Statistics for Particle Physicists Lecture 1

Academic Training Lectures
CERN / Zoom
21-24 June 2021

https://indico.cern.ch/event/1040092/

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

Outline

Lecture 1: Introduction, probability, parameter estimation

Lecture 2: Hypothesis tests, limits

Lecture 3: Systematic uncertainties, experimental sensitivity

Lecture 4: Bayesian methods, Student's t regression

Almost everything is a subset of the University of London course:

http://www.pp.rhul.ac.uk/~cowan/stat_course.html

Some statistics books, papers, etc.

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998

R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989

Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014.

Luca Lista, *Statistical Methods for Data Analysis in Particle Physics*, Springer, 2017.

L. Lyons, *Statistics for Nuclear and Particle Physics*, CUP, 1986

F. James., *Statistical and Computational Methods in Experimental Physics*, 2nd ed., World Scientific, 2006

S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998.

P.A. Zyla et al. (Particle Data Group), Prog. Theor. Exp. Phys. 2020, 083C01 (2020); `pdg.lbl.gov` sections on probability, statistics, MC.

Theory ↔ Statistics ↔ Experiment
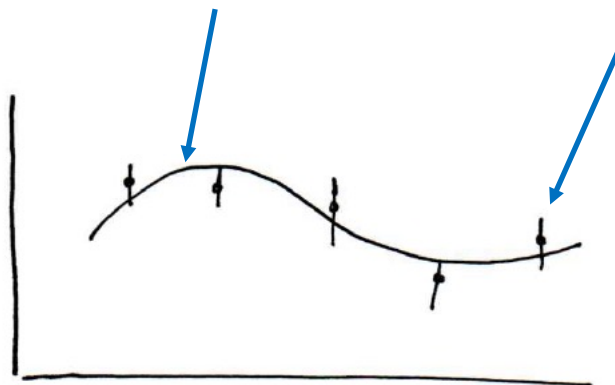
Theory (model, hypothesis):

Experiment (observation):
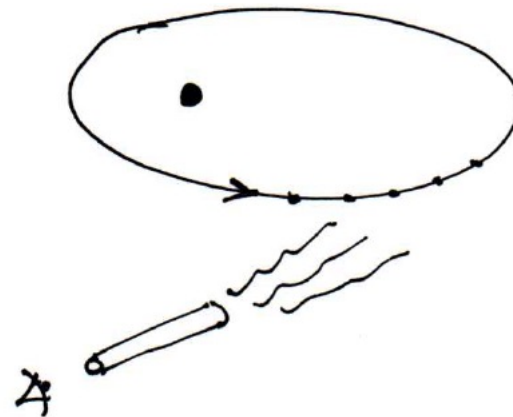
$$F = -G \frac{m_1 m_2}{r^2} \quad , \quad \ldots$$

+ response of measurement apparatus

= model prediction

data

Uncertainty enters on many levels
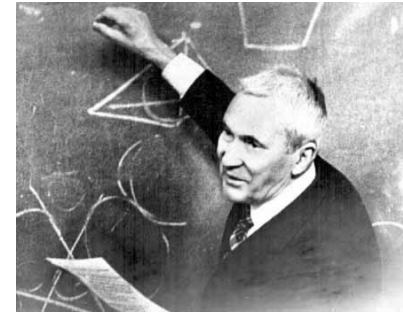
→ quantify with probability

A definition of probability

Consider a set S with subsets $A, B, \ldots$

For all $A \subset S, P(A) \geq 0$

$$P(S) = 1$$

If $A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$

Kolmogorov axioms (1933)

Also define conditional probability of A given B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Subsets A, B independent if: $\quad P(A \cap B) = P(A)P(B)$

If A, B independent, $\quad P(A|B) = \dfrac{P(A)P(B)}{P(B)} = P(A)$

Interpretation of Probability

I. Relative frequency   ($\rightarrow$ "frequentist statistics")
   A, B, ... are outcomes of a repeatable experiment

$$P(A) = \lim_{n \to \infty} \frac{\text{times outcome is } A}{n}$$

II. Subjective probability   ($\rightarrow$ "Bayesian statistics")
   A, B, ... are hypotheses (statements that are true or false)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

• Both interpretations consistent with Kolmogorov axioms.

• In particle physics frequency interpretation often most useful, but subjective probability can provide more natural treatment of non-repeatable phenomena: systematic uncertainties, probability that magnetic monopoles exist,...

Bayes' theorem

From the definition of conditional probability we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

but $P(A \cap B) = P(B \cap A)$, so

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$
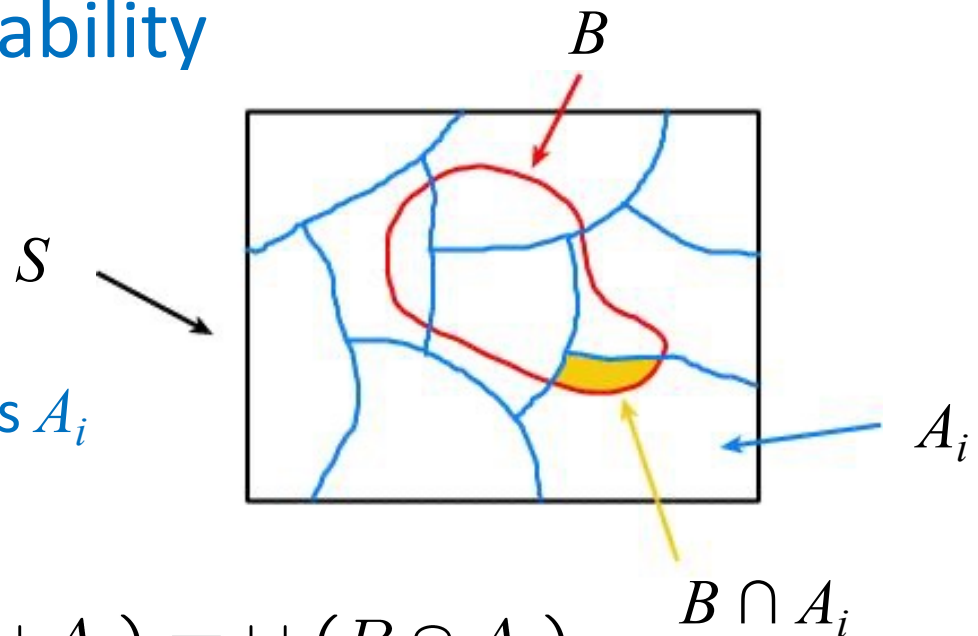
Bayes' theorem

