3.4 M-estimates and robust location estimates. M-estimators, as defined in Sec. 3.3, are sometimes called M-estimators *of ρ type* where the function $h(\theta, x)$ may also be called $\rho(\theta, x)$. Such estimators include maximum likelihood estimators as noted there. Another class of M-estimators is a class of estimators of location in $\mathbb{R}$ which are robust, meaning that they are not sensitive to contamination of the data by a few erroneous values, as will be seen.

Let $\psi(\theta, x)$ be a jointly measurable function, of θ in a parameter space Θ and x in a sample space X, where ψ has values in a Euclidean space $\mathbb{R}^k$. A statistic $T_n := T_n(X_1, \cdots, X_n)$ will be called an *M-estimator of ψ type* if $\sum_{i=1}^n \psi(T_n, X_i) \equiv 0$. Such an estimator is not necessarily an M-estimator of ρ type, but it is related in that if $\rho(\theta, x)$ has continuous first partial derivatives with respect to θ, then a necessary condition for T_n to be an M-estimator (of ρ type) is that it be one of ψ type with ψ = gradient of ρ. Under some rather special conditions, as for exponential families in Theorem 3.1.2, an M-estimator of ψ type where ψ is the gradient of ρ must also be one of ρ type. M-estimators of ψ type, where ψ is not necessarily a gradient, and the definition need only hold approximately as $n \to \infty$, will be further treated in the next two sections.

A class of examples of M-estimators other than maximum likelihood estimators is provided by some location estimators in $\mathbb{R}$ as follows.

For any probability distribution P on $\mathbb{R}$ with $\int x^2 dP < \infty$, $\mu := \int x dP$ is the unique m for which $\int (x-m)^2 dP(x)$ is minimized. To see this, note that the latter integral is a quadratic function of m which is minimized where its derivative is 0. So, for observations in $\mathbb{R}$, the sample mean $\overline{X}$ is an M-estimator for $h(\theta, x) := (x-\theta)^2$. It is an estimator of the true mean μ, which gives a consistent sequence of estimators (by the strong law of large numbers) if and only if μ is defined and finite. In this case, if we set $a(x) := x^2$, then h is adjustable and $a(\cdot)$ is an adjustment function for any P with $\int_{-\infty}^{\infty} |x| dP(x) < \infty$, in other words whenever μ is defined and finite.

Suppose we have a location family in $\mathbb{R}$, in other words a family of laws P_θ, $\theta \in \mathbb{R}$, where $dP_\theta(x) \equiv dP_0(x-\theta)$. If $\int |x| dP_0(x) = +\infty$, then the sample means a.s. do not converge as $n \to \infty$ to a finite limit (RAP, Theorem 8.3.5). For example if P_0 is the Cauchy distribution, $dP_0(x) := dx/[\pi(1+x^2)]$, then $\overline{X}$ has the same distribution as X_1. (This follows from the fact that the characteristic function $\int_{-\infty}^{\infty} e^{itx} dP_0(x) = e^{-|t|}$ for all real t; it is easy to verify that the inverse Fourier transform of $e^{-|t|}$ is the Cauchy density.)

A measure of location which is finite for any law P on $\mathbb{R}$ is the median, where m is a *median* of a random variable X (with law P) iff both $P(X \le m) \ge 1/2$ and $P(X \ge m) \ge 1/2$. If there is only one median it is called *the* median. If X has no atoms in its distribution, in other words $P(X = x) = 0$ for all x, then m is a median of P if and only if $P(X \le m) = 1/2$. A median always exists but may not be unique: for example if $P(X = 0) = P(X = 1) = 1/2$, then all x in $[0, 1]$ are medians. Let F be the distribution function of X or P, $F(t) = P(X \le t)$ for all real t. Then F is always right-continuous and has left limits

$$F(t-) := \lim_{s \uparrow t} F(s) = P(X < t)$$

for all t. We can equivalently define a median of X or P as a t such that $F(t-) \le 1/2 \le$

