing $F_n(x) \xrightarrow{a.s.} F_\theta(x) = \varepsilon a(x-\mu) + \Phi\left(\frac{x-\mu}{\sigma_2}\right)$ uniformly in x. By dominated convergence and (7.9)

$$E[\hat{\varepsilon}_n] = E[Z] = \varepsilon .$$

That is the proportion estimator is unbiased. The usual C.L.T. applies

so that

$$\sqrt{n}(\hat{\varepsilon}_n - \varepsilon) \xrightarrow{D} N\{0,\sigma_{11}(\theta)/\lambda_{11}^2(\theta)\} ,$$

where

$$\lambda_{11}(\theta) = \int_{-\infty}^{+\infty} a^2(x)dx = \Delta\left[2\Phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) - 1\right] - \frac{\sigma_1+\sigma_2}{\sqrt{\pi}} + 2\sqrt{\sigma_1^2+\sigma_2^2}\ \phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right).$$

Observe that $L_1(x) = A(x)$ so that

$$\sigma_{11}(\theta) = 2\Delta\int_{-\infty}^{+\infty} a(x)F_\theta(x)dx - \Delta^2 + \int_{-\infty}^{+\infty} A^2(x)dF_\theta(x) - \left\{\int_{-\infty}^{+\infty} a(x)F_\theta(x)dx\right\}^2.$$

We observe by (2) and (3) that

$$\int_{-\infty}^{+\infty} a(x)F_\theta(x)dx = \Delta - \Delta\Phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) - \sqrt{\sigma_1^2+\sigma_2^2}\ \phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) + \frac{\sigma_2}{\sqrt{\pi}}$$

$$+ \varepsilon\left\{\Delta\left(2\Phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) - 1\right) - \frac{\sigma_1+\sigma_2}{\sqrt{\pi}} + 2\sqrt{\sigma_1^2+\sigma_2^2}\ \phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)\right\}.$$

The variance can then be evaluated. By writing

$$\int_{-\infty}^{+\infty} A(x)^2 dF_\theta(x) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}}\{\varepsilon A^2(\sqrt{2}\sigma_1 y - \Delta + \mu) + (1-\varepsilon)A^2(\sqrt{2}\sigma_2 y + \mu)\}\exp(-y^2)dy$$

this term can be evaluated accurately by Gauss Hermite integration since $A^2(x)$ is infinitely differentiable. Later examples use ten interpolation points.

To implement the M.M.S.E. estimator the corresponding nonlinear M-estimating equations are solved using a nonlinear equation solving routine. Here we used ZSYSRB in the A.N.U. library of subroutines ANULIB2. This is based on a method of Brent (1973). The normal distribution is evaluated quickly and efficiently using Hastings approximation detailed in (26.2.17) of Abramowitz and Stegun (1970).

Multiple roots can and often do exist. Should $\omega_n^2(\theta)$ be used as the selection statistic then it must be evaluated numerically at each root of the equations. Alternatively one may wish to minimize $\omega_n^2(\theta)$ in a search for the local minima, rather than resorting to the minimizing equations. However this latter approach appears numerically time consuming given the numerical integration required at successive iterations. To evaluate $\omega_n^2(\theta)$ it is necessary to truncate the integral at appropriate points b_1, b_2 so that an absolute error of integration is less than a prescribed $\delta > 0$. Then b_1, b_2 depend on δ and θ. First order the data.

$$-\infty < X_{(1)} < X_{(2)} < \ldots\ldots < X_{(n)} < \infty .$$

Then write

$$\omega_n^2(\theta) = \int_{-\infty}^{X_{(1)}} F_\theta(x)^2 dx + \sum_{i=1}^{n-1} \int_{X_{(i)}}^{X_{(i+1)}} \left(\frac{i}{n} - F_\theta(x)\right)^2 dx + \int_{X_{(n)}}^{\infty} (1 - F_\theta(x))^2 dx .$$

The truncation of the first integral is at some point $b_1(\delta,\theta)$ which satisfies

$$\int_{-\infty}^{b_1} F_\theta^2(x)dx < \delta/4 . \tag{7.11}$$

LEMMA 7.3: The value $k(\delta) = -\sqrt{-\ell n(2\pi\delta)}$ satisfies

$$\left|\int_{-\infty}^{k} x\,\Phi(x)^2 dx\right| < \delta, \text{ whenever } \delta < (20\pi)^{-1} .$$

COROLLARY 7.1: $\int_{-\infty}^{k(\delta)} \Phi(x)^2 dx < \delta$ whenever $k(\delta) < -1$.

PROOF OF LEMMA 7.3: Integration by parts gives

$$\int_{-\infty}^{x} \Phi(x)dx = x\Phi(x) + \phi(x) > 0 .$$

That implies

$$|x\Phi(x)| < \phi(x) \quad \text{for} \quad x < 0 .$$

Then

$$\left|\int_{-\infty}^{k} x\Phi(x)^2 dx\right| = \int_{-\infty}^{k} |x\Phi(x)|\Phi(x)dx$$

$$< \int_{-\infty}^{k} \phi(x)\Phi(x)dx .$$

Since $\phi(x)$ is increasing on $(-\infty,0]$

$$< \phi(k)[k\Phi(k) + \phi(k)]$$

$$< \phi(k)^2$$

$$= \delta .$$

The lemma is proved.

If we consider the inequality

$$\int_{-\infty}^{b_1} F_\theta(x)^2 dx < 2\int_{-\infty}^{a} \varepsilon^2 \Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right)^2 + (1-\varepsilon)^2 \Phi\left(\frac{x-\mu}{\sigma_2}\right)^2 dx$$

$$\leq 2\{\varepsilon^2\sigma_1 + (1-\varepsilon)^2\sigma_2\} \int_{-\infty}^{\max\left(\frac{b_1-\mu+\Delta}{\sigma_1}, \frac{b_1-\mu}{\sigma_2}\right)} \Phi(x)^2 dx ,$$

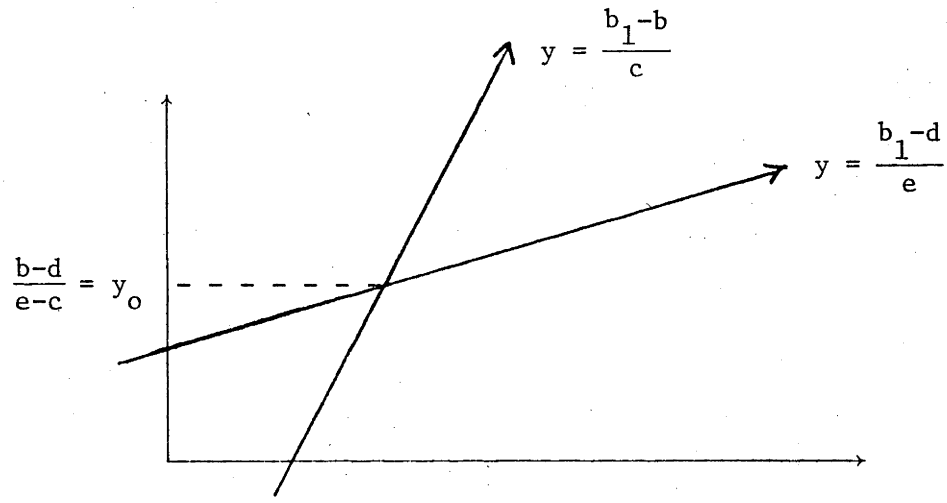
and choose $b_1(\delta,\theta)$ so that

$$\max\left(\frac{b_1-\mu+\Delta}{\sigma_1}, \frac{b_1-\mu}{\sigma_2}\right) = k\left(\frac{\delta}{8\{\varepsilon^2\sigma_1+(1-\varepsilon)^2\sigma_2\}}\right) \tag{7.12}$$

then (7.11) holds.

We solve

$$\max\left\{\frac{b_1-b}{c}\ ,\ \frac{b_1-d}{e}\right\} = k \quad \text{when} \quad c > e > 0$$



For $k > y_o$ it can be seen from the diagram that the solution to the equation is $b_1 = ek+d$. If $k < y_o$ it is $b_1 = ck+b$. So the solution to (7.12) where $y_o = \Delta/(\sigma_1-\sigma_2)$ is the following.

If $\sigma_1 < \sigma_2$, the value of

$$b_1 = \begin{cases} k\sigma_2 + \mu & k > \Delta/(\sigma_1-\sigma_2) \\ k\sigma_1 + \mu + \Delta & k < \Delta/(\sigma_1-\sigma_2) \end{cases}$$

and if $\sigma_1 < \sigma_2$

$$b_1 = \begin{cases} k\sigma_2 + \mu & k < \Delta/(\sigma_1-\sigma_2) \\ k\sigma_1 + \mu - \Delta & k > \Delta/(\sigma_1-\sigma_2)\ . \end{cases}$$

Note that if $\sigma_1 = \sigma_2$

$$b_1 = \begin{cases} k\sigma_1 + \mu - \Delta & \Delta \geq 0 \\ k\sigma_2 + \mu & \Delta \leq 0\ . \end{cases}$$

Analogously to set b_2 so that

$$\int_{b_2}^{\infty} (1 - F_\theta(x))^2 dx < \delta/4$$

consider the inequality

$$\int_{b_2}^{\infty} (1 - F_\theta(x))^2 dx \leq 2\{\varepsilon^2\sigma_1 + (1-\varepsilon^2)\sigma_2\} \int_{-\infty}^{-\min\left(\frac{b_2-\mu+\Delta}{\sigma_1},\ \frac{b_2-\mu}{\sigma_2}\right)} \Phi(x)^2 dx.$$

If $\sigma_1 > \sigma_2$ the solution for b_2 is

$$b_2 = \begin{cases} \mu - \Delta - \sigma_1 k & k \leq \dfrac{\Delta}{\sigma_2-\sigma_1} \\ \mu - \sigma_2 k & k \geq \dfrac{\Delta}{\sigma_2-\sigma_1} \end{cases}.$$

If $\sigma_1 < \sigma_2$ then

$$b_2 = \begin{cases} \mu - \Delta - \sigma_1 k & k \geq \Delta/(\sigma_2-\sigma_1) \\ \mu - \sigma_2 k & k \leq \Delta/(\sigma_2-\sigma_1) \end{cases}$$

and if $\sigma_1 = \sigma_2$

$$b_2 = \begin{cases} \mu - \Delta - \sigma_1 k & \Delta \leq 0 \\ \mu - \sigma_2 k & \Delta \geq 0 \end{cases},$$

where k is given by Lemma 7.3 and

$$k = k[\delta/\{8(\varepsilon^2\sigma_1 + (1-\varepsilon)^2\sigma_2)\}] .$$

The integral is then evaluated by integration over the finite interval

$$[\min\{b_1, X_{(1)}\}, \max\{b_2, X_{(n)}\}] .$$

An absolute error of integration is then specified to be less than $\delta/2$.

CHAPTER 8

SMALL SAMPLE BEHAVIOUR

§8.1 Small Sample Comparison of Least Squares Estimators of ε

While Hill (1963) proposed that large numbers of observations were required for reasonable accuracy in estimating the proportion parameter in small separations of two normal distributions, moderate separations allow simple application of the many "least squares" or minimal distance estimators. The M.L.E. is included in this class by virtue of the fact that the maximum likelihood equations can be derived by formal minimization of

$$\int_{-\infty}^{+\infty} \frac{d(F_n - F_\theta)^2}{dW} , \tag{8.1}$$

where W has derivative $\omega(x) = f_\theta(x)$. This example was given by Bartlett and Macdonald (1968). It is another case where an unnatural weight function (for the distance is infinite in this case) will yield useful estimating equations.

Some methods of estimation in mixtures have already been observed to fail a robustness criterion because of unbounded influence curves. But to obtain further information on remaining distance estimators it is useful to examine small sample behaviour of statistics, particularly for the mixing proportion which has important applications as noted in Odell and Basu (1976) and James (1978).

