Statistical Data Analysis 2025/26
Lecture Week 11

London Postgraduate Lectures on Particle Physics

University of London MSc/MSci course PH4515

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

Course web page via RHUL moodle (PH4515) and also

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

Statistical Data Analysis
Lecture 11-1

- Bayesian model selection
- Bayes factors

Bayesian model selection

Fundamentally the probability of a hypothesis H_i in the Bayesian approach is given by its posterior probability given the data: $P(H_i|\boldsymbol{x})$.

Finding this requires assignment of prior probabilities to all hypotheses that are considered.

We can give the posterior *odds* (ratio of probabilities) for any pair of hypotheses H_i and H_j (use Bayes' theorem; factors of $P(\boldsymbol{x})$ cancel):

$$\frac{P(H_i|\mathbf{x})}{P(H_j|\mathbf{x})} = \frac{P(\mathbf{x}|H_i)}{P(\mathbf{x}|H_j)} \frac{\pi(H_i)}{\pi(H_j)}$$

posterior odds

Bayes factor

prior odds

See: Kass and Raftery, *Bayes Factors*, J. Am Stat. Assoc 90 (1995) 773.

The Bayes factor

The Bayes factor B_{ij} is the likelihood ratio of the two hypotheses:

$$B_{ij} = \frac{P(\mathbf{x}|H_i)}{P(\mathbf{x}|H_j)}$$

= posterior odds if one takes prior odds equal to one.

The Bayes factor is regarded as measuring the weight of evidence of the data in support of H_i over H_j. and can be used much like a p-value (or Z value).

The Jeffreys scale, analogous to the 5σ rule in Particle Physics:

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Marginal likelihood (evidence)

If the model H_i contains internal parameters $\boldsymbol{\theta}_i$, then these must be characterized by a prior pdf $\pi_i(\boldsymbol{\theta}_i|H_i)$ and marginalized:

$$P(\mathbf{x}|H_i) = \int P(\mathbf{x}, \boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i = \int P(\mathbf{x}|H_i, \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i$$

This is called the "marginal likelihood" or "evidence" of H_i.

It is independent of the overall prior probability of H_i

$$\pi(H_i) = \int \pi(H_i, \boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i$$

but it depends on the prior pdf for the model's internal parameters $\boldsymbol{\theta}_i$:

$$\pi_i(\boldsymbol{\theta}_i|H_i) = \frac{\pi(H_i, \boldsymbol{\theta}_i)}{\pi(H_i)}$$

Bayes factor for models with internal parameters

The Bayes factor is thus the ratio of marginal likelihoods for the two models:

$$B_{ij} = \frac{P(\mathbf{x}|H_i)}{P(\mathbf{x}|H_j)} = \frac{\int P(\mathbf{x}|H_i, \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{\int P(\mathbf{x}|H_j, \boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j}$$

Simplifying the notation, the numerator and denominator are both of the form

$$m = \int P(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

For high-dimensional $\boldsymbol{\theta}$ these integrals can be very difficult to compute (more on this later).

Priors for Bayes factors

Prior pdfs for the marginal likelihoods used in Bayes factors cannot be improper, i.e., they cannot be defined only up to an arbitrary normalization constant, in which case B_{ij} would not be well defined.

Suppose we try to take a "non-informative" prior to be constant out to some large cut-off, in the hope that the Bayes factor will decouple from it:

In such cases we find that the Bayes factor remains sensitive to the cut-off even for $a \to \infty$.

So all priors used for Bayes factors must reflect a meaningful degrees of uncertainty about the parameters.

Bayes factor for Poisson counting experiment

Suppose $n \sim \text{Poisson}(s + b)$ with b known. We want to compare

$$H_0 \; : \; s = 0 \; ,$$

$$H_1 \; : \; s > 0 \; .$$

The likelihoods of H_0 and H_1 are

$$L(n|H_0) = \frac{b^n}{n!} e^{-b}$$

$$L(n|s, H_1) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Bayes factor for Poisson counting experiment (2)

Suppose the prior pdf for the parameter s in H_1 is:

$$\pi(s|H_1) = \frac{1}{s_{\max}} \qquad (0 \leq s \leq s_{\max})$$

The posterior probability for s given n is, assuming H_1,

$$p(s|n, H_1) = \frac{L(n|s, H_1)\pi(s|H_1)}{\int L(n|s, H_1)\pi(s|H_1)\, ds}$$

$$= \frac{(s+b)^n e^{-(s+b)}}{\int_0^{s_{\max}} (s+b)^n e^{-(s+b)}\, ds} \qquad (0 \leq s \leq s_{\max})$$

$$= \frac{(s+b)^n e^{-(s+b)}}{\gamma(n+1, s_{\max}+b) - \gamma(n+1, b)}$$

γ = lower incomplete gamma function

Bayes factor for Poisson counting experiment (3)

In the limit $s_{\max} \to \infty$ this goes to

$$p(s|n, H_1) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(n+1) - \gamma(n+1, b)}$$

where $$\gamma(a, x) = \int_0^x t^{a-1} e^{-t}\, dt$$

is the lower incomplete gamma function.

Thus the posterior pdf for s given n under assumption of H_1 decouples from $s_{\max}$ in the limit $s_{\max} \to \infty$, and hence we can use this limiting case e.g. for finding an upper limit (credibility interval) for s.

Bayes factor for Poisson counting experiment (4)

The hypothesis H_0 has no internal parameters so its marginal likelihood is simply $m_0 = L(n|\, H_0)$.

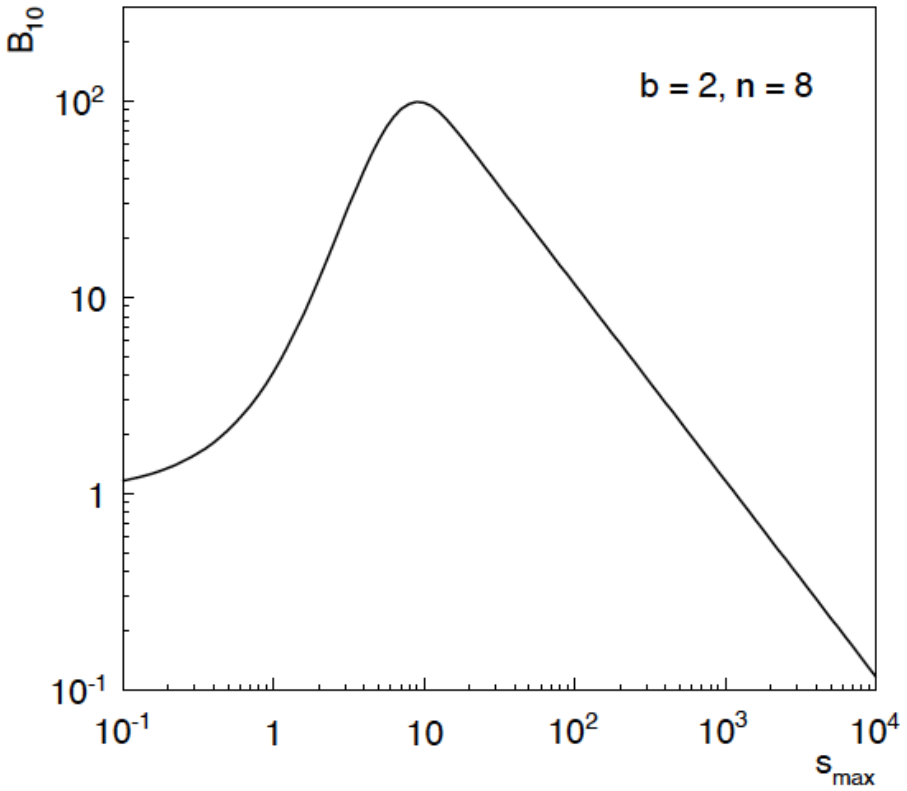
The marginal likelihood of H_1 is

$$
\begin{aligned}
m_1 &= \int L(n|s, H_1)\pi(s|H_1)\, ds \\
&= \frac{1}{n! s_{\max}} \int_0^{s_{\max}} (s+b)^n e^{-(s+b)}\, ds \\
&= \frac{1}{n! s_{\max}} \left(\gamma(n+1, s_{\max}+b) - \gamma(n+1, b)\right)
\end{aligned}
$$

Bayes factor for Poisson counting experiment (5)

So the Bayes factor is

$$B_{10} = \frac{m_1}{m_0} = \frac{1}{s_{\max}} \frac{\gamma(n+1, s_{\max}+b) - \gamma(n+1, b)}{b^n e^{-b}}$$



Example: $b = 2, n = 8$

As $s_{\max}$ increases the data start to favour H_1.

As $s_{\max}$ increases further, the larger volume of H_1's parameter space penalizes it (Ockham's razor).

Bayes Factors and Ockham's Razor

We want to compute a Bayes factor $B_{10} = \dfrac{\int P_1(\boldsymbol{x}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{P_0(\boldsymbol{x})}$

Suppose the full parameter space $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is very large, e.g., the set of coefficients of a high-order polynomial, and that the prior $\pi_1(\boldsymbol{\theta})$ is approximately constant over Θ.

Then $P(\boldsymbol{x}|\boldsymbol{\theta})$ is low for most $\boldsymbol{\theta}$ except in a region not too far from the best fit, and therefore $\int P_1(\boldsymbol{x}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta})\, d\boldsymbol{\theta} \rightarrow$ small

This is the "Ockham's razor" of the Bayes factor.

Models with a high degree of complexity (i.e., a large parameter space) are automatically disfavoured.

Caveat regarding Ockham's Razor

Suppose for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\rm NS}$ the measurement is not sensitive to the difference between H_0 and H_1, i.e,. $P_1(\boldsymbol{x}|\boldsymbol{\theta}) \approx P_0(\boldsymbol{x})$ and in $\boldsymbol{\Theta}_{\rm S} = \boldsymbol{\Theta} - \boldsymbol{\Theta}_{\rm NS}$ H_1 is strongly disfavoured. Then,

$$B_{10} = \frac{\int_{\Theta_{\rm NS}} P_1(\boldsymbol{x}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta})\, d\boldsymbol{\theta} + \int_{\Theta_{\rm S}} P_1(\boldsymbol{x}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{P_0(\boldsymbol{x})}$$

$$\approx \int_{\Theta_{\rm NS}} \pi_1(\boldsymbol{\theta})\, d\boldsymbol{\theta} + \frac{\int_{\Theta_{\rm S}} P_1(\boldsymbol{x}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{P_0(\boldsymbol{x})} \approx f_{\rm NS}$$

(fraction of prior in $\Theta_{\rm NS}$.)