$F(t)$. Let $m_\ell := \inf\{x : F(x) \geq 1/2\}$ and $m_u := \sup\{y : F(y-) \leq 1/2\}$. For any $y < m_\ell$, we have $F(y-) \leq F(y) < 1/2$, thus $m_u \geq m_\ell$. It is then clear that the set of all medians of X or P is exactly the non-empty closed interval $[m_\ell, m_u]$. If $m_\ell < m_u$ then we have $F(t) = 1/2$ for $m_\ell < t < m_u$, and, by right-continuity, also for $t = m_\ell$.

Let P_n be the empirical measure $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_x is a point mass at x. Then a *sample median* is any median of P_n. If $n = 2r$ is even, then a sample median is any point in the interval between order statistics $[X_{(r)}, X_{(r+1)}]$, which is the interval $[m_\ell, m_u]$ for P_n in that case. If $n = 2r - 1$ is odd, then the sample median is unique and equals $X_{(r)}$.

If $\int |x|dP < \infty$, then m is a median for P if and only if it minimizes $\int |x - m|dP$, so a sample median is an M-estimator for $h(\theta, x) := |x - \theta|$, and the condition $\int |x|dP(x) < \infty$ can be removed by an "adjustment," replacing $h(\theta, x)$ by $h(\theta, x) - a(x)$ for $a(x) := |x|$ in this case, as was mentioned in Sec. 3.3 and will now be shown.

3.4.1 Proposition. (a) For any law P on $\mathbb{R}$, m is a median of P if and only if $\int |x - \theta| - |x|dP(x)$ is minimized for $\theta = m$. (b) If $\int |x|dP(x) < \infty$, then m is a median of P if and only if $\int |x - \theta|dP(x)$ is minimized for $\theta = m$.

Proof. Clearly (b) follows from (a) by adding the finite constant $\int |x|dP(x)$ to the quantity being minimized, so it will suffice to prove (a).

Let $h(\theta, x) := |x - \theta| - |x|$. For any fixed value of x, clearly $h(\cdot, x)$ is a convex function of θ. We have $|h(\theta, x)| \leq |\theta|$ for all x and θ. For fixed θ, $h(\theta, \cdot)$ is a bounded, continuous function of x. Let $\gamma(\theta) := E(|X - \theta| - |X|) = Eh(\theta, X)$. The expectation is finite for each θ and preserves the inequalities in the definition of convex function, so $\gamma(\cdot)$ is a convex function. Any convex real function of a real variable on an open interval (in this case, the whole line) has left and right derivatives at every point (RAP, 6.3.3).

Whenever $\phi < \theta$,

$$h(\theta, x) - h(\phi, x) = |x - \theta| - |x - \phi| = \begin{cases} \theta - \phi, & \text{for } x \leq \phi, \\ \theta + \phi - 2x, & \text{for } \phi \leq x \leq \theta, \\ \phi - \theta, & \text{for } x \geq \theta. \end{cases}$$

We have $|h(\theta, x) - h(\phi, x)| \leq |\theta - \phi|$ for all x, θ and ϕ.

Let F be the distribution function of P and $\phi < \theta$. Let

$$g(\phi, \theta, x) := \frac{|x - \theta| - |x - \phi|}{\theta - \phi}.$$

Then $g(\phi, \theta, x) = 1$ for $x \leq \phi$ and -1 for $x \geq \theta$. For $\phi < x < \theta$ we have $|g(\phi, \theta, x)| \leq 1$, so the same holds for all x. If we fix ϕ and let $\theta \downarrow \phi$ then $g(\phi, \theta, x)$ converges boundedly to $1_{x \leq \phi} - 1_{x > \phi}$. Thus by dominated convergence, $\gamma'(\phi+) = F(\phi) - (1 - F(\phi)) = 2F(\phi) - 1$. Or, if we fix θ and let $\phi \uparrow \theta$ then $g(\phi, \theta, x)$ converges boundedly to $1_{x < \theta} - 1_{x \geq \theta}$. Thus $\gamma'(\theta-) = F(\theta-) - (1 - F(\theta-)) = 2F(\theta-) - 1$.

