The Annals of Statistics
1985, Vol. 13, No. 2, 768–774

RATES OF CONVERGENCE OF MINIMUM DISTANCE ESTIMATORS AND KOLMOGOROV'S ENTROPY[1]

BY YANNIS G. YATRACOS[2]

University of California, Berkeley

Let $(\mathscr{X}, \mathscr{A})$ be a space with a σ-field, $M = \{P_s; s \in \Theta\}$ be a family of probability measures on $\mathscr{A}$ with Θ arbitrary, $X_1, \cdots, X_n$ i.i.d. observations on P_θ. Define $\mu_n(A) = (1/n) \sum_{i=1}^n I_A(X_i)$, the empirical measure indexed by $A \in \mathscr{A}$. Assume Θ is totally bounded when metrized by the L_1 distance between measures. Robust minimum distance estimators $\hat{\theta}_n$ are constructed for θ and the resulting rate of convergence is shown naturally to depend on an entropy function for Θ.

1. Introduction. A common problem in statistics is the following: given n independent identically distributed observations with joint distribution $P_{\theta,n}$ try to estimate θ, when θ is an element of a finite dimensional space Θ. There are cases where the parameter of interest θ is an element of an infinite dimensional space however, such as in problems of density estimation and nonparametric regression. It is well known that parametric methods (e.g. maximum likelihood, method of moments) fail in this situation.

The aim of this paper is to provide uniformly consistent robust minimum distance estimators $\{\hat{\theta}_n\}$ for the true parameter θ when Θ has no structure or is infinite dimensional, and to show that the rate of convergence depends on the "massiveness" of the space of measures. A particular application of the method is considered for the problem of estimation of smooth densities. The resulting rate of convergence is the one established by Farrell (1972), Birgé (1983) and others.

In Section 2 we give notations, definitions and a description of the minimum distance principle. In Section 3 we construct the estimates and give applications to density estimation. In Section 4 we discuss robustness of the proposed estimator. In Section 5 a few remarks are made on the validity of the model when Θ is finite dimensional and on the optimality of the resulting rate of convergence.

2. Notations. Definitions. The minimum distance principle. Let $M = \{P_s; s \in \Theta\}$ be a family of measures on a set $\mathscr{X}$ with σ-field $\mathscr{A}$. No structure is assumed for the index set Θ. Let $X_1, \cdots, X_n$ be independent identically distributed P_θ random variables, $\mu_n(A) = (1/n) \sum_{i=1}^n I_A(X_i)$ be the empirical measure indexed by $A \in \mathscr{A}$. Define P_θ^n to be the nth product measure on $(\mathscr{X}^n, \mathscr{A}^n)$. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$ be an estimator of θ.

Received October 1983; revised October 1984.

[1] This research was partially supported by the Hellenic Government and UNESCO. In 1982 it was also partially supported by National Science Foundation Grant MCS80-02698.

[2] The author is currently at Rutgers University.

AMS 1980 *subject classifications.* Primary 62G05, secondary 62G30.

Key words and phrases. Minimum distance estimation, rates of convergence, Kolmogorov's entropy, density estimation.

Since Θ has no structure, its role is artificial and what really counts are the measures. So let us metrize Θ with the L_1 distance d between measures. That is, $d(s, t) = \| P_s - P_t \| = 2 \cdot \sup\{| P_s(A) - P_t(A) |; A \in \mathscr{A}\}$. If both P and Q are dominated by μ then

$$\| P - Q \| = 2[P(\{x: (dP/d\mu)(x) > (dQ/d\mu)(x)\}) - Q(\{x: (dP/d\mu)(x) > (dQ/d\mu)(x)\})].$$

Make the usual identifiability assumption.

DEFINITION. A sequence of estimators $\{\hat{\theta}_n(X_1, \cdots, X_n)\}$ is uniformly consistent for θ with rate of convergence δ_n with respect to d if for every $\varepsilon > 0$ there is $b(\varepsilon) > 0$ such that $\sup_{\theta \in \Theta} P_\theta^n[(X_1, \cdots, X_n): d(\hat{\theta}_n, \theta) > b(\varepsilon) \cdot \delta_n] < \varepsilon$ for every $n \geq 1$.

