The Annals of Statistics
1977, Vol. 5, No. 3, 431–444

ROBUST LOCATION ESTIMATES[1]

BY RUDOLF BERAN

University of California at Berkeley

Measures of location differentiable at every density in the Hellinger metric are constructed in this paper. Differentiability entitles these location functionals to the label "robust," even though their influence curves need not be bounded and continuous. The latter properties are, in fact, associated with functionals differentiable in the Prokhorov metric. A Hellinger metric concept of minimax robustness of a location measure at a density shape f is developed. Asymptotically optimal estimators are found for minimax robust location measures. Since, at f, their asymptotic variance equals the reciprocal of Fisher information, asymptotic efficiency at f and robustness near f prove compatible.

1. Introduction. In his fundamental paper on M-estimators, Huber (1964) demonstrated the existence of an M-estimator of location with minimax asymptotic variance over a specified neighbourhood of a given distribution shape. Although suggestive, the minimax property does not strictly imply robustness of the corresponding M-estimator: uniform convergence in distribution over the specified neighbourhood was not established and the bias caused by an asymmetric data distribution was not fully analyzed.

Hampel (1968, 1974) assessed robustness of an estimator by viewing it as a functional of the empirical cdf and examining the behaviour of the first Volterra derivative of this functional at the ideal model cdf; this approach generated the influence curve concept. Hampel argued, heuristically, that the influence curve should have certain properties, such as boundedness and continuity, to ensure robustness of the corresponding functional and estimator. This idea led him to the construction of M-estimators which sacrifice some asymptotic precision in return for qualitative robustness.

The first to consider an estimator as a functional of the empirical cdf and systematically draw conclusions from the nature of the functional was von Mises (1947). His paper established a link between the asymptotic distribution of an estimator and the Volterra derivatives of the corresponding functional. The well-known relation between the influence curve of an M, L or R location estimator and its asymptotic variance can be viewed as a special case of von Mises' idea (see Huber (1972) for details).

Takeuchi (1967) was the first to study robustness from a functional viewpoint. His unpublished paper included the following ideas, presented heuristically:

Received December 1974; revised October 1976.

[1] This research was partially supported by National Science Foundation Grant GP-31091X.

AMS 1970 *subject classifications*. Primary 62G35; Secondary 62E20.

Key words and phrases. Robust location measures, robust estimates, differentiable functionals, minimax robust, Hellinger metric, asymptotic efficiency, Fisher information.

characterize those functionals of the data cdf which correspond to location parameters, scale parameters, etc.; restrict attention to those functions which are continuous and Fréchet differentiable with respect to some metric on the space of all cdf's; study the sensitivity of functionals to small changes in the underlying distribution through the Fréchet derivatives and identify as robust those functionals which appear most insensitive to changes in distribution; estimate parameters by the value of the corresponding functional of the empirical cdf. Unresolved questions in Takeuchi's paper were: choice of topology and precise definition of the Fréchet derivative; clarifying what is meant by an optimally robust functional; construction of functionals with specified derivative; justification for the estimators used.

A further development of robustness ideas related to Takeuchi's and Hampel's work was carried out by Bickel and Lehmann (1975). Their paper systematically studied location functionals of various types; in particular it identified sub-classes of M, L and R functionals which are Prokhorov continuous and have estimators with globally well-behaved asymptotic variance. Differentiability of functionals was not considered but it was noted that any theory of optimal estimation for location functionals must take into account the existence of super-efficient estimators.

A basic question remains: Is it possible, in general, to reconcile the requirement that a location estimator be robust in a neighbourhood of a specified distribution shape with the requirement that the estimator be asymptotically efficient at the given distribution shape? The robustness papers cited above strongly suggest that the answer is negative if attention is restricted to M, L or R estimators and if Prokhorov continuity or bounded continuous influence curve is the robustness criterion; the best known instances occur when the ideal model distribution is normal or asymmetric.

The present paper shows (subject to technicalities) that robustness and asymptotic efficiency of a location estimator are compatible requirements. The result is established by introducing a new class of location estimators and by using a slightly weaker concept of robustness (which, however, is still strong enough to deal with common models of data contamination). Other results include a minimax concept of optimally robust location functionals and a theory of asymptotically optimal estimation for such functionals.

The specific statistical model considered is as follows. We observe random variables $X_1, X_2, \cdots, X_n$ and assume that, apart from possible data contamination, the random variables are i.i.d. with density f which is known up to translation. We wish to construct a location measure which is translation equivariant, coincides in value with a prescribed location functional (such as mean or a quantile) if the data density is actually some translation of f, is quantitatively insensitive to small perturbations of f (data contamination), and possesses a robust and efficient estimator. Admittedly, this model is too simple to be of much practical use—for that purpose it would be preferable to consider a location-scale model,

at least. However, the pure location model provides the fewest technical obstacles to theoretical insight. Extensions to more complex models are possible and will be treated elsewhere. The assumption of a data density is not so restrictive as it might appear; although we shall assume densities with respect to Lebesgue measure, other choices of measure are possible in the theory.

Formally, let $\mathscr{T}$ denote the set of all densities defined with respect to Lebesgue measure on the real line. For any function r whose domain is the real line, let $r * h$ denote the translation of r defined by $(r * h)(x) = r(x - h)$. A real-valued functional T defined on $\mathscr{T}$ will be called a *measure of location* (or *location functional*) if it is equivariant under arbitrary location shifts:

$$T(g * h) = T(g) + h \tag{1.1}$$

for all $g \in \mathscr{T}$ and all real h. Our first goal is to identify those location functionals which are continuous and differentiable in a suitable topology at the model density f and its translations.

