PREPRINT

SEPT 1971

5

Hans Brøns and Niels Keiding

Consistency of Maximum Likelihood Estimators in the Finite Case

UNIVERSITY OF COPENHAGEN
INSTITUTE OF
MATHEMATICAL STATISTICS

Hans Brøns and Niels Keiding

CONSISTENCY OF MAXIMUM LIKELIHOOD ESTIMATORS IN THE FINITE CASE

Preprint 1971 No. 5

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

September 1971

POLYTEKNISK FORLAG OG TRYKKERI, SØLVGADE 83, 1307 K, TLF. 11 98 22

Introduction.

This note contains an attempt to provide a systematic account of some basic facts concerning estimability of parameter functions and consistency of the maximum likelihood estimators for statistical models on finite sample spaces.

An exposition of the existing theory is found in Rao [1965] .

1. Probabilities on a finite set.

Definition 1.1

Let Ω be a finite set.

$$P(\Omega) = \{(x_\omega)_{\omega\in\Omega} \in \mathbb{R}^\Omega \mid x_\omega \geqq 0 \quad \forall\omega, \quad \Sigma x_\omega = 1\}$$

is the set of probabilities on Ω. A $P \in \mathcal{P}(\Omega)$ is thus given by its values $P(\omega)$, $\omega \in \Omega$.

The relative Euclidean topology on $\mathcal{P}(\Omega)$ is called the weak topology. The strong topology on $[0,1]$ is defined by letting 0 and 1 be isolated points whereas the relative topology on $]0,1[$ is the usual Euclidean topology. The strong topology on $\mathcal{P}(\Omega)$ is defined as the relative topology of the strong product topology.

Remark 1.1 $P_n \to P$ in the strong topology if and only if $P_n \to P$ in the weak topology and there exists a number N such that $P_n \ll P$ (P_n is absolutely continuous with respect to P) for $n > N$.

Remark 1.2 $\mathcal{P}(\Omega)$ is compact in the weak topology, but not compact in the strong topology.

Definition 1.2 If $A \subseteqq \Omega$, $f: A \to \mathbb{R}$ is a real function, the integral

$$\int f dP = \sum_{\omega \in A} f(\omega)P(\omega)$$

is defined for $P \in \mathcal{P}_A = \{P \in \mathcal{P}(\Omega) \mid PA = 1\}$. (Geometrically speaking, P_A is a surface of $P(\Omega)$.)

Remark 1.3 $P_n \to P$ in the weak topology if and only if $\int f dP_n \to \int f dP$ for all functions $f: \Omega \to \mathbb{R}$.

(If Ω is considered in the discrete topology all f are bounded and continuous so that the weak topology corresponds to ordinary weak convergence of probabilities).

Definition 1.3 Let $\acute{R}^* = \acute{R} \cup \{-\infty\} \cup \{\infty\}$ be the extended real axis. If $f:A \to \acute{R}^*$ is an extended real function and $P \in \acute{P}_A \cap \acute{S}_f$, where $\acute{S}_f = \{P \in \acute{P}(\Omega) | P\{f = -\infty\} \cdot P\{f = \infty\} = 0\}$, the integral

$$\int f dP = \sum_{\omega \in A} f(\omega)P(\omega)$$

is well-defined.

2. The information function.

Definition 2.1

Consider two probabilities $P \in \acute{P}(\Omega)$ and $Q \in \acute{P}(\Omega)$. The derivative dQ/dP is given by $Q(\omega)/P(\omega)$ when $\omega \in A_P = \{\omega | P(\omega) > 0\}$.

Assume $\log 0 = -\infty$. The function

$$-\log \frac{dQ}{dP} : A_P \to R^*$$

is then defined by

$$\omega \mapsto -\log \frac{Q(\omega)}{P(\omega)} \quad .$$

Since $P(A_P) = 1$ and $P\{-\log \frac{dQ}{dP} > -\infty\} = 1$ for all $Q \in P(\omega)$, the integral

$$I(P,Q) = \int -\log \frac{dQ}{dP} \, dP = \sum_{\omega \in A_P} (-\log \frac{Q(\omega)}{P(\omega)}) P(\omega)$$

is well-defined (see Definition 1.3). The information function I is thus well-defined on $\acute{P}(\Omega) \times \acute{P}(\Omega)$.