So a potentially large volume of parameter space does not decrease the Bayes factor if the region provides no sensitivity to H_1

E.g., the Minimal Supersymmetric Standard Model has more than 100 parameters, and naively one would think its Bayes factor relative to the SM would be small. But current data are not sensitive to most parts of its parameter space, so $B_{10} \to f_{\rm NS} \to 1$.

Ockham

Statistical Data Analysis
Lecture 11-2

- Numerical determination of Bayes factors

Numerical determination of Bayes factors

Both numerator and denominator of B_{ij} are of the form

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})\, d\vec{\theta}$$ ← 'marginal likelihood'

Various ways to compute these, e.g., using sampling of the posterior pdf (which we can do with MCMC).

- Harmonic Mean (and improvements)
- Importance sampling
- Parallel tempering (~thermodynamic integration)
- Nested Sampling (MultiNest), ...

Kass and Raftery, *Bayes Factors*, J. Am. Stat. Assoc. 90 (1995) 773-795.

Cong Han and Bradley Carlin, *Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review*, J. Am. Stat. Assoc. 96 (2001) 1122-1132.

Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press, 2005.

Harmonic mean estimator

E.g., consider only one model and write Bayes theorem as:

$$\frac{\pi(\boldsymbol{\theta})}{m} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})}$$

$\pi(\boldsymbol{\theta})$ is normalized to unity so integrate both sides,

posterior expectation

$$m^{-1} = \int \frac{1}{L(\mathbf{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p[1/L]$$

Therefore sample $\boldsymbol{\theta}$ from the posterior via MCMC and estimate m with one over the average of $1/L$ (the harmonic mean of L).

M.A. Newton and A.E. Raftery, *Approximate Bayesian Inference by the Weighted Likelihood Bootstrap*, Journal of the Royal Statistical Society B 56 (1994) 3-48.

Called the "worst Monte Carlo method ever"

https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/

Improvements to harmonic mean estimator

The harmonic mean estimator is numerically very unstable; formally infinite variance (!). A variant (cf. Gelfand and Dey):

Rearrange Bayes thm; multiply both sides by arbitrary pdf $f(\boldsymbol{\theta})$:

$$\frac{f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{f(\boldsymbol{\theta})}{m}$$

Integrate over $\boldsymbol{\theta}$:

$$m^{-1} = \int \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p\left[\frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right]$$

Improved convergence if tails of $f(\boldsymbol{\theta})$ fall off similar to $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

Note harmonic mean estimator is special case $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$.

A.E. Gelfand and D.K. Dey, *Bayesian model choice: asymptotics and exact calculations*, Journal of the Royal Statistical Society B 56 (1994) 501-514.

Adaptive Harmonic Mean Integration

A. Caldwell et al., International Journal of Modern Physics A Vol. 35, No. 24 (2020) 2050142

Want to compute $I \equiv \int_{\Omega} f(\lambda) d\lambda$ (Ω = support of f)

E.g. $f(\lambda) = L(\lambda)\,\pi(\lambda)$ = unnormalized target density; we can sample from this with MCMC.

Define integral over subvolume Δ of Ω with volume V_Δ

$$I_\Delta \equiv \int_{\Delta} f(\lambda) d\lambda \qquad r \equiv \frac{I_\Delta}{I}$$

Adaptive Harmonic Mean Integration (2)

If $f(\lambda)$ not small in Δ, then we can find I_Δ from harmonic mean:

$$E\left[\frac{1}{f(\lambda)}\right]_{\lambda\in\Delta} = \int_\Delta \frac{1}{f(\lambda)}\frac{f(\lambda)}{I_\Delta}\,d\lambda = \frac{1}{I_\Delta}\int_\Delta d\lambda = \frac{V_\Delta}{I_\Delta} \approx \frac{1}{N_\Delta}\sum_{\lambda_i\in\Delta}\frac{1}{f(\lambda_i)}$$

Sample λ from $f(\lambda)$ using MCMC, estimate $r = I_\Delta/I$ with fraction of points found in Δ:

$$\hat{r} = \frac{N_\Delta}{N_\Omega}$$

Use these to estimate I:

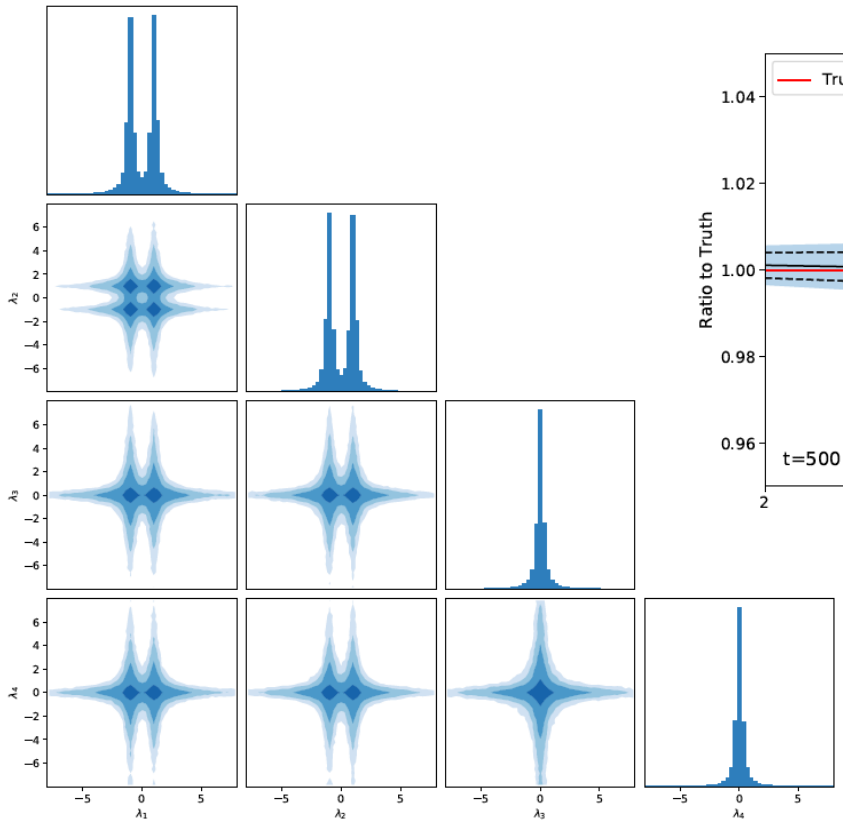
$$\hat{I} = \frac{\hat{I}_\Delta}{\hat{r}} = \frac{N_\Omega V_\Delta}{\sum_{\lambda_i\in\Delta}\frac{1}{f(\lambda_i)}}$$

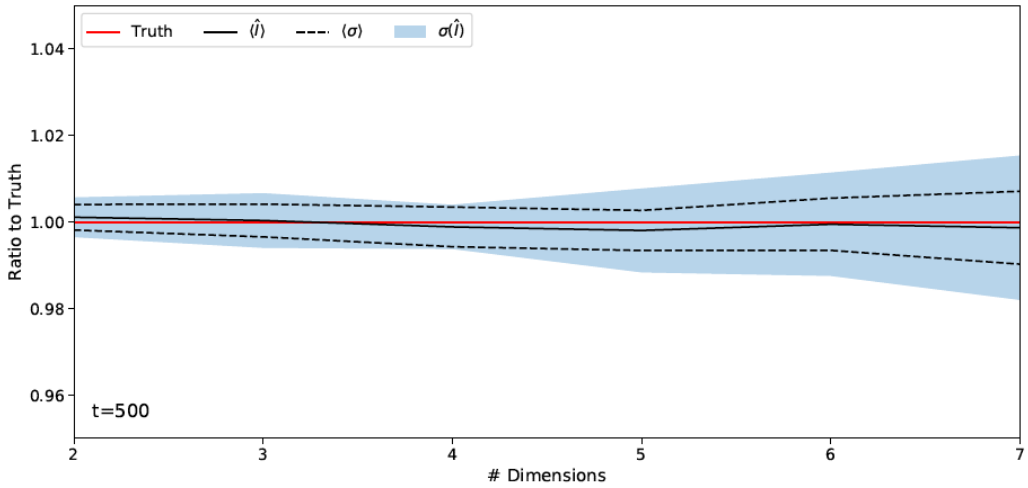
"The task of estimating our integral, therefore reduces to choosing one or several subspaces Δ — typically small regions around local modes of $f(\lambda)$. The full space Ω over which the integration ought to be performed can be large or even infinite, while this does not affect the outcome of our integral estimate."

A. Caldwell et al., IJMP A Vol. 35, No. 24 (2020) 2050142

Adaptive Harmonic Mean Integration (3)

Testing AHMI with multimodal multidimensional Cauchy pdf





Challenging pdf because of long tails.

Good results for up to 7 dimensions for MCMC sample size of 10^6.

Software: Bayesian Analysis Toolkit

https://github.com/bat/BAT.jl

BAT
BAYESIAN ANALYSIS TOOLKIT

Importance sampling

Need pdf $f(\boldsymbol{\theta})$ which we can evaluate at arbitrary $\boldsymbol{\theta}$ and also sample with MC.

The marginal likelihood can be written

$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E_f\left[\frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})}\right]$$

Sample $\boldsymbol{\theta} \sim f(\boldsymbol{\theta})$, compute average of $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/f(\boldsymbol{\theta})$.

Best convergence when $f(\boldsymbol{\theta})$ approximates shape of $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Use for $f(\boldsymbol{\theta})$ e.g. multivariate Gaussian with mean and covariance estimated from posterior.

Nested sampling

J. Skilling, Bayesian Analysis, No. 4, pp. 833-860 (2006)

We want to compute $Z = \text{evidence} = \int L\, dX$ $\qquad L = L(\theta)$

$dX = \pi(\theta) d\theta$

Can add up portions of X (equivalently, θ) space in any order. Use

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta) d\theta$$ = portion of prior for which $L(\theta) > \lambda$.
(Low λ means X near 1, all of θ space included.)

Write inverse function as $L(X(\lambda)) \equiv \lambda$ so that the desired result is

$$Z = \int_0^1 L(X)\, dX$$

Elements of θ space are sorted by decreasing likelihood.