First published (posthumously) by the
Reverend Thomas Bayes (1702–1761)

*An essay towards solving a problem in the doctrine
of chances*, Philos. Trans. R. Soc. **53** (1763) 370

The law of total probability

Consider a subset B of the sample space S,

divided into disjoint subsets A_i such that $\cup_i A_i = S$,

$$\rightarrow \quad B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i),$$

$$\rightarrow \quad P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$

$$\rightarrow \quad P(B) = \sum_i P(B|A_i)P(A_i)$$

Bayes' theorem becomes

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

An example using Bayes' theorem

Suppose the probability (for anyone) to have a disease D is:

$$P(\text{D}) = 0.001$$
$$P(\text{no D}) = 0.999$$

← prior probabilities, i.e., before any test carried out

Consider a test for the disease: result is + or −

$$P(+|\text{D}) = 0.98$$
$$P(-|\text{D}) = 0.02$$

← probabilities to (in)correctly identify a person with the disease

$$P(+|\text{no D}) = 0.03$$
$$P(-|\text{no D}) = 0.97$$

← probabilities to (in)correctly identify a healthy person

Suppose your result is +. How worried should you be?

Bayes' theorem example (cont.)

The probability to have the disease given a + result is

$$p(\mathrm{D}|+) = \frac{P(+|\mathrm{D})P(\mathrm{D})}{P(+|\mathrm{D})P(\mathrm{D}) + P(+|\mathrm{no}\ \mathrm{D})P(\mathrm{no}\ \mathrm{D})}$$

$$= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999}$$

$$= 0.032 \qquad \leftarrow \text{posterior probability}$$

i.e. you're probably OK!

Your viewpoint: my degree of belief that I have the disease is 3.2%.

Your doctor's viewpoint: 3.2% of people like this have the disease.

Frequentist Statistics – general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations (shorthand: x).

Probability = limiting frequency

Probabilities such as

P (string theory is true),
P ($0.117 < \alpha_s < 0.119$),
P (Biden wins in 2024),

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

Preferred theories (models, hypotheses, …) are those that predict a high probability for data "like" the data observed.

Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability for hypotheses:

probability of the data assuming
hypothesis H (the likelihood)

prior probability, i.e.,
before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e.,
after seeing the data

normalization involves sum
over all possible hypotheses

Bayes' theorem has an "if-then" character: If your prior
probabilities were $\pi(H)$, then it says how these probabilities
should change in the light of the data.

No general prescription for priors (subjective!)

Hypothesis, likelihood

Suppose the entire result of an experiment (set of measurements) is a collection of numbers x.

A (simple) hypothesis is a rule that assigns a probability to each possible data outcome:

$$P(\mathbf{x}|H) \quad = \quad \text{the likelihood of } H$$

Often we deal with a family of hypotheses labeled by one or more undetermined parameters (a composite hypothesis):

$$P(\mathbf{x}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}) \quad = \quad \text{the "likelihood function"}$$

Note:

1) For the likelihood we treat the data x as fixed.

2) The likelihood function $L(\boldsymbol{\theta})$ is not a pdf for $\boldsymbol{\theta}$.

The likelihood function for i.i.d.* data

* i.i.d. = independent and identically distributed

Consider n independent observations of x: $x_1, ..., x_n$, where x follows $f(x; \theta)$. The joint pdf for the whole data sample is:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

In this case the likelihood function is

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad (x_i \text{ constant})$$

Parameter estimation

The parameters of a pdf are any constants that characterize it,

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$$

i.e., θ indexes a set of hypotheses.

r.v.   parameter

Suppose we have a sample of observed values: $\boldsymbol{x} = (x_1, ..., x_n)$

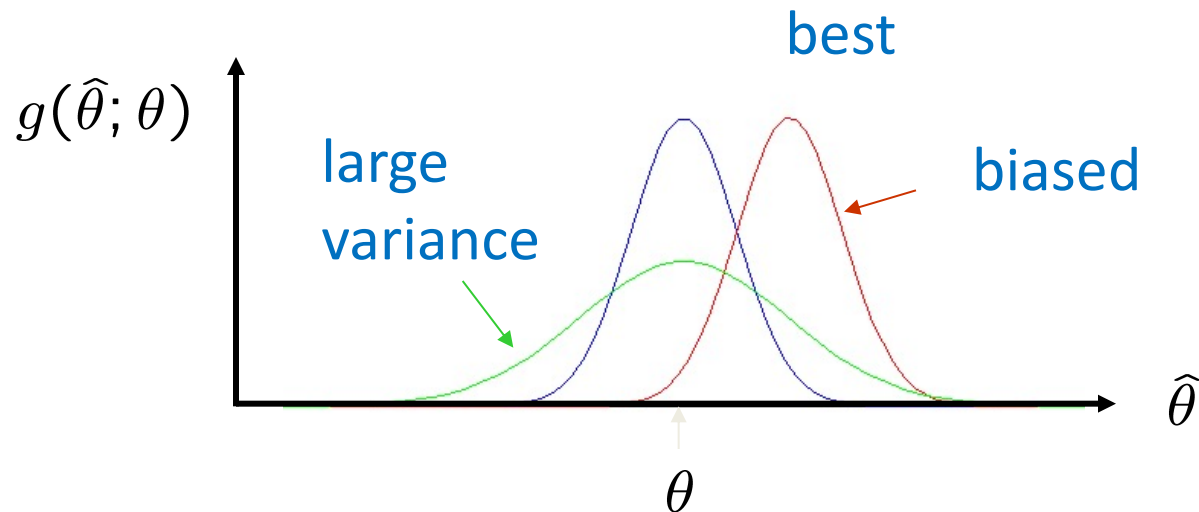We want to find some function of the data to estimate the parameter(s):

