The Annals of Statistics
1988, Vol. 16, No. 2, 552–586

THE "AUTOMATIC" ROBUSTNESS OF MINIMUM DISTANCE FUNCTIONALS

By David L. Donoho[1] and Richard C. Liu

University of California, Berkeley

The minimum distance (MD) functional defined by a distance μ is automatically robust over contamination neighborhoods defined by μ. In fact, when compared to other Fisher-consistent functionals, the MD functional was no worse than twice the minimum sensitivity to μ-contamination, and at least half the best possible breakdown point. In invariant settings, the MD functional has the best attainable breakdown point against μ-contamination among equivariant functionals. If μ is Hilbertian (e.g., the Hellinger distance), the MD functional has the smallest sensitivity to μ-contamination among Fisher-consistent functionals.

The robustness of the MD functional is inherited by MD estimates, both estimates based on "weak" distances and estimates based on "strong" distances, when the empirical distribution is appropriately smoothed.

These facts are general and apply not just in simple location models, but also in multivariate location–scatter and in semiparametric settings.

Of course, this robustness is formal because μ-contamination neighborhoods may not be large enough to contain realistic departures from the model. For the metrics we are interested in, robustness against μ-contamination is stronger than robustness against gross errors contamination; and for "weak" metrics (e.g., μ = Cramér–von Mises, Kolmogorov), robustness over μ-neighborhoods implies robustness over Prohorov neighborhoods.

1. Introduction. An attractive feature of the maximum likelihood estimator is that it is "automatically" efficient. While there are exceptions, a useful rule of thumb in applied work is that the MLE has the smallest possible variance among reasonable estimators when the model holds.

Is there a class of estimators that is "automatically" robust in the same sense, i.e., which is generically optimal according to some quantitative measure of robustness? We will show one sense in which minimum distance (MD) estimators form such a class.

Of the many notions of robustness, we can identify two of a quantitative nature:

(1) *Stability of variance*. The asymptotic variance of the estimator should stay small, uniformly over neighborhoods of the model.

Received October 1985; revised August 1987.

[1] Work supported by the Mathematical Sciences Research Institute (1983) and an NSF Postdoctoral Fellowship in the Mathematical Sciences.

AMS 1980 *subject classifications*. Primary 62F35; secondary 62F12.

Key words and phrases. Quantitative robustness; gross-error sensitivity; breakdown point; Kolmogorov, Lévy, Prohorov, variation, Hellinger distances; Cramér–von Mises discrepancy.

(2) *Stability of quantity estimated.* The quantity being estimated (i.e., the limiting value of the estimate under increasing sample size) should change as little as possible, uniformly over neighborhoods of the model.

A large body of work exists investigating robustness properties of estimators with respect to criterion (1). Huber (1964) showed how to design M-estimators (generalizations of maximum likelihood estimators) satisfying criterion (1) in an optimal fashion. In this sense, the class of M-estimators "automaticaliy" contains robust estimates. However, much work has gone into the problem of finding the optimally robust ones in this class; solving Huber's design problem is by no means an automatic procedure. [See Rousseeuw (1981) for a somewhat different approach to (1).]

Less study has been devoted to robustness criterion (2), which in some senses is the more intuitive of the two criteria. One does not invoke difficult stochastic concepts; one only asks, what quantity has the most stable meaning under departures of a certain kind from the model.

It is with respect to this second notion of robustness that MD estimators are "automatically" robust.

We introduce some notation. Let $\{P_\theta\}$ be a family of probabilities indexed by θ (the parametric model) and let μ be a metric between probabilities. Let $\hat{\theta}(P)$ denote the corresponding *minimum distance functional*, i.e., a solution to

$$\mu(P, P_{\hat{\theta}}) = \min_{\theta} \mu(P, P_\theta). \tag{1.1}$$

(More details will be provided in Section 2.)

The MD functional is the quantity the MD estimator is trying to estimate; it is defined by the model $\{P_\theta\}$ and the metric μ. This functional is automatically robust over μ-neighborhoods of the model.

- $\hat{\theta}$ has within a factor of 2 the smallest sensitivity of small μ-perturbations among all Fisher-consistent functionals, i.e., functionals T satisfying $T(P_\theta) = \theta$.
- It has within a factor of 2 the best breakdown point with respect to μ-contamination among Fisher-consistent functionals.

Thus the value of the functional changes very little over small μ-neighborhoods of the model (subject to Fisher consistency), and the value cannot be distorted arbitrarily away from its value at a particular model distribution without moving very far away from that model.

Often the MD functional behaves even better than this, and the factor of 2 may be removed.

- In situations of invariance [e.g., where μ is invariant under translation (or scaling) and θ is a location (or scale) parameter], $\hat{\theta}$ has the largest possible breakdown point (with respect to μ-contamination) among equivariant functionals [a translation-equivariant functional satisfies $T(P_h) = T(P) + h$, where P_h denotes the shift of P by h: $P_h(A) = P(A - h)$].

- When μ is a Hilbertian distance (e.g., the Hellinger distance), $\hat{\theta}$ has the smallest possible sensitivity (to μ-contamination) among Fisher-consistent functionals.

In principle, this means that to satisfy the robustness criterion (2) is easy: One simply chooses a metric μ generating neighborhoods one would like to be robust over; then the corresponding MD estimates gives good or even optimal quantitative robustness over that neighborhood. Admittedly, one hardly knows what sort of neighborhood "one would like to be robust over." The question has not been systematically addressed in the literature. In Tukey's phrase, we do not know what we should "choose to fear."

It seems worth pointing out the conflict between robustness criteria (1) and (2). Whereas M-estimators can be robust according to (1), they can fail to be robust according to (2). And minimum distance estimators, which are robust according to (2), can fail to be robust according to (1).

Consider first the case of M-estimators. Huber (1976) and Maronna (1976) found that in d-dimensions, an affine-equivariant M-estimator of multivariate location and scatter has a breakdown point not exceeding $1/(d+1)$. Under a relatively small contamination by outliers of a certain kind, the estimator would cease to estimate the location and scatter of the "bulk" of the data, and estimate something else instead; and the amount of contamination necessary to cause this to happen could be quite small in high dimensions.

For some MD estimators, on the other hand, the results of Donoho and Liu (1988) show that one can have an asymptotic variance that is arbitrarily large at some distributions arbitrarily near the model. However, a specific subclass of the MD estimators—those based on Hilbertian metrics—seems to avoid this problem.

Contents of this paper. Sections 2 and 3 of the paper introduce some notation and give background information on MD estimates and on quantitative robustness, respectively.

Section 4 establishes some basic facts about MD estimates in one-parameter models. These include the starred results mentioned previously. Section 5 extends these results to the multivariate location–scatter problem, to semiparametric models and to minimum discrepancy estimates. Thus the facts appear to be somewhat general.

Section 6 considers the properties of MD estimates, and shows that the bounds of Sections 4 and 5 on the MD functional do apply to the limit points of MD estimates. Consequently, the limiting value of the MD estimate based on μ is insensitive, in some cases optimally so, to μ-contaminations. Also, the automatic consistency, and even root-n consistency, of some MD estimates is established as a simple byproduct of our results on the MD functional. Also, we show that for MD estimates at least, the robustness of the MD functional actually implies finite-sample robustness [in the sense of Donoho and Huber (1983)], at least for large enough sample sizes.

An application. Of course, the sort of robustness we are describing—robustness over μ-neighborhoods of the model—is rather formal. However, we focus attention on certain metrics μ, and for these the notion is more than formal. All of the functionals we consider are robust over Huber neighborhoods; and those based on "weak" metrics are robust over Prohorov neighborhoods. These results can therefore establish the existence of consistent estimators with good sensitivity and breakdown point in a variety of estimation problems. As an application, we collect together results from Sections 5.1, 6.2 and 6.4 in the following.

PROPOSITION 1.1. *The MD estimate of multivariate normal location and scatter based on either the halfspace or strip metrics (defined in the following discussion) (has a version which) is affine-equivariant, is consistent and root-n consistent when the model holds, has a finite-sample breakdown point approaching $\frac{1}{2}$ in large samples from the multivariate normal and its limiting value has a finite gross-error sensitivity.*

This is in contrast to the Huber–Maronna phenomenon for M-estimates, which break down easily in high dimensions.

In another sense, however, our results *are* formal. It is not at all obvious how to compute some of the minimum distance estimators we discuss. For example, even computing the halfspace distance between an empirical and true distribution in dimension $d > 1$ seems to require a d-dimensional nonlinear optimization—"projection pursuit." We leave the interesting question of how to efficiently compute minimum distance estimates for further work.

2. Background or minimum distance estimation.

In this paper, we assemble a minimum distance estimator using three components:

(1) A distance measure $\mu(P, Q)$ between probabilities. As a metric, μ satisfies the triangle inequality.
(2) A parametric family $\{P_\theta\}$ to be fitted.
(3) An estimated probability $\hat{P}_n$ based on n observations $\{X_1, \dots, X_n\}$.

The components are combined to produce an estimator by the rule

$$\hat{\theta}(\hat{P}_n) = \arg\min_{\theta} \mu(\hat{P}_n, P_\theta). \tag{2.1}$$

That is, $\hat{\theta}$ is a value of the parameter that produces the best approximation to $\hat{P}_n$ from the family $\{P_\theta\}$. This value need not, of course, be unique, either in the sample or in the population.

We use several combinations of the three components in this paper, and should point out in advance the variety of methods possible.

Metrics. When the data X_i are real-valued observations, so that μ is a metric between probabilities on the real line, we consider μ chosen from the

TABLE 1

	Weak	Strong	Hilbertian	Normed	Location invariant	Scale invariant
Kolmogorov	y			y	y	y
Kuiper	y			y	y	y
Lévy	y				y	
Prohorov	y				y	
Variation		y		y	y	y
Hellinger		y	y	y	y	y

following list:

Distance	$\mu(P, Q)$
Kolmogorov	$\sup_{A=(-\infty, t]} \lvert P(A) - Q(A) \rvert$
Kuiper	$\sup_{A=(a, b]} \lvert P(A) - Q(A) \rvert$
Lévy	$\inf\{\delta\colon P(-\infty, t] \le Q(-\infty, t+\delta] + \delta\}$
Prohorov	$\inf\{\delta\colon P(A) \le Q(A^\delta) + \delta$ for all measurable $A\}$
Variation	$\sup_{\text{measurable } A} \lvert P(A) - Q(A) \rvert$
Hellinger	see (8.4) et seq.

For later reference we list in Table 1 some descriptive information about these metrics. Here *weak* means "based on the distribution function," whereas *strong* means "based on the density function"; *Hilbertian* means "based on a quadratic measure of deviation"; *normed* means $\mu(P, Q)$ arises as the norm of the difference between P and Q in an appropriate sense (e.g., the Kolmogorov–Smirnov distance is the norm of $\Delta(t) = P(-\infty, t] - Q(-\infty, t]$, viewed as an element of $L^\infty(\mathbf{R})$).

We also consider other possibilities. In Section 5.1 we introduce metrics for the case where observations are multivariate. In Section 5.3 we give some results when μ is the Cramér–von Mises goodness-of-fit measure, not satisfying the triangle inequality, and thus not a proper metric.

Estimated probability. Once the choice of μ is made, we select our estimate $\hat{P}_n$ of P as follows. If μ is a weak distance, we use the empirical measure $P_n = n^{-1}\Sigma\delta_{X_i}$. If μ is a strong distance, we smooth the empirical measure, producing an estimate with a density. In detail, we let K_h be a distribution with smooth density and bandwidth (i.e., scale) h, and put

$$\hat{P}_n = K_{h_n} * P_n, \tag{2.2}$$

where h_n depends on the sample size in an appropriate fashion. (See also Section 6.)

Models. In this paper, we consider one-parameter models for real-valued observations, such as the normal location model $P_\theta = N(\theta, 1)$, as well as a multiparameter model—multivariate normal location–scatter. We also briefly consider the semiparametric model $\{P_\theta\} = \{\text{all symmetric distributions}\}$.

Many of the estimators we study here have been discussed in the literature. A partial list of references includes Beran (1977), Holm (1976), Millar (1981), Parr and Schucany (1980) and Rao, Schuster and Littell (1975).

