PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 5, Fasc. 1 (1985), p. 21-29

NON-ASYMPTOTIC MINIMAX RISK FOR HELLINGER BALLS

BY

LUCIEN BIRGÉ (NANTERRE)

Abstract. The following intuitively evident result is shown.

Given a probability P and a radius r, assume that we have to estimate an unknown law belonging to a sphere with centre P and radius r for the Hellinger distance using n independent identically distributed observations. If the risk is measured by the square of the Hellinger distance, then the observations carry no information and the best estimator is just the centre P of the sphere.

1. Introduction. A number of recent papers have been devoted to computations of the minimax risk when we want to estimate some parameter or some density by observing n i.i.d. variables, the loss function being the square of Hellinger distance, (see e.g. [1], [2], [5], and [7]). When we choose to consider such problems from a robustness point of view we are led to consider parameter spaces which contain full Hellinger balls. Then the global risk is certainly larger than the risk computed on such a ball. That is why it may be interesting to compute the exact minimax risk for a Hellinger ball. Actually, with such an enormous parameter space, which is highly infinite-dimensional, it is intuitively obvious that a finite number of observations can give no information about the true underlying law but, as far as I know, there is no such proof in the statistical literature. The purpose of the paper is therefore to give a proof of this result with simple corollaries. But before going to it, I shall first recall a few definitions.

The set of all probability measures on the measured space $(\Omega, \mathcal{A})$ will be denoted by $\mathcal{P}_1(\Omega)$. If P, Q belong to $\mathcal{P}_1(\Omega)$ and are absolutely continuous with respect to some positive measure μ, the *Hellinger distance* between P and Q is defined by

$$h^2(P, Q) = \tfrac{1}{2} \int \left(\sqrt{\frac{dP}{d\mu}} - \sqrt{\frac{dQ}{d\mu}} \right)^2 d\mu$$

and independent of μ. Closely related to it is the *Hellinger affinity*:

$$\varrho(P, Q) = \int \sqrt{\frac{dP}{d\mu}\frac{dQ}{d\mu}}\, d\mu = 1 - h^2(P, Q).$$

For sake of simplicity we shall generally omit μ and write

$$\varrho(P, Q) = \int \sqrt{dP\, dQ}.$$

We shall denote by $\mathscr{B}(P, r)$ the closed Hellinger ball of center P and radius r, i.e.

$$\mathscr{B}(P, r) = \{Q \in \mathscr{P}_1(\Omega) | h(P, Q) \leqslant r\},$$

and notice that $\mathscr{B}(P, 1) = \mathscr{P}_1(\Omega)$ for any P.

The *Kullback information number* $K(P, Q)$ is defined by

$$K(P, Q) = \begin{cases} \int \operatorname{Log} \frac{dP}{dQ}\, dP & \text{if } P \ll Q, \\ +\infty & \text{in the other cases.} \end{cases}$$

Suppose we are given an increasing function g of x, for $x \geqslant 0$, with $g(0) = 0$, and a subset $\mathscr{P}$ of $\mathscr{P}_1(\Omega)$ (the parameter space). Given n i.i.d. observations $X_1, \ldots, X_n$ of law $P \in \mathscr{P}$, an estimator of P is any measurable function $T_n(X_1, \ldots, X_n)$ with values in the metric space $(\mathscr{P}_1(\Omega), h)$. The *minimax risk* is defined by

$$R_n(\mathscr{P}, g) = \inf_{T_n} \sup_{P \in \mathscr{P}} \mathbb{E}_P [g \circ h(P, T_n)].$$

We want to compute quantities of the type

$$R_n(\mathscr{B}(P_0, r), g), \quad 0 < r \leqslant 1.$$

2. Construction of special nets of probabilities on an infinite space. We shall always suppose that Ω and $\mathscr{A}$ are infinite. In this case we can find a sequence $\{A_i\}_{i \geqslant 1}$ of measurable subsets of Ω which do not intersect.

