The Annals of Statistics
1977, Vol. 5, No. 3, 445-463

MINIMUM HELLINGER DISTANCE ESTIMATES FOR PARAMETRIC MODELS[1]

BY RUDOLF BERAN

University of California, Berkeley

This paper defines and studies for independent identically distributed observations a new parametric estimation procedure which is asymptotically efficient under a specified regular parametric family of densities and is minimax robust in a small Hellinger metric neighborhood of the given family. Associated with the estimator is a goodness-of-fit statistic which assesses the adequacy of the chosen parametric model. The fitting of a normal location-scale model by the new procedure is exhibited numerically on clear and on contaminated data.

1. Introduction. The statistical problem which motivates this paper can be described as follows. Random variables $X_1, X_2, \cdots, X_n$ are observed. We postulate that the $\{X_i\}$ are independent identically distributed with density belonging to a specified parametric family $\{f_\theta : \theta \in \Theta\}$. At the same time, we recognize that lack of information, data contamination, and other factors beyond our control make it virtually certain that the model is not strictly correct; we hope that it may be close in some sense. How are we to estimate θ in order to investigate the fit of the model to the data?

A good estimator of θ would have two essential properties: it would be efficient if the postulated model for the data were in fact true and its distribution would not be greatly perturbed if the assumed model were only approximately true. The latter property reflects our disinterest in very small deviations from the assumed model—they are expected in any case.

It has long been known that for many parametric families of interest in applications, the maximum likelihood estimator of θ has full asymptotic efficiency among regular estimators. More recently, it has been recognized that maximum likelihood estimators do not, in general, possess the property of stability under small perturbations in the underlying model (see Huber's (1972) review paper for a discussion of location models and earlier references). To remove the instability, Hampel (1974), building upon Huber's earlier work with location models, suggested replacing the maximum likelihood estimator of θ with a related M-estimator whose asymptotic mean is θ under the model density f_θ and is close to θ under small changes in the underlying data distribution. Unfortunately this procedure typically entails a loss of asymptotic efficiency under the model density f_θ.

Received September 1975; revised July 1976.

[1] This research was partially supported by National Science Foundation Grant MPS74-18967.

AMS 1970 *subject classifications.* Primary 62G35; Secondary 62F10.

Key words and phrases. Robust estimates, minimum Hellinger distance estimates, asymptotically efficient estimates, minimax robust.

This paper introduces a new efficient parametric estimator which is intrinsically stable under small perturbations. Apart from technicalities, the proposed estimator of θ is that value (or values) $\hat{\theta}_n$ in the parameter space Θ which minimizes the Hellinger distance between $f_{\hat{\theta}_n}$ and $\hat{g}_n$, where $\hat{g}_n$ is a suitable nonparametric estimator of the density of X_i. If $||\cdot||$ denotes the L_2-norm, the Hellinger distance in question is defined as $||f_{\hat{\theta}_n}^{\frac{1}{2}} - \hat{g}_n^{\frac{1}{2}}||$.

It is interesting to note that this estimator $\hat{\theta}_n$ is related heuristically to the maximum likelihood estimator of θ if the data density is in fact some f_{θ_0} (but not otherwise!). For with this assumption and n sufficiently large, the MLE should be close to θ_0 and the density estimator $\hat{g}_n$ should be close to f_{θ_0}. Finding the maximum likelihood estimator amounts to maximizing $\int \log(f_\theta(x))\, dG_n(x)$ over Θ, where G_n is the empirical cdf of the data. Arguing formally, we expect that this procedure is nearly the same as maximizing over θ near θ_0 the quantity

$$
\begin{aligned}
&\int \log(f_\theta(x)/\hat{g}_n(x))\hat{g}_n(x)\,dx \\
(1.1)\qquad &= 2\int \log[1 + (f_\theta^{\frac{1}{2}}(x)/\hat{g}_n^{\frac{1}{2}}(x) - 1)]\hat{g}_n(x)\,dx \\
&\cong 2\int [(f_\theta^{\frac{1}{2}}(x)/\hat{g}_n^{\frac{1}{2}}(x) - 1) - 2^{-1}(f_\theta^{\frac{1}{2}}(x)/\hat{g}_n^{\frac{1}{2}}(x) - 1)^2]\hat{g}_n(x)\,dx \\
&= -2||f_\theta^{\frac{1}{2}} - \hat{g}_n^{\frac{1}{2}}||^2\,.
\end{aligned}
$$

Thus, it is not unreasonable to suppose that the minimum Hellinger distance estimator $\hat{\theta}_n$ will be asymptotically efficient under f_θ.

Robustness of $\hat{\theta}_n$ is also intuitively plausible. A moment's reflection reveals that if $\hat{g}_n$ is a kernel density estimator whose kernel has compact support (say), then the addition of a sufficiently distant outlier to the data will scarcely affect the value of $\hat{\theta}_n$. Moreover, if $\hat{g}_n^{\frac{1}{2}}$ is perturbed arbitrarily (by data contamination) in such a way that the new value of $\hat{g}_n^{\frac{1}{2}}$ is close to the old value as measured by the L_2-metric, then $||f_\theta^{\frac{1}{2}} - \hat{g}_n^{\frac{1}{2}}||$ viewed as a function of θ will be uniformly relatively unaffected; hence $\hat{\theta}_n$ will be relatively unchanged too.

The minimum Hellinger distance estimator (MHDE) may be regarded as a particular minimum distance estimator that is distinguished by being asymptotically efficient in regular models. As a class, minimum distance estimators have been studied by Wolfowitz (1952, 1954, 1957), who established their strong consistency under general assumptions and considered the use of the minimized distance in testing goodness-of-fit. In continuous models, asymptotic distributions have been found for a few minimum distance estimators (Blackman (1955), Rao et al. (1975)) and for a special case of the minimized distance (Kac et al. (1955)). However, the MHDE and the corresponding minimized Hellinger distance have apparently been studied only in discrete models (Matusita (1955), Rao (1963)), where there exists a close link with the minimum chi-square methods of Neyman (1949).

The results in this paper are organized as follows. Section 2 establishes some basic continuity and differentiability properties of a functional associated with the estimator $\hat{\theta}_n$. Asymptotic distribution theory for $\hat{\theta}_n$ is developed in Section 3 and the sense in which $\hat{\theta}_n$ is a robust estimator of θ is examined in Section 4.

Section 5 studies the natural goodness-of-fit statistic $\|f^{\frac{1}{2}}_{\hat{\theta}_n} - \hat{g}_n^{\frac{1}{2}}\|^2$ while Section 6 reports the outcome of a modest numerical experiment with $\hat{\theta}_n$.

2. Minimum Hellinger distance functionals.

The approach that will be followed in studying the properties of $\hat{\theta}_n$ is to view it as the value at $\hat{g}_n$ of a functional T. Let $\mathscr{F}$ denote the set of all densities with respect to Lebesgue measure on the real line. The functional T is defined on $\mathscr{F}$ by the requirement that for every $g \in \mathscr{F}$,

$$\|f^{\frac{1}{2}}_{T(g)} - g^{\frac{1}{2}}\| = \min\nolimits_{t \in \Theta} \|f_t^{\frac{1}{2}} - g^{\frac{1}{2}}\|\,. \tag{2.1}$$

Since $T(g)$ may be multiple-valued, we will use the notation $T(g)$ to indicate any one of the possible values, chosen arbitrarily. To ensure existence of $T(g)$, some assumptions are needed on the parametric family $\{f_\theta : \theta \in \Theta\}$. For notational convenience in the remainder of the paper, let $s_t = f_t^{\frac{1}{2}}$.

THEOREM 1. *Suppose that Θ is a compact subset of R^p, $\theta_1 \neq \theta_2$ implies $f_{\theta_1} \neq f_{\theta_2}$ on a set of positive Lebesgue measure, and for almost every x, $f_\theta(x)$ is continuous in θ. Then*

(i) *For every $g \in \mathscr{F}$, there exists $T(g) \in \Theta$ satisfying* (2.1).

(ii) *If $T(g)$ is unique, the functional T is continuous at g in the Hellinger topology.*

(iii) *$T(f_\theta) = \theta$ uniquely for every $\theta \in \Theta$.*

PROOF. (i) Existence. Let $h(t) = \|s_t - g^{\frac{1}{2}}\|$. For any sequence $\{t_n : t_n \in \Theta, t_n \to t\}$,

$$|h^2(t_n) - h^2(t)| = 2|\textstyle\int [s_{t_n}(x) - s_t(x)]g^{\frac{1}{2}}(x)\,dx| \leqq 2\|s_{t_n} - s_t\|\,. \tag{2.2}$$

By Vitali's theorem and the pointwise continuity assumption of this theorem, the right side of (2.2) converges to zero. Hence h is continuous and achieves a minimum for $t \in \Theta$.

