NONSMOOTH ANALYSIS AND FRÉCHET DIFFERENTIABILITY

OF

M-FUNCTIONALS

by

Brenton R. Clarke
Murdoch University

Key Words: M-estimators, Robustness, Gross error sensitivity, Asymptotic expansions, Asymptotic normality, Weak continuity, Selection functional, Local uniqueness, Empirical distribution function.

Mathematics subject classification (Amer. Math. Soc.)
primary 62E20
secondary 62G35

Abbreviated Title: Relaxing Conditions for Fréchet Differentiability.

1. INTRODUCTION

In a recent paper the author showed that contrary to popular opinion, strict Fréchet differentiability of the class of M-functionals is frequently possible. A necessary requirement for existence of the Fréchet derivative is that the defining psi function is uniformly bounded, and this naturally excludes those nonrobust estimators such as the maximum likelihood estimator in normal parametric models. On the other hand, in that paper, smoothness assumptions were imposed on the defining psi function which are not appropriate for many common robust proposals in M-estimation theory, such as Huber's(1964) minimax solution and Hampel's(1974) three part redescender used in estimating location. A host of robust solutions for more general parametric families are obtained through Hampel's(1968) lemma 5, and generalizations of it (cf. Hampel 1978), and these almost invariably are functions with "sharp corners". Indeed the problem that is presented by failure of psi functions to have continuous partial derivatives has been the focus of papers by Huber(1967), Carroll(1978) with respect to proofs of asymptotic normality. While Fréchet differentiability of the M-functional apriori gives asymptotic normality of the M-estimator, at least for real valued observation spaces, it also gives a direct expansion by which the degree of robustness can be directly measured through the gross error sensitivity. The latter quantity is the supremum of the absolute value of the influence curve of Hampel(1968,1974), and Huber(1977) assuming existence of the Fréchet derivative shows that the maximum asymptotic bias in contaminated neighbourhoods of a parametric distribution is proportional to the gross error sensitivity. Subsequently Fréchet differentiability of a statistical functional is an important tool in the robust description of an estimator, and complements the definition of a robust functional as one that is weakly continuous (cf. Hampel 1971).

In this paper the methods of nonsmooth analysis , described in the book by F.H. Clarke(1983), are introduced to the theory of statistical expansions, and are used here in the proofs of weak continuity and Fréchet differentiability of M-functionals. Subsequently the conditions for Fréchet differentiability given in Clarke(1983) can be relaxed to include most popular M-functionals.

The M-estimator is a solution of equations

$$\int_R \psi(x,\tau)\, dF_n(x) = 0 \,, \tag{1.1}$$

where F_n is that distribution which attributes atomic mass 1/n to each of n independent identically distributed observations $X_1,\ldots,X_n$, having common distribution $F \in \mathcal{G}$, the space of probability distributions defined on some separable metrizeable observation space R. For the applications in this paper it is only necessary to consider R = E, the real line. The parameter $\tau \in \Theta$, an open subset of Euclidean r-space E^r, and $\mathcal{F} = \{F_\tau : \tau \in \Theta\}$ is a parametric family where the usual assumption is that $F = F_\theta$ for some $\theta \in \Theta$. The function ψ: $R \times \Theta \to E^r$ can be defined through minimization of some loss function, or obtained by some other optimal criteria.. The theory of robustness makes use of the M-functional T defined on $\mathcal{G}$, so that more generally T[G] is a solution of equations

$$K_G(\tau) = \int_R \psi(x,\tau)\, dG(x) = 0 \tag{1.2}$$

if a solution exists, T[G] = ∞ otherwise. Thus the estimator is given by the functional T evaluated at F_n , and its asymptotic properties follow from continuity and differentiability of T at F with respect to suitable metrics defined on $\mathcal{G}$. This approach to asymptotic theory for statistics was first considered by Von Mises(1947).

To avoid ambiguity, and also for good statistical practice, the concept of a selection functional ρ was introduced by Clarke(1983), in order to identify in the event of several solutions of the equations (1.2), that root which is to be the estimator. That is, the M-functional is defined by ψ,ρ so that

$$\inf_{I(\psi,G)} \rho(G,\tau) = \rho(G, T[\psi,\rho,G]),$$

where

$$I(\psi,G) = \{\tau \mid \int_R \psi(x,\tau)\, dG(x) = 0, \quad \tau\in\Theta\}$$

,if a solution exists. Otherwise $T[\psi,\rho,G] = \infty$. The functional T is then Fréchet differentiable at F with respect to the pair $(\mathcal{G},d_*)$, for suitable metrics d_* on $\mathcal{G}$, if T can be approximated by a linear functional T_F' which is defined on the linear space spanned by the differences G - H of members of $\mathcal{G}$, so that

$$\left| T[G] - T[F] - T_F'(G - F) \right| = o(d_*(G,F)) \tag{1.3}$$

as $d_*(G,F) \to 0$, $G\in\mathcal{G}$. Essentially the expansion for Fréchet differentiability is dependent on a local expansion of equations(1.2), and a robust selection functional will automatically select the Fréchet differentiable root, whenever one exists. To the latter end one uses an auxilliary functional $\rho(G,\tau) = |\tau - \theta|$ to prove existence of a unique Fréchet differentiable root in a local neighbourhood of the parameter θ when considering the derivative at F_θ. Also it is sufficient to consider the expansion (1.3) for T defined on $\mathcal{G}$, and the usual mathematical extension of the domain of T to the linear space of signed measures is of little importance here.

