

# Parameter Estimation

## Lecture 2



INFN School of Statistics  
Paestum, 12-18 May 2024

<https://agenda.infn.it/event/36980/>



Glen Cowan  
Physics Department  
Royal Holloway, University of London  
[g.cowan@rhul.ac.uk](mailto:g.cowan@rhul.ac.uk)  
[www.pp.rhul.ac.uk/~cowan](http://www.pp.rhul.ac.uk/~cowan)

# Outline

Lecture 1: Introduction, Maximum Likelihood

→ Lecture 2: Least squares, Bayesian approach, unfolding

For some slides/links with exercises and code you can see:

[https://www.pp.rhul.ac.uk/~cowan/stat/exercises/cowan\\_stat\\_exercises.pdf](https://www.pp.rhul.ac.uk/~cowan/stat/exercises/cowan_stat_exercises.pdf)

[https://www.pp.rhul.ac.uk/~cowan/stat/exercises/cowan\\_stat\\_exercises\\_full.pdf](https://www.pp.rhul.ac.uk/~cowan/stat/exercises/cowan_stat_exercises_full.pdf)

Most material here is taken from the University of London course:

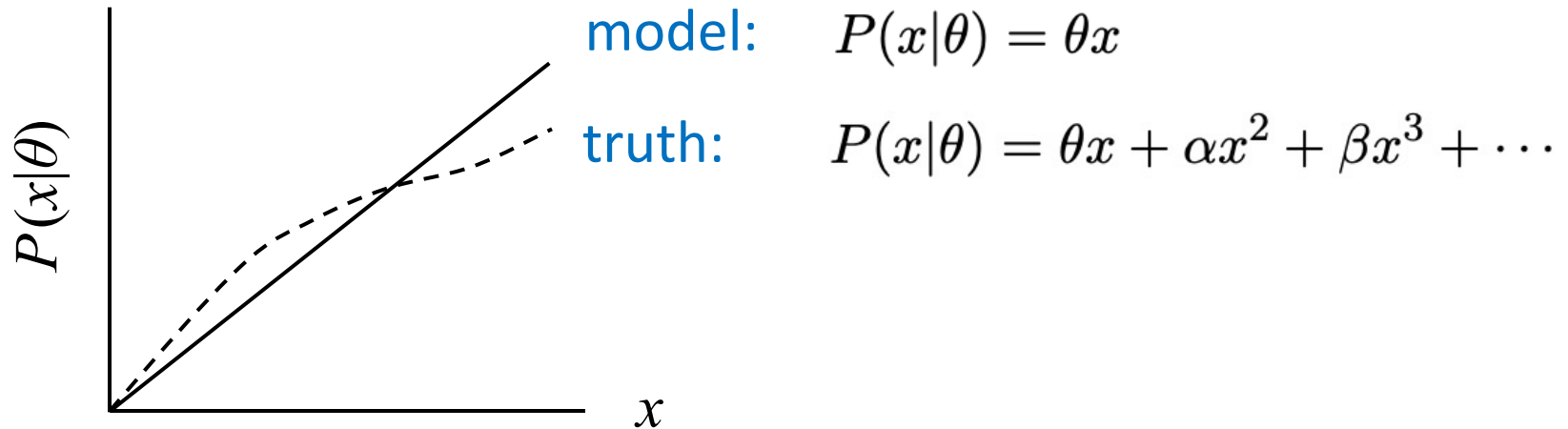
[https://www.pp.rhul.ac.uk/~cowan/stat\\_course.html](https://www.pp.rhul.ac.uk/~cowan/stat_course.html)

# Parameter Estimation 2-1

- Nuisance parameters, systematic uncertainties
- From Maximum Likelihood to Least Squares
- Bayesian parameter estimation
- Marginalization of posterior pdf
- Markov Chain Monte Carlo

# Systematic uncertainties and nuisance parameters

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



Can improve model by including additional adjustable parameters.

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

Nuisance parameter  $\leftrightarrow$  systematic uncertainty. Some point in the parameter space of the enlarged model should be “true”.