If $F(\theta) < 1/2$, then $\gamma'(\theta+) < 0$, so γ is not minimized at θ. If $F(\theta-) > 1/2$, then $\gamma'(\theta-) > 0$, so γ is not minimized at θ. In either case, θ is not a median of X. The remaining case is that $F(\theta-) \leq 1/2 \leq F(\theta)$, in other words, θ is a median of X. In that case $\gamma'(\theta-) \le 0 \le \gamma'(\theta+)$. If the median is unique it is also the unique minimum of γ. Or if $m_\ell < m_u$, then by the non-decreasing property of the one-sided derivatives of γ (RAP, 6.3.3) it follows that $\gamma'(t) = 0$ for $m_\ell < t < m_u$ and γ is constant on $[m_\ell, m_u]$, so it attains its minimum exactly on this interval. $\square$

For a set of data, arranged as order statistics $X_{(1)} \le \cdots \le X_{(n)}$, if $n \ge 3$ the sample median is unchanged if $X_{(1)}$ is replaced by any value $s < X_{(1)}$, or if $X_{(n)}$ is replaced by any $t > X_{(n)}$. On the other hand the sample mean $\overline{X}$ can be affected by a large amount for extreme values of s or t. A very large value of $X_{(n)}$, or a very negative value of $X_{(1)}$, may be considered an "outlier" and one may prefer estimates that are relatively insensitive to outliers, in other words robust estimates. A more precise definition of robustness can be based on the notion of breakdown point, which in turn will require a definition of distance between probability measures.

Definitions. Let (S, d) be a separable metric space and P, Q any two laws (probability measures on the Borel σ-algebra) on S. Then the *Prokhorov metric* $d_0(P, Q)$ is defined by

$$d_0(P,Q) \ := \ \inf\{\varepsilon > 0 : \ P(A) \le Q(A^\varepsilon) + \varepsilon \ \text{ for all Borel sets } \ A\}$$

where $A^\varepsilon \ := \ \{y : \ d(x,y) < \varepsilon \ \text{ for some } \ x \in A\}$.

A real-valued function f on S is *Lipschitz* if

$$\|f\|_L \ := \ \sup\{|f(x) - f(y)|/d(x,y) : \ x \ne y\} \ < \ \infty.$$

Let $\|f\|_\infty \ := \ \sup\{|f(x)| : \ x \in S\}$. Let $\|f\|_{BL} \ := \ \|f\|_L + \|f\|_\infty$ for any bounded, Lipschitz function f. Let

$$d_{BL}(P,Q) \ := \ \sup\{|\int f d(P-Q)| : \ \|f\|_{BL} \le 1\}.$$

It is known that both the Prokhorov metric d_0 and the "dual-bounded-Lipschitz" metric d_{BL} or $d_1 \ := \ d_{BL}/2$ define metrics, for the same topology, on the set of all laws on S, and metrize weak convergence: for any sequence Q_n of laws on S, as $n \to \infty$, $d_0(Q_n, Q) \to 0$ if and only if $d_1(Q_n, Q) \to 0$ if and only if $\int f dQ_n \to \int f dQ$ for every bounded continuous real-valued function f on S (RAP, Sec. 11.3). We have $d_1 \le d_0$ (RAP, Corollary 11.6.5). If Q is a law on S and $X_1, X_2, \cdots$ are i.i.d. variables with values in S and distribution Q (such X_i exist, for example as coordinates on a product S^∞: RAP, Sec. 8.2), let Q_n be the empirical measure which is the sum of measures $1/n$ at X_i for $i = 1, \cdots, n$. Then almost surely $d_0(Q_n, Q) \to 0$ as $n \to \infty$ (a theorem of V. S. Varadarajan, given in RAP, Sec. 11.4).

