An Inconsistent Maximum Likelihood Estimate

THOMAS S. FERGUSON*

An example is given of a family of distributions on $[-1, 1]$ with a continuous one-dimensional parameterization that joins the triangular distribution (when $\theta = 0$) to the uniform (when $\theta = 1$), for which the maximum likelihood estimates exist and converge strongly to $\theta = 1$ as the sample size tends to infinity, whatever be the true value of the parameter. A modification that satisfies Cramér's conditions is also given.

KEY WORDS: Maximum likelihood estimates; Inconsistency; Asymptotic efficiency; Mixtures.

1. INTRODUCTION

There are many examples in the literature of estimation problems for which the maximum likelihood principle does not yield a consistent sequence of estimates, notably Neyman and Scott (1948), Basu (1955), Kraft and LeCam (1956), and Bahadur (1958). In this article a very simple example of inconsistency of the maximum likelihood method is presented that shows clearly one danger to be wary of in an otherwise regular-looking situation. A recent article by Berkson (1980) followed by a lively discussion shows that there is still interest in these problems.

The discussion in this article is centered on a sequence of independent, identically distributed, and, for the sake of convenience, real random variables, $X_1, X_2, \ldots$, distributed according to a distribution, $F(x \mid \theta)$, for some θ in a fixed parameter space Θ. It is assumed that there is a σ-finite measure with respect to which densities, $f(x \mid \theta)$, exist for all $\theta \in \Theta$. The maximum likelihood estimate of θ based on $X_1, \ldots, X_n$ is a value, $\hat{\theta}_n(x_1, \ldots, x_n)$ of $\theta \in \Theta$, if any, that maximizes the likelihood function

$$L_n(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

The maximum likelihood method of estimation goes back to Gauss, Edgeworth, and Fisher. For historical points, see LeCam (1953) and Edwards (1972). For a general survey of the area and a large bibliography, see Norton (1972).

The starting point of our discussion is the theorem of Cramér (1946, p. 500), which states that under certain regularity conditions on the densities involved, if θ is real valued and if the true value θ_0 is an interior point of Θ, then there exists a sequence of roots, $\hat{\theta}_n$, of the likelihood equation,

$$\frac{\partial}{\partial\theta} \log L_n(\theta) = 0,$$

that converges in probability to θ_0 as $n \to \infty$. Moreover, any such sequence $\hat{\theta}_n$ is asymptotically normal and asymptotically efficient. It is known that Cramér's theorem extends to the multiparameter case.

To emphasize the point that this is a local result and may have nothing to do with maximum likelihood estimation, we consider the following well-known example, a special case of some quite practical problems mentioned recently by Quandt and Ramsey (1978). Let the density $f(x \mid \theta)$ be a mixture of two normals, $N(0, 1)$ and $N(\mu, \sigma^2)$, with mixing parameter $\frac{1}{2}$,

$$f(x \mid \mu, \sigma) = \tfrac{1}{2}\varphi(x) + \tfrac{1}{2}\varphi((x - \mu)/\sigma)/\sigma,$$

where φ is the density of the standard normal distribution, and the parameter space is $\Theta = \{(\mu, \sigma) : \sigma > 0\}$. It is clear that for any given sample, $X_1, \ldots, X_n$, from this density the likelihood function can be made as large as desired by taking $\mu = X_1$, say, and σ sufficiently small. Nevertheless, Cramér's conditions are satisfied and so there exists a consistent asymptotically efficient sequence of roots of the likelihood equation even though maximum likelihood estimates do not exist.

A more disturbing example is given by Kraft and LeCam (1956), in which Cramér's conditions are satisfied, the maximum likelihood estimate exists, is unique, and satisfies the likelihood equation, but is not consistent. In such examples, it is possible to find the asymptotically efficient sequence of roots of the likelihood equation by first finding a consistent extimate and then finding the closest root or improving by the method of scoring as in Rao (1965). See Lehmann (1980) for a discussion of these problems.

