1 Maximum Likelihood Estimation

Assume $\mathbf{X} \sim P_\theta, \theta \in \Theta$, with joint pdf (or pmf) $f(\mathbf{x} \mid \theta)$. Suppose we observe $\mathbf{X} = \mathbf{x}$. The Likelihood function is

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

as a function of θ (with the data $\mathbf{x}$ held fixed). The likelihood function $L(\theta \mid \mathbf{x})$ and joint pdf $f(\mathbf{x} \mid \theta)$ are the same except that $f(\mathbf{x} \mid \theta)$ is generally viewed as a function of $\mathbf{x}$ with θ held fixed, and $L(\theta \mid \mathbf{x})$ as a function of θ with $\mathbf{x}$ held fixed. $f(\mathbf{x} \mid \theta)$ is a density in $\mathbf{x}$ for each fixed θ. But $L(\theta \mid \mathbf{x})$ is not a density (or mass function) in θ for fixed $\mathbf{x}$ (except by coincidence).

1.1 The Maximum Likelihood Estimator (MLE)

A point estimator $\hat{\theta} = \hat{\theta}(\mathbf{x})$ is a MLE for θ if

$$L(\hat{\theta} \mid \mathbf{x}) = \sup_\theta L(\theta \mid \mathbf{x}),$$

that is, $\hat{\theta}$ maximizes the likelihood. In most cases, the maximum is achieved at a unique value, and we can refer to "the" MLE, and write

$$\hat{\theta}(\mathbf{x}) = \text{argmax}_\theta L(\theta|\mathbf{x}).$$

(But there are cases where the likelihood has flat spots and the MLE is not unique.)

1.2 Motivation for MLE's

Note: We often write $L(\theta \mid \mathbf{x}) = L(\theta)$, suppressing $\mathbf{x}$, which is kept fixed at the observed data. Suppose $\mathbf{x} \in \mathbb{R}^n$.
Discrete Case: If $f(\cdot \mid \theta)$ is a mass function ($\mathbf{X}$ is discrete), then

$$L(\theta) = f(\mathbf{x} \mid \theta) = P_\theta(\mathbf{X} = \mathbf{x}).$$

$L(\theta)$ is the probability of getting the observed data $\mathbf{x}$ when the parameter value is θ.
Continuous Case: When $f(\cdot \mid \theta)$ is a continuous density $P_\theta(\mathbf{X} = \mathbf{x}) = 0$, but if $B \subset \mathbb{R}^n$ is a very, very small ball (or cube) centered at the observed data $\mathbf{x}$, then

$$P_\theta(\mathbf{X} \in B) \approx f(\mathbf{x} \mid \theta) \times \text{Volume}(B) \propto L(\theta).$$

$L(\theta)$ is proportional to the probability the random data $\mathbf{X}$ will be close to the observed data $\mathbf{x}$ when the parameter value is θ. Thus, the MLE $\hat{\theta}$ is the value of θ which makes the observed data $\mathbf{x}$ "most probable".

To find $\hat{\theta}$, we maximize $L(\theta)$. This is usually done by calculus (finding a stationary point), but **not** always. If the parameter space Θ contains endpoints or boundary points, the maximum can be achieved at a boundary point without being a stationary point. If $L(\theta)$ is not "smooth" (continuous and everywhere differentiable), the maximum does not have to be achieved at a stationary point.

Cautionary Example: Suppose $X_1, \ldots, X_n$ are iid Uniform$(0, \theta)$ and $\Theta = (0, \infty)$. Given data $\mathbf{x} = (x_1, \ldots, x_n)$, find the MLE for θ.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \theta^{-1} I(0 < x_i < \theta) = \theta^{-n} I(0 \le x_{(1)}) I(x_{(n)} \le \theta) \\ &= \begin{cases} \theta^{-n} & \text{for} \quad \theta \ge x_{(n)} \\ 0 & \text{for} \quad 0 < \theta < x_{(n)} \end{cases} \end{aligned}$$

which is maximized at $\theta = x_{(n)}$, which is a point of discontinuity and certainly not a stationary point. Thus, the MLE is $\hat{\theta} = x_{(n)}$.