The Fréchet derivative may be considered strong in the sense that existence of the Fréchet derivative for statistical functionals implies existence of the weaker Hadamard or compact derivatives of Reeds(1976), Fernholz(1983), and the Gâteaux derivative discussed by Kallianpur(1963), a special case of which is the influence curve

$$IC(x,F,T) = \lim_{\varepsilon \to 0} \frac{T[(1-\varepsilon)F+\varepsilon\delta_x] - T[F]}{\varepsilon}$$

; here δ_x is the distribution attributing mass 1 to the point x.
The Gâteaux derivative is given by

$$\int IC(x,F,T)\ d(G-F)(x),$$

which coincides with the Fréchet derivative when the latter exists. Unfortunately comments by Kallianpur(1963) which were in specific relation to the maximum likelihood estimator (mle) led other researchers to believe the derivative too strong to obtain. Indeed Huber(1981) states " Unfortunately the concept of Fréchet differentiability appears too strong : in too many cases, the Fréchet derivative does not exist, and even if it does, the fact is difficult to establish. " .
In Clarke(1983) simple conditions for Fréchet differentiability of M-functionals were given together with a counterexample to the comments of Kallianpur.

Boos and Serfling(1980) introduce the related notion of a quasi-differential which assumes the same expansion (1.3), but restricts $G=F_n$ and allows for small order errors in probability with respect to the Kolmogorov distance between F_n and F. This expansion does not offer the same properties of robust description of the estimating functional, and even the mean functional satisfies this stochastic form of differentiability. Beran(1977) also adopts a differential approach using the Hellinger metric, though this appears to be for more specific application.

A weaker set of conditions than conditions A of Clarke(1983) are introduced in section 2, though for smooth psi functions conditions A of that paper are easier to apply. Theorem 2.1 of this paper is necessary to show condition A_4' , introduced here, holds for the popular nonsmooth psi functions. It can be considered as a variation or a generalization of the Glivenko Cantelli result. Conditions A' are used in sections 3 and 4 in the theorems that give existence of a unique continuous and Fréchet differentiable root of equations (1.2). In particular the arguments for weak continuity follow when either of Lévy or Prokhorov metrics are used. Important examples of application are given in section 5, together with the conclusion.

2. A DISCUSSION OF DEFINITIONS AND CONDITIONS A'

Suppose f maps E^r to itself and θ is a point near which f is Lipschitz. Denote Ω_f to be the set of points at which f fails to be differentiable, which by Radermachers theorem is known to be a set of Lebesgue measure zero. Let $Jf(\tau)$ be the usual $r \times r$ matrix of partial derivatives whenever $\tau \notin \Omega_f$.

Definition 2.1: *The generalized Jacobian of* f *at* θ *, denoted by* $\partial f(\theta)$ *, is the convex hull of all* $r \times r$ *matrices* Z *obtained as the limit of a sequence of the form* $Jf(\tau_i)$ *where* $\tau_i \to \theta$ *and* $\tau_i \in \Omega_f$.

The generalized Jacobian $\partial f(\theta)$ is said to be of maximal rank provided every matrix in $\partial f(\theta)$ is of maximal rank (i.e. nonsingular). The following proposition is proved on page 71 of F.H. Clarke (1983).

Proposition 2.1: *The generalized Jacobian* $\partial f(\theta)$ *is upper semicontinuous, which means, given* $\epsilon > 0$ *there exists a* $\delta > 0$ *such that for* $\tau \in U_\delta(\theta)$ *, the open ball of radius* δ *centered at* θ ,

$$\partial f(\tau) \subset \partial f(\theta) + \epsilon B_{r\times r} .$$

Here $B_{r\times r}$ *is the unit ball of matrices for which* $B \in B_{r\times r}$ *implies* $||B|| \le 1$.

Remark 2.1: Without loss of generality we can assume $||B||$ to be the least upper bound of $|By|$ where $|y| \leq 1$.

Frequently several solutions of equations (1.1), (1.2) can exist whereupon a robust selection of the functional root is obtained using the idea of a selection functional ρ introduced in Clarke (1983). The robust selection functional retains the continuity properties of the selected root in small enough neighbourhoods $n(\epsilon,F) \subset G$ of a distribution F , which can be considered here to be defined by metrices d_* . The M-functional is then defined by ψ and ρ as $T[\psi,\rho,.]$. Typical choices for d_* include d_k, d_L , d_p the Kolmogorov, Lévy and Prokhorov metrics respectively.

<u>Conditions A′</u>:

A_0' : $T[\psi,\rho,F_\theta] = \theta$,

A_1' : $\psi(x,\tau)$ is an $r \times 1$ vector function on $R \times \Theta$ which is continuous and bounded on $R \times D$ where $D \subset \Theta$ is some nondegenerate compact interval containing θ in its interior, and R is some separable metrizable space

A_2' : $\psi(x,\tau)$ is locally Lipschitz in τ about θ in the sense that for some constant α

$$|\psi(x,\tau) - \psi(x,\theta)| < |\tau - \theta|$$

uniformly in $x \in R$ and for all τ in a neighbourhood of θ

A_3' : Letting differentiation be with respect to the argument in parentheses $\partial K_{F_\theta}(\tau)$ is of maximal rank at $\tau = \theta$

A_4' : Given $\delta > 0$ there exists an $\epsilon > 0$ such that for all $G \in n(\epsilon,F_\theta)$

$$\sup_{\tau\in D} |K_G(\tau) - K_{F_\theta}(\tau)| < \delta$$

and

$$\partial K_G(\tau) \subset \partial K_{F_\theta}(\tau) + \delta\, B_{r\times r} \quad \text{uniformly in } \tau \in D .$$

Remark 2.2: $A_0' \equiv A_0$

Remark 2.3: For a function ψ satisfying A_1' it follows from remark 2.2 and theorem 6.1 in Clarke (1983) that given $\delta > 0$ there exists an $\epsilon > 0$ such that for all $G \in n(\epsilon, F_\theta)$

$$\sup_{\tau \in D} |K_G(\tau) - K_{F_\theta}(\tau)| < \delta ,$$

whenever $n(\epsilon, F_\theta)$ is generated by metrics d_k, d_L, d_p .
This establishes the first part of condition A_4' .

