Computing and Statistical Data Analysis
Stat 7: Parameter Estimation, ML, LS

London Postgraduate Lectures on Particle Physics;

University of London 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:

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

The likelihood function

Suppose the entire result of an experiment (set of measurements) is a collection of numbers ***x***, and suppose the joint pdf for the data ***x*** is a function that depends on a set of parameters ***θ***:

$$f(\vec{x};\vec{\theta})$$

Now evaluate this function with the data obtained and regard it as a function of the parameter(s). This is the likelihood function:

$$L(\vec{\theta}) = f(\vec{x};\vec{\theta})$$

(***x*** constant)

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:

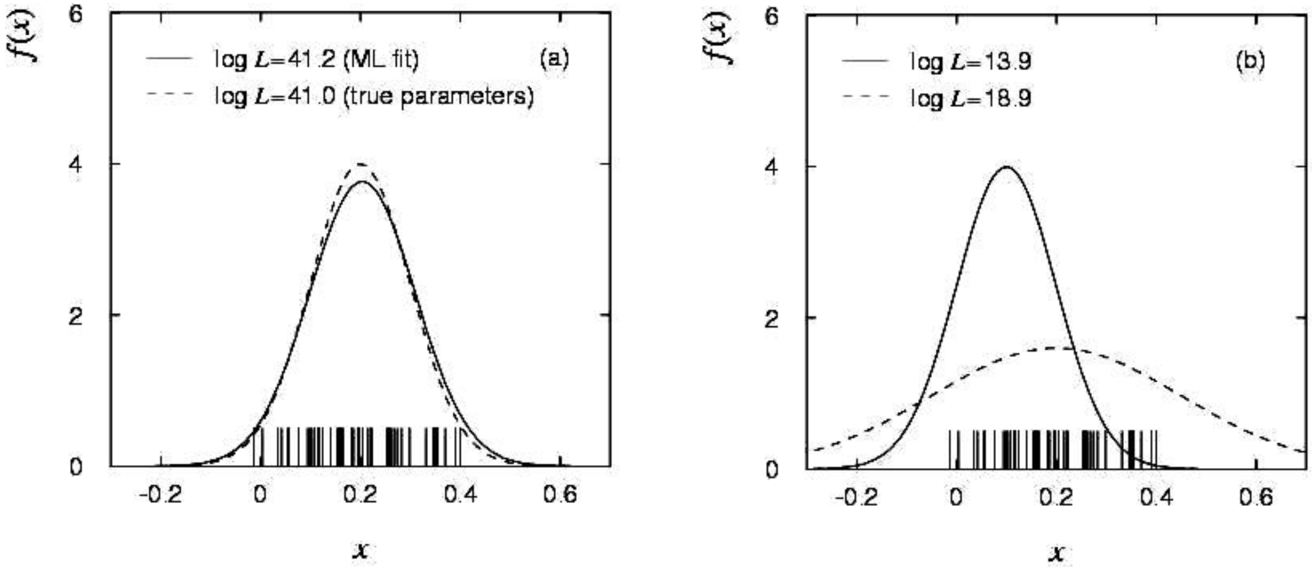
$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

In this case the likelihood function is

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

Maximum likelihood estimators

If the hypothesized θ is close to the true value, then we expect a high probability to get data like that which we actually found.



