Statistical Data Analysis 2020/21
Lecture Week 7

London Postgraduate Lectures on Particle Physics

University of London MSc/MSci course PH4515

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Course web page via RHUL moodle (PH4515) and also

`www.pp.rhul.ac.uk/~cowan/stat_course.html`

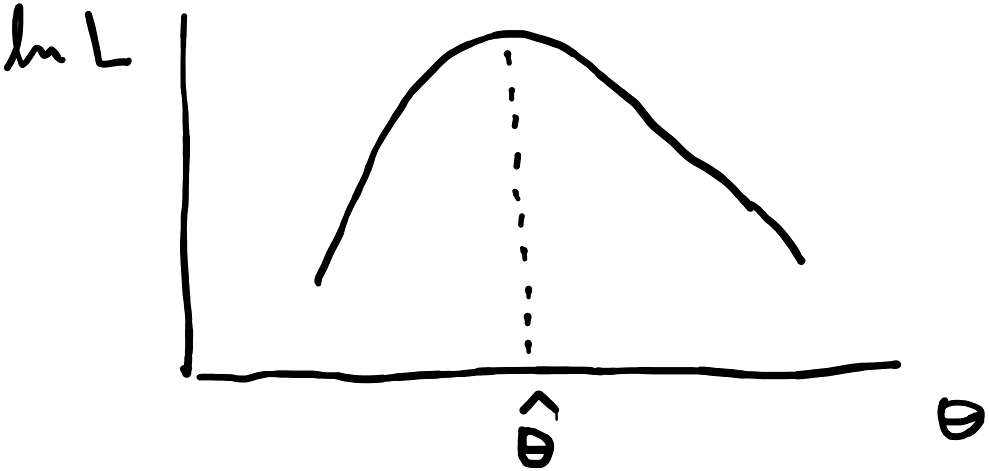
Statistical Data Analysis
Lecture 7-1

- Reminder of maximum likelihood
- The information inequality
- Large-sample properties of MLEs

Reminder of maximum likelihood

The estimators for parameters $\boldsymbol{\theta}$ are defined to be the values that maximize the likelihood function $L(\boldsymbol{\theta}) = P(\boldsymbol{x}|\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$$



Reminder of MLE for exponential

Exponential pdf, $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$, i.i.d. data $t_1, \ldots, t_n$

Likelihood function: $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

Log-likelihood function: $\ln L(\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$

Set $\frac{\partial \ln L}{\partial \tau} = \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0$ and solve for τ.

MLE: $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$ →

Bias: $b = E[\hat{\tau}] - \tau = 0$

Variance: $V[\hat{\tau}] = \frac{\tau^2}{n}$

The information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only MLE). For a single parameter:

$$(b = E[\hat{\theta}] - \theta)$$

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

= MVB (Minimum Variance Bound)

where $E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = \int \frac{\partial^2 \ln P(\mathbf{x}|\theta)}{\partial \theta^2} P(\mathbf{x}|\theta)\, d\mathbf{x}$

Proof in Exercise 6.6 of SDA, http://www.pp.rhul.ac.uk/~cowan/sda/prob/prob_6.pdf

"Efficiency" of an estimator = MVB / actual variance.

An estimator whose variance equals the MVB is said to be efficient.

MVB for MLE of exponential parameter

Find $$\mathrm{MVB} = -\left(1 + \frac{\partial b}{\partial \tau}\right)^2 \Bigg/ E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right]$$

We found for the exponential parameter the MLE $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$

and we showed $b = 0$, hence $\partial b/\partial\tau = 0$.

We find $$\frac{\partial^2 \ln L}{\partial \tau^2} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - \frac{2t_i}{\tau^3}\right)$$

and since $E[t_i] = \tau$ for all i, $E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right] = -\frac{n}{\tau^2}$,

and therefore $\mathrm{MVB} = \frac{\tau^2}{n} = V[\hat{\tau}]$. So here the MLE is efficient.

Large-sample (asymptotic) properties of MLEs

Suppose we have an i.i.d. data sample of size n: $x_1,\ldots,x_n$

In the large-sample (or "asymptotic") limit ($n \to \infty$) and assuming regularity conditions one can show that the likelihood and MLE have several important properties.

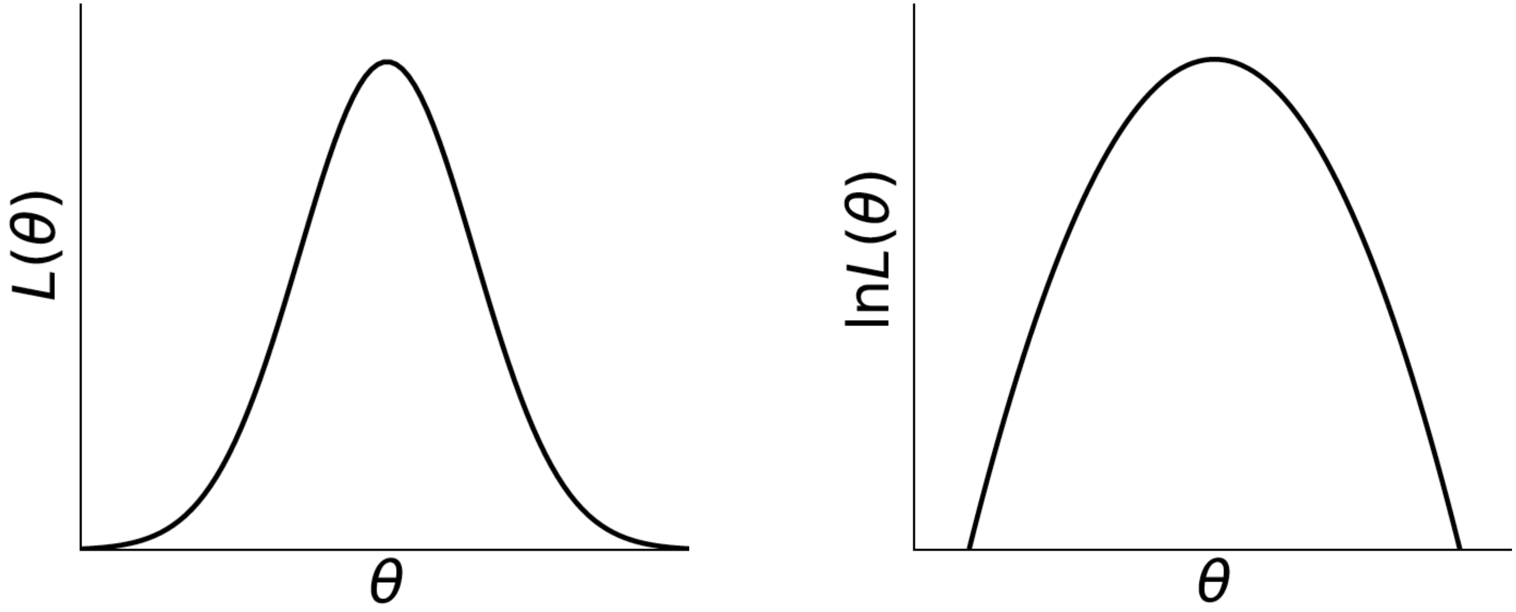
The regularity conditions include:

- the boundaries of the data space cannot depend on the parameter;
- the parameter cannot be on the edge of the parameter space;
- $\ln L(\theta)$ must be differentiable;
- the only solution to $\partial \ln L / \partial\theta = 0$ is $\hat{\theta}$.

