Z. Wahrscheinlichkeitstheorie verw. Gebiete
55, 73–89 (1981)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete
© Springer-Verlag 1981

# Robust Estimation Via Minimum Distance Methods

P. Warwick Millar*
Statistics Department, University of California, Berkeley, CA 94720, USA

**Summary.** Given a fixed parametric family $\{P_\theta\}$ it is desired to estimate $\theta$. However, due to contaminations of various sorts, the data actually collected by the statistician follow a distribution that is close to, but possibly distinct from, the $P_\theta$'s. It is proved that, under these conditions and in an appropriate asymptotic framework, the minimum distance estimators of the Cramér-von Mises type are robust. Specification of a Cramér-von Mises weight function $H$ defines a notion of distance; each such choice of $H$ then delineates the kind of contamination possible, and leads to an estimator which defends optimally against it. When the theory is specialized to location models, various choices of $H$ lead to estimators asymptotically equivalent to such familiar ones as trimmed mean, median, Hodges-Lehmann estimator, and so forth. The framework developed herein provides some guidelines for choosing among the possible estimators, and suggests that the standard Cramér-von Mises estimator of location is probably as good a robust estimator as any.

## 1. Introduction

Let $\{P_\theta, \theta \in \Theta\}$ be a parametric family of probabilities on the line, indexed by $\Theta$, an open subset of $R^d$. Let $\hat{F}_n$ be the sample distribution obtained from $n$ identically distributed observations. Let $d(\cdot, \cdot)$ be some "distance" or "discrepancy" defined on distribution functions; for example, $d(F, G) = \|F-G\|$ where $\|\cdot\|$ is a norm on the collection of right continuous functions, or $d(F, G) = \int (F-G)^2 \, dG$, the Cramér-von Mises "distance". Let $\pi_0 \hat{F}_n$ be an element of $\{P_\theta\}$ which is "closest" to $\hat{F}_n$; i.e., an element of $\{P_\theta\}$ that achieves $\inf_{P_\theta} d(\hat{F}_n, P_\theta)$. Assuming for the purposes of this introduction that questions of existence and uniqueness have been settled, call $\pi_0 \hat{F}_n$ a minimum distance estimator. If selection of $P_{\theta_0} \in \{P_\theta\}$ is tantamount to selecting $\theta_0$, then evidently $\pi_0 \hat{F}_n$ can be regarded as an estimator of the parameter $\theta$. The main result of this paper is that certain minimum distance estimators are robust, in the sense made precise in Sect. 2.

* Supported by NSF grant MCS 75-10376

0044-3719/81/0055/0073/$03.40

The basic thought behind the minimum distance concept has no doubt been implicit in many statistical investigations from the very start; an impressive demonstration of its power and utility occurred, for example, in the pioneering work of J. Neyman on minimum chisquare (Neyman, 1949). The idea was taken up by Wolfowitz who, in a series of papers (1953, 1957 for example), elevated it explicitly to a general principle. Since then, such estimators have been studied by many authors. It has long been part of the statistical folklore that such estimators should have good robustness properties. For example, Knüsel, 1969, argued for the robustness of certain such estimators in some location models; unfortunately, the notion of robustness he used is decidedly unsatisfactory. Some very interesting and suggestive remarks on the relationship between minimum distance and robustness were made by Holm (discussion of Bickel, 1976). A more penetrating study was undertaken by Beran, 1977a, who argued that minimum Hellinger distance estimators should be robust; however, as acknowledged by that author, the robustness properties were only incompletely established - indeed, it was Beran who first called my attention to the fact that problem of the robustness of minimum distance estimators was still open.

Recently, Parr and Schucany, 1979, undertook Monte Carlo studies of several such estimators and found empirically that, in certain location models, they enjoy robustness properties that compare favorably with those of currently fashionable robust procedures. Apparently it is even an empirical fact that minimum distance estimators are robust. A main contribution of this paper is the construction of a reasonable mathematical framework in which this empirical phenomenon becomes understandable. As such, this paper appears to be the first to provide general theoretical justification for the widespread beliefs in the robustness of minimum distance estimators.

The results of this paper are derived within a clearly articulated decision theoretic framework. This is explained rigorously in Sect. 2, but since this point of view appears to be somewhat at variance with the current mode in robustness investigations, it is perhaps best to give an heuristic account of it here. The family $\{P_\theta, \theta \in \Theta\}$ is to be regarded as a theoretically correct probabilistic model of the phenomenon at hand; one could imagine it perhaps forced on us by accepted principles of theoretical physics. However, due to noise, roundoff errors, incompetence, or other contaminations, the random variables actually presented to the statistician follow a distribution $G$ which may be distinct from any of the $P_\theta$'s. If the incompetence has not been too great, presumably $G$ will lie "close" to some $P_\theta$.

If so, the basic task is then to discover what $\theta$ is, even though the true distribution of the data follows some distribution $G$ different from any $P_\theta$. Since a large number of probability distributions are involved in this framework, it is most natural to parametrize by the measures themselves. In decision theoretic terms, the suggested approach to robust estimation is to adopt as parameter set some (fairly large) collection of distributions $G$, and as decision space the collection of measures $\{P_\theta, \theta \in \Theta\}$; loss will then be expressed essentially in terms of some "distance" between $G$ and $P_\theta$, if $G$ is the 'true' distribution of the data given the statistician, and $P_\theta$ is selected from the decision space. The loss functions will be selected in such a way that a) if $G$ is very close to $\{P_\theta\}$, and an inept choice of

$P_\theta$ is made, then the statistician is penalized; b) if $G$ is relatively far from $\{P_\theta\}$, then no matter what choice of $P_\theta$ is made, there will still be some penalty – this is a reasonable property since one has no right to any confidence in the decision made if the data has been badly contaminated.

Technically, robustness is given a 'local asymptotic' definition. Let $H$ be a measure on the line. At time $n$, the observations are $n$ i.i.d. random variables from a distribution $G$ satisfying $d(G, P_{\theta_0}) \leq cn^{-\frac{1}{2}}$, $\theta_0 \in \Theta$ (here $d$ is $L^2(H)$ distance between cdf's – see Sect. 2). The set of such $G$ delineates the possible contamination at time $n$, and forms the enlarged parameter set described in the preceding paragraph. If $P_\theta$ is then selected from the decision space, loss will essentially be a function of $n^{\frac{1}{2}} d(G, P_\theta)$ – see Sect. 2 for the precise description. Any estimator that is asymptotically minimax in this decision theoretic framework is then called robust. In short, the essential feature of robustness – namely, reasonable behavior of the estimator over a neighborhood of the model – is to be automatically guaranteed by the choice of decision theoretic optimality criteria, and not by the more common criteria of 'continuity' or 'differentiability' (cf., Hampel, 1971; Beran, 1977b).

Studies of robustness via local neighborhoods shrinking at rate $n^{-\frac{1}{2}}$ have been undertaken by many authors (e.g., Beran, 1979; Huber-Carol, 1970; Jaeckel, 1971; Rieder, 1978; Bickel, 1979); see the survey by Bickel, 1979, for a description of its advantages, statistical significance, and for detailed discussion of references. The usual shrinking neighborhoods have heretofore employed spheres in the $L_1$, Hellinger, Kolmogorov-Smirnov (K-S), or Prohorov metrics on measures. We introduce here for the first time neighborhoods based on the Hilbertian $L^2(H)$ metric on cdf's. This technical innovation has several advantages. First, the possibility of varying the measure $H$ gives the present set-up a useful flexibility. Because the basic notion of distance is Hilbertian, the resulting theory is fairly simple; $L_1$ and K-S neighborhood systems are more difficult to deal with. In particular, the minimum distance theory developed in our framework will be an asymptotically normal one- unlike that based on the K-S metric. Finally, the contamination neighborhoods here (like K-S) are quite large, so optimal procedures will deal effectively with severe contamination. Unlike $L_1$ and Hellinger systems, ours encompasses contamination that is singular with respect to the model. Because these neighborhoods are so broad, a rationale for restriction to $n^{-\frac{1}{2}}$ balls can be advanced: namely, one can first assess the plausibility of the model via a Kolmogorov-Smirnov test; should the model be accepted, restriction to $n^{-\frac{1}{2}}$ neighborhoods makes sense since the K-S balls of given radius are contained in the $L^2(H)$ balls. No such interpretation is possible if $L_1$ or Hellinger distances between measures are used to construct the contamination neighborhoods (cf. Bickel, 1979).

