Stat 5102 Notes: Maximum Likelihood

Charles J. Geyer

February 2, 2007

1 Likelihood

Given a parametric model specified by a p. f. or p. d. f. $f(\mathbf{x} \mid \boldsymbol{\theta})$, where either $\mathbf{x}$ or $\boldsymbol{\theta}$ may be a vector, the *likelihood* is the same function thought of as a function of the parameter (possibly a vector) rather than a function of the data, possibly with multiplicative terms not containing the parameter dropped.

We write

$$L(\boldsymbol{\theta}) = f(\mathbf{x} \mid \boldsymbol{\theta})$$

to denote the likelihood. Note that the left hand side is also a function of the data $\mathbf{x}$ even though the notation does not indicate this. So when the data is considered random, the likelihood is a random function. Sometimes we write $L_n(\boldsymbol{\theta})$ to indicate the likelihood for sample size n when it is important to keep track of n.

1.1 Binomial Likelihood

Suppose x is Binomial(n, θ). Then the p. f. is

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \qquad 0 \le \theta \le 1.$$

Since the binomial coefficient does not contain the parameter θ we can drop it from the likelihood.

$$L(\theta) = \theta^x (1 - \theta)^{n-x}, \qquad 0 \le \theta \le 1.$$

1.2 Normal Location-Scale Likelihood

Suppose x_1, x_2, ..., x_n are i. i. d. normal(μ, σ^2). Then the p. d. f. is

$$\begin{aligned} f(\mathbf{x} \mid \mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \end{aligned}$$

Since the $\sqrt{2\pi}$ terms do not involve the parameters μ and σ^2, we can drop them from the likelihood

$$L(\mu, \sigma^2) = \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right), \qquad -\infty < \mu < +\infty,\ \sigma > 0.$$

We are free to consider the parameter to be either standard deviation or variance, and variance is theoretically more tractable. So let us write $\nu = \sigma^2$ and

$$L(\mu, \nu) = \nu^{-n/2} \exp\left(-\frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2 \right), \qquad -\infty < \mu < +\infty,\ \nu > 0.$$

1.3 Normal Location Likelihood

If we consider the variance known in the preceding section, then μ is the only parameter, the likelihood is a function of μ only and we may drop multiplicative terms not containing μ, giving

$$L(\mu) = \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right), \qquad -\infty < \mu < +\infty.$$

1.4 Uniform Likelihood

Suppose x_1, x_2, ..., x_n are i. i. d. Uniform$(0, \theta)$. Then the p. d. f. is

$$\begin{aligned} f(\mathbf{x} \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(x_i) \\ &= \theta^{-n} I_{[0,\theta]}(x_{(n)}) \end{aligned}$$

where $x_{(n)}$ denotes the maximum $x_1, \ldots, x_n$. This follows because $I_{[0,\theta]}(x_i)$ is one when $x_i \le \theta$ and zero otherwise, the product of these terms is one when $x_i \le \theta$ holds for all i (and zero otherwise), and this in turn happens when $x_{(n)} \le \theta$. Thus the likelihood is

$$L(\theta) = \begin{cases} \theta^{-n}, & \theta \ge x_{(n)} \\ 0, & \text{otherwise} \end{cases} \tag{1.1}$$

2 Maximum Likelihood and Optimization

2.1 Local and Global Optimization

If g is a real-valued function on a set Θ, then we say a point θ is a *global maximum* if

$$g(\theta) \le g(\varphi), \qquad \varphi \in \Theta.$$

If Θ is also a metric space with distance function d, then we say a point θ is a *local maximum* if there exists an $r > 0$ such that

$$g(\theta) \le g(\varphi), \qquad \varphi \in \Theta \text{ and } d(\theta, \varphi) \le r.$$

2.2 Maximum Likelihood Estimation

There are two notions of maximum likelihood

- A point $\boldsymbol{\theta}$ that is the global maximum of the likelihood is called the *maximum likelihood estimate* (MLE).
- A point $\boldsymbol{\theta}$ that is a local maximum of the likelihood near to another good estimate of the parameter is also called the *maximum likelihood estimate*.

