The Annals of Probability
1979, Vol. 7, No. 6, 989–1002

ROBUSTNESS OF ESTIMATORS ON STATIONARY OBSERVATIONS

By P. Papantoni-Kazakos and Robert M. Gray[1]

University of Connecticut and Stanford University

Hampel's general qualitative definition of robustness of sequences of estimators on memoryless observation processes is generalized to stationary ergodic processes by substituting the generalized Ornstein (or $\bar{\rho}$) distance for the marginal Prohorov distance as the measure of "closeness" of observations. More general sequences of estimators are also allowed. The approach yields results analogous to those of Hampel for the more general case considered, often provides strict generalizations of Hampel's results, and in some cases yields simpler proofs.

1. Introduction. In his classic paper, Hampel (1971) introduced a definition of robustness in parameter estimation that accurately reflected the intuitive notion that a sequence of estimates of a parameter was robust for an observation process μ if another process ν that was "close" to μ yielded a "close" distribution on the parameter estimates. Hampel considered memoryless or independent, identically distributed (i.i.d.) observation processes and measured their "closeness" by the Prohorov distance on the marginal probability measures. As he considered i.i.d. processes, his underlying parameter depended implicitly only on these unknown marginals. Hampel then proved that weak* continuous functionals on the space of probability distributions defined robust sequences of estimators under his assumptions. He also showed his results could be adapted via an alternative notion of robustness to weakly dependent observations, in particular observations that were close to memoryless in a Prohorov sense.

A critical part of his derivation was the fact that if two i.i.d. processes μ and ν are close in a marginal Prohorov sense, then one could construct a pair process p having μ and ν as coordinate processes and such that under p the sample distributions of two coordinate n-tuples x^n produced by μ and y^n produced by ν were close in a Prohorov sense with high probability. During the past few years, a generalization of Ornstein's $\bar{d}$ distance of ergodic theory (called the $\bar{\rho}$, "rho-bar," or generalized Ornstein distance) has been shown to provide a similar control for sample distributions for general stationary and ergodic processes and, largely as a result, has found several applications in information theory (see, e.g., Gray, Neuhoff and Shields (1975), Gray, Neuhoff, and Omura (1975)). In this paper we

Received November 1, 1977; revised July 28, 1978.

[1] This research was supported by the Air Force Office of Scientific Research under AFOSR Grant 77-3156 and by the National Science Foundation under NSF Grant ENG 76-02276.

AMS 1970 *subject classifications.* Primary 62G35; secondary 60B05, 60H99.

Key words and phrases. Robust estimation, stationary data, ergodicity, stochastic distance measures.

show that using the $\bar{\rho}$ distance as a measure of closeness of the observation processes, there is a natural qualitative definition of robustness for all stationary ergodic processes, that a weakened version of Hampel's weak*-continuous estimator sequence implies robustness, and that all of Hampel's results have analogs in this more general case. Our formalism does not quite contain Hampel's in the case of i.i.d. processes and parameters depending only on the marginal probabilities, but is a strict generalization in some cases such as when the metric on the observation alphabet is bounded or when the class of probability measures considered is constrained to have a finite second moment (see Lemma 2.1).

We also note that we need not confine estimates to take values in R^k as Hampel does, but instead we only require that the parameter alphabet be a complete, separable metric (Polish) space. Hence function valued parameter spaces are allowed.

As a side result, some easy generalizations of the convergence of sample distributions (Parthasarathy (1967)) for stationary and ergodic processes are developed.

2. Preliminaries.

Let $(\Omega, \mathcal{B}_\Omega)$ be a measurable space such that Ω is a complete, separable metric space (or Polish space) with metric ρ and $\mathcal{B}_\Omega$ is the Borel σ-field generated by the open sets under ρ. Since Ω is separable, there is a countable collection of sets $\mathcal{G}_\Omega = \{G_i; i = 1, 2, \cdots\}$ such that $\mathcal{B}_\Omega = \sigma(\mathcal{G}_\Omega)$, that is, $\mathcal{B}_\Omega$ is the σ-field generated by $\mathcal{G}_\Omega$. Let Ω^n be the space of n-tuples with coordinates in Ω and Ω^∞ the space of sequences $\omega = (\cdots, \omega_{-1}, \omega_0, \omega_1, \cdots)$, $\omega_i \in \Omega$ all i. Let $\mathcal{B}_\Omega^n$ be the σ-field of subsets generated by all rectangles of the form $\times_{i=0}^{n-1} B_i$, $B_i \in \mathcal{B}_\Omega$ (since Ω is Polish $\mathcal{B}_\Omega^n = \sigma(\mathcal{G}_\Omega^n)$, the σ-field generated by rectangles with $B_i \in \mathcal{G}_\Omega$). Let $\mathcal{B}_\Omega^\infty$ be the σ-field generated by all rectangles of the form $B = \{\omega : \omega_i \in B_i, n \leqslant i \leqslant m\}$, $B_i \in \mathcal{B}_\Omega$. Let μ be a probability measure on the measurable space $(\Omega^\infty, \mathcal{B}_\Omega^\infty)$ yielding a probability space $(\Omega^\infty, \mathcal{B}_\Omega^\infty, \mu)$. The sequence of coordinate functions $X_n : \Omega^\infty \to \Omega$ defined by $X_n(\omega) = \omega_n$, $n = \cdots, -1, 0, I, \cdots$ on $(\Omega^\infty, \mathcal{B}_\Omega^\infty, \mu)$ forms a random process and is denoted either by $[\Omega, \mu, X]$ to emphasize alphabet Ω, measure μ, and name X, or simply by μ to emphasize measure, or by $\{X_n\}$ to emphasize name.

Let $T : \Omega^\infty \to \Omega^\infty$ denote the shift transformation defined by $X_n(T\omega) = X_{n+1}(\omega)$. The process μ is stationary if $\mu(TF) = \mu(F)$ for all $F \in \mathcal{B}_\Omega^\infty$. The process is ergodic if $TF = F$ implies $\mu(F) = 0$ or 1.

Denote $(\omega_0, \cdots, \omega_{n-1})$ by ω^n and define $X^n : \Omega^\infty \to \Omega^n$ by $X^n(\omega) = (X_0(\omega), X_1(\omega), \cdots, X_{n-1}(\omega)) = \omega^n$. Let μ^n denote the restriction of μ to $(\Omega^n, \mathcal{B}_\Omega^n)$, that is, if $F \in \mathcal{B}_\Omega^n$, then $\mu^n(F) = \mu(X^n)^{-1}(F) = \mu(\omega : \omega^n \in F)$.

