CHAPTER 3. MAXIMUM LIKELIHOOD AND M-ESTIMATION

3.1 Maximum likelihood estimates — in exponential families. Let $(X, \mathcal{B})$ be a measurable space and $\{P_\theta,\ \theta \in \Theta\}$ a measurable family of laws on $(X, \mathcal{B})$, dominated by a σ-finite measure v. Let $f(\theta, x)$ be a jointly measurable version of the density $(dP_\theta/dv)(x)$ by Theorem 1.3.3. For each $x \in X$, a *maximum likelihood estimate (MLE)* of θ is any $\hat{\theta} = \hat{\theta}(x)$ such that $f(\hat{\theta}, x) = \sup\{f(\phi, x) :\ \phi \in \Theta\}$. In other words, $\hat{\theta}(x)$ is a point at which $f(\cdot, x)$ attains its maximum. In general, the supremum may not be attained, or it may be attained at more than one point. If it is attained at a unique point $\hat{\theta}$, then $\hat{\theta}$ is called *the* maximum likelihood estimate of θ. A measurable function $\hat{\theta}(\cdot)$ defined on a measurable subset B of X is called a *maximum likelihood estimator* if for all $x \in B$, $\hat{\theta}(x)$ is a maximum likelihood estimate of θ, and for v-almost all x not in B, the supremum of $f(\cdot, x)$ is not attained at any point.

Examples. (i) For each $\theta > 0$ let P_θ be the uniform distribution on $[0, \theta]$, with $f(\theta, x) := 1_{[0,\theta]}(x)/\theta$ for all x. Then if $X_1, \ldots, X_n$ are observed, i.i.d. (P_θ), the MLE of θ is $X_{(n)} := \max(X_1, \ldots, X_n)$. Note however that if the density had been defined as $1_{[0,\theta)}(x)$, its supremum for given $X_1, \ldots, X_n$ would not be attained at any θ. The MLE of θ is the smallest possible value of θ given the data, so it is not a very reasonable estimate in some ways. For example, it is not Bayes admissible.
(ii). For $P_\theta = N(\theta, 1)^n$ on $\mathbb{R}^n$, with usual densities, the sample mean $\overline{X}$ is the MLE of θ. For $N(0, \sigma^2)^n,\ \sigma > 0$, the MLE of σ^2 is $\sum_{j=1}^n X_j^2/n$. For $N(m, \sigma^2)^n,\ n \geq 2$, the MLE of (m, σ^2) is $(\overline{X}, \sum_{j=1}^n (X_j - \overline{X})^2/n)$. Here recall that the usual, unbiased estimator of σ^2 has $n-1$ in place of n, so that the MLE is biased, although the bias is small, of order $1/n^2$ as $n \to \infty$. The MLE of σ^2 fails to exist (or equals 0, if 0 were allowed as a value of σ^2) exactly on the event that all X_j are equal for $j \leq n$, which happens for $n = 1$, but only with probability 0 for $n \geq 2$. On this event, $f((\overline{X}, \sigma^2), x) \to +\infty$ as $\sigma \downarrow 0$.

In general, let Θ be an open subset of $\mathbb{R}^k$ and suppose $f(\theta, x)$ has first partial derivatives with respect to θ_j for $j = 1, \ldots, k$, forming the gradient vector

$$\nabla_\theta f(\theta, x) \ :=\ \{\partial f(\theta, x)/\partial \theta_j\}_{j=1}^k.$$

If the supremum is attained at a point in Θ, then the gradient there will be 0, in other words the *likelihood equations* hold,

$$\partial f(\theta, x)/\partial \theta_j \ =\ 0 \ \text{ for } \ j = 1, \ldots, k. \tag{3.1.1}$$

If the supremum is not attained on Θ, then it will be approached at a sequence of points $\theta^{(m)}$ approaching the boundary of Θ, or which may become unbounded if Θ is unbounded.

The equations (3.1.1) are called "maximum likelihood equations" in a number of statistics books and papers, but that is unfortunate terminology because in general a solution of (3.1.1) could also be (a) only a local, not a global maximum of the likelihood, (b) a local or global minimum of the likelihood, or (c) a saddle point, as in an example to be given below.

