Statistical Methods: Parameter Estimation

`https://agenda.infn.it/conferenceDisplay.py?confId=12288`

INFN School of Statistics

Ischia, 7-10 May 2017

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

Rough outline

Parameter estimation

The parameters of a pdf are constants that characterize its shape, e.g.

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

r.v.   parameter

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

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

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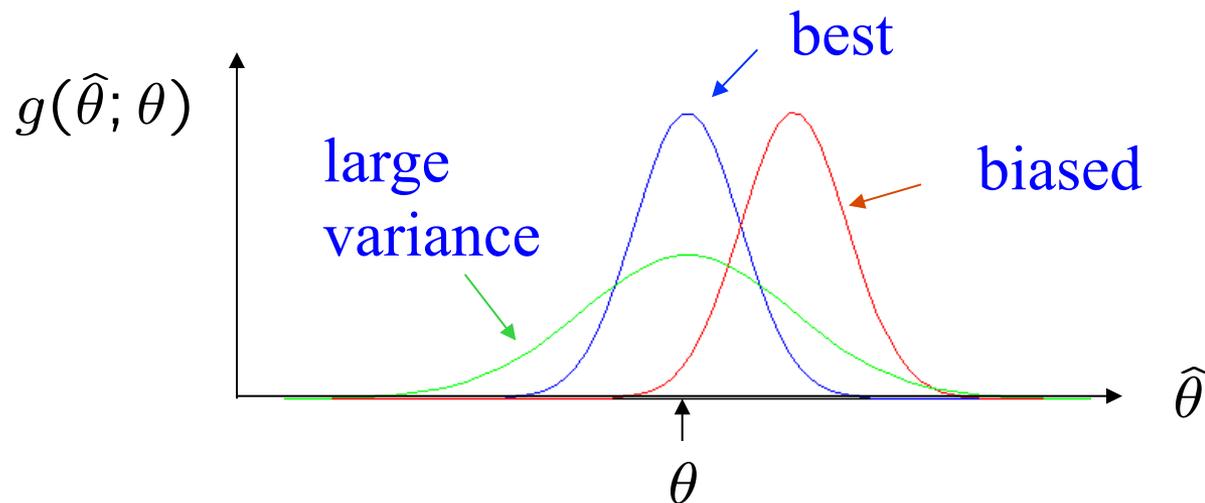
We want small (or zero) bias (systematic error): $b = E[\hat{\theta}] - \theta$

→ 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

An estimator for the mean (expectation value)

Parameter: $\mu = E[x] = \langle x \rangle = \int_{-\infty}^{\infty} x f(x)\, dx$

Estimator: $\widehat{\mu} = \dfrac{1}{n} \sum_{i=1}^{n} x_i \equiv \overline{x}$ ('sample mean')

We find: $b = E[\widehat{\mu}] - \mu = 0$

$$V[\widehat{\mu}] = \frac{\sigma^2}{n} \qquad \left(\sigma_{\widehat{\mu}} = \frac{\sigma}{\sqrt{n}} \right)$$

An estimator for the variance

Parameter: $\sigma^2 = V[x] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$

Estimator: $\widehat{\sigma^2} = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \equiv s^2$ ('sample variance')

We find:

$$b = E[\widehat{\sigma^2}] - \sigma^2 = 0 \quad \text{(factor of } n{-}1 \text{ makes this so)}$$

$$V[\widehat{\sigma^2}] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2 \right), \quad \text{where}$$

$$\mu_k = \int (x - \mu)^k f(x)\, dx$$

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*:

$$P(\mathbf{x}|\boldsymbol{\theta})$$

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

$$L(\boldsymbol{\theta}) = P(\mathbf{x}|\boldsymbol{\theta})$$ (*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.
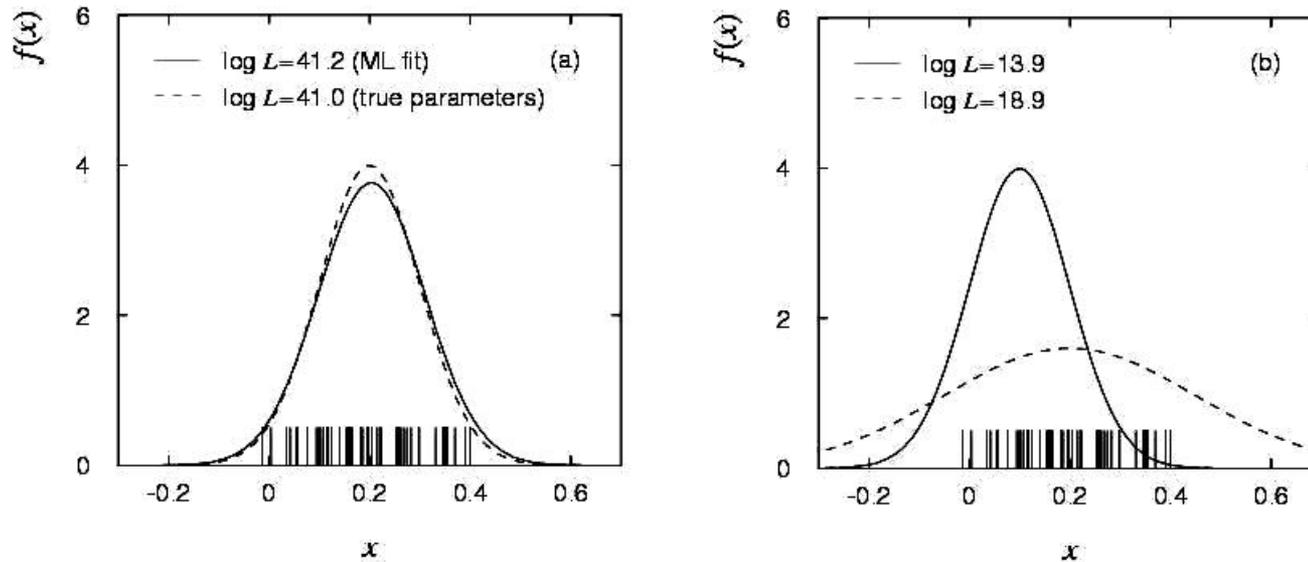
So we define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

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

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

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

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 $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0$,

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

Monte Carlo test:
    generate 50 values
    using $t = 1$:

We find the ML estimate:

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

INFN School of Statistics, Ischia, 7-10 May 2017
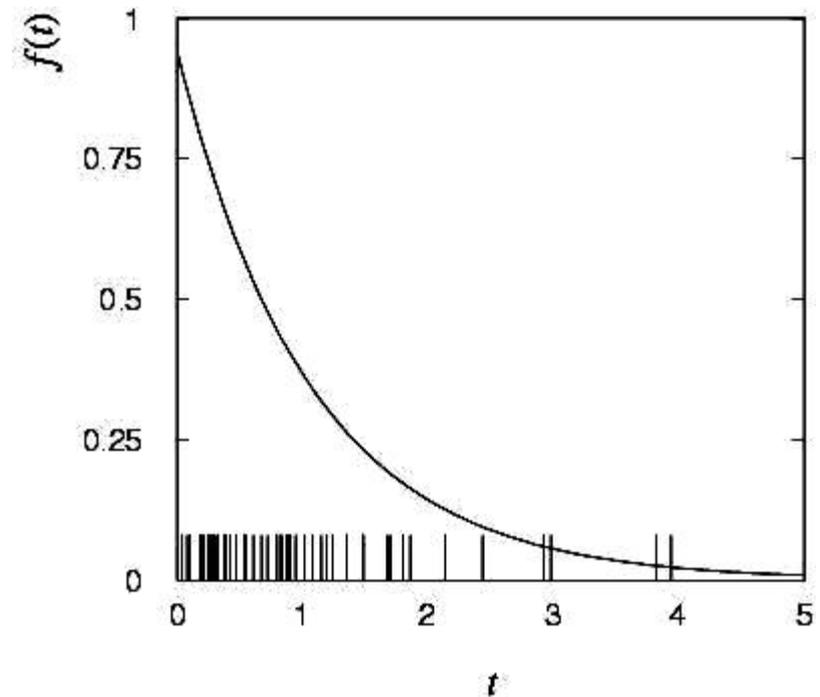
ML example: parameter of exponential pdf (3)

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 ML estimator $\quad \hat{\tau} = \dfrac{1}{n} \displaystyle\sum_{i=1}^n t_i \quad$ we therefore find

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

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

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 (with unique inverse) $\lambda(\tau)$ of a parameter τ, it doesn't matter whether we express L as a function of λ or τ.

The ML estimator of a function $\lambda(\tau)$ is simply $\hat{\lambda} = \lambda(\hat{\tau})$

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

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

Can show $E[\hat{\lambda}] = \lambda \dfrac{n}{n-1}$. (bias $\to 0$ for $n \to \infty$)

Example of ML: parameters of Gaussian pdf

Consider independent $x_1, ..., x_n$, 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

$$\ln L(\mu, \sigma^2) = \sum_{i=1}^{n} \ln f(x_i; \mu, \sigma^2)$$

$$= \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \ln \frac{1}{\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right) .$$

Example of ML: parameters of Gaussian pdf (2)

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

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

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

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

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.
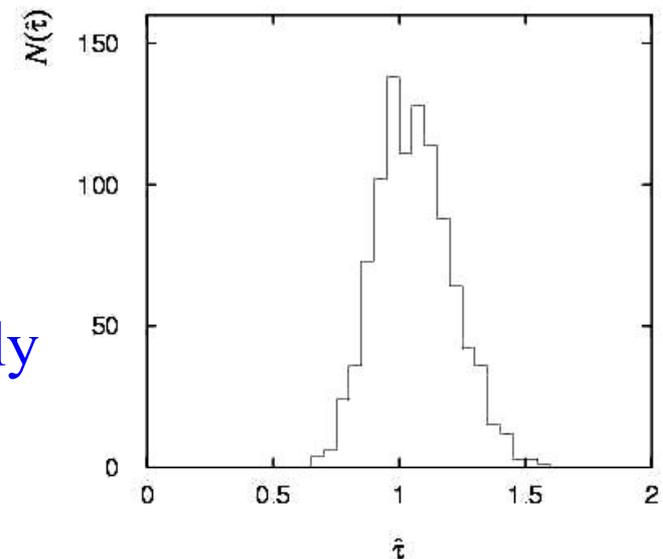
Variance of estimators from information inequality

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

Minimum Variance Bound (MVB)

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

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

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:

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

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):

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

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

$\rightarrow$ 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:

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

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

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

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$
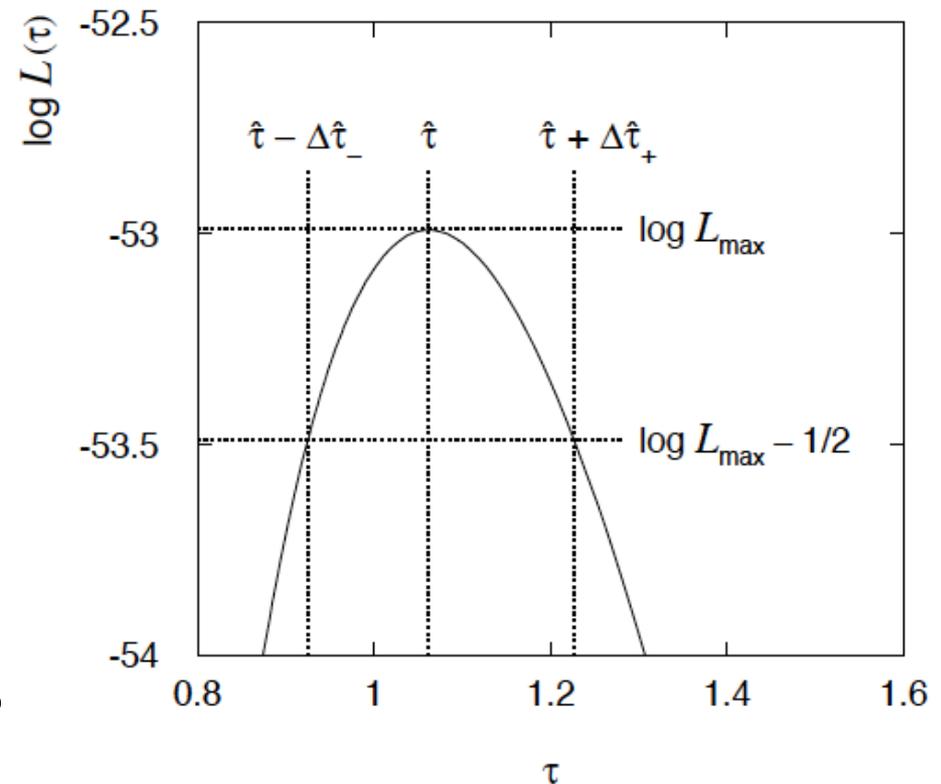
Not quite parabolic ln L since finite sample size ($n = 50$).

Information inequality for *n* parameters

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

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

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

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 $x = \cos\theta$,

$$f(x; \alpha, \beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta/3}$$
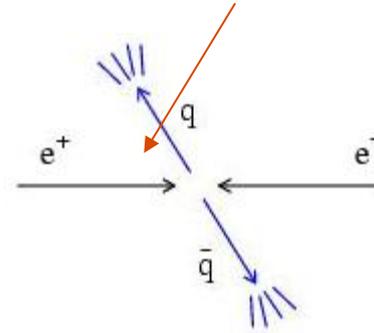
or if $x_{min} < x < x_{max}$, need always to normalize so that

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

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).
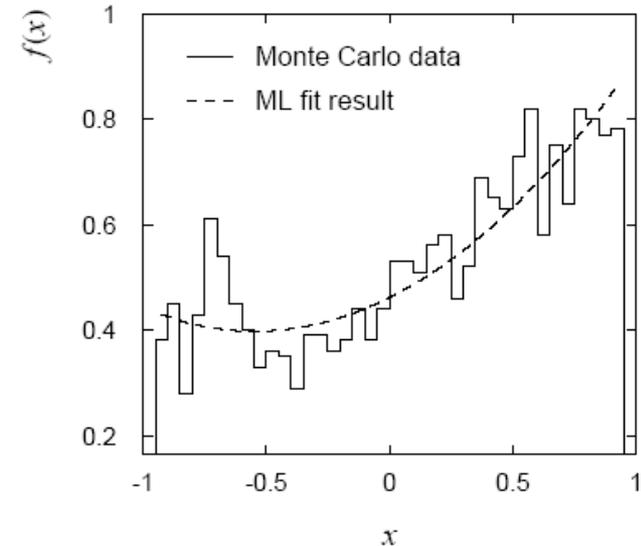


(Co)variances from $(\widehat{V^{-1}})_{ij} = -\dfrac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\Big|_{\vec{\theta}=\hat{\vec{\theta}}}$ (**MINUIT** routine **HESSE**)

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

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

Two-parameter fit: MC study

Repeat ML fit with 500 experiments, all with $n = 2000$ events:
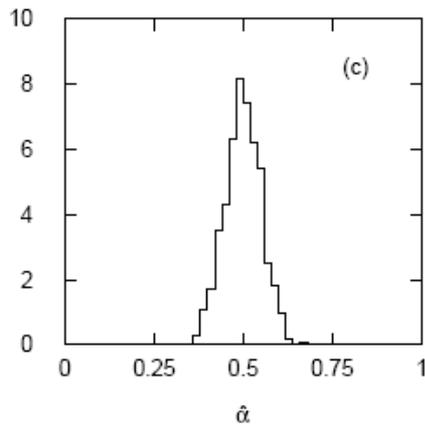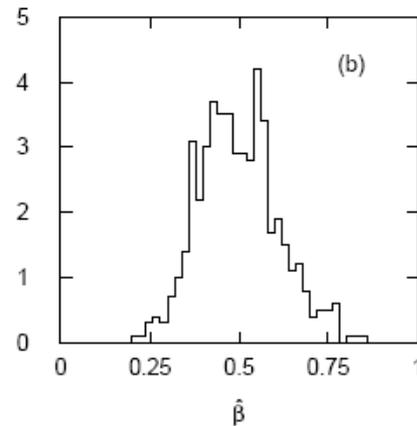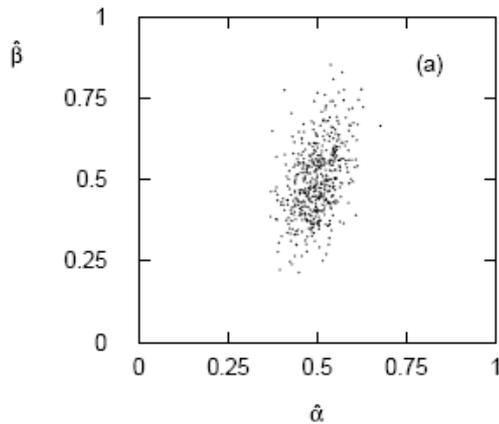
$$\overline{\hat{\alpha}} = 0.499$$

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

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

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

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

$$r = 0.42$$

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

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

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

$$\ln L(\alpha, \beta) \approx \ln L_{\mathsf{max}}$$

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

The contour $\ln L(\alpha, \beta) = \ln L_{\mathsf{max}} - 1/2$ 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
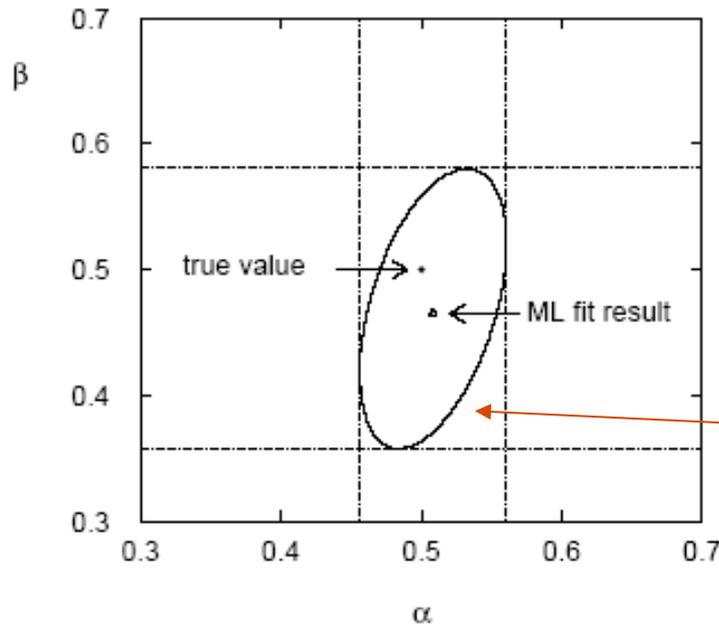
The α, β plane for the first MC data set

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

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

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

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

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 $\rightarrow$ smaller errors for $\widehat{\vec{\theta}}$

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:

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

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

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:

$$\mu_\mathsf{s} = 6$$

$$\mu_\mathsf{b} = 60$$

Maximize log-likelihood in terms of μ_s and μ_b:

$$\widehat{\mu}_\mathsf{s} = 8.7 \pm 5.5$$

$$\widehat{\mu}_\mathsf{b} = 54.3 \pm 8.8$$

Here errors reflect total Poisson fluctuation as well as that in proportion of signal/background.
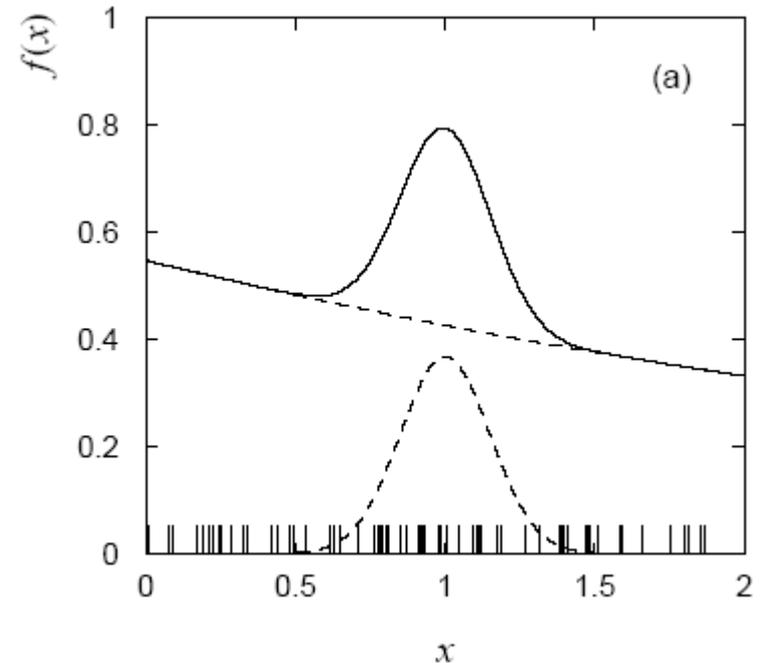
Extended ML example: an unphysical estimate

A downwards fluctuation of data in the peak region can lead to even fewer events than what would be obtained from background alone.