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

# Example: fitting a straight line

Data:  $(x_i, y_i, \sigma_i)$ ,  $i = 1, \dots, n$ .

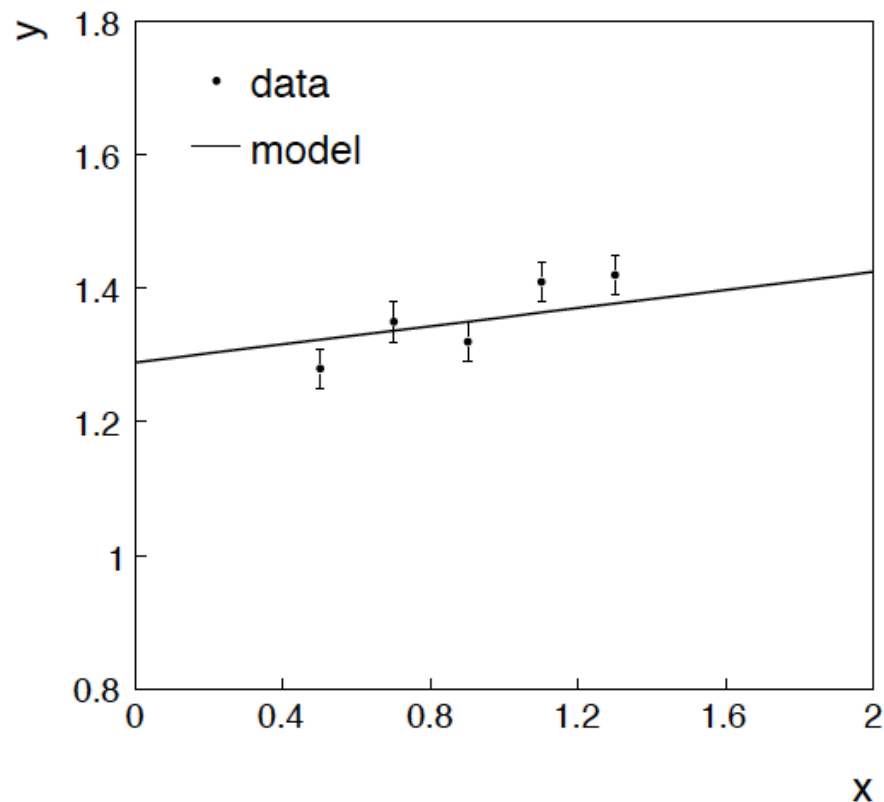
Model:  $y_i$  independent and all follow  $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

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

assume  $x_i$  and  $\sigma_i$  known.

Goal: estimate  $\theta_0$

Here suppose we don't care about  $\theta_1$  (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)

# $\theta_1$ known a priori

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

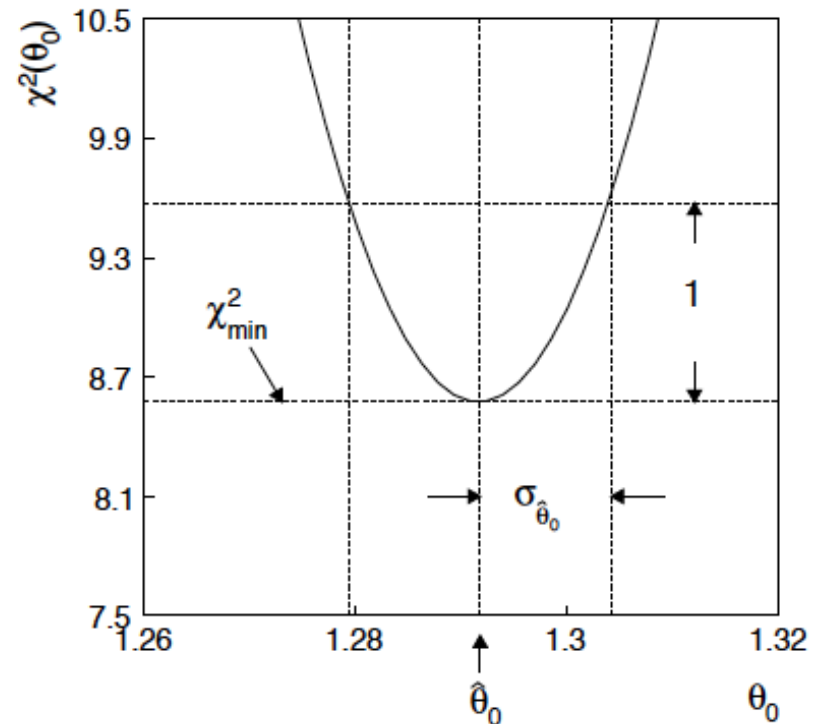
$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

