A Note on the Consistency and Maxima of the Roots of Likelihood Equations

K. C. Chanda

Biometrika, Vol. 41, No. 1/2. (Jun., 1954), pp. 56-61.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28195406%2941%3A1%2F2%3C56%3AANOTCA%3E2.0.CO%3B2-I

Biometrika is currently published by Biometrika Trust.

A NOTE ON THE CONSISTENCY AND MAXIMA OF THE ROOTS OF LIKELIHOOD EQUATIONS

By K. C. CHANDA, *Bombay University*

1. Introduction

An exhaustive study of the properties of the maximum-likelihood estimates was made earlier by Huzurbazar (1948). He did, however, confine his discussions to distributions with only one unknown parameter. Although the extension of his results to multi-parametric distributions is conceptually easy, the algebraic details of such derivations are sometimes complex and at times quite interesting to follow. The consistency of the maximum-likelihood estimates for the general case was, however, deduced later by Wald (1949), but on the assumption that the estimate really maximizes the likelihood function; it is thus different from the maximum-likelihood estimate as defined by Fisher. No such assumptions are made here; but the basic assumptions underlying the proofs are completely different from, and in a sense stronger than, those of Wald.

2. Assumptions

Suppose $f(x, \theta)$ be the probability density law; $\theta = (\theta_1, \ldots, \theta_k)$ is the unknown parameter vector and let $x_1, \ldots, x_n$ be n independent observations on x. The likelihood equations for estimating θ are then given by

$$\frac{\partial \log \phi}{\partial \theta_r} = 0 \quad (r = 1, 2, \ldots, k)$$

where

$$\log \phi = \sum_{i=1}^{n} \log f(x_i, \theta).$$

For brevity's sake we shall henceforth write f for $f(x, \theta)$ and f_i for $f(x_i, \theta)$.

Let us now assume that:

(i) The point represented by the vector θ lies in a k-dimensional interval Ω; for almost all x and for all $\theta \in \Omega$

$$\frac{\partial \log f}{\partial \theta_r}, \quad \frac{\partial^2 \log f}{\partial \theta_r \partial \theta_s} \quad \text{and} \quad \frac{\partial^3 \log f}{\partial \theta_r \partial \theta_s \partial \theta_t}$$

exist for all $r, s, t = 1, 2, \ldots, k$.

(ii) For almost all x and for every point $\theta \in \Omega$

$$\left|\frac{\partial f}{\partial \theta_r}\right| < F_r(x), \quad \left|\frac{\partial^2 f}{\partial \theta_r \partial \theta_s}\right| < F_{rs}(x) \quad \text{and} \quad \left|\frac{\partial^3 \log f}{\partial \theta_r \partial \theta_s \partial \theta_t}\right| < H_{rst}(x),$$

while

$$\int_{-\infty}^{\infty} H_{rst}(x) f \, dx < M$$

for all $\theta \in \Omega$ and for all $r, s, t = 1, 2, \ldots, k$, M being a finite positive constant.

(iii) For all $\theta \in \Omega$ the matrix $J = ((J_{rs}(\theta)))$, where

$$J_{rs}(\theta) = \int_{-\infty}^{\infty} \frac{\partial \log f}{\partial \theta_r} \frac{\partial \log f}{\partial \theta_s} f \, dx$$

is positive-definite and that $|J|$ is finite.

Assumption (iii) forms the pivot of later proofs; but it is not, as it stands, very much restrictive. The reason is that, whenever we are studying the properties of different estimates (which include maximum-likelihood estimates as well), we have always at the background the concept of efficiency which in turn is measured by a simple function of the information-determinant $|J|$. For the class of unbiased estimates, the generalized variance is bounded below by a quantity which assumes the trivial value, zero, when $|J| = \infty$. This assumption is thus felt to be of quite general importance. As J is a dispersion matrix it is always positive-definite, except perhaps in degenerate cases. This may be proved very simply as follows:

Let $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ be any real row vector. Clearly

$$E\left\{\left[\lambda_1\frac{\partial \log f}{\partial \theta_1}+\lambda_2\frac{\partial \log f}{\partial \theta_2}+\ldots+\lambda_k\frac{\partial \log f}{\partial \theta_k}\right]\right\}^2 \geqslant 0,$$

i.e.

$$\Sigma\Sigma\lambda_r\lambda_s J_{rs} \geqslant 0,$$

which proves $[J]$ to be positive-definite.