Other more practical examples of inconsistency in the maximum likelihood method involve an infinite number of parameters. Neyman and Scott (1948) show that the maximum likelihood estimate of the common variance of a sequence of normal populations with unknown means based on a fixed sample size k taken from each population converges to a value lower than the true value as the number of populations tends to infinity. This example led directly to the paper of Kiefer and Wolfowitz (1956) on the consistency and efficiency of the maximum likelihood

* Thomas S. Ferguson is Professor, Department of Mathematics, University of California, Los Angeles, CA 90024. Research was supported in part by the National Science Foundation under Grant MCS77-2121. The author wishes to acknowledge the help of an exceptionally good referee whose very detailed comments benefited this article substantially.

© Journal of the American Statistical Association
December 1982, Volume 77, Number 380
Theory and Methods Section

estimates with infinitely many nuisance parameters. Another example, mentioned in Barlow et al. (1972), involves estimating a distribution known to be star-shaped (i.e., $F(\lambda x) \le \lambda F(x)$ for all $0 \le \lambda \le 1$ and all x such that $F(x) < 1$). If the true distribution is uniform on (0, 1), the maximum likelihood estimate converges to $F(x) = x^2$ on (0, 1).

The central theorem on the global consistency of maximum likelihood estimates is due to Wald (1949). This theorem gives conditions under which the maximum likelihood estimates and approximate maximum likelihood estimates (values of θ that yield a value of the likelihood function that comes within a fixed fraction c, $0 < c < 1$, of the maximum) are strongly consistent. Other formulations of Wald's Theorem and its variants may be found in LeCam (1953), Kiefer and Wolfowitz (1956), Bahadur (1967), and Perlman (1972). A particularly informative exposition of the problem may be found in Chapter 9 of Bahadur (1971).

The example contained in Section 2 has the following properties:

1. The parameter space Θ is a compact interval on the real line.
2. The observations are independent identically distributed according to a distribution $F(x \mid \theta)$ for some $\theta \in \Theta$.
3. Densities $f(x \mid \theta)$ with respect to some σ-finite measure (Lebesgue measure in the example) exist and are continuous in θ for all x.
4. (Identifiability) If $\theta \ne \theta'$, then $F(x \mid \theta)$ is not identical to $F(x \mid \theta')$.

It is seen that whatever the true value, θ_0, of the parameter, the maximum likelihood estimate, which exists because of 1, 2, and 3, converges almost surely to a fixed value (1 in the example) independent of θ_0.

Example 2 of Bahadur (1958) (Example 9.2 of Bahadur 1971) also has the properties stated previously, and the example of Section 2 may be regarded as a continuous version of Bahadur's example. However, the distributions in Bahadur's example seem rather artificial and the parameter space is countable with a single limit point. The example presented here is more natural; the sample space is $[-1, +1]$, the parameter space is [0, 1], and the distributions are familiar, each being a mixture of the uniform distribution and a triangular one.

In Section 3, it is seen how to modify the example using beta distributions so that Cramér's conditions are satisfied. This gives an example in which asymptotically efficient estimates exist and may be found by improving any convenient $0(\sqrt{n})$-consistent estimate by scoring, and yet the maximum likelihood estimate exists and eventually satisfies the likelihood equation but converges to a fixed point with probability 1 no matter what the true value of the parameter happens to be. Such an example was announced by LeCam in the discussion of Berkson's (1980) paper.

2. THE EXAMPLE

The following densities on $[-1, 1]$ provide a continuous parameterization between the triangular distribution (when $\theta = 0$) and the uniform (when $\theta = 1$) with parameter space $\Theta = [0, 1]$:

$$f(x \mid \theta) = (1 - \theta)\frac{1}{\delta(\theta)}\left(1 - \frac{|x - \theta|}{\delta(\theta)}\right) \times I_A(x) + \frac{\theta}{2} I_{[-1,1]}(x),$$

where A represents the interval $[\theta - \delta(\theta), \theta + \delta(\theta)]$, $\delta(\theta)$ is a continuous decreasing function of θ with $\delta(0) = 1$ and $0 < \delta(\theta) \le 1 - \theta$ for $0 < \theta < 1$, and $I_S(x)$ represents the indicator function of the set S. For $\theta = 1$, $f(x \mid \theta)$ is taken to be $\frac{1}{2} I_{[-1,1]}(x)$. It is assumed that independent identically distributed observations $X_1, X_2, \ldots$ are available from one of these distributions. Then conditions 1 through 4 of the introduction are satisfied. These conditions imply the existence of a maximum likelihood estimate for any sample size because a continuous function defined on a compact set achieves its maximum on that set.