2. Differentiable functionals. Let $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ denote, respectively, the usual norm and inner product in L_2. If f and g belong to $\mathscr{T}$, the Hellinger distance between f and g is defined as $\|f^{\frac{1}{2}} - g^{\frac{1}{2}}\|$. Convergence in the Hellinger metric clearly implies convergence in the Kolmogorov and Prokhorov metrics, but the converse is not true. We will use the Hellinger metric in discussing continuity and differentiability of location measures for three main reasons. First, differentiability in the Hellinger metric is relatively easy to check for specific functionals. Secondly, the Hellinger metric underlies substantial portions of classical large sample estimation theory. It is not surprising, therefore, that Hellinger differentiable functionals prove to have an accessible asymptotic estimation theory. Thirdly, the Hellinger metric makes possible and useful a minimax robustness concept (Theorem 5) and the reconciliation of estimator robustness with estimator efficiency.

DEFINITION. A real-valued functional T defined on $\mathscr{T}$ is said to be differentiable at $g \in \mathscr{T}$ if there exists a function $\rho_g \in L_2$, depending on both g and T, such that

$$\lim_{n\to\infty} \|g_n^{\frac{1}{2}} - g^{\frac{1}{2}}\|^{-1}[T(g_n) - T(g) - \langle \rho_g, g_n^{\frac{1}{2}} - g^{\frac{1}{2}} \rangle] = 0 \tag{2.1}$$

for every sequence of densities $\{g_n\}$ converging to g in the Hellinger topology.

The function ρ_g is called the Fréchet derivative of T at g with respect to the Hellinger metric (see Luenberger (1969) for more details on Fréchet derivatives). It is easily seen that the property (2.1) does not define ρ_g uniquely; the function $\rho_g + \alpha g^{\frac{1}{2}}$ will also work in (2.1) for arbitrary choice of real α. To avoid ambiguity and simplify later formulae, we will construct the derivative ρ_g to be orthogonal to $g^{\frac{1}{2}}$. Under this convention, the function $\rho_g/(2g^{\frac{1}{2}})$ coincides with Takeuchi's derivative or Hampel's influence curve when all of these exist. It is clear from (2.1) that differentiability implies continuity in the Hellinger metric.

For arbitrary g_n and g in $\mathscr{T}$, we have, by projection, the representation

$$g_n^{\frac{1}{2}} = \cos(\theta_n) g^{\frac{1}{2}} + \sin(\theta_n)\delta_n , \tag{2.2}$$

where $\cos(\theta_n) = \langle g_n^{\frac{1}{2}}, g^{\frac{1}{2}} \rangle$, $\theta_n \in [0, \pi/2]$, $||\delta_n|| = 1$ and $\langle g^{\frac{1}{2}}, \delta_n \rangle = 0$. The following is immediate:

THEOREM 1. *A functional T is differentiable at g if and only if there exists $\rho_g \in L_2$ depending upon g and T such that for every sequence $\{\theta_n : \theta_n \to 0, \theta_n \in [0, \pi/2]\}$ and every sequence $\{\delta_n : ||\delta_n|| = 1, \langle g^{\frac{1}{2}}, \delta_n \rangle = 0\}$*

$$\lim_{n\to\infty} \theta_n^{-1}[T(g_n) - T(g) - \theta_n \langle \rho_g, \delta_n \rangle] = 0 , \tag{2.3}$$

where g_n is defined through (2.2).

A location functional T will be called *robust* at g if it is differentiable at g, and therefore at all translations of g. Since differentiability implies continuity, the robustness label seems justified. Moreover, the local linear approximation entailed by differentiability makes possible quantitative comparisons of robustness among different location functionals (cf. Section 3).

Analogous definitions of differentiability and robustness can of course be made using other metrics, such as the L_1 or Prokhorov metrics. However, the class of functionals differentiable in the Hellinger metric is substantially larger than the differentiable classes generated by the other metrics; this fact is essential to the reconciliation of robustness with asymptotic efficiency and to the statement of Theorem 5. Though less restrictive as a robustness concept, Hellinger metric differentiability can handle simple data contamination models, such as the mixture model. Indeed, if $g_\alpha(x) = (1 - \alpha)f(x) + \alpha h(x)$ for $f, h \in \mathscr{F}$ and $\alpha \in [0, 1]$, then by Vitali's theorem, $\lim_{\alpha\to 0} ||g_\alpha^{\frac{1}{2}} - f^{\frac{1}{2}}|| = 0$ i.e., a small mixing fraction α corresponds to a small perturbation of f in the Hellinger metric.

THEOREM 2. *Let T be a measure of location differentiable at g. Then*

(i) *T is differentiable at $\{g * h : h \in R^1\}$ with $\rho_{g*h} = \rho_g * h$.*
(ii) *If g is absolutely continuous and $g'/g^{\frac{1}{2}} \in L_2$, then $\langle \rho_g, -g'/(2g^{\frac{1}{2}}) \rangle = 1$.*

PROOF. Let $\{g_n \in \mathscr{T}\}$ converge to g in the Hellinger metric. Since $T(g_n * h) - T(g * h)$ equals $T(g_n) - T(g)$ for all real h and since $\langle \rho_g, \delta_n \rangle$ equals $\langle \rho_g * h, \delta_n * h \rangle$, conclusion (i) follows from Theorem 1 and (2.2).

Under the assumptions of (ii), a standard argument shows that

$$\lim_{n\to 0} ||h^{-1}[g^{\frac{1}{2}} * h - g^{\frac{1}{2}}] + g'/(2g^{\frac{1}{2}})|| = 0 , \tag{2.4}$$

which implies

$$\lim_{n\to 0} \langle \rho_g, h^{-1}[g^{\frac{1}{2}} * h - g^{\frac{1}{2}}] \rangle = \langle \rho_g, -g'/(2g^{\frac{1}{2}}) \rangle . \tag{2.5}$$

Conclusion (ii) follows from the fact that $h^{-1}[T(g * h) - T(g)] = 1$, from differentiability of T, and from (2.5).