Estimate for μ_s here pushed negative (unphysical).

We can let this happen as long as the (total) pdf stays positive everywhere.
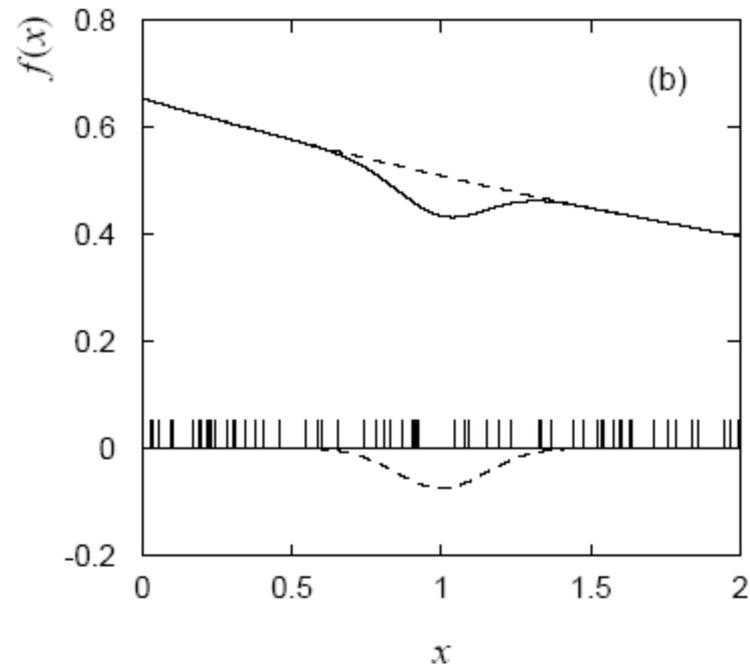
Unphysical estimators (2)

Here the unphysical estimator is unbiased and should nevertheless be reported, since average of a large number of unbiased estimates converges to the true value (cf. PDG).

Repeat entire MC experiment many times, allow unphysical estimates:
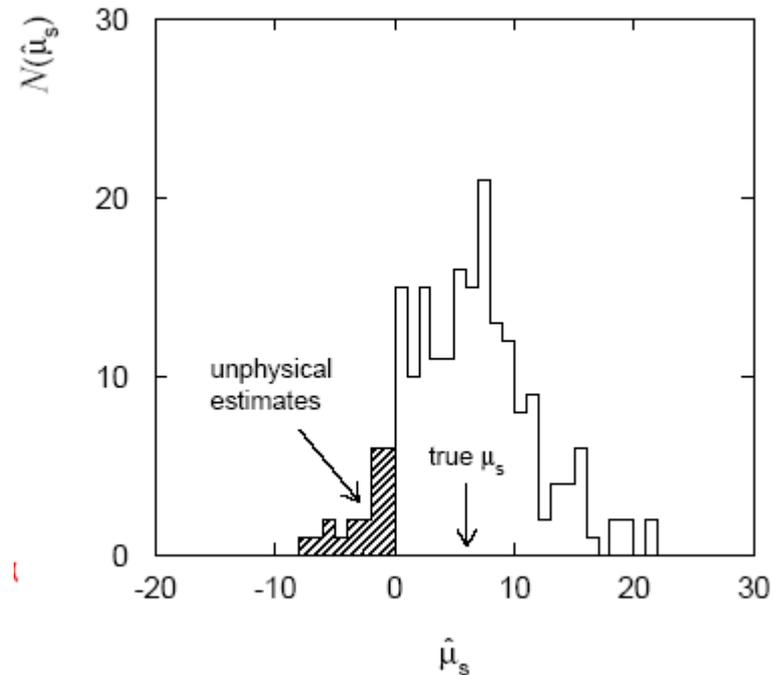
ML with binned data

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

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

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

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

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

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

ML example with binned data

Previous example with exponential, now put data into histogram:
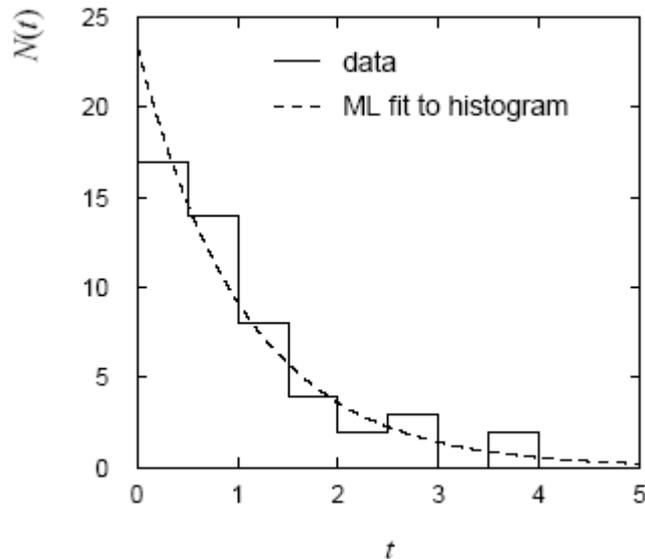
$\hat{\tau} = 1.07 \pm 0.17$

$(1.06 \pm 0.15$ for unbinned

ML with same sample$)$

Limit of zero bin width $\rightarrow$ usual unbinned ML.

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

In Bayesian statistics, both θ and x are random variables:

$$L(\theta) = L(\vec{x}|\theta) = f_{\text{joint}}(\vec{x}|\theta)$$

Recall the Bayesian method:

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 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,

$\rightarrow$ summarize using an estimator $\hat{\theta}_{\text{Bayes}}$

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

$$\hat{\theta}_{\text{Bayes}} = \hat{\theta}_{\text{ML}}$$

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

'Complete prior ignorance' is not well defined.

Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called "objective priors"
Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In a Subjective Bayesian analysis, using objective priors can be an important part of the sensitivity analysis.

Priors from formal rules (cont.)

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes' theorem as a recipe to produce an interval with certain coverage properties. For a review see:

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in HEP, but there is recent interest in this direction; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, arxiv:1002.1111 (Feb 2010)

Jeffreys' prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

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

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters.

For a Gaussian mean, the Jeffreys' prior is constant; for a Poisson mean μ it is proportional to $1/\sqrt{\mu}$.

"Invariance of inference" with Jeffreys' prior

Suppose we have a parameter θ, to which we assign a prior $\pi_\theta(\theta)$.

An experiment gives data x, modeled by $L(\theta) = P(x|\theta)$.

Bayes' theorem then tells us the posterior for θ:

$$P(\theta|x) \propto P(x|\theta)\pi_\theta(\theta)$$

Now consider a function $\eta(\theta)$, and we want the posterior $P(\eta|x)$.

This must follow from the usual rules of transformation of random variables:

$$P(\eta|x) = P(\theta(\eta)|x) \left| \frac{d\theta}{d\eta} \right|$$

"Invariance of inference" with Jeffreys' prior (2)

Alternatively, we could have just starting with η as the parameter in our model, and written down a prior pdf $\pi_\eta(\eta)$.

Using it, we express the likelihood as $L(\eta) = P(x|\eta)$ and write Bayes' theorem as

$$P(\eta|x) \propto P(x|\eta)\pi_\eta(\eta)$$

If the priors really express our degree of belief, then they must be related by the usual laws of probability $\pi_\eta(\eta) = \pi_\theta(\theta(\eta)) |d\theta/d\eta|$, and in this way the two approaches lead to the same result.

But if we choose the priors according to "formal rules", then this is not guaranteed. For the Jeffrey's prior, however, it does work!

Using $\pi_\theta(\theta) \propto \sqrt{I(\theta)}$ and transforming to find $P(\eta|x)$ leads to the same as using $\pi_\eta(\eta) \propto \sqrt{I(\eta)}$ directly with Bayes' theorem.

Jeffreys' prior for Poisson mean

Suppose $n \sim \text{Poisson}(\mu)$. To find the Jeffreys' prior for μ,

$$L(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \qquad\qquad \frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\mu^2}$$

$$I = -E\left[\frac{\partial^2 \ln L}{\partial \mu^2}\right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for $\mu = s + b$, this means the prior $\pi(s) \sim 1/\sqrt{(s + b)}$, which depends on b. But this is not designed as a degree of belief about s.

The method of least squares

Suppose we measure N values, $y_1, ..., y_N$, assumed to be independent Gaussian r.v.s with

$$E[y_i] = \lambda(x_i; \theta) \ .$$

Assume known values of the control variable $x_1, ..., x_N$ and known variances

$$V[y_i] = \sigma_i^2 \ .$$

We want to estimate θ, i.e., fit the curve to the data points.

The likelihood function is

$$L(\theta) = \prod_{i=1}^{N} f(y_i; \theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - \lambda(x_i; \theta))^2}{2\sigma_i^2}\right]$$
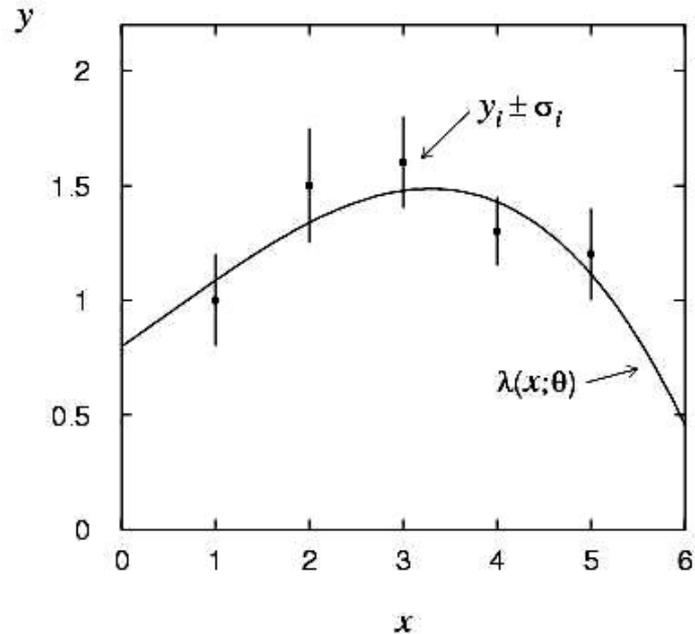
The method of least squares (2)

The log-likelihood function is therefore

$$\ln L(\theta) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \theta))^2}{\sigma_i^2} + \text{ terms not depending on } \theta$$

So maximizing the likelihood is equivalent to minimizing

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \theta))^2}{\sigma_i^2}$$

Minimum defines the least squares (LS) estimator $\hat{\theta}$.

Very often measurement errors are ~Gaussian and so ML and LS are essentially the same.

Often minimize χ^2 numerically (e.g. program **MINUIT**).

LS with correlated measurements

If the y_i follow a multivariate Gaussian, covariance matrix V,

$$g(\vec{y}, \vec{\lambda}, V) = \frac{1}{(2\pi)^{N/2} |V|^{1/2}} \exp\left[-\frac{1}{2} (\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda}) \right]$$

Then maximizing the likelihood is equivalent to minimizing

$$\chi^2(\vec{\theta}) = \sum_{i,j=1}^{N} (y_i - \lambda(x_i; \vec{\theta})) (V^{-1})_{ij} (y_j - \lambda(x_j; \vec{\theta}))$$

Linear LS problem

LS has particularly simple properties if $\lambda(x; \vec{\theta})$ linear in $\vec{\theta}$:

$$\lambda(x; \vec{\theta}) = \sum_{i=1}^{m} a_j(x) \theta_j$$

where $a_j(x)$ are any linearly independent functions of x.

Matrix notation: let $A_{ij} = a_j(x_i)$,

$$\chi^2(\vec{\theta}) = (\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda})$$

$$= (\vec{y} - A\vec{\theta})^T V^{-1} (\vec{y} - A\vec{\theta})$$

Linear LS problem (2)

Set derivitives with respect to θ_i to zero,

$$\nabla \chi^2 = -2(A^T V^{-1} \vec{y} - A^T V^{-1} A \vec{\theta}) = 0$$

Solve to get the LS estimators,

$$\hat{\vec{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \vec{y} \equiv B \vec{y}$$

N.B. estimators $\hat{\theta}_i$ are linear functions of the measurements y_i.

Linear LS problem (3)

Error propagation (exact for linear problem) for $U_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$:

$$U = B\,V\,B^T = (A^T V^{-1} A)^{-1}$$

Equivalently, use

$$(U^{-1})_{ij} = \frac{1}{2}\left[\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j}\right]_{\vec{\theta}=\hat{\vec{\theta}}}$$

Equals MVB if y_i Gaussian)

Example of least squares fit

Fit a polynomial of order p: $\quad \lambda(x; \theta_0, \ldots, \theta_p) = \displaystyle\sum_{n=0}^{p} \theta_n x^n$
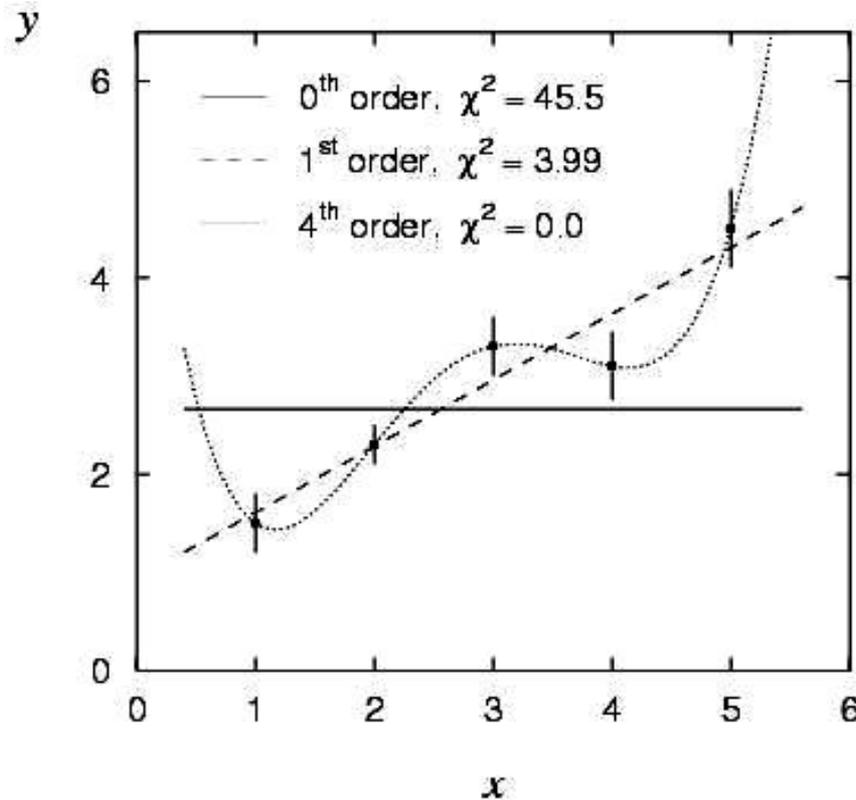
Variance of LS estimators

In most cases of interest we obtain the variance in a manner similar to ML. E.g. for data ~ Gaussian we have

$$\chi^2(\theta) = -2 \ln L(\theta) + C$$

and so

$$\widehat{\sigma^2}_{\widehat{\theta}} \approx 2 \left[\frac{\partial^2 \chi^2}{\partial \theta^2} \right]^{-1}_{\theta = \widehat{\theta}}$$

or for the graphical method we take the values of θ where

$$\chi^2(\theta) = \chi^2_{min} + 1$$
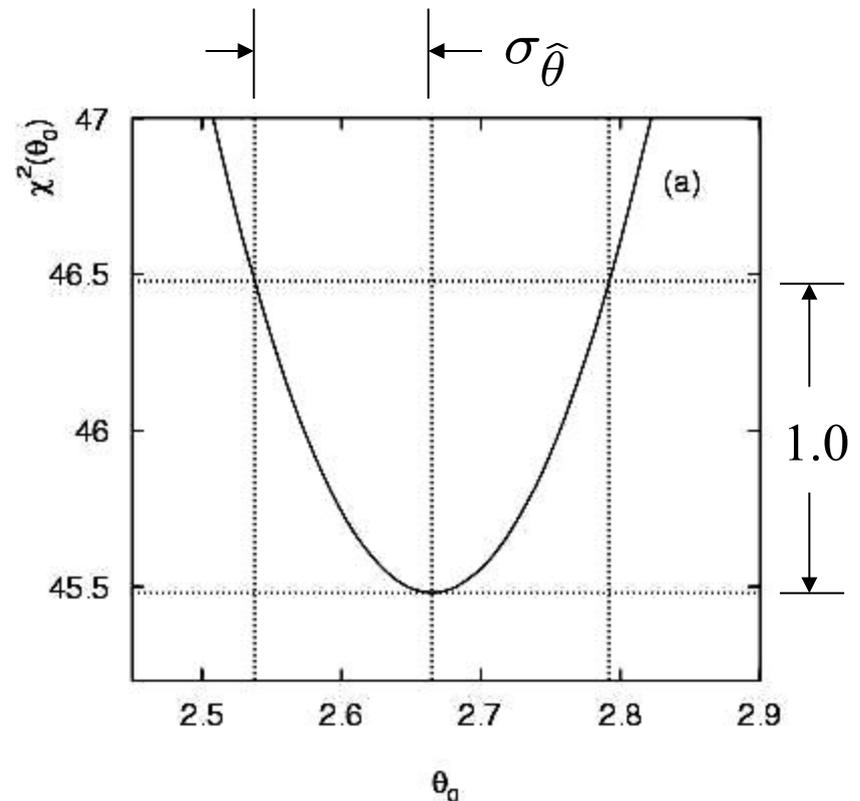
Two-parameter LS fit

2-parameter case (line with nonzero slope):

$$\hat{\theta}_0 = 0.93 \pm 0.30,$$

$$\hat{\theta}_1 = 0.68 \pm 0.10$$

$$\widehat{\text{cov}}[\hat{\theta}_0, \hat{\theta}_1] = -0.028$$

$$r = -0.90$$

$$\chi^2 = 3.99$$
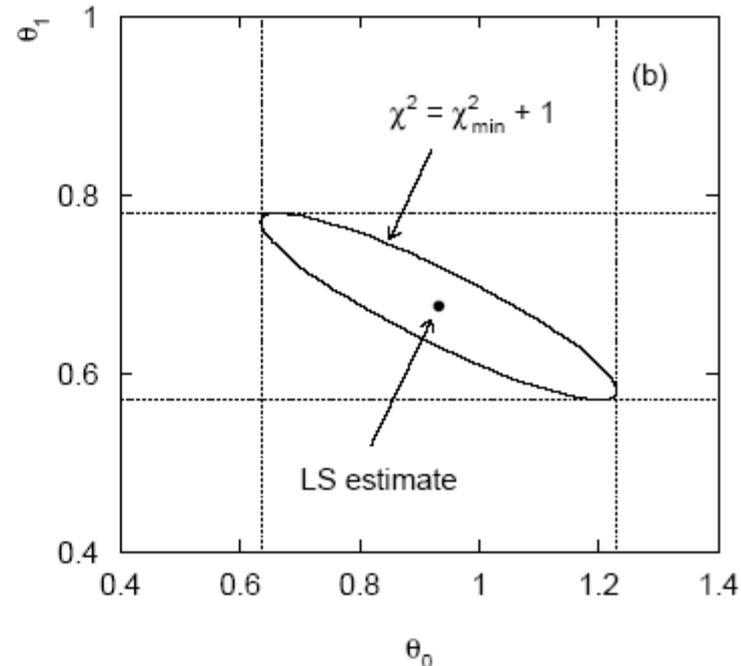
Tangent lines $\rightarrow \sigma_{\hat{\theta}_0}, \sigma_{\hat{\theta}_1}$.