J. Skilling, Bayesian Analysis, No. 4, pp. 833-860 (2006)

Nested sampling (2)

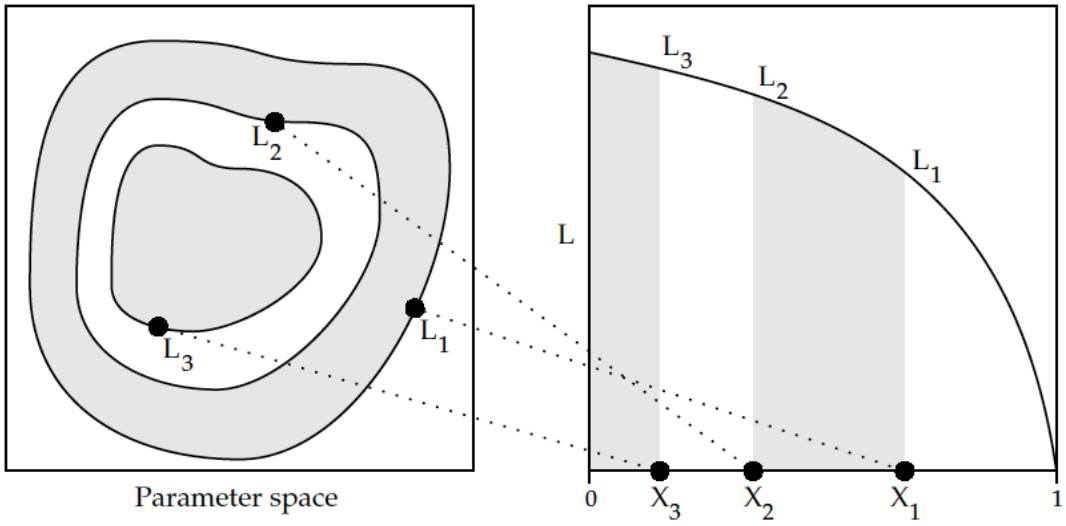


Figure 3: Nested likelihood contours are sorted to enclosed prior mass X.

The evidence Z is the area under the curve of $L(X)$.

Computational challenge is to sample $\boldsymbol{\theta}$ space from prior subject to constraint $L(\boldsymbol{\theta}) > \lambda$. Software: MultiNest

Farhan Feroz, Mike Hobson, Mon. Not. Roy. Astron. Soc., 384, 2, 449-463 (2008); arXiv:0704.3704,

F. Feroz, M.P. Hobson, M. Bridges, Mon. Not. Roy. Astron. Soc. 398: 1601-1614,2009; arXiv:0809.3437

F. Feroz, M.P. Hobson, E. Cameron, A.N. Pettitt, arXiv:1306.2144

Statistical Data Analysis
Lecture 11-3

- Statistical models with uncertain error parameters (part 1)

"Errors on Errors"

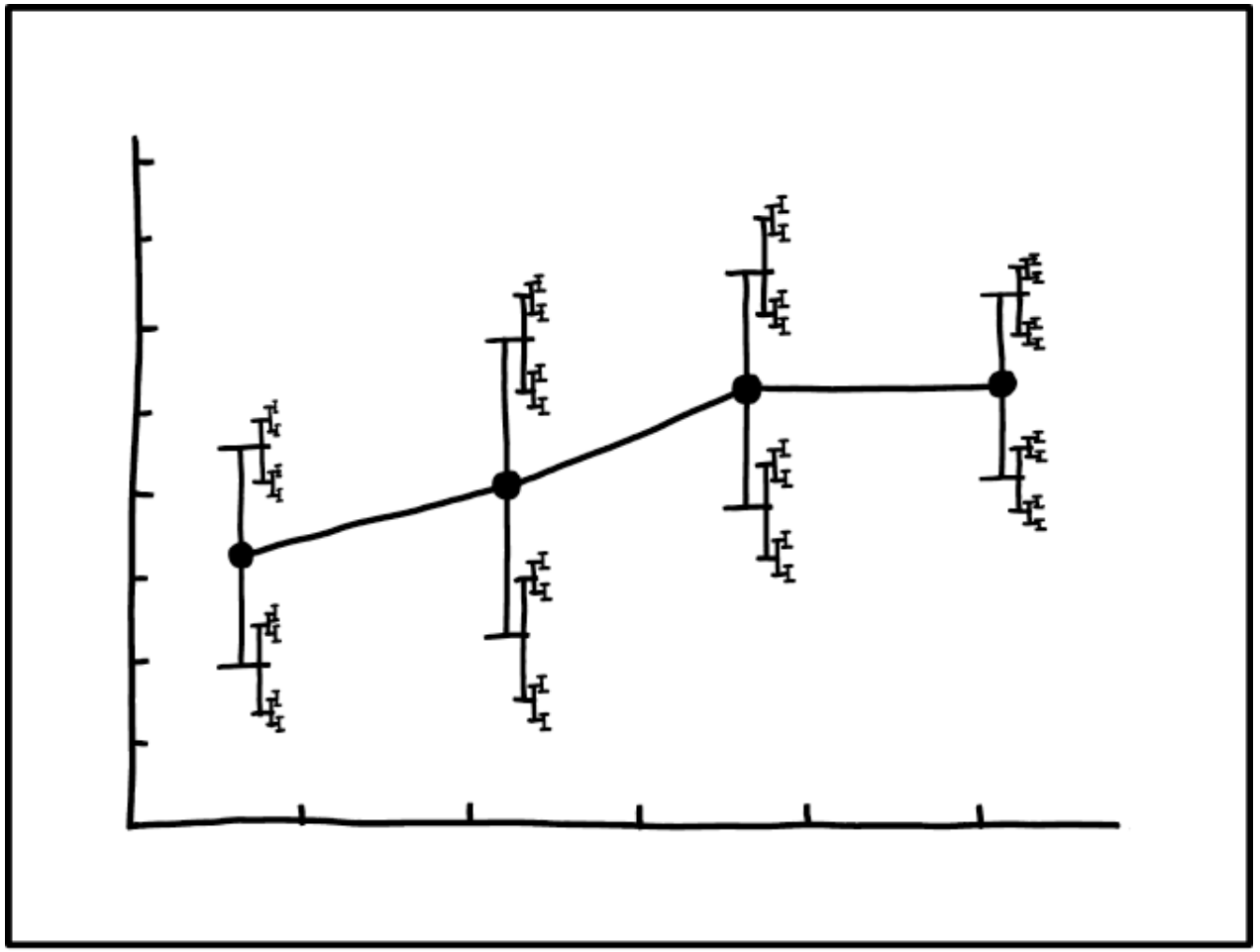
The final part of the lectures constitutes a "special seminar" on material that will not be on the exam. Details in:

G. Cowan, *Statistical Models with Uncertain Error Parameters*, Eur. Phys. J. C (2019) 79:133, arXiv:1809.05778

G. Cowan, *Effect of Systematic Uncertainty Estimation on the Muon g − 2 Anomaly*, EPJ Web of Conferences 258, 09002 (2022), arXiv:2107.02652

E. Canonero, A. Brazzale and G. Cowan, *Higher-order asymptotic corrections and their application to the Gamma Variance Model*, Eur. Phys. J. C (2023) 83:1100, arXiv:2304.10574

E. Canonero and G. Cowan, *Correlated Systematic Uncertainties and Errors-on-Errors in Measurement Combinations: Methodology and Application to the 7-8 TeV ATLAS-CMS Top Quark Mass Combination*, Eur. Phys. J. C (2025) 85: 156, arXiv:2407.05322



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

Errors on errors: the issue

The method of least squares requires the standard deviations of the measured quantities, but often these are poorly known.

The uncertainty (e.g. confidence interval) of an LS average does not reflect goodness of fit:

LS average of 9 ± 1 and 11 ± 1 is 10 ± 0.71

LS average of 5 ± 1 and 15 ± 1 is 10 ± 0.71

LS estimators are equivalent to maximum-likelihood assuming Gaussian distributed measurements; but the tails of a Gaussian fall off very fast, not always an appropriate model.

$\rightarrow$ Outliers in LS average have very large influence.

Solution: incorporate the uncertainty in the standard deviations of the measurements into the analysis.

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

Often the θ_i could represent a systematic bias and its best estimate u_i in the real measurement is zero.

The $\sigma_{u,i}$ are the corresponding "systematic errors".

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 mean and width for $f(v)$:

$$E[v] = \sigma_u^2 = \alpha/\beta,$$

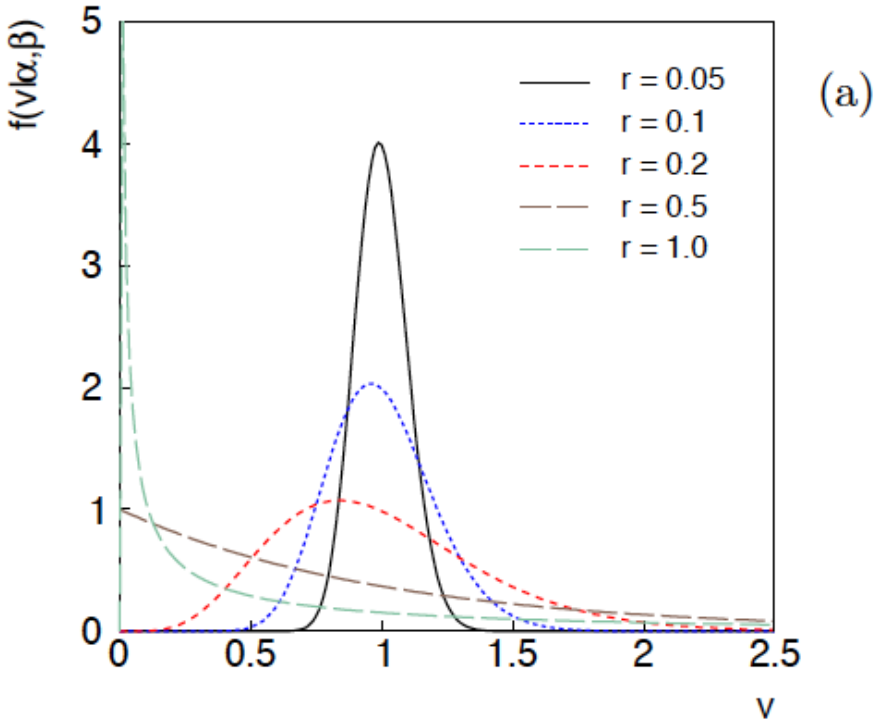
$r = 1/2\sqrt{\alpha} \approx$ relative "error on the error" = $\sigma_s/E[s]$.