Also of interest is the question converse to Theorem 2: given a density f and a function $\rho \in L_2$ which is orthogonal to $f^{\frac{1}{2}}$ and satisfies either $\langle \rho, -f'/(2f^{\frac{1}{2}}) \rangle = 1$

or $\langle \rho', f^{\frac{1}{2}} \rangle = 1$, according to which assumptions are preferred, does there exist a location functional differentiable at f such that $\rho_f = \rho$? The answer proves to be affirmative; moreover, the construction yields a functional which is differentiable at all $g \in \mathscr{T}$. One plausible approach is to consider the possible derivatives at f of the usual M, L and R location functionals. Unfortunately these functionals solve our converse problem only for restricted choices of $\rho \in L_2$. For example, an M-functional cannot have derivative ρ at f unless the function $\rho/f^{\frac{1}{2}}$ is uniformly bounded. To obtain an answer for more general ρ, we introduce a new class of location measures.

Let T_0 be any location functional which is differentiable at every density $g \in \mathscr{T}$ (cf. Theorem 4 for one such class of functionals) and define a functional T by

$$T(g) = T_0(g) + \int \rho(x) g^{\frac{1}{2}}[x + T_0(g) - T_0(f)]\, dx\,, \tag{2.6}$$

for every $g \in \mathscr{T}$. It is evident that T is a location functional and that $T(f) = T_0(f)$ since ρ is assumed orthogonal to $f^{\frac{1}{2}}$.

THEOREM 3. *Suppose that $\rho \in L_2$ is orthogonal to $f^{\frac{1}{2}}$ and is absolutely continuous with derivative $\rho' \in L_2$ which satisfies $\langle \rho', f^{\frac{1}{2}} \rangle = 1$. Suppose that T_0 is a location functional differentiable at g with derivative $\rho_{0,g}$. Then*

(i) *T defined by* (2.6) *is a location functional differentiable at g with derivative*

$$\begin{aligned} \rho_g(x) = {} & [1 - \int \rho'(t) g^{\frac{1}{2}}(t + T_0(g) - T_0(f))\, dt] \rho_{0,g}(x) \\ & + \rho(x - T_0(g) + T_0(f)) \\ & - [\int \rho(t - T_0(g) + T_0(f)) g^{\frac{1}{2}}(t)\, dt] g^{\frac{1}{2}}(x)\,. \end{aligned} \tag{2.7}$$

(ii) *In particular $\rho_f = \rho$.*

PROOF. Let $\{g_n\}$ be defined as in (2.2), with $\theta \to 0$ as $n \to \infty$ and $\{\delta_n\}$ arbitrary apart from $\|\delta_n\| = 1$. Write D_n for $T_0(g_n) - T_0(f)$ and D for $T_0(g) - T_0(f)$. By assumption,

$$D_n - D = T_0(g_n) - T_0(g) = \theta_n \langle \rho_{0,g}, \delta_n \rangle + o(\theta_n)\,. \tag{2.8}$$

From the assumptions on ρ,

$$\lim_{h \to 0} \|h^{-1}(\rho * h - \rho) + \rho'\| = 0\,. \tag{2.9}$$

It follows from (2.8) and (2.9) that

$$\begin{aligned} & \int \rho(x) g_n^{\frac{1}{2}}(x + D_n)\, dx \\ & \quad = \cos(\theta_n) \int \rho(x) g^{\frac{1}{2}}(x + D_n)\, dx + \sin(\theta_n) \int \rho(x) \delta_n(x + D_n)\, dx \\ & \quad = \int \rho(x) g^{\frac{1}{2}}(x + D)\, dx - (D_n - D) \int \rho'(x) g^{\frac{1}{2}}(x + D)\, dx \\ & \qquad + \theta_n \int \rho(x) \delta_n(x + D)\, dx + o(\theta_n) \\ & \quad = \int \rho(x) g^{\frac{1}{2}}(x + D)\, dx - \theta_n \int \rho_{0,g}(x) \delta_n(x)\, dx \int \rho'(t) g^{\frac{1}{2}}(t + D)\, dt \\ & \qquad + \theta_n \int \rho(x - D) \delta_n(x)\, dx + o(\theta_n)\,. \end{aligned} \tag{2.10}$$

Hence $T(g_n) = T(g) + \theta_n \langle \rho_g, \delta_n \rangle + o(\theta_n)$, with ρ_g defined by (2.7); note that ρ_g is orthogonal to $g^{\frac{1}{2}}$.

If $g \equiv f$, the first and third terms in (2.7) vanish under the assumptions on ρ, proving (ii).

Initial location measures T_0 differentiable at every density $g \in \mathscr{T}$ can be found within the class of Huber's M-functionals. Define T_0 implicitly through the equation

$$\int \psi[x - T_0(g)]g(x)\,dx = 0\,. \tag{2.11}$$

THEOREM 4. *Suppose that the function ψ is strictly monotone and bounded, $\lim_{x\to\infty}\psi(x) > 0$, $\lim_{x\to-\infty}\psi(x) < 0$, and ψ has a continuous bounded derivative ψ'. Then the location functional T_0 defined by* (2.11) *is differentiable at every $g \in \mathscr{T}$ with derivative*

$$\rho_{0,g}(x) = \frac{2\psi[x - T_0(g)]g^{\frac{1}{2}}(x)}{\int \psi'[x - T_0(g)]g(x)\,dx}\,. \tag{2.12}$$

PROOF. For fixed g, $H_g(z) = \int \psi(x - z)g(x)\,dx$ is a continuous, bounded, strictly monotone function of z, with $\lim_{z\to\infty} H_g(z) < 0$, $\lim_{z\to-\infty} H_g(z) > 0$. Hence, $T_0(g)$ is well defined by (2.11) for all $g \in \mathscr{T}$. If $g_n \to g$ in the Hellinger topology, then (2.11) and boundedness of ψ imply that

$$\lim_{n\to\infty} \int \psi[x - T_0(g_n)]g(x)\,dx = 0\,, \quad \text{i.e.,} \quad \lim_{n\to\infty} H_g[T_0(g_n)] = H_g[T_0(g)]\,.$$

Since H_g^{-1} is continuous the functional T_0 is continuous at every $g \in \mathscr{T}$.

