Consistency and Asymptotic Distribution of MLE

Preliminary fact:

(from derivation of alternate formula for $I(\theta)$)

$$\int_{-\infty}^{\infty}\left[\frac{\partial}{\partial\theta}\left(\log f(x|\theta)\right)\right] f(x|\theta)\,dx = 0$$

equivalently

$$E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = 0$$

when X has density $f(x|\theta)$.

Heuristic argument for consistency of MLE

Notation:

$X_1, X_2, \ldots, X_n$ IID $f(x|\theta_0)$

$\uparrow$ true value.

$f(x|\theta)$ is regular one-parameter family of pdf's (or pmf's).

$\hat{\theta}$ denotes MLE of θ.

Consistent: $\hat{\theta} \approx \theta_0$ for large n $\left(\hat{\theta} \to \theta_0 \text{ as } n \to \infty\right)$

$\hat{\theta}$ is the value θ maximizing

log-likelihood $\ell(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$

or equivalently

$$\frac{1}{n}\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta)$$

This is a sample average (of something).
For large n, will be close to population average. (by LLN)

$\boxed{\begin{array}{l}\text{LLN: If } V_1, V_2, \ldots, V_n \text{ IID,} \\ \quad \text{then } \bar{V} \approx EV_i \text{ for large } n \\ \quad\quad (\bar{V} \to EV_i \text{ as } n \to \infty)\end{array}}$

Take $V_i = \log f(X_i|\theta)$ in LLN. (fix θ arbitrary)

For large n expect

$$\frac{1}{n}\ell(\theta) \approx E \log f(X_i|\theta)$$

$$= \int (\log f(x|\theta))\, f(x|\theta_0)\, dx$$

(assume density case.

$$= \Psi(\theta). \quad \text{(definition)}$$

If $\frac{1}{n}\ell(\theta) \approx \Psi(\theta)$ for "all" θ,

then $\hat{\theta}$ (the value maximizing $\frac{1}{n}\ell(\theta)$) should be close to the value of θ which maximizes $\Psi(\theta)$.

But $\Psi(\theta)$ is maximized when $\theta = \theta_0$:

$$\frac{\partial \Psi(\theta)}{\partial \theta} = \int \frac{\partial}{\partial \theta}(\log f(x|\theta))\, f(x|\theta_0)\, dx = 0 \text{ when } \theta = \theta_0.$$

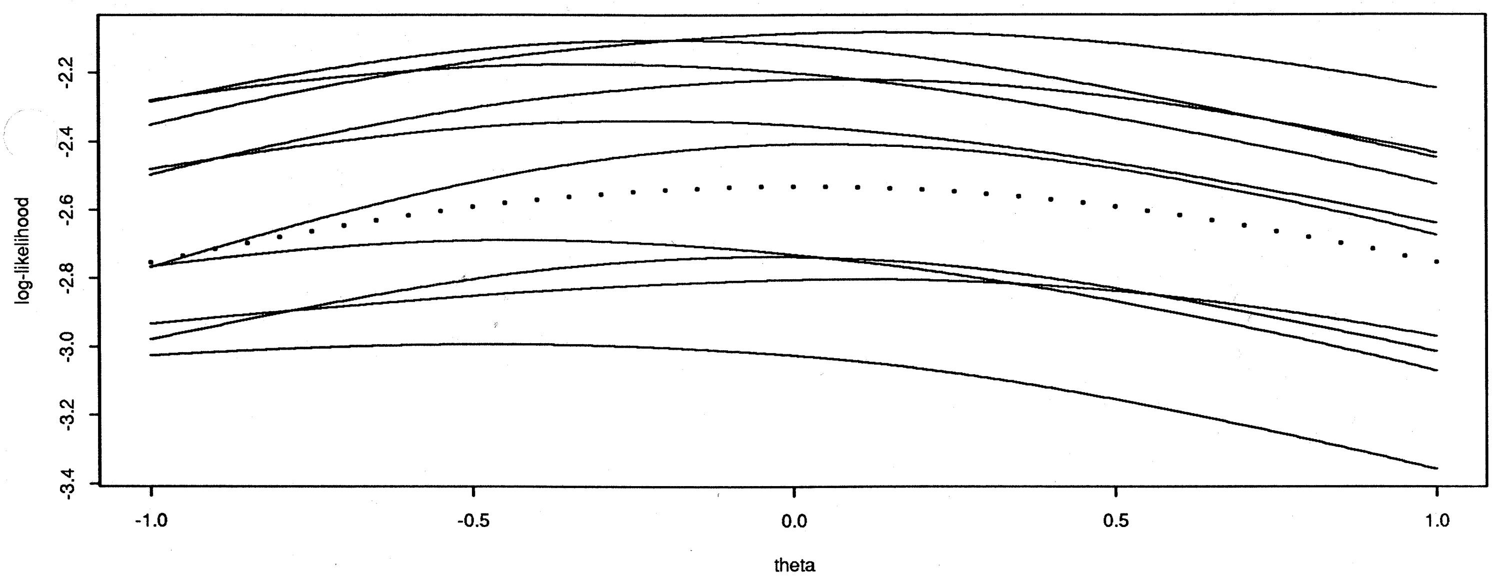
(at least θ_0 is a stationary point.)