Properties of the information function are given in the following propositions.

Proposition 2.1 $I \geqq 0$. $I(P,Q) = 0 \iff P = Q$.

Proof. The function $-\log : [0, \infty[\to \acute{R} \cup \{\infty\}$ is convex. Hence by Jensen's inequality

$$I(P,Q) = \sum_{P(\omega)>0} -P(\omega)\log \frac{Q(\omega)}{P(\omega)}$$

$$\geqq -\log \sum_{P(\omega)>0} P(\omega)\,\frac{Q(\omega)}{P(\omega)} = -\log \sum_{P(\omega)>0} Q(\omega) \geqq 0,$$

since $\sum_{P(\omega)>0} Q(\omega) \leqq 1$. The last inequality becomes an equality if and only if $Q \ll P$.

Under this assumption the first inequality is an equality if and only if

$$\log \frac{Q(\omega)}{P(\omega)} = c, \text{ i.e. } Q(\omega) = kP(\omega)$$

when $P(\omega) > 0$. Since $Q \ll P$, k must be 1, so that $Q = P$.

<u>Proposition 2.2</u> $I(P,Q) < \infty \Leftrightarrow P \ll Q$.

<u>Proof</u>. $I(P,Q) = \infty$ if and only if there exists an ω with

$$-\log \frac{Q(\omega)}{P(\omega)} = \infty \quad \text{and } P(\omega) > 0.$$

This is equivalent to saying that $Q(\omega) = 0$ and $P(\omega) > 0$.

<u>Proposition 2.3</u> I is continuous in the topology on $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega)$ given by the product topology of the strong topology on $\mathcal{P}(\Omega)$ and the weak topology on $\mathcal{P}(\Omega)$, respectively.

<u>Proof.</u> It is possible to choose a strong neighbourhood U of P such that $P' \equiv P$ (<u>i.e.</u> $P' \ll P$ and $P \ll P'$) for $P' \in U$ and a weak neighbourhood V of Q such that $Q \ll Q'$ for $Q' \in V$. If $I(P,Q) < \infty \Leftrightarrow P \ll Q$ we may then infer that $P' \ll Q' \Leftrightarrow I(P',Q') < \infty$ for $(P',Q') \in U \times V$ which is a neighbourhood of (P,Q). The continuity at (P,Q) is then obvious from the definition of I. If $I(P,Q) = \infty$, the set $A = \{\omega | P(\omega) > 0 \text{ and } Q(\omega) = 0\} \neq \emptyset$. As $(P',Q') \to (P,Q)$, $Q'(\omega) \to 0$ for $\omega \in A$. Hence since $\{\omega | P(\omega) > 0\} = \{\omega | P'(\omega) > 0\}$,

$$I(P',Q') = -\sum_{P'(\omega)>0} P'(\omega) \log \frac{Q'(\omega)}{P'(\omega)} \to \infty = I(P,Q).$$

The following result is well-known.

Lemma 2.1 If $f\colon \acute{X} \times \acute{Y} \to \acute{R}$ is continuous and $K \subseteq \acute{Y}$ is compact then

$$f_K\colon \acute{X} \to \acute{R}$$

given by $f_K(x) = \inf_{y \in K} f(x,y)$ is continuous.

Proposition 2.4 For all $\Pi \subseteq P(\Omega)$, $\inf_{\pi \in \Pi} I(\cdot,\pi)$ is continuous in the strong topology.

Proof. The weak closure $\overline{\Pi}$ of Π is compact since $P(\Omega)$ is compact in the weak topology. As then

$$\inf_{\pi \in \Pi} I(P,\pi) = \inf_{\pi \in \overline{\Pi}} I(P,\pi)$$

the result is a corollary of Proposition 2.3 and Lemma 2.1.

Proposition 2.5 (Birch[1965])

$$\sum_{\omega \in \Omega} (Q(\omega) - P(\omega))^2 \leq 2I(P,Q).$$

Proof. If P is not absolutely continuous with respect to Q, $I(P,Q) = \infty$ and the inequality is trivially satisfied.