Now

$$\begin{aligned} 0 &= \int \psi[x - T_0(g_n)]g_n(x)\,dx \\ &= \int \psi[x - T_0(g)]g_n(x)\,dx - [T_0(g_n) - T_0(g)] \int \psi'(x - \xi_n)g_n(x)\,dx\,, \end{aligned} \tag{2.13}$$

where ξ_n lies between $T_0(g_n)$ and $T_0(g)$ and hence converges to $T_0(g)$ as $n \to \infty$. Since

$$\int \psi[x - T_0(g)]g_n(x)\,dx = 2\theta_n \int \psi[x - T_0(g)]g^{\frac{1}{2}}(x)\delta_n(x)\,dx + O(\theta_n^2)\,, \tag{2.14}$$

in view of (2.11) and (2.2), and

$$\int \psi'(x - \xi_n)g_n(x)\,dx = \int \psi'(x - \xi_n)g(x)\,dx + O(\theta_n) \tag{2.15}$$

and $\int \psi'(x - z)g(x)\,dx > 0$ for all real z, the theorem follows from (2.13).

REMARKS. Many well-known functions ψ fulfill the requirements of Theorem 4: for instance, $\psi(x) = \arctan(x)$.

If ψ is assumed to have a bounded second derivative, the conclusions of Theorem 4 can be strengthened to

$$T_0(g_n) = T_0(g) + \theta_n\langle\rho_{0,g}, \delta_n\rangle + O(\theta_n^2)\,. \tag{2.16}$$

Similarly, if ρ' is absolutely continuous with $\rho'' \in L_2$ and if the preliminary location functional T_0 satisfies (2.16), then the remainder term in Theorem 3 is also $O(\theta_n^2)$. This fact will be used in Section 5.

An argument similar to the proof of Theorem 4 establishes Hellinger metric differentiability of the median functional at every density g which is continuous

and positive at its median. The derivative at g can still be expressed in a form resembling (2.12) by letting T_0 denote the median functional, setting $\psi(x) = \operatorname{sgn}(x)$, and replacing the denominator with $2g[T_0(g)]$. If g has a finite derivative at its median, then (2.16) also holds.

It can be shown readily that the mean functional is nowhere Hellinger differentiable or continuous. This remains true even when attention is restricted to those densities that have a finite mean. Thus the boundedness assumption on ψ in Theorem 4 cannot be abandoned.

3. Minimax location functionals. Consider anew the estimation of location problem described in the introduction. If the effect of data contamination is represented by a small perturbation (Hellinger metric) in the expected density shape f, it is reasonable for the sake of robustness to restrict attention to those location functionals which are at least continuous at f and its translations. If differentiability is postulated, quantitative comparisons of robustness at f become possible.

Indeed, suppose T_1 and T_2 are any two location measures. Then the functional T_2^* defined by $T_2^*(g) = T_2(g) + [T_1(f) - T_2(f)]$ for all $g \in \mathscr{T}$ is also a location measure and coincides in value with T_1 at all translations of f. Differentiability of T_2 implies the same for T_2^* and the derivatives coincide. Thus, in principle, there exist a vast number of location measures which coincide in value at all translations of f but have different derivatives at these points. The robustness of any particular location measure belonging to this multitude depends upon the behavior of the linear term in the definition of a differentiable functional.

The situation described can be viewed as a zero-sum game between the statistician and nature with payoff function $L(\rho, \delta) = |\langle \rho, \delta \rangle|$. The quantity $2\theta L(\rho, \delta)$ is a first order approximation to the change $|T(g) - T(f)|$ that occurs in the value of T when f is perturbed, in "direction" δ, into the density g defined by $g^{\frac{1}{2}}(x) = \cos(\theta) f^{\frac{1}{2}}(x) + \sin(\theta)\delta(x)$; it is assumed that T is differentiable at f with derivative ρ and that $\theta > 0$ is small. To achieve quantitative robustness, the statistician attempts to minimize $L(\rho, \delta)$ by choice of derivative ρ, subject to the constraints satisfied by the derivative at f of a location functional: $\rho \in L_2$, $\rho \perp f^{\frac{1}{2}}$, $\langle \rho, -f'/(2f^{\frac{1}{2}}) \rangle = 1$. Nature, assumed malevolent, seeks to maximize the payoff by choice of perturbation "direction" δ, subject to the constraints $\|\delta\| = 1$ and $\delta \perp f^{\frac{1}{2}}$. Fortunately for our analysis, the game has a saddle point in pure strategies.

THEOREM 5. *Suppose that f is absolutely continuous with finite Fisher information $I(f) = \|f' f^{-\frac{1}{2}}\|^2$, $\rho \in L_2$, $\langle \rho, -f'/(2f^{\frac{1}{2}}) \rangle = 1$, $\delta \perp f^{\frac{1}{2}}$ and $\|\delta\| = 1$. Then*

$$\max_\delta \min_\rho |\langle \rho, \delta \rangle| = \min_\rho \max_\delta |\langle \rho, \delta \rangle| = |\langle \rho_0, \delta_0 \rangle| , \tag{3.1}$$

where

$$\begin{aligned} \rho_0 &= -2I^{-1}(f) f' f^{-\frac{1}{2}} \\ \delta_0 &= -I^{-\frac{1}{2}}(f) f' f^{-\frac{1}{2}} . \end{aligned} \tag{3.2}$$

PROOF. Since $\max_\delta \min_\rho |\langle \rho, \delta \rangle| \leqq \min_\rho \max_\delta |\langle \rho, \delta \rangle|$, it suffices to prove the reverse inequality. By the Cauchy–Schwarz inequality, $\max_\delta |\langle \rho, \delta \rangle| = ||\rho||$, the maximizing choice of δ being $||\rho||^{-1}\rho$. Because of the constraints on it, ρ may be represented in the form $\rho = \rho_0 + \sigma$, where ρ_0 is defined in (3.2) and $\sigma \in L_2$ is orthogonal to both ρ_0 and $f^{\frac{1}{2}}$. Hence

$$\min_\rho \max_\delta |\langle \rho, \delta \rangle| = \min_\sigma ||\rho_0 + \sigma|| = ||\rho_0|| = 2I^{-\frac{1}{2}}(f) . \tag{3.3}$$

On the other hand, for δ_0 defined in (3.2)

$$\max_\delta \min_\rho |\langle \rho, \delta \rangle| \geqq \min_\rho |\langle \rho, \delta_0 \rangle| = 2I^{-\frac{1}{2}}(f) , \tag{3.4}$$

the last step using the equality constraint on ρ. The desired inequality and therefore the theorem follow from (3.3) and (3.4). Evidently, (ρ_0, σ_0) is a saddle point for the game.