LEMMA 1 (Hoeffding, 1963). *Let $X_1, \cdots, X_n$ be independent random variables such that $0 \leq X_j \leq 1$, $j = 1, \cdots, n$. Let $S_n = \sum_{j=1}^n X_j$, $ES_n = n \cdot p$. Assume that $p \leq \frac{1}{2}$. Then $P[S_n \geq np + K] \leq \exp\{-K^2/2(np + K)\}$ and $P[S_n \leq np - K] \leq \exp\{-K^2/2np(1 - p)\}$.*

DEFINITION. Let (Y, d) be a totally bounded metric space. For $a > 0$ let $N(a)$ be the smallest number of d-balls of radius a that cover Y. The function $\log_2 N(a)$ is called the d-Kolmogorov entropy of the space Y.

The minimum distance principle was formalized to a general principle by Wolfowitz (1957). LeCam (1966), Beran (1977a), Pollard (1980) used the principle to construct estimators for the case where Θ is finite dimensional. Pfanzagl (1968) considered the problem for Θ infinite dimensional. Millar (1981, 1983) treated a parametric estimation problem in a nonparametric situation. He proved that minimum distance estimates enjoy optimality properties.

For a distance d_M defined on the set of measures the minimum distance estimator $\hat{\theta}_n$ is such as

$$d_M(P_{\hat{\theta}_n}, \mu_n) = \inf_\theta d_M(P_\theta, \mu_n).$$

Without loss of generality we have assumed that the infimum is achieved. If not, use any $\hat{\theta}_n$ that brings $d_M(P_\theta, \mu_n)$ within γ_n of its infimum, with γ_n decreasing rapidly to zero.

3. Construction of estimates.

LEMMA 2. *Let $M = \{P_s; s \in \Theta\}$ be an L_1-totally bounded family of probability measures on $(\mathscr{X}, \mathscr{A})$. Then for every $a > 0$ there exists a class of sets $F_a \subseteq \mathscr{A}$ of cardinality $| F_a | \leq N^2(a)$ such that $\| P_s - P_t \| \leq 4a + 2 \cdot \sup\{| P_s(A) - P_t(A) |; A \in F_a\}$ for every P_s, P_t in M.*

PROOF. We will use the triangle inequality and properties of the L_1-norm.

Fix $a > 0$. By hypothesis there are $N(a)$ L_1-balls covering M. Let $P_1, \cdots, P_{N(a)}$ be the centers of the balls, P_s, P_t be in M. Assume without loss of generality

$P_s(P_t)$ belongs to the balls with center $P_1(P_2)$. Let $A_{1,2} = \{x\colon (dP_1/d\mu)(x) > (dP_2/d\mu)(x)\}$. Then we have:

$$\begin{aligned}\| P_s - P_t \| &\le \| P_s - P_1 \| + \| P_1 - P_2 \| + \| P_2 - P_t \| \\ &\le 4a + 2 \cdot \sup\{ | P_s(A) - P_t(A) | ; A \in F_a\}\end{aligned}$$

where $F_a = \{\{x\colon (dP_i/d\mu)(x) > (dP_j/d\mu)(x)\}$, for $1 \le i \le j \le N(a)\}$. Obviously $| F_a | = N(a)(N(a) - 1)/2 \le N^2(a)$.