Macdonald (1971) proposed the parameter minimizing the Cramér von Mises statistic

$$S_n^2(\theta) = \int_{-\infty}^{+\infty} \{F_n(x) - F_\theta(x)\}^2 dF_\theta(x)$$

as an estimator in mixtures. He was motivated by the significant bias in the estimates of mixing proportion in sample sizes as large as 200 that was exhibited in Table 1. of Choi and Bulgren (1968). The latter authors minimized the statistic

$$\int_{-\infty}^{+\infty} (F_n(x) - F_\theta(x))^2 dF_n(x)$$

which in terms of the Heaviside functions $I_j(x)$, that are equivalent to $I_{(-\infty,x]}(X_j)$, is written

$$\frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{+\infty} \left(\frac{1}{n} \sum_{j=1}^{n} I_j(x) - F_\theta(x) \right)^2 dI_j(x) \ .$$

Macdonald noted that with the interpretation of $I_j(x)dI_j(x)$ as $\frac{1}{2}dI_j(x)$ minimizing this distance was equivalent to minimizing $S_n^2(\theta)$. The Cramer-von Mises distance had been used by Blackman (1955) to estimate location and possessed the convenient representation

$$S_n^2(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(F_\theta(X_{(i)}) - \frac{(i-\frac{1}{2})}{n} \right)^2 + \frac{1}{12n^2} \ . \qquad (8.2)$$

Monte Carlo work of Macdonald revealed a significant reduction in bias in the proportion estimator using $S_n^2(\theta)$. The estimating equations for this statistic are

$$0 = \nabla' S_n^2(\theta) = \int \{F_n(x)-F_\theta(x)\}^2 \nabla' f_\theta(x) dx -$$

$$- 2 \int \{F_n(x)-F_\theta(x)\}\{\nabla' F_\theta(x)\} f_\theta(x) dx \ .$$

The first term is $O_p(n^{-1})$ at F_θ and is asymptotically negligible in the estimation, not contributing to the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ at the model. The degree to which bias is incurred can be

studied through the usual Taylor expansion. For a univariate parameter we observe

$$0 = \nabla S_n^2(\hat{\theta}_n) = \nabla S_n^2(\theta) + \nabla^2 S_n^2(\theta)(\hat{\theta}_n - \theta) + \tfrac{1}{2}\nabla^3 S_n^2(\theta)(\hat{\theta}_n - \theta)^2 + O_p(n^{-3/2}) \ .$$

Letting

$$\eta_\theta(x) = \int_{-\infty}^{x} \{\nabla F_\theta(y)\} f_\theta(y)dy$$

it can be observed (see Appendix 4) that

$$E_\theta[\nabla S_n^2(\theta)] = n^{-1}\{\eta_\theta(\infty) - 2\int \eta_\theta(x) f_\theta(x)dx\} \ ,$$

and

$$E_\theta[\nabla^2 S_n^2(\theta)] = \int \{\nabla F_\theta(x)\}^2 f_\theta(x)dx + n^{-1}\int \{\nabla^2 F_\theta(x)\}\{F_\theta(x) - \tfrac{1}{2}\} f_\theta(x)dx \ .$$

These extra terms of order n^{-1} suggest the possibility of a still significant small sample bias in the Macdonald Cramér Von Mises (M.C.V.M.) statistic. We contrast this with the M-estimator where each of the terms in the corresponding Taylor expansion is unbiased toward its asymptotic value. The approximating equation for normality of the M.C.V.M. statistic is

$$\int \{\nabla F_\theta(x)\}^2 f_\theta(x)dx(\hat{\theta}_n - \theta)\sqrt{n} = -\nabla S_n^2(\theta)\sqrt{n} \ .$$

Here it can be shown that

$$\begin{aligned} n \ \mathrm{var}_\theta[\nabla S_n^2(\theta)] = 4 \Bigg[& \int \eta_\theta(x)^2 f_\theta(x)dx - 4\eta_\theta(\infty)\int \eta_\theta(x) f_\theta(x)dx \\ & + 3\left(\int \eta_\theta(x) f_\theta(x)dx \right)^2 + \eta_\theta^2(\infty) \Bigg] + 4n^{-1}\Bigg[-\frac{5}{2}\eta_\theta^2(\infty) \\ & + 7\left(\eta_\theta(\infty) - \int \eta_\theta(x) f_\theta(x)dx\right) \int \eta_\theta(x) dF_\theta(x) \\ & - \int \eta_\theta(x)^2 dF_\theta(x) \Bigg] + O(n^{-2}) \ . \end{aligned}$$

The bias term of $O(n^{-1})$ contrasts with that of the M-estimating equations where

$$n \operatorname{var}_\theta[K_n(\theta)] = \operatorname{var}_\theta[\psi(X,\theta)] \ .$$

In fact the M.M.S.E. estimator and I.S.E. estimators are unbiased for the proportion parameter. But the M.L.E. is still known to incur a small sample bias of

$$b(\varepsilon,F_\theta) = \frac{-2\int f_\theta'^3/f_\theta^2 + \int f_\theta'^2/f_\theta \int f_\theta'^3/f_\theta^3}{2\int f_\theta'^2/f_\theta}$$

where $f_\theta' = (\partial/\partial\varepsilon)f_\theta(x)$. This is nonzero.

In view of the fact that there appears to be no optimal estimator it is instructive to compare various forms of minimal distance estimators. Letting $W(t) = \exp\{-\eta^2 t^2/2\}$ the I.S.E. estimator for proportion can be derived from Lemma 7.1. It has the explicit representation

$$\hat{\varepsilon}_n = A/B \ , \tag{8.3}$$

where

$$A = [(\sigma_1^2+\eta^2)^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} \phi\left(\frac{X_i-\mu+\Delta}{\sqrt{\sigma_1^2+\eta^2}}\right) - (\sigma_2^2+\eta^2)^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} \phi\left(\frac{X_i-\mu}{\sqrt{\sigma_2^2+\eta^2}}\right)$$

$$- (\sigma_1^2+\sigma_2^2+\eta^2)^{-\frac{1}{2}} \phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2+\eta^2}}\right) + \frac{1}{\sqrt{2\pi}} (2\sigma_2^2+\eta^2)^{-\frac{1}{2}}]$$

and

$$B = \left[\frac{1}{\sqrt{2\pi}} (2\sigma_1^2+\eta^2)^{-\frac{1}{2}} - 2(\sigma_1^2+\sigma_2^2+\eta^2)^{-\frac{1}{2}} \phi\left(\frac{\Delta}{\sqrt{\sigma_1^2+\sigma_2^2+\eta^2}}\right) + \frac{1}{\sqrt{2\pi}} (2\sigma_2^2+\eta^2)^{-\frac{1}{2}}\right].$$

To concentrate the weight function near the origin, η should be chosen sufficiently large. In the absence of any criterion for choosing η we let $T = [t_1,t_2]$ be the region that is most weighted (relatively) by $W(t)$. Then the information to be gained about F_θ should be gleaned if T is restricted so that $T.X$ lies within approximately one period of $\exp\{i \cdot\}$ with probability greater than .9. With this criterion

and the knowledge that $W(t)$ is a normal weighting distribution the value of η was chosen to be $(3/\pi)\beta.x_o$, with $\beta = 2$ and

$$x_o = \max\{|F_\theta^{-1}(.05)|, |F_\theta^{-1}(.95)|\}.$$

In practice η would need to be estimated, for instance by using the order statistics $X_{([n\alpha])}$, $X_{([n(1-\alpha)])}$, $\alpha = .05$ to estimate the 5 and 95% quantiles. Investigations showed the I.S.E. to be stable should the continuous weighting distribution be estimated this way. But this was not the case with a discrete weighting distribution, whose atoms were given by four equally weighted points

$$t_1 = -\pi/\beta x_o \ , \quad t_2 = \pi/\beta x_o$$

$$t_3 = -\pi/2\beta x_o, \quad t_4 = \pi/2\beta x_o \ , \ \beta = 2 \ .$$

Estimating the weighting distribution affects the asymptotic distribution of the statistics and even the unbiasedness of the proportion estimator.

In a Monte Carlo study with 500 replications of each experiment with data generated from the distribution

$$F_\theta(x) = \varepsilon\Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) + (1-\varepsilon)\Phi\left(\frac{x-\mu}{\sigma_2}\right) \ ,$$

for various parameter values and different sample sizes, five "least squares" estimators of proportion were compared. Samples were generated from GGNML in the International Mathematical Statistical Libraries. The M.M.S.E. and I.S.E. estimators of proportion were evaluated from their explicit expressions (7.10) and (8.3) respectively. Also the I.S.E. estimator with discrete weighting distribution (whose atoms were chosen assuming F_θ to be known) was evaluated from its explicit expression. The M.C.V.M. statistic was obtained by numerical minimization of the expression (8.2), truncating at zero or one if necessary. The M.L.E.

was obtained iteratively from its equation maximizing the log likelihood, again with truncation at zero or one if necessary.

With small separations and small samples sizes the M.M.S.E. and I.S.E. estimators given by explicit representations ranged outside the parameter space. Only then the truncated M.C.V.M. and M.L.E. statistics showed significantly smaller mean squared errors. For two equivalent populations separated by a location shift the M.M.S.E. and I.S.E. with continuous weight function performed slightly better in terms of small sample bias than did the M.L.E. or M.C.V.M., but the M.L.E. appeared to have smaller mean squared error. When the two populations differed in scale the M.C.V.M. performed poorly having a significant small sample bias. This was while the M.M.S.E. and I.S.E. with continuous weight function remained relatively unaffected. The M.L.E. dominated the other statistics when a large separation of the populations was apparent.

The I.S.E. estimator with discrete weighting points performed near an acceptable level when F_θ was known. But large samples were required to successfully estimate the proportion when weighting points were estimated.

The M.M.S.E. estimator performed consistently well. Asymptotic variances corresponded closely with the mean squared errors which lends support to both the theory and the Monte Carlo. This estimator showed a decided advantage when the locations of the two populations were close but dispersion were different. This is the advantage of weighting by Lebesgue measure. It appears then that the M.M.S.E. must be considered alone amongst the least squares estimators considered here as providing a satisfactory estimation procedure for estimating parameters in the mixture of two normal distributions.

