

SOME PATHOLOGIES ASSOCIATED WITH MAXIMUM LIKELIHOOD ESTIMATION

Eric B. Hall
Department of Electrical Engineering
Southern Methodist University
Dallas, Texas 75275

Gary L. Wise
Department of Electrical and Computer Engineering
and
Department of Mathematics
The University of Texas at Austin
Austin, Texas 78712

Abstract

Various aspects of maximum likelihood estimation are presented which should be of interest to its many proponents in the area of information sciences and systems.

Development

The theoretical statistical community has long been aware of the many problems associated with the concept of maximum likelihood estimation. In fact, as indicated in [1], Sir Ronald Fisher's 1922 revival of the maximum likelihood method is seen by many to have been a historical retrogression since the method had been considered and rejected years earlier by none other than Carl Friedrich Gauss. The popularity of maximum likelihood estimation today is due primarily, in the view of Le Cam, to Fisher's "propaganda." In fact, Le Cam states in [2, p. 622] that "In view of Fisher's vast influence, it is perhaps not surprising that the presumed superiority of the method is still for many an article of faith promoted with religious fervor. This state of affairs remains, in spite of a long accumulation of evidence to the effect that maximum likelihood estimators are often useless, or grossly misleading." It is precisely this evidence with which this paper is concerned, since, judging by its popularity, many of the problems associated with the maximum likelihood method appear to have been overlooked

by many in the area of information sciences and systems. We begin with a definition of the maximum likelihood method.

Let \mathbf{N} denote the set of positive integers, and for a topological space T , let $B(T)$ denote the family of Borel subsets of T . Let Θ be a topological space. Let $\{P_\theta: \theta \in \Theta\}$ be a family of probability measures on $(\mathbf{R}, B(\mathbf{R}))$ such that there exists a σ -finite measure μ on $(\mathbf{R}, B(\mathbf{R}))$ such that P_θ is absolutely continuous with respect to μ for each $\theta \in \Theta$. For each $\theta \in \Theta$ let $f_\theta(\cdot): \mathbf{R} \rightarrow [0, \infty)$ denote the Radon-Nikodym derivative of P_θ with respect to μ . For $n \in \mathbf{N}$ let X_1, X_2, \dots, X_n denote n mutually independent random variables each having the distribution P_θ for some $\theta \in \Theta$. Let

$$L(\theta, X_1, X_2, \dots, X_n) = \prod_{i=1}^n f_\theta(X_i).$$

We say that $\hat{\theta}_n: \mathbf{R}^n \rightarrow \Theta$ is a maximum likelihood estimator of θ if $\hat{\theta}_n^{-1}(B(\Theta)) \subset B(\mathbf{R}^n)$, if

$\hat{\theta}_n(X_1, X_2, \dots, X_n) \in \Theta$, and if

$$L(\hat{\theta}_n(X_1, X_2, \dots, X_n), X_1, X_2, \dots, X_n) = \sup_{\theta \in \Theta} L(\theta, X_1, X_2, \dots, X_n) < \infty \text{ a.e. } [\mu].$$

To begin, the following example points out that a maximum likelihood estimator need not in general even exist. Consider a collection $\{X_1, \dots, X_n\}$ of mutually independent, identically distributed random variables each possessing a probability density function given by

$$f_{\theta}(x) = \frac{9}{10\sigma} h\left(\frac{x-\mu}{\sigma}\right) + \frac{1}{10}h(x-\mu)$$

where $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$ and where $h(\cdot)$ is the continuous zero mean, unit variance Gaussian probability density function. The likelihood function for this situation is given by

$$L(\theta, x_1, \dots, x_n) = \prod_{i=1}^n \left[\frac{9}{10\sigma} h\left(\frac{x_i-\mu}{\sigma}\right) + \frac{1}{10}h(x_i-\mu) \right]$$

Notice, in particular, that

$$L(\theta, x_1, \dots, x_n) > \frac{9}{10\sigma} h\left(\frac{x_1-\mu}{\sigma}\right) \prod_{j=2}^n \frac{1}{10}h(x_j-\mu).$$

Further, notice that this lower bound is unbounded for θ in Θ since it approaches infinity as $\sigma \rightarrow 0$ when $\mu = x_1$. Thus, a maximum likelihood estimate for θ does not exist since the supremum of the likelihood function over all θ in Θ is not finite.

