Discussion on fitting and averages

including "errors on errors"

LAL Orsay, 23 March 2018

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

Outline

Review of parameter estimation (pedagogical)

Some examples of averaging measurements

Averaging with errors on errors

Distribution, likelihood, model

Suppose the outcome of a measurement is x. (e.g., a number of events, a histogram, or some larger set of numbers).

The probability density (or mass) function or 'distribution' of x, which may depend on parameters θ, is:

$P(x|\theta)$ (Independent variable is x; θ is a constant.)

If we evaluate $P(x|\theta)$ with the observed data and regard it as a function of the parameter(s), then this is the likelihood:

$L(\theta) = P(x|\theta)$ (Data x fixed; treat L as function of θ.)

We will use the term 'model' to refer to the full function $P(x|\theta)$ that contains the dependence both on x and θ.

Maximum likelihood

The most important frequentist method for constructing estimators is to take the value of the parameter(s) that maximize the likelihood:

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

The resulting estimators are functions of the data and thus characterized by a sampling distribution with a given (co)variance:

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

In general they may have a nonzero bias:

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

Under conditions usually satisfied in practice, bias of ML estimators is zero in the large sample limit, and the variance is as small as possible for unbiased estimators.

ML estimator may not in some cases be regarded as the optimal trade-off between these criteria (cf. regularized unfolding).

Example: fitting a straight line

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

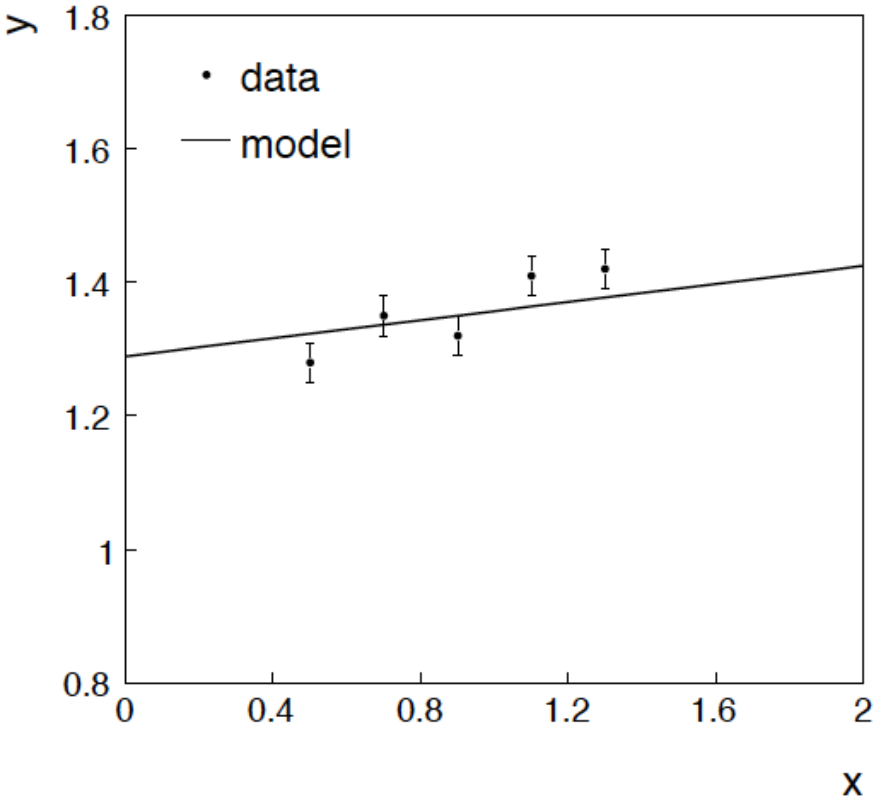
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

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

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

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

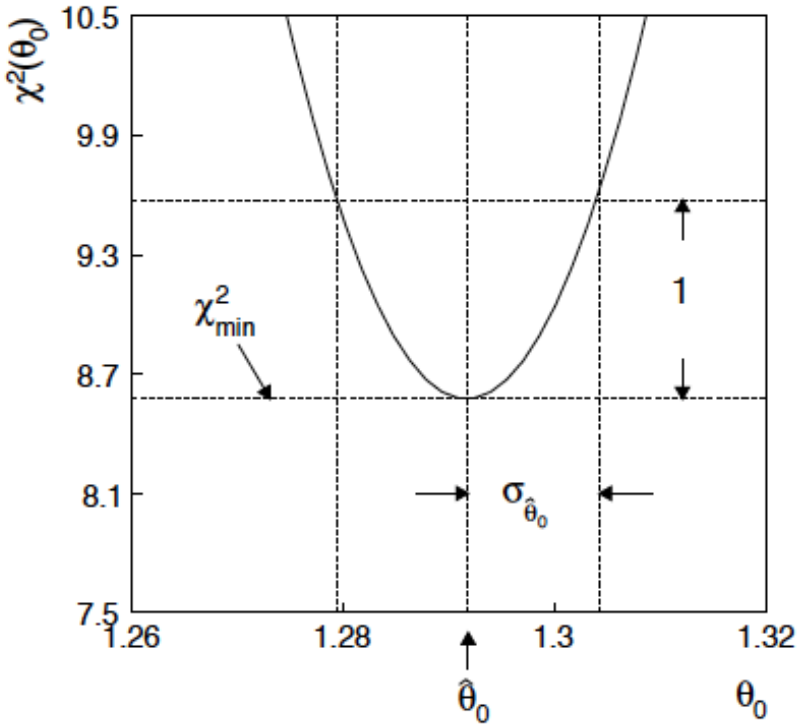
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

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



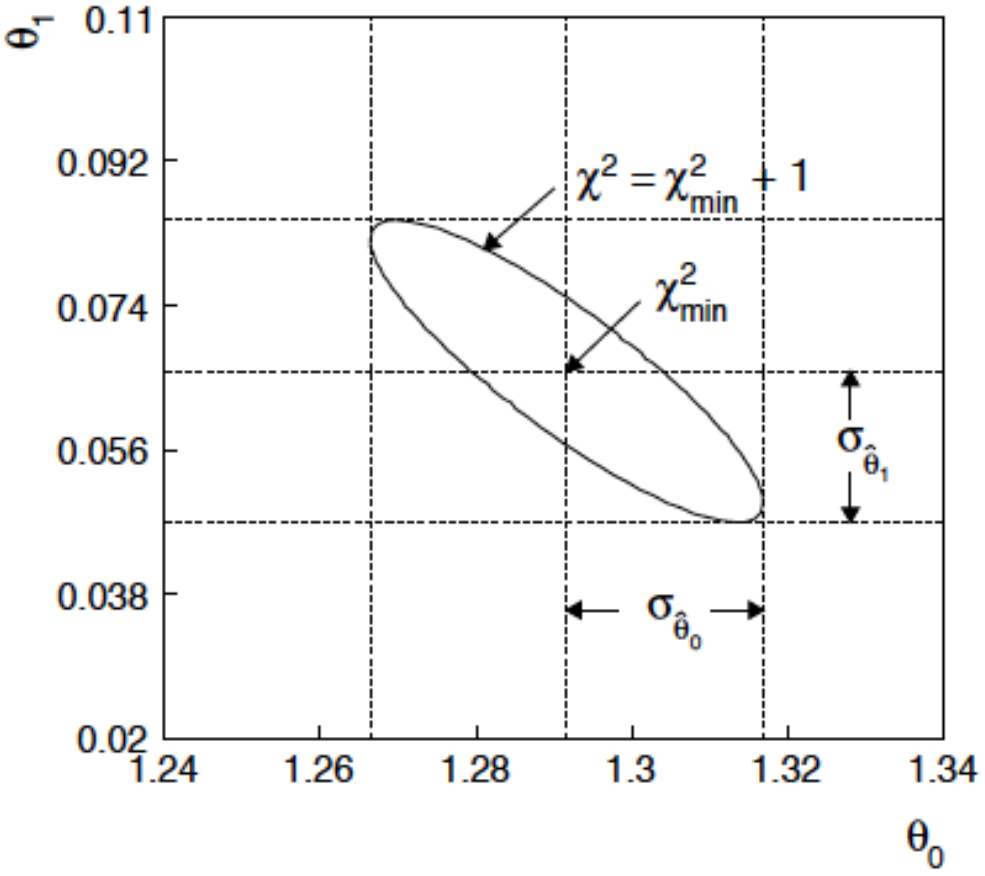
ML (or LS) fit of θ_0 and θ_1

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

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.