The decision theoretic point of view advanced above is, of course, not new; it can be discerned in Beran, 1979, and in Holm (op. cit.), and it is implicit in a large number of other papers. We differ here mainly in the militancy with which we insist on its explicitness. Reparametrization of the problem in terms of the measures themselves has ample historical precedent; it occurs with especial force in the work of LeCam, 1973.

The decision theoretic framework just described does not include all possible notions of robustness, but it does cover a fair number of situations arising in practice. Besides, it has several advantages. First, it is not limited to locationscale models: fairly general models $\{P_\theta\}$ can be employed. For example, Sect. 3 applies the theory to families differentiable in quadratic mean, to exponential families, and others. Second, the kind of 'contamination' is essentially arbitrary. In this framework, the observed distributions can actually be singular with respect to the $P_\theta$'s; this allows, without strained artifices, contaminations due to discretizations of the data (roundoff, for example). Third, no restrictions are made a priori on the class of estimators to be considered; they are not required to be solutions of favored functional equations, nor are they required to have a preferred asymptotic expansion; they are not even required to be asymptotically normal. The local minimax optimality criterion employed here apparently cuts out unreasonable competitors, as it does in purely parametric estimation problems. Fourth, this particular framework sidesteps the debilitating relativism involved in the frequent suggestion (Huber, 1972; Bickel-Lehmann, 1975; Hampel, 1974) that the true parameter of interest is whatever one's favorite estimator estimates. Such relativism arose historically as an attempt to cope with the immense conceptual difficulties in understanding what 'center of symmetry' should mean when a symmetric location model is subject to asymmetric contamination. The present framework does not depend on the existence of symmetries, and so here the difficulty does not arise. Fifth, the abstract setup here excludes from the select circle of 'robust estimators' those estimators that have long incurred opprobrium. To illustrate, Example 3A in Sect. 3 presents a simple result to the effect that the notorious 'sample mean' is essentially never robust in the present framework. In short, elevated status is denied to known troublemakers.

Finally, and perhaps most important, the theory developed here appears to provide a coherent rationale that explains the robustness of a number of very popular procedures that have proved reliable in practice. Each choice of the measure $H$ leads, according to the basic theory, to a minimum distance estimator which depends on $H$. Each such estimator will be robust according to the local asymptotic definition of this paper. So in this way one arrives at a whole family of robust minimum distance estimators. In pure location models, various choices of $H$ (see Sect. 3 for the details) lead, under certain conditions, to such familiar estimators as median, Hodges-Lehmann estimator, trimmed mean, and so forth. That is, a number of estimators that have been proposed on what seem to be ad hoc grounds, but which nevertheless have demonstrated empirical robustness, fit neatly into one general framework. Of course, some choices of $H$ are more reasonable than others, and the general development here provides some guidelines. For example, choice of $H$ also determines the size of the possible $n^{-\frac{1}{2}}$ contamination neighborhoods, and the larger the neighborhood, the better the estimator can handle contamination. Finite $H$ yield very full neighborhoods, infinite $H$ result in smaller neighborhoods. The Hodges-Lehmann estimator in location models results from an infinite $H$, and consequently would perhaps be downrated slightly in favor of, say, a standard Cramér-von Mises estimator. These matters are discussed at length in Sect. 3, Subsect. (B). It is our belief that the structure developed in this paper provides helpful conceptual guidelines for picking one's way through the heap of ad hoc procedures, guidelines that permit intelligent assessments as to the viable possibilities before plunging into the Monte Carlo simulations. At the very least, there is the guarantee that the estimator selected fulfills minimal theoretical prerequisites.

The organization of the paper is as follows. Section 2 contains the technical statement of the main results, and the definition of robustness used throughout. Section 3 analyzes a number of common examples, including location-scale models and exponential families, and explains how the basic theory provides sensible robust estimators. Section 4 contains the proofs, which are extensions of those of Millar, 1979, involving tools of abstract decision theory, Hilbert space parametrizations, and concomitant notions from the theory of abstract Wiener spaces. Section 5 presents a few extensions of the main theory.

## 2. Statement of Main Results

Let $\Theta$ be an open subset of $R^d$. For $\theta \in \Theta$, let $P_\theta$ be a probability on the line, absolutely continuous with respect to a sigma finite measure $J$. Let $F_\theta$ be the cumulative distribution function (cdf) of $P_\theta$. Single out a distinguished point $\theta_0 \in \Theta$; for simplicity, take $\theta_0 = 0$. Let $H$ be a finite measure on $R^1$, and define $|\ |_H, \langle,\rangle_H$ to be norm and inner product of $L^2(H)$, while $|\ |$, and $\langle,\rangle$, shall be ordinary Euclidean norm and inner product for $R^d$.

The parametric family $\{P_\theta\}$ and the measure $H$ shall satisfy the following assumptions:

(2.1) Differentiability: there exists a function $\xi$ on $R^1$ with values in $R^d$ such that $\langle \theta, \xi \rangle \in L^2(H)$ for all $\theta$ and

$$|F_\theta - F_0 - \langle \theta, \xi \rangle|_H = o(|\theta|) \qquad \text{as } \theta \text{ goes to } 0$$

(2.2) Identifiability: if $|F_{\theta_n} - F_\theta|_H \to 0$ for some sequence $\theta_n$, then $\theta_n \to \theta$.

Condition (2.1) is Frêchet differentiability of $F_\theta$ at $F_0$ as elements of $L^2(H)$; the derivative $\xi$ of course depends on $\theta_0$ and $H$. Although the analysis of this paper is local, proper interpretation of the results requires these hypotheses to hold at each $\theta_0 \in \Theta$.

Differentiability conditions on parametric families appear in many asymptotic studies. In classical parametric theory (LeCam, 1969) as well as in modern robustness studies (Beran, Bickel, Rieder, op. cit.), the usual condition is quadratic mean differentiability. The differentiability condition (2.1) on cdf's is weaker than q.m. differentiability and applies to more parametric families (cf., Sect. 3E, 3F). Systematic use of it in robustness theory appears here for the first time.

Define $\sum = \sum_{\theta_0} = \{\langle \theta, \xi \rangle : \theta \in R^d\}$; assume that

(2.3) dimension $\sum$ in $L^2(H)$ is $d$.

For *a* cdf $G$, define $\pi_0 G$ to be any cdf $F_{\theta'}, \theta' \in \Theta$ which satisfies $\inf_\theta |G - F_\theta|_H = |G - F_{\theta'}|_H$. Define $\pi$ to be orthogonal projection in $L^2(H)$ to the subspace $\sum$.

Let $N(c) = \{q \in L^2(H) : |q|_H \leqq c\}$, and let $G_{nq}$ be the cdf defined by $G_{nq} = F_0 + n^{-\frac{1}{2}} q$, $q \in N(c)$ (assuming $q$ chosen so that this is indeed a cdf). Let $G^n_{nq}(dx)$ be the product measure $\prod_i G_{nq}(dx_i)$, $x=(x_1,\ldots,x_n)$. The measures $G^n_{nq}$, $q\in N(c)$, give the possible distributions of the contaminated data at stage $n$; that is, the data consists of $n$ iid observations from some cdf lying in a ball of radius $cn^{-\frac{1}{2}}$ of $F_0$ in the metric $|\ |_H$. The collection of $G_{nq}$, $q\in N(c)$ is called a contamination neighborhood of $F_0$. Different choices of $H$ give different contamination neighborhoods; if $H$ is finite, however, each such neighborhood contains all distributions $G$ such that $|G-F_0|_K \equiv \sup_t |G(t)-F_0(t)| \leqq cn^{-\frac{1}{2}}H(R^d)$ - i.e., contains a $n^{-\frac{1}{2}}$ ball in the Kolmogorov distance, a fact whose significance is discussed later in this section.