Next, consider a sequence $\{X_1, X_2, \dots\}$ of mutually independent, identically distributed random variables each possessing a Gaussian distribution with unit variance and mean given by θ . Let S_n denote the sum of the first n of these random variables and note that $\frac{S_n}{n}$ is a maximum likelihood estimator of θ . Further, notice that $\frac{S_n}{n}$ is Gaussian with mean θ and variance $\frac{1}{n}$. Now, for a real number α , the Hodges-Le Cam estimator is given by:

$$T_n\left(\frac{S_n}{n}\right) = \begin{cases} \frac{S_n}{n} & \text{if } \frac{|S_n|}{n} \geq \left(\frac{1}{n}\right)^{\frac{1}{4}} \\ \frac{\alpha S_n}{n} & \text{if } \frac{|S_n|}{n} < \left(\frac{1}{n}\right)^{\frac{1}{4}} \end{cases}$$

Notice first that

$$\lim_{n \rightarrow \infty} P\left[\frac{|S_n|}{n} < \left(\frac{1}{n}\right)^{\frac{1}{4}}\right]$$

is equal to 0 if θ is nonzero via the strong law of large numbers and the dominated convergence theorem and is

equal to 1 if $\theta = 0$ via Chebyshev's inequality. Hence, if θ is nonzero then it follows that T_n and $\frac{S_n}{n}$ each converge in distribution to the constant θ . Now, notice that for $\theta = 0$, we have that

$$\text{VAR}(T_n) \leq \frac{\alpha^2}{n} + \frac{1}{n} P\left[\frac{|S_n|}{n} \geq \left(\frac{1}{n}\right)^{\frac{1}{4}}\right] \leq \frac{\alpha^2}{n} + \left(\frac{1}{n}\right)^{\frac{3}{2}},$$

and the variance of $\frac{S_n}{n}$ is equal to $\frac{1}{n}$. Thus, for $|\alpha| < 1$ and for n sufficiently large, the variance of T_n is strictly smaller than the variance of $\frac{S_n}{n}$. The point $\theta = 0$ is called a point of super-efficiency. Recalling that the maximum likelihood estimator $\frac{S_n}{n}$ achieves the Cramer-Rao lower bound, it follows that, at the point $\theta = 0$, and for n sufficiently large, the variance of $T_n\left(\frac{S_n}{n}\right)$ is in fact smaller than the Cramer-Rao lower bound when $|\alpha| < 1$.

Consider now a collection X_1, X_2, \dots, X_n of mutually independent random variables each possessing a probability density function given by

$$f_{\theta}(x) = \frac{1-\alpha}{\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2}\right] + \frac{\alpha\gamma}{\sqrt{2\pi(\gamma^2+\epsilon)}} \exp\left[\left(\frac{-(x-\mu)^2}{2}\right) \left(\frac{\gamma}{\gamma^2+\epsilon}\right)^2\right]$$

for $x \in \mathbb{R}$, $0 < \alpha < 1$, $\epsilon > 0$, and $\theta = (\mu, \gamma)$ where $\mu \in \mathbb{R}$ and $\gamma > 0$. When α is small, $f_{\theta}(x)$ behaves very much like a Gaussian density function with unit variance and mean μ . Recall that if f were such a density function then the maximum likelihood estimator of μ would be given by the sample mean, $\frac{X_1+X_2+\dots+X_n}{n}$. It is easy to see, however, that for the above case, the sample mean is not a maximum likelihood estimator of the mean even though the above density is very close to being a Gaussian density function.