3. Background on quantitative robustness. We are given a functional T and are interested in quantifying its robustness with respect to small changes in$P$. There are several ways of doing this, depending on which robustness criterion one is interested in—i.e., criterion (1) or criterion (2) of the Introduction. In what follows, we focus on (2): We measure the change in $T(P)$ caused by small changes in P.

We need the concept of an "ideal" distribution P_0 (which holds for physical or other reasons); the real data we are able to obtain have a distribution P distorted through gross errors, nonlinearities of measurement, rounding errors and other factors outside our control. To make a quantitative assessment of the effects of such distortions, we employ a measure of distortion $\delta = \delta(P; P_0)$. δ may be one of the metrics mentioned earlier, or else a discrepancy such as the Huber contamination discrepancy,

$$\delta_{\text{Huber}}(P; P_0) = \inf\{\varepsilon\colon P(A) \geq (1-\varepsilon)P_0(A) \text{ for all measurable } A\}.$$

We can then measure how much T changes under a δ-distortion of size $\leq \varepsilon$. A formal measure of this is the *bias-distortion curve*

$$b(\varepsilon) = \sup\{|T(P) - T(P_0)|\colon \delta(P; P_0) \leq \varepsilon\}.$$

$b(\varepsilon)$ depends on T, P_0 and δ for its definition; although we shall usually suppress these as they will be obvious from context.

If the neighborhoods $\{P\colon \delta(P; P_0) \leq \varepsilon\}$ increase with increasing ε, $b(\varepsilon)$ is increasing with ε. Our main interest is in how fast it increases, and in finding procedures for which it does not increase too fast. The "classical" work of Hampel (1968) can be viewed as designing estimators to make $b(\varepsilon)$ optimally small for small ε, subject to a constraint on the asymptotic variance at the model. Huber (1964) covered, in passing, the problem of finding an estimator with minimal $b(\varepsilon)$. Our approach is analogous, but we work with a variety of distortion measures; Huber and Hampel worked principally with the Huber discrepancy.

Useful information can be read off from a graph of $b(\varepsilon)$. Figure 1 portrays three important descriptive parameters. If $\lim_{\varepsilon \to 0} b(\varepsilon)$ is 0, then small distortions away from P_0 do not affect the value of T very much. Second, if $b(\varepsilon)$ is nearly *linear* near 0, with slope γ^*, a small distortion away from P_0 changes T in a fashion that is essentially no worse than linear in the contamination measure δ:

$$|T(P) - T(P_0)| \leq \gamma^* \delta(P; P_0)(1 + o(1)),$$

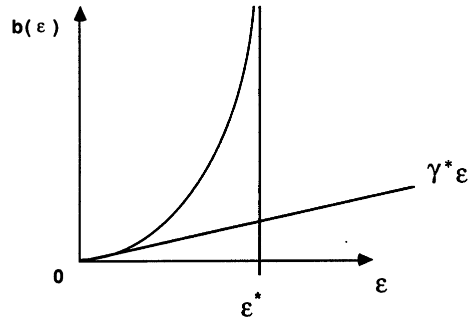

FIG. 1.

as $\delta(P, P_0) \to 0$. Third, if the vertical asymptote of this curve (at ε^*, say) is far from $\varepsilon = 0$, then it takes a fairly large distortion away from P_0 to make T blow up completely.

For certain δ, the properties just mentioned were studied by Hampel (1968). For δ = Prohorov metric, the condition $b(\varepsilon) \to 0$ as $\varepsilon \to 0$ is Hampel's condition for "qualitative robustness" of T. For δ = Huber discrepancy, γ^* was called the "gross-error sensitivity" and ε^* the "breakdown point," respectively. In the following we use the same names even when different discrepancies or metrics are involved. All three concepts are really quite intuitive when δ is a metric. Indeed, they measure properties of the graph of T near P_0. If the graph is continuous at P_0, then $b(\varepsilon) \to 0$ as $\varepsilon \to 0$; if it is locally Lipschitz at P_0, then $\gamma^* = \lim_{\varepsilon \to 0} \sup(b(\varepsilon)/\varepsilon) < \infty$; and ε^* is the largest ε such that the graph of T has no "singularities" in a ball of radius $\leq \varepsilon$ about P_0.

Much more could be said about quantitative robustness, particularly as concerns robustness criterion (1). The reader in pursuit of further information should consult the books of Huber (1981) and of Hampel, Ronchetti, Rousseeuw and Stahel (1986). Finally, Bickel (1981) discusses a way to measure quantitative robustness with respect to both (1) and (2) simultaneously, by the "shrinking neighborhoods" technique.

4. Bias optimality of MD. In this section $\{P_\theta\}$ is a parameter family indexed by a real-valued parameter θ. For example, P_θ might be $N(\theta, 1)$. We study the $b(\varepsilon)$ behavior of the minimum distance functional under the assumption $\mu = \delta$, so the estimation metric and the contamination discrepancy are the same. We obtain bounds on $b(\varepsilon)$ from above and below by various approaches.

In detail, we define $b(\varepsilon)$ as

$$b(\varepsilon) = \sup_{\mu(P, P_0) \leq \varepsilon} \; \sup_{\text{solutions to (1.1)}} \left|\hat{\theta}(P) - \theta_0\right|. \tag{4.1}$$

That is, the bound we are calculating applies to every version of the MD functional, i.e., every solution of the MD equation *in the population*,

$$\mu(P, P_{\hat{\theta}}) = \min_\theta \mu(P, P_\theta). \tag{4.2}$$

For the moment, we leave aside the question of whether this also bounds the limiting value of the MD statistic under sampling. Section 6 shows that our results do have rigorous application.

4.1. *Basic bounds on* $b(\varepsilon)$. Some basic inequalities for $b(\varepsilon)$ in the $\mu = \delta$ case are easy to obtain. They all involve the following *gauge* function, used to convert from distance between probabilities (units of μ) to distance between parameter values (units of θ):

$$b_0(\varepsilon) = \sup\{|\theta - \theta_0| : \mu(P_\theta, P_0) \le \varepsilon\}.$$

This function says how far apart two "labels" (parameter values) can get while the "models" (probability distributions) stay within ε of each other. It depends implicitly on $\{P_\theta\}$, μ and θ_0, although we suppress this dependence. For regular families [e.g., the $N(\theta, 1)$] and regular metrics $b_0(\varepsilon) \to 0$ as $\varepsilon \to 0$ and, in fact, b_0 is nearly linear in ε near $\varepsilon = 0$.

A basic observation concerning b_0 is: *For every Fisher-consistent functional* T [i.e., $T(P_\theta) = \theta$] $b_T(\varepsilon) \ge b_0(\varepsilon)$. To see this, pick θ so that $\mu(P_\theta, P_{\theta_0}) = \varepsilon$ and $|\theta - \theta_0| = b_0(\varepsilon)$. By Fisher consistency $T(P_\theta) = \theta$, so $|T(P_\theta) - \theta_0| = b_0(\varepsilon)$: The largest value of $|T(P) - \theta_0|$ over the entire ε-neighborhood must be at least this big.

A basic fact for the MD functional is

$$b(\varepsilon) \le b_0(2\varepsilon)$$

(assuming $\mu = \delta$). This follows directly from the triangle inequality. Let $\mu(P, P_{\theta_0}) \le \varepsilon$. Since $\hat{\theta}$ is a solution of the MD equation,

$$\mu(P, P_{\hat{\theta}}) \le \mu(P, P_{\theta_0}) \le \varepsilon,$$

we have

$$\mu(P_{\theta_0}, P_{\hat{\theta}}) \le 2\varepsilon.$$

Now by definition of the gauge b_0,

$$|\hat{\theta} - \theta_0| \le b_0(2\varepsilon).$$

It follows from this inequality that $\hat{\theta}$ is Fisher-consistent—i.e., $\hat{\theta}(P_\theta) = \theta$—whenever $b_0(0; \theta_0) = 0$ for all θ_0. Examining definitions, this last condition is the same as *identifiability*,

$$\theta_1 \ne \theta_0 \textit{ implies } \mu(P_{\theta_1}, P_{\theta_0}) > 0.$$

Thus *the MD functional* $\hat{\theta}$ *is Fisher-consistent whenever the family* $\{P_\theta\}$ *is identifiable.*

Fisher consistency implies, as we have seen, the inequality $b(\varepsilon) \ge b_0(\varepsilon)$. Thus $b(\varepsilon)$ is bracketed between $b_0(2\varepsilon)$ and $b_0(\varepsilon)$. Actually, the lower bound is true even without Fisher consistency. For

$$b(\varepsilon) \ge \sup_{\mu(P_\theta, P_0) \le \varepsilon} \ \sup_{\text{solution to (1.1)}} |\hat{\theta}(P_\theta) - \theta_0|$$

and since the set of solutions of (1.1) always contains θ when $P = P_\theta$,

$$\sup_{\text{solutions to (1.1)}} |\hat{\theta}(P_\theta) - \theta_0| \geq |\theta - \theta_0|.$$

Thus

$$b(\varepsilon) \geq \sup_{\mu(P_\theta, P_0) \leq \varepsilon} |\theta - \theta_0| = b_0(\varepsilon).$$

We summarize this discussion formally.

PROPOSITION 4.1. *For the MD functional based on metric* μ,

$$b_0(\varepsilon) \leq b(\varepsilon) \leq b_0(2\varepsilon), \tag{4.3}$$

whenever the distortion measure $\delta = \mu$.

This easy result has two basic corollaries.

COROLLARY 1. *If the gauge* $b_0(\varepsilon) \sim C\varepsilon$ *as* $\varepsilon \to 0$, *then for the MD functional we have*

$$\gamma^*(\hat{\theta}) \leq 2 \inf_T \gamma^*(T), \tag{4.4}$$

where the infimum is over all Fisher-consistent T.

COROLLARY 2. *For the MD functional*

$$\varepsilon^*(\hat{\theta}) \geq \tfrac{1}{2} \sup_T \varepsilon^*(T), \tag{4.5}$$

where the supremum is over all Fisher-consistent T.

These results say that for general parameter families, the MD functional has within a factor of 2 of the best gross-error sensitivity and breakdown point. We will show that the MD often has the best possible breakdown point and the best possible gross-error sensitivity.

4.2. *The minimax functional.* Computation of $b(\varepsilon)$ has a simple, game-theoretic interpretation. Think of a two-person game where nature chooses a θ and a P with $\mu(P, P_\theta) \leq \varepsilon$ and a statistician chooses a functional T with loss to the statistician of

$$|T(P) - \theta|.$$

Then $b(\varepsilon)$ bounds the loss of the statistician who plays strategy $\hat{\theta}$. We will see that this strategy is nearly minimax in many cases.

The minimax strategy for this game is as follows. The statistician, presented with P (and knowing ε), computes the set $\mathbf{S}_\varepsilon$ of all parameter values θ such that $\mu(P, P_\theta) \leq \varepsilon$. Let T_ε be the *center* of this set, that is the θ-value minimizing $\max_{t \in \mathbf{S}_\varepsilon} |\theta - t|$. For example, if $\mathbf{S}_\varepsilon$ is an interval, then the center is the midpoint of that interval.

T_ε is the statistician's minimax strategy. Let $b_-(\varepsilon)$ be its value, i.e., $b_{T_\varepsilon}(\varepsilon)$.

PROPOSITION 4.2. *For any functional T,*

$$\max_{\theta_0} b_T(\varepsilon;\theta_0) \geq \max_{\theta_0} b_-(\varepsilon;\theta_0). \tag{4.6}$$

The main interest of this result is in cases where an invariance is present, so that the "$\max_{\theta_0}$" is unnecessary. Consider the case where $\{P_\theta\}$ is a location model [e.g., $N(\theta, 1)$] and the metric μ is translation-invariant. Then the T_ε functional will be translation-equivariant and $b_-(\varepsilon)$ will not depend on θ_0.

In such cases of invariance $b_-(\varepsilon)$ can often be computed explicitly. It turns out to involve the gauge b_0 in an explicit way. Here are two examples.

PROPOSITION 4.3. *Let μ be a translation-invariant metric defined by a norm: $\mu(P, Q) = \|P - Q\|$. Examples include total-variation, Kuiper and Kolmogorov–Smirnov. Then, if $\{P_\theta\}$ is a translation family,*

$$b_-(\varepsilon) = b_0(2\varepsilon)/2. \tag{4.7}$$

PROPOSITION 4.4. *Let μ be Hellinger distance. Then, if $\{P_\theta\}$ is a translation family,*

$$b_-(\varepsilon) = b_0\Big(2\varepsilon\sqrt{1 - (\varepsilon/2)^2}\Big)/2. \tag{4.8}$$

It is possible to compute b_- in a few other invariant cases although we do not pursue this here.