Finally, a loss function will be introduced which assesses penalty when $G^n_{nq}$ is the distribution of the data, and $\theta$ is selected from the decision space. Because of (2.2), knowledge of $F_\theta$ determines $\theta$ so it is reasonable to reparametrize, taking the cdf's $F_\theta$ as the basic objects to be estimated. In this situation, a natural candidate for loss might be an increasing function of $n|G_{nq}-F_\theta|^2_H$; however, simple considerations show the problem degenerate for this choice. Instead, introduce the closely related loss $L_n(G_{nq}, F_\theta)$:

$$L_n(G_{nq}, F_\theta)=l[n|\pi_0 G_{nq}-F_\theta|^2_H+(n|G_{nq}-\pi_0 G_{nq}|^2_H)\wedge a]$$

where $l$ is an increasing function on $[0,\infty)$. This loss evidently has the two properties mentioned in the introduction: it penalizes both for bad data (but only boundedly), and also for inept choice of $\theta$ when $G_{nq}$ is close to the parametric family. Such loss functions, of course, are not 'philosophically necessary' (as one referee complained); I confess never having met any philosophically necessary loss function. The point is, however, that the losses introduced here are intuitively reasonable, lead to a simple theory, and the procedures derived within the resultant decision theoretic structure accord well with those that have been found best in practice.

Let $W$ be the usual Brownian bridge stochastic process. For theorem (2.4) below, the estimators $T$ in question are estimators of the cdf's $F_\theta$, according to the reparametrization discussed in the preceding paragraph.

(2.4) **Theorem.** *Under hypotheses* (2.1)–(2.3) *and the technical hypothesis* (4.4), *if* $c_n\uparrow\infty$, *then*

$$\liminf_n \inf_T \sup_{q\in N(c_n)} \int L_n(G_{nq}, T)\, dG^n_{nq} \geqq E\,l[|\pi W\circ F_0|^2_H+a].$$

(2.5) *Definition.* Any sequence of estimators $T_n$ for which the limiting minimax risk, $\lim_n \sup_{q\in N(c_n)} \int L_n(G_{nq}; T_n)\, dG^n_{nq}$, equals the lower bound of (2.4) will be called $H$-robust. If $H$ is finite, then the contamination neighborhoods contain a ball in the Kolmogorov metric, and in this case call the estimators $T_n$ fully robust. This terminology is reasonable because a preliminary Kolmogorov-Smirnov test can then assess the plausibility of the model; if the model be accepted, then restriction to our Hilbertian $n^{-\frac{1}{2}}$ neighborhoods is reasonable. As mentioned in the introduction, no such interpretation is possible if $L^1$, Hellinger, or $L^2(H)$ ($H$ only sigma finite) distances are used.

Define $\hat{F}_n$ to be the empirical distribution function. Assume that $l$ satisfies the growth condition

(2.6) $y^{-2}\log l(y)=o(1)$ as $y\to\infty$.

(2.7) **Theorem.** *Under assumptions* (2.1)–(2.3), (2.6), *there exist* $c_n \uparrow \infty$ *such that*

$$\lim_n \sup_{q\in N(c_n)} \int L_n(G_{nq}; \pi_0 \hat{F}_n)\, dG^n_{nq} = E\, l[|\pi W \circ F_0|^2_H + a]$$

*Moreover, relative to the metric* $|\ |_H$:

$$\pi_0 \hat{F}_n = F_0 + \pi(\hat{F}_n - F_0) + o(n^{-\frac{1}{2}}) \qquad \textit{under } \{G^n_{nq}\}$$

In short, the minimum distance estimator $\pi_0 \hat{F}_n$ is robust in the sense of Definition (2.6).

**Subsection 2A: Local Structure.** The estimators discussed in Theorems (2.4), (2.7) are estimators of cdf's and only indirectly are they estimators of $\theta$. It is important to realize that locally the manner in which $\theta$ is actually selected by the estimator $\pi_0 \hat{F}_n$ is quite transparent. By (2.7), $\pi_0 \hat{F}_n$ is locally like orthogonal projection of $\hat{F}_n$ (in $L^2(H)$) to $F_0 + \sum$, so it estimates $\theta$ by picking an element of the form $F_0 + \langle \hat{\theta}_n, \xi \rangle$, where the estimator $\hat{\theta}_n$ is easily determined from well known properties of orthogonal projection in Hilbert space. For example if $\sum$ is one dimensional, for example, $\langle \hat{\theta}_n, \xi \rangle$ is $\hat{\theta}_n \xi$, and so by the usual characterization of projection of projection in Hilbert space

(2.8) $\hat{\theta}_n = \langle \hat{F}_n - F_0, \xi \rangle_H / |\xi|^2_H$

The asymptotic normality of the estimate $\hat{\theta}_n$ of $\theta$ is hence immediate, as it is in the $d$-dimensional case as well. Calculation of (2.8) in special cases can be found in Sect. 3.

Asymptotically, the loss at time $n$, when $G_{nq}$ is true and $P_\theta$ guessed, is an increasing function of $|\pi q - \langle \theta, \xi \rangle|^2_H + |(1-\pi)\, q|^2_H \wedge a$. But $\pi q = \langle \theta_q, \xi \rangle$ for some $\theta_q \in R^d$, so this becomes $|\langle \theta_q - \theta, \xi \rangle|^2_H + |(1-\pi)\, q|^2_H \wedge a$. If there is no data contamination, then the second term in the last expression is zero, so the loss is essentially a squared error loss between the 'true' $\theta_q$ and our guess for it. In general it is clear that, even with data contamination, the loss still preserves its character as locally a kind of squared error loss.

**Subsection 2*B*: Extensions.** Proper understanding of the examples of Sect. 3 requires two extensions of the main result. Other extensions are mentioned in Sect. 5.

*(a) Sigma finite H.* Suppose that $H$ is a sigma finite measure satisfying

(2.9) $\int F_0(t)\, [1 - F_0(t)]\, H(dt) < \infty$

in addition to (2.1) – (2.3). If $l$ is bounded (say) then Theorems (2.4), (2.7) continue to hold, provided the sets $N(c)$ are narrowed to (e.g.) $\{q\colon |q|_H \leqq c, \int |q|\, dH \leqq c\}$. Notice that when $H$ is infinite, the sets $N(c)$ will typically be quite skimpy – they will not, for example, contain K-S balls as they do when $H$ is finite. Therefore, though the estimator $\pi_0 \hat{F}_n$ will be $H$-robust according to Definition (2.5), it is clear by looking at $N(c)$ that these estimators can not defend against heavy contamination as well as those coming from finite $H$. See the examples of Sect. 3 for illustrations of this point.

*(b) Families of Measures.* For each $\theta \in \Theta$, let $H_\theta$ be a measure on the line; denote by $| \ |_\theta$ the $L^2(H_\theta)$ norm. Amend the definition of $\pi_0 G$ to mean any element $F_\theta$ which achieves $\inf_\theta |G - F_\theta|_\theta$. The choice $H_\theta = F_\theta$ makes $\pi_0 \hat{F}_n$ the standard Cramér-von Mises estimator of $F_\theta$. Under reasonable smoothness assumptions on $\{H_\theta\}$, the conclusions of Theorems (2.4), (2.7) still hold. Here is one workable set of such assumptions: (i) each $H_\theta$ is a probability (ii) $\theta \to H_\theta$ is continuous in the usual weak* sense (iii) each $F_\theta$ is continuous (iv) if $|F_{\theta_n} - F_{\theta_0}|_{\theta_n} \to 0$, then $\theta_n \to \theta_0$. (v) differentiability assumption (2.1) holds, but with $| \ |_K$ replacing $| \ |_H$ (iv) dimension $\sum$ is $d$ in $L^2(H_\theta)$, all $\theta$. We shall omit the proof; discussion, contemporaneous with the present effort, of the asymptotic normality of $\pi_0 \hat{F}_n$ under various assumptions may be found in Parr and de Wet, 1979, and in Pollack, 1979. After preliminary global considerations, the main thrust of these discussions is to replace the norms $| \ |_\theta$ by $| \ |_{\theta_0}$, and so reduce the problem locally to the case of just one measure, $H_{\theta_0}$.