As an example, choose $\alpha = 10^{-10^0}$ and note that

$$L(\theta, x_1, x_2, \dots, x_n) = \prod_{i=1}^n [g(x_i, \mu) + h(x_i, \mu, \gamma)]$$

where

$$g(x_i, \mu) = \frac{1-\alpha}{\sqrt{2\pi}} \exp \left[\frac{-(x_i - \mu)^2}{2} \right]$$

and,

$$h(x_i, \mu, \gamma) = \frac{\alpha\gamma}{\sqrt{2\pi(\gamma^2 + \epsilon)}} \exp \left[\left(\frac{-(x_i - \mu)^2}{2} \right) \left(\frac{\gamma}{\gamma^2 + \epsilon} \right)^2 \right]$$

Further, notice that

$$L(\theta, x_1, x_2, \dots, x_n) > h(x_1, \mu, \gamma) \prod_{i=2}^n g(x_i, \mu).$$

In particular, if $\mu = x_1$ and $\gamma = \sqrt{\epsilon}$ then

$$L(\theta, x_1, x_2, \dots, x_n) > \frac{\alpha}{2\sqrt{2\pi\epsilon}} \prod_{i=2}^n g(x_i, \mu)$$

which may be made to exceed any preassigned real number by choice of ϵ . Let $n = 10^{10^0}$. Then, for ϵ sufficiently small, simply estimating the mean by the first observation and discarding the others will provide a larger value for L than would be obtained by taking the sample mean of the 10^{10^0} observations even though, as noticed above, the density governing the observations is for all practical purposes Gaussian. Note that this example also raises serious concerns regarding robustness properties of maximum likelihood estimators.

The following example, inspired by [3, p. 151], points out that a maximum likelihood estimator need not be consistent. Fix a positive integer k and consider a collection $\{X_{ij}; i \in \mathbb{N}, 1 \leq j \leq k\}$ of mutually independent random variables with X_{ij} having a Gaussian distribution with mean μ_i and positive variance σ^2 for each $i \in \mathbb{N}$ and $1 \leq j \leq k$. Let p be an integer greater than 1 and consider a subset of $\{X_{ij}; i \in \mathbb{N}, 1 \leq j \leq k\}$ given by $\{X_{ij}; 1 \leq i \leq p, 1 \leq j \leq k\}$. Likelihood functions for μ_1, \dots, μ_p and σ^2 based upon elements from this subset are given by

$$L_1(\mu_1, x_{11}, \dots, x_{1k})$$

$$= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{k}{2}} \exp \left[\frac{-1}{2\sigma^2} ((x_{11} - \mu_1)^2 + \dots + (x_{1k} - \mu_1)^2) \right]$$

and

$$L_2(\sigma^2, x_{11}, \dots, x_{pk})$$

$$= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{kp}{2}} \exp \left[\frac{-1}{2\sigma^2} ((x_{11} - \mu_1)^2 + \dots + (x_{pk} - \mu_p)^2) \right],$$

respectively. It follows easily that a maximum likelihood estimator for μ_i where $1 \leq i \leq p$ is given by

$$T(X_{i1}, \dots, X_{ik}) = \frac{1}{k} (X_{i1} + \dots + X_{ik}).$$

Notice that

$$\log(L_2(\sigma^2, x_{11}, \dots, x_{pk})) =$$

$$\log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{kp}{2}} + \left(\frac{-1}{2\sigma^2} ((x_{11} - \mu_1)^2 + \dots + (x_{pk} - \mu_p)^2) \right) \right]$$

and that

$$\begin{aligned} & \frac{\partial}{\partial \sigma^2} \log(L_2(\sigma^2, x_{11}, \dots, x_{pk})) \\ &= \frac{-kp}{2\sigma^2} + \left[\frac{1}{2\sigma^4} ((x_{11} - \mu_1)^2 + \dots + (x_{pk} - \mu_p)^2) \right] \end{aligned}$$

from which it follows that a maximum likelihood estimator for σ^2 is given by

$$S(X_{11}, \dots, X_{pk}) = \frac{1}{kp} \sum_{i=1}^p \sum_{j=1}^k [X_{ij} - T(X_{i1}, \dots, X_{ik})]^2.$$

For a moment, consider a collection $\{Z_1, \dots, Z_n\}$ of mutually independent Gaussian random variables each with mean μ and positive variance σ^2 . Let $\bar{Z} = \frac{1}{n} (Z_1 + \dots + Z_n)$. In addition, let A be an $n \times n$ matrix (a_{ij}) such that the columns of A form an orthogonal basis for \mathbb{R}^n and such that $a_{i1} = \frac{1}{\sqrt{n}}$ for $1 \leq i \leq n$.