TABLE 8.1

Monte Carlo comparison of least squares proportion estimators

					M.S.E.E.		M.C.V.M.	
n	Δ/σ_1	σ_2/σ_1	ε	as var/n	Mean	M.S.E.	Mean	M.S.E.
50	.25	1.	.5	.3408	.5565 ± .0514	.3460	.5252 ± .0333	.1447
10	.5	1.	.5	.4365	.5421 ± .0554	.4007	.5279 ± .0341	.1517
50	.5	1.	.5	.0873	.5068 ± .0356	.0854	.5071 ± .0237	.0728
10	1.	1.	.5	.1234	.5193 ± .0277	.1022	.5214 ± .0249	.0813
50	1.	1.	.5	.0247	.4956 ± .0124	.0201	.4943 ± .0127	.0210
20	1.	1.	.75	.0585	.7652 ± .0200	.0520	.7471 ± .0180	.0423
20	5.	1.	.5	.0135	.4999 ± .0029	.0011	.4995 ± .0045	.0011
20	.5	2.	.75	.0678	.7506 ± .0222	.0638	.7002 ± .0230	.0710
10	0	2.	.5	.1600	.4787 ± .0336	.1468	.3576 ± .0294	.1324
50	0	2.	.5	.0322	.4814 ± .0142	.0265	.4287 ± .0165	.0402
10	.5	2.	.5	.1950	.4869 ± .0392	.1994	.4402 ± .0315	.1325

TABLE 8.1 (Continued)

				M.L.E.		I.S.E. (exponential weight function)	
n	Δ/σ_1	σ_2/σ_1	ε	Mean	M.S.E.	Mean	M.S.E.
50	.25	1.	.5	.5269 ± .0322	.1436	.5569 ± .0502	.3312
10	.5	1.	.5	.5346 ± .0340	.1510	.5472 ± .0554	.4533
50	.5	1.	.5	.5080 ± .0236	.0726	.5077 ± .0251	.0820
10	1.	1.	.5	.5240 ± .0264	.0909	.5190 ± .0284	.1050
50	1.	1.	.5	.4980 ± .0125	.0203	.4974 ± .0124	.0198
20	1.	1.	.75	.7514 ± .0181	.0427	.7685 ± .0198	.0515
20	5.	1.	.5	.5008 ± .0017	.0004	.5040 ± .0038	.0019
20	.5	2.	.75	.7444 ± .0179	.0415	.7514 ± .0274	.8977
10	0	2.	.5	.4939 ± .0290	.1091	.4734 ± .0373	.1818
50	0	2.	.5	.4854 ± .0139	.0255	.4800 ± .0163	.0349
10	.5	2.	.5	.4965 ± .0295	.1134	.5138 ± .0581	.4382

TABLE 8.1 (Continued)

					(Fixed Weighting Points)		(Estimated Weighting Points)	
n	Δ/σ_1	σ_2/σ_1	ε	as var/n	Mean	M.S.E.	Mean	M.S.E.
50	.25	1.	.5	1.1753	.5737 ± .0971	1.2302	.7148 ± .1143	1.744
10	.5	1.	.5	1.1130	.5050 ± .0919	1.0973	+ ∞	
50	.5	1.	.5	.2227	.5010 ± .0420	.2295	.5674 ± .0535	.3759
10	1.	1.	.5	.2056	.5212 ± .0362	.1711	+ ∞	
50	1.	1.	.5	.0411	.4868 ± .0167	.0365	.5005 ± .0184	.0440
20	1.	1.	.75	.0902	.7549 ± .0262	.0892	.8181 ± .0342	.1561
20	5.	1.	.5	.0135	.4989 ± .0028	.0010	.4982 ± .0027	.0010
20	.5	2.	.75	.1341	.7498 ± .0311	.1256	.8611 ± .1272	2.114
10	0	2.	.5	.2521	.4869 ± .0433	.2437	+ ∞	
50	0	2.	.5	.0504	.4800 ± .0155	.0316	.4658 ± .0160	.0343
10	.5	2.	.5	.2239	.4404 ± .0348	.1609	.3986 ± .0796	.8331

§8.2 Application of a Fréchet Differentiable M-functional to Seismic Data

The complexities inherrant in estimating parameters of a mixture of two normal distributions using minimal distance methods make it important to investigate applicability of these procedures in real data situations. Frechet differentiability of the M-estimator that is a solution to the minimizing equations of the Mean Squared Error distance indicates a robustness of this statistic. This combined with the consistently good small sample behaviour under the model suggests that it is a useful statistic to investigate and apply to real data. Other established statistics, the M.L.E., M.G.F., and method of moments are discarded on the grounds of nonrobustness. The M.C.V.M. statistic has poor small sample behaviour when the dispersion parameters differ which leads us to drop this statistic even though it appears to have robust asymptotic behaviour. Finally the I.S.E. which seems to be the most appealing of the statistics other than the M.M.S.E. is dropped since the behaviour of the estimator depends much on the relationship between the true parameter value and the weighting distribution. There is no clear guide as to an overall good weight function.

A particular area in which data may be considered "rough" is seismology. Influences of many extraneous factors cause this. They are exemplified by the nonhomogeneous nature of the earth, and the measuring techniques used (Jeffreys 1970, P.71 and 1967, P.214). The application of interest to geologists is the following:

Recordings of all earthquakes of reasonable magnitude have been kept in recent years, for it is known that their attributes give insight into the structure of the earth. The focus of a quake is deduced, and time taken for a primary or longitudinal wave to travel through an epicentral distance to an observation point is recorded.

DIAGRAM 8.1

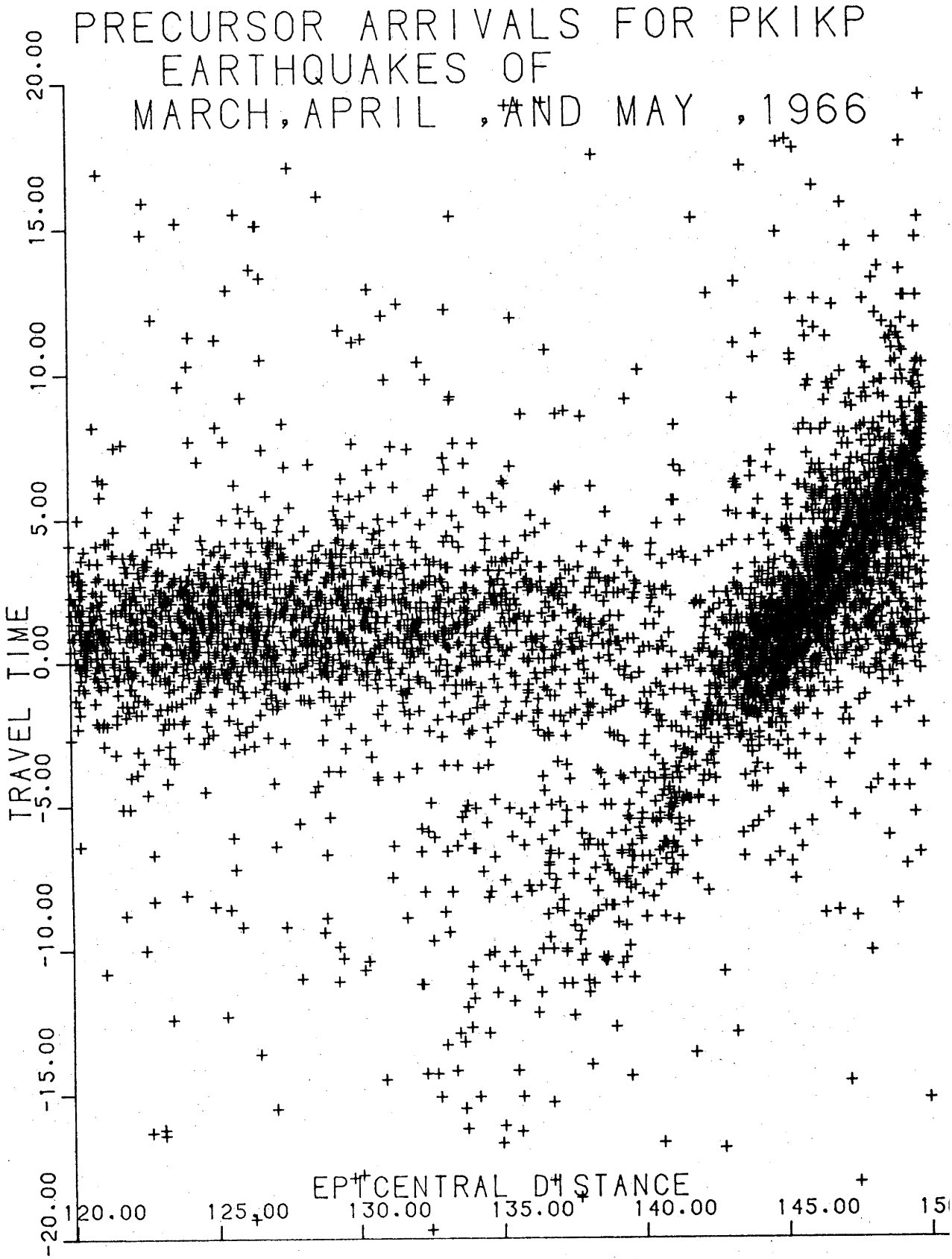

PRECURSOR ARRIVALS FOR PKIKP
EARTHQUAKES OF
MARCH, APRIL , AND MAY , 1966
TRAVEL TIME
20.00
15.00
10.00
5.00
0.00
-5.00
-10.00
-15.00
-20.00
EPICENTRAL DISTANCE
120.00
125.00
130.00
135.00
140.00
145.00

Diagram 8.1 illustrates a plot of such observations for data collected over the months March-May (1966). There exist two main streams. The horizontal and major stream corresponds to PKIKP waves. They are waves that have travelled from the surface through the outer layer of the earth, the core of the earth, and then the inner core. They then travel back to the surface. At each interface they are refracted due to changes in velocity as the waves pass through material of different densities. The diagonal stream ascending upwards from 145° epicentral distance corresponds to PKPP waves, which are waves that have travelled in the same way as the PKIKP waves but on reaching the surface are reflected once more through the outer layer to the observation point. For a background on the theory of such occurrences the interested reader is referred to Jeffreys (1970). Figure 16. of that gook gives plots of established paths of these distance time plots. Assuming a model of a perfectly spherical earth and core, both with uniform density, and assuming perfect observation, the data would follow two continuous curves. But actual irregularities in the earth and in particular the interface between the mantle and core allow distortions to appear. These add to observational error. In the latter context we note that the established method of determining the epicenter of an earthquake is the classical estimation method of least squares which is a nonrobust procedure.

Anderssen (1979) communicated the existence of the extra diagonal stream of points extending below 145° which had not been fully investigated. Cleary and Hadden (1972) put forward the hypothesis that the observations are the result of scattering of PKP waves from irregularities in the vicinity of the core mantle boundary. An individual investigation of the seismograph of any one earthquake can lead to a

correct classification of the scattered wave from the PKIKP wave but the scattered waves are sometimes mistaken as the PKIKP waves and duly recorded as such. It is important to identify the locations of the two streams of data and the proportion of each. The proportion will indicate the amount of false classification.

The data is examined at fixed epicentral distances using the mixture model. This avoids any complicated modelling of randomness of epicentral distance. The assumed model is then of the form

$$F_{\theta_\tau}(x) = \varepsilon_\tau \Phi\left(\frac{x-\mu_\tau+\Delta_\tau}{\sigma_{1\tau}}\right) + (1-\varepsilon_\tau)\Phi\left(\frac{x-\mu_\tau}{\sigma_{2\tau}}\right) \quad ,$$

where $\theta_\tau = (\varepsilon_\tau, \mu_\tau, \Delta_\tau, \sigma_{1\tau}, \sigma_{2\tau})$ is the parameter corresponding to the epicentral distance τ, to be estimated for each given τ. This is done by taking increments of $(\tau-\delta, \tau+\delta)$ and solving the M.M.S.E. equations with $F_n = F_{n\tau}$, the empirical distribution function given by the observed travel time differences in that region of epicentral distance. So $(\mu_\tau - \Delta_\tau, \sigma_{1\tau})$ represents the location and scale of the minor diagonal stream, while $(\mu_\tau, \sigma_{2\tau})$ describes that of the major stream. Then ε_τ represents the proportion of scattered waves identified as PKIKP waves.

The nonlinear equation solver ZSYSRB was employed on the M.S.E. equations with a search being carried out with varying initial estimates in order to obtain extrema of $\omega^2_{n\tau}(\theta)$, the M.S.E. distance. To assist in this regard it was contemplated that the main horizontal stream would be reasonably homogeneous. Then some idea for the choice of initial estimates of $(\mu_\tau, \sigma_{2\tau})$ is attained by considering parameters in the model $\Phi\left(\frac{x-\mu_\tau}{\sigma_\tau}\right)$ in the region of epicentral distance $120. < \tau < 132.5$. While scattered waves are known to be in this region (Cleary and Hadden Fig. 2., or Adams and Randall, 1964) few appeared to be recorded as

PKIKP waves. The M.M.S.E. was used to estimate location and scale when the single normal distribution was modelled. This M-estimator was used in preference to the constructions of §6.4 as no inference was warranted, and a unique solution is obtained.