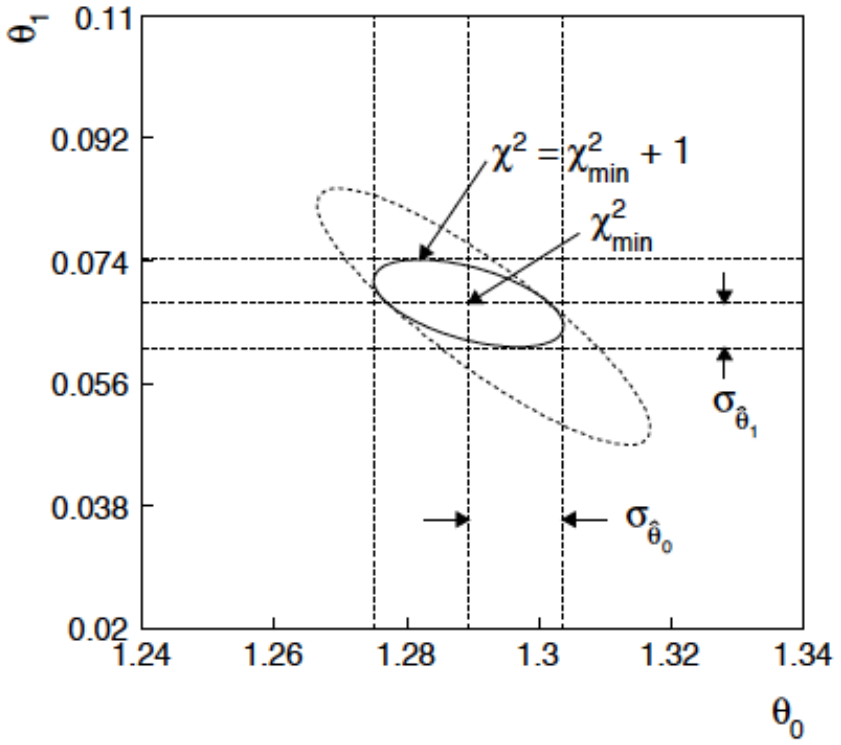


If we have a measurement $t_1 \sim \text{Gauss}\,(\theta_1, \sigma_{t_1})$

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

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

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



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, → 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 $L(\theta_0)$

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

← 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 ∝ likelihood × prior

Bayesian method (continued)

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

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

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

$$\begin{aligned} p(\theta_0|x) &= \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma^2_{\theta_0}} \quad \text{with} \\ \hat{\theta}_0 &= \text{same as ML estimator} \\ \sigma_{\theta_0} &= \sigma_{\hat{\theta}_0} \ \text{(same as before)} \end{aligned}$$

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 $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

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

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

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

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.

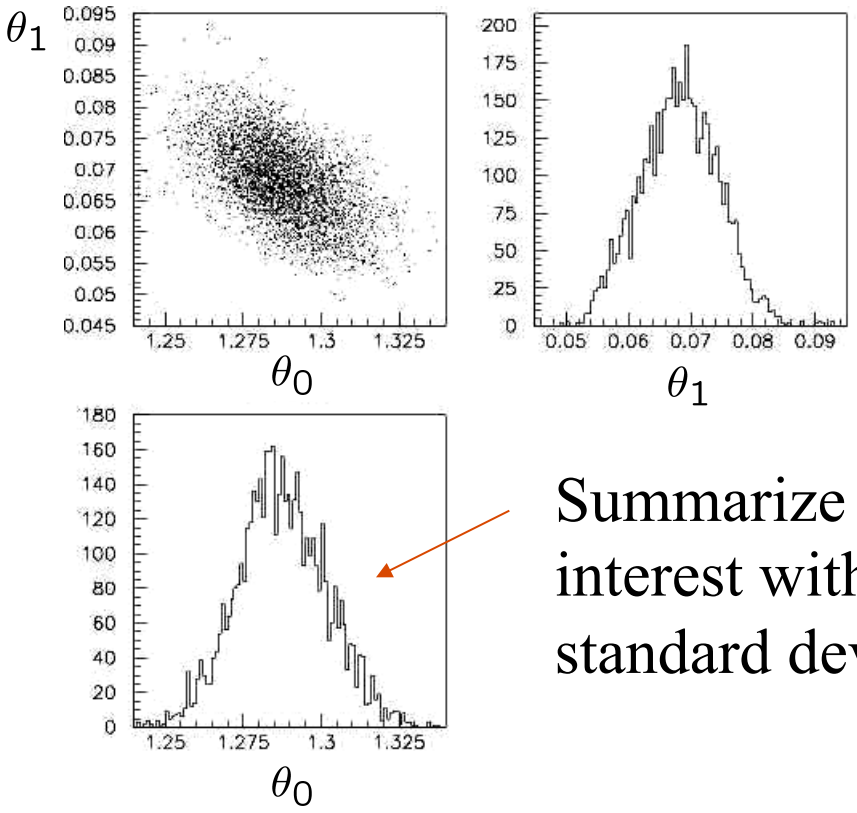
The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

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

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

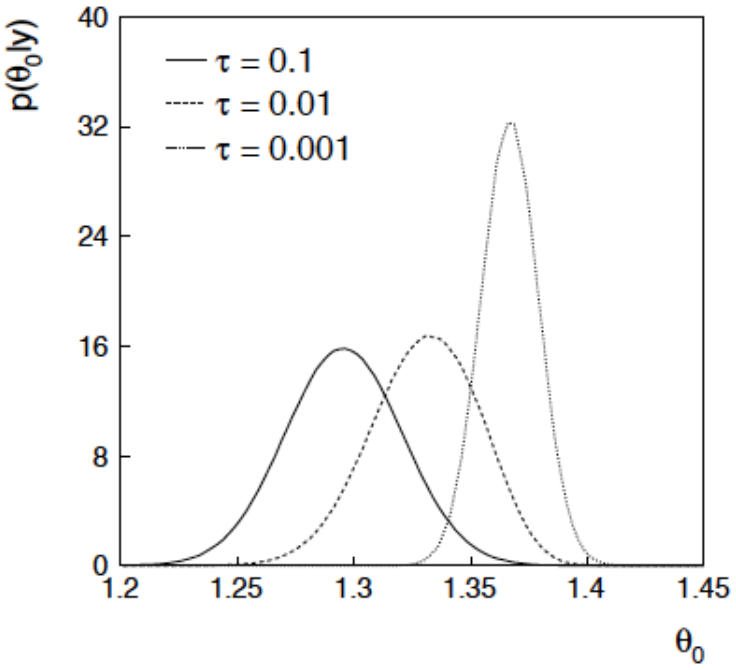
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

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

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

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

control variable parameters bias

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

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

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 $\boldsymbol{b}$:

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

→ 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'.