Notes: $L(\theta) = 0$ for $\theta < x_{(n)}$ is just saying that these values of θ are absolutely ruled out by the data (which is obvious). A strange property of the MLE in this example (not typical):

$$P_\theta(\hat{\theta} < \theta) = 1$$

The MLE is biased; it is always less than the true value.

A Similar Example: Let $X_1, \ldots, X_n$ be iid Uniform(α, β) and $\Theta = \{(\alpha, \beta) : \alpha < \beta\}$. Given data $\mathbf{x} = (x_1, \ldots, x_n)$, find the MLE for $\theta = (\alpha, \beta)$.

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} (\beta - \alpha)^{-1} I(\alpha < x_i < \beta) = (\beta - \alpha)^{-n} I(\alpha \le x_{(1)}) I(x_{(n)} \le \beta) \\ &= \begin{cases} (\beta - \alpha)^{-n} & \text{for} \quad \alpha \le x_{(1)}, x_{(n)} \le \beta \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

which is maximized by making $\beta - \alpha$ as small as possible without entering "0 otherwise" region. Clearly, the maximum is achieved at $(\alpha, \beta) = (x_{(1)}, x_{(n)})$. Thus the MLE is $\theta = (\hat{\alpha}, \hat{\beta}) = (x_{(1)}, x_{(n)})$. Again, $P_{\alpha,\beta}(\alpha < \hat{\alpha}, \hat{\beta} < \beta) = 1$.

2 Maximizing the Likelihood (one parameter)

2.1 General Remarks

Basic Result: A continuous function $g(\theta)$ defined on a closed, bounded interval J attains its supremum (but might do so at one of the endpoints). (That is, there exists a point $\theta_0 \in J$ such that $g(\theta_0) = \sup_{\theta \in J} g(\theta)$.)
Consequence: Suppose $g(\theta)$ is a continuous, non-negative function defined on an open interval $J = (c, d)$ (where perhaps $c = -\infty$ or $d = \infty$). If

$$\lim_{\theta \to c} g(\theta) = \lim_{\theta \to d} g(\theta) = 0,$$

then g attains its supremum. Thus, MLEs usually exist when the likelihood function is continuous.

3 Maxima at Stationary Points

Suppose the function $g(\theta)$ is defined on an interval Θ (which may be open or closed, infinite or finite). If g is differentiable and attains its supremum at a point θ_0 in the interior of Θ, that point must be a stationary point (that is, $g'(\theta_0) = 0$).

1. If $g'(\theta_0) = 0$ and $g''(\theta_0) < 0$, then θ_0 is a local maximum (but might not be the global maximum).

2. If $g'(\theta_0) = 0$ and $g''(\theta) < 0$ for all $\theta \in \Theta$, then θ_0 is a global maximum (that is, it attains the supremum). The condition in (1) is necessary (but not sufficient) for θ_0 to be a global maximum. Condition (2) is sufficient (but not necessary).
   A function satisfying $g''(\theta) < 0$ for all $\theta \in \Theta$ is called strictly concave. It lies below any tangent line. Another useful condition (sufficient, but not necessary) is:

3. If $g'(\theta) > 0$ for $\theta < \theta_0$ and $g'(\theta) < 0$ for $\theta > \theta_0$, then θ_0 is a global maximum.

4 Maximizing the Likelihood (multi-parameter)

4.1 Basic Result:

A continuous function $g(\theta)$ defined on a closed, bounded set $J \subset \mathbb{R}^k$ attains its supremum (but might do so on the boundary).

4.2 Consequence:

Suppose $g(\theta)$ is a continuous, non-negative function defined for all $\theta \in \mathbb{R}^k$. If $g(\theta) \to 0$ as $||\theta|| \to \infty$, then g attains its supremum. Thus, MLEs usually exist when the likelihood function is continuous.

Suppose the function $g(\theta)$ is defined on a convex set $\Theta \subset \mathbb{R}^k$ (that is, the line segment joining any two points in Θ lies entirely inside Θ). If g is differentiable and attains its supremum at a point θ_0 in the interior of Θ, that point must be a stationary point:

$$\frac{\partial g(\theta_0)}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, k.$$

Define the vector D and Hessian matrix H:

$$D(\theta) = \left(\frac{\partial g(\theta)}{\partial \theta_i} \right)_{i=1}^{k} \quad (\text{a } k \times 1 \text{ vector}).$$

$$H(\theta) = \left(\frac{\partial^2 g(\theta)}{\partial \theta_i \partial \theta_j} \right)_{i,j=1}^{k} \quad (\text{a } k \times k \text{ matrix})$$

4.3 Maxima at Stationary Points

1. If $D(\theta_0) = 0$ and $H(\theta_0)$ is negative definite, then θ_0 is a local maximum (but might not be the global maximum).