TABLE 8.2

Location and scale ω^2-estimates of main population

Epicentral Distance	Location	Scale
120.-121.25	1.0280	1.6111
121.25-122.5	.9100	2.0304
122.5-123.75	.8695	2.2100
123.75-125.	1.2460	1.5867
125.-126.25	1.2737	1.9523
126.25-127.5	1.4092	2.0478
127.5-128.75	1.3779	1.6733
128.75-130.	1.3899	2.2653
130.-131.25	1.4914	2.2806
131.25-132.5	1.0383	2.3299

All solutions $\hat{\theta}_{n\tau}$ to the M.S.E. equations for the mixture model were sought out. Having prior knowledge of the data, the model and the results of Table 8.2 it was possible to make perspicacious choices of regions in which initial estimates to the equation solving algorithm were chosen. With as many as 500 initial estimates for any given epicentral distance the algorithm was permitted to run for 150 iterations on each. The algorithm had then either; diverged; was deemed as not making good progress (by the criteria of the algorithm); had converged in the sense that the absolute value of the residual was less than 10^{-6}, or the relative error between two successive estimates was less than 10^{-6}, or both; or the algorithm was still iterating. In the latter case parameters at the last iteration were used as initial estimates and allowed to iterate till one of the former states had been achieved. The

TABLE 8.3

Solutions to minimal mean squared error estimating equations using earthquake data together with comparisons of selection statistics

Epicentral Distance	No. of Observations	Solutions	Distances			
τ	$n \in [\tau-\delta, \tau+\delta]$	$\hat{\theta}_{n\tau} = (\varepsilon, \mu, \Delta, \sigma_1, \sigma_2)$	$\omega^2_{n\tau}(\hat{\theta}_{n\tau})$	$S_{n\tau}(\hat{\theta}_{n\tau})$	$I_{n\tau}(\hat{\theta}_{n\tau})$	η
133.0	91	(.1595, 1.9488, 9.8792, 5.7126, 2.2529) †(.1248, 1.8685, 11.6981, 3.9046, 2.3383) (.3899, 1.6778, 2.4494, 8.2291, 1.4085)	.02103 .02027 .00393 +	.000985 .000989 .000273 o	1.9183 1.9181 1.9076 *	18.53
133.5	104	(.2301, 1.6652, 9.8293, 5.9302, 2.2079) (.4609, 1.4051, 3.6598, 8.2496, 1.2617) †(.1698, 1.5132, 12.0936, 3.7744, 2.3877)	.01999 .00507 + .01901	.001016 .000232 o .001053	.05166* .05302 .05184	24.64
134.0	86	(.4291, 1.0559, 5.3637, 7.2114, 1.4090) (.1, -.9818, 0.5342, 4.2261, .5432)	.00653 + .11453	.000264 o .005276	1.3160 * 1.3334	24.64
134.5	81	(1., 4.3563, 4.5243, 3.9356, 1.5107) (.3790, 1.3686, 5.9610, 6.0502, 1.6753)	.08782 .00244 +	.004275 .000198 o	.7780 .7633 *	21.96
135.0	91	(.2552, 1.9638, 9.2088, 5.3182, 1.7762) †(.2292, 1.9212, 9.9895, 4.5239, 1.8340) (.3970, 1.8836, 5.4437, 7.2195, 1.3730)	.00729 .00725 .00474 +	.000445 .000456 .000311 o	.02335* .02339 .02360	21.96

TABLE 8.3 (Continued)

Epicentral Distance	No. of Observations	Solutions	Distances			
τ	$n \in [\tau-\delta, \tau+\delta]$	$\hat{\theta}_{n\tau} = (\varepsilon, \mu, \Delta, \sigma_1, \sigma_2)$	$\omega^2_{n\tau}(\hat{\theta}_{n\tau})$	$S_{n\tau}(\hat{\theta}_{n\tau})$	$I_{n\tau}(\hat{\theta}_{n\tau})$	η
135.5	78	(.2576, 2.1172, 10.2204, 5.8908, 1.9721) †(.2083, 2.0107, 11.8705, 4.1947, 2.1120) (.4699, 1.8432, 4.2540, 8.8450, 1.1680)	.02675 .02605 .00768 +	.000771 .000812 .000238 o	.004538 .004519 .00417 *	27.12
136.0	80	(.4546, 1.3160, 5.0919, 7.3140, 1.7550) †(.2214, 1.2835, 10.6588, 3.6955, 2.3558) (.3105, 1.4836, 8.4102, 5.6930, 2.1594) (.5, -1.2247, 9., 5.2201, 5.2201)	.00820 + .00995 .01087 .14200	.000352 .000265 o .000293 .009265	.70890 .70628 .70581 .69944 *	21.96
136.5	98	†(.26789, .8422, 8.8044, 2.7478, 2.2806) (.4214, 1.1682, 6.5545, 4.9162, 2.0049) (.5396, 1.0256, 4.7510, 5.7701, 1.6894)	.00782 + .00924 .00858	.000332 o .000438 .000451	4.5928 4.5903 * 4.595	19.86
137.0	93	(.4877, 1.7949, 7.0809, 4.5607, 2.0743) (.7155, 1.6035, 4.4007, 5.8152, 1.1651) †(.3381, 1.3382, 8.6355, 3.0373, 2.4534)	.01304 .00789 + .01215	.000391 .000387 .000386 o	4.5229* 4.5313 4.5261	19.29
137.5	84	(.4342, 1.5853, 7.4650, 4.8958, 1.7378) †(.3044, 1.2742, 9.3642, 3.0307, 2.0902) (.6061, 1.4126, 4.8647, 6.1772, 1.1531)	.01088 .00917 .00725 +	.000303 .000319 .000269 o	3.3886* 3.3911 3.3961	19.86

TABLE 8.3 (Continued)

Epicentral Distance	No. of Observations	Solutions	Distances			
τ	$n \in [\tau-\delta, \tau+\delta]$	$\hat{\theta}_{n\tau} = (\varepsilon, \mu, \Delta, \sigma_1, \sigma_2)$	$\omega^2_{n\tau}(\hat{\theta}_{n\tau})$	$S_{n\tau}(\hat{\theta}_{n\tau})$	$I_{n\tau}(\hat{\theta}_{n\tau})$	η
138.0	77	(.5426, 1.3072, 7.0791, 5.4295, 1.3846) †(.3532, .9694, 9.5729, 2.6840, 1.9024) (.6231, 1.1949, 5.8756, 6.0661, 1.0897)	.01589 .00940 + .01539	.000564 .000341 o .000559	5.5178 * 5.5212 5.5229	21.96
138.5	81	†(.3585, 1.020, 8.9717, 2.4820, 2.2594) (.5000, -1.9023, -.0002, 5.3443, 5.3444) (1., -1.34994, .5523, 5.344, 0.) (1., 4.8446, 6.7469, 5.3444, 1.9253)	.00543 + .05833 .05833 .05833	.000220 o .003181 .003181 .003181	7.0748 * 7.0968 7.0968 7.0968	19.86
139.0	85	†(.4086, .8814, 8.5049, 2.3923, 2.1768)	.00365 +	.002923	12.4286	19.86
139.5	74	†(.5795, 1.2117, 8.1599, 2.4662, 2.0311) (1., 5.9755, 8.6611, 6.1983, 0.) (1., -.0247, 3.5631, 5.1983, 2.9706) (1., 5.8143, 9.3999, 5.1983, 2.7491)	.00563 + .03523 .03523 .03234	.000317 o .001971 .001972 .001972	15.0987 15.0982 * 15.0984 15.0984	21.01
140.0	67	†(.5506, 1.2400, 7.4784, 1.8988, 2.5028)	.00692 +	.000228 o	15.0004 *	17.00
140.5	76	†(.5684, .7794, 6.0973, 2.0913, 2.4143) (1., 1.1419, 3.9469, 4.0353, 2.2600) (1., .9573, 3.7623, 4.0353, 4.3775)	.00428 + .01535 .01535	.000208 o .000990 .000990	22.0871 22.0738 * 22.0738 *	15.66

TABLE 8.3 (Continued)

Epicentral Distance	No. of Observations	Solutions	Distances			
τ	$n \in [\tau-\delta, \tau+\delta]$	$\hat{\theta}_{n\tau} = (\varepsilon, \mu, \Delta, \sigma_1, \sigma_2)$	$\omega^2_{n\tau}(\hat{\theta}_{n\tau})$	$S_{n\tau}(\hat{\theta}_{n\tau})$	$I_{n\tau}(\hat{\theta}_{n\tau})$	η
141.0	91	(.4234, -.9372, 4.2989, 1.2163, 4.5075) (1., -3.1345, .0025, 3.9126, 3.9013) (1., 2.1143, 5.2512, 3.9126, 0.)	.00242 + .04362 .04936	.000162 o .003204 .003204	46.0180 * 45.9227 45.9227	14.13
141.5	78	†(.5507, .1225, 4.6193, 1.6578, 5.1159) (1., -2.5072, .3122, 3.7829, 1.5166) (.5, -2.9194, 0., 3.7829, 3.7829)	.00396 + .05504 .05504	.000325 o .003036 .003036	27.3109 27.2103 * 27.2103	14.13
142.0	74	†(.8758, 1.3206, 3.0377, 2.2176, 9.0597) (.5, -1.5619, .0001, 2.6883, 2.6883) (1., -1.5498, .0123, 2.6883, 2.6821)	.00654 + .01619 .01619	.000640 o .000961 .000960	13.673 13.6429 13.6427 *	12.03
142.5	88	(.5001, -1.0206, .0003, 1.9861, 1.9860) (.9148, -2.2425, -1.2445, 1.729, 9.6173) (.0852, -.9976, 1.2445, 9.6173, 1.729)	.01114 .00508 + .00508	.000786 .000709 o .000709	28.74 * 28.76 28.76	8.02
143.0	127	†(.7778, 1.3949, 1.6042, 1.4537, 5.9371) (1., -.0278, .0047, 2.0742, 2.0716) (.5, -.0325, 0., 2.0742, 2.0745)	.00321 + .01867 .01867	.000187 o .000793 .000793	.2546* .2583 .2583	9.93

extensive searching ensured the detection of all local extrema of $\omega^2_{n\tau}(\theta)$ as this is important whenever the data is rough. In general initial estimates found in large regions of plausible values in the parameter space converged to a minima quickly but epicentral distances $\tau = 134., 134.5$ and $\tau \geq 141.$ were not conducive to this behaviour.

Three and sometimes four local extrema of $\omega^2_{n\tau}(\theta)$ were detected, although one or two often degenerated to a single normal component. When two solutions representing nondegenerate mixtures had to be distinguished one chooses that parameter which separates the two populations clearly. In this manner the solutions indicated by † were selected by the author. The discarded parameter often represented a mixture where one component had accommodated both populations by allowing the location to be centered between the two streams and having a large dispersion parameter. For example at $\tau = 136.5$ the estimate

$$F_\theta(x) = .27\,\Phi\left(\frac{x-7.96}{2.75}\right) + .73\,\Phi\left(\frac{x-.84}{2.28}\right)$$