For Gaussian  $y_i$ , ML same as LS

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

Come up one unit from  $\chi_{\min}^2$

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



# ML (or LS) fit of $\theta_0$ and $\theta_1$

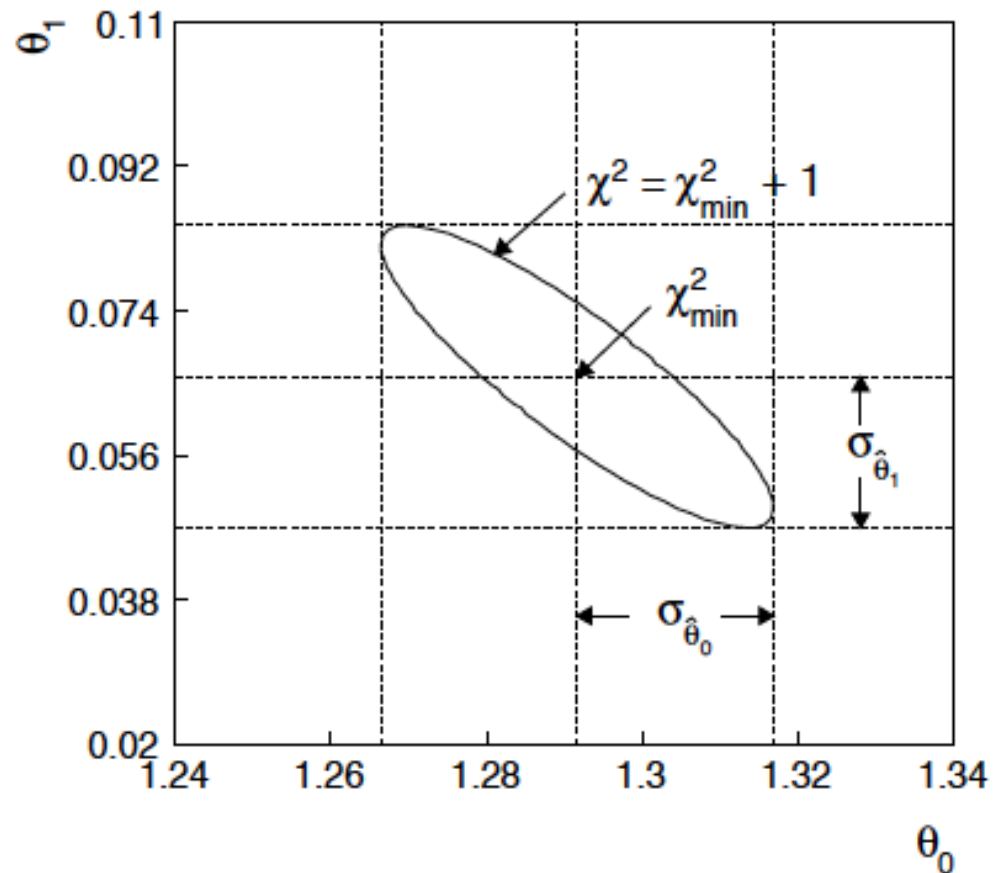
$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from  
tangent lines to contour