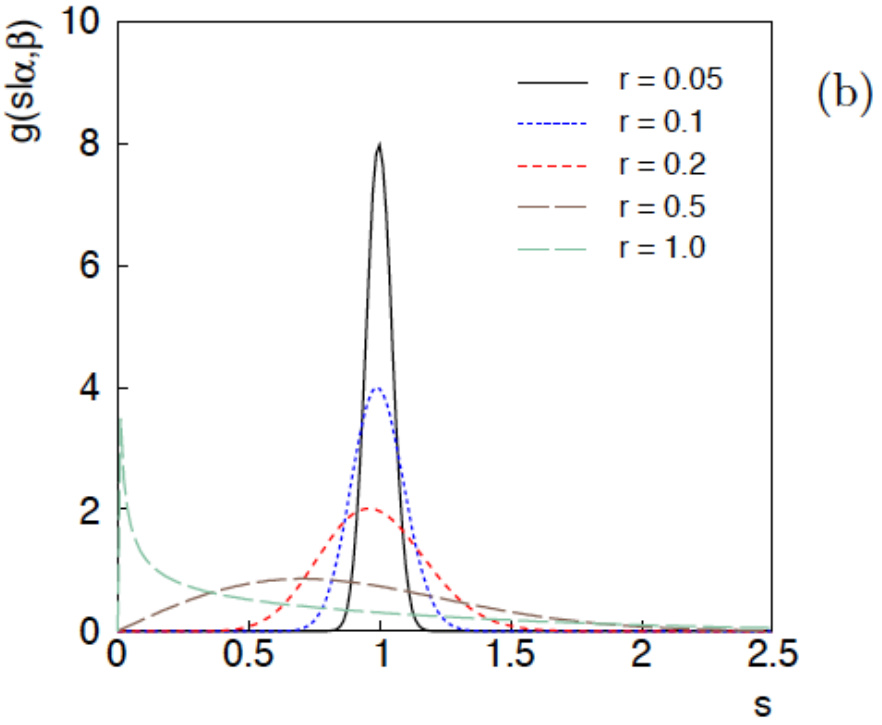
Distributions of v and $s = \sqrt{v}$

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

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

relative "error on error"





Motivation for gamma model

If one were to have n independent observations $u_1,..,u_n$, with all $u \sim \text{Gauss}(\theta, \sigma_u^2)$, 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 $(n-1)v/\sigma_u^2$ 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.

Likelihood for gamma error model

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

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

$$\alpha_i = 1/4\mathrm{r}_i^2$$
$$\beta_i = \alpha_i/\sigma_{ui}^2$$

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:

$\mu_1,\ldots,\mu_M$	(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.)

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

In limit of small r_i and $v_i \to {\sigma_{u,i}}^2$, 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

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

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

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

Statistical Data Analysis
Lecture 11-4

- Statistical models with uncertain error parameters (part 2)

Recall the profile likelihood from Lecture 11-3:

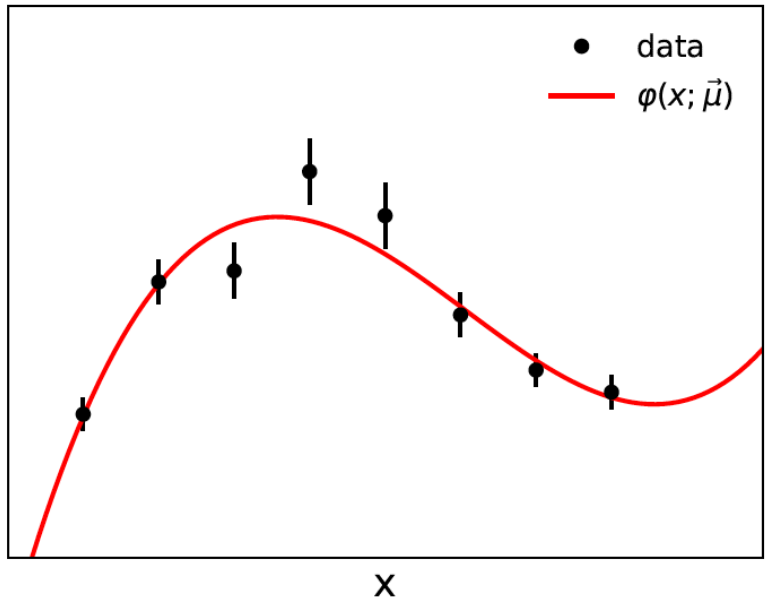
$$\begin{aligned}\ln L'(\mu, \boldsymbol{\theta}) &= \ln L(\mu, \boldsymbol{\theta}, \widehat{\widehat{\boldsymbol{\sigma}^2}}_{\mathbf{u}}) \\ &= \ln P(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) - \frac{1}{2}\sum_{i=1}^{N}\left(1 + \frac{1}{2r_i^2}\right)\ln\left[1 + 2r_i^2\frac{(u_i - \theta_i)^2}{v_i}\right]\end{aligned}$$

Curve fitting, averages

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

$$E[y_i] = \varphi(x_i;\boldsymbol{\mu}) + \theta_i \ ,$$

$$V[y_i] = \sigma_{y_i}^2 \quad \text{(known)}.$$



$\boldsymbol{\mu}$ 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:

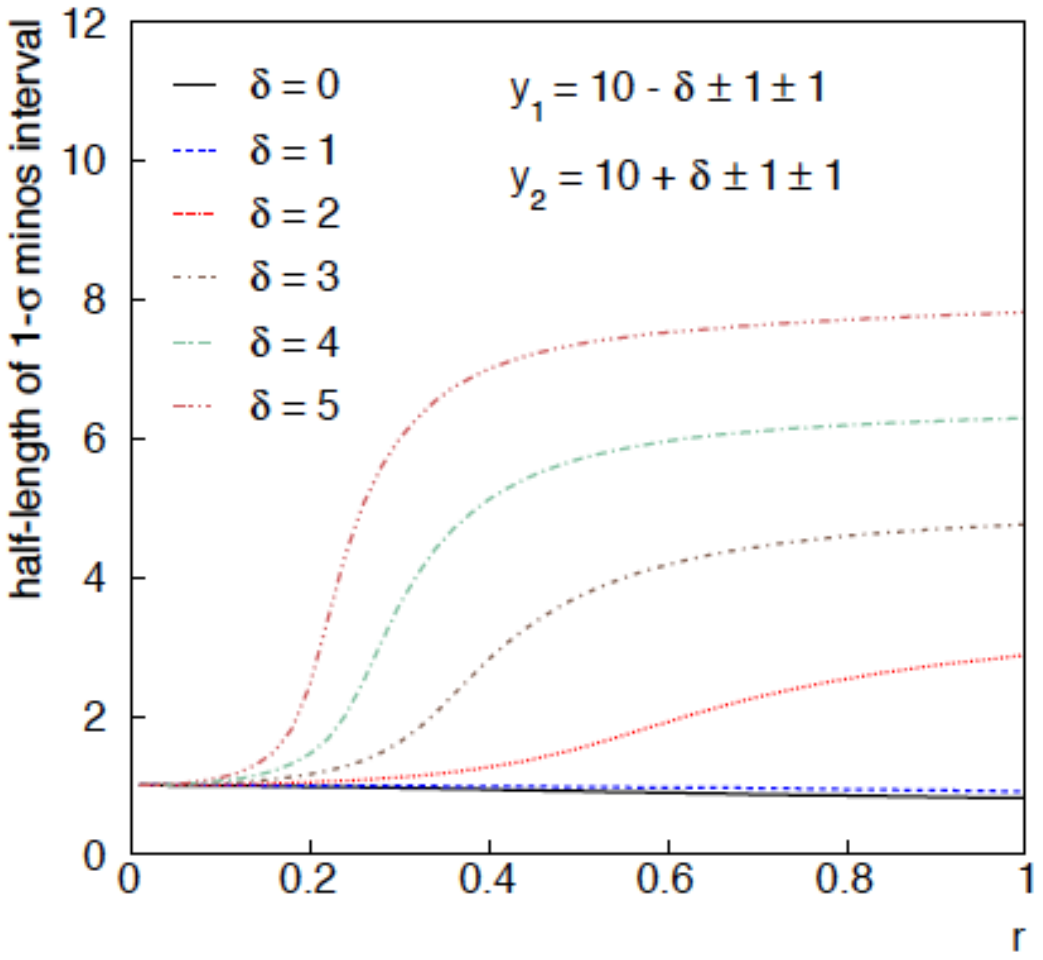
$$\theta_i^3 + [-2u_i - y_i + \varphi_i]\,\theta_i^2 + \left[\frac{v_i+(1+2r_i^2)\sigma_{y_i}^2}{2r_i^2} + 2u_i(y_i-\varphi_i) + u_i^2\right]\theta_i$$

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

Example: average of two measurements

Approximate ("MINOS") confidence interval based on

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



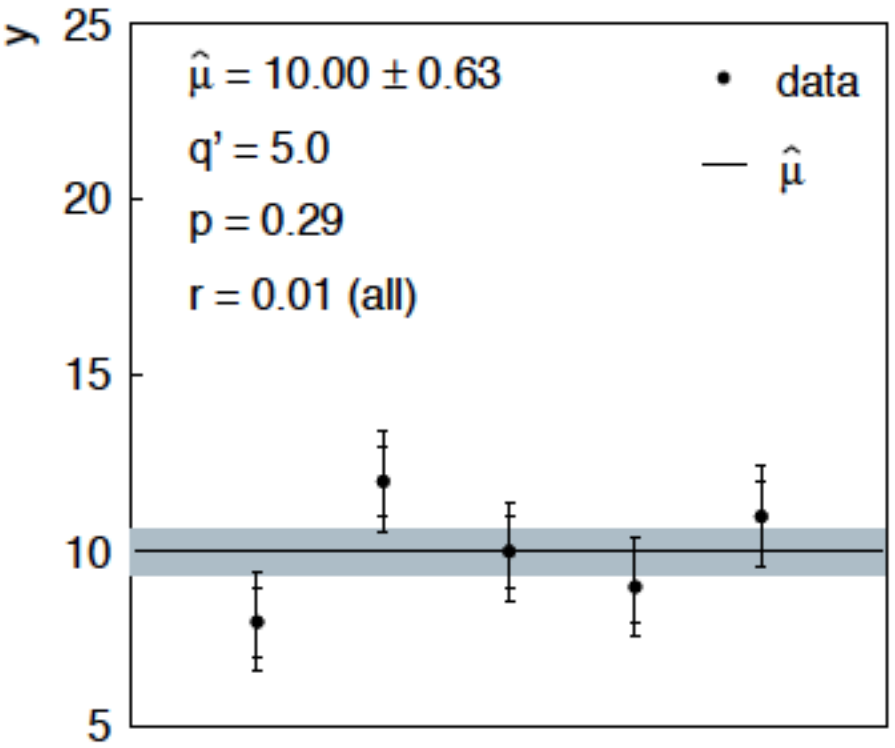
Increased discrepancy between values to be averaged gives larger interval.

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

relative error on sys. error

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).