Assume then $P \ll Q$. Taylor expansion of x log x around 1 yields

$$x \log x = x-1+ \frac{1}{y}\,\frac{(x-1)^2}{2} \;,$$

where y is between 1 and x.

If $P(\omega) > 0$ and hence $Q(\omega) > 0$, put $x = \frac{P(\omega)}{Q(\omega)}$ to get

$$-P(\omega) \log \frac{Q(\omega)}{P(\omega)} = P(\omega)-Q(\omega) + \frac{1}{2Q(\omega)y}\,(P(\omega)-Q(\omega))^2$$

$$\geq P(\omega)-Q(\omega) + \tfrac{1}{2}(P(\omega)-Q(\omega))^2$$

since $Q(\omega)y$ is between 1 and $P(\omega)$, i.e. ≤ 1.

By summation we obtain

$$I(P,Q) \geqq 1 - \sum_{P(\omega)>0} Q(\omega) + \frac{1}{2} \sum_{P(\omega)>0} (P(\omega)- Q(\omega))^2$$

$$= \frac{1}{2} \sum_{\omega \in \Omega} (P(\omega)- Q(\omega))^2 + \sum_{P(\omega)=0} [Q(\omega)- \frac{1}{2}(Q(\omega))^2]$$

$$\geqq \frac{1}{2} \sum_{\omega \in \Omega} (P(\omega) - Q(\omega))^2.$$

3. Continuous maximum likelihood estimation.

Definition 3.1 *A statistical problem* (Ω,Θ,ψ) is given by a *parameter set* Θ and a mapping $\psi\colon \Theta \to \mathcal{P}(\Omega)$.

Let $\mathcal{P}_\psi = \{P \mid \exists\ \theta\colon P << \psi(\theta)\}$, let Θ' be a topological space and $\phi\colon \Theta \to \Theta'$ a mapping. Define the function $\alpha\colon$ $P(\Omega) \times \Theta' \to [0,\infty[$ by

$$\alpha(P,\theta') = \inf_{\phi(\theta)=\theta'} I(P,\psi(\theta)).$$

Note that

$$\alpha(P,\theta') < \infty \Rightarrow (P,\theta') \in \mathcal{P}_\psi \times \phi(\Theta).$$

Definition 3.2 ϕ is said to be *continuously estimable* and $f\colon \mathcal{P}_\psi \to \Theta'$ a *continuous maximum likelihood estimator* of ϕ if

1) $\forall P \in \mathcal{P}_\psi \quad \forall \theta' \in \Theta'\colon \ \alpha(P,\theta') \geqq \alpha(P,f(P))$.

2) $\forall\ \theta \in \Theta\colon f(\psi(\theta)) = \phi(\theta)$

3) f is continuous in the strong topology on $\mathcal{P}_\psi$ at each $P \in \psi(\Theta)$.

Remark 3.1 From the remark above about α it follows that $f(\mathcal{P}_\psi) \subseteqq \phi(\Theta)$. But from 2), $f(\mathcal{P}_\psi) \supseteqq \phi(\Theta)$. It is thus seen that $f(\mathcal{P}_\psi) = \phi(\Theta)$.

Remark 3.2 Since by Proposition 2.1 $I(P,Q) = 0 <=> P = Q$ it follows that for all $\theta \in \Theta$

$$\alpha(\psi(\theta),\phi(\theta)) = \inf_{\bar\theta:\phi(\bar\theta)=\phi(\theta)} I(\psi(\theta),\psi(\bar\theta)) = 0$$

and that the infimum is attained if and only if $\psi(\bar{\theta}) = \psi(\theta)$. From 1) we have

$$\alpha(\psi(\theta), f(\psi(\theta))) = \inf_{\theta' \in \Theta'} \alpha(\psi(\theta), \theta') = 0$$

since this value is attained if and only if $\theta' = \phi(\bar{\theta})$ with $\psi(\bar{\theta}) = \psi(\theta)$. If ψ is injective it follows that $\bar{\theta} = \theta$ and hence $\phi(\theta) = f(\psi(\theta))$. It is concluded that if ψ is injective, 1) implies 2).