The influence curve that corresponds to ρ_0 is $\psi_0 = -I^{-1}(f)f'f^{-1}$. Unless ψ_0 is uniformly bounded, there does not exist an M-functional that has derivative ρ_0 at f; the M-functional having ψ_0 as its score function is not even continuous in the Hellinger topology when ψ_0 is unbounded, hence is not robust. A well-known example occurs when f is $N(0, 1)$ so that $\rho_0(x) = 2(2\pi)^{-\frac{1}{4}}x \exp(-x^2/4)$ and $\psi_0(x) = x$. However, in this example, the construction of Theorems 3 and 4 does yield a location functional having derivative ρ_0 at the $N(0, 1)$ density. This functional is minimax robust at all translations of the $N(0, 1)$.

4. Asymptotically optimal estimators. Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables each distributed according to density $g \in \mathscr{T}$. Let T be a location functional differentiable at g with derivative ρ_g. We pose the question: how well can $T(g)$ be estimated on the basis of the sample $(X_1, X_2, \cdots, X_n)$? To exclude pathological super-efficient estimators (cf. Bickel and Lehmann (1975) regarding the existence of such), it is necessary to impose some regularity assumptions upon the class of estimators considered. Theorem 6 in this section provides an asymptotic answer to our question under restrictions on the estimators which are consistent with the goal of robustness. This theorem is an extension of Hájek's (1970) representation for limiting distributions of regular estimators in parametric families of distributions.

Let $\mathscr{C}(g, \beta)$ denote the set of all sequences of densities $\{g_n\}$ such that

$$\lim_{n\to\infty} ||n^{\frac{1}{2}}(g_n^{\frac{1}{2}} - g^{\frac{1}{2}}) - \beta|| = 0 , \tag{4.1}$$

where $\beta \in L_2$ and $g \in \mathscr{T}$. Note that (4.1) implies that β is orthogonal to $g^{\frac{1}{2}}$. Let $\mathscr{C}(g)$ denote the union over β of all sets $\{\mathscr{C}(g, \beta) : \beta \in L_2, \beta \perp g^{\frac{1}{2}}\}$. Let $\hat{T}_n$ be any estimator of the location measure T which is *regular* at g in the sense that, for every sequence $\{g_n\} \in \mathscr{C}(g)$ and for $(X_1, X_2, \cdots, X_n)$ independent and identically distributed with density g_n, the distribution of $n^{\frac{1}{2}}[\hat{T}_n - T(g_n)]$ converges weakly to a distribution $\mathscr{D}(g)$ that depends only on g and not upon the particular sequence $\{g_n\}$. This assumption excludes super-efficient estimators and is, in fact, a robustness property because it entails $\lim_{n\to\infty} T(g_n) = T(g)$ for every sequence of densities $\{g_n\} \in \mathscr{C}(g)$. Indeed $\hat{T}_n - T(g) \to_P 0$ under g, $\hat{T}_n - T(g_n) \to_P 0$

under every sequence $\{g_n\} \in \mathscr{C}(g)$ and the density sequences $\{\prod_{i=1}^n g_n(x_i)\}$, $\{\prod_{i=1}^n g(x_i)\}$ are contiguous, as can be verified with the aid of (4.3) below.

THEOREM 6. *Suppose that T is differentiable at g with derivative ρ_g and $\hat{T}_n$ is an estimator regular at g. Then $\mathscr{D}(g)$ can be represented as the convolution of a $N(0, 4^{-1}||\rho_g||^2)$ distribution with $\mathscr{D}_1(g)$, a distribution depending only upon g.*

Thus, for no regular estimator $\hat{T}_n$ can $\mathscr{D}(g)$ be less dispersed than $N(0, 4^{-1}||\rho_g||^2)$. It is not immediately clear whether the distribution $\mathscr{D}_1(g)$ can be made degenerate at zero by suitable choice of regular $\hat{T}_n$. However, in Section 5, we shall see that this is the case, at least under supplementary assumptions.

Let L_n be a random variable defined by

$$L_n = 2 \log \{\prod_{i=1}^n [g_n^{\frac{1}{2}}(X_i)/g^{\frac{1}{2}}(X_i)]\} \tag{4.2}$$

whenever (4.2) is finite. The following result is needed for the proof of Theorem 6.

LEMMA. *Let $\{g_n\}$ be a sequence of densities such that* (4.1) *holds for some $\beta \in L_2$. Then for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P_g[|L_n - 2n^{-\frac{1}{2}} \sum_{i=1}^n \beta(X_i) g^{-\frac{1}{2}}(X_i) + 2||\beta||^2| > \varepsilon] = 0 . \tag{4.3}$$

The proof of this lemma is implicit in arguments given by Le Cam (cf. Le Cam (1968), for example). The lemma implies that for every $\{g_n\} \in \mathscr{C}(g)$, the density sequences $\{\prod_{i=1}^n g_n(x_i)\}$ and $\{\prod_{i=1}^n g(x_i)\}$ are contiguous. Note that (4.3) can also be used to deduce the more familiar expansion of log-likelihood for a suitably regular parametric family of densities.

