Statistics for Particle Physicists

Lecture 2

https://pages.lip.pt/data-science/school/

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

Hypothesis tests

From hypothesis tests to machine learning

→ Lecture 2: Parameter estimation

Confidence limits

Systematic uncertainties

Hypothesis, likelihood

Suppose the entire result of an experiment (set of measurements) is a collection of numbers $\boldsymbol{x}$.

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):

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

Note:

1) For the likelihood we treat the data $\boldsymbol{x}$ 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: $x_1, \ldots, x_n$, 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,

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

r.v. parameter

i.e., θ indexes a set of hypotheses.

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

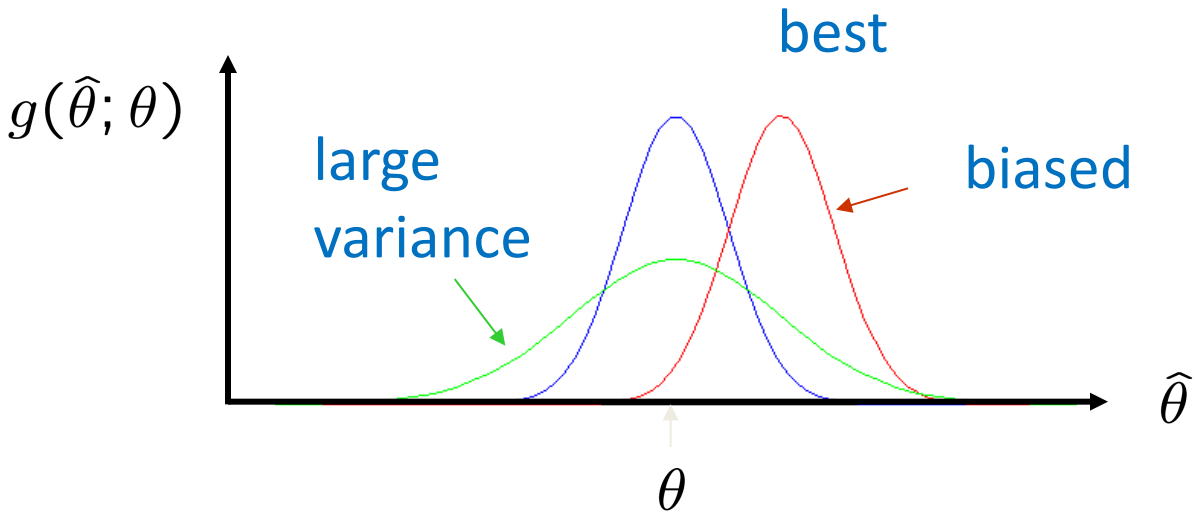
We want to find some function of the data to estimate the parameter(s):

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

Sometimes we say 'estimator' for the function of $x_1, \ldots, x_n$; '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): $b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

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

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

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

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

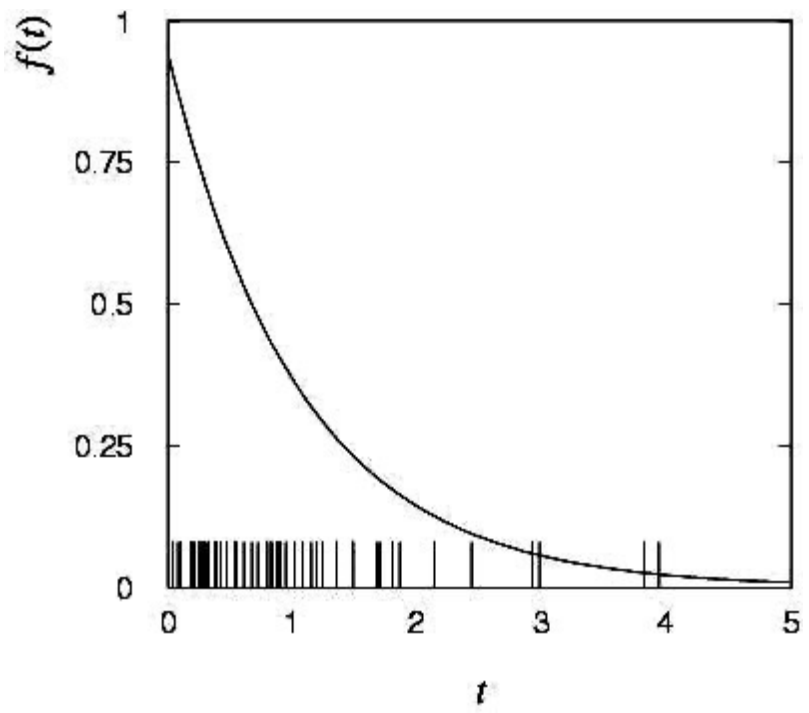
The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

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

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 $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$ we therefore find

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

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

Variance of estimators: Monte Carlo method

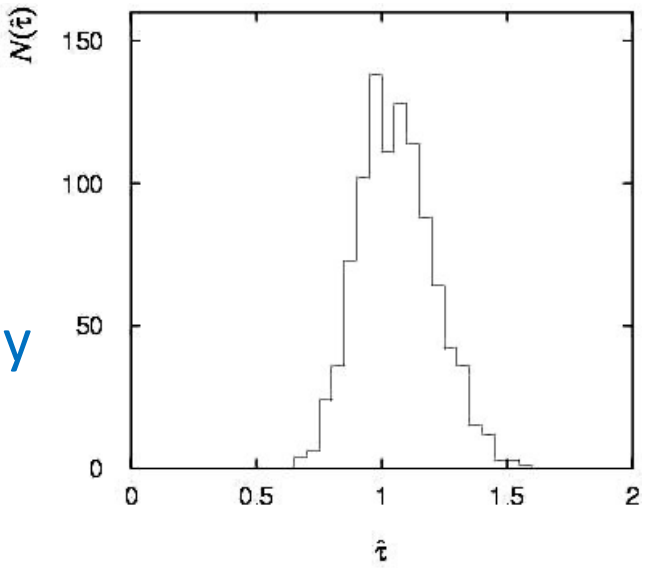
Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

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:

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

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):

Minimum Variance Bound (MVB)

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

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

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

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

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

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 $$\mathrm{MVB} = \frac{\tau^2}{n} = V[\hat{\tau}]\,.$$ (Here MLE is "efficient").