For exponential families, it will be shown that MLE's can be found from the likelihood equations, as follows:

3.1.2 Theorem. Let $\{P_\theta,\ \theta \in \Theta\}$ be an exponential family of order k, where Θ is the natural parameter space in a minimal representation (2.5.3). Let U be the interior of Θ and $j(\theta) := -\log C(\theta)$. Then for any n and observations $X_1, \ldots, X_n$ i.i.d. (P_θ), there is at most one MLE $\hat{\theta}$ in U. The likelihood equations have the form

$$\partial j/\partial\theta_i = \sum_{j=1}^n T_i(X_j)/n \quad \text{for} \quad i = 1, \ldots, k, \tag{3.1.3}$$

and have at most one solution in U, which if it exists is the MLE. Conversely, any MLE in U must be a solution of the likelihood equations. If an MLE exists in U for v-almost all x, it is a sufficient statistic for θ.

Proof. Maximizing the likelihood is equivalent to maximizing its logarithm (the log likelihood), which is

$$-nj(\theta) + \sum_{j=1}^n \sum_{i=1}^k \theta_i T_i(X_j),$$

and the gradient of the likelihood is 0 if and only if the gradient of the log likelihood is 0, which evidently gives the equations (3.1.3). Then $K = e^j$ is a smooth function of θ on U by Theorem 2.5.8, hence so is j, and the other summand in the log likelihood is linear in θ, so the log likelihood is a smooth (C^∞) function of θ. So at a maximum in U, the gradient must be 0, in other words (3.1.3) holds.

A real-valued function f on a convex set in $\mathbb{R}^k$ is called *concave* if $-f$ is convex and *strictly concave* if $-f$ is strictly convex. It is easily seen that a strictly concave function on a convex open set has at most one local maximum, which then must be a strict absolute maximum. Adding a linear function preserves (strict) convexity or concavity. So, to show that the log likelihood is strictly concave on U is equivalent to showing that j is strictly convex on U.

By Corollary 2.5.9, the matrix of second partial derivatives of j is the covariance matrix of the components of T. Since the representation is minimal, the covariance matrix is non-singular: otherwise there would be a non-zero linear combination of the T_i with 0 variance and so equal to a constant almost surely. Since a covariance matrix is always nonnegative definite, in this case it is positive definite. Then, restricted to any line segment included in U, j has a strictly positive second derivative along the segment by the chain rule, which implies that j is strictly convex. For any strictly concave function h on a convex open set, if the gradient of h exists and is 0 at a point θ, then h has a strict local maximum at θ, as can be seen first in the one-dimensional case, then taking all lines through θ in general. Then θ is a strict global maximum of h as desired.

If for almost all x, (3.1.3) has a solution $\theta = \theta(x)$ in U, which is then unique, then by (3.1.3), the vector $\{\sum_{j=1}^n T_i(X_j)\}_{i=1}^k$, which is a sufficient k-dimensional statistic as noted in Section 2.5, is a function of $\theta(x)$ which thus must also be sufficient. □

Next is an example to show that if a maximum likelihood estimate exists almost surely but may be on the boundary of the parameter space, it may not be sufficient. Let v be the law with density $2x^{-3}$ on $[1, \infty)$ and 0 elsewhere. Consider the exponential family of order 1 having densities $C(\theta)e^{\theta x}$ with respect to v, where $C(\theta)$ as usual is the normalizing

constant. Then the natural parameter space is $(-\infty, 0]$. According to Barndorff-Nielsen (1978, p. 153), if $x \geq 2$ is observed the MLE is $\theta = 0$. This can be proved as follows. We have $K(\theta) = 2\int_1^\infty e^{\theta x} x^{-3} dx$. Since the density is nonincreasing in x for any $\theta \leq 0$ it is enough to show that $e^{2\theta} < K(\theta)$ for any $\theta < 0$. Some classical special functions are defined for $t > 0$ and $n = 0, 1, 2, ...$, by $E_n(t) := \int_1^\infty e^{-tx} x^{-n} dx$ (Gautschi and Cahill, hereafter G&C, (5.1.4)). Thus $g(t) := K(-t) = 2E_3(t)$ for all $t > 0$ and we want to show this is larger than e^{-2t}. Both functions equal 1 at $t = 0$.