Theorem. Let $\hat{\theta}_n$ denote a maximum likelihood estimate of θ based on a sample of size n. If $\delta(\theta) \to 0$ sufficiently fast as $\theta \to 1$ (how fast is noted in the proof), then $\hat{\theta}_n \to 1$ with probability 1 as $n \to \infty$, whatever be the true value of $\theta \in [0, 1]$.

Proof. Continuity of $f(x \mid \theta)$ in θ and compactness of Θ implies that the maximum likelihood estimate, $\hat{\theta}_n$, some value of θ that maximizes the log-likelihood function

$$l_n(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

exists. Since for $\theta < 1$

$$f(x \mid \theta) \le \frac{1 - \theta}{\delta(\theta)} + \frac{\theta}{2} < \frac{1}{\delta(\theta)} + \tfrac{1}{2},$$

we have that for each fixed positive number $\alpha < 1$,

$$\max_{0 \le \theta \le \alpha} \frac{1}{n} l_n(\theta) \le \frac{1}{\delta(\alpha)} + \tfrac{1}{2} < \infty$$

since $\delta(\theta)$ is decreasing. We complete the proof by showing that whatever be the true value of θ,

$$\max_{0 \le \theta \le 1} \frac{1}{n} l_n(\theta) \to \infty \quad \text{with probability one}$$

provided $\delta(\theta) \to 0$ sufficiently fast as $\theta \to 1$, since then $\hat{\theta}_n$ will eventually be greater than α for any preassigned $\alpha < 1$. Let $M_n = \max\{X_1, \ldots, X_n\}$. Then $M_n \to 1$ with probability one whatever be the true value of θ, and since $0 < M_n < 1$ with probability one,

$$\max_{0 \le \theta \le 1} \frac{1}{n} l_n(\theta) \ge \frac{1}{n} l_n(M_n)$$

$$\ge \frac{n - 1}{n} \log \frac{M_n}{2} + \frac{1}{n} \log \frac{1 - M_n}{\delta(M_n)}.$$

Therefore, with probability one

$$\lim_{n\to\infty}\inf \max_{0\le\theta\le 1}\frac{1}{n} l_n(\theta) \ge \log\frac{1}{2} + \lim_{n\to\infty}\inf\frac{1}{n}\log\frac{1-M_n}{\delta(M_n)}.$$

Whatever be the value of θ, M_n converges to 1 at a certain rate, the slowest rate being for the triangular ($\theta = 0$) since this distribution has smaller mass than any of the others in sufficiently small neighborhoods of 1. Thus we can choose $\delta(\theta) \to 0$ so fast as $\theta \to 1$ that $(1/n)\log((1 - M_n)/\delta(M_n)) \to \infty$ with probability one for the triangular and hence for all other possible true values of θ, completing the proof.

How fast is fast enough? Take $\theta = 0$ and note that if $0 < \epsilon < 1$,

$$\begin{aligned}\sum_n P_0(\sqrt[4]{n}(1 - M_n) > \epsilon) &= \sum_n P_0(M_n < 1 - \epsilon n^{-1/4})\\ &= \sum_n P_0(X < 1 - \epsilon n^{-1/4})^n\\ &= \sum_n (1 - \tfrac{1}{2}\epsilon^2 n^{-1/2})^n\\ &\le \sum_n \exp(-\tfrac{1}{2}\epsilon^2\sqrt{n}) < \infty\end{aligned}$$

so that by the Borel-Cantelli Lemma, $\sqrt[4]{n}(1 - M_n) \to 0$ with probability one. Therefore, the choice

$$\delta(\theta) = (1 - \theta)\exp(-(1 - \theta)^{-4}) + 1$$

gives a $\delta(\theta)$ that is continuous, decreasing, with $\delta(0) = 1$, $0 < \delta(\theta) < 1 - \theta$ for $0 < \theta < 1$, and

$$\frac{1}{n}\log\frac{1 - M_n}{\delta(M_n)} = \frac{1}{n(1 - M_n)^4} - \frac{1}{n} \to \infty$$

with probability one.

