The Annals of Statistics
1988, Vol. 16, No. 2, 587–608

PATHOLOGIES OF SOME MINIMUM DISTANCE ESTIMATORS

BY DAVID L. DONOHO[1] AND RICHARD C. LIU

University of California, Berkeley

Minimum distance estimates are studied at the $N(\theta, 1)$ model. Estimates based on a non-Hilbertian distance μ (μ = Kolmogorov–Smirnov, Lévy, Kuiper, variation and Prohorov) can exhibit very large variances, or even outright inconsistency, at distributions arbitrarily close to the model in terms of μ-distance. For Hilbertian distances (μ = Cramér–von Mises, Hellinger) this problem does not seem to occur. Geometric motivation for these results is provided.

1. Introduction. Folklore has it that minimum distance (MD) estimators are "automatically" consistent and root-n consistent when the model holds. For some general results to this effect, see Section 6 of Donoho and Liu (1988). This paper shows that this "automatic" niceness need not hold when the model is only approximately true. We will give examples where the asymptotic variance of some MD estimators is arbitrarily large at distributions arbitrarily near the normal $(\theta, 1)$ model. We give examples of other MD estimators which are actually *inconsistent* at distributions arbitrarily close to the model, in the sense of not having an asymptotic limit. The distributions causing these pathologies are, in each case, symmetric about θ.

These pathologies all involve non-Hilbertian distances, such as the Kolmogorov and Prohorov distances, for which the minimum distance projection is highly nonlinear. This nonlinearity can be understood geometrically, by considering the faceted "shape" of the corresponding neighborhoods. Similar geometric considerations show that these pathologies do not generally occur for MD estimators based on Hilbertian distances, where the MD projection is generally linear. We give two examples of Hilbertian MD estimates—the Cramér–von Mises and Hellinger—where these pathologies do not occur.

It is perhaps worth remarking that for good M-estimators these pathologies do not occur. For example, the Huber M-estimators are consistent at every symmetric distribution and have bounded asymptotic variance over various neighborhoods of the model. Compare Huber (1964) and Bickel (1981). Freedman and Diaconis (1982) have shown that for some poorly tuned M-estimates, e.g., Tukey biweights with c less than the recommended values, a similar inconsistency phenomenon will occur. However, even these M-estimators are consistent near the model: The Freedman–Diaconis phenomenon occurs only at distributions very far from the model.

Received October 1985; revised August 1987.

[1] Supported by an NSF Postdoctoral Fellowship in the Mathematical Sciences.

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

Key words and phrases. Inconsistency, Hadamard and Fréchet differentiability of statistical functionals.

Some notation. We use the same notation as in Donoho and Liu (1988). We observe $X_1, X_2, \ldots$ i.i.d. according to some unknown probability P. We form an empirical estimate $\hat{P}_n$ of P based on $X_1, \ldots, X_n$ (more details follow). P_θ denotes the Normal $(\theta, 1)$ model and we form a minimum distance estimate by finding any solution $\hat{\theta}_n$ to

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

where μ is one of the distance functions chosen from the list:

Distance	$\mu(P, Q)$	Weak?	Hilbertian?
Kolmogorov	$\sup_{A=(-\infty, t]} \lvert P(A) - Q(A) \rvert$	yes	no
Cramér–von Mises	see (4.12)	yes	yes
Kuiper	$\sup_{A=(a, b]} \lvert P(A) - Q(A) \rvert$	yes	no
Lévy	$\inf\{\delta: P(-\infty, t] \le Q(-\infty, t+\delta] + \delta\}$	yes	no
Prohorov	see (4.8)	yes	no
Variation	$\sup_{\text{measurable } A} \lvert P(A) - Q(A) \rvert$	no	no
Hellinger	see (4.13)	no	yes

Of course, other choices of μ are possible, but we do not study them in this paper.

The form of $\hat{P}_n$ depends on μ. If μ is a "weak" metric, $\hat{P}_n$ will be just the empirical measure $P_n = n^{-1}\Sigma_1^n \delta_{x_i}$. If μ is a strong metric, $\hat{P}_n$ will be a smoothed empirical

$$\hat{P}_n = K_{h_n}^* P_n, \tag{1.2}$$

where K_h denotes a kernel scaled so that it is a density supported on $[-h, h]$, and the "bandwidth" h_n is chosen so that $h_n \to 0$ and $nh_n \to \infty$. In either case, $\hat{P}_n$ is μ-consistent for P if the model holds, and as in Donoho and Liu (1988) this implies consistency of $\hat{\theta}$ when the model holds.

These MD estimators have been discussed in one or another of the following: Beran (1977), Holm (1976), Kozek (1982), Millar (1981), Parr and Schucany (1980), Rao, Schuster and Littell (1975).

After Millar, we call the Cramér–von Mises discrepancy and the Hellinger distance *Hilbertian* and the other metrics *non-Hilbertian*. Without going into technicalities, it is clear that there is one major difference between these two classes of metrics: A Hilbertian neighborhood is round like a sphere in Euclidean space, while the non-Hilbertian neighborhoods are faceted like a cube or diamond.

2. Pathologies of some non-Hilbertian MD estimators.

There are two kinds of pathologies we shall demonstrate: asymptotic variance unbounded over any small μ-neighborhood of the model and inconsistency at some distributions in any small μ-neighborhood.

Unbounded variance. Consider the minimum Kolmogorov–Smirnov distance (MKSD) estimator of $N(\theta, 1)$ location. At the normal model its asymptotic distribution involves a nonlinear functional of the Brownian bridge [Rao, Schuster and Littell (1975)], so the asymptotic distribution is non-Gaussian. It has a Monte Carlo variance of 1.08 at the standard normal, as reported by Parr and Schucany (1980), so it is 92% efficient when the model holds, and its sampling behavior appears quite reasonable.

THEOREM 1. *In a Kolmogorov–Smirnov (KS) neighborhood of the $N(\theta, 1)$ model, there are many distributions at which the minimum KS estimator is asymptotically normal. Among these, there are distributions at which the asymptotic variance is as large as desired. Consequently, the asymptotic variance of the MKSD estimator is unbounded in every KS neighborhood of the model.*

All proofs are given in Section 4. The proof shows that the phenomenon is caused by distributions which agree with the model except in the tails. It suggests that the variance of the estimator would be adversely affected by outliers and by clumps of observations located slightly beyond the $\pm 3\sigma$ points of the normal curve. Tukey calls such slightly aberrant data "fringeliers." The proof is based on the Hadamard differentiability technique developed by Reeds and explained in Fernholz (1983).

This phenomenon also occurs for MD estimates based on Lévy and Kuiper distances.

THEOREM 2. *In every Lévy (resp. Kuiper) neighborhood of $N(\theta, 1)$ there are distributions at which the minimum Lévy (resp. Kuiper) distance estimate of location is asymptotically normal with an arbitrarily large variance.*

It is proved in the same way as Theorem 1 so we do not give the proof here.

This phenomenon can be understood via Figure 1. The figure portrays a situation in $\mathbb{R}^2$ analogous to minimum distance estimation. Think of the "empirical distribution" as a random point in $\mathbb{R}^2$, the "parameter family" as a straight line in $\mathbb{R}^2$ and the "estimated distribution" as the point on that line

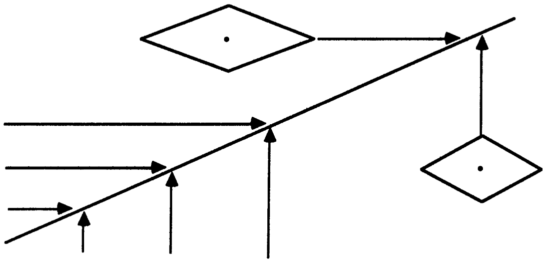

FIG. 1.

closest to the "observed distribution." Here closest is measured in terms of a norm on $\mathbb{R}^2$ which, like the Kolmogorov–Smirnov norm, has a unit ball which is faceted rather than rotund. The figure shows the level sets of the MD functional in this case. On one side of the parameter family, the level sets are horizontal lines; on the other, they are vertical lines. Thus the "MD functional" is piecewise linear, but with very different linear pieces joining together at the parameter family. Note that the level sets on the "upper side" of the parameter family make a very small angle with the family, so that small variations in the position of the "observed distribution" lead to large variation in the functional's value. The proof of Theorem 1 rigorously establishes much of the intuition fostered by this sketch. It shows that the MD functional is Hadamard differentiable (i.e., nearly linear) near (but not on) the parameter family, that very different differentials occur near (but not on) the family and that some of these differentials represent linear approximations to the MD functional which make arbitrary small angles with the parameter family, i.e., have arbitrarily large asymptotic variances.

