The Annals of Statistics
1981, Vol. 9, No 1, 225–228

NOTE ON THE CONSISTENCY OF THE MAXIMUM LIKELIHOOD ESTIMATE FOR NONIDENTIFIABLE DISTRIBUTIONS

BY RICHARD REDNER[1]

NASA, Johnson Space Center

The results of Wald on the consistency of the maximum likelihood estimate are extended. Applications are made to mixture distributions and to clustering when the number of clusters is not known.

1. Introduction. The question of consistency of the maximum likelihood estimate has been investigated by several authors (see, for example, Chanda [1], Cramér [2], Wald [4] and Wolfowitz [5]). In this note we observe that Wald's result can be extended to cover parameterizations for which the true distribution is represented by more than one parameter. This includes families of distributions for which the distributions are not identifiable and the case where the nonuniqueness is caused by the particular parameterization. In this note we use this fact to establish the consistency of the maximum likelihood estimate for mixture distributions with compact parameter space. An application is given in the area of cluster analysis.

Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed n-dimensional random variables. We will assume that the distribution of X_1 is known except for some parameter θ. The set of all parameter points Ω is called the parameter space and θ_0 will denote the true parameter. It will also be assumed that there is a σ-finite measure μ such that for each $\theta \in \Omega$ the probability measure μ_θ is absolutely continuous with respect to μ. We let $f(x, \theta)$ denote any representative of the density of μ_θ with respect to μ.

The following are the general assumptions made by Wald.

ASSUMPTION 1. The parameter space Ω is a metric space with metric $\delta(\cdot, \cdot)$ and has the property that every closed and bounded subset of Ω is compact.

Before stating the next assumption we need to define two functions. Let $N_r(\theta)$ denote the closed ball of radius r about θ. For $\theta \in \Omega$ and any positive real numbers r and s let

$$f(x, \theta, r) = \sup_{\phi \in N_r(\theta)} f(x, \phi), \qquad f^*(x, \theta, r) = \max(1, f(x, \theta, r))$$

$$h(x, s) = \sup_{\phi \notin N_s(\theta_0)} f(x, \phi), \qquad h^*(x, s) = \max(1, h(x, s)).$$

ASSUMPTION 2. For each θ and for sufficiently small r and sufficiently large s, $f(\cdot, \theta, r)$ is measurable and

(a) $$\int \log f^*(x, \theta, r)\, du_{\theta_0}$$

and

(b) $$\int \log h^*(x, s)\, du_{\theta_0}$$

are finite.

ASSUMPTION 3. If $\delta(\theta_0, \theta_i) \to +\infty$ then $f(x, \theta_i) \to 0$ except on a set A which does not depend on θ_i and has μ_{θ_0} measure zero.

Received February 1979; revised November 1979.

[1] National Research Council Associate.

AMS 1970 *subject classification.* 62F10.

Key words and phrases. Clustering, consistency, maximum likelihood estimation, mixture densities.

ASSUMPTION 4.

$$\int |\log f(x, \theta_0)| \, du_{\theta_0} < \infty.$$

ASSUMPTION 5. If $\theta_i \to \theta$ then $f(x, \theta_i) \to f(x, \theta)$ except on a set A which does not depend on the sequence θ_i and has μ_{θ_0} measure zero.

ASSUMPTION 6. If $\theta \neq \theta_0$ then $\mu_\theta \neq \mu_{\theta_0}$.

The following two theorems have been proven by Wald.

THEOREM 1. (Wald). *Suppose that Assumptions* 1–6 *are satisfied and let* S *be any closed subset of the parameter space which does not contain the true parameter point* θ_0. *Then*

$$P\left\{\lim_{N\to\infty} \sup_{\theta\in S} \frac{\prod_{i=1}^N f(x_i, \theta)}{\prod_{i=1}^N f(x_i, \theta_0)} = 0\right\} = 1.$$

THEOREM 2. (Wald). *If Assumptions* 1–6 *are satisfied and* $\bar{\theta}_N(x_1, \ldots, x_N)$ *is any function of the observations* $x_1, \ldots, x_N$ *such that*

$$\prod_{i=1}^N \frac{f(x_i, \bar{\theta}_N)}{f(x_i, \theta_0)} \geq c > 0 \quad \textit{for all} \quad N$$

then $P\{\lim_{N\to\infty} \bar{\theta}_N = \theta_0\} = 1$.

The maximum likelihood estimate is an obvious example of such a function.