Angle of ellipse $\rightarrow$ correlation (same as for ML)

Goodness-of-fit with least squares

The value of the χ^2 at its minimum is a measure of the level of agreement between the data and fitted curve:

$$\chi^2_{\min} = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \hat{\theta}))^2}{\sigma_i^2}$$

It can therefore be employed as a goodness-of-fit statistic to test the hypothesized functional form $\lambda(x; \theta)$.

We can show that if the hypothesis is correct, then the statistic $t = \chi^2_{\min}$ follows the chi-square pdf,

$$f(t; n_{\mathrm{d}}) = \frac{1}{2^{n_{\mathrm{d}}/2} \Gamma(n_{\mathrm{d}}/2)} t^{n_{\mathrm{d}}/2 - 1} e^{-t/2}$$

where the number of degrees of freedom is

$$n_{\mathrm{d}} = \text{number of data points} - \text{number of fitted parameters}$$

Goodness-of-fit with least squares (2)

The chi-square pdf has an expectation value equal to the number of degrees of freedom, so if $\chi^2_{min} \approx n_d$ the fit is 'good'.

More generally, find the p-value:
$$p = \int_{\chi^2_{min}}^{\infty} f(t; n_d)\, dt$$

This is the probability of obtaining a χ^2_{min} as high as the one we got, or higher, if the hypothesis is correct.

E.g. for the previous example with 1st order polynomial (line),

$$\chi^2_{min} = 3.99\,, \qquad n_d = 5-2 = 3\,, \qquad p = 0.263$$

whereas for the 0th order polynomial (horizontal line),

$$\chi^2_{min} = 45.5\,, \qquad n_d = 5-1 = 4\,, \qquad p = 3.1 \times 10^{-9}$$

Goodness-of-fit vs. statistical errors

Small statistical error does not mean a good fit (nor vice versa).

Curvature of χ^2 near its minimum $\rightarrow$ statistical errors $(\sigma_{\hat{\theta}})$

Value of $\chi^2_{\min}$ $\rightarrow$ goodness-of-fit

Horizontal line fit, move the data points, keep errors on points same:

$\hat{\theta}_0 = 2.84 \pm 0.13$

$\chi^2_{\min} = 4.48$

Variance same as before,

now $\chi^2_{\min}$ 'good'.
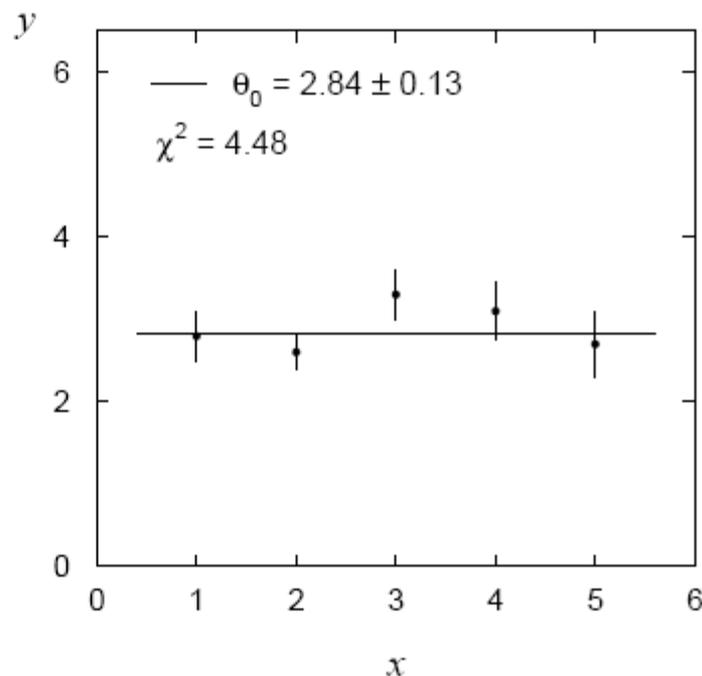
Goodness-of-fit vs. stat. errors (2)

$\rightarrow \chi^2(\theta_0)$ shifted down, same curvature as before.

Variance of estimator (statistical error) tells us:

if experiment repeated many times, how wide is the distribution of the estimates $\hat{\theta}$. (Doesn't tell us whether hypothesis correct.)

P-value tells us:

if hypothesis is correct and experiment repeated many times, what fraction will give equal or worse agreement between data and hypothesis according to the statistic $\chi^2_{\min}$.

Low P-value $\rightarrow$ hypothesis may be wrong $\rightarrow$ systematic error.

LS with binned data

Histogram:

N bins, n entries.

Hypothesized pdf:

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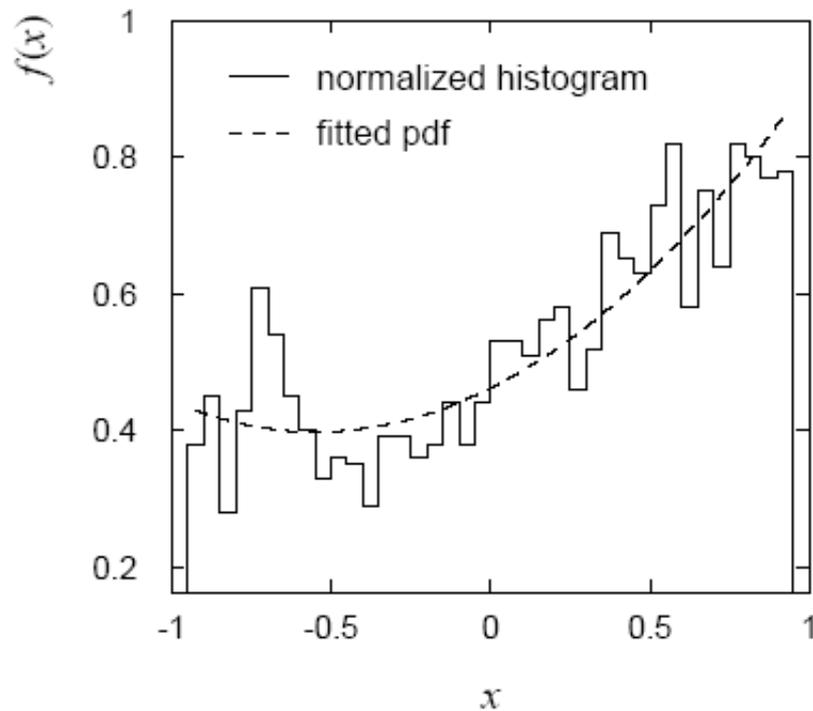
We have

$$y_i = \text{number of entries in bin } i,$$

$$\lambda_i(\vec{\theta}) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = n p_i(\vec{\theta})$$

LS with binned data (2)

LS fit: minimize

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\vec{\theta}))^2}{\sigma_i^2}$$

where $\sigma_i^2 = V[y_i]$, here not known a priori.

Treat the y_i as Poisson r.v.s, in place of true variance take either

$$\sigma_i^2 = \lambda_i(\vec{\theta}) \qquad \text{(LS method)}$$

$$\sigma_i^2 = y_i \qquad \text{(Modified LS method)}$$

MLS sometimes easier computationally, but $\chi^2_{\min}$ no longer follows chi-square pdf (or is undefined) if some bins have few (or no) entries.

LS with binned data — normalization

Do not 'fit the normalization':

$$\lambda_i(\vec{\theta}, \nu) = \nu \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = \nu p_i(\vec{\theta})$$

i.e. introduce adjustable ν, fit along with $\vec{\theta}$.

$\hat{\nu}$ is a bad estimator for n (which we know, anyway!)

$$\hat{\nu}_{\text{LS}} = n + \frac{\chi^2_{\min}}{2}$$

$$\hat{\nu}_{\text{MLS}} = n - \chi^2_{\min}$$

LS normalization example

Example with $n = 400$ entries, $N = 20$ bins:
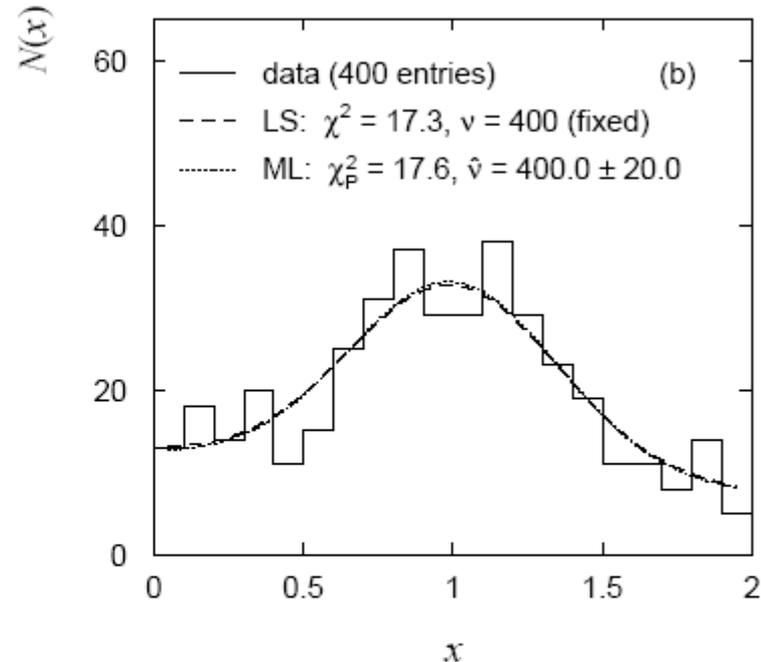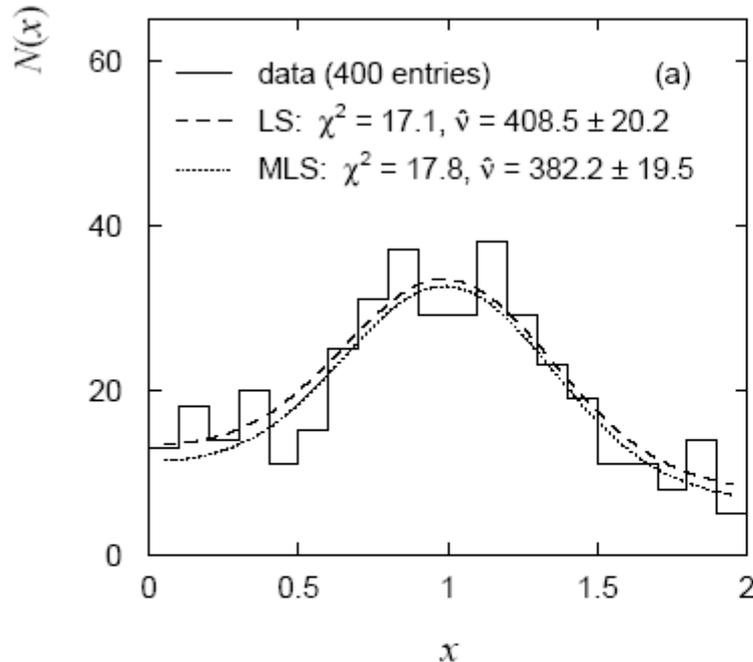
Expect $\chi^2_{\min}$ around $N - m$,

$\rightarrow$ relative error in $\hat{\nu}$ large when N large, n small

Either get n directly from data for LS (or better, use ML).

Goodness of fit from the likelihood ratio

Suppose we model data using a likelihood $L(\boldsymbol{\mu})$ that depends on N parameters $\boldsymbol{\mu} = (\mu_1,..., \mu_N)$. Define the statistic

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})}$$

Value of $t_{\boldsymbol{\mu}}$ reflects agreement between hypothesized $\boldsymbol{\mu}$ and the data.

Good agreement means $\hat{\boldsymbol{\mu}} \approx \boldsymbol{\mu}$, so $t_{\boldsymbol{\mu}}$ is small;

Larger $t_{\boldsymbol{\mu}}$ means less compatibility between data and $\boldsymbol{\mu}$.

Quantify "goodness of fit" with p-value: $p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\text{obs}}}^{\infty} f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}) \, dt_{\boldsymbol{\mu}}$

Likelihood ratio (2)

Now suppose the parameters $\boldsymbol{\mu} = (\mu_1,...,\mu_N)$ can be determined by another set of parameters $\boldsymbol{\theta} = (\theta_1,...,\theta_M)$, with $M < N$.

E.g. in LS fit, use $\mu_i = \mu(x_i; \boldsymbol{\theta})$ where x is a control variable.

Define the statistic

fit M parameters

$$q_\mu = -2\ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})}$$
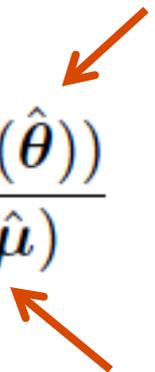
fit N parameters

Use $q_{\boldsymbol{\mu}}$ to test hypothesized functional form of $\mu(x; \boldsymbol{\theta})$.

To get p-value, need pdf $f(q_{\boldsymbol{\mu}}|\mu)$.

Wilks' Theorem (1938)

Wilks' Theorem: if the hypothesized parameters $\boldsymbol{\mu} = (\mu_1,..., \mu_N)$ are true then in the large sample limit (and provided certain conditions are satisfied) $t_{\boldsymbol{\mu}}$ and $q_{\boldsymbol{\mu}}$ follow chi-square distributions.

For case with $\boldsymbol{\mu} = (\mu_1,..., \mu_N)$ fixed in numerator:

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})} \qquad f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}) \sim \chi^2_N$$

Or if M parameters adjusted in numerator,

degrees of freedom

$$q_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})} \qquad f(q_{\boldsymbol{\mu}}|\boldsymbol{\mu}) \sim \chi^2_{N-M}$$

S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

Goodness of fit with Gaussian data

Suppose the data are N independent Gaussian distributed values:

$$y_i \sim \text{Gauss}(\mu_i, \sigma_i), \qquad i = 1, \ldots, N$$

want to estimate   known

Likelihood:
$$L(\boldsymbol{\mu}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu_i)^2 / 2\sigma_i^2}$$

Log-likelihood:
$$\ln L(\boldsymbol{\mu}) = -\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} + C$$

ML estimators:
$$\hat{\mu}_i = y_i \qquad i = 1, \ldots, N$$

Likelihood ratios for Gaussian data

The goodness-of-fit statistics become

$$t_\mu = -2 \ln \frac{L(\mu)}{L(\hat{\mu})} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} \qquad f(t_\mu | \mu) \sim \chi_N^2$$

$$q_\mu = -2 \ln \frac{L(\mu(\hat{\theta}))}{L(\hat{\mu})} = \sum_{i=1}^{N} \frac{(y_i - \mu_i(\hat{\theta}))^2}{\sigma_i^2} \qquad f(q_\mu | \mu) \sim \chi_{N-M}^2$$

So Wilks' theorem formally states the well-known property of the minimized chi-squared from an LS fit.

Likelihood ratio for Poisson data

Suppose the data are a set of values $n = (n_1,..., n_N)$, e.g., the numbers of events in a histogram with N bins.

Assume $n_i \sim \text{Poisson}(\nu_i)$, $i = 1,..., N$, all independent.

Goal is to estimate $\nu = (\nu_1,..., \nu_N)$.

Likelihood:
$$L(\nu) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

Log-likelihood:
$$\ln L(\nu) = \sum_{i=1}^{N} [n_i \ln \nu_i - \nu_i] + C$$

ML estimators:
$$\hat{\nu}_i = n_i, \qquad i = 1, \ldots, N$$

Goodness of fit with Poisson data

The likelihood ratio statistic (all parameters fixed in numerator):

$$
\begin{aligned}
t_{\boldsymbol{\nu}} &= -2 \ln \frac{L(\boldsymbol{\nu})}{L(\hat{\boldsymbol{\nu}})} \\
&= -2 \sum_{i=1}^{N} \left[n_i \ln \frac{\nu_i}{\hat{\nu}_i} - \nu_i + \hat{\nu}_i \right] \\
&= -2 \sum_{i=1}^{N} \left[n_i \ln \frac{\nu_i}{n_i} - \nu_i + n_i \right]
\end{aligned}
$$

Wilks' theorem: $f(t_{\boldsymbol{\nu}} | \boldsymbol{\nu}) \sim \chi_N^2$

Goodness of fit with Poisson data (2)

Or with *M* fitted parameters in numerator:

$$q_\nu = -2 \ln \frac{L(\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\nu}})} = -2 \sum_{i=1}^{N} \left[n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i} - \nu_i(\hat{\boldsymbol{\theta}}) + n_i \right]$$

Wilks' theorem: $f(q_\nu | \boldsymbol{\nu}) \sim \chi^2_{N-M}$

Use t_μ, q_μ to quantify goodness of fit (*p*-value).

Sampling distribution from Wilks' theorem (chi-square).

Exact in large sample limit; in practice good approximation for surprisingly small n_i (~several).

Goodness of fit with multinomial data

Similar if data $\boldsymbol{n} = (n_1,..., n_N)$ follow multinomial distribution:

$$P(\mathbf{n}|\mathbf{p}, n_{\text{tot}}) = \frac{n_{\text{tot}}!}{n_1! n_2! \ldots n_N!} p_1^{n_1} p_2^{n_2} \ldots p_N^{n_N}$$

E.g. histogram with N bins but fix: $n_{\text{tot}} = \sum_{i=1}^{N} n_i$

Log-likelihood: $\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} n_i \ln \frac{\nu_i}{n_{\text{tot}}} + C$ $(\nu_i = p_i n_{\text{tot}})$

ML estimators: $\hat{\nu}_i = n_i$ (Only $N-1$ independent; one is n_{tot} minus sum of rest.)

