Combinations in Statistical Data Analysis
From basics to errors on errors

DESY 2nd Pan-European Advanced School of Statistics

https://indico.desy.de/event/32536/

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

Outline

Basic formalism

- Frequentist vs Bayesian approach
- Systematic uncertainties

Combination based on a full likelihood

Simplified approaches

Errors on errors

Discussion and conclusions

Basic formalism: likelihood

Suppose the outcome of a measurement is a collection of numbers $\boldsymbol{x}$ (scalar or vector) – here, the “data”.

And suppose a model (hypothesis H) predicts the probability for the data:

$$P(\mathbf{x}|H)$$

Often a family of models is indexed by a set of parameters, i.e.,

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

If we view this as a function of the model (or of the parameters), then this is the likelihood; often written

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

Frequentist approach

Frequentist statistics: probability only associated with data x, not hypotheses or parameters.

Hypothesis (or model or parameter value) is "preferred" if the model predicts a high probability for data like what we got.

Important tools:

Maximum likelihood estimator for parameters

Hypothesis test of size α

Reject H if data found in critical region w with $P(x \text{ in } w \mid H) \leq \alpha$

p-value of hypothesis H

= $P(x$ equally or more incompatible with $H \mid H)$

Confidence interval at CL = $1 - \alpha$

= set of parameter values with p-value > α

Bayesian approach

Probability associated with both data and hypotheses

probability of the data assuming hypothesis H (the likelihood)

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

$$P(\theta|x) = \frac{P(x|\theta)\pi(\theta)}{\int P(x|\theta)\pi(\theta)\,d\theta}$$

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

normalization involves sum over all possible hypotheses

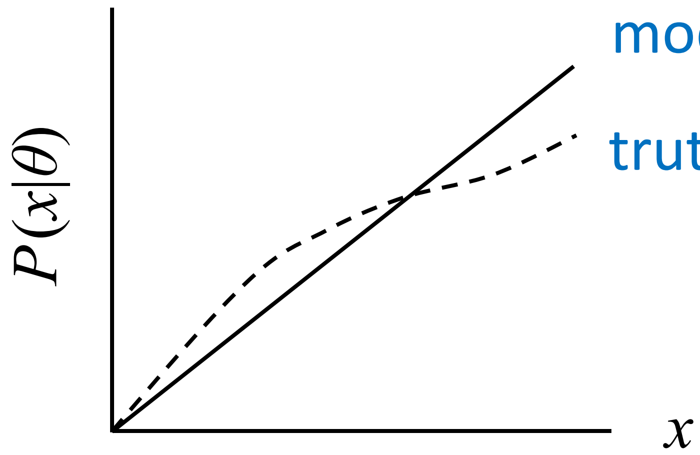
Requires prior probabilities for all relevant parameters/hypotheses.

Inference follows from the posterior probability, e.g., point estimate from mode of posterior, credible intervals,...

For both Bayesian and Frequentist approaches, the model $P(x|\theta)$ is a fundamental ingredient.

Systematic uncertainties and nuisance parameters

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



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

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

Can improve model by including additional adjustable parameters.

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

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

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

Combinations: simplest case

Suppose two measurements yield independent data whose model contains a parameter μ understood to be the same for both.

$$x \sim P(x|\mu)$$

$$y \sim P(y|\mu)$$

Goal: combine the information from x and y to estimate/test μ.

If x and y are independent, then the joint probability for the data is

$$P(x, y|\mu) = P(x|\mu)\, P(y|\mu)$$

$$= L(\mu) \qquad \leftarrow \text{ the likelihood}$$

So use this for e.g.

Frequentist: maximum likelihood, *p*-value, conf. interval

Bayesian: use in Bayes' theorem $\rightarrow$ posterior $P(\mu|x, y)$

Combo with nuisance parameters

Suppose that the models contains a nuisance parameters θ, λ, ξ, in addition to the parameter of interest μ, where θ is common to both models but λ and ξ are not.

$$x \sim P(x | \mu, \theta, \lambda)$$

$$y \sim P(y | \mu, \theta, \xi)$$

Auxiliary measurements to constrain the nuisance parameters:

$$\boldsymbol{u} \sim P(\boldsymbol{u} | \theta, \lambda, \xi)$$

If the primary and auxiliary measurements are independent, then the joint probability for x, y and u is

$$P(x, y, u \mid \mu, \theta, \lambda, \xi) = P(x | \mu, \theta, \lambda) \, P(y | \mu, \theta, \xi) \, P(\boldsymbol{u} \mid \theta, \lambda, \xi)$$

Having θ common to the models for both x and y corresponds to a “correlated systematic”.

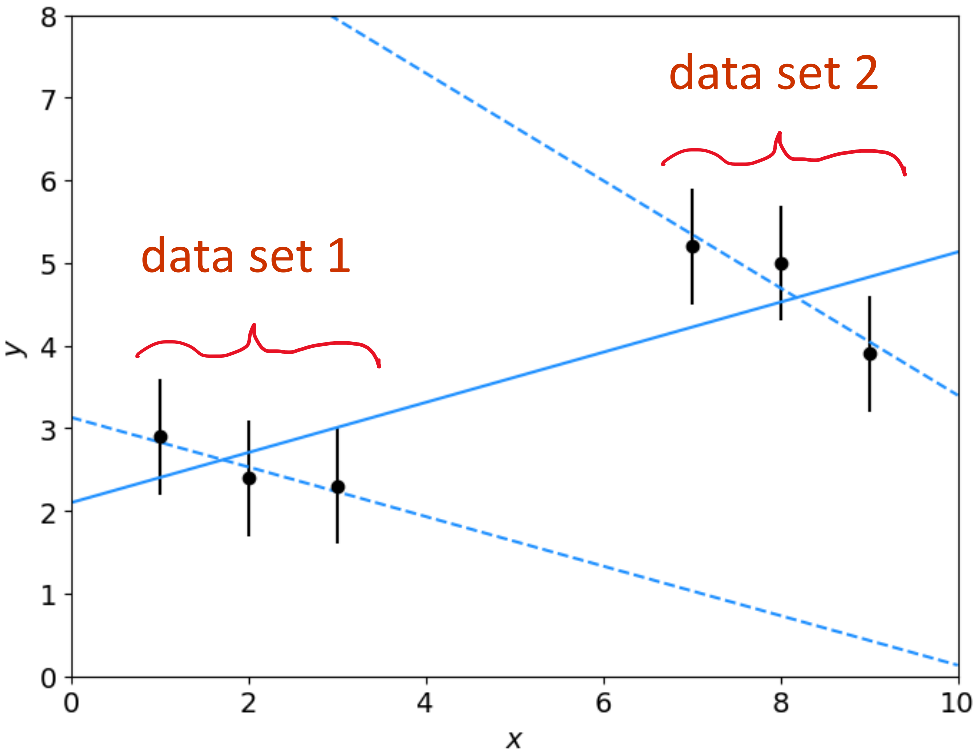
Example of a combination

Inspiration from Louis Lyons via Olaf Behnke

Fit a straight line: $f(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x$

Minimize $$-2\ln L(\boldsymbol{\theta}) = \sum_i \frac{(y_i - f(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} \equiv \chi^2(\boldsymbol{\theta})$$

2 data sets with 3 measurements each:



Data set 1 only:

$\theta_0 = 3.13 \pm 1.07$

$\theta_1 = -0.30 \pm 0.49$

p-value = 0.82

Data set 2 only:

$\theta_0 = 9.90 \pm 3.98$

$\theta_1 = -0.65 \pm 0.49$

p-value = 0.60

Combination:

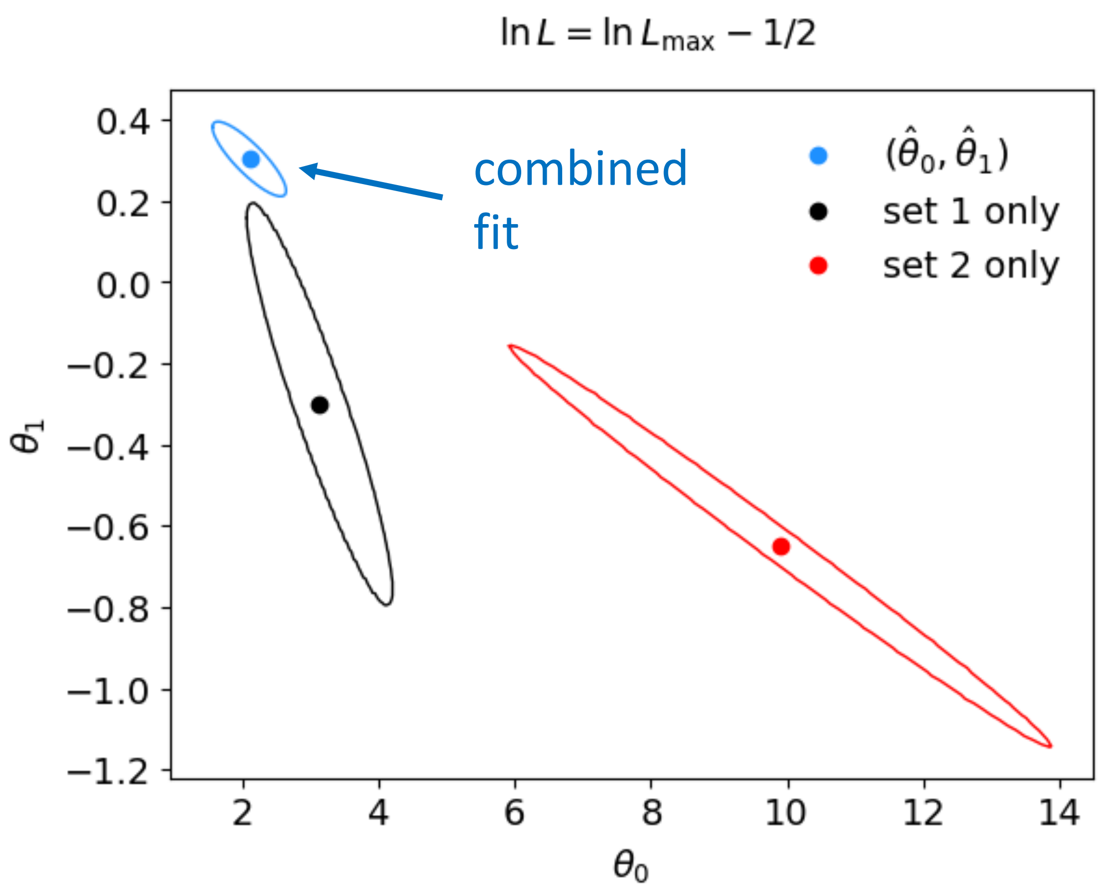
$\theta_0 = 2.10 \pm 0.54$

$\theta_1 = 0.303 \pm 0.092$

p-value = 0.21

Results from combination

In this example the combination leads to a very large reduction in the uncertainties.