In the slides immediately following the properties are shown without proof for a single parameter; the corresponding properties hold also for the multiparameter case, $\boldsymbol{\theta} = (\theta_1,\ldots,\ \theta_m)$.

log-likelihood becomes quadratic

The likelihood function becomes Gaussian in shape, i.e. the log-likelihood becomes quadratic (parabolic).



The MLE becomes increasingly precise as the (log)-likelihood becomes more tightly concentrated about its peak,

but $L(\theta) = P(\boldsymbol{x}|\theta)$ is the probability for $\boldsymbol{x}$, not a pdf for θ.

The MLE converges to the true parameter value

In the large-sample limit, the MLE converges in probability to the true parameter value.

That is, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

The MLE is said to be *consistent*.

MLE is asymptotically unbiased

In general the MLE can be biased, but in the large-sample limit, this bias goes to zero:

$$\lim_{n \to \infty} E[\hat{\theta}] - \theta = 0$$

(Recall for the exponential parameter we found the bias was identically zero regardless of the sample size n.)

The MLE's variance approaches the MVB

In the large-sample limit, the variance of the MLE approaches the minimum-variance bound, i.e., the information inequality becomes an equality (and bias goes to zero):

$$\lim_{n\to\infty} V[\hat{\theta}] = -\frac{1}{E\left[\frac{\partial^2 \ln L}{\partial\theta^2}\right]}$$

The MLE is said to be *asymptotically efficient*.

The MLE's distribution becomes Gaussian

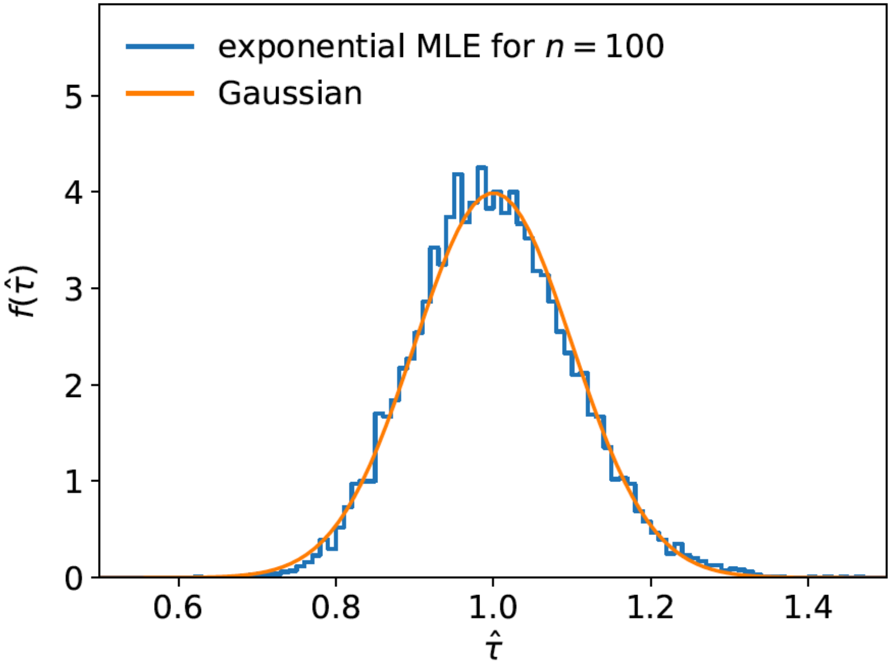
In the large-sample limit, the pdf of the MLE becomes Gaussian,

$$f(\hat{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{\theta}}} e^{-(\hat{\theta}-\theta)^2/2\sigma_{\hat{\theta}}^2}$$

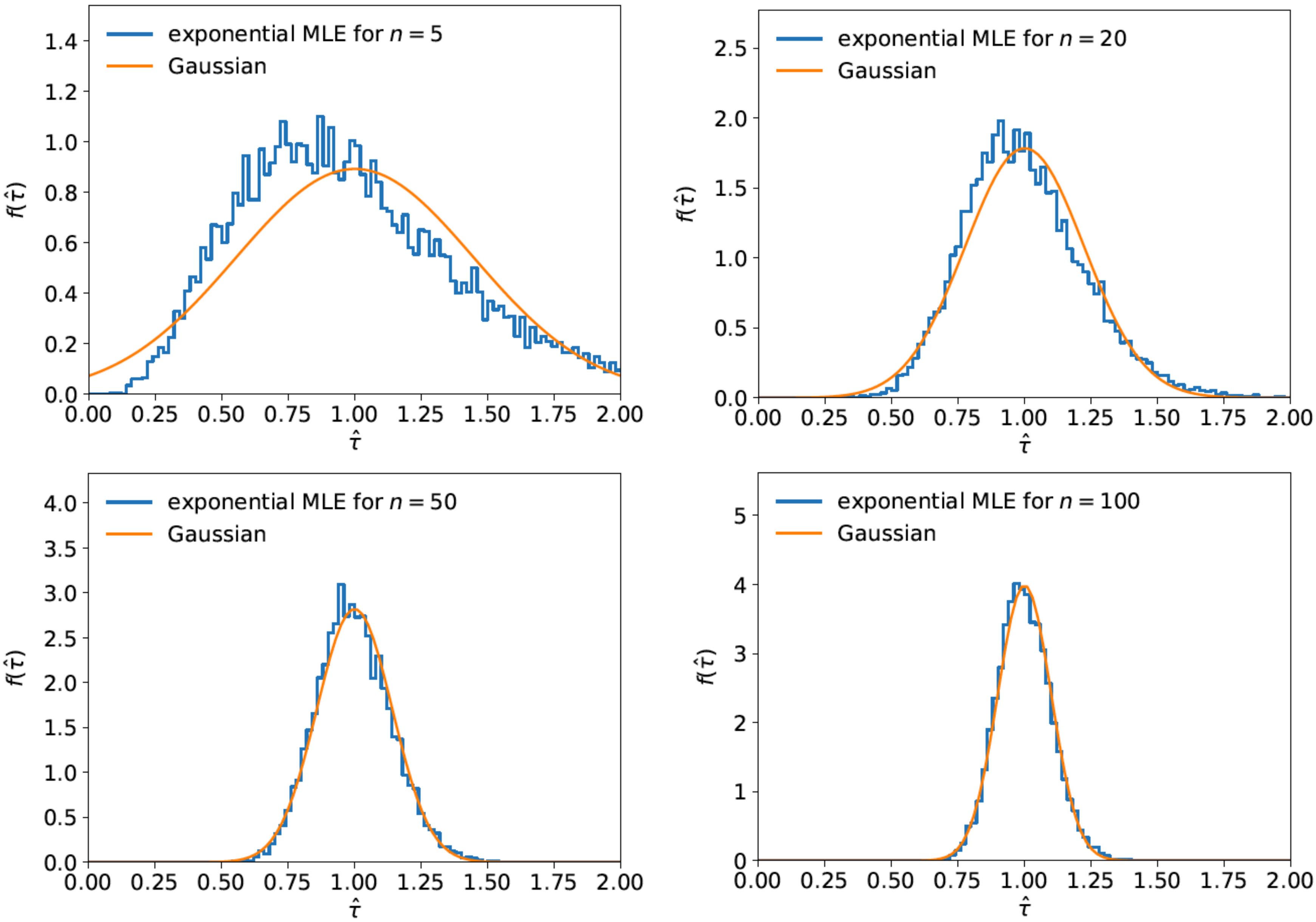
where $\sigma_{\hat{\theta}}^2$ is the minimum variance bound (note bias is zero).