Case 1: $0 < t \leq 1/2$. It will be enough to show that $g'(t) = -2E_2(t) > -2e^{-2t}$, or equivalently $E_2(t) < e^{-2t}$ for $0 < t \leq 1/2$. We have $E_2(t) \equiv e^{-t} - tE_1(t)$ (integration by parts, G&C 5.1.14), so we want to show that $tE_1(t) > e^{-t} - e^{-2t}$. By G&C, (5.1.20), $E_1(t) > (e^{-t}/2)\log(1 + \frac{2}{t})$, so it will be enough to show that

$$h(t) := \frac{t}{2}\log\left(1 + \frac{2}{t}\right) - 1 + e^{-t} > 0,$$

$0 < t \leq 0.5$. Note that $e^{-t} \geq 1 - t + \frac{1}{2}t^2 - \frac{1}{6}t^3$ for all $t \geq 0$, since both sides are equal at $t = 0$, then taking derivatives and iterating. Thus it's enough to show that

$$\frac{t}{2}\log\left(1 + \frac{2}{t}\right) - t + \frac{1}{2}t^2 - \frac{1}{6}t^3 > 0$$

for $0 < t \leq 1/2$, which is equivalent to

$$\frac{t}{2}\left[\log\left(1 + \frac{2}{t}\right) - 2 + t - \frac{1}{3}t^2\right] > 0, \quad \text{or} \quad \log\left(1 + \frac{2}{t}\right) - 2 + t - \frac{1}{3}t^2 > 0.$$

Let $f(t) := \log\left(1 + \frac{2}{t}\right)$ and $g(t) := 2 - t + \frac{1}{3}t^2$. We can check that $f(1/2) = \log 5 \doteq 1.6094 > g(1/2) = 19/12 \doteq 1.5833$. So it suffices to check that $f'(t) = -2/[t/(t+2)] < g'(t) = \frac{2t}{3} - 1$ for $0 < t < 1/2$. So we need to check that $6t < 2t^3 + t^2 + 6$ which holds since $6t < 6$.

Case 2: $t \geq 0.33$. On this half-line we use the continued fraction for $E_3(t)$ given by G&C, (5.1.22), and the fact (noticed by Euler) that for continued fractions with all terms positive, successive convergents are alternately above and below the value of the continued fraction. This gives us a lower bound $2E_3(t) \geq 2(t+5)e^{-t}/[t^2 + 8t + 12]$ which we want to prove larger than e^{-2t}. Equivalently, we want to prove $j(t) := 2(t+5)e^t - t^2 - 8t - 12 > 0$ for $t \geq 0.33$. This can be directly verified for $t = 0.33$. We have $j'(t) = (2t+12)e^t - 2t - 8 = 2t(e^t - 1) + 4(3e^t - 2) > 0$ for all $t > 0$, so $j(t) > 0$ for all $t \geq 0.33$ as desired.

So, 0 is the MLE for any observation $x \geq 2$ as stated. But, the identity function x is a Lehmann-Scheffé sufficient statistic by factorization and Theorem 2.5.10, therefore minimal sufficient by Theorem 2.3.3, so the maximum likelihood estimator is not sufficient in this case although it is defined almost surely.

PROBLEMS

1. Let $X_1, ..., X_n$ be i.i.d. Poisson with unknown parameter λ, $0 < \lambda < \infty$.
   (a) As a function of λ, find the probability that a MLE exists in the parameter space.
   (b) When the MLE exists, show that it can be found from the likelihood equation and evaluate it.
2. In the example where for $x \geq 2$ the MLE is 0, evaluate $c_0 := \int_{-\infty}^{0} K(\theta)d\theta$. Using $d\pi(\theta) = K(\theta)d\theta/c_0$, evaluate the Bayes estimator of θ for squared-error loss.

NOTES

For the example in which for $x \geq 2$ the MLE is 0, M. Manstavičius provided some steps in the proof.

REFERENCES

Barndorff-Nielsen, O. (1978). See Sec. 3.5.

Gautschi, Walter, and Cahill, William F. Exponential integral and related functions. Chap. 5 of *Handbook of Mathematical Functions*, ed. M. Abramowitz and I. A. Stegun, Government Printing Office, 1964, 9th Dover printing, date of publication not given (after 1972).