THEOREM 1. *If M is L_1-totally bounded, there exists a uniformly consistent estimator $\hat{\theta}_n$ for θ whose rate of convergence a_n satisfies the equation $a_n = [(\log N(a_n))/n]^{1/2}$ (provided the entropy is such that $a_n \to 0$).*

PROOF. Fix a sequence $\{a_n\}$. Let $P_1, \cdots, P_{N(a_n)}$ be the centers of L_1-balls of radius a_n covering M and F_{a_n} be the family of sets of cardinality $| F_{a_n} | < N^2(a_n)$ as defined in Lemma 2. Define pseudo-distances $d_{M,n}$ by $d_{M,n}(P, Q) = 2 \cdot \sup\{ | P(A) - Q(A) | ; A \in F_{a_n}\}$. Let $\hat{\theta}_n$ be the minimum distance estimator for $d_{M,n}$ chosen among the centers of the balls.

We have to prove that $d(\hat{\theta}_n, \theta) \to 0$ uniformly in P_θ^n probability and find the rate of convergence.

From Lemma 2,

$$d(\hat{\theta}_n, \theta) = \| P_{\hat{\theta}_n} - P_\theta \| \le 4a_n + 2d_{M,n}(\mu_n, P_\theta). \tag{1}$$

From Lemma 1,

$$P_\theta^n[d_{M,n}(\mu_n, P_\theta) > K_n] \le 2 \cdot N^2(a_n) \cdot \exp\{-(nK_n^2/(2K_n + 1))\}. \tag{2}$$

By choosing $K_n = 10(\log N(a_n)/n)^{1/2}$ the optimal rate of convergence becomes a_n where a_n satisfies the equation $a_n = [(\log N(a_n))/n]^{1/2}$.

We offer now the form of $b(\varepsilon)$ (defined in Section 2) for the proposed estimator.

PROPOSITION 1. *Under the conditions of Theorem 1, for the derived minimum distance estimators $\{\hat{\theta}_n\}$ for $0 < \varepsilon < \varepsilon_0 < 1$, $b(\varepsilon) = C_{\varepsilon_0} \cdot (\log(1/\varepsilon))^{1/2}$.*

PROOF. Following the proof of the theorem, we have that

$$\begin{aligned}P_\theta^n[\| P_{\hat{\theta}_n} - P_\theta \| &> b(\varepsilon) \cdot [(\log N(a_n))/n]^{1/2}] \\ &\le 2 \cdot N^2(a_n) \cdot \exp\{-b^2(\varepsilon) \cdot \log N(a_n)/C_1\} = C/N(a_n)^{(b^2(\varepsilon)/C_1)-2},\end{aligned}$$

with C, C_1 positive constants.

We require the right-hand side of the above inequality to be less than ε uniformly in n. Observe that $N(a_n)$ increases as n increases, so it is enough to define $b(\varepsilon)$ in such a way that $C/N(a_1)^{(b^2(\varepsilon)/C_1)-2} < \varepsilon$. This implies that for $0 < \varepsilon < \varepsilon_0 < 1$, $b(\varepsilon) = C_{\varepsilon_0} \cdot (\log(1/\varepsilon))^{1/2}$.

Some applications. (i) Let $\mathscr{X} = [0, 1]^s$, $\Theta = \{f\colon [0, 1]^s \to \mathbb{R}^+, f$ is a density with p derivatives and pth derivative satisfying a Lipschitz condition $| f^{(p)}(x) - f^{(p)}(y) | \le L \cdot | x - y |^b$ (or: $\int_0^1 | f^{(p)}(x + h) - f^{(p)}(x) | dx \le L \cdot h^b$), $q = p + b\}$.

Kolmogorov and Tihomirov (1959) have shown that Θ metrized with sup norm $\| \cdot \|_\infty$ (which bounds the L_1 norm) is totally bounded, that for every $a > 0$, the number $N_\infty(a)$ of $\| \cdot \|_\infty$-balls of radius a is $N_\infty(a) \sim 2^{(1/a)^{(s/q)}}$ and gave a construction for the centers of balls. Clements (1963) has shown that Θ metrized with L_1 distance can be covered by $N_{L_1}(a) \sim 2^{(1/a)^{(s/q)}}$ balls of radius a. Performing minimum distance estimation we get as rate of convergence $a_n = n^{-(q/(2q+s))}$. By the Borel-Cantelli theorem the convergence holds almost surely.