$$\widehat{\theta}(\vec{x})$$ ← estimator written with a hat

Sometimes we say 'estimator' for the function of $x_1, ..., x_n$; 'estimate' for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:

We want small (or zero) bias (systematic error): $b = E[\widehat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\widehat{\theta}]$

→ small bias & variance are in general conflicting criteria

Maximum Likelihood Estimators (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L equivalent to maximizing $\log L$

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

Could have multiple maxima (take highest).

MLEs not guaranteed to have any 'optimal' properties, (but in practice they're very good).

MLE example: parameter of exponential pdf

Consider exponential pdf, $f(t; \tau) = \dfrac{1}{\tau} e^{-t/\tau}$

and suppose we have i.i.d. data, $t_1, \ldots, t_n$

The likelihood function is $L(\tau) = \displaystyle\prod_{i=1}^{n} \dfrac{1}{\tau} e^{-t_i/\tau}$

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i; \tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

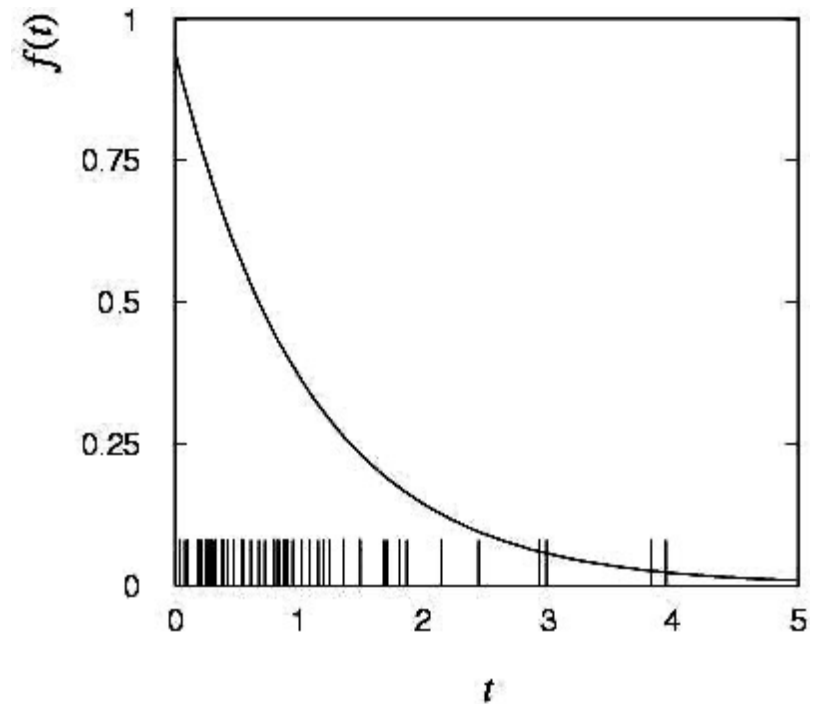# MLE example: parameter of exponential pdf (2)

Find its maximum by setting $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0$,

$$\rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Monte Carlo test:
generate 50 values
using $\tau = 1$:

We find the ML estimate:

$$\hat{\tau} = 1.062$$

MLE example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$$

$$V[t] = \int_0^\infty (t - \tau)^2 \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$$

For the MLE $\quad \hat{\tau} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} t_i \quad$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n} \sum_{i=1}^{n} t_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n} \sum_{i=1}^{n} t_i \right] = \frac{1}{n^2} \sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

The information inequality

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

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

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \bigg/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] \quad = \text{MVB} \quad \text{(Minimum Variance Bound)}$$

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

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

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

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 $\text{MVB} = \dfrac{\tau^2}{n} = V[\hat{\tau}]$. So here the MLE is efficient.

Large-sample (asymptotic) properties of MLEs

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

In the large-sample (or "asymptotic") limit ($n \rightarrow \infty$) 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, $\boldsymbol{\theta} = (\theta_1,...,\theta_m)$.