Let $\mathcal{M}_s$ denote the class of all stationary processes with alphabet Ω and let $\mathcal{M}_e$ denote the class of all stationary and ergodic processes with alphabet Ω. To avoid confusion we will often use different names with different measures, e.g., typical members of $\mathcal{M}_e$ are $[\Omega, \mu, X]$ and $[\Omega, \nu, Y]$.

A process $[\Omega, \mu, X]$ is said to be i.i.d. if for every rectangle $B = \times_{i=0}^{n-1} B_i$, $B_i \in \mathcal{B}_\Omega$, we have $\mu^n(B) = \Pi_{i=0}^{n-1} \mu^1(B_i)$. Let $\mathfrak{M}_m$ denote the collection of all i.i.d. or memoryless processes and note that $\mathfrak{M}_m \subset \mathfrak{M}_e \subset \mathfrak{M}_s$.

Given two processes $\mu, \nu \in \mathfrak{M}_s$ the generalized Ornstein distance or $\bar{\rho}$ distance between μ and ν can be defined as follows: for $x^n, y^n \in \Omega^n$ set

$$\rho_n(x^n, y^n) = n^{-1} \Sigma_{i=0}^{n-1} \rho(x_i, y_i)$$

and define $\mathcal{P}(\mu^n, \nu^n)$ as the set of all measures p^n on $(\Omega^n \times \Omega^n, \mathcal{B}_\Omega^n \times \mathcal{B}_\Omega^n)$ having μ^n and ν^n as coordinates, that is, $p^n(\Omega^n \times F) = \nu^n(F)$, $p^n(F \times \Omega^n) = \mu^n(F)$, all $F \in \mathcal{B}_\Omega^n$. Define the nth order distance

$$\bar{\rho}_n(\mu^n, \nu^n) = \inf_{p \in \mathcal{P}(\mu^n, \nu^n)} E_p \rho_n \tag{2.1}$$

and the $\bar{\rho}$ distance by

$$\bar{\rho}(\mu, \nu) = \sup_n \bar{\rho}_n(\mu^n, \nu^n). \tag{2.2}$$

If with a slight abuse of notation we also let X^n and Y^n denote coordinate functions on $\Omega^n \times \Omega^n$ so that if $z = (x^n, y^n) \in \Omega^n \times \Omega^n$, then $X^n(z) = x^n$, $Y^n(y) = y^n$, then (2.1) also can be written

$$\bar{\rho}_n(\mu^n, \nu^n) = \inf_{p \in \mathcal{P}(\mu^n, \nu^n)} E_p \rho_n(X^n, Y^n).$$

Thus $\bar{\rho}_n(\mu^n, \nu^n)$ measures the smallest possible expected "distortion" between X^n and Y^n over all stochastic links preserving the probabilistic description of each. We note $\bar{\rho}_n$ is the Vasershtein-distance between the random vectors X^n and Y^n described by μ^n and ν^n (Vasershtein, (1969)). The following are some useful properties of $\bar{\rho}$ for later use.

Properties of $\bar{\rho}$ (*Gray* et al., (1975)):

(i) $\lim_{n \to \infty} \bar{\rho}_n(\mu^n, \nu^n)$ exists and equals $\sup_n \bar{\rho}_n(\mu^n, \nu^n)$.

(ii) If μ and ν are i.i.d., then $\bar{\rho}(\mu, \nu) = \bar{\rho}_1(\mu^1, \nu^1)$.

(iii) $\bar{\rho}(\mu, \nu) \leqslant \bar{\rho}(\mu, \eta) + \bar{\rho}(\eta, \nu)$ (triangle inequality).

(iv) The distance can also be defined as follows: let $\mathcal{P}_s(\mu, \nu)$ be the collection of all stationary pair processes with coordinate processes μ and ν, that is, all measures p on $(\Omega^\infty \times \Omega^\infty, \mathcal{B}_\Omega^\infty \times \mathcal{B}_\Omega^\infty)$ such that $p(\Omega^\infty \times F) = \nu(F)$, $p(F \times \Omega^\infty) = \mu(F)$, all $F \in \mathcal{B}_\Omega^\infty \times \mathcal{B}_\Omega^\infty$ (where we use T to denote the shift on $\Omega^\infty \times \Omega^\infty$ as well as on Ω^∞). In a similar fashion let $\mathcal{P}_e(\mu, \nu)$ denote the class of all stationary and ergodic pair processes with μ and ν as coordinates. Define the coordinate functions $(X_n, Y_n) : \Omega^\infty \times \Omega^\infty \to \Omega \times \Omega$ by $(X_n, Y_n)(x, y) = (X_n(x), Y_n(y)) = (x_n, y_n)$. We have that

$$\bar{\rho}(\mu, \nu) = \inf_{p \in \mathcal{P}_s(\mu, \nu)} E_p \rho(X_0, Y_0) \tag{2.3a}$$

and if $\mu, \nu \in \mathfrak{M}_e$,

$$\bar{\rho}(\mu, \nu) = \inf_{p \in \mathcal{P}_e(\mu, \nu)} E_p \rho(X_0, Y_0). \tag{2.3b}$$

We note that (2.3b) follows from (2.3a) via the ergodic decomposition of stationary processes (see Oxtoby (1952) or Rohlin (1949)).

Another important property of $\bar{\rho}$ is that it is the closest that generic (typical, regular) sequences of μ and ν (those sequences whose sample averages converge to expectations of enough functions to determine the measure) can be made to each other in a limiting ρ_n sense (Gray et al. (1975)). In the next section we develop a result for sample distributions similar to that of Hampel and Parthasarathy: the existing $\bar{\rho}$ result is not directly useful here because it involves a different type of sample average. The basic idea is that $\bar{\rho}$ closeness of two processes will imply that with high probability the process will produce close sample distributions.

Hampel used the Prohorov metric between μ^1 and ν^1 to measure the distance between i.i.d. processes μ and ν. We can define a Prohorov distance between processes using a generalization of Moser et al. (1975) and this distance can be easily related to $\bar{\rho}$ by using the Strassen-Dudley form for the Prohorov distance (Strassen (1965), Dudley (1968)): Define the nth order Prohorov distance

$$\Pi_n(\mu^n, \nu^n) = \inf_{p \in \mathcal{P}(\mu^n, \eta^n)} \inf\{\gamma : p(x^n, y^n : \rho_n(x^n, y^n) > \gamma) \leqslant \gamma\}, \tag{2.4}$$

which is the Prohorov metric between μ^n and ν^n with respect to the metric ρ_n (which generates the product topology).