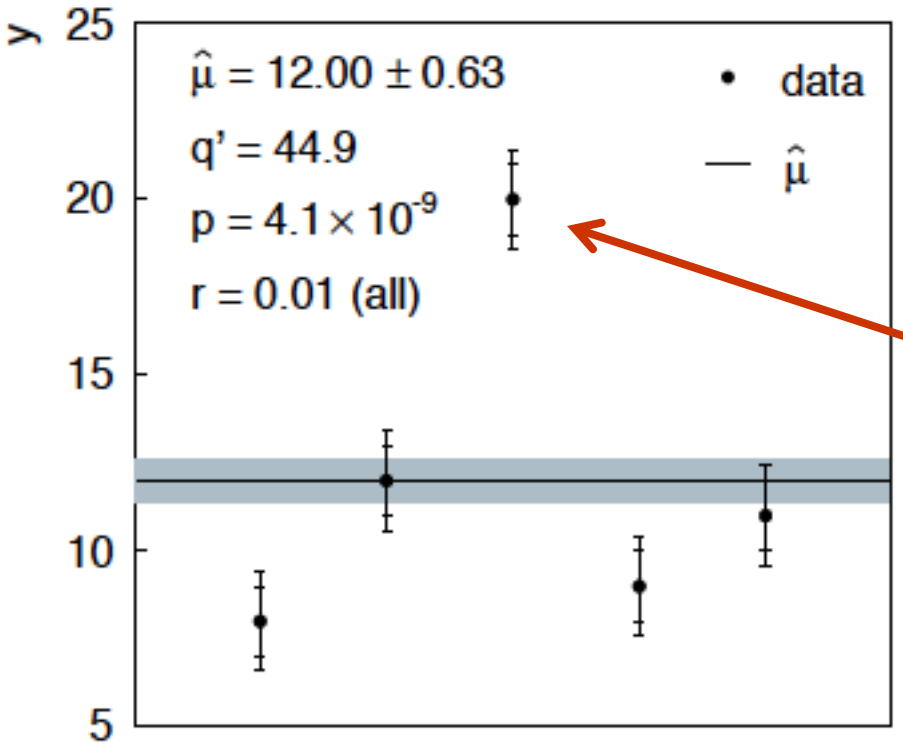


inner error bars
$= \sigma_{y,i}$

outer error bars
$= (\sigma_{y,i}^2 + \sigma_{u,i}^2)^{1/2}$

Sensitivity of average to outliers (2)

Now suppose the measurement at 10 had come out at 20:

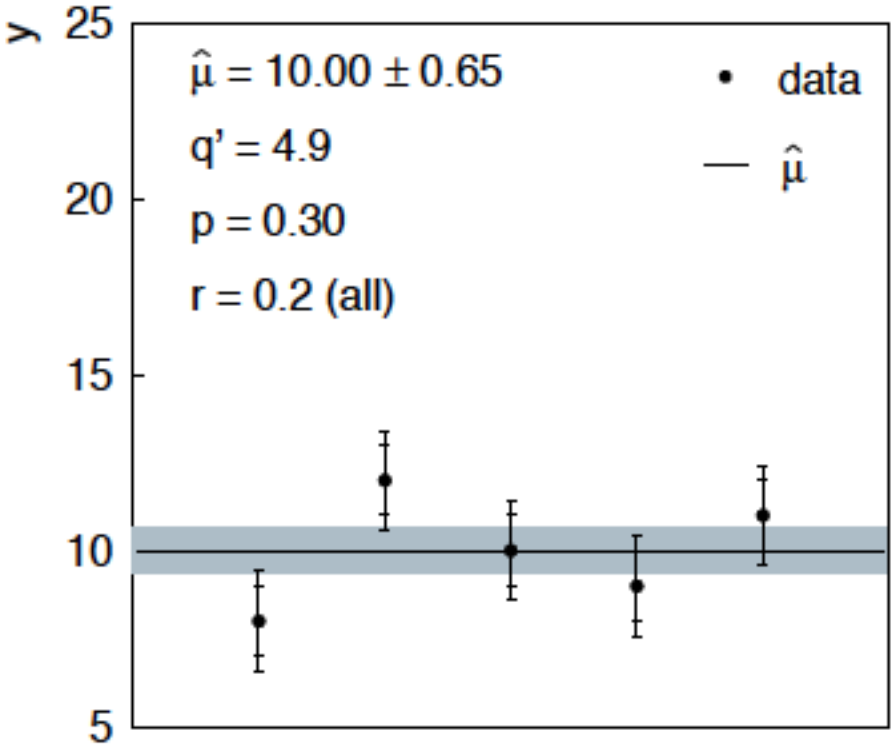


"outlier"

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

Average with all $r = 0.2$

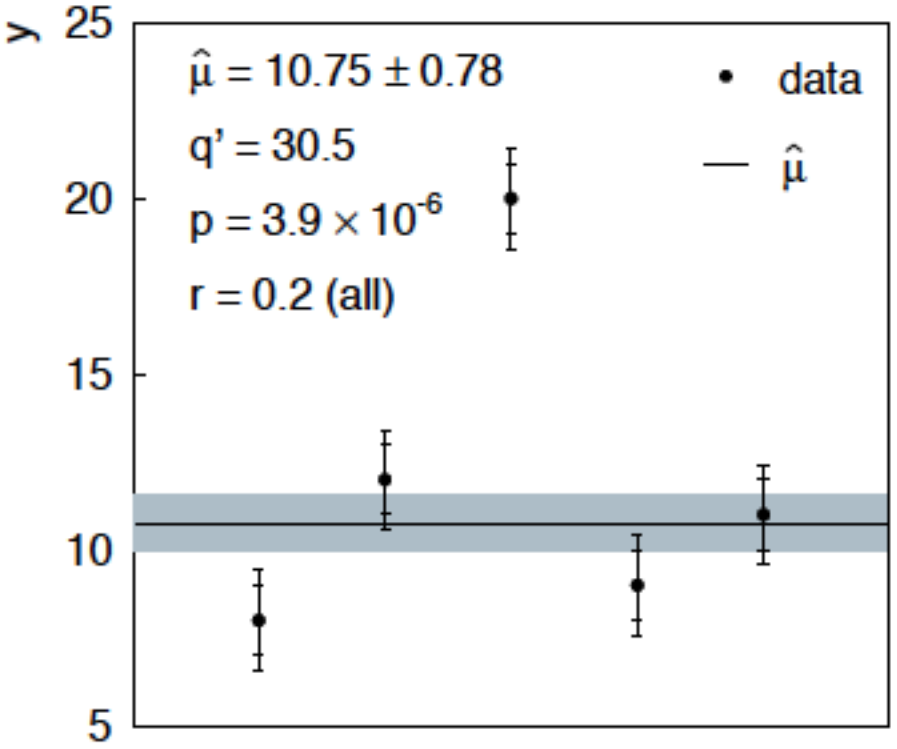
If we assign to each measurement $r = 0.2$,



Estimate still at 10.00, size of interval moves 0.63 → 0.65

Average with all $r = 0.2$ with outlier

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



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

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

Discussion / Conclusions

Gamma model for variance estimates gives confidence intervals that increase in size when the data are internally inconsistent, and gives decreased sensitivity to outliers (known property of Student's t based regression).

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

Simple profile likelihood – quadratic terms replaced by logarithmic:

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

Discussion / Conclusions (2)

Asymptotics can break for increased error-on-error, may need Bartlett correction, higher-order asymptotics or MC.*

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

Alternatively, one could try to fit a global r to all systematic errors, analogous to PDG scale factor method or meta-analysis à la DerSimonian and Laird. ($\longrightarrow$ future work).

Could also use e.g. as "stress test" – crank up the r_i values until significance of result degrades and ask if you really trust the assigned systematic errors at that level.

* see E. Canonero et al., Eur. Phys. J. C (2023) 83:1100, arXiv:2304.10574

Extra Slides

Single-measurement model

As a simplest example consider

$y \sim \text{Gauss}(\mu, \sigma^2)$,

$v \sim \text{Gamma}(\alpha, \beta)$, $\qquad \alpha = \frac{1}{4r^2}, \quad \beta = \frac{1}{4r^2\sigma^2}$

$$L(\mu, \sigma^2) = f(y, v|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} \frac{\beta^\alpha}{\Gamma(\alpha)} v^{\alpha-1} e^{-\beta v}$$

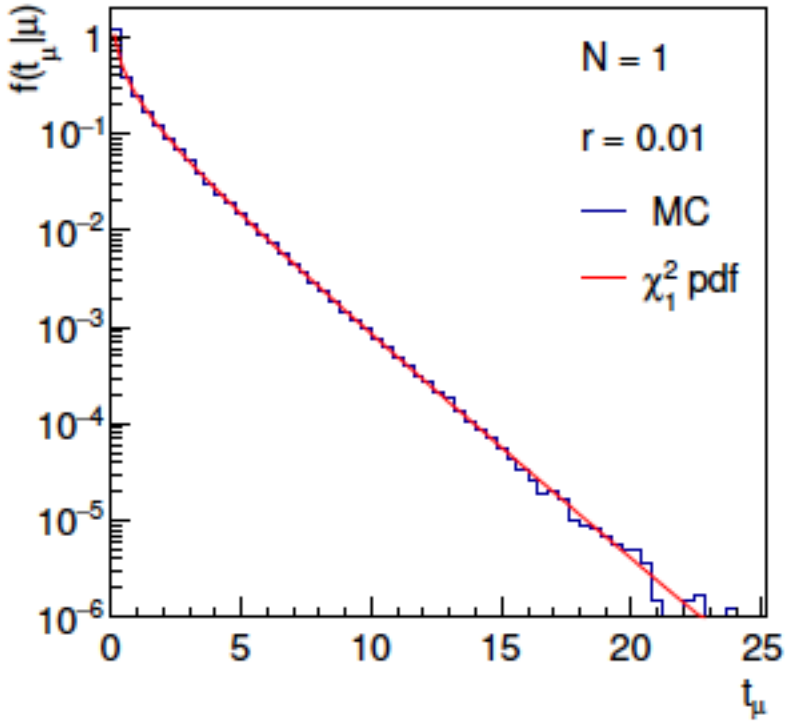
Test values of μ with $t_\mu = -2 \ln \lambda(\mu)$ with $\quad \lambda(\mu) = \dfrac{L(\mu, \widehat{\widehat{\sigma^2}}(\mu))}{L(\hat{\mu}, \widehat{\sigma^2})}$

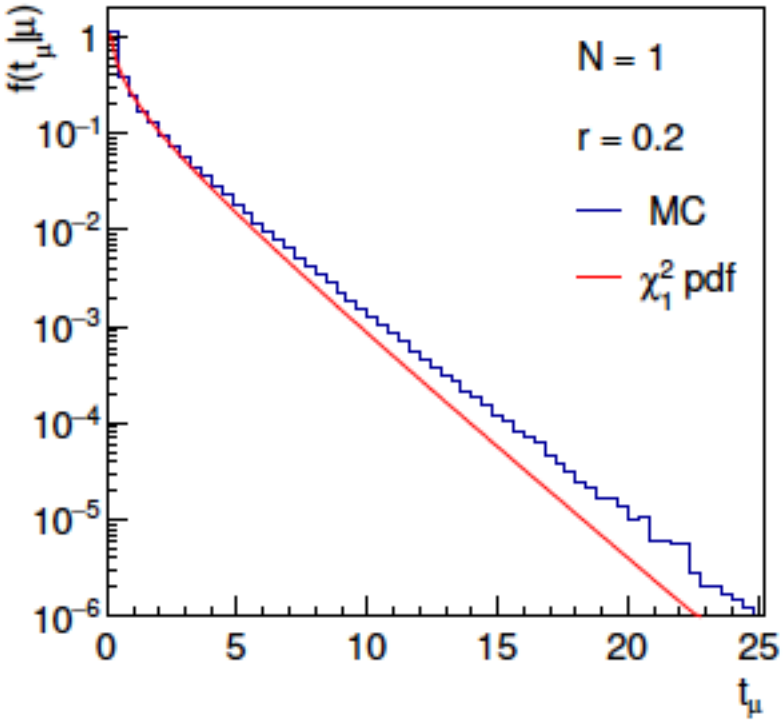
$$t_\mu = \left(1 + \frac{1}{2r^2}\right) \ln \left[1 + 2r^2 \frac{(y-\mu)^2}{v}\right]$$

Distribution of t_μ

From Wilks' theorem, in the asymptotic limit we should find $t_\mu \sim$ chi-squared(1).

