Invariance Principle for 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: If $\tau(\theta)$ is a 1-1 function, this is a trivial theorem.

If $\tau(\theta)$ is <u>not</u> 1-1, this is essentially true by definition of induced likelihood. (see later)

<u>Example</u>: $\underbrace{X_1, X_2, \ldots, X_n}_{X}$ iid $N(\underbrace{\mu, \sigma^2}_{\theta})$

The usual parameters $\theta = (\mu, \sigma^2)$ are related to the natural parameters $\eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$ of the 2pef by a 1-1 function: $\eta = \tau(\theta)$.

The likelihood in terms of θ is

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

where $T_1 = \sum X_i$, $T_2 = \sum X_i^2$

Simple Example

$\underbrace{X_1, X_2, \ldots, X_n}_{X}$ iid Bernoulli(p)

It is known that
MLE of p is $\hat{p} = \bar{X}$.

Thus
(a) MLE of p^2 is $(\hat{p})^2 = \bar{X}^2$

(b) MLE of $p(1-p)$ is $\bar{X}(1-\bar{X})$

The function of p in (a) is 1-1, but not 1-1 in (b).

Definition of Induced Likelihood:

If $\eta = \tau(\theta)$, then

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

(dependence on x is implicit)

If the MLE $\hat{\eta}$ of η is defined to be the value which maximizes $L^*(\eta)$, 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

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

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 (SS/n)

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{x}, \frac{1}{n}\sum(x_i - \bar{x})^2\right)$$

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 $\alpha = \mu + \sigma^2$?

$\alpha = g(\mu, \sigma^2) = \mu + \sigma^2$ (not 1-1)

so that

$$\hat{\alpha} = g(\hat{\mu}, \hat{\sigma}^2) = \hat{\mu} + \hat{\sigma}^2 = \bar{x} + \frac{SS}{n}$$

What is the MLE of μ? σ^2?

with $g_1(x,y) = x$, $g_2(x,y) = y$

we have

$$\mu = g_1(\theta)$$

$$\sigma^2 = g_2(\theta)$$

so that the MLE's are

$$\hat{\mu} = g_1(\hat{\theta}) = \bar{x}$$

$$\hat{\sigma}^2 = g_2(\hat{\theta}) = SS/n .$$

Thus, the invariance principle implies:

$$\widehat{(\mu, \sigma^2)} = (\hat{\mu}, \hat{\sigma}^2)$$

$\widehat{(\mu, \sigma^2)}$: MLE of pair

$\hat{\mu}$: MLE of μ

$\hat{\sigma}^2$: MLE of σ^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

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

Natural param: $\eta = w(\theta)$

$$f(x|\eta) = c(\eta)h(x)\exp\{\eta\, t(x)\}$$

$c(\eta)$ → really $c^*(\eta) = c(w^{-1}(\eta))$

→ really $f^*(x|\eta) = f(x|w^{-1}(\eta))$

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

$$\ell(\eta) = N\log c(\eta) + \sum_{i=1}^{N}\log h(X_i) + \eta \underbrace{\sum_{i=1}^{N} t(X_i)}_{T(X)}$$

$$\ell'(\eta) = N \underbrace{\frac{\partial}{\partial \eta} \log c(\eta)}_{-E\,t(X_i)} + \sum_{i=1}^{N} t(X_i) \qquad (*)$$

by 3.32(a)

$$= -E\left[\sum_{i=1}^{N} t(X_i)\right] + \sum_{i=1}^{N} t(X_i)$$

(X_i: rv's; x_i: observed values)

$$= -E\,T(X) + T(x)$$

$x = (x_1, \ldots, x_n)$ = observed data

$X = (X_1, \ldots, X_n)$ = random data

so that the condition for a stationary point is equivalent to:

$$E_\eta T(X) = T(x) \qquad \ell'(\eta) = 0$$

Note that (using (*))

$$\ell''(\eta) = N \frac{\partial^2}{\partial \eta^2} \log c(\eta)$$

$$= N\left(-\mathrm{Var}_\eta\, t(X_i)\right) < 0$$

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 the fine print,
For a lpef, the log-likelihood will have a unique stationary point which will be the MLE.

kpef

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

Natural param:

$$\eta = (\eta_1, \ldots, \eta_k) = (w_1(\theta), \ldots, w_k(\theta)),$$

that is, $\eta_j = w_j(\theta)$.

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

$f \to f^*$, $c \to c^*$

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

$$\ell(\eta) = N\log c(\eta) + \sum_{i=1}^{N} \log h(x_i) + \sum_{j=1}^{k} \eta_j \left(\sum_{i=1}^{N} t_j(x_i)\right)$$

$$\frac{\partial \ell}{\partial \eta_j} = N \underbrace{\frac{\partial}{\partial \eta_j}\log c(\eta)}_{= -E\,t_j(X_i)} + \sum_{i=1}^{N} t_j(x_i)$$

$$= -E\sum_{i=1}^{N} t_j(X_i) + \sum_{i=1}^{N} t_j(x_i)$$

$$\frac{\partial^2 \ell}{\partial \eta_j \partial \eta_\ell} = N\left(\frac{\partial^2}{\partial \eta_j \partial \eta_\ell} \log c(\eta)\right)$$