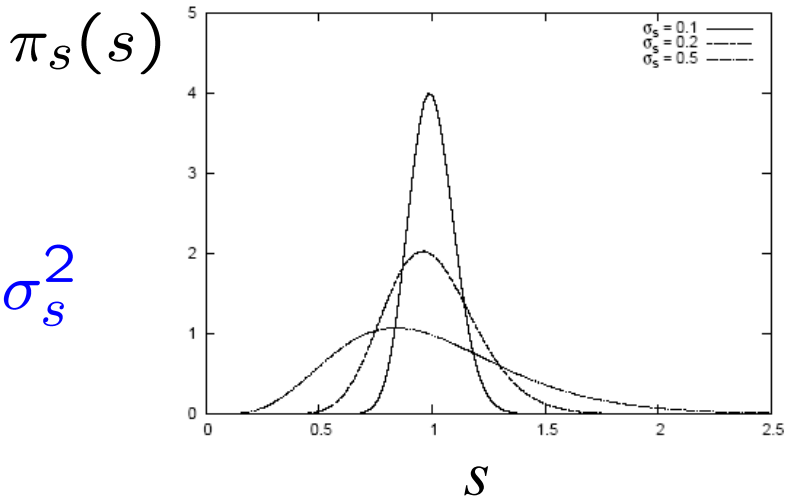
Error-on-error function $\pi_s(s)$

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

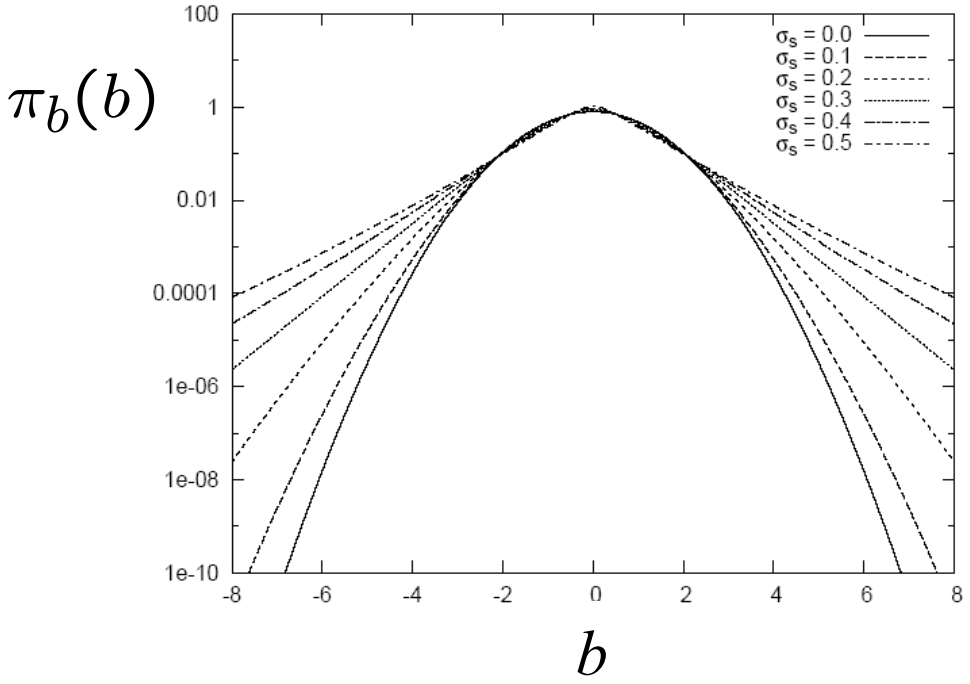
Want e.g. expectation value of 1 and adjustable standard Deviation σ_s , i.e., $a = b = 1/\sigma_s^2$



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

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



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

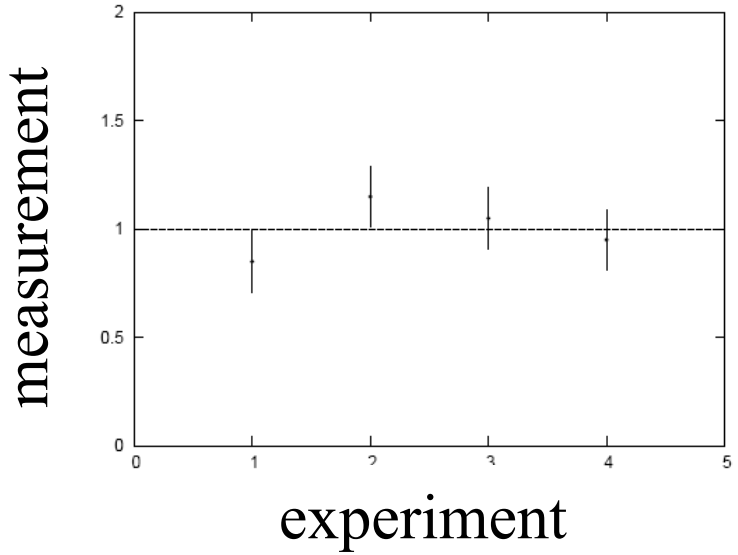
$\sigma_s = 0.5$ $\quad P(|b| > 4\sigma_{\text{sys}}) = 0.65\%$

A simple test

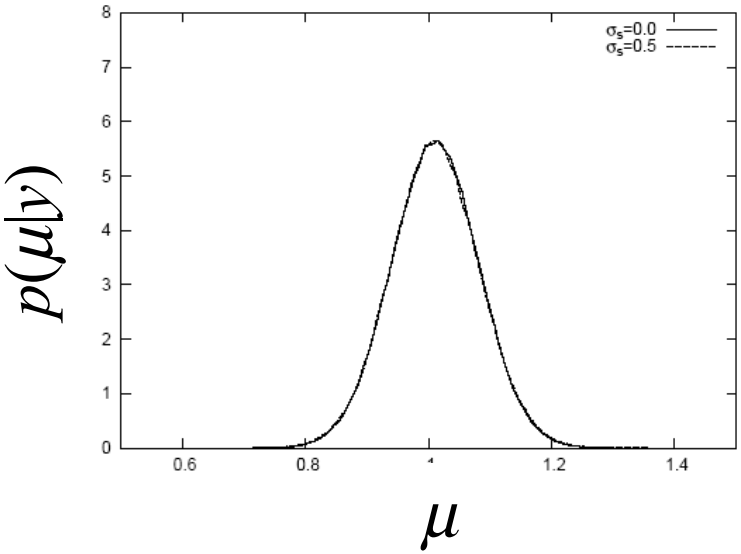
Suppose a fit effectively averages four measurements.

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

Case #1: data appear compatible



Posterior $p(\mu|y)$:



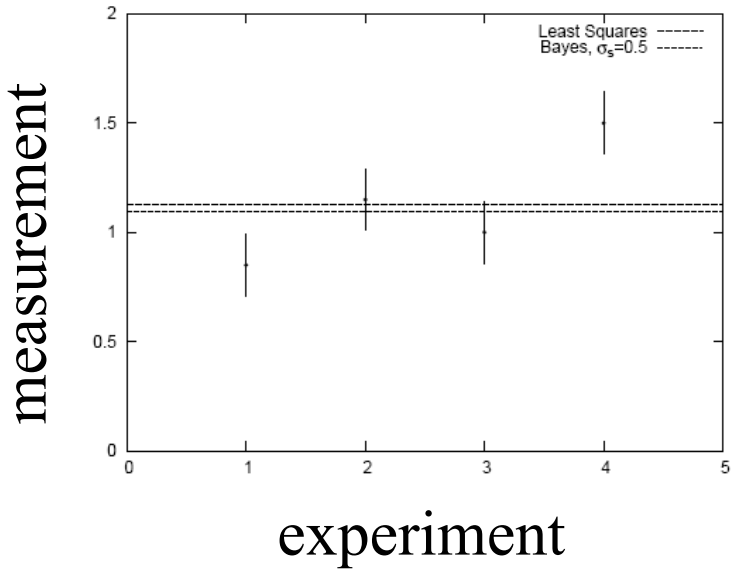
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

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

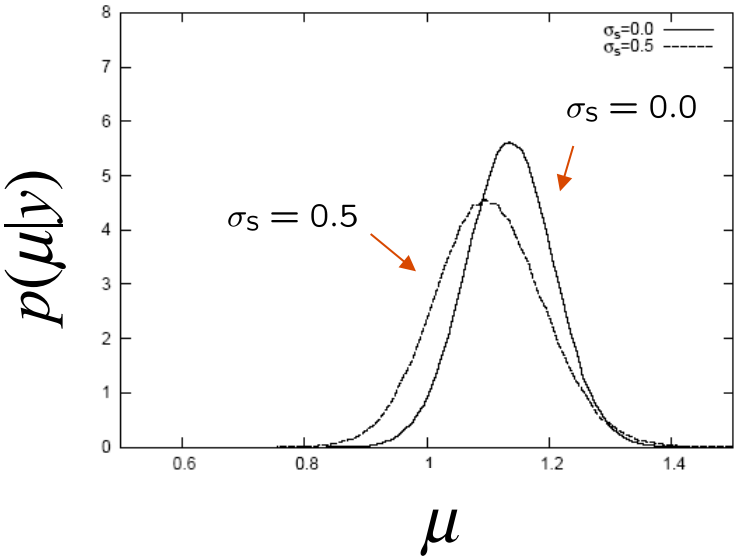
$\sigma_s = 0.5: \quad \hat{\mu} = 1.000 \pm 0.072$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



$$\sigma_s = 0.0: \quad \hat{\mu} = 1.125 \pm 0.071$$

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

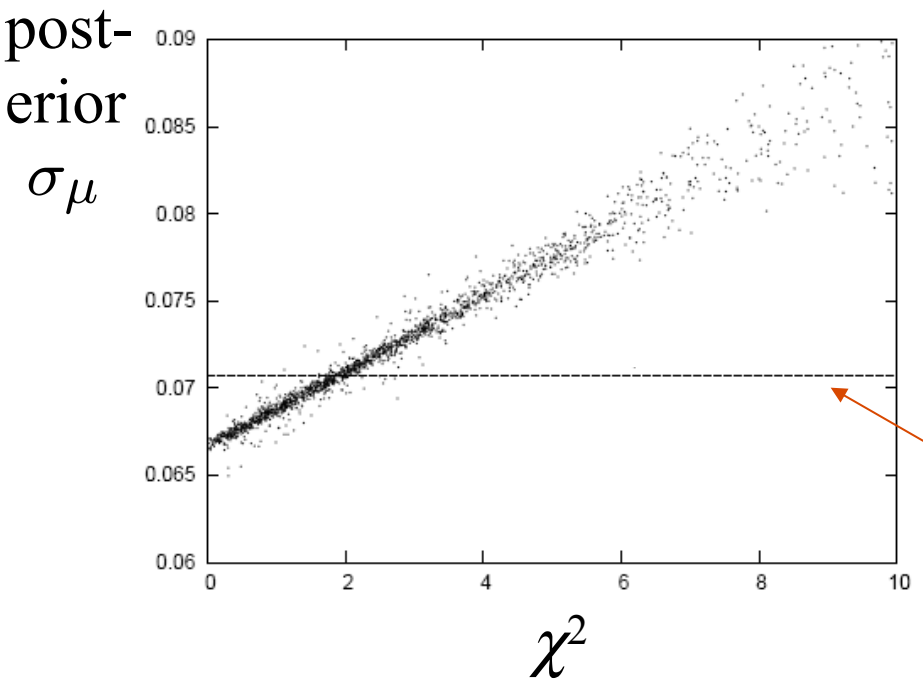
→ Bayesian fit less sensitive to outlier.