Thus expect

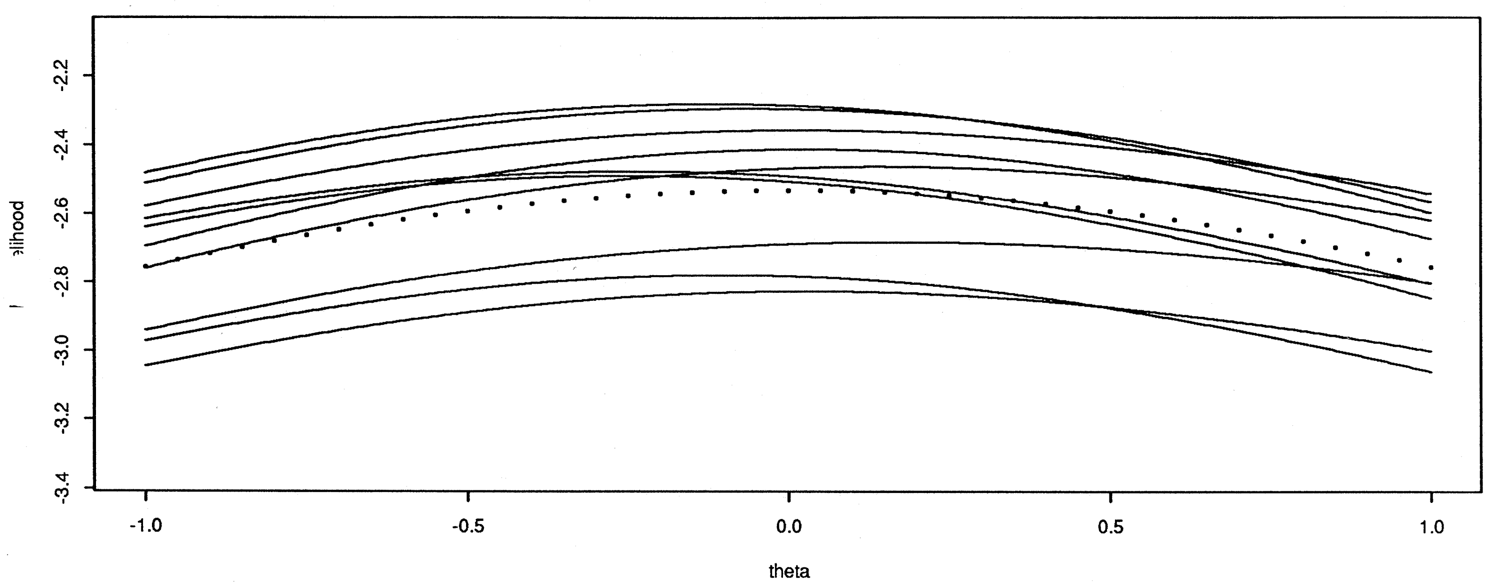
$$\hat{\theta} \approx \theta_0 \text{ for large } n. \qquad \text{QED}$$

$\hat{\theta}$ is consistent estimator.