Choose four numbers a, b, k, and m in the following way: $k, m \in N$; $m \geqslant 1$, $k \geqslant 2$; $0 < b < 1 < a < k$; $a, b \in \mathbf{R}^+$ and

$$a + (k-1)b = k. \tag{1}$$

Put $l = mk$ and select a probability μ such that $\mu(A_i) = l^{-1}$ for $i = 1, \ldots, l$. For any subset $I = \{i_1; \ldots; i_m\}$ of cardinal m of $L = \{1; 2; \ldots; l\}$ define A_I to be $\bigcup_{i \in I} A_i$ and A_{I^c} in an obvious way. We shall define P_I by its density with respect to μ:

$$\frac{dP_I}{d\mu}(x) = \begin{cases} a & \text{if } x \in A_I, \\ b & \text{if } x \in A_{I^c}. \end{cases} \tag{2}$$

P_I is easily seen to be a probability because of (1). $\mathscr{I}_m$ is the set of all different I's and its cardinal is equal to $\binom{l}{m}$. It is easily seen from (2) that if P_I and P_J are such that Card $(I \cap J) = p$, $0 \leqslant p \leqslant m$, we get

$$\begin{aligned} \varrho(P_I, P_J) &= l^{-1}[pa + 2(m-p)\sqrt{ab} + (l+p-2m)b] \\ &= 1 - \frac{m-p}{l}(\sqrt{a}-\sqrt{b})^2; \end{aligned} \tag{3}$$

$$h^2(P_I, P_J) = \frac{m-p}{l}(\sqrt{a}-\sqrt{b})^2.$$

Let us denote by $\mathscr{P}_m$ the set $\{P_I\}_{I \in \mathscr{I}_m}$ and, for sake of simplicity, since $\mathscr{I}_m$ is finite, just identify it to $\{1; 2; \ldots; \binom{l}{m}\} = M$, so that P_I becomes P_s for some integer s. We are now able to prove the following

PROPOSITION 1. *For arbitrary* ε, $0 < \varepsilon < 1$, *there exist values* a, b, k, *depending only of* ε *and such that for all large* m *there exists a subset* S_m *of* $M = \{1; 2; \ldots; \binom{km}{m}\}$ *with*:

$$h(P_s, P_t) > 1 - \varepsilon \qquad \forall\, s, t \in S_m,\ s \neq t; \tag{4}$$

$$\frac{dP_s}{dP_t} \leqslant \frac{a}{b} \qquad P_s + P_t \text{ p.s. } \forall\, s, t \in S_m; \tag{5}$$

$$\text{Log [Card } S_m] \geqslant \frac{m}{100}. \tag{6}$$

Proof. First choose k large enough for

$$\varepsilon \text{ Log } \varepsilon + 2(1-\varepsilon) \text{ Log } (1-\varepsilon) + \varepsilon \text{ Log } k > \frac{1}{50}, \qquad k\varepsilon > 1, \tag{7}$$

to hold which is always possible since $\varepsilon < 1$. Since $\sqrt{a} - \sqrt{b} = \sqrt{k-(k-1)b} - \sqrt{b}$, we may choose b small enough to get $\sqrt{a} - \sqrt{b} = \sqrt{k(1-\varepsilon)}$. Then, using (3), this shows that

$$h(P_I, P_J) > (1-\varepsilon) \Leftrightarrow \text{Card } (I \cap J) < m\varepsilon. \tag{8}$$

Now suppose a, b, k are fixed in that way and that S_m is a subset of maximal cardinality in $\mathscr{I}_m$ satisfying (4). Because of this maximality property the set $\{P_s\}_{s \in S_m}$ is a $(1-\varepsilon)$-net of $\mathscr{P}_m$, which means that

$$\begin{aligned} &\mathscr{P}_m \subset \bigcup_{s \in S_m} [\mathscr{B}(P_s, 1-\varepsilon) \cap \mathscr{P}_m], \\ &\text{Card } \mathscr{P}_m \leqslant \text{Card } S_m \times \sup_{s \in S_m} \text{Card } [\mathscr{B}(P_s, 1-\varepsilon) \cap \mathscr{P}_m]. \end{aligned} \tag{9}$$