3. Derivation of results

Let θ^0 be the unknown true value of the parameter vector θ, where we suppose that θ^0 is an inner point of Ω. Consider the following expansion:

$$\frac{\partial \log f}{\partial \theta_r} = \left(\frac{\partial \log f}{\partial \theta_r}\right)_{\theta=\theta^0} + \sum_{s=1}^{k}(\theta_s-\theta_s^0)\left(\frac{\partial^2 \log f}{\partial \theta_r \partial \theta_s}\right)_{\theta=\theta^0} + \frac{1}{2}\sum_{s,t=1}^{k}(\theta_s-\theta_s^0)(\theta_t-\theta_t^0)\left(\frac{\partial^3 \log f}{\partial \theta_r \partial \theta_s \partial \theta_t}\right)_{\theta=\theta'}, \quad (3\cdot1)$$

where $\theta' = \theta'(x)$ is a value depending on x, but for all x, lying inside the hyper-cell, of which the vector $\theta-\theta^0$ is the diagonal. Multiplying both sides by $1/n$ and summing the corresponding expressions for x_i's over $i = 1, 2, \ldots, n$ we may rewrite (3·1) as

$$L_r(\theta) = L_r(\theta^0) - \sum_{s=1}^{k}\delta_s L_{rs}(\theta^0) + \frac{1}{2}\sum_{s,t=1}^{k}\delta_s\delta_t L_{rst} \quad (r = 1, 2, \ldots, k), \quad (3\cdot2)$$

where

$$\left.\begin{aligned} \delta_s &= (\theta_s-\theta_s^0), \\ L_r(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \log f_i}{\partial \theta_r}\right) = \frac{1}{n}\frac{\partial \log \phi}{\partial \theta_r}, \\ L_{rs}(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \log f_i}{\partial \theta_r \partial \theta_s}\right) = -\frac{1}{n}\frac{\partial^2 \log \phi}{\partial \theta_r \partial \theta_s}, \\ L_{rst} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^3 \log f_i}{\partial \theta_r \partial \theta_s \partial \theta_t}\right)_{\theta=\theta'(x_i)} \end{aligned}\right\} \quad (3\cdot3)$$

We note that

$$\left.\begin{aligned} &\text{(i)} \quad E\left(\frac{\partial \log f}{\partial \theta_r}\right) = 0 && (r = 1, 2, \ldots, k), \\ &\text{(ii)} \quad E\left(-\frac{\partial^2 \log f}{\partial \theta_r \partial \theta_s}\right) = J_{rs}(\theta) && (r, s = 1, 2, \ldots, k), \\ &\text{(iii)} \quad E\left(\left|\frac{\partial^3 \log f}{\partial \theta_r \partial \theta_s \partial \theta_t}\right|\right) < M && (r, s, t = 1, 2, \ldots, k), \end{aligned}\right\} \quad (3\cdot4)$$

for all $\theta \in \Omega$.

From (iii) it follows immediately that

$$E\left|\frac{\partial^3 \log f}{\partial\theta_r \partial\theta_s \partial\theta_t}\right|_{\theta=\theta'(x)} < M \quad \text{for all } x$$

(since $\theta'(x) \in \Omega$ for all x). Also

$$|L_{rst}| \leqslant \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3 \log fi}{\partial\theta_r \partial\theta_s \partial\theta_t}\right|_{\theta=\theta'(x_i)} < \frac{1}{n}\sum_{i=1}^{n} H_{rst}(x_i)$$

and

$$E\{H_{rst}(x_i)\} < M$$

for all $i = 1, 2, \ldots, n$; so by Khintchine's theorem,

$$\left.\begin{aligned} &\text{(i)}\ L_r(\theta^0) \to 0, \\ &\text{(ii)}\ L_{rs}(\theta^0) \to J_{rs}(\theta^0), \\ &\text{(iii)}\ |L_{rst}| < \frac{1}{n}\sum_{i=1}^{n} H_{rst}(x_i) \to E\{H_{rst}(x_i)\} < M, \quad \text{i.e. } |L_{rst}| \to B < M, \end{aligned}\right\} \tag{3·5}$$

where B is a finite-positive constant, with probability tending to unity. That is to say, we can choose an $n_0 = n_0(\eta, \epsilon)$, such that for all $n > n_0(\eta, \epsilon)$

$$\Pr[|L_r(\theta^0)| < \eta;\ |L_{rs}(\theta^0) - J_{rs}(\theta^0)| < \eta;\ |L_{rst}| < N] > 1 - \epsilon, \tag{3·6}$$