(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).



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

σ_μ from least squares

Frequentist errors on errors

Despite the nice features of the Bayesian treatment, it has some important drawbacks:

Bayesian gamma-distributed error-on-error requires numerical integration. (Inverse-gamma prior for s gives Student's *t*, but this allows very large errors).

The Particle Physics community does almost every analysis within a frequentist framework – best if one can include errors on errors without changing entire paradigm.

Recently I have been studying the frequentist treatment of errors-on-errors; outcome is very similar to the Bayesian approach (work in progress).

So first review some properties of frequentist averages...

ML example

Suppose we measure uncorrelated $y_i \sim \text{Gauss}(\mu, \sigma_{y_i}^2)$, $i = 1, \ldots, N$ so that the likelihood is

$$L(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{y_i}} e^{-(y_i-\mu)^2/2\sigma_{y_i}^2}$$

Maximizing the likelihood is equivalent to minimizing

$$\chi^2(\mu) = \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{\sigma_{y_i}^2}$$

This gives a linear unbiased estimator with minimum variance (i.e., equivalent to BLUE):

$$\hat{\mu} = \frac{\sum_{i=1}^{N} \frac{y_i}{\sigma_{y_i}^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_{y_i}^2}} \qquad V[\hat{\mu}] = \frac{1}{\sum_{i=1}^{N} \frac{1}{\sigma_{y_i}^2}}$$

ML example with systematics

Suppose now $y_i \sim \text{Gauss}(\mu + b_i, \sigma_i^2)$, and we have estimates of bias parameter b_i, $u_i \sim \text{Gauss}(b_i, \sigma_{u,i}^2)$, so that the likelihood becomes

$$L(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\frac{(y_i - \mu - \theta_i)^2}{\sigma_i^2}\right] \frac{1}{\sqrt{2\pi\sigma_{u,i}^2}} \exp\left[-\frac{1}{2}\frac{(u_i - \theta_i)^2}{\sigma_{u,i}^2}\right]$$

After profiling over the nuisance parameters $\boldsymbol{b}$, one obtains the same result as before but with

$$\sigma_{y_i}^2 \to \sigma_{y_i}^2 + \sigma_{u_i}^2$$

So again this is the same as BLUE, extended to use addition of the statistical and systematic errors in quadrature.

Example extension: PDG scale factor

Suppose we do not want to take the quoted errors as known constants. Scale the variances by a factor ϕ,

$$\sigma_i^2 \to \phi\sigma_i^2$$

The likelihood function becomes

$$L(\mu,\phi) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\phi\sigma_i^2}} \exp\left[-\frac{1}{2}\frac{(y_i-\mu)^2}{\phi\sigma_i^2}\right]$$

The estimator for μ is the same as before; for ϕ ML gives

$$\hat{\phi}_{\mathrm{ML}} = \frac{\chi^2(\hat{\mu})}{N}$$ which has a bias; $$\hat{\phi} = \frac{\chi^2(\hat{\mu})}{N-1}$$ is unbiased.

The variance of $\hat{\mu}$ is inflated by ϕ: $$V[\hat{\mu}] = \frac{\phi}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

Frequentist errors on errors

Suppose we want to treat the systematic errors as uncertain, so let the $\sigma_{u,i}$ be adjustable nuisance parameters.

Suppose we have estimates s_i for $\sigma_{u,i}$ or equivalently $v_i = s_i^2$, is an estimate of $\sigma_{u,i}^2$.

Model the v_i as independent and gamma distributed:

$$f(v;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} v^{\alpha-1} e^{-\beta v}$$

$$E[v] = \frac{\alpha}{\beta}$$

$$V[v] = \frac{\alpha}{\beta^2}$$

We can set α and β so that they give e.g. a desired relative uncertainty r in σ_u.

Gamma model for estimates of variance

Suppose the estimated variance v was obtained as the sample variance from n observations of a Gaussian distributed bias estimate u.

In this case one can show v is gamma distributed with

$$\alpha = \frac{n-1}{2} \qquad \beta = \frac{n-1}{2\sigma_u^2}$$

We can relate α and β to the relative uncertainty r in the systematic uncertainty as reflected by the standard deviation of the sampling distribution of s, σ_s

$$r = \frac{\sigma_s}{E[s]} = \frac{1}{2}\frac{\sigma_v}{E[v]}$$

Full likelihood with errors on errors

$$L(\mu, \mathbf{b}, \boldsymbol{\sigma}_\mathbf{u}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{y_i}} e^{-(y_i-\mu-b_i)^2/2\sigma_{y_i}^2}$$

$$\times \frac{1}{\sqrt{2\pi}\sigma_{u_i}} e^{-(u_i-b_i)^2/2\sigma_{u_i}^2} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} v_i^{\alpha_i-1} e^{-\beta v_i}$$

Treated like data:

$y_1,\ldots,y_N$	(the real measurements)
$u_1,\ldots,u_N$	(estimates of biases)
$v_1,\ldots,v_N$	(estimates of variances of estimates of biases)

Parameters:

μ	(parameter of interest)
$b_1,\ldots,b_N$	(bias parameters)
$\sigma_{u1},\ldots,\ \sigma_{uN}$	(sys. errors = std. dev. of of bias estimates)

Full log-likelihood with errors on errors

Setting the parameters of the gamma distributions in terms of the relative uncertainty in the systematic errors r_i and the systematic errors themselves σ_{ui},