COROLLARY 3. *In the cases covered by Propositions* 4.3 *and* 4.4,

$$b(\varepsilon) \leq 2b_-(\varepsilon).$$

COROLLARY 4. *In the cases covered by Propositions* 4.3 *and* 4.4,

$$b_-(\varepsilon)/b_0(\varepsilon) \to 1, \quad \text{as } \varepsilon \to 0.$$

4.3. *Breakdown point.* An implication of Propositions 4.3 and 4.4 is that in these cases, *the MD functional has the best possible breakdown point among all translation-equivariant functionals.*

Indeed, since in these cases we have $b_-(\varepsilon) \geq b(\varepsilon)/2$, it follows that the MD functional is within a factor of 2 of being bias-minimax *for each* ε. Consequently, the bias of the MD functional is finite whenever that of the minimax functional is finite. The two functionals therefore have the same breakdown point.

4.4. *Sensitivity.* In certain cases, the MD functional has a sensitivity that is actually optimal rather than within a factor of 2 of being optimal [as is always the case, by (4.4)]. The optimality extends over all Fisher consistent functionals, and does not depend for example on invariance.

The basic idea is simple geometry. Consider Figure 2, which presents a straight "parameter family" and a spherical "contamination neighborhood." It is a simple fact of Euclidean geometry, that for every "contaminated" point lying

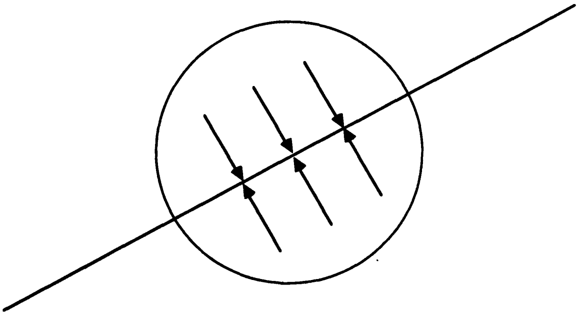

FIG. 2.

inside the ball but not on the "family," the closest point on the family lies inside the ball. Indeed, the point on the ball that projects to the point farthest from its center is the member of the "parameter family" at the point where the "family" exits the ball. Accordingly, for a "flat" parameter family and $\delta = \mu =$ Euclidean distance, the least favorable contamination is one which stays within the parameter family. As points in the parametric family within an ε-ball about 0 cannot have a parameter value larger than what the gauge permits, $b_0(\varepsilon)$ (this is after all the definition of the gauge), it follows that in this setting,

$$b(\varepsilon) = b_0(\varepsilon),$$

so that MD is bias-minimax among all Fisher-consistent functionals in this special model.

This geometric fact applies to our setting in the following way. First, Euclidean distance can be replaced by any *Hilbertian* distance [e.g., the $L_2(H)$ or the Hellinger distances]. Second, the parameter family, although globally curved, must be locally *flat* (i.e., differentiable). Then $b(\varepsilon) \approx b_0(\varepsilon)$ for small ε; i.e., γ^* is optimal.

PROPOSITION 4.5. *Let $\mu(P, Q) = \|P - Q\|$, where $\|\cdot\|$ is the norm of a Hilbert space. If the curve $\theta \to P_\theta$ is Fréchet-differentiable at θ_0, and if $b_0(\varepsilon)$ is differentiable at* 0 *with $b_0' > 0$, then*

$$\lim_{\varepsilon \to 0} \left(b(\varepsilon)/b_0(\varepsilon) \right) = 1,$$

i.e.,

$$\gamma^*(\hat{\theta}) = \inf_T \gamma^*(T),$$

where the infimum is over all Fisher-consistent functionals.

We can make two easy applications of this result in the location model.

COROLLARY 5 (Millar's MD functional). *Let μ be the $L_2(P_0)$ distance between c.d.f.'s, as in Millar* (1981). *Let $\{P_\theta\}$ be a location model with density p_θ,* *and suppose that $\theta \to p_\theta$ is continuous in $L_2(P_0)$ quadratic mean at $\theta = \theta_0$. (For example, suppose the density is bounded.) Then the MD functional based on this distance has the smallest sensitivity to $L_2(P_0)$ perturbations of the model of any Fisher-consistent functional.*

This optimal property of the MD functional was not mentioned by Millar (1981).

COROLLARY 6 (Beran's MD functional). *Let μ be the Hellinger distance; suppose again that $\{P_\theta\}$ is a location model, and that P_0 has finite Fisher information. Then the minimum Hellinger distance functional has the least sensitivity to Hellinger perturbations of the model among all Fisher-consistent functionals.*

This last result was first obtained by Beran (1977).

In short, for Hilbertian MD functionals, $\gamma^*(\hat{\theta})$ will typically be optimal.

What happens if μ is not Hilbertian? For example, if it is the Kolmogorov or variation distance? Typically, in these cases, the factor 2 in (4.4) cannot be avoided.

PROPOSITION 4.6. *If $\{P_\theta\}$ is a location family with continuous c.d.f., then for $\mu = \delta =$ Kolmogorov distance*

$$b(\varepsilon) = b_0(2\varepsilon)\ (= 2b_-(\varepsilon)).$$

PROPOSITION 4.7. *If $\{P_\theta\}$ is a location family with density having a continuous, integrable derivative, then for $\mu = \delta =$ variation distance*

$$b(\varepsilon) = b_0(2\varepsilon) + o(\varepsilon).$$

Again, some understanding can be gleaned from finite-dimensional geometry. Figure 3 shows two flat "parameter families" and an unit ball which is not round —the l_1 ball. Note that in the first case, some points in the ball project outside the ball under minimum distance projection in the l_1 distance. In the second

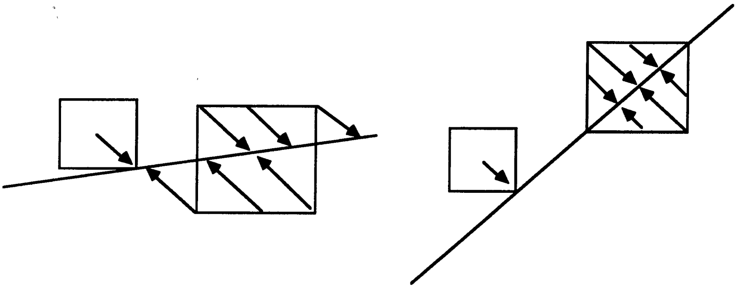

FIG. 3.

case, this is not true: All points inside the ball project inside the ball (or there exists a *version* of the projection with this behavior).

In this non-Hilbertian case, the MD functional is suboptimal by a factor of 2 for some families; for other families it is optimal. At least in the case of Kolmogorov and variation distances, it appears that only the suboptimal cases occur within the "regular" (i.e., smooth) families.

This distinction between Hilbertian and non-Hilbertian distances is also important in Donoho and Liu (1988).

5. Generalizations. Although the results of Section 4 are stated in a restricted setting—simple one-dimensional parametric models—the reasoning behind them is in fact quite abstract and general. We will give in this section examples showing that

- θ can be allowed to range over a k-dimensional or even an abstract set;
- $|\theta_1 - \theta_2|$ can be replaced by quite general norms or discrepancies;
- the distance measure μ can be replaced by a discrepancy;

and the main conclusions of Section 4 will continue to hold.

Our three examples cover: mulivariate location and scatter models, semiparametric models and minimum discrepancy estimates.

5.1. *Multivariate location and scatter.* Let $\theta = (t, C)$ where t is a vector in $\mathbf{R}^d$ (a location parameter) and C is a $d \times d$ covariance matrix (scatter parameter). Let $P_\theta = P_0(C^{-1/2}(\cdot - t))$, where P_0 is the standard d-dimensional Gaussian measure. Define the following discrepancy between parameter values:

$$D_{\mathrm{aff}}(\theta_0, \theta_1) = \log\left[\left(\det\left(C_0 C_1^{-1}\right) + \det\left(C_1 C_0^{-1}\right)\right)/2\right] + \left(\|t_0 - t_1\|_{C_0} + \|t_0 - t_1\|_{C_1}\right)/2,$$

where det is the usual determinant function, and

$$\|u - v\|_\Sigma^2 = (u - v)^{\mathrm{T}} \Sigma^{-1} (u - v).$$

Note that $D_{\mathrm{aff}}(\theta_0, \theta_1)$ is invariant under affine transformations. Indeed, if $\theta_i = (t_i, C_i)$, $i = 0, 1$, and $\tilde{\theta}_i = (At_i + b, AC_iA^{\mathrm{T}})$, $i = 0, 1$, it is easy to check that $D_{\mathrm{aff}}(\theta_0, \theta_1) = D_{\mathrm{aff}}(\tilde{\theta}_0, \tilde{\theta}_1)$.

We will be interested in the following metrics between probability distributions on $\mathbf{R}^d$. First, the d-dimensional variation and Hellinger distances, generalized in the obvious way from the one-dimensional case. Second, the following generalizations of the Kolmogorov and Kuiper distances. Let $H_{u,t}$ denote the halfspace $\{x\colon u^{\mathrm{T}} x \le t\}$ of $\mathbf{R}^d$. Put

$$\text{halfspace}(P, Q) = \sup_{u,t} \left| P(H_{u,t}) - Q(H_{u,t}) \right|, \tag{5.1}$$

$$\text{strip}(P, Q) = \sup_{u,t} \left| P(H_{u,s} \cap H_{u,r}) - Q(H_{u,s} \cap H_{u,r}) \right|. \tag{5.2}$$

These are the "halfspace" and "strip" metrics in that they measure the largest discrepancy between P and Q over halfspaces $H_{u,x}$ or strips $H_{u,r} \cap H_{u,s}$. They are generalizations of Kolmogorov and Kuiper distances in that they measure the largest Kolmogorov (Kuiper) distance between any one-dimensional projection of the two probabilities. Thus they may be evaluated by "projection pursuit." See also Donoho (1982), Section 6. We note that *all these metrics are affine invariant.*

This machinery allows one to define minimum distance estimators of location and scatter in the obvious way. Note that by construction, the set of solutions to the MD equation in the population is affine-equivariant. Thus in all these cases there are versions of the MD functional that are affine-equivariant.

To measure the robustness of the MD functional in this settting, it makes sense to define b_0, b_- and b as in Section 4, only with respect to the discrepancy D_{aff}. For example, in this setting we would define the gauge via

$$b_0(\varepsilon) = \left\{D_{\text{aff}}(\theta_0, \theta_1) \colon \mu\left(P_{\theta_0}, P_{\theta_1}\right) \leq \varepsilon\right\}.$$

Doing this leads to the following generalization of the results of Section 4.

PROPOSITION 5.1. *In the multivariate location and scatter problem with parameter discrepancy D_{aff}, let μ be one of the metrics: variation, Hellinger, halfspace, strip. If $\mu = \delta$, we have that b_0, b_- and b are all affine-invariant;*

$$b_0(\varepsilon) \leq b(\varepsilon) \leq b_0(2\varepsilon), \qquad b_-(\varepsilon) = b_0(2\varepsilon)/2,$$

except for $\mu = \delta =$ Hellinger where

$$b_-(\varepsilon) = b_0\left(2\varepsilon\sqrt{1 - (\varepsilon/2)^2}\right)/2.$$

Thus in each case the MD functional has the best breakdown point with respect to μ-contamination of all equivariant functionals. For the Hellinger case we also have, if P_θ has an absolutely continuous density p_θ with $\theta \to p_\theta^{1/2}$ differentiable in quadratic mean,

$$b(\varepsilon) = b_0(\varepsilon) + o(\varepsilon).$$

Thus the MD functional has the best resistance against small Hellinger perturbations among Fisher-consistent functionals.

We conjecture that, as in Section 4.4, the other metrics, all of which are non-Hilbertian, have twice the best attainable γ^* at regular models.