2. If $D(\theta_0) = 0$ and $H(\theta)$ is negative definite for all $\theta \in \Theta$, then θ_0 is a global maximum (that is, it attains the supremum).

(1) is necessary (but not sufficient) for θ_0 to be a global maximum. (2) is sufficient (but not necessary).
A function for which $H(\theta)$ is negative definite for all $\theta \in \Theta$ is called strictly concave. It lies below any tangent plane.

4.4 Positive and Negative Definite Matrices

Suppose M is a $k \times k$ symmetric matrix.
Note: Hessian matrices and covariance matrices are symmetric.
Definitions:

1. M is positive definite if $x'Mx > 0$ for all $x \neq 0$ ($x \in \mathbb{R}^k$).

2. M is negative definite if $x'Mx < 0$ for all $x \neq 0$.

3. M is non-negative definite (or positive semi-definite) if $x'Mx \geq 0$ for all $x \in \mathbb{R}^k$.

Facts:

1. M is p.d. iff all its eigenvalues are positive.
2. M is n.d. iff all its eigenvalues are negative.
3. M is n.n.d. iff all its eigenvalues are non-negative.
4. M is p.d. iff $-M$ is n.d.
5. If M is p.d., all its diagonal elements must be positive.
6. If M is n.d., all its diagonal elements must be negative.
7. The determinant of a symmetric matrix is equal to the product of its eigenvalues.
   2×2 Symmetric Matrices:

$$\begin{aligned} M &= (m_{ij}) = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}, m_{12} = m_{21} \\ |M| &= m_{11}m_{22} - m_{12}m_{21} = m_{11}m_{22} - m_{12}^2. \end{aligned}$$

   A 2×2 matrix is p.d. when the determinant is positive and the diagonal elements are positive. A 2×2 matrix is n.d. when the determinant is positive and the diagonal elements are negative.
   The bare minimum you need to check:
   M is p.d. if $m_{11} > 0$ (or $m_{22} > 0$) and $|M| > 0$.
   M is n.d. if $m_{11} < 0$ (or $m_{22} < 0$) and $|M| > 0$.

Example: Observe $X_1, X_2, \ldots, X_n$ be iid Gamma(α, β).
Preliminaries:

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{x_i^{\alpha-1} e^{-x_i/\beta}}{\beta^\alpha \Gamma(\alpha)}$$

Maximizing L is same as maximzing $l = \log L$ given by

$$l(\alpha, \beta) = (\alpha - 1)T_1 - T_2/\beta - n\alpha \log \beta - n \log \Gamma(\alpha)$$

where $T_1 = \sum_i \log x_i, T_2 = \sum_i x_i$. Note that $T = (T_1, T_2)$ is the natural sufficient statistic of this 2pef.

$$\begin{aligned}
\frac{\partial l}{\partial \alpha} &= T_1 - n\log\beta - n\psi(\alpha), \quad \psi(\alpha) \equiv \frac{d}{d\alpha}\log\Gamma(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\
\frac{\partial l}{\partial \beta} &= \frac{T_2}{\beta^2} - \frac{n\alpha}{\beta} = \frac{1}{\beta^2}(T_2 - n\alpha\beta) \\
\frac{\partial^2 l}{\partial \alpha^2} &= -n\psi'(\alpha) \\
\frac{\partial^2 l}{\partial \beta^2} &= \frac{-2T_2}{\beta^3} + \frac{n\alpha}{\beta^2} = \frac{-1}{\beta^3}(2T_2 - n\alpha\beta) \\
\frac{\partial^2 l}{\partial \alpha \partial \beta} &= \frac{-n}{\beta}
\end{aligned}$$