$$\alpha_i = \frac{1}{4r_i^2} \qquad \beta_i = \frac{1}{4r_i^2\sigma_{u_i}^2}$$

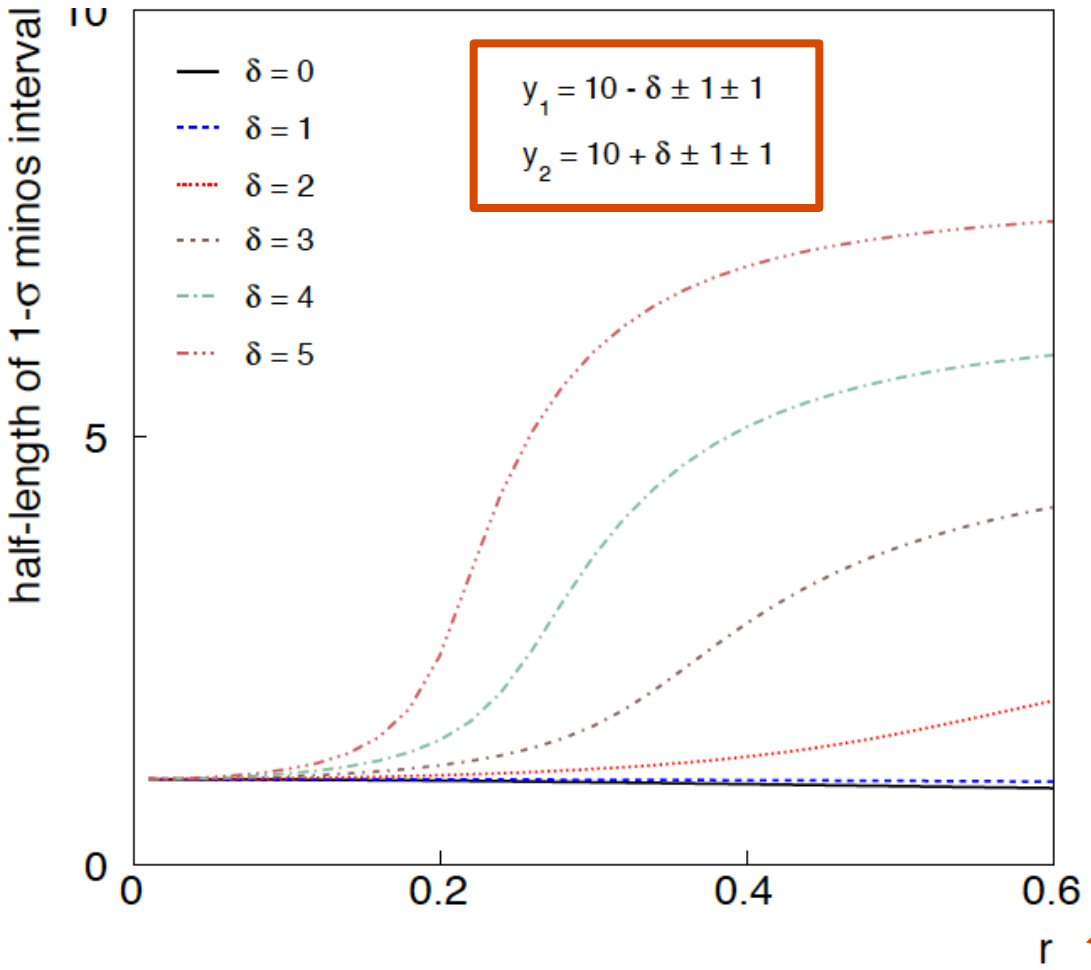
which gives the log-likelihood

$$\ln L(\mu, \mathbf{b}, \boldsymbol{\sigma}_\mathbf{u}) = -\frac{1}{2}\sum_{i=1}^{N}\left[\frac{(y_i - \mu - b_i)^2}{\sigma_{y_i}^2} + \ln\sigma_{u_i}^2 + \frac{(u_i - b_i)^2}{\sigma_{u_i}^2} + \frac{1}{2r_i^2}\left(\ln\sigma_{u_i}^2 + \frac{v_i}{\sigma_{u_i}^2}\right)\right] + C$$

Toy study with errors on errors

MINOS interval (= approx. confidence interval) based on

$$\ln L(\mu, \mathbf{b}, \boldsymbol{\sigma}_\mathbf{u}) = \ln L_{\max} - \frac{Q_\alpha}{2} \quad \text{with} \quad Q_\alpha = F^{-1}_{\chi^2}(1-\alpha; n)$$



Increased discrepancy between values to be averaged gives larger interval.

Interval length saturates at ~level of absolute discrepancy between input values.

relative error on sys. error

Goodness of fit with errors on errors

Because we now treat the σ_{ui} as adjustable parameters, $-2\ln L$ is no longer a sum of squares; "usual" χ^2 not usable for g.o.f.

To assess the goodness of fit, one can define a statistic based on the profile likelihood ratio

$$q = -2 \ln \frac{L(\hat{\mu}, \hat{\mathbf{b}}, \hat{\boldsymbol{\sigma}}_\mathbf{u})}{L(\hat{\boldsymbol{\mu}}, \hat{\mathbf{b}}, \hat{\boldsymbol{\sigma}}_\mathbf{u})}$$

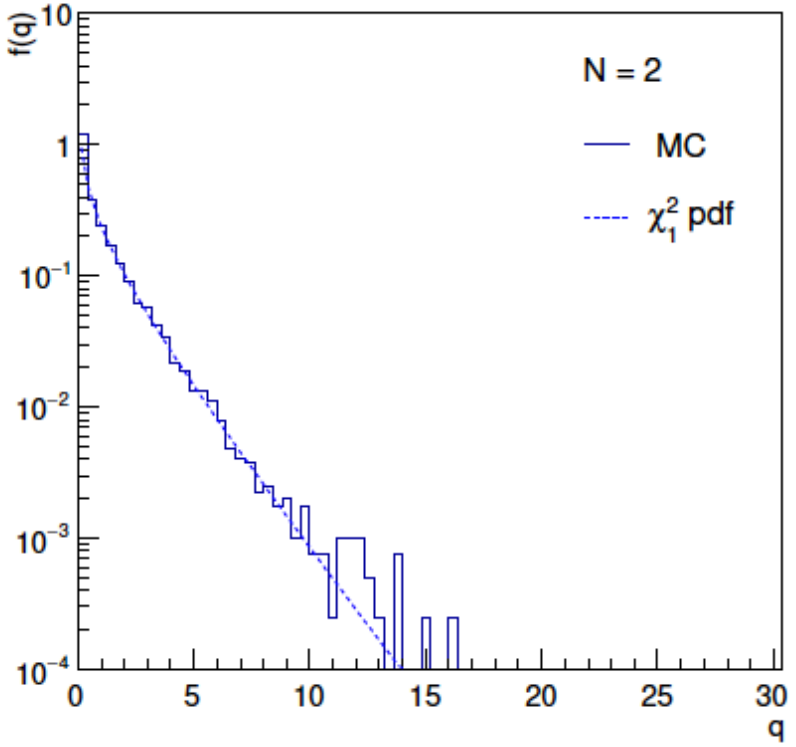
For q in the numerator one assumes a single value of mu for all measurements (the model being tested).

The denominator is the "saturated model", i.e., an adjustable μ_i for each measurement, i.e., $\boldsymbol{\mu} = (\mu_1, ..., \mu_N)$.

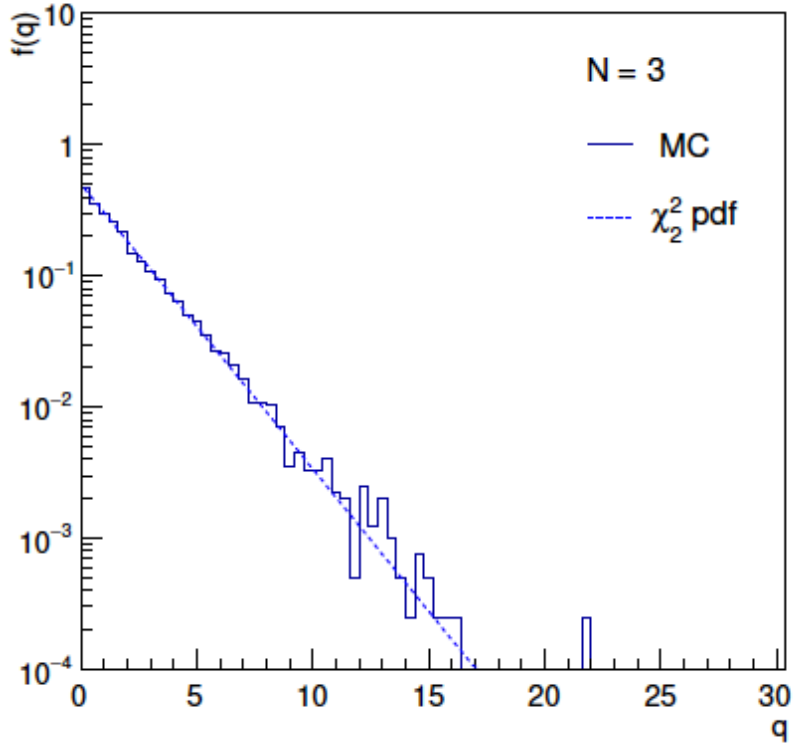
Asymptotically from Wilks' theorem $q \sim$ chi-square($N-1$)

Goodness of fit with errors on errors ($N = 2,3$)

Asymptotic distribution works well with $r = 0.2$

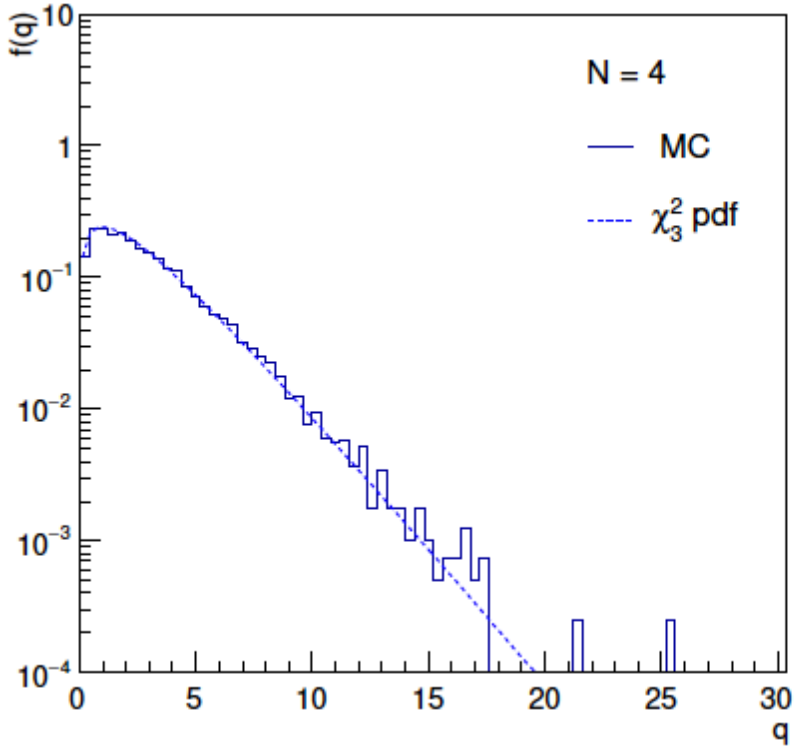


(a)