One reason for studying this example is the fact, mentioned in the Introduction, that any affine-equivariant M-estimator of location and scatter has a breakdown point in d-dimensions not exceeding $1/(d+1)$. In this sense, the M-estimators get less and less robust in high dimensions. Stahel (1981) and Donoho (1982) showed that a variety of affine-equivariant procedures attain high breakdown point even in high dimensions—among the examples of Donoho (1982) was an MD estimator based on the halfspace metric. Proposition 5.1 shows us that we can generally expect affine-equivariant MD functionals to have the best attainable breakdown point with respect to μ-contamination—"automatically."

5.2. *Semiparametric models.* Now let $\theta = (t, S)$, where $t \in \mathbf{R}$ is a location parameter and S is any distribution symmetric about 0. Put $P_\theta(\cdot) = S(\cdot - t)$. Thus $\{P_\theta\}$ is the family of all symmetric distributions. Let $D_{\mathrm{sp}}(\theta_0, \theta_1) = |t_0 - t_1|$. This discrepancy only pays attention to the difference in location parameters.

Although the set $\{\theta\}$ is now infinite dimensional, the following lemma shows that the MD functional is sometimes available from a one-dimensional minimization.

LEMMA 5.2. *Let μ be a vector norm, which is both translation- and reflection-invariant. The components of the MD functional are given by*

$$\hat{t}(P) = \arg\min_t \mu\left(\tfrac{1}{2}P(\cdot) + \tfrac{1}{2}P(2t - \cdot), P\right),$$

$$\hat{S}(P) = \tfrac{1}{2}P(\cdot) + \tfrac{1}{2}P(2\hat{t} - \cdot).$$

Such μ include the Kolmogorov, Kuiper and variation distances. [If μ is Hellinger distance, $\frac{1}{2}(P(\cdot) + P(2t - \cdot))$ in these expressions is replaced by the measure equidistant between $P(\cdot)$ and $P(2t - \cdot)$ along the Hellinger geodesic connecting them; see, e.g., the proof of Proposition 4.4.]

In this setting, we can define a semiparametric gauge $b_0^{(\mathrm{sp})}$ and semiparametric bias–distortion curve $b^{(\mathrm{sp})}$, using the parameter discrepancy D_{sp}. The analog of b_- does not make sense because the problem is not invariant under changes in the shape component S of θ.

PROPOSITION 5.3. *If μ is a translation-invariant, reflection-invariant vector norm,*

$$b_0^{(\mathrm{sp})}(\varepsilon) = b_0(2\varepsilon)/2, \tag{5.3}$$

where b_0 is the gauge of the parameter family generated by shifts of P_{θ_0}.

[For Hellinger distance, a similar expression holds, with a factor $\sqrt{1 - \varepsilon^2/4}$ multiplying the argument to b_0 on the right-hand side of (5.3).]

Thinking of b_0 as a measure of the "intrinsic" limit on the robustness attainable in a problem (because $b \geq b_0$), this proposition says robustness is not intrinsically more difficult (for small ε) in the semiparametric situation.

PROPOSITION 5.4. *For each of $\mu =$ Kolmogorov, Kuiper, variation and Hellinger, we have*

$$\gamma^*(\textit{semiparametric MD}) = \gamma^*(\textit{parametric MD}),$$

where the parametric family is generated by shifts of P_{θ_0}, at each θ_0 where the appropriate regularity conditions of Propositions 4.4, 4.5 *or* 4.6 *apply.*

As far as the breakdown point goes, we can only conclude from (5.3) that ε^*(semiparametric MD) $\geq \varepsilon^*$(parametric MD)/2. However, we believe that something better is true.

CONJECTURE. *For each of μ = Kolmogorov, Kuiper, variation and Hellinger,*

$$\varepsilon^*(\textit{semiparametric MD}) = \varepsilon^*(\textit{parametric MD}),$$

where now the parameter family for the parametric MD is generated by translation and scaling of θ_0.

Thus for both small and large ε, we believe that this semiparametric problem is no more difficult than the ordinary parametric problem.

We remark that, as in Donoho (1982), one can also consider a *multivariate* symmetric location problem. That is, $\theta = (t, S)$ where now t is a vector in $\mathbf{R}^d$ and S is a distribution on $\mathbf{R}^d$ *centrosymmetric* about 0. In that setting, results similar to Propositions 5.2–5.3 hold for the multivariate variation, Hellinger, halfspace and strip metrics.

5.3. *Minimum discrepancy functionals.* Results similar to those of Section 4 can continue to hold if μ is replaced by a discrepancy that does not obey the triangle inequality. The gauge is of less use, however.

Consider for example the minimum discrepancy procedure defining $\hat{\theta}$ via

$$\delta(P; P_{\hat{\theta}}) = \min_{\theta} \delta(P; P_\theta).$$

For concreteness, let δ be the Cramér–von Mises discrepancy

$$\delta(P; P_0) = \|P - P_0\|_{L_2(P_0)}. \tag{5.4}$$

This discrepancy does not satisfy the triangle inequality, and is not even symmetric in its arguments. Nevertheless, one can compute $b(\varepsilon)$, and, as the proof of the next proposition shows, $b_-(\varepsilon)$. Note that this discrepancy is invariant under location and scale changes. Thus in, say, a location model, there exists a translation-equivariant version of the minimum CvM functional. Moreover, as the CvM discrepancy is locally Hilbertian, the idea of Section 4.4 can be applied.

PROPOSITION 5.5. *For the minimum CvM functional in the location model both $b(\varepsilon)$ and $b_-(\varepsilon)$ are invariant under θ_0, $b(\varepsilon) \le 2b_-(\varepsilon)$ and so the minimum CvM functional has the largest breakdown point with respect to CvM contamination of any translation-equivariant functional. Similar conclusions hold in the location–scale model, although the value of the optimal breakdown point for affine-equivariant functionals is half as large as that for translation-equivariant ones. If $b_0'(0)$ exists and is positive and if $\{P_\theta\}$ is a location family satisfying the conditions* (8.25) *given in the following discussion, then $b(\varepsilon) = b_0(\varepsilon) + o(\varepsilon)$, and the minimum CvM functional has the smallest γ^* with respect to CvM contamination of any Fisher-consistent functional.*

There are extensions to cover general one-parameter models and even the semiparametric model, but we do not pursue these here.

6. Behavior of MD estimates. This section shows that the formal results of Sections 4 and 5 have rigorous application. That is, we will show that the $b(\varepsilon)$ bound for the behavior, in the population, of the MD *functional* also bounds the large-sample behavior of the MD *estimate*. We will also give some simple applications of $b(\varepsilon)$ computations to proving consistency, rates of convergence and finite-sample robustness of MD estimates.

6.1. *Bias behavior under sampling.* We are still assuming $\delta = \mu$. Recall equations (2.1) and (4.2). By our definition of $b(\varepsilon)$ in Section 4, *any* solution of (2.1) satisfies

$$(6.1) \qquad \left|\hat{\theta}(\hat{P}_n) - \theta_0\right| \le b\left(\mu\left(\hat{P}_n, P_0\right)\right).$$

Therefore, in a sampling situation where $X_1, X_2, \ldots$ are iid P, $\mu(P, P_0) \le \varepsilon$ and

$$(6.2) \qquad \limsup_{n\to\infty} \mu\left(\hat{P}_n, P_0\right) \le \varepsilon \quad \text{a.s.},$$

we have immediately

$$(6.3) \qquad \limsup_{n\to\infty} \sup_{\text{solutions to (2.1)}} \left|\hat{\theta}(\hat{P}_n) - \theta_0\right| \le b(\varepsilon^+) \quad \text{a.s.}$$

Actually, (6.2) holds in all the cases that interest us. The only subtletly is that we need to take some care in choosing the smoothing procedure when μ is a strong metric.

PROPOSITION 6.1. *Let $X_1, X_2, \ldots$ be iid P, where $\mu(P, P_0) \le \varepsilon$. We have (6.2) in any of these cases.*

(1) *Weak metrics/discrepancies. Here $\hat{P}_n$ is the empirical measure. The X_i are real-valued, and μ is either Kolmogorov–Smirnov, Kuiper, Lévy, Prohorov or Cramér–von Mises. The X_i are vector valued, and μ is either the halfspace or strip metric.*

(2) *Strong metrics. Here $\hat{P}_n$ is a smoothed empirical $\hat{P}_n = K_{h_n} * P_n$, where P_n denotes the empirical distribution. The model P_0 is absolutely continuous. The X_i are $\mathbf{R}^d$-valued. The bandwidth of the kernel goes to 0 slowly enough:*

$$(6.4) \qquad h_n \to 0, \qquad nh_n^d \to \infty.$$

COROLLARY 7. *For each of the situations mentioned in Sections 4 and 5, and each metric μ mentioned in connection with those models, we have that in the sampling situation described in Proposition 6.1,*

$$(6.5) \qquad \limsup_{n\to\infty} \sup_{\text{solutions to (2.1)}} D\left(\hat{\theta}\left(\hat{P}_n\right), \theta_0\right) \le b(\varepsilon^+),$$

when $\hat{P}_n$ is chosen as described in that proposition. Here D is the parameter discrepancy appropriate to the problem (e.g., $|\theta_0 - \theta_1|$, D_{aff} or D_{sp}), and b is computed assuming $\mu = \delta$ and using D as a parameter discrepancy.

In short, when (6.2) holds, our formal computation of $b(\varepsilon)$ gives a rigorous upper bound on how far limit points of $\hat{\theta}(\hat{P}_n)$ can be from θ_0 under an ε-distortion of P_0—it is only necessary to replace the formally computed b by its right-continuous version, b is, however, often continuous and so provides a bound with which to begin.

If we regard the right-hand side of (6.5) as the "formal" $b(\varepsilon)$ and the left-hand side of (6.5) as the "rigorous" $b(\varepsilon)$, we have that

$$\text{rigorous } b(\varepsilon) \leq \text{formal } b(\varepsilon^+),$$

so that

$$\text{rigorous } \gamma^* \leq \text{formal } \gamma^*, \tag{6.6}$$

$$\text{rigorous } \varepsilon^* \geq \text{formal } \varepsilon^*. \tag{6.7}$$

We have checked a number of cases without finding any instance where the rigorous quantities and the formal quantities differ. We presume that the inequality (6.5) is actually an equality, but have no argument to cover all the cases mentioned in this paper.

6.2. *Automatic consistency.* Minimum distance estimators have the reputation for being "automatically" consistent, and even "automatically" $n^{-1/2}$ consistent. Actually, this follows directly from properties of the gauge and the metric.

PROPOSITION 6.2. *If the gauge and the metric satisfy $b_0(\varepsilon) \to 0$ as $\varepsilon \to 0$ and $\mu(\hat{P}_n, P) \to 0$ a.s., then the MD estimator based on μ is consistent when the model holds.*

PROOF. Combining inequality (6.1) with the inequality $b(\varepsilon) \leq b_0(2\varepsilon)$, we have

$$D(\hat{\theta}_n, \theta_0) \leq b_0(2\mu(\hat{P}_n, P_0)). \tag{6.8}$$

Here θ_0 is the true parameter value, and D is the appropriate parameter discrepancy. By hypothesis, the right-hand side tends to 0 almost surely. □

The result can be strengthened with rates of convergence.

PROPOSITION 6.3. *If $b_0(\varepsilon) = O(\varepsilon)$ as $\varepsilon \to 0$ and $\mu(\hat{P}_n, P_0) = O_P(n^{-1/2})$, then the MD estimator based on μ is $n^{-1/2}$ consistent when the model holds.*

PROOF. In (6.8) the right-hand side is $O_P(n^{-1/2})$ by hypothesis. □

The first result applies immediately to all the cases we have been considering. For any of the weak metrics, and either of the strong metrics smoothed appropriately, the condition $\mu(\hat{P}_n, P_0) \to 0$ a.s. holds (for the strong metrics, we must assume that the model P_0 has a density). The condition $b_0(\varepsilon) \to 0$ is just a restatement of the Wolfowitz identifiability condition. The second result is more restricted. It applies only to some of the weak metrics: Kolmogorov, Kuiper, Lévy, $L_2(P_0)$ and the halfspace and strip metrics. However, the condition $b_0(\varepsilon) = O(\varepsilon)$ generally holds for smooth models.

For a simple application of these results, consider the multivariate location and scatter problem of Section 5.1. It is routine to calculate that $b_0(\varepsilon) = O(\varepsilon)$ at the Gaussian model with mean 0 and identity covariance. As the halfspace and strip metrics are both $n^{-1/2}$ consistent (this is an application of Vapnik–Cervonenkis theory), the two propositions imply that the MD estimate of location and scatter is consistent and root-n consistent at the Gaussian model.