For example, exponential MLE with sample size $n = 100$.

Note that for exponential, MLE is arithmetic average, so Gaussian MLE seen to stem from Central Limit Theorem.



Distribution of MLE of exponential parameter



Statistical Data Analysis
Lecture 7-2

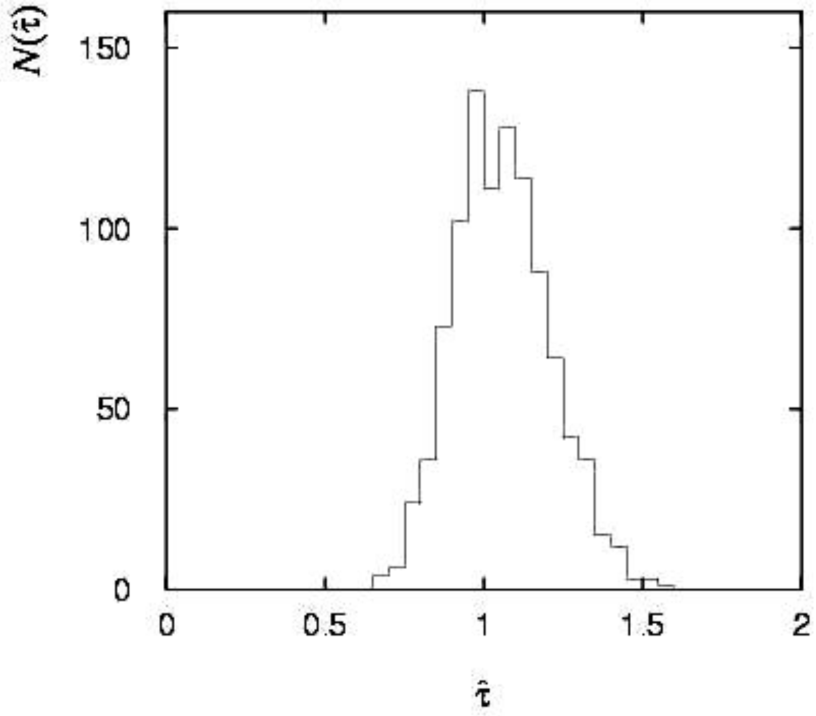
- Finding the variance of MLEs
- Information inequality for multiple parameters
- MLE for mean and variance of Gaussian

Variance of estimators: Monte Carlo method

Having estimated our parameter we now need to report its 'statistical error', using e.g. the estimator's standard deviation, or (co)variance.

It is usually not possible to do this with an exact calculation.

Another way is to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).



For exponential example (n=50), from sample variance of estimates we find:

$$\hat{\sigma}_{\hat{\tau}} = 0.151$$

Variance of estimators from information inequality

Recall the information inequality (RCF) sets a lower bound on the variance of any estimator (not only MLE):

MVB

$$V[\widehat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] \qquad (b = E[\widehat{\theta}] - \theta)$$

Often the bias b is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\widehat{\theta}] \approx -1 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of $\ln L$ at its maximum:

$$\widehat{V}[\widehat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\widehat{\theta}}$$

Variance of estimators: graphical method

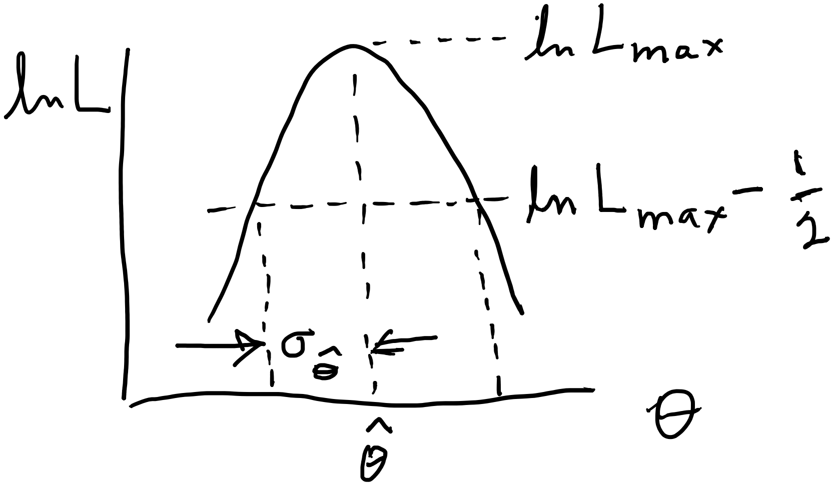
Expand $\ln L(\theta)$ about its maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

First term is $\ln L_{\max}$, second term is zero, for third term use information inequality (assume equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}}$$

$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

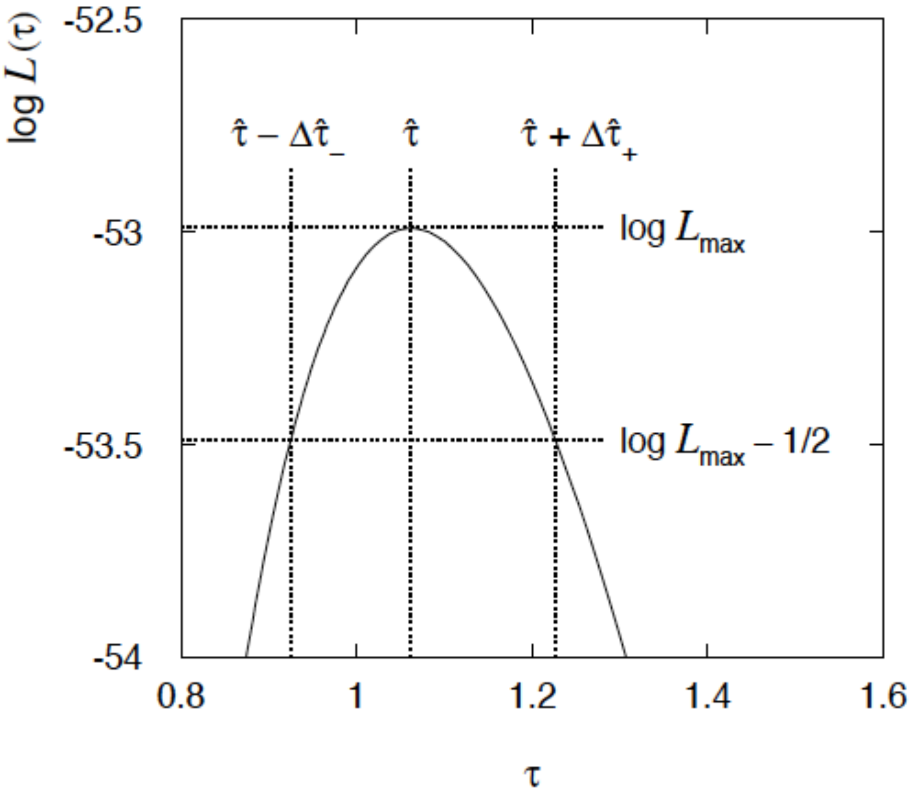


→ to get $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until $\ln L$ decreases by $1/2$.