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

$\rightarrow$ 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

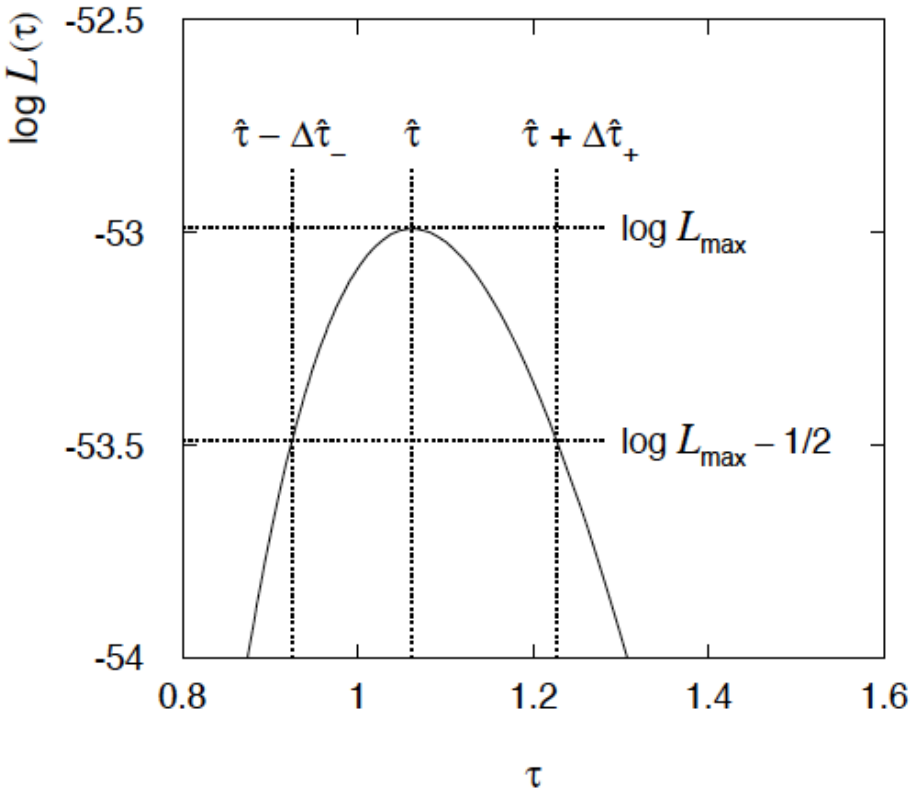
ML example with exponential:

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

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

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

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



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

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

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 $\rightarrow 0$

inequality $\rightarrow$ 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 $$\widehat{V}_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\hat{\boldsymbol{\theta}}}$$

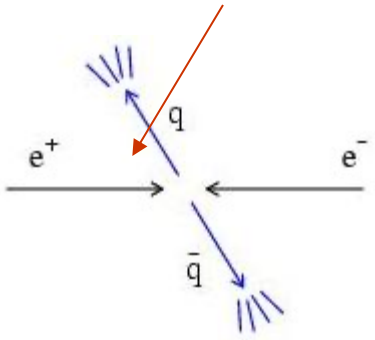
Find the matrix V^{-1} numerically (or with automatic differentiation), then invert to get the covariance matrix of the estimators

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

Example of 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.

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

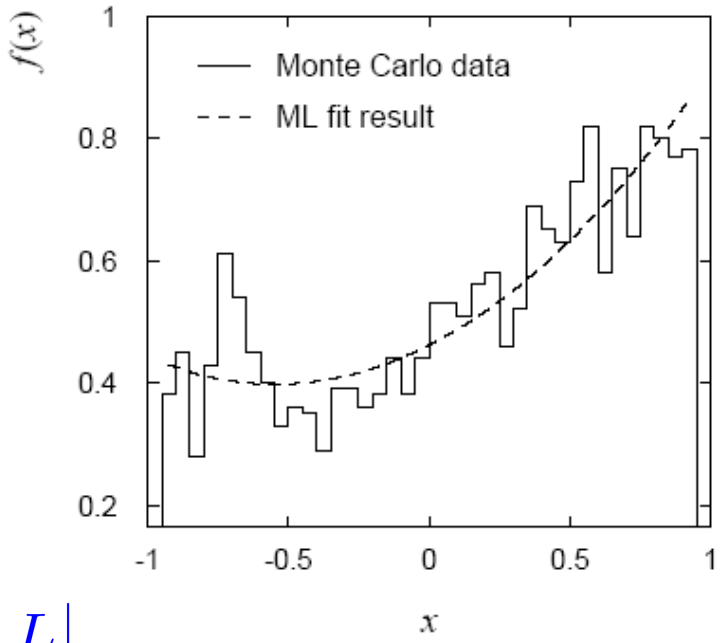
need to find maximum numerically

Example of ML with 2 parameters: fit result

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

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

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



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

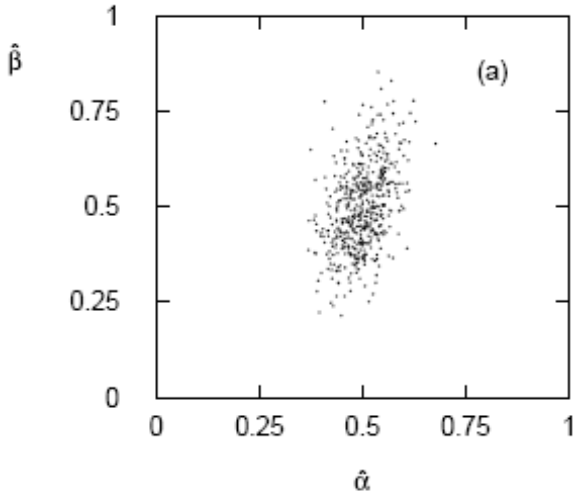
(Co)variances from $(\widehat{V^{-1}})_{ij} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\hat{\vec{\theta}}}$