log-likelihood becomes quadratic

The likelihood function becomes Gaussian in shape, i.e. the log-likelihood becomes quadratic (parabolic).
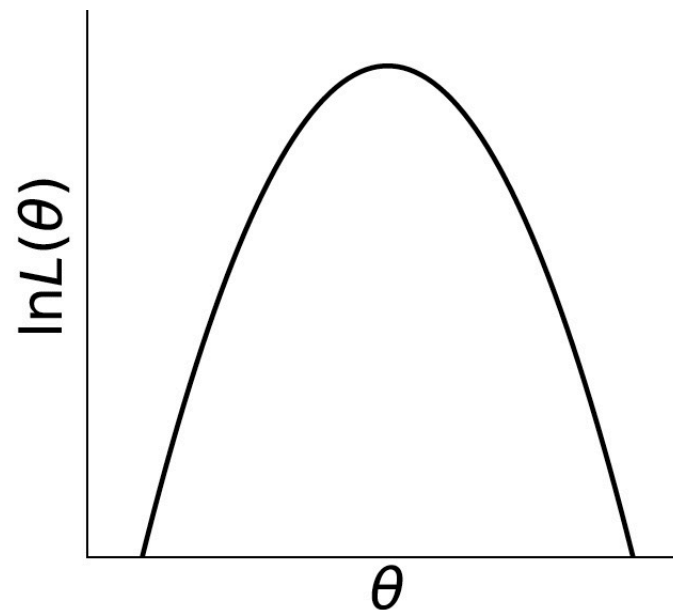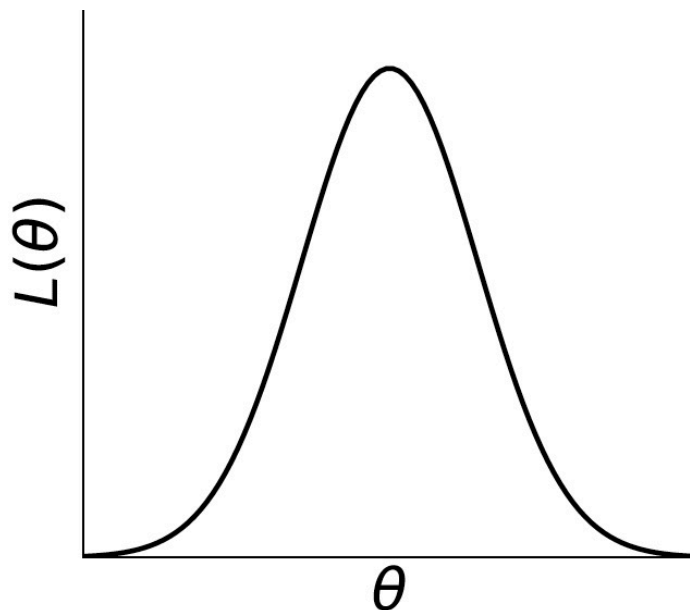
The MLE becomes increasingly precise as the (log)-likelihood becomes more tightly concentrated about its peak, but $L(\theta) = P(\boldsymbol{x}|\theta)$ is the probability for $\boldsymbol{x}$, not a pdf for θ.

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*.
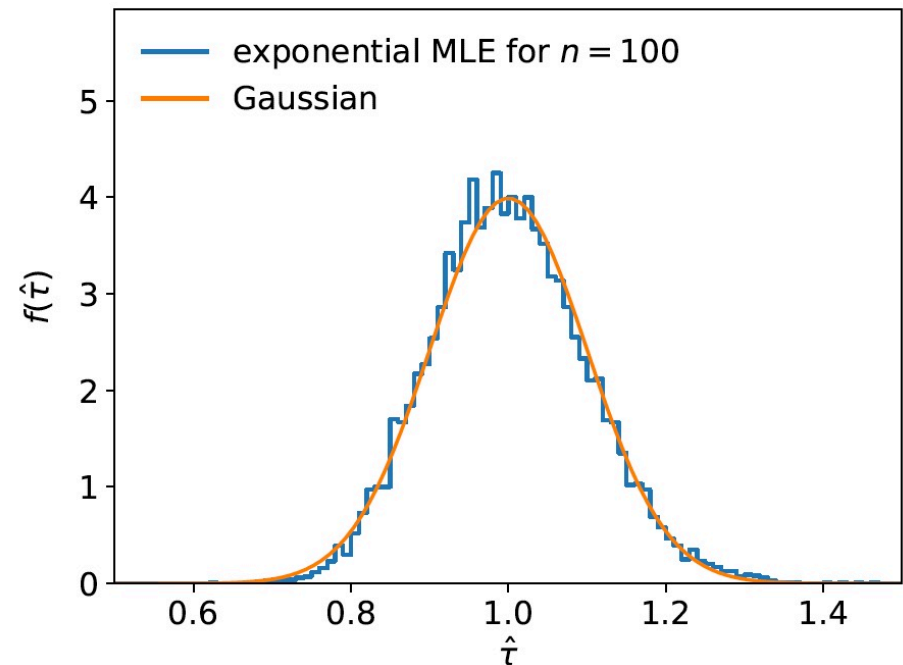
The MLE's distribution becomes Gaussian