Remark 3.3 Let the mapping $\gamma \colon \Theta \to \psi(\Theta)$ be defined by $\gamma(\theta) = \psi(\theta), \theta \in \Theta$. If ϕ is continuously estimable, then there exists one and only one mapping $\delta \colon \psi(\Theta) \to \Theta'$, continuous in the relative strong topology on $\psi(\Theta)$, so that $\phi = \delta \circ \gamma$. (Define δ as the restriction of an arbitrary continuous maximum likelihood estimator f.)

Theorem 3.1 Consider $\mathcal{P}(\Omega)$ in the weak topology. $\psi \colon \Theta \to \mathcal{P}(\Omega)$ is continuously estimable if and only if $\psi(\Theta)$ is compact.

Proof. Assume first that $\psi(\Theta)$ is weakly compact. It is seen that for $Q \in \psi(\Theta)$

$$\alpha(P,Q) = \inf_{\psi(\theta)=Q} I(P,\psi(\theta)) = I(P,Q).$$

Since $I(P,\cdot)$ is continuous in the weak topology by Proposition 2.3, $\inf_{Q \in \psi(\Theta)} I(P,Q)$ is attained by some $\psi(\theta)$ so that the set

$$F(P) = \{\psi(\theta) \mid I(P,\psi(\theta)) = \inf_{\bar{\theta} \in \Theta} I(P,\psi(\bar{\theta}))\} \neq \emptyset.$$

Define f: $P_\psi \to P(\Omega)$ by choosing for f(P) some value in F(P). Then for all $P \in \mathcal{P}_\psi$ and all $Q \in \psi(\Theta)$

$$\alpha(P,Q) = I(P,Q) \geqq \inf_{\theta \in \Theta} I(P,\psi(\theta)) = I(P,f(P)) = \alpha(P,f(P))$$

and for $Q \notin \psi(\Theta)$, $\alpha(P,Q) = \infty$ as remarked above.

POLYTEKNISK FORLAG OG TRYKKERI SØLVGADE 83, 1307 K. TLF. 11 93 22

Hence $\alpha(P,Q) \geqq \alpha(P,f(P))$ for all $P \in \mathcal{P}_\psi$ and all $Q \in P(\Omega)$ and 1) in Definition 3.2 is verified.

To show 2), note that for all $\theta \in \Theta$

$$F(\psi(\theta)) = \{\psi(\theta^*) \mid I(\psi(\theta),\psi(\theta^*)) = \inf_{\bar{\theta} \in \Theta} I(\psi(\theta),\psi(\bar{\theta}))\} = \{\psi(\theta)\}$$

since $I(P,Q) = 0 \iff P = Q$.

Hence $f(\psi(\theta)) = \psi(\theta)$ for all $\theta \in \Theta$.

Finally it is to be shown that $f\colon \mathcal{P}_\psi \to \mathcal{P}(\Omega)$ considered in the <u>strong</u> topology on $\mathcal{P}_\psi$ and the <u>weak</u> topology on $\mathcal{P}(\Omega)$ is continuous. Let then $P_n \to \psi(\theta_0)$ strongly.

From Proposition 2.4 it follows that the function $\inf_{Q \in \psi(\Theta)} I(\cdot,Q)$ is strongly continuous which implies

$$I(P_n,f(P_n)) = \inf_{\theta \in \Theta} I(P_n,\psi(\theta)) \to \inf_{\theta \in \Theta} I(\psi(\theta_0),\psi(\theta)) = 0$$

and according to Proposition 2.5

$$\sum_{\omega \in \Omega} (P_n(\omega) - f(P_n)(\omega))^2 \leqq \frac{1}{2} I(P_n,f(P_n)) \to 0.$$

Since $P_n \to \psi(\theta_0)$ in the strong topology, and hence in the weak topology, it follows that in the weak topology

$$f(P_n) \to \psi(\theta_0) = f(\psi(\theta_0))$$

as noted above. This completes the proof of the "if"-part of the theorem.

Suppose, conversely, that $P_0 \in \overline{\psi(\Theta)}$, the weak closure of $\psi(\Theta)$. We shall show that $P_0 \in \psi(\Theta)$, <u>i.e.</u> that $\psi(\Theta)$ is closed. Since $\mathcal{P}(\Omega)$ is compact in the weak topology it will follow that $\psi(\Theta)$ is compact.