More nuisance parameters

But what if we allowed for systematic biases in each of the two data sets, i.e.,

nuisance parameters

for data set 1: $f_1(x;\theta_0,\theta_1,\lambda_1) = \lambda_1 + \theta_0 + \theta_1 x$

for data set 2: $f_2(x;\theta_0,\theta_1,\lambda_2) = \lambda_2 + \theta_0 + \theta_1 x$

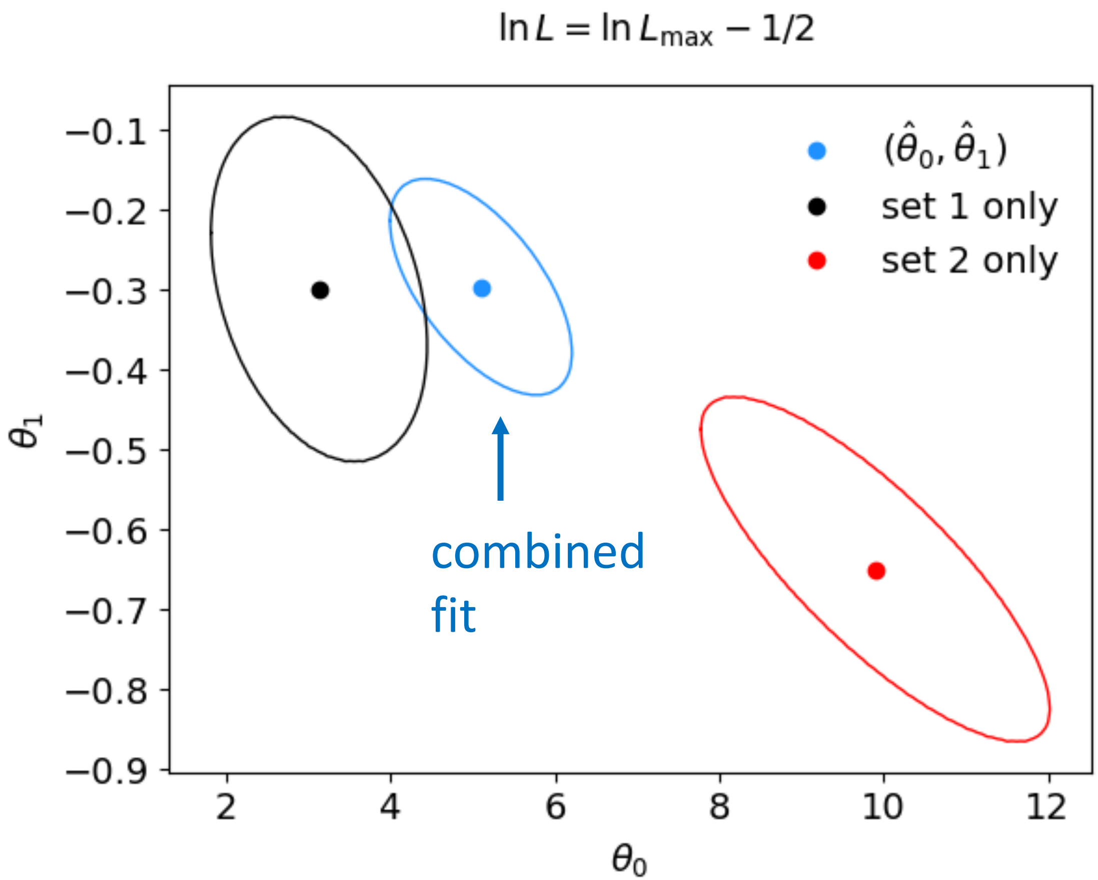
$-2\ln L(\boldsymbol{\theta},\boldsymbol{\lambda})$ becomes:

auxiliary measurements

$$\chi^2(\boldsymbol{\theta},\boldsymbol{\lambda}) = \sum_{i=1,2,3} \frac{(y_i - f_1(x_i;\boldsymbol{\theta},\lambda_1))^2}{\sigma_i^2} + \sum_{i=4,5,6} \frac{(y_i - f_2(x_i;\boldsymbol{\theta},\lambda_2))^2}{\sigma_i^2} + \sum_{j=1}^{2} \frac{(u_j - \lambda_j)^2}{\sigma_{u,j}^2}$$

Independent nuisance parameters

Separately adjustable λ_1 and λ_2 each with independent Gaussian distributed estimate u_i, ~Gauss(λ_1, σ_u)



Combination now prefers negative slope parameter θ_1, since each data set can tolerate some separate vertical shift.

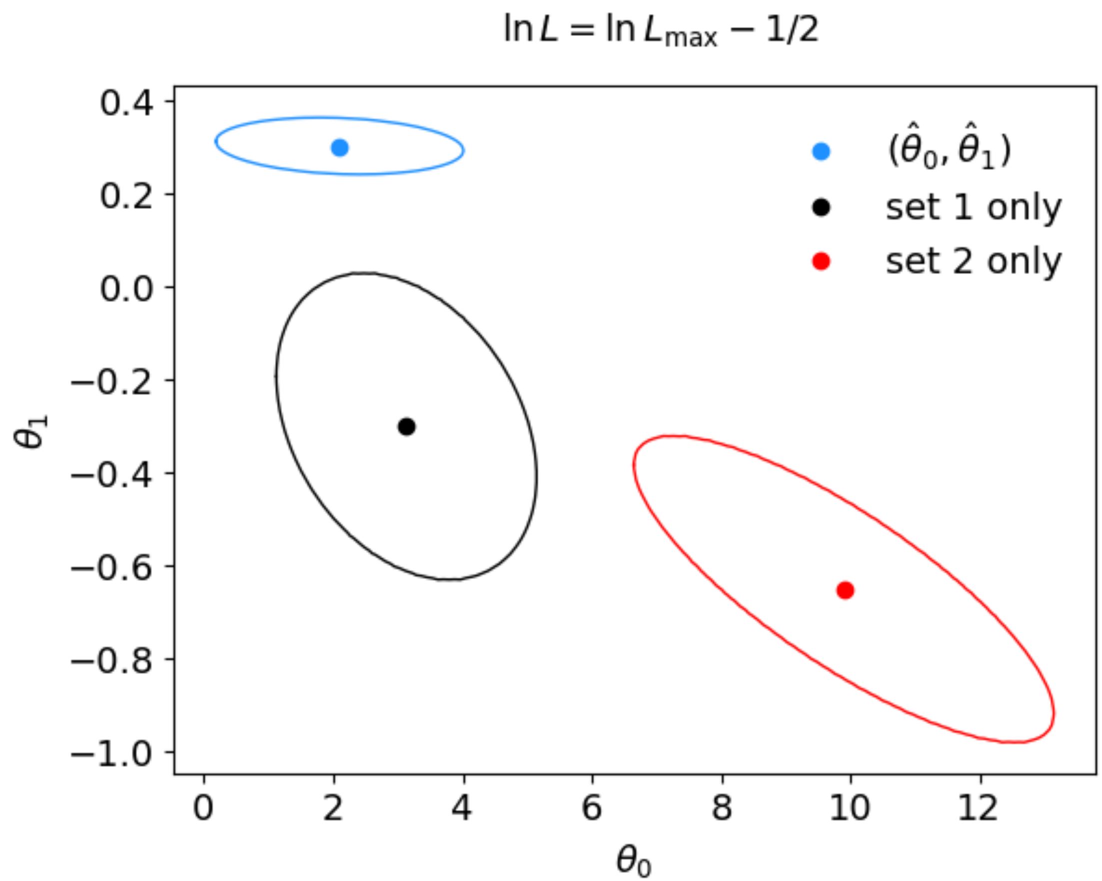
Common nuisance parameter

Alternatively, we might have $\lambda_1 = \lambda_2 \equiv \lambda$, so minimize

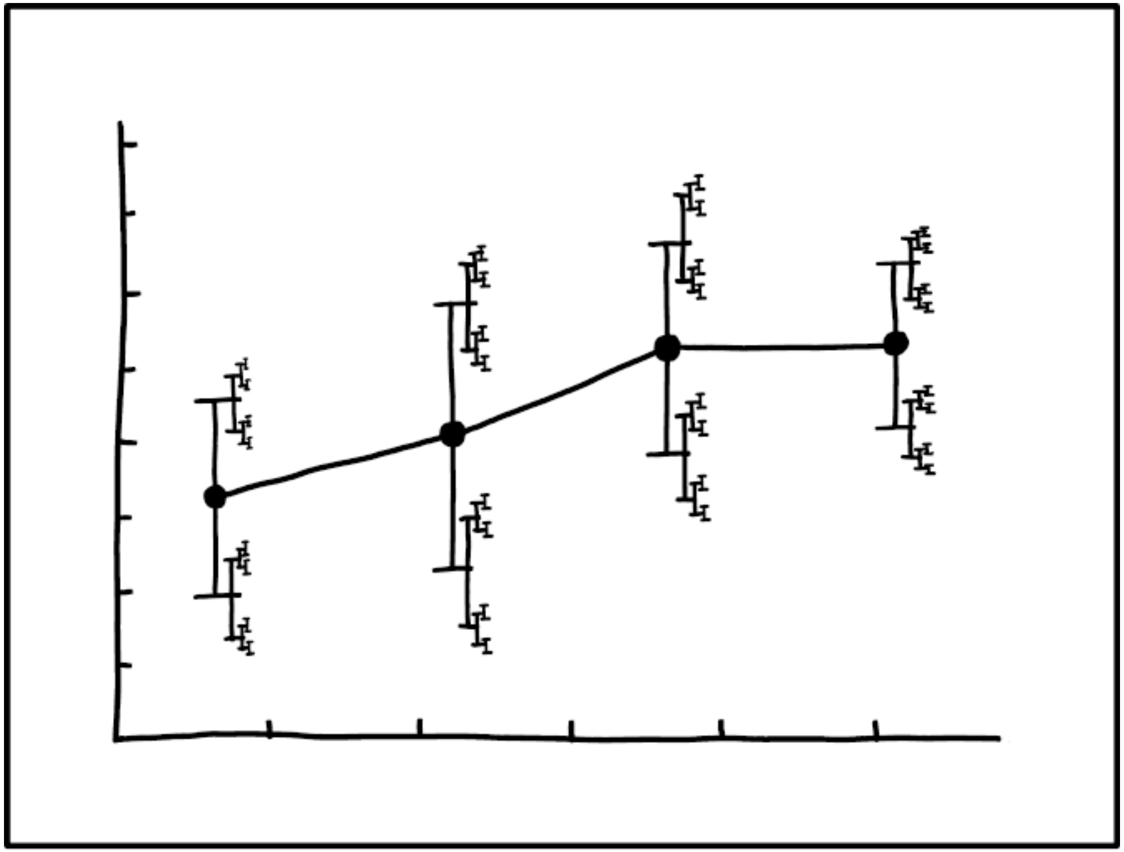
$$\chi^2(\boldsymbol{\theta}, \lambda) = \sum_i \frac{(y_i - f(x_i; \boldsymbol{\theta}, \lambda))^2}{\sigma_i^2} + \frac{(u - \lambda)^2}{\sigma_u^2}$$

Now the two data sets can only move up and down coherently, so the slope parameter θ_1 from the combination is again very accurate.