So we define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

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

ML example: parameter of exponential pdf

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

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

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

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

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

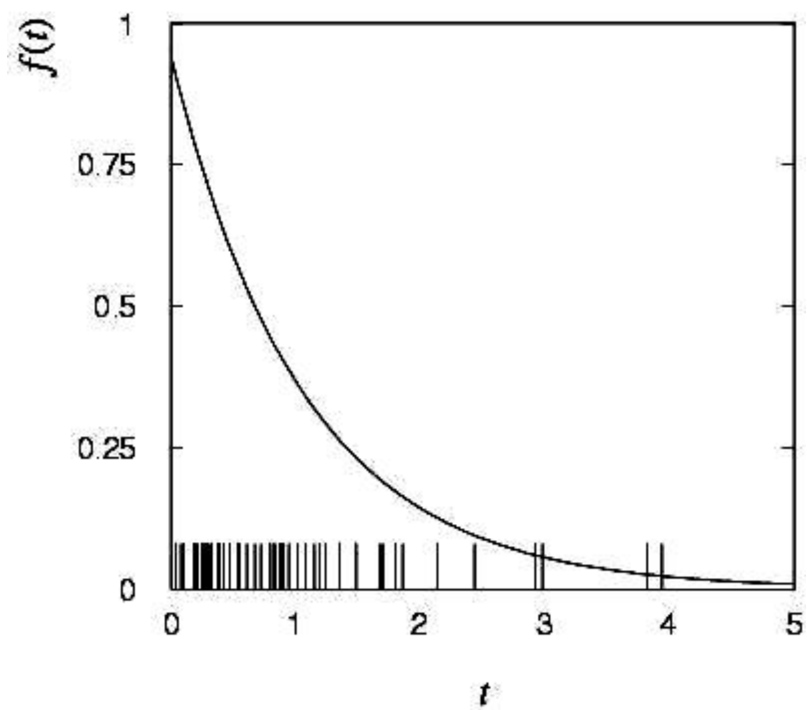
ML example: parameter of exponential pdf (2)

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

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

Monte Carlo test:

generate 50 values using $\tau = 1$:



We find the ML estimate:

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

Functions of ML estimators

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 ML estimator for λ?

For a function $\alpha(\theta)$ of a parameter θ, it doesn't matter whether we express L as a function of α or θ.

The ML estimator of a function $\alpha(\theta)$ is simply $\hat{\alpha} = \alpha(\hat{\theta})$.

So 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 $\rightarrow 0$ for $n \rightarrow \infty$)

Example of ML: parameters of Gaussian pdf

Consider independent $x_1, \ldots, x_n$, with $x_i \sim$ Gaussian (μ,σ^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 ML estimator for σ^2 has a bias, but $b \to 0$ for $n \to \infty$. 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.

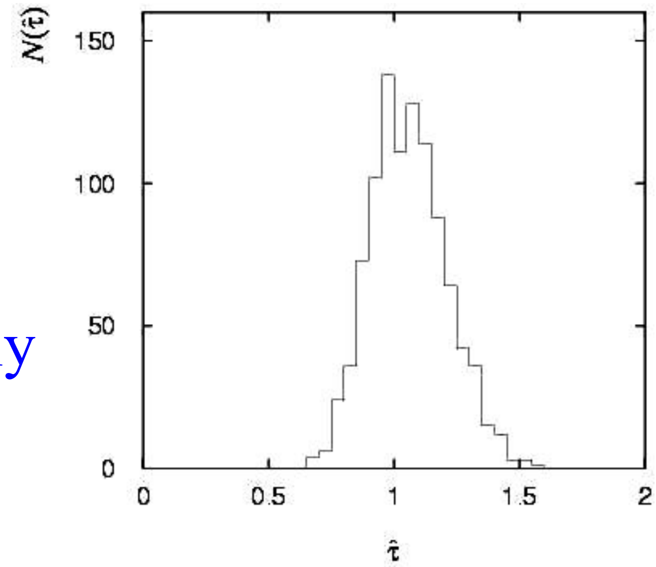
Variance of estimators: Monte Carlo method

Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

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



Note distribution of estimates is roughly Gaussian − (almost) always true for ML in large sample limit.

Variance of estimators from information inequality

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

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

Minimum Variance Bound (MVB)

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

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):