Fix s in S_m. Going back to the previous notation we have $P_s = P_I$. The

number of different $J \subset L$ such that Card $(I \cap J) = j$ is exactly $\binom{m}{j}\binom{l-m}{m-j}$ so that, if we denote by p the smallest integer such that $p \geqslant \varepsilon m$,

$$\text{(10)} \qquad \text{Card } [\mathscr{B}(P_I, 1-\varepsilon) \cap \mathscr{P}_m] = \sum_{j=p}^{m} \binom{m}{j}\binom{l-m}{m-j} = \sum_{i=0}^{m-p} \binom{m}{i}\binom{l-m}{i}.$$

But from

$$\binom{m}{i}\binom{l-m}{i} = \binom{m}{i-1}\binom{l-m}{i-1} \times \frac{(m-i+1)(l-m-i+1)}{i^2}$$

and the fact that the ratio on the right is obviously decreasing with i, we can bound (10) by

$$\binom{m}{p}\binom{l-m}{m-p} \sum_{i=0}^{m-p} \left[\frac{(m-p)^2}{(p+1)(l-2m+p+1)}\right]^i$$

and, finally,

$$\text{(11)} \qquad \text{Card } [\mathscr{B}(P_I, 1-\varepsilon) \cap \mathscr{P}_m] \leqslant \binom{m}{p}\binom{l-m}{m-p}\left[1 - \frac{(1-\varepsilon)^2}{\varepsilon(k-2+\varepsilon)}\right]^{-1},$$

provided that $(1-\varepsilon)^2 < \varepsilon(k-2+\varepsilon)$ which is the case because of (7). Going back to (9) we find, as a lower bound for Card S_m,

$$\text{(12)} \qquad \text{Card } S_m \geqslant \frac{\binom{l}{m}}{\binom{m}{p}\binom{l-m}{m-p}} \frac{k\varepsilon - 1}{\varepsilon(k-2+\varepsilon)}.$$

Using Stirling approximation, many factors cancel and this, finally, becomes

$$\text{(13)} \qquad \text{Card } S_m \geqslant \frac{k\varepsilon - 1}{\varepsilon(k-2+\varepsilon)} \sqrt{\frac{2\pi}{m}} \frac{k^{mk}(1-\varepsilon)^{2m(1-\varepsilon)} \varepsilon^{m\varepsilon} (k-2+\varepsilon)^{m(k-2+\varepsilon)}}{(k-1)^{2m(k-1)}},$$

hence

$$\begin{aligned} \text{Log Card } S_m &\geqslant C - \tfrac{1}{2} \text{Log } m + m[k \text{ Log } k + 2(1-\varepsilon) \text{ Log } (1-\varepsilon) + \\ &\quad + \varepsilon \text{ Log } \varepsilon + (k-2+\varepsilon) \text{ Log } (k-2+\varepsilon) - 2(k-1) \text{ Log } (k-1)] \\ &> C - \tfrac{1}{2} \text{Log } m + m[\varepsilon \text{ Log } \varepsilon + 2(1-\varepsilon) \text{ Log } (1-\varepsilon) + \varepsilon \text{ Log } k]. \end{aligned}$$

Then from (7) we deduce that

$$\text{Log Card } S_m > C - \tfrac{1}{2} \text{Log } m + \frac{m}{50}$$

and (6) follows for large m. (5) is obvious from the definition of P_I, q.e.d.

Let us now consider on $\mathscr{P}_1(\Omega)$ the ball $\mathscr{B}(Q, r)$ of center Q and radius r with $0 < r \leqslant 1$. Choose some arbitrary infinite set $A \in \mathscr{A}$ and some $\varepsilon > 0$. Then from Proposition 1 we may find, for all large m, a set of probabilities $\{P_s\}_{s \in S_m}$ with support on A ($P_s(A) = 1$) satisfying (4), (5) and (6). If θ is such that $r = \sqrt{2} \sin \frac{1}{2}\theta$, $0 < \theta \leqslant \pi/2$, we define Q_s by

$$Q_s = Q \cos^2 \theta + P_s \sin^2 \theta, \tag{14}$$

getting a new set of probabilities $\{Q_s\}_{s \in S_m}$ of cardinality larger than $\exp [m/100]$ with the following properties:

PROPOSITION 2. *The probabilities Q_s, defined by* (14), *satisfy*:

$$h(Q, Q_s) \leqslant r \qquad \forall s \in S_m, \tag{15}$$

$$K(Q_s, Q_t) \leqslant C(\varepsilon) \qquad \forall\, s, t \in S_m, \tag{16}$$

C being a constant independent of r, m, Q and A.