PROOF OF THEOREM 6. We use a characteristic function approach developed for the parametric case by Bickel (cf. Roussas (1972) for a published version of that proof). Since T is differentiable at g and $n^{\frac{1}{2}}[g_n^{\frac{1}{2}} - g^{\frac{1}{2}}] \to_{L_2} \beta$ as $n \to \infty$,

$$\lim_{n\to\infty} n^{\frac{1}{2}}[T(g_n) - T(g)] - \langle \rho_g, \beta \rangle = 0 . \tag{4.4}$$

Therefore, the characteristic function of $n^{\frac{1}{2}}[\hat{T}_n - T(g_n)]$ under g_n is

$$\begin{aligned} & E_{g_n} \exp[iun^{\frac{1}{2}}(\hat{T}_n - T(g_n))] \\ & \quad = E_{g_n} \exp[iun^{\frac{1}{2}}(\hat{T}_n - T(g)) - iu\langle \rho_g, \beta \rangle] + o(1) \\ & \quad = E_g \exp[iun^{\frac{1}{2}}(\hat{T}_n - T(g)) + L_n - iu\langle \rho_g, \beta \rangle] + o(1) , \end{aligned} \tag{4.5}$$

this being true for all $\beta \in L_2$ orthogonal to $g^{\frac{1}{2}}$. Choose $\beta = h\rho_g||\rho_g||^{-1}$, h being an arbitrary real constant. By considering only a subsequence if necessary, we may assume that under g, the random vectors

$$\{(n^{\frac{1}{2}}[\hat{T}_n - T(g)], n^{-\frac{1}{2}} \sum_{i=1}^n \rho_g(X_i)||\rho_g||^{-1} g^{-\frac{1}{2}}(X_i))\}$$

converge weakly to a random vector (S, Z), depending on ρ_g, such that Z has a $N(0, 1)$ distribution. It follows from (4.3) and the choice of β that the random vectors $\{(n^{\frac{1}{2}}[\hat{T}_n - T(g)], L_n)\}$ converge weakly under g to the random vector $(S, 2hZ - 2h^2)$.

In addition.

$$\lim_{n\to\infty} E_g \exp[iun^{\frac{1}{2}}(\hat{T}_n - T(g)) + L_n] = E\exp[iuS + 2hZ - 2h^2]. \tag{4.6}$$

For, on the one hand

$$E_g|\exp[iun^{\frac{1}{2}}(\hat{T}_n - T(g)) + L_n]| = 1 = E|\exp[iuS + 2hZ - 2h^2]|. \tag{4.7}$$

On the other hand, there exists a probability space and versions of $\{(n^{\frac{1}{2}}[\hat{T}_n - T(g)], L_n)\}$ and $(S, 2hZ - 2h^2)$ defined on that space such convergence w.p. 1. holds as well as weak convergence. Since (4.7) remains true for these versions, Vitali's theorem gives (4.6).

From (4.5), (4.6) and regularity of $\hat{T}_n$,

$$E\exp[iuS] = E\exp[iuS + 2hZ]\exp[-iuh||\rho_g|| - 2h^2] \tag{4.8}$$

for all real h. Let $\varphi(u, v) = E[\exp iuS + ivZ]$ denote the characteristic function of (S, Z). Equation (4.8) becomes

$$\varphi(u, 0) = E\exp[iuS + 2hZ]\exp[-iuh||\rho_g|| - 2h^2]. \tag{4.9}$$

The right side of (4.9) is analytic in h, constant for all real h, hence constant for all complex h. In particular, the choice of $h = 2^{-1}iv$ yields, for all real u, v,

$$\begin{aligned}\varphi(u, 0) &= \varphi(u, v)\exp[2^{-1}uv||\rho_g|| + 2^{-1}v^2] \\ &= \varphi(u, v)\exp[2^{-1}(v + 2^{-1}||\rho_g||u)^2]\exp[-8^{-1}||\rho_g||^2u^2].\end{aligned} \tag{4.10}$$

The special choice $v = -2^{-1}||\rho_g||u$ gives

$$\varphi(u, 0) = \varphi(u, -2^{-1}||\rho_g||u)\exp[-8^{-1}||\rho_g||^2u^2] \tag{4.11}$$

for all real u. Since $\varphi(u, 0)$ is the characteristic function of $\mathscr{D}(g)$ and the first factor on the right side of (4.11) is the characteristic function of $S - 2^{-1}||\rho_g||Z$ while the second factor is the characteristic function of a $N(0, 4^{-1}||\rho_g||^2)$ distribution, the theorem follows.

The above proof makes use of the apparently arbitrary choice $\beta = h\rho_g||\rho_g||^{-1}$, h a real number. In fact, this choice yields the strongest possible theorem statement; other choices of β would result in smaller variances for the normal component of $\mathscr{D}(g)$, hence would transfer probability mass into the component $\mathscr{D}_1(g)$.

5. Robust location estimators. A natural estimator for a location measure $T(g)$ is $T(\hat{g}_n)$, where $\hat{g}_n$ is the kernel density estimator

$$\hat{g}_n(x) = (nc_n)^{-1}\sum_{i=1}^n w\left(\frac{x - X_i}{c_n}\right), \tag{5.1}$$

$\{c_n\}$ being a sequence of positive constants converging to zero at a suitable rate and w being a smooth member of $\mathscr{T}$.

THEOREM 7. *Suppose*

(i) *w is absolutely continuous and w' is integrable.*

(ii) *g is continuous.*

(iii) $\lim_{n\to\infty} c_n = 0$ *and* $\lim_{n\to\infty} n^{\frac{1}{2}}c_n = \infty$.

(iv) *T is a functional continuous in the Hellinger metric.*

Then $T(\hat{g}_n) \to_P T(g)$ *as* $n \to \infty$.