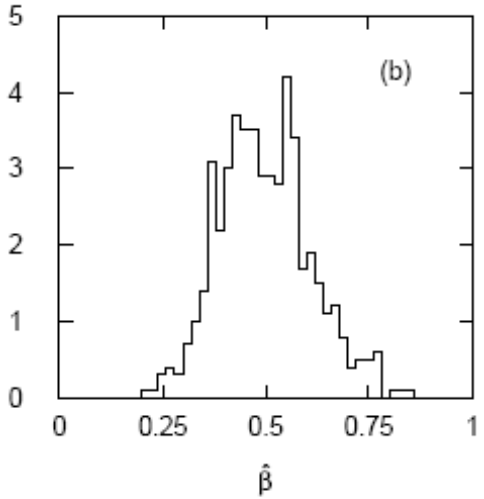
$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \qquad \mathrm{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$

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

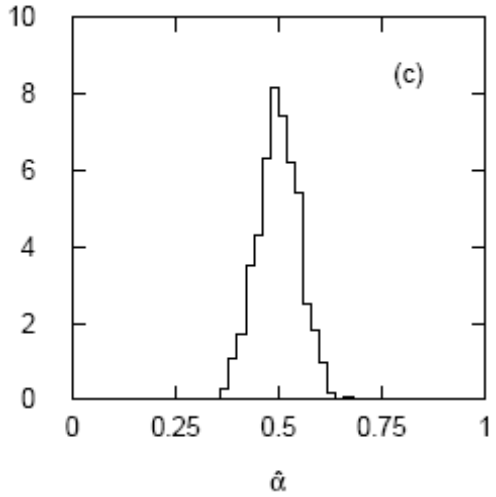
Two-parameter fit: MC study

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





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



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

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

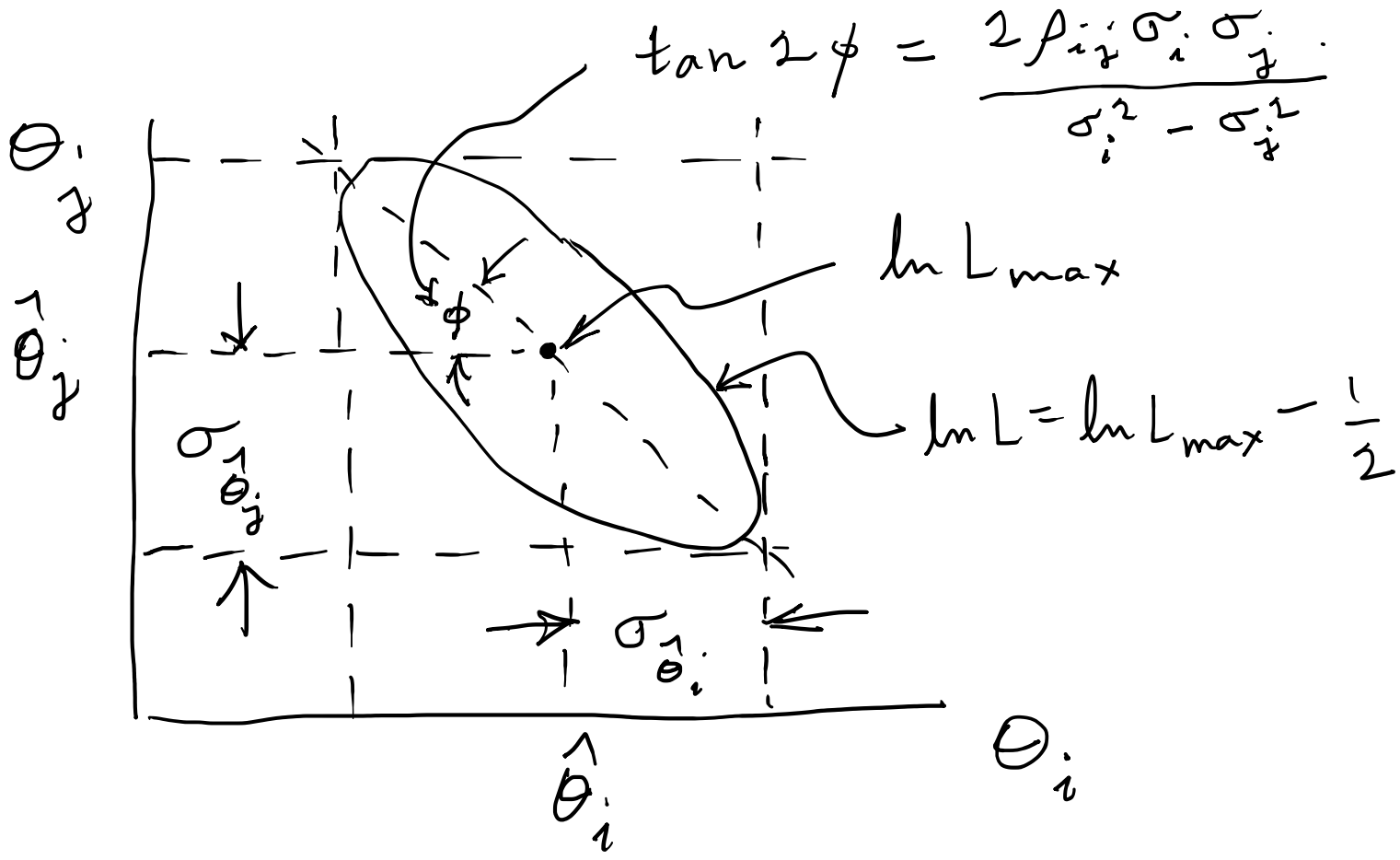
relate to covariance matrix of MLEs using information (in)equality.