Situation #1: Suppose $\alpha = \alpha_0$ is known. Find MLE for β. (Drop α from arguments: $l(\beta) = l(\alpha_0, \beta)$ etc.)
$l(\beta)$ is continuous and differentiable.
$l(\beta)$ has a unique stationary point:

$$\begin{aligned}
l'(\beta) &= \frac{\partial l}{\partial \beta} = \frac{1}{\beta^2}(T_2 - n\alpha_0\beta) = 0 \\
&\text{iff} \quad T_2 = n\alpha_0\beta, \text{ iff } \beta = \frac{T_2}{n\alpha_0} (\equiv \beta^*)
\end{aligned}$$

Now we check the second derivative.

$$l''(\beta) = \frac{\partial^2 l}{\partial \beta^2} = \frac{-1}{\beta^3}(2T_2 - n\alpha\beta) = \frac{-1}{\beta^3}\{T_2 + (T_2 - n\alpha\beta)\}.$$

Note $l''(\beta^*) < 0$ since $T_2 - n\alpha_0\beta^* = 0$, but $l''(\beta) > 0$ for $\beta > 2T_2/(n\alpha_0)$. Thus, the stationary point satisfies the necessary condition for a global maximum, but not the sufficient condition (i.e., $l(\beta)$ is not a strictly concave function). How can we be sure that we have found the global maximum, and not just a local maximum? In this case, there is a simple argument: The stationary point β^* is unique, and $l'(\beta) > 0$ for $\beta < \beta^*$,and $l'(\beta) < 0$ for $\beta > \beta^*$. This ensures β^* is the unique global maximizer.
Conclusion: $\hat{\beta} = \frac{T_2}{n\alpha_0}$. (This is a function of T_2, which is a sufficient statistic for β when α is known.)
Situation #2: Suppose $\beta = \beta_0$ is known. Find MLE for α. (Drop β from arguments: $l(\alpha) = l(\alpha, \beta_0)$ etc.)
Note: $l'(\alpha)$ and $l''(\alpha)$ involve $\psi(\alpha)$. The function ψ is infinitely differentiable on the interval $(0, \infty)$, and satisfies $\psi'(\alpha) > 0$ and $\psi''(\alpha) < 0$ for all $\alpha > 0$. (The function is strictly increasing and strictly concave.)

Also,

$$\lim_{\alpha \to 0^+} \psi(\alpha) = -\infty, \lim_{\alpha \to \infty} \psi(\alpha) = \infty.$$

Thus $\psi^{-1} : \mathbb{R} \to (0, \infty)$ exists. $l(\alpha)$ is continuous and differentiable. $l(\alpha)$ has a unique stationary point:

$$\begin{aligned} & l'(\alpha) = T_1 - n \log \beta_0 - n\psi(\alpha) = 0 \\ \text{iff} \quad & \psi(\alpha) = T_1/n - \log \beta_0 \\ \text{iff} \quad & \alpha = \psi^{-1}(T_1/n - \log \beta_0) \end{aligned}$$

This is the unique global maximizer since

$$l''(\alpha) = -n\psi'(\alpha) < 0, \quad \forall \alpha > 0.$$

Thus $\hat{\alpha} = \psi^{-1}(T_1/n - \log \beta_0)$ is the MLE. (This is a function of T_1, which is a sufficient statistic for α when β is known.)

Situation #3: Find MLE for $\theta = (\alpha, \beta)$. $l(\alpha, \beta)$ is continuous and differentiable. A stationary point must satisfy the system of two equations:

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= T_1 - n \log \beta - n\psi(\alpha) = 0 \\ \frac{\partial l}{\partial \beta} &= \frac{1}{\beta^2}(T_2 - n\alpha\beta) = 0. \end{aligned}$$

Solving the second equation for β gives

$$\beta = \frac{T_2}{n\alpha}$$

Plugging this into the first equation, and rearranging a bit leads to

$$\frac{T_1}{n} - \log\left(\frac{T_2}{n}\right) = \psi(\alpha) - \log \alpha \equiv H(\alpha)$$

The function $H(\alpha)$ is continuous and strictly increasing from $(0, \infty)$ to $(-\infty, 0)$, so that it has an inverse mapping $(-\infty, 0)$ to $(0, \infty)$. Thus, the solution to the above equation can be written:

$$\alpha = H^{-1}\left\{\frac{T_1}{n} - \log\left(\frac{T_2}{n}\right)\right\}$$

Thus the unique stationary point is:

$$\begin{aligned} \hat{\alpha} &= H^{-1}\left\{\frac{T_1}{n} - \log\left(\frac{T_2}{n}\right)\right\} \\ \hat{\beta} &= \frac{T_2}{n\hat{\alpha}} \end{aligned}$$