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

Correlation between

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



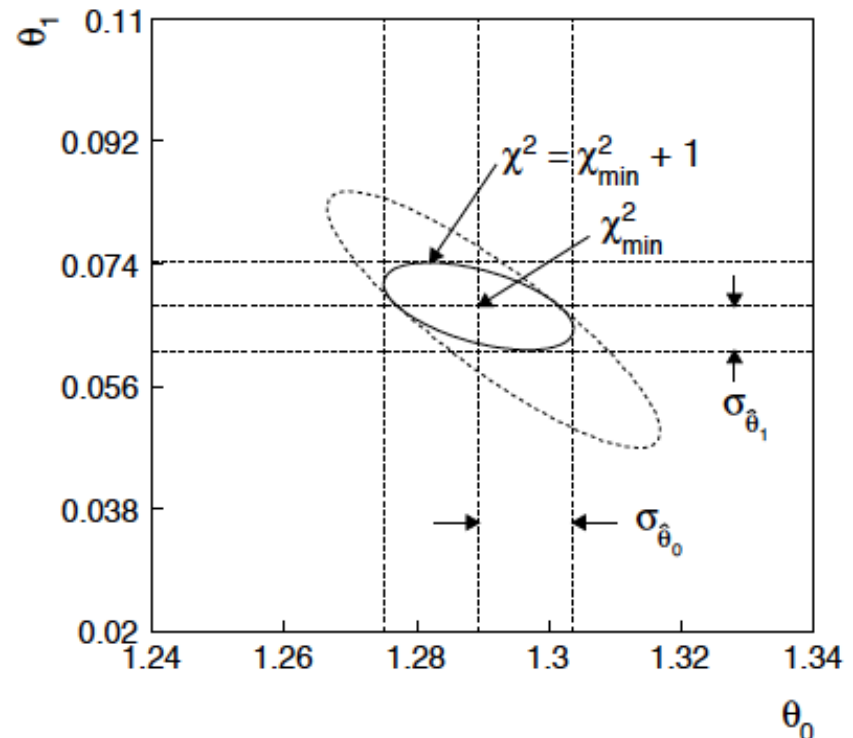


If we have a measurement  $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t_1})$

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

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

The information on  $\theta_1$   
improves accuracy of  $\hat{\theta}_0$ .



# Reminder of Bayesian approach

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

Interpret probability of  $\theta$  as ‘degree of belief’ (subjective).

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

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

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

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

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

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

We need to associate prior probabilities with  $\theta_0$  and  $\theta_1$ , e.g.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\pi_1(\theta_1) \quad \leftarrow \text{suppose knowledge of } \theta_0 \text{ has no influence on knowledge of } \theta_1$$

$$\pi_0(\theta_0) = \text{const.} \quad \leftarrow \text{'non-informative', in any case much broader than } L(\theta_0)$$

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

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

Ur = "primordial"  
prior

Likelihood for control  
measurement  $t_1$

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

Putting the ingredients into Bayes' theorem gives:

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

posterior  $\propto$  likelihood  $\times$  prior

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

The information from the control measurement  $t_1$  has been put into the prior for  $\theta_1$ .

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

# Marginalizing the posterior pdf

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

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

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

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

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

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

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

# Marginalization with MCMC

Bayesian computations involve integrals like

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

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

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

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

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

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

# MCMC basics: Metropolis-Hastings algorithm

Goal: given an  $n$ -dimensional pdf  $p(\theta)$ , generate a sequence of points  $\theta_1, \theta_2, \theta_3, \dots$

Proposal density  $q(\theta; \theta_0)$   
e.g. Gaussian centred  
about  $\theta_0$

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

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

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

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

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

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

6) Iterate

# Metropolis-Hastings (continued)

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

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

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric:  $q(\theta; \theta_0) = q(\theta_0; \theta)$