Example of variance by graphical method

ML example with exponential:

$$\begin{aligned}\hat{\tau} &= 1.062 \\ \Delta\hat{\tau}_- &= 0.137 \\ \Delta\hat{\tau}_+ &= 0.165 \\ \hat{\sigma}_{\hat{\tau}} &\approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15\end{aligned}$$



Not quite parabolic $\ln L$ since finite sample size ($n = 50$).

Information inequality for N parameters

Suppose we have estimated N parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_N)$

The *Fisher information matrix* is

$$I_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} P(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}$$

and the covariance matrix of estimators $\hat{\boldsymbol{\theta}}$ is $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

The information inequality states that the matrix

$$M_{ij} = V_{ij} - \sum_{k,l} \left(\delta_{ik} + \frac{\partial b_i}{\partial\theta_k}\right) I^{-1}_{kl} \left(\delta_{lj} + \frac{\partial b_l}{\partial\theta_j}\right)$$

is positive semi-definite:

$\boldsymbol{z}^{\mathrm{T}} M \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \neq 0$, diagonal elements ≥ 0

Information inequality for N parameters (2)

In practice the inequality is ~always used in the large-sample limit:

bias → 0

inequality → equality, i.e, $M = 0$, and therefore $V^{-1} = I$

That is, $V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$

This can be estimated from data using $\widehat{V}_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\hat{\boldsymbol{\theta}}}$

Find the matrix V^{-1} numerically (or with automatic differentiation), then invert to get the covariance matrix of the estimators

$$\widehat{V}_{ij} = \widehat{\mathrm{cov}}[\hat{\theta}_i, \hat{\theta}_j]$$

Example of MLE: parameters of Gaussian pdf

Consider independent $x_1, \ldots, x_n$, with $x_i \sim \mathrm{Gauss}(\mu, \sigma^2)$

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

The log-likelihood function is

$$\begin{aligned} \ln L(\mu,\sigma^2) &= \sum_{i=1}^{n} \ln f(x_i;\mu,\sigma^2) \\ &= \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \ln \frac{1}{\sigma^2} - \frac{(x_i-\mu)^2}{2\sigma^2} \right) . \end{aligned}$$

Example of ML: parameters of Gaussian pdf (2)

Set derivatives with respect to μ, σ^2 to zero and solve,

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \, , \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2 \, .$$

We already know that the estimator for μ is unbiased.

But we find, however, $E[\widehat{\sigma^2}] = \frac{n-1}{n}\sigma^2$, so the MLE for σ^2 has a bias, but $b \rightarrow 0$ for $n \rightarrow \infty$. Recall, however, that

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

is an unbiased estimator for σ^2. Usually not important whether one uses s^2 or the MLE to estimate σ^2.

Example of ML: parameters of Gaussian pdf (3)

Use 2nd derivatives of $\ln L$ to find covariance.

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2} \quad \longrightarrow \quad E\left[\frac{\partial^2 \ln L}{\partial \mu^2}\right] = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\longrightarrow \quad E\left[\frac{\partial^2 \ln L}{\partial (\sigma^2)^2}\right] = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n} E[(x_i - \mu)^2] = -\frac{n}{2\sigma^4}$$

$$\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) \quad \longrightarrow \quad E\left[\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu}\right] = -\frac{1}{\sigma^4}\sum_{i=1}^{n} E[x_i - \mu] = 0$$

Example of ML: parameters of Gaussian pdf (4)

So the Fisher information matrix is

$$I_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \approx V_{ij}^{-1}$$

Invert to find covariance matrix $V \approx I^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$

That is, $V[\hat{\mu}] = \dfrac{\sigma^2}{n}\,, \quad V[\widehat{\sigma^2}] = \dfrac{2\sigma^4}{n}\,, \quad \mathrm{cov}[\hat{\mu}, \widehat{\sigma^2}] = 0\,.$

From error prop., $V[\hat{\sigma}] = \left(\dfrac{\partial\hat{\sigma}}{\partial\widehat{\sigma^2}}\right)^2 \Bigg|_{\widehat{\sigma^2}=\sigma^2} V\left[\widehat{\sigma^2}\right] = \dfrac{\sigma^2}{2n} \longrightarrow \sigma_{\hat{\sigma}} = \dfrac{\sigma}{\sqrt{2n}}$

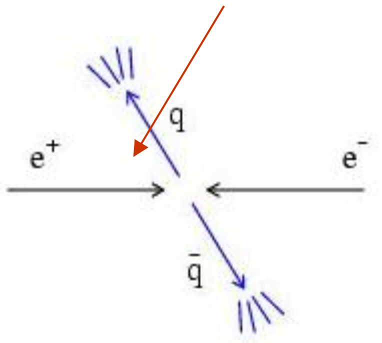
Statistical Data Analysis
Lecture 7-3

- Numerical example of 2-D MLE
- The $\ln L = \ln L_{\max} - ½$ contour
- MLE for function of a parameter

Example of ML with 2 parameters

Consider a scattering angle distribution with $x = \cos\theta$,

$$f(x;\alpha,\beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta/3}$$



or if $x_{\min} < x < x_{\max}$, need to normalize so that

$$\int_{x_{\min}}^{x_{\max}} f(x;\alpha,\beta)\, dx = 1 \; .$$

Example: $\alpha = 0.5$, $\beta = 0.5$, $x_{\min} = -0.95$, $x_{\max} = 0.95$, generate $n = 2000$ events with Monte Carlo.

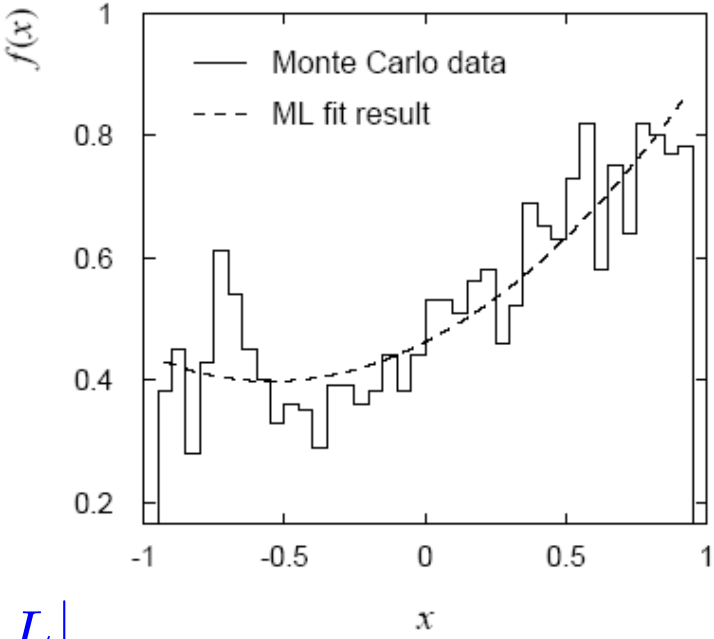
$$\ln L(\alpha,\beta) = \sum_{i=1}^{n} \ln f(x_i;\alpha,\beta) \longleftarrow$$

need to find maximum numerically

Example of ML with 2 parameters: fit result

Finding maximum of $\ln L(\alpha, \beta)$ numerically gives

$$\begin{aligned}\hat{\alpha} &= 0.508 \\ \hat{\beta} &= 0.47\end{aligned}$$



N.B. No binning of data for fit, but can compare to histogram for goodness-of-fit (e.g. 'visual' or χ^2).