Define Y_i for $1 \leq i \leq n$ via

$$(Y_1 \dots Y_n) = (Z_1 \dots Z_n) A$$

and note that, since the columns of A are orthogonal, the Y_i 's are mutually independent Gaussian random variables with variances equal to σ^2 and means given by $E[Y_1] = \mu\sqrt{n}$ and $E[Y_i] = 0$ for $1 < i \leq n$. Finally notice that

$$\begin{aligned} \sum_{k=1}^n (Z_k - \bar{Z})^2 &= \left(\sum_{k=1}^n Z_k^2 \right) - n\bar{Z}^2 \\ &= \left(\sum_{k=1}^n Z_k^2 \right) - Y_1^2 = \left(\sum_{k=1}^n Y_k^2 \right) - Y_1^2 = \sum_{k=2}^n Y_k^2 \end{aligned}$$

which has a chi-square density with $n-1$ degrees of freedom. Hence, we see that

$$\frac{1}{\sigma^2} \sum_{j=1}^k [X_{ij} - T(X_{i1}, \dots, X_{ik})]^2$$

has a chi-square distribution with $k-1$ degrees of freedom for each positive integer $i \leq p$. Thus, $E[S(X_{i1}, \dots, X_{pk})] = \left(\frac{k-1}{k} \right) \sigma^2$. Letting $p \rightarrow \infty$ and applying the weak law of large numbers to the sequence

$$\left\{ \frac{1}{k} \sum_{j=1}^k [X_{ij} - T(X_{i1}, \dots, X_{ik})]^2 : i \in \mathbf{N} \right\}$$

implies that $S(X_{i1}, \dots, X_{pk})$ converges in probability to $\left(\frac{k-1}{k} \right) \sigma^2$. Thus, the maximum likelihood estimator $S(X_{i1}, \dots, X_{pk})$ is not consistent.

Now consider a collection $\{X_1, X_2, \dots, X_n\}$ of mutually independent random variables each with a probability density function given by $f_\theta(x) = \frac{1}{\theta} I_{(0,\theta]}(x)$ where $\theta \in \Theta = (0, \infty)$. Note that

$$L(\theta, x_1, x_2, \dots, x_n) = \frac{1}{\theta^n} I_{(0,\theta]} \left(\max_{i \leq n} x_i \right).$$

This function possesses a unique maximum at $\theta = \max_{i \leq n} x_i$. Hence, it follows that the maximum likelihood estimator of θ is given by

$$\hat{\theta}(x_1, x_2, \dots, x_n) = \max_{i \leq n} x_i.$$

Notice that this estimator of θ is less than θ with probability one.

The next example shows that a maximum likelihood estimator need not be unique. In particular, consider a family of probability density functions $\{f_\theta(\cdot)\}$ where $\theta \in \Theta = \mathbf{R}$ and $f_\theta(x) = I_{[\theta-\frac{1}{2}, \theta+\frac{1}{2}]}(x)$. For a fixed positive integer n , let X_1, X_2, \dots, X_n be a collection of

mutually independent, identically distributed random variables each with a probability density function given by $f_\theta(x)$ for some fixed, yet unknown, value of θ . Further, for $1 \leq i \leq n$, let Y_i denote the i -th order statistic of the X_i 's. Notice that

$$L(\theta, x_1, x_2, \dots, x_n) = 1 \text{ if } x_i \in \left[\theta - \frac{1}{2}, \theta + \frac{1}{2} \right]$$

for each positive integer $i \leq n$ and is equal to zero otherwise. That is,

$$L(\theta, x_1, x_2, \dots, x_n) = 1$$

$$\text{if } \max_{i \leq n} x_i - \frac{1}{2} \leq \theta \leq \min_{i \leq n} x_i + \frac{1}{2}$$

and is equal to zero otherwise. Therefore, any statistic $T(X_1, X_2, \dots, X_n)$ for which

$$Y_n - \frac{1}{2} \leq T(X_1, X_2, \dots, X_n) \leq Y_1 + \frac{1}{2}$$

holds is a maximum likelihood estimator of the parameter θ . It follows easily that for any $\lambda \in [0, 1]$,

$$(1-\lambda) Y_n + \lambda Y_1 + \left(\lambda - \frac{1}{2} \right)$$

is a maximum likelihood estimator for θ . Hence, not only are maximum likelihood estimators not in general unique, but, as in this case, there may exist uncountably many distinct maximum likelihood estimators for a single parameter.

