Stat 5102 Lecture Slides: Deck 3
Likelihood Inference

Charles J. Geyer
School of Statistics
University of Minnesota

Likelihood Inference

We have learned one very general method of estimation: the method of moments.

Now we learn another: the method of maximum likelihood.

Likelihood

Suppose have a parametric statistical model specified by a PMF or PDF. Our convention of using boldface to distinguish between scalar data x and vector data $\mathbf{x}$ and a scalar parameter θ and a vector parameter $\boldsymbol{\theta}$ becomes a nuisance here. To begin our discussion we write the PMF or PDF as $f_\theta(x)$. But it makes no difference in likelihood inference if the data x is a vector. Nor does it make a difference in the fundamental definitions if the parameter θ is a vector.

You may consider x and θ to be scalars, but much of what we say until further notice works equally well if either x or θ is a vector or both are.

Likelihood (cont.)

The PMF or PDF, considered as a function of the unknown parameter or parameters rather than of the data is called the *likelihood function*

$$L(\theta) = f_\theta(x)$$

Although $L(\theta)$ also depends on the data x, we suppress this in the notation. If the data are considered random, then $L(\theta)$ is a random variable, and the function L is a random function. If the data are considered nonrandom, as when the observed value of the data is plugged in, then $L(\theta)$ is a number, and L is an ordinary mathematical function. Since the data X or x do not appear in the notation $L(\theta)$, we cannot distinguish these cases notationally and must do so by context.

Likelihood (cont.)

For all purposes that likelihood gets used in statistics — it is the key to both likelihood inference and Bayesian inference — it does not matter if multiplicative terms not containing unknown parameters are dropped from the likelihood function.

If $L(\theta)$ is a likelihood function for a given problem, then so is

$$L^*(\theta) = \frac{L(\theta)}{h(x)}$$

where h is any strictly positive real-valued function.

Log Likelihood

In frequentist inference, the *log likelihood function*, which is the logarithm of the likelihood function, is more useful. If L is the likelihood function, we write

$$l(\theta) = \log L(\theta)$$

for the log likelihood.

When discussing asymptotics, we often add a subscript denoting sample size, so the likelihood becomes $L_n(\theta)$ and the log likelihood becomes $l_n(\theta)$.

Note: we have yet another capital and lower case convention: capital L for likelihood and lower case l for log likelihood.

Log Likelihood (cont.)

As we said before (slide 5), we may drop multiplicative terms not containing unknown parameters from the likelihood function. If

$$L(\theta) = h(x)g(x, \theta)$$

we may drop the term $h(x)$. Since

$$l(\theta) = \log h(x) + \log g(x, \theta)$$

this means we may drop additive terms not containing unknown parameters from the log likelihood function.

Examples

Suppose X is Bin(n, p), then the likelihood is

$$L_n(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

but we may, if we like, drop the term that does not contain the parameter, so

$$L_n(p) = p^x (1-p)^{n-x}$$

is another (simpler) version of the likelihood.

The log likelihood is

$$l_n(p) = x \log(p) + (n-x) \log(1-p)$$

Examples (cont.)

Suppose X_1, ..., X_n are IID $\mathcal{N}(\mu, \nu)$, then the likelihood is

$$\begin{aligned} L_n(\mu, \nu) &= \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\nu}} e^{-(x_i-\mu)^2/(2\nu)} \\ &= (2\pi)^{-n/2} \nu^{-n/2} \exp\left(-\frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2 \right) \end{aligned}$$

but we may, if we like, drop the term that does not contain parameters, so

$$L_n(\mu, \nu) = \nu^{-n/2} \exp\left(-\frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

Examples (cont.)

The log likelihood is

$$l_n(\mu, \nu) = -\frac{n}{2}\log(\nu) - \frac{1}{2\nu}\sum_{i=1}^{n}(x_i - \mu)^2$$

We can further simplify this using the empirical mean square error formula (slide 7, deck 1)

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = v_n + (\bar{x}_n - \mu)^2$$

where $\bar{x}_n$ is the mean and v_n the variance of the empirical distribution. Hence

$$l_n(\mu, \nu) = -\frac{n}{2}\log(\nu) - \frac{nv_n}{2\nu} - \frac{n(\bar{x}_n - \mu)^2}{2\nu}$$

Log Likelihood (cont.)

What we consider the log likelihood may depend on what we consider the unknown parameters.

If we say μ is unknown but ν is known in the preceding example, then we may drop additive terms not containing μ from the log likelihood, obtaining

$$l_n(\mu) = -\frac{n(\bar{x}_n - \mu)^2}{2\nu}$$

If we say ν is unknown but μ is known in the preceding example, then every term contains ν so there is nothing to drop, but we do change the argument of l_n to be only the unknown parameter

$$l_n(\nu) = -\frac{n}{2}\log(\nu) - \frac{1}{2\nu}\sum_{i=1}^{n}(x_i - \mu)^2$$

Maximum Likelihood Estimation

The *maximum likelihood estimate* (MLE) of an unknown parameter θ (which may be a vector) is the value of θ that maximizes the likelihood in some sense.

It is hard to find the global maximizer of the likelihood. Thus a local maximizer is often used and also called an MLE. The global maximizer can behave badly or fail to exist when the right choice of local maximizer can behave well. More on this later.

$\widehat{\theta}_n$ is a global maximizer of L_n if and only if it is a global maximizer of l_n. Same with local replacing global.

Local Maxima

Suppose W is an open interval of $\mathbb{R}$ and $f : W \to \mathbb{R}$ is a differentiable function. From calculus, a necessary condition for a point $x \in W$ to be a local maximum of f is

$$f'(x) = 0 \qquad (*)$$

Also from calculus, if f is twice differentiable and $(*)$ holds, then

$$f''(x) \leq 0$$

is another necessary condition for x to be a local maximum and

$$f''(x) < 0$$

is a sufficient condition for x to be a local maximum.

Global Maxima

Conditions for global maxima are, in general, very difficult. Every known procedure requires exhaustive search over many possible solutions.

There is one special case — concavity — that occurs in many likelihood applications and guarantees global maximizers.

Concavity

Suppose W is an open interval of $\mathbb{R}$ and $f : W \to \mathbb{R}$ is a twice-differentiable function. Then f is *concave* if

$$f''(x) \leq 0, \qquad \text{for all } x \in W$$

and f is *strictly concave* if

$$f''(x) < 0, \qquad \text{for all } x \in W$$

Concavity (cont.)

From the fundamental theorem of calculus

$$f'(y) = f'(x) + \int_x^y f''(s)\,ds$$

Hence concavity implies f' is nonincreasing

$$f'(x) \geq f'(y), \qquad \text{whenever } x < y$$

and strict concavity implies f' is decreasing

$$f'(x) > f'(y), \qquad \text{whenever } x < y$$

Concavity (cont.)

Suppose $f'(x) = 0$. Another application of the fundamental theorem of calculus gives

$$f(y) = f(x) + \int_x^y f'(s)\,ds$$

Suppose f is concave. If $x < y$, then $0 = f'(x) \geq f'(s)$ when $s > x$, hence

$$\int_x^y f'(s)\,ds \leq 0$$

and $f(y) \leq f(x)$. If $y < x$, then $f'(s) \geq f'(x) = 0$ when $s < x$, hence

$$-\int_x^y f'(s)\,ds = \int_y^x f'(s)\,ds \geq 0$$

and again $f(y) \leq f(x)$.

Concavity (cont.)