Further generalization can be considered for the case $\mathscr{X} = \mathbb{R}$. We will be dealing with increasing sequences of compacts K_n of length ℓ_n. For every K_n,

$$N_{K_n}(a_n) \sim 2^{(1/a_n)^{1/q} \cdot \ln}$$

By choosing ℓ_n appropriately we can get a rate as close as we like to $n^{-(q/(2q+1))}$ (but not uniformly).

(ii) Let $\mathscr{X} = [0, 1]$, $\Theta = \{f: [0, 1] \to \mathbb{R}^+, f$ is density with a modulus of continuity $\omega_f(\delta) = \sup\{|f(x + h) - f(x)|; x \in [0, 1], |h| < \delta\} \leq \omega(\delta)\}$. By Lorentz (1966) Θ metrized with the sup norm $\| \cdot \|_\infty$ is totally bounded, $N_\infty(a) \leq K/\delta(\gamma \cdot a)$ for K, γ fixed constants and $\delta = \delta(a)$ defined as any root of the equation $\omega(\delta) = a$.

(iii) Let $\Theta(K, G, C) = \{f: G \to \mathbb{R}^+, f$ is a density and analytic, G is an arbitrary region of a one-dimensional complex space, K is a simply connected continuum, $K \subseteq G$, $\sup\{f(z); z \in G\} \leq C\}$ be metrized with the sup norm. Then $\log N_\infty(a) \sim (\log(1/a))^2$. The resulting rate of convergence is $\log n/n^{1/2}$.

4. Robustness properties of the proposed estimators. There are many different definitions of robustness. A great deal of those are devoted to the robustness of different estimates of a translation parameter in the independent identically distributed case against some departures from the underlying distribution. We should mention that Beran (1977a), Millar (1981) and Parr and Schucany (1980) examine robustness properties of minimum distance estimators.

We will follow Sture Holm's idea as it appears in Bickel (1976). This is the situation where randomness is due to variations between individuals and there is no observational error. The judgement of the estimates should then be based not on the errors in estimating the parameters themselves but in estimating the probability distributions.

So we suppose that the true measure P is in a neighborhood of the parametric model $M = \{P_s; s \in \Theta\}$. Define $M(\varepsilon) = \{Q: \| Q - P_s \| < \varepsilon$ for some $s \in \Theta$, Q probability measure on $\mathscr{A}\}$. Note that $(1 - t) \cdot P_s + t \cdot H \in M(\varepsilon)$ for H a probability measure on $\mathscr{A}$, $0 < t < \varepsilon/2$.

DEFINITION. A sequence of estimators $\{P_{\hat{s}_n}\}$ is robust if for a fixed constant ℓ and every $\varepsilon > 0$,

$$\sup\{E_{P^n} \| P_{\hat{s}_n} - P \|; P \in M(\varepsilon)\} < \ell \cdot \varepsilon + \gamma_n,$$

for all n, where γ_n tends to 0 at the rate at which the estimator $P_{\hat{s}_n}$ converges to the true measure P.

Note that ε is allowed to depend on n and tend to 0.

PROPOSITION 2. *The proposed sequence of minimum distance estimators is robust.*

PROOF. Using the triangle inequality, the method of construction of $\hat{\theta}_n$ and that $P \in M(\varepsilon)$, we have:

$$(1) \qquad E_{P^n} \| P_{\hat{\theta}_n} - P \| \leq 3\varepsilon + 4a_n + 4E_{P^n}\sup\{|\mu_n(A) - P(A)|; A \in F_{a_n}\}.$$

Let us consider now the last term of (1). Let

$$W_n = \sup\{|\mu_n(A) - P(A)|; A \in F_{a_n}\}.$$

Note $W_n \leq 1$. Then

$$E_{P^n} W_n \leq K_n + N^2(a_n) \cdot \exp\{-nK_n^2/(2K_n + 1)\}$$

by Hoeffding. So (1) becomes

$$(2) \qquad E_{P^n} \| P_{\hat{\theta}_n} - P \| \leq 3\varepsilon + 4a_n + 4K_n + 2N^2(a_n) \cdot \exp\{-nK_n^2/(2K_n + 1)\}.$$