(ii) Continuity of T. Suppose $g_n^{\frac{1}{2}} \to g^{\frac{1}{2}}$ in L_2 and put $h_n(t) = \|s_t - g_n^{\frac{1}{2}}\|$. Write $\theta_0 \equiv T(g)$ and $\theta_n \equiv T_n(g_n)$ for convenience (any one of the possible values in the latter case). By Minkowski's inequality

$$\lim\nolimits_{n\to\infty} \sup\nolimits_t |h_n(t) - h(t)| = 0 \tag{2.3}$$

which implies that $|\min_t h_n(t) - \min_t h(t)| \to 0$ or, equivalently, $h_n(\theta_n) \to h(\theta_0)$. Since (2.3) also implies that $|h_n(\theta_n) - h(\theta_n)| \to 0$, we conclude that

$$\lim\nolimits_{n\to\infty} h(\theta_n) = h(\theta_0)\,. \tag{2.4}$$

If $\theta_n \not\to \theta_0$, compactness of Θ ensures existence of a subsequence $\{\theta_m\} \subset \{\theta_n\}$ such that $\theta_m \to \theta_1 \neq \theta_0$, implying $h(\theta_m) \to h(\theta_1)$ by continuity of h. By (2.4), $h(\theta_1) = h(\theta_0)$, which contradicts the assumed uniqueness of $\theta_0 \equiv T(g)$.

(iii) $T(f_\theta) = \theta$ uniquely. This is immediate from the identifiability assumption on the parametrization.

Theorem 1 is also useful for parametric families $\{f_\theta : \theta \in \Theta\}$ where Θ is not compact but can be embedded within a compact set. We illustrate this point for a location-scale family $\{\sigma^{-1}f(\sigma^{-1}(x-\mu)) : \sigma > 0, -\infty < \mu < \infty\}$ where f is continuous. Write $\mu = \tan(t_1)$, $\sigma = \tan(t_2)$, $t = (t_1, t_2)$, $\Theta = (-\pi/2, \pi/2) \times (0, \pi/2)$, and $f_t(x) = [\tan(t_2)]^{-1}f([\tan(t_2)]^{-1}(x - \tan(t_1)))$; the location-scale family can thus be represented as $\{f_t(x) : t \in \Theta\}$. As $t_1 \to \pm\pi/2$ and $t_2 \to 0$ or $\pi/2$, $h(t) = ||f_t^{\frac{1}{2}} - g^{\frac{1}{2}}|| \to 2^{\frac{1}{2}}$ by a simple argument. Therefore h can be extended to a continuous function on $\bar{\Theta} = [-\pi/2, \pi/2] \times [0, \pi/2]$, which is compact, and the extended function achieves a minimum in $\bar{\Theta}$. In fact the minimum must occur in Θ since $0 \leqq h(t) \leqq 2^{\frac{1}{2}}$ for every $t \in \Theta$ and $h(t) \equiv 2^{\frac{1}{2}}$ is impossible. Consequently the conclusions of Theorem 1 remain valid for this location-scale model.

With further assumptions on $s_t = f_t^{\frac{1}{2}}$, the functional T becomes differentiable, a property that is fundamental for further developments in this paper. For specified $t \in \Theta \subset R^p$, we will typically assume that there exist a $p \times 1$ vector $\dot{s}_t(x)$ with components in L_2 and a $p \times p$ matrix $\ddot{s}_t(x)$ with components in L_2 such that for every $p \times 1$ real vector e of unit euclidean length and for every scalar α in a neighborhood of zero,

$$s_{t+\alpha e}(x) = s_t(x) + \alpha e^T \dot{s}_t(x) + \alpha e^T u_\alpha(x) \tag{2.5}$$

$$\dot{s}_{t+\alpha e}(x) = \dot{s}_t(x) + \alpha \ddot{s}_t(x) e + \alpha v_\alpha(x) e\,, \tag{2.6}$$

where $u_\alpha(x)$ is $p \times 1$, $v_\alpha(x)$ is $p \times p$, and the components of u_α and of v_α individually tend to zero in L_2 as $\alpha \to 0$. Some convenient sufficient conditions for (2.5) and (2.6) are established at the end of this section. In the following theorem, $T(g)$ is viewed as a $p \times 1$ vector.

THEOREM 2. *Suppose that* (2.5) *and* (2.6) *hold for every* $t \in \operatorname{int}(\Theta)$, $T(g)$ *exists, is unique, and lies in* $\operatorname{int}(\Theta)$, $\int \ddot{s}_{T(g)}(x) g^{\frac{1}{2}}(x)\,dx$ *is a nonsingular matrix, and the functional T is continuous at g in the Hellinger topology. Then for every sequence of densities* $\{g_n\}$ *converging to g in the Hellinger metric,*

$$\begin{aligned} T(g_n) = T(g) &+ \int \rho_g(x)[g_n^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x)]\,dx \\ &+ a_n \int \dot{s}_{T(g)}(x)[g_n^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x)]\,dx\,, \end{aligned} \tag{2.7}$$

where

$$\rho_g(x) = -[\int \ddot{s}_{T(g)}(x) g^{\frac{1}{2}}(x)\,dx]^{-1}\dot{s}_{T(g)}(x) \tag{2.8}$$

and a_n *is a real* $p \times p$ *matrix which tends to zero as* $n \to \infty$. *In particular, for* $g = f_\theta$,

$$\begin{aligned} \rho_{f_\theta}(x) &= -[\int \ddot{s}_\theta(x) s_\theta(x)\,dx]^{-1}\dot{s}_\theta(x) \\ &= [\int \dot{s}_\theta(x)\dot{s}_\theta^T(x)\,dx]^{-1}\dot{s}_\theta(x)\,. \end{aligned} \tag{2.9}$$

PROOF. As in the previous proof, write $\theta_0 \equiv T(g)$ and $\theta_n \equiv T(g_n)$. Since $\theta_0 \in \operatorname{int}(\Theta)$ maximizes $\int s_t(x) g^{\frac{1}{2}}(x)\,dx$ and since from (2.5),

$$\lim_{\alpha \to 0} \alpha^{-1} \int [s_{t+\alpha e}(x) - s_t(x)] g^{\frac{1}{2}}(x)\,dx = e^T \int \dot{s}_t(x) g^{\frac{1}{2}}(x)\,dx \tag{2.10}$$

for every unit vector e and every $t \in \operatorname{int}(\Theta)$, it follows that $\int \dot{s}_{\theta_0}(x) g^{\frac{1}{2}}(x)\,dx = 0$. A similar conclusion applies to $\dot{s}_{\theta_n}$. Hence, using also (2.6),

$$\begin{aligned} 0 &= \int \dot{s}_{\theta_n}(x) g_n^{\frac{1}{2}}(x)\,dx \\ &= \int [\dot{s}_{\theta_0}(x) + \ddot{s}_{\theta_0}(x)(\theta_n - \theta_0) + v_n(x)(\theta_n - \theta_0)] g_n^{\frac{1}{2}}(x)\,dx\,, \end{aligned} \tag{2.11}$$

where the components of the $p \times p$ matrix $v_n(x)$ converge in L_2 to zero as $n \to \infty$ since $\theta_n \to \theta_0$. Thus, for n sufficiently large,

$$\begin{aligned} \theta_n - \theta_0 &= -[\int (\ddot{s}_{\theta_0}(x) + v_n(x))g_n^{\frac{1}{2}}(x)\,dx]^{-1} \int \dot{s}_{\theta_0}(x)g_n^{\frac{1}{2}}(x)\,dx \\ &= -[\int \ddot{s}_{\theta_0}(x)g^{\frac{1}{2}}(x)\,dx]^{-1} \int \dot{s}_{\theta_0}(x)[g_n^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x)]\,dx \\ &\quad + a_n \int \dot{s}_{\theta_0}(x)[g_n^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x)]\,dx\,, \end{aligned} \tag{2.12}$$

as was to be shown.

When $g = f_\theta$, θ_0 equals θ and (2.10) now shows $\int \dot{s}_\theta(x)s_\theta(x)\,dx = 0$ for every $\theta \in \operatorname{int}(\Theta)$. Thus, for every sufficiently small real $\alpha \neq 0$ and every unit vector e,

$$\begin{aligned} 0 &= \int \alpha^{-1}[\dot{s}_{\theta+\alpha e}(x)s_{\theta+\alpha e}(x) - \dot{s}_\theta(x)s_\theta(x)]\,dx \\ &= \int \alpha^{-1}\{[\dot{s}_{\theta+\alpha e}(x) - \dot{s}_\theta(x)]s_\theta(x) + [s_{\theta+\alpha e}(x) - s_\theta(x)]\dot{s}_{\theta+\alpha e}(x)\}\,dx \\ &= [\int \ddot{s}_\theta(x)s_\theta(x)\,dx + \int \dot{s}_\theta(x)\dot{s}_\theta^T(x)\,dx]e + o(1)\,, \end{aligned} \tag{2.13}$$

which yields (2.9).

Some relatively accessible conditions under which (2.5) and (2.6) hold are provided by the following two lemmas.