Summarizing, if $f'(x) = 0$ and f is concave, then

$$f(y) \leq f(x), \qquad \text{whenever } y \in W$$

hence x is a global maximizer of f.

By a similar argument, if $f'(x) = 0$ and f is strictly concave, then

$$f(y) < f(x), \qquad \text{whenever } y \in W \text{ and } y \neq x$$

hence x is the unique global maximizer of f.

Concavity (cont.)

The first and second order conditions are almost the same with and without concavity. Suppose we find a point x satisfying

$$f'(x) = 0$$

This is a candidate local or global optimizer. We check the second derivative.

$$f''(x) < 0$$

implies x is a local maximizer.

$$f''(y) < 0, \qquad \text{for all } y \text{ in the domain of } f$$

implies x is a the unique global maximizer. The only difference is whether we check the second derivative only at x or at all points.

Examples (cont.)

For the binomial distribution the log likelihood

$$l_n(p) = x \log(p) + (n - x) \log(1 - p)$$

has derivatives

$$\begin{aligned} l_n'(p) &= \frac{x}{p} - \frac{n - x}{1 - p} \\ &= \frac{x - np}{p(1 - p)} \\ l_n''(p) &= -\frac{x}{p^2} - \frac{n - x}{(1 - p)^2} \end{aligned}$$

Setting $l'(p) = 0$ and solving for p gives $p = x/n$. Since $l''(p) < 0$ for all p we have strict concavity and $\widehat{p}_n = x/n$ is the unique global maximizer of the log likelihood.

Examples (cont.)

The analysis on the preceding slide doesn't work when $\widehat{p}_n = 0$ or $\widehat{p}_n = 1$ because the log likelihood and its derivatives are undefined when $p = 0$ or $p = 1$. More on this later.

Examples (cont.)

For IID normal data with known mean μ and unknown variance ν the log likelihood

$$l_n(\nu) = -\frac{n}{2} \log(\nu) - \frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2$$

has derivatives

$$l_n'(\nu) = -\frac{n}{2\nu} + \frac{1}{2\nu^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$l_n''(\nu) = \frac{n}{2\nu^2} - \frac{1}{\nu^3} \sum_{i=1}^{n} (x_i - \mu)^2$$

Examples (cont.)

Setting $l_n'(\nu) = 0$ and solving for ν we get

$$\widehat{\sigma}_n^2 = \widehat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

(recall that μ is supposed known, so this is a statistic).

Since

$$\begin{aligned} l_n''(\widehat{\nu}_n) &= \frac{n}{2\widehat{\nu}_n^2} - \frac{1}{\widehat{\nu}_n^3}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= -\frac{n}{2\widehat{\nu}_n^2} \end{aligned}$$

we can say this MLE is a local maximizer of the log likelihood.

Examples (cont.)

Since

$$l_n''(\nu) = \frac{n}{2\nu^2} - \frac{1}{\nu^3}\sum_{i=1}^{n}(x_i - \mu)^2$$
$$= \frac{n}{2\nu^2} - \frac{n\widehat{\nu}_n}{\nu^3}$$

is not negative for all data and all $\nu > 0$, we cannot say the MLE is the unique global maximizer, at least not from this analysis. More on this later.

MLE on Boundary of Parameter Space

All of this goes out the window when we consider possible maxima that occur on the boundary of the domain of a function. For a function whose domain is a one-dimensional interval, this means the endpoints of the interval.

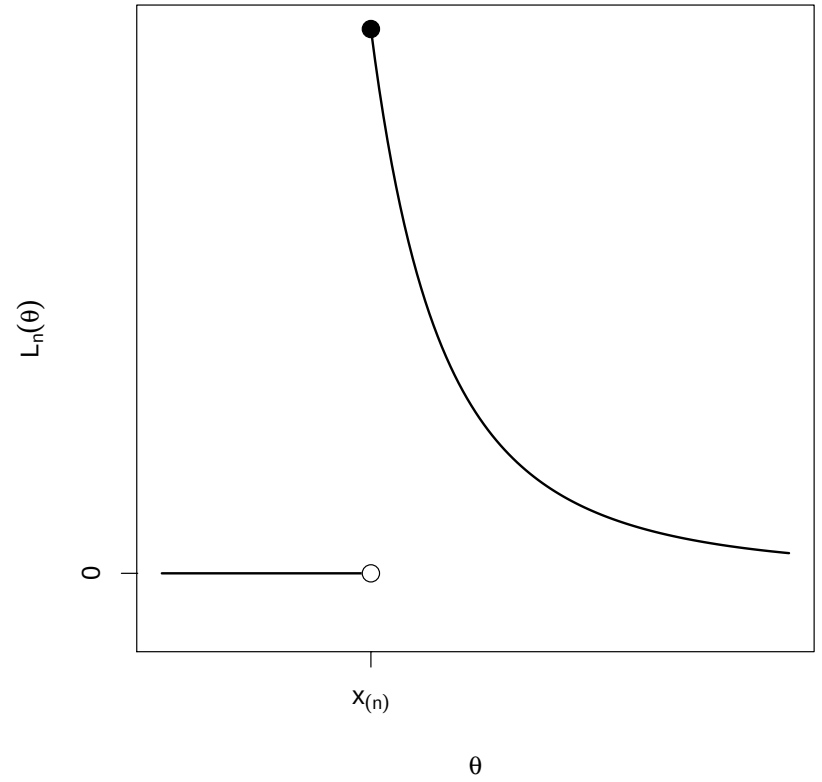
MLE on Boundary of Parameter Space (cont.)

Suppose X_1, ..., X_n are IID Unif$(0, \theta)$. The likelihood is

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} f_\theta(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(x_i) \\ &= \theta^{-n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i) \\ &= \theta^{-n} I_{[x_{(n)},\infty)}(\theta) \end{aligned}$$

The indicator functions $I_{[0,\theta]}(x_i)$ are all equal to one if and only if $x_i \le \theta$ for all i, which happens if and only if $x_{(n)} \le \theta$, a condition that is captured in the indicator function on the bottom line.

MLE on Boundary of Parameter Space (cont.)



Likelihood for Unif$(0, \theta)$ model.

MLE on Boundary of Parameter Space (cont.)

It is clear from the picture that the unique global maximizer of the likelihood is

$$\widehat{\theta}_n = x_{(n)}$$

the n-th order statistic, which is the largest data value.

For those who want more math, it is often easier to work with the likelihood rather than log likelihood when the MLE is on the boundary. It is clear from the picture that $\theta \mapsto \theta^{-n}$ is a decreasing function, hence the maximum must occur at the lower end of the range of validity of this formula, which is at $\theta = x_{(n)}$.

MLE on Boundary of Parameter Space (cont.)

If one doesn't want to use the picture at all,

$$L_n'(\theta) = -n\theta^{-(n+1)}, \qquad \theta > x_{(n)}$$

shows the derivative of L_n is negative, hence L_n is a decreasing function when $\theta > x_{(n)}$, which is the interesting part of the domain.

MLE on Boundary of Parameter Space (cont.)

Because of the way we defined the likelihood at $\theta = x_{(n)}$, the maximum is achieved. This came from the way we defined the PDF

$$f_\theta(x) = \frac{1}{\theta} I_{[0,\theta]}(x)$$

Recall that the definition of a PDF at particular points is arbitrary. In particular, we could have defined it arbitrarily at 0 and θ. We chose the definition we did so that the value of the likelihood function at the discontinuity, which is at $\theta = x_{(n)}$, is the upper value as indicated by the solid and hollow dots in the picture of the likelihood function.

MLE on Boundary of Parameter Space (cont.)