2. An extension of Wald's theorems. We now want to consider the case where $C \equiv \{\theta \in \Omega \mid \mu_\theta = \mu_{\theta_0}\}$ is not a singleton set. Under the Assumptions 1–5 of Wald, C is a compact set and the following theorem is immediate from Wald's proof of Theorem 1.

THEOREM 3. *Let Assumptions* 1–5 *be satisfied and* $C \equiv \{\theta \in \Omega \mid \mu_\theta = \mu_{\theta_0}\}$. *If* S *is any closed subset of* Ω *not intersecting* C *then*

$$P\left\{\lim_{N\to\infty} \sup_{\theta\in S} \frac{\prod_{i=1}^N f(x_i, \theta)}{\prod_{i=1}^N f(x_i, \theta_0)} = 0\right\} = 1.$$

Hence if N_C is any open neighborhood containing C then the probability that $\bar{\theta}_N$ is cofinally in N_C is one. Letting $\hat{\Omega}$ be the quotient topological space obtained from Ω by identifying C to a point denoted $\hat{\theta}_0$, we have the following.

THEOREM 4. *If Assumptions* 1–5 *are satisfied then* $P\{\bar{\theta}_N \to \hat{\theta}_0\} = 1$.

It should also be noted that if $\tilde{\Omega}$ is the quotient topological space obtained from $\hat{\Omega}$ by identifying those parameters whose related densities are equal almost everywhere, then it follows from the theory of quotient spaces that the maximum likelihood estimate also converges to the true parameter in this topological space. We also observe that since C is compact, it follows from Theorem 4 and Assumption 5 that $f(x, \bar{\theta}_N) \to f(x, \hat{\theta}_0)$ in measure. With some further assumptions about the exceptional sets, one gets stronger convergence results.

3. Applications to mixture families. Let $\{\mu_\theta\}_{\theta\in\Omega}$ be a dominated family of probability measures on R^n. Let $J = \{(\alpha_1, \ldots, \alpha_m) \mid \sum_{i=1}^m \alpha_i = 1, \alpha_i \geq 0, i = 1, \ldots, m\}$, and let $\Gamma = J \times$

Ω^m. The set of all measures γ of the form $\gamma = \sum_{i=1}^m \alpha_i \mu_{\theta_i}$ where $(\alpha_1, \ldots, \alpha_m) \in J$ and $(\theta_1, \ldots, \theta_m) \in \Omega^m$ is the set of all mixtures of order m from $\{\mu_\theta\}_{\theta \in \Omega}$ and we will let $g(x, \gamma)$ denote this mixture density. We observe that Γ is a natural and convenient parameterization of this mixture family and has a natural topology; however, elements of Γ are not identifiable. The only natural topological spaces which identifiably parameterize this mixture family are quotient spaces related to Γ and we will show that for compact Γ, the maximum likelihood estimate is strongly consistent in the topological space $\tilde{\Gamma}$. First we will need to modify one of the assumptions put forward in Section 1.

ASSUMPTION 4′. For $(\theta, \phi) \in \Omega \times \Omega$, $\int |\log f(x, \theta)| \, du_\phi < \infty$.

THEOREM 5. *If Assumptions* 1, 2a, 4′, *and* 5 *are satisfied for* Ω *and if* Γ' *is a compact subset of* Γ *containing* γ_0, *then the maximum likelihood estimate of* γ_0 *in* Γ' *is strongly consistent in the topological space* $\hat{\Gamma}'$.

PROOF. The theorem will be established by showing that conditions 1–5 are satisfied for Γ'.

The following inequalities show that $\int \log g^*(x, \gamma, r) \, du_{\gamma_0} < \infty$.

$$\int \log g^*(x, \gamma, r) \, du_{\gamma_0} = \int \log \max[1, \sup_{\gamma' \in N_r(\gamma)} \textstyle\sum_{i=1}^m \alpha_i f(x, \theta_i)] \, du_{\gamma_0}$$

$$\leq \int \log \max[1, \sup_{\gamma' \in N_r(\gamma)} f(x, \theta_1), \ldots, \sup_{\gamma' \in N_r(\gamma)} f(x, \theta_m)] \, du_{\gamma_0}$$

$$\leq \textstyle\sum_{i=1}^m \displaystyle\int \log \max[1, \sup_{\gamma' \in N_r(\gamma)} f(x, \theta_i)] \, du_{\gamma_0}.$$

We now show that $\int |\log g(x, \gamma_0)| \, du_{\gamma_0} < \infty$.