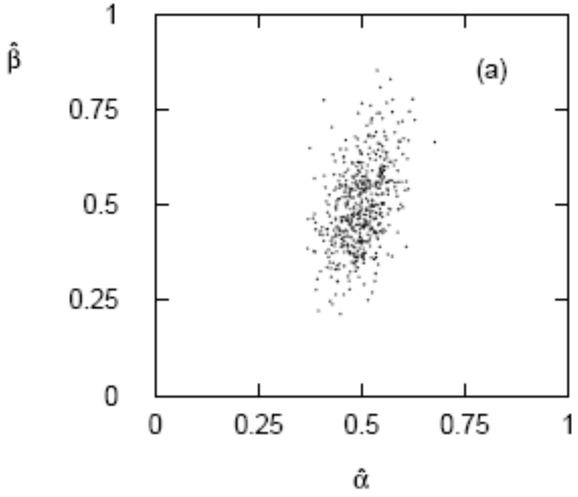
(Co)variances from $(\widehat{V^{-1}})_{ij} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\hat{\vec{\theta}}}$

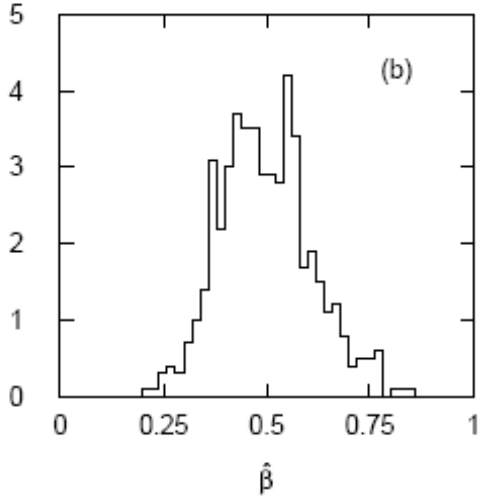
$$\begin{aligned}\hat{\sigma}_{\hat{\alpha}} &= 0.052 & \mathrm{cov}[\hat{\alpha}, \hat{\beta}] &= 0.0026 \\ \hat{\sigma}_{\hat{\beta}} &= 0.11 & r &= 0.46\end{aligned}$$

= correlation coefficient

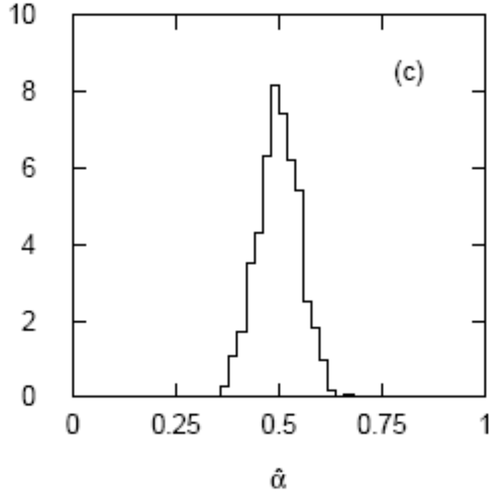
Two-parameter fit: MC study

Repeat ML fit with 500 experiments, all with $n = 2000$ events:





$$\begin{aligned} \overline{\hat{\alpha}} &= 0.499 \\ s_{\hat{\alpha}} &= 0.051 \\ \overline{\hat{\beta}} &= 0.498 \\ s_{\hat{\beta}} &= 0.111 \\ \widehat{\mathrm{cov}}[\hat{\alpha}, \hat{\beta}] &= 0.0024 \\ r &= 0.42 \end{aligned}$$



Estimates average to ~true values;
(Co)variances close to previous estimates;
marginal pdfs approximately Gaussian.

Multiparameter graphical method for variances

Expand $\ln L(\boldsymbol{\theta})$ to 2nd order about MLE:

$$\ln L(\boldsymbol{\theta}) \approx \ln L(\hat{\boldsymbol{\theta}}) + \sum_i \left.\frac{\partial \ln L}{\partial \theta_i}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i) + \frac{1}{2!} \sum_{i,j} \left.\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

$\ln L_{\max}$

zero

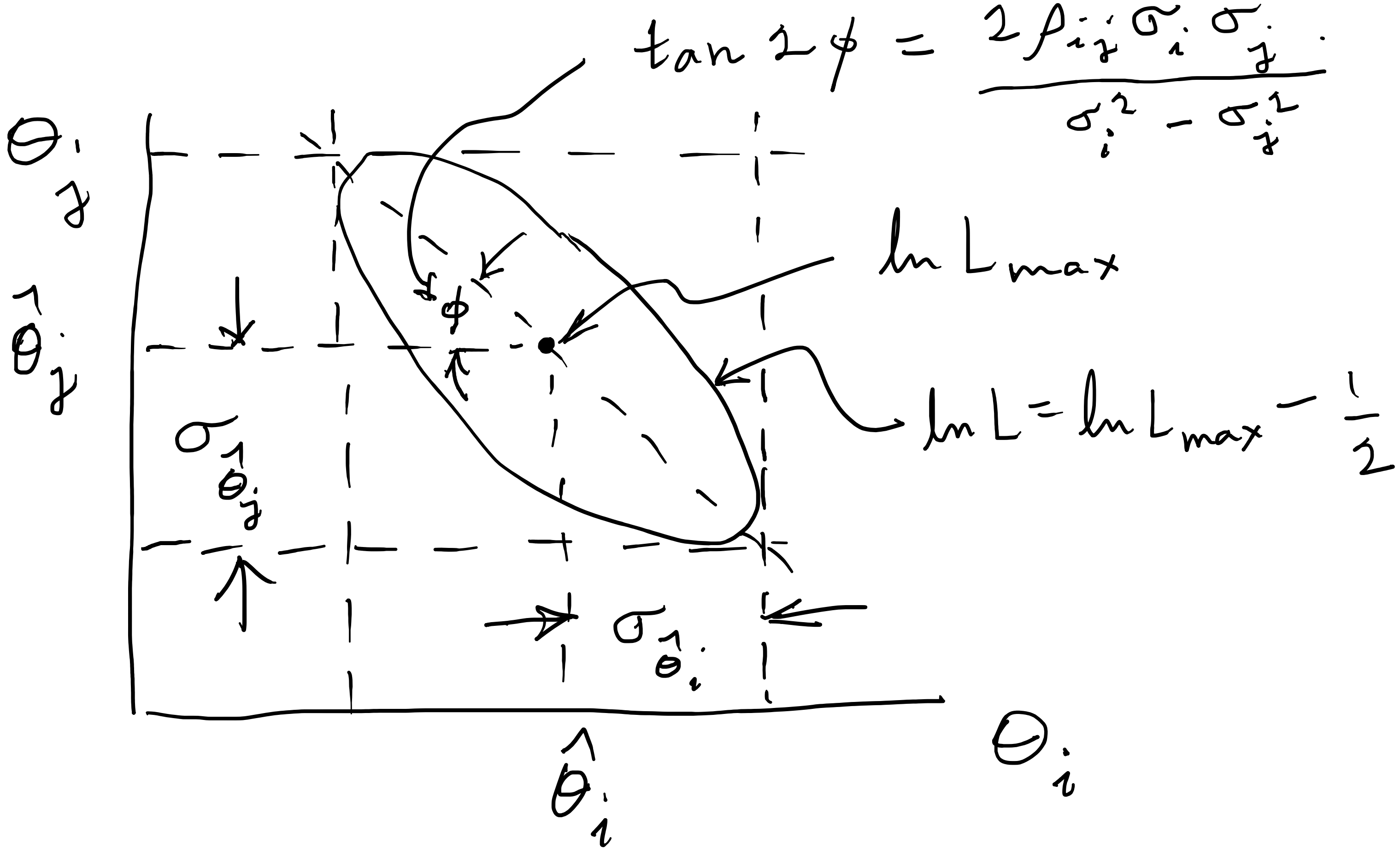
relate to covariance matrix of MLEs using information (in)equality.