Inconsistency. A second striking phenomenon happens for the MD estimators based on variation distance. As in Donoho and Liu (1988), the estimator is consistent when $\hat{P}_n$ is a smoothed empirical as described in the introduction and the model holds.

THEOREM 3. *Let μ = variation distance. In every μ-neighborhood of $N(0,1)$ there is a density with a nonunique μ-closest member of the $N(\theta,1)$ family. The density is symmetric and there are two solutions of the MD equation in the population. These are situated symmetrically about zero. When $\hat{P}_n$ is smoothed as in* (1.2), *these two solutions are the limit points of the sequence of MD estimates: The estimates oscillate with increasing sample size and do not converge in probability or almost surely.*

Minimum Prohorov distance estimates exhibit similar behavior.

THEOREM 4. *Let μ = Prohorov distance. In every μ-neighborhood of $N(0,1)$ there is a distribution with a nonunique μ-closest member of the $N(\theta,1)$ family. The distribution is symmetric and there are two solutions of the MD equation in the population. The solutions are symmetrically placed about zero and are separated from zero. Thus zero is not a limit point of the MD estimate with increasing sample size; in fact, the estimate oscillates with increasing sample size and does not converge either almost surely or in probability.*

This phenomenon can be understood from Figure 2, which sketches a curve portraying the parameter family. The closest point on this family is found with a norm on $\mathbb{R}^2$ which has a nonrotund unit ball. Because of the curvature of the parameter family, the closest points on the curve to the point marked P are actually closer to P than any points in between the two in terms of parameter

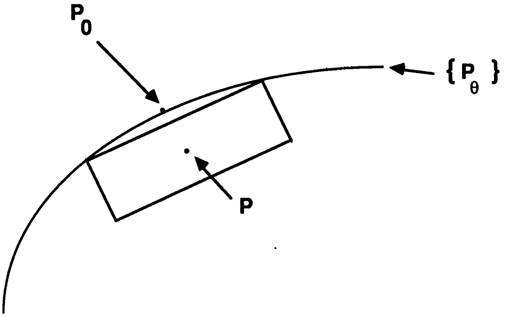

FIG. 2.

values. The proofs of Theorems 3 and 4 show that exactly this sort of phenomenon occurs for the minimum variation and minimum Prohorov estimates.

Interpretation. The practical effect in all these cases is that the estimator does not converge at a uniform root-n rate to an asymptotic limit over a neighborhood of the model. There is no constant C such that, for example,

$$\lim_n P\left\{\left|\hat{\theta}(\hat{P}_n) - \hat{\theta}(P)\right| \leq C/\sqrt{n}\right\} \geq 1 - \alpha$$

for all P in a small μ neighborhood of the model. Minute variations in the underlying distributions can cause enormous changes in the length of confidence intervals.

3. Hilbertian discrepancies. We have shown, then, that all the non-Hilbertian MD estimates in our list can have terrible sampling behavior quite near the model. For Hilbertian distances, it appears that better properties hold. We give two results in this direction.

THEOREM 5. *Throughout every small enough Cramér–von Mises (CvM) neighborhood of the $N(\theta, 1)$ model, the minimum CvM distance functional is uniquely defined, continuous and Fréchet differentiable. The MD estimate is root-n consistent and asymptotically normal at every distribution in such a neighborhood. The asymptotic variance of the estimator over an ε-neighborhood of the model is bounded by $v(\varepsilon)$ satisfying*

$$\lim_{\varepsilon \to 0} v(\varepsilon) = v(0) \tag{3.1}$$

where $v(0)$ denotes the variance at the model.

We note that the Fréchet differentiability mentioned in the proof is with respect to the Kolmogorov distance. It turns out to be rather messy to give an explicit evaluation of $v(\varepsilon)$ in the CvM case.

As for minimum Hellinger distance estimation, we know of no proof that at each distribution in a small Hellinger neighborhood about the $N(\theta, 1)$ model, the Hellinger MD is asymptotically normal. However, we can say the following.

THEOREM 6. *Throughout every (small enough) Hellinger neighborhood of the* $N(\theta, 1)$ *model, the minimum Hellinger distance functional is uniquely defined, continuous and Fréchet differentiable. The MHDE is consistent at every density in such a neighborhood of the model. The formal expression for the asymptotic variance of the MHDE is bounded above over Hellinger* ε*-neighborhoods by a* $v(\varepsilon)$ *satisfying* (3.1).

The technical components of this result are all present in Beran (1977). Explicit evaluation of $v(\varepsilon)$ and comparison with the best possible (among regular estimators) $v(\varepsilon)$ behavior is given in Liu (1987).

In both proofs, a key condition is that $\theta \to \mu(P, P_\theta)$ has a unique global minimum and nonsingular quadratic behavior at the minimum. As the proofs show, when this condition holds, the MD functional is continuous and differentiable.

It would be interesting to have some quantitative idea of the size of the neighborhoods of the model over which this condition holds. In the case where μ = Hellinger distance, geometric insight into this question is available. Let θ be the one-dimensional parameter of the (not necessarily location) parameter family $\{P_\theta\}$ and let p_θ denote the density of P_θ. We introduce two geometrically derived quantities: ρ, the minimum radius of curvature of $\theta \to p_\theta^{1/2}$ viewed as a curve in $\mathbb{L}_2$ and χ, the minimum distance between members $P_{\theta_1}, P_{\theta_2}$ of the parameter family at which the distance $\mu(P_{\theta_1}, P_{\theta_2})$ has a critical point (i.e., both partial derivatives $\partial/\partial\theta_1$ and $\partial/\partial\theta_2$ vanish). It turns out that $\theta \to \mu(P, P_\theta)$ has a unique nonsingular quadratic global minimum at $\hat{\theta}(P)$ whenever

$$\mu(P, P_{\hat{\theta}}) < \min(\rho, \chi/2).$$

This can be motivated by Figures 3 and 4. The first shows that if $\mu(P, P_{\hat{\theta}}) < \rho$, the ball of radius $\mu(P, P_{\hat{\theta}})$ makes only first-order contact with the parameter family at $\hat{\theta}$, and so the distance function has a nonsingular quadratic local minimum at $\hat{\theta}$. The second shows that P cannot have a nonunique closest

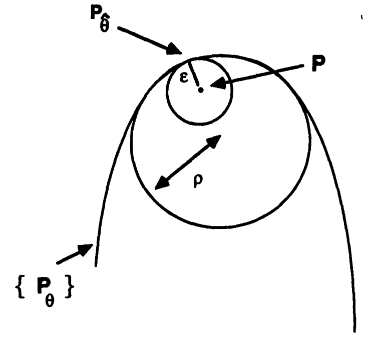

FIG. 3. *The distance function* $\mu(P, P_\theta)$ *has a nonsingular quadratic minimum if* $\varepsilon < \rho$. *The large circle makes second-order contact to the family at the projection of P. Its radius* ρ *is the radius of curvature. The distance from P to its projection is* ε. *The distance function* $\mu(P, P_\theta)$ *has a nonsingular quadratic minimum if* $\varepsilon < \rho$.