Is this the MLE?
Let us examine the Hessian.

$$\begin{aligned} H(\alpha,\beta) &= \begin{pmatrix} \frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial \beta} \\ \frac{\partial^2 l}{\partial \alpha \partial \beta} & \frac{\partial^2 l}{\partial \beta^2} \end{pmatrix} \\ &= \begin{pmatrix} -n\psi'(\alpha) & \frac{-n}{\beta} \\ \frac{-n}{\beta} & \frac{-1}{\beta^3}(2T_2 - n\alpha\beta) \end{pmatrix} \\ H(\hat{\alpha},\hat{\beta}) &= \begin{pmatrix} -n\psi'(\hat{\alpha}) & \frac{-n^2\hat{\alpha}}{T_2} \\ \frac{-n^2\hat{\alpha}}{T_2} & \frac{-n^3\hat{\alpha}^3}{T_2^2} \end{pmatrix} \end{aligned}$$

The diagonal elements are both negative, and the determinant is equal to

$$\frac{n^4\hat{\alpha}^2}{T_2^2}(\hat{\alpha}\psi'(\hat{\alpha}) - 1).$$

This is positive since $\alpha\psi'(\alpha) - 1 > 0$ for all $\alpha > 0$. This guarantees that $H(\hat{\alpha},\hat{\beta})$ is negative definite so that $(\hat{\alpha},\hat{\beta})$ is at least a local maximum.

5 Invariance principle for the MLE's

If $\eta = \tau(\theta)$ and $\hat{\theta}$ is the MLE of θ, then $\hat{\eta} = \tau(\hat{\theta})$ is the MLE of η.
Comments

1. If $\tau(\theta)$ is a 1-1 function, this is a trivial theorem.
2. If $\tau(\theta)$ is not 1-1, this is essentially true by definition of induced likelihood. (see later).

Example: $X = (X_1, X_2, \ldots, X_n)$ iid $\mathrm{N}(\mu, \sigma^2)$. The usual parameters $\theta = (\mu, \sigma^2)$ are related to the natural parameters $\eta = (\mu/\sigma^2, -1/(2\sigma^2))$ of the 2pef by a 1-1 function: $\eta = \tau(\theta)$. The likelihood in terms of θ is

$$L_1(\theta) = (2\pi\sigma^2)^{-n/2} e^{-n\mu^2/2\sigma^2} e^{\mu/\sigma^2 T_1 - (1/2\sigma^2)T_2}$$

where $T_1 = \sum X_i, T_2 = \sum X_i^2$.
Simple Example: $X = (X_1, X_2, \ldots, X_n)$ iid Bernoulli(p). It is known that MLE of p is $\hat{p} = \bar{X}$. Thus

1. MLE of p^2 is $\hat{p}^2 = \bar{X}^2$.
2. MLE of $p(1-p)$ is $\bar{X}(1-\bar{X})$.

The function of p in 1. is 1-1, but not 1-1 in 2.

5.1 Induced Likelihood

Definition 1. *If $\eta = \tau(\theta)$, then*

$$L^*(\eta) \equiv \sup_{\theta:\tau(\theta)=\eta} L(\theta).$$

Go back to the example $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ iid. If the MLE $\hat{\eta}$ of η is defined to be the value which maximized $L^*(\eta)$, then it is easily seen that $\hat{\eta} = \tau(\hat{\theta})$. The likelihood in terms of η is

$$L_2(\eta) = (-\pi/\eta_2)^{-n/2} e^{n\eta_1^2/4\eta_2} e^{\eta_1 T_1 + \eta_2 T_2}$$

obtained by substituting in $L_1(\theta)$

$$\mu = -\eta_1/(2\eta_2), \quad \sigma^2 = -1/(2\eta_2),$$

that is, evaluating L_1 at

$$\theta = (\mu, \sigma^2) = (-\eta_1/(2\eta_2), -1/(2\eta_2)) = \tau^{-1}(\eta).$$