For the binomial distribution, there were two cases we did not do: $x = 0$ and $x = n$. If $\widehat{p}_n = x/n$ is also the correct MLE for them, then the MLE is on the boundary.

Again we use the likelihood

$$L_n(p) = p^x(1-p)^{n-x}, \qquad 0 < p < 1$$

In case $x = 0$, this becomes

$$L_n(p) = (1-p)^n, \qquad 0 \le p \le 1$$

Now that we no longer have to worry about 0^0 being undefined, we can extend the domain to $0 \le p \le 1$. It is easy to check that L_n is a decreasing function: draw the graph or check that $L_n'(p) < 0$ for $0 < p < 1$. Hence the unique global maximum occurs at $p = 0$. The case $x = n$ is similar. In all cases $\widehat{p} = x/n$.

Usual Asymptotics of MLE

The method of maximum likelihood estimation is remarkable in that we can determine the asymptotic distribution of estimators that are defined only implicitly — the maximizer of the log likelihood — and perhaps can only be calculated by computer optimization. In case we do have an explicit expression of the MLE, the asymptotic distribution we now derive must agree with the one calculated via the delta method, but is easier to calculate.

Asymptotics for Log Likelihood Derivatives

Consider the identity

$$\int f_\theta(x)\,dx = 1$$

or the analogous identity with summation replacing integration for the discrete case. We assume we can differentiate with respect to θ under the integral sign

$$\frac{d}{d\theta}\int f_\theta(x)\,dx = \int \frac{d}{d\theta} f_\theta(x)\,dx$$

This operation is usually valid. We won't worry about precise technical conditions.

Asymptotics for Log Likelihood Derivatives (cont.)

Since the derivative of a constant is zero, we have

$$\int \frac{d}{d\theta} f_\theta(x)\, dx = 0$$

Also

$$\begin{aligned} l'(\theta) &= \frac{d}{d\theta} \log f_\theta(x) \\ &= \frac{1}{f_\theta(x)} \frac{d}{d\theta} f_\theta(x) \end{aligned}$$

Hence

$$\frac{d}{d\theta} f_\theta(x) = l'(\theta) f_\theta(x)$$

and

$$0 = \int l'(\theta) f_\theta(x)\, dx = E_\theta\{l'(\theta)\}$$

Asymptotics for Log Likelihood Derivatives (cont.)

This gives us the first log likelihood derivative identity

$$E_\theta\{l_n'(\theta)\} = 0$$

which always holds whenever differentiation under the integral sign is valid (which is usually).

Note that it is important that we write E_θ for expectation rather than E. The identity holds when the θ in $l_n'(\theta)$ and the θ in E_θ are the same.

Asymptotics for Log Likelihood Derivatives (cont.)

For our next trick we differentiate under the integral sign again

$$\int \frac{d^2}{d\theta^2} f_\theta(x)\, dx = 0$$

Also

$$\begin{aligned} l''(\theta) &= \frac{d}{d\theta} \frac{1}{f_\theta(x)} \frac{d}{d\theta} f_\theta(x) \\ &= \frac{1}{f_\theta(x)} \frac{d^2}{d\theta^2} f_\theta(x) - \frac{1}{f_\theta(x)^2} \left(\frac{d}{d\theta} f_\theta(x) \right)^2 \end{aligned}$$

Hence

$$\frac{d^2}{d\theta^2} f_\theta(x) = l''(\theta) f_\theta(x) + l'(\theta)^2 f_\theta(x)$$

and

$$0 = \int l''(\theta) f_\theta(x)\, dx + \int l'(\theta)^2 f_\theta(x)\, dx = E_\theta\{l''(\theta)\} + E_\theta\{l'(\theta)^2\}$$

Asymptotics for Log Likelihood Derivatives (cont.)

This gives us the second log likelihood derivative identity

$$\text{var}_\theta\{l_n'(\theta)\} = -E_\theta\{l_n''(\theta)\}$$

which always holds whenever differentiation under the integral sign is valid (which is usually). The reason why

$$\text{var}_\theta\{l_n'(\theta)\} = E_\theta\{l_n'(\theta)^2\}$$

is the first log likelihood derivative identity $E_\theta\{l_n'(\theta)\} = 0$.

Note that it is again important that we write E_θ for expectation and var_θ for variance rather then E and var.

Asymptotics for Log Likelihood Derivatives (cont.)

Summary:

$$E_\theta\{l_n'(\theta)\} = 0$$
$$\text{var}_\theta\{l_n'(\theta)\} = -E_\theta\{l_n''(\theta)\}$$

These hold whether the data is discrete or continuous (for discrete data just replace integrals by sums in the preceding proofs).

Fisher Information

Either side of the second log likelihood derivative identity is called *Fisher information*

$$\begin{aligned} I_n(\theta) &= \mathrm{var}_\theta\{l_n'(\theta)\} \\ &= -E_\theta\{l_n''(\theta)\} \end{aligned}$$

Asymptotics for Log Likelihood Derivatives (cont.)

When the data are IID, then the log likelihood and its derivatives are the sum of IID terms

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

$$l_n'(\theta) = \sum_{i=1}^{n} \frac{d}{d\theta} \log f_\theta(x_i)$$

$$l_n''(\theta) = \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f_\theta(x_i)$$

From either of the last two equations we see that

$$I_n(\theta) = nI_1(\theta)$$

Asymptotics for Log Likelihood Derivatives (cont.)

If we divide either of the equations for likelihood derivatives on the preceding overhead by n, the sums become averages of IID random variables

$$n^{-1} l_n'(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\theta} \log f_\theta(x_i)$$

$$n^{-1} l_n''(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f_\theta(x_i)$$

Hence the LLN and CLT apply to them.

Asymptotics for Log Likelihood Derivatives (cont.)

To apply the LLN we need to know the expectation of the individual terms

$$\begin{aligned} E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(x_i) \right\} &= E_\theta\{l_1'(\theta)\} \\ &= 0 \\ E_\theta \left\{ \frac{d^2}{d\theta^2} \log f_\theta(x_i) \right\} &= E_\theta\{l_1''(\theta)\} \\ &= -I_1(\theta) \end{aligned}$$

Asymptotics for Log Likelihood Derivatives (cont.)

Hence the LLN applied to log likelihood derivatives says

$$n^{-1} l_n'(\theta) \xrightarrow{P} 0$$
$$n^{-1} l_n''(\theta) \xrightarrow{P} -I_1(\theta)$$

It is assumed here that θ is the true unknown parameter value, that is, X_1, X_2, ... are IID with PDF or PMF f_θ.

Asymptotics for Log Likelihood Derivatives (cont.)

To apply the CLT we need to know the mean and variance of the individual terms

$$\begin{aligned} E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(x_i) \right\} &= E_\theta\{l_1'(\theta)\} \\ &= 0 \\ \text{var}_\theta \left\{ \frac{d}{d\theta} \log f_\theta(x_i) \right\} &= \text{var}_\theta\{l_1'(\theta)\} \\ &= I_1(\theta) \end{aligned}$$

We don't know the variance of $l_n''(\theta)$ so we don't obtain a CLT for it.

Asymptotics for Log Likelihood Derivatives (cont.)

Hence the CLT applied to log likelihood first derivative says

$$\sqrt{n}\Big(n^{-1} l_n'(\theta) - 0\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, I_1(\theta)\Big)$$

or (cleaning this up a bit)

$$n^{-1/2} l_n'(\theta) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, I_1(\theta)\Big)$$

It is assumed here that θ is the true unknown parameter value, that is, X_1, X_2, ... are IID with PDF or PMF f_θ.