PROOF. Let G_n denote the empirical cdf based upon $(X_1, X_2, \cdots, X_n)$, let G denote the cdf of g, and let

$$\tilde{g}_n(x) = c_n^{-1} \int w\left(\frac{x-y}{c_n}\right) dG(y) . \tag{5.2}$$

Integration by parts yields

$$|\hat{g}_n(x) - \tilde{g}_n(x)| \leqq n^{-\frac{1}{2}}c_n^{-1} \sup_x |B_n \cdot G(x)| \cdot \int |w'(x)|\, dx , \tag{5.3}$$

where B_n is the empirical Brownian bridge process. Also,

$$\tilde{g}_n(x) - g(x) = \int [g(x - c_n z) - g(x)]w(z)\, dz , \tag{5.4}$$

which tends to zero for every x as $n \to \infty$. From (5.3), there exist versions of the $\{\hat{g}_n\}$, defined on a suitable probability space, such that $\sup_x |\hat{g}_n(x) - \tilde{g}_n(x)| \to 0$ w.p. 1. Hence $P[\lim_{n\to\infty} \hat{g}_n^{\frac{1}{2}}(x) = g^{\frac{1}{2}}(x)$ for every $x] = 1$. Since $||g_n^{\frac{1}{2}}|| = 1 = ||g^{\frac{1}{2}}||$, we conclude by Vitali's theorem that $\lim_{n\to\infty} ||\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}|| = 0$ w.p. 1 for these versions. The theorem follows by continuity of T.

Suppose that T is differentiable at g. The following theorem shows that, under some additional assumptions, $T(\hat{g}_n)$ is an asymptotically optimal estimator of $T(g)$ in the sense of Theorem 6. We expect that weaker assumptions would suffice, but do not have a proof in that case.

THEOREM 8. *Suppose*

(i) *w is symmetric about 0, square integrable, and* $\int x^2 w(x)\, dx < \infty$.

(ii) *w is absolutely continuous and w' is integrable.*

(iii) *g is absolutely continuous, g' is absolutely continuous, and g'' is uniformly bounded.*

(iv) *$g(x)$ vanishes outside a closed, bounded interval $I \subset (-\infty, \infty)$ and $g(x) \geqq \delta > 0$ for $x \in I$.*

(v) $\lim_{n\to\infty} n^{\frac{1}{2}}c_n = \infty$ *and* $\lim_{n\to\infty} n^{\frac{1}{2}}c_n^2 = 0$.

(vi) $T(g_n) = T(g) + \langle \rho_g, g_n^{\frac{1}{2}} - g^{\frac{1}{2}} \rangle + O(||g_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2)$ *for g_n in a Hellinger neighborhood of g.*

(vii) *ρ_g is continuous on I and vanishes outside I.*

Then the limiting distribution of $n^{\frac{1}{2}}[T(\hat{g}_n) + T(g)]$ *as* $n \to \infty$ *is* $N(0, 4^{-1}||\rho_g||^2)$.

PROOF. Under assumption (vi), it suffices to examine the asymptotic behavior of $\langle \rho_g, \hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}} \rangle$ and $||\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2$. By Taylor expansion,

$$\begin{aligned} n^{\frac{1}{2}}\langle \rho_g, \hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}} \rangle &= n^{\frac{1}{2}} \int_I \frac{\rho_g(x)[\hat{g}_n(x) - g(x)]}{2g^{\frac{1}{2}}(x)}\, dx \\ &\quad - n^{\frac{1}{2}} \int_I \frac{\rho_g(x)[\hat{g}_n(x) - g(x)]^2}{8\xi_n^{\frac{3}{2}}(x)}\, dx \\ &= T_{1n} + T_{2n} , \quad \text{say,} \end{aligned} \tag{5.5}$$

where $\xi_n(x)$ lies between $\hat{g}_n(x)$ and $g(x)$; since g, $\hat{g}_n$, and ρ_g are continuous, ξ_n can be constructed measurable. From assumptions (i) and (ii), for $\tilde{g}_n$ defined by (5.2),

$$\sup_x |\tilde{g}_n(x) - g(x)| \leqq 2^{-1}c_n^2 \sup_x |g''(x)| \int x^2 w(x)\,dx = O(c_n^2)\,. \tag{5.6}$$

Write

$$T_{1n} = n^{\frac{1}{2}} \int_I \psi(x)[\hat{g}_n(x) - \tilde{g}_n(x)]\,dx + n^{\frac{1}{2}} \int_I \psi(x)[\tilde{g}_n(x) - g(x)]\,dx \tag{5.7}$$

where $\psi(x) = \rho_g(x)/(2g^{\frac{1}{2}}(x))$. The second integral on the right is $O(n^{\frac{1}{2}}c_n^2)$ because of (5.6). The first integral can be expressed as $V_n = \int_I dB_n(y) \int \psi(y + c_n z)w(z)dz$, where $B_n(y) = n^{\frac{1}{2}}[G_n(y) - G(y)]$. Now

$$E[V_n - \int_I \psi(y)\,dB_n(y)]^2 \leqq \int_I dG(y)[\int \{\psi(y + c_n z) - \psi(y)\}w(z)\,dz]^2 \tag{5.8}$$

tends to zero as $n \to \infty$, since ψ is continuous and bounded. We conclude from (5.7), (5.8) and assumption (v) that T_{1n} is asymptotically $N(0, 4^{-1}||\rho_g||^2)$.