6.3. *Finite sample robustness.* Donoho (1982) introduced notions of finite-sample robustness somewhat different from the asymptotic ones used here. Let observations $X_1, \dots, X_n$ be given; we can imagine contaminating the dataset $X = \{X_1, \dots, X_n\}$ in several ways. Two of these are the *augmenting* model and the *replacement* model. In the first, we adjoin an arbitrary dataset Y of size m to X, resulting in an augmented dataset $X^a = X \cup Y$ of size $n + m$. In the second, we change m of the n values in X arbitrarily, replacing them with new values, and producing a dataset X^r of size n. These are finite-sample analogs of Huber and variation contamination, respectively.

Are MD estimates robust against this type of contamination? An argument can be made that as contamination acting on samples can be made conditional on the actual sample drawn, finite-sample contamination is more disrupting than contamination of the probability distribution, followed by sampling.

However, it turns out that the MD estimators we are studying *are* robust against this type of contamination, at least in large samples. Again, the basic insight comes from (6.1) and some simple inequalities. Suppose that $\mu(\cdot, P)$ is convex in its first argument and bounded by 1 (all the metrics except the Hellinger satisfy this assumption). Then, letting P^a_{n+m} denote the empirical distribution of X^a,

$$P^a_{n+m} = \frac{n}{n+m} P_n + \frac{m}{n+m} Q_m,$$

where P_n is the empirical distribution of X and Q_m is the empirical of Y. Consequently,

$$\mu(P^a_{n+m}, P) \le \frac{n}{n+m}\mu(P_n, P) + \frac{m}{n+m}\mu(Q_m, P). \tag{6.9}$$

Apply now (6.1). Put $\varepsilon_a = m/n + m$, and obtain for the MD estimate based on μ:

$$\left|\hat{\theta}(\hat{P}^a_n) - \hat{\theta}(P)\right| \le b\big((1 - \varepsilon_a)\mu(\hat{P}_n, P) + \varepsilon_a\big). \tag{6.10}$$

(We remark that this conclusion even holds if μ is the TV distance and $\hat{P}_n$ is obtained by kernel smoothing.)

A similar bound can be obtained for the case of replacement contamination by assuming that μ is bounded by the total variation distance (again this holds for all the measures except the Hellinger). These bounds can be used to give connections between the finite-sample and asymptotic contamination models.

PROPOSITION 6.4. *Let $X_1, X_2, \dots$ be iid P_0. Suppose that the metric μ is convex and bounded by* 1. *Suppose that μ is one of the metrics covered by Proposition* 6.1. *The finite-sample augmenting contamination breakdown point of $\hat{\theta}$, $\varepsilon_a^*(\hat{\theta}, \{X_1, \dots, X_n\})$, satisfies*

$$\liminf_{n\to\infty} \varepsilon_a^*\big(\hat{\theta}, \{X_1, \dots, X_n\}\big) \geq \varepsilon^*\big(\hat{\theta}, P_0\big) \quad a.s.,$$

where $\varepsilon^*(\hat{\theta}, P_0)$ is the asymptotic breakdown point of the MD functional defined by μ. Suppose that the metric μ is bounded by the total variation distance. Then the finite-sample replacement contamination breakdown point of $\hat{\theta}$, $\varepsilon_r^*(\hat{\theta}, \{X_1, \dots, X_n\})$, satisfies

$$\liminf_{n\to\infty} \varepsilon_r^*\big(\hat{\theta}, \{X_1, \dots, X_n\}\big) \geq \varepsilon^*\big(\hat{\theta}, P_0\big) \quad a.s.$$

We mention that these inequalities can be slack. In fact, the breakdown point on the right-hand side may be attained, in some cases, only by μ-contamination that cannot be realized via augmenting or replacement contamination.

This proposition can be used to recover the results of Donoho (1982) on the finite-sample breakdown point of the location–scatter estimators described in Section 5.1. Also, a variation on the approach will cover Hellinger distance, and so recover the results of Boos and Tamura (1985) showing that the finite-sample breakdown point of the MHDE of multivariate location and scatter is at least $\frac{7}{16}$. However, a direct argument shows that the breakdown point is actually $\frac{1}{2}$; this illustrates the slack underlying (6.10).

7. Discussion.

7.1. *Implications.* Some of the results obtained in this paper would be rather difficult to get from any other point of view.

The authors were surprised by Corollary 3 and its generalizations: *In settings of invariance, there are estimators with $b(\varepsilon)$ never greater than twice the minimax value, for all ε.* The surprising aspect of this result is how it addresses both local quantities like the sensitivity and global ones like the breakdown point. Before this result, it was not easy to see that one can do well according to both sorts of measures, simultaneously. For example, one can construct affine-equivariant M-estimates of multivariate location and scatter with good gross-error sensitivity, but they will have a low breakdown point in high dimensions. On the other hand, comments in Donoho, Rousseeuw and Stahel (1988) reveal that one can have equivariant estimators with a high breakdown point together with very bad sensitivity to small departures from the model. So that it did not appear evident that local and global viewpoints could be reconciled.

Another insight is that minimum distance estimators are related to the median in several ways. First, Huber (1964) showed that the median is bias-minimax over Huber neighborhoods. So the median possesses the same sort of $b(\varepsilon)$-optimality with respect to a certain neighborhood structure that we have exhibited here for MD estimates. Second, neither MD estimators nor the median seem to have good variance stability over large neighborhoods. Donoho and Liu (1988) show that the non-Hilbertian MD estimators have bad asymptotic variances near the model.

On the other hand, the Hilbertian MD estimators have good variance properties over neighborhoods [see again Donoho and Liu (1988)], so they are not similar to the median in this last respect.

7.2. *Innovations.* The paper has introduced two notions which may be of broader interest. The first is the noting of *gauge* b_0. Inspection of the results of Sections 4–6 will alert the reader to the fact that criteria for identifiability and root-n consistency are easily posed in terms of the gauge, and that the minimum possible sensitivity and the best possible breakdown point (with respect to μ-contamination) at a given model "pop out" of the gauge.

The second notion is the bias-minimax functional T_ε of Section 4.2. This is not much harder to compute than $\hat{\theta}$ in some cases, and has reasonable $b(\varepsilon)$ behavior (optimal at one particular value of ε, by construction). It appears to have better variability properties than $\hat{\theta}$ when μ is not Hilbertian; for example, the pathologies described in Donoho and Liu (1988) do not appear to happen for T_ε.

7.3. *Relation to earlier work.* The closest connection to earlier work is to Theorem 6 of Beran (1977), which is essentially a proof of our Proposition 4.5 presented in different language. Compare also Huber (1964) for a discussion of bias minimaxity over gross-errors neighborhoods. Huber (1981), Theorems 1.4.1 and 1.4.2, gives results similar to Corollary 7 and equations (6.6) and (6.7).

Millar (1981) gives a result that he interprets in a similar way to our results: Namely, that MD estimators are robust over the contamination neighborhoods defined by their own metric. However, close inspection of his paper will reveal that he is addressing robustness criterion (1) in his paper, and that his results have to do with a special subclass of MD estimators, and not general MD estimators as we discuss here.

8. Proofs of results.

Proof of Proposition 4.2. Let **P** denote the set of all P within an ε distance of *some* P_θ (thus **P** is a "tube" about $\{P_\theta\}$). The proof is an exercise in slicing up the set **P** in different ways.

As a preliminary, we note that for any function $l(P, \theta)$,

$$\sup_{\theta} \sup_{\{P:\, \mu(P, P_\theta) \le \varepsilon\}} l(P, \theta) = \sup_{P \in \mathbf{P}} \sup_{\theta \in \mathbf{S}_\varepsilon(P)} l(P, \theta). \tag{8.1}$$

Indeed every (P, θ)-pair that appears in one expression appears in the other. Now

$$\begin{aligned}\sup_{\theta} b_T(\varepsilon;\theta) &= \sup_{\theta}\ \sup_{\{P:\ \mu(P,P_\theta)\le\varepsilon\}} |T(P)-\theta| \\ &= \sup_{P\in\mathbf{P}}\ \sup_{\theta\in\mathbf{S}_\varepsilon(P)} |T(P)-\theta| \\ &\ge \sup_{P\in\mathbf{P}} \operatorname{radius}\mathbf{S}_\varepsilon(P) \\ &= \sup_{P\in\mathbf{P}}\ \sup_{\theta\in\mathbf{S}_\varepsilon(P)} |T_\varepsilon(P)-\theta| \\ &= \sup_{\theta}\ \sup_{\{P:\ \mu(P,P_\theta)\le\varepsilon\}} |T_\varepsilon(P)-\theta| \\ &= \sup_{\theta} b_-(\varepsilon;\theta).\end{aligned}\tag{8.2}$$

The first and last equalities are by definition of b_T and b_-; the second and next-to-last are by (8.1). The steps surrounding (8.2) are, in more detail: For each $P\in\mathbf{P}$,

$$\begin{aligned}\sup_{\theta\in\mathbf{S}_\varepsilon(P)} |t-\theta| &\ge \operatorname{radius}\mathbf{S}_\varepsilon(P) \\ &= \inf_{t}\ \sup_{\theta\in\mathbf{S}_\varepsilon(P)} |t-\theta| \\ &= \sup_{\theta\in\mathbf{S}_\varepsilon(P)} |T_\varepsilon(P)-\theta|.\end{aligned}\tag{8.3}$$

The second line follows by definition of the radius, the third, by definition of T_ε. □

Remark 8.1. The proof carries over to settings where θ is a multiparameter and the measure $|\theta_1-\theta_2|$ is replaced by a more general parameter discrepancy $D(\theta_1,\theta_2)$. The key condition is the nesting property,

$$\{\theta_2:\ D(\theta_1,\theta_2)\le\delta\}\subset\{\theta_2:\ D(\theta_1,\theta_2)\le\delta+\varepsilon\},\quad \text{for all } \delta\ge\varepsilon\ge 0.$$

When this holds, the proof goes through even if θ ranges over an *abstract* set.

Remark 8.2. The only catch in such a generalization is as follows. Over an abstract set, the infimum in (8.3) may not be attained. Then there will be no functional T_ε with the minimax property. However, if we define

$$b_-(\varepsilon) = \sup_{\mu(P,P_\theta)\le\varepsilon} \operatorname{radius}\mathbf{S}_\varepsilon(P),$$

the theorem still holds, and for each $\delta>0$ there is a functional $T_{\varepsilon,\delta}$ which is δ-minimax:

$$b_{T_{\varepsilon,\delta}}(\varepsilon)\le b_-(\varepsilon)+\delta.$$

Remark 8.3. Close inspection reveals that the sup in (4.6) need not be over all θ; the proof actually gives the stronger result

$$\max_{|\theta-\theta_0|\le b_0(2\varepsilon)} b_T(\varepsilon)\ge \max_{|\theta-\theta_0|\le b_0(2\varepsilon)} b(\varepsilon).$$

Proof of Proposition 4.3. It is easy to see that $b_0(2\varepsilon)/2$ gives an upper bound on $b_-(\varepsilon)$. If $b_-(\varepsilon) = \rho$, say, then for each $\delta > 0$ there must be two parameter values θ_1 and θ_2 with $|\theta_1 - \theta_2| \geq 2\rho - \delta$ and a P with $\mu(P, P_{\theta_1}) \leq \varepsilon$; $\mu(P, P_{\theta_2}) \leq \varepsilon$. By the triangle inequality we have $\mu(P_{\theta_1}, P_{\theta_2}) \leq 2\varepsilon$ and so the value of ρ is certainly no larger than $b_0(2\varepsilon)/2$.

This upper bound is actually sharp. Indeed, $P = \frac{1}{2}(P_{\theta_1} + P_{\theta_2})$ is a probability and

$$\mu(P, P_{\theta_1}) = \left\|\tfrac{1}{2}(P_{\theta_1} + P_{\theta_2}) - P_{\theta_1}\right\| = \tfrac{1}{2}\|P_{\theta_2} - P_{\theta_1}\| = \tfrac{1}{2}\mu(P_{\theta_2}, P_{\theta_1})$$

and similarly for $\mu(P, P_{\theta_2})$. Consequently, if $\mu(P_{\theta_1}, P_{\theta_2}) \leq 2\varepsilon$ but $|\theta_1 - \theta_2| \geq b_0(2\varepsilon) - \delta$, then radius $\mathbf{S}_\varepsilon(P) \geq b_0(2\varepsilon)/2 - \delta/2$. Since $b_-(\varepsilon) \geq$ radius $\mathbf{S}_\varepsilon(P)$, the conclusion follows. □

Remark 8.4. The proof goes through even if θ is a multiparameter and $|\theta_1 - \theta_2|$ is replaced by a discrepancy with the nesting property. See Remark 8.1.