Now estimators T_n will be considered as functions $T_n(P_n)$ taking values in a locally compact non-compact separable metric space U such as $\mathbb{R}^k$.

Definition. Given a separable metric space S and a metric e for laws on S, a sequence $\{T_n\}$ of estimators has *breakdown point* at P for e given by $b(\{T_n\}, P, e) \ :=$
$\sup\{\varepsilon > 0 :$ for some compact set $K \subset U$, $e(P,Q) < \varepsilon$ implies $\Pr^*(T_n(Q_n)) \notin K) \to 0$ as $n \to \infty\}$, where the probability $\Pr$ is for $X_1, X_2, \cdots$ i.i.d. with distribution Q, and $\Pr^*(C) \ := \ \inf\{\Pr(D) : \ D \ \text{ measurable}, D \supset C\}$ for any set C.

An estimator with low breakdown point, specifically, one with breakdown point 0, is not robust. An estimator as robust as possible should have a breakdown point as high as possible. In estimating location in $\mathbb{R}$ we will have $U = \mathbb{R}$. The sample mean functional is $\overline{X} := T_n(Q_n) := \int x dQ_n$. It is easily seen that it has breakdown point 0 for e either of the given metrics d_0 or d_1.

In general, breakdown points are not expected to be larger than 1/2. To see why, suppose that $Q = (P + \mu)/2$ where P and μ are any two laws on S. Then $d_0(P, Q) \leq 1/2$, and also $d_1(P, Q) \leq 1/2$. Suppose $\{T_n\}$ had a breakdown point $> 1/2$ at P and at μ. Then $T_n(Q_n)$ should with high probability take values in one compact set for the definition of breakdown point at P, and another compact set for μ. If T_n are to be consistent for some parameter, which is very different for P and μ, then the compact set must be large to contain both parameters. This is possible in general, but can be ruled out in certain cases, for example as follows.

A real-valued statistic T_n based on real observations $X_1, \cdots, X_n$ is called *equivariant* (for translation) iff

$$T_n(X_1 + c, \cdots, X_n + c) \equiv T_n(X_1, \cdots, X_n) + c$$

for any real constant c. For example, $\overline{X}$ is equivariant for translation.

3.4.2 Proposition. Let $\{T_n\}$ be a sequence of equivariant real-valued statistics. Let $dP_\theta(x) := dP_0(x - \theta)$, $\theta \in \mathbb{R}$, be a family of laws on $\mathbb{R}$. Let e be d_0 or d_1. Then if $\{T_n\}$ has breakdown point $> 1/2$ at P_θ for all θ, it cannot be a consistent sequence of estimators of θ for any θ.

Proof. If $\{T_n\}$ is consistent for some θ, i.e. $T_n \to \theta$ as $n \to \infty$ for $T_n = T_n(X_1, \cdots, X_n)$ and X_i i.i.d. P_θ, then $\{T_n\}$ is consistent for all θ by equivariance. If $\{T_n\}$ has breakdown point $r > 1/2$ for $\theta = 0$, take the compact set K from the definition of breakdown point for some ε with $1/2 < \varepsilon < r$. Then for some t large enough, $K + t := \{x + t : x \in K\}$ is disjoint from K. By equivariance, $\{T_n\}$ has breakdown point r also for $\theta = t$ with the compact set $K + t$ working for ε. Let $Q := (P_0 + P_t)/2$. Then by definition of breakdown point, for n large enough, $\Pr^*\{T_n(Q_n) \notin L\} < 1/2$ for $L = K$ and for $L = K + t$, but this implies that $\Pr^*$ of the whole probability space is less than 1, a contradiction. $\square$