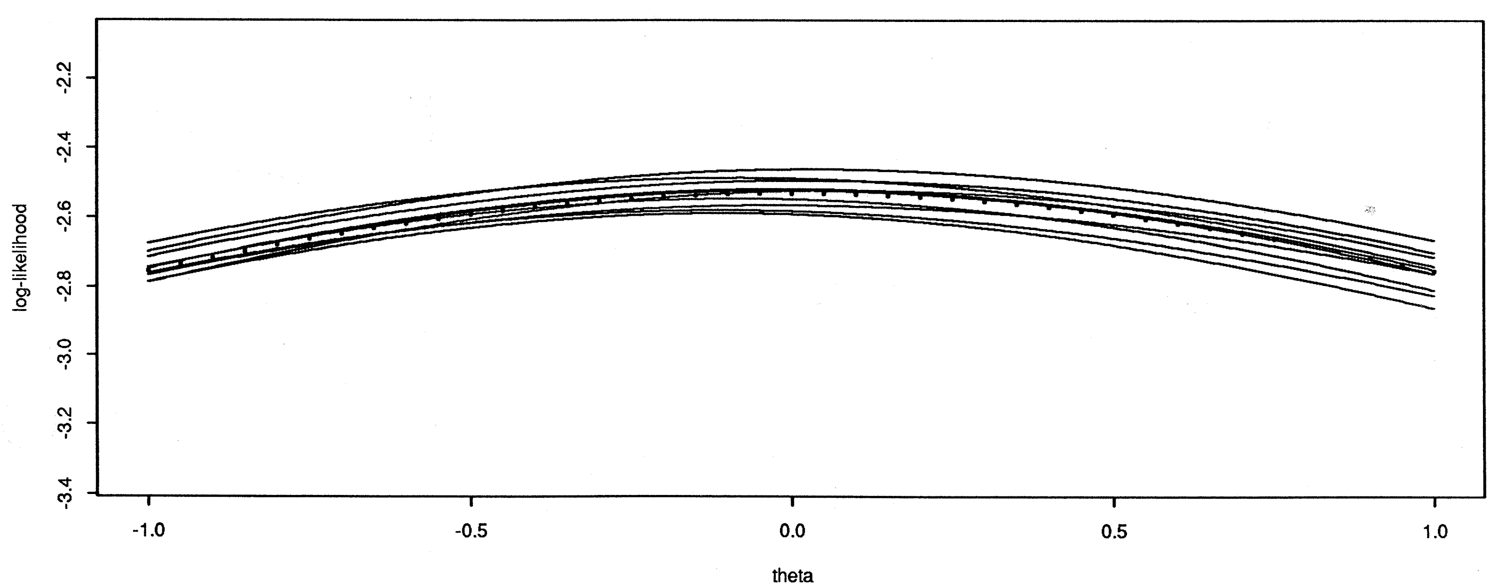
Log-Likelihood Functions for 10 Samples of Size 20



Log-Likelihood Functions for 10 Samples of Size 100



Log-Likelihood Functions for 10 Samples of Size 500



Asymptotic distn. of MLE

follows (by an elaborate argument) from the CLT applied to

$$\underbrace{\frac{1}{n}\dot{\ell}(\theta_0)}_{\text{This is } \bar{V}.} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\partial}{\partial\theta}\log f(X_i|\theta)\Big|_{\theta=\theta_0}}_{\text{Call this } V_i.}$$

CLT says (for large n)

$$\bar{V} \sim \text{approx } N\left(EV_i, \frac{Var\, V_i}{n}\right).$$

Since

$$EV_i = E\frac{\partial}{\partial\theta}\log f(X_i|\theta)\Big|_{\theta=\theta_0} = 0,$$

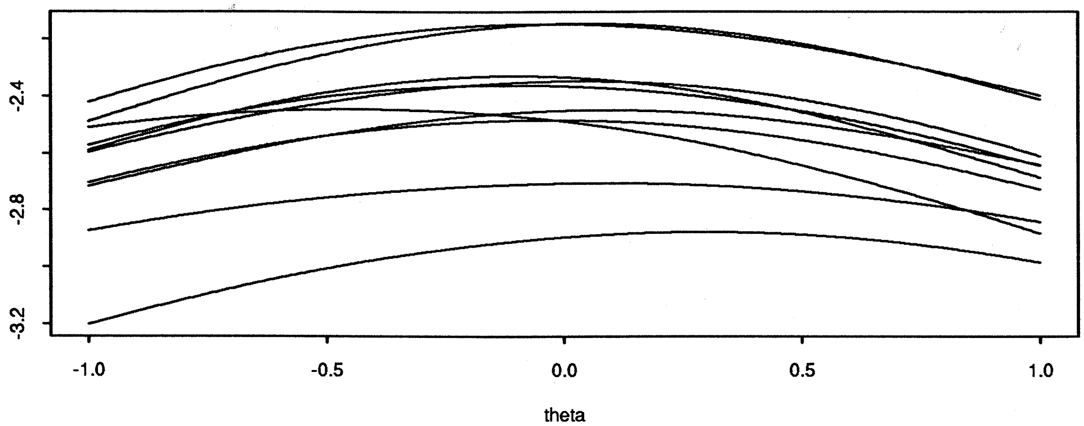
and $Var\, V_i = Var\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)\Big|_{\theta=\theta_0} = I(\theta_0)$,