Stated abstractly $L_2(\eta) = L_1(\tau^{-1}(\eta))$, so that L_2 is maximized when $\tau^{-1}(\eta) = \hat{\theta}$, that is, by $\eta = \tau(\hat{\theta})$. The MLE of θ is known to be

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{X}, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right)$$

so the invariance principle says the MLE of η is

$$\hat{\eta} = \tau(\hat{\theta}) = \left(\frac{\hat{\mu}}{\hat{\sigma}^2}, \frac{-1}{2\hat{\sigma}^2} \right).$$

Continuation of example: What is the MLE of $\alpha = \mu + \sigma^2$? Note that

$$\alpha = g(\mu, \sigma^2) = \mu + \sigma^2$$

is not a 1-1 function, but

$$\hat{\alpha} = g(\hat{\mu}, \hat{\sigma^2}) = \hat{\mu} + \hat{\sigma^2} = \bar{X} + \mathrm{SS}/n$$

where $\mathrm{SS} = \sum_{i=1}^{n} (X_i - \bar{X})^2$.
What is MLE of μ?, σ^2 ?
With $g_1(x, y) = x, g_2(x, y) = y$, we have

$$\mu = g_1(\theta), \quad \sigma^2 = g_2(\theta)$$

so that the MLEs are

$$\hat{\mu} = g_1(\hat{\theta}) = \bar{X}, \quad \hat{\sigma}^2 = g_2(\hat{\theta}) = \mathrm{SS}/n.$$

Thus, the invariance principle implies:

$$(\mu, \hat{\sigma^2})(\text{MLE as a pair}) = (\hat{\mu}(\text{MLE of } \mu), \hat{\sigma}^2(\text{MLE of } \sigma^2))$$

5.2 MLE for Exponential Families

The invariance principle for MLEs allows us to work with the natural parameter η (which is a 1-1 function of θ).
1pef:

$$f(x \mid \theta) = c(\theta)h(x)\exp\{w(\theta)t(x)\}$$

Natural parameter: $\eta = w(\theta)$.
With a little abuse of notation (writing $f(x \mid \eta)$ for $f^*(x \mid \eta) = f(x \mid w^{-1}(\eta))$ and $c(\eta)$ for $c^*(\eta) = c(w^{-1}(\eta))$, we can write

$$f(x \mid \eta) = c(\eta)h(x)\exp\{\eta t(x)\}.$$

For clarity of notations, we will use $x = (x_1, \ldots, x_N)$ as the observed data and $X = (X_1, X_2, \ldots, X_N)$ as the random data. If $X_1, \ldots, X_N$ iid from $f(x \mid \eta)$, then

$$l(\eta) = N\log\{c(\eta)\} + \sum_{i=1}^{N}\log h(x_i) + \eta\sum_{i=1}^{N} t(x_i)$$

Since by 3.32(a) $Et(X_i) = -\frac{\partial}{\partial\eta}\log\{c(\eta)\}$, we have

$$\begin{aligned} l'(\eta) &= N\frac{\partial}{\partial\eta}\log\{c(\eta)\} + \sum_{i=1}^{N} t(x_i) \qquad (1)\\ &= -E\left[\sum_{i=1}^{N} t(X_i)\right] + \sum_{i=1}^{N} t(x_i)\\ &= -ET(X) + T(x) \end{aligned}$$

where $T(X) = \sum_{i=1}^{N} t(X_i)$. Hence the condition for a stationary point is equivalent to:

$$E_\eta T(X) = T(x)$$

Note that using (1),

$$l''(\eta) = N\frac{\partial^2}{\partial\eta^2}\log\{c(\eta)\} = N\{-\mathrm{Var}_\eta t(X_i)\} < 0$$

for all η. Thus any interior stationary point (not on the boundary of $\Theta^* = \{w(\theta) : \theta \in \Theta\}$) is automatically a global maximum so long as Θ^* is convex. In one dimension ($\Theta \subset \mathbb{R}$), this means Θ^* must be an interval of some sort (can be infinite). Ignoring this fine point,

for a 1pef, the log-likelihood will have a unique stationary point which will be the MLE.
k-pef:

$$f(x \mid \theta) = c(\theta)h(x)\exp\{\sum_{j=1}^{k} w_j(\theta)t_j(x)\}$$

Natural parameter: $\eta = (\eta_1, \ldots, n_k) = (w_1(\theta), \ldots, w_k(\theta))$, that is $\eta_j = w_j(\theta)$.

