Z. Wahrscheinlichkeitstheorie 4, 69—73 (1965)

Extension to Markov Processes of a Result by A. Wald about the Consistency of the Maximum Likelihood Estimate

By

GEORGE G. ROUSSAS*

Summary. In this note the proof of the consistency of a maximum likelihood estimate (MLE) obtained by WALD in [7] in the case of independent and identically distributed random variables is extended to the case of Markov processes.

There is an extensive literature about the existence of a MLE and its consistency, most of which includes the assumption of the existence of derivatives of the densities with respect to the parameter involved. (See, for example, [2] and other references cited there.) Even under the rather strong assumption of pointwise differentiability of densities, and other additional regularity conditions, the problem of existence and consistency of a MLE has not been solved satisfactorily. (See, for example, [1], [2], [4], [6].) On the other hand, there have appeared papers like [3], where the consistency of a MLE is proved for processes with dependent random variables, and without the usual differentiability assumptions. The conditions used in the present paper are, however, of a different nature from those imposed in [3], and also are slightly different from WALD's assumptions in [7]. To our knowledge, a proof of consistency of a MLE under conditions similar to the ones used here has not appeared in the literature.

I would like to take this opportunity to thank Professor L. LECAM for a number of remarks in connection with this paper.

1. Introduction, notation and assumptions

We consider a measurable space $(\mathcal{X}, \mathcal{A})$ and for each $\theta \in \Theta$ let P_θ be a probability measure on $\mathcal{A}$. It is assumed that for every $\theta \in \Theta \{X_n, n \geqq 0\}$ is a Markov process defined on the probability space $(\mathcal{X}, \mathcal{A}, P_\theta)$. In fact, without loss of generality, we may assume that $(\mathcal{X}, \mathcal{A})$ is the infinite Cartesian product $\prod_{i=0}^{\infty} (R, \mathcal{B})$ of the Borel real line $(R, \mathcal{B})$, and P_θ is the probability measure induced in $\mathcal{A}$ by a set of transition probabilities $p_\theta(\cdot, \cdot)$ defined on $R \times \mathcal{B}$, and a probability distribution $p_\theta(\cdot)$ on $\mathcal{B}$, according to Kolmogorov's Consistency Theorem. In such a case the process $\{X_n, n \geqq 0\}$ will be taken to be the coordinate process, and then it will be a Markov process with initial distribution $p_\theta(\cdot)$ and (stationary) transition probabilities $p_\theta(\cdot, \cdot)$.

We denote by $P_{n,\theta}$ the restriction of P_θ to the σ-field $\mathcal{A}_n = \mathcal{B}(X_0, X_1, \ldots, X_n)$, $n \geqq 0$, and we will assume in the following that the probability measures $\{P_{n,\theta}, \theta \in \Theta\}$, $n \geqq 0$ are absolutely continuous with respect to one another. Therefore, for any $\theta, \theta' \in \Theta$ we will have $[dP_{0,\theta'}/dP_{0,\theta}] = q(X_0; \theta, \theta')$, $[dP_{1,\theta'}/dP_{1,\theta}] = q(X_0, X_1; \theta, \theta')$, and if we set $q(X_1 | X_0; \theta, \theta') = [q(X_0, X_1; \theta, \theta'/q(X_0; \theta, \theta')]$ we will then have for the joint measures $P_{n,\theta'}$, $P_{n,\theta}$: $[dP_{n,\theta'}/dP_{n,\theta}] = q(X_0; \theta, \theta') \prod_{j=1}^{n} q(X_j | X_{j-1}; \theta, \theta')$. Clearly, $[dP_{n,\theta'}/dP_{n,\theta}]$ is well defined except possibly on P_θ-null sets for all $\theta \in \Theta$. In what follows, we will always work outside

* Prepared with the partial support of the National Science Foundation, Grant GP—10.

of these null sets, although we will not always point it out explicitly. Actually, we will fix an arbitrary $\theta^* \in \Theta$, and the various likelihoods will be formed with respect to P_{n,θ^*}. In doing so we will find it convenient to write $f(X_0; \theta)$ and $f(X_0, X_1; \theta)$ instead of $q(X_0; \theta^*, \theta)$ and $q(X_1 | X_0; \theta^*, \theta)$, respectively.

The following set of assumptions will be used in various places in this paper.