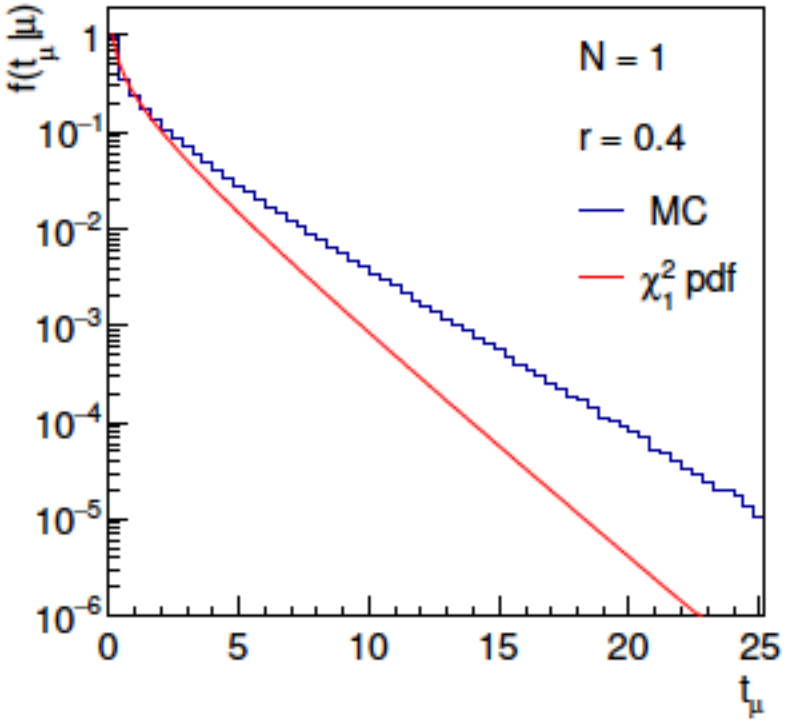
Here "asymptotic limit" means all estimators ~Gauss, which means $r \to 0$. For increasing r, clear deviations visible:

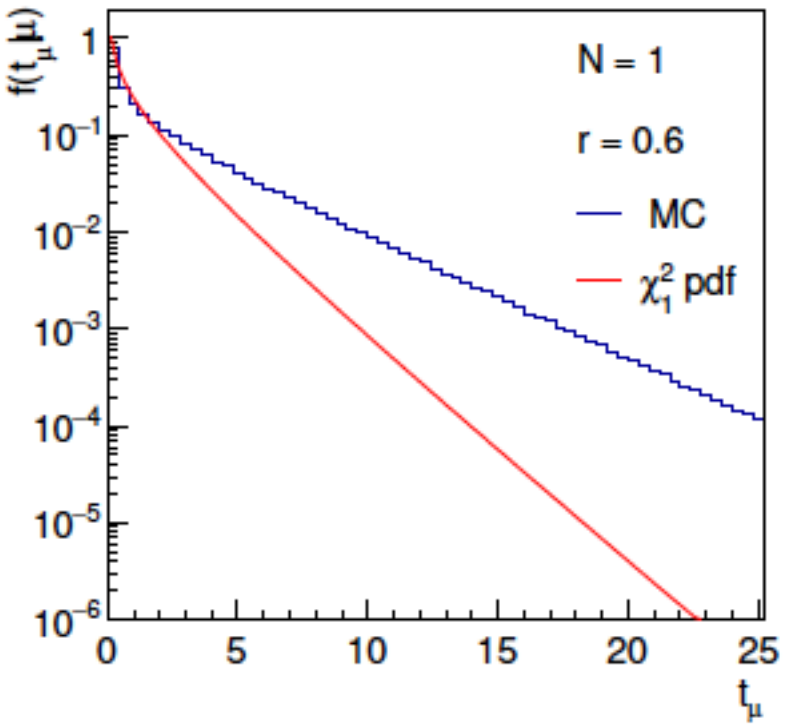




Distribution of t_μ (2)

For larger r, breakdown of asymptotics gets worse:





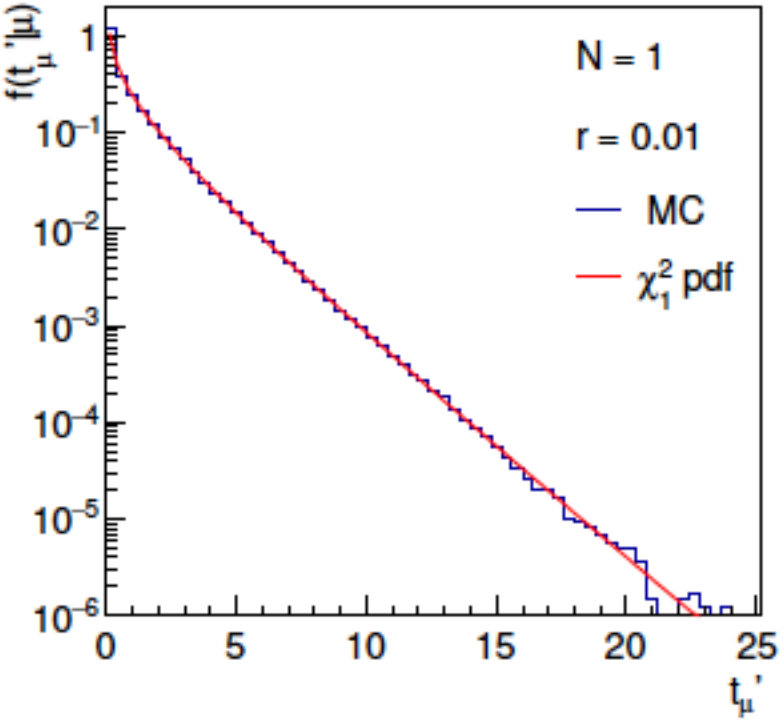
Values of $r \sim$ several tenths are relevant so we cannot in general rely on asymptotics to get confidence intervals, p-values, etc.

Bartlett corrections

One can modify t_μ defining $$t'_\mu = \frac{n_\mathrm{d}}{E[t_\mu]} t_\mu$$

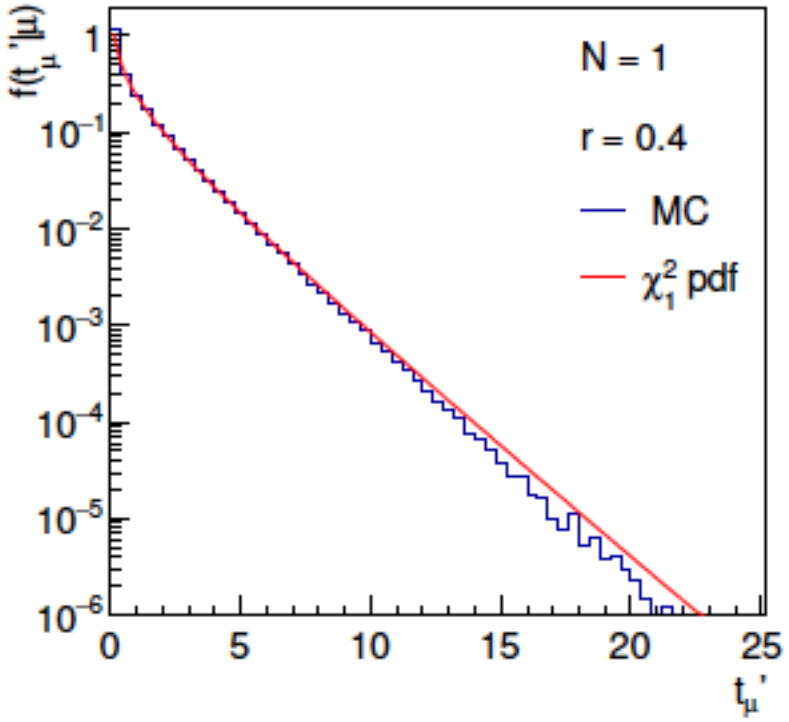
such that the new statistic's distribution is better approximated by chi-squared for n_d degrees of freedom (Bartlett, 1937).