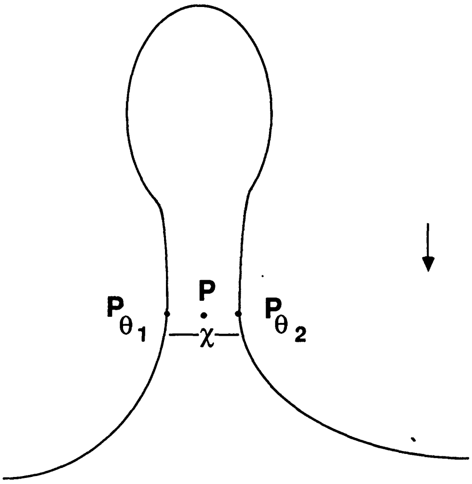

FIG. 4. *The distance function* $\mu(P_{\theta_1}, P_{\theta_2})$ *has a critical point at distance* χ. *P, equidistant between* P_{θ_1} *and* P_{θ_2}, *has a nonunique projection.*

member of the family P_θ unless P is farther away from the family than $\sqrt{2 - \sqrt{4 - \chi^2}}$; so in this case the global minimum is unique.

In short, throughout any tube of radius $\varepsilon < \min(\rho, \sqrt{2 - \sqrt{4 - \chi^2}}$ about the parameter family $\{P_\theta\}$, the key condition for continuity and differentiability of the MHD functional will hold. At the normal location model, we can compute $\chi = \sqrt{2}$ and $\rho = 1/\sqrt{3}$, so the MHD functional is well-behaved over large neighborhoods of the model in that case.

We would like to get some idea of the size of the CvM neighborhoods over which the minimum CvM functional is continuous and differentiable. However, we have not been able to obtain any geometric insight into this question.

4. Proofs.

PROOF OF THEOREM 1. It will be convenient to state the proof in the language of distribution functions. Let Φ denote the distribution function of $N(\theta, 1)$. First, note that there are distribution functions in an ε KS-neighborhood of Φ equal to Φ in the middle of their range, and so that $F - \Phi$ has a unique locally quadratic maximum in the left tail and a unique locally quadratic minimum in the right tail.

For example, let $z = \Phi^{-1}(\varepsilon)$ and define the distribution F with density f via

$$f(x) = \begin{cases} \dfrac{\varepsilon}{1-\varepsilon}\phi(x - 2z), & x < z, \\ \phi(x), & x \in [z, -z], \\ \dfrac{\varepsilon}{1-\varepsilon}\phi(x + 2z), & x > -z, \end{cases}$$

where ϕ denotes the $N(0, 1)$ density.

By Lemma 4.1, it makes sense to define $\hat{\theta}(G)$ as the unique root of

$$\Lambda(G, \theta) = 0;$$

the minimum KS distance estimator is $\hat{\theta}_n = \hat{\theta}(F_n)$, where F_n is the empirical distribution function.

We follow the approach of Reeds to prove that $\hat{\theta}_n$ is asymptotically normal; that is, we prove that $\hat{\theta}$ is a Hadamard differentiable functional at the F introduced previously.

By Lemma 4.2, Λ is Hadamard differentiable at (F, θ_0), where $\theta_0 = 0$ is the root of $\Lambda(F, \theta_0) = 0$ and differentiability is in the sense of Fernholz (1983) with respect to $C[0,1]$. The Lipschitz bounds established in Lemma 4.3 allow us to apply the implicit function theorem of Fernholz [(1983), Theorem 6.2.1] to conclude that $\hat{\theta}(G)$ is Hadamard differentiable at F over $C[0,1]$. Let $F_n^* \in C(\mathbb{R})$ be the smoothed version of the empirical distribution defined in Lemma 4.4. Define $\hat{\theta}_n^*$ as the root

$$\Lambda\left(F_n^*, \hat{\theta}_n^*\right) = 0.$$

The differentiability of $\hat{\theta}$ over $C[0,1]$ implies that

$$\left(\hat{\theta}_n^* - \theta_0\right) = \mu_F(F_n^* - F) + o_p\left(|F_n^* - F|\right).$$

Here μ_F is the linear functional

$$\mu_F(F_n^* - F) = \frac{(F_n^* - F)(x_0) - (F_n^* - F)(x_1)}{f(x_0) + f(x_1)},$$

where x_0 is the point near $2z$ at which the maximum $(F - \Phi)(x)$ is attained and x_1 is the point near $-2z$ at which the minimum $(F - \Phi)(x)$ is attained. This means that $\hat{\theta}_n^*$ is asymptotically normal. Applying Lemma 4.4, we conclude that $\hat{\theta}_n^*$ approximates $\hat{\theta}_n$ so well that $\hat{\theta}_n$ is also asymptotically normal. Thus

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to_D N\left(0, \sigma_F^2\right)$$

with

$$\sigma_F^2 = \frac{\varepsilon(1 - 2\varepsilon)}{4f(x_0)^2}$$

[using the symmetry $f(x_0) = f(x_1)$]. Now as $f(x_0) < \varepsilon/(1 - \varepsilon)\phi(0)$ by construction, we have

$$\sigma_F^2 \geq \frac{\pi}{\varepsilon}(1 - 2\varepsilon)(1 - \varepsilon)^2$$

which is arbitrarily large for small ε. □

LEMMA 4.1. *Put*

$$D^+(H) = \sup_x H(x), \qquad |H| = \sup_x |H(x)|$$

and

$$\Lambda(F, \theta) = D^+(F - F_\theta) - D^+(F_\theta - F).$$

Suppose that $\{F_\theta\}$ is a location family $F_\theta(x) = F_0(x - \theta)$. Then:

(i) *$\Lambda(F, \theta)$ is nondecreasing in θ. It tends from -1 at $\theta = -\infty$ to 1 at $\theta = \infty$. It is strictly increasing in θ if $F_\theta(x)$ is strictly increasing in x.*

(ii) *If F_θ is uniformly continuous in x, Λ is a uniformly continuous function of θ.*

(iii) *If F_θ is uniformly continuous and strictly increasing, the minimum KS distance estimator is the unique root of*

$$\Lambda(F_n, \theta) = 0.$$

Proof. Fix F. Note that for a fixed x, $(F - F_\theta)(x)$ is increasing in θ, strictly so if $F_\theta(x)$ is strictly increasing in x.

(i) Note that as $(F - F_\theta)(x)$ is increasing in x for each fixed x, the supremum and infimum over all x are both increasing as well and strictly increasing if F_θ is strictly increasing in x. The limits at $\theta = \pm\infty$ are evident.

(ii) As $|F - G| = \max(D^+(F - G), D^+(G - F))$,

$$|D^+(F - F_\theta) - D^+(F - F_{\theta+t})| \le |F_\theta - F_{\theta+t}|,$$

$$|\Lambda(F, \theta) - \Lambda(F, \theta + t)| \le 2|F_\theta - F_{\theta+t}|.$$

As F_θ is a location model, the uniform continuity of F_0 in x bounds the right side of these inequalities.