The key (and the hard part) in a combination is identifying common nuisance parameters.



Errors on Errors



I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.

Details in G. Cowan, Eur. Phys. J. C (2019) 79:133, arXiv:1809.05778

Collaborators include: Enzo Canonero (RHUL), Alessandra Brazzale (U. Padova)

Motivation

Analyses that are limited by systematic uncertainties become sensitive to the assigned values of systematic errors.

But these error estimates are also uncertain ($\rightarrow$ errors on errors)

Could just try inflating the systematic error estimates, but this turns out not to be enough, especially if the analysis uses least squares (equivalent to assuming Gaussian pdfs in likelihood).

Need for "errors on errors" most visible when measurements are not internally consistent within their estimated uncertainties.

Candidate use cases in particle physics:

Combinations of inconsistent measurements

Analyses where systematic error assigned by ad hoc recipe

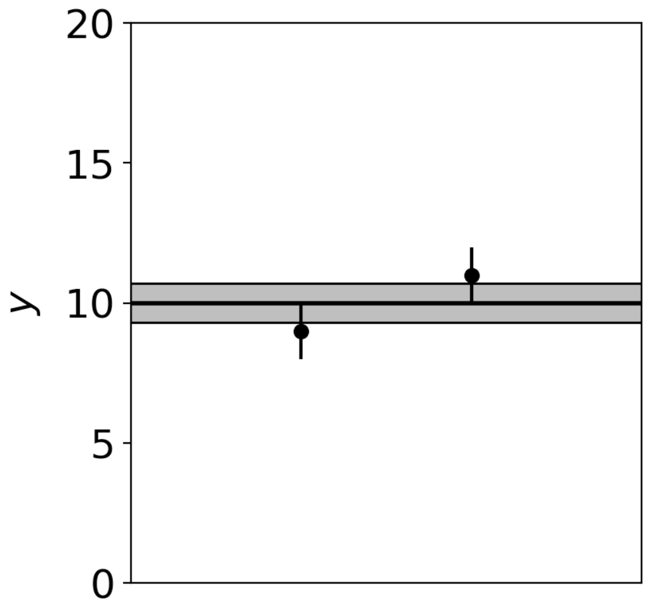
Any analysis where assigned systematic error is uncertain

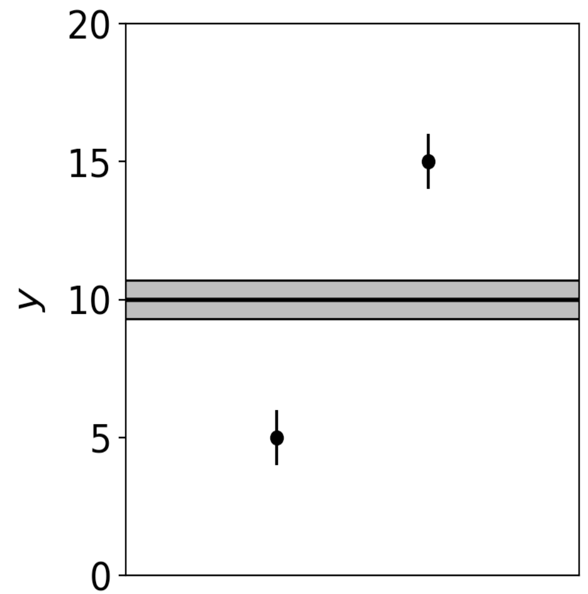
Motivation (2)

Assuming known standard deviations for least squares, uncertainty (e.g. confidence interval) does not reflect goodness of fit:

Least squares average of 9 ± 1 and 11 ± 1 is 10 ± 0.71

Least squares average of 5 ± 1 and 15 ± 1 is 10 ± 0.71





Width of confidence interval for the mean does not reflect the consistency of the values being averaged.

Formulation of the problem

Suppose measurements $\boldsymbol{y}$ have probability (density) $P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta})$,

$\boldsymbol{\mu}$ = parameters of interest

$\boldsymbol{\theta}$ = nuisance parameters

To provide info on nuisance parameters, often treat their best estimates $\boldsymbol{u}$ as indep. Gaussian distributed r.v.s., giving likelihood

$$L(\boldsymbol{\mu},\boldsymbol{\theta}) = P(\mathbf{y},\mathbf{u}|\boldsymbol{\mu},\boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta})P(\mathbf{u}|\boldsymbol{\theta})$$

$$= P(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}) \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{u_i}} e^{-(u_i-\theta_i)^2/2\sigma_{u_i}^2}$$

or log-likelihood (up to additive const.)

$$\ln L(\boldsymbol{\mu},\boldsymbol{\theta}) = \ln P(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}) - \frac{1}{2}\sum_{i=1}^{N} \frac{(u_i-\theta_i)^2}{\sigma_{u_i}^2}$$

Systematic errors and their uncertainty

Sometimes $\sigma_{u,i}$ is well known, e.g., it is itself a statistical error known from sample size of a control measurement.

Other times the u_i are from an indirect measurement, Gaussian model approximate and/or the $\sigma_{u,i}$ are not exactly known.

Or sometimes $\sigma_{u,i}$ is at best a guess that represents an uncertainty in the underlying model ("theoretical error").

In any case we can allow that the $\sigma_{u,i}$ are not known in general with perfect accuracy.

Gamma model for variance estimates

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:

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

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

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

Set α and β so that they give desired relative uncertainty r in σ_u.

Other ”bell-shaped” models tried; qualitatively similar results.

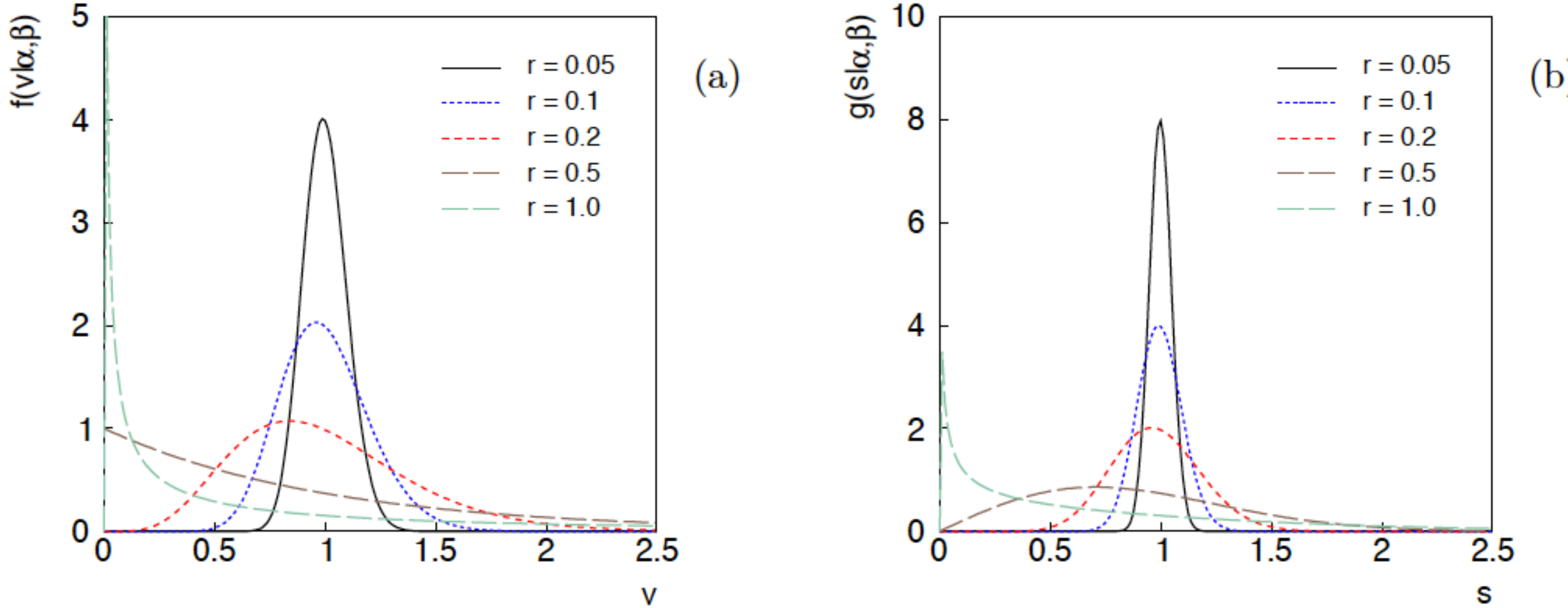
Gamma pdf leads to important mathematical simplifications.

Distributions of v and $s = \sqrt{v}$

For α, β of gamma distribution, $\alpha_i = \frac{1}{4r_i^2}\,, \quad \beta_i = \frac{1}{4r_i^2 \sigma_{u_i}^2}$

$$r_i \equiv \frac{1}{2}\frac{\sigma_{v_i}}{E[v_i]} = \frac{1}{2}\frac{\sigma_{v_i}}{\sigma_{u_i}^2} \approx \frac{\sigma_{s_i}}{E[s_i]}$$

relative “error on error”



Likelihood for Gamma Variance Model

$$L(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\sigma}_{\mathbf{u}}^2) = P(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{u_i}^2}} e^{-(u_i-\theta_i)^2/2\sigma_{u_i}^2}$$

$$\times \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} v_i^{\alpha_i - 1} e^{-\beta_i v_i} .$$

$$\alpha_i = \frac{1}{4r_i^2} ,$$

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

Treated like data: $y_1,\ldots,y_L$ (the primary measurements)

$u_1,\ldots,u_N$ (estimates of nuisance par.)