Remark 2.4: If $K_{F_\theta}(\tau)$ is continuously differentiable in τ at θ then $A_3 \equiv A_3'$, where condition A_3 is that of Clarke (1983).

Conditions A_0' - A_3' can be considered fairly straightforward, whereas the condition A_4' is not so obvious. When $R = E$, the real line, it can be shown to be a consequence of the following theorem, a proof of which is detailed in the appendix. It is sufficient here to establish the result for the Kolmogorov distance d_k .

Theorem 2.1 : *Let* A *be a class of continuous functions defined on* E *with the following properties: (1)* A *is uniformly bounded, that is, there exists a constant* H *such that* $|f(x)| \leq H < \infty$ *for all* $f \in A$ *and* $x \in E$ *; and (2)* A *is equicontinuous. Let* $F_\theta \in G$ *be given. Then,*

for every $\delta > 0$ *there is an* $\epsilon > 0$ *such that* $d_k(F_\theta, G) \leq \epsilon$ *implies*

$$\sup_{f \in A} \sup_{x \in E \cup \{+\infty\}} \left| \int_{I_x} f dG = \int_{I_x} f dF_\theta \right| < \delta , \qquad (2.1)$$

where integration is performed over the intervals I_x *which can be either open or closed of the form* $(-\infty, x)$ *or* $(-\infty, x]$

Remark 2.5: A similar proof yields the same result with d_L replaced by d_k . In some instances Fréchet differentiability with respect to d_k implies that with respect to d_L , d_p following (6.2) of Clarke (1983).

Consider ψ with continuous partial derivatives bar on a finite set of points $S(\tau)$. From F.H. Clarke (1983 pp. 75-83) it follows that

$$\partial K_G(\tau) = \partial \int \psi(y,\tau)dG(y) \subset \int \partial\psi(y,\tau)dG(y) \ , \qquad (2.2)$$

from which the right hand side can be expanded to a finite summation

$$\sum_{j=1}^{m} \int_{I_j} f_j(y,\tau)dG(y) + \sum_{x \in S(\tau)} \partial\psi(x,\tau)G\{x\} \ .$$

Here $f_j \in A$ and $\frac{\partial\psi}{\partial\tau}(y,\tau) = f_j(y,\tau)$ on the connected interval I_j , for $j = 1,\ldots m$. Since ψ is Lipschitz in τ and $\partial\psi(x,\tau)$ bounded, theorem 2.1 implies condition A_4' .

3. UNIQUENESS OF FUNCTIONAL SOLUTIONS TO EQUATIONS

For those psi functions which do not admit a unique root of the equations, at least a unique root of the equations in a local region of the parameter space about θ can be shown to exist for small enough neighbourhoods of F_θ . If conditions A' are with respect to Lévy or Prokhorov neighbourhoods, existence of a weakly continuous root is shown, for which the global argument of Clarke (1983) can be used to select it if more than one root exists. When the Kolmogorov distance is used only consistency is directly established.

The following propositions are established on pp.252-255 of F.H. Clarke (1983), and obviate the condition of continuous derivatives in the argument for the inverse function theorem.

Proposition 3.1: *Suppose* f *satisfies properties described in Section 2 and*

$$4\lambda_f \le \inf_{\partial f(\theta)} \|M(\theta,f)\| ,$$

where the infimum is taken over all matrices $M(\theta,f) \in \partial f(\theta)$ *, and for some* $\delta > 0, \tau \in U_\delta(\theta)$ *implies*

$$2\lambda_f \le \inf_{\partial f(\tau)} \|M(\tau,f)\| .$$

Then for arbitrary τ_1 , $\tau_2 \in \overline{U}_\delta(\theta)$ *, the closure of the ball* $U_\delta(\theta)$,

$$|f(\tau_1) - f(\tau_2)| \ge 2\lambda_f |\tau_1-\tau_2| .$$

Proposition 3.2: *Under the conditions of Proposition 3.1* $f(U_\delta(\theta))$ *contains* $U_{\lambda_f\delta}(f(\theta))$.

Remark 3.1: For $v \in U_{\lambda_f\delta}(f(\theta))$ we can define $f^{-1}(v)$ to be the unique $\tau \in U_{\lambda_f\delta}(\theta)$ such that $f(\tau) = v$ and Proposition 3.1 implies f^{-1} is Lipschitz with Lipschitz constant $1/(2\lambda_f)$.

Lemma 3.1: *Let conditions* Λ' *hold for some* ψ,ρ . *Then there is a* $\delta_1 > 0$ *and an* $\epsilon_1 > 0$ *such that for all* $G \in n(\epsilon_1,F_\theta)$ *any matrix*

$$M(\tau,G) \in \partial K_G(\tau)$$

will satisfy $\|M(\tau,G)\| > 2\lambda$ *where* λ *is defined to be a value for which* $M(\theta,F_\theta) \in \partial K_{F_\theta}(\theta)$ *implies* $\|M(\theta,F_\theta)\| > 4\lambda$.

Remark 3.2: If $K_{F_\theta}(\tau)$ is continuously differentiable in τ then the choice of $\lambda = 1/(4\|M(\theta,F_\theta)^{-1}\|)$ satisfies the criterion of Lemma 3.1.