Proof of Proposition 4.4. The proof is, in outline, the same as that for Proposition 4.3 except that it is not in a vector space setting, so new details arise.

Fix $\delta > 0$. There are θ_1, θ_2 and P so that

$$\mu(P, P_{\theta_1}) \leq \varepsilon, \qquad \mu(P, P_{\theta_2}) \leq \varepsilon \tag{8.4}$$

and

$$|\theta_1 - \theta_2| \geq 2b_-(\varepsilon) - \delta. \tag{8.5}$$

Lemma 8.1 shows that (8.4) implies

$$\mu(P_{\theta_1}, P_{\theta_2}) \leq 2\varepsilon\sqrt{1 - (\varepsilon/2)^2}. \tag{8.6}$$

Assuming this to be true, we have by definition of b_0,

$$|\theta_1 - \theta_2| \leq b_0\left(2\varepsilon\sqrt{1 - (\varepsilon/2)^2}\right); \tag{8.7}$$

from this and (8.5) it follows that

$$b_-(\varepsilon) \leq b_0\left(2\varepsilon\sqrt{1 - (\varepsilon/2)^2}\right)/2. \tag{8.8}$$

This inequality is actually an equality; pick any θ_1 and θ_2 satisfying (8.6) and also

$$|\theta_1 - \theta_2| \geq b_0\left(2\varepsilon\sqrt{1 - (\varepsilon/2)^2}\right) - \delta. \tag{8.9}$$

Lemma 8.2 shows there is a P satisfying (8.4) for this choice of θ_1, θ_2 and so

$$b_-(\varepsilon) \geq \text{radius}\, \mathbf{S}_\varepsilon(P) \geq |\theta_1 - \theta_2|/2.$$

Together with (8.9) this gives

$$b_-(\varepsilon) \geq b_0\left(2\varepsilon\sqrt{1 - (\varepsilon/2)^2}\right)/2$$

and so (4.7). □

Remark 8.5. Remark 8.4 also applies here.

It remains to establish the lemmas. The following machinery seems the shortest way to get these results.

Let P and Q be probabilities and $\nu = (P+Q)/2$. Put $\eta(P,Q) = \arccos(\langle (dP/d\nu)^{1/2}, (dQ/d\nu)^{1/2}\rangle_\nu)$, where $\langle f, g\rangle_\nu = \int fg\,d\nu$. η is the angular distance ($\eta \in [0, \pi/2]$) between $(dP/d\nu)^{1/2}$ and $(dQ/d\nu)^{1/2}$ viewed as points on the sphere $\|f\|_\nu = 1$. Thus the reader may be willing to accept, without proof, the statement that: *$\eta(P,Q)$ is geodesic distance between P and Q*, i.e., the length of the shortest path between $(dP/d\nu)^{1/2}$ and $(dQ/d\nu)^{1/2}$ that stays on the surface of the sphere $\|f\|_\nu = 1$. Formally,

$$\eta(P,Q) = \inf\int_0^1 \|\sigma'\|_\nu\,dt, \tag{8.10}$$

where the infimum is over all $L_2(\nu)$-valued $\sigma(t)$ satisfying

$$\sigma(0) = \left(\frac{dP}{d\nu}\right)^{1/2}, \qquad \sigma(1) = \left(\frac{dQ}{d\nu}\right)^{1/2}, \qquad \|\sigma\|_\nu = 1,$$

$t \to \sigma(t)$ is differentiable in $L_2(\nu)$ quadratic mean.

Indeed, the path $\sigma^*(t)$ achieving the infimum is the great circle

$$\sigma^*(t) = \frac{(1-t)\left(\dfrac{dP}{d\nu}\right)^{1/2} + t\left(\dfrac{dQ}{d\nu}\right)^{1/2}}{\sqrt{(1-t)^2 + t^2 + 2t(1-t)\cos(\eta)}}.$$

We need two facts based on η.

(1) η satisfies the triangle inequality. This is immediate from formula (8.10).

(2) If $\eta(P,Q) = \delta$, the probability $dR = (\sigma^*(1/2))^2\,d\nu$ is equidistant between P and Q,

$$\eta(P,R) = \eta(R,Q) = \delta/2.$$

We also need one fact connecting η and μ.

(3) The function κ: $[0, 2^{1/2}] \to [0, \pi/2]$ defined by

$$\kappa(t) = \arccos\left(\tfrac{1}{2}(2 - t^2)\right)$$

is smooth, monotone increasing and

$$\kappa(\mu(P,Q)) = \eta(P,Q),$$

where μ denotes Hellinger distance. This can be derived by direct calculations.

Armed with these we can prove

Lemma 8.1. *Let P, P_1, P_2 satisfy*

$$\mu(P_1, P) \le \varepsilon, \qquad \mu(P, P_2) \le \varepsilon.$$

Then

$$\mu(P_1, P_2) \le 2\varepsilon\sqrt{1 - (\varepsilon/2)^2}. \tag{8.11}$$

PROOF. Using κ, we have

$$\eta(P, P_1) \le \kappa(\varepsilon), \qquad \eta(P, P_2) \le \kappa(\varepsilon),$$

so that

$$\eta(P_1, P_2) \le 2\kappa(\varepsilon),$$

by the triangle inequality for η. Using κ^{-1}, we have

$$\mu(P_1, P_2) \le \kappa^{-1}(2\kappa(\varepsilon)),$$

which yields (8.11) by direct calculations. □

LEMMA 8.2. *If $\mu(P_1, P_2) \le \kappa^{-1}(2\kappa(\varepsilon))$ there is a probability P with*

$$\mu(P_1, P) = \mu(P, P_2) \le \varepsilon.$$

PROOF. Put $dP = (\sigma^*(1/2))^2\,d\nu$, where ν and σ^* are defined in the computation of $\eta(P_1, P_2)$. Then

$$\eta(P, P_1) = \eta(P, P_2) \le \kappa(\varepsilon)$$

and the result follows by applying κ^{-1}. □

PROOF OF PROPOSITION 4.5. We will actually prove a more general proposition. Let $\theta \in \mathbf{R}^d$ be a multiparameter. Say that P_θ is differentiable in quadratic mean at θ_0 if there exists $\boldsymbol{\eta}$, a d-tuple $(\eta_1, \dots, \eta_d)$ with $\|\eta_i\| < \infty$ such that

$$\left\| P_\theta - P_0 - (\theta - \theta_0)^{\mathrm{T}}\boldsymbol{\eta} \right\| = o(|\theta - \theta_0|),$$

where now $|\cdot|$ denotes d-dimensional Euclidean distance.

We will show that the result holds in this multiparameter setting.

Note that the existence of $b_0'(0)$ implies two things. Since b_0 is continuous at 0, we know that $\{P_\theta\}$ does not curve back on itself globally. Indeed, parameter values very far away from θ_0 do not index distributions close to P_0. Second, since the derivative of b_0 is positive, we conclude that the elements of $\boldsymbol{\eta}$ are linearly independent. Actually, from the differentiability of P_θ, we can see that b_0' is the minimal eigenvalue of the Gram matrix $\Sigma = [(\eta_i, \eta_j)]$. This eigenvalue can only be 0 if the components of η are linearly dependent. Taken together, these two things mean that P_θ is both locally and globally identifiable.

Let $Q_\theta = P_0 + (\theta - \theta_0)^{\mathrm{T}}\boldsymbol{\eta}$ be the linear approximation to $\{P_\theta\}$ at θ_0. Note that the MD functional $\theta^*(P)$ for the family $\{Q_\theta\}$ gives the coordinates of a *projection* onto the linear manifold spanned by the components of $\boldsymbol{\eta}$. Thus we have

$$Q_{\theta^*(P)} = \pi^*P,$$

where π^*P denotes the projection onto Q_θ.

$$\|Q_{\theta^*(P)} - Q_0\| = \|\pi^*P - Q_0\| = \|\pi^*P - \pi^*Q_0\| \le \|P - P_0\|,$$

where the last inequality is due to $\|\pi^*\| = 1$. Thus the bias-distortion curve of θ^* is bounded by the gauge of P_θ:

$$b_{\theta^*}(\varepsilon) \leq b_0(\varepsilon). \tag{8.12}$$

On the other hand, if $\|P - P_0\| \leq \varepsilon$,

$$\|P - P_\theta\|^2 = \|P - Q_\theta\|^2 + o(|\theta - \theta_0|^2).$$

Since $Q_{\theta^*(P)} = \pi^* P$, we can use the Pythagorean theorem:

$$\|P - Q_\theta\|^2 = \|P - Q_{\theta^*}\|^2 + \|Q_{\theta^*} - Q_\theta\|^2.$$

It follows that any minimizer $\hat{\theta}(P)$ of $\|P - P_\theta\|^2$ satisfies

$$\|P - P_{\hat{\theta}}\|^2 \leq \|P - Q_{\theta^*}\|^2 + o(|\hat{\theta} - \theta_0|^2),$$

$$\|P - Q_{\hat{\theta}}\|^2 \leq \|P - Q_{\theta^*}\|^2 + o(|\hat{\theta} - \theta_0|^2),$$

so that

$$\|Q_{\hat{\theta}} - Q_{\theta^*}\|^2 = o(|\hat{\theta} - \theta_0|^2).$$

Now as b_0 is differentiable at 0 the inequality (4.3) gives $|\hat{\theta}(P) - \theta_0|^2 = O(\|P - P_0\|^2) = O(\varepsilon^2)$. We conclude that

$$\|Q_{\hat{\theta}} - Q_{\theta^*}\|^2 = o(\varepsilon^2).$$

Note that the quadratic form $R(s, t) = (Q_s - Q_0, Q_t - Q_0)$ is positive definite: It can be represented as $\mathbf{a}^{\mathrm{T}}\Sigma\mathbf{b}$, where $\mathbf{a}$ and $\mathbf{b}$ give the coefficients of $Q_s - Q_0$ and $Q_t - Q_0$ in terms of the components of $\boldsymbol{\eta}$. But Σ is nonsingular as mentioned previously. We conclude from this fact and the last display that

$$|\hat{\theta} - \theta^*| = o(\varepsilon).$$

Consequently,

$$b(\varepsilon) \leq b_{\theta^*}(\varepsilon) + o(\varepsilon),$$

which with (8.12) gives

$$b(\varepsilon) = b_0(\varepsilon) + o(\varepsilon). \qquad \square$$

Remark 8.6. The proof goes through if the Euclidean norm $|\theta_1 - \theta_2|$ is replaced by any smooth discrepancy satisfying

$$D(\theta_1, \theta_2)^2 = (\theta_1 - \theta_2)^{\mathrm{T}}\Sigma(\theta_1 - \theta_2) + o(|\theta_1 - \theta_2|^2).$$

Here Σ is a positive definite symmetric matrix. If b_0 and b are defined relative to this discrepancy, the same result holds:

$$b(\varepsilon) = b_0(\varepsilon) + o(\varepsilon).$$

Proof of Proposition 4.6. Let $\theta_{2\varepsilon}$ denote the θ-value attaining $|F_0 - F_{\theta_{2\varepsilon}}| = 2\varepsilon$, $|\theta_{2\varepsilon} - \theta_0| = b_0(2\varepsilon)$. Without loss of generality assume $\theta_{2\varepsilon} > \theta_0$. The proof is accomplished by showing how to construct a sequence of distribution functions G_n satisfying $|G_n - F_0| \leq \varepsilon$ and $\hat{\theta}(G_n) \to \theta_{2\varepsilon}$. It then follows that $b(\varepsilon) = b_0(2\varepsilon)$ as claimed.