N being a finite-positive quantity greater than M and η and ϵ being two arbitrarily chosen small positive quantities. The likelihood equations are therefore given by putting $L_r(\theta) = 0$ $(r = 1, 2, \ldots, k)$ in (3·2), i.e.

$$L_r(\theta^0) - \sum_{s=1}^{k} \delta_s L_{rs}(\theta^0) + \frac{1}{2}\sum_{s,t=1}^{k} \delta_s \delta_t L_{rst} = 0 \quad (r = 1, 2, \ldots, k), \tag{3·7}$$

i.e.

$$\sum_{s=1}^{k} \delta_s L_{rs}(\theta^0) = L_r(\theta^0) + \frac{1}{2}\sum_{s,t=1}^{k} \delta_s \delta_t L_{rst} \quad (r = 1, 2, \ldots, k).$$

Again, since $E(L_{rs}(\theta^0)) = J_{rs}(\theta^0)$ and since $((J_{rs}(\theta^0)))^{-1} = J_0^{-1}$ (say) exists, $((L_{rs}(\theta^0)))^{-1} = L^{-1}$ (say) exists, too, and so we have from (3·7)

$$\delta_r = \beta_r + \sum_{s,t=1}^{k} \delta_s \delta_t a_{rst}, \tag{3·8}$$

where

$$\beta_r = \sum_{p=1}^{k} L_p(\theta^0) L^{pr}(\theta^0),$$

$$a_{rst} = \sum_{p=1}^{k} L_{pst} L^{pr}(\theta^0),$$

$$((L^{pr}(\theta^0))) = L^{-1}.$$

Since L^{-1} is a matrix with finite elements, we can choose our $n_0(\eta, \epsilon)$ for arbitrary η, ϵ such that for all $n > n_0(\eta, \epsilon)$

$$\Pr\left[|\beta_r| < \sum_{p=1}^{k} |L_p(\theta^0)|\,|L^{pr}(\theta^0)| < \eta;\right.$$
$$\left.|a_{rst}| < \frac{1}{2}\sum_{p=1}^{k} |L_{pst}|\,|L^{pr}(\theta^0)| < T \quad \text{for all } r = 1, 2, \ldots, k\right]$$
$$> 1 - \epsilon \quad \text{for} \quad \infty > T > N. \tag{3·9}$$

Consider now the set of equations

$$\delta_r = \beta_r + \sum_{s,t=1}^{k} \delta_s \delta_t a_{rst} \quad (r = 1, 2, \ldots, k).$$

It is easily seen that this set of equations admits a solution $\hat{\delta} = (\hat{\delta}_1, \ldots, \hat{\delta}_k)$ which are of the same order as η. For if we take η to be arbitrarily small, then in view of (3·9) it follows that the contribution due to the second term on the right-hand side can be made to be equal to a quantity of smaller order than η; and so

$$\Pr[|\hat{\delta}| < \eta] > 1 - \epsilon$$

for a sufficiently large value of n and for all values larger than that. It follows immediately, therefore, that there exists at least one solution of the likelihood equations, which is a consistent estimate of the true parameter vector θ^0. It is not, however, at once evident that this equation does not admit any other solution which may be inconsistent. We propose, however, to show that the consistent solution is unique. But before that we shall prove the following

THEOREM 1. The matrix $\Theta = \left(\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\hat{\theta}}\right)$, where $\hat{\theta}$ is a consistent root of the likelihood equation, is negative-definite with probability tending to unity.