To distinguish between $\boldsymbol{\theta}$ considered as an ordinary variable that ranges over the parameter space and the MLE, we denote the latter by $\hat{\boldsymbol{\theta}}$ or by $\hat{\boldsymbol{\theta}}_n$ when it is important to keep track of the sample size n. Since the likelihood is a random function, the MLE is a random variable or random vector (since it depends on the data, although this is not indicated by the notation).

The two notations of maximum likelihood are important. Global optimization is difficult, especially when there are several parameters. So the first notion is often not useful. Moreover, there are situations where the global maximum is not well behaved or does not even exist but a good local maximum does exist and is well behaved.

2.3 Log Likelihood

For many purposes it is simpler to maximize the log likelihood

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$$

or

$$l_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta}).$$

Since the log function is order-preserving and infinitely differentiable points are local or global maximizers of the log likelihood if and only if they are are local or global maximizers (respectively) of the likelihood. Moreover the log likelihood has just as many derivatives as the likelihood.

2.4 Optimization in One Variable

Consider a real-valued function g on a set Θ that is an interval in the real line (possibly unbounded in one or both directions). In our examples g will be either log likelihood or likelihood, usually the former. We consider all of the facts in this section facts from calculus and will not prove them.

2.4.1 Necessary Conditions

- If θ is an interior point of Θ and a local maximum of g, then $g'(\theta) = 0$.
- If θ is an interior point of Θ and a local maximum of g, then $g''(\theta) \leq 0$.

Note that if Θ is a bounded interval, then if an endpoint is a local maximum, these necessary conditions need not apply: the first derivative need not be zero and the second derivative need not be nonpositive. These tests work only for interior points of the domain. Thus we may need to check whether the endpoints are local or global maxima, and that check needs to use the function g itself, because derivatives are not helpful.

2.4.2 Sufficient Conditions

- If $g'(\theta) = 0$, then we say θ is a *stationary point* of g.
- If $g'(\theta) = 0$ and $g''(\theta) < 0$, then θ is a *local maximum* of g.

2.4.3 Concavity Conditions

If g is continuous on Θ and $g''(\theta) < 0$ for all θ that are interior points of Θ, then we say g is a *strictly concave function*. In this case, any stationary point of g is the unique global maximum of g.

2.4.4 Binomial Likelihood

It is easier to use log likelihood. In this case

$$\begin{aligned} l(\theta) &= x \log(\theta) + (n - x) \log(1 - \theta) \\ l'(\theta) &= \frac{x}{\theta} - \frac{n - x}{1 - \theta} \\ l''(\theta) &= -\frac{x}{\theta^2} - \frac{n - x}{(1 - \theta)^2} \end{aligned}$$

Note that the derivatives only exist for $0 < \theta < 1$.

Since l is continuous on $0 \leq \theta \leq 1$ and $l''(\theta)$ is negative for $0 < \theta < 1$, we conclude that l is strictly concave and hence any local maximizer that we find will be the unique global maximizer.

Setting the first derivative equal to zero and solving for θ we find the solution x/n. Thus the MLE is

$$\hat{\theta} = \frac{x}{n} \tag{2.1}$$

if this solution satisfies $0 < \hat{\theta} < 1$, otherwise, this is not a solution to $l'(\theta) = 0$.

Thus we have not yet analyzed the cases $x = 0$ and $x = n$. When $x = 0$ we have

$$l(\theta) = n \log(1 - \theta).$$

Since log is an increasing function, $\theta \mapsto \log(1-\theta)$ is a decreasing function. Hence $l(\theta)$ is maximized when θ takes the smallest possible value, here zero. Thus (2.1) also gives the MLE in the case $x = 0$. A similar analysis of the case $x = n$ shows that (2.1) gives the MLE in that case too.

2.4.5 Normal Location Likelihood

Again, it is easier to use log likelihood. In this case

$$\begin{aligned} l(\mu) &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \\ l'(\mu) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ l''(\mu) &= -\frac{n}{\sigma^2} \end{aligned}$$

Since l is continuous and l'' is negative on $-\infty < \mu < \infty$, we conclude that l is strictly concave and hence any local maximizer that we find will be the unique global maximizer.