Assumptions. (A1) *Θ is a compact metric space with metric d.*

(A2) *For each $\theta \in \Theta$ the Markov process $\{X_n, n \geqq 0\}$ is stationary and metrically transitive (ergodic).*

(A3) *The probability measures $\{P_{n,\theta}, \theta \in \Theta\}$, $n \geqq 0$, are mutually absolutely continuous.*

(A4) *For each $x \in R$ $\sup\{f(x; \theta); \theta \in \Theta\}$ is finite.*

(A5) *For any $x, y \in R$ $f(x, y; \theta)$ is upper semi-continuous* (u.s.c) *in θ.*

(A6) *For each $\theta \in \Theta$ there is a neighborhood $W(\theta)$ of θ such that for any open set V with $\theta \in V \subset W(\theta)$ the $\sup\{f(x, y; \theta); \theta \in V\}$ is $\mathcal{B} \times \mathcal{B}$-measurable.*

(A7) *Let $h(x, y; t, \theta) = \log[f(x, y; \theta)/f(x, y; t)]$. Then for each $\theta \in \Theta$ there exists an open set $W^*(\theta)$ such that $\theta \in W^*(\theta) \subset W(\theta)$, with the property that $|\int h(x, y; W^*(\theta)) dP_{1,\theta}| < \infty$, where $h(x, y; W^*(\theta)) = \inf\{h(x, y; t, \theta); t \in W^*(\theta)\}$.*

(A8) *For any $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$ implies $\int |f(x, y; \theta_1) - f(x, y; \theta_2)| . dP_{1,\theta} > 0$.*

2. Main result

In formulating the main result of this paper the concepts introduced below will be needed.

Definitions. Any mapping $\hat{\theta}_n = \hat{\theta}_n(X_0, X_1, \ldots, X_n)$ on $\mathcal{X}$ into Θ which is $\mathcal{A}_n$-measurable is called an *estimate*. An estimate $\hat{\theta}_n$ such that $\sup\{f(X_0; \theta) \prod_{j=1}^{n} f(X_{j-1}, X_j; \theta); \theta \in \Theta\} c \leqq f(X_0; \hat{\theta}_n) \prod_{j=1}^{n} f(X_{j-1}, X_j; \hat{\theta}_n)$ is called a *quasi-maximum likelihood estimate* (q-MLE) *with coefficient* $c \in (0, 1]$. For $c = 1$ we get a MLE in the usual sense. The estimates $\{\hat{\theta}_n\}$, $n > 0$ are *consistent* at $\theta \in \Theta$ if $\hat{\theta}_n \to \theta$ in P_θ-probability, as $n \to \infty$, and they are *strongly consistent* if $\hat{\theta}_n \to \theta$ a. s. $[P_\theta]$, as $n \to \infty$.

Remark. It is customary to define an estimate $\hat{\theta}_n$ the way we did above, i. e., as an $\mathcal{A}_n$-measurable map of $\mathcal{X}$ into Θ; and in this paper we use this definition of an estimate. Sometimes, however, as in the case of a q-MLE with coefficient $c < 1$, there may be nonmeasurable maps of $\mathcal{X}$ into Θ. In such cases the probability measure P_θ is to be replaced by the inner probability measure $P_{*,\theta}$ in proving consistency. Such a proof would be rather long and uninteresting, and we choose not to present it here.

Theorem. *Under assumptions* (A1) *to* (A8), *quasi-maximum likelihood estimates $\{\hat{\theta}_n\}$ with coefficient $c \in (0, 1]$ are strongly consistent at $\theta \in \Theta$*, i. e.,

$$\hat{\theta}_n \to \theta \text{ a. s. } [P_\theta], \quad \text{as} \quad n \to \infty, \ \theta \in \Theta \tag{3.1}$$

The proof of this theorem will follow after we formulate and prove three lemmas.

We denote by θ_0 the (unknown) true value of the parameter, and let θ and t vary over Θ. By (A7) we have $h(X_0, X_1; \theta, \theta_0) = \log[f(X_0, X_1; \theta_0)/f(X_0, X_1; \theta)]$. Then the following result is easily established, namely.

Lemma 1. *As* $n \to \infty$, $\limsup 1/n \sum_{j=1}^{n} h(X_{j-1}, X_j; \hat{\theta}_n, \theta_0) \leqq 0$.

Proof. We have $[f(X_0; \theta)\prod_{j=1}^{n} f(X_{j-1}, X_j; \theta)]c \leqq f(X_0; \hat{\theta}_n) \cdot \prod_{j=1}^{n} f(X_{j-1}, X_j; \hat{\theta}_n)$ for all $\theta \in \Theta$. Dividing both sides by $f(X_0; \theta_0)\prod_{j=1}^{n} f(X_{j-1}, X_j; \theta_0)$, and taking logarithms, we get $\log[f(X_0; \theta)/f(X_0; \theta_0)] - \sum_{j=1}^{n} h(X_{j-1}, X_j; \theta, \theta_0) + \log c \leqq$ $\leqq \log[f(X_0; \hat{\theta}_n)/f(X_0; \theta_0)] - \sum_{j=1}^{n} h(X_{j-1}, X_j; \hat{\theta}_n, \theta_0)$ for all $\theta \in \Theta$. This will be true, in particular, for $\theta = \theta_0$, i. e., $\sum_{j=1}^{n} h(X_{j-1}, X_j; \hat{\theta}_n, \theta_0) - \log[f(X_0; \hat{\theta}_n)/f(X_0; \theta_0)] \leqq -\log c$. Dividing throughout by n and letting $n \to \infty$ we get, by means of (A4), $\limsup 1/n \sum_{j=1}^{n} h(X_{j-1}, X_j; \hat{\theta}_n, \theta_0) \leqq 0$, as was to be seen.