While the d_0 and d_1 metrics are defined for laws on any metric space, there is another metric for the special case of the line $\mathbb{R}$, called the *Lévy metric* L, defined as follows. If laws P and Q on $\mathbb{R}$ have respective distribution functions F and G, let

$$L(P, Q) := \inf\{\varepsilon > 0 : G(x - \varepsilon) - \varepsilon \leq F(x) \leq G(x + \varepsilon) + \varepsilon \text{ for all } x\}.$$

It is easy to check that $L(\cdot, \cdot)$ is indeed a metric for laws on $\mathbb{R}$ and that $L(P, Q) \leq d_0(P, Q)$, considering half-lines as sets A in the definition of the Prokhorov metric d_0. Here d_0 may be substantially larger than L: if P has mass $1/n$ at each of $2, 4, \cdots, 2n$, and Q has mass $1/n$ at each of $1, 3, \cdots, 2n - 1$, then $L(P, Q) = 1/n$ but $d_0(P, Q) \equiv 1$. Such P and Q have mass "escaping to ∞" and don't converge to any law. The Lévy metric does metrize the same topology as that of d_0 and d_1: for any laws P_n and P on $\mathbb{R}$, $d_0(P_n, P) \to 0$

is equivalent to $L(P_n, P) \to 0$, which is equivalent to the distribution functions of P_n converging to that of P at all points where the latter is continuous (Helly-Bray property, e.g. RAP, 11.1.2, in light of 11.1.1 and 11.3.3).

Before treating specific robust estimators further, a little general theory of robustness will be developed. Let e be a metric for laws and T a real-valued functional on the class of all laws which is continuous for e at a law P_0. Thus for $\varepsilon > 0$ small enough,

$$b_1(\varepsilon) \;:=\; \sup\{|T(P) - T(P_0)| : \; e(P, P_0) < \varepsilon\} \;<\; \infty.$$

Let as before $T_n = T_n(P_n)$ where P_n is an nth empirical measure for P. Let $M(P, T_n)$ be the median, specifically, the smallest median of the distribution of $T_n - T(P)$. Let

$$\underline{b}(\varepsilon) \;:=\; \liminf_{n\to\infty} \sup\{M(P, T_n) : \; e(P, P_0) < \varepsilon\},$$

$$\overline{b}(\varepsilon) \;:=\; \limsup_{n\to\infty} \sup\{M(P, T_n) : \; e(P, P_0) < \varepsilon\}.$$

If $\underline{b}(\varepsilon) = \overline{b}(\varepsilon)$, call their common value $b(\varepsilon)$.

3.4.3 Proposition. For any $\varepsilon > 0$, if $T(P_n) \to T(P)$ in probability as $n \to \infty$ for every law P with $e(P, P_0) < \varepsilon$, then $\overline{b}(\varepsilon) \geq b_1(\varepsilon)$.

Proof. We can assume $T(P_0) = 0$. For each P in the given neighborhood, since $T(P_n) \to T(P)$ in probability, the median $M(P, T_n) \to T(P)$ also. Thus $\overline{b}(\varepsilon) \geq T(P)$ for each such P, so $\overline{b}(\varepsilon) \geq b_1(\varepsilon)$ by definition of b_1. □

Under a stronger, uniformity assumption there is a converse:

3.4.4 Proposition. Assume that for a given $\varepsilon > 0$, as $n \to \infty$, $e(P_n, P) \to 0$ in probability as $n \to \infty$ uniformly for P such that $e(P_0, P) < \varepsilon$, in other words for any $\delta > 0$,

$$\sup\{\Pr\{e(P_n, P) > \delta\} : \; e(P, P_0) < \varepsilon\} \;\to\; 0$$

as $n \to \infty$. Then $\overline{b}(\varepsilon) \leq b_1(\varepsilon+) \;:=\; \lim_{\gamma\downarrow 0} b_1(\varepsilon + \gamma)$.