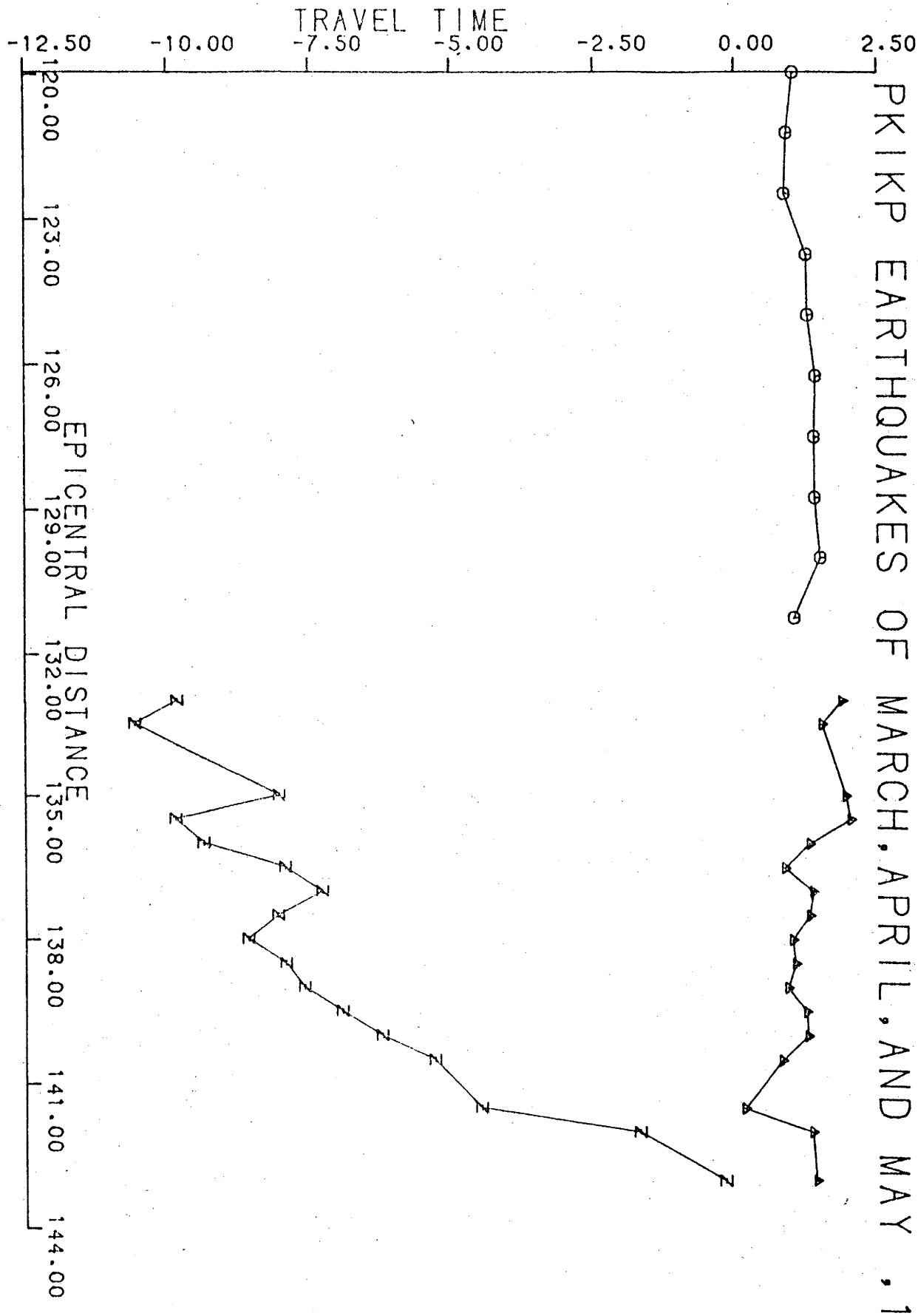
is chosen for it clearly separates the populations. For the other solutions the estimate for the first population envelopes the other as well. Solutions for which the value of μ became negative were neglected. Selected minima are plotted in the plot 8.1.

It is now instructive to examine the performance of distances $\omega^2_{n\tau}$, $S^2_{n\tau}$, and $I_{n\tau}$ as selection statistics. By the plots 8.2 and 8.3 of locations of the estimated populations it is clear that these statistics do not select solutions in an always continuous manner. The Cramér-Von Mises distance which is more robust against tail contamination is only slightly better in this regard. If we exclude that solution with largest first component distribution scale, which corresponds to overlapped component populations, and any degenerate solutions the resulting M.S.E. estimate corresponds to that chosen by the author. This leads

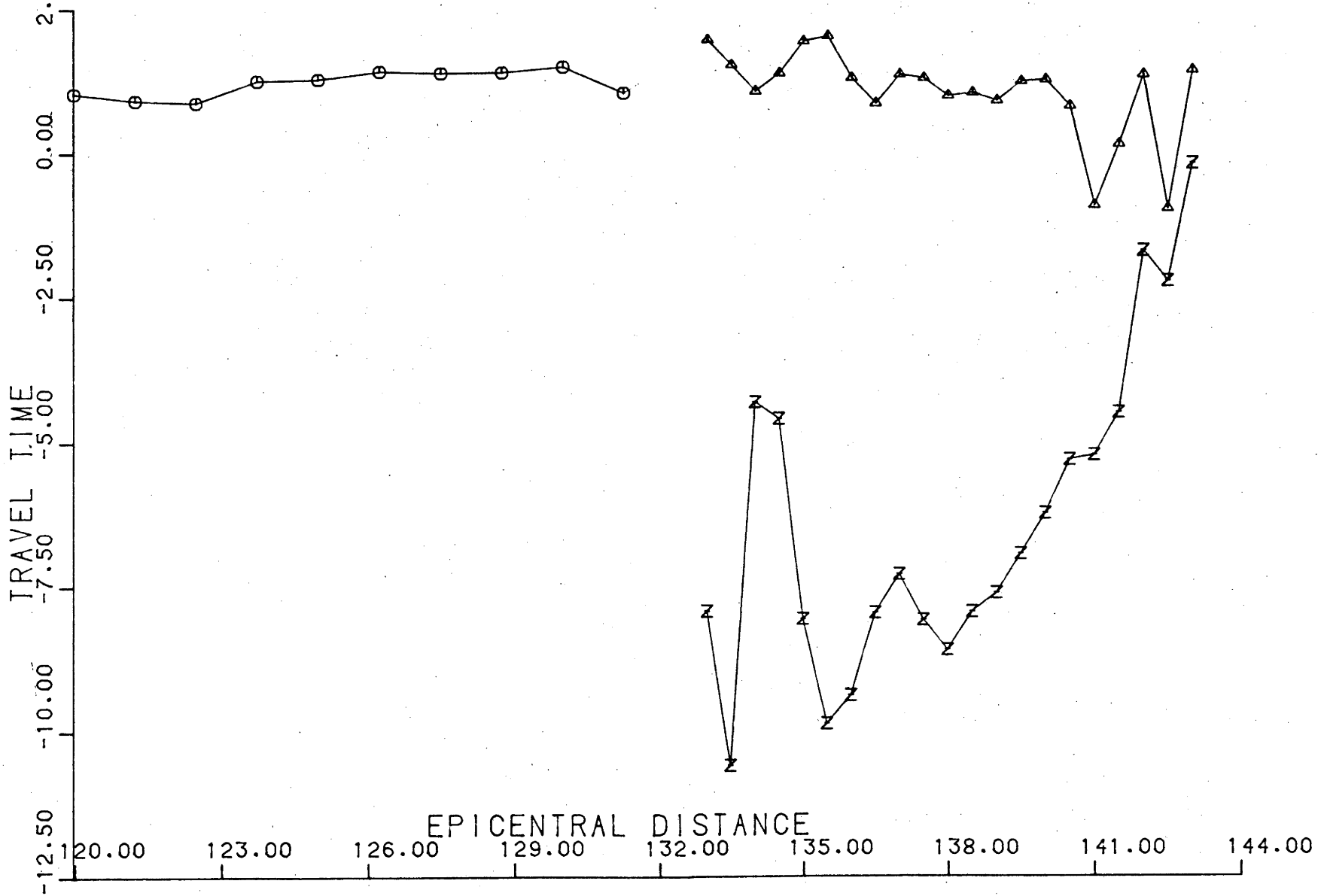
to a roughly continuous plot. Even using this strategy $S^2_{n_\tau}(\theta)$ would still not completely agree with the selection of roots. But it would seem that even though the M.S.E. is not Prokhorov continuous it can still be an effective selection statistic. The principal is though that each of the local minima should be investigated thoroughly.

Clearly the I.S.E. should not be used as a selection statistic because of its erratic behaviour. Co-incidental values of the estimated η in the weight function of the I.S.E. distance reflects the large rounding of errors employed in the computation of the data, accuracy being taken to only one decimal place for travel time. This is a major reason for using a Fréchet differentiable functional in the analysis.

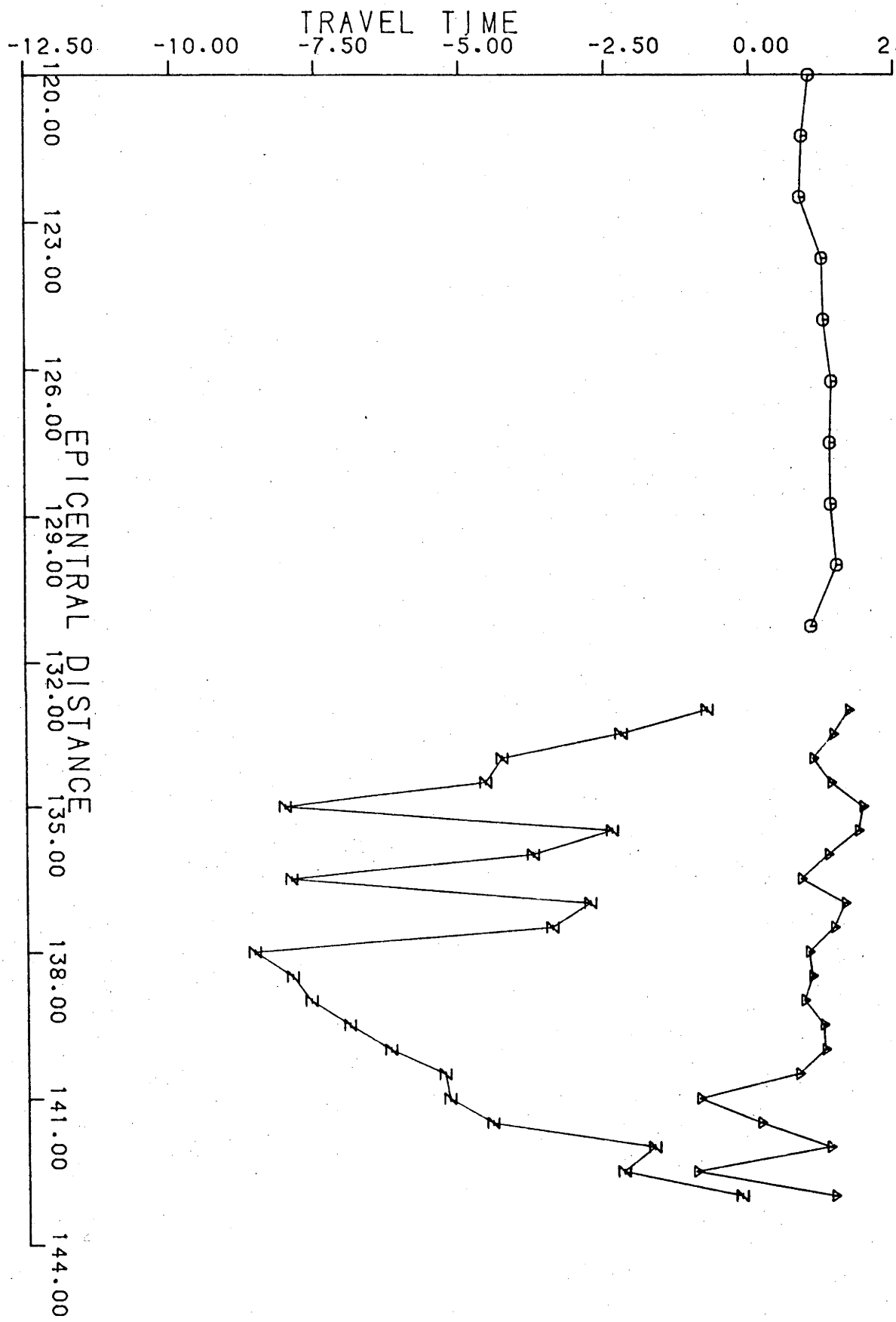
Work is being proceeded with on detailed investigation of the geological significance of the descending stream and its path. This will be expounded upon elsewhere.