Setting the first derivative equal to zero and solving for μ we find the solution $\bar{x}_n$. Thus the MLE is

$$\hat{\mu}_n = \bar{x}_n. \tag{2.2}$$

Here there are no endpoints of the parameter space to worry about so (2.2) always gives the MLE.

2.4.6 Uniform Likelihood

Here it does not matter whether we use likelihood or log likelihood. Direct examination of the likelihood (1.1) shows that it is a decreasing function of θ where it is not zero. Where it is zero, the derivative is zero, but these points are minima not maxima. Thus calculus is no help in finding the maximum. Since the $L(\theta)$ increases as θ decreases from ∞ to $x_{(n)}$ and is zero for $\theta < x_{(n)}$, it is clear that the maximum occurs at $x_{(n)}$. Thus the MLE is

$$\hat{\theta}_n = x_{(n)}. \tag{2.3}$$

2.5 Optimization in Many Variables

Consider a function g of many variables $\theta_1, \ldots, \theta_k$ which we can also consider a function of a vector variable $\boldsymbol{\theta}$. Again let the domain be denoted Θ, which we take to be a subset of $\mathbb{R}^k$. Again, we consider all of the facts in this section facts from multivariable calculus and will not prove them.

2.5.1 Multivariate Derivatives

Now we no longer have a first derivative, rather we have k partial derivatives. We assemble them into a vector

$$\nabla g(\boldsymbol{\theta}) = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1} \cdots \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_k} \right)$$

which is called the *first derivative vector* or the *gradient vector*. And we no longer have a second derivative, rather we have k^2 second partial derivatives, including mixed partial derivatives. We assemble them into a matrix

$$\nabla^2 g(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_2} & \cdots & \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_k} \end{pmatrix}$$

which is called the *second derivative matrix* or the *Hessian matrix*.

2.5.2 Necessary Conditions

- If $\boldsymbol{\theta}$ is an interior point of Θ and a local maximum of g, then $\nabla g(\boldsymbol{\theta}) = 0$.
- If $\boldsymbol{\theta}$ is an interior point of Θ and a local maximum of g, then $\nabla^2 g(\boldsymbol{\theta})$ is a negative semi-definite matrix.

Note that if Θ has boundary points and a local maximum occurs on the boundary, these necessary conditions need not apply: the first derivative vector need not be zero and the second derivative need not be negative semi-definite. These tests work only for interior points of the domain.

Unlike the one-parameter case where the boundary has at most two points, checking all of the points on the boundary in higher dimensions is very difficult, so difficult that it takes weeks to cover in a course on constrained optimization, which is not a prerequisite for this course. Thus we will punt this issue. It is enough to know that the issue exists.

The notion of a negative semi-definite (or negative definite) matrix will be explained in Section 2.5.5 below after we state the rest of the conditions.

2.5.3 Sufficient Conditions

- If $\nabla g(\boldsymbol{\theta}) = 0$, then we say $\boldsymbol{\theta}$ is a *stationary point* of g.
- If $\nabla g(\boldsymbol{\theta}) = 0$ and $\nabla^2 g(\boldsymbol{\theta})$ is a negative definite matrix, then $\boldsymbol{\theta}$ is a *local maximum* of g.

2.5.4 Concavity Conditions

A set Θ in $\mathbb{R}^k$ is *convex* if for every two points in Θ the entire line segment connecting them is contained in Θ. Now assume that the domain Θ of the function g we are optimizing is a convex set.

If g is continuous on Θ and $\nabla^2 g(\boldsymbol{\theta})$ is a negative definite matrix for all $\boldsymbol{\theta}$ that are interior points of Θ, then we say g is a *strictly concave function*. In this case, any stationary point of g is the unique global maximum of g.

2.5.5 Negative Definite and Negative Semi-Definite Matrices

A square symmetric matrix $\mathbf{M}$ is said to be *positive semi-definite* if

$$t'\mathbf{M}t \geq 0, \qquad t \in \mathbb{R}^k,$$

where the prime in t' denotes the transpose. And $\mathbf{M}$ is said to be *positive definite* if

$$t'\mathbf{M}t > 0, \qquad t \in \mathbb{R}^k \text{ and } t \neq 0.$$

A matrix $\mathbf{M}$ is said to be *negative semi-definite* or *negative definite* if $-\mathbf{M}$ is positive semi-definite or positive definite (respectively).