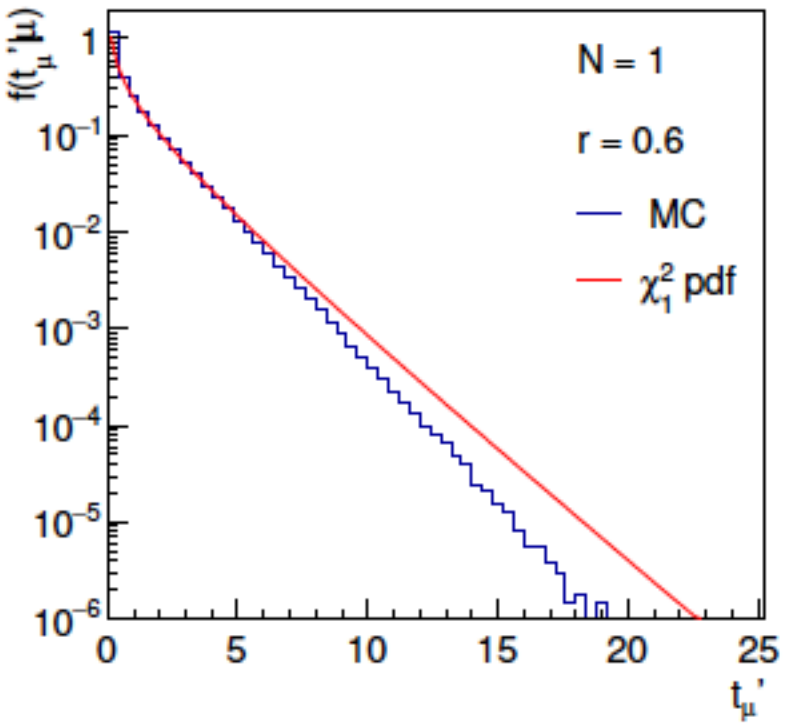
For this example $E[t_\mu] \approx 1 + 3r^2 + 2r^4$ works well:



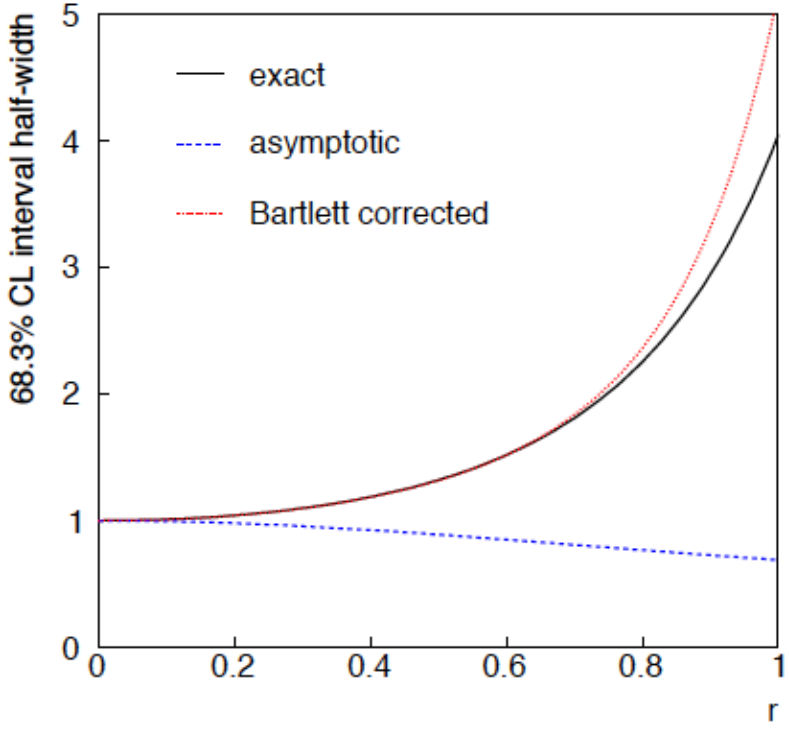
Bartlett corrections (2)

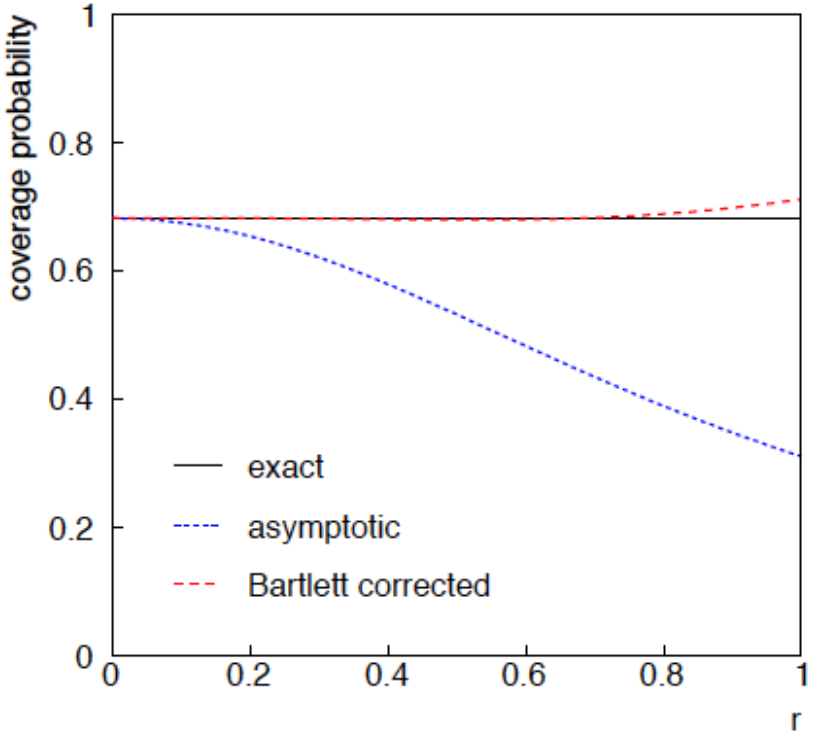
Good agreement for $r \sim$ several tenths out to $\sqrt{t_\mu}' \sim$ several, i.e., good for significances of several sigma:





68.3% CL confidence interval for μ





Goodness of fit

Can quantify goodness of fit with statistic

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

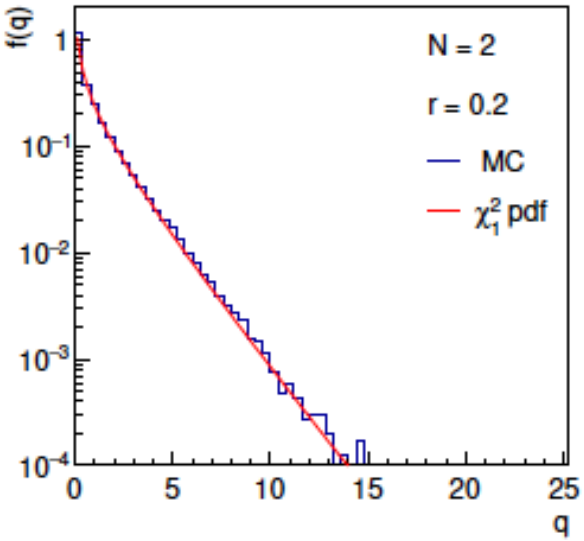
$$= \min_{\boldsymbol{\mu}, \boldsymbol{\theta}} \sum_{i=1}^{N} \left[\frac{(y_i - \varphi(x_i; \boldsymbol{\mu}) - \theta_i)^2}{\sigma_{y_i}^2} + \left(1 + \frac{1}{2r_i^2}\right) \ln \left(1 + 2r_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right) \right]$$

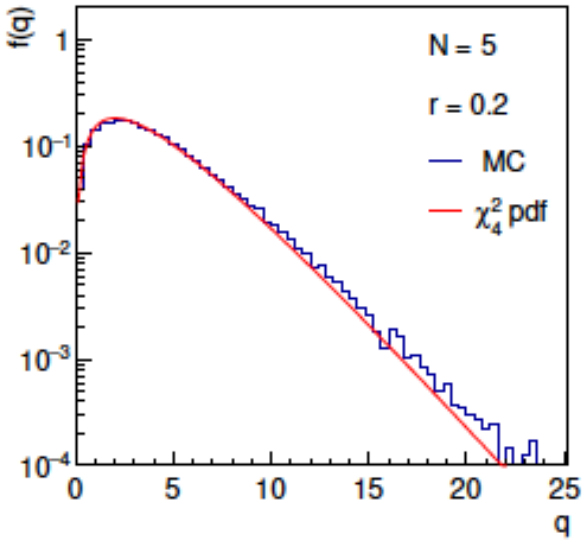
where $L'(\boldsymbol{\varphi},\boldsymbol{\theta})$ has an adjustable φ_i for each y_i (the saturated model).

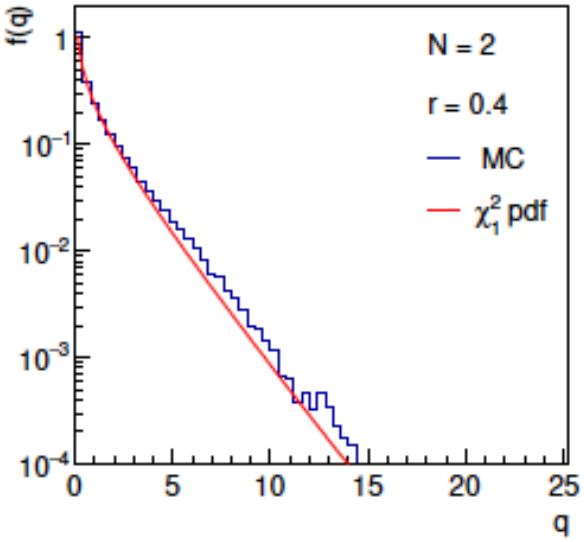
Asymptotically should have $q \sim$ chi-squared(N–M).

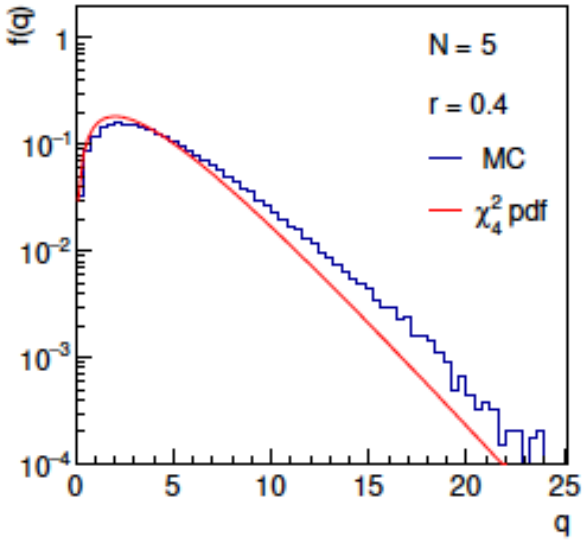
For increasing r_i, may need Bartlett correction or MC.