PLOT 8.1
AUTHOR SELECTED M.S.E. ESTIMATES FOR
PKIKP EARTHQUAKES OF MARCH, APRIL, AND MAY, 1966
TRAVEL TIME
-12.50
-10.00
-7.50
-5.00
-2.50
0.00
2.50
EPICENTRAL DISTANCE
120.00
123.00
126.00
129.00
132.00
135.00
138.00
141.00
144.00

PLOT 8.2 CRAMER VON MISES SELECTED M.S.E. ESTIMATES FOR PKIKP EARTHQUAKES OF MARCH, APRIL, AND MAY , 1966



PLOT 8.3 M.S.E. ESTIMATES FOR PKIKP EARTHQUAKES OF MARCH, APRIL, AND MAY, 1966



APPENDIX 1

TWO MATHEMATICAL THEOREMS

The two mathematical theorems that we give here are the Brouwer fixed point theorem and the inverse function theorem. We also remark on the Taylor expansions that are used in the thesis.

THEOREM (Brouwer)

If ϕ is a continuous mapping of the closed unit sphere

$$S = \{x \in E^n \mid \|x\| \leq 1\} ,$$

of Euclidean n-space into itself, then there is a point $y \in S$ such that $\phi(y) = y$.

A proof of this theorem can be found in Dunford and Schwarz (1958, P.467). The following theorem can be found in Rudin (1964).

Inverse Function Theorem

Suppose ϕ is a mapping from an open set Θ in E^r into E^r, the partial derivatives of ϕ exist and are continuous on Θ, and the matrix of derivatives $\phi'(\theta^*)$ has inverse $\phi'(\theta^*)^{-1}$ for some $\theta^* \in \Theta$. Write $\lambda = 1/(4\|\phi'(\theta^*)^{-1}\|)$. Use the continuity of the elements of $\phi'(\theta)$ to fix a neighbourhood U_δ of θ^* of sufficiently small radius $\delta > 0$ to ensure that $\|\phi'(\theta) - \phi'(\theta^*)\| < 2\lambda$, whenever $\theta \in U_\delta$. Then

(a) for every $\theta_1, \theta_2 \in U_\delta$

$$\|\phi(\theta_1) - \phi(\theta_2)\| \geq 2\lambda\|\theta_1 - \theta_2\| ;$$

and

(b) the image set $\phi(U_\delta)$ contains the open neighbourhood with radius $\lambda\delta$ about $\phi(\theta^*)$.

Conclusion (a) ensures ϕ is one-to-one on U_δ and that ϕ^{-1} is well defined on the image set $\phi(U_\delta)$.

Taylor Expansions: Two Taylor expansions are used. Given functions $f : E^r \to E'$ that are continuously differentiable, a two term Taylor expansion based on the mean value theorem is possible (cf. Hoffman 1975, P.391)

$$f(\theta) = f(\theta_o) + \nabla f(\xi)(\theta-\theta_o) ,$$

where $\nabla f(\theta)$ is evaluated at a point ξ on the diagonal between θ and θ_o. Then for an $r\times 1$ function $\nabla f(\xi)$ represents an $r\times r$ matrix where the rows are evaluated at possibly different points ξ on the diagonal.

The second type of Taylor expansion concerning a twice continuously differentiable real valued function f is

$$f(\theta) = f(\theta_o) + \nabla f(\theta_o)(\theta-\theta_o) + \tfrac{1}{2}(\theta-\theta_o)'\nabla\nabla' f(\xi)(\theta-\theta_o) ,$$

where ξ is on the diagonal between θ and θ_o (Hoffman 1975, P.393).

APPENDIX 2

A UNIFORM CONVERGENCE RESULT

We remark here that for a general i.i.d. univariate sequence with distribution G_o on E

$$\sup_{-\infty<x<+\infty} \sqrt{n}|F_n(x) - G_o(x)| = 0_p(1) .$$

This is a result from Shorack's (1972) presentation. For let $\underset{\sim}{\xi}$ be a sequence of random variables defined on (Ω,A,P) that are independent Uniform $(0,1)$. U denotes the Brownian bridge on Ω. That is $\{U(t)|0 \leq t \leq 1\}$ is a normal process with all sample paths continuous, $E[U(t)] = 0$ for $0 \leq t \leq 1$, and covariance function of the U process is $\min(s,t) - st$. Let Γ_n be the empirical distribution function of the sample $(\xi_1,\ldots,\xi_n)$.

For $n \geq 1$ we define the "uniform empirical process" U_n by

$$U_n(t) = \sqrt{n}(\Gamma_n(t) - t) \qquad \text{for} \qquad 0 \leq t \leq 1 .$$

For functions f_1, f_2 on $(0,1)$ let $\rho(f_1,f_2) = \sup_{0<t<1}|f_1(t) - f_2(t)|$. Then $\rho(U_n,U) \xrightarrow[e]{} 0$ as $n \to \infty$ $\oplus$

Every sample path of the U_n process converges uniformly as $n \to \infty$ to the corresponding sample path of the U process. Set $X = G_o^{-1}(\xi) = \inf\{y|G_o(y) \geq \xi\}$. See that $\{X \leq x\} = \{\xi \leq G_o(x)\}$. By Proposition A1 of Shorack X has distribution function G_o and $F_n(x) = \Gamma_n(G_o(x))$. So

$$\begin{aligned}\sup_{-\infty<x<+\infty}\sqrt{n}|F_n(x) - G_o(x)| &= \sup_{-\infty<x<+\infty} \sqrt{n}|\Gamma_n(G_o(x)) - G_o(x)| \\ &\leq \sup_{t\in(0,1)} |U_n(t)| \\ &= 0_p(1)\end{aligned}$$

by $\oplus$ and since $\sup_{t\in(0,1)} |U(t)| = 0_p(1)$.

APPENDIX 3

TABLES ON LOCATION AND SCALE ESTIMATES

Extensions to the tables of §6.4 giving results for a wider variety of parameter values are given here. The first table is an extension of Tables 6.8 and 6.9. Added is a column which details the number of times out of the 500 replications that the equation solving algorithm failed to converge from any one of the twenty five initial starting grid points. In small samples there appears to exist a small probability that no solution exists when redescending influence functions are used. This decreases with n. Small sample bias for the scale estimate is also more prevalent. The data is generated from the standard normal distribution.

The second table if anything exhibits the stability of inferences that are made from M-estimators that are formed from the redescending influence functions. Under wide ranging conditions they perform close to the model values. The table shows that heavy contamination in the region where the curve descends (i.e. on [b,c]) increases bias and variance markedly, but to values that are comparable with Proposal 2. Also it is verified that contamination outside the null set results in zero bias and that variance increase is of order $(1-\varepsilon)^{-1}$. Again v_{11}, v_{22} represent the asymptotic variance of location and scale estimators. In Table 1 these are evaluated at the standard normal distribution. The symbols $\hat{mv}_{11}$, $\hat{mv}_{22}$ are the corresponding mean squared errors from the Monte Carlo simulation. Again $Z_{.05} = \Phi^{-1}(.95)$ and the 90% E.C.I. is calculated from the ordered estimates of location.

TABLE 1

Estimator	Sample Size n	ν_{11}/n	ν_{22}/n	$m\hat{\nu}_{11}$	$m\hat{\nu}_{22}$	$\pm z_{.025}\sqrt{\nu_{11}/n}$	95% E.C.I.	$\pm z_{.05}\sqrt{\nu_{22}/n}$	90% E.C.I.	Failures	$(2\Phi(c)-1)^n$
Maximum Likelihood	10	.10+0	.50	.97-1	.56-1	.620	-.590,.652	.520	-.506,.544	-	
	20	.50-1	.25	.53-1	.24-1	.438	-.413,.478	.368	-.349,.398	-	
	50	.20-1	.10	.21-1	.11-1	.277	-.298,.266	.233	-.235,.226	-	
	100	.10	.50-3	.96-2	.52-2	.196	-.196,.195	.165	-.149,.158	-	
Proposal 2 k = 1.96	10	.10+0	.57-1	.86-1	.59-1	.624	-.598,.588	.523	-.454,.490	2	
	20	.51-1	.29-1	.42-1	.31-1	.441	-.400,.379	.370	-.314,.316	0	
	50	.20-1	.1101	.20-1	.13-1	.279	-.289,.267	.234	-.240,.224	0	
	100	.10-1	.57-2	.88-1	.67-2	.197	-.197,.172	.166	-.174,.154	0	
Proposal 2 k = 1.645	10	.10+0	.64-1	.86-1	.65-1	.628	-.579,.558	.527	-.466,.496	1	
	20	.51-1	.32-1	.51-1	.34-1	.44	-.476,.384	.373	-.402,.321	0	
	50	.21-1	.13-1	.20-1	.14-1	.281	-.289,.265	.236	-.241,.229	0	
	100	.10-1	.64-2	.10-1	.76-2	.199	-.199,.195	.167	-.173,.164	0	
R.E. a = 1.645 b = 2. c = 3.3	10	.10+0	.78-1	.15+0	.12+0	.648	-.700,.767	.544	-.535,.684	8	.9904
	20	.55-1	.39-1	.76-1	.58-1	.458	-.424,.608	.440	-.340,.517	1	.9808
	50	.22-1	.16-1	.38-1	.18-1	.290	-.254,.450	.243	-.216,.401	0	.9528
	100	.11-1	.78-2	.19-1	.76-2	.205	-.144,.311	.172	-.122,.260	1	.9078
R.E. a = 1.645 b = 2.4 c = 3.	10	.11+0	.75-1	.12+0	.12+0	.644	-.752,.704	.541	-.541,.577	6	.9733
	20	.54-1	.38-1	.63-1	.63-1	.456	-.517,.426	.382	-.438,.364	3	.9473
	50	.22-1	.15-1	.23-1	.18-1	.288	-.283,.309	.242	-.241,.269	0	.8736
	100	.11-1	.75-2	.11-1	.92-2	.204	-.191,.209	.171	-.172,.185	0	.7631
R.E. a = 1.96 b = 2.5 c = 3.	10	.11+0	.67-1	.13+0	.13+0	.638	-.738,.727	.536	-.583,.582	7	.9733
	20	.53-1	.34-1	.56-1	.49-1	.451	-.504,.400	.379	-.414,.348	4	.9473
	50	.21-1	.19-1	.23-1	.16-1	.286	-.288,.299	.240	-.243,.268	0	.8736
	100	.11-1	.67-2	.11-1	.78-2	.202	-.199,.198	.69	-.171,.182	0	.7631
R.E. a = 2. b = 2.91 c = 3.1	10	.10+0	.62-1	.12+0	.98-1	.631	-.699,.639	.530	-.546,.528	7	.9808
	20	.52-1	.31-1	.55-1	.40-1	.446	-.498,.400	.374	-.418,.347	0	.9620
	50	.21-1	.10-1	.21-1	.14-1	.282	-.289,.277	.237	-.255,.232	1	.9077
	100	.10-1	.62-2	.10-1	.72-2	.200	-.198,.190	.167	-.177,.157	0	.8239

TABLE 2

Asymptotic bias and variance incurred by estimators of location and scale when $G = (1-\varepsilon)\Phi(x) + \varepsilon H(x)$

$H(x) = \Phi(\frac{x+\Delta}{\alpha})$

ε = .05		Maximum Likelihood				Proposal 2 : c = 1.96			
Δ	α	$T[\psi,G]$	v_{11}	v_{22}	v_{12}	$T[\psi,G]$	v_{11}	v_{22}	v_{12}
0.	1.	0.	1.	1.	.5 0.	0.	1.	1.012	.571 0.
0.	.1	0.	.975	.951	.512 0.	0.	.968	.958	.593 0.
0.	3.	0.	1.183	1.4	2.33 0.	0.	1.061	1.163	.75 0.
1.5	1.	-.075	1.052	1.107	.592 .272	-.069	1.045	1.111	.65 -.062
1.5	3.	-.075	1.228	1.507	3.06 .93	-.04	1.067	1.185	.781 -.104
2.5	.1	-.125	1.117	1.247	.599 .462	-.119	1.126	1.425	1.165 -.469
2.5	1.	-.125	1.139	1.297	.904 .608	-.101	1.092	1.255	.841 -.226
3.	3.	-.15	1.352	1.828	5.064 2.01	-.073	1.083	1.242	.867 -.213
3.5	.1	-.175	1.238	1.532	1.113 .948	-.12	1.127	1.439	1.206 -.495