Proof of Lemma 3.1: Since ∂K_{F_θ} is upper semicontinuous, choose by Proposition 2.1 $\delta_1 > 0$ such that $\partial K_{F_\theta}(\tau) \subset \partial K_{F_\theta}(\theta) + \lambda B_{r\times r}$ whenever $\tau \in U_{\delta_1}(\theta)$. By condition Λ_4' there exists an $\epsilon_1 > 0$ such that $G \in n(\epsilon_1,F_\theta)$ implies

$$\partial K_G(\tau) \subset \partial K_{F_\theta}(\tau) + \lambda B_{r\times r} \quad \text{uniformly in } \tau \in D .$$

Hence given $M(\tau,G) \in \partial K_G(\tau)$ for $\tau \in U_{\delta_1}(\theta)$ there exists $M(\theta,F_\theta) \in \partial K_{F_\theta}(\theta)$ such that

$$\|M(\tau,G) - M(\theta,F_\theta)\| < 2\lambda ,$$

whence by Proposition 3.1 $\|M(\tau,G)\| > 2\lambda$.

It is now possible to state and prove the uniqueness argument of Theorem 3.1 of Clarke (1983) using weakened conditions A′ . The result also implies existence of a weakly continuous root for either Lévy or Prokhorov neighbourhoods. As usual the following selection functional is only used as an auxilliary device.

<u>Theorem 3.1</u>: *Let* $\rho(G,\tau) = |\tau-\theta|$ *and suppose conditions* A′ *hold. Then given* $\kappa > 0$ *there exists an* $\epsilon > 0$ *such that* $G \in n(\epsilon,F_\theta)$ *implies* $T[\psi,\rho,G]$ *exists and is an element of* $U_\kappa(\theta)$. *Further for this* ϵ *there is a* $\kappa^* > 0$ *such that*

$$I(\psi,G) \cap U_{\kappa^*}(\theta) = T[\psi,\rho,G] ,$$

and $\partial K_G(\tau)$ *is of maximal rank for* $\tau \in U_{\kappa^*}(\theta)$. *For any null sequence of positive numbers* $\{\epsilon_n\}$ *let* $\{G_k\}$ *be an arbitrary sequence for which* $G_k \in n(\epsilon_k,F_\theta)$. *Then*

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

Proof of Theorem 3.1 : Since $\partial K_{F_\theta}(\tau)$ is upper semicontinuous in τ choose $0 < \kappa^* < \min(\delta_1, \kappa)$ such that $\tau \in U_{\kappa^*}(\theta)$ implies

$$\inf_{\partial K_G(\tau)} ||M(\tau,G)|| > 2\lambda \quad \text{for all} \quad G \in n(\epsilon_1, F_\theta)$$

where the infimum is taken over all matrices $M(\tau,G) \in \partial K_G(\tau)$. Here δ_1 , ϵ_1 , and λ are defined in Lemma 3.1. Hence

$$4\lambda(G) = \inf_{\partial K_G(\theta)} ||M(\theta,G)|| > 2\lambda \ .$$

Choose $0 < \epsilon^* \le \epsilon_1$ so that the following relations hold

$$\begin{aligned} \partial K_G(\tau) &\subset \partial K_{F_\theta}(\tau) + (\lambda/4)B_{r\times r} && \text{by } A'_4 \\ &\subset \partial K_{F_\theta}(\theta) + (\lambda/2)B_{r\times r} && \text{by Proposition 2.1} \\ &\subset \partial K_G(\theta) + \lambda B_{r\times r} && \text{by } A'_4 \end{aligned}$$

Then for every $M(\tau,G) \in \partial K_G(\tau)$ there exists an $M(\theta,G) \in \partial K_G(\theta)$ such that

$$||M(\tau,G) - M(\theta,G)|| < \lambda < 2\lambda(G) \ ,$$

whenever $G \in n(\epsilon^*, F_\theta)$ and uniformly in $\tau \in U_{\kappa^*}(\theta)$.

By Proposition 3.1 $K_G(.)$ is a one-to-one function from $U_{\kappa^*}(\theta)$ onto $K_G(U_{\kappa^*}(\theta))$ and by Proposition 3.2 the image set contains the open ball of radius $\lambda\kappa^*/2$ about $K_G(\theta)$. The argument for uniqueness now proceeds as in Clarke (1983).

4. FRÉCHET DIFFERENTIABILITY

It will be assumed in this section that $K_{F_\theta}(\tau)$ has at least a continuous derivative $K_{F_\theta}(\tau)$ at $\tau = \theta$, which is denoted $M(\theta)$. This is common with absolutely continuous parametric families. With this restriction Fréchet differentiability follows.

Theorem 4.1: *Let* $\rho(G,\tau) = |\tau-\theta|$ *and assume conditions* A′ *hold with respect to this functional and neighbourhoods generated by the metrics* d_* *on* G . *Suppose for all* $G \in G$

$$\int_R \psi(x,\theta)d(G-F_\theta)(x) = O(d_*(G,F_\theta)) \tag{4.1}$$

Then $T[\psi,\rho,\cdot]$ *is Fréchet differentiable at* F_θ *with respect to* (G,d_*) *and has derivative*

$$T'_{F_\theta}(G-F_\theta) = -M(\theta)^{-1} \int_R \psi(x,\theta)d(G-F_\theta)(x) \ .$$

To prove the theorem it is necessary to introduce the following generalization of the mean value result described as Proposition 2.6.5 in F.H. Clarke (1983)

Proposition 4.1 *Let* f *be Lipschitz on an open convex set* U *in* E^r *and let* τ_1 *and* τ_2 *be points in* U . *Then one has*

$$f(\tau_1) - f(\tau_2) \in \text{co } \partial f([\tau_1,\tau_2]) \ (\tau_2-\tau_1)$$