It is known (Strassen (1965), Dudley (1968)) that a p_n achieving the infimum exists. We have immediately using Chebychev's inequality (as in Dobrushin (1970)) that if p^n achieves $\bar{\rho}_n$ (i.e., $E\ p_n \rho_n = \bar{\rho}_n$; in the Appendix it is shown that the infimum is a minimum for Polish alphabets), then

$$p^n(x^n, y^n : \rho_n(x^n, y^n) > \varepsilon) \leqslant E p^n \rho_n / \varepsilon \leqq \bar{\rho}_n(\mu^n, \nu^n)/\varepsilon$$

and hence choosing $\bar{\rho}_n(\mu^n, \nu^n) = \varepsilon^2$ yields

$$p^n\big(x^n, y^n : \rho_n(x^n, y^n) > \bar{\rho}_n(\mu^n, \nu^n)^{1/2}\big) \leqslant \bar{\rho}_n(\mu^n, \nu^n)^{1/2}$$

whence

$$\begin{aligned} \Pi_n(\mu^n, \nu^n)^2 &\leqslant \bar{\rho}_n(\mu^n, \nu^n) \\ &\leqslant \bar{\rho}(\mu, \nu), \qquad \text{all} \quad n, \end{aligned} \tag{2.5}$$

so that closeness in $\bar{\rho}$ is stronger than closeness in Prohorov. In some cases the two distances generate the same topology, however, as the following easy lemma shows.

LEMMA 2.1. (a) *If*

$$\rho(x_0, y_0) \leqslant \rho_{\max} < \infty, \qquad \textit{all} \quad x_0, y_0, \tag{2.6}$$

then

$$\bar{\rho}_n(\mu, \nu) \leqslant \Pi_n(\mu^n, \nu^n)(1 + \rho_{\max}).$$

(b) *If there exists an* a^* *so that*

$$E_\mu \rho(X_0, a^*)^2 \leqslant \rho^* < \infty, \tag{2.7}$$

$$E_\nu \rho(Y_0, a^*)^2 \leqslant \rho^* < \infty, \tag{2.8}$$

then

$$\bar{\rho}_n(\mu^n, \nu^n) \leqslant \Pi_n(\mu^n, \nu^n) + 2(\rho^* \Pi_n(\mu^n, \nu^n))^{\frac{1}{2}}.$$

PROOF. (a) Let p^n yield Π_n, then

$$\begin{aligned}\bar{\rho}_n(\mu^n, \nu^n) &\leqslant E_{p^n}\rho_n(X^n, Y^n) \leqslant \Pi_n(\mu^n, \nu^n) + p^n(x^n, y^n : \rho_n(x^n, y^n) \\ &\quad > \Pi^n(\mu^n, \nu^n))\rho_{\max} \\ &= \Pi_n(\mu^n, \nu^n)(1 + \rho_{\max}).\end{aligned}$$

(b) As in (2.9) we have that

$$\bar{\rho}_n(\mu^n, \nu^n) \leqslant \Pi_n(\mu^n, \nu^n) + \int_G dp^n(x^n, y^n)\rho_n(x^n, y^n),$$

where

$$G = \{x^n, y^n : \rho_n(x^n, y^n) > \Pi^n(\mu^n, \nu^n)\}.$$

Let 1_G be the indicator function for G. Let E denote expectation with respect to p^n. Since ρ_n is a metric, we have from the triangle inequality and the Cauchy-Schwarz inequality

$$\begin{aligned}\bar{\rho}_n(\mu, \nu) &\leqslant E\rho_n(X^n, Y^n) \leqslant \Pi_n(\mu^n, \nu^n) + E\, \rho_n(X^n, a^{*n})1_G + E\, \rho_n(Y^n, a^{*n})1_G \\ &\leqslant \Pi_n(\mu^n, \nu^n) + \left(E\, \rho_n(X^n, a^{*n})^2\right)^{\frac{1}{2}}(E\, 1_G^2)^{\frac{1}{2}} + \left(E\, \rho_n(Y^n, a^{*n})^2\right)^{\frac{1}{2}}(E\, 1_G^2)^{\frac{1}{2}}.\end{aligned}$$

Applying the Cauchy-Schwarz inequality for sums and (2.7) yields

$$E\, \rho_n(X^n, a^{*n})^2 = E_\mu\{n^{-1}\Sigma_{i=0}^{n-1}\rho(X_i, a^*)\}^2 \leqslant E_\mu\{n^{-1}\Sigma_{i=0}^{n-1}\rho(X_i, a^*)^2\} \leqslant \rho^*$$

and hence

$$\bar{\rho}_n(\mu, \nu) \leqslant \Pi_n(\mu^n, \nu^n) + 2(\rho^*)^{\frac{1}{2}}(p_n(G))^{\frac{1}{2}} = \Pi_n(\mu^n, \nu^n) + 2(\rho^*)^{\frac{1}{2}}(\Pi_n(\mu^n, \nu^n))^{\frac{1}{2}}.$$

Lemma 2.1 and (2.5) imply that if $\mathfrak{M}$ is a space of processes for which either (a) the metric is bounded, or (b) there is an a^* such that (2.7) holds for all $\mu \in \mathfrak{M}$, then $\bar{\rho}_n$ and Π_n generate the same topology on nth order distributions.

We use $\bar{\rho}$ as our distance measure on processes primarily since it permits a simple demonstration that close processes likely produce sample functions with close sample distributions (as in the next section). Hampel's Prohorov approach worked in the i.i.d. case because he was able to produce an i.i.d. pair process p with the correct coordinate processes by simply taking the product measure with marginal yielding $\Pi_1(\mu^1, \nu^1)$. If μ and ν were not i.i.d., p constructed in this way would not have μ and ν as coordinates. The $\bar{\rho}$-distance avoids this problem since it has an equivalent definition in terms of processes.

An additional advantage of the $\bar{\rho}$-distance is that it is often amenable to explicit evaluation or bounding.

Even though $\bar{\rho}$ is the distance measure used on processes, the Prohorov metric is quite adequate as a measure of distance of random variables, and hence for many intermediate steps we will use the weaker Prohorov distance in order to follow Hampel's basic approach where possible.