ve have $\frac{1}{n}\dot{\ell}(\theta_0) \sim \text{approx } N\left(0, \frac{1}{n}I(\theta_0)\right)$.

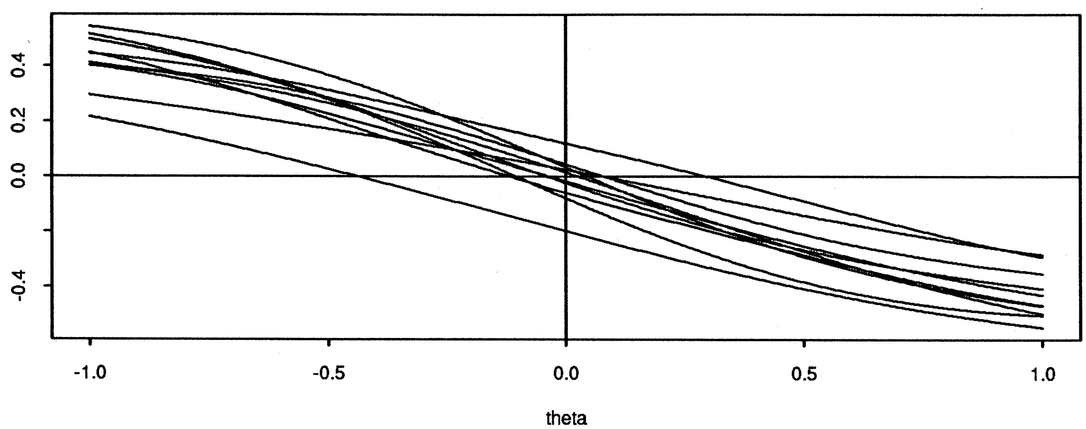
(There is much more.)

10 Samples of Size 50 from the Cauchy Distn.

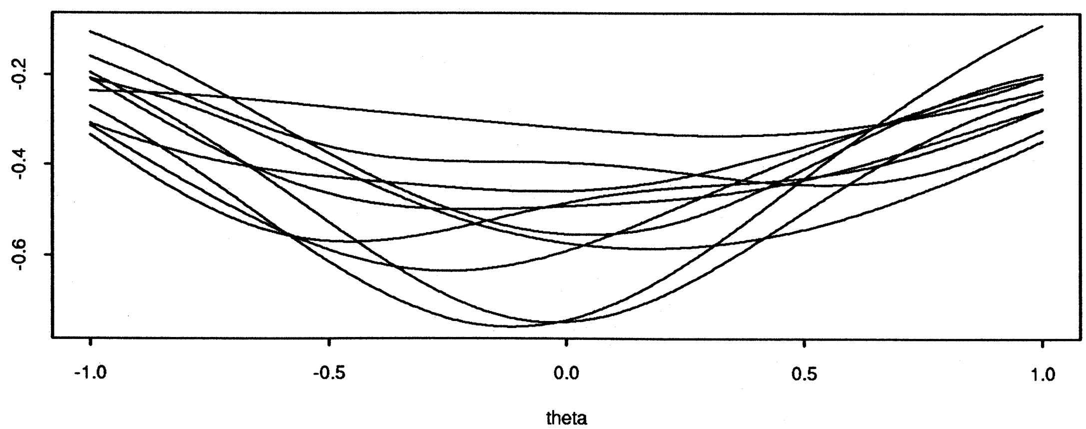
Log-likelihood



First Derivative of Log-likelihood



Second Derivative of Log-likelihood



Heuristic argument for asymptotic distn. of MLE

$X_1, X_2, \ldots, X_n$ IID $f(x|\theta_0)$.

true value

$\hat{\theta}$ denotes MLE.

$\ell(\theta)$ = log-likelihood function.

Notation

MLE $\hat{\theta}$ is the solution of $\frac{\partial \ell}{\partial \theta} = 0$.