Asymptotics for MLE

The MLE $\widehat{\theta}_n$ satisfies

$$l_n'(\widehat{\theta}_n) = 0$$

because the MLE is a local maximizer (at least) of the log likelihood.

Expand the first derivative of the log likelihood in a Taylor series about the true unknown parameter value, which we now start calling θ_0

$$l_n'(\theta) = l_n'(\theta_0) + l_n''(\theta_0)(\theta - \theta_0) + \text{higher order terms}$$

Asymptotics for MLE (cont.)

We rewrite this

$$n^{-1/2} l_n'(\theta) = n^{-1/2} l_n'(\theta_0) + n^{-1} l_n''(\theta_0) n^{1/2}(\theta - \theta_0) + \text{higher order terms}$$

because we know the asymptotics of $n^{-1/2} l_n'(\theta_0)$ and $n^{-1} l_n''(\theta_0)$.

Then we assume the higher order terms are negligible when $\widehat{\theta}_n$ is plugged in for θ

$$0 = n^{-1/2} l_n'(\theta_0) + n^{-1} l_n''(\theta_0) n^{1/2}(\widehat{\theta}_n - \theta_0) + o_p(1)$$

Asymptotics for MLE (cont.)

This implies

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{n^{-1/2} l_n'(\theta_0)}{n^{-1} l_n''(\theta_0)} + o_p(1)$$

and by Slutsky's theorem

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} -\frac{Y}{I_1(\theta_0)}$$

where

$$Y \sim \mathcal{N}\Big(0, I_1(\theta_0)\Big)$$

Asymptotics for MLE (cont.)

Since

$$E\left\{-\frac{Y}{I_1(\theta_0)}\right\} = 0$$

$$\begin{aligned}\text{var}\left\{-\frac{Y}{I_1(\theta_0)}\right\} &= \frac{I_1(\theta_0)}{I_1(\theta_0)^2} \\ &= I_1(\theta_0)^{-1}\end{aligned}$$

we get

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I_1(\theta_0)^{-1}\right)$$

Asymptotics for MLE (cont.)

It is now safe to get sloppy

$$\widehat{\theta}_n \approx \mathcal{N}\Big(\theta_0, n^{-1} I_1(\theta_0)^{-1}\Big)$$

or

$$\widehat{\theta}_n \approx \mathcal{N}\Big(\theta_0, I_n(\theta_0)^{-1}\Big)$$

This is a remarkable result. Without knowing anything about the functional form of the MLE, we have derived its asymptotic distribution.

Examples (cont.)

We already know the asymptotic distribution for the MLE of the binomial distribution, because it follows directly from the CLT (5101, deck 7, slide 36)

$$\sqrt{n}(\widehat{p}_n - p) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, p(1-p)\Big)$$

but let us calculate this using likelihood theory

$$\begin{aligned} I_n(p) &= -E_p\left\{l_n''(p)\right\} \\ &= -E_p\left\{-\frac{X}{p^2} - \frac{n-X}{(1-p)^2}\right\} \\ &= \frac{np}{p^2} + \frac{n-np}{(1-p)^2} \\ &= \frac{n}{p(1-p)} \end{aligned}$$

(the formula for $l_n''(p)$ is from slide 20)

Examples (cont.)

Hence

$$I_n(p)^{-1} = \frac{p(1-p)}{n}$$

and

$$\widehat{p}_n \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

as we already knew.

Examples (cont.)

For the IID normal data with known mean μ and unknown variance ν

$$\begin{aligned} I_n(\nu) &= -E_\nu \left\{ l_n''(\nu) \right\} \\ &= -E_\nu \left\{ \frac{n}{2\nu^2} - \frac{1}{\nu^3} \sum_{i=1}^{n} (X_i - \mu)^2 \right\} \\ &= -\frac{n}{2\nu^2} + \frac{1}{\nu^3} \cdot n\nu \\ &= \frac{n}{2\nu^2} \end{aligned}$$

(the formula for $l_n''(\nu)$ is from slide 22). Hence

$$\widehat{\nu}_n \approx \mathcal{N}\left(\nu, \frac{2\nu^2}{n} \right)$$

Examples (cont.)

Or for IID normal data

$$\hat{\sigma}_n^2 \approx \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n}\right)$$

because $\nu = \sigma^2$.

We already knew this from homework problem 4-9.

Examples (cont.)

Here's an example we don't know. Suppose X_1, X_2, ... are IID Gam(α, λ) where α is unknown and λ is known. Then

$$\begin{aligned} L_n(\alpha) &= \prod_{i=1}^{n} \frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \\ &= \left(\frac{\lambda^\alpha}{\Gamma(\alpha)} \right)^n \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\lambda x_i} \\ &= \left(\frac{\lambda^\alpha}{\Gamma(\alpha)} \right)^n \left(\prod_{i=1}^{n} x_i \right)^{\alpha-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i \right) \end{aligned}$$

and we can drop the term that does not contain α.

Examples (cont.)

The log likelihood is

$$l_n(\alpha) = n\alpha \log \lambda - n \log \Gamma(\alpha) + (\alpha - 1) \log \left(\prod_{i=1}^{n} x_i \right)$$

Every term except $n \log \Gamma(\alpha)$ is linear in α and hence has second derivative with respect to α equal to zero. Hence

$$l_n''(\alpha) = -n \frac{d^2}{d\alpha^2} \log \Gamma(\alpha)$$

and

$$I_n(\alpha) = n \frac{d^2}{d\alpha^2} \log \Gamma(\alpha)$$

because the expectation of a constant is a constant.

Examples (cont.)

The second derivative of the logarithm of the gamma function is not something we know how to do, but is a "brand name function" called the *trigamma function*, which can be calculated by R or Mathematica.

Again we say, this is a remarkable result. We have no closed form expression for the MLE, but we know its asymptotic distribution is

$$\widehat{\alpha}_n \approx \mathcal{N}\left(\alpha, \frac{1}{n\,\mathsf{trigamma}(\alpha)}\right)$$

Plug-In for Asymptotic Variance

Since we do not know the true unknown parameter value θ_0, we do not know the Fisher information $I_1(\theta_0)$ either. In order to use the asymptotics of MLE for confidence intervals and hypothesis tests, we need plug in. If $\theta \mapsto I_1(\theta)$ is a continuous function, then

$$I_1(\hat{\theta}_n) \xrightarrow{P} I_1(\theta_0)$$

by the continuous mapping theorem. Hence by the plug-in principle (Slutsky's theorem)

$$\sqrt{n} \cdot \frac{\hat{\theta}_n - \theta_0}{I_1(\hat{\theta}_n)^{-1/2}} = (\hat{\theta}_n - \theta_0) I_n(\hat{\theta}_n)^{1/2} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

is an asymptotically pivotal quantity that can be used to construct confidence intervals and hypothesis tests.

Plug-In for Asymptotic Variance (cont.)

If z_α is the $1 - \alpha$ quantile of the standard normal distribution, then

$$\widehat{\theta}_n \pm z_{\alpha/2} I_n(\widehat{\theta}_n)^{-1/2}$$

is an asymptotic $100(1 - \alpha)\%$ confidence interval for θ.

Plug-In for Asymptotic Variance (cont.)