Goodness of fit with multinomial data (2)

The likelihood ratio statistics become:

$$t_{\nu} = -2 \sum_{i=1}^{N} n_i \ln \frac{\nu_i}{n_i} \qquad f(t_{\nu}|\boldsymbol{\nu}) \sim \chi^2_{N-1}$$

$$q_{\nu} = -2 \sum_{i=1}^{N} n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i} \qquad f(q_{\nu}|\boldsymbol{\nu}) \sim \chi^2_{N-M-1}$$

One less degree of freedom than in Poisson case because effectively only $N-1$ parameters fitted in denominator.

Estimators and g.o.f. all at once

Evaluate numerators with $\boldsymbol{\theta}$ (not its estimator):

$$\chi^2_{\mathrm{P}}(\boldsymbol{\theta}) = -2\sum_{i=1}^{N}\left[n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i} - \nu_i(\boldsymbol{\theta}) + n_i\right] \qquad \text{(Poisson)}$$

$$\chi^2_{\mathrm{M}}(\boldsymbol{\theta}) = -2\sum_{i=1}^{N} n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i} \qquad \text{(Multinomial)}$$

These are equal to the corresponding $-2 \ln L(\boldsymbol{\theta})$, so minimizing them gives the usual ML estimators for $\boldsymbol{\theta}$.

The minimized value gives the statistic $q_{\boldsymbol{\mu}}$, so we get goodness-of-fit for free.

Steve Baker and Robert D. Cousins, *Clarification of the use of the chi-square and likelihood functions in fits to histograms*, NIM **221** (1984) 437.

Using LS to combine measurements

Use LS to obtain weighted average of N measurements of λ:

$$y_i = \text{result of measurement } i, \; i = 1, \ldots, N;$$

$$\sigma_i^2 = V[y_i], \text{ assume known;}$$

$$\lambda = \text{true value (plays role of } \theta).$$

For uncorrelated y_i, minimize

$$\chi^2(\lambda) = \sum_{i=1}^{N} \frac{(y_i - \lambda)^2}{\sigma_i^2},$$

Set $\frac{\partial \chi^2}{\partial \lambda} = 0$ and solve,

$$\rightarrow \quad \hat{\lambda} = \frac{\Sigma_{i=1}^{N} y_i/\sigma_i^2}{\Sigma_{j=1}^{N} 1/\sigma_j^2} \qquad V[\hat{\lambda}] = \frac{1}{\Sigma_{i=1}^{N} 1/\sigma_i^2}$$

Combining correlated measurements with LS

If $\text{cov}[y_i, y_j] = V_{ij}$, minimize

$$\chi^2(\lambda) = \sum_{i,j=1}^{N} (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda),$$

$$\rightarrow \quad \hat{\lambda} = \sum_{i=1}^{N} w_i y_i, \quad w_i = \frac{\sum_{j=1}^{N}(V^{-1})_{ij}}{\sum_{k,l=1}^{N}(V^{-1})_{kl}}$$

$$V[\hat{\lambda}] = \sum_{i,j=1}^{N} w_i V_{ij} w_j$$

LS $\hat{\lambda}$ has zero bias, minimum variance (Gauss–Markov theorem).

Example: averaging two correlated measurements

Suppose we have y_1, y_2, and $V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

$$\rightarrow \quad \hat{\lambda} = wy_1 + (1-w)y_2, \quad w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$V[\hat{\lambda}] = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \sigma^2$$

The increase in inverse variance due to 2nd measurement is

$$\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} = \frac{1}{1-\rho^2}\left(\frac{\rho}{\sigma_1} - \frac{1}{\sigma_2}\right)^2 > 0$$

$\rightarrow$ 2nd measurement can only help.

Negative weights in LS average

If $\rho > \sigma_1/\sigma_2$, $\rightarrow w < 0$,

$\rightarrow$ weighted average is not between y_1 and y_2 (!?)

Cannot happen if correlation due to common data, but possible for shared random effect; very unreliable if e.g. ρ, σ_1, σ_2 incorrect.

See example in SDA Section 7.6.1 with two measurements at same temperature using two rulers, different thermal expansion coefficients: average is outside the two measurements; used to improve estimate of temperature.

G. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.

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")
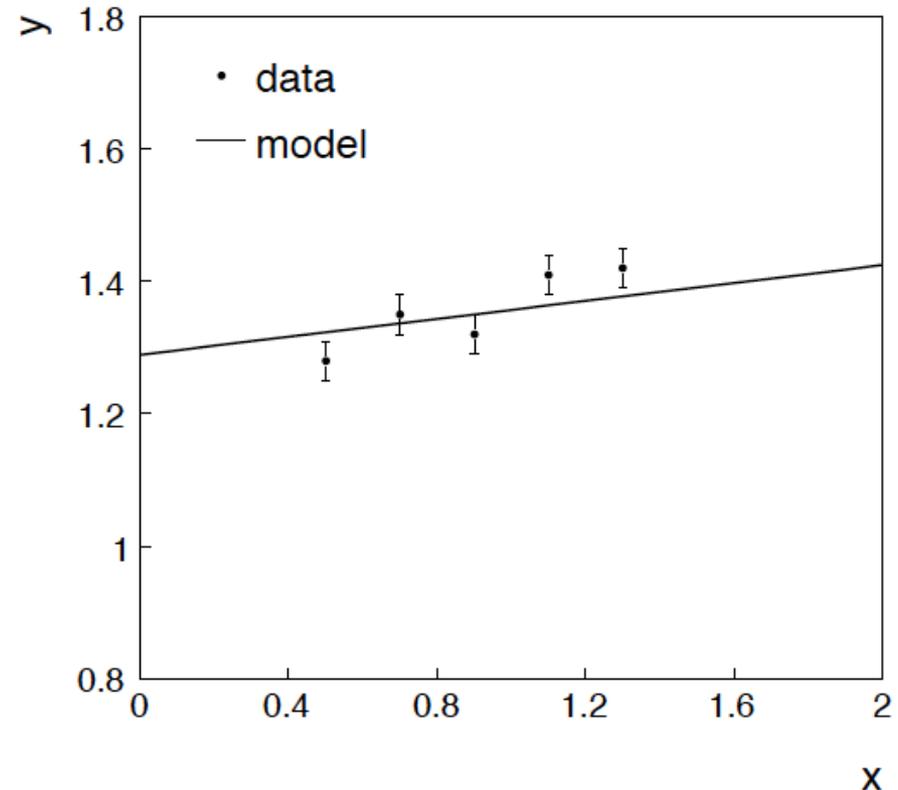
Maximum likelihood fit with Gaussian data

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

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

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

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 $\sigma_{\hat{\theta}_0}$.
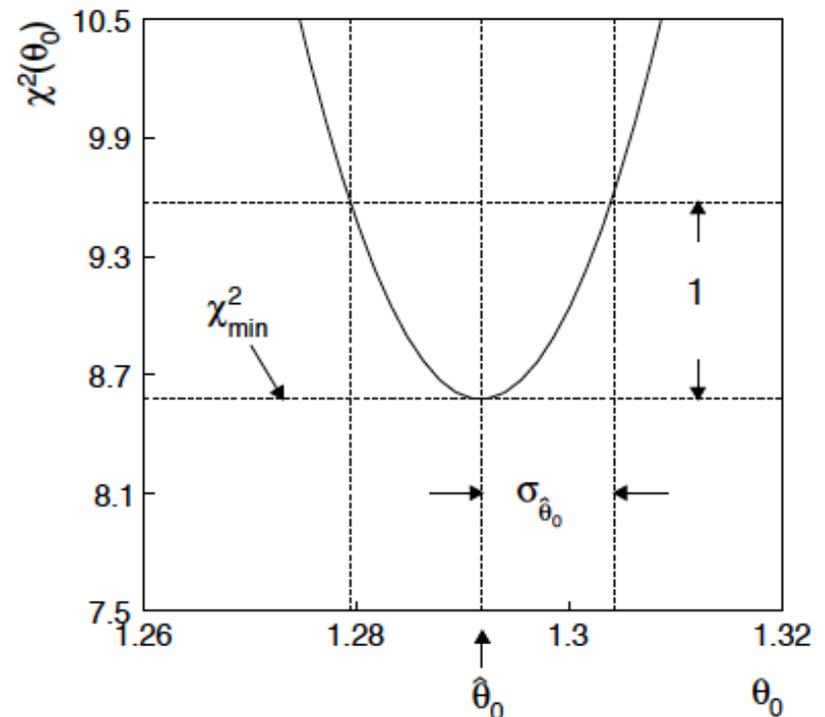
ML (or LS) fit of θ_0 and θ_1

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

Standard deviations from tangent lines to contour

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

Correlation between $\hat{\theta}_0, \ \hat{\theta}_1$ causes errors to increase.
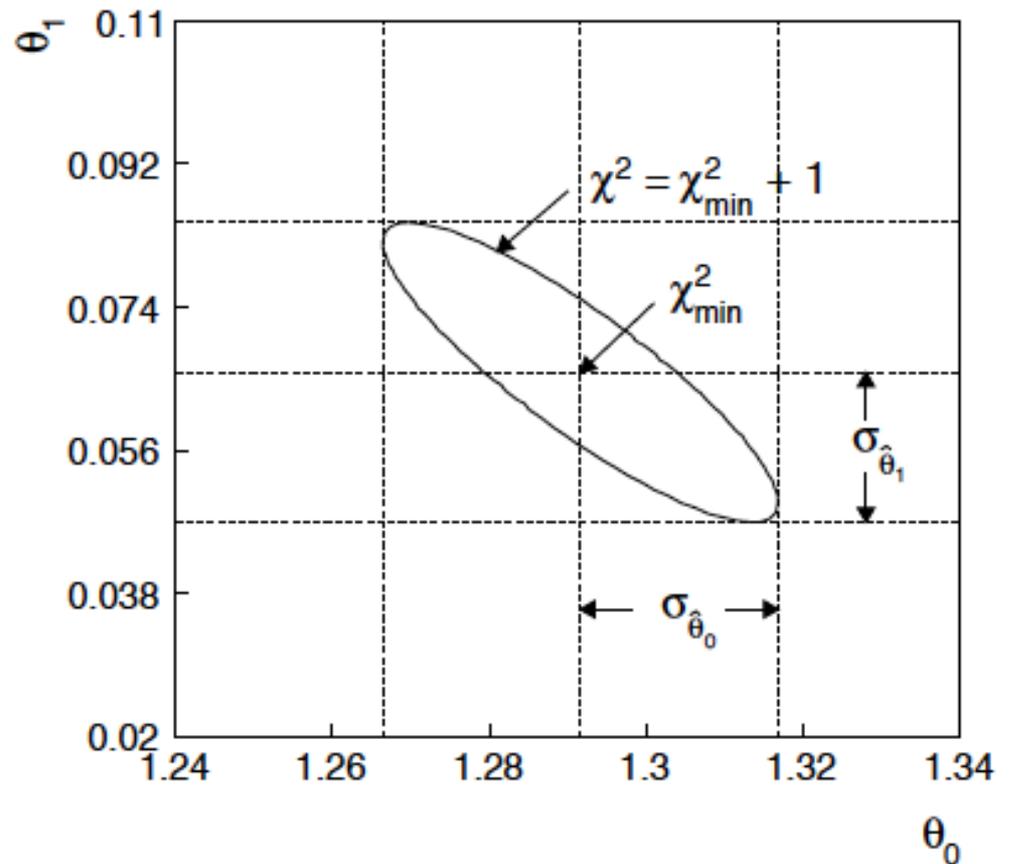
If we have a measurement $t_1 \sim$ Gauss (θ_1, σ_{t_1})

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

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

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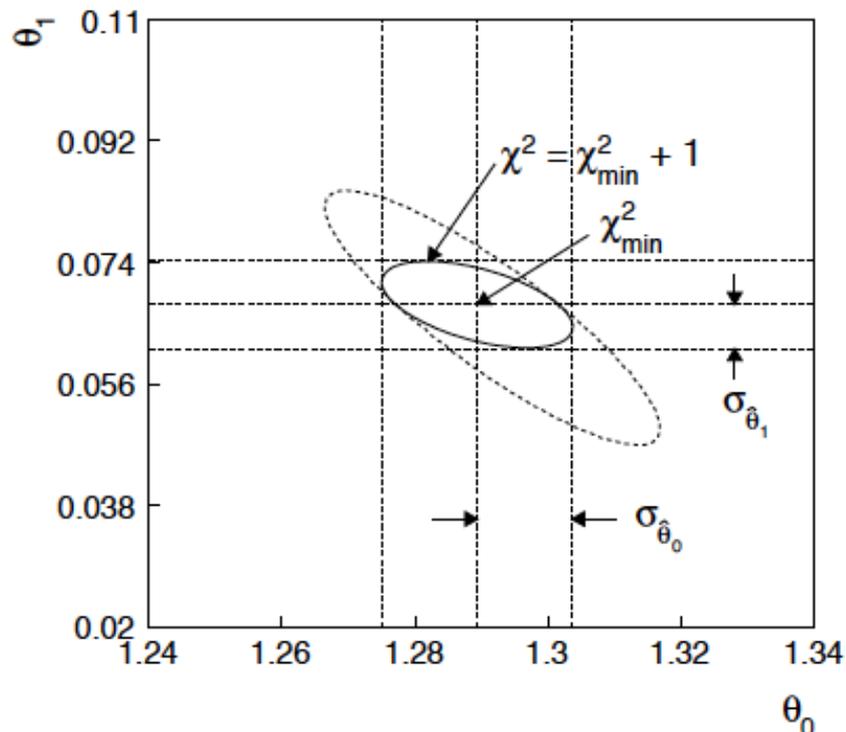
The Bayesian approach

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 function $L(x|\theta)$.

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 method

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

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

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

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

$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_{t_1}}e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

$\leftarrow$ based on previous measurement

Putting this 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

Bayesian method (continued)

We then integrate (marginalize) $p(\theta_0, \theta_1 \mid x)$ to find $p(\theta_0 \mid x)$:

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

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

$$p(\theta_0|x) \;=\; \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0-\hat{\theta}_0)^2/2\sigma_{\theta_0}^2} \quad \text{with}$$

$$\hat{\theta}_0 \;=\; \text{same as ML estimator}$$

$$\sigma_{\theta_0} \;=\; \sigma_{\hat{\theta}_0} \; \text{(same as before)}$$

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

Digression: 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 $\vec{\theta}$, look, e.g., only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \dots$

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

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 $u \sim \mathsf{Uniform}[0, 1]$

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).

For our purposes this correlation is not fatal, but statistical errors larger than if points were independent.

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

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

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.

If proposed step rejected, hop in place.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:
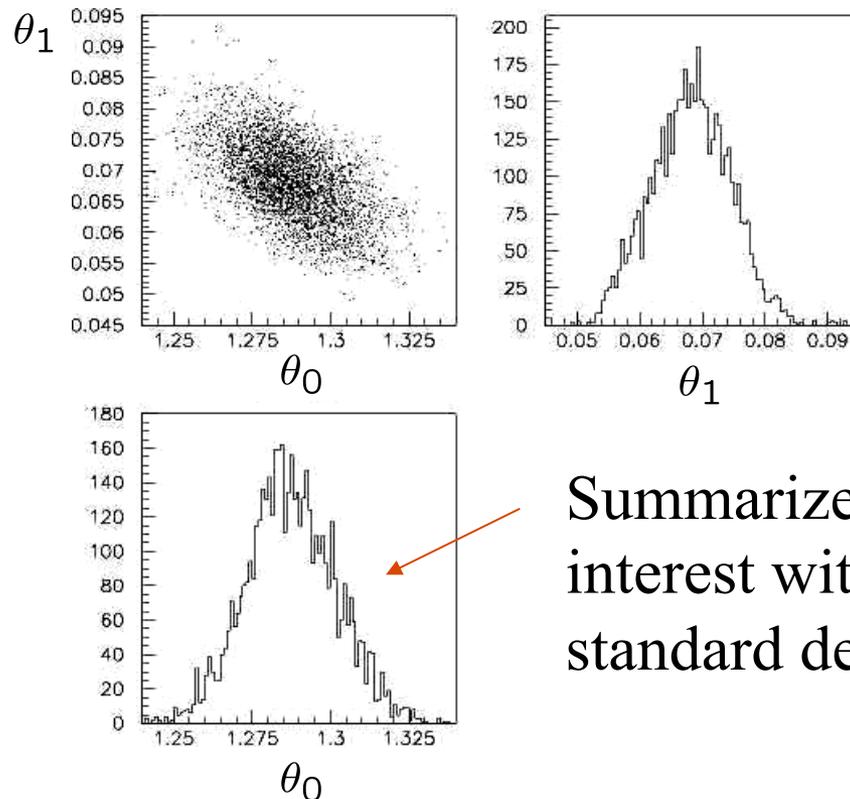
Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, e.g., a theorist 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:
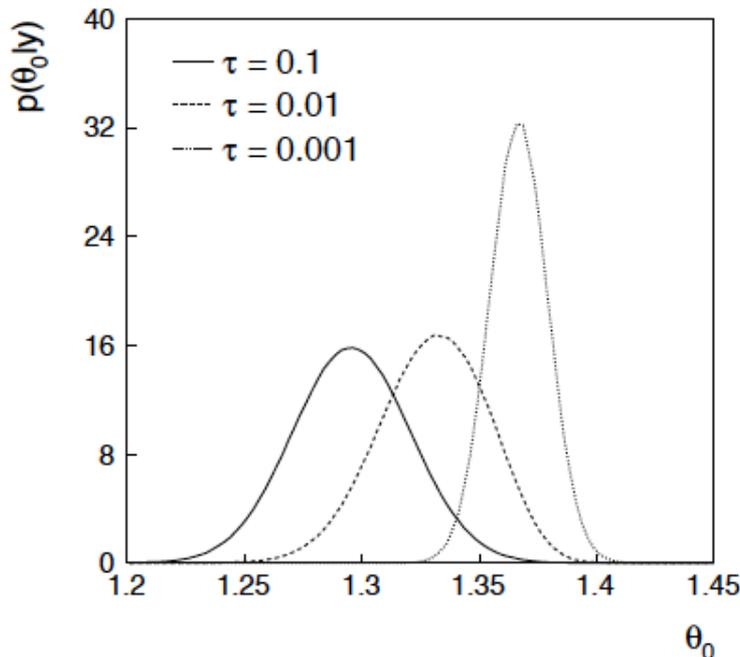
This summarizes all knowledge about θ_0.

Look also at result from variety of priors.

A more general fit (symbolic)

Given measurements: $\qquad y_i \pm \sigma_i^{\text{stat}} \pm \sigma_i^{\text{sys}}, \quad i = 1, \ldots, n$,

and (usually) covariances: $V_{ij}^{\text{stat}}, \ V_{ij}^{\text{sys}}$.

Predicted value: $\mu(x_i; \theta)$, $\quad$ expectation value $\quad E[y_i] = \mu(x_i; \theta) + b_i$

$\qquad$ control variable $\qquad$ parameters $\qquad\qquad\qquad\qquad\qquad$ bias

Often take: $\quad V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

Minimize $\quad \chi^2(\theta) = (\vec{y} - \vec{\mu}(\theta))^T V^{-1} (\vec{y} - \vec{\mu}(\theta))$

Equivalent to maximizing $L(\theta) \sim e^{-\chi^2/2}$, i.e., least squares same as maximum likelihood using a Gaussian likelihood function.