$$f(x \mid \eta) = c(\eta)h(x)\exp\{\sum_{j=1}^{k} \eta_j t_j(x)\}$$

If $X_1, X_2, \ldots, X_N$ iid from $f(x \mid \eta)$, then

$$\begin{aligned} l(\eta) &= N\log c(\eta) + \sum_{i=1}^{N}\log h(x_i) + \sum_{j=1}^{k}\eta_j\Big\{\sum_{i=1}^{N} t_j(x_i)\Big\} \\ \frac{\partial l}{\partial \eta_j} &= N\underbrace{\frac{\partial}{\partial \eta_j}\log c(\eta)}_{-t_j(X_i)} + \sum_{i=1}^{N} t_j(x_i) \\ &= -E\sum_{i=1}^{N} t_j(X_i) + \sum_{i=1}^{N} t_j(x_i) \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \eta_j \partial \eta_l} &= N\left(\frac{\partial^2}{\partial \eta_j \partial \eta_l}\log c(\eta)\right) \\ &= N(-\mathrm{Cov}(t_j(X_i), t_l(X_i))) \end{aligned}$$

Thus, the equations for a stationary point is

$$\frac{\partial l}{\partial \eta_j} = 0, \quad j = 1, \ldots, k$$

are equivalent to

$$E_\eta T_j(\mathbf{X}) = T_j(\mathbf{x}), \quad j = 1, \ldots, k \tag{2}$$

where $T_j(\mathbf{X}) = \sum_{i=1}^{N} t_j(X_i)$ and $T_j(\mathbf{x}) = \sum_{i=1}^{N} t_j(x_i)$ or in vector notation,

$$E_\eta T(\mathbf{X}) = T(\mathbf{x}), \quad j = 1, \ldots, k$$

where $T(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))$ and $T(\mathbf{x}) = (T_1(\mathbf{x}), \ldots, T_k(\mathbf{x}))$.

The Hessian matrix $H(\eta) = \left(\frac{\partial^2 l}{\partial \eta_i \partial \eta_j} \right)_{i,j=1}^{k}$ is given

$$H(\eta) = -N\Sigma(\eta)$$

where $\Sigma(\eta)$ is the $k \times k$ covariance matrix of $(T_1(X_1), T_2(X_1), \ldots, T_k(X_1))$. A covariance matrix will be positive definite (except in degenerate cases), so that $H(\eta)$ will be negative definite for all η.
Conclusion: An interior stationary point (i.e., a solution of (2)) must be the unique global maximum, and hence the MLE. This result also holds in the original parameterization with (2) restated as

$$E_\theta T_j(\mathbf{X}) = T_j(\mathbf{x}), \quad j = 1, \ldots, k.$$

Connection with MOM: For a 1pef with $t(x) = x$, MOM and MLE agree. For a $k$pef with $t_j(x) = x^j$, MOM and MLE agree. Why? Because then (2) is equivalent to the equations for the MOM estimator.

6 Revisiting Gamma Example:

The system of equations for the MLE of (α, β) may be easily derived directly from (2).

$$ET_1(\mathbf{X}) = T_1(\mathbf{x})$$
$$ET_2(\mathbf{X}) = T_2(\mathbf{x})$$

which becomes

$$E \sum_{i=1}^{n} \log X_i = nE \log X_1 = n(\log \beta + \psi(\alpha)) = T_1(\mathbf{x})$$
$$E \sum_{i=1}^{n} X_i = nEX_1 = n\alpha\beta = T_2(\mathbf{x})$$

The equations are the same as the equations for a stationary point derived earlier. For $X \sim \text{Gamma}(\alpha, \beta)$, we have used:

$$\begin{aligned}
E \log X &= \int_0^\infty \log x \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} dx \\
&= \int_0^\infty (\log(x/\beta) + \log \beta) \frac{(x/\beta)^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)} \frac{dx}{\beta} \\
&= \int_0^\infty (\log z + \log \beta) \frac{z^{\alpha-1} e^{-z}}{\Gamma(\alpha)} dz \\
&= \log \beta + \frac{1}{\Gamma(\alpha)} \int_0^\infty \underbrace{(z^{\alpha-1} \log z)}_{\frac{\partial}{\partial \alpha} z^{\alpha-1}} e^{-z} dz \\
&= \log \beta + \frac{1}{\Gamma(\alpha)} \frac{\partial}{\partial \alpha} \int_0^\infty z^{\alpha-1} e^{-z} dz \\
&= \log \beta + \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \log \beta + \psi(\alpha).
\end{aligned}$$