Let $c = \min(\alpha_1, \ldots, \alpha_m)$, $A_1 = \{x \in R^n \mid \sum_{i=1}^m \alpha_i f(x, \theta_i) \geq 1\}$ and $A_2 = R^n \backslash A_1$. Then we have that

$$\int |\log \textstyle\sum_{i=1}^m \alpha_i f(x, \theta_i)| \, du_{\gamma_0} \leq \displaystyle\int_{A_1} |\log \max_i f(x, \theta_i)| \, du_{\gamma_0} + \int_{A_2} |\log c \textstyle\sum_{i=1}^m f(x, \theta_i)| \, du_{\gamma_0}$$

$$\leq \left\{ \textstyle\sum_{i=1}^m \displaystyle\int_{A_1} |\log f(x, \theta_i)| \, du_{\gamma_0} \right\} + \textstyle\sum_{i=1}^m \left\{ \displaystyle\int_{A_2} |\log f(x, \theta_i)| \, du_{\gamma_0} \right\} + c'$$

$$= \textstyle\sum_{i=1}^m \displaystyle\int |\log f(x, \theta_i)| \, du_{\gamma_0} + c'.$$

This establishes Assumption 4. All of the other assumptions are obviously satisfied and this concludes our proof.

In the particular the above assumptions hold if $\{\mu_\theta\}_{\theta \in \Omega}$ is a subset of the family of multivariate normal distributions with the property that Ω is compact. This last result should be compared to the results of Peters and Walker [3] on the strong consistency of the solution to the likelihood equations for mixtures of normals.

The problem which we now address is one in cluster analysis. The problem of cluster analysis is to take a set of observations and to form groups of data points which are known as clusters. The hope is that these clusters will represent some natural partitioning of the observations. In other words it is hoped that each cluster reflects some property of the objects which are being observed. It is therefore natural to assume that the observations come from a mixture of distributions. The goal then becomes not only to estimate the parameters for these individual distributions but to determine the number of mixing distributions. The problem of determining the number of mixing distributions or classes has received some attention but with little result. The next theorem offers one solution to this problem.

Let $\beta: \Omega \times \Omega \to R$ be a continuous function satisfying $\beta(\theta, \phi) = 0$ if $f_\theta = f_\phi$ $\{\mu\}$.

THEOREM 6. *Let Ω satisfy Assumptions* 1, 2a, 4′ *and* 5 *and let each of ε_1 and ε_2 be positive numbers. Suppose that $\{\mu_\theta\}_{\theta \in \Omega}$ is an identifiable set of distributions. Let Γ' be a subset of Γ which satisfies the following conditions.*

(a) *Γ' is compact.*

(b) *If $\gamma \in \Gamma'$ then $\alpha_i = 0$ or $\alpha_i \in [\varepsilon_1, 1]$, $i = 1, \ldots, m$.*

(c) *If $\gamma \in \Gamma'$ then $\beta(\theta_i, \theta_j) \geq \varepsilon_2$, $i \neq j$, $\theta_i, \theta_j \in \gamma$.*

(d) $\gamma_0 \in \Gamma'$.

It follows that $\bar{\gamma}_N$ is strongly consistent in $\hat{\Gamma}'$ and furthermore the number of nonzero prior probabilities for $\bar{\gamma}_N$ is the correct number of classes with probability one as $N \to \infty$.

PROOF. From Theorem 3 it follows that $\bar{\gamma}_N$ is consistent. Conditions (b) and (c) guarantee that every representative of $f(x, \gamma_0)$ has the same number of nonzero prior probabilities and in fact the continuity of β implies that there is an open neighborhood N_0 of $\hat{\gamma}_0$ in $\hat{\Gamma}'$ such that if $\gamma \in N_0$ then the number of nonzero prior probabilities for γ is the same as the number of nonzero prior probabilities for $\hat{\gamma}_0$. Since $\bar{\gamma}_N$ is eventually in this neighborhood it follows that the number of classes is eventually determined. This concludes our proof.

Again this theorem applies to the case that $\{\mu_\theta\}_{\theta \in \Omega}$ is a subset of the family of multivariate normal distributions with Ω a compact set. Although the function in Theorem 4 need only satisfy certain conditions, one obvious choice of β is the Bhattacharya coefficient. This would be a good choice because of its relation to the probability of misclassification.

REFERENCES

[1] CHANDA, K. C. (1954). A note on the consistency and maxima of the roots of the likelihood equations. *Biometrika* **41** 56–61.

[2] CRAMÉR, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

[3] PETERS, B. C. and WALKER, H. F. (1978). An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.* **35** 362–378.

[4] WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

[5] WOLFOWITZ, J. (1949). On Wald's proof of the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 601–602.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
LYNDON B. JOHNSON SPACE CENTER
HOUSTON, TX 77058