Assume $K_n = c \cdot [(\log N(a_n))/n]^{1/2}$, $c > 2$ being a constant fixed for all n. By (2) and the choice of K_n,

$$E_{P^n} \| P_{\hat{\theta}_n} - P \| \leq 3\varepsilon + 4a_n + 4ca_n + \frac{2N^2(a_n)}{N(a_n)^{c^2(2ca_n+1)}}$$

with $c^2/(2ca_n + 1) > 2$. The last term is $\sim (N(a_n))^{-(c^2-2-4ca_n)/(2ca_n+1)} \leq N(a_n)^{-c^*}$, c^* being a fixed constant. So what we really want is $N(a_n)^{-c^*} \leq a_n$.

By the relation $a_n = (\log N(a_n)/n)^{1/2}$ we get that $N(a_n) = \exp(na_n^2)$ so $\exp(-c^* n a_n^2) \leq a_n$ for n big.

5. Some more remarks.

5.1. A natural question is what kind of results one can get with the proposed method when Θ is a subset of $\mathbb{R}^k$. To answer this question we should remember that we provide estimators uniformly consistent in L_1 distance, so if there is a nice continuity relation between the Euclidean distance and L_1, our estimator will behave well for the Euclidean distance.

To illustrate, we consider now a family of densities of the form $f_\theta(x) = .5 \cdot \exp(-|x - \theta|)$, $x \in \mathbb{R}$, $\theta \in [0, 1]$. It is known that for this case we can construct, with the classical methods, $n^{-1/2}$ uniformly consistent estimators. Since $N(\varepsilon) \sim 1/\varepsilon$ for ε small, Theorem 1 gives a rate of convergence $(\log n/n)^{1/2}$, which is not as good as the achievable $n^{-1/2}$. The reason is partly due to the fact that the present method, intended for situations where little is assumed about the structure of the family of measures, does not take into account the special features available in the example.

By using instead LeCam's notion of "dimension" and his estimation method (1975), the right $n^{-1/2}$ rate is achieved. As LeCam communicated to me, the use in a method of the notion of either entropy $N(a)$ or dimension $D(a)$ provides equivalent results if $N(a)$ increases rapidly as a tends to 0.

More precisely, in all our derivations, we have used Kolmogorov's entropy $H(a) = \log_2 N(a)$. Birgé (1983) and LeCam (1975), use instead the concept of

dimension $D(a)$ (with respect to Hellinger distance) where $2^{D(a)}$ is the number of sets of diameter a needed to cover a set of diameter $2a$. This allows refinements that would give the usual $n^{-1/2}$ rate for the double exponential used above as an example.

The refinements may be useful in some situations. However one should note that the a-entropy, $H(a) = \log_2 N(a)$ satisfies the inequality

$$H(a) \leq \sum_{k=0}^{m} D(a \cdot 2^k)$$

where m is $[\log(1/a)] + 1$. Thus the difference between using "entropy" and using "dimension" is not going to affect the rate of convergence unless $D(a)$ (or $H(a)$) increase rather slowly as a tends to 0. The preceding inequality can be used to show that replacement of $D(a)$ by $H(a)$ will not change rates of convergence unless $a^\gamma \cdot D(a)$ converges to 0 as a tends to 0 for every $\gamma > 0$.

Although it is easy to construct families with $D(a)$ converging to ∞ and $a^\gamma \cdot D(a)$ converging to 0 for every $\gamma > 0$, none seem to have occurred in ordinary practice. The ones where $D(a)$ remains bounded are plentiful. Most of the usual parametric families are of that nature. For them, other more specialized techniques can always be used.