If $Q \in \psi(\Theta)$, $\alpha(P_0,Q) = I(P_0,Q)$ as noted above, and we therefore get

$$\alpha(P_0,f(P_0)) = \inf_{Q \in P(\Omega)} \alpha(P_0,Q) = \inf_{Q \in \psi(\Theta)} \alpha(P_0,Q)$$

$$= \inf_{Q \in \psi(\Theta)} I(P_0,Q) = \underline{\inf_{Q \in \psi(\Theta)}} I(P_0,Q) = 0$$

the infimum being attained if and only if $Q = P_0$. Hence $\alpha(P_0,f(P_0)) = I(P_0,f(P_0)) = 0$ implying that $P_0 = f(P_0) \in \psi(\Theta)$.

Remark 3.4 To a given mapping $\phi: \Theta \to \Theta'$ we define the mapping $\nu: \Theta \to \phi(\Theta)$ by $\nu(\theta) = \phi(\theta)$, $\theta \in \Theta$. It is obvious that ν is continuously estimable if and only if ϕ is continuously estimable.

Theorem 3.2 Let $\phi:\Theta \to \Theta'$ be continuously estimable and f a continuous maximum likelihood estimator of ϕ . Let $\lambda: \Theta' \to \Theta''$ be continuous. Then $\lambda \circ f$ is a continuous maximum likelihood estimator of $\lambda \circ \phi$.

Proof. Fix an arbitrary $P \in \acute{P}_\psi$ and define $\beta(P,\theta'') = \inf_{\lambda\circ\phi(\theta)=\theta''} I(P,\psi(\theta))$.

It is seen that

$$\beta(P,\theta'') = \inf_{\substack{\lambda(\theta')=\theta'' \\ \theta'\in\phi(\Theta)}} \left(\inf_{\phi(\theta)=\theta'} I(P,\psi(\theta)) \right) = \inf_{\lambda(\theta')=\theta''} \alpha(P,\theta')$$

since $\alpha(P,\theta') = \infty$ for $\theta' \notin \phi(\Theta)$.

Hence $\beta(P,\theta'') \geqq \inf_{\theta'\in\Theta'} \alpha(P,\theta') = \alpha(P,f(P))$ and conversely

$$\beta(P, \lambda \circ f(P)) = \inf_{\lambda(\theta')=\lambda\circ f(P)} \alpha(P,\theta') \leqq \alpha(P,f(P))$$

so that for all $\theta'' \in \Theta''$

$$\beta(P,\theta'') \geqq \beta(P, \lambda \circ f(P))$$

showing that $\lambda \circ f$ meets 1) of Definition 3.2. 2) and 3) are trivially satisfied.

Corollary. Let $\delta: \psi(\Theta) \to \Theta'$ be continuous in the relative weak topology on $\psi(\Theta)$. If $\psi(\Theta)$ is compact then $\delta \circ \gamma$ is continuously estimable.

Proof. This is obvious from Theorems 3.1 and 3.2 and Remark 3.4.

4. Independent identically distributed observations.

In the present section the theory of section 3 will be applied to the simple statistical situation where the observations consist of independent identically distributed replications. An observation is then a point $\eta^{(n)} = (\eta_1,\dots,\eta_n) \in \Omega^n$ for some $n = 1,2,\dots$.

The product probabilities P^n on Ω^n and $P^{\mathbb{N}}$ on $\Omega^{\mathbb{N}}$ are defined in the usual way. A statement valid with $P^n(P^{\mathbb{N}})$-probability one is said to be true a.s. [P].

Definition 4.1 Let $n \in \mathbb{N}$ and let $N_\omega(\eta^{(n)}) = \#\{i = 1,\dots,n \mid \eta_i = \omega\}$. The mapping $E_n : \Omega^n \to P(\Omega)$ given by $E_n(\eta^{(n)})(\omega) = N_\omega(\eta^{(n)})/n$ is called the empirical distribution (of order n).