3. Sample distributions. Hampel (1975) following Parthasarathy (1967) considers only marginal sample distributions of the following kind: Given an n-tuple $x^n \in \Omega^n$, define the measure $\mu^1_{x^n}$ on $(\Omega, \mathfrak{B}_\Omega)$ by assigning probability n^{-1} to each x_i, $i = 0, 1, \cdots, n-1$ (if, say, k of the x_i are identical, this point gets probability k/n). This assignment gives a measure $\mu^1_{x^n}$ on $(\Omega, \mathfrak{B}_\Omega)$, via

$$\mu^1_{x^n}(F) = \Sigma_{i\,:\,x_i \in F} n^{-1}.$$

Parthasarathy (1967) proves that for an i.i.d. process μ,

$$\Pi_1(\mu^1_{x^n}, \mu^1) \to 0 \quad \text{as } n \to \infty, \mu\text{-a.e.} \tag{3.1}$$

We shall wish to consider more general processes and parameters depending on the whole process and not just the marginal μ^1. Hence we wish to estimate more than just the marginal μ^1 from x^n. Given an n-tuple $x^n \in \Omega^n$ form an estimate of the entire underlying process as follows: form the periodic string $\bar{x} = (\cdots, x^n, x^n, x^n, \cdots)$, that is, $\bar{x}_k = x_{k \bmod n}$. Define the measure μ_{x^n} on $(\Omega^\infty, \mathfrak{B}^\infty_\Omega)$ by placing probability n^{-1} on each string $T^i\bar{x}$, $i = 0, 1, \cdots, n-1$ (grouping together identical strings as before), that is,

$$\mu_{x^n}(F) = \Sigma_{i\,:\,T^i\bar{x} \in F} n^{-1}, \qquad \text{all } F \in \mathfrak{B}^\infty_\Omega. \tag{3.2}$$

The process is periodic as defined by Parthasarathy (1961) since $\mu_{x^n}(F \cap T^nF) = \mu_{x^n}(F)$, all $F \in \mathfrak{B}^\infty_\Omega$. It is also easily seen to be stationary from (3.2). Furthermore, if $TG = G$ and hence $T^{-1}G = G$, then if $T^i\bar{x} \in G$ for any i, $T^j\bar{x} \in G$ for all j; hence $\mu_{x^n}(G) = 0$ or 1 and the process is ergodic. The process μ_{x^n} has restrictions $\mu^k_{x^n}$ which assign measure n^{-1} to each k-tuple obtained by viewing k adjacent symbols within x^n or an "overlap" k-tuple constructed by $(x_i, \cdots, x_{n-1}, x_0, \cdots, x_{k+i-n})$, $i = n-k+1, \cdots, n-1$. In particular, $\mu^1_{x^n}$ is the same as the Parthasarathy marginal sample distribution. Note that only if $k \leqslant n$ are the sample distributions "trustworthy," but it is in fact the sample distributions $\mu^k_{x^n}$, $n \geqslant k$, that will be most important.

Given a stationary and ergodic source $[\Omega, \mu, X]$, then the ergodic theorem implies that for any fixed k

$$\lim_{n\to\infty} \Pi_k(\mu^k_{x^n}, \mu^k) = 0, \qquad \mu\text{-a.e.} \tag{3.3}$$

If, in addition, there exists a reference letter a^* such that (2.7) holds, then for any fixed k (3.3) and Lemma 2.1 imply

$$\lim_{n\to\infty} \bar{\rho}_k(\mu^k_{x^n}, \mu^k) = 0, \qquad \mu\text{-a.e.} \tag{3.4}$$

One might hope that a stronger result would hold to the effect that $\bar{\rho}(\mu_{x^n}, \mu) \to 0$, μ-a.e. That $\bar{\rho}(\mu_{x^n}, \mu) \to 0$ is impossible, however, even for general finite alphabet processes since in that case with ρ being the Hamming (discrete) metric convergence in $\bar{\rho}$ (in this case called $\bar{d}$ and being Ornstein's distance) implies convergence in entropy (Shields (1975)), yet periodic processes have entropy zero and hence cannot converge in $\bar{\rho}$ to a process with nonzero entropy. Roughly speaking, sample distributions can describe the kth order restrictions of a process to arbitrary accuracy as $n \to \infty$ and any fixed k, but they cannot approximate the kth order restrictions for *all* k simultaneously. This observation leads to some of the definitions generalizing those of Hampel to stationary ergodic processes.

4. Sequences of estimators. A sequence of estimators $\{S_n\}$ is a sequence of measurable mappings $S_n : \Omega^n \to \Lambda$, $n = 1, 2, \cdots$, where the parameter space Λ is a Polish space with metric d and $\mathcal{B}_\Lambda$ is the Borel σ-field of subsets of Λ. Unlike Hampel, we do not consider S_n to depend on its argument x^n only through $\mu^1_{x^n}$; that is, $S_n(x^n)$ is not assumed to be invariant under permutations of x^n. In addition, Λ need not be $\mathcal{R}^k$ with the Euclidean metric as in Hampel, allowing more general function spaces. In some cases there will exist a "true" value $S(\mu)$ of the parameter of the process μ being estimated by the sequence $\{S_n\}$. Analogous to a special case considered by Hampel, if $S : \mathfrak{M}_e \to \Lambda$ is the mapping giving the "true" parameter, one candidate for the sequence of estimators is $S_n(x^n) = S(\mu_{x^n})$, the parameter associated with the periodic process obtained from the sample n-tuple. Examples are the sample mean $(S_n(x^n) = n^{-1}\Sigma_{i=0}^{n-1}x_i)$ and sample correlation $(S_n(x^n) = n^{-1}\Sigma_{i=0}^{n-1}x_{i \bmod n}x_{(i+\tau)\bmod n})$ which are simply the mean and correlation of the process μ_{x^n}. Certain results analogous to those of Hampel will be proved for this special case.

DEFINITION. (i) A parameter $S : \mathfrak{M}_S \to \Lambda$ is said to be weakly continuous at μ with respect to the $\bar{\rho}$ distance if given $\varepsilon > 0$ there exists a $\delta > 0$ such that $\bar{\rho}(\mu, \nu) < \delta$ implies $d(S(\mu), S(\nu)) < \varepsilon$.

(ii) A parameter $S : \mathfrak{M}_S \to \Lambda$ is said to be strongly continuous with respect to the $\bar{\rho}$ distance if given $\varepsilon > 0$ there exists a positive integer k and a $\delta > 0$ such that if $\bar{\rho}_k(\mu^k, \nu^k) < \delta$, then $d(S(\mu), S(\nu)) < \varepsilon$.

(iii) A parameter $S : \mathfrak{M}_S \to \Lambda$ is said to be, simply, strongly continuous (or strongly continuous with respect to the Prohorov distance) if given $\varepsilon > 0$ there exists a positive integer k and a $\delta > 0$ such that if $\Pi_k(\mu^k, \nu^k) < \delta$, then $d(S(\mu), S(\nu)) < \varepsilon$.

It follows from the properties of the distance that strong continuity $\Rightarrow$ strong continuity with respect to the $\bar{\rho}$ distance $\Rightarrow$ weak continuity with respect to the $\bar{\rho}$ distance.

The strong notions of continuity are required when considering sample distributions, as there the conditions of Π_k or $\bar{\rho}_k$ being small can be met, while the condition of small $\bar{\rho}$ in general cannot.