(The right hand side above denotes the convex hull of all points of the form $Z(\tau_2-\tau_1)$ *where* $Z \in \partial f(u)$ *for some point* u *in* $[\tau_1,\tau_2]$. *Since* $[\text{co } \partial F([\tau_1,\tau_2])](\tau_2-\tau_1) = \text{co}[\partial f([\tau_1,\tau_2])(\tau_2-\tau_1)]$, *there is no ambiguity.)*

Proof of Theorem 4.1: Abbreviate $T[\psi,\rho,.] = T[.]$ and let κ^*,ϵ be given by Theorem 2. Let $\{\epsilon_k\}$ be so that $\epsilon_k \downarrow 0^+$ as $k \to \infty$ and let $\{G_k\}$ be any sequence such that $G_k \in n(\epsilon_k,F_\theta)$. By theorem 2, $T[G_k]$ exists and is unique in $U_{\kappa^*}(\theta)$ for $k > k_0$ where $\epsilon_{k_0} \le \epsilon$. By A'_4 see that for arbitrary $\delta > 0$

$$\partial K_{G_k}(\tau) \subset \partial K_{F_\theta}(\tau) + \delta B_{r\times r} \quad \text{uniformly in } \tau \in D \tag{4.2}$$

for sufficiently large k. Consider the two term expansion,

$$0 = K_{G_k}(T[G_k]) = K_{G_k}(\theta) + M(\tilde{\tau}_k, G_k)(T[G_k]-\theta) , \tag{4.3}$$

where $|\tilde{\tau}_k - \theta| < |T[G_k] - \theta|$, which tends to zero as $k \to \infty$ by theorem 3.1, and $\tilde{\tau}_k$ is evaluated at different points for each component function expansion obtained as a consequence of Proposition 4.1 (i.e. $\tilde{\tau}_k$ takes different values in each row of matrix M). See from (4.3), (4.1) and Lemma 3.1 that

$$|T[G_k] - \theta| = O(K_{G_k}(\theta)) = O(\epsilon_k) .$$

Also,

$$T[G_k] - \theta = -M(\theta)^{-1} K_{G_k}(\theta) - M(\theta)^{-1}\{M(\tilde{\tau}_k, G_k) - M(0)\}(T[G_k] - \theta) .$$

By upper semicontinuity of $K_G(\tau)$ in τ and (4.2)

$$||M(\tilde{\tau}_k, G_k) - M(\theta)|| = o(1) .$$

So

$$|T[G_k] - \theta - T'_{F_\theta}(G_k - F_\theta)| = o(1)\ 0\ (d*(G_k, F_\theta)) = o(\epsilon_k)$$

5. EXAMPLES AND CONCLUSION

Huber (1964, 1981) introduced a proposal for estimation of location and scale of the normal distribution defined as a solution of

$$\int \psi\left(\frac{x - \tau_1}{\tau_2}\right) dF_n(x) = 0$$

where $\Theta = \{(\tau_1, \tau_2): -\infty < \tau_1 < \infty , \tau_2 > o\}$ and the vector function $\psi = (\psi_1, \psi_2)'$ where

$$\psi_1(x) = \max[-k , \min(k,x)]$$

$$\psi_2(x) = \psi_1(x)^2 - \beta(k) ,$$

and $\beta(k) = \int \min(k^2, x^2) d\Phi(x)$. Here Φ denotes the normal distribution. setting $\underset{\sim}{\theta} = (\theta_1, \theta_2)$, where $\sim$ now distinguishes the vector parameter, it follows that since $K_\Phi(\underset{\sim}{\theta})$ is continuously differentiable, the Jacobian

$$M(\underset{\sim}{\theta}) = -\frac{1}{\theta_2}\begin{pmatrix} \int \psi_1'(y)\,d\Phi(y) & 0 \\ 0 & \int y\psi_2'(y)\,d\Phi(y) \end{pmatrix} \tag{5.1}$$

Condition A_0' follows since $E_\Phi[\psi] = 0$. A_1' , A_2' hold by inspection, and A_3' holds since $M(\underset{\sim}{\theta})$ is nonsingular. Remark 2.2 suffices for the first part of A_4' . To apply theorem 2.1 consider the function

$$f(x,\underset{\sim}{\tau}) = I_{(\tau_1-k\tau_2,\tau_1+k\tau_2)}(x)\begin{pmatrix} -\frac{1}{\tau_2} & \frac{-(x-\tau_1)}{\tau_2^2} \\ \frac{-(x-\tau_1)}{\tau_2^2} & \frac{-(x-\tau_1)^2}{\tau_2^3} \end{pmatrix}$$

$$+\frac{1}{\tau_2}\left(I_{(-\infty,\tau_1-k\tau_2]}(x) + I_{[\tau_1+k\tau_2,\infty)}(x)\right)\begin{pmatrix} -1 & -k \\ -k & -k^2 \end{pmatrix}$$

It is clear that $A = \{f(.,\underset{\sim}{\tau}):\underset{\sim}{\tau} \in D\}$ forms a bounded equicontinuous class of functions on E . Also

$$\partial K_G(\underset{\sim}{\tau}) = \int_{(\tau_1-k\tau_2,\tau_1+k\tau_2)} f(x,\underset{\sim}{\tau})\,dG(x)$$

$$+ \partial\psi(\tau_1-k\tau_2,\underset{\sim}{\tau})G\{\tau_1-k\tau_2\} + \partial\psi(\tau_1+k\tau_2,\underset{\sim}{\tau})G\{\tau_1+k\tau_2\} ,$$

where differentiation of ψ is with respect to the second argument, while

$$\partial K_{F_\theta}(\underset{\sim}{\tau}) = \int_{(\tau_1-k\tau_2,\tau_1+k\tau_2)} f(x,\underset{\sim}{\tau})\,dF_{\underset{\sim}{\theta}}(x) \text{ , where } F_{\underset{\sim}{\theta}}(x) = \Phi\left(\frac{x-\theta_1}{\theta_2}\right)$$