Define $s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$ ($= \dot{\ell}(\theta)$ earlier)

Then $\hat{\theta}$ is solution of $s(\theta) = 0$.

For θ near θ_0

$$s(\theta) \approx s(\theta_0) + s'(\theta_0)(\theta - \theta_0).$$

If $\hat{\theta}$ is close to θ_0 (we can show $\hat{\theta}$ is a consistent estimator of θ),

$\hat{\theta}$ will be $\approx$ equal to the solution of

$$s(\theta_0) + s'(\theta_0)(\theta - \theta_0) = 0$$

which is

$$\theta - \theta_0 = \frac{-s(\theta_0)}{s'(\theta_0)}$$

$$\rightarrow \theta = \theta_0 - \frac{s(\theta_0)}{s'(\theta_0)}.$$

Thus $\hat{\theta} \approx \theta_0 - \dfrac{s(\theta_0)}{s'(\theta_0)}$.

$$s(\theta_0) = \frac{\partial l(\theta_0)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log f(X_i|\theta) \Big|_{\theta=\theta_0}$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i|\theta_0)$$

$$s'(\theta_0) = \frac{\partial s(\theta_0)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta_0).$$

$$\frac{s(\theta_0)}{s'(\theta_0)} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i|\theta_0) \longrightarrow \text{Apply CLT } ②}{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta_0) \longrightarrow \text{Apply LLN } ①}$$

Define $\mathcal{N}$ and $\mathcal{D}$ to be the numerator and denominator of this fraction.

Suppose n is large.

By the LLN,

$$\mathcal{D} \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(X_i \,|\, \theta_0) \approx E\left(\frac{\partial^2}{\partial\theta^2} \log f(X_1 \,|\, \theta_0)\right) = -I(\theta_0)\,.$$

Thus

$$\widehat{\theta} \approx \theta_0 - \frac{s(\theta_0)}{s'(\theta_0)} \;=\; \theta_0 - \frac{\mathcal{N}}{\mathcal{D}} \;\approx\; \theta_0 + I(\theta_0)^{-1}\mathcal{N}$$

Note that $\mathcal{N} = n^{-1}s(\theta_0) = n^{-1}\dot{\ell}(\theta_0)$.

As shown earlier using the CLT,

$$\mathcal{N} \sim \text{approx}\, N(0\,,\, n^{-1}I(\theta_0))\,.$$

Thus we conclude

$$\widehat{\theta} \sim \text{approx}\, N\left(\theta_0\,,\, \frac{1}{nI(\theta_0)}\right)\,.$$

Here we have used the fact:

$$X \sim N(\mu, \sigma^2) \;\;\Rightarrow\;\; aX + b \sim N(a\mu + b\,,\, a^2\sigma^2)\,.$$

Fisher Information, CR-bound, Asymp. distn. of MLE's in the Multi-parameter case

Notation: $\underset{\sim}{X} \sim f(\underset{\sim}{x} \mid \theta)$

$\theta = (\theta_1, \theta_2, \ldots, \theta_p)$

$$\frac{\partial}{\partial \theta} = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{pmatrix}$$

$$\underset{p \times 1 \text{ vector}}{S} = \text{vector of scores} = \frac{\partial}{\partial \theta} \log f(\underset{\sim}{X} \mid \theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \log f(\underset{\sim}{X} \mid \theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \log f(\underset{\sim}{X} \mid \theta) \end{pmatrix}$$

<u>Definition</u>

computed at θ
evaluated at θ } same θ!

$$\underset{p \times p \text{ matrix}}{I_{\underset{\sim}{X}}(\theta)} = E(\underbrace{\overset{p\times 1}{S}\ \overset{1 \times p}{S^t}}_{p \times p})$$

For a vector or matrix, we define expected value in this way:

$$E\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} EY \\ EZ \end{pmatrix}$$

$$E\begin{pmatrix} W & X \\ Y & Z \end{pmatrix} = \begin{pmatrix} EW & EX \\ EY & EZ \end{pmatrix}$$

<u>Properties</u>:

→ $\underset{p\times 1}{E_\theta S} = \underset{p\times 1}{0}$

→ $I_{\underset{\sim}{X}}(\theta) = Cov(S)$ (The variance-covariance matrix of S.)