We should mention also that the proposed method, using rough approximations (as one can see from (2) of Section 3 where we approximate the probability of a union by a sum of probabilities) and not taking into account special features of the measures, cannot compete with methods based on a finite dimensional parametrization. In the same context, another weakness comes from the fact that problems finite dimensional in Euclidean distance might be infinite dimensional in Hellinger or L_1 distance. Such an example (Dacunha-Castelle, 1978) is given by the translation family $f(x - \theta)$, $\theta \in (-1, 1)$, $x \in \mathbb{R}$ with

$$f(x) = c(a) \cdot \exp\{-x^2\}/|x| \cdot |\log|x||^{1+a}, \quad a > 0$$

which is one-dimensional in Euclidean distance but infinite dimensional in Hellinger distance h, since

$$h^2(\theta, 0) = \frac{1}{2}\int_{-\infty}^{+\infty} (f^{1/2}(x) - f^{1/2}(x - \theta))^2 \, dx \sim \frac{1}{a(\log|\theta|)^a}.$$

In this example and similar ones, there exist estimates converging at a much better rate than the rates obtainable through methods not relying on special features of the parametric family.

5.2. We would like at this point to stress that the described method offers upper bounds for the L_1 risk for each fixed sample size n. In the literature, people are usually interested in proving that this upper bound is actually the same with the minimax risk modulo a constant independent of the sample size. This is true for the estimators obtained with the described method when the measures have smooth densities on a compact as results from comparison of rates with other estimates (e.g. kernel estimates) possessing this property. It is easy to prove that the same result holds for L_1 totally bounded families of measures satisfying regularity conditions, such as those of Birgé (1983), insuring the existence for

each n of at least one density which is difficult to detect due to a number $m(n)$ of other densities in the family which are a perturbation of it.

Acknowledgment. This work is part of the author's Ph.D. Thesis at the University of California, Berkeley, written under the supervision of Professor Lucien M. LeCam whose guidance and suggestions are gratefully acknowledged and appreciated. Thanks are also due to the referee and especially the associate editor for useful suggestions and the prompt processing of this paper.

REFERENCES

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

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

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. verw. Gebiete* **65** 181–237.

CLEMENTS, G. F. (1963). Entropies of several sets of functions. *Pacific J. Math.* **13** 1085–1097.

DACUNHA-CASTELLE, D. (1978). Vitesse de convergence pour certains problèmes statistiques. *École d'été de Saint-Flour VII. Lecture Notes in Math.* Springer-Verlag, New York.

FARRELL, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43** 170–180.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–31.

KOLMOGOROV, A. N. and TIHOMIROV, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspehi Mat. (N.S.)* **14** 2(86), 3–86 (in Russian); (1961) *Amer. Math. Soc. Transl. 2* **17** 277–364.

LECAM, L. M. (1966). Likelihood functions for large numbers of independent observations. *Festschrift for J. Neyman.* 167–187. Editor F. N. David. Wiley, New York.

LECAM, L. M. (1975). On local and global properties in the theory of asymptotic normality of experiments. *Stochastic Processes and Related Topics 1*, 13–54. Academic, New York.

LECAM, L. M. (1985). *Asymptotic Methods in Statistical Decision Theory.* Forthcoming book.

LORENTZ, G. G. (1966). *Approximation of Functions.* Holt, Rinehart and Winston, New York.

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

MILLAR, P. W. (1983). A general approach to the optimality of minimum distance estimators. Preprint.

PARR, W. C. and SCHUCANY, W. R. (1980). Minimum distance and robust estimation. *J. Amer. Statist. Assoc.* **75** 616–625.

PFANZAGL, J. (1968). On the existence of consistent estimates and tests. *Z. Wahrsch. verw. Gebiete* **10** 43–62.

POLLARD, D. (1980). The minimum distance method for testing. *Metrika* 43–71.

WOLFOWITZ, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28** 75–88.

YATRACOS, Y. G. (1983). Uniformly consistent estimates and rates of convergence via minimum distance methods. Ph.D. Thesis, University of California, Berkeley.

DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
NEW BRUNSWICK, NEW JERSEY 08903