Now we define $H(\theta)$ by $H(\theta) = \mathscr{E}_{\theta_0}[h(X_0, X_1; \theta, \theta_0)] = \int \log[f(x, y; \theta_0)/f(x, y; \theta)]\, f(x, y; \theta_0) f(x; \theta_0) dP_{1,\theta^*}$, and then the following lemma is true;

Lemma 2. *With the above notation,* $H(\theta) \geqq 0$ *and* $H(\theta) = 0$ *if and only if* $f(x, y; \theta) = f(x, y; \theta_0)$ a. s. $[P_{1,\theta_0}]$.

Proof. We use the inequality $\exp(z) \geqq 1 + z$, $z \in R$, and $\exp(z) = 1 + z$ if and only if $z = 0$.

Replacing z by $\log[f(x, y; \theta)/f(x, y; \theta_0)]$ we get $[f(x, y; \theta)/f(x, y; \theta_0)] \geqq 1 + \log[f(x, y; \theta)/f(x, y; \theta_0)]$, and equality holds if and only if $f(x, y; \theta) = f(x, y; \theta_0)$. Then

$$[f(x, y; \theta)/f(x, y; \theta_0)] - 1 - \log[f(x, y; \theta)/f(x, y; \theta_0)] \geqq 0, \tag{3.2}$$

with equality holding if and only if $f(x, y; \theta) = f(x, y; \theta_0)$. But $\int [f(x, y; \theta)/f(x, y; \theta_0)]\, dP_{1,\theta_0} = \mathscr{E}_{\theta_0}[f(X_0, X_1; \theta)/f(X_0, X_1; \theta_0)] = \mathscr{E}_{\theta_0}\{\mathscr{E}_{\theta_0}[f(X_0, X_1; \theta)/f(X_0, X_1; \theta_0)] \mid X_0\} = \int [f(x, y; \theta) dP_{0,\theta^*}] \cdot dP_{0,\theta_0} = 1$. Therefore, integrating both sides of (3.2) with respect to P_{1,θ_0} we get the required inequality.

An application of the Ergodic Theorem gives

$$1/n \sum_{j=1}^{n} h(X_{j-1}, X_j; \theta, \theta_0) \to H(\theta) \text{ a. s. } [P_{\theta_0}], \text{ as } n \to \infty. \tag{3.3}$$

This relation will be used in the following.

Lemma 3. *For any neighborhood* $U = U(\theta_0)$ *of* θ_0 *there exists a* $\delta = \delta(U(\theta_0)) > 0$ *such that, as* $n \to \infty$, $\liminf [\inf\{1/n \sum_{j=1}^{n} h(X_{j-1}, X_j; \theta, \theta_0); \theta \in U^c\}] > \delta$ a. s. $[P_{\theta_0}]$.

Proof. For $t \in U^c$ let $V_k(t) = \{\theta; d(\theta, t) < 1/k\}$. Then for k large enough $V_k(t) \subset W(t)$, and hence, $\sup\{f(x, y; \theta); \theta \in V_k(t)\}$ is measurable by (A6). We set $h(x, y; V_k(t)) = \inf\{h(x, y; \theta, \theta_0); \theta \in V_k(t)\}$. Then from the measurability of

$\sup\{f(x, y; \theta); \theta \in V_k(t)\}$ and the definition of $h(x, y; \theta, \theta_0)$ it follows that $h(x, y; V_k(t))$ is measurable. Clearly, $h(x, y; V_k(t))$ does not decrease as $k \to \infty$. On the other hand, the u. s. c. assumption (A5) of $f(x, y; \theta)$ in θ implies the u. s. c. in θ of $\log[f(x, y; \theta)/f(x, y; \theta_0)]$, and this, in turn, implies the lower semicontinuity (l. s. c.) in θ of $h(x, y; \theta, \theta_0)$. Therefore, $h(x, y; V_k(t))$ converges to $h(x, y; t, \theta_0)$, as $k \to \infty$. To summarize: For k sufficiently large $h(x, y; V_k(t))$ is measurable, and as $k \to \infty$, $h(x, y; V_k(t)) \to h(x, y; t, \theta_0)$ non-decreasingly.