Proposition 4.1 For all $P \in \mathcal{P}(\Omega)$ and $n \in \mathbb{N}$, $E_n \ll P$ a.s. [P] .

Proof. The statement is written out as

$$P^n\{\eta^{(n)} \mid \exists\ \omega \in \Omega:\ E_n(\eta^{(n)})(\omega) > 0 \text{ and } P(\omega) = 0\} = 0.$$

This is, however, obvious from the fact that $P(\omega) = 0 \Rightarrow P^n\{N_\omega(\eta^{(n)}) > 0\} = 0$.

Proposition 4.2 For all $P \in \mathcal{P}(\Omega)$ $E_n \to P$ a.s. [P] as $n \to \infty$ in the strong topology on $\mathcal{P}(\Omega)$.

Proof. E_n is canonically thought of as a mapping on $\Omega^{\mathbb{N}}$ depending only on the first n coordinates. By the strong law of large numbers, $P^{\mathbb{N}}\ \{\eta \in \Omega^{\mathbb{N}} \mid E_n(\eta)(\omega) \to P(\omega)\} = 1$, where the convergence is in the ordinary Euclidean topology on $[0,1]$. By Proposition 4.1 $P(\omega) = 0 \Rightarrow P^{\mathbb{N}}\{E_n(\eta)(\omega)=0\} = 1$ so that the convergence holds in fact in the strong topology on $[0,1]$ (see Remark 1.1). The result then follows from the definition of the strong topology as the relative strong product topology on $[0,1]^\Omega$.

Definition 4.2 Assume that a statistical problem (Ω,Θ,ψ) is given. Consider a mapping $\phi: \Theta \to \Theta'$ and a sequence $F_n = f \circ E_n : \Omega^n \to \Theta'$, where $f: \mathcal{P}(\Omega) \to \Theta'$ is a mapping . Let $\bar{f}: \mathcal{P}_\psi \to \Theta'$ be the restriction of f. If $\bar{f}$ is a continuous maximum likelihood estimator in the sense of Definition 3.2, then ϕ is called empirically estimable and F_n an empirical maximum likelihood estimator.

Remark 4.1 Proposition 4.1 implies that $E_n \in P_\psi$ a.s. $[\psi(\theta)]$ for all $\theta \in \Theta$.

Remark 4.2 From the properties of f noted in section 3 we immediately deduce that $F_n(\Omega^n) \subseteq \phi(\Theta)$ a.s. $[\psi(\theta)]$ for all $\theta \in \Theta$ and $n \in N$.

Theorem 4.1 $F_n \to \phi(\theta)$ a.s. $[\psi(\theta)]$ as $n \to \infty, \theta \in \Theta$, i.e. F_n is consistent.

Proof. From Proposition 4.2 it follows that $E_n \to \psi(\theta)$ a.s. $[\psi(\theta)]$ in the strong topology as $n \to \infty$. Since f is continuous in the strong topology at $\psi(\theta)$, it follows that

$$F_n = f(E_n) \to f(\psi(\theta)) = \phi(\theta) \quad \text{a.s.}[\psi(\theta)]$$

as $n \to \infty$.

The following properties of empirical maximum likelihood estimators are easily deduced from the results of section 3.

Theorem 4.2 If ϕ is empirically estimable then there exists one and only one mapping $\delta: \psi(\Theta) \to \Theta'$, continuous in the relative strong topology on $\psi(\Theta)$, so that $\phi = \delta \circ \gamma$.

If $P(\Omega)$ is considered in the weak topology $\psi: \Theta \to P(\Omega)$ is empirically estimable if and only if $\psi(\Theta)$ is compact.

If ϕ is empirically estimable, if $\lambda: \Theta' \to \Theta''$ is continuous, and if F_n is an empirical maximum likelihood estimator of ϕ, then $\lambda \circ F_n$ is an empirical maximum likelihood estimator of $\lambda \circ \phi$.

References.

M.W. Birch [1964] : A new proof of the Pearson-Fisher theorem. Ann. Math. Statist. 35, 817-824. Acknowledgement of Priority: Ann. Math. Statist. 36 (1965), 344.

C.R. Rao [1965] : Linear statistical inference and its applications. Wiley, New York.