If under μ a sequence of estimators $\{S_n\}$ converges in probability (under μ) to a value $S_\infty(\mu)$, that is, if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mu(x : d(S_n(x^n), S_\infty(\mu)) > \varepsilon) = 0, \tag{4.1}$$

then we say $\{S_n\}$ is consistent for $S_\infty(\mu)$ under μ. As pointed out by Hampel, $S_\infty(\mu)$ need not be the same as the "true" parameter value $S(\mu)$, but in such a case $S_\infty(\mu)$ might be a better definition of the "true" parameter given the S_n.

A sequence of estimators $\{S_n\}$ on a process μ induces a family of probability measures $\mu^n S_n^{-1}$ on $(\Lambda, \mathcal{B}_\Lambda)$ defined by

$$(4.2) \qquad \mu^n S_n^{-1}(F) = \mu^n(S_n^{-1}(F)), \qquad \text{all } F \in \mathcal{B}_\Lambda.$$

From (3.4)-(3.5) we have that if $S : \mathcal{M}_S \to \Lambda$ is either (i) a strongly continuous parameter at $\mu \in \mathcal{M}_e$, or (ii) a strongly continuous parameter at $\mu \in \mathcal{M}_e$ with respect to the $\bar{\rho}$ distance and there exists a reference letter in the sense of (2.7), then the sequence of estimators $\{S_n\}$ given by $S_n(x^n) = S(\mu_{x^n})$ is consistent for S at μ.

Lastly, let Π_d denote the Prohorov distance between measures on $(\Lambda, \mathcal{B}_\Lambda)$ with respect to the metric d.

5. Robust sequences.

Definition. Given a collection of processes $\mathcal{M} \subset \mathcal{M}_s$, a sequence of estimators $\{S_n\}$ is robust for $\mathcal{M}$ at a process μ if given $\varepsilon > 0$ there is a $\delta > 0$ such that for all n and all processes $\nu \in \mathcal{M}$

$$(5.1) \qquad (\mathrm{A}) \bar{\rho}(\mu, \nu) < \delta \Rightarrow \Pi_d(\mu^n S_n^{-1}, \nu^n S_n^{-1}) < \varepsilon.$$

The definition is intuitively the same as Hampel's: a robust sequence is one for which close observation processes imply uniformly (over n) close estimate distributions. Hampel defines robustness only at i.i.d. processes and only for $\mathcal{M}_m$, the class of all i.i.d. processes. In the case of $\mathcal{M}_m$, (A) is equivalent to $\bar{\rho}(\mu, \nu) = \bar{\rho}_1(\mu^1, \nu^1)$, the marginal distance, being small. Since $\Pi_1(\mu^1, \nu^1)^2 \leqslant \bar{\rho}_1(\mu^1, \nu^1)$, robustness at an i.i.d. process for $\mathcal{M}_m$ in our sense is slightly weaker than Hampel's robustness. If ρ is bounded or we add the constraint to $\mathcal{M}_m$ that there exist a reference letter as in Lemma 2.1, then for $\mathcal{M}_m$ the two notions for robustness at an i.i.d. process are equivalent.

The following auxiliary definitions will prove useful.

Definition. (i) A sequence of estimators $\{S_n\}$ is asymptotically robust for a collection $\mathcal{M} \subset \mathcal{M}_s$ at μ if given $\varepsilon > 0$ there is a $\delta > 0$ and an n_0 such that for all $n \geqslant n_0$ and processes $\nu \in \mathcal{M}$ (A) holds true.

(ii) A sequence of estimators $\{S_n\}$ is small sample robust for a collection $\mathcal{M} \subset \mathcal{M}_s$ at μ if for any integer n_0 and any $\varepsilon > 0$ there is a $\delta > 0$ such that (A) holds for all $n = 1, 2, \cdots, n_0$.

Obviously if a sequence $\{S_n\}$ is both asymptotically robust and small sample robust for $\mathcal{M}$ at μ, then it is robust for $\mathcal{M}$ at μ.

Definition. Condition (B) is said to be asymptotically satisfied for a sequence of estimators $\{S_n\}$ and a process μ if given $\varepsilon > 0$, $\eta > 0$ there exist positive integers k and n_0 and a $\delta > 0$ and for all $n \geqslant n_0$ a set $F_n \in \mathcal{B}_\Omega^n$ such that

$$(5.2) \qquad \mu^n(F_n) > 1 - \eta$$

and if $x^n \in F_n$, $y^n \in \Omega^n$, and

$$(5.3) \qquad \Pi_k(\mu_{x^n}^k, \mu_{y^n}^k) < \delta,$$

where Π_k is the Prohorov distance with respect to ρ_k as in (2.4), then

$$d(S_n(x^n), S_n(y^n)) < \varepsilon. \tag{5.4}$$

If we forced $n_0 = 1$, then the above condition would be identical to Hampel's except for the fact that we allow a general k (which may depend on ε and n) while he requires $k = 1$. Hence our condition is weaker (his condition (B) implies ours, but not conversely). The following is analogous to Hampel's Lemma 1.

LEMMA 5.1. *If $\mu \in \mathfrak{M}_s$ and $\{S_n\}$ asymptotically satisfy condition* (B), *then $\{S_n\}$ is asymptotically robust at μ.*

PROOF. Choose ε as in (A). For (B) use the same ε, set $\eta = \varepsilon/2$ and let k, δ_B, n_0, F_n be the promised objects for $n \geqslant n_0$. Choose $\delta = \min(\delta_B^4, \varepsilon^2/4)$. Analogous to the construction of the coordinate sample distribution $\mu_{x^n}^k$, define

$$p_{x^n, y^n}(F) = \Sigma_{i\,:\,T^i(\bar{x}, \bar{y}) \in F}\, n^{-1}, \qquad F \in \mathfrak{B}_\Omega^\infty \times \mathfrak{B}_\Omega^\infty,$$

where $\bar{y} = (\cdots, y^n, y^n, \cdots)$, and its restriction $p' = p_{x^n, y^n}^k$ to $\mathfrak{B}_\Omega^k \times \mathfrak{B}_\Omega^k$. By construction $p' \in \mathfrak{P}(\mu_{x^n}^k, \mu_{y^n}^k)$ and p' places probability n^{-1} on each of the first n k-windows in $(\bar{x}, \bar{y})$. We therefore have that

$$\begin{aligned}\bar{\rho}(\mu_{x^n}^k, \mu_{y^n}^k) \leqslant E_{p'}\rho_n = n^{-1}&\Sigma_{i=0}^{n-1}\rho_k(x_i^k, y_i^k)\\ &+ n^{-1}\Sigma_{i=n-k+1}^{n-1}\rho_k((x_i, \cdots, x_{n-1}, x_0, \cdots, x_{k+i-n-1}),\\ &\times (y_i, \cdots, y_{n-1}, y_0, \cdots, y_{k+i-n-1}))\\ = n^{-1}&\Sigma_{i=0}^{n-1}\rho(x_i, y_i) = \rho_n(x^n, y^n)\end{aligned}$$