Although the maximum likelihood method fails asymptotically in this example, other methods of estimation can yield consistent estimates. Bayes methods, for example, would be strongly consistent for *almost all* θ with respect to the prior distribution, as implied by a general argument of Doob (1948). Simpler computationally, but not generally as accurate, are the estimates given by the method of moments or minimum χ^2 based on a finite number of cells, and such methods can be made to yield consistent estimates. Estimates that are consistent may also be constructed by the minimum distance method of Wolfowitz (1957).

If one simple condition were added to conditions 1 through 4 of the introduction, the argument of Wald (1949) would imply the strong consistency of the maximum likelihood estimates. This is a uniform boundedness condition that may be stated as follows: Let θ_0 denote the true value of the parameter. Then the maximum likelihood estimate $\hat{\theta}_n$ converges to θ_0 with probability one provided conditions 1 through 4 hold and

5. There is a function $K(x) \ge 0$ with finite expectation,

$$E_{\theta_0}K(x) = \int K(x)f(x \mid \theta_0)\,dx < \infty,$$

such that

$$\log\frac{f(x \mid \theta)}{f(x \mid \theta_0)} < K(x) \quad \text{for all} \quad x \quad \text{and all} \quad \theta.$$

(To get global consistency this assumption must be made for all $\theta_0 \in \Theta$, but $K(x)$ may depend on θ_0.) This condition is therefore not satisfied in the example. It would be satisfied if the parameter space were limited to, say, $[0, 1 - \epsilon]$ since the density would then be bounded.

3. A DIFFERENTIABLE MODIFICATION

Without much difficulty, this example can be modified so that the densities satisfy Cramér's conditions for the existence of an asymptotically efficient sequence of roots of the likelihood equation. This amounts to modifying the distributions so that the resulting density, $f(x \mid \theta)$, (a) has two continuous derivatives that may be passed beneath the integral sign in $\int f(x \mid \theta)dx \equiv 1$, (b) has finite and positive Fisher information at all points θ interior to Θ, and (c) satisfies $| \partial^2/\partial\theta^2 f(x \mid \theta) | < K(x)$ in some neighborhood of the true θ_0, where $K(x)$ is θ_0-integrable. The simplest modification is to use the family of beta densities on [0, 1] as follows. Let g denote the density of the $Be(\alpha, \beta)$ distribution,

$$g(x \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1} I_{[0,1]}(x),$$

and let f be the density of the mixture of a $Be(1, 1)$ (uniform) and a $Be(\alpha, \beta)$,

$$f(x \mid \theta) = \theta g(x \mid 1, 1) + (1 - \theta)g(x \mid \alpha(\theta), \beta(\theta)),$$

where $\alpha(\theta)$ and $\beta(\theta)$ are chosen to be twice continuously differentiable and to give the density a very sharp peak close to θ, say mean θ and variance tending to 0 sufficiently fast as $\theta \to 1$. Thus we take $\Theta = [\frac{1}{2}, 1]$,

$$\alpha(\theta) = \theta\delta(\theta), \quad \text{and} \quad \beta(\theta) = (1 - \theta)\delta(\theta).$$

The particular form of $\delta(\theta)$ is not important. What is important is that

1. $\delta(\theta)$ is twice continuously differentiable,
2. $(1 - \theta)\delta(\theta)$, and hence $\delta(\theta)$, is increasing on $[\frac{1}{2}, 1)$,
3. $\delta(\frac{1}{2}) > 2$ (to obtain identifiability), and
4. $\delta(\theta)$ tends to ∞ sufficiently fast as $\theta \to 1$.