LEMMA 1. *Suppose that $s_\theta = f_\theta^{\frac{1}{2}}$ has the following properties*:

(i) *For every $x \notin N$* (a Lebesgue null set) *and for every θ in some neighborhood of t, $s_\theta(x)$ has first partial derivatives $\{\dot{s}_\theta^{(j)}(x);\ 1 \leqq j \leqq p\}$ with respect to θ which are continuous in θ at $\theta = t$.*

(ii) *For every j, $\dot{s}_\theta^{(j)} \in L_2$ and $\|\dot{s}_\theta^{(j)}\|$ is continuous in θ at $\theta = t$.*

Then expansion (2.5) *holds for $\dot{s}_t = (\dot{s}_t^{(1)}, \dot{s}_t^{(2)}, \cdots, \dot{s}_t^{(p)})^T$.*

PROOF. For $x \notin N$, α sufficiently small, arbitrary unit vector $e \in R^p$, and $\dot{s}_t$ defined as above,

$$\alpha^{-1}[s_{t+\alpha e}(x) - s_t(x) - e^T\dot{s}_t(x)] = \alpha^{-1} \int_0^\alpha e^T[\dot{s}_{t+ce}(x) - \dot{s}_t(x)]\,dc\,. \tag{2.14}$$

But for every j,

$$\|\alpha^{-1} \int_0^\alpha [\dot{s}_{t+ce}^{(j)}(x) - \dot{s}_t^{(j)}(x)]\,dc\|^2 \leqq \alpha^{-1} \int_0^\alpha \|\dot{s}_{t+ce}^{(j)}(x) - \dot{s}_t^{(j)}\|^2\,dc\,, \tag{2.15}$$

which converges to zero as $\alpha \to 0$ because as $c \to 0$, $\dot{s}_{t+ce}^{(j)}(x) \to \dot{s}_t^{(j)}(x)$ for $x \notin N$ and $\|\dot{s}_{t+ce}^{(j)}\| \to \|\dot{s}_t^{(j)}\|$. The lemma follows.

A similar argument establishes

LEMMA 2. *Suppose that s_θ has the following properties*:

(i) *For every $x \notin N$* (a Lebesgue null set) *and for every θ in some neighborhood of t, $s_\theta(x)$ has first partial derivatives $\{\dot{s}_\theta^{(j)}(x);\ 1 \leqq j \leqq p\}$ and second partial derivatives $\{\ddot{s}_\theta^{(j,k)}(x);\ 1 \leqq j, k \leqq p\}$ with respect to θ; the latter are continuous in θ at $\theta = t$.*

(ii) *For every (j, k), $\ddot{s}_\theta^{(j,k)} \in L_2$ and $\|\ddot{s}_\theta^{(j,k)}\|$ is continuous at $\theta = t$.*

Then expansion (2.6) *holds for $\dot{s}_t = (\dot{s}_t^{(1)}, \dot{s}_t^{(2)}, \cdots, \dot{s}_t^{(p)})^T$ and $\ddot{s}_t = \{\ddot{s}_t^{(j,k)}\}$.*

3. Asymptotic distributions. The minimum Hellinger distance estimator $\hat{\theta}_n$ is defined as $T(\hat{g}_n)$, where T is the functional studied in the previous section and $\hat{g}_n$ is a suitable density estimator. In this section we examine the large sample behavior of $T(\hat{g}_n)$ when $\hat{g}_n$ is a kernel density estimator

$$\hat{g}_n(x) = (nc_n s_n)^{-1} \textstyle\sum_{i=1}^{n} w[(c_n s_n)^{-1}(x - X_i)] , \tag{3.1}$$

$\{c_n\}$ being a sequence of constants converging to zero at an appropriate rate, $s_n = s_n(X_1, X_2, \cdots, X_n)$ being a robust scale estimator, and w being a smooth density on the real line. For computational reasons it is convenient to consider densities w that have compact support. The observed random variables $\{X_i\}$ are assumed independent identically distributed, the density of each being g. Evidently, $\hat{g}_n$ is a location-scale invariant estimator of g.

THEOREM 3. *Suppose*

(i) *w is absolutely continuous and has compact support; w' is bounded.*
(ii) *g is uniformly continuous.*
(iii) $\lim_{n\to\infty} c_n = 0$, $\lim n^{\frac{1}{2}} c_n = \infty$.
(iv) *As $n \to \infty$, $s_n \to_p s$ a positive finite constant depending on g.*

Then $\|\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}\| \to_p 0$ as $n \to \infty$. If T is a functional continuous at g in the Hellinger metric, then $T(\hat{g}_n) \to_p T(g)$.

PROOF. Let G_n denote the empirical cdf of $(X_1, X_2, \cdots, X_n)$, which are assumed i.i.d. with density g and cdf G. Let

$$\tilde{g}_n(x) = (c_n s_n)^{-1} \int w[(c_n s_n)^{-1}(x - y)]\, dG(y) . \tag{3.2}$$

Integration by parts gives

$$|\hat{g}_n(x) - \tilde{g}_n(x)| \leqq n^{-\frac{1}{2}}(c_n s_n)^{-1} \sup_x |B_n(x)| \cdot \int |w'(x)|\, dx \tag{3.3}$$

where $B_n(x) = n^{\frac{1}{2}}[G_n(x) - G(x)]$. Moreover, if $a > 0$ is such that the interval $[-a, a]$ contains the support of w, then

$$|\tilde{g}_n(x) - g(x)| \leqq \sup_{|t| \leqq a} |g(x - c_n s_n t) - g(x)| . \tag{3.4}$$

From (3.3) and (3.4), there exist versions of the $\{\hat{g}_n\}$, defined on a suitable probability space, such that $\sup_x |\hat{g}_n(x) - g(x)| \to 0$ w.p. 1; hence $P[\lim_{n\to\infty} \hat{g}_n^{\frac{1}{2}}(x) = g^{\frac{1}{2}}(x)$ for all $x] = 1$. Since $\|\hat{g}_n^{\frac{1}{2}}\| = \|g^{\frac{1}{2}}\| = 1$, $\lim_{n\to\infty} \|\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}\| = 0$ w.p. 1 for these versions and the theorem follows.

The next theorem shows, under stronger assumptions, that $T(\hat{g}_n)$ has an asymptotically normal distribution about $T(g)$. We expect that substantially weaker assumptions would suffice, but do not have a proof in that case.

THEOREM 4. *Suppose*

(i) *w is symmetric about* 0 *and has compact support.*
(ii) *w is twice absolutely continuous; w'' is bounded.*
(iii) *T satisfies* (2.7) *and ρ_g has compact support K on which it is continuous.*

(iv) *$g > 0$ on K; g is twice absolutely continuous and g'' is bounded.*

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

(vi) *There exists a positive finite constant s depending on g such that $n^{\frac{1}{2}}(s_n - s)$ is bounded in probability.*

Then the limiting distribution of $n^{\frac{1}{2}}[T(\hat{g}_n) - T(g)]$ under g as $n \to \infty$ is $N(0, 4^{-1} \int \rho_g(x)\rho_g^T(x)\,dx)$. In particular, if $g = f_\theta$, the limiting distribution of $n^{\frac{1}{2}}[T(\hat{g}_n) - \theta]$ is $N(0, 4^{-1}[\int \dot{s}_\theta(x)\dot{s}_\theta^T(x)\,dx]^{-1})$.

PROOF. An argument similar to that for Theorem 3 shows $||\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}|| \to_p 0$. From (2.7),

$$T(\hat{g}_n) = T(g) + \int \rho_g(x)[\hat{g}_n^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x)]\,dx + V_n \int \dot{s}_{T(g)}(x)[\hat{g}_n^{\frac{1}{2}}(x) - \hat{g}^{\frac{1}{2}}(x)]\,dx\,, \tag{3.5}$$

where $V_n \to_p 0$. Hence it suffices to prove that the limiting distribution of $n^{\frac{1}{2}} \int \sigma(x)[\hat{g}_n^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x)]\,dx$, with $\sigma \in L_2$, $\sigma \perp g^{\frac{1}{2}}$ and σ supported on K is $N(0, 4^{-1}||\sigma||^2)$.

For $b \geqq 0$, $a > 0$ we have the algebraic identity

$$b^{\frac{1}{2}} - a^{\frac{1}{2}} = (b-a)/(2a^{\frac{1}{2}}) - (b-a)^2/[2a^{\frac{1}{2}}(b^{\frac{1}{2}} + a^{\frac{1}{2}})^2]\,. \tag{3.6}$$

Thus

$$n^{\frac{1}{2}} \int \sigma(x)[\hat{g}_n^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x)]\,dx = n^{\frac{1}{2}} \int_K \sigma(x)[\hat{g}_n(x) - g(x)]/(2g^{\frac{1}{2}}(x))\,dx + R_n \tag{3.7}$$

where, for $\delta = \min_{x \in K} g(x) > 0$ and for $\tilde{g}_n(x)$ defined by (3.2),

$$\begin{aligned} |R_n| &\leqq n^{\frac{1}{2}} \int |\sigma(x)|[\hat{g}_n(x) - g(x)]^2/(2g^{\frac{3}{2}}(x))\,dx \\ &\leqq \delta^{-\frac{3}{2}}\{n^{\frac{1}{2}} \int |\sigma(x)|[\hat{g}_n(x) - \tilde{g}_n(x)]^2\,dx + n^{\frac{1}{2}} \int |\sigma(x)|[\tilde{g}_n(x) - g(x)]^2\,dx\} \\ &= \delta^{-\frac{3}{2}}\{W_{1n} + W_{2n}\}\,, \quad \text{say.} \end{aligned} \tag{3.8}$$

With $B_n(x) = n^{\frac{1}{2}}[G_n(x) - G(x)]$, the difference $\hat{g}_n(x) - \tilde{g}_n(x)$ can be expressed as the sum of two terms, $T_{1n}(x)$ and $T_{2n}(x)$:

$$\begin{aligned} T_{1n}(x) &= n^{-\frac{1}{2}}(c_n s)^{-1} \int w[(c_n s)^{-1}(x-y)]\,dB_n(y)\,, \\ T_{2n}(x) &= -n^{\frac{1}{2}} \int dB_n(y) \int_{c_n s}^{c_n s_n} t^{-2}\{w[t^{-1}(x-y)] \\ &\qquad + [t^{-1}(x-y)]w'[t^{-1}(x-y)]\}\,dt \\ &= n^{-\frac{1}{2}} \int_{c_n s}^{c_n s_n} t^{-2}\,dt \int B_n(x - tz)[2w'(z) + zw''(z)]\,dz\,. \end{aligned} \tag{3.9}$$

Evidently,

$$\begin{aligned} E[T_{1n}^2(x)] &\leqq (nc_n s)^{-1} \int w^2(z)g(x - c_n sz)\,dz \\ \sup_x |T_{2n}(x)| &= O_p(n^{-1}c_n^{-1})\,, \end{aligned} \tag{3.10}$$

which implies that $W_{1n} \to_p 0$. Since

$$\sup_x |\tilde{g}_n(x) - g(x)| \leqq 2^{-1} c_n^2 s_n^2 \sup_x |g''(x)| \cdot \int x^2 w(x)\,dx\,, \tag{3.11}$$

$W_{2n} \to_p 0$ also; consequently $R_n \to_p 0$.