Proof. It will be enough to show that for each fixed $\gamma > 0$, $\overline{b}(\varepsilon) \leq b_1(\varepsilon + \gamma)$. By the triangle inequality, as $n \to \infty$

$$\sup\{\Pr\{e(P_n, P_0) \geq \varepsilon + \gamma\} : \; e(P, P_0) < \varepsilon\} \;\to\; 0.$$

As soon as this supremum of probabilities is less than $1/2$, the median $M(P, T_n)$ satisfies

$$M(P, T_n) \;\leq\; \sup\{|T(P) - T(Q)| : \; e(Q, P_0) \leq \varepsilon + \gamma\} \;=\; b_1(\varepsilon + \gamma). \qquad \square$$

In the case of the line and Lévy metric, we have:

3.4.5 Proposition. If the sample space is $\mathbb{R}$ and e is the Lévy metric, then the hypothesis of Prop. 3.4.4 holds.

Proof. The Glivenko-Cantelli theorem (RAP, Theorem 11.4.2) implies that empirical distribution functions F_n almost surely converge uniformly on $\mathbb{R}$ to the true distribution function. Moreover, the proof of the Glivenko-Cantelli theorem shows that the convergence for all distributions on $\mathbb{R}$ reduces to the case of $U[0,1]$ and so is uniform over all laws on $\mathbb{R}$. Also, the Lévy distance between two laws on $\mathbb{R}$ is less than or equal to the supremum of the absolute difference of their distribution functions. This implies that on $\mathbb{R}$, $L(P_n, P) \to 0$ as $n \to \infty$ (almost surely) at a rate which is uniform over all laws P on $\mathbb{R}$. □

One can say that "the median is equivariant" in the sense that it is equivariant (for translation) whenever it is unique, and when there is an interval of more than one median, a translation of the data gives a translation of the interval of medians. The median is sometimes defined as the midpoint $(m_\ell + m_u)/2$ of the interval of medians or as the smallest median m_ℓ.

There is a class of M-estimators of location having some properties like those of the median, including robustness, but which are more often, or always, unique. A non-constant function $\rho(\theta, x)$ for x and θ real will be called a *wide-sense Huber function* if $\rho(\theta, x) \equiv \rho(|x-\theta|)$ where $\rho(x) \equiv \rho(-x)$, ρ is convex, and $\rho(x)/|x|$ is bounded as $|x| \to \infty$. The convexity and symmetry properties imply that ρ attains its absolute minimum at 0 (and perhaps elsewhere). Examples of wide-sense Huber functions include

(a) $\rho(x) := |x|$,

(b) $\rho(x) := (c^2 + x^2)^{1/2}$ for any real c, and

(c) $\rho(x) := x^2$ for $|x| \le b$ and $\rho(x) := c|x| - d$ for $|x| > b$ where $b > 0$ and the other constants are chosen to make ρ continuously differentiable, so that $cb - d = b^2$ and $2b = c$, so $d = b^2$ and for $|x| > b$, $\rho(x) = b(2|x| - b)$.

Since Huber especially studied functions defined by (c), they might be called "narrow-sense Huber functions."

The functions in (b) and (c) are strictly convex in neighborhoods of 0, and the ones in (b) are strictly convex everywhere. Note that if ρ is convex, then the sum $\sum_{i=1}^n \rho(X_i - \theta)$ is convex in θ for any X_i. Also, if ρ is strictly convex for $|x| \le b$, then the sum is strictly convex in a neighborhood of θ for any θ such that $|X_i - \theta| < b$ for some i. For n large and b not too small, the set of such θ will often include the set on which the sum takes its smallest values, so that the minimum (M-estimator) will be unique. We will always have uniqueness if ρ is strictly convex everywhere, as $(c^2 + x^2)^{1/2}$ is. Clearly, whenever an M-estimator (of ρ type) is unique for a wide-sense Huber function ρ, it will be equivariant, so the following shows that such estimators have the highest possible breakdown point:

3.4.6 Proposition. Any M-estimator sequence in $\mathbb{R}$ where ρ is a wide-sense Huber function has breakdown point 1/2 for the Lévy metric or the Prokhorov metric at any law P_0.