The following example presents a situation in which the number of roots of the likelihood equation is proportional to the number of observations. In particular, consider a collection $\{X_1, X_2, \dots, X_n\}$ of mutually independent random variables each with a Cauchy density function given by

$$f_\theta(x) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}$$

where $\theta \in \Theta = \mathbf{R}$. For this case it follows that

$$\begin{aligned} \log(L(\theta, x_1, x_2, \dots, x_n)) \\ = -n \log(\pi) - \sum_{i=1}^n \log(1+(x_i-\theta)^2), \end{aligned}$$

and, hence that

$$\frac{\partial}{\partial \theta} \log(L(\theta, x_1, x_2, \dots, x_n)) = \sum_{i=1}^n \frac{2(x_i-\theta)}{1+(x_i-\theta)^2}.$$

Notice that this expression may be written as a polynomial in θ of degree $2n - 1$. As the number of observations n increases, the computational demands of finding the roots of this polynomial increase dramatically. Not only must $2n - 1$ roots in general be considered, but a large number of roots must be checked to determine whether they correspond to local or global extrema.

The next example was suggested by [4, p. 208]. Consider a random variable X which has a Bernoulli distribution such that $P(X=1) = \theta$ and $P(X=0) = 1-\theta$ where $\theta \in \Theta = [\frac{1}{3}, \frac{2}{3}]$. The likelihood function resulting from a single observation is then given by $L(\theta, x) = (\theta)^x (1-\theta)^{1-x}$ which implies that a maximum likelihood estimator of θ is given by $\hat{\theta}(x) = \frac{x+1}{3}$ for $x = 0, 1$. Note that $E[\hat{\theta}(X)] = \frac{\theta+1}{3}$ which implies that this maximum likelihood estimator is not unbiased. Consider for a moment a family of estimators each of the form

$$\theta_\alpha(x) = \begin{cases} \alpha & \text{if } x = 0 \\ 1-\alpha & \text{if } x = 1 \end{cases}$$

where $\frac{1}{3} \leq \alpha \leq \frac{1}{2}$, and note that $\hat{\theta}(x) = \theta_{\frac{1}{3}}(x)$.

Further, note that

$$\begin{aligned} R(\alpha, \theta) &= E[(\theta_\alpha(X) - \theta)^2] \\ &= (\alpha-\theta)^2(1-\theta) + ((1-\alpha)-\theta)^2\theta. \end{aligned}$$

For the maximum likelihood estimator $\hat{\theta}$ given above we see that $R(\frac{1}{3}, \theta) = \frac{\theta^2}{3} - \frac{\theta}{3} + \frac{1}{9}$. Further,

$$R(\alpha, \theta) - R(\frac{1}{3}, \theta) = (\alpha - \frac{1}{3}) \left[(2\theta-1)^2 + (\alpha - \frac{2}{3}) \right] < 0$$

for all $\alpha \in [\frac{1}{3}, \frac{1}{2}]$ and all $\theta \in \Theta$ since $\theta \in \Theta$ implies that $(2\theta-1)^2 \leq \frac{1}{9}$ and $\alpha \in [\frac{1}{3}, \frac{1}{2}]$ implies that $(\alpha - \frac{2}{3}) \leq \frac{-1}{6}$ and $(\alpha - \frac{1}{3}) \geq 0$. Thus, not only is this maximum likelihood estimator not admissible, but it has the uniformly largest mean square error over all of the uncountably

many estimators in the family $\{\theta_\alpha: \alpha \in [\frac{1}{3}, \frac{1}{2}]\}$.