Let $\psi(x) = \sigma(x)/(2g^{\frac{1}{2}}(x))$ and write

$$\begin{aligned} n^{\frac{1}{2}} \int \psi(x)[\hat{g}_n(x) - g(x)]\, dx &= n^{\frac{1}{2}} \int \psi(x)T_{1n}(x)\, dx + n^{\frac{1}{2}} \int \psi(x)T_{2n}(x)\, dx \\ &\quad + n^{\frac{1}{2}} \int \psi(x)[\tilde{g}_n(x) - g(x)]\, dx \\ &= \textstyle\sum_{i=1}^{3} U_{in}\,, \quad \text{say.} \end{aligned} \tag{3.12}$$

Both U_{2n} and $U_{3n} \to_p 0$ as $n \to \infty$ because of (3.10) and (3.11) respectively. The first integral U_{1n} can be expressed as $\int dB_n(y) \int \psi(y + c_n sz)w(z)\, dz$. Since

$$E[U_{1n} - \int \psi(y)\, dB_n(y)]^2 \leqq \int w(z)\, dz \int [\psi(y + c_n sz) - \psi(y)]^2 g(y)\, dy \tag{3.13}$$

tends to zero as $n \to \infty$, the limiting distribution of U_{1n} is $N(0, 4^{-1}||\sigma||^2)$, from which the theorem follows.

Under the model f_θ, the asymptotic covariance matrix of $n^{\frac{1}{2}}[T(\hat{g}_n) - \theta]$ is $4^{-1}[\int \dot{s}_\theta(x)\dot{s}_\theta^T(x)\, dx]^{-1}$, which is the reciprocal of the Fisher information matrix. In general, $T(\hat{g}_n)$ is a distinguished estimator of $T(g)$ under g wherever the limiting distribution of Theorem 4 is applicable. This statement is clarified by the next theorem, which is an extension of Hájek's (1970) representation theorem for limiting distributions of regular estimators.

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

$$\lim_{n\to\infty} ||n^{\frac{1}{2}}(g_n^{\frac{1}{2}} - g^{\frac{1}{2}}) - \beta|| = 0\,, \tag{3.14}$$

where $\beta \in L_2$ and $g \in \mathscr{F}$. Note that (3.14) 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 functional T which is *regular* at g in the sense that, under every sequence $\{g_n\} \in \mathscr{C}(g)$, 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 on the particular sequence $\{g_n\}$. This assumption excludes superefficient estimators, for which naive asymptotics can be misleading.

THEOREM 5. *Suppose $\hat{T}_n$ is an estimator of T which is regular at g. Then $\mathscr{D}(g)$ can be represented as the convolution of a $N(0, 4^{-1} \int \rho_g(x)\rho_g^T(x)\, dx)$ distribution with $\mathscr{D}_1(g)$, a distribution depending only upon g and $\hat{T}_n$.*

A proof of this result can be had by modifying the argument for Theorem 6 in Beran (1977), which dealt with one dimensional functionals of a related kind. Under the assumptions required for Theorem 4, the estimator $T(\hat{g}_n)$ is regular because, under g,

$$n^{\frac{1}{2}}[T(\hat{g}_n) - T(g)] = \int \rho_g(x)/g^{\frac{1}{2}}(x)\, dB_n(x) + o_p(1) \tag{3.15}$$

and the log-likelihood ratio $L_n = \log\,[\prod_{i=1}^n g_n(X_i)/g(X_i)]$ can be approximated by

$$L_n = 2n^{-\frac{1}{2}} \textstyle\sum_{i=1}^n \beta(X_i)g^{-\frac{1}{2}}(X_i) - 2||\beta||^2 + o_p(1) \tag{3.16}$$

for every $\{g_n\} \in \mathscr{C}(g, \beta)$ (cf. Le Cam (1969) for the essential argument). In particular, (3.16) entails contiguity of $\{g_n\}$ to g. Thus the estimator $T(\hat{g}_n)$ is distinguished by having the least dispersed limiting distribution allowed regular estimators of T by Theorem 5.

4. Robustness properties. Robustness of the estimator $\hat{\theta}_n = T(\hat{g}_n)$ would ideally be studied by considering what happens to the distribution of $T(\hat{g}_n)$ as the distribution of the data is varied. Specifically, as one qualitative criterion of robustness, Hampel (1971) proposed Prohorov continuity of the estimator distribution under Prohorov metric perturbations of the data distribution. Since the exact distribution of $T(\hat{g}_n)$ is not available and the Prohorov topology is too weak in this setting, we will study instead the Hellinger continuity of the approximate distribution for $T(\hat{g}_n)$ suggested by Theorem 4, $N(T(g), (4n)^{-1} \int \rho_g(x)\rho_g{}^T(x)dx)$, as g varies within a small Hellinger neighborhood of f_θ. This approach is not rigorous because uniform convergence to the approximating normal distribution in Hellinger-metric neighborhoods of f_θ has not been established. However, the choice of metric and the restriction to small Hellinger-metric neighborhoods of f_θ can be supported: on the one hand, a variety of plausible data contamination models can be expressed, exactly or approximately, as Hellinger-metric perturbations of the data density; on the other hand, the goodness-of-fit test developed in Section 5 of this paper helps to identify situations where the actual g is far from any of the $\{f_\theta : \theta \in \Theta\}$ in the Hellinger metric.

Let $\Sigma(g) = 4^{-1} \int \rho_g(x)\rho_g{}^T(x)\, dx$ with ρ_g defined by (2.8). Under the assumptions of Theorem 2, $\Sigma(g)$ is positive definite because $\int \ddot{s}_{T(g)}(x)g^{\frac{1}{2}}(x)\, dx$ is negative definite and $\int \dot{s}_t(x)\dot{s}_s{}^T(x)\, dx$ is positive definite for every $t \in \operatorname{int}(\Theta)$. The normal approximation to the distribution of $\hat{\theta}_n = T(\hat{g}_n)$ under g has density

$$(4.1) \qquad \varphi(x; g) = n^{p/2}(2\pi)^{-p/2}|\Sigma(g)|^{-\frac{1}{2}} \exp[-n^{\frac{1}{2}}2^{-1}(x - T(g))^T\Sigma^{-\frac{1}{2}}(g)(x - T(g))]$$

on R^p. Still under the assumptions of Theorem 2, both $T(g)$ and $\Sigma(g)$ are Hellinger continuous at f_θ, which implies pointwise convergence of $\varphi^{\frac{1}{2}}(x; g)$ to $\varphi^{\frac{1}{2}}(x; f_\theta)$ as $g^{\frac{1}{2}} \to f_\theta{}^{\frac{1}{2}}$ in L_2. Since $\|\varphi^{\frac{1}{2}}(\cdot\,; g)\| = \|\varphi^{\frac{1}{2}}(\cdot\,; f_\theta)\| = 1$ as well, the convergence $g^{\frac{1}{2}} \to f_\theta{}^{\frac{1}{2}}$ in L_2 entials $\|\varphi^{\frac{1}{2}}(\cdot\,; g) - \varphi^{\frac{1}{2}}(\cdot\,; f_\theta)\| \to 0$. Thus the normal approximation to the distribution of $\hat{\theta}_n$ is itself Hellinger continuous at f_θ (hence also Prohorov continuous for Hellinger-metric perturbations). This result at least encourages the belief that $\hat{\theta}_n$ is a robust estimator under data contamination which corresponds to a small Hellinger-metric perturbation of some f_θ.

Another way to appreciate the robustness of $\hat{\theta}_n = T(\hat{g}_n)$ is simply to note that a small Hellinger-metric change in $\hat{g}_n$, induced by data recording errors or other mechanisms, will typically induce a correspondingly small change in the value of $T(\hat{g}_n)$, by virtue of the continuity of T.