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

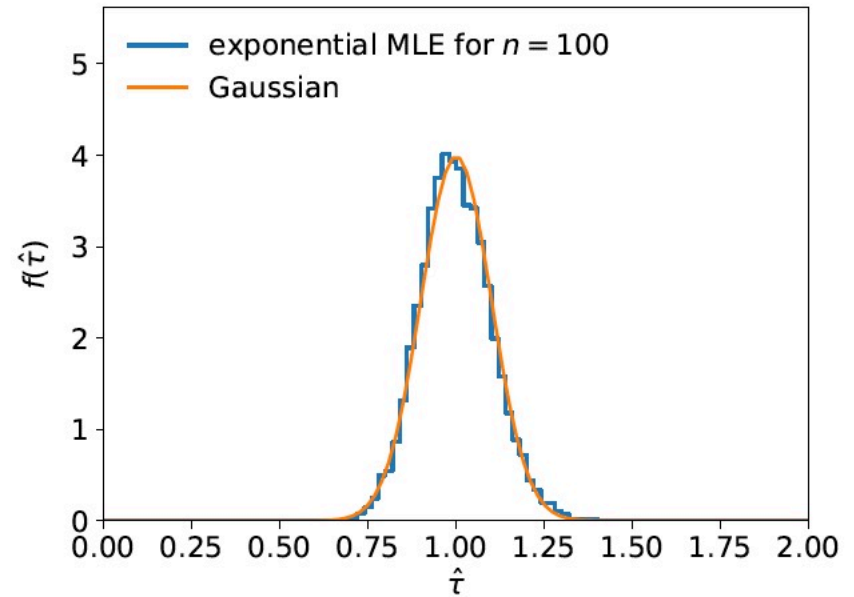$$f(\hat{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{\theta}}} e^{-(\hat{\theta}-\theta)^2/2\sigma_{\hat{\theta}}^2}$$
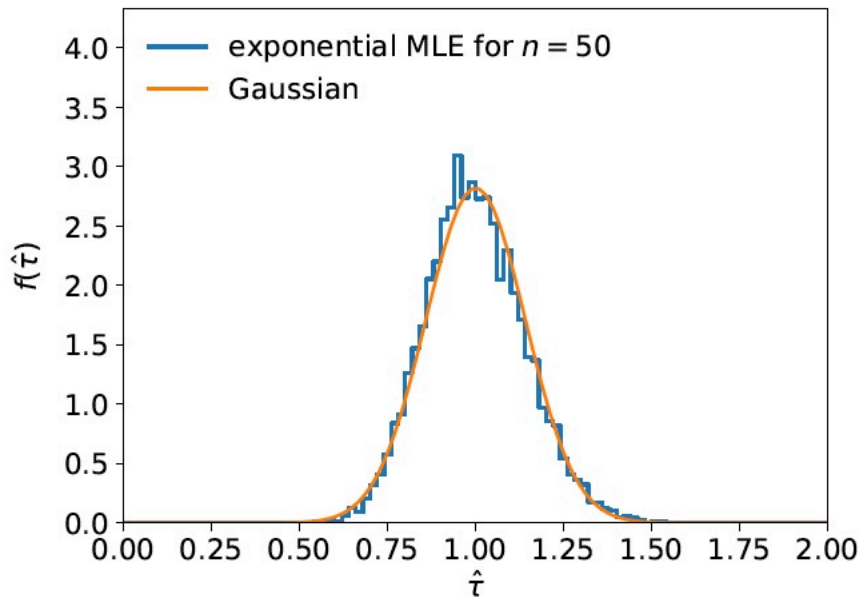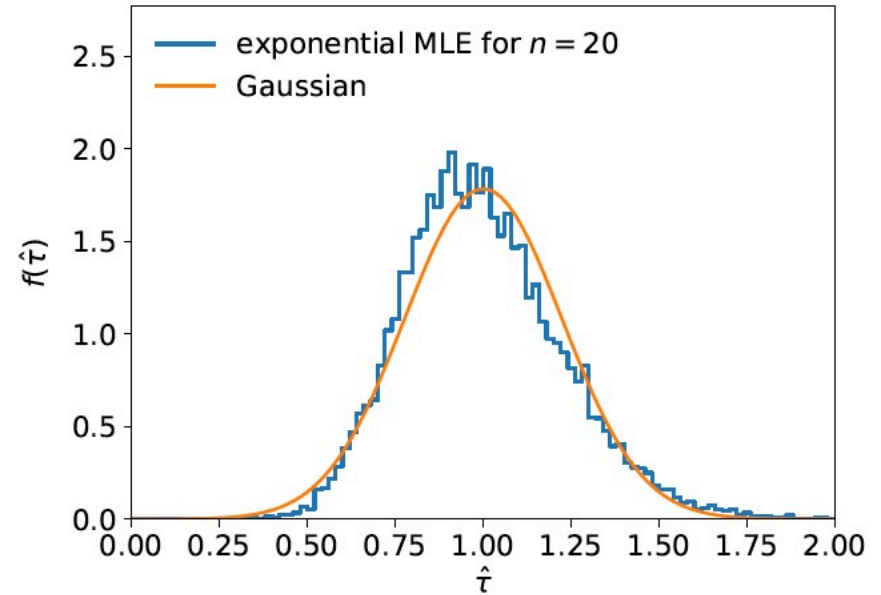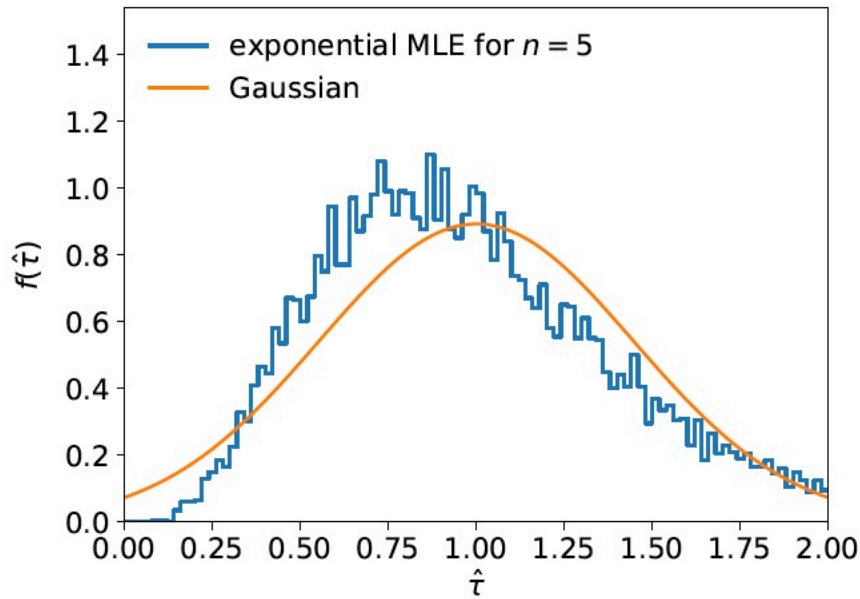
where $\sigma_{\hat{\theta}}^2$ is the minimum variance bound (note bias is zero).