Result: $$\ln L(\boldsymbol{\theta}) = \ln L_{\max} - \frac{1}{2} \sum_{i,j} (\theta_i - \hat{\theta}_i) V_{ij}^{-1} (\theta_j - \hat{\theta}_j)$$

So the surface $\ln L(\boldsymbol{\theta}) = \ln L_{\max} - \dfrac{1}{2}$ corresponds to

$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T V^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = 1$, which is the equation of a (hyper-) ellipse.

Multiparameter graphical method (2)


$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2}$.
θ_j
$\hat{\theta}_j$
ϕ
$\ln L_{max}$
$\ln L = \ln L_{max} - \frac{1}{2}$
$\sigma_{\hat{\theta}_j}$
$\sigma_{\hat{\theta}_i}$
$\hat{\theta}_i$
θ_i

Distance from MLE to tangent planes gives standard deviations.

The ln $L_{\max}$ – 1/2 contour for two parameters

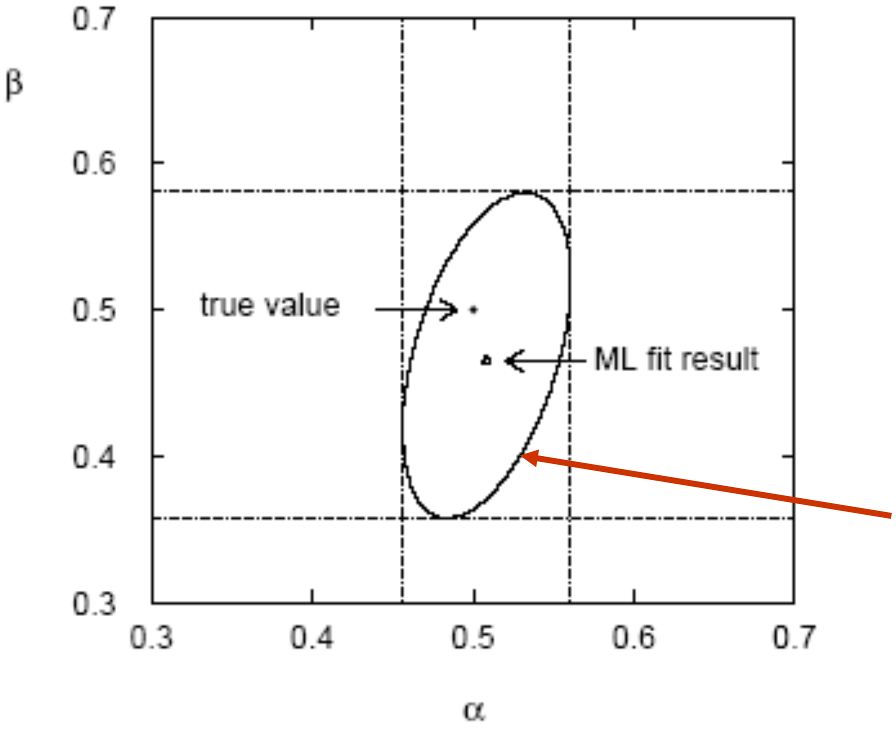
For large n, $\ln L$ takes on quadratic form near maximum:

$$\ln L(\alpha, \beta) \approx \ln L_{\max}$$
$$-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)^2+\left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)^2-2\rho\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)\left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)\right]$$

The contour $\ln L(\alpha, \beta) = \ln L_{\max} - 1/2$ is an ellipse:

$$\frac{1}{(1-\rho^2)}\left[\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)^2+\left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)^2-2\rho\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)\left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)\right]=1$$

(Co)variances from ln *L* contour



The α, β plane for the first MC data set

$$\ln L(\alpha, \beta) = \ln L_{\max} - 1/2$$

$\longrightarrow$ Tangent lines to contours give standard deviations.

$\longrightarrow$ Angle of ellipse φ related to correlation: $\tan 2\phi = \dfrac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma^2_{\hat{\alpha}} - \sigma\hat{\beta}^2}$

Functions of maximum-likelihood estimators

Suppose likelihood has a parameter θ.

Define a new parameter α given by function $\alpha = a(\theta)$.

What is the MLE of α?

For now suppose $a(\theta)$ has a unique inverse, so $\theta = a^{-1}(\alpha)$.

The likelihood is $L(\theta) = L(a^{-1}(\alpha))$.

The maximum of the likelihood is $L_{\max} = L(\hat{\theta})$.

So to maximize L, find $\alpha \equiv \hat{\alpha}$ such that

$$a^{-1}(\hat{\alpha}) = \hat{\theta} \quad \longrightarrow \quad \hat{\alpha} = a(\hat{\theta})$$

MLE of a function is the function of the MLE.

Still works when function is not one-to-one. Very useful result.

Functions of MLEs: exponential example

Suppose we had written the exponential pdf as $f(t;\lambda) = \lambda e^{-\lambda t}$, i.e., we use $\lambda = 1/\tau$. What is the MLE estimator for λ?

For the decay constant we have $$\hat{\lambda} = \frac{1}{\hat{\tau}} = \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right)^{-1} .$$

Caveat: $\hat{\lambda}$ is biased, even though $\hat{\tau}$ is unbiased.

Can show $E[\hat{\lambda}] = \lambda \frac{n}{n-1} .$ (bias →0 for n →∞)

In general MLE for a function of an unbiased estimator stays unbiased only for a linear function.

Statistical Data Analysis
Lecture 7-4

- Extended maximum likelihood
- Maximum likelihood with a histogram of data
- Relationship between MLE and Bayesian estimator

Extended ML

We observe n independent values of $x \sim f(x;\boldsymbol{\theta})$.

Suppose we regard n not as fixed, but as a Poisson r.v., mean ν.

Result of experiment defined as: $n, x_1, \ldots, x_n$.

$P(n,\boldsymbol{x}) = P(n)P(\boldsymbol{x}|n)$, so the (extended) likelihood function is:

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i; \vec{\theta})$$

Suppose theory gives $\nu = \nu(\boldsymbol{\theta})$, then the log-likelihood is

$$\ln L(\vec{\theta}) = -\nu(\vec{\theta}) + \sum_{i=1}^{n} \ln(\nu(\vec{\theta}) f(x_i; \vec{\theta})) + C$$

where C represents terms not depending on $\boldsymbol{\theta}$.

Extended ML (2)

Example: expected number of events $\nu(\vec{\theta}) = \sigma(\vec{\theta}) \int L\, dt$ where the total cross section $\sigma(\boldsymbol{\theta})$ is predicted as a function of the parameters of a theory, as is the distribution of a variable x.

Extended MLE uses information both from the number of events n as well as the observed values of x.