→ If $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ has independent components, then

$$I_{\underset{\sim}{X}}(\theta) = \underset{p\times p}{I_{X_1}(\theta)} + \underset{p\times p}{I_{X_2}(\theta)} + \cdots + \underset{p\times p}{I_{X_n}(\theta)}$$

$\rightarrow$ If $X_1, X_2, \ldots, X_n$ are iid,

then $I_{\underset{\sim}{X}_n}(\theta) = n \underset{p \times p}{I_{X_1}(\theta)}$

$\rightarrow$ $I_{\underset{\sim}{X}}(\theta) = E\left(-\frac{\partial^2}{\partial \theta^2} \log f(\underset{\sim}{X} | \theta)\right)$

where we define

$$\frac{\partial^2}{\partial \theta^2} \log f(\underset{\sim}{X} | \theta) = \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\underset{\sim}{X} | \theta)\right)$$

= the $p \times p$ matrix whose (i, j) entry is $\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\underset{\sim}{X} | \theta)$.

Asymptotic distn. of MLE (of θ)

If $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$ is the sequence of MLE's (based on progressively larger samples), then

$$\hat{\theta}_n \sim \underset{\text{asymptotically}}{A} \underset{\text{multivariate normal}}{N}\left(\theta, \left(I_{\underset{\sim}{X}}(\theta)\right)^{-1}\right)$$

which means roughly that

$$\hat{\theta}_n \sim \text{approx } N\left(\theta, \left(I_{\underset{\sim}{X}}(\theta)\right)^{-1}\right)$$

for large n.

Recall: In iid case $I_{\underset{\sim}{X}}(\theta) = n\, I_{X_1}(\theta)$.

Estimate $I_{\underset{\sim}{X}}(\theta)$ by $I_{\underset{\sim}{X}}(\hat{\theta}_n)$ or

$$\left(-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\underset{\sim}{X} \mid \theta)\right)\Bigg|_{\theta = \hat{\theta}_n}$$

Multi-parameter CRLB

$\theta = (\theta_1, \dots, \theta_p)^T$.

① If $E\hat{\theta} = \theta$ (implies $E\hat{\theta}_i = \theta_i$ for all i)
then $Var(\hat{\theta}) - (I_{\underset{\sim}{X}}(\theta))^{-1}$ is non-negative definite.

so that $Var(\hat{\theta}_i) \geq (I^{-1})_{ii}$

where $I = I_{\underset{\sim}{X}}(\theta)$.

② If $E\hat{\theta}_i = \theta_i$ for a particular i,
then (by fixing θ_j for $j \neq i$ and using the CRLB for the one-parameter case)

$$Var(\hat{\theta}_i) \geq (I_{ii})^{-1} = \frac{1}{E\left(\frac{\partial}{\partial\theta_i} \log f(\underset{\sim}{X}|\theta)\right)^2}$$

③ If $E\hat{\theta} = \theta$, then

$$Var(\hat{\theta}_i) \geq (I^{-1})_{ii} \geq (I_{ii})^{-1}$$

Best you can do if ... ↓ All parameters are <u>un</u>known.

↓ θ_j are known for $j \neq i$.

Example: Normal$(\mu, \underset{\underset{\xi}{\uparrow}}{\sigma^2})$ dist.

$$f(x|\underbrace{\mu,\xi}_{\theta}) = \frac{1}{\sqrt{2\pi\xi}}\, e^{-(x-\mu)^2/2\xi}$$

$$\log f = -\frac{1}{2}\log(2\pi\xi) - \frac{(x-\mu)^2}{2\xi}$$

$$\underbrace{\begin{pmatrix} \frac{\partial}{\partial\mu}\log f \\ \frac{\partial}{\partial\xi}\log f \end{pmatrix}}_{\frac{\partial}{\partial\theta}\log f(X|\theta)} = \begin{pmatrix} \frac{X-\mu}{\xi} \\ -\frac{1}{2\xi} + \frac{(X-\mu)^2}{2\xi^2} \end{pmatrix}$$

$$\underbrace{\begin{pmatrix} \frac{\partial^2 l}{\partial\mu^2} & \frac{\partial^2 l}{\partial\mu\partial\xi} \\ \frac{\partial^2 l}{\partial\xi\partial\mu} & \frac{\partial^2 l}{\partial\xi^2} \end{pmatrix}}_{\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\right)^T l} = \begin{pmatrix} -\frac{1}{\xi} & -\frac{(X-\mu)}{\xi^2} \\ -\frac{(X-\mu)}{\xi^2} & \frac{1}{2\xi^2} - \frac{(X-\mu)^2}{\xi^3} \end{pmatrix}$$

$$I(\theta) = -E(\,\cdot\,) = \begin{pmatrix} \frac{1}{\xi} & 0 \\ 0 & \frac{1}{2\xi^2} \end{pmatrix}.$$

$$I^{-1} = \begin{pmatrix} \xi & 0 \\ 0 & 2\xi^2 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}$$