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
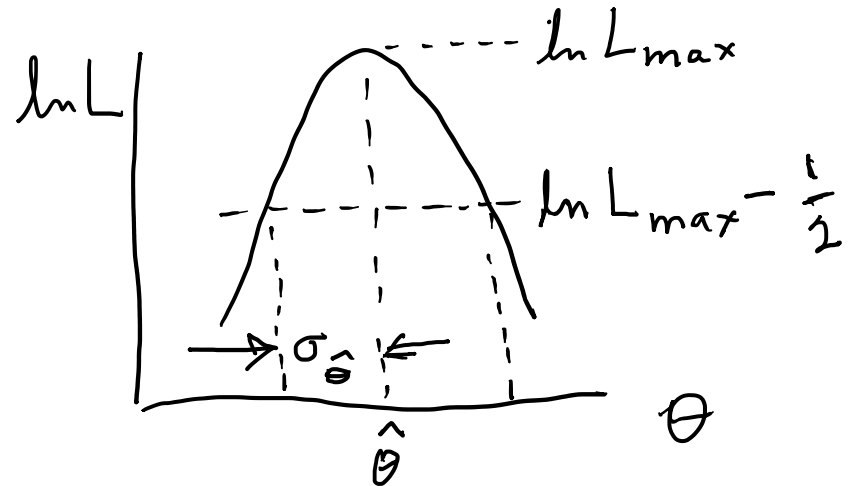
Variance of estimators: graphical method

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

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

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

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

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

$\rightarrow$ to get $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until $\ln L$ decreases by $1/2$.
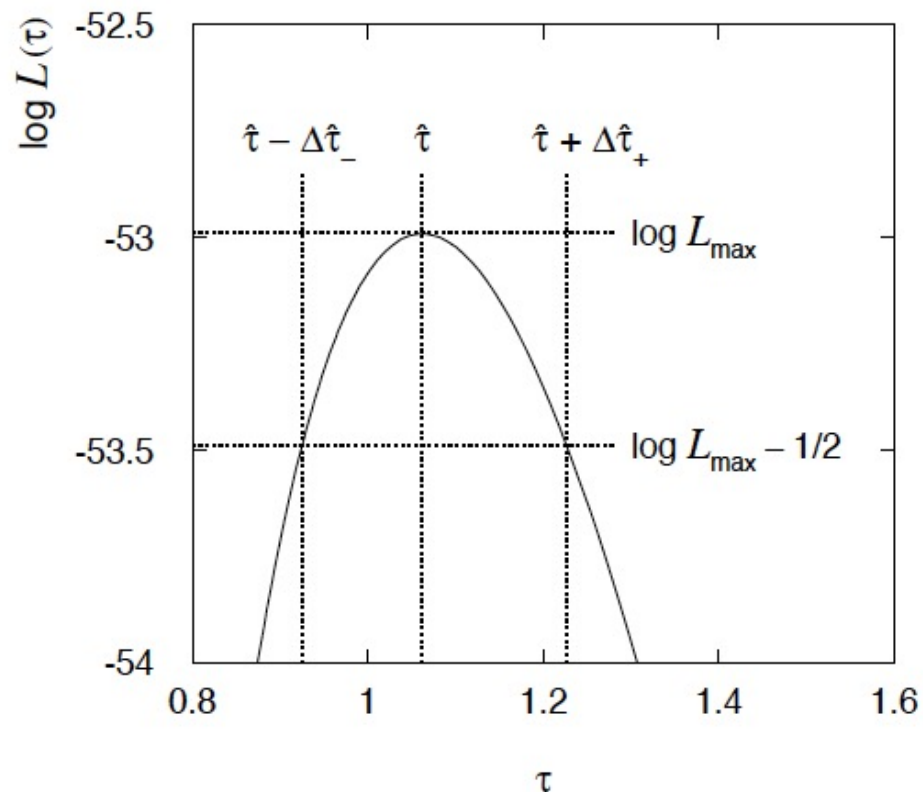
Example of variance by graphical method

ML example with exponential:

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$

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

Next time...

Lecture 1: Introduction, probability, parameter estimation

→ Lecture 2: Hypothesis tests, limits

Lecture 3: Systematic uncertainties, experimental sensitivity

Lecture 4: Bayesian methods, Student's t regression

Extra slides

Information inequality for N parameters

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

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 $\quad V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

The information inequality states that the matrix

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

is positive semi-definite:

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

Information inequality for N parameters (2)

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

    bias $\rightarrow 0$

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

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

This can be estimated from data using $\quad \widehat{V}_{ij}^{-1} = -\left.\dfrac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right|_{\hat{\boldsymbol{\theta}}}$
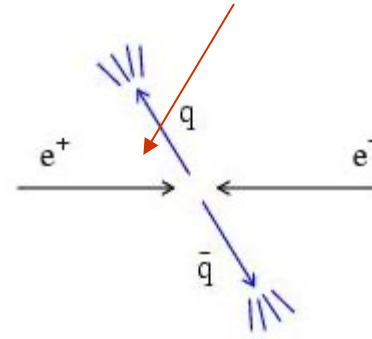
Find the matrix V^{-1} numerically (or with automatic differentiation), then invert to get the covariance matrix of the estimators

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

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.