→ smaller errors for $\hat{\vec{\theta}}$ (compared to using x alone).

If ν does not depend on $\boldsymbol{\theta}$ but remains a free parameter, extended ML gives:

$$\hat{\nu} = n$$

$$\hat{\theta} = \text{same as ML}$$

ML with binned data

Often put data into a histogram: $\vec{n} = (n_1, \ldots, n_N), \; n_{\mathrm{tot}} = \sum_{i=1}^{N} n_i$

Hypothesis specifies $\vec{\nu} = (\nu_1, \ldots, \nu_N), \; \nu_{\mathrm{tot}} = \sum_{i=1}^{N} \nu_i$ where

$$\nu_i(\vec{\theta}) = \nu_{\mathrm{tot}} \int_{\mathrm{bin}\; i} f(x; \vec{\theta}) \, dx$$

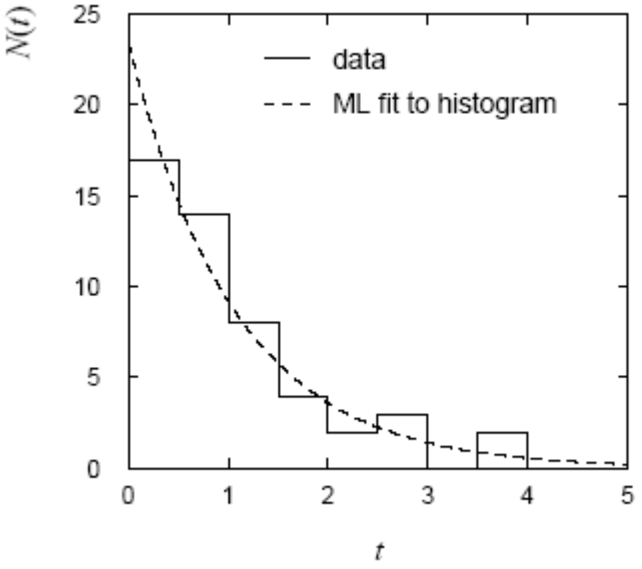
If we model the data as multinomial (n_{tot} constant),

$$f(\vec{n}; \vec{\nu}) = \frac{n_{\mathrm{tot}}!}{n_1! \ldots n_N!} \left(\frac{\nu_1}{n_{\mathrm{tot}}} \right)^{n_1} \cdots \left(\frac{\nu_N}{n_{\mathrm{tot}}} \right)^{n_N}$$

then the log-likelihood function is: $\ln L(\vec{\theta}) = \sum_{i=1}^{N} n_i \ln \nu_i(\vec{\theta}) + C$

Example with binned data

Previous example with exponential, now put data into histogram:



$\hat{\tau} = 1.07 \pm 0.17$

$(1.06 \pm 0.15$ for unbinned ML with same sample)

Binning results in loss of information, increased std. dev. of MLE.

Limit of zero bin width → usual unbinned MLE.

If n_i treated as Poisson, we get extended log-likelihood:

$$\ln L(\nu_{\text{tot}}, \vec{\theta}) = -\nu_{\text{tot}} + \sum_{i=1}^{N} n_i \ln \nu_i(\nu_{\text{tot}}, \vec{\theta}) + C$$

Relationship between ML and Bayesian estimators

Recall the Bayesian approach:

Both $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are random variables.

Use subjective probability for hypotheses (θ);

before experiment, knowledge summarized by prior pdf $\pi(\theta)$;

use Bayes' theorem to update prior in light of data:

$$p(\theta|\vec{x}) = \frac{L(\vec{x}|\theta)\pi(\theta)}{\int L(\vec{x}|\theta')\pi(\theta')\,d\theta'}$$

Posterior pdf (conditional pdf for θ given $\boldsymbol{x}$)

ML and Bayesian estimators (2)

Purist Bayesian: $p(\theta|x)$ contains all knowledge about θ.

Pragmatist Bayesian: $p(\theta|x)$ could be a complicated function,

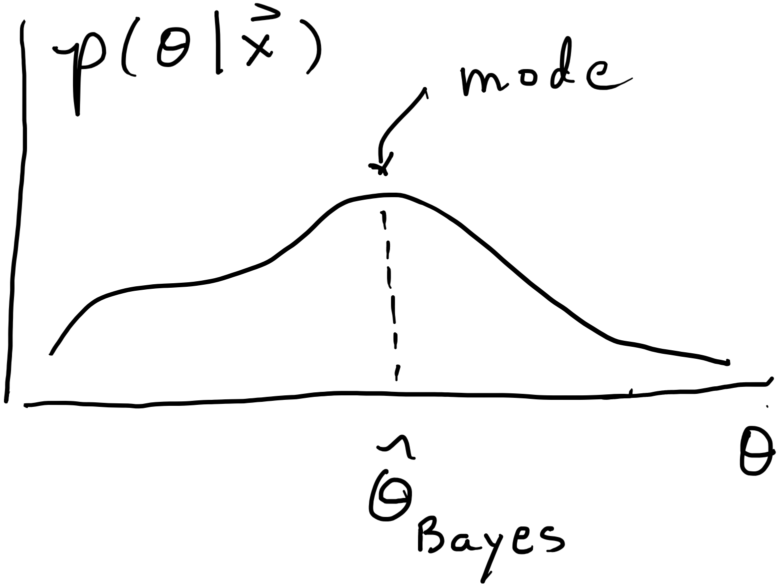
→ summarize using an estimator $\hat{\theta}_{\mathrm{Bayes}}$

Take mode of $p(\theta|x)$, (could also use e.g. expectation value)

What do we use for $\pi(\theta)$?
No golden rule (subjective!),
often represent ‘prior ignorance’
by $\pi(\theta) =$ constant, in which case

$$p(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta) = L(\theta)$$

$$\longrightarrow \quad \hat{\theta}_{\mathrm{Bayes}} = \hat{\theta}_{\mathrm{ML}}$$



ML and Bayesian estimators (3)

Note $\pi_\theta(\theta)$ = const. cannot be normalized – "improper prior".

Can be allowed for some problems; prior always appears multiplied by likelihood, so product $L(\theta)\pi_\theta(\theta)$ can result in normalizable posterior probability.

But... we could have used a different parameter, e.g., $\lambda = 1/\theta$, and if prior $\pi_\theta(\theta)$ is constant, then $\pi_\lambda(\lambda)$ is not:

$$\pi_\lambda(\lambda) = \pi_\theta(\theta) \left| \frac{d\theta}{d\lambda} \right| \propto \frac{1}{\lambda^2}$$

Maybe we know say we nothing about λ, so take $\pi_\lambda(\lambda) = $ const.

Then $\hat{\lambda}_{\rm Bayes} = \hat{\lambda}_{ML} \neq \frac{1}{\hat{\theta}_{\rm Bayes}}$

'Complete prior ignorance' is not well defined.

Extra slides

MLE for number of taxis

The number plate of taxis in every canton in Switzerland ends with a number N from 1 to N_{tot}, where N_{tot} is the total number of taxis.

Model the probability for observing plate number N with

$$P(N|N_{\mathrm{tot}}) = \frac{1}{N_{\mathrm{tot}}}, \quad 1 \leq N \leq N_{\mathrm{tot}}$$

MLE for N_{tot}

Suppose you observe one taxi at random with plate number N.