Its Bayesian equivalent

Take
$$L(\vec{y}|\vec{\theta},\vec{b}) \sim \exp\left[-\frac{1}{2}(\vec{y} - \vec{\mu}(\theta) - \vec{b})^T V_{\text{stat}}^{-1}(\vec{y} - \vec{\mu}(\theta) - \vec{b})\right]$$

$$\pi_b(\vec{b}) \sim \exp\left[-\frac{1}{2}\vec{b}^T V_{\text{sys}}^{-1} \vec{b}\right]$$

$$\pi_\theta(\theta) \sim \text{const.}$$

Joint probability for all parameters

and use Bayes' theorem:
$$p(\theta,\vec{b}|\vec{y}) \propto L(\vec{y}|\theta,\vec{b})\pi_\theta(\theta)\pi_b(\vec{b})$$

To get desired probability for θ, integrate (marginalize) over b:

$$p(\theta|\vec{y}) = \int p(\theta,\vec{b}|\vec{y})\, d\vec{b}$$

$\rightarrow$ Posterior is Gaussian with mode same as least squares estimator, σ_θ same as from $\chi^2 = \chi^2_{\text{min}} + 1$. (Back where we started!)

The error on the error

Some systematic errors are well determined

Error from finite Monte Carlo sample

Some are less obvious

Do analysis in n 'equally valid' ways and
extract systematic error from 'spread' in results.

Some are educated guesses

Guess possible size of missing terms in perturbation series;

vary renormalization scale $(\mu/2 < Q < 2\mu \ ?)$

Can we incorporate the 'error on the error'?

(cf. G. D'Agostini 1999; Dose & von der Linden 1999)

Motivating a non-Gaussian prior $\pi_b(b)$

Suppose now the experiment is characterized by

$$y_i, \quad \sigma_i^{\text{stat}}, \quad \sigma_i^{\text{sys}}, \quad s_i, \quad i = 1, \ldots, n \,,$$

where s_i is an (unreported) factor by which the systematic error is over/under-estimated.

Assume correct error for a Gaussian $\pi_b(b)$ would be $s_i \sigma_i^{\text{sys}}$, so

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi} s_i \sigma_i^{\text{sys}}} \exp\left[-\frac{1}{2} \frac{b_i^2}{(s_i \sigma_i^{\text{sys}})^2}\right] \pi_s(s_i) \, ds_i$$

Width of $\sigma_s(s_i)$ reflects 'error on the error'.

Error-on-error function $\pi_s(s)$

A simple unimodal probability density for $0 < s < 1$ with adjustable mean and variance is the Gamma distribution:

$$\pi_s(s) = \frac{a(as)^{b-1}e^{-as}}{\Gamma(b)}$$

mean $= b/a$
variance $= b/a^2$

$\pi_s(s)$

Want e.g. expectation value of 1 and adjustable standard Deviation σ_s, i.e., $a = b = 1/\sigma_s^2$

s
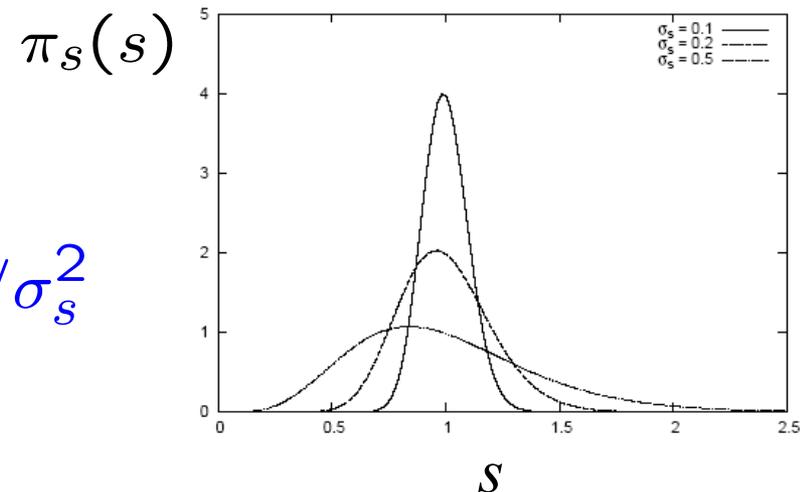
In fact if we took $\pi_s(s) \sim$ *inverse Gamma*, we could integrate $\pi_b(b)$ in closed form (cf. D'Agostini, Dose, von Linden). But Gamma seems more natural & numerical treatment not too painful.

Prior for bias $\pi_b(b)$ now has longer tails

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi}s_i\sigma_i^{\text{sys}}} \exp\left[-\frac{1}{2}\frac{b_i^2}{(s_i\sigma_i^{\text{sys}})^2}\right] \pi_s(s_i)\,ds_i$$

$\pi_b(b)$
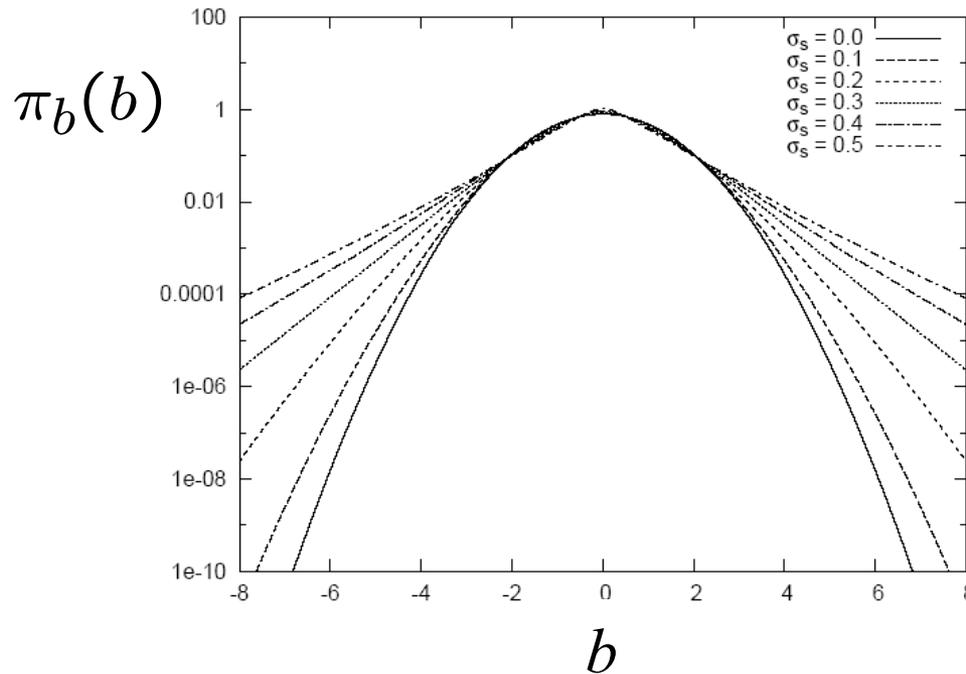
Gaussian ($\sigma_s = 0$) $P(|b| > 4\sigma_{\text{sys}}) = 6.3 \times 10^{-5}$

$\sigma_s = 0.5$ $P(|b| > 4\sigma_{\text{sys}}) = 0.65\%$

# A simple test

Suppose a fit effectively averages four measurements.

Take $\sigma_{sys} = \sigma_{stat} = 0.1$, uncorrelated.

Case #1: data appear compatible
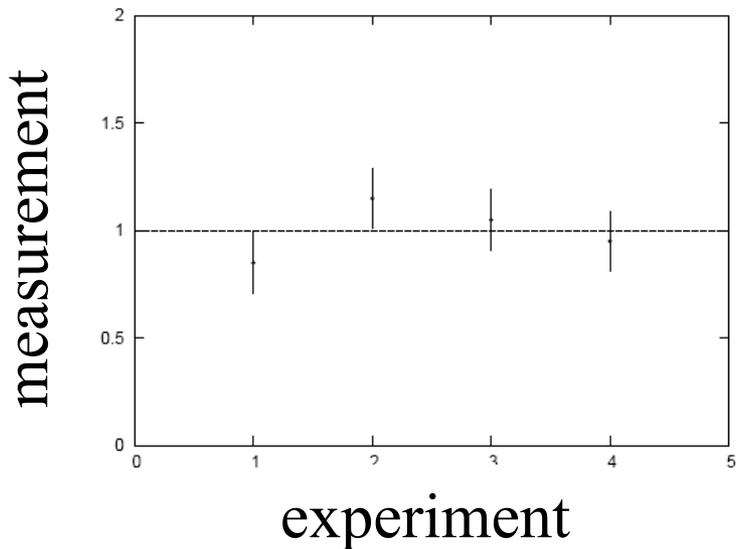
Posterior $p(\mu|y)$:
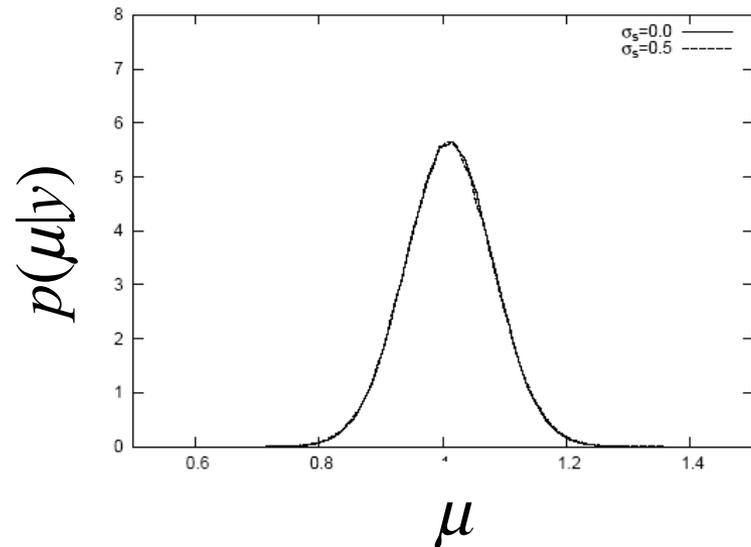
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$\sigma_s = 0.0 : \quad \hat{\mu} = 1.000 \pm 0.071$

$\sigma_s = 0.5 : \quad \hat{\mu} = 1.000 \pm 0.072$

Simple test with inconsistent data

Case #2: there is an outlier

Posterior $p(\mu|y)$:
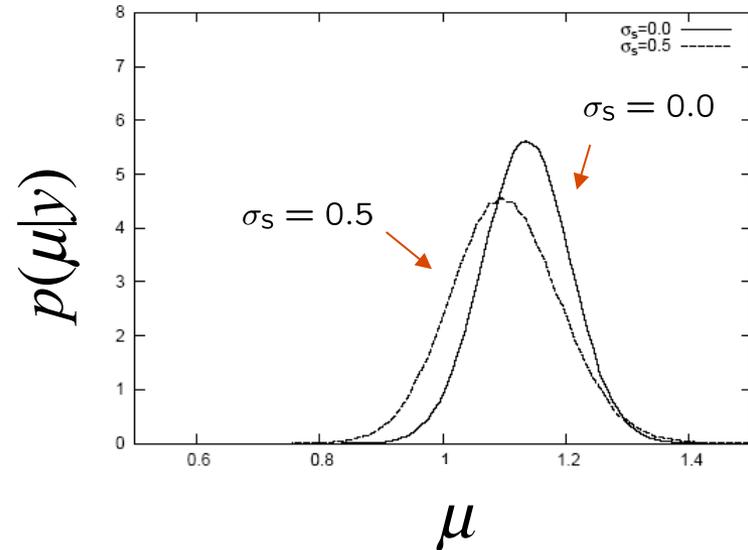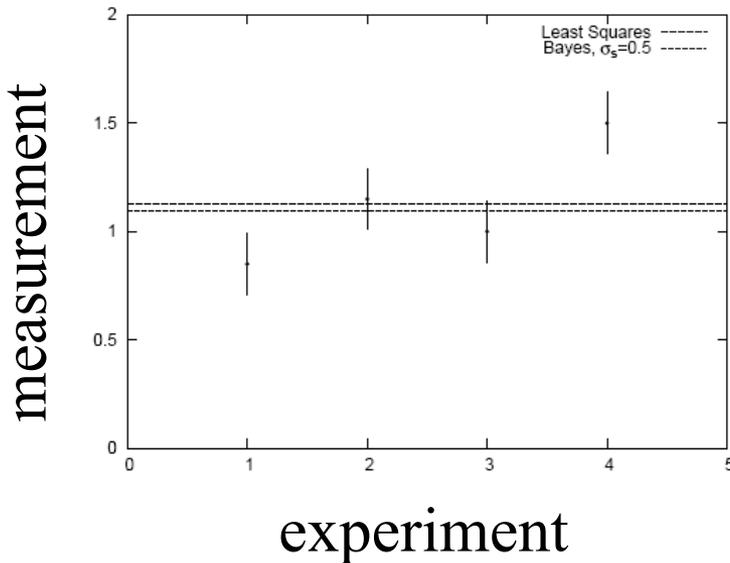
$$\sigma_S = 0.0 : \quad \hat{\mu} = 1.125 \pm 0.071$$

$$\sigma_S = 0.5 : \quad \hat{\mu} = 1.093 \pm 0.089$$

→ Bayesian fit less sensitive to outlier.

(See also D'Agostini 1999; Dose & von der Linden 1999)

Goodness-of-fit vs. size of error

In LS fit, value of minimized χ^2 does not affect size of error on fitted parameter.

In Bayesian analysis with non-Gaussian prior for systematics, a high χ^2 corresponds to a larger error (and vice versa).

posterior σ_μ
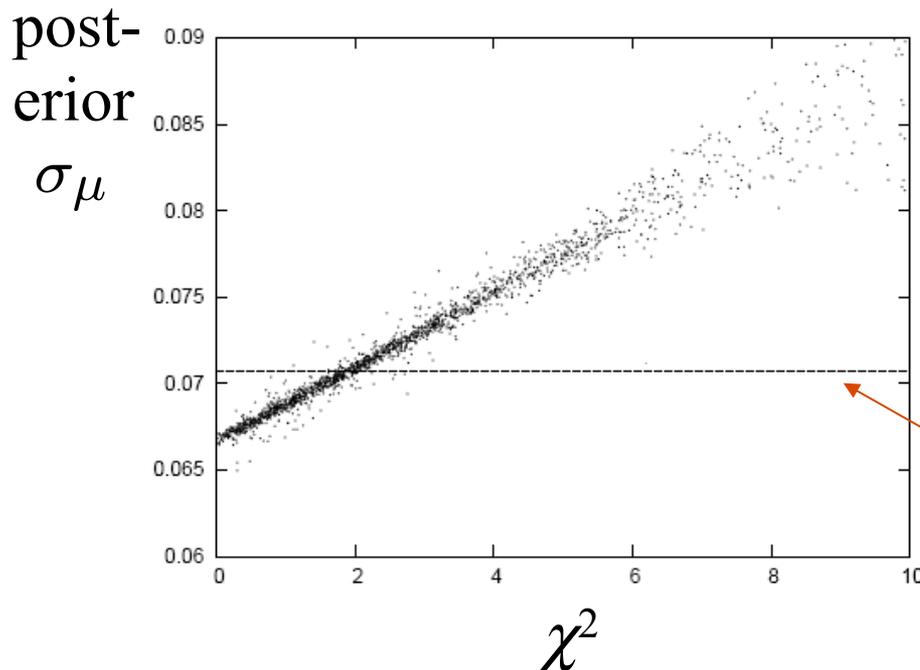
2000 repetitions of experiment, $\sigma_s = 0.5$, here no actual bias.

σ_μ from least squares

χ^2

Formulation of the unfolding problem

Consider a random variable y, goal is to determine pdf $f(y)$.

If parameterization $f(y;\theta)$ known, find e.g. ML estimators $\hat{\theta}$.

If no parameterization available, construct histogram:
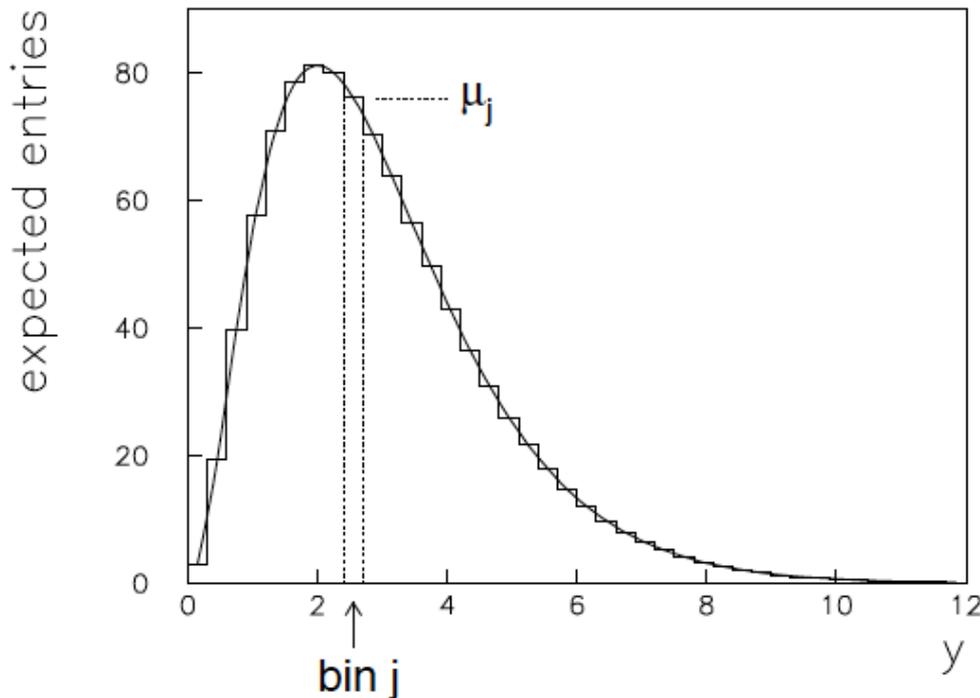
$$p_j = \int_{\text{bin } j} f(y)\, dy$$

$$\mu_j = \mu_{\text{tot}} p_j$$

"true" histogram

New goal: construct estimators for the μ_j (or p_j).

Migration

Effect of measurement errors: y = true value, x = observed value,

migration of entries between bins,

$f(y)$ is 'smeared out', peaks broadened.

$$f_{\text{meas}}(x) = \int R(x|y) f_{\text{true}}(y) \, dy$$

discretize: data are $\mathbf{n} = (n_1, \ldots, n_N)$

$$\nu_i = E[n_i] = \sum_{j=1}^{M} R_{ij} \mu_j, \quad i = 1, \ldots, N$$

response matrix

$$R_{ij} = P(\text{observed in bin } i \,|\, \text{true in bin } j)$$

Note $\boldsymbol{\mu}$, $\boldsymbol{v}$ are constants; $\boldsymbol{n}$ subject to statistical fluctuations.