A'_4 follows by Theorem 2.1 and because $\partial\psi(\tau_1 - k\tau_2, \underset{\sim}{\tau})$ and $\partial\psi(\tau_1 + k\tau_2, \underset{\sim}{\tau})$ are bounded. Assumption (4.1) holds for the Kolmogorov distance through integration by parts and noting that ψ is a function of total bounded variation. Thus by Theorems 3.1 and 4.1 there exists a root that is Fréchet differentiable at $F_{\underset{\sim}{\theta}} = \Phi\left(\frac{x-\theta_1}{\theta_2}\right)$ with respect to d_k. Since Φ has a bounded density T is Fréchet differentiable with respect to d_L, d_p also. Consequently, the infinitesimal robustness of this M-estimator at the normal parametric distribution is evident through Fréchet differentiability. It is also Fréchet differentiable at the distribution $F_0\left(\frac{x-\theta_1}{\theta_2}\right)$ for which the density function of F_0 is

$$f_0(x) = \begin{cases} \dfrac{(1-\epsilon)}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} & \text{for } |x| \le k \\[2ex] \dfrac{(1-\epsilon)}{\sqrt{2\pi}}\, e^{\frac{k^2}{2} - k|x|} & \text{for } |x| > k \end{cases}$$

with k and ϵ connected through

$$\frac{2\phi(k)}{k} - 2\,\Phi(-k) = \frac{\epsilon}{1-\epsilon}$$

($\phi = \Phi'$ being the standard normal density). Then the M-estimator coincides with the mle, and provides another example of a robust and asymptotically efficient estimator.

Examples where multiple roots of the equations exist include Hampel's 3-part redescender M-estimator for location dependent on three parameters a,b,c;

$$\psi_{a,b,c}(x) = \begin{cases} x & |x| \le a \\ a\,\mathrm{sign}(x) & a \le |x| \le b \\ a\,\dfrac{c-|x|}{c-b}\,\mathrm{sign}(x) & b \le |x| \le c \\ 0 & c \le |x| \end{cases}$$

With the choice of selection functional $\rho(G,\tau) = |\tau - G^{-1}(\frac{1}{2})|$, whereby the root closest to the median is selected, the functional $T[\psi_{a,b,c},\rho,\cdot]$ is Fréchet differentiable at $\Phi(\frac{x-\theta_1}{\theta_2})$.

In a sense weak continuity and Fréchet differentiability of the functional at the empirical distribution function are also important. Weak continuity at F_n indicates stability of the estimate in the presence of rounding errors in the recording of observations, and at least for sufficiently large n the effects of gross errors can be considered blunted. Fréchet differentiability at F_n on the other hand, could be used to justify asymptotics involved in Edgeworth type expansions and bootstrapping, for example as considered in Hampel(1982), Beran(1982). When the psi function is smooth, the only change to the arguments of Clarke(1983) for Fréchet differentiability at F_n , is to replace F_θ by F_n in conditions A_1-A_4.

Similarly the same substitution of conditions can be made in the results of this paper, however if it should occur that an observation X falls exactly at the point where $\psi(X,\tau)$ does not have a continuous partial derivative at $\tau = T[F_n]$ then the generalized gradient $\partial K_{F_n}(T[F_n])$ does not reduce to a single matrix. Even though such an event would occur with probability zero in most forseeable examples in which the underlying distribution was absolutely continuous, it can be said nevertheless that the proof used in theorem 4.1 does not follow through. In this instance the question of whether T is Fréchet differentiable at F_n is then left open. At least in the domain of M-functionals defined through (1.2), it can be concluded that Huber's(1981) remarks should not be interpreted in the sense that Fréchet differentiability is too strong. This is only the case for nonrobust M-functionals, and consequently we should consider Fréchet differentiability an advantage.

The problems induced by nonsmooth psi functions are not unique to proofs of Fréchet differentiability, and are applicable to many asymptotic proofs. More frequently it is the case, that rather than consider the difficulties, the appropriate smoothness assumptions are made in the proofs, but somehow the results are expected to be applicable to those continuous but nonsmooth functions also. Nonsmooth analysis can the be considered as one possible avenue of justifying such an approach.

6. ACKNOWLEDGEMENT

The author wishes to express gratitude to Professor F.H.Hampel and Professor R.J.Carroll for encouragement and the opportunity to pursue this research. This research has received partial support from a US Airforce Grant No. F 49620 82 C 009, while the author was visiting the University of North Carolina at Chapel Hill.

6. APPENDIX

The proof of theorem 2.1 is preceded by some necessary lemmas. The notational abbreviation $G(x^-) = \lim_{h \downarrow 0} G(x - h)$ is used.