for all k (and hence $\bar{\rho}(\mu_{x^n}, \mu_{y^n})$ is small if $\rho_n(x^n, y^n)$ is). Let p be the stationary process yielding $E_p\rho(X_0, Y_0) = \bar{\rho}(\mu, \nu)$. We have from (5.5) that

$$E_p\bar{\rho}_k(\mu_{X^n}^k, \mu_{Y^n}^k) \leqslant E_p(n^{-1}\Sigma_{i=0}^{n-1}\rho(X_i, Y_i)) = \bar{\rho}(\mu, \nu) \leqslant \delta, \tag{5.6}$$

and hence from Chebychev's inequality

$$p\left(x, y : \bar{\rho}_k(\mu_{x^n}^k, \mu_{y^n}^k) > \delta^{\frac{1}{2}}\right) < \delta^{\frac{1}{2}};$$

whence

$$p\left(x, y : \Pi_k(\mu_{x^n}^k, \mu_{y^n}^k) > \delta_B\right) \leqslant p\left(x, y : \Pi_k(\mu_{x^n}^k, \mu_{y^n}^k) > \delta^{\frac{1}{4}}\right) < \delta^{\frac{1}{2}}$$

and

$$p\left(x, y : x^n \in F_n, \Pi_k(\mu_{x^n}^k, \mu_{y^n}^k) < \delta_B\right) > 1 - \eta - \delta^{\frac{1}{2}} \geqslant 1 - \varepsilon/2 - \varepsilon/2 = 1 - \varepsilon,$$

which from (B) implies that with probability $1 - \varepsilon \cdot d(S_n(x^n), S_n(y^n)) < \varepsilon$ and hence $\Pi_d(\mu^n S_n^{-1}, \nu^n S_n^{-1}) \leqslant \varepsilon$, completing the proof.

The following definition is a weakened version of one of Hampel's corresponding definitions.

DEFINITION. A sequence of estimators $\{S_n\}$ is continuous at μ if given $\varepsilon > 0$, there exist positive integers k, n_0 and a $\delta > 0$ such that if $n, m \geqslant n_0$, $x^n \in \Omega^n$,

$y^m \in \Omega^m$, and

$$\text{(5.7)} \qquad \Pi_k(\mu^k_{x^n}, \mu^k) < \delta$$
$$\Pi_k(\mu^k_{y^m}, \mu^k) < \delta$$

then

$$\text{(5.8)} \qquad d(S_n(x^n), S_m(y^m)) < \varepsilon.$$

If a single k works for all ε, we say $\{S_n\}$ is continuous of order k at μ (or continuous at μ^k).

Hampel's definition of continuity of an estimator sequence is what we call continuity of order 1 (or at μ^1). Hampel essentially restricts his estimator sequence to depend only on the marginal properties of the process. Analogous to our strong continuity of parameters, we allow the estimator sequence to depend on higher order properties, but for a given $\varepsilon > 0$ there must be a finite k such that matching sample distributions of order k to the underlying μ^k forces the estimators to match up for long observation sequences.

Analogous to Hampel's special case, if a parameter $S : \mathfrak{M}_e \to \Lambda$ is strongly continuous, then the sequence of estimators $\{S_n\}$ defined by $S_n(x^n) = S(\mu_{x^n})$ is continuous.

The following lemma is a strict generalization of Hampel's Lemma 2 since our continuity notion for $\{S_n\}$ is weaker than his.

LEMMA 5.2. *If $\{S_n\}$ is continuous at $\mu \in \mathfrak{M}_e$, then, under μ, $\{S_n\}$ is consistent for some $S_\infty(\mu)$, that is, for any $\delta > 0$*

$$\lim_{n\to\infty} \mu(x : d(S_n(x^n), S_\infty(\mu)) > \delta) = 0.$$

PROOF. For a sequence $\varepsilon_i \downarrow 0$ choose $\delta_i \downarrow 0$ and $n_i \uparrow \infty$ such that the continuity condition is fulfilled for $n, m > n_i$ (for each i). Define for positive integers k, n and $\delta > 0$ the set

$$B_n(k, \delta) = \{x^n : \Pi_k(\mu^k_{x^n}, \mu^k) < \delta\}$$

and note from (3.3) that for fixed k, δ

$$\text{(5.9)} \qquad \lim_{n\to\infty} \mu^n(B_n(k, \delta)) = 1.$$

From the continuity condition, if $x^n \in B_n(k_i, \delta_i)$, $y^m \in B_m(k_i, \delta_i)$, $n, m \geqslant n_i$, then $d(S_n(x^n), S_m(y^m)) < \varepsilon$; and hence the set

$$\text{(5.10)} \qquad G_i = \cup_{n \geqslant n_i} \cup_{x^n \in B_n(k_i, \delta_i)} S_n(x^n) \subset \Lambda$$

has diameter $\text{diam}(G_i) \leqslant 2\varepsilon_i$. Defining the set $S_n(B_n(k_i, \delta_i)) = \cup_{x^n \in B_n(k_i, \delta_i)} S_n(x^n)$, (5.10) can also be written

$$G_i = \cup_{n \geqslant n_i} S_n(B_n(k_i, \delta_i)).$$

Define the set

$$A_i' = \cap_{j=1}^i G_j = \cap_{j=1}^i \cup_{n \geqslant n_i} S_n(B_n(k_j, \delta_j))$$

and let A_i denote the closure of A_i' (A_i will play the role of Hampel's A_i). The A_i are closed and monotone decreasing since $A_i > A_{i+1}$ and diam $A_i \leqslant 2\varepsilon_i \downarrow 0$. Furthermore, the sets A_i are nonempty as can be seen as follows: for fixed i and $n \geqslant n_i$, we have from (5.9) that

$$\mu(x : S_n(x^n) \in A_i) \geqslant \mu(x : S_n(x^n) \in A_i') \geqslant \mu\big(x : S_n(x^n) \in \cap_{j=1}^{i} S_n\big(B_n(k_j, \delta_j)\big)\big)$$
$$\geqslant \mu\big(x : x^n \in \cap_{j=1}^{i} B_n(k_j, \delta_j)\big) \to 1 \quad \text{as } n \to \infty$$

and hence A_i cannot be empty. Since Λ is complete and the A_i are closed, monotone decreasing, and nonempty, from the Cantor intersection theorem, there exists a single point, say $S_\infty(\mu)$, such that $A_i \downarrow S_\infty(\mu)$. Coupled with (5.11), this proves the lemma.

The lemma immediately yields the following.