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 $\ln L(\boldsymbol{\theta}) = \ln L_{\max} - \frac{1}{2}$ 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)



Distance from MLE to tangent planes gives standard deviations.

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

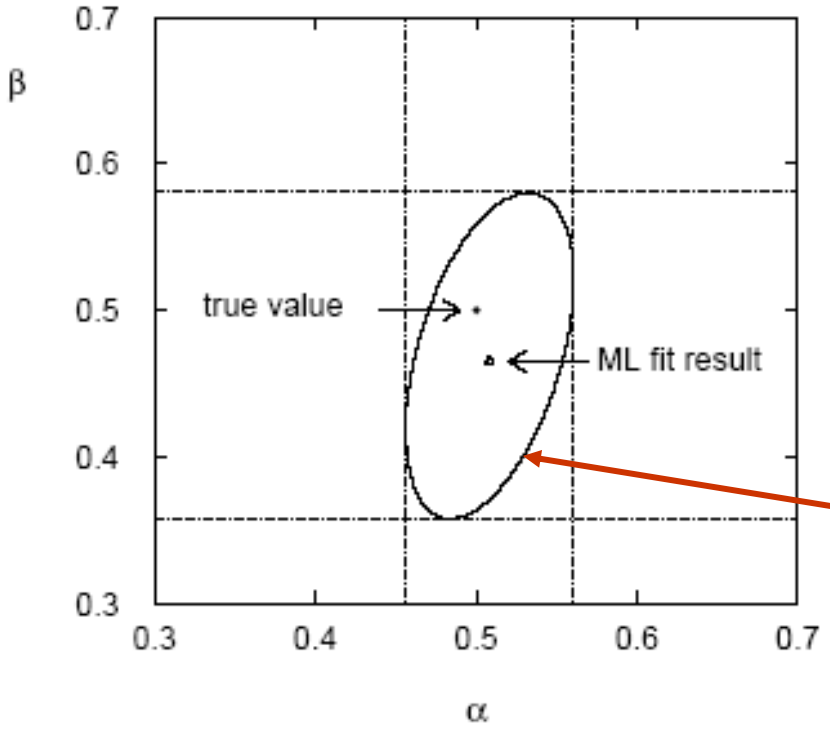
For large n, $\ln L$ takes on quadratic form near maximum:

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

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: $\tan 2\phi = \dfrac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma_{\hat{\beta}}^2}$

Confidence intervals by inverting a test

In addition to a 'point estimate' of a parameter we should report an interval reflecting its statistical uncertainty.

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

Specify values of the data that are 'disfavoured' by θ (critical region) such that $P(\text{data in critical region}|\theta) \leq \alpha$ for a prespecified α, e.g., 0.05 or 0.1.

If data observed in the critical region, reject the value θ.

Now invert the test to define a confidence interval as:

set of θ values that are not rejected in a test of size α (confidence level CL is $1-\alpha$).

Relation between confidence interval and p-value

Equivalently we can consider a significance test for each hypothesized value of θ, resulting in a p-value, p_θ.

If $p_\theta \leq \alpha$, then we reject θ.

The confidence interval at CL $= 1 - \alpha$ consists of those values of θ that are not rejected.

E.g. an upper limit on θ is the greatest value for which $p_\theta > \alpha$.

In practice find by setting $p_\theta = \alpha$ and solve for θ.

For a multidimensional parameter space $\boldsymbol{\theta} = (\theta_1, \ldots \theta_M)$ *use* same idea – result is a confidence "region" with boundary determined by $p_{\boldsymbol{\theta}} = \alpha$.

Coverage probability of confidence interval

If the true value of θ is rejected, then it's not in the confidence interval. The probability for this is by construction (equality for continuous data):

$$P(\text{reject } \theta | \theta) \leq \alpha = \text{type-I error rate}$$

Therefore, the probability for the interval to contain or "cover" θ is

$$P(\text{conf. interval "covers" } \theta | \theta) \geq 1 - \alpha$$

This assumes that the set of θ values considered includes the true value, i.e., it assumes the composite hypothesis $P(\boldsymbol{x}|H,\theta)$.

Frequentist upper limit on Poisson parameter

Consider again the case of observing $n \sim \text{Poisson}(s + b)$.

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

Relevant alternative is $s = 0$ (critical region at low n)