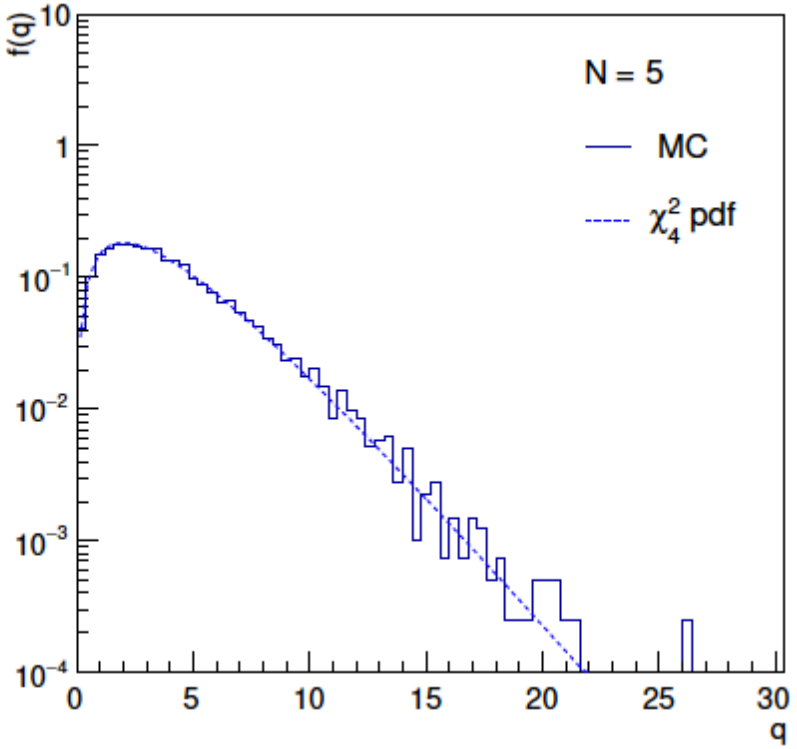


Goodness of fit with errors on errors ($N = 4,5$)

Asymptotic distribution works well with $r = 0.2$

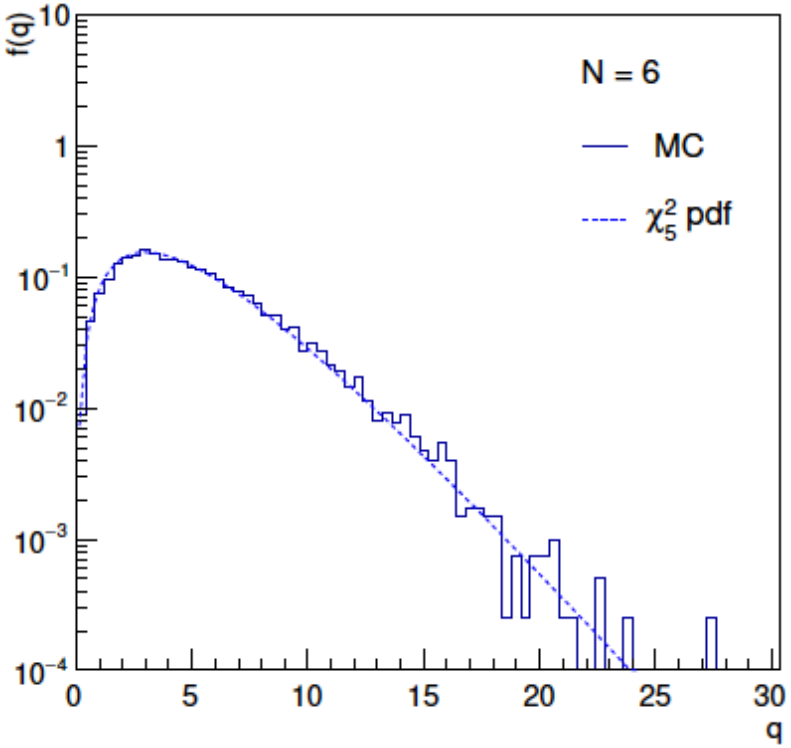


(c)

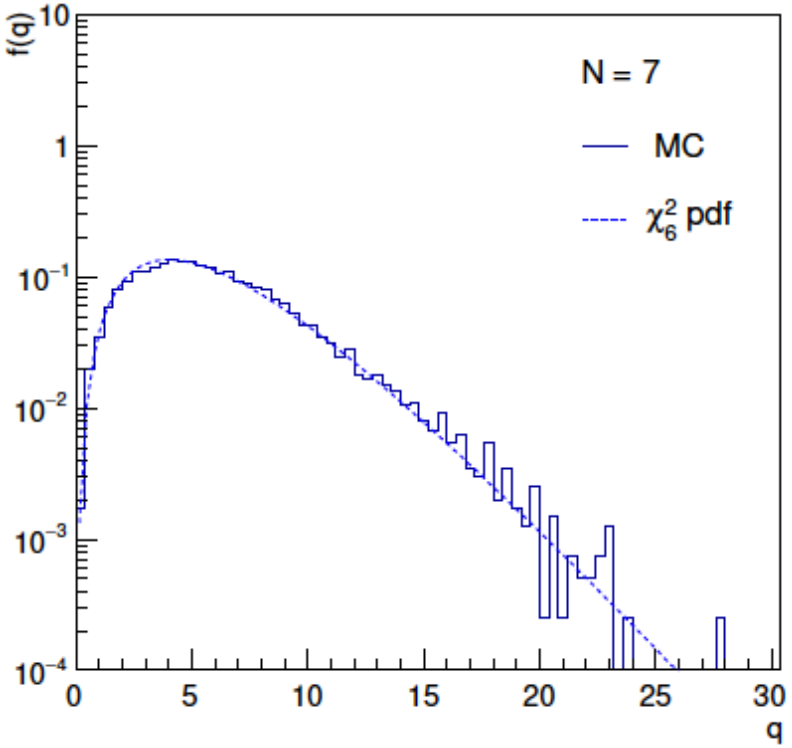


Goodness of fit with errors on errors ($N = 6,7$)

Asymptotic distribution works well with $r = 0.2$



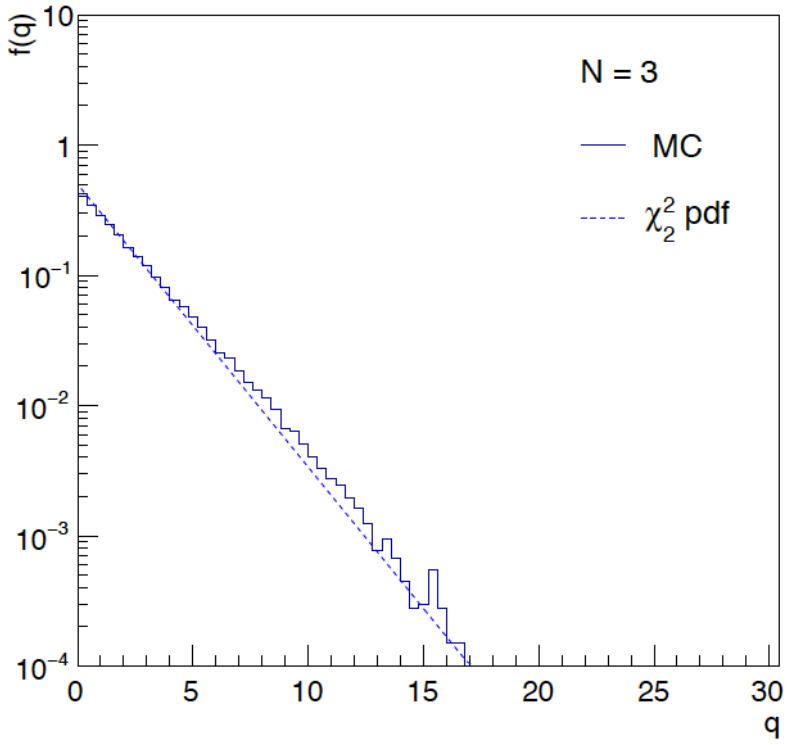
(e)