COROLLARY 5.1. *Given* $\{S_n\}$, μ, $S_\infty(\mu)$ *as in Lemma* 5.2, *given* $\varepsilon > 0$ *there exists a* δ, k, n_0 *such that if* $n \geqslant n_0$ *and*

$$\Pi_k(\mu_{x^n}^k, \mu^k) < \delta,$$

then $d\{S_\infty(\mu), S_n(x^n)) < \varepsilon$.

The following theorem is the main result of this paper and is the analog to Hampel's theorem for stationary and ergodic processes and the general sequence of estimators here considered. We show that continuity of $\{S_n\}$ implies asymptotically robust and that continuity of the S_n considered as point functions implies small sample robust.

THEOREM 5.1. *Let a sequence of estimators* $\{S_n\}$ *and a* $\mu \in \mathfrak{M}_e$ *be such that*

(i) S_n *is continuous as a point function on* Ω^n *for every* n, *that is, given* n, $x^n \in \Omega^n$, $\varepsilon > 0$, *there exists a* $\delta = \delta(n, x^n, \varepsilon)$ *such that* $\rho_n(x^n, y^n) \leqslant \delta$ *implies* $d(S_n(x^n), S_n(y^n)) < \varepsilon$.

(ii) $\{S_n\}$ *is continuous at* μ, μ *stationary and ergodic.*

Then $\{S_n\}$ *is robust for* $\mathfrak{M}_e$ *at* μ.

COMMENTS. Condition (i) might appear different from that of Hampel since we use $\rho_n(x^n, y^n) = n^{-1}\Sigma_{i=0}^{n-1}\rho(x_i, y_i)$ and he uses $\rho_n'(x^n, y^n) = \max_i \rho(x_i, y_i)$. These metrics generate the same (product) topology, however, and hence the notions are equivalent. Recall also that (ii) is weaker than Hampel's corresponding assumption and the observation processes are far more general, but that our conclusion is in general slightly weaker. We also note that for large n our proof parallels Hampel's by proving condition (B). For small n, however, robustness is proved directly from (ii) and our proof is simpler than Hampel's.

PROOF. First choose $\varepsilon > 0$, $n > 0$ for property (B). From Lemma 5.2 and its corollary and (3.3) there exists $S_\infty(\mu)$, $\delta_0 > 0$, $n_0 \geqslant n_1$, k such that for $n \geqslant n_1$

$$\Pi_k(\mu_{x^n}^k, \mu^k) < 2\delta_0 \Rightarrow d(S_\infty(\mu), S_n(x^n)) < \varepsilon/2. \tag{5.12}$$
$$\mu(x : \Pi_k(\mu_{x^n}^k, \mu^k) > \delta_0) < \eta.$$

For $n \geqslant n_0$ define $F_n = \{x^n : \Pi_k(\mu^k_{x^n}, \mu^k) < \delta_0\}$ so that if $x^n \in F_n$ and $\Pi_k(\mu^k_{x^n}, \mu^k_{y^n}) < \delta_0$, then

$$\Pi_k(\mu^k_{y^n}, \mu^k) \leqslant \Pi_k(\mu^k_{y^n}, \mu^k_{x^n}) + \Pi_k(\mu^k_{x^n}, \mu^k) \leqslant 2\delta_0.$$

Hence, from (5.12), $d(S_\infty(\mu), S_n(y^n)) < \varepsilon/2$ and therefore

$$d(S_n(x^n), S_n(y^n)) \leqslant d(S_n(x^n), S_\infty(\mu)) + d(S_\infty(\mu), S_n(y^n)) \leqslant \varepsilon,$$

proving condition (B) is asymptotically satisfied and hence by Lemma 5.1 $\{S_n\}$ is asymptotically robust at μ. Next, given $\varepsilon > 0$ as before and any n, there exists from Parthasarathy ((1967), Theorem 3.2, Chapter 3) a compact set K_n such that

$$\mu^n(K_n) > 1 - \varepsilon/4, \nu^n(K_n) > 1 - \varepsilon/4.$$

Since $S_n : \Omega^n \to \Lambda$, it is uniformly continuous on K_n and hence there is a δ_n such that for $x^n, y^n \in K_n$, $\rho_n(x^n, y^n) < \delta_n$ implies $d(S_n(x^n), S_n(y^n)) < \varepsilon$. Choose δ so that $\delta \leqslant \min(\delta_i^2, i = 1, \cdots, n_0, \varepsilon^2/4)$ and let $p \in \mathcal{P}_e(\mu, \nu)$ yield $\bar{\rho}(\mu, \nu) = E_p\rho(X_0, Y_0) \leqslant \delta$. We have using the Chebychev inequality that

$$p(x, y : d(S_n(x^n), S_n(y^n)) > \varepsilon) \leqslant \mu^n(K_n^c) + \nu^n(K_n^c)$$
$$+ p(x, y : \rho_n(x^n, y^n) > \delta_n) \leqslant \varepsilon$$

and hence

$$\Pi_d(\mu^n S_n^{-1}, \nu^n S_n^{-1}) \leqslant \varepsilon,$$

so that $\{S_n\}$ is also small sample robust.

The only point in the preceding development where ergodicity was required was in the use of (3.3) in Lemma 5.2 ensuring sample distributions of the process μ converged to the actual distribution of μ. The resulting consistency of $\{S_n\}$ at μ was then in turn used to prove asymptotic robustness at μ. In particular, if the process μ is ergodic and we allow the processes ν of Theorem 5.1 to be stationary but not necessarily ergodic, then the entire proof goes through as before giving the following.

COROLLARY 5.2. *Given the conditions of Theorem* 5.1, *then* $\{S_n\}$ *is robust for* $\mathcal{M}_s$ *at* μ.

That robustness for the class of ergodic processes implies robustness for the class of stationary processes also can be seen from the ergodic decomposition theorem of Rohlin (1949). The theorem states, roughly, that every stationary nonergodic process is a mixture of ergodic processes, that is, can be viewed as nature first selecting an ergodic process (unknown to the observer) and then sending a sample function from the ergodic process. Thus, if ν is stationary, the observer will actually see some unknown ergodic component, say ν_θ, of ν and hence robustness for ergodic processes will ensure robustness for stationary nonergodic processes.

COROLLARY 5.3. *Let* $S : \mathcal{M}_s \to \Lambda$ *be such that* S *is strongly continuous at* $\mu \in \mathcal{M}_e$ *and* $S_n(x^n) = S(\mu_{x^n})$ *is a continuous mapping from* Ω^n *to* Λ. *Then* $\{S_n\}$ *is robust for* $\mathcal{M}_e$ *at* μ.

Note that if S *is strongly continuous for all* μ, *then* $S_n(x^n) = S(\mu_{x^n})$ *is automatically continuous as a point function from* (2.5).