Efficiency, background

Sometimes an event goes undetected:

$$\sum_{i=1}^{N} R_{ij} = \sum_{i=1}^{N} P(\text{observed in bin } i \mid \text{true value in bin } j)$$

$$= P(\text{observed anywhere} \mid \text{true value in bin } j)$$

$$= \varepsilon_j \qquad \longleftarrow \qquad \text{efficiency}$$

Sometimes an observed event is due to a background process:

$$\nu_i = \sum_{j=1}^{M} R_{ij}\mu_j + \beta_i$$

β_i = expected number of background events in *observed* histogram.

For now, assume the β_i are known.

The basic ingredients
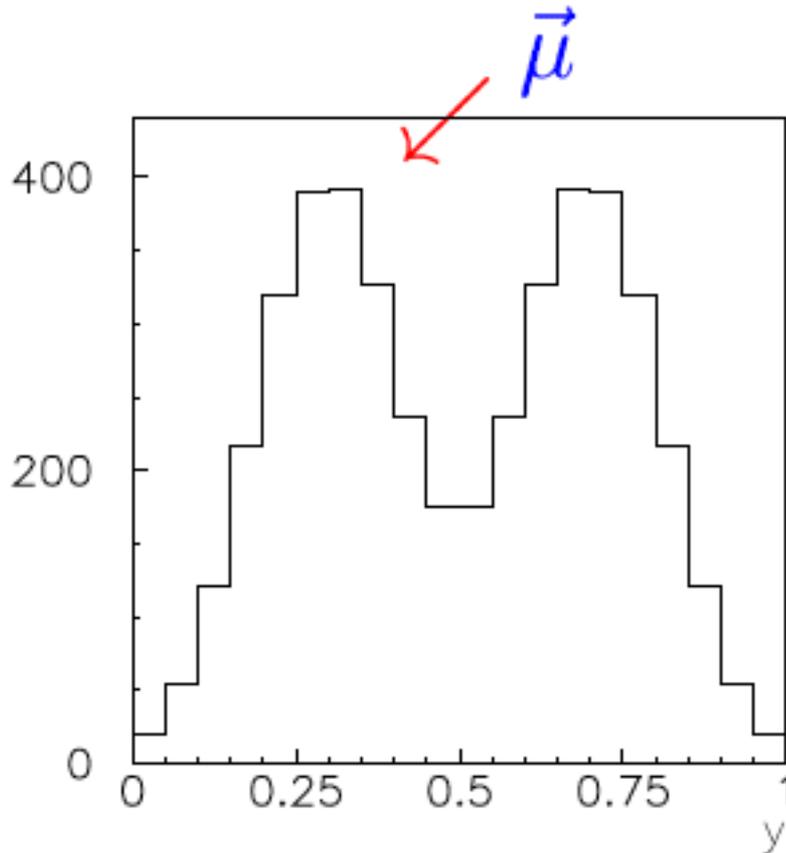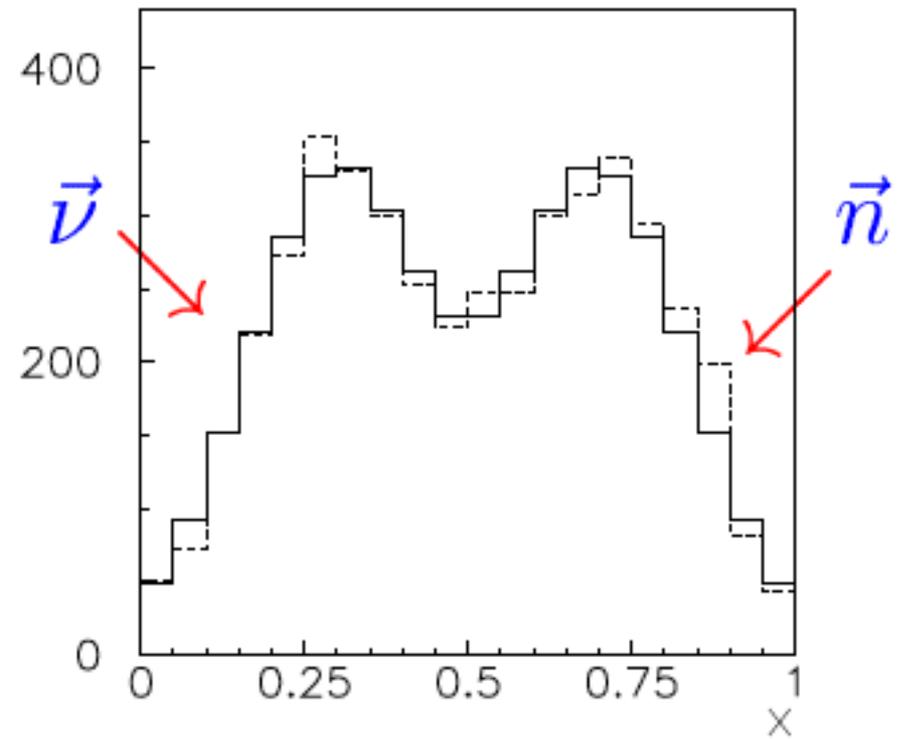
"true"   "observed"

Summary of ingredients

'true' histogram: $\quad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_M), \quad \mu_{\text{tot}} = \sum_{j=1}^{M} \mu_j$

probabilities: $\quad \mathbf{p} = (p_1, \ldots, p_M) = \boldsymbol{\mu}/\mu_{\text{tot}}$

expectation values for observed histogram: $\quad \boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$

observed histogram: $\quad \mathbf{n} = (n_1, \ldots, n_N)$

response matrix: $\quad R_{ij} = P(\text{observed in bin } i \,|\, \text{true in bin } j)$

efficiencies: $\quad \varepsilon_j = \sum_{i=1}^{N} R_{ij}$

expected background: $\quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$

These are related by: $\quad \boxed{E[\mathbf{n}] = \boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}}$

Maximum likelihood (ML) estimator from inverting the response matrix

Assume $\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$ can be inverted: $\boldsymbol{\mu} = R^{-1}(\boldsymbol{\nu} - \boldsymbol{\beta})$

Suppose data are independent Poisson: $P(n_i; \nu_i) = \dfrac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$

So the log-likelihood is $\ln L(\boldsymbol{\mu}) = \sum\limits_{i=1}^{N} (n_i \ln \nu_i - \nu_i)$

ML estimator is $\hat{\boldsymbol{\nu}} = \mathbf{n}$

$\longrightarrow \quad \hat{\boldsymbol{\mu}} = R^{-1}(\mathbf{n} - \boldsymbol{\beta})$

Example with ML solution
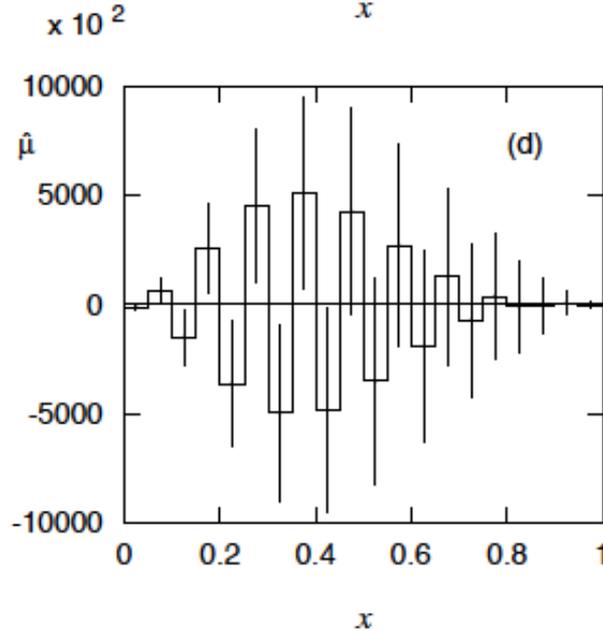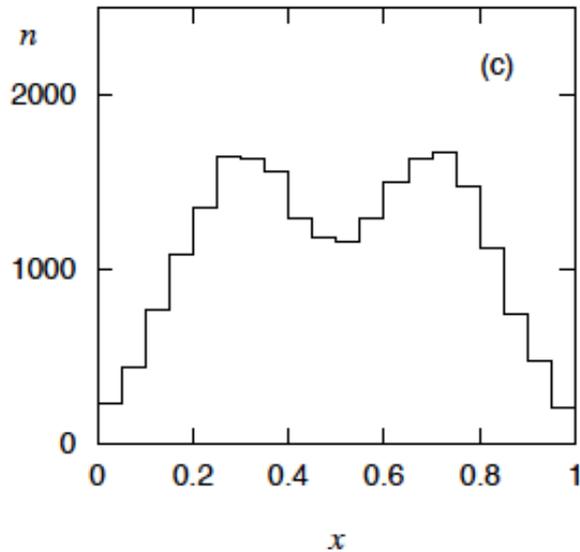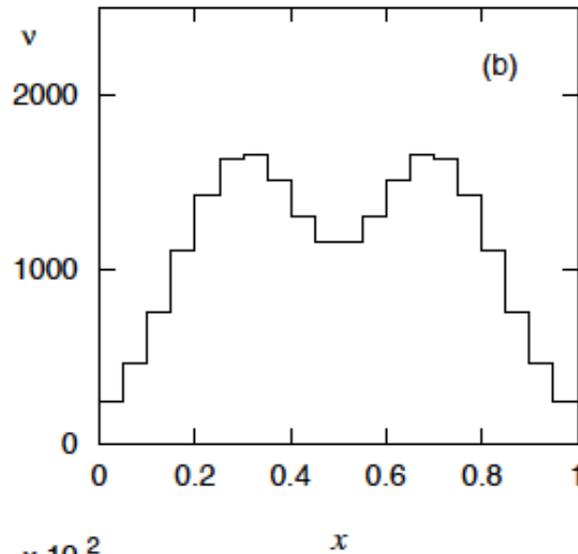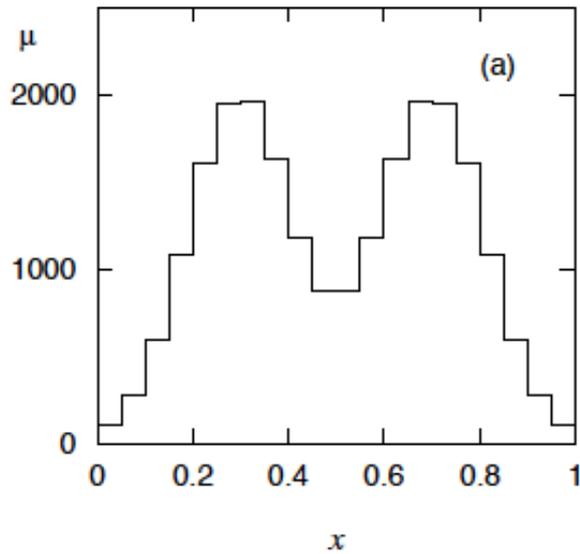
Catastrophic failure???

What went wrong?
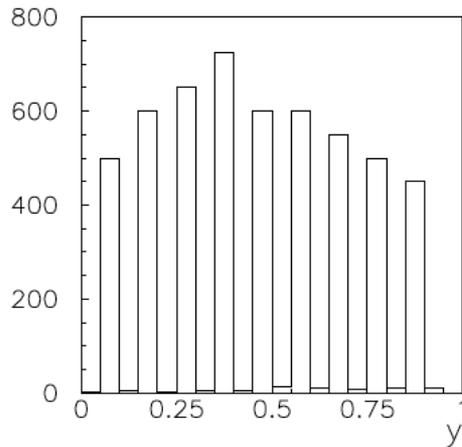
Suppose μ really had a lot of fine structure.

$\vec{\mu}$

Applying R washes this out, but leaves a residual structure:

$\vec{\nu} = R\vec{\mu}$
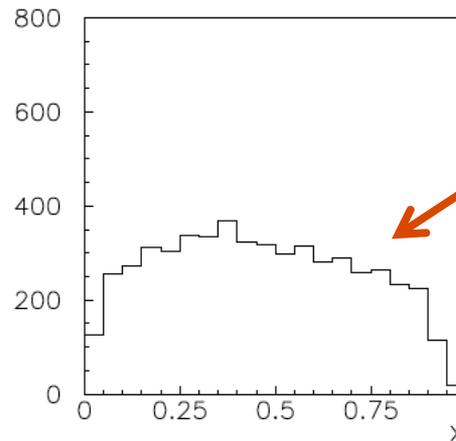
Applying R^{-1} to $\vec{\nu}$ puts the fine structure back: $\vec{\mu} = R^{-1}\vec{\nu}$.

But we don't have v, only n. R^{-1} "thinks" fluctuations in n are the residual of original fine structure, puts this back into $\hat{\mu}$.

ML solution revisited

For Poisson data the ML estimators are unbiased:

$$E[\hat{\boldsymbol{\mu}}] = R^{-1}(E[\mathbf{n}] - \boldsymbol{\beta}) = \boldsymbol{\mu}$$

Their covariance is:

$$U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j] = \sum_{k,l=1}^{N} (R^{-1})_{ik}(R^{-1})_{jl}\, \text{cov}[n_k, n_l]$$

$$= \sum_{k=1}^{N} (R^{-1})_{ik}(R^{-1})_{jk}\, \nu_k$$

(Recall these statistical errors were huge for the example shown.)

ML solution revisited (2)

The information inequality gives for unbiased estimators the minimum (co)variance bound:

$$(U^{-1})_{kl} = -E\left[\frac{\partial^2 \log L}{\partial \mu_k \, \partial \mu_l}\right] = \sum_{i=1}^{N} \frac{R_{ik}\,R_{il}}{\nu_i}$$

invert $\rightarrow$ $\quad U_{ij} = \sum_{k=1}^{N} (R^{-1})_{ik}\,(R^{-1})_{jk}\,\nu_k$

This is the same as the actual variance! I.e. ML solution gives smallest variance among all unbiased estimators, even though this variance was huge.

In unfolding one must accept some bias in exchange for a (hopefully large) reduction in variance.

Correction factor method

Use equal binning for $\vec{\mu}$, $\vec{\nu}$ and take $\hat{\mu}_i = C_i(n_i - \beta_i)$, where

$$C_i = \frac{\mu_i^{\mathrm{MC}}}{\nu_i^{\mathrm{MC}}}$$

ν_i^{MC} and μ_i^{MC} from Monte Carlo simulation (no background).

$$U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j] = C_i^2 \, \mathrm{cov}[n_i, n_j]$$

Often $C_i \sim O(1)$ so statistical errors far smaller than for ML.

But the bias $b_i = E[\hat{\mu}_i] - \mu_i$ is

$$b_i = \left(\frac{\mu_i^{\mathrm{MC}}}{\nu_i^{\mathrm{MC}}} - \frac{\mu_i}{\nu_i^{\mathrm{sig}}} \right)$$

Nonzero bias unless MC = Nature.

$$\nu_i^{\mathrm{sig}} = \nu_i - \beta_i$$

Reality check on the statistical errors

Suppose for some bin *i* we have:

$$C_i = 0.1 \qquad \beta_i = 0 \qquad n_i = 100$$

$$\longrightarrow \qquad \hat{\mu}_i = C_i n_i = 10 \qquad \sigma_{\hat{\mu}_i} = C_i \sqrt{n_i} = 1.0 \qquad \text{(10\% stat. error)}$$

But according to the estimate, only 10 of the 100 events found in the bin belong there; the rest spilled in from outside.

How can we have a 10% measurement if it is based on only 10 events that really carry information about the desired parameter?

Discussion of correction factor method

As with all unfolding methods, we get a reduction in statistical error in exchange for a bias; here the bias is difficult to quantify (difficult also for many other unfolding methods).

The bias should be small if the bin width is substantially larger than the resolution, so that there is not much bin migration.

So if other uncertainties dominate in an analysis, correction factors may provide a quick and simple solution (a "first-look").

Still the method has important flaws and it would be best to avoid it.

Regularized unfolding

Consider 'reasonable' estimators such that for some $\Delta \log L$,

$$\log L(\vec{\mu}) \geq \log L_{\text{max}} - \Delta \log L$$

Out of these estimators, choose the 'smoothest', by maximizing

$$\Phi(\vec{\mu}) = \alpha \, \log L(\vec{\mu}) + S(\vec{\mu}),$$

$S(\vec{\mu}) =$ regularization function (measure of smoothness),

$\alpha =$ regularization parameter (choose to give desired $\Delta \log L$)

Regularized unfolding (2)

In addition require $\sum\limits_{i=1}^{N} \nu_i = \sum\limits_{i,j} R_{ij}\mu_j = n_{\text{tot}}$, i.e. maximize

$$\varphi(\vec{\mu}, \lambda) = \alpha \log L(\vec{\mu}) + S(\vec{\mu}) + \lambda \left[n_{\text{tot}} - \sum_{i=1}^{N} \nu_i \right]$$

where λ is a Lagrange multiplier, $\quad \partial\varphi/\partial\lambda = 0 \rightarrow \sum\limits_{i=1}^{N} \nu_i = n_{\text{tot}}$.

$\alpha = 0$ gives smoothest solution (ignores data!),

$\alpha \rightarrow \infty$ gives ML solution (variance too large).

We need: regularization function $S(\vec{\mu})$,

        a prescription for setting α.

Tikhonov regularization

Take measure of smoothness = mean square of kth derivative,

$$S[f_{\text{true}}(y)] = - \int \left(\frac{d^k f_{\text{true}}(y)}{dy^k} \right)^2 dy \text{ , where } k = 1, 2, \dots$$

If we use Tikhonov ($k = 2$) with $\log L = -\frac{1}{2}\chi^2$,

$$S(\boldsymbol{\mu}) = - \sum_{i=1}^{M-2} \left(-\mu_i + 2\mu_{i+1} - \mu_{i+2} \right)^2$$

$$\varphi(\vec{\mu}, \lambda) = -\frac{\alpha}{2}\chi^2(\vec{\mu}) + S(\vec{\mu}) \quad \text{quadratic in } \mu_i,$$

$\rightarrow$ setting derivatives of φ equal to zero gives linear equations.

Solution using Singular Value Decomposition (SVD).

SVD implementation of Tikhonov unfolding

A. Hoecker, V. Kartvelishvili, NIM A372 (1996) 469; (TSVDUnfold in ROOT).

Minimizes $$\chi^2(\boldsymbol{\mu}) + \tau \sum_i \left[(\mu_{i+1} - \mu_i) - (\mu_i - \mu_{i-1})^2 \right]$$

Numerical implementation using Singular Value Decomposition.

Recommendations for setting regularization parameter τ:

Transform variables so errors ~ Gauss(0,1);
number of transformed values significantly different
from zero gives prescription for τ;
or base choice of τ on unfolding of test distributions.