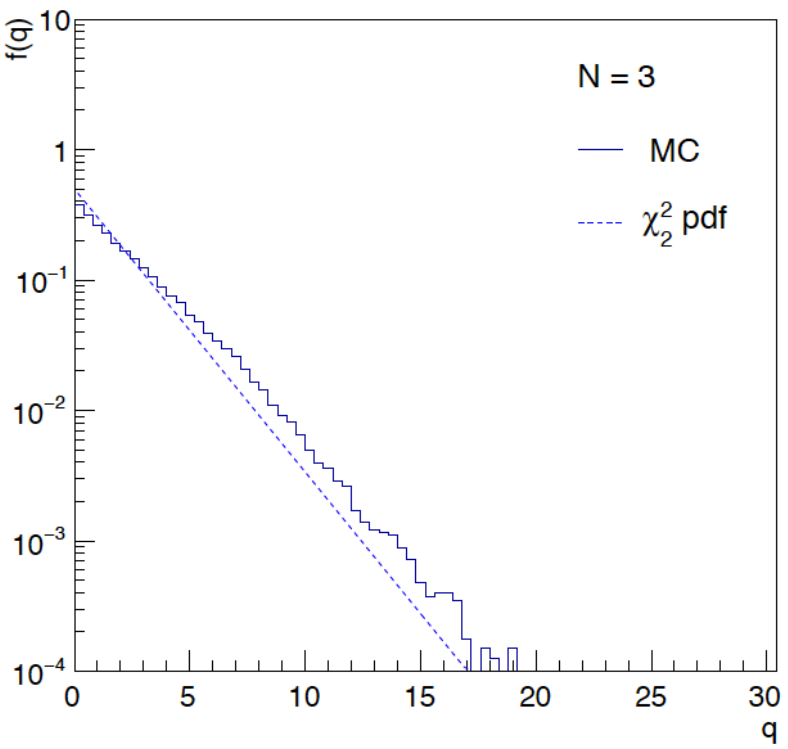
Goodness-of-fit with large errrors on errors

Asymptotic distribution starts to break for large r

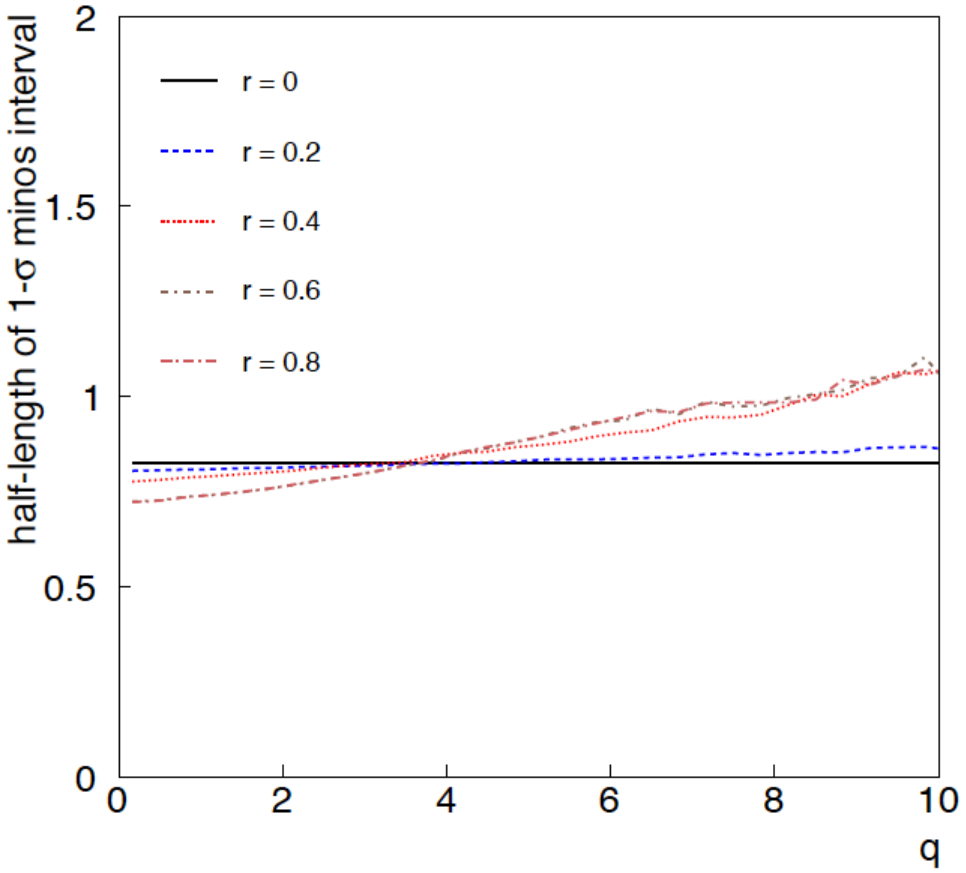
$r = 0.4$



$r = 0.6$



Connection between goodness-of-fit and size of confidence interval



Similar to Bayesian case, the length of the confidence interval increases if the goodness-of-fit is bad (high q), but only if one includes enough error-on-error.

Further work

Next:

Study sensitivity of average to outliers with errors-on-errors

Extend to analyses beyond averages, e.g.,

$$\mu \rightarrow \boldsymbol{\mu} = (\mu_1, \ldots, \mu_N) = \boldsymbol{\mu}(\boldsymbol{\theta})$$

Still a number of software issues to overcome, e.g., problems with fit convergence, determining MINOS interval (thanks to Lorenzo Moneta for help with this)

Extra slides

Using LS to combine measurements

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

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

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

λ = true value (plays role of θ).

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,

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

Combining correlated measurements with LS

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

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

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

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

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 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

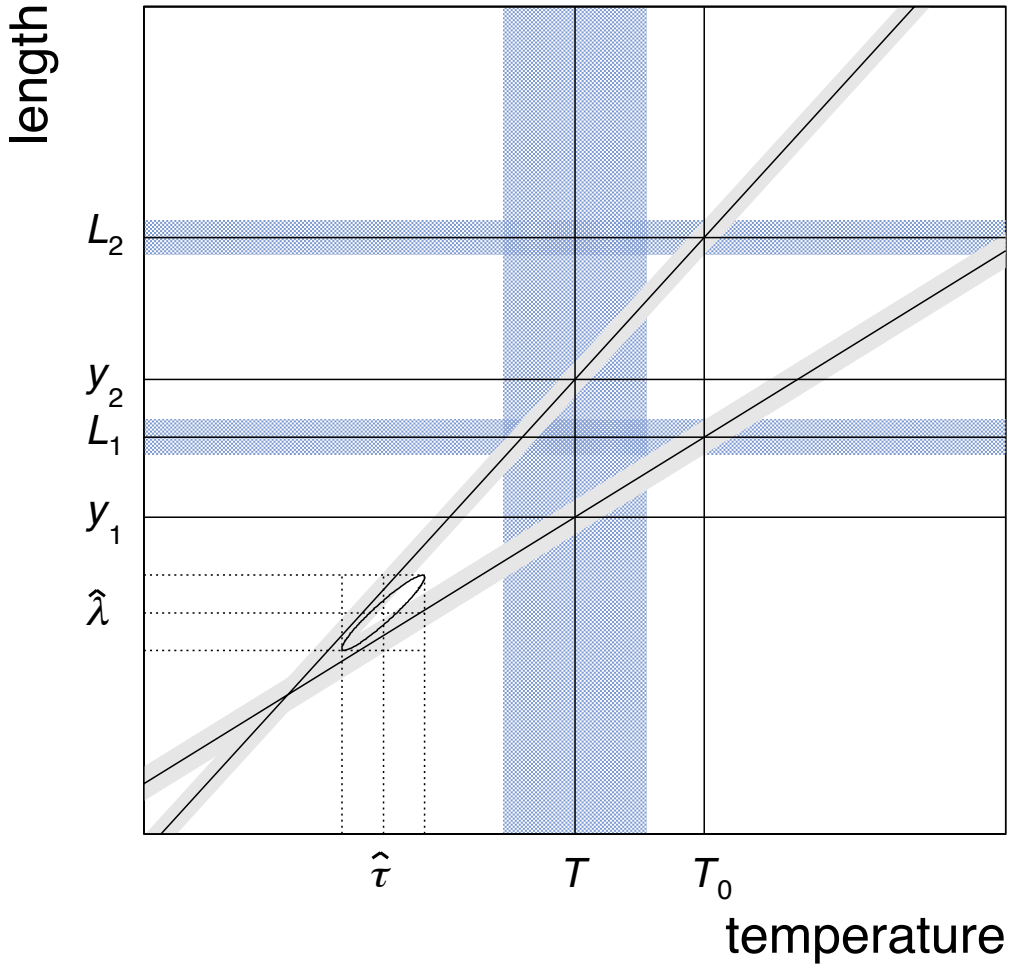
$$\mathrm{cov}[y_1, y_2] = \alpha_1\alpha_2\sigma_T^2 \qquad \mathrm{cov}[y_i, T] = \alpha_i\sigma_T^2$$

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.