The likelihood function is $$L(N_{\text{tot}}) = \frac{1}{N_{\text{tot}}} \,, \qquad N_{\text{tot}} \geq N$$

which is maximized for $\widehat{N}_{\text{tot}} = N$

The expectation value and bias of the MLE are

$$E[\widehat{N}_{\text{tot}}] = E[N] = \sum_{N=1}^{N_{\text{tot}}} \frac{N}{N_{\text{tot}}} = \frac{N_{\text{tot}} + 1}{2} \qquad\qquad b = \frac{1 - N_{\text{tot}}}{2}$$

For better estimators, see similar problem with tanks in WW2:
https://en.wikipedia.org/wiki/German_tank_problem

E.g. the minimum-variance unbiased estimator is: $\widehat{N}_{\text{tot}} = 2N - 1$

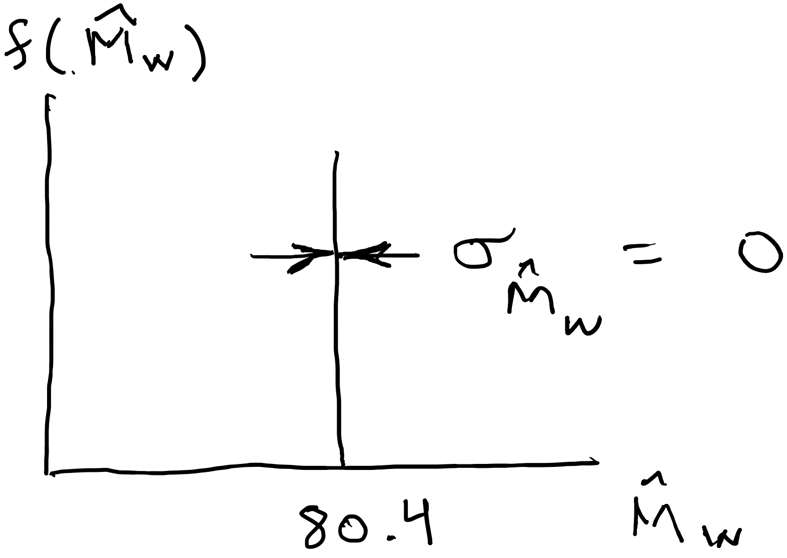
Cheap estimator for mass of W boson

The Particle Physics community has spent huge sums trying to estimate the mass of the W boson with the smallest possible statistical and systematic uncertainty.

Here is an estimator with zero statistical uncertainty. And it's free!

$$\widehat{M}_W = 80.4\,\mathrm{GeV}$$

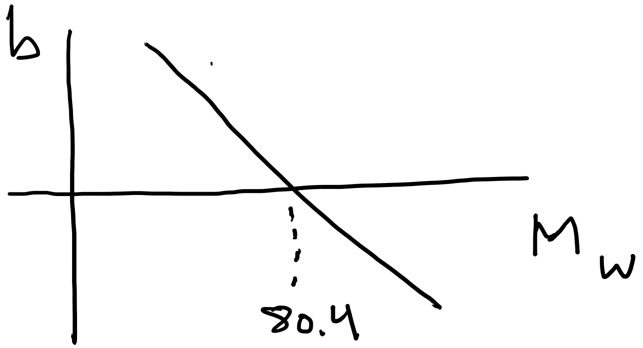
Here is its sampling distribution:



Does this violate the information inequality?

Cheap estimator for mass of W boson (2)

This estimator’s bias is $b = E[\widehat{M_W}] - M_W \;= 80.4\,\mathrm{GeV} \;- M_W$



Note current best estimate of M_W is 80.379±0.012 GeV, so the numerical value of the bias may be fairly small.

But we have $\dfrac{\partial b}{\partial M_W} = -1$ and so

$$\mathrm{MVB} = -\left(1 + \frac{\partial b}{\partial M_W}\right)^2 \Big/ E\left[\frac{\partial^2 \ln L}{\partial M_W^2}\right] = 0$$

So the information inequality is still satisfied.

Extended ML example

Consider two types of events (e.g., signal and background) each of which predict a given pdf for the variable x: $f_s(x)$ and $f_b(x)$.

We observe a mixture of the two event types, signal fraction = θ, expected total number = ν, observed total number = n.

Let $\mu_{\rm s} = \theta\nu$, $\mu_{\rm b} = (1-\theta)\nu$, goal is to estimate μ_s, μ_b.

$$f(x;\mu_{\rm s},\mu_{\rm b}) = \frac{\mu_{\rm s}}{\mu_{\rm s}+\mu_{\rm b}} f_{\rm s}(x) + \frac{\mu_{\rm b}}{\mu_{\rm s}+\mu_{\rm b}} f_{\rm b}(x)$$

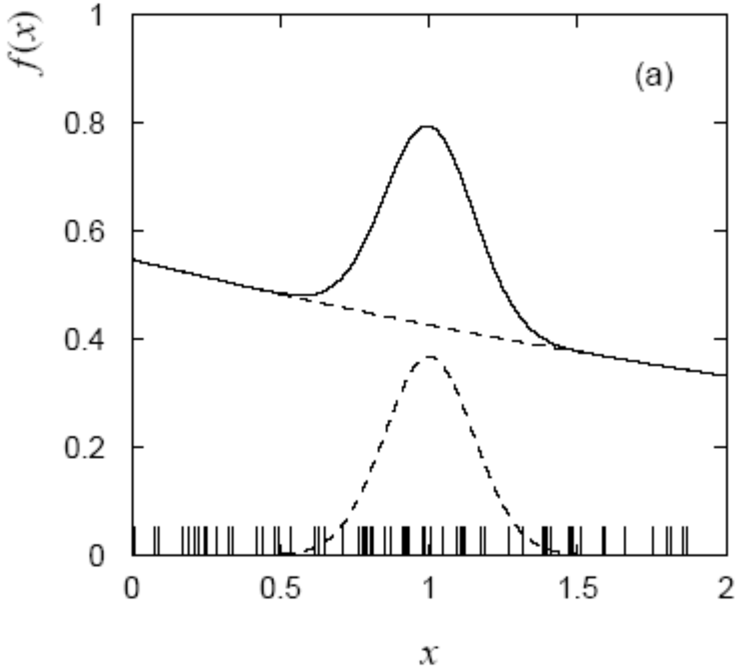
$$P(n;\mu_{\rm s},\mu_{\rm b}) = \frac{(\mu_{\rm s}+\mu_{\rm b})^n}{n!} e^{-(\mu_{\rm s}+\mu_{\rm b})}$$

$$\rightarrow \quad \ln L(\mu_{\rm s},\mu_{\rm b}) = -(\mu_{\rm s}+\mu_{\rm b}) + \sum_{i=1}^{n} \ln\left[(\mu_{\rm s}+\mu_{\rm b}) f(x_i;\mu_{\rm s},\mu_{\rm b})\right]$$

Extended ML example (2)

Monte Carlo example with combination of exponential and Gaussian:

$$\begin{aligned} \mu_{\rm s} &= 6 \\ \mu_{\rm b} &= 60 \end{aligned}$$



Maximize log-likelihood in terms of μ_s and μ_b:

$$\begin{aligned} \hat{\mu}_{\rm s} &= 8.7 \pm 5.5 \\ \hat{\mu}_{\rm b} &= 54.3 \pm 8.8 \end{aligned}$$

Here errors reflect total Poisson fluctuation as well as that in proportion of signal/background.