Pick θ_n so close to $\theta_{2\varepsilon}$ that

$$|F_{\theta_n} - F_{\theta_{2\varepsilon}}| \le 1/n. \tag{8.13}$$

From the triangle inequality and the fact that $F_{\theta_n}(x) \le F_0(x)$ for all x,

$$D^-(F_0, F_{\theta_n}) = |F_0 - F_{\theta_n}| \ge 2\varepsilon - 1/n,$$

where $D^-(G, H) = -\inf_x G(x) - H(x)$ [and $D^+(G, H) = D^-(H, G)$].

As $F_0 - F_{\theta_n}$ is a continuous function of x, there is a point y at which $F_0 - F_{\theta_n} < 1/n - 2\varepsilon$. As $F_0 - F_{\theta_n}(x)$ tends to 0 at $-\infty$ and $+\infty$, there is a largest $x < y$ at which $(F_0 - F_{\theta_n})(x) = -\varepsilon$—call this x_n—and there is a smallest $x > y$ at which $(F_0 - F_{\theta_n})(x) = \varepsilon - D^-(F_0, F_n)$—call it z_n.

Now define

$$\begin{aligned} G_n(x) &= F_{\theta_n}(x), && x \le x_n, \\ &= F_0(x) - \varepsilon, && x_n < x < z_n, \\ &= F_{\theta_n}, && z_n < x. \end{aligned}$$

Note that $F_0 \ge G_n \ge F_0 - \varepsilon$, so $|F_0 - G_n| \le \varepsilon$. Also, note that by construction

$$|G_n - F_{\theta_n}| = D^+(F_{\theta_n}, G_n) = F_{\theta_n}(z_n) - (F_0(z_n) - \varepsilon).$$

For each $\theta < \theta_n$, $F_\theta(z_n) > F_{\theta_n}(z_n)$, so

$$\begin{aligned} |F_\theta - G_n| &\ge D^+(F_\theta, G_n) \\ &\ge F_\theta(z_n) - G_n(z_n^-) \\ &\ge F_{\theta_n}(z_n) - G_n(z_n^-) \\ &= D^+(F_{\theta_n}, G_n) \\ &= |F_{\theta_n} - G_n|. \end{aligned}$$

Thus any θ minimizing $|F_\theta - G_n|$ must lie in the interval $[\theta_n, \theta_{2\varepsilon}]$. As (8.13) requires $\theta_n \to \theta_{2\varepsilon}$, this shows that

$$\sup_n \left|\hat{\theta}(G_n) - \theta_0\right| = b_0(2\varepsilon; \theta_0). \qquad \square$$

Proof of Proposition 4.7. We shall show that for each $\delta > 0$ there is a density $g_{\varepsilon,\delta}$ within ε distance of f_0 and satisfying

$$\hat{\theta}(g_{\varepsilon,\delta}) - \theta \ge (2 - \delta) b_0(\varepsilon) + o(\varepsilon). \tag{8.14}$$

As the variation distance is $\frac{1}{2}$ the L_1 distance, we may do computations in L_1 distance.

Let $g_{\varepsilon,\delta}$ be the density constructed via Lemma 8.5 and 8.6 and let ν_δ and $\{k_\theta\}$ be as in Lemma 8.5. Then

$$\int |g_{\varepsilon,\delta} - f_{\hat{\theta}}| = \int |g_{\varepsilon,\delta} - k_{\hat{\theta}}| + o(|\hat{\theta} - \theta_0|). \tag{8.15}$$

We know from Lemma 8.4 and (4.2) that $|\hat{\theta} - \theta_0| = O(\varepsilon)$, so the remainder in

(8.15) is $o(\varepsilon)$. From (8.21) there is a constant $c = c(\delta)$ so that

$$(8.16) \qquad \int |f_0 + \varepsilon\nu_\delta - k_\theta| \geq \int |f_0 + \varepsilon\nu_\delta - k_{\theta^*}| + c|\theta - \theta^*|,$$

where θ^* is as in (8.21). By (8.22)

$$\int |g_{\varepsilon,\delta} - k_{\hat\theta}| = \int |f_0 + \varepsilon\nu_\delta - k_{\hat\theta}| + o(\varepsilon)$$

and the fact that $\hat\theta$ minimizes the left-hand side of (8.15) implies that

$$\int |g_{\varepsilon,\delta} - f_{\hat\theta}| = \int |f_0 + \varepsilon\nu_\delta - k_{\theta^*}| + o(\varepsilon).$$

We conclude from (8.16) that

$$|\hat\theta - \theta^*| = o(\varepsilon).$$

On the other hand, applying the result (8.20) for θ^*, we get (8.14). □

REMARK. The proof obviously goes through for nontranslation families, where $\partial/\partial\theta\, f$ is a continuous function and in L_1.

LEMMA 8.3. *If f is a density with a derivative f' that is a continuous function in L_1, then*

$$(8.17) \qquad \int |f(t-\theta) - f(t-\theta_0) - (\theta-\theta_0)f'(t-\theta_0)| = o(|\theta-\theta_0|).$$

PROOF. As $f(t-\theta) - f(t-\theta_0) = \int_{\theta_0}^{\theta} f'(t)\,dt$, putting $h = \theta - \theta_0$ and $K_h(t) = (1/h)I_{[0,h]}(t)$, the result is equivalent to

$$\int |K_h^* f' - f'| = o(1),$$

as $h \to 0$. This is standard and follows by, e.g., the Lebesgue density theorem. □

LEMMA 8.4. *Under the same conditions,*

$$(8.18) \qquad b_0(\varepsilon) = \varepsilon/\beta + o(\varepsilon),$$

where

$$\beta = \int |f'|.$$

PROOF. Because of (8.17)

$$\int |f_\theta - f_{\theta_0}| = (\theta-\theta_0)\int |f'| + o(|\theta-\theta_0|).$$

Thus for $\int |f_\theta - f_{\theta_0}| = O(\varepsilon)$, $|\theta - \theta_0| = O(\varepsilon)$ and putting $\int |f_\theta - f_{\theta_0}| = \varepsilon$, we get

$$(\theta - \theta_0) = \varepsilon/\beta + o(\varepsilon),$$

which is (8.18). □

Lemma 8.5. *Fix $\delta > 0$. Define η by*

$$\int |f'| = (2-\delta)\int |f'| I_{\{|f'| \geq \eta\}}$$

(η exists by continuity of f'). Define

$$\nu_\delta = f' I_{\{|f'| \geq \eta\}}(2-\delta)/\beta$$

(so that $\int |\nu_\delta| = 1$). Put $k_\theta(t) = f_{\theta_0}(t) + (\theta - \theta_0) f'(t)$ (so that $\{k_\theta\}$ is a linear approximation to $\{f_\theta\}$ near $\theta = \theta_0$).

The θ^* minimizing

$$\int |f_{\theta_0} + \varepsilon\nu_\delta - k_\theta| \tag{8.19}$$

is

$$\theta^* = \theta_0 + (2-\delta)\varepsilon/\beta. \tag{8.20}$$

Proof. The quantity to be minimized can be written as $\int |\varepsilon\nu_\delta - (\theta - \theta_0) f'|$, which in more explicit terms is

$$(-\theta_0)(1 - (1/(2-\delta)))\beta + |(2-\delta)\varepsilon/\beta - (\theta - \theta_0)|\beta/(2-\delta). \tag{8.21}$$

This has its minimum at θ^* given by (8.20). □

Lemma 8.6. *For each $\varepsilon > 0$ there is a density g_ε satisfying*

$$\int |g_\varepsilon - (f_0 + \varepsilon\nu_\delta)| = o(\varepsilon), \tag{8.22}$$

$$\int |g_\varepsilon - f_0| \leq \varepsilon. \tag{8.23}$$

Proof. Put $h_\varepsilon = |\nu_\delta| I_{\{\varepsilon\nu_\delta < -f\}}$. We claim that $h_\varepsilon(x) \to 0$ a.e. as $\varepsilon \to 0$. Indeed, on the set $A = \{f = 0\}$, by continuity of f, we must have $f' = 0$ as well (because of positivity $f \geq 0$). Then $h_\varepsilon(x) = 0$ for all ε and all $x \in A$. On the set $B = \{f > 0\}$, use the fact that, by continuity, f' is bounded on compacts; for each x where $f > 0$, there is an n where

$$f(x) + (1/n) f'(x) > 0;$$

then $h_\varepsilon(x) = 0$ as soon as $\varepsilon < 1/n$. Thus $h_\varepsilon(x) \to 0$ a.e. Now $h_\varepsilon \leq |\nu_\delta|$; since $\int |\nu_\delta| = 1$, we can apply the Lebesgue dominated convergence theorem to conclude

$$\omega(\varepsilon) = \int |h_\varepsilon| = o(1).$$

To apply this, note that

$$g_\varepsilon^{(1)} = f_0 + \varepsilon\nu_\delta I_{\{\varepsilon\nu_\delta > -f\}}$$

is a positive function which is closer to f_0 than $f_0 + \varepsilon\nu$. It may not integrate to 1, but for an appropriate $B > 0$,

$$f_0 + \varepsilon\nu_\delta I_{\{\varepsilon\nu_\delta \geq -f\}} I_{\{\nu_\delta < B\}}$$

will integrate to 1 and be even closer to f_0. Call the resulting function g_ε. This is a density and satisfies (8.23) by construction. Now $\int|f_0 + \varepsilon\nu_\delta - g_\varepsilon^{(1)}| = \varepsilon\omega(\varepsilon)$. Also, $\int g_\varepsilon^{(1)} \le 1 + \varepsilon\omega(\varepsilon)$. It follows that $\int|g_\varepsilon - g_\varepsilon^{(1)}| \le \varepsilon\omega(\varepsilon)$ and so

$$\int \left| g_\varepsilon - (f_0 + \varepsilon\nu_\delta) \right| \le 2\varepsilon\omega(\varepsilon).$$

As $\omega(\varepsilon) = o(1)$, this establishes (8.22). □

Proof of Proposition 5.1. The discrepancy D_{aff} is easily seen to have the nesting property of Remark 8.1. Therefore, relations (4.3) and (4.6) continue to hold in this multiparameter setting. The results (4.7) and (4.8) hold also in this setting; see Remark 8.4. The breakdown properties of the estimate follow from these remarks.

Think of the parameter $\theta = (t, S)$ as a vector in $\mathbf{R}^{d+d^2}$ gotten by concatenating the coefficients of t and S in the standard basis together into one long vector. The function $d(\theta) = D_{\text{aff}}(\theta, (0, I))$ can then be viewed as a function on $\mathbf{R}^{d+d^2}$, which is smooth at $\theta = (0, I)$ and admits the expansion

$$d(\theta)^2 = (\theta - (0, I))^{\mathrm{T}}\Sigma(\theta - (0, I)) + o\left(|\theta - (0, I)|^2\right),$$

where Σ is a positive definite symmetric matrix and $|\theta_1 - \theta_2|$ represents the Euclidean norm on $\mathbf{R}^{d+d^2}$. Remark 8.6 applies, and we conclude that the result established in the proof of Theorem 4.5 applies here. Invoking regularity of certain derivatives of $\phi_\theta^{1/2}$, we conclude that if μ is Hellinger distance,

$$b(\varepsilon) = b_0(\varepsilon) + o(\varepsilon),$$

as promised. □

Proof of Lemma 5.2. If S is symmetric about t and the norm has reflection and translation invariance, then

$$\begin{aligned}\mu(S, P) &= \tfrac{1}{2}\left[\|S - P\| + \|S - P(2t - \cdot)\|\right]\\ &\ge \tfrac{1}{2}\|P - P(2t - \cdot)\|.\end{aligned}$$

So we have a lower bound on how close to P S can be. Now $S_t = (P + P(2t - \cdot))/2$ is symmetric about t and attains the lower bound

$$\|S_t - P\| = \|S_t - P(2t - \cdot)\| = \tfrac{1}{2}\|P - P(2t - \cdot)\|.$$

It follows that if one minimizes this last expression over t, then (t, S_t) gives the best fitting symmetric distribution to P. □

Proof of Proposition 5.3. By definition,

$$b_0^{(\mathrm{sp})}(\varepsilon) = \sup\{|t - \theta_0| : \text{there is a } P \text{ symmetric about t with } \mu(P, P_0) \le \varepsilon\}.$$

By Lemma 5.2, for each t, the P symmetric about t that is closest to P_0 is given by $S_t = (P_0 + P_0(2t - \cdot))/2$ with

$$\mu(S_t, P_0) = \mu(P_0, P_{2t})/2,$$

so

$$b_0^{(\mathrm{sp})}(\varepsilon) = \sup\{|t - \theta_0| : \mu(P_0, P_{2t}) \le 2\varepsilon\}$$
$$= b_0(2\varepsilon)/2. \qquad \square$$

PROOF OF PROPOSITION 5.4 (sketch). The result for the μ = Hellinger case is proved in Donoho and Liu (1987). The result for the others can be gotten by showing $\gamma^* = 2b_0'(0)$. To do this, one only needs to exhibit an example approaching the bound closely. We give an idea how to do this for variation distance. We will work with L_1 distance, which is proportional.