For k large enough it is also true that $V_k(t) \subset W^*(t)$ and hence $h(x, y; V_k(t))$ is measurable and also bounded below by the integrable function $h(x, y; W^*(t))$, according to (A7). Therefore, the FATOU-LEBESGUE Theorem ([5], p. 125) applies and gives $\int h(x, y; V_k(t)) dP_{1, \theta_0} \to \int h(x, y; t, \theta_0) dP_{1, \theta_0}$ as $k \to \infty$. Equivalently, $\mathscr{E}_{\theta_0}[h(X_0, X_1; V_k(t)] \to H(t) > 0$, as $k \to \infty$. Thus, for every $t \in U^c$ there exists an open set $V_k(t)$ containing t, and a positive integer $N(t)$ such that $\mathscr{E}_{\theta_0}[h(X_0, X_1; V_k(t)] > 1/2\, H(t)$ for $k > N(t)$. Take U to be open. Then U^c is closed, hence compact. So there is a finite number of sets $V_k(t)$ covering U^c. Let them be $V_k(t_i)$, $i = 1, \ldots, m$. Thus, if $k > N = \max\{N(t_i), i = 1, \ldots, m\}$ we see that $\mathscr{E}_{\theta_0}[h(X_0, X_1; V_k(t_i)] > \delta > 0$, where $\delta = \min\{1/2\, H(t_i), i = 1, \ldots, m\}$. Now, for every $\theta \in U^c$ there is an i such that $\theta \in V_k(t_i)$. Then $h(x, y; \theta, \theta_0) \geqq h(x, y; V_k(t_i))$. Next, as $n \to \infty$, $1/n \sum_{j=1}^{n} h(X_{j-1}, X_j; V_k(t_i)) \to \mathscr{E}_{\theta_0}[h(X_0, X_1; V_k(t_i))] > 1/2\, H(t_i)$ for $k > N$. Therefore, as $n \to \infty$,

$$\liminf[\min\{1/4 \sum_{j=1}^{n} h(X_{j-1}, X_j; V_k(t_i)); i = 1, \ldots, n\}] > \delta \text{ a. s. } [P_{\theta_0}].$$

But

$$\liminf[\inf\{1/n \sum_{j=1}^{n} h(X_{j-1}, X_j; \theta, \theta_0); \theta \in U^c\}] \geqq$$

$$\liminf[\min\{1/n \sum_{j=1}^{n} h(X_{j-1}, X_j; V_k(t_i); i = 1, \ldots, m\}], \quad \text{as} \quad n \to \infty.$$

That is,

$$\liminf[\inf\{1/n \sum_{j=1}^{n} h(X_{j-1}, X_j; \theta, \theta_0); \theta \in U^c\}] \geqq \delta \text{ a. s. } [P_{\theta_0}], \quad \text{as} \quad n \to \infty.$$

This completes the proof of the lemma.

Lemma 1 and Lemma 3 taken together imply that for n sufficiently large $\hat{\theta}_n$ will lie a. s. $[P_{\theta_0}]$ in $U(\theta_0)$. In other words, $\hat{\theta}_n \to \theta_0$ a. s. $[P_{\theta_0}]$, as $n \to \infty$, and this concludes the proof of the theorem.

3. Extension

Assumption (A1) includes compactness of the metric space Θ. In practice, however, there occur interesting problems where Θ is locally compact, but not compact. We assert that under this weaker assumption about Θ the result still holds true. In fact, the compactness of Θ was used only in proving Lemma 3. If we merely assume that Θ is locally compact, Lemma 3 is still true, provided we take U^c to be the complement of U with respect to a compact neighborhood $C(\theta)$ of θ, where U is an open neighborhood of θ contained in $C(\theta)$. This is all we needed from Lemma 3, together with Lemma 1, to establish the Theorem.

References

[1] Basu, D.: An inconsistency of the method of maximum likelihood. Ann. math. Statistics **26**, 144—146 (1955).

[2] Billingsley, P.: Statistical Inferences for Markov Processes. The University of Chicago Press (1961).

[3] Kraft, C.: Some conditions of consistency and uniform consistency of statistical procedures. University of California Publications in Statistics **2**: **6**, 125—242 (1955).

[4] —, and L. Lecam: A remark on the roofs of the maximum likelihood equation. Ann. math. Statistics **27**, 1174—1176 (1956).

[5] Loève, M.: Probability Theory, 3rd ed., Princeton, N. J.: Van Nostrand 1963.

[6] Neyman, J., and E. Scott: Consistent estimates based on partially consisted observations. Econometrica **16**, 1—32 (1948).

[7] Wald, A.: Note on the consistency of the maximum likelihood estimate. Ann. math. Statistics **20**, 595—601 (1949).

Mathematics Department
San Jose State College,
San Jose 14, California

(Received November 16, 1964)