SVD example

A. Höcker, V. Kartvelishvili, NIM A**372** (1996) 469.
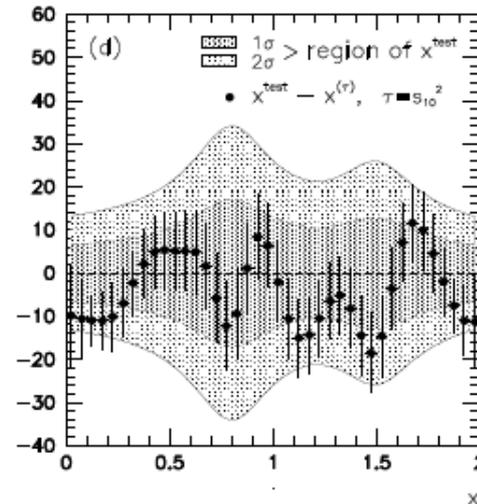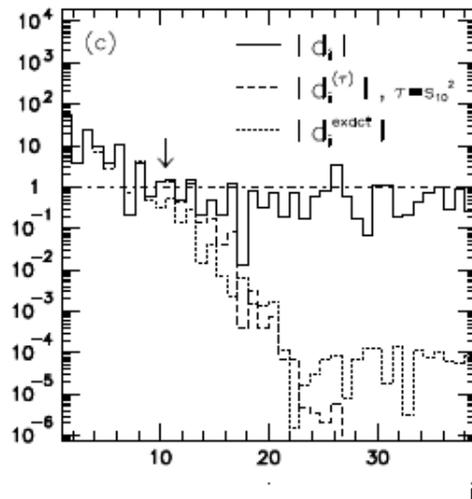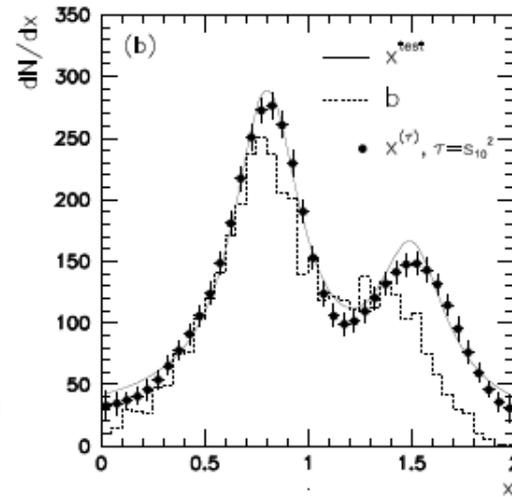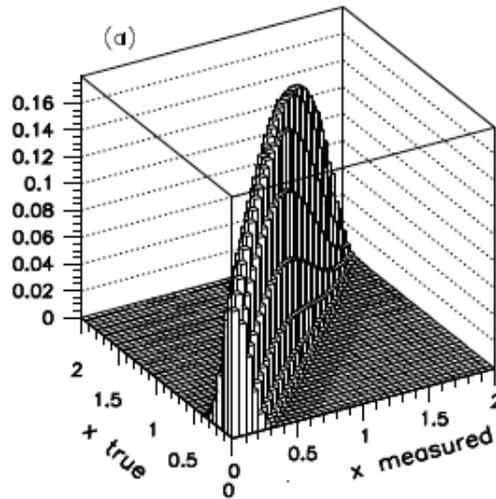
Regularization function based on entropy

Shannon entropy of a set of probabilities is

$$H = - \sum_{i=1}^{M} p_i \log p_i$$

All p_i equal $\rightarrow$ maximum entropy (maximum smoothness)

One $p_i = 1$, all others $= 0$ $\rightarrow$ minimum entropy

Use entropy as regularization function,

$$S(\vec{\mu}) = H(\vec{\mu}) = - \sum_{i=1}^{M} \frac{\mu_i}{\mu_{\text{tot}}} \log \frac{\mu_i}{\mu_{\text{tot}}}$$

$\propto \log(\text{number of ways to arrange } \mu_{\text{tot}} \text{ entries in } M \text{ bins})$

Can have Bayesian motivation: $S(\vec{\mu}) \rightarrow$ prior pdf for $\vec{\mu}$

Example of entropy-based unfolding
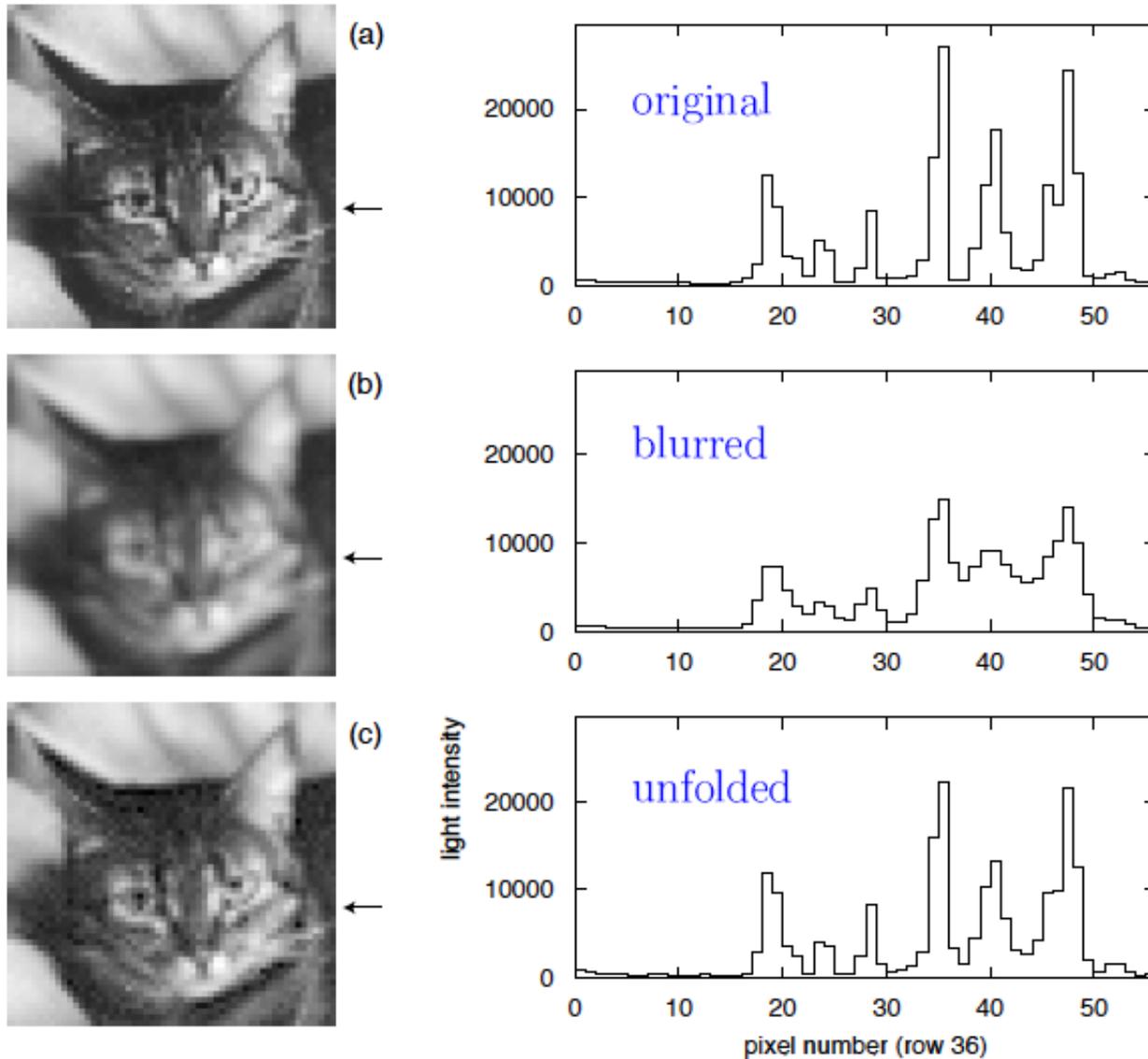
Estimating bias and variance

In general, the equations determining $\hat{\vec{\mu}}(\vec{n})$ are nonlinear.

Expand $\hat{\vec{\mu}}(\vec{n})$ about $\vec{n}_{\text{obs}}$ (observed data set),

Use error propagation to get covariance $U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j]$,

and estimators for the bias, $b_i = E[\hat{\mu}_i] - \mu_i$,

$$\hat{b}_i = \sum_{j=1}^{N} \frac{\partial \hat{\mu}_i}{\partial n_j} (\hat{\nu}_j - n_j),$$

where $\hat{\vec{\nu}} = R\hat{\vec{\mu}} + \vec{\beta}$.   (N.B. $\hat{\vec{\nu}} \neq \vec{n}$.)

Choosing the regularization parameter

$\alpha = 0 \longrightarrow \hat{\vec{\mu}}$ maximally smooth (ignores data).

$\alpha \longrightarrow \infty \longrightarrow$ ML solution (no bias, very large variance).

Possible criteria for best trade-off between bias and variance:

Minimize mean squared error,

$$\mathrm{MSE} = \frac{1}{M} \sum_{i=1}^{M} \left(U_{ii} + \hat{b}_i^2 \right), \quad \text{or}$$

$$\mathrm{MSE}' = \frac{1}{M} \sum_{i=1}^{M} \frac{U_{ii} + \hat{b}_i^2}{\hat{\mu}_i}.$$

Choosing the regularization parameter (2)

Or look at changes in χ^2 from unregularized (ML) solution,

$$\Delta\chi^2 = 2\Delta \log L = N$$

Or require that bias be consistent with zero to within its own error,

$$\chi_b^2 = \sum_{i=1}^{M} \frac{\hat{b}_i^2}{W_{ii}} = M \quad \text{where } W_{ij} = \text{cov}[\hat{b}_i, \hat{b}_j].$$

i.e. if bias significantly different from zero, we would subtract it;

$\rightarrow$ equivalent to going to smaller $\Delta \log L$ or larger α (less bias).

Some examples with Tikhonov regularization
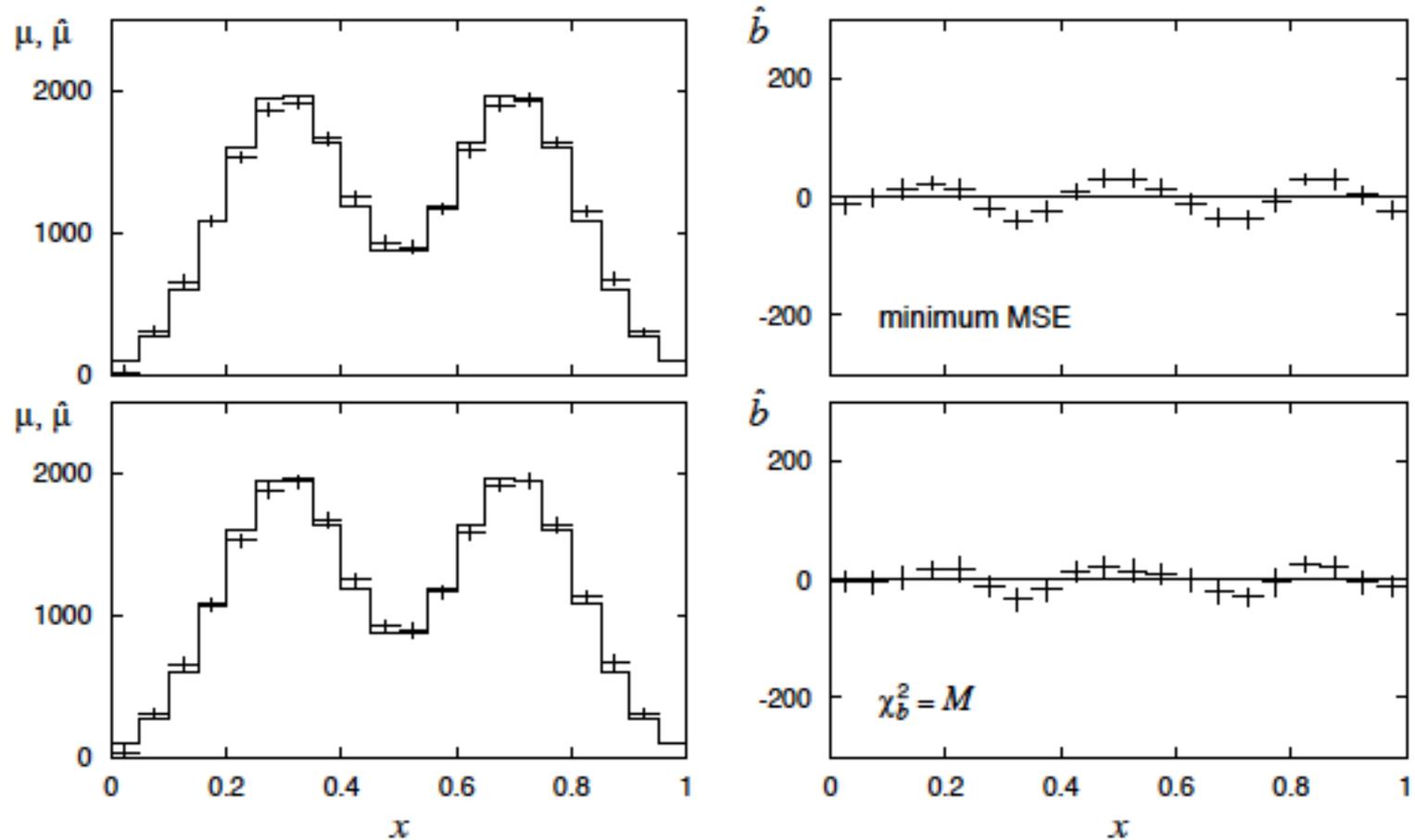
Some examples with entropy regularization
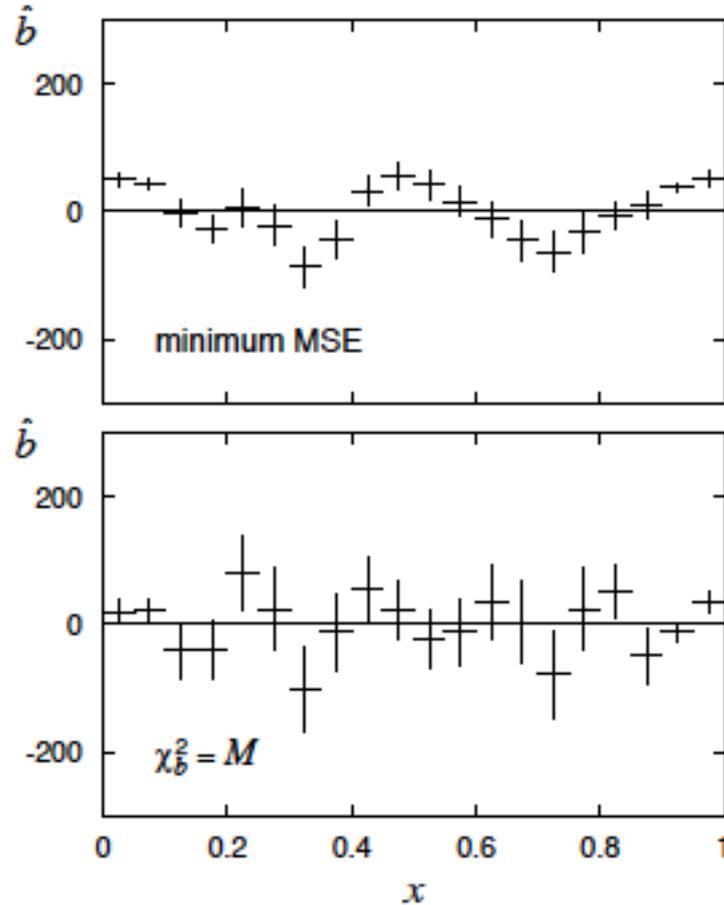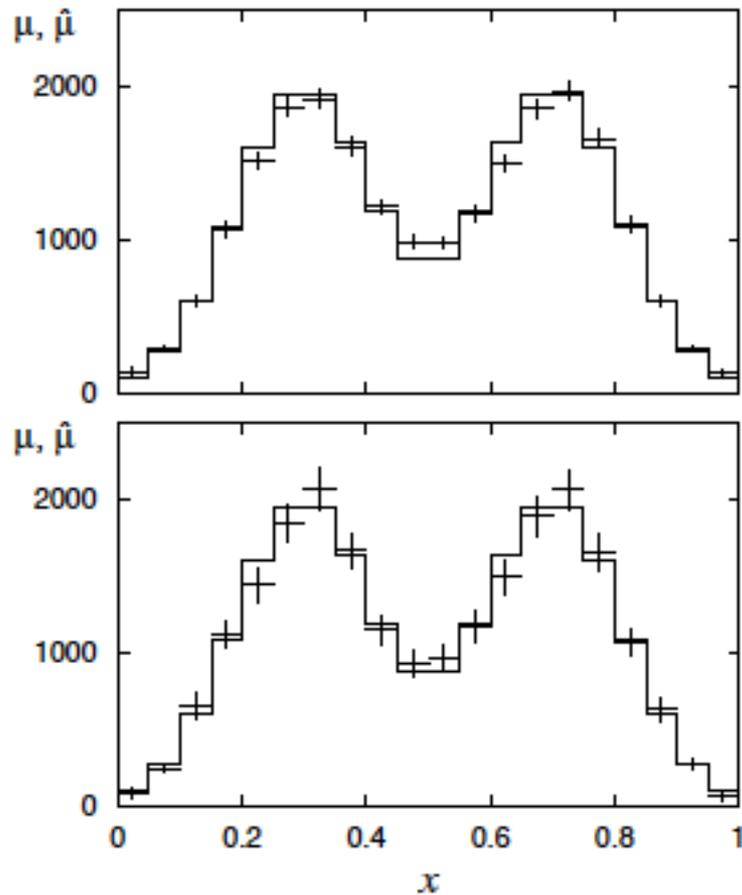
Stat. and sys. errors of unfolded solution

In general the statistical covariance matrix of the unfolded estimators is not diagonal; need to report full

$$U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j]$$

But unfolding necessarily introduces biases as well, corresponding to a systematic uncertainty (also correlated between bins).

This is more difficult to estimate. Suppose, nevertheless, we manage to report both U_{stat} and U_{sys}.

To test a new theory depending on parameters $\boldsymbol{\theta}$, use e.g.

$$\chi^2(\boldsymbol{\theta}) = (\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}})^T (U_{\text{stat}} + U_{\text{sys}})^{-1} (\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}})$$

Mixes frequentist and Bayesian elements; interpretation of result can be problematic, especially if U_{sys} itself has large uncertainty.

Folding

Suppose a theory predicts $f(y) \rightarrow \boldsymbol{\mu}$ (may depend on parameters $\boldsymbol{\theta}$).

Given the response matrix R and expected background $\boldsymbol{\beta}$, this predicts the expected numbers of observed events:

$$\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$$

From this we can get the likelihood, e.g., for Poisson data,

$$L(\mathbf{n}|\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

And using this we can fit parameters and/or test, e.g., using the likelihood ratio statistic

$$q = -2\ln\frac{L(\mathbf{n}|\boldsymbol{\nu})}{L(\mathbf{n}|\hat{\boldsymbol{\nu}})} \sim \chi_N^2$$

Versus unfolding

If we have an unfolded spectrum and full statistical and systematic covariance matrices, to compare this to a model $\boldsymbol{\mu}$ compute likelihood

$$L(\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}) \sim e^{-\chi^2/2}$$

where

$$\chi^2 = (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^T (U_{\text{stat}} + U_{\text{sys}})^{-1}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})$$

Complications because one needs estimate of systematic bias U_{sys}.

If we find a gain in sensitivity from the test using the unfolded distribution, e.g., through a decrease in statistical errors, then we are exploiting information inserted via the regularization (e.g., imposed smoothness).

ML solution again

From the standpoint of testing a theory or estimating its parameters, the ML solution, despite catastrophically large errors, is equivalent to using the uncorrected data (same information content).