Analogous to Hampel's Lemma 3 and corollary we have the following:

LEMMA 5.3. *If $\{S_n\}$ is robust at $\mu \in \mathfrak{M}_e$ and consistent for $S_\infty(\mu)$ at all $\nu \in \mathfrak{M}_s$ in a $\bar{\rho}$ neighborhood of μ, then $S_\infty(\mu)$ is weakly continuous at μ.*

COROLLARY 5.4. *If $\{S_n\}$ is robust and continuous for all $\mu \in \mathfrak{M}_e$, then $S_\infty(\mu)$ is weakly continuous at all μ.*

6. Discussion and applications. Our approach allows the construction of robust estimators for parameters included in the kth order (K finite, fixed) restriction $(\Omega^K, \mathfrak{B}_\Omega^K, \mu^K)$ of an ergodic stationary process $[\Omega, \mu, X]$. Such parameters are the moments of order less than or equal to K.

The M-estimation $S_\infty(\mu)$ of a scalar parameter S included in $(\Omega^K, \mathfrak{B}_\Omega^K, \mu^K)$ will be now the solution (if it exists) of the expression (Huber (1964), Huber (1972))

$$\int_{\mathfrak{B}_\Omega^K} \psi(x_1, \cdots, x_k, S_\infty(\mu))\mu^K(dx_1, \cdots, dx_K) = 0. \tag{6.1}$$

As in the i.i.d. case, the sequence of estimators $\{S_n\}$ defined by ψ $(S_n : \Sigma_{i=1}^{n-K}\psi(x_i, \cdots, x_{i+K}, S_n) = 0)$ is robust if the solution in (6.1) is unique and ψ is bounded. For example, a solution exists if ψ is such that the mapping $s \mapsto \int\psi(x_1, \cdots, x_k; s)\mu^k(dx_1, \cdots, dx_k)$ is bijective and has a continuous inverse.

For the robust estimation of a location parameter, in particular, M-estimators, L-estimators or R-estimators, can be used again (Huber (1972)), where the first order restriction $[\Omega^1, \mathfrak{B}_\Omega^1, \mu^1]$ of the ergodic stationary process $[\Omega, \mu, X]$ is considered. For the M-estimators, we may use the Kth restriction $[\Omega^K, \mathfrak{B}_\Omega^K, \mu^K]$ instead and recover the estimate from the expression:

$$\int_{\mathfrak{B}_\Omega^K} \psi(x_i - S_\infty(\mu), \cdots, x_K - S_\infty(\mu))\mu^K(dx_1, \cdots, dx_K) = 0.$$

The asymptotic distribution of the estimate $S_\infty(\mu)$ can be found by methods similar to the ones used by Huber (1964).

New estimators determined through new functionals of the data may be considered, where the properties of the functionals may be determined through the conditions in Theorem 5.1.

APPENDIX

Equations (2.1) and (2.3a) are actually minima. (The proof is due to P. C. Shields.)

Since Ω and hence Ω^∞ are complete, separable metric spaces, any measure μ on $(\Omega^\infty, \mathfrak{B}_\Omega^\infty)$ is tight, that is, for any $\varepsilon > 0$ there is a compact set F such that $\mu(F) \geqslant 1 - \varepsilon$ (Parthasarathy (1967), Theorem 3.2, page 29). If one has a family of measures such that given ε there is a compact set F such that all members of the family place measure at least $1 - \varepsilon$ on F, then the family is compact in the weak topology (Parthasarathy (1967), Theorem 6.7, page 47). Given μ, ν choose compact $F \in \mathfrak{B}_\Omega^\infty$ such that $\mu(F) \geqslant 1 - \varepsilon/2$, $\nu(F) \geqslant 1 - \varepsilon/2$; then if $p \in \mathcal{P}_s(\mu, \nu)$, $p(F \times F) \geqslant 1 - \varepsilon$ and $F \times F$ is compact. Thus $\mathcal{P}_s(\mu, \nu)$ is compact in the weak topology and a sequence $p_n \in \mathcal{P}_s(\mu, \nu)$ such that

$$E_{\tau_n}\rho(X_0, Y_0) \leqslant \bar{\rho}(\mu, \nu) + 1/n$$

will have a subsequence—say p_{n_k}— that converges in the weak topology to a limiting p. The limit $p \in \mathcal{P}_s(\mu, \nu)$ and $E_p\rho(X_0, Y_0) = \bar{\rho}(\mu, \nu)$, completing the proof. The same argument applied to $(\Omega^n, \mathcal{B}_\Omega^n)$ shows that $\bar{\rho}_n$ is also actually a minimum.

Acknowledgment. The authors gratefully acknowledge the numerous constructive comments and corrections of Robert Fontana of Carnegie Mellon University and of an anonymous referee.

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

DUDLEY, R. M. (1968). Distances of probability measures and random variables. *Ann. Math. Statist.* **39** 1563–1572.

GRAY, R. M., NEUHOFF, D. L. and OMURA, J. K. (1975). Process definitions of distortion-rate functions and source coding theorems. *IEEE Trans. Information Theory* **IT-21** 524–532.

GRAY, R. M., NEUHOFF, D. L. and SHIELDS, P. C. (1975). A generalization of Ornstein's $\bar{d}$ distance with applications to information theory. *Ann. Probability* **3** 315–328.

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

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

HUBER, P. J. (1972). Robust statistics. A review. *Ann. Math. Statist.* **43** 1041–1067.

MOSER, J., PHILLIPS, E. and VARADHAN, S. (1975). *Ergodic Theory: A Seminar.* Courant Institute of Math. Sciences, New York.

OXTOBY, J. (1952). Ergodic sets. *Bull. Amer. Math. Soc.* **58** 116–136.

PARTHASARATHY, K. R. (1961). On the category of ergodic measures. *Ill. J. Math.* **5** 648–656.

PARTHASARATHY, K. R. (1968). *Probability Measures on Metric Spaces.* Academic Press, New York.

ROHLIN, V. A. (1949). Selected topics from the metric theory of dynamical systems. *Uspehi Mat. Nauk.* **4** 57–128. (*AMS Translations* (2), **49** 171–240.)

SHIELDS, P. C. (1975). *The Theory of Bernoulli Shifts.* Univ. of Chicago Press.

STRASSEN, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.* **36** 423–429.

VASHERSTEIN, L. N. (1969). Markov processes on countable product space describing large systems of automata. *Problemy Peredači Informacii* **5** 64–73.

DEPARTMENT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
UNIVERSITY OF CONNECTICUT
STORRS, CONNECTICUT 06268

DEPARTMENT OF ELECTRICAL
ENGINEERING, DURAND 133
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94035