The test statistic

$$T = (\hat{\theta}_n - \theta_0) I_n(\hat{\theta}_n)^{1/2}$$

is asymptotically standard normal under the null hypothesis

$$H_0 : \theta = \theta_0$$

As usual, the approximate P-values for upper-tail, lower-tail, and two-tail tests are, respectively,

$$\begin{aligned}
\Pr_{\theta_0}(T \geq t) &\approx 1 - \Phi(t) \\
\Pr_{\theta_0}(T \leq t) &\approx \Phi(t) \\
\Pr_{\theta_0}(|T| \geq |t|) &\approx 2\Big(1 - \Phi(|t|)\Big) = 2\Phi(-|t|)
\end{aligned}$$

where Φ is the DF of the standard normal distribution.

Plug-In for Asymptotic Variance (cont.)

Sometimes the expectation involved in calculating Fisher information is too hard to do. Then we use the following idea. The LLN for the second derivative of the log likelihood (slide 43) says

$$n^{-1} l_n''(\theta_0) \xrightarrow{P} -I_1(\theta_0)$$

which motivates the following definition:

$$J_n(\theta) = -l_n''(\theta)$$

is called *observed Fisher information*. For contrast $I_n(\theta)$ or $I_1(\theta)$ is called *expected Fisher information*, although, strictly speaking, the "expected" is unnecessary.

Plug-In for Asymptotic Variance (cont.)

The LLN for the second derivative of the log likelihood can be written sloppily

$$J_n(\theta) \approx I_n(\theta)$$

from which

$$J_n(\hat{\theta}_n) \approx I_n(\hat{\theta}_n) \approx I_n(\theta_0)$$

should also hold, and usually does (although this requires more than just the continuous mapping theorem so we don't give a proof).

Plug-In for Asymptotic Variance (cont.)

This gives us two asymptotic $100(1 - \alpha)\%$ confidence intervals for θ

$$\widehat{\theta}_n \pm z_{\alpha/2} I_n(\widehat{\theta}_n)^{-1/2}$$
$$\widehat{\theta}_n \pm z_{\alpha/2} J_n(\widehat{\theta}_n)^{-1/2}$$

and the latter does not require any expectations. If we can write down the log likelihood and differentiate it twice, then we can make the latter confidence interval.

Plug-In for Asymptotic Variance (cont.)

Similarly, we have two test statistics

$$T = (\hat{\theta}_n - \theta_0) I_n(\hat{\theta}_n)^{1/2}$$
$$T = (\hat{\theta}_n - \theta_0) J_n(\hat{\theta}_n)^{1/2}$$

which are asymptotically standard normal under the null hypothesis

$$H_0 : \theta = \theta_0$$

and can be used to perform hypothesis tests (as described on slide 60).

Again, if we can write down the log likelihood and differentiate it twice, then we can perform the test using latter test statistic.

Plug-In for Asymptotic Variance (cont.)

Sometimes even differentiating the log likelihood is too hard to do. Then we use the following idea. Derivatives can be approximated by "finite differences"

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}, \qquad \text{when } h \text{ is small}$$

When derivatives are too hard to do by calculus, they can be approximated by finite differences.

Plug-In for Asymptotic Variance (cont.)

The R code on the computer examples web page about maximum likelihood for the gamma distribution with shape parameter α unknown and rate parameter $\lambda = 1$ known,

```
Rweb> n <- length(x)
Rweb> mlogl <- function(a) sum(- dgamma(x, a, log = TRUE))
Rweb> out <- nlm(mlogl, mean(x), hessian = TRUE, fscale = n)
Rweb> ahat <- out$estimate
Rweb> z <- qnorm(0.975)
Rweb> ahat + c(-1, 1) * z / sqrt(n * trigamma(ahat))
[1] 1.271824 2.065787
Rweb> ahat + c(-1, 1) * z / sqrt(out$hessian)
[1] 1.271798 2.065813
```

The Information Inequality

Suppose $\widehat{\theta}_n$ is any unbiased estimator of θ. Then

$$\text{var}_\theta(\widehat{\theta}_n) \geq I_n(\theta)^{-1}$$

which is called the *information inequality* or the *Cramér-Rao lower bound*.

Proof:

$$\begin{aligned} \text{cov}_\theta\{\widehat{\theta}, l'(\theta)\} &= E_\theta\{\widehat{\theta} \cdot l'(\theta)\} - E_\theta(\widehat{\theta}) E_\theta\{l'(\theta)\} \\ &= E_\theta\{\widehat{\theta} \cdot l'(\theta)\} \end{aligned}$$

because of the first log likelihood derivative identity.

The Information Inequality (cont.)

And

$$\begin{aligned} E_\theta\{\widehat{\theta} \cdot l'(\theta)\} &= \int \widehat{\theta}(x) \left[\frac{1}{f_\theta(x)} \cdot \frac{d}{d\theta} f_\theta(x) \right] f_\theta(x)\, dx \\ &= \int \widehat{\theta}(x) \left[\frac{d}{d\theta} f_\theta(x) \right] dx \\ &= \frac{d}{d\theta} \int \widehat{\theta}(x) f_\theta(x)\, dx \\ &= \frac{d}{d\theta} E_\theta(\widehat{\theta}) \end{aligned}$$

assuming differentiation under the integral sign is valid.

The Information Inequality (cont.)

By assumption $\widehat{\theta}$ is unbiased, which means

$$E_\theta(\widehat{\theta}) = \theta$$

and

$$\frac{d}{d\theta} E_\theta(\widehat{\theta}) = 1$$

Hence

$$\text{cov}_\theta\{\widehat{\theta}, l'(\theta)\} = 1$$

The Information Inequality (cont.)

From the correlation inequality (5101, deck 6, slide 61)

$$\begin{aligned} 1 &\geq \mathrm{cor}_\theta\{\hat{\theta}, l'(\theta)\}^2 \\ &= \frac{\mathrm{cov}_\theta\{\hat{\theta}, l'(\theta)\}^2}{\mathrm{var}_\theta(\hat{\theta})\,\mathrm{var}_\theta\{l'(\theta)\}} \\ &= \frac{1}{\mathrm{var}_\theta(\hat{\theta})I(\theta)} \end{aligned}$$

from which the information inequality follows immediately.

The Information Inequality (cont.)

The information inequality says no unbiased estimator can be more efficient than the MLE. But what about biased estimators? They can be more efficient.

An estimator that is better than the MLE in the ARE sense is called *superefficient*, and such estimators do exist.

The Hájek convolution theorem says no estimator that is asymptotically unbiased in a certain sense can be superefficient.

The Le Cam convolution theorem says no estimator can be superefficient except at a set of true unknown parameter points of measure zero.

The Information Inequality (cont.)

In summary, the MLE is as about as efficient as an estimator can be.

For exact theory, we only know that no unbiased estimator can be superefficient.

For asymptotic theory, we know that no estimator can be super-efficient except at a negligible set of true unknown parameter values.

Multiparameter Maximum Likelihood

The basic ideas are the same when there are multiple unknown parameters rather than just one. We have to generalize each topic

- conditions for local and global maxima,
- log likelihood derivative identities,
- Fisher information, and
- asymptotics and plug-in.

Multivariate Differentiation

This topic was introduced last semester (5101, deck 7, slides 96–98). Here we review and specialize to scalar-valued functions.

If W is an open region of $\mathbb{R}^p$, then $f : W \to \mathbb{R}$ is *differentiable* if all partial derivatives exist and are continuous, in which case the vector of partial derivatives evaluated at $\mathbf{x}$ is called the *gradient* vector at x and is denoted $\nabla f(\mathbf{x})$.