(iii) Note that

$$|F - F_\theta| = \max(D^+(F - F_\theta), D^-(F_\theta - F)). \tag{4.1}$$

As $D^+(F - F_\theta)$ is strictly increasing from 0 at $\theta = -\infty$ to 1 at $+\infty$, and $D^+(F_\theta - F)$ is strictly decreasing from 1 at $-\infty$ to 0 at ∞ and both are continuous in θ, there is exactly one value of θ, $\hat{\theta}$, at which they are equal. $\hat{\theta}$ evidently minimizes (4.1), but it also satisfies

$$\Lambda(F, \hat{\theta}) = 0$$

as well. As Λ is continuous and monotone, this is the unique root of $\Lambda(F, \cdot)$. □

Lemma 4.2. *Let F and F_0 be continuous, and suppose that $\Delta = F - F_0$ has unique maxima and minima (at x_0 and x_1 say). Suppose that for small enough $\varepsilon > 0$, there is $\delta > 0$ with*

$$\Delta(x) \ge \Delta(x_0) - \delta \Rightarrow |x - x_0| \le \varepsilon$$

and similarly for x_1. Suppose that F_θ is a location family $[F_\theta(x) = F_0(x - \theta)]$ and that F_0 has a uniformly continuous density. Then the functionals $D^+(H - F_0)$, $D^-(H - F_0)$ and $\Lambda(H, \theta)$ are Hadamard differentiable at F [resp. (F, θ_0)] in the sense described in Fernholz. That is, the induced functionals on $C[0, 1]$,

$$\tau^+(G) = D^+(G \circ F - F_0),$$

$$\tau^-(G) = D^-(G \circ F - F_0),$$

$$\sigma(G, \theta) = D^+(G \circ F - F_\theta) - D^-(G \circ F - F_\theta),$$

are Hadamard differentiable at $G = U$ $[G = (U, \theta_0)]$, *where* U *denotes the uniform distribution on* $[0, 1]$.

PROOF. We will reduce all the cases to the Hadamard differentiability of τ^+ at U. Consider

$$\sigma^+(G, \theta) = D^+(G \circ F - F_\theta).$$

This can be written as

$$\sigma^+(G, \theta) = D^+(\Delta(G, \theta))$$

where $\Delta(G, \theta) = G \circ F - F_\theta$. We claim that $\Delta \colon D[0,1] \times \mathbb{R} \to D[0,1]$ is Fréchet differentiable with respect to the sup norm. Indeed, Δ depends in a bounded linear fashion on G; for the θ-dependence, note that

$$\begin{aligned} &\sup_u |F_\theta(u) - F_0(u) - (\theta - \theta_0) f_\theta(u)| \\ &\qquad = |\theta - \theta_0| \sup_x | f_0(x + \zeta(x)) - f_0(x)|, \end{aligned} \tag{4.2}$$

where $|\zeta(x)| \le |\theta - \theta_0|$. But f_0 is assumed uniformly continuous, so that

$$\sup_x | f_0(x + \zeta(x)) - f_0(x)| \le \sup_x \sup_{h \le |\theta - \theta_0|} | f_0(x + h) - f_0(x)| = o(1)$$

as $\theta \to \theta_0$. Thus (4.2) is $o(|\theta - \theta_0|)$ and so $\theta \to F_\theta$ is Fréchet differentiable.

As Δ is Fréchet differentiable at (U, θ_0), the chain rule of Fernholz (1983) will imply that σ^+ is Hadamard differentiable at (U, θ_0) if D^+ is at $\Delta(U, \theta_0)$. Similarly, σ^- is Hadamard differentiable under a condition on D^-. Because of the symmetry in the hypotheses of the lemma, a proof for D^+ will apply to D^- as well.

Let $u_0 = F(x_0)$. By hypothesis and continuity of F^{-1} at u_0, we have a modulus $\omega_\Delta(h)$ with

$$|u - u_0| \le \omega_\Delta(\Delta(u) - \Delta(u_0))$$

and $\omega_\Delta(h) \to 0$ as $h \to 0$ [where $\Delta \equiv \Delta(U, 0) \equiv F - F_0$]. Let $\mathbb{K}$ be a compact set in $C[0, 1]$, that is, a set of uniformly bounded and equicontinuous functions, with bound b_K and modulus

$$\omega_K(h) = \sup_{g \in \mathbb{K}} \sup_u |g(u + h) - g(u)|.$$

Take any g in $\mathbb{K}$ and consider $D^+(\Delta + tg)$. Now

$$D^+(\Delta + tg) \le D^+(\Delta) + tb_K.$$

Let v_0 be any maximizer of $\Delta + tg$, and note

$$\begin{aligned} (\Delta + tg)(v_0) &\ge (\Delta + tg)(u_0) \ge \Delta(u_0) - tb_K, \\ (\Delta + tg)(v_0) &\le \Delta(u_0) + tb_K, \\ \Delta(v_0) &\ge \Delta(u_0) - 2tb_K. \end{aligned}$$

This bound does not depend on g: It is uniform over $\mathbb{K}$.

Now as $g(v_0) - g(u_0) \le \omega_K(|v_0 - u_0|)$ we have

$$(\Delta + tg)(v_0) \le (\Delta + tg)(u_0) + (\Delta(v_0) - \Delta(u_0)) + t\omega_K(\omega_\Delta(2tb_K)).$$

Using $\Delta(v_0) \le \Delta(u_0)$ and $(\Delta + tg)(v_0) \ge (\Delta + tg)(u_0)$, we get

$$0 \le (\Delta + tg)(v_0) - (\Delta + tg)(u_0) \le t\omega_K(\omega_\Delta(2tb_K)).$$

Again, this bound does not depend on g: It is uniform over $\mathbb{K}$. Using now $D^+(\Delta + tg) = (\Delta + tg)(v_0)$, etc., the bound can be written

$$\sup_{g \in \mathbb{K}} |D^+(\Delta + tg) - D^+(\Delta) - tg(u_0)| = o(t),$$

which establishes the Hadamard differentiability of $D^+(H)$ at $H = \Delta$, with derivative $H \to (H - \Delta)(u_0)$. □

LEMMA 4.3. (i) *Let F_θ be the Gaussian translation family $F_\theta = \Phi(\cdot - \theta)$:*

$$|\Lambda(G_0, \theta) - \Lambda(G_1, \theta)| \le 8|G_0 - G_1|. \tag{4.3}$$

(ii) *Let F be as in Lemma 4.2. Let θ_0 be the root of $\Lambda(F, \theta) = 0$. For small enough $\delta > 0$, there are A and $B > 0$ with*

$$A|\theta - \eta| \le |\Lambda(G, \theta) - \Lambda(G, \eta)| \le B|\theta - \eta| \tag{4.4}$$

for all G such that $|G - F| \le \delta$ and θ, η such that $|\theta - \theta_0| \le \delta$, $|\eta - \theta_0| \le \delta$.

PROOF. (i) Note that D^+ is Lipschitz with constant 2. As Λ is a sum of two D^+ terms, the expression (4.3) involves four D^+ terms and so has a Lipschitz constant no bigger than 8.

(ii) From the Lipschitz property of D^+ we have

$$|\Lambda(G, \theta) - \Lambda(G, \eta)| \le 8|F_\theta - F_\eta|.$$

But

$$|F_\theta - F_\eta| \le \left\{\sup_x \phi(x)\right\}|\theta - \eta|,$$

where ϕ is the density of the Gaussian. It follows that we may take $B = 8/\sqrt{2\pi}$ in (4.4).

As for the lower bound, note that if $|G - F| < \delta$ and $|\theta - \theta_0| < \delta$, etc., $|G - F_\theta - F - F_{\theta_0}| < c\delta$ for some $c \le B + 1$. Put $\delta' = c\delta$. As in Lemma 4.2 we have that the maximum $D^+(G - F_\theta)$ is attained in some interval $I = [x_0 - \varepsilon', x_0 + \varepsilon']$ where $\varepsilon' = \varepsilon'(\delta')$. Now as $d/d\theta(G - F_\theta)(x) = -\phi(x - \theta)$, we have that for each $x \in I$ and each $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$,

$$\frac{d}{d\theta}(G - F_\theta)(x) < -c_0,$$

where

$$c_0 = \inf\{\phi(x) \colon x_0 - \varepsilon' - \delta \le x \le x_0 + \varepsilon' + \delta\} > 0.$$

For $\theta > \eta$ we have

$$D^+(G - F_\eta) - D^+(G - F_\theta) \geq c_0|\eta - \theta|.$$

We can argue similarly for $D^+(F_\theta - G)$, getting

$$D^+(F_\theta - G) - D^+(F_\eta - G) \geq c_1|\theta - \eta|$$

and so concluding that we may take $A = c_0 + c_1$. □

Lemma 4.4. *Let U_n^* be the distribution which places a uniform distribution of mass $(n+1)^{-1}$ in each of the $n+1$ intervals $[y_{i-1}, y_i]$ [where $y_0 = 0$, $y_{n+1} = 1$, $y_k = F(X_k)$ and $X_1, \dots, X_n$ are i.i.d. F]. Let $F_n^* = U_n^* \circ F$ and $\hat{\theta}_n^*$ be the root $\Lambda(F_n^*, \hat{\theta}_n^*) = 0$. Suppose*

$$\sqrt{n}\left(\hat{\theta}_n^* - \theta_0\right) \to_D N(0, \sigma^2).$$

Then also

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to_D N(0, \sigma^2),$$

where $\hat{\theta}_n^*$ satisfies $\Lambda(F_n^*, \hat{\theta}_n^*) = 0$.