Redescending estimators

		a = 1.96 b = 2.5 c = 3.				a = 1.645 b = 2. c = 3.3			
0.	1.	0.	1.	1.061	.674 0.	0.	1.	1.094	.784 0.
0.	.1	0.	.965	1.012	.726 0.	0.	.957	1.040	.858 0.
0.	3.	0.	1.025	1.164	.801 0.	0.	1.023	1.186	.899 0.
1.5	1.	-.063	1.040	1.170	.781 -.088	-.059	1.038	1.200	.891 -.084
1.5	3.	-.084	1.028	1.181	.814 -.092	-.077	1.025	1.199	.911 -.084
2.5	.1	-.115	1.14	1.440	1.272 -.479	-.090	1.095	1.773	1.799 -.831
2.5	1.	-.066	1.06	1.340	1.040 -.299	-.057	1.050	1.320	1.084 -.238
3.	3.	-.127	1.029	1.023	.828 -.144	-.115	1.025	1.216	.923 -.129
3.5	.1	0.	1.	1.116	.710 0.	0.	1.	1.156	.835 -.007

TABLE 2 (Continued)

ε = .1		Maximum Likelihood				Proposal 2 : c = 1.645				
Δ	α	$T[\psi,G]$	v_{11}	v_{22}	v_{12}	$T[\psi,G]$	v_{11}	v_{22}	v_{12}
0.	1.	0.	1.	1.	.5	0.	0.	1.	1.026	.640	0.
0.	.1	0.	.949	.901	.524	0.	0.	.9229	.906	.702	0.
0.	3.	0.	1.342	1.8	3.300	0.	0.	1.110	1.309	.961	0.
1.5	1.	-.15	1.097	1.203	.645	.483	-.137	1.084	1.221	.792	-.114
1.5	3.	-.15	1.415	2.003	4.150	1.519	-.076	1.122	1.357	1.028	-.199
2.5	.1	-.25	1.210	1.464	.597	.758	-.238	1.249	2.063	2.213	-1.228
2.5	1.	-.25	1.250	1.463	1.040	1.001	-.199	1.176	1.538	1.190	-.467
3.	3.	-.3	1.616	2.610	6.261	3.124	-.141	1.153	1.494	1.228	-.432
3.5	.1	-.35	1.415	2.004	1.156	1.455	-.238	1.250	2.090	2.295	-1.280

Redescending estimators

		a = 1.645 b = 2.4 c = 3.					a = 2. b = 2.9 c = 3.1				
0.	1.	0.	1.	1.081	.752	0.	0.	1.	1.036	.619	0.
0.	.1	0.	.914	.976	.920	0.	0.	.931	.936	.705	0.
0.	3.	0.	1.049	1.291	1.024	0.	0.	1.064	1.287	.958	0.
1.5	1.	-.127	1.079	1.306	.967	-.172	-.137	1.085	1.256	.802	-.159
1.5	3.	-.157	1.060	1.360	1.090	-.231	-.178	1.079	1.404	1.063	-.313
2.5	.1	-.235	1.267	2.052	2.313	-1.229	-.062	1.272	1.502	.796	-.159
2.5	1.	-.145	1.128	1.850	1.718	-.819	-.189	1.172	1.875	1.619	-.864
3.	3.	-.245	1.073	1.475	1.198	-.41	-.289	1.101	1.645	1.282	-.623
3.5	.1	0.	1.	1.201	.841	0.	0.	1.	1.151	.688	0.

APPENDIX 4

SMALL SAMPLE BIAS OF THE CRAMER VON MISES ESTIMATOR

The small sample bias of the estimator derived from the Cramer Von Mises distance, $S_n(\theta)$ of (7.2), is investigated through the small sample bias of the terms in the Taylor expansion on the minimizing equations. The expansion that is truncated to two terms is usually given

$$0 = \nabla' S_n(\hat{\theta}_n) = \nabla' S_n(\theta_o) + \nabla\nabla' S_n(\theta_o)(\hat{\theta}_n - \theta_o) + O_p(n^{-1}) .$$

Here

$$\nabla' S_n(\theta) = 2 \int_{-\infty}^{+\infty} \nabla' F_\theta(x)(F_\theta(x) - F_n(x))dF_n(x) ,$$

where the interpretation of $I_i(x)dI_i(x) = \frac{1}{2}dI_i(x)$ is made on the Heaviside functions, and $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_i(x)$. Then abbreviating $F_\theta(x)$ to F see that

$$\tfrac{1}{2}E_{\theta_o}[\nabla' S_n(\theta_o)] = \int \nabla F.F.dF - E_{\theta_o}\left[\int \nabla' F.F_n dF_n\right] .$$

But

$$E_{\theta_o}\left[\int \nabla' F.F_n dF_n\right] = \sum_{i\neq j} \frac{1}{n^2} E_{\theta_o} \int \nabla' F.I_i dI_j + \sum_{i=1}^{n} \frac{1}{n^2} E_{\theta_o} \int \nabla F\, I_i dI_i$$

$$= \frac{n(n-1)}{n^2} \int \nabla' F.FdF + \frac{1}{2n} \int \nabla' FdF .$$

So

$$E_{\theta_o}[\nabla' S_n(\theta_o)] = \frac{2}{n} \int_{-\infty}^{+\infty} \nabla' F_{\theta_o}(x)(F_{\theta_o}(x) - \tfrac{1}{2})dF_{\theta_o}(x) ,$$

and letting

$$\eta_\theta(x) = \int_{-\infty}^{x} (\nabla' F_\theta(y))f_\theta(y)dy ,$$

integration by parts gives

$$E_\theta[\nabla' S_n^2(\theta)] = n^{-1}\{\eta_\theta(\infty) - 2\int \eta_\theta(x) f_\theta(x)dx\} .$$

The second term of the expansion is also biased towards its expectation

$$E_{\theta_o}[\nabla\nabla' S_n(\theta_o)] = 2\int_{-\infty}^{+\infty} \{\nabla' F_{\theta_o}(x)\}\{\nabla F_{\theta_o}(x)\}dF_{\theta_o}(x)$$
$$+ \frac{2}{n}\int_{-\infty}^{+\infty} \{\nabla\nabla' F_{\theta_o}(x)\}(F_{\theta_o}(x) - \tfrac{1}{2})dF_{\theta_o}(x) .$$

Calculations of variance terms also reveal bias. The variance of $\sqrt{n}\ \nabla' S_n(\theta_o)$ is biased from its asymptotic value. To calculate the bias terms set $X_a = (\partial/\partial\theta_a)S_n(\theta_o)$ and $X_b = (\partial/\partial\theta_b)S_n(\theta_o)$ and use the formula $\mathrm{cov}(X_a, X_b) = E[X_a\ X_b] - E[X_a]E[X_b]$. Then from the expression

$$E_\theta[X_a\ X_b] = \frac{1}{n^4}\sum_{i,j,k,\ell} E_\theta\left[\iiint F_a(x)F_b(y)(F(x) - I_i(x))(F(y) - I_j(y))dI_k(x)dI_\ell(y)\right],$$

particular consideration of combinations of $i,j,k,$ and ℓ allows one to evaluate this quantity. It is a lengthy exercise and the final result is given in §8.1. The bias of the actual estimator is not evaluated. But bias in these terms of the Taylor expansion will contribute to the overall bias term of order $1/n$ that is given by the expansion method of Cox and Hinkley (1974, P.310).

REFERENCES

Abramowitz, M. and Stegun, I.A. (1970). *Handbook of Mathematical Functions*, Dover Publications Inc., New York.

Adams, R.D. and Randall, M.J. (1964). The fine structure of the earth's core, *Bull. Seism. Soc. Amer.*, 54, 1299-1313.

Aitchison, J. and Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints, *Ann. Math. Statist.*, 29, 813-828.

Alexandroff, A.D. (1943). Additive set functions in abstract spaces, *Mat. Sbornik. New Series*, 13, 169-234.

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

Anderssen, R.S. (1979). Personal communication, Australian National University.

Andrews, D.F. (1974). A robust method for multiple linear regression, *Technometrics*, 16, 523-531.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972). *Robust Estimates of Location*, Princeton University Press, Princeton, New Jersey.

Bartlett, M.S. and MacDonald, P.D.M. (1968). "Least squares" estimation of distribution mixtures, *Nature, Lond.*, 217, 195-196.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models, *Ann. of Statist.*, 5, 445-463.

Bickel, P.J. and Lehman, E.L. (1976). Descriptive statistics for non-parametric models III. Dispersion, *Ann. of Statist.*, 4, 1139-1158.

Billingsley, P. (1956). The invariance principle for dependent random variables, *Trans. Amer. Math. Soc.*, 83, 250-268.

Billingsley, P. and Topsoe, F. (1967). Uniformity in weak convergence, *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 7, 1-16.

Blackman, J. (1955). On the approximation of a distribution function by an empiric distribution, *Ann. Math. Statist.*, 26, 256-267.

Blackwell, D. (1956). On a class of probability spaces, *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, ed. Neyman, J., 2, 1-6.

Boes, D.C. (1966). On the estimation of mixing distributions, *Ann. Math. Statist.*, 37, 177-188.

Bolthausen, F. (1977). Convergence in distribution of minimum distance estimators, *Metrika*, 24, 215-227.

Boos, D.D. (1977). The differential approach in statistical theory and robust inference, Ph.D. dissertation, Florida State University, Microfilms International, Ann Arbor.

Boos, D.D. and Serfling, R.J. (1980). A note on differentials and the C.L.T. and L.I.L. for statistical functions, with application to M-estimates, *Ann. of Statist.*, 6, to appear.

Bowmen, K.O. and Shenton, L.R. (1973). Space of solutions for a normal mixture, *Biometrika*, 60, 629-636.

Breiman, L. (1968). *Probability*, Addison-Wesley Publishing Co., London.

Brent, R.P. (1973). Some efficient algorithms for solving systems of nonlinear equations, *Siam J. Numer. Anal.*, 10, 327-344.

Brown, L.D. and Purves, R. (1973). Measurable selections of extrema, *Ann. of Statist.*, 5, 1, 902-912.

Carroll, R.J. (1978a). On almost sure expansions for M-estimates, *Ann. of Statist.*, 6, 314-318.

Carroll, R.J. (1978b). On the asymptotic distribution of multivariate M-estimates, *Jour. Mult. Anal.*, 8, 361-371.

Carroll, R.J. (1979). On estimating variances of robust estimators when the errors are asymmetric, *Jour. Amer. Statist. Assoc.*, 74, 674-679.

Chanda, K.C. (1954). A note on the consistency and maxima of the roots of likelihood equations, *Biometrika*, 41, 56-61.

Chandra, S. (1969). On mixtures of probability distributions, M.S. thesis, Department of Statistics, University of Chicago.

Choi, K. and Bulgren, W.G. (1968). An estimation procedure for mixtures of distributions, *Jour. Roy. Statist. Soc., Ser. B*, 30, 444-460.

Chung, K.L. (1968). *A Course in Probability Theory*, Academic Press, New York.

Clarke, B.R. and Heathcote, C.R. (1978). Comment on "Estimating mixtures of normal distributions and switching regressions", by Quandt and Ramsey, *Jour. Amer. Statist. Assoc.*, 73, 749-750.

Cleary, J.R. and Hadden, R.A.W. (1972). Seismic wave scattering near the core-mantle boundary: a new interpretation to precursors to PKP, *Nature*, 240, 549-551.

Cohen, A.C. (1967). Estimation in mixtures of two normal distributions, *Technometrics*, 9, 15-28.

Collins, J.R. (1976). Robust estimation of a location parameter in the presence of asymmetry, *Ann. of Statist.*, 4, 68-86.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

Cramér, H. (1946). *Mathematical Methods in Statistics*, Princeton University Press, Princeton, New Jersey.

Cramér, H. and Wold, H. (1936). Some theorems on distribution functions *Jour. Lond. Math. Soc.*, 11, 290-295.