In an infinitesimal neighborhood of f_θ, the minimum Hellinger distance functional T proves to be optimally insensitive to perturbations of its argument. To make this precise, consider the set of all functionals U defined on $\mathscr{F}$ that have the following two properties for every $\theta \in \operatorname{int}(\Theta)$

$$(4.2) \qquad \begin{aligned} U(f_\theta) &= \theta \\ U(g) - U(f_\theta) &= \int \rho(x)[g^{\frac{1}{2}}(x) - f_\theta{}^{\frac{1}{2}}(x)]\, dx + o(\|g^{\frac{1}{2}} - f_\theta{}^{\frac{1}{2}}\|)\,, \end{aligned}$$

where ρ is a p-dimensional vector whose components belong to L_2 and the remainder term is a p-dimensional vector each of whose components are $o(||g^{\frac{1}{2}} - f_\theta^{\frac{1}{2}}||)$. Orthogonality of $g^{\frac{1}{2}}$ and every component of ρ can be assumed without loss of generality. For if this is not the case, ρ can be replaced by $\tilde{\rho}(x) = \rho(x) - [\int \rho(x) g^{\frac{1}{2}}(x)\, dx] g^{\frac{1}{2}}(x)$ and the difference between $\int \rho(x)[g^{\frac{1}{2}}(x) - f_\theta^{\frac{1}{2}}(x)]\, dx$ and $\int \tilde{\rho}(x)[g^{\frac{1}{2}}(x) - f_\theta^{\frac{1}{2}}(x)]\, dx$, being $O(||g^{\frac{1}{2}} - f_\theta^{\frac{1}{2}}||^2)$ componentwise, can be absorbed into the remainder term in (4.2). Evidently, when Theorem 2 applies, the functional T belongs to this class of functionals.

The first requirement in (4.2) imposes a further constraint on ρ: for every $\theta \in \operatorname{int}(\Theta)$, $\int \rho(x)\dot{s}_\theta^T(x)\, dx = I$, the $p \times p$ identity matrix. Indeed, for every unit vector $e \in R^p$ and for every $\alpha \neq 0$,

$$\begin{aligned} e &= \alpha^{-1}[U(f_{\theta+\alpha e}) - U(f_\theta)] \\ &= \alpha^{-1} \int \rho(x)[s_{\theta+\alpha e}(x) - s_\theta(x)]\, dx + \alpha^{-1} o(||s_{\theta+\alpha e} - s_\theta||) \\ &\rightarrow [\int \rho(x)\dot{s}_\theta^T(x)\, dx] e \qquad \text{as} \quad \alpha \rightarrow 0 \end{aligned} \tag{4.3}$$

provided (2.5) and (2.6) apply.

We pose the question: which functional U within the class just described is least affected by infinitesimal perturbations of f_θ? To answer this, we will examine the behavior of $c^T[U(g) - \theta]$ for every constant vector $c \in R^p$, assuming that g is near f_θ in the Hellinger metric. By projection, $g^{\frac{1}{2}}$ can be represented as follows: $g^{\frac{1}{2}}(x) = \cos(\gamma) f_\theta^{\frac{1}{2}}(x) + \sin(\gamma)\, \delta(x)$ where $\gamma \in [0, \pi/2]$, $||\delta|| = 1$, and $\delta \perp f_\theta^{\frac{1}{2}}$. The second equation in (4.2) can be rewritten as

$$U(g) - \theta = \gamma \int \rho(x)\, \delta(x)\, dx + o(\gamma) \tag{4.4}$$

since the components of ρ are orthogonal to $f_\theta^{\frac{1}{2}}$. For γ small and fixed (which is equivalent to $||g^{\frac{1}{2}} - f_\theta^{\frac{1}{2}}||$ small and fixed), the behavior of $|c^T[U(g) - \theta]|$ is determined primarily by the term $|\int c^T \rho(x)\, \delta(x)\, dx| = L_c(\rho, \delta)$.

The situation here can be described loosely as a zero-sum game between the statistician and nature with payoff functional $L_c(\rho, \delta)$. The statistician attempts to minimize the payoff by choice of U, i.e., by choice of the function ρ subject to the constraints established above: $\rho \in L_2$, $\rho \perp f_\theta^{\frac{1}{2}}$ and $\int \rho(x)\dot{s}_\theta^T(x)\, dx = I$. Nature, pessimistically viewed, seeks to maximize the payoff by choice of perturbation direction δ, subject to the constraint $||\delta|| = 1$. As proved in the next theorem, which extends a result in Beran (1977), this game has a saddle point in pure strategies; moreover, the statistician's minimax strategy does not depend on c and corresponds to use of the functional T (when that belongs to the class of functionals U being considered here).

THEOREM 6. *Suppose that s_θ satisfies* (2.5), $\rho \in L_2$, $\rho \perp s_\theta$, $\int \rho(x)\dot{s}_\theta^T(x)\, dx = I$, $\delta \in L_2$, $\delta \perp s_\theta$, *and* $||\delta|| = 1$. *Then for every* $c \in R^p$,

$$\max_\delta \min_\rho L_c(\rho, \delta) = \min_\rho \max_\delta L_c(\rho, \delta) = L_c(\rho_0, \delta_{0,c}) \tag{4.5}$$

where

$$\begin{aligned} \rho_0 &= [\int \dot{s}_\theta(x)\dot{s}_\theta^T(x)\, dx]^{-1}\dot{s}_\theta \\ \delta_{0,c} &= ||c^T\rho_0||^{-1} c^T\rho_0\,. \end{aligned} \tag{4.6}$$

PROOF. It suffices to show that $\max_\delta \min_\rho L_c(\rho, \delta) \geqq \min_\rho \max_\delta L_c(\rho, \delta)$ since the reverse inequality is trivial. By the Cauchy-Schwarz inequality, $\max_\delta L_c(\rho, \delta) = \|c^T\rho\|$, the maximizing choice of δ being $\|c^T\rho\|^{-1}c^T\rho$. Since every component of ρ can be represented as the sum of a function in the subspace spanned by the components of $\dot{s}_\theta$ and a function orthogonal to that subspace, we can write $\rho = A\dot{s}_\theta + \sigma$, each component of σ being orthogonal to every component of $\dot{s}_\theta$ as well as to s_θ. The constraint $\int \rho(x)\dot{s}_\theta^T(x)\,dx = I$ implies that $A = [\int \dot{s}_\theta(x)\dot{s}_\theta^T(x)\,dx]^{-1}$, so that $\rho = \rho_0 + \sigma$. Hence

$$\min_\rho \max_\delta L_c(\rho, \delta) = \min_\sigma \|c^T\rho_0 + c^T\sigma\| = \|c^T\rho_0\| \,. \tag{4.7}$$

On the other hand, for $\delta_{0,c}$ defined by (4.6),

$$\begin{aligned}\max_\delta \min_\rho L_c(\rho, \delta) &\geqq \min_\rho L_c(\rho, \delta_{0,c}) \\ &= \|c^T\rho_0\|^{-1} \min_\rho c^T[\textstyle\int \rho(x)\rho_0^T(x)\,dx]c = \|c^T\rho_0\| \,,\end{aligned} \tag{4.8}$$

the last inequality using $\rho = \rho_0 + \sigma$. The theorem follows now from (4.7) and (4.8).

Theorem 6 invites more speculation: since the functional T is (typically) locally minimax robust at f_θ in the sense of (4.5) and since $\|\hat{g}_n^{\frac{1}{2}} - f_\theta^{\frac{1}{2}}\| \to_p 0$ under the assumptions of Theorem 3, it is already evident, heuristically, that the estimator $\hat{\theta}_n = T(\hat{g}_n)$ should be asymptotically efficient in some sense under f_θ. In fact, Theorems 4 and 5 showed this to be the case under some additional technical assumptions. What is of particular interest, however, is a philosophical point: local minimax robustness at f_θ entails asymptotic efficiency at f_θ but not conversely (as the method of maximum likelihood demonstrates).

The robustness properties considered so far have been based upon a Hellinger metric model of data contamination. It is of interest also, though less convenient mathematically, to examine the behavior of T under a mixture model for gross errors; the results confirm our belief that T is robust and reveal the limitations of Hampel's (1974) influence curve in assessing robustness.

Let δ_z denote the uniform density on the interval $(z - \varepsilon, z + \varepsilon)$, where $\varepsilon > 0$ is very small, and let $f_{\theta,\alpha,z} = (1 - \alpha)f_\theta + \alpha\delta_z$ for $\theta \in \Theta$, $\alpha \in [0, 1)$, and real z. The density $f_{\theta,\alpha,z}$ models an experiment where independent observations distributed according to f_θ are mixed with approximately $100\alpha\%$ gross errors located near z. The following theorem compares $T(f_{\theta,\alpha,z})$ with $T(f_\theta) = \theta$.

THEOREM 7. *For every $\alpha \in (0, 1)$, every $\theta \in \Theta$, and under the assumptions of Theorem* 1, *$T(f_{\theta,\alpha,z})$ is a continuous bounded function of z such that*

$$\lim_{z\to\infty} T(f_{\theta,\alpha,z}) = \theta \,. \tag{4.9}$$

If $f_\theta(x)$ is a positive density continuous in x and if the conclusions of Theorem 2 *hold for $g = f_\theta$, then*

$$\lim_{\alpha\to 0} \alpha^{-1}[T(f_{\theta,\alpha,z}) - \theta] = \int [2s_\theta(x)]^{-1}\rho_{f_\theta}(x)\delta_z(x)\,dx \tag{4.10}$$

for every real z, with ρ_{f_θ} defined by (2.9).