Proof. Let $\Omega_{n,\delta}$ denote the event that $|\hat{\theta}| \leq \delta$, $|\hat{\theta}_n^*| \leq \delta$ and $|F_n - F| \leq \delta + 1/n$. Conditioning on this event, we can apply (4.4) to conclude that

$$\begin{aligned}\left|\hat{\theta}_n - \hat{\theta}_n^*\right| &\leq A^{-1}\left|\Lambda\left(F_n, \hat{\theta}_n^*\right) - \Lambda\left(F_n, \hat{\theta}_n\right)\right| \\ &= A^{-1}\left|\Lambda\left(F_n, \hat{\theta}_n^*\right)\right| \\ &= A^{-1}\left|\Lambda\left(F_n, \hat{\theta}_n^*\right) - \Lambda\left(F_n^*, \hat{\theta}_n^*\right)\right|,\end{aligned}$$

where A is the constant introduced by Lemma 4.3 in connection with (4.4). By (4.3), we have

$$\begin{aligned}\left|\hat{\theta}_n - \hat{\theta}_n^*\right| &< \frac{8}{A}|F_n - F_n^*| \\ &= \frac{8}{A}|U_n - U_n^*| \\ &< \frac{8}{nA},\end{aligned}$$

where $U_n = F_n \circ F^{-1}$.

Since by Glivenko–Cantelli $P\{\Omega_{n,\delta}\} \to 1$, we have

$$\sqrt{n}\left(\hat{\theta}_n - \hat{\theta}_n^*\right) = o_p(1).$$

The result now follows by Slutzky's theorem. □

Proof of Theorem 3. We will state this proof in the language of density functions, and do computations in L_1 distance, which is proportional to variation distance. Let ϕ denote the $N(0,1)$ density.

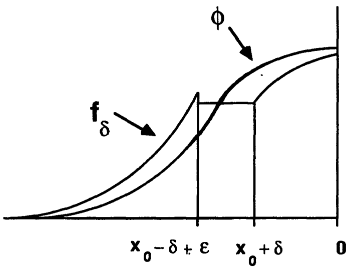

FIG. 5.

Let $\delta > 0$ and let $\varepsilon < \delta$ and $x_0 = x_0(\delta, \varepsilon)$ be numbers to determine. Define f_δ by

$$f_\delta(x) = \begin{cases} \phi(x+\delta), & x < x_0 - \delta + \varepsilon, \\ \phi(x_0), & x_0 - \delta + \varepsilon \le x < x_0 + \delta, \\ \phi(x-\delta), & x_0 + \delta < x \le 0, \end{cases}$$

for $x < 0$ and by symmetry for $x > 0$. This is a positive integrable function, actually a density when ε and x_0 are chosen appropriately. Its general appearance with the values of ε and x_0 we intend is portrayed in Figure 5. Note that f_δ is above ϕ in the tails, below ϕ in the center and flat where the two cross.

According to Lemma 4.5, with appropriate choice of ε and x_0, we can insure that f_δ is a density. The lemma also shows that

$$\Lambda(\theta) = \int |f_\delta(t) - \phi(t-\theta)|\,dt$$

is an even function, is constant on $[0, 2\delta - \varepsilon]$, is *decreasing* on $(2\delta - \varepsilon, \delta)$ and increasing on $[2\delta, \infty)$. Consequently, there are two minimizers of $\Lambda(\theta)$: $\theta = 2\delta$ and $\theta = -2\delta$. Thus there are two solutions of the MD equation in the population, as claimed.

By Lemma 4.6, the function

$$\Lambda_n(\theta) = \int |\hat{f}_n - \phi_\theta|$$

has, with high probability in large samples, its minima near either one of the two minimizers in the population. Moreover, in a given large sample, all minimizers occur (with high probability) near only one of the two population minimizers. By symmetry of f_δ and of ϕ, the law of $\Lambda_n(t)$ is invariant under reflection $t \to -t$, and so any sample minimizer is equally likely to be close to one population minimizer as the other. Let $\hat{\theta}_n$ be defined by a rule which selects the sample minimizer or which in the case of ties selects from among ties in some (measurable) way. Then the distribution of $\hat{\theta}_n$ converges to a pair of equal point masses located at the two population minimizers and $\hat{\theta}$ does not converge in probability. But $\hat{\theta}$ cannot converge almost surely either, for it would have to converge to 2δ with probability $\frac{1}{2}$ and -2δ with probability $\frac{1}{2}$; then the tail field would contain events of probability neither zero nor one—contradicting the zero–one law for independent observations. Thus the sequence of minima alternates between -2δ and 2δ and does not converge. □

Lemma 4.5. *For any value of* $\varepsilon < \delta$ *and* x_0, Λ *is minimized at* $\theta = 2\delta$ *and* $\theta = -2\delta$. *For appropriate values of* $\varepsilon < \delta$ *and* x_0, f_δ *is a density; as* $\delta \to 0$, $f_\delta \to \phi$ *in* L_1.

Proof. $\Lambda(\theta) = (2 - 2A(\theta))/2$ where $A(\theta)$ is the area under the graph of $\min(f_\delta, \phi(\cdot - \theta))$. By inspection of Figure 5, one sees that this area is constant for $\theta \in [-2\delta + \varepsilon, 2\delta - \varepsilon]$, is decreasing (increasing) on $(2\delta - \varepsilon, 2\delta]$ ($[-2\delta, -2\delta + \varepsilon)$) and increasing (decreasing) on $(2\delta, \infty)$ $[(-\infty, 2\delta)]$.

The integral of f_δ may be calculated as

$$1 + 2(2\delta - \varepsilon)\phi(x_0) + 2\int_{x_0}^{x_0+\varepsilon} \phi(t)\,dt - \int_{-\delta}^{\delta} \phi(t)\,dt.$$

Fixing δ and $\varepsilon < \delta$, this integral is a continuous function of x_0 which is greater than 1 for $x_0 > -\sqrt{\log(4)}$ and is less than 1 for x_0 large. It follows that at some $x_0^* = x_0(\varepsilon, \delta)$ it is 1. To see that as $\delta \to 0$, $f_\delta \to_{L_1} \phi$, note that

$$\int |f_\delta - \phi| = 4\int_{x_0}^{0} \phi - f_\delta$$
$$= \left(\tfrac{1}{2} - \Phi(x_0) - (\phi(x_0)\delta + \Phi(\delta) - \Phi(x_0))\right) \text{ by calculation,}$$

which is $O(\delta)$ as $\delta \to 0$. □

Lemma 4.6. Λ_n *converges uniformly to* Λ *with probability* 1. *For large enough* n, *any minimizer of* Λ_n *is close to either* 2δ *or* -2δ. *For large* n, *the global minimizers of* Λ_n *all occur near* 2δ *with probability approaching* $\frac{1}{2}$; *similarly, they all occur near* -2δ *with probability approaching* $\frac{1}{2}$.