Crowder, M.J. (1977). Maximum likelihood estimation for dependent observations with applications to nonhomogeneous Markov chains, Ph.D. dissertation, University of Surrey.

Day, N.E. (1969). Estimating the components of a mixture of normal distributions, *Biometrika*, 56, 463-474.

Doob, J.L. (1953). *Stochastic Processes*, Wiley, New York.

Dunford, N. and Schwarz, J.T. (1958). *Linear Operators, Vol. I*, Interscience, New York.

Durbin, J., Knott, M. and Taylor, C.C. (1975). Components of Cremér-von Mises statistics II, *Jour. Roy. Statist. Soc. Ser. B*, 37, 216-237.

Elker, J., Pollard, D. and Stute, W. (1979). Glivenko Cantelli theorems for classes of convex sets, *Adv. Appl. Probability*, 11, 820-833.

Fabian, V. (1970). On uniform convergence of measures, *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 15, 139-143.

Fillipova, A.A. (1962). Mises' theorem of the asymptotic behaviour of functionals of empirical distribution functions and its statistical applications, *Theor. Probability Appl.*, 7, 24-57.

Finkelstein, H.F. (1971). The law of the iterated logarithm for empirical distributions, *Ann. Math. Statist.*, 42, 607-615.

Foutz, R.V. (1977). On the unique consistent solution to the likelihood equations, *Jour. Amer. Statist. Assoc.*, 72, 147-148.

Foutz, R.V. and Srivastava, R.C. (1979). Statistical inference for Markov processes when the model is incorrect, *Adv. Appl. Probability*, 11, 737-749.

Fowlkes, E.B. (1979). Some methods for studying the mixture of two normal (log normal) distributions, *Jour. Amer. Statist. Assoc.*, 74, 561-575.

Graves, L.M. (1946). *Theory of Functions of Real Variables*, McGraw-Hill, New York.

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*, Academic Press, New York.

Hall, P. (1981). A comedy of errors: the canonical form for a stable characteristic function, *Bull. Lond. Math. Soc.*, to appear.

Hampel, F.R. (1968). Contributions to the theory of robust estimation, Ph.D. Thesis, University of California, Berkeley.

Hampel, F.R. (1971). A general qualitative definition of robustness, *Ann. Math. Statist.*, 42, 1887-1896.

Hampel, F.R. (1973). Some small sample asymptotics, *Proc. Prague Symposium on Asymptotic Statistics.*

Hampel, F.R. (1974). The influence curve and its role in robust estimation, *Jour. Amer. Statist. Assoc.*, 69, 383-393.

Hampel, F.R. (1978). Modern trends in the theory of robustness, *Math. Operationforsch. Statist. Ser. Statistics*, 9, 425-442.

Hannan, E.J. (1970). *Multiple Time Series*, Wiley, New York.

Heathcote, C.R. (1977). The integrated squared error estimation of parameters, *Biometrika*, 64, 255-264.

Heathcote, C.R. and Silvapulle, P.M.J. (1980). Minimum mean squared estimation of location and scale parameters under misspecification of the model, *Biometrika*, to appear.

Hill, B.M. (1963). Information for estimating the proportions in mixtures of exponential and normal distributions, *Jour. Amer. Statist. Assoc.*, 58, 918-932.

Hodges, J.L. and Lehman, E.L. (1963). Estimates of location based upon rank tests, *Ann. Math. Statist.*, 34, 598-611.

Hoffman, K. (1975). *Analysis in Euclidean Space*, Prentice-Hall Inc., Englewood Cliffs, N.J.

Hogg, R.V. (1977). An introduction to robust procedures, *Communications in Statistcs, A6*, 789-794.

Hogg, R.V. (1979). Statistical robustness: one view of its use in applications today, *The Amer. Statistician*, 33, 108-115.

Huber, P.J. (1964). Robust estimation of a location parameter, *Ann. Math. Statist.*, 35, 1, 73-101.

Huber, P.J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions, *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Ed. Le Cam, L. and Neyman, J., 1, 221-233.

Huber, P.J. (1972). The 1972 Wald lecture, robust statistics: a review, *Ann. Math. Statist.*, 43, 2, 1041-1067.

Huber, P.J. (1977). *Robust Statistical Procedures*, Soc. for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

James, B.R. (1975). A functional law of the iterated logarithm for weighted empirical distributions, *Ann. Prob.*, 3, 762-772.

James, I.R. (1978). Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements, *Biometrics*, 34, 265-275.

Jeffreys, H. (1967). *Theory of Probability*, Oxford, Clarendon Press.

Jeffreys, H. (1970). *The Earth: Its Origin, History and Physical Constitution,* Cambridge University Press.

Kallianpur, G. (1963). Von Mises functionals and maximum likelihood estimation, *Sankhya, Ser. A,* 23, 149-158.

Kallianpur, G. and Rao, C.R. (1955). On Fisher's lower bound to asymptotic variance of a consistent estimate, *Sankhya,* 15, 331-342.

Kaufman, R. and Phillip, W. (1978). A uniform law of the iterated logarithm for classes of functions, *Ann. Prob.,* 6, 930-952.

Kelley, J.L. (1967). *General Topology,* D. Van Nostrand Co., New York.

Kiefer, N. (1978). Discrete parameter variation: efficient estimation of a switching regression model, *Econometrica,* 46, 427-434.

Knüsel, L.F. (1969). Ueber minimum-distance-schätzungen, Ph.D. thesis, Swiss Federal Institute of Technology, Zurich.

Kumar, K.D., Nicklin, E.H., and Paulson, A.S. (1979). Comment on "Estimating mixtures of normal distributions and switching regressions" by Quandt and Ramsey, *Jour. Amer. Statist. Assoc.,* 74, 52-55.

Landers, D. (1972). Existence and consistency of modified minimum contrast estimates, *Ann. Math. Statist.,* 43, 74-83.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Baye's estimates, *Univ. of Calif. Publ. Statist.,* 1, 227-330.

Loève, M. (1950). On sets of probability laws and their limit elements, *Univ. Calif. Publ. Statist.,* 1, 53-88.

Loève, M. (1955). *Probability Theory,* D. Van Nostrand Co., New York.

Macdonald, P.D.M. (1971). Comment on "An estimation procedure for mixtures of distributions" by Choi and Bulgren, *Jour. Roy. Statist. Soc., Ser B,* 33, 326-329.

Macdonald, P.D.M. (1975). Estimation of finite distribution mixtures in *Applied Statistics*, ed. R.P. Gupta, 231-245, North Holland Publ. Co.

Odell, P.L. and Basu, J.P. (1976). Concerning several methods for estimating crop acreages using remote sensing data, *Communications in Statistics, Theory and Methods,* A5, 1091-1114.

Ostrowski, A.M. (1960). *Solutions of Equations and Systems of Equations,* Academic Press, New York.

Padgett, W.J. and Taylor, R.L. (1973). *Laws of Large Numbers for Normed Linear Spaces and Certain Fréchet Spaces,* Lecture Notes in Mathematics, 360, Springer-Verlag, New York.

Parzen, E. (1954). On uniform convergence of families of sequences of random variables, *Univ. Calif. Publ. Statist.,* 2, 23-54.

Paulson, A.S., Holcomb, E.W. and Leitch, R.A. (1975). The estimation of parameters of the stable laws, *Biometrika*, 62, 163-70.

Pearson, K. (1894). Contribution to the mathematical theory of evolution: I. On the dissection of asymmetrical frequency curves, *Phil. Trans. R. Soc. Lond., A*, 185, 1-40.

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates, *Metrika*, 14, 249-272.

Pollard, D. (1980). The minimum distance method of testing, *Metrika*, to appear.

Prokhorov, Yu. V. (1956). Convergence of random processes and limit theorems in probability theory, *Theor. Probability Appl.*, 1, 157-214.

Quandt, R.E. and Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regressions, *Jour. Amer. Statist. Assoc.*, 73, 730-752.

Quenoville, M.H. (1956). Notes on bias in estimation, *Biometrika*, 43, 353-60.

Rao, C.R. (1957). Maximum likelihood estimation for the multinomial distribution, *Sankhya*, 18, 139-148.

Rao, C.R. (1963). Criteria of estimation in large samples, *Sankhya, Ser. A*, 25, 189-206.

Rao, R.R. (1962). Relations between weak and uniform convergence of measures with applications, *Ann. Math. Statist.*, 33, 659-680.

Reeds, J.A. (1976). On the definition of Von Mises functionals, Ph.D. Thesis, Dept. of Statistics, Harvard Univ., Cambridge MA.

Reeds, J.A. (1978). Jackknifing maximum likelihood estimates, *Ann. Statist.*, 6, 727-739.

Reiss, R.D. (1978). Consistency of minimum contrast estimators in non-standard cases, *Metrika*, 25, 129-142.

Rey, W.J.J. (1977). M-estimates in robust regression, a case study, in Recent Developments in Statistics, ed. Barra, J.R., Brodeau, F., Romier, G., and Van Custem, B.

Roussas, G.G. (1969). Asymptotic normality of the maximum likelihood estimate in Markov processes, *Metrika*, 14, 62-70.

Rudin, W. (1964). *Principles of Mathematical Analysis*, McGraw-Hill Inc., New York.

Sahler, W. (1970). Estimation by minimum discrepancy methods, *Metrika*, 16, 85-106.

Shorack, G.R. (1972). Functions of order statistics, *Ann. Math. Statist.*, 43, 412-427.

Silvapulle, P.M.J. (1980). The minimum ω^2-method of estimation, Ph.D. thesis, Australian National University, Canberra.

Staudte, R.G. (1978). *Lecture Notes on Robust Estimation*, Dept. of Math. Statist., La Trobe University.

Teicher, H. (1960). On the mixture of distributions, *Ann. Math. Statist.*, 31, 360-369.

Teicher, H. (1961). Identifiability of mixtures, *Ann. Math. Statist.*, 32, 244-248.

Teicher, H. (1963). Identifiability of finite mixtures, *Ann. Math. Statist.*, 34, 1265-1269.

Thall, P.F. (1979). Huber sense robust M-estimation of a scale parameter with application to the exponential distribution, *Jour. Amer. Statist. Assoc.*, 74, 147-152.

Thornton, J.C. and Paulson, A.S. (1977). Asymptotic distribution of characteristic function-based estimators for the stable laws, *Sankhya*, 39, 341-354.

Topsoe, F. (1970). On the Glivenko Cantelli theorem, *Z. Wahrscheinlichkeitstheorie und verw. Geb.*, 14, 239-250.

Tucker, H.G. (1959). A generalization of the Glivenko-Cantelli theorem, *Ann. Math. Statist.*, 30, 828-830.

Tukey, J.W. (1960). A survey of sampling from contaminated distributions, *Contributions to Probability and Statistics*, I. Olkin, ed., Stanford University Press, Stanford, CA.

Varadarajan, V.S. (1958). On the convergence of probability distributions, *Sankhya*, 19, 23-26.

Von Mises, R. (1947). On the asymptotic distributions of differentiable statistical functions, *Ann. Math. Statist.*, 18, 309-348.

Wald, A. (1941). Asymptotically most powerful tests of statistical hypothesis, *Ann. Math. Statist.*, 12, 1-19.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Trans. Amer. Math. Soc.*, 54, 426-482.

Wald, A. (1949). A note on the consistency of maximum likelihood estimation, *Ann. Math. Statist.*, 20, 595-601.

Wolfowitz, J. (1952). Consistent estimation of the parameter of a linear structural relation, *Skand. Aktuarie Tidskr.*, 35, 132-157.

Wolfowitz, J. (1954). Estimation by the minimum distance method in nonparametric difference equations, *Ann. Math. Statist.*, 25, 203-217.

Wolfowitz, J. (1957). The minimum distance method, *Ann. Math. Statist.*, 28, 75-88.

Yakowitz, S. (1969). A consistent estimator for the identification of finite mixtures, *Ann. Math. Statist.*, 40, 1728-1735.