For an unbiased estimate of μ

$(E_{\mu,\sigma^2} W = \mu)$

$\text{Var}\, W \geq \frac{\sigma^2}{n}$ (achieved by $W = \bar{X}$)

For an unbiased estimate of σ^2

$\text{Var}\, W \geq \frac{2\sigma^4}{n}$ (not achieved exactly.

s^2 is best unbiased

and

$s^2 = \frac{\sigma^2}{n-1} \chi^2_{n-1}$

so that

$\text{Var}\, s^2 = \frac{2\sigma^4}{(n-1)}$.

$$\begin{pmatrix} \bar{X} \\ \hat{\sigma}^2 \end{pmatrix} \sim AN\left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} \right)$$

Limiting distn. of MLE

Note:

$$\mathrm{Var}\left(\frac{1}{n}\sum (X_i-\mu)^2\right) = \frac{2\sigma^4}{n}$$

$$E\,\frac{1}{n}\sum (X_i-\mu)^2 = \sigma^2.$$

Achieves CR-bound, but not legitimate estimator if μ is unknown.

<u>Example</u>: Gamma(α, β)

$$f(x|\underbrace{\alpha,\beta}_{\theta}) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}$$

$$\log f = -\log\Gamma(\alpha) - \alpha\log\beta + (\alpha-1)\log x - x/\beta$$

$$\begin{pmatrix} \frac{\partial}{\partial\alpha}\log f \\ \frac{\partial}{\partial\beta}\log f \end{pmatrix} = \begin{pmatrix} -\psi(\alpha) - \log\beta + \log X \\ -\frac{\alpha}{\beta} + \frac{X}{\beta^2} \end{pmatrix}$$

$$\begin{pmatrix} \frac{\partial^2}{\partial\alpha^2} & \frac{\partial}{\partial\alpha}\frac{\partial}{\partial\beta} \\ \frac{\partial^2}{\partial\alpha\partial\beta} & \frac{\partial}{\partial\beta} \end{pmatrix} \log f = \begin{pmatrix} -\psi'(\alpha) & -1/\beta \\ -1/\beta & \frac{\alpha}{\beta^2} - \frac{2X}{\beta^3} \end{pmatrix}$$

$$I(\theta) = -E(\ \ ) = \begin{pmatrix} +\psi'(\alpha) & +1/\beta \\ +1/\beta & \alpha/\beta^2 \end{pmatrix}$$

[Note: $EX = \alpha\beta$]

$$I^{-1}(\theta) = \begin{pmatrix} \alpha/\beta^2 & -1/\beta \\ -1/\beta & \psi'(\alpha) \end{pmatrix} \Big/ \left(\frac{\alpha\psi'(\alpha) - 1}{\beta^2}\right)$$

$$I^{-1}(\theta) = \begin{pmatrix} \alpha & -\beta \\ -\beta & \beta^2\psi'(\alpha) \end{pmatrix} \Big/ (\alpha\psi'(\alpha)-1)$$

CR bound for unbiased estimates of β.

$$\mathrm{Var}(\hat{\beta}) \geq \frac{1}{n}\left(I^{-1}(\theta)\right)_{22} \geq \frac{1}{n}\left(I(\theta)_{22}\right)^{-1}$$