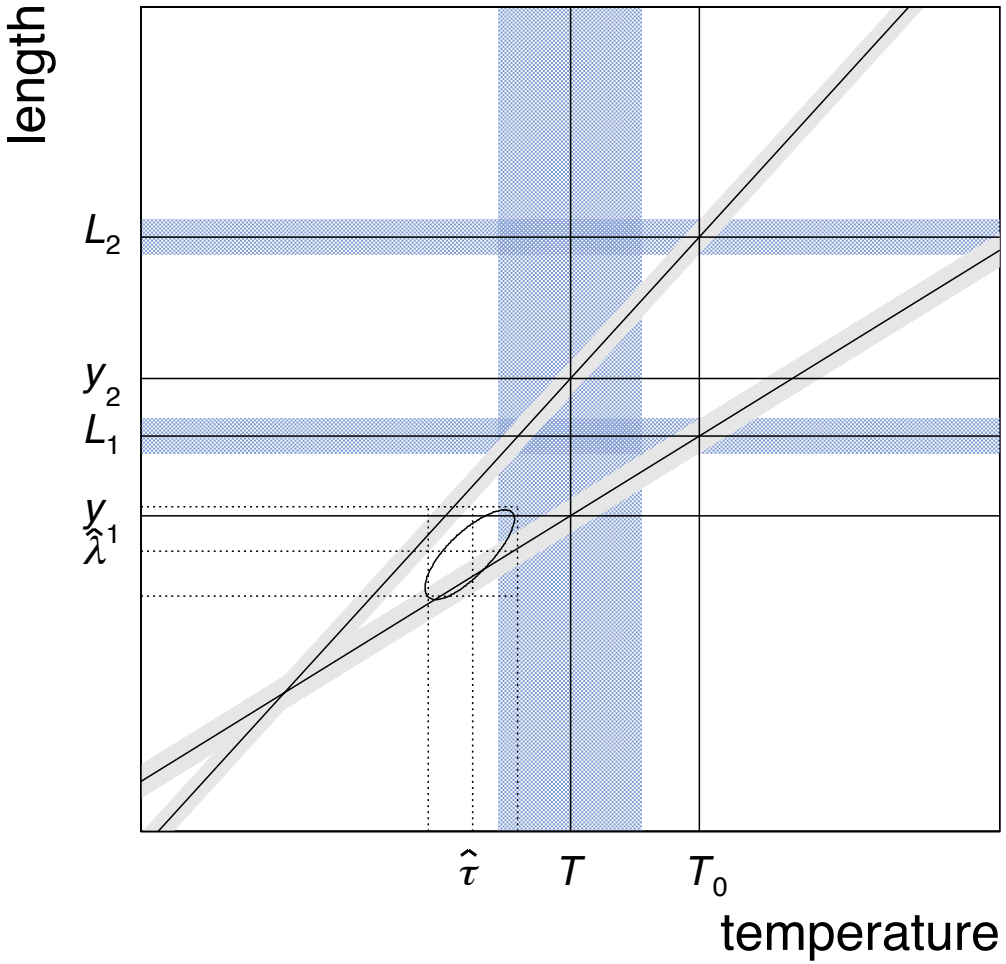
$$L(L_1, L_2, T, \tilde{\tau}_0, \tilde{\alpha}_1, \tilde{\alpha}_2 | \lambda, \tau, \tau_0, \alpha_1, \alpha_2) =$$

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

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

Two rulers (5)

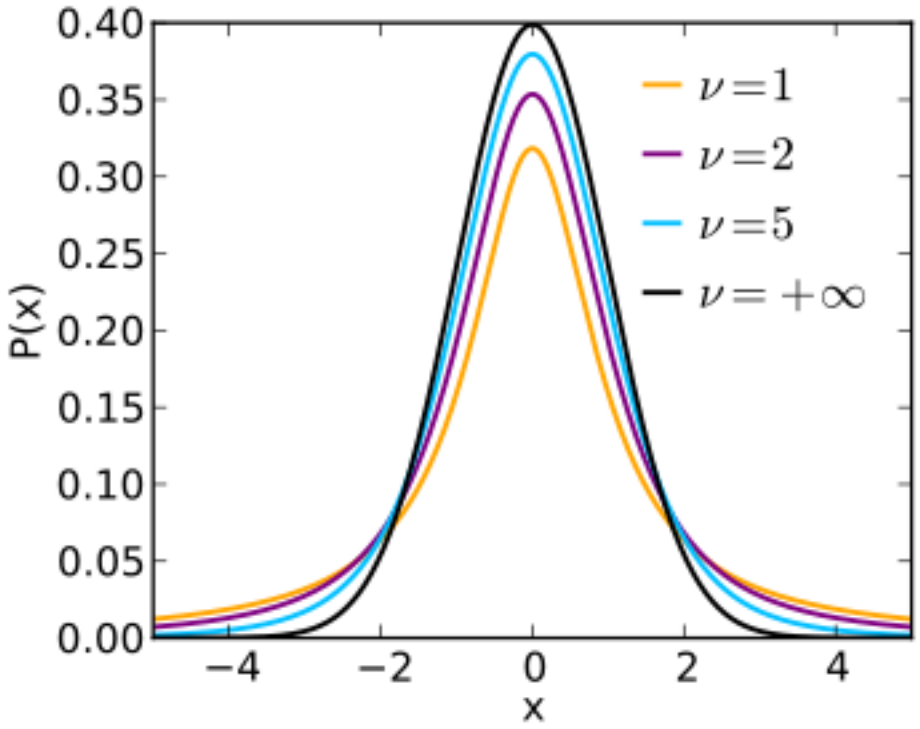
The outcome changes; some surprises may be "reduced".



Non-Gaussian likelihoods

Tails of Gaussian fall of very quickly; not always most realistic model esp. for systematics. Can try e.g. Student's *t*,

$$f(x;\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\frac{1}{\sigma}\left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}$$



$v = 1$ gives Cauchy,

v large gives Gaussian.

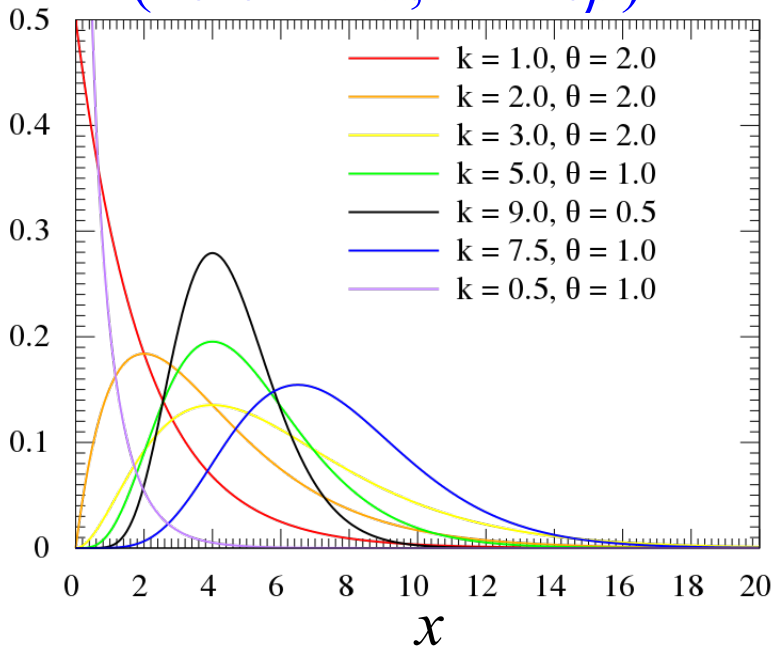
Can either fix v or constrain like other nuisance parameters.

ML no longer equivalent to least-squares, BLUE.

Gamma distribution for sys. errors

First attempt to treat estimates of sys. errors as log-normal distributed has long tail towards large errors; maybe more realistic to use gamma distribution:

(here $k = \alpha$, $\theta = 1/\beta$)



$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

$$E[x] = \frac{\alpha}{\beta} \qquad V[x] = \frac{\alpha}{\beta^2}$$

Take $s \sim \text{Gamma}(\alpha, \beta)$ as "measurement" of sys. error σ_u with relative unc. in sys. error r, $\alpha = 1/r^2$, $\beta = \alpha/\sigma_u$.