It is often difficult to use the definition to check whether a matrix is negative definite. There are two other kinds of conditions that can be used to check. One is useful for computer checks, the other useful (somewhat) when checking by hand.

Any square symmetric matrix $\mathbf{M}$ has a spectral decomposition $\mathbf{M} = \mathbf{ODO}'$, where $\mathbf{O}$ is an orthogonal matrix, meaning $\mathbf{O}' = \mathbf{O}^{-1}$, and $\mathbf{D}$ is a diagonal matrix.

The diagonal elements of $\mathbf{D}$ are called the *eigenvalues* of $\mathbf{M}$ and the columns of $\mathbf{O}$ are called the *eigenvectors* of $\mathbf{M}$. $\mathbf{M}$ is negative definite if all its eigenvalues are negative. $\mathbf{M}$ is negative semi-definite if all its eigenvalues are nonpositive.

Eigenvalues and eigenvectors are very difficult for hand calculation (we will never try to do them), but they are easy for the computer. Thus this is an easy check when doing calculations with a computer.

The *principal minors* of a square symmetric matrix $\mathbf{M}$ are $\mathbf{M}$ itself and matrices obtained by deleting a set of rows and the corresponding columns of $\mathbf{M}$. $\mathbf{M}$ is positive definite if the determinants of its principal minors are all

positive. $\mathbf{M}$ is positive semi-definite if the determinants of its principal minors are all nonnegative.

If $\mathbf{M}$ is a $k \times k$ matrix, then the determinant of $-\mathbf{M}$ is $(-1)^k$ times the determinant of $\mathbf{M}$. So principal minors of odd dimension of a negative definite matrix are negative, while those of even dimension are positive.

2.5.6 Normal Location-Scale Likelihood

It is easier to use log likelihood. In this case

$$\begin{aligned}
l_n(\mu, \nu) &= -\frac{n}{2} \log(\nu) - \frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2 \\
\frac{\partial l_n(\mu, \nu)}{\partial \mu} &= \frac{1}{\nu} \sum_{i=1}^{n} (x_i - \mu) \\
\frac{\partial l_n(\mu, \nu)}{\partial \nu} &= -\frac{n}{2\nu} + \frac{1}{2\nu^2} \sum_{i=1}^{n} (x_i - \mu)^2 \\
\frac{\partial^2 l_n(\mu, \nu)}{\partial \mu^2} &= -\frac{n}{\nu} \\
\frac{\partial^2 l_n(\mu, \nu)}{\partial \mu \partial \nu} &= -\frac{1}{\nu^2} \sum_{i=1}^{n} (x_i - \mu) \\
\frac{\partial^2 l_n(\mu, \nu)}{\partial \nu^2} &= +\frac{n}{2\nu^2} - \frac{1}{\nu^3} \sum_{i=1}^{n} (x_i - \mu)^2
\end{aligned}$$

Now l is not strictly concave because $\partial^2 l_n(\mu, \nu)/\partial \nu^2$ is not negative for all possible parameter values (for sufficiently large ν the negative term will be smaller than the positive term).

Setting $\partial l_n(\mu, \nu)/\partial \mu$ to zero and solving for μ, we again obtain (2.2) as the MLE for μ. Then setting $\partial l_n(\mu, \nu)/\partial \nu$ to zero, plugging in (2.2) for μ and solving for ν gives

$$\hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \tag{2.4}$$

If we now plug in the MLE's into the second derivatives we obtain

$$\nabla^2 l_n(\hat{\mu}_n, \hat{\nu}_n) = \begin{pmatrix} -\frac{n}{\hat{\nu}_n} & 0 \\ 0 & -\frac{n}{2\hat{\nu}_n} \end{pmatrix}$$

Since this matrix is diagonal, its diagonal elements are its eigenvalues. Since its eigenvalues are negative, it is negative definite, and we see that the solution obtained is a local maximum.