## 3. Examples

The theory of this paper functions more or less like a scientific theory of empirical robustness: it predicts in known cases what experience acknowledges to be the case, and in unknown cases suggests sensible estimators whose empirical performance can be checked. To illustrate this we first check (Subsect. 3A) that the sample mean is never fully robust in the sense of Definition (2.5); examination of the argument shows that the theory here reproduces the well known empirical fact that the sample mean cannot defend against heavy tailed contamination. Next, in Subsect. 3B, 3C, 3D, location models, scale models, and locationscale models are studied. Simple choices of bounded $H$ lead quickly to procedures asymptotically equivalent to empirically reliable procedures. On the other hand, it is shown that there is an $H$ (not bounded) for which the corresponding minimum distance estimator is asymptotically equivalent to the sample mean: the contamination neighborhoods in this case are extremely thin, of course, but nevertheless the theory predicts that the sample mean can cope with certain limited kinds of contamination. This seems to be borne out in empirical investigations of the behaviour of the sample mean vis a vis fairly accurate (but discernibly non-normal) data from the physical sciences. (cf., Stigler, 1977). To illustrate the breadth of the theory, Subsect. (3F) shows that it applies to general quadratic mean families. These include, in particular, exponential families whose importance in modelling in applied statistics is well known. Subsection 3G applies the theory to the uniform $[0, \theta]$ families. This perhaps has amusement value: the classical parametric analysis of $U(0, \theta)$ leads to an (annoying, actually) non-normal theory; the present framework, involving contamination, forces it back into the normal fold. Seasoned practitioners could no doubt have guessed this, but it is gratifying that the theory does it so effortlessly.

**(3A) Non-robust Estimators.** This subsection demonstrates that the sample mean is never *fully* robust in the theory of this paper. Similar analyses easily show, for example, that the sample variance cannot be robust either.

To see that the sample mean is not robust, suppose given a family $\{P_\theta\}$ with $\theta$ real. We shall localize about $\theta=0$. Suppose that the characteristic function $\varphi$ of $P_0$ satisfies $\varphi(t)=1+\varphi'(0)\,t+o(t), t\to 0$. Assume there exist $r>0, s>0$, such that $|P_\theta-P_0|_H^2>r$ whenever $|\theta|>s$. The standard real normal shift family satisfies these conditions for example. Let $C_n$ be the Cauchy distribution with ch.f. $\exp\{-n|t|\}$. Set $G_n=(1-n^{-\frac{1}{2}})\,F_0+n^{-\frac{1}{2}}C_n$. If $H$ is assumed bounded, then $G_n$ is in the contamination neighborhood. If $\varphi_n(t)$ is the ch.f. of the sample mean under $G_n$, then $\varphi_n$ is easily computed and $\lim\limits_n \varphi_n(t)=0$ for every $t \neq 0$. It then follows from standard identities involving ch.f. and cdf that $\lim\limits_n G_n^n(|\overline{X}|>a)=1$ for any $a>0$. This easily implies that the limiting minimax risk is $l(\infty)$, proving the non-robustness of $\overline{X}$.

Obviously, it does not help to center the sample mean by addition of a bounded function. One reason the sample mean fails to be robust here is that the contamination neighborhoods are quite broad; if extremely narrow neighborhoods are employed, the robustness or not of the sample mean is more delicate - see Beran, 1979, and Example (3B.5) below.

**(3B) Location Model.** Let $\Theta=R^1$; fix a cdf $F_0$. Define $F_\theta$ by $F_\theta(t)=F_0(t-\theta)$. The parametric family $\{F_\theta\}$ is the so-called location model, the most intensively studied model in robustness theory to date. In this subsection, we apply (2.10) to see what familiar estimators arise from particular choices of $H$. The thrust of the discussion is that the fully robust estimators of the present theory yield estimators asymptotically equivalent to well-known reliable estimators of location; those that are not fully robust (e.g. sample mean) are downrated because of the narrowness of the contamination neighborhoods. Since it is easy to check the Hypotheses (2.1) - (2.3), we omit discussion of this point.

Assume $F_0$ has a density $f$ with respect to Lebesgue measure. Then under mild conditions, $\xi(t)=-f(t)$. If $G(t)=\int^t f(s)\,H(ds)$ then integration by parts yields $\langle \hat{F}_n-F_0, \xi\rangle_H=\int G(t)\,d(\hat{F}_n-F_0)(t)$. Set $c=\int G(t)\;F_0(dt)$, $b^{-1}=\int f^2(t)\,H(dt)$, $\psi(t)=b[G(t)-c]$. Then by (2.8) asymptotically about $\theta_0=0$:

$$\hat{\theta}_n=n^{-1}\sum_1^n \psi(X_i)$$

Notice that the $\psi$ functions here (the so-called 'influence curves') satisfy: (i) $\psi$ is monotone increasing and bounded if $H$ is finite (ii) if $H$, $f$ are symmetric about 0, then $\psi$ is odd (iii) if $H$ is nonatomic, then $\psi$ is continuous. Properties (i), (iii) are often considered desirable for location estimators with robust pretentions.

Next, let us evaluate the influence curves $\psi$ for various choices of $H$. Remember that the localization is about $\theta_0=0$; localizations about arbitrary $\theta$ should involve the families $H_\theta$, $H_\theta(dt)=H(dt-\theta)$.

**(3B.1) Hodges-Lehmann Estimator (H/L).** Take $H$ to be Lebesgue measure on the line. Then $\psi(t)=[F_0(t)-\frac{1}{2}]/\int f^2(t)\,dt$. When $F_0$ is symmetric, $\psi$ is the influence curve of the familiar Hodges-Lehmann estimator (cf. Hodges, Lehmann, 1963; Andrews, et al., 1972). Therefore (under symmetry) in the case $H(dx)=dx$, the minimum distance estimator is asymptotically equivalent to the $H/L$ estimator;

one way of looking at this particularly interesting case is to view the $H/L$ estimator as an adaptive version of the minimum distance estimator since its recipe does not depend on knowing $\{F_\theta\}$ (remark of P. Bickel). The $H/L$ estimator is not fully robust, of course; however, the contamination neighborhoods contain, if $F_0$ is normal, contamination by stable distributions of index $\alpha > 1$. That is, $H/L$ can, in the present theory, still deal with very heavy tailed contamination – a fact which perhaps explains the excellent performance of $H/L$ under standard Monte Carlo analysis.

**(3B.2) Normal Distributions.** If $H$ is $N(0, \sigma^2)$ and if $F_0$ is $N(0,1)$, then $\psi(t) = (2\pi(1+q^2))^{\frac{1}{2}} (\Phi(tq^{-1}) - \frac{1}{2})$ where $q^2 = \sigma^2(\sigma^2+1)^{-1}$, $\Phi = \text{cdf of } N(0,1)$. The case $\sigma^2 = 1$ gives the standard Cramér-von Mises estimator, while $\sigma^2 = \infty$ corresponds to $H =$ Lebesgue measure (cf. 3B.1).

**(3B.3) Median.** If $H$ is unit mass at 0, then $\psi(t) = [I_{[0,\infty)}(t) - (1 - F_0(0))]/f(0)$, $I$ being the indicator of the interval $[0, \infty)$. If $F_0$ is symmetric, $\psi(t) = \frac{1}{2}(f(0))^{-1}$ sign $t$, the influence curve of the median.