Proof. We begin with (4.9). For simpler notation, write $\theta_z \equiv T(f_{\theta,\alpha,z})$, $s_{t,\alpha,z} \equiv f_{t,\alpha,z}^{\frac{1}{2}}$, and $s_t \equiv f_t^{\frac{1}{2}}$. Let $m_z(t) = \int s_t(x) s_{\theta,\alpha,z}(x)\, dx$ and let $k_z(t) = \int s_t(x)[(1-\alpha)^{\frac{1}{2}} s_\theta(x) + \alpha^{\frac{1}{2}} \delta_z^{\frac{1}{2}}(x)]\, dx$. By calculation, using Cauchy–Schwarz,

$$\lim_{z\to\infty} \sup_{t\in\Theta} |k_z(t) - m_z(t)| = 0\,. \tag{4.11}$$

Suppose $\theta_z \nrightarrow \theta$ as $z \to \infty$. Without loss of generality, by going to a subsequence if necessary, we may assume that $\lim_{z\to\infty} \theta_z = \theta_1 \neq \theta$. Then, much as in the proof of Theorem 1,

$$\begin{aligned} \lim_{z\to\infty} m_z(\theta_z) = \lim_{z\to\infty} k_z(\theta_z) &= (1-\alpha)^{\frac{1}{2}} \int s_{\theta_1}(x) s_\theta(x)\, dx \\ &< (1-\alpha)^{\frac{1}{2}} = \lim_{z\to\infty} m_z(\theta)\,. \end{aligned} \tag{4.12}$$

On the other hand, since $t = \theta_z$ maximizes $m_z(t)$ over Θ,

$$\lim_{z\to\infty} m_z(\theta_z) \geqq \lim_{z\to\infty} m_z(\theta)\,, \tag{4.13}$$

which contradicts (4.12). Hence $\lim_{z\to\infty} \theta_z = \theta$ as asserted in (4.9).

Continuity of θ_z in z is a consequence of Hellinger continuity of T (Theorem 1). Boundedness of θ_z in z follows with the aid of (4.9).

Clearly $\lim_{\alpha\to 0} \|s_{\theta,\alpha,z} - s_\theta\| = 0$ for every real z. To prove (4.10), therefore, it suffices to show that for every $\sigma \in L_2$, $\sigma \perp s_\theta$,

$$\lim_{\alpha\to 0} \alpha^{-1} \int \sigma(x)[s_{\theta,\alpha,z}(x) - s_\theta(x)]\, dx = \int [2s_\theta(x)]^{-1} \sigma(x) \delta_z(x)\, dx\,. \tag{4.14}$$

By calculation,

$$\lim_{\alpha\to 0} \alpha^{-1} \int_{|x-z|\geqq\varepsilon} \sigma(x)[s_{\theta,\alpha,z}(x) - s_\theta(x)]\, dx = (-2)^{-1} \int_{|x-z|\geqq\varepsilon} \sigma(x) s_\theta(x)\, dx \tag{4.15}$$

and

$$\begin{aligned} &\lim_{\alpha\to 0} \alpha^{-1} \int_{|x-z|<\varepsilon} \sigma(x)[s_{\theta,\alpha,z}(x) - s_\theta(x)]\, dx \\ &\quad = \lim_{\alpha\to 0} \int_{|x-z|<\varepsilon} [s_{\theta,\alpha,z}(x) + s_\theta(x)]^{-1} \sigma(x)[\delta_z(x) - f_\theta(x)]\, dx \\ &\quad = \int_{|x-z|<\varepsilon} [2s_\theta(x)]^{-1} \sigma(x) \delta_z(x)\, dx + (-2)^{-1} \int_{|x-z|<\varepsilon} \sigma(x) s_\theta(x)\, dx\,, \end{aligned} \tag{4.16}$$

which imply (4.14).

The limit evaluated in (4.10), viewed as a function of z, is the influence curve of the functional T at f_θ. Actually, we have modified Hampel's (1974) definition slightly to make it suitable for functionals with domain in $\mathscr{F}$; however, the change is otherwise unimportant. Since $\lim_{x\to\infty} s_\theta^{-1}(x) \rho_{f_\theta}(x)$ need not be finite for many parametric families $\{f_\theta : \theta \in \Theta\}$ (such as the normal location-scale family), the right side of (4.10) can be an unbounded function of z. On the other hand, the first part of Theorem 7 shows: for every $\alpha \in (0, 1)$, the difference quotient (or α-influence curve) $\alpha^{-1}[T(f_{\theta,\alpha,z}) - \theta]$ is a bounded continuous function of z such that $\lim_{z\to\infty} \alpha^{-1}[T(f_{\theta,\alpha,z}) - \theta] = 0$. Hence the functional T is robust at f_θ against $100\alpha\%$ contamination by gross errors at arbitrary real z; whether or not the influence curve of T is bounded is irrelevant to the matter.

In mathematical terms, the convergence of the α-influence curves of T to the influence curve need not be uniform in z, so that the influence curve of T can differ dramatically in shape from each of the α-influence curves. This observation suggests two conclusions. First, to assess the robustness of a functional with respect to the gross-error model, it is necessary to examine the α-influence curves rather than the influence curve, except in those cases where the latter provides a uniform approximation to the former. Secondly, since a functional with well-behaved α-influence curves can have an unbounded influence curve, there is no intrinsic conflict between robustness of an estimator and asymptotic efficiency.

That an estimator with unbounded influence curve can possess some degree of robustness was recognized in Sections 5.3 and 8 of Hampel (1974), an example proposed being the normal scores rank estimator of location in contaminated normal distributions. However, Hampel's Section 8 also observed that the normal scores estimator is not very robust quantitatively; his Section 6 advocated bounding the influence curve and compromising on asymptotic efficiency so as to gain quantitative robustness in parametric estimation. Our results for the MHDE suggest that such an approach may be too pessimistic, since the MHDE is asymptotically efficient at the parametric model and is quantitatively robust in the local minimax sense of Theorem 6. It is noteworthy that, under the mixture model of contamination, the breakdown point of a rank estimator is typically fairly low whereas the breakdown point of a MHDE for location is $\frac{1}{2}$.

5. Goodness-of-fit. Two major aims in estimation for parametric models are

(i) To fit a model which explains the bulk of the data by a procedure which is insensitive to occasional divergent observations and is highly efficient at and near the assumed model.

(ii) To identify as clearly as possible divergent observations for further investigation.

The minimum Hellinger distance estimator studied in this paper provides a method for achieving the first goal provided the specified family $\{f_\theta : \theta \in \Theta\}$ contains a density which is close in the Hellinger metric to the actual data density g. We need a way to check the plausibility of this proviso for given data, a way to modify the parametric family if it does not appear to fit, and a way to identify those possibly interesting observations which are not well explained by the fitted model.

A plot of the residual process $f^{\frac{1}{2}}_{\hat{\theta}_n}(x) - \hat{g}_n^{\frac{1}{2}}(x)$, $\hat{\theta}_n$ being the minimum Hellinger distance estimator, is a useful starting point in considering these questions. For fixed x, $n^{\frac{1}{2}}c_n \to \infty$, $n^{\frac{1}{2}}c_n^{\,2} \to 0$, and some regularity assumptions, the limiting distribution under f_θ of $(nc_n)^{\frac{1}{2}}[f^{\frac{1}{2}}_{\hat{\theta}_n}(x) - \hat{g}_n^{\frac{1}{2}}(x)]$ is $N(0, 4^{-1}||w||^2)$. Since this distribution does not depend on x and since, for n large, the covariance function of the process typically tends to zero rapidly as distance between its arguments increases, it is possible to assess the gross features of the residual plot visually: an occasional sharp peak marks the presence of divergent observations while an underlying trend casts doubt upon the fit of the parametric family $\{f_\theta : \theta \in \Theta\}$ to the data set. The nature of any systematic trend in the residual plot may point to a more appropriate model; however, the goal is not to explain every observation ($\hat{g}_n$ does that already) but rather to find a plausible model that fits the bulk of the data. The suspended or hanging rootogram described by Tukey (1971, Chapter 26) is related to the residual plot $f^{\frac{1}{2}}_{\hat{\theta}_n}(x) - \hat{g}_n^{\frac{1}{2}}(x)$, but differs in several technical respects including the estimator of θ and the density estimator.

For assessing goodness-of-fit more formally, the summary statistic $||f^{\frac{1}{2}}_{\hat{\theta}_n} - \hat{g}_n^{\frac{1}{2}}||^2$ seems particularly apt, since it estimates the Hellinger distance squared between the actual data density g and the nearest density in the parametric family $\{f_\theta : \theta \in \Theta\}$. The magnitude of the latter distance affects the local minimax robustness and asymptotic efficiency properties of $\hat{\theta}_n$ as well as the meaningfulness of fitting an f_θ to the data. It is important to note that the statistic $||f^{\frac{1}{2}}_{\hat{\theta}_n} - \hat{g}_n^{\frac{1}{2}}||^2$ does not respond much to minor failures in fit such as a few outliers (cf. Section 4). This selective insensitivity is precisely what makes the statistic valuable in deciding whether the bulk of the data can be reasonably fitted by an f_θ.