Proof. Two applications of the triangle inequality give

$$(4.5) \qquad |\Lambda_n(\theta) - \Lambda(\theta)| \le \varepsilon_n = \int |\hat{f}_n - f_\delta|.$$

Inspection of Λ shows that there is a modulus ω_Λ with

$$(4.6) \quad \operatorname{dist}(x, \{-2\delta, 2\delta\}) \equiv \min(|x - 2\delta|, |x + 2\delta|) \le \omega_\Lambda(\Lambda(x) - \Lambda(2\delta))$$

and $\omega_\Lambda(0^+) = 0$. From (4.5) we have $\Lambda(\hat{\theta}_n) \le \Lambda(2\delta) + 2\varepsilon_n$. Therefore $\operatorname{dist}(\hat{\theta}_n, \{2\delta, -2\delta\}) \le \omega_\Lambda(2\varepsilon_n)$. By Theorem 3.1 of Devroye and Gyorfi (1984), $\varepsilon_n \to 0$ almost surely and in probability; this proves the first and second statements.

For the third statement, we only give a sketch. Similar arguments occur in Freedman and Diaconis (1982). From the last paragraph, any sample minimum of Λ_n will happen to be near either 2δ or -2δ. Pick a small neighborhood I^+ of 2δ and let M_n^+ be any minimum of Λ_n on I^+; let M_n^- be the analogous quantity for a neighborhood I^- of -2δ. We wish to show that $P\{\Lambda_n(M_n^+) \ne \Lambda_n(M_n^-)\} \to$

1. This will establish that all global minimizers of Λ_n occur only in one of I^+ or I^-. Using Hadamard differentiability of the functional $\min_{\theta<0}\Lambda(\theta)$ at f_δ and stochastic equicontinuity of the (suitably normed) sequence of processes $\{\Lambda_n(t) - \Lambda_n(2\delta), t \in I^+\}$ and $\{\Lambda_n(t) - \Lambda_n(-2\delta), t \in I^-\}$ one can justify the following. There exists a sequence of normalizing constants α_n (depending on f_δ, on the bandwidth h_n and on the kernel k) so that:

(i) The behavior of $\Lambda_n(M_n^+) - \Lambda_n(M_n^-)$ tracks that of $\Lambda_n(2\delta) - \Lambda_n(-2\delta)$ to within $o_p(\alpha_n^{-1/2})$.

(ii) We have the limit theorem

$$\alpha_n^{1/2}\{\Lambda_n(2\delta) - \Lambda_n(-2\delta)\} \to N(0,1).$$

Using these two claims, we have that for any $\eta > 0$,

$$\begin{aligned}\liminf_n P\{\Lambda_n(M_n^+) \neq \Lambda_n(M_n^-)\} &\geq \liminf_n P\{|\Lambda_n(2\delta) - \Lambda_n(-2\delta)| > \eta\alpha_n^{-1/2}\}\\ &= P\{|Z| > \eta\},\end{aligned}$$

where Z is distributed $N(0,1)$. But as $\eta \to 0$, $P\{|Z| > \eta\} \to 1$ and so

$$\lim_{n\to\infty} P\{\Lambda_n(M_n^+) \neq \Lambda_n(M_n^-)\} = 1.$$

The verification of the two claims is routine but tedious. The key idea is to establish an analog of Lemma 4.2 for the functional $\Lambda_n(\delta)$ showing that $\Lambda_n(\delta) - \Lambda(\delta)$ is approximately a linear functional of $\hat{f}_n - f_\delta$ and to invoke central limit theorem for the smoothed density estimate. □

Proof of Theorem 4. Let μ denote Prohorov distance (4.8b). Fix $\varepsilon > 0$ and put

$$F(x) = \begin{cases}\Phi(x+c), & x \leq -c,\\ \frac{1}{2}, & x \in [-c, c],\\ \Phi(x-c), & x \geq c,\end{cases}$$

where $c = c(\varepsilon)$ is chosen so that $\mu(F, \Phi) = \varepsilon$.

We claim that any μ-closest points to F among the normal distributions Φ_t are at $t = -c$ and $+c$, even though F is symmetric about 0. Also, we claim that any μ-closest point to the empirical distribution F_n in a sample of size n is very near either c or $-c$. These claims are established by Lemmas 4.7 and 4.8. From this point on the argument is the same as that of Theorem 3. □

Lemma 4.7. *Put* $\Lambda(t) = \mu(F, \Phi_t)$. *The minimizers of* $\Lambda(t)$ *are at* $-c$ *and* $+c$.

Proof. Define

$$\Delta(t,\delta) = \inf_A F[A^\delta] - \Phi_t[A], \tag{4.7}$$

where the infimum is over all measurable sets A and $A^\delta = \{x: \mathrm{dist}(x, A) \leq \delta\}$.

We claim that Prohorov distance $\mu(F, \Phi_t)$ is given by

$$\delta^* = \inf\{\delta: \Delta(t, \delta) \leq -\delta\}. \tag{4.8a}$$

This can be seen by comparing (4.7) with the definition of Prohorov distance

$$\mu(F, \Phi_t) = \inf\{\delta: \Phi[A] \leq F[A^\delta] + \delta \text{ for all } A \text{ measurable}\}. \tag{4.8b}$$

Note that $-\Delta(t, 0)$ is the variation distance between F and Φ_t and that $\Delta(t, \delta)$ is monotone increasing in δ for t fixed. Moreover, as $\delta \to \infty$, $\Delta \to 0$ (with this choice of F and Φ_t) for t fixed, so (4.8) makes sense.

To compute $\Delta(t, \delta)$ note that the densities f and ϕ_t compare as

$$f(x) \geq \phi_t(x) \quad \text{on supp}(f),\ t \in [-c, c]. \tag{4.9}$$

Let A be a set and let B be the largest subset of A with B^δ disjoint from supp(f). We claim that

$$F[A^\delta] - \Phi[A] \geq F[B^\delta] - \Phi[B]. \tag{4.10}$$

To see this, let C be a subset of $[c - \delta, \infty)$ and use (4.9) to get

$$F[C^\delta] \geq \Phi_c[C] \geq \Phi_t[C], \qquad t \leq c.$$

Now let $C = (A - B) \cap \mathbb{R}^+$ and conclude

$$F\left[(A - B)^\delta\right] \geq \Phi_t[A - B], \qquad t \in [-c, c].$$

But

$$F[A^\delta - B^\delta] = F[A^\delta] \geq F\left[(A - B)^\delta\right],$$

where the equality follows because $F[B^\delta] = 0$ [i.e. $B^\delta \cap \text{supp}(f) = \varnothing$]. Combining the last two displays we have

$$F[A^\delta - B^\delta] \geq \Phi_t[A - B], \qquad t \in [-c, c],$$

which gives (4.10). Comparing (4.10) with (4.8), we see that in computing $\Delta(t, \delta)$ it suffices to consider the infimum over only sets B of the form $F[B^\delta] = 0$. The infimum will be attained by any B solving

$$\max_B \Phi_t[B] \quad \text{subject to} \quad F[B^\delta] = 0.$$

Clearly this maximum is attained when $B^\delta = \mathbb{R} - \text{supp}(f)$, i.e., by $B = [-c + \delta, c - \delta]$. Then

$$\Delta(t, \delta) = -\Phi_t[-c + \delta, c - \delta] \tag{4.11}$$

for $t \in [-c, c]$. For a fixed value of δ, $\Delta(t, \delta)$ is maximized at $t = 0$ and minimized at $t = \pm c$. It follows that δ^* defined by (4.6) is largest at $t = 0$ and smallest at $t = \pm c$. Consequently, Λ is minimized on $[-c, c]$ at $\pm c$ rather than at 0. The argument that $\pm c$ are global minimizers of Λ is routine and we omit it. □