For $\theta = 1$, $f(x \mid 1)$ is defined to be $g(x \mid 1, 1)$. Then $f(x \mid \theta)$ is continuous in $\theta \in [\frac{1}{2}, 1]$ for each x, and for the true $\theta_0 \in (\frac{1}{2}, 1)$, Cramér's conditions are satisfied.

The proof that every maximum likelihood sequence converges to 1 with probability one as $n \to \infty$ no matter what the true value of $\theta \in [\frac{1}{2}, 1]$ is completely analogous to the corresponding proof in Section 2, except that in the inequalities, Stirling's formula in the form

$$\sqrt{2\pi}\,\alpha^{\alpha-(1/2)}e^{-\alpha} \le \Gamma(\alpha) \le \sqrt{2\pi}\,\alpha^{\alpha-(1/2)}\exp(-\alpha + (1/12\alpha))$$

as in Feller (1950, p. 44) is useful. In this example, the slowest rate of convergence of $\max_{i\le n}X_i$ to 1 occurs for $\theta = \frac{1}{2}$. By the method of Section 2, it may be calculated that the function

$$\delta(\theta) = (1-\theta)^{-1}\exp((1-\theta)^{-2})$$

converges to ∞ sufficiently fast and satisfies conditions 1 to 4 of this section.

[*Received October 1980. Revised April 1982.*]

REFERENCES

BAHADUR, R.R. (1958), "Examples of Inconsistency of Maximum Likelihood Estimates," *Sankhya*, 20, 207–210.

——— (1967), "Rates of Convergence of Estimates and Test Statistics," *Annals of Mathematical Statistics*, 38, 303–324.

——— (1971), *Some Limit Theorems in Statistics*, Regional Conference Series in Applied Mathematics, 4, Philadelphia: SIAM.

BARLOW, R.E., BARTHOLOMEW, D.J., BREMNER, J.M., and BRUNK, H.D. (1972), *Statistical Inference Under Order Restrictions*, New York: John Wiley.

BASU, D. (1955), "An Inconsistency of the Method of Maximum Likelihood," *Annals of Mathematical Statistics*, 26, 144–145.

BERKSON, J. (1980), "Minimum Chi-Square, not Maximum Likelihood!" *Annals of Statistics*, 8, 457–487.

CRAMÉR, H. (1946), *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

DOOB, J. (1948), "Application of the Theory of Martingales," *Le Calcul des Probabilités et ses Applications. Colloques Internationaux du Centre National de la Researche Scientifique*, Paris, 23–28.

EDWARDS, A.W.F. (1972), *Likelihood*, Cambridge: Cambridge University Press.

FELLER, W. (1950), *An Introduction to Probability Theory and its Applications*, (Vol. 1, 1st Ed.), New York: John Wiley.

KIEFER, J., and WOLFOWITZ, J. (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *Annals of Mathematical Statistics*, 27, 887–906.

KRAFT, C.H., and LECAM, L.M. (1956), "A Remark on the Roots of the Maximum Likelihood Equation," *Annals of Mathematical Statistics*, 27, 1174–1177.

LECAM, L.M. (1953), "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates," *University of California Publications in Statistics*, 1, 277–328.

LEHMANN, E.L. (1980), "Efficient Likelihood Estimators," *The American Statistican*, 34, 233–235.

NEYMAN, J., and SCOTT, E. (1948), "Consistent Estimators Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.

NORTON, R.H. (1972), "A Survey of Maximum Likelihood Estimattion," *Review of the International Statistical Institute*, 40, 329–354, and part II (1973), 41, 39–58.

PERLMAN, M.D. (1972), "On the Strong Consistency of Approximate Maximum Likelihood Estimates," Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, 1, 263–281.

QUANDT, R.E., and RAMSEY, J.L. (1978), "Estimating Mixtures of Normal Distributions and Switching Regressions," *Journal of the American Statistical Association*, 73, 730–738.

RAO, C.R. (1965), *Linear Statistical Inference and Its Applications*, New York: John Wiley.

WALD, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate," *Annals of Mathematical Statistics*, 20, 595–601.

WOLFOWITZ, J. (1957), "The Minimum Distance Method," *Annals of Mathematical Statistics*, 28, 75–88.