$v_1,\ldots,v_N$ (estimates of variances of estimates of NP)

Adjustable parameters: $\mu_1,\ldots,\mu_{\mathrm{M}}$ (parameters of interest)

$\theta_1,\ldots,\theta_N$ (nuisance parameters)

$\sigma_{u,1},\ldots,\sigma_{u,N}$ (sys. errors = std. dev. of of NP estimates)

Fixed parameters: $r_1,\ldots,r_N$ (rel. err. in estimate of $\sigma_{u,i}$)

Profiling over systematic errors

We can profile over the $\sigma_{u,i}$ in closed form

$$\widehat{\widehat{\sigma^2}}_{u_i} = \underset{\sigma^2_{u_i}}{\operatorname{argmax}} L(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2_{\mathbf{u}}) = \frac{v_i + 2r_i^2 (u_i - \theta_i)^2}{1 + 2r_i^2}$$

which gives the profile log-likelihood (up to additive const.)

$$\boxed{\begin{aligned} \ln L'(\mu, \boldsymbol{\theta}) &= \ln L(\mu, \boldsymbol{\theta}, \widehat{\widehat{\boldsymbol{\sigma}^2}}_{\mathbf{u}}) \\ &= \ln P(\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\theta}) - \frac{1}{2} \sum_{i=1}^{N} \left(1 + \frac{1}{2r_i^2} \right) \ln \left[1 + 2r_i^2 \frac{(u_i - \theta_i)^2}{v_i} \right] \end{aligned}}$$

In limit of small r_i and $v_i \rightarrow \sigma_{u,i}^2$, the log terms revert back to the quadratic form seen with known $\sigma_{u,i}$.

Equivalent likelihood from Student's t

We can arrive at same likelihood by defining $z_i \equiv \dfrac{u_i - \theta_i}{\sqrt{v_i}}$

Since $u_i \sim$ Gauss and $v_i \sim$ Gamma, $z_i \sim$ Student's t

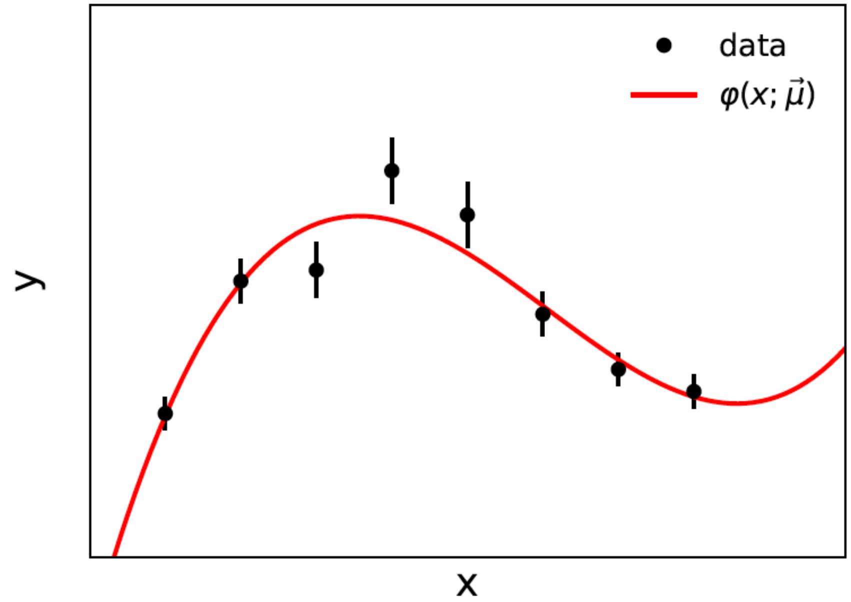
$$f(z_i|\nu_i) = \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\sqrt{\nu_i \pi}\,\Gamma(\nu_i/2)} \left(1 + \frac{z_i^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}}$$

with $\nu_i = \dfrac{1}{2r_i^2}$

Resulting likelihood same as profile $L'(\boldsymbol{\mu},\boldsymbol{\theta})$ from gamma model

$$L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \prod_{i=1}^{N} \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\sqrt{\nu_i \pi}\,\Gamma(\nu_i/2)} \left(1 + \frac{z_i^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}}$$

Curve fitting, averages



Suppose independent
$y_i \sim$ Gauss, $i = 1,\ldots,N$, with

$$E[y_i] = \varphi(x_i;\boldsymbol{\mu}) + \theta_i \ ,$$
$$V[y_i] = \sigma_{y_i}^2 \ .$$

$\boldsymbol{\mu}$ are the parameters of interest in the fit function $\varphi(x;\boldsymbol{\mu})$,

$\boldsymbol{\theta}$ are bias parameters constrained by control measurements $u_i \sim$ Gauss(θ_i, $\sigma_{u,i}$), so that if $\sigma_{u,i}$ are known we have

$$-2\ln L(\boldsymbol{\mu},\boldsymbol{\theta}) = \sum_{i=1}^{N}\left[\frac{(y_i - \varphi(x_i;\boldsymbol{\mu}) - \theta_i)^2}{\sigma_{y_i}^2} + \frac{(u_i - \theta_i)^2}{\sigma_{u_i}^2}\right]$$

Profiling over θ_i with known $\sigma_{u,i}$

Profiling over the bias parameters θ_i for known $\sigma_{u,i}$ gives usual least-squares (BLUE)

$$-2\ln L'(\boldsymbol{\mu}) = \sum_{i=1}^{N} \frac{(y_i - \varphi(x_i;\boldsymbol{\mu}) - u_i)^2}{\sigma_{y_i}^2 + \sigma_{u_i}^2} \equiv \chi^2(\boldsymbol{\mu})$$

Widely used technique for curve fitting in Particle Physics.

Generally in real measurement, $u_i = 0$.

Generalized to case of correlated y_i and u_i by summing statistical and systematic covariance matrices.

Curve fitting with uncertain $\sigma_{u,i}$

Suppose now $\sigma_{u,i}^2$ are adjustable parameters with gamma distributed estimates v_i.

Retaining the θ_i but profiling over $\sigma_{u,i}^2$ gives

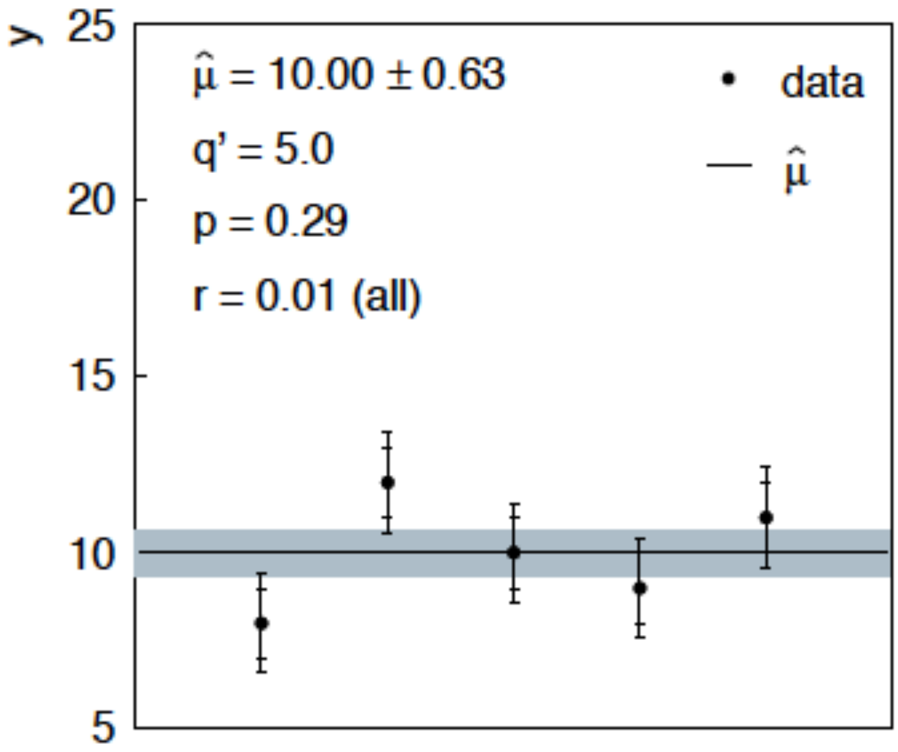
$$-2\ln L'(\boldsymbol{\mu},\boldsymbol{\theta}) = \sum_{i=1}^{N}\left[\frac{(y_i-\varphi(x_i;\boldsymbol{\mu})-\theta_i)^2}{\sigma_{y_i}^2} + \left(1+\frac{1}{2r_i^2}\right)\ln\left(1+2r_i^2\frac{(u_i-\theta_i)^2}{v_i}\right)\right]$$

Profiled values of θ_i from solution to cubic equations

$$\begin{aligned}\theta_i^3 \quad &+ \quad [-2u_i - y_i + \varphi_i]\,\theta_i^2 + \left[\frac{v_i + (1+2r_i^2)\sigma_{y_i}^2}{2r_i^2} + 2u_i(y_i-\varphi_i) + u_i^2\right]\theta_i \\ &+ \quad \left[(\varphi_i - y_i)\left(\frac{v_i}{2r_i^2} + u_i^2\right) - \frac{(1+2r_i^2)\sigma_{y_i}^2 u_i}{2r_i^2}\right] = 0\,, \quad i = 1,\ldots,N\,,\end{aligned}$$

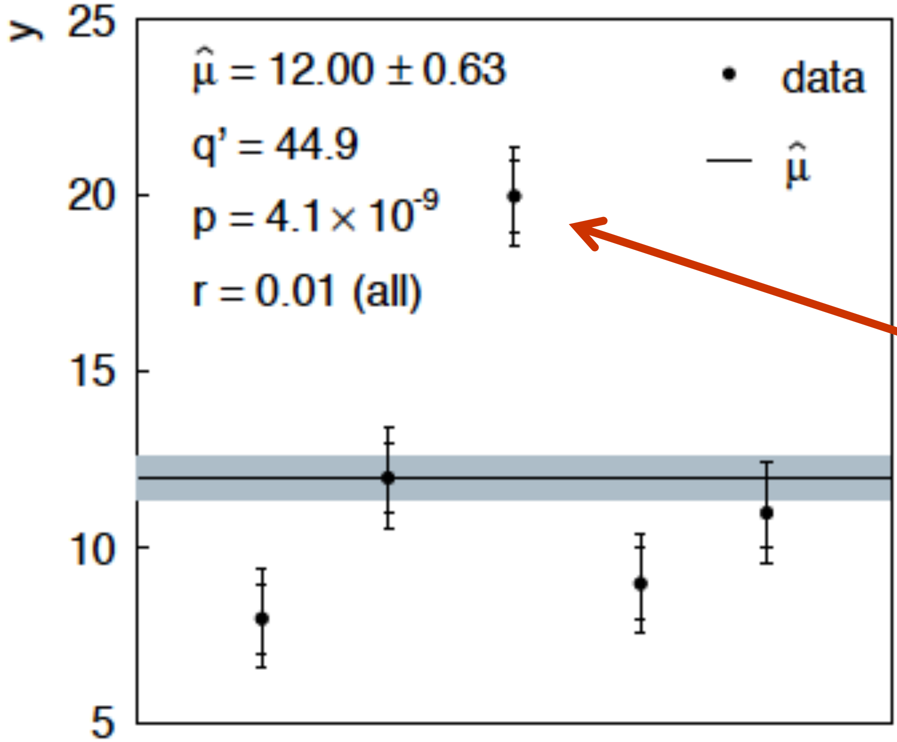
Sensitivity of average to outliers

Suppose we average 5 values, $y = 8, 9, 10, 11, 12$, all with stat. and sys. errors of 1.0, and suppose negligible error on error (here take $r = 0.01$ for all).