The following example, based on [2, p. 624], presents a situation in which a maximum likelihood estimator $\hat{\theta}_n$ exists for n observations yet is such that $\hat{\theta}_n \rightarrow \infty$ as $n \rightarrow \infty$ with probability one even though the actual value of θ is a fixed positive integer. In particular, let $\alpha_0 = 1$ and define a decreasing sequence of positive numbers via

$$\int_{\alpha_{k+1}}^{\alpha_k} \left(\exp\left(\frac{1}{x^2}\right) - \frac{1}{2} \right) dx = \frac{1}{2}$$

for each nonnegative integer k . Further, define a family of probability density functions $\{f_\theta(x) : \theta \in \Theta\}$ on $(0, 1]$ where $\Theta = \mathbb{N}$ and

$$f_\theta(x) = \frac{1}{2} + \left(\exp\left(\frac{1}{x^2}\right) - \frac{1}{2} \right) I_{(\alpha_\theta, \alpha_{\theta-1})}(x),$$

and let $\{X_n: n \in \mathbb{N}\}$ be a sequence of mutually independent random variables each with the same probability density function $f_\theta(x)$ for some fixed yet unknown value of θ from Θ . Note that

$$\log L(\theta, x_1, \dots, x_n) = \sum_{i=1}^n \Lambda(\theta, x_i)$$

where

$$\Lambda(\theta, x_i) = \begin{cases} \frac{1}{x_i^2} & \text{if } x_i \in (\alpha_\theta, \alpha_{\theta-1}] \\ -\log(2) & \text{otherwise.} \end{cases}$$

Hence, a maximum likelihood estimator $\hat{\theta}_n$ exists and is equal to an integer j for which the following term is maximized:

$$\left\{ \sum_{\{m: \alpha_j < x_m \leq \alpha_{j-1}\}} \frac{1}{x_m^2} \right\}$$

Note that for any $\beta > 0$, it follows from the Borel-Cantelli lemma that infinitely many terms from the sequence $\{X_n : n \in \mathbb{N}\}$ will fall in the interval $(0, \beta]$ almost surely. Thus it follows immediately that $\hat{\theta}_n$ tends almost surely to infinity even though the actual value of θ is a fixed positive integer.

The following example addresses an additional problem of uniqueness which may exist in maximum likelihood estimation. Let $\Theta = \mathbb{R}$ and let

$$f_{\theta}(x) = \left(\frac{1}{2}\right) \exp(-|x-\theta|)$$

for each θ in Θ and x in \mathbb{R} . Note that

$$L(\theta, x_1, \dots, x_n) = -\sum_{i=1}^n |x_i - \theta|$$

which thus implies that a maximum likelihood estimator for θ is given by a median of $\{X_1, \dots, X_n\}$. Further, notice that the observations X_1, \dots, X_n are almost surely distinct. Thus, we see that for an even number of observations, there exist almost surely uncountably many distinct maximum likelihood estimators for θ , yet for an odd number of observations, there exists almost surely a unique maximum likelihood estimator.

Conclusion

This paper has considered a few of the many problems which may arise in the use of maximum likelihood estimation. The prominent position of maximum likelihood estimation in information sciences and systems leads us to wonder if researchers in this area are fully aware of difficulties such as the ones which we have touched upon. In closing, we believe, as did Gauss, that maximum likelihood estimation should have been rejected long ago.

Acknowledgement

This research was supported partially by the Air Force Office of Scientific Research under Grant AFOSR-86-0026.

Several aspects of this paper were developed while the second author was visiting the Department of Statistics at the University of California at Berkeley, and he would like to acknowledge the helpful influence of many of his colleagues there.

References

- [1] Berkson, J., "Estimation by Least Squares and by Maximum Likelihood," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 1-11, (University of California Press, Berkeley, 1956).
- [2] Le Cam, L., *Asymptotic Methods in Statistical Decision Theory*, (Springer-Verlag, New York, 1986).
- [3] Bickel, P. J. and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, (Holden-Day, Oakland, California, 1977).
- [4] Kiefer, J. C., *Introduction to Statistical Inference*, (Springer-Verlag, New York, 1987).