Let $g = f_0 + \varepsilon h$, where f_0 is a symmetric density, $\int|h| = 1$, f_0 and h both have continuous derivatives in L_1.

The closest symmetric density to g is gotten by solving for t in

$$\min_t \int |g - S_t|,$$

where $S_t(x) = (g(x) + g(2t - x))/2$. Now a linear approximation to the family $\{S_t\}$ at $t = \theta_0$ is

$$\tilde{k}_\theta = S_0 + (\theta - \theta_0)\, \partial/\partial t S_t.$$

An analysis similar to that of Theorem 4.7 will show that the minimizer θ^{**} of

$$\int |g - \tilde{k}_\theta|$$

is only $o(\varepsilon)$ away from $\hat{\theta}(g)$.

Consider now the case where k is a very good approximation to the function ν_δ introduced in Lemma 8.5: h being asymmetric (like ν_δ), smooth (unlike ν_δ) and

$$\int |h - \nu_\delta| < \varepsilon^2.$$

In this case $\tilde{k}_\theta = f_0 - (\theta - \theta_0)[f'(x) + \varepsilon h'(-x)]$. Define $k_\theta = f_0 + (\theta - \theta_0)f'(x)$. One can show, although we do not, that the minimizer of

$$\int |g - k_\theta|$$

is only $o(\varepsilon)$ away from that for $\int|g - \tilde{k}_\theta|$. But for this last expression Lemma 8.5 implies that the minimizer θ^* of

$$\int |f_0 + \varepsilon\nu_\delta - k_\theta|$$

satisfies

$$\theta^* = \theta_0 + 2b_0(\varepsilon) + o(\varepsilon).$$

As in Lemma 8.6, one then shows that h can be chosen so that in addition to being close to ν_δ, $f_0 + \varepsilon h$ is a density. $\square$

Proof of Proposition 5.5. The invariance of b and b_- is obvious. Recall that

$$b_-(\varepsilon) = \tfrac{1}{2} \sup_{\theta_0, \theta_1} \bigl\{ |\theta_0 - \theta_1| : \text{there is a P such that } \delta\bigl(P; P_{\theta_0}\bigr) \le \varepsilon \text{ and } \delta\bigl(P; P_{\theta_1}\bigr) \le \varepsilon \bigr\}. \tag{8.24}$$

Now let P satisfy $\delta(P; P_0) \le \varepsilon$. If $\hat{\theta}$ is a version of the minimum CvM functional, then

$$\delta(P; P_{\hat{\theta}}) \le \delta(P; P_0) \le \varepsilon.$$

Comparing this with (8.24), we see that $|\hat{\theta} - \theta_0| \le 2b_-(\varepsilon)$, i.e., $b \le 2b_-$.

The regularity conditions referred to in the statement of the result are that f_0 has two continuous derivatives and

$$\int |f_0 f_0'| < \infty, \tag{8.25a}$$

$$\int |f_0''| < \infty, \tag{8.25b}$$

$$\int |f_0 f_0'' + (f_0')^2| < \infty, \tag{8.25c}$$

$$\int |f_0'| < \infty. \tag{8.25d}$$

We remark that these conditions are not independent, nor best possible.

By the assumption on $b_0'(0)$, for small enough ε, if $\delta(P; P_0) \le \varepsilon$, the minimum CvM functional is the unique root of $\lambda_{\mathrm{CvM}}(\theta) = 0$ on a small interval I about θ_0. Here

$$\lambda_{\mathrm{CvM}}(\theta) = 2\int (F - F_\theta) f_\theta^2 - \int (F - F_\theta)^2 f_\theta'.$$

Compare Theorem 5 in Donoho and Liu (1985). Now a standard Taylor series argument will give that for $|F - F_0| \le \varepsilon$, and all small enough ε,

$$\hat{\theta}(F) = \theta_0 + \frac{\lambda_{\mathrm{CvM}}(F, \theta_0)}{l(F_0, \theta_0)} + R(F),$$

where

$$l(F, t) = 2\int f_t^3 - 6\int (F - F_t) f_t f_t' + \int (F - F_t)^2 f_t''$$

and

$$|R(F)| \le R_1(\varepsilon) + R_2(\varepsilon), \tag{8.26}$$

where

$$R_1(\varepsilon) = C_1 \sup_{|t - \theta_0| \le 2b_0(\varepsilon)} \left| \frac{l(F, t) - l(F_0, \theta_0)}{l(F_0, \theta_0)} \right|^2,$$

$$R_2(\varepsilon) = \sup_{|t - \theta_0| \le 2b_0(\varepsilon)} \left| \frac{\lambda_{\mathrm{CvM}}(F, \theta_0)}{l(F_0, \theta_0)} \frac{l(F, t) - l(F_0, \theta_0)}{l(F_0, \theta_0)} \right|.$$

Analysis of $l(F, t) - l(F_0, \theta_0)$ shows that it is

$$\leq |F - F_0| \int |f_0 f_0'| + |t - \theta_0| \int \left| (f_0)^2 + f_0 f_0'' \right|$$

$$+ |F - F_0|^2 \int |f_0''|,$$

where $|F - G|$ denote the Kolmogorov distance, so that both R_1 and R_2 are bounded by $C_2 |F - F_0|^2$ for $|F - F_0|$ small enough.

Thus

$$\hat{\theta}(F) = \theta_0 + \frac{\lambda_{\mathrm{CvM}}(F, \theta_0)}{l(F_0, \theta_0)} + O\left(|F - F_0|^2\right).$$

Now using the inequality $\delta(F, F_0) \geq 3^{-0.5} |F - F_0|^{1.5}$ [see Donoho and Liu (1988)]

$$\hat{\theta}(F) = \theta_0 + \frac{\lambda_{\mathrm{CvM}}(F, \theta_0)}{l(F_0, \theta_0)} + o(\delta(F; F_0)).$$

Define

$$\lambda_0(F, \theta) = 2 \int (F - F_\theta) f_0^2.$$

Then, using

$$\int (F - F_\theta)^2 f_\theta' \leq |F - F_\theta|^2 \int |f_\theta'|,$$

we have by (8.25d) that

$$\lambda_{\mathrm{CvM}}(F, \theta_0) = \lambda_0(F, \theta_0) + o(\delta(F; F_0)),$$

so that

$$\hat{\theta}(F) = \theta_0 + \frac{\lambda_0(F, \theta_0)}{l(F_0, \theta_0)} + o(\delta(F; F_0)).$$

Now notice that $\delta(F; F_0) = \|F - F_0\|_{L_2(P_0)}$. Therefore

$$b_0(\varepsilon; \theta_0, \mathrm{CvM}) = b_0(\varepsilon; \theta_0, L_2(P_0)).$$

We know from Corollary 5 that θ^*, the minimum $L_2(P_0)$ estimator satisfies

$$b_{\theta^*}(\varepsilon) = b_0(\varepsilon) + o(\varepsilon).$$

But we can also establish

$$\theta^*(F) = \theta_0 + \frac{\lambda_0(F, \theta_0)}{l(F_0, \theta_0)} + o\left(\|F - F_0\|_{L_2(P_0)}\right).$$

It then follows that

$$\theta^*(F) - \hat{\theta}(F) = o(\delta(F; F_0))$$

and so

$$b_{\hat{\theta}}(\varepsilon) = b_0(\varepsilon) + o(\varepsilon). \qquad \square$$

PROOF OF PROPOSITION 6.1. For all the weak metrics, we have a Glivenko–Cantelli result,

$$\mu(P_n, P) \to 0 \quad \text{a.s.}$$

This combined with the triangle inequality and the fact that $\mu(P, P_0) \le \varepsilon$ gives (6.2).

For the strong metrics, a little more subtlety is necessary. Let μ be the one- or d-dimensional variation distance. Note that

$$\mu(K_{h_n} * P_n, K_{h_n} * P) \to 0 \quad \text{a.s.,} \tag{8.27}$$

for every P, not just those with density, provided (6.4) holds. Devroye and Gyorfi (1984) only claim (8.27) for all P with density, but a close inspection of the proof reveals that they have actually proved (8.27) for all P. Now

$$\mu(K_{h_n} * P_n, P_0) \le \mu(K_{h_n} * P_n, K_{h_n} * P) + \mu(K_{h_n} * P, K_{h_n} * P_0) + \mu(K_{h_n} * P_0, P_0).$$

By (8.27) the first term on the right-hand side tends to 0. By convexity and translation invariance of μ, the second term is bounded by $\mu(P, P_0)$, which by assumption is bounded by ε. If P_0 has a density, then $\mu(I_{h_n} * P_0, P_0) \to 0$ [see again Devroye and Gyorfi (1984)]. Combining these facts, the lim sup of the left-hand side is not larger than ε. The result for Hellinger distance follows from the relation Hellinger^2 ≤ variation. □

PROOF OF PROPOSITION 6.4. Using (6.10) gives a finite bound on $\hat{\theta}(\hat{P}_n^a) - \hat{\theta}(P)$, unless

$$\varepsilon_a \ge \frac{\varepsilon^* - \mu(\hat{P}_n, P_0)}{1 - \mu(\hat{P}_n, P_0)}.$$

So the finite-sample breakdown point is at least as big as the right-hand side of this display. But as $\mu(\hat{P}_n, P_0) \to 0$ a.s. we have

$$\liminf_n \varepsilon_a^* \ge \varepsilon^*$$

as claimed.

The result for ε_r^* is proved in a similar fashion, replacing the bound (6.10) by

$$\hat{\theta}(\hat{P}_n^r) - \hat{\theta}(P) \le b(\mu(\hat{P}_n, P_0) + \varepsilon_r). \qquad \square$$

Acknowledgments. We would like to thank the organizers of the meeting on robustness at Oberwolfach 1984 for the opportunity to present this work. The authors would like to thank L. Le Cam and P. J. Bickel for helpful discussions and Paul Robertson for help in preparing the manuscript.

REFERENCES

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

BICKEL, P. J. (1981). Quelques aspects de la statisque robuste. *Ecole d'Été de Probabilités de Saint Flour IX, 1979. Lecture Notes in Math.* **876** 1–72. Springer, Berlin.

BOOS, D. J. and TAMURA, R. (1985). Minimum Hellinger distance estimates of multivariate location and covariance. Unpublished.

DEVROYE, L. P. and GYORFI, L. (1984). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.

DONOHO, D. L. (1982). Breakdown properties of multivariate location estimates. Ph.D. qualifying paper, Harvard Univ.

DONOHO, D. L. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 157–184. Wadsworth, Belmont, Calif.

DONOHO, D. L. and LIU, R. C. (1987). Geometrizing rates of convergence. I. Unpublished.

DONOHO, D. L. and LIU, R. C. (1988). Pathologies of some minimum distance estimators. *Ann. Statist.* **16** 587–608.

DONOHO, D. L., ROUSSEEUW, P. and STAHEL, W. (1988). The breakdown point and the exact fit property. Unpublished.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis, Univ. California, Berkeley.

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

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

HOLM, S. (1976). Discussion of Bickel, P. J. *Scand. J. Statist.* **3** 158–161.

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

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

KOZEK, A. (1982). Minimum Lévy distance estimation of a translation parameter. Technical Report No. 70, Univ. Cologne.

MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.

MILLAR, P. W. (1981). Robust estimation via minimum distance methods. *Z. Wahrsch. verw. Gebiete* **55** 73–89.

PARR, W. C. and SCHUCANY, W. R. (1980). Minimum distance and robust estimation. *J. Amer. Statist. Assoc.* **75** 616–624.

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

ROUSSEEUW, P. J. (1981). A new infinitesimal approach to robust estimation. *Z. Warhsch. verw. Gebiete* **56** 127–132.

STAHEL, W. (1981). Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. thesis, Swiss Federal Institute of Technology, Zurich.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720