The following theorem and corollary establish an asymptotic approximation to the distribution of the statistic when the observations are i.i.d. with density f_θ. In principle the result can be used to approximate the significance level of an observed value of $||f^{\frac{1}{2}}_{\hat{\theta}_n} - \hat{g}_n^{\frac{1}{2}}||^2$; however, the accuracy of the approximation is unknown. We retain the notation of Section 3.

THEOREM 8. *Suppose*

(i) *w is symmetric about* 0 *and has compact support*;

(ii) *w is twice absolutely continuous*; *w'' is bounded*;

(iii) *g is twice absolutely continuous and g'' is bounded*; *g is supported and positive on a compact interval I*;

(iv) $\lim_{n\to\infty} nc_n^{\frac{7}{2}} = 0$, $\lim_{n\to\infty} nc_n^{3} = \infty$;

(v) *there exists a positive finite constant s depending on g such that $n^{\frac{1}{2}}(s_n - s) = O_p(1)$ under g.*

Let $R_n = \max_{1\leqq i\leqq n} X_i - \min_{1\leqq i\leqq n} X_i$, *let* $\mu_n = 4^{-1}R_n||w||^2$, *and let* $\sigma_n^2 = 8^{-1}c_n R_n||w * w||^2$, *where* $*$ *denotes convolution. Then the limiting distribution of* $\sigma_n^{-1}(nc_n||\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2 - \mu_n)$ *under g as $n \to \infty$ is $N(0, 1)$.*

PROOF. Initially we shall suppose that the scale estimator s_n occurring in $\hat{g}_n$ has been replaced by its stochastic limit s; after establishing the theorem for this case, we will show that the substitution makes no difference asymptotically.

Since g has support I,

$$\begin{aligned} nc_n^{\frac{1}{2}}||\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2 &= -2nc_n^{\frac{1}{2}} \int_I [\hat{g}_n^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x)]g^{\frac{1}{2}}(x)\,dx \\ &= nc_n^{\frac{1}{2}} \int_I \{-g^{-\frac{1}{2}}(x)[\hat{g}_n(x) - g(x)] \\ &\quad + 4^{-1}g^{-\frac{3}{2}}(x)[\hat{g}_n(x) - g(x)]^2 \\ &\quad - 8^{-1}\xi_n^{-\frac{5}{2}}(x)[\hat{g}_n(x) - g(x)]^3\}g^{\frac{1}{2}}(x)\,dx\,, \end{aligned} \tag{5.1}$$

where $\xi_n(x)$ lies between $\hat{g}_n(x)$ and $g(x)$ and is continuous w.p. 1 for n sufficiently large. The first integral in the last expression differs from zero by $O_p(nc_n^{\frac{7}{2}})$. Corollary 1 in Rosenblatt (1975) implies under the present assumptions that the limiting distribution of $c_n^{-\frac{1}{2}}[nc_n \int_I g^{-\frac{1}{2}}(x)[\hat{g}_n(x) - g(x)]^2\,dx - \mu(I)||w||^2]$ is $N(0, 2\mu(I)||w * w||^2$; here $\mu(I)$ denotes the length of the interval I. The third integral in the last expansion (5.1) is bounded in absolute value by a constant multiple of $\sup_x |\hat{g}_n(x) - g(x)|nc_n^{\frac{1}{2}} \int [\hat{g}_n(x) - g(x)]^2\,dx$; hence from (3.3), (3.10) and (3.11), the third integral tends to zero in probability. Therefore, the limiting distribution of $c_n^{-\frac{1}{2}}[nc_n||\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2 - 4^{-1}\mu(I)||w||^2]$ is $N(0, 8^{-1}\mu(I)||w * w||^2)$.

The substitution of R_n for $\mu(I)$ in the centering of $||\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2$ is justified because the difference between R_n and $\mu(I)$ is $O_p(n^{-1})$. Indeed, suppose $I = [a, b]$. Let $U_{(1)} \leqq U_{(2)} \leqq \cdots \leqq U_{(n)}$ denote an ordered sample from the $U(0, 1)$ distribution. Then $b - \max_{1 \leqq i \leqq n} X_i = G^{-1}(1) - G^{-1}(U_{(n)})$, which is bounded by a constant multiple of $1 - U_{(n)} = O_p(n^{-1})$ since $g(x) \geqq \delta > 0$ on I. Similarly $\min_{1 \leqq i \leqq n} X_i - a = O_p(n^{-1})$.

The use of s_n instead of s in defining $\hat{g}_n$ perturbs the value of $\hat{g}_n(x) - g(x)$ by a term whose supremum over x is $O_p(n^{-1}c_n^{-1} + n^{-\frac{1}{2}}c_n^2)$; this emerges from (3.10), (3.11) and assumption (v). The corresponding effect on $nc_n^{\frac{1}{2}} \int [\hat{g}_n(x) - g(x)]^2\,dx$ is therefore negligible asymptotically. It follows from examination of (5.1) that replacing s_n by s does not change the limiting distribution of $nc_n||\hat{g}_n^{\frac{1}{2}} - g^{\frac{1}{2}}||^2$.

COROLLARY. *Suppose the assumptions of Theorem* 8 *are satisfied for* $g = f_\theta$ *and, in addition,*

(vi) $\hat{\theta}_n$ *is an estimator of* θ *such that* $n^{\frac{1}{2}}(\hat{\theta}_n - \theta) = O_p(1)$ *under* f_θ;
(vii) $s_t = f_t^{\frac{1}{2}}$ *satisfies* (2.5) *for* t *in a neighborhood of* θ; $\theta \in \operatorname{int}(\Theta)$.

Then the limiting distribution of $\sigma_n^{-1}[nc_n||\hat{g}_n^{\frac{1}{2}} - f_{\hat{\theta}_n}^{\frac{1}{2}}|| - \mu_n]$ *under* f_θ *as* $n \to \infty$ *is* $N(0, 1)$.

PROOF. It suffices to show that

$$D_n = nc_n^{\frac{1}{2}}(||s_{\hat{\theta}_n} - \hat{g}_n^{\frac{1}{2}}||^2 - ||s_\theta - \hat{g}_n^{\frac{1}{2}}||^2) \to_p 0 \qquad \text{as} \quad n \to \infty\,.$$

From (2.5),

$$nc_n^{\frac{1}{2}}||s_{\hat{\theta}_n} - \hat{g}_n^{\frac{1}{2}}||^2 = nc_n^{\frac{1}{2}}||(s_\theta - \hat{g}_n^{\frac{1}{2}}) + (\hat{\theta}_n - \theta)^T\dot{s}_\theta + (\hat{\theta}_n - \theta)r_n||^2 \tag{5.2}$$

where $||r_n|| = o_p(1)$. Squaring out the right side of (5.2) yields an expression for D_n whose terms are clearly $o_p(1)$ with the possible exception of the cross product

$$nc_n^{\frac{1}{2}}(\hat{\theta}_n - \theta)^T \int \dot{s}_\theta(x)[s_\theta(x) - \hat{g}_n^{\frac{1}{2}}(x)]\,dx\,.$$

However, a series of approximations for the integral like those used in the proof of Theorem 4 ultimately shows that this cross product term also tends to zero in probability.

It should be note that any sequence $\{c_n\}$ which satisfies assumption (v) of Theorem 4 also satisfies assumption (iv) of Theorem 8. Thus, the corollary proved above can be applied to the minimized Hellinger distance statistic.

6. Trial by numbers. For numerical work, it is useful to note that the minimum Hellinger distance estimator $\hat{\theta}$ (we drop the subscript n for notational convenience here) can also be defined as that value (or values) of $t \in \Theta$ which maximizes $\int s_t(x)\hat{g}_n^{\frac{1}{2}}(x)\,dx$, where $s_t = f_t^{\frac{1}{2}}$. If $\hat{\theta}^{(0)}$ denotes a reasonable initial guess at $\hat{\theta}$, Newton's method applied to the problem yields the iterative algorithm

$$\hat{\theta}^{(m+1)} = \hat{\theta}^{(m)} - [\int \ddot{s}_{\hat{\theta}^{(m)}}(x)\hat{g}_n^{\frac{1}{2}}(x)\,dx]^{-1} \int \dot{s}_{\hat{\theta}^{(m)}}(x)\hat{g}_n^{\frac{1}{2}}(x)\,dx\,, \quad m \geqq 0\,, \tag{6.1}$$

for $\dot{s}_t$, $\ddot{s}_t$ defined by (2.5) and (2.6). Choice of a density estimator $\hat{g}_n$ with compact support simplifies numerical approximation of the integrals in (6.1) and is desirable on logical grounds as well.