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

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

→ 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

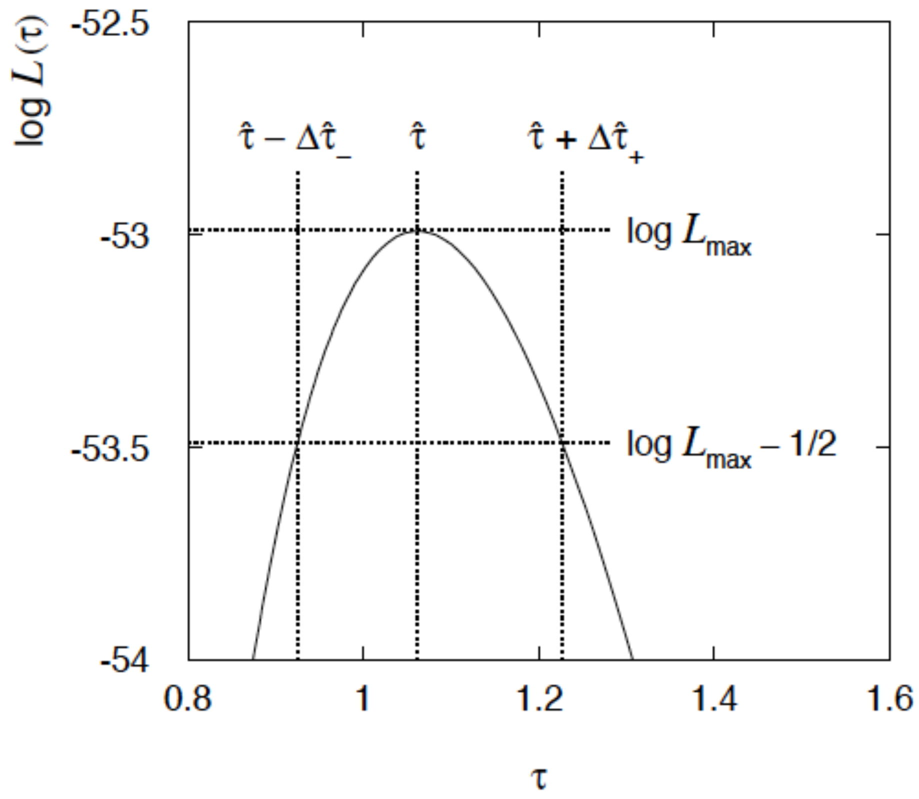
ML example with exponential:

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

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

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

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



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

Information inequality for *n* parameters

Suppose we have estimated *n* parameters $\vec{\theta} = (\theta_1, \ldots, \theta_n)$.

The (inverse) minimum variance bound is given by the Fisher information matrix:

$$I_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right] = -n \int f(x;\vec{\theta}) \frac{\partial^2 \ln f(x;\vec{\theta})}{\partial \theta_i \partial \theta_j} \, dx$$

The information inequality then states that $V - I^{-1}$ is a positive semi-definite matrix, where $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$. Therefore

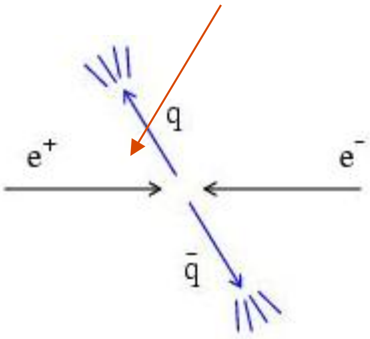
$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use I^{-1} as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of L.

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 always 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.

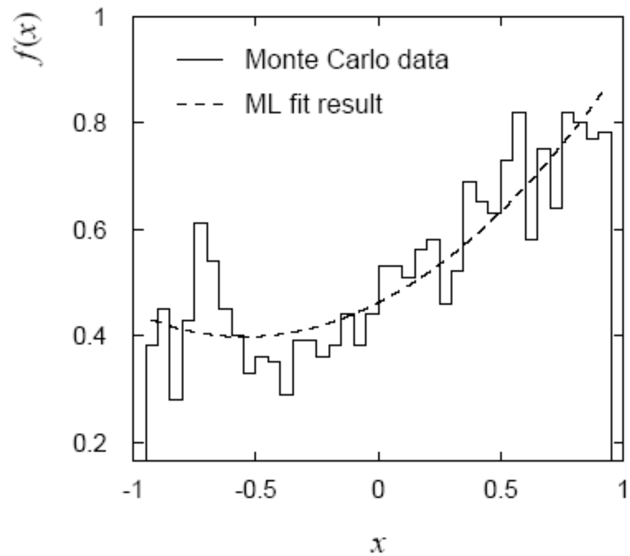
Example of ML with 2 parameters: fit result