Suppose $Q(A) = \lambda^2$ and M is a probability such that

$$h(M, Q_s) \leqslant r - \eta; \quad \eta > 0; \ \lambda \leqslant \frac{\eta}{2} \frac{r}{1-r^2} \text{ if } r < 1. \tag{17}$$

Then $M(A) > 0$ and the probability M_A, defined on A by

$$M_A(B) = \frac{M(A \cap B)}{M(A)},$$

satisfies

$$h(M_A, P_s) \leqslant \begin{cases} 1 - \dfrac{\eta}{2} \dfrac{1-r^2}{r(2-r^2)} & \text{if } r < 1, \\ 1 - \eta & \text{if } r = 1. \end{cases} \tag{18}$$

Proof. Since Log (dQ_s/dQ_t) is smaller than Log (dP_s/dP_t) if it is positive, we easily get from (5) that $K(Q_s, Q_t) \leqslant$ Log (a/b), a and b depending only on ε, which is (16). We also have $\varrho(Q, Q_s) \geqslant \cos \theta$ and (15) follows from the definition of θ. Let us now consider $\varrho(M, Q_s)$:

$$\begin{aligned} \varrho(M, Q_s) &= \int \sqrt{\cos^2 \theta \, dQ dM + \sin^2 \theta \, dP_s \, dM} \\ &\leqslant \cos \theta \Big[\int_{A^c} \sqrt{dQ dM} + \int_A \sqrt{dQ dM} \Big] + \sin \theta \int_A \sqrt{dP_s \, dM}. \end{aligned}$$

If $M(A) = 0$, $\varrho(M, Q_s) \leqslant \cos \theta$ and $h(M, Q_s) \geqslant r$, which is impossible by assumption. Then put $M(A) = \sin^2 \varphi$, $0 < \varphi \leqslant \pi/2$. We get

$$\varrho(M, Q_s) \leqslant \cos \theta [\cos \varphi + \lambda] + \sin \theta \sin \varphi \int_A \sqrt{dP_s \, dM_A}$$

or, taking $\varrho = \varrho(P_s, M_A)$,

$$\varrho(M, Q_s) \leqslant \cos\theta[\cos\varphi + \lambda] + \varrho \sin\theta \sin\varphi.$$

We may now maximize this expression with respect to φ and, with $\operatorname{tg}\varphi = \varrho \operatorname{tg}\theta$ if $r < 1$ ($\cos\theta > 0$), find easily

$$\varrho(M, Q_s) \leqslant \cos\theta[\lambda + \sqrt{1+\varrho^2 \operatorname{tg}^2\theta}]. \tag{19}$$

But by assumption $h(M, Q_s) \leqslant r - \eta$, which implies $\varrho(M, Q_s) \geqslant 1 - r^2 + \eta r$. Using (17), (19) and $\cos\theta = 1 - r^2$, we find

$$1 + \frac{\eta}{2}\frac{r}{1-r^2} \leqslant \sqrt{1+\varrho^2 \operatorname{tg}^2\theta} \quad \text{or} \quad \eta \frac{r}{1-r^2} \leqslant \varrho^2 \operatorname{tg}^2\theta,$$

which leads finally to

$$\varrho^2(P_s, M_A) \geqslant \eta \frac{1-r^2}{r(2-r^2)}.$$

We easily deduce (18) for the case $r < 1$. If $r = 1$, $\cos\theta = 0$ and we find $\varrho(M, Q_s) \leqslant \varrho(M_A, P_S)$, which concludes the proof.

3. Minimax risk for Hellinger balls and related sets. Before we prove the main result, we shall need a few technical lemmas. The first one is a version of the result known in information theory as Fano's lemma (see [1] or [5]).