Verifying Stationary Point is Global Maximum: The Gamma family is a 2pef (or a 1pef if α or β is held fixed). Switching to the natural parameters $\eta_1 = \alpha - 1$, $\eta_2 = -1/\beta$(or just making the substitution $\lambda = 1/\beta$) simplifies the second derivatives w.r.t. η_2 (or λ). The Hessian matrix is now negative definite for all $\theta = (\eta_1, \eta_2)$, which is a sufficient condition for the stationary point to be the global maximum.

6.1 MLEs for More General Exponential Families

Proposition 1. *If $\mathbf{X} \sim P_\theta, \theta \in \Theta$, where P_θ has a joint pdf (pmt) from an n-variate k-parameter exponential family*

$$f(\mathbf{x} \mid \theta) = c(\theta) h(\mathbf{x}) \exp\left\{ \sum_{j=1}^k w_j(\theta) T_j(\mathbf{x}) \right\}$$

for $\mathbf{x} \in \mathbb{R}^n, \theta \in \Theta \subset \mathbb{R}^k$, then the MLE of θ based on the observed data $\mathbf{x}$ is the solution of the system of equations

$$E_\theta T_j(\mathbf{X}) = T_j(\mathbf{x}), j = 1, \ldots, k, \quad \textit{Solve for } \theta.$$

providing the solution (call it $\hat{\theta}$) satisfies

$$w(\hat{\theta}) \in \textit{interior of} \{w(\theta) : \theta \in \Theta\}.$$

Proof. Essentially the same as for the ordinary $k$pef. □

Example: Simple Linear Regression with known variance: $Y_1, Y_2, \ldots, Y_n$ are independent with

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma_0^2), \quad \theta = (\beta_0, \beta_1)$$

Joint distribution of $\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_n)$ forms exponential family. Natural sufficient statistic is

$$t(\underset{\sim}{Y}) = (\sum_i Y_i, \sum_i x_i Y_i).$$

$E_\theta t(\underset{\sim}{Y}) = t(\underset{\sim}{y})$ has the form

$$\begin{aligned} E(\sum Y_i) &= \sum y_i \\ E(\sum x_i Y_i) &= \sum x_i y_i \end{aligned}$$

Thus the MLE $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1)$ is solution of

$$\begin{aligned} \sum_i (\beta_0 + \beta_1 x_i) &= \sum_i y_i \\ \sum_i x_i (\beta_0 + \beta_1 x_i) &= \sum_i x_i y_i \end{aligned}$$

6.2 Sufficient statistics and MLEs

If $T = T(X)$ is a sufficient statistic for θ, then there is an MLE which is a function of T. (If the MLE is unique, then we can say the MLE is a function of T).

Proof. By FC,

$$f(x \mid \theta) = g(T(x), \theta) h(x).$$

Assume for convenience the MLE is unique. Then the MLE is

$$\begin{aligned} \hat{\theta}(x) &= \mathrm{argmax}_\theta f(x \mid \theta) \\ &= \mathrm{argmax}_\theta g(T(x), \theta) \end{aligned}$$

which is clearly a function of $T(x)$. □

MLE coincides with "Least Squares". For independent normal rv's with constant variance σ^2 (known or unknown).
$Y_1, Y_2, \ldots, Y_n$ are independent with

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma_0^2), \theta = (\beta_0, \beta_1)$$

or more generally,

$$Y_i \sim \mathrm{N}(g(x_i, \beta), \sigma_0^2),$$

where β is possibly a vector. Then

$$L(\underbrace{\beta, \sigma^2}_{\theta}) = f(\underset{\sim}{y} \mid \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - g(x_i, \beta)^2 \right\}.$$

For any σ^2 (fixed arbitrary value), maximizing $L(\beta, \sigma^2)$ with respect to β is equivalent to minimizing $\sum_{i=1}^{n}(y_i - g(x_i, \beta))^2$ with respect to β. Hence MLE and Least squares give same estimates of β parameters.