LEMMA 4.8. *Put $\Lambda_n(t) = \mu(F_n, \Phi_t)$. Λ_n converges uniformly to Λ, almost surely and in probability. Any minimizer of Λ_n is, for all large enough sample sizes, close to either c or $-c$. As $n \to \infty$, the chance that all minimizers occur* *near c approaches* 0.5, *and the chance that they all occur near* $-c$ *approaches* 0.5 *as well.*

PROOF. With two applications of the triangle inequality we get

$$|\mu(F_n, \Phi_t) - \mu(F, \Phi_t)| \le \mu(F_n, F)$$

so that $|\Lambda_n(t) - \Lambda(t)| \le \mu(F_n, F)$. By Glivenko–Cantelli, $\mu(F_n, F) \to 0$ almost surely and in probability, so the first statement is proved. Inspection of Λ shows that there is a modulus ω_Λ with

$$\mathrm{dist}(x, \{c, -c\}) \le \omega_\Lambda(\Lambda(x) - \Lambda(c)).$$

Combining the two, we get that any minimizer $\hat{\theta}_n^*$ of Λ_n satisfies

$$\mathrm{dist}(\hat{\theta}_n, \{c, -c\}) \le \omega_\Lambda(2\mu(F_n, F)).$$

As $\mu(F_n, F) \to 0$ a.s. this completes the proof of the second statement.

For the third statement, we only indicate the idea of proof. Let $\delta > 0$; suppose we could show that

$$\text{(A)} \qquad \min_{|t-c|\le\delta} \Lambda_n(t) - \min_{|t+c|\le\delta} \Lambda_n(t) = \Lambda_n(c) - \Lambda_n(-c) + o_P(|\Lambda_n(c) - \Lambda_n(-c)|)$$

and

$$\text{(B)} \qquad P\{\Lambda_n(c) \ne \Lambda_n(-c)\} \to 1.$$

Then it would follow that for M_n^+, any minimizer of Λ_n on $|t - c| \le \delta$ and M_n^-, any minimizer of Λ_n on $|t + c| \le \delta$,

$$P\{\Lambda_n(M_n^+) \ne \Lambda_n(M_n^-)\} \to 1.$$

Thus only one of M_n^+ and M_n^- can be a global minimizer.

Now assertion (A) is a form of stochastic equicontinuity of Λ_n at $\pm c$; the second assertion is a weak form of "limit theorem" for $(\Lambda_n(c), \Lambda_n(-c))$. Unlike the case in Lemma 4.6, no central limit theorem will be available here. Rather, as we have approached it, one compares $\Lambda_n(c) - \Lambda_n(c + \varepsilon)$ $(\varepsilon < \delta)$ with $\Lambda_n(c) - \Lambda(-c)$ by explicit computations. These however are tedious and unenlightening. Compare also Kersting (1978). □

PROOF OF THEOREM 5. Our proof will actually show that the minimum CvM distance estimator has the indicated properties over Kolmogorov neighborhoods of the model. However, we have the inequalities

$$|F - G|^{3/2}/3^{1/2} \le \mu(F, G) \le |F - G|,$$

where $|F - G| = \sup_t|F(t) - G(t)|$ and μ denotes the Cramér–von Mises discrepancy

$$\mu(F, G)^2 = \int (F(t) - G(t))^2\, dG. \tag{4.12}$$

This inequality is valid whenever G is continuous; see Choi and Bulgren (1968).

Putting $G = \Phi$, we see that a Kolmogorov neighborhood of size ε about Φ contains a Cramér–von Mises neighborhood of size $\varepsilon^{3/2}/\sqrt{3}$.

Lemma 4.9 states that the minimum CvM functional is $|\cdot|$-continuous and differentiable, and the minimum CvM estimator is asymptotically normal, throughout a $|\cdot|$-open set about any F satisfying the "key condition" referred to in the text, with asymptotic variance given by

$$\operatorname{Var}(\hat{\theta}, F) = E_F IC^2_{\hat{\theta}, F}.$$

Lemma 4.10 shows that the mapping $F \to IC_{\hat{\theta}, F}$ is a continuous mapping from the set of distributions on $\mathbb{R}$ equipped with $|\cdot|$-norm, to $(L_\infty(\mathbb{R}), |\cdot|)$. But if two influence curves are closer than ε in the uniform metric, their corresponding $L_2(F)$ norms (i.e., the asymptotic standard deviations of their respective functionals) are closer than ε, no matter what F is. Consequently, the asymptotic variance of $\hat{\theta}$ is $|\cdot|$-continuous at the model.

The relation between CvM and Kolmogorov distance implies that $\operatorname{Var}(\hat{\theta}, F)$ is CvM-continuous at the model, and the theorem is established. □

LEMMA 4.9. *Suppose that F has a unique μ-closest point $\hat{\theta}(F)$ on $\{\Phi_\theta\}$ and that $\Lambda(\theta) = \mu(F, \Phi_\theta)$ has a nonsingular quadratic minimum at $\hat{\theta}$. Then*:

(i) *These conditions continue to hold in a $|\cdot|$-open set about F.*
(ii) *$\hat{\theta}$ is continuous and Fréchet differentiable at F.*
(iii) *$\hat{\theta}(F_n)$ is asymptotically normal at F with asymptotic variance $E_F IC^2_{\hat{\theta}, F}$.*

PROOF. Put $\Lambda(F, \theta) = \mu(F, \Phi_\theta)$, $\lambda(F, \theta) = (\partial/\partial\theta)\Lambda(F, \theta)$ and $l(F, \theta) = (\partial/\partial\theta)\lambda(F, \theta)$. Explicit computations [of the same kind as underlying the analysis of $L(f) = l(F, \hat{\theta}(F))$ in Lemma 4.10] imply that all these functions are jointly continuous in F and θ, where the $|\cdot|$ topology on F is used. Consequently for G in a small neighborhood N around F, on a small interval I containing $\hat{\theta}(F)$, we have that $l(G, t)$ is bounded away from zero and that $\lambda(G, \cdot)$ has one zero crossing on I. Thus $\Lambda(G, \cdot)$ has a single nonsingular quadratic minimum on I for all $G \in N$.

Actually, for all G near enough F, $\Lambda(G, t)$ has its global minimum on I. Define

$$\Delta_G(t) = \Lambda(G, t) - \Lambda(F, \hat{\theta}(F)).$$

Now for each $\eta > 0$ there is a $\mathbb{K} = [-k, k]$ with

$$\Delta_F > 1 - \Lambda(F, \hat{\theta}(F)) - \delta, \qquad t \in \mathbb{K}.$$

Note that

$$|\Lambda(G, t) - \Lambda(G, t + h)| \leq |\Phi_t - \Phi_{t+h}|$$

so that Δ_F is uniformly continuous. Now by assumption, $\Lambda(F, \hat{\theta}(F)) < \lambda(F, t)$, $t \notin I$; as Δ is uniformly continuous this implies that

$$\Delta_F(t) \geq \rho > 0, \qquad t \in \mathbb{K} - I,$$

so we have

$$\Delta_F(t) \geq \rho, \qquad t \notin I.$$

Now let $|G - F| < \rho/2$. Then from

$$|\Lambda(G,t) - \Lambda(F,t)| \le |F - G|$$

we have

$$\Delta_G(t) \ge \rho/2, \qquad t \notin I.$$

However

$$|\Lambda(G,\hat{\theta}(F)) - \Lambda(F,\hat{\theta}(F))| < \rho/2$$

so $\min_t \Lambda(G,t)$ is attained only in I. This, combined with remarks of the preceding paragraph show that $\Lambda(G,\cdot)$ has a unique global minimum which is a nonsingular quadratic, at least over

$$N_0 = \{G: G \in N \text{ and } |G - F| < \rho/2\}.$$

This establishes part (i) of the lemma.