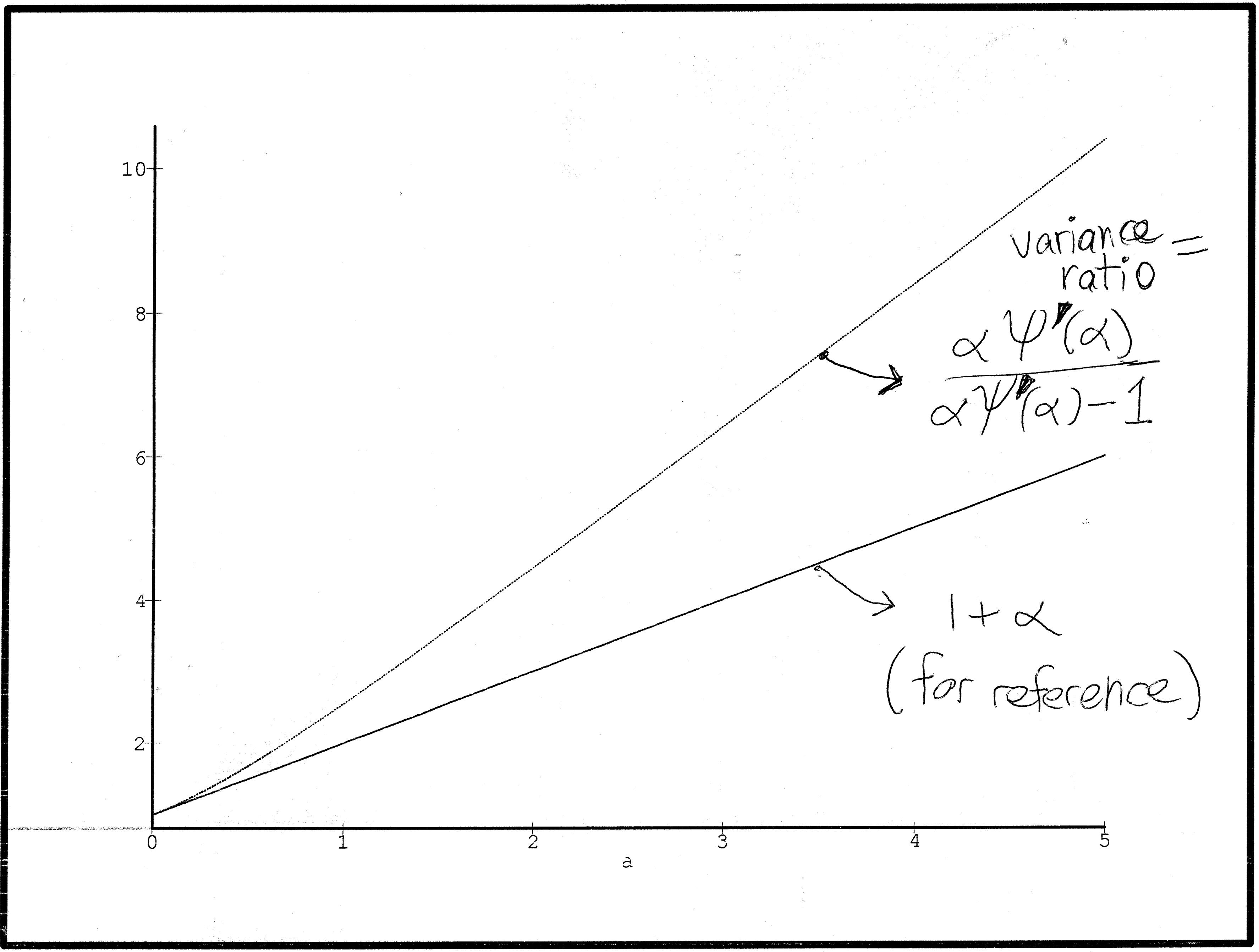
$$\geq \underbrace{\frac{1}{n}\left(\frac{\beta^2\psi'(\alpha)}{\alpha\psi'(\alpha)-1}\right)}_{\frac{\beta^2}{\alpha n}\cdot\left[\frac{\psi'(\alpha)}{\psi'(\alpha)-\frac{1}{\alpha}}\right]} > \underbrace{\frac{1}{n}\frac{1}{\alpha/\beta^2}}_{\frac{\beta^2}{\alpha n}}$$

If α is known, lower lower bound is achieved.

$$E\left(\frac{\overline{X}}{\alpha}\right) = \beta$$

$$\mathrm{Var}\left(\frac{\overline{X}}{\alpha}\right) = \frac{1}{\alpha^2}\frac{\mathrm{Var}\,X}{n} = \frac{\alpha\beta^2}{n\alpha^2} = \frac{\beta^2}{n\alpha}$$

If α must be estimated, there is a variance penalty which does <u>not</u> vanish asymptotically $(n\to\infty)$.



10
8
6
4
2
0
1
2
3
4
5
a
variance ratio =
$\frac{\alpha\psi'(\alpha)}{\alpha\psi'(\alpha)-1}$
$1+\alpha$
(for reference)