$$= N\left(-\operatorname{Cov}\left(t_j(X_i), t_\ell(X_i)\right)\right)$$

proved case $j = \ell$ in homework

Thus, the equations for a stationary point

$$\frac{\partial \ell}{\partial \eta_j} = 0 \quad \text{for} \quad j = 1, \ldots, k$$

are equivalent to

$$E_\eta T_j(\boldsymbol{X}) = T_j(\boldsymbol{x}) \quad \text{for } j = 1, \ldots, k \qquad (\ddagger)$$

$$\text{where} \quad T_j(\boldsymbol{X}) = \sum_{i=1}^{N} t_j(X_i) \quad \text{and} \quad T_j(\boldsymbol{x}) = \sum_{i=1}^{N} t_j(x_i)$$

or in vector notation

$$E_\eta T(\boldsymbol{X}) = T(\boldsymbol{x}) \quad \text{where}$$
$$T(\boldsymbol{X}) = (T_1(\boldsymbol{X}), \ldots, T_k(\boldsymbol{X})) \text{ and } T(\boldsymbol{x}) = (T_1(\boldsymbol{x}), \ldots, T_k(\boldsymbol{x}))\,.$$

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

$$H(\eta) = -N \boldsymbol{\Sigma}(\eta)$$

where $\boldsymbol{\Sigma}(\eta)$ 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 ($\ddagger$)) must be the unique global maximum, and hence the MLE.

This result also holds in the original parameterization with ($\ddagger$) restated as $E_\theta T_j(\boldsymbol{X}) = T_j(\boldsymbol{x})$, $j = 1, \ldots, k$.

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

Revisiting Gamma Example:

The Gamma family is a 2pef (or a 1pef if α or β is held fixed).

Switching to the natural parameters $\eta_1 = \alpha - 1$ and $\eta_2 = -1/\beta$ (or just making the substitution $\lambda = 1/\beta$) simplifies the second derivatives w.r.t. η_2 (or λ) and makes the sufficient condition for a stationary point to be the global max hold.

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

MLE's for More General Exponential Families

Proposition: If $\boldsymbol{X} \sim P_\theta$, $\theta \in \Theta$ where P_θ has a joint pdf (pmf) from an n-variate k parameter exponential family (nvkpef):

$$f(\boldsymbol{x} \,|\, \theta) \;=\; c(\theta)h(\boldsymbol{x}) \exp\left\{ \sum_{j=1}^{k} w_j(\theta) T_j(\boldsymbol{x}) \right\}$$

$$\text{for } \boldsymbol{x} \in \mathbb{R}^n,\ \theta \in \Theta \subset \mathbb{R}^k\,,$$

then the MLE of θ based on the observed data $\boldsymbol{x}$ is the solution of the system of equations

$$E_\theta T_j(\boldsymbol{X}) = T_j(\boldsymbol{x}) \quad \text{for } j = 1, \ldots, k \qquad \text{(Solve for } \theta\text{)}$$

providing this solution (call it $\widehat{\theta}$) satisfies

$$w(\widehat{\theta}) \in \text{interior of } \{w(\theta) : \theta \in \Theta\}.$$

Proof: Essentially the same as for the ordinary kpef.

Example: Simple Linear Regression with Known Variance

$Y_1, Y_2, \ldots, Y_n$ independent

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

Joint distn. of $\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_n)$ forms exponential family.

Natural sufficient statistic is

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

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

$$E\left(\sum Y_i\right) = \sum y_i$$

$$E\left(\sum x_i Y_i\right) = \sum x_i y_i$$

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

$$\sum_i (\beta_0 + \beta_1 x_i) = \sum_i y_i$$

$$\sum_i x_i (\beta_0 + \beta_1 x_i) = \sum_i x_i y_i$$

Sufficient Statistics and MLE's

If $T = T(X)$ is a sufficient stat. for θ, then there is an MLE which is a function of T.

(If the MLE is unique, then we can say <u>the</u> MLE is a function of T.)

<u>Proof</u>: By FC,

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

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

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmax}}\ f(x|\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\ g(T(x), \theta)$$

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)

$Y_1, Y_2, \ldots, Y_n$ independent

$Y_i \sim N(\underbrace{\beta_0 + \beta_1 x_i}, \sigma^2)$

or $\beta_0 + \beta_1 x_i + \beta_2 x_i^2$

or $\beta_0 \sin \beta_1 x_i$

or $\dfrac{\beta_0}{1 - \beta_1 x_i}$

or $g(x_i, \beta)$ → possibly vectors

$L(\underbrace{\beta, \sigma^2}_{\theta}) = f(y \mid \theta)$

$= \left(\dfrac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left\{ -\dfrac{1}{2\sigma^2} \sum \right.$ ↗

$\left. (y_i - g(x_i, \beta))^2 \right\}$

For any σ^2, (fixed arbitrary value),

Maximizing $L(\beta, \sigma^2)$ with respect to β

$\iff$ Minimizing $\sum_{i=1}^{n} (y_i - g(x_i, \beta))^2$ with respect to β.

$\therefore$ MLE and Least squares give same estimates of β parameters.