Now we have that for $G \in N_0$, $\hat{\theta}(G)$ is the unique solution to

$$\lambda(G,t) = 0, \qquad t \in I.$$

Straightforward calculations as in Lemma 4.10 will give the Lipschitz bounds

$$A|\theta - \eta| \le \lambda(G,\theta) - \lambda(G,\eta) \le B|\theta - \eta|$$

for $G \in N_0$, $\theta > \eta$ and $\theta, \eta \in I$. [Here $A = \inf_{G \in N_0, t \in I} l(G,t)$ will do, for example.] And $\lambda(G,\theta) - \lambda(F,\theta) \le C|F - G|$.

Combining these bounds with Fréchet differentiability of $\lambda(G,t)$ at $G = F$, $t = \hat{\theta}(F)$, we can apply the implicit function theorem of Fernholz [(1983), Theorem 6.1.2] to conclude part (ii) of the lemma.

[The argument that $\lambda(G,t)$ is Fréchet differentiable goes as follows:

$$\lambda(G,t) = 2\int (G - \Phi_t)\phi_t^2 - 2\int (G - \Phi_t)^2 \phi_t'.$$

This expression is made up out of elementary parts which are easily seen to be Fréchet differentiable. In detail

$$\lambda(G,t) = \langle \Delta(G,t), \phi_t^2 \rangle - \langle \Delta(G,t)^2, \phi_t' \rangle,$$

where $\Delta(G,t) = (G - \Phi_t)$ and $\langle H, f \rangle$ denotes the linear pairing $\int Hf$. Now Δ and Δ^2 are obviously Fréchet differentiable mappings between $L_\infty(\mathbb{R})$ and itself. The functionals

$$(H,t) \to \langle H, \phi_t \rangle, \qquad (H,t) \to \langle H, \phi_t' \rangle$$

are Fréchet differentiable at (H,t) for all $H \in L_\infty$ and in particular for $H = \Delta(F, \hat{\theta}(F))$ (resp. $H = \Delta(F, \hat{\theta}(F))^2$).]

Part (iii) of the lemma is an automatic consequence of part (ii), as explained in Fernholz (1983). □

LEMMA 4.10. *Suppose the same conditions on F as in Lemma 4.9. For each $\varepsilon > 0$, there is a $\delta > 0$ so that if*

$$|F - G| \le \delta,$$

then

$$|IC_{\hat\theta, F} - IC_{\hat\theta, G}| < \varepsilon.$$

PROOF. By calculation, the Gâteaux derivative of $\hat\theta$ is

$$IC_{\hat\theta, F}(t) = \Psi(F, t)/L(F),$$

where

$$\Psi(F, t) = \int_{-\infty}^{t} \phi_{\hat\theta}^2 - 2(F - \Phi_{\hat\theta})\phi_{\hat\theta}',$$

$\hat\theta = \hat\theta(F)$ and

$$L(F) = l(F, \hat\theta)$$

$$= 2\int \phi_{\hat\theta}^3 - 6\int (F - \Phi_{\hat\theta})\phi_{\hat\theta}\phi_{\hat\theta}' + \int (F - \Phi_{\hat\theta})^2 \phi_{\hat\theta}''.$$

Now $F \to L(F)$ is $|\cdot|$-continuous. Write it as $L(F) = 2\int\phi^3 - 6L_A(F) + L_B(F)$. Then

$$L_A(F) - L_A(G) = \int (F_{\hat\theta(F)} - G_{\hat\theta(G)})\phi\phi'$$

[where F_t denotes the translate $F(\cdot - t)$]

$$= \int (F_{\hat\theta(F)} - G_{\hat\theta(F)})\phi\phi' + \int (G_{\hat\theta(F)} - G_{\hat\theta(G)})\phi\phi'$$

$$\le |F - G| \int |\phi\phi'| + |\hat\theta(F) - \hat\theta(G)| \int |(\phi')^2 + \phi\phi''|.$$

As $\hat\theta$ is continuous at F, this establishes the continuity of L_A. L_B is continuous by a similar argument.

As for Ψ, we have

$$|\Psi(F, t) - \Psi(G, t)| \le 4\Big(\sup_x |\phi'(x)|\Big)|\hat\theta(F) - \hat\theta(G)|$$

$$+ 4\Big(\int |\phi'|\Big)\Big(|G - F| + |\hat\theta(F) - \hat\theta(G)| \sup_x \phi(x)\Big)$$

so that $F \to \Psi(F, \cdot)$ is a continuous mapping from distribution functions equipped with $|\cdot|$ to $\mathbf{C}(\mathbb{R})$, also equipped with that norm.

By hypothesis, $L(F) \neq 0$ [i.e., $\Lambda(F, \cdot)$ has a nonsingular quadratic minimum at $\hat\theta$], so

$$G \to \Psi(G, \cdot)/L(G)$$

is likewise a continuous mapping at $G = F$. □

PROOF OF THEOREM 6. A proof can be given paralleling that of Theorem 5 practically line-by-line. Let μ = Hellinger distance

$$\mu(P, \Phi_t)^2 = 2 - 2\int p^{1/2}\phi_t^{1/2}, \tag{4.13}$$

where p is the density of the absolutely continuous part of P. Put $\Lambda(P, \theta) = \mu(P, P_\theta)$. The first part of the theorem follows by noting that $\Lambda(\Phi, \cdot)$ has a unique global minimum at 0 at which the behavior is nonsingular quadratic. Applying Lemma 4.11, $\hat{\theta}$ is continuous and differentiable in a neighborhood of Φ.

For the second part, note that if P has a density, $\mu(\hat{P}_n, P) \to 0$. Indeed, by Theorem 3.1 of Devroye and Gyorfi (1984), the empirical density estimate converges in variation norm to P. Together with the relation $\mu^2 \leq$ variation and the continuity of $\hat{\theta}$, this establishes the second part.

The formal expression for the asymptotic variance of $\hat{\theta}$ in the $N(\theta, 1)$ model is [see Beran (1977), Theorem 4]

$$\operatorname{Var}(\hat{\theta}, P) = \left(\int \left(\phi_{\hat{\theta}}^{1/2} \right)''' p^{1/2} \right)^{-1},$$

which is obviously continuous in the Hellinger topology at the model. This establishes the last part of the theorem.

Lemma 4.11. *Suppose that* $\Lambda(P, \theta)$ *has a unique global minimum at* $\hat{\theta}$ *at which* $\Lambda(P, t)$ *is locally quadratic in* t. *Then*:

(i) *This condition continues to hold in a Hellinger open set about* P.
(ii) $\hat{\theta}$ *is continuous and Fréchet differentiable at* P.

A proof of (i) can be given, paralleling that in Lemma 4.9 line-by-line. As for (ii), we note that the result is essentially contained already in Theorems 1 and 2 of Beran (1977), so we omit the argument.

Acknowledgment. The authors would like to thank A. Kozek and H. Rieder for interesting discussions at the Oberwolfach meeting on robustness, September 1984.

REFERENCES

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

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

Choi, K. and Bulgren, W. (1968). An estimation procedure for mixtures of distributions. *J. Roy. Statist. Soc. Ser. B* **30** 444–460.

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

Donoho, D. L. and Liu R. C. (1988). The "automatic" robustness of minimum distance functionals. *Ann. Statist.* **16** 552–586.

Fernholz, L. T. (1983). *von Mises Calculus for Statistical Functionals. Lecture Notes in Statist.* **19.** Springer, Berlin.

Freedman, D. A. and Diaconis, P. (1982). On inconsistent M estimators. *Ann. Statist.* **10** 454–461.

Holm, S. (1976). In 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.

KERSTING, G. D. (1978). Die Geschwindigkeit der Glivenko–Cantelli-Konvergenz gemessen in der Prohorov-Metrik. *Math. Z.* **163** 65–102.

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

LIU, R. C. (1987). Geometry and topology in robustness and nonparametrics. Thesis, Univ. California, Berkeley.

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–Smirov statistics. *Ann. Statist.* **3** 862–873.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720