**(3B.4) Trimmed Means.** Choose $H(dx) = (f(x))^{-1} dx$ if $|x| < \alpha$, $H(dx) = 0$ otherwise. If $f$ is bounded away from 0 on $(-\alpha, \alpha)$. the resulting minimum distance estimator will be fully robust; otherwise it will not. Here $\psi(t) = -\alpha/p$, $t < -\alpha$; $= t/p$, $-\alpha \leqq t \leqq \alpha$; $= \alpha/p$, $t > \alpha$ where $p = \int_{-\alpha}^{\alpha} f(t)\,dt$. This is the influence curve of the trimmed mean (at $\frac{1}{2}(1-p)$ percent; cf. Andrews, et al. 1972). That is, for this choice of $H$, the minimum distance estimator is asymptotically equivalent to the trimmed mean estimator of location.

**(3.B5) Sample Mean.** Choose $H(dx) = (f(x))^{-1} dx$ (the measure is defined to be zero where $f(x) = 0$). The existence of the required derivatives can be established if $\int (f'(x))^2/f(x)\,dx < \infty$ (i.e., if $f$ has finite Fisher information.) Assume in addition that $\int |x| f(x)\,dx < \infty$ and support $F_0 = R^1$. Then $\psi(t) = t - c$, $-\infty < t < \infty$, where $c = \int x f(x)\,dx$; this is the familiar $\psi$ function of the sample mean. Of course $H$ here is not finite so the sample mean is not *fully* robust. The contamination neighborhoods contain only such cdf's $G$ that satisfy $\int (G(t) - F_0(t))^2 (f(t))^{-1} dt \leqq c n^{-\frac{1}{2}}$. The presence of $1/f(t)$ in this integral forces the tails of the allowable $G$'s to mimic the tails of $F_0$ extremely closely. The development here therefore suggests the fact, often noted in practice, that the mean cannot be robust against "heavy tailed" contamination, while at the same time our development shows in a striking quantitative way that the mean does have some robustness against very limited kinds of contamination. Of course, the fact that the allowable contamination neighborhoods are so small indicates that the mean will not generally be a reliable estimator of location. It is interesting to note that if $F_0$ is $N(0,1)$ then these contamination neighborhoods do *not* contain the contaminated normal distributions $(1 - n^{-\frac{1}{2}}) F_0 + n^{-\frac{1}{2}} G$ where $G$ is $N(0, \sigma^2)$, $\sigma^2 > 1$. Thus the present theory appears to reproduce the suspicion of some applied statisticians that the misbehaviour of the sample mean is already apparent even for this type contamination (cf. Tukey, 1960).

**(3B.6) Outer Mean.** The choice $H(dt) = 1/f(t)\,dt$ if $|t| > \alpha$, $= 0$ otherwise leads to $\psi(t) = t + \alpha$, $t < -\alpha$; $= 0$, $-\alpha \leqq t \leqq \alpha$; $= t - \alpha$, $t > \alpha$. This is the influence curve of the outer mean; as expected, examination of the contamination neighborhoods shows that the tails of the contaminating $G$'s must follow the tails of $F_0$ closely, but the middle part of $G$ can be fairly arbitrary.

**(3B.7) Winsorized Means.** Choose $H=H_1+H_2$ where $H_1(dt)=(1/f(t))\,dt$ if $\alpha<t<\beta$, $H_1=0$ otherwise; and $H_2$ is a two point distribution on $\{\alpha\}$, $\{\beta\}$. One is led to a $\psi$ of the form $\psi(t)=a_1, t<\alpha;\ =a_2+a_3(t-\alpha), \alpha\leqq t<\beta=a_4, t\geqq\beta$. The function $\psi$ will jump at $\alpha$ and $\beta$. Various specializations of this produce the influence curve of the Winsorized mean (see Hampel, 1974 for the definition). It would seem that the unnatural choice of $H$ would in itself call this estimator into question.

**(3B.8) Which *H*?** The basic theory of this paper delineates a broad class of robust estimators (one for each finite $H$), and also damns a host of other possible estimators. Among the fully robust estimators, the theory does NOT lead inexorably to a single estimator whose excellence towers over all others- instead a whole spectrum of good estimators is produced. There are some guidelines for selection among these. For example, since $|\ \ |_H$ is supposed to measure distance between measures, it seems clear that in general an $H$ with full support would do a better job than a point mass. One could also try to choose $H$ so that the resulting minimum distance estimator has as high efficiency as possible (in the classical sense) at the model; see Parr, deWet, 1980, for some attempts in this direction. Finally, if one knows something about the kind of contamination one could attempt to choose the estimator whose contamination neighborhoods have the most appropriate shape. None of these further suggestions are entirely convincing, and they fall distinctly outside the purview of the decision theoretic structure. Since (rightly) neither practitioners nor theoreticians can decide for all time on (e.g.) the 'proper' percentage for a trimmed mean, it is perhaps a good thing that our theory does not rigidly dictate the amount either.

**(3C) Scale Models.** Assume $\Theta=(0,\infty)$ and that the family $\{F_\theta\}$ is given by $F_\theta(x)=F_0(x/\theta)$ for a fixed cdf $F_0$. Assume that $F_0$ has a density $f$ with respect to Lebesgue measure. Reparametrize by $\sigma$: $F_\sigma(x)=F_0(x/1+\sigma)$. We intend to see what the estimator $\pi_0\hat{F}_n$ amounts to for $\sigma$ near 0. Denote by $\hat{\sigma}_n$ the choice of $\sigma$ made by the estimator $\pi_0\hat{F}_n$. Proceeding as in the location case, one finds under mild assumptions, $\xi(t)=-tf(t)$ so that $\hat{\sigma}_n=n^{-1}\sum\psi(X_i)$ where $\psi(t)=M(t)-c$, $M(t)=\int^t sf(s)\,H(ds)$, $c=\int M(t)\,F_0(dt)$. If $H, f$ are symmetric about 0 then the $\psi$ curves here satisfy (i) $\psi$ decreases as $t$ goes from $-\infty$ to 0 and increases as $t$ goes from 0 to $\infty$; (ii) $\psi$ is a bounded, even function if $\int|s|f(s)\,H(ds)<\infty$. Again, the shape of these $\psi$ curves is typically what one might expect of robust estimators in the scale case.

**(3C.1) Normal Distributions.** Let $F_0$ be $N(0,1)$ and let $H$ be $N(0,c^2)$. Straightforward calculation gives $\psi(t)=K[p(p^2+1)^{-\frac{1}{2}}-\exp(-t^2/2p^2)]$ where $K$ is a constant and $p^2=c^2/(c^2+1)$.

**(3C.2) Step Function $\psi$.** Take $H$ to be point mass 1 at each of the points $\{-1\}$ and $\{1\}$. If $f$ is symmetric, then

$$\begin{aligned}\psi(t)&=(2f(1))^{-1}(F_0(1)-F_0(-1)) \quad \text{if } |t|>1;\\ &=(2f(1))^{-1}(F_0(1)-F_0(-1)-1), \quad |t|\leqq 1.\end{aligned}$$

Slight variants of this example produce the influence curves of the familiar *median absolute deviation.*

**(3D) Location-Scale Model.** Here the parameter set $\Theta$ is $R \times (0, \infty)$; $P_\theta$ is given by $F_\theta(x) = F_0(x - \theta_1/\theta_2)$ if $\theta = (\theta_1, \theta_2)$. Reparametrize by $F_0(x - \theta/1 + \sigma), \theta, \sigma$ real. Then $\pi_0(\hat{F}_n - F_0)$ is given locally (near the parameter point $(0, 0)$) by projection onto the span $\sum$ of $f(t), tf(t)$ where $f$ is the density of $F_0$; every point in $\sum$ has a representation $\langle(\theta, \sigma), \xi\rangle$, $\xi = (-f(t), -tf(t))$ and so the point $(\hat{\theta}, \hat{\sigma})$ that gives $\pi(\hat{F}_n - F_0)$ is the estimator of $(\theta, \sigma)$. Performing this calculation gives the estimator $S(\hat{F}_n - F_0)$, where $S(x)$ is the vector $(S_1(x), S_2(x))$ with $x \in L^2(H)$ and

$$S_1(x) = (a_{12}\, a_2(x) + a_{22}\, a_1(x))/c$$
$$S_2(x) = (a_{11}\, a_2(x) + a_{12}\, a_1(x))/c$$

where

$$a_{ij} = \langle e_i, e_j\rangle_H, \quad a_i(x) = \langle x, e_i\rangle_H \quad i = 1, 2;\ j = 1, 2$$
$$e_1 = -f(t), \quad e_2 = -tf(t)$$
$$c = a_{11}\, a_{22} - (a_{12})^2$$

Notice that if $f(t)$ is perpendicular to $tf(t)$ then $S(x)$ simplifies to

$$S_1(x) = a_1(x)/a_{11}$$
$$S_2(x) = a_2(x)/a_{22}$$

so in this case the estimation of location and scale can be accomplished by estimating each separately, according to the recipes given earlier in this section. This separation of cases occurs, for example, when $f$ and $H$ are symmetric about 0.