p-value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$

Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from

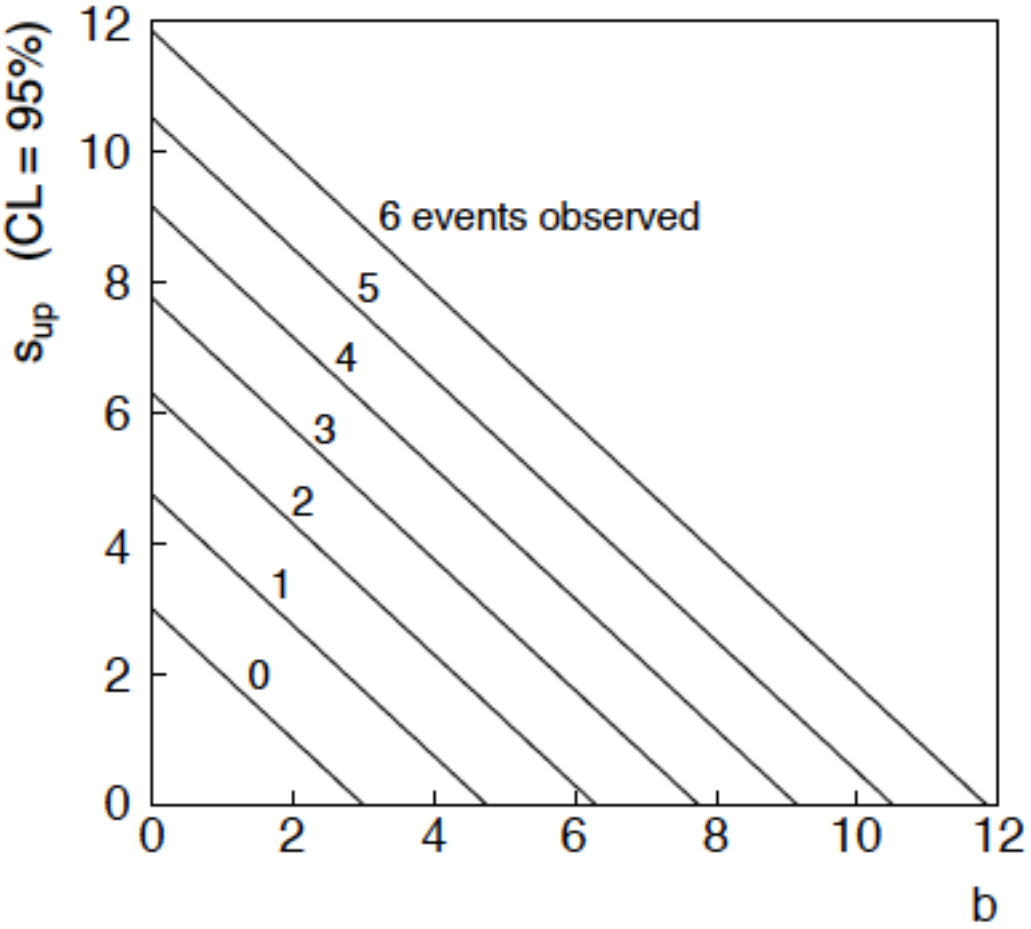
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

$$s_{\text{up}} = \frac{1}{2} F^{-1}_{\chi^2}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F^{-1}_{\chi^2}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

$n \sim$ Poisson(s+b): frequentist upper limit on s

For low fluctuation of n, formula can give negative result for $s_{\rm up}$; i.e. confidence interval is empty; all values of $s \geq 0$ have $p_s \leq \alpha$.



Limits near a boundary of the parameter space

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL $= 0.9$, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (\text{CL} = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when testing parameter values for which one has very little experimental sensitivity, e.g., very small s.

Expected limit for $s = 0$

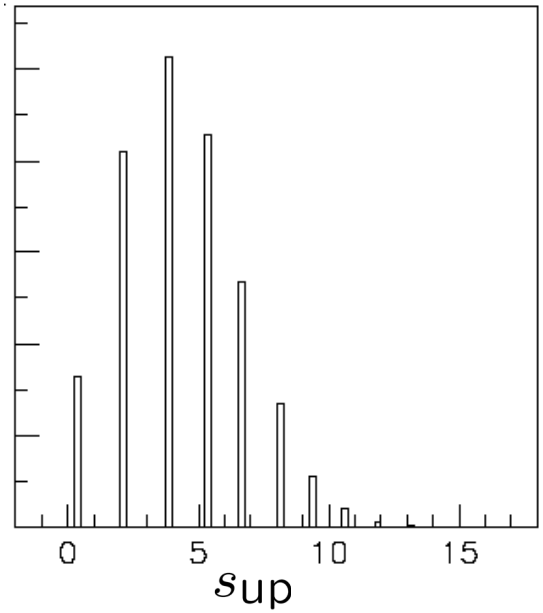
Physicist: I should have used CL = 0.95 — then $s_{\rm up} = 0.496$

Even better: for CL = 0.917923 we get $s_{\rm up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

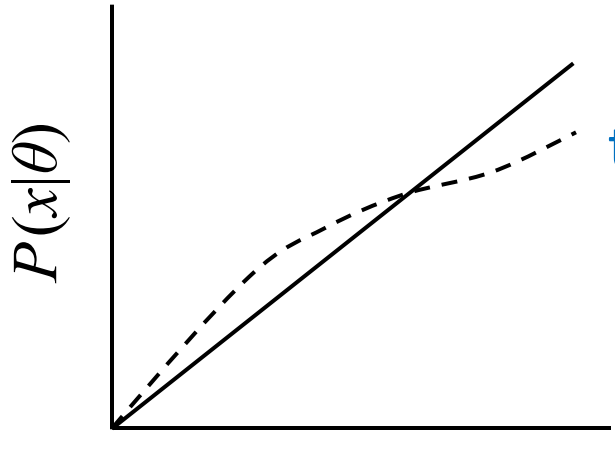
Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



Systematic uncertainties and nuisance parameters

In general, our model of the data is not perfect:



model: $P(x|\theta) = \theta x$

truth: $P(x|\theta) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

Can improve model by including additional adjustable parameters.

$$P(x|\theta) \to P(x|\theta, \nu)$$

Nuisance parameter ↔ systematic uncertainty. Some point in the parameter space of the enlarged model should be "true".

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

Example: fitting a straight line

Data: (x_i, y_i, σ_i) , $i = 1, \ldots, n$.