Proof. Difference-quotients $(\rho(b) - \rho(a))/(b-a)$ for $0 < a < b$ are positive and nondecreasing as $a \to +\infty$ and/or $b \to +\infty$, and they are bounded above by the assumption that $\rho(x)/|x|$ is bounded. Thus ρ is a Lipschitz function with $\gamma := \|\rho\|_L < \infty$ and as $|x| \to \infty$, $\rho(x)/(\gamma|x|) \to 1$ where $\gamma > 0$. Letting $a(x) := \rho(x)$, and recalling $\rho(\theta, x) \equiv \rho(|x-\theta|)$, it follows that $|\rho(\theta, x) - a(x)| \le \gamma|\theta|$ for all x and θ.

Let $\psi(x) := \rho'(x)$. This is defined except for at most countably many values of x: ρ, being convex, has left and right derivatives everywhere (RAP, Corollary 6.3.3), with possibly $\rho'(x-) < \rho'(x+)$ for x in a countable set. To be specific, let $\psi(x) := \rho'(x-)$

for all x. Then ψ is non-decreasing. Now, our estimators of ρ type will also be of ψ type except that the value 0 may not be attained. Letting $\psi(t,x) := \psi(x-t)$, an estimator T_n will be said to be of *extended* ψ *type* if $\sum_{i=1}^n \psi(\theta, X_i)$ is ≥ 0 for all $\theta < T_n$ and ≤ 0 for all $\theta > T_n$.

Let $\lambda(t,P) := \int \psi(x-t)dP(x) = E\psi(X-t)$ where X has law P. Then $\lambda(\cdot, P)$ is nonincreasing in t. Also, $\lambda(t,P)$ increases if X becomes stochastically larger, in other words if X is replaced by $X+Y$ with $Y \geq 0$. We have $\psi(u) < 0$ as $u \to -\infty$ and $\psi(u) > 0$ as $u \to +\infty$. Let $T^*(P) := \sup\{t : \lambda(t,P) > 0\}$ and $T^{**}(P) := \inf\{t : \lambda(t,P) < 0\}$. Then $T^*(P) \leq T^{**}(P)$. If $T^*(P) < T^{**}(P)$ then $\lambda(t,P) = 0$ for $T^*(P) < t < T^{**}(P)$.

If F_0 is the distribution function of P_0 and F is that of P, then to make the distribution P stochastically as large as possible in the neighborhood $L(P, P_0) \leq \varepsilon$, take $F = G$ where $G(x) := \max(0, F_0(x-\varepsilon) - \varepsilon)$ for all x. The maximum value of G on $\mathbb{R}$ is $1-\varepsilon$. G can be viewed as a probability distribution function on $(-\infty, \infty]$ placing mass ε at $+\infty$; or, if desired, one can consider a law putting mass ε at some large M and let $M \to +\infty$. The result is the same for integrals of ψ, which approaches limits as $x \to \infty$ or $x \to -\infty$.

Let $x_0 := \inf\{x : F_0(x) > \varepsilon\}$. Then $F_0(x_0) = \varepsilon$, unless P_0 has an atom at x_0. We have $G(x) = 0$ for $x < x_0 + \varepsilon$ and $G(x) = F_0(x-\varepsilon) - \varepsilon$ for $x \geq x_0 + \varepsilon$. Then for any F with $L(P_0, P) < \varepsilon$,

$$\lambda(t,F) \leq \lambda(t,G) = \int_{x_0+\varepsilon}^{\infty} \psi(x-t)dF_0(x-\varepsilon) + \varepsilon\psi(+\infty)$$

$$= \int_{x_0}^{\infty} \psi(x+\varepsilon-t)dF_0(x) + \varepsilon\psi(+\infty),$$