LEMMA 1. *Suppose we are given $p+1$ probabilities $P_0, \ldots, P_p$ satisfying $K(P_i, P_j) \leqslant K$ for $0 \leqslant i, j \leqslant p$, then for any estimate $\hat{\psi}$ with values in the set $\{0; 1; \ldots; p\}$ the Bayes risk has the following lower bound:*

$$\frac{1}{p+1}\sum_{i=0}^{p} P_i[\hat{\psi} \neq i] \geqslant 1 - \frac{K + \operatorname{Log} 2}{\operatorname{Log} p}. \tag{20}$$

The second lemma is connected with the diameter of convex sets (Jung's theorem) and the fact that Hellinger distance is of Hilbertian type and that, with this distance, a finite set of probabilities may be considered as a set of points on a Euclidian sphere.

LEMMA 2. *Suppose we are given $p+1$ probabilities $P_0, \ldots, P_p$ such that $h(P_i, P_j) \geqslant \delta \ \forall\, i \neq j$. Then for any arbitrary probability P we have*

$$\sup_{i=0,\ldots,p} h^2(P, P_i) \geqslant 1 - \sqrt{1 - \frac{p}{p+1}\delta^2}. \tag{21}$$

As an easy consequence we find the following

LEMMA 3. *Suppose we are given two numbers r and δ with $0 < r < 1$; $r\sqrt{2-r^2} < \delta < 1$ and a set S of probabilities such that $h(Q_1, Q_2) \geqslant \delta$ for any Q_1, Q_2 in S. For any arbitrary probability P we have*

$$\text{Card}\,[S \cap \mathscr{B}(P, r)] \leqslant \frac{1}{\delta^2 - r^2(2-r^2)}. \tag{22}$$

The proof of Lemma 2 being purely geometrical and related to known results, will not be given here.

THEOREM 1. *Let us consider on some infinite space* $(\Omega, \mathscr{A})$ *the Hellinger ball* $\mathscr{Q} = \mathscr{B}(Q, r)$. *Then for any estimate* $\hat{Q}_n$ *depending on n i.i.d. observations, with values on the set* $\mathscr{P}_1(\Omega)$ *we have, for* $\eta > 0$,

$$\sup_{Q' \in \mathscr{Q}} Q'^n [h(\hat{Q}_n, Q') \geqslant r - \eta] = 1. \tag{23}$$

Proof. Suppose this is not true, then there exists η, $\eta' > 0$, and some estimate $\hat{Q}_n$ with

$$Q'^n [h(\hat{Q}_n, Q') \leqslant r - \eta] \geqslant \eta' > 0 \quad \forall\, Q' \in \mathscr{Q}. \tag{24}$$

Fix

$$\gamma = \begin{cases} \dfrac{\eta}{2} \dfrac{1-r^2}{r(2-r^2)} & \text{if } r < 1, \\ \eta & \text{if } r = 1. \end{cases}$$

Choose Γ, $\varepsilon > 0$ in such a way that

$$1 - \varepsilon > (1-\gamma)\sqrt{2-(1-\gamma)^2}, \quad \Gamma^{-1} = (1-\varepsilon)^2 - (1-\gamma)^2[2-(1-\gamma)^2]. \tag{25}$$

Now select an infinite subset A of Ω with

$$\sqrt{Q(A)} \leqslant \frac{\eta}{2} \frac{r}{1-r^2},$$

which is always possible, and, using A as shown in the preceding paragraph, consider the sets $\{Q_s\}_{s \in S_m} \subset \mathscr{B}(Q, r)$ which satisfy Proposition 2. From (24) we get

$$Q_s^n [h(\hat{Q}_n, Q_s) \leqslant r - \eta] \geqslant \eta' \quad \forall\, s \in S_m.$$

But $h(\hat{Q}_n, Q_s) \leqslant r - \eta$ implies that $\hat{Q}_{nA}$ is well defined and satisfies $h(\hat{Q}_{nA}, P_s) \leqslant 1 - \gamma$.