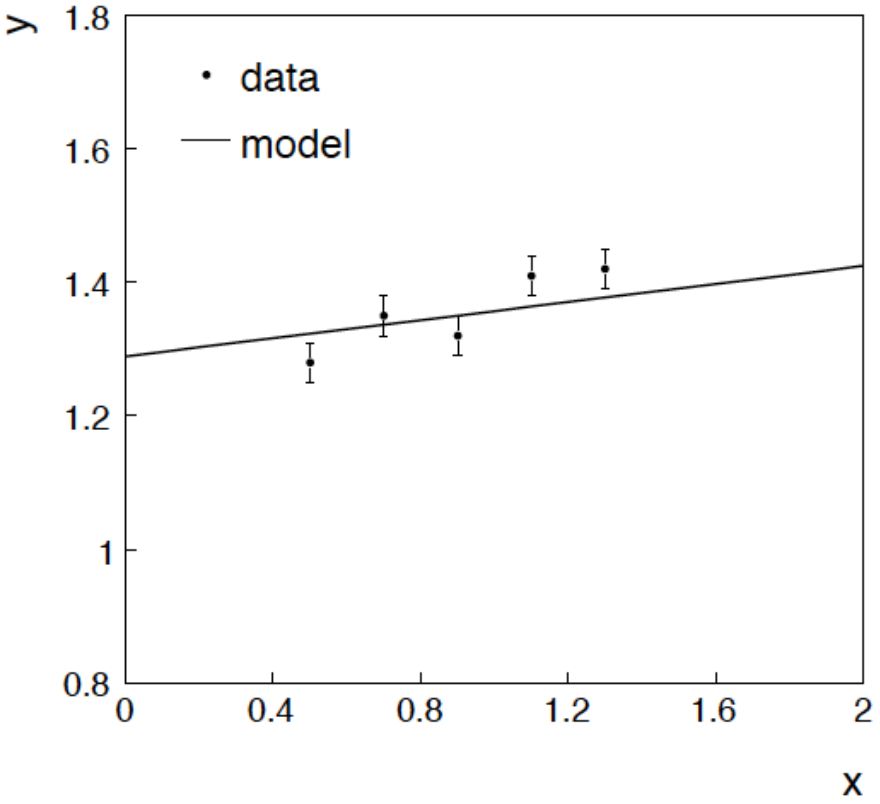
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x \ ,$$

assume x_i and σ_i known.

Goal: estimate θ_0

Here suppose we don't care about θ_1 (example of a "nuisance parameter")



Maximum likelihood fit with Gaussian data

In this example, the y_i are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right] ,$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

θ_1 known a priori

$$L(\theta_0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right] .$$

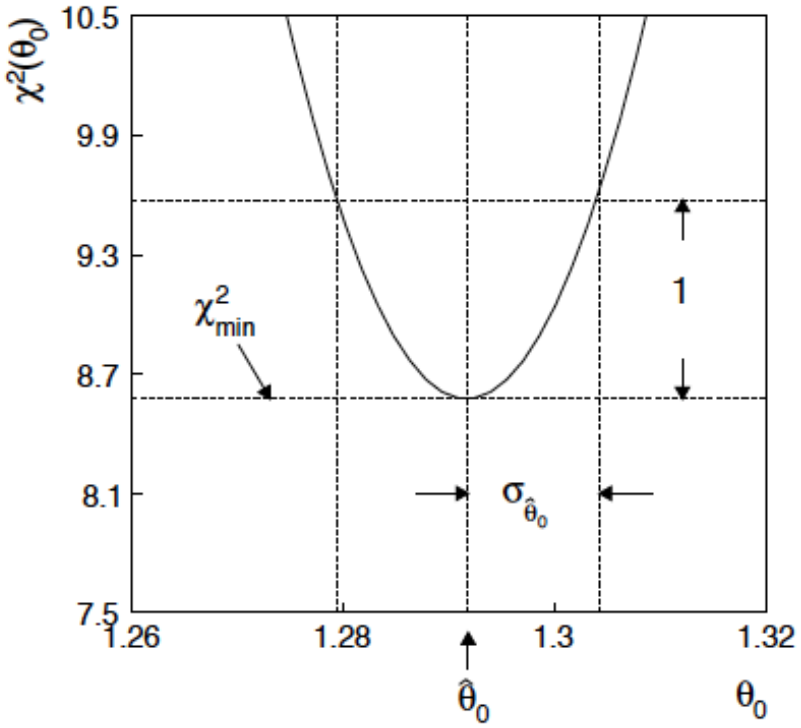
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i, ML same as LS

Minimize $\chi^2 \rightarrow$ estimator $\hat{\theta}_0$.

Come up one unit from χ^2_{min}

to find $\sigma_{\hat{\theta}_0}$.