Proof. Consider, in fact, the relation

$$\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\hat{\theta}} = \left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\theta^0} + \sum_{t=1}^{k} (\hat{\theta}_t - \theta_t^0) \left(\frac{\partial^3 \log \phi}{\partial\theta_r \partial\theta_s \partial\theta_t}\right)_{\theta=\theta''}, \tag{3·10}$$

where as before $\theta'' = \theta''(x_1, \ldots, x_n)$ is, for all $(x_1, \ldots, x_n)$ an inner point of the hyper-cell of which the vector $\hat{\theta} - \theta^0$ is the diagonal.

We have

$$\left|\frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\hat{\theta}} - \frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\theta^0}\right| \leqslant \sum_{t=1}^{k} |\hat{\theta}_t - \theta_t^0| \left|\frac{1}{n}\left(\frac{\partial^3 \log \phi}{\partial\theta_r \partial\theta_s \partial\theta_t}\right)_{\theta=\theta''}\right|. \tag{3·11}$$

Since $\hat{\theta}$ is a consistent estimate of θ^0, and since $\theta''(x_1, \ldots, x_n)$ is an interior point of Ω,

$$\left.\begin{aligned} \Pr\Bigg[|\hat{\theta}_t - \theta_t^0| < \eta/2kN \quad (t = 1, 2, \ldots, k); \\ \frac{1}{n}\left|\frac{\partial^3 \log \phi}{\partial\theta_r \partial\theta_s \partial\theta_t}\right|_{\theta=\theta''} < N\Bigg] > (1 - \epsilon) \quad \text{for all } n > n_1(\eta, \epsilon). \end{aligned}\right\} \tag{3·12}$$

Hence for $n > n_1(\eta, \epsilon)$, we have

$$\Pr\left[\left|\frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\hat{\theta}} - \frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\theta^0}\right| < \tfrac{1}{2}\eta\right] > 1 - \epsilon. \tag{3·13}$$

Also $E\left(\frac{\partial^2 \log f}{\partial\theta_r \partial\theta_s}\right)_{\theta=\theta^0} = -J_{rs}(\theta^0)$, and so, by Khintchine's theorem, it follows that

$$\Pr\left[\left|\frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\theta^0} + J_{rs}(\theta^0)\right| < \tfrac{1}{2}\eta\right] > 1 - \epsilon \tag{3·14}$$

for $n > n_2(\eta, \epsilon)$.

Again we have

$$\left|\frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\hat{\theta}} + J_{rs}(\theta^0)\right| \leqslant \left|\frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\hat{\theta}} - \frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\theta^0}\right| + \left|\frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\theta^0} + J_{rs}(\theta^0)\right|.$$

Hence for $n > n_3 = \max(n_1, n_2)$, we have

$$\Pr\left[\left|\frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\hat{\theta}} + J_{rs}(\theta^0)\right| < \tfrac{1}{2}\eta + \tfrac{1}{2}\eta = \eta\right] > 1-\epsilon. \tag{3·15}$$

Consider the positive-definite quadratic form

$$Q_0 = \sum_{r,s=1}^{k} u_r u_s J_{rs}(\theta^0).$$

Let further

$$Q = \sum_{r,s=1}^{k} u_r u_s \Theta_{rs},$$

where

$$\Theta_{rs} = \frac{1}{n}\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\hat{\theta}}.$$

We have for $n > n_3$,

$$\Pr\left[-Q_0 - \eta(\Sigma u_r)^2 < Q < -Q_0 + \eta(\Sigma u_r)^2\right] > 1-\epsilon. \tag{3·16}$$

Since

$$Q_0 = \sum_{r,s=1}^{k} u_r u_s J_{rs}(\theta^0) \quad \text{and} \quad Q_0' = \left(\sum_{r=1}^{k} u_r\right)^2$$

are respectively of rank k and 1, it follows that we can by a non-singular linear transformation $u_r \to U_r$ convert Q_0 and Q_0' to $\sum_{r=1}^{k} U_r^2$ and μU_1^2 respectively, where $\mu > 0$ is the only non-zero root of the characteristic equation $|A - \mu J_0| = 0$, A being the matrix with all unit elements. Evidently $0 < \mu < \infty$, and so

$$\frac{Q_0}{\left(\sum_{r=1}^{k} u_r\right)^2} = \frac{1}{\mu} + \frac{\sum_{r=2}^{k} U_r^2}{\mu U_1^2} > \frac{1}{\mu} > \xi \text{ (say)}. \tag{3·17}$$