Distributions of q

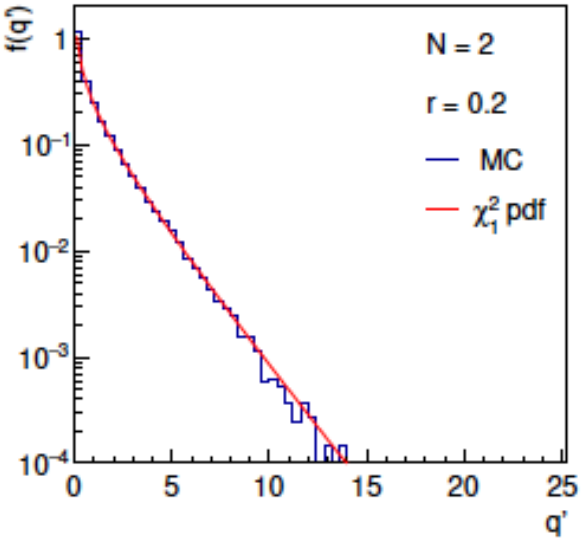


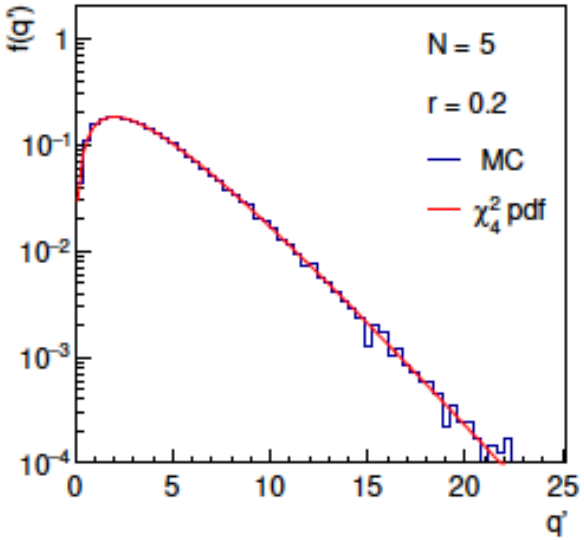


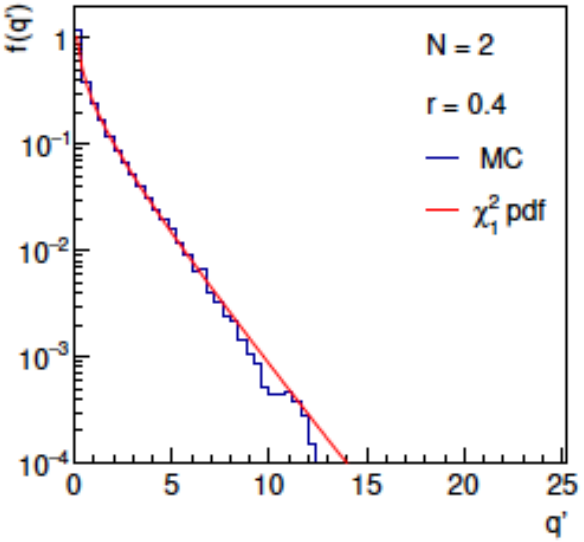


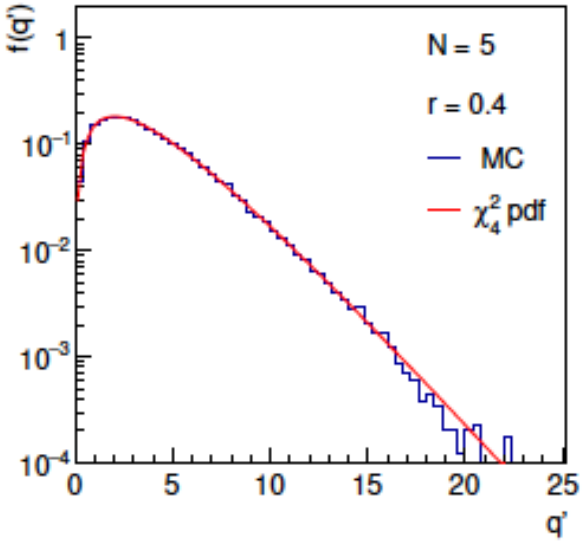


Distributions of Bartlett-corrected q'

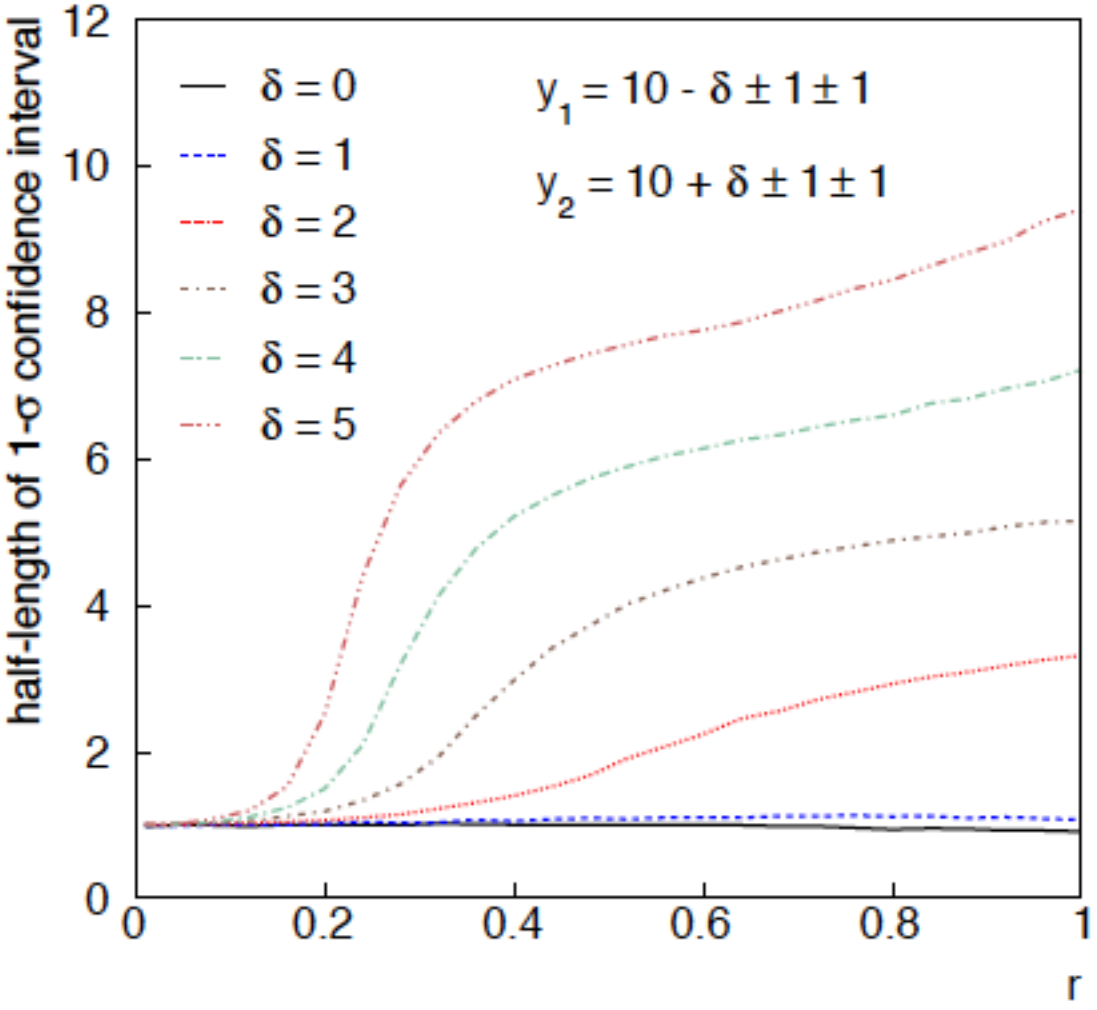








Same with interval from $p_\mu = \alpha$ with nuisance parameters profiled at μ



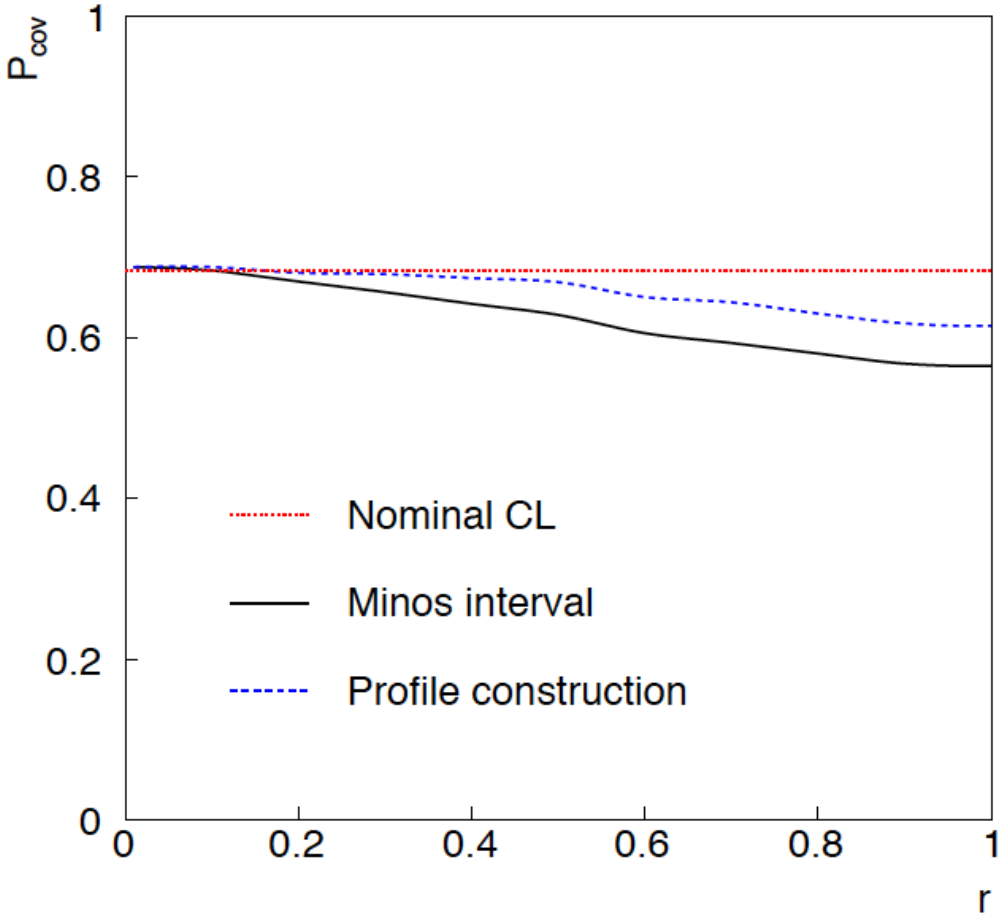
Coverage of intervals

Consider previous average of two numbers but now generate for $i = 1, 2$ data values

$$y_i \sim \text{Gauss}(\mu, \sigma_{y,i})$$
$$u_i \sim \text{Gauss}(0, \sigma_{u,i})$$
$$v_i \sim \text{Gamma}(\sigma_{u,i}, r_i)$$
$$\sigma_{y,i} = \sigma_{u,i} = 1$$

and look at the probability that the interval covers the true value of μ.

Coverage stays reasonable to $r \sim 0.5$, even not bad for Profile Construction out to $r \sim 1$.



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.

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.