There is no bias (at least from unfolding), so use

$$\chi^2(\boldsymbol{\theta}) = (\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}}_{\mathrm{ML}})^T U_{\mathrm{stat}}^{-1}(\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}}_{\mathrm{ML}})$$

The estimators of $\boldsymbol{\theta}$ should have close to optimal properties: zero bias, minimum variance.

The corresponding estimators from any unfolded solution cannot in general match this.

Crucial point is to use full covariance, not just diagonal errors.

Summary/discussion

Unfolding can be a minefield and is not necessary if goal is to compare measured distribution with a model prediction.

Even comparison of uncorrected distribution with *future* theories not a problem, as long as it is reported together with the expected background and response matrix.

In practice complications because these ingredients have uncertainties, and they must be reported as well.

Unfolding useful for getting an actual estimate of the distribution we think we've measured; can e.g. compare ATLAS/CMS.

Model test using unfolded distribution should take account of the (correlated) bias introduced by the unfolding procedure.

Finally...

Estimation of parameters is usually the "easy" part of statistics:

> Frequentist: maximize the likelihood.

> Bayesian: find posterior pdf and summarize (e.g. mode).

> Standard tools for quantifying precision of estimates: Variance of estimators, confidence intervals,...

But there are many potential stumbling blocks:

> bias versus variance trade-off (how many parameters to fit?);

> goodness of fit (usually only for LS or binned data);

> choice of prior for Bayesian approach;

> unexpected behaviour in LS averages with correlations,...

We can practice this later with MINUIT.

Extra Slides

Error propagation

Suppose we measure a set of values $\vec{x} = (x_1, \ldots, x_n)$

and we have the covariances $V_{ij} = \text{cov}[x_i, x_j]$

which quantify the measurement errors in the x_i.

Now consider a function $y(\vec{x})$.

What is the variance of $y(\vec{x})$?

The hard way: use joint pdf $f(\vec{x})$ to find the pdf $g(y)$,

then from $g(y)$ find $V[y] = E[y^2] - (E[y])^2$.

Often not practical, $f(\vec{x})$ may not even be fully known.

Error propagation (2)

Suppose we had $\vec{\mu} = E[\vec{x}]$

in practice only estimates given by the measured $\vec{x}$

Expand $y(\vec{x})$ to 1st order in a Taylor series about $\vec{\mu}$

$$y(\vec{x}) \approx y(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)$$

To find $V[y]$ we need $E[y^2]$ and $E[y]$.

$$E[y(\vec{x})] \approx y(\vec{\mu}) \quad \text{since} \quad E[x_i - \mu_i] = 0$$

Error propagation (3)

$$E[y^2(\vec{x})] \approx y^2(\vec{\mu}) + 2y(\vec{\mu}) \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} E[x_i - \mu_i]$$

$$+ E\left[\left(\sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)\right)\left(\sum_{j=1}^{n} \left[\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j)\right)\right]$$

$$= y^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Putting the ingredients together gives the variance of $y(\vec{x})$

$$\sigma_y^2 \approx \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Error propagation (4)

If the x_i are uncorrelated, i.e., $V_{ij} = \sigma_i^2 \delta_{ij}$, then this becomes

$$\sigma_y^2 \approx \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}}^2 \sigma_i^2$$

Similar for a set of m functions $\vec{y}(\vec{x}) = (y_1(\vec{x}), \ldots, y_m(\vec{x}))$

$$U_{kl} = \text{cov}[y_k, y_l] \approx \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

or in matrix notation $U = AVA^T$, where

$$A_{ij} = \left[\frac{\partial y_i}{\partial x_j} \right]_{\vec{x}=\vec{\mu}}$$

Error propagation (5)

The 'error propagation' formulae tell us the covariances of a set of functions $\vec{y}(\vec{x}) = (y_1(\vec{x}), \ldots, y_m(\vec{x}))$ in terms of the covariances of the original variables.

Limitations: exact only if $\vec{y}(\vec{x})$ linear. Approximation breaks down if function nonlinear over a region comparable in size to the σ_i.
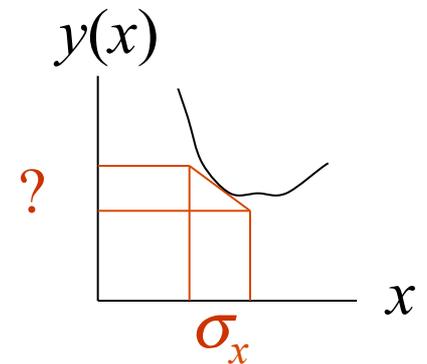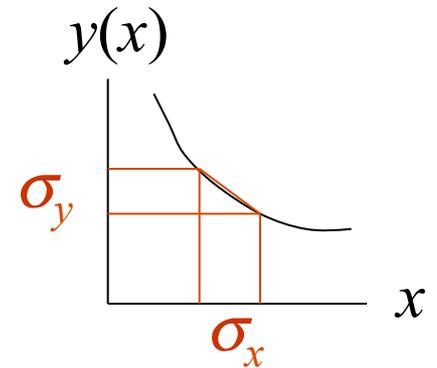
N.B. We have said nothing about the exact pdf of the x_i, e.g., it doesn't have to be Gaussian.

Error propagation − special cases

$$y = x_1 + x_2 \quad \rightarrow \quad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\text{cov}[x_1, x_2]$$

$$y = x_1 x_2 \quad \rightarrow \quad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{\text{cov}[x_1, x_2]}{x_1 x_2}$$

That is, if the x_i are uncorrelated:

add errors quadratically for the sum (or difference),

add relative errors quadratically for product (or ratio).

But correlations can change this completely...

Error propagation − special cases (2)

Consider $y = x_1 - x_2$ with

$$\mu_1 = \mu_2 = 10, \quad \sigma_1 = \sigma_2 = 1, \quad \rho = \frac{\text{cov}[x_1, x_2]}{\sigma_1 \sigma_2} = 0 \ .$$

$$V[y] = 1^2 + 1^2 = 2, \ \rightarrow \ \sigma_y = 1.4$$

Now suppose $\rho = 1$. Then

$$V[y] = 1^2 + 1^2 - 2 = 0, \ \rightarrow \ \sigma_y = 0$$

i.e. for 100% correlation, error in difference $\rightarrow 0$.

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ using the ratio

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \qquad\qquad 0 \leq \lambda(\boldsymbol{\theta}) \leq 1$$

Lower $\lambda(\boldsymbol{\theta})$ means worse agreement between data and hypothesized $\boldsymbol{\theta}$. Equivalently, usually define

$$t_{\boldsymbol{\theta}} = -2 \ln \lambda(\boldsymbol{\theta})$$

so higher $t_{\boldsymbol{\theta}}$ means worse agreement between $\boldsymbol{\theta}$ and the data.

p-value of $\boldsymbol{\theta}$ therefore

$$p_{\boldsymbol{\theta}} = \int_{t_{\boldsymbol{\theta},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta})\, dt_{\boldsymbol{\theta}}$$

need pdf

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and providing certain conditions hold...)

$$f(t_\theta|\boldsymbol{\theta}) \sim \chi_n^2$$

chi-square dist. with # d.o.f. = # of components in $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$.

Assuming this holds, the *p*-value is

$$p_{\boldsymbol{\theta}} = 1 - F_{\chi_n^2}(t_\theta)$$

To find boundary of confidence region set $p_\theta = \alpha$ and solve for t_θ:

$$t_\theta = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Recall also

$$t_\theta = -2\ln\frac{L(\theta)}{L(\hat{\theta})}$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in $\boldsymbol{\theta}$ space is where

$$\ln L(\boldsymbol{\theta}) = \ln L(\hat{\boldsymbol{\theta}}) - \tfrac{1}{2} F^{-1}_{\chi^2_n}(1 - \alpha)$$

For example, for $1 - \alpha = 68.3\%$ and $n = 1$ parameter,

$$F^{-1}_{\chi^2_1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

Example of interval from ln $L(\theta)$

For $n=1$ parameter, CL $= 0.683$, $Q_\alpha = 1$.
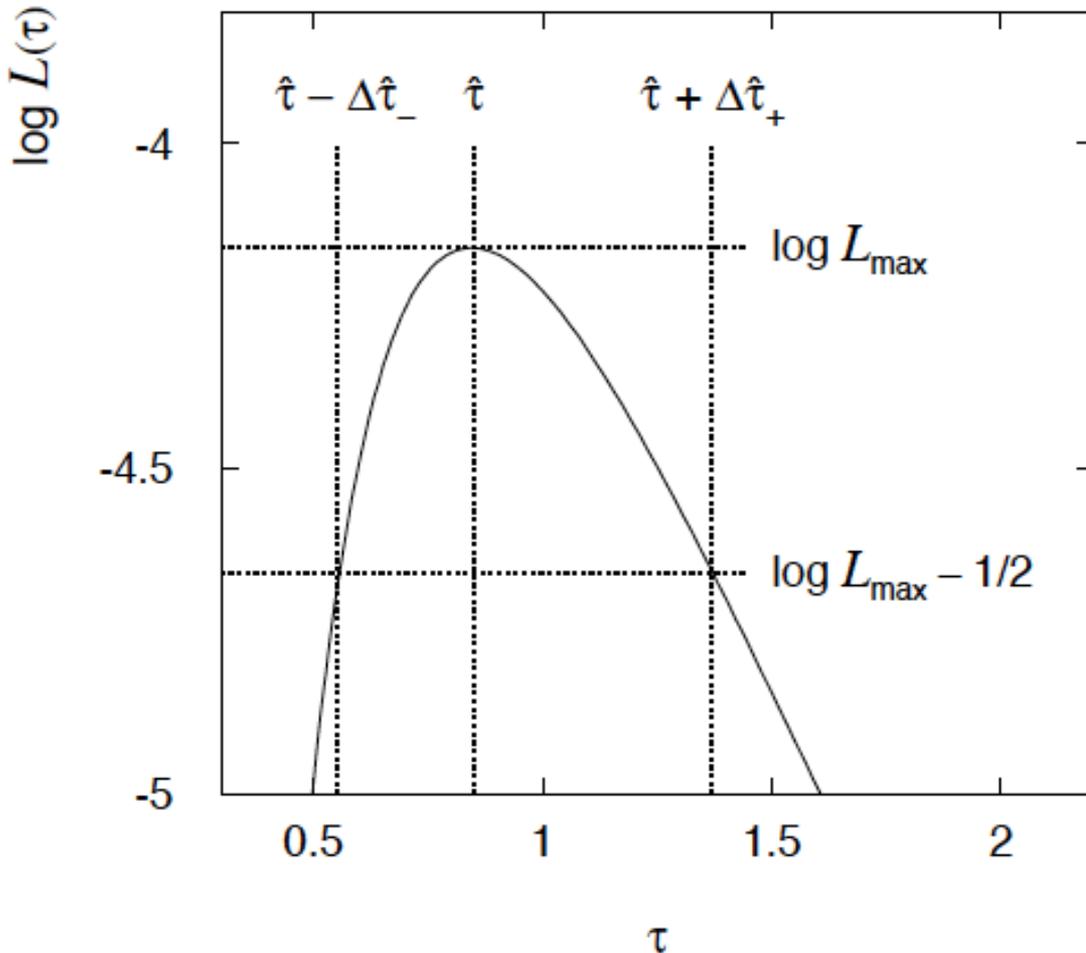
Our exponential example, now with only $n = 5$ events.

Can report ML estimate with approx. confidence interval from ln $L_{max} - 1/2$ as "asymmetric error bar":

$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

Multiparameter case

For increasing number of parameters, CL = $1 - \alpha$ decreases for confidence region determined by a given

$$Q_\alpha = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Q_α	$1 - \alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

Multiparameter case (cont.)

Equivalently, Q_α increases with *n* for a given CL $= 1 - \alpha$.

$1 - \alpha$	Q_α				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Ingredients for a test / interval

Note that these confidence intervals can be found using only the likelihood function evaluated with the observed data. This is because the statistic

$$t_\theta = -2 \ln \frac{L(\theta)}{L(\hat\theta)}$$

approaches a well-defined distribution independent of the distribution of the data in the large sample limit.

For finite samples, however, the resulting intervals are approximate.

In general to carry out a test we need to know the distribution of the test statistic $t(x)$, and this means we need the full model $P(x|\boldsymbol{\theta})$.

Covariance, correlation, etc.

For a pair of random variables x and y, the covariance and correlation are

$$\text{cov}[x, y] = E[xy] - E[x]E[y] \qquad \rho_{xy} = \frac{\text{cov}[x, y]}{\sigma_x \sigma_y}$$

One only talks about the correlation of two quantities to which one assigns probability (i.e., random variables).

So in frequentist statistics, estimators for parameters can be correlated, but not the parameters themselves.

In Bayesian statistics it does make sense to say that two parameters are correlated, e.g.,

$$\text{cov}[\theta_i, \theta_j] = \int \theta_i \theta_j p(\boldsymbol{\theta}|x) \, d\boldsymbol{\theta} - \int \theta_i p(\boldsymbol{\theta}|x) \, d\boldsymbol{\theta} \int \theta_j p(\boldsymbol{\theta}|x) \, d\boldsymbol{\theta}$$

Example of "correlated systematics"

Suppose we carry out two independent measurements of the length of an object using two rulers with diferent thermal expansion properties.

Suppose the temperature is not known exactly but must be measured (but lengths measured together so T same for both),

$$T \sim \text{Gauss}(\tau, \sigma_T)$$

The expectation value of the measured length L_i ($i = 1, 2$) is related to true length λ at a reference temperature τ_0 by

$$E[L_i] = \lambda - \alpha_i(T - \tau_0), \qquad i = 1, 2$$

and the (uncorrected) length measurements are modeled as

$$L_i \sim \text{Gauss}(\lambda - \alpha_i(\tau - \tau_0), \sigma_i)$$

Two rulers (2)

The model thus treats the measurements T, L_1, L_2 as uncorrelated with standard deviations σ_T, σ_1, σ_2, respectively:

$$L(T, L_1, L_2 | \lambda, \tau) = \frac{1}{\sqrt{2\pi}\sigma_T} e^{-(T-\tau)^2/2\sigma_T^2} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(L_i - \lambda + \alpha_i(\tau - T_0))^2/2\sigma_i^2}$$

Alternatively we could correct each raw measurement:

$$y_i = L_i + \alpha_i(T - \tau_0)$$

which introduces a correlation between y_1, y_2 and T

$$\text{cov}[y_1, y_2] = \alpha_1 \alpha_2 \sigma_T^2 \qquad\qquad \text{cov}[y_i, T] = \alpha_i \sigma_T^2$$

But the likelihood function (multivariate Gauss in T, y_1, y_2) is the same function of τ and λ as before.

Language of y_1, y_2: temperature gives correlated systematic.
Language of L_1, L_2: temperature gives "coherent" systematic.

Two rulers (3)

Outcome has some surprises:
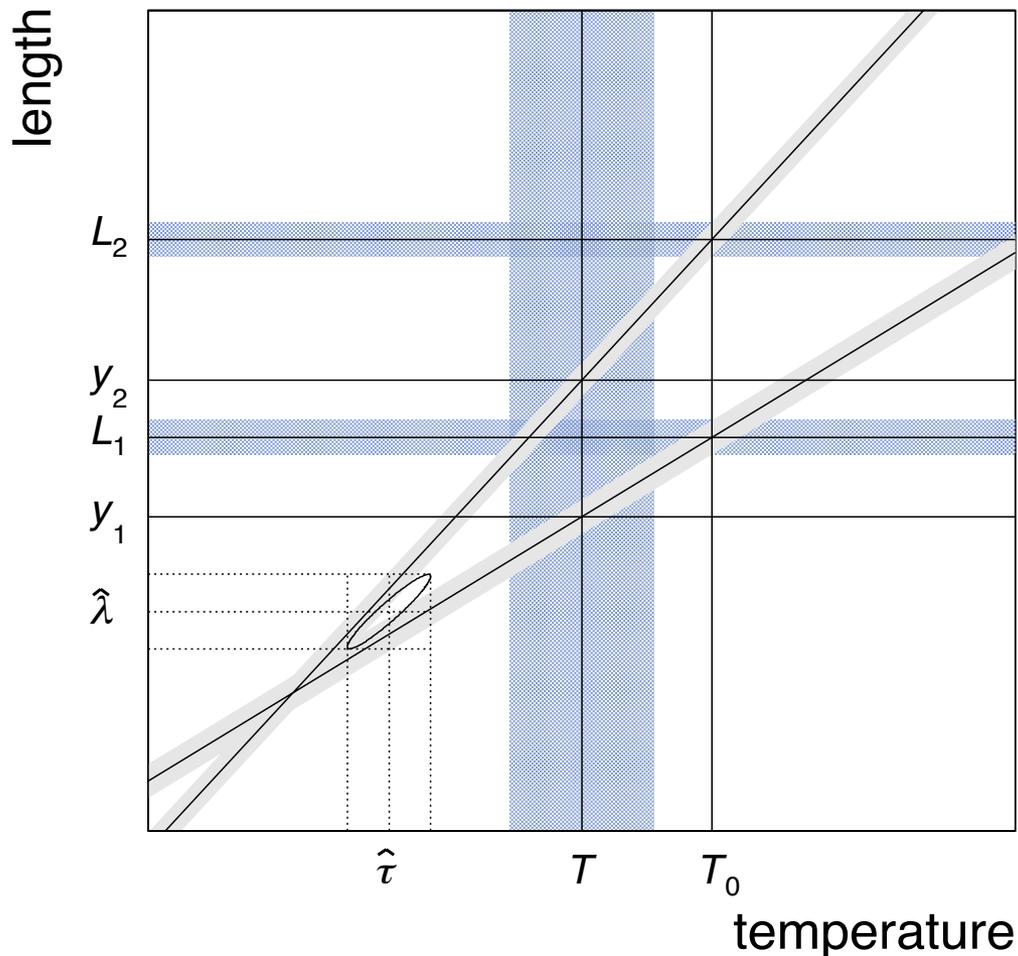
Estimate of λ does not lie between y_1 and y_2.

Stat. error on new estimate of temperature substantially smaller than initial σ_T.

These are features, not bugs, that result from our model assumptions.

Two rulers (4)

We may re-examine the assumptions of our model and conclude that, say, the parameters α_1, α_2 and τ_0 were also uncertain.

We may treat their nominal values as measurements (need a model; Gaussian?) and regard α_1, α_2 and τ_0 as as nuisance parameters.

$$L(L_1, L_2, T, \tilde{\tau}_0, \tilde{\alpha}_1, \tilde{\alpha}_2 | \lambda, \tau, \tau_0, \alpha_1, \alpha_2) =$$

$$\frac{1}{\sqrt{2\pi}\sigma_T} e^{-(T-\tau)^2/2\sigma_T^2} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(L_i - \lambda + \alpha_i(\tau-\tau_0))^2/2\sigma_i^2}$$

$$\times \frac{1}{\sqrt{2\pi}\sigma_{\tilde{\tau}_0}} e^{-(\tilde{\tau}_0 - \tau_0)^2/2\sigma_{\tilde{\tau}_0}^2} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_{\tilde{\alpha}_i}} e^{-(\tilde{\alpha}_i - \alpha_i)^2/2\sigma_{\tilde{\alpha}_i}^2}$$

Two rulers (5)

The outcome changes; some surprises may be "reduced".