**(3E) General Parametric Families.** An important feature of the present approach to robustness is that it does not depend on notions of symmetry for its implementation – quite general models can be employed. As an example of this, suppose $P_\theta$ is absolutely continuous with respect to $dJ$, and $f_\theta = dP_\theta/dJ$. Assume the family $\{P_\theta\}$ is quadratic mean differentiable (qmd) at $\theta_0 (=0)$: there exists $\eta = (\eta_1, \eta_2, \ldots, \eta_d)$, $\eta_i \in L^2(J)$, depending on $\theta_0$, such that $\int |f_\theta^{\frac{1}{2}} - f_\theta^{\frac{1}{2}} - \langle\theta, \eta\rangle|^2\, dJ = o(|\theta|^2)$ as $\theta \to 0$. Then (2.1) holds with $\xi$ defined by $\langle\theta, \xi(t)\rangle = 2\int^t f_\theta^{\frac{1}{2}} \langle\theta, \eta\rangle\, dJ$, and the differentiability of (2.1) is actually in the sense of the sup norm $|\ |_K$. Let $C$ be the Banach space of continuous functions obtained as the span of all elements $\int^t f^{\frac{1}{2}}\ h\, dJ$, $h \in L^2(J)$. Then $\sum$ is $d$-dimensional in $C$ if the matrix with entries $\int \eta_i(t)\, \eta_j(t)\, F_0(dt)$ is non-singular. The notion of qmd is fundamental for certain problems of classical estimation theory; see LeCam, 1969, for discussion of its statistical significance.

A great many of the common parametric families satisfy the qmd assumption. Perhaps of greatest importance for applied statistics are the exponential families. Here, in standard form, $dP_\theta/dJ = a(\theta) \exp(\langle\theta, h\rangle)$ where $a(\theta)$ is a normalizing constant, $h = (h_1, \ldots, h_d)$, $h_i$ a Borel function on some measure space, $\Theta = \operatorname{int}\{\theta \in R^d : 0 < a(\theta) < \infty\}$. Then $\eta = \frac{1}{2}(h - E(h|P_0))$; $\sum$ will be $d$-dimensional (in $C$) if $\operatorname{Cov}(h|P_0)$ is non-singular. Identifiability assumption (2.2) is also satisfied; see Berk, 1972.

**(3F) Uniform distributions.** The uniform distributions on $[0, \theta]$ are in general somewhat of a headache for classical estimation theory, because if the model is correct, the 'best' estimator of $\theta$ has an asymptotic theory that does not lead to Gaussian distributions. The deviant theory is trivial, but the departure from normality is annoying. The optimality properties of the maximum as an estimator of $\theta$ depend critically on peculiarities of the $\mathrm{U}(0, \theta)$ distribution; it is obvious, however that such an estimator cannot be (empirically) robust under any substantial data contamination. One amusing feature of our robustness theory is that it forces (perhaps brutally) the $U(0, \theta)$ model back into the asymptotically Gaussian framework. Indeed, if the model is reparametrized to $U(0, 1+\theta)$ and local analysis is undertaken at $\theta_0=0$, then it is easy to see that (2.1) holds, with $\xi(t)=0,\ t<0;\ =-t,\ 0\leqq t\leqq 1;\ =0,\ t>1$. Assumptions (2.2), (2.3) are immediate, and so for finite $H$ the minimum distance estimator of $\theta$ is robust. Its asymptotic form, given by (2.8), shows it asymptotically normal. That is, unlike the purely parametric situation, the best (robust) estimator is indeed asymptotically normal. It is easy to see that the maximum cannot be robust in the present framework. The uniform $(\theta-\frac{1}{2},\ \theta+\frac{1}{2})$ models also satisfy the assumptions (2.1)-(2.3), and could have been treated under Subsect. (3 B); of course the best (robust) estimators here will be asymptotically normal too.

## 4. Proofs

This section supplies proofs of theorems (2.4), (2.7), including the extension to sigma finite $H$ in Subsect. 2B. The proofs will be undertaken in three steps.

**Step (i). Weak Convergence of Empirical cdf in $L^2(H)$.** Let $W$ be the standard Brownian bridge stochastic process.

(4.1) **Proposition.** *Let $F_n$ be a sequence of cdf's on the line, $F_n$ converging in $H$ measure to a cdf $F$. Assume $\int F_n(1-F_n)dH<\infty$, $\int F(1-F)dH<\infty$ and $\lim_n \int F_n(1-F_n)dH =\int F(1-F)dH$. Then under the product measures $\{F_n^n\}$, $n^{\frac{1}{2}}(\hat{F}_n^n-F_n)$ is an $L^2(H)$-valued random variable which converges in distribution on $L^2(H)$ to $W\circ F$.*

*Proof.* If $H$ is finite, then $|\ |_H\leqq \text{const}\ |\ |_K$, so (4.1) is a trivial consequence of the well known weak convergence of the empirical cdf on the usual Banach space of continuous functions. Set $Y_n=n^{\frac{1}{2}}(\hat{F}_n-F_n)$, $Y=W\circ F$. Let $e_i, i=1,2,\ldots$ be an orthonormal basis of $L^2(H)$ such that each $e_i$ is bounded and $H(S_i)<\infty$, where $S_i$ is the support of $e_i$. By the usual weak convergence criterion in Hilbert space (cf. Prohorov, 1956) it suffices to show that for each $\varepsilon>0$, there exists $k$ for which

$$(4.2)\quad \sup_n \sum_{i\geqq k} E\langle Y_n, e_i\rangle_H^2<\varepsilon$$

Routine calculation, using the hypotheses and choice of $e_i$, show $\lim_n \sum_{i\geqq 1} E\langle Y_n, e_i\rangle_H^2=\sum_{i\geqq 1} E\langle Y, e_i\rangle_H^2$ and, for each $i$, $\lim_n\ E\langle Y_n,\ e_i\rangle_H^2=E\langle Y,\ e_i)_H^2$, which are sufficient for (4.2). Since the finite dimensional distributions $\langle Y_n, e_1\rangle_H,\ldots,\langle Y_n, e_k\rangle_H$ converge to $\langle Y, e_1\rangle_H,\ldots,\langle Y, e_k\rangle_H$, Proposition (4.1) follows.

Application of this result will employ $F_n$ of the form $F_n=F_0+n^{-\frac{1}{2}}q_n$, $q_n\in N(c)$.