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \partial f(\mathbf{x})/\partial x_1 \\ \partial f(\mathbf{x})/\partial x_2 \\ \vdots \\ \partial f(\mathbf{x})/\partial x_p \end{pmatrix}$$

Multivariate Differentiation (cont.)

If W is an open region of $\mathbb{R}^p$, then $f : W \to \mathbb{R}$ is *twice differentiable* if all second partial derivatives exist and are continuous, in which case the matrix of second partial derivatives evaluated at $\mathbf{x}$ is called the *Hessian* matrix at x and is denoted $\nabla^2 f(\mathbf{x})$.

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_p \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_p^2} \end{pmatrix}$$

Local Maxima

Suppose W is an open region of $\mathbb{R}^p$ and $f : W \to \mathbb{R}$ is a twice-differentiable function. A necessary condition for a point $\mathbf{x} \in W$ to be a local maximum of f is

$$\nabla f(\mathbf{x}) = 0$$

and a sufficient condition for $\mathbf{x}$ to be a local maximum is

$$\nabla^2 f(\mathbf{x}) \text{ is a negative definite matrix}$$

Positive Definite Matrices

A symmetric matrix $\mathbf{M}$ is *positive semi-definite* (5101, deck 2, slides 68–69 and deck 5, slides 103–105) if

$$\mathbf{w}^T\mathbf{M}\mathbf{w} \geq 0, \qquad \text{for all vectors } \mathbf{w}$$

and *positive definite* if

$$\mathbf{w}^T\mathbf{M}\mathbf{w} > 0, \qquad \text{for all nonzero vectors } \mathbf{w}.$$

A symmetric matrix $\mathbf{M}$ is *negative semi-definite* if $-\mathbf{M}$ is positive semidefinite, and $\mathbf{M}$ is *negative definite* $-\mathbf{M}$ is positive definite.

Positive Definite Matrices (cont.)

There are two ways to check that the Hessian matrix is negative semi-definite.

First, one can try to verify that

$$\sum_{i=1}^{p} \sum_{j=1}^{p} w_i w_j \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} < 0$$

holds for all real numbers w_1, ..., w_p, at least one of which is nonzero (5101, deck 2, slides 68–69). This is hard.

Second, one can verify that all the eigenvalues are negative (5101, deck 5, slides 103–105). This can be done by computer, but can only be applied to a numerical matrix that has particular values plugged in for all variables and parameters.

Local Maxima (cont.)

The first-order condition for a local maximum is not much harder than before. Set all first partial derivatives to zero and solve for the variables.

The second-order condition is harder when done by hand. The computer check that all eigenvalues are negative is easy.

Examples (cont.)

The log likelihood for the two-parameter normal model is

$$l_n(\mu, \nu) = -\frac{n}{2} \log(\nu) - \frac{n v_n}{2\nu} - \frac{n(\bar{x}_n - \mu)^2}{2\nu}$$

(slide 10). The first partial derivatives are

$$\frac{\partial l_n(\mu, \nu)}{\partial \mu} = \frac{n(\bar{x}_n - \mu)}{\nu}$$

$$\frac{\partial l_n(\mu, \nu)}{\partial \nu} = -\frac{n}{2\nu} + \frac{n v_n}{2\nu^2} + \frac{n(\bar{x}_n - \mu)^2}{2\nu^2}$$

Examples (cont.)

The second partial derivatives are

$$\frac{\partial^2 l_n(\mu,\nu)}{\partial \mu^2} = -\frac{n}{\nu}$$
$$\frac{\partial^2 l_n(\mu,\nu)}{\partial \mu \partial \nu} = -\frac{n(\bar{x}_n - \mu)}{\nu^2}$$
$$\frac{\partial^2 l_n(\mu,\nu)}{\partial \nu^2} = +\frac{n}{2\nu^2} - \frac{n v_n}{\nu^3} - \frac{n(\bar{x}_n - \mu)^2}{\nu^3}$$

Examples (cont.)

Setting the first partial derivative with respect to μ equal to zero and solving for μ gives

$$\mu = \bar{x}_n$$

Plugging that into the first partial derivative with respect to ν set equal to zero gives

$$-\frac{n}{2\nu} + \frac{nv_n}{2\nu^2} = 0$$

and solving for ν gives

$$\nu = v_n$$

Examples (cont.)

Thus the MLE for the two-parameter normal model are

$$\widehat{\mu}_n = \bar{x}_n$$
$$\widehat{\nu}_n = v_n$$

and we can also denote the latter

$$\widehat{\sigma}_n^2 = v_n$$

Examples (cont.)

Plugging the MLE into the second partial derivatives gives

$$\begin{aligned}
\frac{\partial^2 l_n(\widehat{\mu}_n, \widehat{\nu}_n)}{\partial \mu^2} &= -\frac{n}{\widehat{\nu}_n} \\
\frac{\partial^2 l_n(\widehat{\mu}_n, \widehat{\nu}_n)}{\partial \mu \partial \nu} &= 0 \\
\frac{\partial^2 l_n(\widehat{\mu}_n, \widehat{\nu}_n)}{\partial \nu^2} &= +\frac{n}{2\widehat{\nu}_n^2} - \frac{n v_n}{\widehat{\nu}_n^3} \\
&= -\frac{n}{2\widehat{\nu}_n^2}
\end{aligned}$$

Hence the Hessian matrix is diagonal, and is negative definite if each of the diagonal terms is negative (5101, deck 5, slide 106), which they are. Thus the MLE is a local maximizer of the log likelihood.

Global Maxima

A region W of $\mathbb{R}^p$ is *convex* if

$s\mathbf{x}+(1-s)\mathbf{y} \in W,$ whenever $\mathbf{x} \in W$ and $\mathbf{y} \in W$ and $0 < s < 1$

Suppose W is an open convex region of $\mathbb{R}^p$ and $f : W \to \mathbb{R}$ is a twice-differentiable function. If

$$\nabla^2 f(\mathbf{y}) \text{ is a negative definite matrix for all } \mathbf{y} \in W,$$

then f is called *strictly concave*. In this case

$$\nabla f(\mathbf{x}) = 0$$

is a sufficient condition for $\mathbf{x}$ to be the unique global maximum.

Log Likelihood Derivative Identities

The same differentiation under the integral sign argument applied to partial derivatives results in

$$E_{\boldsymbol{\theta}}\left\{\frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_i}\right\} = 0$$

$$E_{\boldsymbol{\theta}}\left\{\frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_i} \cdot \frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_j}\right\} = -E_{\boldsymbol{\theta}}\left\{\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right\}$$

which can be rewritten in matrix notation as

$$E_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} = 0$$

$$\text{var}_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} = -E_{\boldsymbol{\theta}}\{\nabla^2 l_n(\boldsymbol{\theta})\}$$

(compare with slide 38).

Fisher Information

As in the uniparameter case, either side of the second log likelihood derivative identity is called *Fisher information*

$$\begin{aligned} \mathbf{I}_n(\boldsymbol{\theta}) &= \text{var}_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} \\ &= -E_{\boldsymbol{\theta}}\{\nabla^2 l_n(\boldsymbol{\theta})\} \end{aligned}$$

Being a variance matrix, the Fisher information matrix is symmetric and positive semi-definite.

Usually the Fisher information matrix is actually positive definite, and we will always assume this.

Examples (cont.)