Lemma 1: Let $G(x)$ 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

$$a = x_0 < x_1 < \ldots\ldots < x_{k'} = b,$$

so that

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

Proof: Define $G^{-1}(t) = \inf \{x \mid G(x) \geq t, \, x \in [a,b]\}$

Since G is right continuous $G(G^{-1}(t)) \geq t$, choose

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

where $k \geq 1$ is chosen so that

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

Then

$$\begin{aligned} G(y_j) - G(y_{j-1}) &\geq G(a) + \tfrac{j}{k}(G(b^-) - G(a)) - G(y_{j-1}) \\ &\geq G(a) + \tfrac{j}{k}(G(b^-) - G(a)) \\ &\quad - \{G(a) + \tfrac{j-1}{k}(G(b^-) - G(a)) + \eta/4\} \\ &= \tfrac{1}{k}(G(b^-) - G(a)) - \eta/4 \\ &\geq \eta/4 \end{aligned}$$

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

$$\begin{aligned} G(y_j) - G(y_j^-) &= G(y_j) - G(y_{j-1}^-) \\ &\geq G(y_j) - G(y_{j-1}) \\ &\geq \eta/4 \end{aligned}$$

But this contradicts the initial assumption. Now since

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

then

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

Note that $y_0 = a$, $y_1 > a$, and if $y_k < b$, then $G(b^-) - G(y_k) = 0$.

Let

$$a = x_0 < x_1 < \ldots . < x_{k'} = b$$

be the partition formed from $\{y_j\}_{j=1}^k \cup \{b\}$

Lemma 6.2: Let F_θ be given. Also for given $c > o$ let $C = (-c, c]$. Then $\forall\, \eta > o\ \exists\, \epsilon' > o$ such that $d_k(G,F_\theta) \le \epsilon'$ implies

$$\sup_{f\in A} \sup_{x\in C} \left| \int_{C\cap I_x} f(y)dG(y) - \int_{C\cap I_x} f(y)dF_\theta(y) \right| < \eta \qquad (6.1)$$

where intervals I_x may represent either open or closed intervals from $-\infty$ to x.

Proof: Given $\eta > o$, let $\{d_i\}_{i=1}^{\ell}$ be the at most finite set of points in C such that $F_\theta(d_i) - F_\theta(d_i^-) \ge \eta/(16H)$, if they exist. Since the family A is equicontinuous and $\overline{C}$, the closure of C, is compact, we may choose a decomposition

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

so that $a_{i-1} \le x < y \le a_i$ implies $|f(x) - f(y)| < \eta/4$, for every $f\in A$, and $i = 1, \ldots .., m$. Let $\{a_i^*\}_{i=o}^k$ be the further decomposition obtained by 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 6.1 whenever $F_\theta(a_i^{*-}) - F_\theta(a_{i-1}^*) > \eta/(4H)$ there exists a finite decomposition $\{x_{ij}\}_{j=o}^{n_i}$ so that

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

for which

$$F_\theta(x_{ij}^-) - F_\theta(x_{i(j-1)}) < \eta/(4H) \qquad j=1,\ldots,n_i \qquad (6.2)$$

If $F_\theta(a_i^{*-}) - F_\theta(a_{i-1}^*) < \eta/(4H)$, set $n_i = 1$, $x_{io} = a_{i-1}^*$ and $x_{i2} = a_i^*$.

That is, no further partitioning is necessary. Let $\{b_i\}_{i=o}^{n'}$ be the set of points that partition $(-c,c]$ formed by combining $\{x_{ij}\}_{j=o}^{n_i}$, $i = 1, \ldots, k$. Denote F^* the possibly improper distribution that attributes weight $F_\theta(b_i) - F_\theta(b_i^-)$ to the points b_i, and weight $F_\theta(b_i^-) - F_\theta(b_{i-1})$ to the points $p_i = \frac{1}{2}(b_i + b_{i+1})$, $i = 1, \ldots, n' - 1$.

Suppose $x \in C$ is given. Then either: (a) there exists an $o \le i_x \le n' - 1$ such that $b_{i_x} < x < b_{i_x+1}$; or (b) there exists an $1 \le i_x \le n'$ for which $x = b_{i_x}$.

For case (a) and $f \in A$

$$\left|\int_{C\cap I_x} f dF^* - \int_{C\cap I_x} f dF_\theta\right| \le \sum_{j=1}^{i_x} \int_{(b_{j-1}, b_j)} \left|f(p_j) - f(y)\right| dF_\theta(y)$$

$$+ \left|\int_{(b_{i_x}, c]\cap I x} f(y) d(F^* - F_\theta)(y)\right|$$

$$\le \frac{\eta}{4} F_\theta\{C\} + 2H\{F_\theta(b_{i_x+1}^-) - F_\theta(b_{i_x})\}$$

$$< \eta/4 + \eta/2 = \frac{3}{4}\eta$$

For case (b) where $x = b_{i_x}$ for some $1 \le i_x \le n'$

$$\left|\int_{C\cap I_x} f dF^* - \int_{C\cap I_x} f dF_\theta\right| \le \sum_{j=1}^{i_x} \int_{(b_{j-1}, b_j)} \left|f(p_j) - f(y)\right| dF_\theta(y)$$

$$< \frac{\delta}{4} F_\theta\{C\} \le \eta/4$$

Hence

$$\sup_{f\in A} \sup_{x\in C} \left|\int_{C\cap I_x} f dF^* - \int_{C\cap I_x} f dF_\theta\right| < \frac{3}{4}\eta$$

This is true for any distribution satisfying the inequalities (6.2).

In particular we can choose ϵ^* such that $d_k(G,F_\theta) < \epsilon^*$ implies

$$G(x_{ij}^-) - G(x_{ij-1}) < \eta/(4H), \quad \text{for} \quad j = 1, \ldots, n_i$$
$$i = 1, \ldots, k.$$

Let G^* be the corresponding improper measure constructed from G.