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

$$\hat{\alpha} = 0.508$$

$$\hat{\beta} = 0.47$$

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



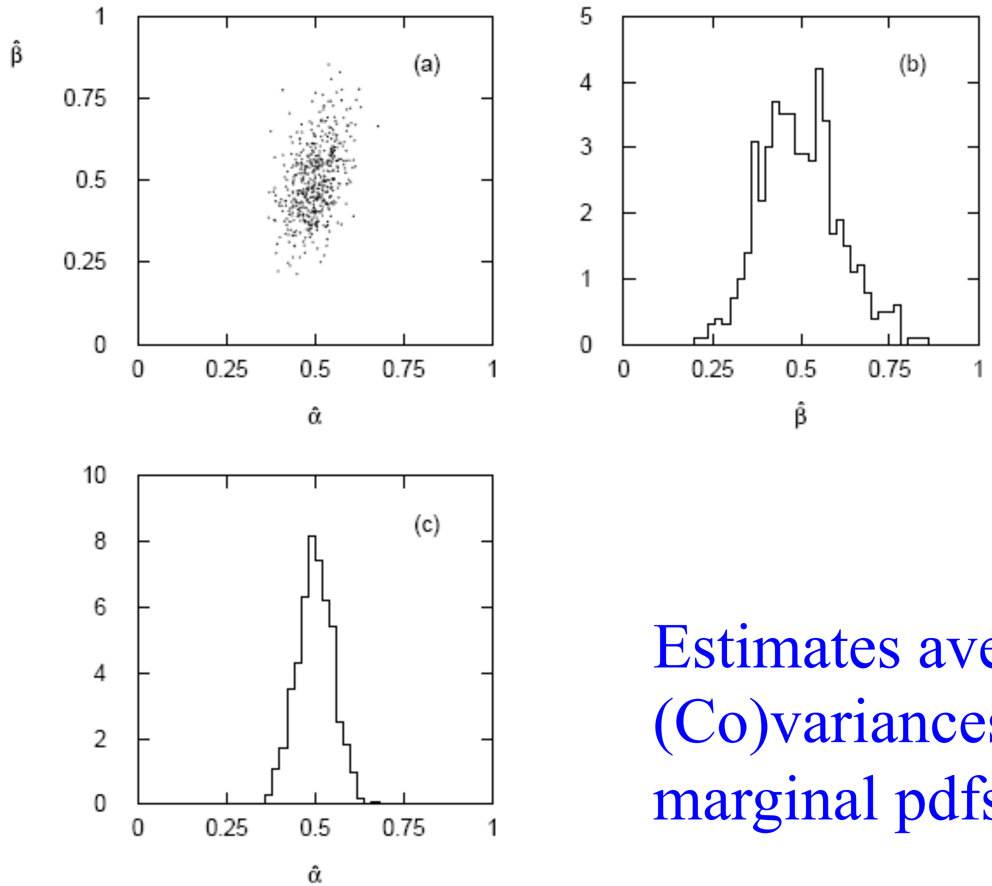
(Co)variances from $\widehat{(V^{-1})}_{ij} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\hat{\vec{\theta}}}$ (**MINUIT** routine **HESSE**)

$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \qquad \mathrm{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$

$$\hat{\sigma}_{\hat{\beta}} = 0.11 \qquad r = 0.46$$

Two-parameter fit: MC study

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



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

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

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

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

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

$$r = 0.42$$

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

The ln L_{max} − 1/2 contour

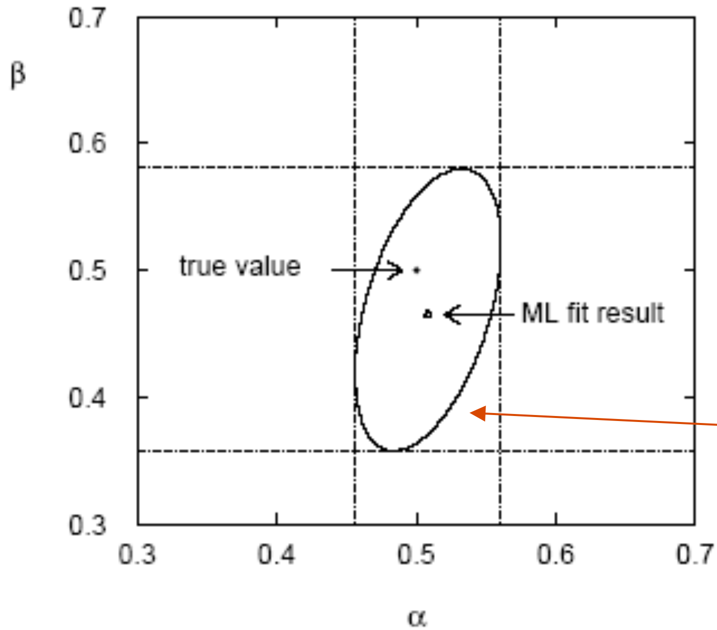
For large n, ln L takes on quadratic form near maximum:

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

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

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

(Co)variances from ln L contour



The α, β plane for the first MC data set

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

→ Tangent lines to contours give standard deviations.

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

Correlations between estimators result in an increase in their standard deviations (statistical errors).

Information inequality for n parameters

Suppose we have estimated n parameters $\vec{\theta} = (\theta_1, \ldots, \theta_n)$.

The (inverse) minimum variance bound is given by the Fisher information matrix:

$$I_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -n \int f(x;\vec{\theta}) \frac{\partial^2 \ln f(x;\vec{\theta})}{\partial\theta_i \partial\theta_j}\, dx$$

The information inequality then states that $V - I^{-1}$ is a positive semi-definite matrix, where $V_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$. Therefore

$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use I^{-1} as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of L.

Extended ML

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

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

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 ML uses more info → smaller errors for $\hat{\vec{\theta}}$

Important e.g. for anomalous couplings in $e^+e^- \rightarrow W^+W^-$

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

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

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

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 $\mu_s = \theta\nu$, $\mu_b = (1-\theta)\nu$, goal is to estimate μ_s, μ_b.

$$f(x; \mu_s, \mu_b) = \frac{\mu_s}{\mu_s + \mu_b} f_s(x) + \frac{\mu_b}{\mu_s + \mu_b} f_b(x)$$

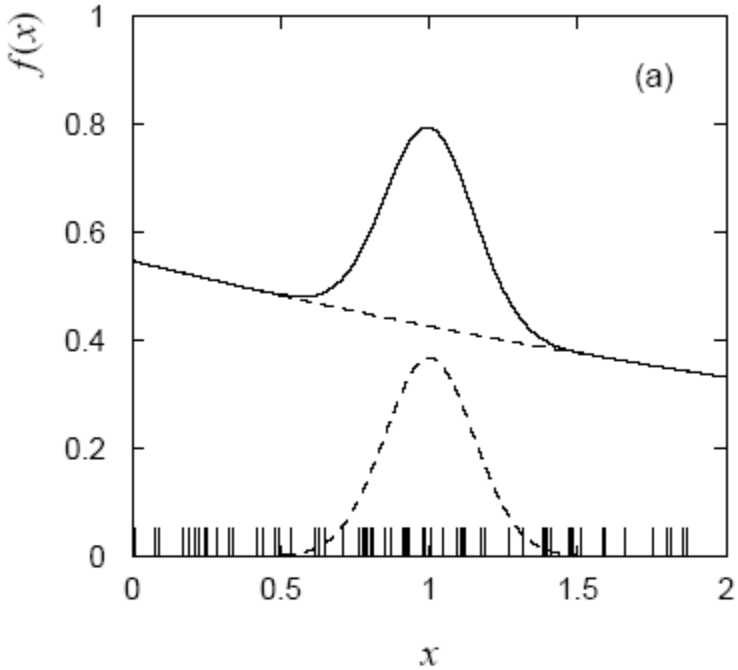
$$P(n; \mu_s, \mu_b) = \frac{(\mu_s + \mu_b)^n}{n!} e^{-(\mu_s + \mu_b)}$$

$$\rightarrow \ln L(\mu_s, \mu_b) = -(\mu_s + \mu_b) + \sum_{i=1}^{n} \ln\left[(\mu_s + \mu_b) f(x_i; \mu_s, \mu_b)\right]$$

Extended ML example (2)

Monte Carlo example with combination of exponential and Gaussian:

$$\mu_{\rm s} = 6$$
$$\mu_{\rm b} = 60$$



Maximize log-likelihood in terms of μ_s and μ_b:

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

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