**Step (ii). Proof of (2.4).** The proof will be carried out under the assumption that $l$ is uniformly continuous; the general case follows on approximating general $l$ from below by uniformly continuous functions. Fix $c$ and $q \in N(c)$. Then because of (2.1) - (2.3), $\pi_0 G_{nq} = F_0 + n^{-\frac{1}{2}} \pi q$ within $o(n^{-\frac{1}{2}})$ in the metric $| \ |_H$. For such $G_{nq}$, the form of $L_n$ forces the elements $F_\theta$ selected from the decision space to lie within a $n^{-\frac{1}{2}}$-ball of $F_0(|\cdot|_H$ metric); so within $o(n^{-\frac{1}{2}})$ such an element has the form $F_0 + n^{-\frac{1}{2}} T$, $T = \langle \theta, \xi \rangle$. Because of the assumed uniform continuity of $l$, the loss function $L_n(G_{nq}, F_\theta)$ may then for asymptotic purposes, be replaced by $l\,(|\pi q - T|_H^2 + (|(1-\pi)\, q|_H^2) \wedge a)$. To lighten the notation, assume $a = 0$. This having been done, we now bound from below

$$(4.3) \quad \liminf_n \inf_T \sup_{q \in N(c)} \int l(|\pi(q-T)|_H^2)\, dG_{nq}^n$$

where procedures $T$ now select elements of $\sum$. In particular, the cdf parametrization $\{F_\theta\}$ described in section 2 has been replaced by $\{\langle \theta, \xi \rangle, \theta \in R^d\}$.

If $f$ is the density of $P_0$ with respect to $J$, define $L_0^2(J)$ to be the Hilbert space consisting of elements $h \in L^2(J)$ orthogonal to $f^{\frac{1}{2}}$. Let $|\cdot|_J$ be the norm of $L^2(J)$. Define densities $f(n; h; x)$, $h \in L_0^2(J)$ by $f^{\frac{1}{2}}(n; h; x) = (1 - n^{-1}|h|_H^2)^{\frac{1}{2}}\, f^{\frac{1}{2}}(x) + n^{-\frac{1}{2}} h(x)$. This will indeed be a density if $|h|_H^2 < n$. Define $\tau$, a mapping on $L_0^2(J)$ by $\tau h(t) = \int^t f^{\frac{1}{2}}\, h dJ$. Since $|\tau h(t)|^2 \leqq |h|_J^2\ F_0(t) \wedge (1 - F_0(t))$, it is clear from Hypothesis (2.9) that $\tau$ maps $L_0^2(J)$ into $L^2(H)$. Let $L_0^2(H)$ be the closure in $L^2(H)$ of $\tau\ L_0^2(J)$. It is not difficult to check that $\tau$ is a Hilbert-Schmidt operator from $L_0^2(J)$ to $L_0^2(H)$. Assume from now on:

$$(4.4) \quad \sum \subset L_0^2(H)$$

Discussion of this hypothesis appears in Remark (4.5) below.

Let $Q_0$ be the image under $\tau$ of the canonical normal distribution on $L_0^2(J)$ (the characteristic functional of $Q_0$ is $\exp\{-\frac{1}{2}|\tau^* g|_J^2\}$, $g \in L_0^2(H)$, $\tau^* =$ adjoint $\tau$; see, e.g., Millar, 1979, for the definition of this concept.) Since $\tau$ is Hilbert-Schmidt, $Q_0$ is of course countably additive; indeed, $Q_0$ is the distribution on $L_0^2(H)$ of $W \circ F$ (cf. (4.1)). Define probabilities $Q_h$ on the Borel sets of $L_0^2(H)$ by $Q_h(A) = Q_0(A - \tau h)$. Then $\prod_{i=1}^n f(n; h; x_i)/f(x_i)$ converges in distribution, under $\prod_{i=1}^n f(x_i) dx_i$, to $dQ_h/dQ_0$. It then follows from a basic asymptotic minimax theorem and (4.4) that (4.3) may be bounded below by

$$\liminf_n \inf_T \sup_{\tau h \in N(c)} \int l(|\pi(\tau h - T)|_H^2)\, dG_{nq}^n \geqq \inf_T \sup_{\tau h \in N(c)} \int l(|\pi(\tau h - T)|_H^2\, dQ_h$$

where now $T$ takes values in $L_0^2(H)$.

See Millar, 1979, for more detail on the argument just indicated, as well as for a version of the minimax theorem immediately applicable to the present situation; see LeCam, 1972, Hájek, 1972, for other variants of the minimax theorem. If now $c_n$ increases to infinity in any way whatsoever, it then follows that

$$\liminf_n \inf_T \sup_{q \in N(c_n)} \int l(|\pi(q-T)|_H^2)\, dG_{nq}^n \geqq \inf_T \sup_h \int l(|\pi(T - \tau h)|_H^2)\, dQ_0$$

Since $x \to l(|\pi x|_H^2)$ is subconvex on $L_0^2(H)$ in the sense of Millar, 1979, section 3, the last expression in the foregoing display can be evaluated as $\int l(|\pi x|_H^2)\, Q_0(dx)$. Q.E.D.

**Step (iii). Proof of (2.7).** The proof will be accomplished first under the assumption that $l$ be uniformly continuous (and bounded, if $H$ is only sigma finite). Fix $c$; define $Y_n = n^{\frac{1}{2}}(\hat{F}_n - F_0)$. If $A(\lambda) = \{|Y_n|_H > \lambda\}$, then by Chebychev, $G_{nq}^n(A(\lambda)) = O(\lambda^{-2})$, where 0 is independent of $q$ as long as it belongs to $N(c)$. It then follows from (2.1) - (2.3) that on complement $A(\lambda)$, $\pi_0 \hat{F}_n = F_0 + \pi(\hat{F}_n - F_0) + o_p(n^{-\frac{1}{2}})$. Since $l$ is uniformly continuous on $[0, \lambda]$, the global loss $L_n(G_{nq}, \pi_0 \hat{F}_n)$ may, for asymptotic purposes, be replaced on $A^c(\lambda)$ by $l(n|\pi(\hat{F}_n - G_{nq})|_H^2)$ (taking $a = 0$ as in step (ii)). If in addition $l$ is uniformly bounded, it is easy to see that this replacement on the entire sample space affects the risk at $G_{nq}^n$ by $0(\lambda^{-2})$(uniformly for $q \in N(c)$); if $l$ satisfies only (2.6) and $H$ is finite a slightly tedious uniform integrability argument (which invokes Kiefer, Wolfowitz, 1958) leads to essentially the same conclusion with $o(1)$ replacing $0(\lambda^{-2})$ as $\lambda \to \infty$.

It is also clear that

$$\lim_n \sup_{q \in N(c)} \int l(n|\pi(\hat{F}_n - G_{nq})|_H^2)\, dG_{nq}^n = El(|\pi W \circ F_0|_H^2)$$

Indeed, if this were not so, then there would exist $q_n \in N(c)$ and $\varepsilon > 0$ such that if $F_n = F_0 + n^{-\frac{1}{2}} q_n$ then (because of step (ii))

$$\lim_n \int l\,(n|\pi(\hat{F}_n - F_n)|_H^2)\, dF_n^n = El(|\pi W \circ F_0|_H^2) + \varepsilon.$$

But if $q_n \in N(c)$, then $F_n$ satisfies the hypotheses of (4.1), leading to a contradiction. Taking the approximation of the preceding paragraph into account, one finds, by a standard argument, sequences $c_n \uparrow \infty$ such that, by letting $\lambda \uparrow \infty$, one obtains (2.7). A modest amount of care is needed to do the entire argument in complete detail.

To remove the hypothesis that $l$ be uniformly continuous, notice that the uniform continuity was needed only for $l$ restricted to $[0, \lambda]$. The extension to general $l$ can be completed by approximating $l$ above on $[0, \lambda]$ by uniformly continuous functions and using the fact the distribution of $|\pi W \circ F_0|_H^2$ is nonatomic. Q.E.D.