where $\psi(+\infty) := \lim_{u\to\infty} \psi(u)$, and functionals of laws are written as functionals of their distribution functions. Let $b_+(\varepsilon) := \sup\{T^{**}(F) : L(F, F_0) \leq \varepsilon\} = T^{**}(G)$. Symmetrically to G, let H be the distribution function for the stochastically smallest random variable with $L(F_0, H) \leq \varepsilon$, so that $H(x) := \min(1, F_0(x+\varepsilon)+\varepsilon)$. Here H yields a distribution on $[-\infty, \infty)$ with mass ε at $-\infty$. Then $b_-(\varepsilon) := \inf\{T^*(F) : L(F, F_0) \leq \varepsilon\} = T^{**}(G)$.

As the set K in the definition of breakdown point consider the interval $[b_-(\varepsilon), b^+(\varepsilon)]$. Let's see for which ε this is compact. Now $b_+(\varepsilon) < \infty$ if for some t, $\lambda(t,G) < 0$, which is true if and only if $(1-\varepsilon)\psi(-\infty) + \varepsilon\psi(+\infty) < 0$. For our case, $\psi(-\infty) = -\psi(+\infty)$ so the condition on ε is $\varepsilon/(1-\varepsilon) < 1$ which is equivalent to $\varepsilon < 1/2$. Consideration of $b_-(\varepsilon)$ leads to the same condition, $\varepsilon < 1/2$.

If $L(F_0, F) < \varepsilon$, then $L(F_0, F_n) < \varepsilon$ with probability $\to 1$ as $n \to \infty$. Then for any M-estimators T_n of ρ type, $T^*(H) = b_-(\varepsilon) \leq T_n \leq b_+(\varepsilon) = T^{**}(G)$. These facts imply that the breakdown point of the sequence $\{T_n\}$ is at least 1/2 for the Lévy metric.

Now since the Prokhorov metric is at least as large as the Lévy metric, $d_0 \geq L$, $d_0(P_0, P) \leq \varepsilon$ implies $L(P_0, P) \leq \varepsilon$ which implies $G(x) \leq F(x) \leq H(x)$ for all x. On the other hand, let Y be a random variable with distribution function G. Such a Y exists having the following joint distribution with a random variable X_0 having distribution P_0: $Y = +\infty$ with probability ε, specifically if $X_0 < x_0$ and if $P_0(x_0) > 0$, then $\Pr(Y = +\infty | X_0 = x_0)$ has whatever value is needed to make $\Pr(Y = +\infty) = \varepsilon$, while $Y < \infty$ for $X_0 > x_0$. In fact when Y is not $+\infty$ it is $X_0+\varepsilon$. So Y is always within ε of X_0, except with probability ε. So the Ky Fan distance between the random variables X_0 and Y is ε (RAP, 9.2.2 and 11.6.4), and it follows that the Prokhorov distance between F_0 and G is $\leq \varepsilon$ and so $= \varepsilon$. So the breakdown point is also at least $1/2$ for the Prokhorov metric.

On the other hand, to show that the breakdown point is at most $1/2$, so that it is exactly $1/2$ for both metrics, suppose $r > 1/2$. For any $M < \infty$, taking a distribution F close to F_0, we will have for some $\delta > 0$, $\lambda(t, F) > \delta$ for all $t \leq M$, and likewise for F_n with high probability for n large. Then $T_n \geq M$. Letting $M \to \infty$, it follows that there is no compact set K as required in the definition of breakdown point, so the breakdown point is $1/2$. □

In some cases one can prove the breakdown point is at least $1/2$ by applying 3.4.2, but not in the generality of Prop. 3.4.6: consider the distribution with density $1/2$ each on $[0, 1]$ and $[2, 3]$, with $\rho(x) := |x|$, so that the estimators are medians. Since sample medians don't converge in this case, consistency and Proposition 3.4.2 don't apply, even if we choose medians to be equivariant estimators.

NOTES

The facts in this section are as in Huber (1981).

REFERENCES

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