Now let us consider a new randomized estimate $\hat{P}_n$ with values in $\{Q_s\}_{s \in S_m}$ and defined as follows: if $\hat{Q}_n(A) = 0$, take for $\hat{P}_n$ anything you want; if $\hat{Q}_{nA}$ is well-defined, consider the uniform distribution on the finite (because S_m is finite) set $\{P_s\}_{s \in S_m} \cap \mathscr{B}(\hat{Q}_{nA}, 1-\gamma)$, choose P_s according to this distribution and take $\hat{P}_n = Q_s$. If Q_s^n is the true underlying distribution, we know that with probability at least η', P_s will belong to the ball $\mathscr{B}(\hat{Q}_{nA}, 1-\gamma)$. But the mutual distance between the P_s being at least $1-\varepsilon$, Lemma 3 proves that

$$\text{Card}\,[\{P_s\}_{s \in S_m} \cap \mathscr{B}(\hat{Q}_{nA}, 1-\gamma)] \leqslant \Gamma$$

and then we have a probability larger than $\eta' \Gamma^{-1}$ to find the correct value P_s.

Finally, we get

$$\inf_{s \in S_m} Q_s^n [\hat{P}_n = Q_s] \geqslant \eta' \Gamma^{-1}. \tag{26}$$

But using Propositions 1 and 2 we find that

$$K(Q_s^n, Q_t^n) \leqslant nC(\varepsilon) \quad \forall\, s, t \in S_m, \qquad \text{Card } S_m \geqslant \exp\left[\frac{m}{100}\right],$$

and applying Lemma 1 with m large enough we get

$$\frac{1}{\text{Card } S_m} \sum_{s \in S_m} Q_s^n [\hat{P}_n \neq Q_s] > 1 - \eta' \Gamma^{-1}$$

which, together with (26), gives the desired contradiction.

COROLLARY 1. *Suppose we have n i.i.d. observations of an unknown probability with continuous (or even $\mathscr{C}^\infty$) density with respect to Lebesgue measure on $[0, 1]$. Then for any estimate $\hat{P}_n$*

$$\sup_{P \in \mathscr{P}} P^n [h(\hat{P}_n, P) \geqslant 1 - \varepsilon] = 1 \qquad \forall\, \varepsilon > 0,$$

$\mathscr{P}$ being the set of all possible continuous (or $\mathscr{C}^\infty$) densities.

Proof. It is just a particular case of Theorem 1. The restriction that the densities must be continuous is not serious because in the construction of the family $\{P_I\}_{I \in \mathscr{I}_m}$ we may take the A_i to be contiguous intervals and then take a smooth version of (2) without changing Proposition 1.

COROLLARY 2. *For any increasing function g the minimax risk $R_n(\mathscr{B}(P_0, r), g)$ is $g(r)$ and one best possible estimate is $T_n(X_1, \ldots, X_n) = P_0$.*

The proof is obvious from Theorem 1, an analogous result holds with the assumptions of Corollary 1. This means that there is no nice estimate in such a case, the parameter space being too large. The only reasonable thing to do is to throw away the observations and take as an estimate the center of the ball. If you want to estimate a continuous density, take anything you like as an estimate, it does not make any difference as long as you want to compute the minimax risk.

REFERENCES

[1] L. Birgé, *Approximation dans les espaces métriques et théorie de l'estimation*, Z. f. Wahr. u. ver. Gebiete 65 (1983), p. 181-237.

[2] J. Bretagnolle et C. Huber, *Estimation des densités: risque minimax*, Z. f. Wahr. u. ver. Gebiete 47 (1979), p. 119-137.

[3] N. N. Čencov, *On correctness of the problem of statistical point estimation*, Theory of Prob. and its Applications 26 (1981), p. 13-29.

[4] H. G. Eggleston, *Convexity*, Cambridge University Press, 1958.
[5] I. A. Ibragimov and R. Z. Khas'minskii, *Statistical Estimation, Asymptotic Theory*, Springer-Verlag, New York 1981.
[6] — *On estimates of the density function*, Investigation in the mathematical statistics IV. Zap. Nauč. Semin. LOMI 98 (1980), p. 61-85 (in Russian).
[7] L. Le Cam, *On local and global properties in the theory of asymptotic normality of experiments*, Stochastic processes and related topics 1 (1975), p. 13-54.
[8] — *Asymptotic methods in statistical decision theory* (to appear).

U.E.R. de Sciences Economiques
Université Paris X — Nanterre
200, Avenue de la République
F-92001 Nanterre Cedex

Received on 15. 11. 1982