$$\ln L(\alpha, \beta) = \sum_{i=1}^{n} \ln f(x_i; \alpha, \beta)$$ ⟵ need to find maximum numerically
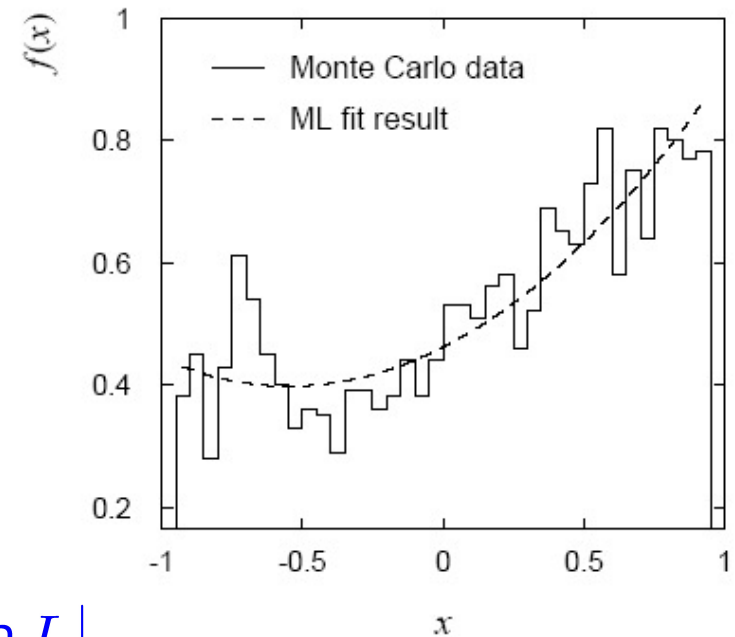
Example of ML with 2 parameters: fit result

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

$$\widehat{\alpha} = 0.508$$

$$\widehat{\beta} = 0.47$$

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

$$\widehat{(V^{-1})}_{ij} = -\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\bigg|_{\vec{\theta}=\widehat{\vec{\theta}}}$$

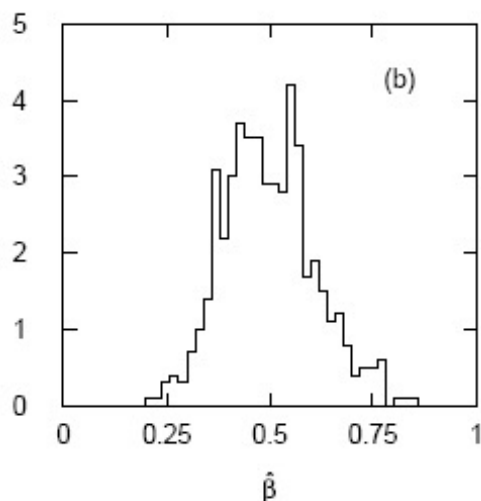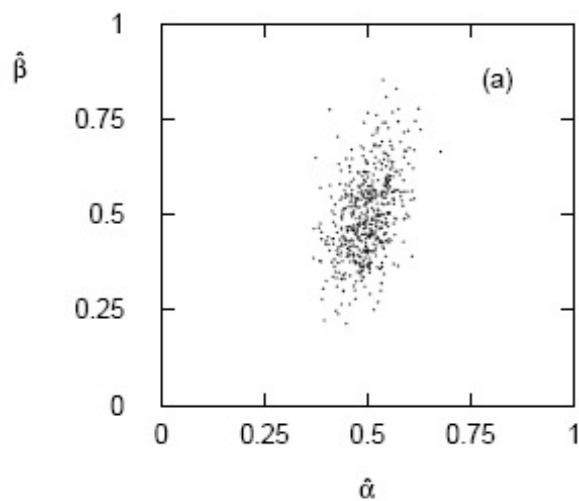$$\widehat{\sigma}_{\widehat{\alpha}} = 0.052 \qquad \text{cov}[\widehat{\alpha}, \widehat{\beta}] = 0.0026$$
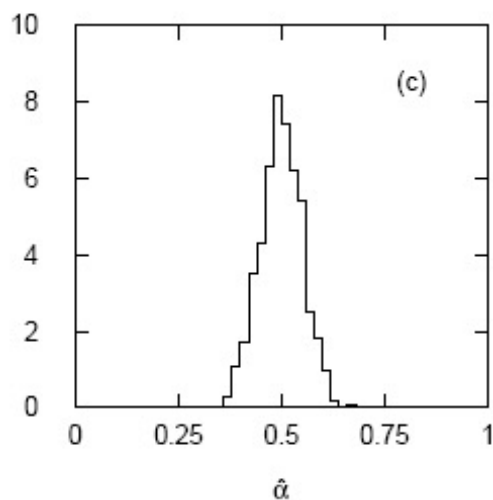
$$\widehat{\sigma}_{\widehat{\beta}} = 0.11 \qquad\qquad r = 0.46 \ = \ \text{correlation coefficient}$$

Two-parameter fit: MC study

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

$$\overline{\hat{\alpha}} = 0.499$$

$$s_{\hat{\alpha}} = 0.051$$

$$\overline{\hat{\beta}} = 0.498$$

$$s_{\hat{\beta}} = 0.111$$

$$\widehat{\text{cov}}[\hat{\alpha}, \hat{\beta}] = 0.0024$$

$$r = 0.42$$

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:

$$\ln L(\boldsymbol{\theta}) \approx \ln L(\hat{\boldsymbol{\theta}}) + \sum_i \left.\frac{\partial \ln L}{\partial \theta_i}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i) + \frac{1}{2!} \sum_{i,j} \left.\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$
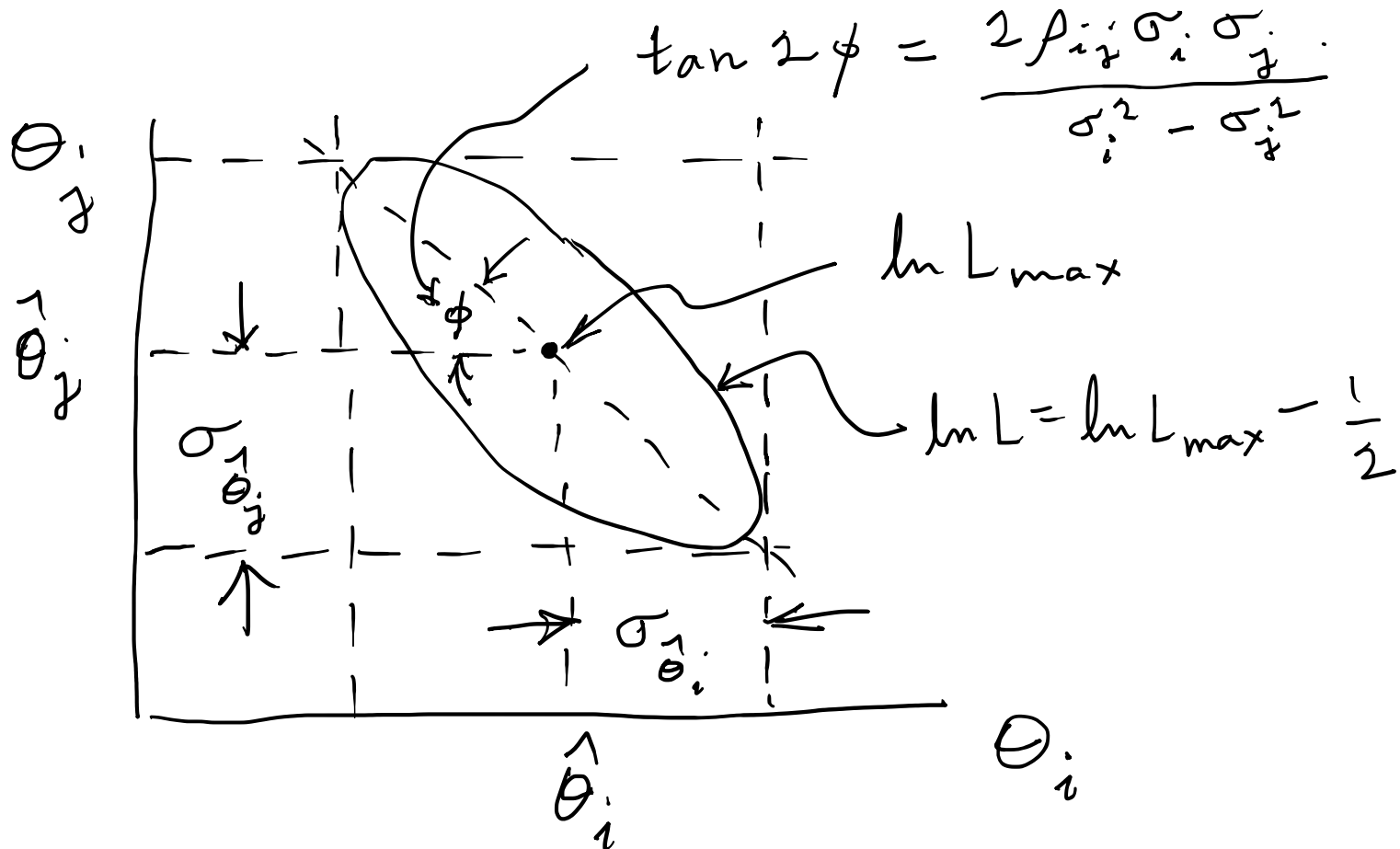
$\ln L_{\text{max}}$ zero relate to covariance matrix of MLEs using information (in)equality.

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

So the surface $\quad \ln L(\boldsymbol{\theta}) = \ln L_{\text{max}} - \frac{1}{2} \quad$ 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)

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2}$$

$\ln L_{max}$

$\ln L = \ln L_{max} - \frac{1}{2}$

Distance from MLE to tangent planes gives standard deviations.

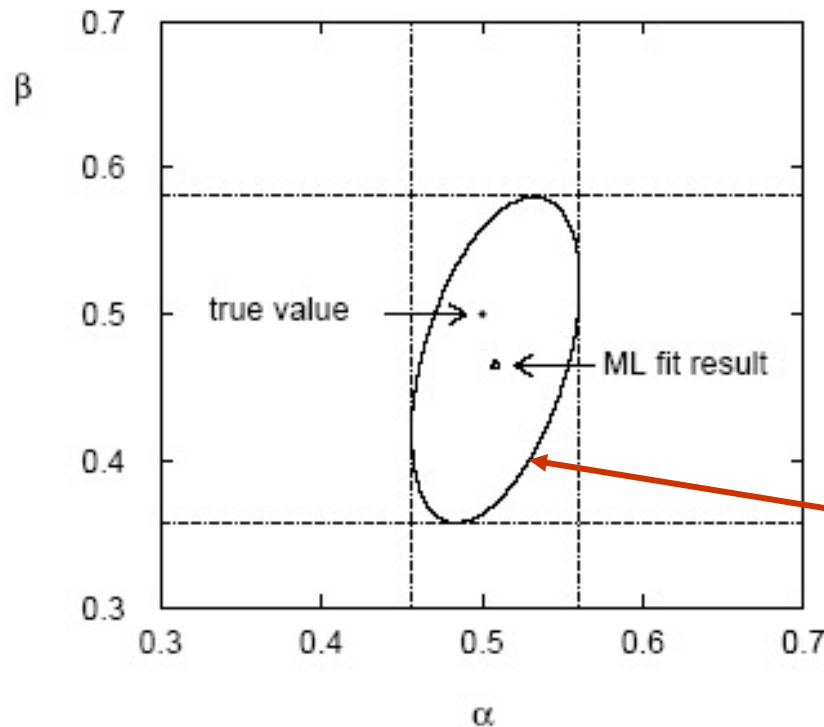# The ln $L_{\max} - 1/2$ contour for two parameters

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

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

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

$\rightarrow$ Angle of ellipse φ related to correlation: $\tan 2\phi = \dfrac{2\rho\sigma_{\widehat{\alpha}}\sigma_{\widehat{\beta}}}{\sigma_{\widehat{\alpha}}^2 - \sigma_{\widehat{\beta}}^2}$