Returning to the two-parameter normal model, and taking expectations of the second partial derivatives gives

$$
\begin{aligned}
E_{\mu,\nu}\left\{\frac{\partial^2 l_n(\mu,\nu)}{\partial\mu^2}\right\} &= -\frac{n}{\nu} \\
E_{\mu,\nu}\left\{\frac{\partial^2 l_n(\mu,\nu)}{\partial\mu\partial\nu}\right\} &= -\frac{nE_{\mu,\nu}(\overline{X}_n-\mu)}{\nu^2} \\
&= 0 \\
E_{\mu,\nu}\left\{\frac{\partial^2 l_n(\mu,\nu)}{\partial\nu^2}\right\} &= +\frac{n}{2\nu^2} - \frac{nE_{\mu,\nu}(V_n)}{\nu^3} - \frac{nE_{\mu,\nu}\{(\overline{X}_n-\mu)^2\}}{\nu^3} \\
&= +\frac{n}{2\nu^2} - \frac{(n-1)E_{\mu,\nu}(S_n^2)}{\nu^3} - \frac{n\,\mathrm{var}_{\mu,\nu}(\overline{X}_n)}{\nu^3} \\
&= -\frac{n}{2\nu^2}
\end{aligned}
$$

Examples (cont.)

Hence for the two-parameter normal model the Fisher information matrix is

$$\mathbf{I}_n(\boldsymbol{\theta}) = \begin{pmatrix} n/\nu & 0 \\ 0 & n/(2\nu^2) \end{pmatrix}$$

Asymptotics for Log Likelihood Derivatives (cont.)

The same CLT argument applied to the gradient vector gives

$$n^{-1/2} \nabla l_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \mathbf{I}_1(\boldsymbol{\theta}_0)\Big)$$

and the same LLN argument applied to the Hessian matrix gives

$$-n^{-1} \nabla^2 l_n(\boldsymbol{\theta}_0) \xrightarrow{P} \mathbf{I}_1(\boldsymbol{\theta}_0)$$

These are multivariate convergence in distribution and multivariate convergence in probability statements (5101, deck 7, slides 73–78 and 79–85).

Asymptotics for MLE (cont.)

The same argument — expand the gradient of the log likelihood in a Taylor series, assume terms after the first two are negligible, and apply Slutsky — used in the univariate case gives for the asymptotics of the MLE

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}\Big)$$

or the sloppy version

$$\hat{\boldsymbol{\theta}}_n \approx \mathcal{N}\Big(\boldsymbol{\theta}_0, \mathbf{I}_n(\boldsymbol{\theta}_0)^{-1}\Big)$$

(compare slides 46–50). Since Fisher information is a matrix, $\mathbf{I}_n(\boldsymbol{\theta}_0)^{-1}$ must be a matrix inverse.

Examples (cont.)

Returning to the two-parameter normal model, inverse Fisher information is

$$\mathbf{I}_n(\boldsymbol{\theta})^{-1} = \begin{pmatrix} \nu/n & 0 \\ 0 & 2\nu^2/n \end{pmatrix} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}$$

Because the asymptotic covariance is zero, the two components of the MLE are asymptotically independent (actually we know they are exactly, not just asymptotically independent, deck 1, slide 58 ff.) and their asymptotic distributions are

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$V_n \approx \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n}\right)$$

Examples (cont.)

We already knew these asymptotic distributions, the former being the CLT and the latter being homework problem 4-9.

Examples (cont.)

Now for something we didn't already know. Taking logs in the formula for the likelihood of the gamma distribution (slide 55) gives

$$\begin{aligned} l_n(\alpha, \lambda) &= n\alpha \log \lambda - n \log \Gamma(\alpha) + (\alpha - 1) \log \left(\prod_{i=1}^{n} x_i \right) - \lambda \sum_{i=1}^{n} x_i \\ &= n\alpha \log \lambda - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i) - \lambda \sum_{i=1}^{n} x_i \\ &= n\alpha \log \lambda - n \log \Gamma(\alpha) + n(\alpha - 1)\bar{y}_n - n\lambda \bar{x}_n \end{aligned}$$

where

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} \log(x_i)$$

Examples (cont.)

$$\frac{\partial l_n(\alpha,\lambda)}{\partial \alpha} = n \log \lambda - n\,\text{digamma}(\alpha) + n\bar{y}_n$$
$$\frac{\partial l_n(\alpha,\lambda)}{\partial \lambda} = \frac{n\alpha}{\lambda} - n\bar{x}_n$$
$$\frac{\partial^2 l_n(\alpha,\lambda)}{\partial \alpha^2} = -n\,\text{trigamma}(\alpha)$$
$$\frac{\partial^2 l_n(\alpha,\lambda)}{\partial \alpha \partial \lambda} = \frac{n}{\lambda}$$
$$\frac{\partial^2 l_n(\alpha,\lambda)}{\partial \lambda^2} = -\frac{n\alpha}{\lambda^2}$$

Examples (cont.)

If we set first partial derivatives equal to zero and solve for the parameters, we find we cannot. The MLE can only be found by the computer, maximizing the log likelihood for particular data.

We do, however, know the asymptotic distribution of the MLE

$$\begin{pmatrix} \widehat{\alpha}_n \\ \widehat{\lambda}_n \end{pmatrix} \approx \mathcal{N}\left(\begin{pmatrix} \alpha \\ \lambda \end{pmatrix}, \mathbf{I}_n(\boldsymbol{\theta})^{-1} \right)$$

where

$$\mathbf{I}_n(\boldsymbol{\theta}) = \begin{pmatrix} n\,\text{trigamma}(\alpha) & -n/\lambda \\ -n/\lambda & n\alpha/\lambda^2 \end{pmatrix}$$

Plug-In for Asymptotic Variance (cont.)

As always, since we don't know $\boldsymbol{\theta}$, we must use a plug-in estimate for asymptotic variance. As in the uniparameter case, we can use either expected Fisher information.

$$\hat{\boldsymbol{\theta}}_n \approx \mathcal{N}\left(\boldsymbol{\theta}, \mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\right)$$

or observed Fisher information.

$$\hat{\boldsymbol{\theta}}_n \approx \mathcal{N}\left(\boldsymbol{\theta}, \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\right)$$

where

$$\mathbf{J}_n(\boldsymbol{\theta}) = -\nabla^2 l_n(\boldsymbol{\theta})$$

Caution

There is a big difference between the Right Thing (standard errors for MLE are square roots of diagonal elements of the **inverse** Fisher information matrix) and the Wrong Thing (inverses of square roots of diagonal elements of the Fisher information matrix)

```
Rweb:> fish
           [,1]       [,2]
[1,]  24.15495 -29.71683
[2,] -29.71683  49.46866
Rweb:> 1 / sqrt(diag(fish)) # Wrong Thing
[1] 0.2034684 0.1421788
Rweb:> sqrt(diag(solve(fish))) # Right Thing
[1] 0.3983007 0.2783229
```

Starting Points for Optimization

When a maximum likelihood problem is not concave, there can be more than one local maximum. Theory says one of those local maxima is the efficient estimator which has inverse Fisher information for its asymptotic variance. The rest of the local maxima are no good.

How to find the right one? Theory says that if the starting point for optimization is a "root n consistent" estimator, that is, $\tilde{\boldsymbol{\theta}}_n$ such that

$$\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + O_p(n^{-1/2})$$

and any CAN estimator satisfies this, for example, method of moments estimators and sample quantiles.

Invariance of Maximum Likelihood

If $\boldsymbol{\psi} = g(\boldsymbol{\theta})$ is an invertible change-of-parameter, and $\hat{\boldsymbol{\theta}}_n$ is the MLE for $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\psi}}_n = g(\hat{\boldsymbol{\theta}}_n)$ is the MLE for $\boldsymbol{\psi}$.