(4.5) **Remark.** Hypothesis (4.4) can be guaranteed in a variety of ways. To illustrate, if $\Theta$ is a subinterval of $R^1$, then in many examples, $\xi(t) = \int^t \dot{f}(\theta_0, s)\, J(ds)$ where $\dot{f}(\theta, s) = \partial/\partial\theta\, f(\theta, s)$. If $\dot{f}(\theta_0, s)/f^{\frac{1}{2}}(\theta_0, s) \in L^2(J)$. Then $\xi = \tau(\dot{f}/f^{\frac{1}{2}})$ so (4.4) evidently holds. For a better condition, write $f^{\frac{1}{2}}(\theta, t) = (1 - |h_\theta|^2)^{\frac{1}{2}} f^{\frac{1}{2}}(0, t) + h_\theta$ where $h_\theta$ is orthogonal to $f^{\frac{1}{2}}(0, \cdot)$ in $L^2(J)$. Then, because of (2.1) it is easy to see that (4.4) will hold if $n^{\frac{1}{2}} \zeta^2(P_{\theta n} - \frac{1}{2}, P_0) \to 0$, for then $\langle \theta, \xi \rangle$ is a limit in $L^2(H)$ of $2 \int^t f_0^{\frac{1}{2}}\, h_{\theta n} - \frac{1}{2}\, dJ$; Here $\zeta$ denotes Hellinger distance between the measures in question. That (4.4) holds for the $U(0, \theta)$ case can be checked directly.

## 5. Extensions

**(5A) Dependent Observations.** The assumption of independent observations can be weakened. If, under the envisioned dependence, (4.1) still holds, and if independence is a special case of the type of dependence under consideration, then the main results may be recovered.

**(5B) Multidimensional Case.** Analogues of the main results continue to hold if $P_\theta$ is permitted to be a probability on $R^d$ (instead of $R^1$). The Brownian bridge, of course, is replaced by another Gaussian process depending on $F_0$, and so forth.

**(5C) Global Variants.** It is possible to write variants of the main result in which the point $\theta_0$ is not chosen in advance. Instead of $n^{-\frac{1}{2}}$ balls about a fixed $\theta_0$, one envisions a $n^{-\frac{1}{2}}$ 'sausage' surrounding the entire $P_\theta$ family; this sausage is the union of $n^{-\frac{1}{2}}$ balls at each $\theta$. To rederive the main result under these circumstances, additional (but mild) compactness properties of the family $\{P_\theta\}$ are brought in; we spare the reader the details.

**(5D) Testing.** It is possible to develop a theory of robust testing based on the estimators of this paper. Roughly speaking, one would reject (say) $\theta=0$ if $\pi_0 \hat{F}_n$ were too far away from $P_0$. This theory, with the definition of robustness (2.5) adapted to the testing problem, has been developed in Millar, 1979.

**(5E) Families of Measures.** The extension of the theory to families $\{H_\theta\}$ was mentioned in Sect. 2. In these circumstances, other minimum distance type estimators - easier to analyse - can be introduced. For example, if $\tilde{\theta}_n$ is a preliminary $n^{\frac{1}{2}}$ consistent estimator of $\theta$ (see LeCam, 1969, for existence of such; his theory can be adapted to robustness frameworks), one can define $\pi_1 \hat{F}_n$ to be any element $F_\theta$ which achieves $\inf_\theta |\hat{F}_n - F_\theta|_{\tilde{\theta}_n}$. Robustness, in the sense of this paper, can be established for this modified estimator. Variants of this turn out to be useful in the testing problem when there are nuisance parameters (see Millar, 1979, for this; see Pollack, 1979, for some asymptotic normality results about these modified estimators).

**(5F) Alternatives toSample cdf.** The minimum distance estimators of this paper are defined in terms of the sample cdf. Instead, one can introduce a kernel $k(\cdot)$, defined on $R^1$, and consider $\hat{F}_n(k, t) = \int k(t-x)\, \hat{F}_n(dx)$. If $k$ is the indicator of $[0, \infty)$, the empirical cdf is recovered. The parameter $\theta$ naturally will be estimated by picking the element $\int k(t-x)\, F_\theta(dx)$ closest to $\hat{F}_n(k, t)$. Under regularity assumptions on $k$, one can obtain analogues of the main result. In the location problem, one may obtain by proper choice of $k$, estimators with redescending influence curves - something not possible with $\pi_0 \hat{F}_n$. This possibility was pointed out to me by R.J. Beran.

*Acknowledgement.* I thank R. Beran for calling my attention to the problem discussed herein, for a number of helpful comments, and for continued interest in the project.

## References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W.: Robust Estimates of Location. Princeton Univ. Press 1972

Beran, R.J.: Minimum Hellinger distance estimates for parametric models. Ann. Statist. **5**, 445–463 (1977a)

Beran, R.J.: Robust location estimates. Ann. Statist. **5**, 431–444 (1977b)

Beran, R.J.: Asymptotic lower bounds for risk in robust estimation. Preprint 1979

Bickel, P.J.: Another look at robustness: a review of reviews and some new developments. Scand. J. Statist. **3**, 145–168 (1976)

Bickel, P.J.: Quelques aspects de la statistique robustes. Ecole d'Ete de St. Flour. Springer. To appear (1979)

Bickel, P.J., Lehmann, E.L.: Descriptive statistics for non parametric models I, II. Ann. Statist. **3**, 1038–1069 (1975)

Berk, R.H.: Consistency and asymptotic normality of MLE's for exponential families. Ann. Math. Stat. **43**, 193–204 (1972)

Hâjek, J.: Local asymptotic minimax and admissibility in estimation. Proc. Sixth Berkeley Symposium **1**, 175–194 (1972)

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

Hampel, F.R.: The influence curve and its role in robust estimation. J. Amer. Stat. Assoc. **69**, 383–393 (1974)

Hodges, J.L., Lehmann, E.L.: Estimates of location based on rank tests. Ann. Math. Stat. **34**, 598–611 (1963)

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

Huber-Carol, C.: Thesis. E.T.H. Zürich (1970)

Jaeckel, L.: Robust estimates of location: symmetry and asymmetric contamination. Ann. Math. Statist. **42**, 1020–1034 (1971)

Kiefer, J., Wolfowitz, J.: On the deviation of the empiric distribution function of vector chance variables. Trans. Amer. Math. Soc. (1958)

Knüsel, L.F.: Über minimum distance Schätzungen. Thesis. ETH (1969)

LeCam, L.: Theorie Asymptotique de la Decision Statistique. Les Presses de l'Universite de Montreal (1968)

LeCam, L.: Limits of experiments. Proc. Sixth Berkeley Symposium **I**, 245–261 (1972)

LeCam, L.: Convergence of estimates under dimensionality restrictions. Ann. Stat. **1**, 38–53 (1973)

Millar, P.W.: Asymptotic minimax theorems for the sample distribution. To appear. Zeit. für Wahrscheinlichkeitstheorie **48**, 233–252 (1979)

Millar, P.W.: Robust tests of statistical hypotheses. Preprint (1979)

Neymann, J.: Contributions to the theory of the $\chi^2$-test, Proc. First Berkeley Symp., 239–273 (1949)

Parr, W.C., Schucany, W.R.: Minimum distance and robust estimation. Preprint (1979)

Parr, W.C., deWet, T.: On minimum weighted Cramer von Mises statistic estimation. Preprint (1979)

Pollack, D.: The minimum distance method of testing. To appear, Metrika (1979)

Prohorov, Yu.V.: Convergence of random processes and limit theorems in probability theory. Theor. Prob. Appl. **1**, 157–214 (1956)

Rieder, H.: Estimates derived from robust tests. Submitted to Ann. Stat. (1978)

Skorokhod, A.V.: Limit theorems for stochastic processes. Theor. Prob. Appl. **1**, 261–290 (1956)

Stigler, S.: Do robust estimators work with real data? Ann. Stat. **6**, 1055–1098 (1977)

Tukey, J.W.: A survey of sampling from contaminated distributions. Contrib. to Prob. and Stat. I. Olkin, ed. pp. 448–485. Stanford Univ. Press. (1960)

Wolfowitz, J.: Estimation by the minimum distance method. Ann. Inst. Stat. Math. **5**, 9–23 (1953)

Wolfowitz, J.: The minimum distance method. Ann. Math. Stat. **28**, 75–88 (1957)

Received June 3, 1980