By taking $\eta = \xi$ and by choosing an $n_3(\eta, \epsilon)$ we can ensure that

$$\Pr[Q \leqslant 0] > 1-\epsilon \quad \text{for all } n > n_3(\eta, \epsilon)$$

and so

$$\Pr[\Theta \text{ is negative-definite}] > 1-\epsilon \quad \text{for all } n > n_3(\eta, \epsilon). \tag{3·18}$$

It follows immediately that the likelihood function has a relative maximum at $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ with probability tending to unity. Now we prove the following

THEOREM 2. Of all possible solutions to equations (3·8), one and only one tends in probability to the true parameter vector θ^0.

Proof. Suppose, if possible, $\hat{\theta}'$ be another consistent estimate and at the same time a solution of (3·8). By hypothesis,

$$\left(\frac{\partial \log \phi}{\partial\theta_r}\right)_{\theta=\hat{\theta}} = 0, \quad \left(\frac{\partial \log \phi}{\partial\theta_r}\right)_{\theta=\hat{\theta}'} = 0 \quad \text{for } r = 1, 2, \ldots, k. \tag{3·19}$$

Now let us consider only one such pair of equations, say the rth. It follows, therefore, in virtue of Rolle's theorem extended to multivariate functions that there exists a point $\hat{\theta}_r''$ (depending on r), lying inside the hyper-cell of which the vector $\hat{\theta} - \hat{\theta}'$ is the diagonal, such that

$$\left(\frac{\partial^2 \log \phi}{\partial\theta_r \partial\theta_s}\right)_{\theta=\hat{\theta}_r''} = 0 \quad \text{for } s = 1, 2, \ldots, k. \tag{3·20}$$

It is therefore immediately evident that the determinant of the matrix $\left(\left(\frac{\partial^2 \log \phi}{\partial\theta_p \partial\theta_s}\right)_{\theta=\hat{\theta}''_r}\right)$ vanishes. This, however, is a contradiction, since $\hat{\theta}''_r$ being a consistent estimate of θ^0 (since it lies inside the hyper-cell with diagonal $\hat{\theta}-\hat{\theta}'$ and since $\hat{\theta}$, $\hat{\theta}'$ are both consistent estimates of θ^0), the Hermitean $\left(\left(\frac{\partial^2 \log \phi}{\partial\theta_p \partial\theta_s}\right)_{\theta=\hat{\theta}''_r}\right)$ must be negative-definite with probability tending to unity, as we have seen earlier; hence the theorem is proved.

Consider now the likelihood equations

$$\sum_{s=1}^{k} \delta_s L_{rs}(\theta^0) = L_r(\theta^0) + \frac{1}{2} \sum_{r,t=1}^{k} \delta_s \delta_t L_{rst} \quad (r = 1, 2, \ldots, k). \tag{3·21}$$

We remember that:

(1) $L_r(\theta^0) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \log f_i}{\partial\theta_r}\right)_{\theta=\theta^0}$, being the average of n independent variables, each having identical distribution, is itself distributed asymptotically normally with zero mean and variance $= J_{rr}(\theta^0)/n$. Again in virtue of the generalization of Liapounoff's central limit theorem, it follows that $L_1(\theta^0), \ldots, L_k(\theta^0)$ are asymptotically jointly distributed as a k-variate normal distribution, with zero means and variance-covariance matrix J_0/n.

(2) $L_{rs}(\theta^0) \to J_{rs}(\theta^0)$ with probability tending to unity.

(3) The second expression on the right-hand side of (3·21) is of a smaller order than $\sum_{s=1}^{k} \delta_s L_{rs}(\theta^0)$.

It follows, therefore, in virtue of Khintchine's theorem, that δ_r's have asymptotically a joint k-variate normal distribution with zero means and the variance-covariance matrix given by $V = ((nJ_{rs}(\theta^0)))^{-1}$.

REFERENCES

HUZURBAZAR, V. S. (1948). The likelihood equation, consistency, and the maxima of the likelihood function. *Ann. Eugen., Lond.*, **14**, 185.

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