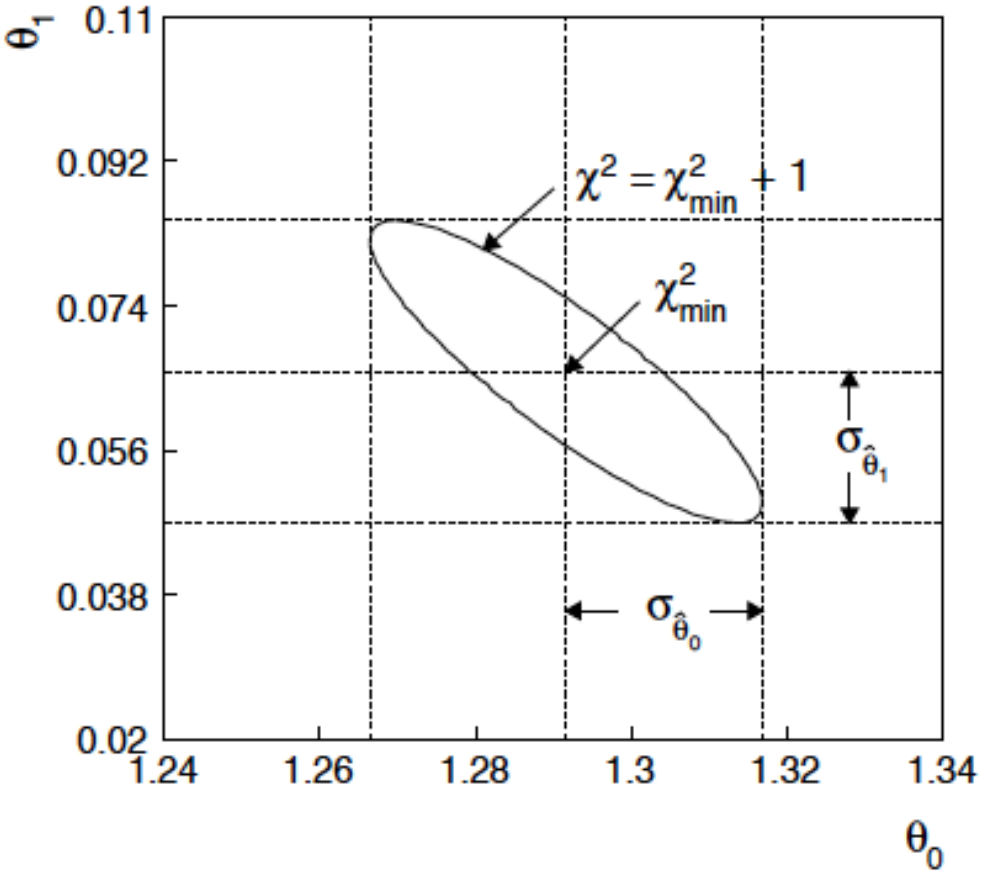
ML (or LS) fit of θ_0 and θ_1

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \, .$$

Standard deviations from tangent lines to contour

$$\chi^2 = \chi^2_{\text{min}} + 1 \, .$$

Correlation between $\hat{\theta}_0$, $\hat{\theta}_1$ causes errors to increase.

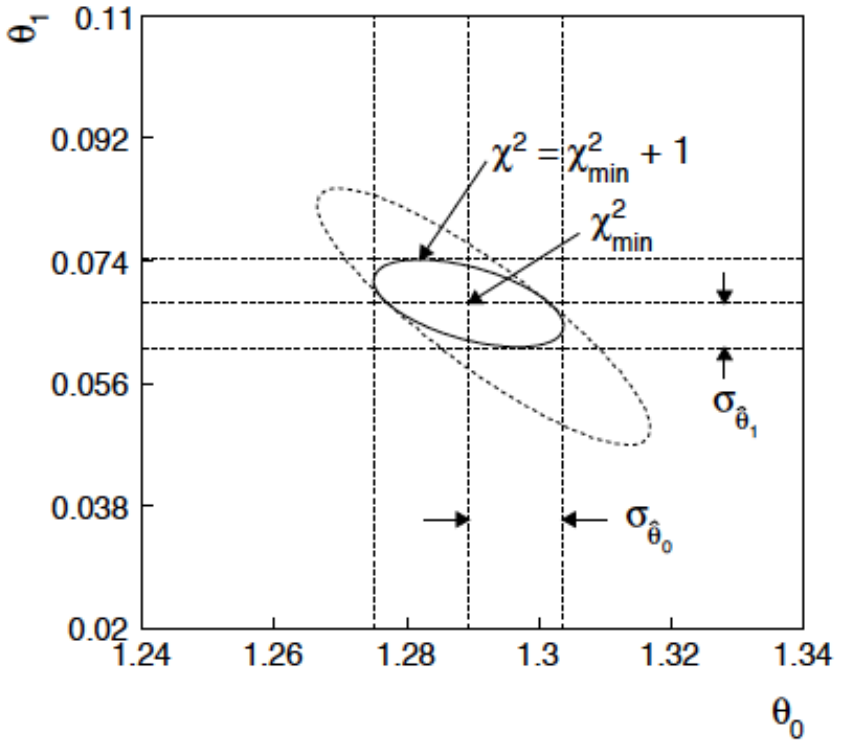


If we have a measurement $t_1 \sim \text{Gauss}\,(\theta_1, \sigma_{t_1})$

$$L(\theta_0, \theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1-\theta_1)^2/2\sigma_{t_1}^2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right]$$

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} + \frac{(t_1 - \theta_1)^2}{\sigma_{t_1}^2}$$

The information on θ_1 improves accuracy of $\hat{\theta}_0$.



Reminder of Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter value θ.

Interpret probability of θ as 'degree of belief' (subjective).

Need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Our experiment has data x, $\rightarrow$ likelihood $L(x|\theta)$.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

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

Posterior pdf $p(\theta|x)$ contains all our knowledge about θ.

Bayesian approach: $y_i \sim \text{Gauss}(\mu(x_i;\theta_0,\theta_1), \sigma_i)$

We need to associate prior probabilities with θ_0 and θ_1, e.g.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\pi_1(\theta_1)$$

← suppose knowledge of θ_0 has no influence on knowledge of θ_1

$$\pi_0(\theta_0) = \text{const.}$$

← 'non-informative', in any case much broader than $L(\theta_0)$

$$\pi_1(\theta_1) = p(\theta_1|t_1) \propto p(t_1|\theta_1)\pi_{\text{Ur}}(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1-\theta_1)^2/2\sigma_t^2} \times \text{const.}$$

prior after t_1, before $\boldsymbol{y}$

Ur = "primordial" prior

Likelihood for control measurement t_1

Bayesian example: $y_i \sim \text{Gauss}(\mu(x_i;\theta_0,\theta_1), \sigma_i)$

Putting the ingredients into Bayes' theorem gives:

$$p(\theta_0,\theta_1|\vec{y}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i-\mu(x_i;\theta_0,\theta_1))^2/2\sigma_i^2} \; \pi_0 \; \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1-t_1)^2/2\sigma_{t_1}^2}$$

posterior $\propto$ likelihood $\times$ prior

Note here the likelihood only reflects the measurements $\boldsymbol{y}$.

The information from the control measurement t_1 has been put into the prior for θ_1.

We would get the same result using the likelihood $P(\boldsymbol{y},t|\theta_0,\theta_1)$ and the constant "Ur-prior" for θ_1.

Marginalizing the posterior pdf

We then integrate (marginalize) $p(\theta_0, \theta_1 | \boldsymbol{y})$ to find $p(\theta_0 | \boldsymbol{y})$:

$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y})\, d\theta_1$$

In this example we can do the integral (rare). We find

$$p(\theta_0|\mathbf{y}) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma_{\theta_0^2}}$$

$$\hat{\theta}_0 = \text{same as MLE}$$

$$\sigma_{\theta_0} = \sigma_{\hat{\theta}_0} \qquad \text{(same as for MLE)}$$

For this example, numbers come out same as in frequentist approach, but interpretation different.

Marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

often high dimensionality and impossible in closed form,
also impossible with 'normal' acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized
Bayesian computation.

MCMC (e.g., Metropolis-Hastings algorithm) generates
correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than if all values independent .

Basic idea: sample multidimensional $\boldsymbol{\theta}$ but look only at
distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\boldsymbol{\theta})$, generate a sequence of points $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots$

Proposal density $q(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$ e.g. Gaussian centred about $\boldsymbol{\theta}_0$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

3) Form Hastings test ratio $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \text{Uniform}[0, 1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

Still works if $p(\boldsymbol{\theta})$ is known only as a proportionality, which is usually what we have from Bayes' theorem: $p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = q(\boldsymbol{\theta}_0; \boldsymbol{\theta})$

Test ratio is (*Metropolis*-Hastings): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

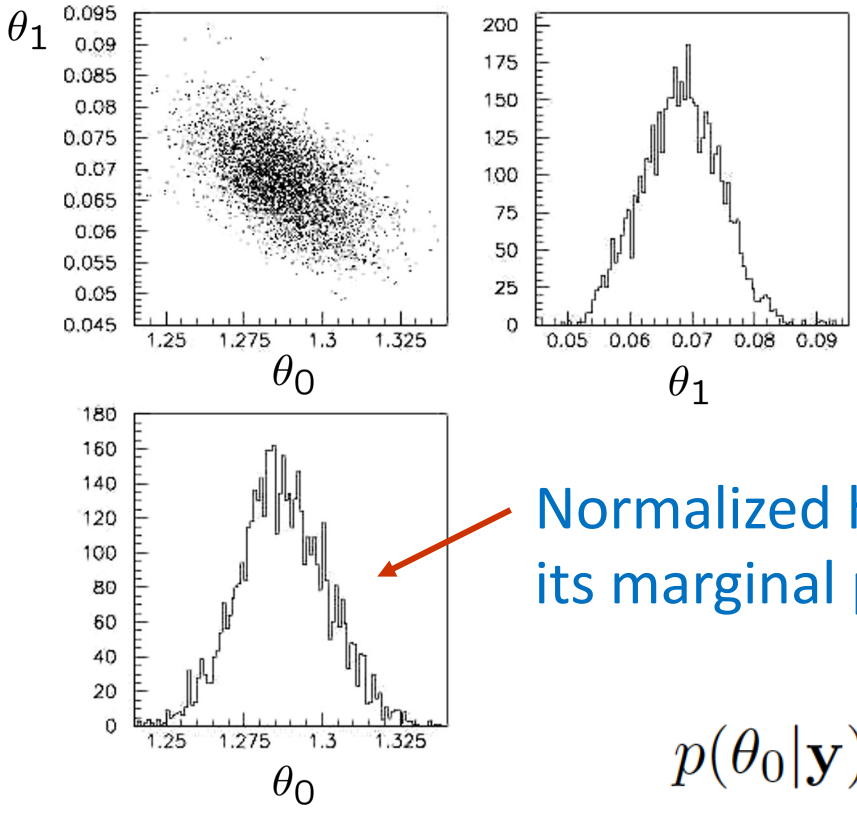
I.e. if the proposed step is to a point of higher $p(\boldsymbol{\theta})$, take it;

if not, only take the step with probability $p(\boldsymbol{\theta})/p(\boldsymbol{\theta}_0)$.

If proposed step rejected, repeat the current point.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Normalized histogram of θ_0 gives its marginal posterior pdf:

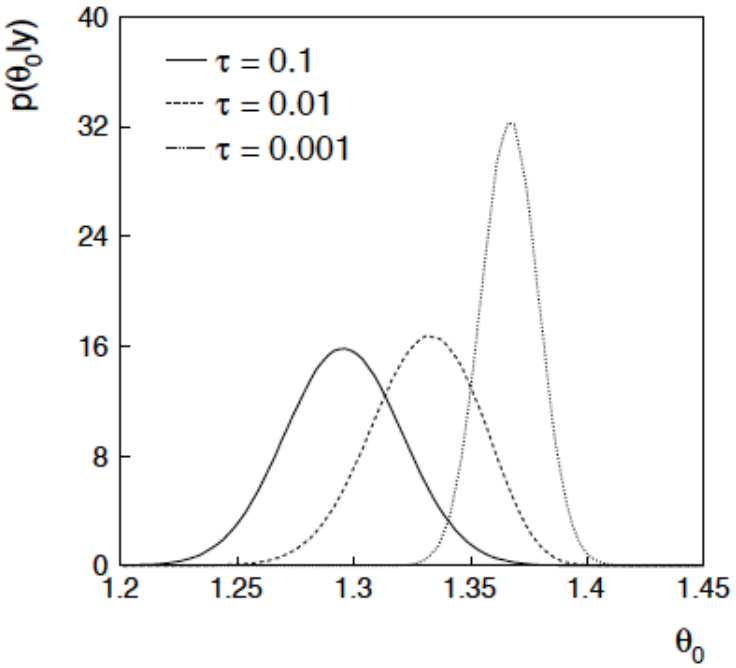
$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y})\, d\theta_1$$

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, an "expert" says it should be positive and not too much greater than 0.1 or so, i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau} , \quad \theta_1 \geq 0 , \quad \tau = 0.1 .$$

From this we obtain (numerically) the posterior pdf for θ_0:



This summarizes all knowledge about θ_0.

Look also at result from variety of priors.

Finally

Two lectures only enough for a brief introduction to:

- Parameter estimation
- Hypothesis tests ($\rightarrow$ path to Machine Learning)
- Limits (confidence intervals/regions)
- Systematics (nuisance parameters)
- A bit beyond... (Bayesian methods, MCMC)

Final thought: once the basic formalism is fixed, most of the work focuses on writing down the likelihood, e.g., $P(\boldsymbol{x}|\boldsymbol{\theta})$, and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches) so often best to invest most of your time with it.

Extra slides