This is obvious if one thinks of $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ as locations in different coordinate systems for the same geometric object, which denotes a probability distribution. The likelihood function, while defined as a function of the parameter, clearly only depends on the distribution the parameter indicates. Hence this invariance.

Invariance of Maximum Likelihood (cont.)

This invariance does not extend to derivatives of the log likelihood.

If $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix for $\boldsymbol{\theta}$ and $\widetilde{\mathbf{I}}(\boldsymbol{\psi})$ is the Fisher information matrix for $\boldsymbol{\psi}$, then the chain rule and log likelihood derivative identities give

$$\mathbf{I}_n(\boldsymbol{\theta}) = \Big[\nabla g(\boldsymbol{\theta})\Big]^T \Big[\widetilde{\mathbf{I}}_n(g(\boldsymbol{\theta}))\Big]\Big[\nabla g(\boldsymbol{\theta})\Big]$$

Invariance of Maximum Likelihood (cont.)

We only do the one-parameter case.

$$
\begin{aligned}
l_n(\theta) &= \tilde{l}_n(g(\theta)) \\
l_n'(\theta) &= \tilde{l}_n'(g(\theta))g'(\theta) \\
l_n''(\theta) &= \tilde{l}_n''(g(\theta))g'(\theta)^2 + \tilde{l}_n'(g(\theta))g''(\theta)
\end{aligned}
$$

Taking expectations, the second term in the second derivative is zero by the first log likelihood derivative identity. This leaves the one-parameter case of what was to be proved.

Non-Existence of Global Maxima

The web page about maximum likelihood does a normal mixture model. The data are IID with PDF

$$f(x) = p\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-p)\phi\left(\frac{x-\mu_2}{\sigma_2}\right)$$

where ϕ is the standard normal PDF. Hence the log likelihood is

$$\begin{aligned} l_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log\left[p\phi\left(\frac{x_i-\mu_1}{\sigma_1}\right) + (1-p)\phi\left(\frac{x_i-\mu_2}{\sigma_2}\right)\right] \\ &= \sum_{i=1}^{n} \log\left[\frac{p}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right)\right. \\ &\qquad \left. + \frac{1-p}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right)\right] \end{aligned}$$

Non-Existence of Global Maxima (cont.)

If we set $\mu_1 = x_i$ for some i, then the i-th term of the log likelihood becomes

$$\log\left[\frac{p}{\sqrt{2\pi}\sigma_1} + \frac{1-p}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right)\right]$$

and this goes to infinity as $\sigma_1 \to 0$. Hence the supremum of the log likelihood is $+\infty$ and no values of the parameters achieve the supremum.

Nevertheless, the good local maximizer is the efficient estimator.

Exponential Families of Distributions

A statistical model is called an *exponential family* if the log likelihood has the form

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{p} t_i(\mathbf{x}) g_i(\boldsymbol{\theta}) - h(\boldsymbol{\theta}) + u(\mathbf{x})$$

and the last term can, of course, be dropped.

The only term in the log likelihood that contains both statistics and parameters is a finite sum of terms that are a product of a function of the data times a function of the parameter

Exponential Families of Distributions (cont.)

Introduce new statistics and parameters

$$y_i = t_i(\mathbf{x})$$
$$\psi_i = g_i(\boldsymbol{\theta})$$

which are components of the *natural statistic* vector $\mathbf{y}$ and the *natural parameter* vector $\boldsymbol{\psi}$.

That $\boldsymbol{\psi}$ is actually a parameter is shown by the fact that the PDF or PMF must integrate or sum to one for all $\boldsymbol{\theta}$. Hence $h(\boldsymbol{\theta})$ must actually be a function of $\boldsymbol{\psi}$.

Exponential Families of Distributions (cont.)

The log likelihood in terms of natural parameters and statistics has the simple form

$$l(\boldsymbol{\psi}) = \mathbf{y}^T \boldsymbol{\psi} - c(\boldsymbol{\psi})$$

and derivatives

$$\nabla l(\boldsymbol{\psi}) = \mathbf{y} - \nabla c(\boldsymbol{\psi})$$
$$\nabla^2 l(\boldsymbol{\psi}) = -\nabla^2 c(\boldsymbol{\psi})$$

Exponential Families of Distributions (cont.)

The log likelihood derivative identities give

$$E_{\boldsymbol{\psi}}(\mathbf{Y}) = \nabla c(\boldsymbol{\psi})$$
$$\operatorname{var}_{\boldsymbol{\psi}}(\mathbf{Y}) = \nabla^2 c(\boldsymbol{\psi})$$

Hence the MLE is a method of moments estimator that sets the observed value of the *natural statistic vector* equal to its expected value.

The second derivative of the log likelihood is always nonrandom, so observed and expected Fisher information for the *natural parameter vector* are the same, and the log likelihood for the natural parameter is always concave and strictly concave unless the distribution of the natural statistic is degenerate. Hence any local maximizer of the log likelihood is the unique global maximizer.

Exponential Families of Distributions (cont.)

By invariance of maximum likelihood, the property that any local maximizer is the unique global maximizer holds for any parametrization.

Brand name distributions that are exponential families: Bernoulli, binomial, Poisson, geometric, negative binomial (p unknown, r known), normal, exponential, gamma, beta, multinomial, multivariate normal.

Exponential Families of Distributions (cont.)

For binomial, the log likelihood is

$$\begin{aligned} l(p) &= x \log(p) + (n - x) \log(1 - p) \\ &= x \Big[\log(p) - \log(1 - p) \Big] + n \log(1 - p) \end{aligned}$$

hence is exponential family with natural statistic x and natural parameter

$$\theta = \operatorname{logit}(p) = \log(p) - \log(1 - p) = \log \left(\frac{p}{1 - p} \right)$$

Exponential Families of Distributions (cont.)

Suppose X_1, ..., X_n are IID from an exponential family distribution with log likelihood for sample size one

$$l_1(\boldsymbol{\theta}) = \sum_{i=1}^{p} t_i(\mathbf{x}) g_i(\boldsymbol{\theta}) - h(\boldsymbol{\theta})$$

Then the log likelihood for sample size n is

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{p} \left(\sum_{j=1}^{n} t_i(\mathbf{x}_j) \right) g_i(\boldsymbol{\theta}) - n h(\boldsymbol{\theta})$$

hence the distribution for sample size n is also an exponential family with the same natural parameter vector as for sample size one and natural statistic vector $\mathbf{y}$ with components

$$y_i = \sum_{j=1}^{n} t_i(\mathbf{x}_j)$$

Exponential Families of Distributions (cont.)

For the two-parameter normal, the log likelihood for sample size one is

$$\begin{aligned} l_1(\mu, \sigma^2) &= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2 \\ &= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2) \\ &= -\frac{1}{2\sigma^2}\cdot x^2 + \frac{\mu}{\sigma^2}\cdot x - \frac{1}{2}\log(\sigma^2) - \frac{\mu^2}{2\sigma^2} \end{aligned}$$

Since this is a two-parameter family, the natural parameter and statistic must also be two dimensional. We can choose

$$\mathbf{y} = \begin{pmatrix} x \\ x^2 \end{pmatrix} \qquad \boldsymbol{\theta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix}$$

Exponential Families of Distributions (cont.)

It seems weird to think of the natural statistic being two-dimensional when we usually think of the normal distribution as being one-dimensional.

But for sample size n, the natural statistics become

$$y_1 = \sum_{i=1}^{n} x_i$$
$$y_2 = \sum_{i=1}^{n} x_i^2$$

and it no longer seems weird.