To check the feasibility and finite sample size behavior of $\hat{\theta}$ and its associated goodness-of-fit statistic $||s_{\hat{\theta}} - \hat{g}_n^{\frac{1}{2}}||^2 = 2 - 2\int s_{\hat{\theta}}(x)\hat{g}_n^{\frac{1}{2}}(x)\,dx$, a modest numerical experiment was performed. A pseudo-random sample of size 40 was drawn by computer from a $N(0, 1)$ distribution. In the first stage of the experiment, a $N(\mu, \sigma^2)$ distribution was fitted to this data for various choices of c_n, the goal being to identify values of c_n suitable for normal samples of size $n = 40$. Initial estimates for μ and σ were $\hat{\mu}^{(0)} = \text{median}\,\{X_i\}$ and $\hat{\sigma}^{(0)} = (.674)^{-1}\,\text{median}\,\{|X_i - \hat{\mu}^{(0)}|\}$. Under normality, both of these statistics are root-n consistent estimators of their respective parameters and both are robust under perturbations of normality. The density estimator $\hat{g}_n(x) = (nc_n s_n)^{-1} \sum_{i=1}^n w[(c_n s_n)^{-1}(x - X_i)]$ was based upon the Epanechnikov kernel $w(x) = .75(1 - x^2)$ for $|x| \leqq 1$, with the scale statistic s_n set equal to $\hat{\sigma}^{(0)}$. The derivatives required in (6.1) are $\dot{s}_\theta(x) = \{\dot{s}_{\mu,\sigma}^{(j)}(x)\}$ and $\ddot{s}_\theta(x) = \{\ddot{s}_{\mu,\sigma}^{(j,k)}(x)\}$ where, writing $z = \sigma^{-1}(x - \mu)$ and $A = 2^{-\frac{5}{4}}\pi^{-\frac{1}{4}}$, we have

$$\begin{aligned} \dot{s}_{\mu,\sigma}^{(1)}(x) &= A\sigma^{-\frac{3}{2}}\, z \exp(-.25z^2) \\ \dot{s}_{\mu,\sigma}^{(2)}(x) &= A\sigma^{-\frac{3}{2}}(-1 + z^2) \exp(-.25z^2) \end{aligned} \tag{6.2}$$

and

$$\begin{aligned} \ddot{s}_{\mu,\sigma}^{(1,1)}(x) &= A\sigma^{-\frac{5}{2}}(-1 + .5z^2) \exp(-25z^2) \\ \ddot{s}_{\mu,\sigma}^{(2,2)}(x) &= A\sigma^{-\frac{5}{2}}(1.5 - 4z^2 + .5z^4) \exp(-.25z^2) \\ \ddot{s}_{\mu,\sigma}^{(1,2)}(x) &= \ddot{s}_{\mu,\sigma}^{(2,1)}(x) = A\sigma^{-\frac{5}{2}}(-2.5z + .5z^3) \exp(-.25z^2)\,. \end{aligned} \tag{6.3}$$

Each numerical integral was evaluated by the trapezoidal rule over a grid of 100 equally spaced points on the support of $\hat{g}_n$. The test integral $\int \hat{g}_n(x)\,dx$ was approximated correctly to three decimal places by this procedure for every value of c_n considered in the first part of this experiment and for every case except one treated in the second part of the experiment.

The 40 realized sample values were:

−.706781, .143266, .123015, −.745385, 2.16105,
.654191, 1.14438, −.118696, .258899, −.154302,
.352057, −1.28269, .885335, 2.51841, −1.09603,
2.04580, .402274, .0431284, −.456585, −2.07226,
−1.64175, −.0192038, 1.70932, .929303, .144781,
−.885728, −.588767, −.169394, .699988, −.162130,
.0621123, .729453, .655040, 1.67987, −.194017,
1.01924, −.927988, −.524994, .133760, −.412047.

TABLE 1

Effects of varying c_n on the MHDE *for the observed normal sample of size* 40

		$c_n = .4$	$c_n = .5$	$c_n = .6$	$c_n = .7$	$c_n = .8$	$c_n = .9$	$c_n = 1.0$
Location estimates	Sample median	.0926	—	—	—	—	—	—
	MHDE of location	.132	.137	.141	.143	.146	.148	.149
	Sample mean	.158	—	—	—	—	—	—
Scale estimates	Rescaled sample median absolute deviation	.909	—	—	—	—	—	—
	MHDE of scale	.962	.977	.992	1.007	1.023	1.039	1.056
	Sample standard deviation	1.012	—	—	—	—	—	—

For every case examined during this numerical trial, the Newton algorithm (6.1) converged in three iterations to at least six significant figures. Examination of the matrix $\int \ddot{s}_{\hat{\theta}}(x)\hat{g}_n^{\frac{1}{2}}(x)\,dx$ showed that a local maximum had been attained. Some of the results are presented in Table 1, to fewer decimals because of the numerical integrations.

As might be expected, increasing the value of c_n spreads out the density estimate $\hat{g}_n$ and therefore increases the minimum Hellinger distance estimate of σ. How should we choose c_n? For appropriate $\{c_n\}$, the minimum Hellinger distance estimator (MHDE) of (μ, σ) is asymptotically equivalent, under normality, to the sample mean and sample standard deviation (cf. proof of Theorem 4). It is reasonable, therefore, to choose c_n so that the corresponding estimates roughly match the classical values in Table 1. On this basis, we selected $c_n = .7$ as roughly suitable for normal samples of size 40.

The second stage of the experiment examined the response of the calibrated MHDE to outliers, essentially by calculating some points on an empirical α-influence curve. Specifically, the observation nearest to zero in the data set listed

TABLE 2

Effects of varying X_{22} on the MHDE *and on the classical estimates*

		Original sample	$X_{22}=1$	$X_{22}=2$	$X_{22}=3$	$X_{22}=4$	$X_{22}=5$	$X_{22}=10$	$X_{22}=15$
Location estimates	MHDE of location ($c_n = .7$)	.143	.173	.191	.218	.194	.156	.150	.151
	Sample mean	.158	.184	.209	.234	.259	.284	.409	.534
Scale estimates	MHDE of scale ($c_n = .7$)	1.007	1.019	1.044	1.091	1.080	1.032	1.020	1.018
	Sample standard deviation	1.012	1.020	1.052	1.106	1.179	1.268	1.855	2.555
Goodness of fit	Fitted squared Hellinger distance	.0176	.0134	.0198	.0219	.0322	.0401	.0418	.0424
	Asymptotic upper .10 critical value for squared Hellinger distance	.0437	.0437	.0437	.0473	.0545	.0616	.0957	.128

above, $X_{22} = -.0192038$, was replaced by a series of increasing positive values, ranging from $X_{22} = 1$ to $X_{22} = 15$. The MHDE for location and scale ($c_n = .7$) was computed in each case, as was the fitted squared Hellinger distance $||s_{\hat\theta} - \hat g_n^{\frac12}||^2$ and the asymptotic .10 critical value (upper tail) provided by the corollary to Theorem 8; for the Epanechnikov kernel, $||w||^2 = \frac{3}{5}$ and $||w * w||^2 = \frac{167}{385}$. The results of the various calculations are reported in Table 2.

For values of X_{22} consistent with the assumption that the entire sample is drawn from a normal distribution, the MHDE follows the classical estimators closely. But for $X_{22} \geqq 4$ in Table 2, the MHDE recognizes X_{22} as a possible outlier and begins to discount it smoothly. For $X_{22} \geqq 10$, the MHDE differs little from what it was at the original value of X_{22}. This behavior is in accordance with the first part of Theorem 7 and with what we might expect from a good robust estimator. It occurs even though, for infinitesimal amounts of gross error, the MHDE has the same unbounded influence curve at the normal distribution as the sample mean and standard deviation. Since only improbable values of X_{22} are discounted, it is likely that the exact efficiency of the MHDE under normality is near that of the classical estimator, at least for sample size 40 or more. (The apparent reversal in the MHDE of location at $X_{22} = 10$ and 15 may be caused by insufficient accuracy in the numerical integrations when $X_{22} = 15$).

The .10 upper critical values calculated from the asymptotic distribution of $||s_{\hat\theta} - \hat g_n^{\frac12}||^2$ are all substantially larger than the corresponding observed values of the statistic, suggesting that the fitted normal distribution is not unreasonable in each case. This too is as it should be, since changing one observation out of 40 does not affect the bulk of the sample. The residual plot $f_{\hat\theta}^{\frac12}(x) - \hat g_n^{\frac12}(x)$ would suggest that the larger values of X_{22} are not consistent with the rest of the sample under a normal model.

REFERENCES

BERAN, R. J. (1977). Robust location estimates. *Ann. Statist.* **5** 431-444.

BLACKMAN, J. (1955). On the approximation of a distribution function by an empiric distribution. *Ann. Math. Statist.* **26** 256-267.

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

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

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

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

KAC, M., KIEFER, J. and WOLFOWITZ, J. (1955). On tests of normality and other tests of goodness-of-fit based an distance methods. *Ann. Math. Statist.* **26** 189-211.

LE CAM, L. (1969). *Théorie Asymptotique de la Décision Statistique.* Les Presses de l'Université de Montréal.

MATUSITA, K. (1955). Decision rules on the distance, for problems of fit, two-samples, and estimation. *Ann. Math. Statist.* **26** 631-640.

NEYMAN, J. (1949). Contributions to the theory of the χ^2 test. *Proc. First Berkeley Symp. Math. Statist. Prob.* 239-273, Univ. of California Press.

RAO, C. R. (1963). Criteria of estimation in large samples. *Sankhyā* **25** 189-206.

RAO, P. V., SCHUSTER, E. F. and LITTEL, R. C. (1975). Estimation of shift and center of symmetry based on Kolmogorov-Smirnov statistics. *Ann. Statist.* **3** 862-873.

ROSENBLATT, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3** 1-14.

TUKEY, J. W. (1971). *Exploratory Data Analysis*. Preliminary edition. Addison-Wesley, Reading, Mass.

WOLFOWITZ, J. (1952). Consistent estimation of the parameter of a linear structural relation. *Skand. Aktuarietidskr.* **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.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720