Sensitivity of average to outliers (2)

Now suppose the measurement at 10 was actually at 20:

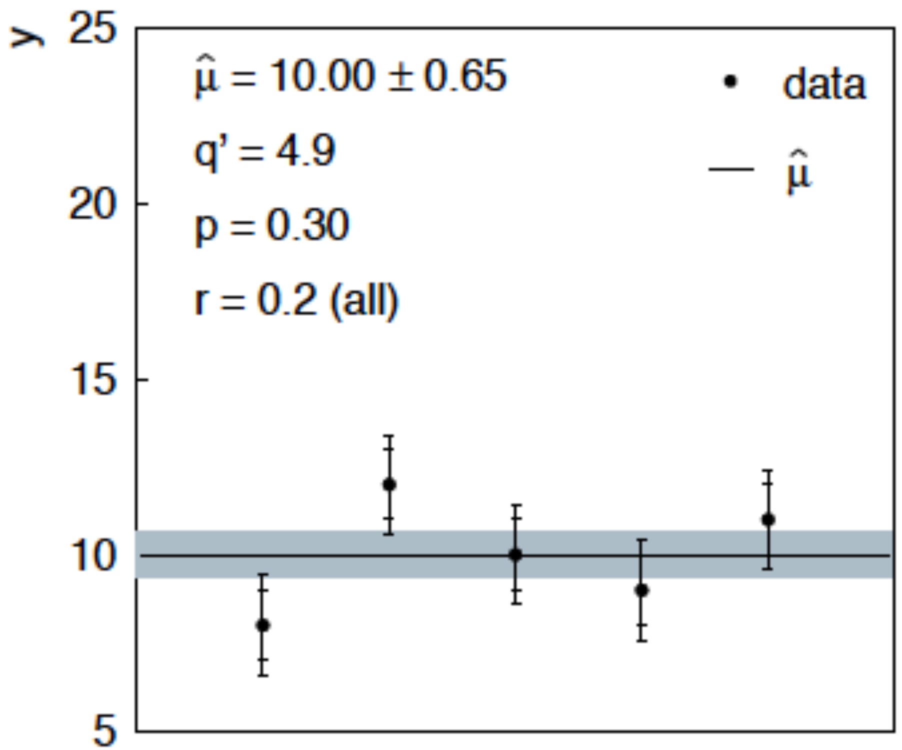


"outlier"

Estimate pulled up to 12.0, size of confidence interval ~unchanged (would be exactly unchanged with $r \to 0$).

Average with all $r = 0.2$

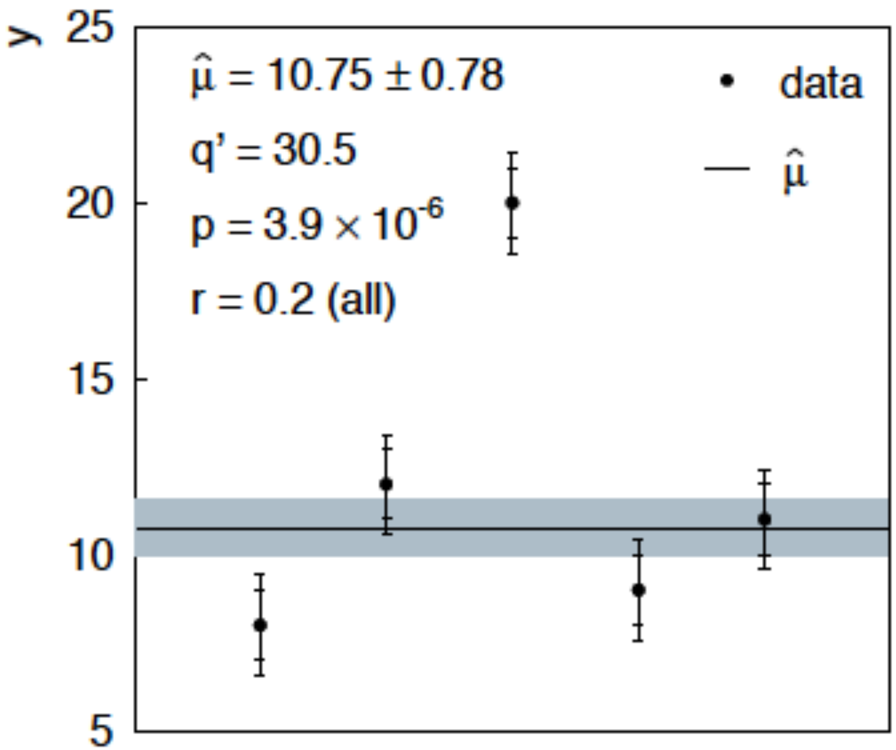
If we assign to each measurement $r = 0.2$,



Estimate still at 10.00, size of interval moves $0.63 \to 0.65$

Average with all $r = 0.2$ with outlier

Same now with the outlier (middle measurement $10 \rightarrow 20$)



Estimate $\rightarrow 10.75$ (outlier pulls much less).

Half-size of interval $\rightarrow 0.78$ (inflated because of bad g.o.f.).

Naive approach to errors on errors

Naively one might think that the error on the error in the previous example could be taken into account conservatively by inflating the systematic errors, i.e.,

$$\sigma_{u_i} \to \sigma_{u_i}(1 + r_i)$$

But this gives

$\hat{\mu} = 10.00 \pm 0.70$ without outlier (middle meas. 10)

$\hat{\mu} = 12.00 \pm 0.70$ with outlier (middle meas. 20)

So the sensitivity to the outlier is not reduced and the size of the confidence interval is still independent of goodness of fit.

Conclusions on errors on errors

Gamma model for variance estimates gives confidence intervals that increase in size when the data are internally inconsistent, and gives decreased sensitivity to outliers.

Method assumes that meaningful r_i values can be assigned and is valuable when systematic errors are not well known but enough "expert opinion" is available to do so.

Equivalence with Student's t model, $\nu = 1/2r^2$ degrees of freedom.

Simple profile likelihood – quadratic terms replaced by logs:

$$\boxed{\frac{(u_i - \theta_i)^2}{\sigma_{u_i}^2} \quad \rightarrow \quad \left(1 + \frac{1}{2r_i^2}\right) \ln\left[1 + 2r_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right]}$$

Discussion / Conclusions on combinations

The fundamental approach to combinations is to construct a likelihood that represents all the measurements.

Need to identify common nuisance parameters that are common.

Sometimes not enough information available to reconstruct a meaningful likelihood (only have *p*-values, confidence intervals,...)

> This can be a difficult situation – best to try to cobble together some approximation to the likelihood; include additional nuisance parameters as appropriate.

Many aspects not treated due to time, e.g., Bayesian methods; see e.g. G. Cowan, arXiv:1012.3589.

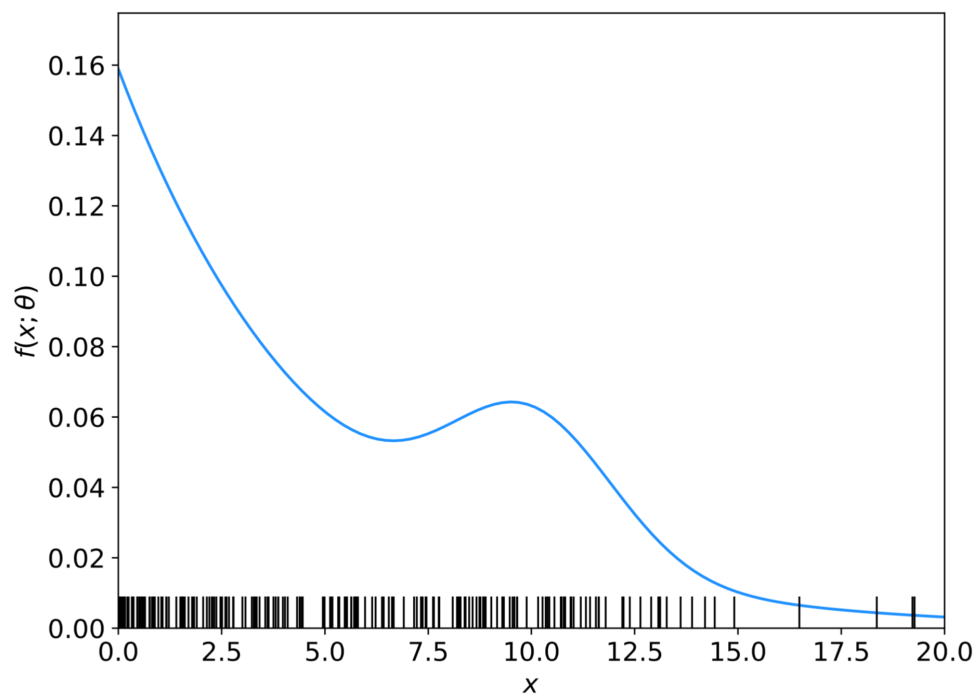
Extra slides

Example with nuisance parameters

Suppose x follows the pdf

$$f(x;\theta,\xi) = \theta\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2} + (1-\theta)\frac{1}{\xi}e^{-x/\xi}$$

and we have an i.i.d. data sample:

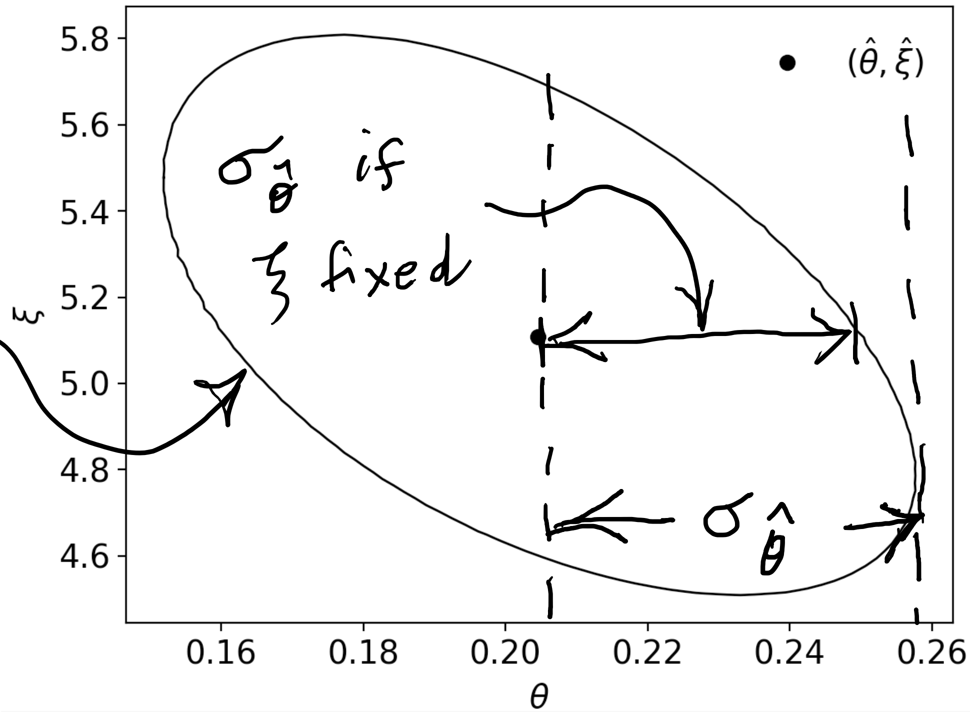


Goal: estimate parameter of interest θ; the rest are nuisance parameters.

Example with nuisance parameters (2)

Standard deviation of estimator $\hat{\theta}$ from tangents to contour

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



Would be smaller if nuisance parameter ξ were to be exactly known.

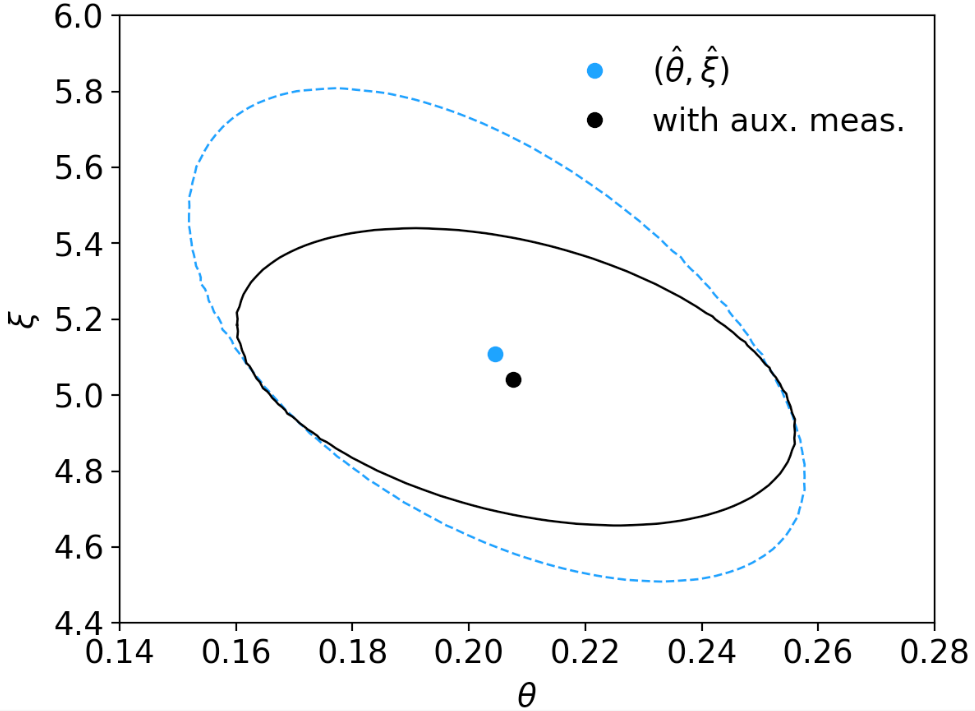
Free	Fixed	sigma_thetaHat
theta	mu, sigma, xi	0.044535
theta, xi	mu, sigma	0.052736
theta, xi, sigma	mu	0.064456
theta, xi, sigma, mu	--	0.085786

Presence of nuisance params. inflates uncertainty on param. of interest

Auxiliary measurement to constrain nuisance param.

So often include an auxiliary measurement that constrains ξ, e.g., suppose $u \sim \mathrm{Gauss}(\xi, \sigma_u)$.

$$L(\theta, \xi) = \frac{1}{\sqrt{2\pi}\sigma_u} e^{-(u-\xi)^2/2\sigma_u^2} \prod_{i=1}^{n} \left[\theta \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} + (1-\theta) \frac{1}{\xi} e^{-x_i/\xi} \right]$$



The aux. measurement u compresses the contour in both the ξ and θ directions and thus decreases the uncertainty on the estimate of θ.

Motivation for gamma model

If one were to have n independent observations $u_1,..,u_n$, with all $u \sim$ Gauss(θ, σ_u^2), and we use the sample variance

$$v = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \overline{u})^2$$

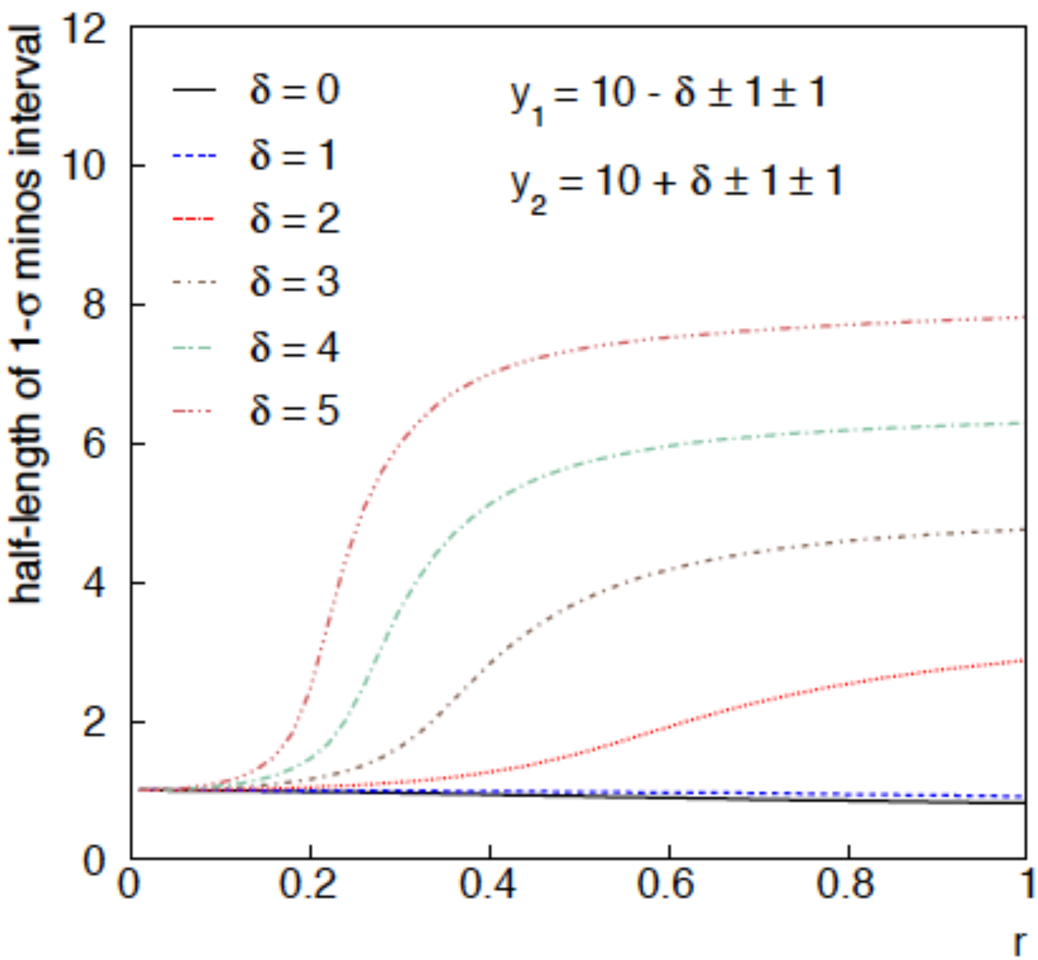
to estimate σ_u^2, then $(n\text{-}1)v/\sigma_u^2$ follows a chi-square distribution for $n-1$ degrees of freedom, which is a special case of the gamma distribution ($\alpha = n/2$, $\beta = 1/2$). (In general one doesn't have a sample of u_i values, but if this were to be how v was estimated, the gamma model would follow.)

Furthermore choice of the gamma distribution for v allows one to profile over the nuisance parameters σ_u^2 in closed form and leads to a simple profile likelihood.