Then

$$\sup_{f\in A} \sup_{x\in C} \left|\int_{C\cap I_x} f dG^* - \int_{C\cap I_x} f dG\right| < \tfrac{3}{4}\eta .$$

It is now convenient to consider case (b) first

$$\sup_{f\in A} \left|\int_{C\cap I_x} f dG^* - \int_{C\cap I_x} f dF_\theta^*\right| \qquad (+)$$

$$\le H \left[\sum_{j=1}^{i_x} \left|\int_{(b_{j-1},b_j)} d(G-F_\theta)\right| + \sum_{j=1}^{i_x} \left|G\{b_j\} - F_\theta\{b_j\}\right|\right]$$

$$\le H \left[\sum_{j=1}^{n'} \left|\int_{(b_{j-1},b_j)} d(G-F_\theta)\right| + \sum_{j=1}^{n'} \left|G\{b_j\} - F_\theta\{b_j\}\right|\right] \qquad (*)$$

Choose $o < \epsilon' < \epsilon^*$ such that $d_k(G,F_\theta) < \epsilon'$ implies $(*) < \eta/4$.

There are two possibilities for case (a). Either

$$b_{i_x} < x < p_{i_x} \quad \text{for some} \quad o \le i_x \le n' - 1,$$

whence

$$(+) \le (*) + \sup_{f\in A} \left|\int_{(b_{i_x},c]\cap I_x} f d(G^*-F_\theta^*)\right| = (*) < \eta/4,$$

or

$$p_{i_x} \le x \le b_{i_x+1}, \quad o \le i_x \le n' - 1,$$

for which

$$(+) \le (*) + H \left|G(b_{i_x+1}^-) - G(b_{i_x}) - F_\theta(b_{i_x+1}^-) + F_\theta(b_{i_x})\right| < \eta/2.$$

For either case if it happens that $d_k(G,F_\theta) < \epsilon'$, then

$$\sup_{f\in A} \sup_{x\in C} \left|\int_{C\cap I_x} f dG^* - \int_{C\cap I_x} f dF_\theta^*\right| < \eta/2.$$

Then the lemma follows.

Proof of Theorem 2.1: Given $\delta > o$ choose $c > o$ so that

$$F_\theta\{E - C\} < \delta/(8H)$$

For any $x \in (-\infty,-c]$ and G within Kolmogorov distance $\delta/(8H)$ from F_θ

$$\left|\int_{I_x} fdG - \int_{I_x} fdF_\theta\right| \le H(G\{I_x\} + F_\theta\{I_x\})$$

$$\le H(G\{E - C\} + F_\theta\{E - C\})$$

$$\le \delta/2$$

Let ϵ' be given by Lemma 6.2 for the choice of $\eta = \delta/2$.

Choose $\epsilon = \min\{\epsilon', \delta/(8H)\}$. Then for arbitrary $x > c$ and G within Kolmogorov distance ϵ from F_θ

$$\left|\int_{I_x} fdG - \int_{I_x} fdF_\theta\right| \le H(G\{E - C\} + F_\theta\{E - C\}) + \sup_{y\in C}\left|\int_{C\cap I_y} fdG - \int_{C\cap I_y} fdF_\theta\right|$$

$$< \delta/2 + \delta/2 \qquad \text{by Lemma 6.2}$$

Hence

$$d_k(G,F_\theta) < \epsilon \text{ implies}$$

$$\sup_{f\in A} \sup_{x\in E} \left|\int_{I_x} fdG - \int fdF_\theta\right| < \delta$$

REFERENCES

Beran, R.: Robust location estimates, Ann. Statist., 5, 431-444 (1977)

Beran, R.: Estimated sampling distributions: the bootstrap and competitors, Ann. Statist., 10, 212-225 (1982)

Boos, D.D. and Serfling, R.J.: A note on differentials and the CLT and LIL for statistical functions with application to M-estimates. Ann. Statist., 8, 618-624 (1980)

Carroll, R.J.: On the asymptotic distribution of multivariate M-estimates, J. of Mult. Analysis, 8, 361-371 (1978)

Clarke, B.R.: Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations, Ann. Statist., 1196-1205 (1983)

Clarke, F.H.: Optimization and Nonsmooth Analysis, Wiley, New York (1983)

Fernholz, L.T.: Von Mises Calculus for Statistical Functionals, Lecture notes in statistics, 19 Springer Verlag, New York, 1983.

Hampel, F.R.: Contributions to the Theory of Robust Estimation, Ph.D. Thesis, University of California, Berkeley, 1968

Hampel, F.R.: A general qualitative definition of robustness. Ann Math. Statist. 42, 1887-1896 (1971)

Hampel, F.R.: The influence curve and its role in robust estimation, J. Amer. Statist. Ass., 62, 1179-1186 (1974)

Hampel, F.R.: Optimally bounding the gross error-sensitivity and the influence of position in factor space; Proc. Amer. Statist. Assoc., Statist. Computing Section, 59-64 (1978)

Hampel, F.R.: Small-sample asymptotic distributions of M-estimators of location, Biometrika, 69, 29-46 (1982)

Huber, P.J.: Robust estimation of a location parameter, Ann. Math. Statist., 35, 73-101 (1964)

Huber, P.J.: The behaviour of maximum likelihood estimates under non standard conditions, in: Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability Vol. 1, University of California Press, Berkeley (1967)

Huber, P.J.: Robust Statistical Procedures, Regional Conference Series in Applied Mathematics No. 27, Soc. Industr. App. Math., Philadelphia, Penn. 1977

Huber, P.J.: Robust Statistics, Wiley, New York, 1981

Kallianpur, G.: Von Mises functions and maximum likelihood estimation. Sankhya, Ser. A., 23, 149-158 (1963)

Reeds, J.A.: On the Definition of von Mises Functions, Ph.D Thesis, Department of Statistics, Harvard University, Cambridge, Mass. 1976.

Von Mises, R.: On the asymptotic distribution of differentiable statistical functions. An.. Math. Statist. 18, 309-348 (1947)