On the other hand, $T_{2n} \to_p 0$ as $n \to \infty$. Indeed, there exist versions of $\{\hat{g}_n\}$ such that $\sup_x |\hat{g}_n(x) - g(x)| \to 0$ w.p. 1 (cf. proof of Theorem 7); hence for sufficiently large n, $\sup_x |\xi_n(x)| \geqq \delta/2$ w.p. 1. Thus w.p. 1

$$\begin{aligned} |T_{2n}| &\leqq 2(2\delta)^{-\frac{3}{2}}\{n^{\frac{1}{2}} \int_I |\rho(x)|[\hat{g}_n(x) - \tilde{g}_n(x)]^2\,dx \\ &\quad + n^{\frac{1}{2}} \int_I |\rho(x)|[\tilde{g}_n(x) - g(x)]^2\,dx\} \\ &= 2(2\delta)^{-\frac{3}{2}}[W_{1n} + W_{2n}]\,, \quad \text{say.} \end{aligned} \tag{5.9}$$

From (5.6) and assumption (vii), $W_{2n} = O(n^{\frac{1}{2}}c_n^4)$. Moreover,

$$\begin{aligned} E|W_{1n}| &\leqq n^{-\frac{1}{2}} \int_I |\rho(x)| \left[c_n^{-2} \int w^2\left(\frac{x - y}{c_n}\right) g(y)\,dy \right] dx \\ &= n^{-\frac{1}{2}}c_n^{-1} \int_I |\rho(x)|\,dx \int w^2(z)g(x - c_n z)\,dz = O(n^{-\frac{1}{2}}c_n^{-1})\,. \end{aligned} \tag{5.10}$$

Hence $T_{2n} \to_p 0$ as $n \to \infty$. The above considerations imply that $n^{\frac{1}{2}}\langle \rho_g, \hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}} \rangle$ is asymptotically $N(0, 4^{-1}||\rho_g||^{-2})$.

To complete the proof of the theorem, it remains to show that $n^{\frac{1}{2}}||g_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2 \to_p 0$ as $n \to \infty$. But this follows from the representation

$$n^{\frac{1}{2}}||\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2 = -2n^{\frac{1}{2}} \int_I g^{\frac{1}{2}}(x)[\hat{g}_n^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x)]\,dx\,, \tag{5.11}$$

since the right side of (5.11) behaves analogously to the left side of (5.5), with $T_{1n} = 0$.

REMARKS. If g_n is represented in the form (2.2), it becomes evident that $O(||g_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2) = O(\theta_n^2)$. Thus the location functional T defined in Theorem 3 satisfies assumption (vi) of Theorem 8 under the conditions described at the end of Section 2. The corresponding location estimator is

$$T(\hat{g}_n) = T_0(\hat{g}_n) + \int \rho(x)\hat{g}_n^{\frac{1}{2}}[x + T_0(\hat{g}_n) - T_0(f)]\,dx\,. \tag{5.12}$$

If Theorem 8 is applicable, a contiguity argument shows that $T(\hat{g}_n)$ is a regular estimator of $T(g)$ in the sense of Theorem 6. If $\rho = \rho_0$ defined in Theorem 5,

the asymptotic variance of $T(\hat{g}_n)$ under f or any translation of f is precisely the reciprocal of the Fisher information of f. Thus, asymptotic efficiency at f is compatible with minimax robustness at f.

The preliminary estimator $T_0(\hat{g}_n)$ appearing in (5.12) may be replaced by any estimator $\hat{T}_{0,n}$ such that $n^{\frac{1}{2}}[T_0(\hat{g}_n) - \hat{T}_{0,n}] \to_p 0$. If T_0 is an M-functional, the corresponding M-estimator often has this property.

EXAMPLE. Suppose that the ideal model for the density of X_i is $f(x - \theta)$, where f is $N(0, 1)$ and θ is unknown. Let T_0 be the median functional, let $\rho(x) = 2xf^{\frac{1}{2}}(x) = 2(2\pi)^{-\frac{1}{4}}x \exp(-x^2/4)$, and define the location functional T by (5.6). As noted at the end of Section 3, T is a minimax robust location functional at all translations of the $N(0, 1)$ distribution.

Estimate the actual data density g by the kernel estimator $\hat{g}_n$ defined in (5.1), using the Epanechnikov kernel $w(x) = 3/4(1 - x^2)$ $|x| \leqq 1$, for computational simplicity. Define the location estimator $T(\hat{g}_n)$ by (5.12), with one difference: replace $T_0(\hat{g}_n)$ by the sample median since that is simpler and asymptotically equivalent (under Theorem 8 at least). The integral in (5.12) can be evaluated numerically since the integrand is continuous and has compact support by choice of w.

The asymptotics of Theorem 8 do not specify c_n in a useful manner. The following procedure is plausible: for given n, simulate several $N(0, 1)$ samples of size n and determine, by trial and error, which choice of c_n will bring the values of $T(\hat{g}_n)$ into close match with the corresponding sample means. The asymptotic equivalence under normality of $T(\hat{g}_n)$ and the sample mean (see the proof of Theorem 8) is the rationale here.

Strictly speaking, Theorem 8 does not apply to this example because neither ρ nor f has compact support; we will overlook this deficiency which, fortunately, appears to be of small practical importance. If the $\{X_i\}$ are i.i.d. $N(\theta, 1)$ random variables, we expect that for sufficiently large n the distribution of $n^{\frac{1}{2}}(\hat{T}(g_n) - \theta)$ is approximately normal with mean 0 and variance $4^{-1} \int \rho^2(x)\,dx = 1$. Thus $\hat{T}(g_n)$ is both robust and efficient at normality as an estimator of θ. It should be noted that the symmetry of the normal model plays no essential role in this example.

REFERENCES

[1] BICKEL, P. J. and LEHMANN, E. L. (1975). Descriptive statistics for nonparametric models II. Location. *Ann. Statist.* **3** 1045–1064.

[2] HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **14** 323–330.

[3] HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Unpublished Ph.D. dissertation, Univ. of California, Berkeley.

[4] HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.

[5] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

[6] HUBER, P. J. (1972). Robust statistics: a review. *Ann. Math. Statist.* **43** 1041–1067.

[7] LE CAM, L. (1969). *Théorie Asymptotique de la Décision Statistique.* Les Presses de l'Université de Montréal.
[8] LUENBERGER, D. G. (1969). *Optimization by Vector Space Methods.* Wiley, New York.
[9] ROUSSAS, G. (1972). *Contiguity of Probability Measures.* Cambridge Univ. Press.
[10] TAKEUCHI, K. (1967). Robust estimation and robust parameter. Unpublished paper presented at 1967 annual I.M.S. meeting, Washington, D.C.
[11] VON MISES, R. (1947). On the asymptotic distributions of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720