Example: average of two measurements

MINOS interval (= approx. confidence interval) based on

$$\ln L'(\mu) = \ln L'(\hat{\mu}) - Q_\alpha/2 \quad \text{with} \quad Q_\alpha = F^{-1}_{\chi^2}(1-\alpha; n)$$



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

Can quantify goodness of fit with statistic

$$q = -2 \ln \frac{L'(\hat{\boldsymbol{\mu}}, \hat{\hat{\boldsymbol{\theta}}})}{L'(\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\theta}})}$$

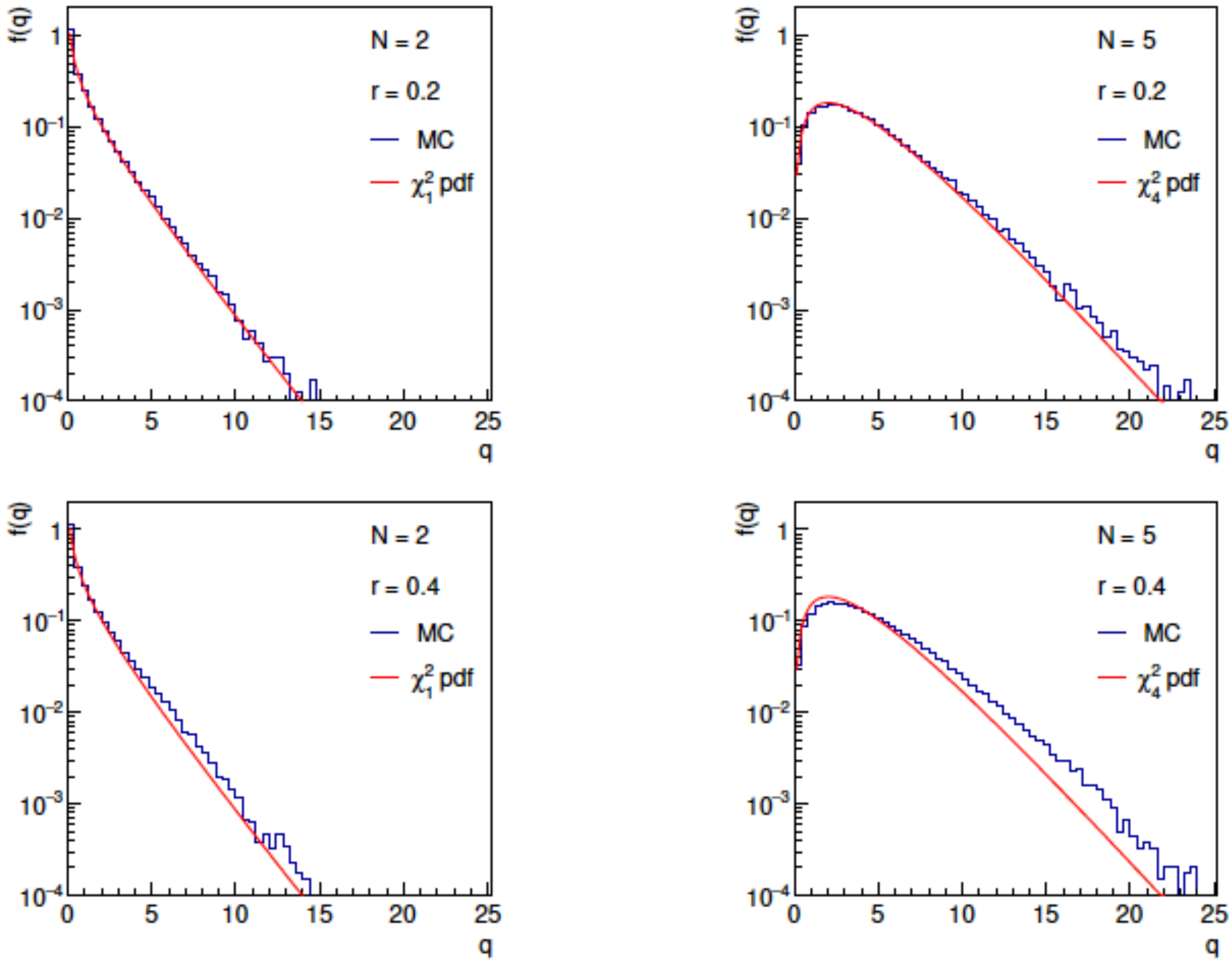
$$= \min_{\boldsymbol{\mu}, \boldsymbol{\theta}} \sum_{i=1}^{N} \left[\frac{(y_i - \varphi(x_i; \boldsymbol{\mu}) - \theta_i)^2}{\sigma_{y_i}^2} + \left(1 + \frac{1}{2r_i^2}\right) \ln\left(1 + 2r_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right)\right]$$

where $L'(\boldsymbol{\varphi}, \boldsymbol{\theta})$ has an adjustable φ_i for each y_i (the saturated model).

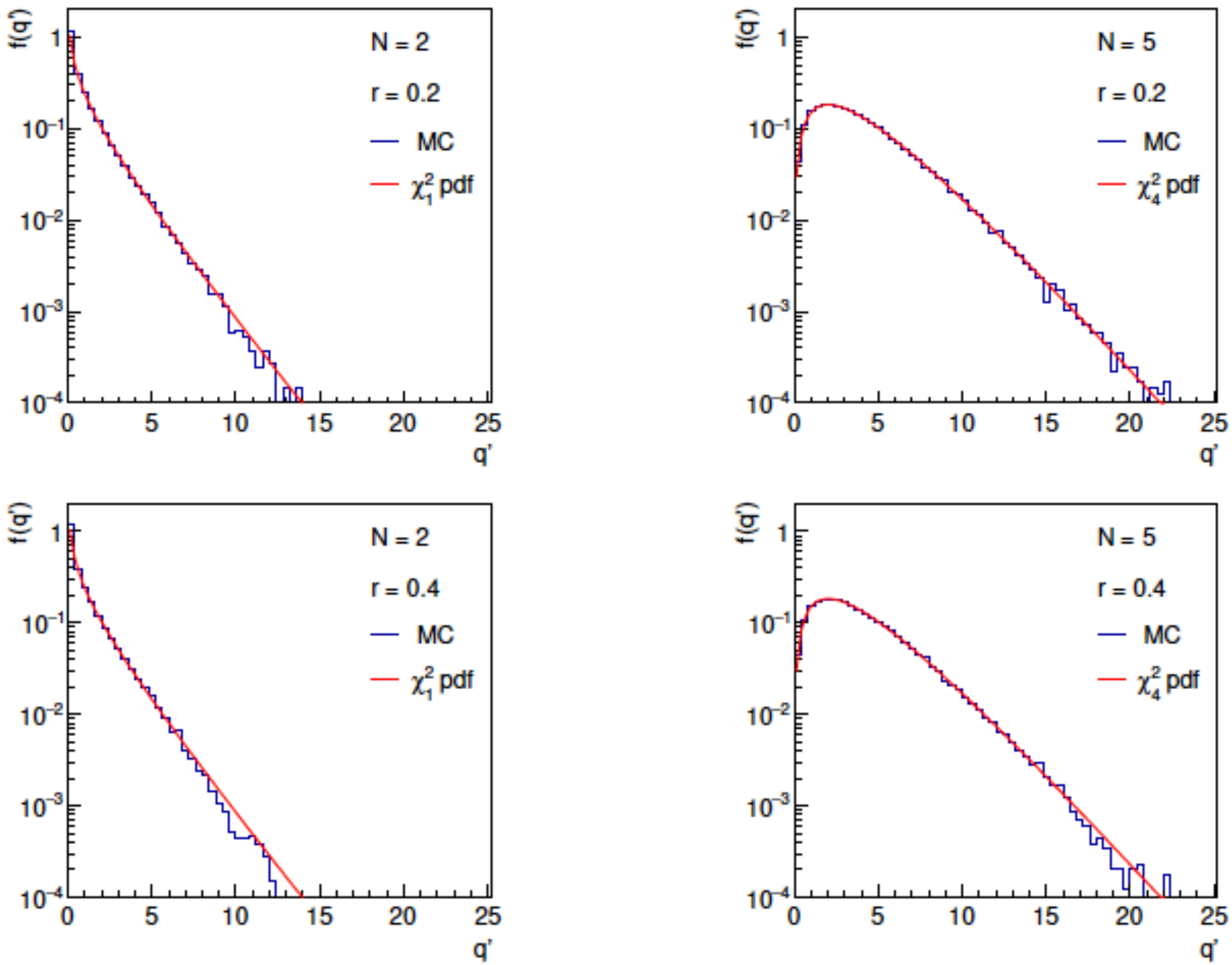
Asymptotically should have $q \sim$ chi-squared($N-M$).

For increasing r_i, may need Bartlett correction or MC.

Distributions of q



Distributions of Bartlett-corrected q'



Correlated uncertainties

The phrase "correlated uncertainties" usually means that a single nuisance parameter affects the distribution (e.g., the mean) of more than one measurement.

For example, consider measurements $\boldsymbol{y}$, parameters of interest $\boldsymbol{\mu}$, nuisance parameters $\boldsymbol{\theta}$ with

$$E[y_i] = \varphi_i(\boldsymbol{\mu}, \boldsymbol{\theta}) \approx \varphi_i(\boldsymbol{\mu}) + \sum_{j=1}^{N} R_{ij}\theta_j$$

That is, the θ_i are defined here as contributing to a bias and the (known) factors R_{ij} determine how much θ_j affects y_i.

As before suppose one has independent control measurements $u_i \sim$ Gauss(θ_i, σ_{ui}).

Correlated uncertainties (2)

The total bias of y_i can be defined as $$b_i = \sum_{j=1}^{N} R_{ij} \theta_j$$

which can be estimated with $$\hat{b}_i = \sum_{j=1}^{N} R_{ij} u_j$$

These estimators are correlated having covariance

$$U_{ij} = \mathrm{cov}[\hat{b}_i, \hat{b}_j] = \sum_{k=1}^{N} R_{ik} R_{jk} V[u_k]$$

In this sense the present method treats “correlated uncertainties”, i.e., the control measurements u_i are independent, but nuisance parameters affect multiple measurements, and thus bias estimates are correlated.

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:

$$V[\hat{\mu}] = \frac{\phi}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

Bayesian approach

G. Cowan, *Bayesian Statistical Methods for Parton Analyses*, in *Proceedings of the 14th International Workshop on Deep Inelastic Scattering (DIS2006)*, M. Kuze, K. Nagano, and K. Tokushuku (eds.), Tsukuba, 2006.

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

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

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

control variable parameters bias

Frequentist approach: $V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

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

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:

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

→ 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!)

Bayesian approach with 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 $\pi_s(s)$

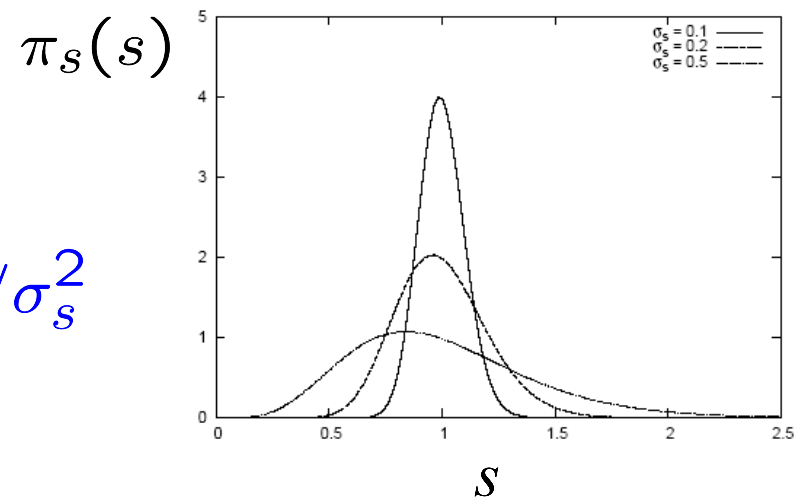
A simple unimodal probability density for $0 < s < 1$ with adjustable mean and variance is the Gamma distribution:

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

mean $= b/a$

variance $= b/a^2$

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

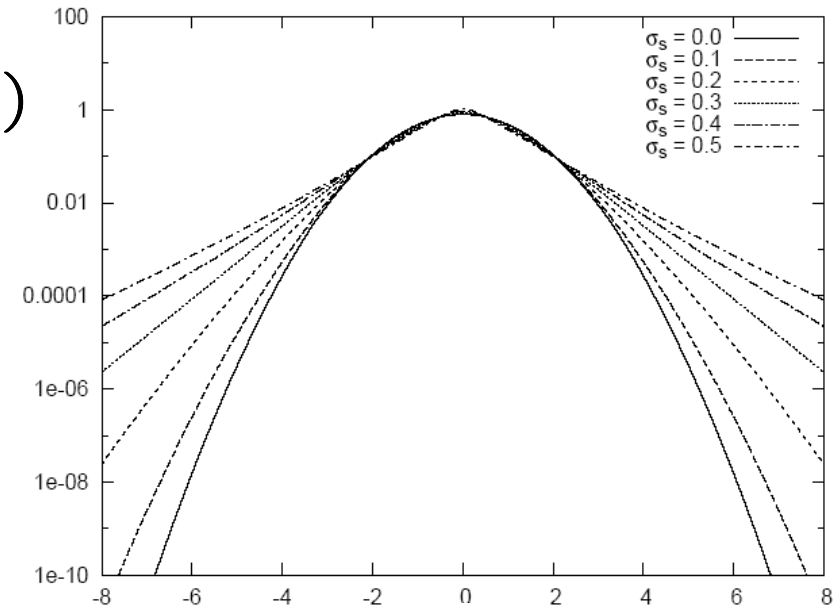


In fact if we took $\pi_s(s) \sim$ *inverse Gamma*, we could find $\pi_b(b)$ in closed form (cf. D'Agostini, Dose, von Linden). But Gamma seems more natural & numerical treatment not too painful.

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

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

$\pi_b(b)$



b

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

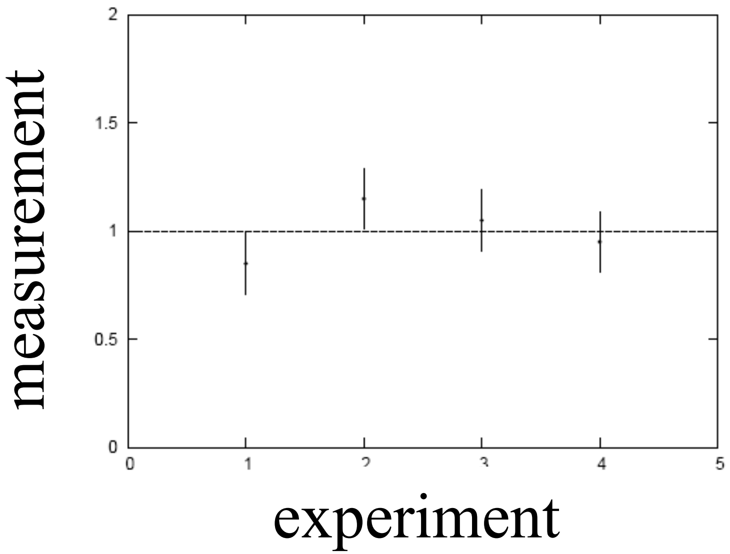
$\sigma_s = 0.5$ $P(|b| > 4\sigma_{\text{sys}}) = 0.65\%$

A simple test

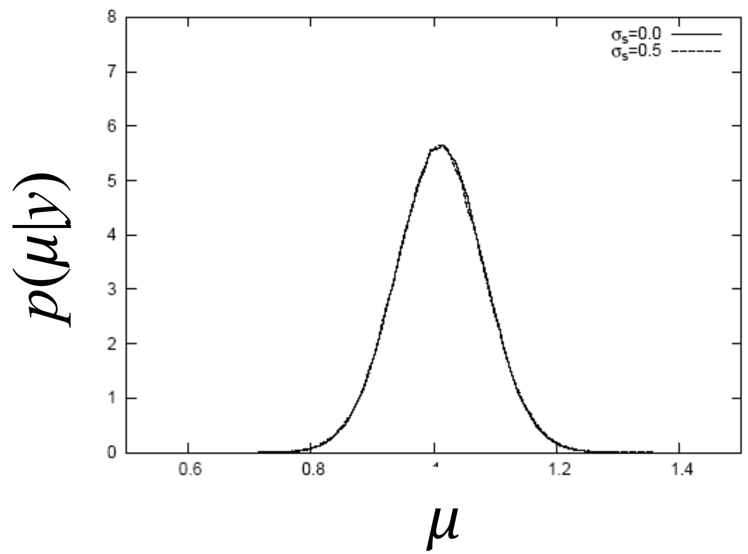
Suppose a fit effectively averages four measurements.

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

Case #1: data appear compatible



Posterior $p(\mu|y)$:



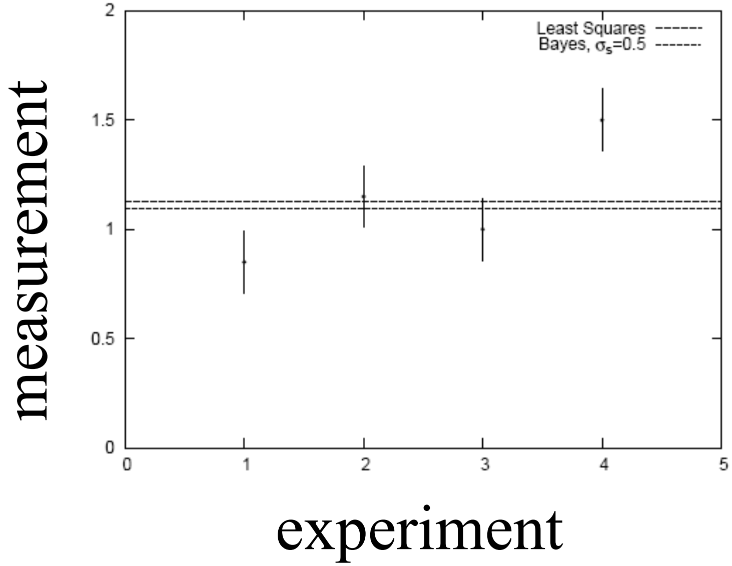
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

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

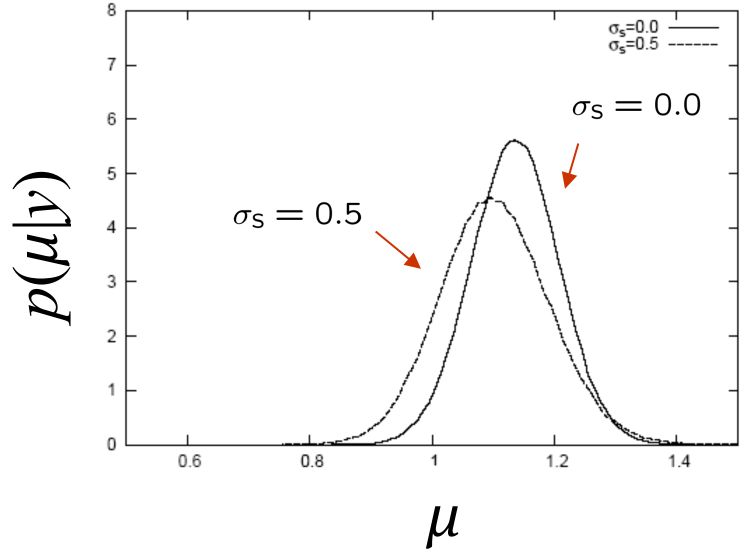
$$\sigma_s = 0.5 : \quad \hat{\mu} = 1.000 \pm 0.072$$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



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

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

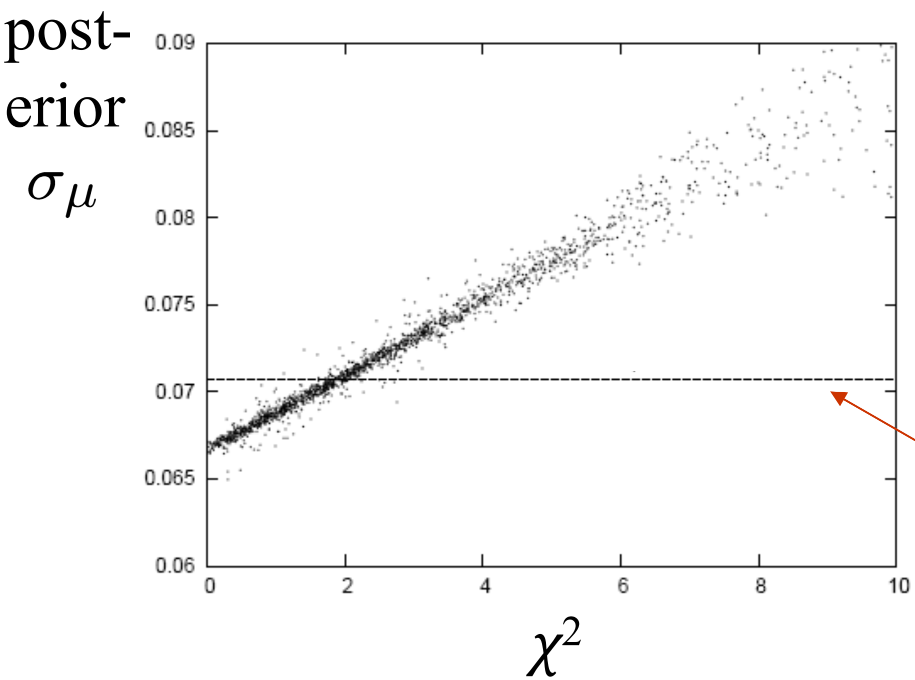
→ Bayesian fit less sensitive to outlier. See also

G. D'Agostini, *Sceptical combination of experimental results: General considerations and application to epsilon-prime/epsilon*, arXiv:hep-ex/9910036 (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