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

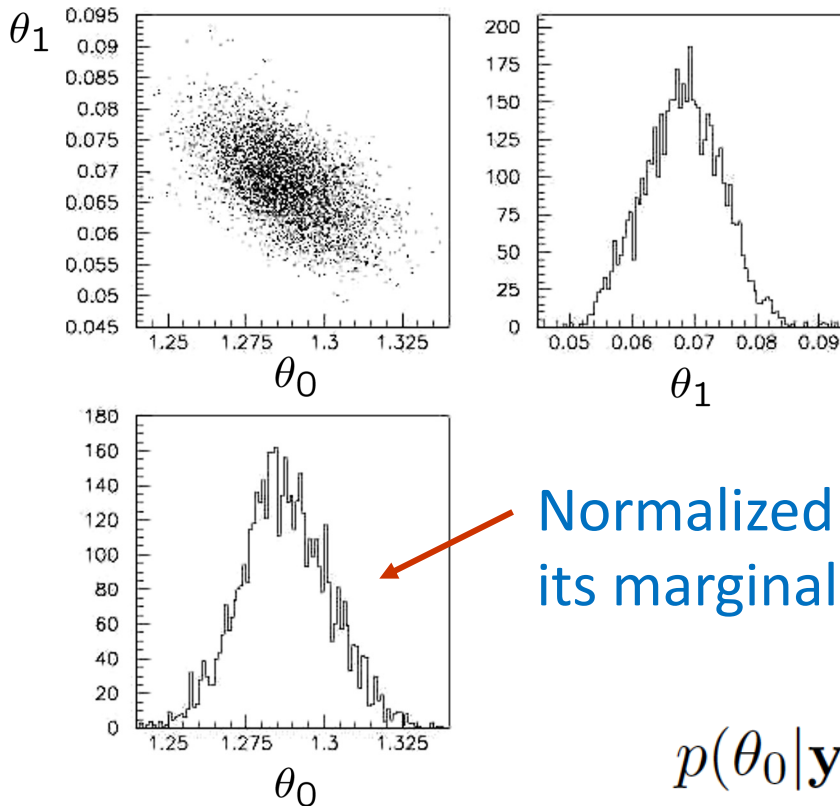
I.e. if the proposed step is to a point of higher  $p(\theta)$ , take it; if not, only take the step with probability  $p(\theta)/p(\theta_0)$ .

If proposed step rejected, repeat the current point.



# Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Normalized histogram of  $\theta_0$  gives its marginal posterior pdf:

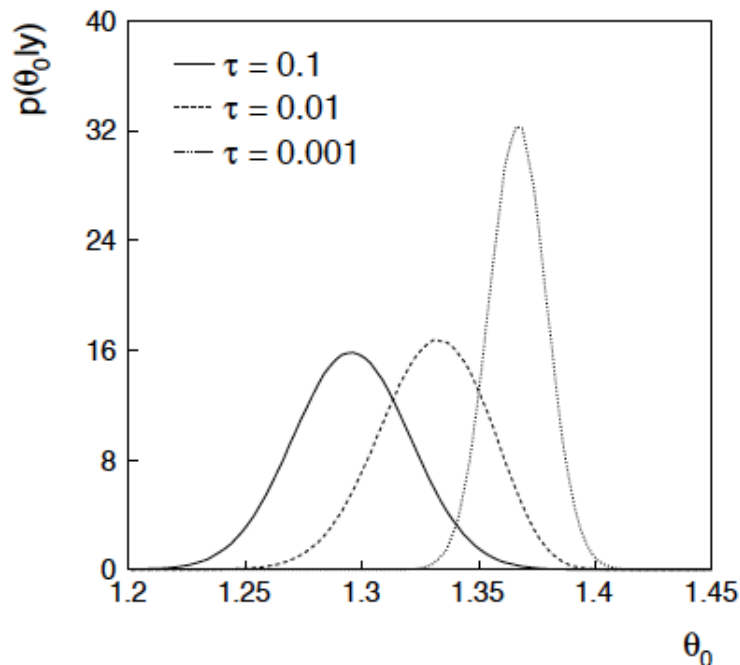
$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y}) d\theta_1$$

# Bayesian method with alternative priors

Suppose we don't have a previous measurement of  $\theta_1$  but rather, an “expert” says it should be positive and not too much greater than 0.1 or so, i.e., something like

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

From this we obtain (numerically) the posterior pdf for  $\theta_0$ :



This summarizes all knowledge about  $\theta_0$ .

Look also at result from variety of priors.

# Parameter Estimation 2-2

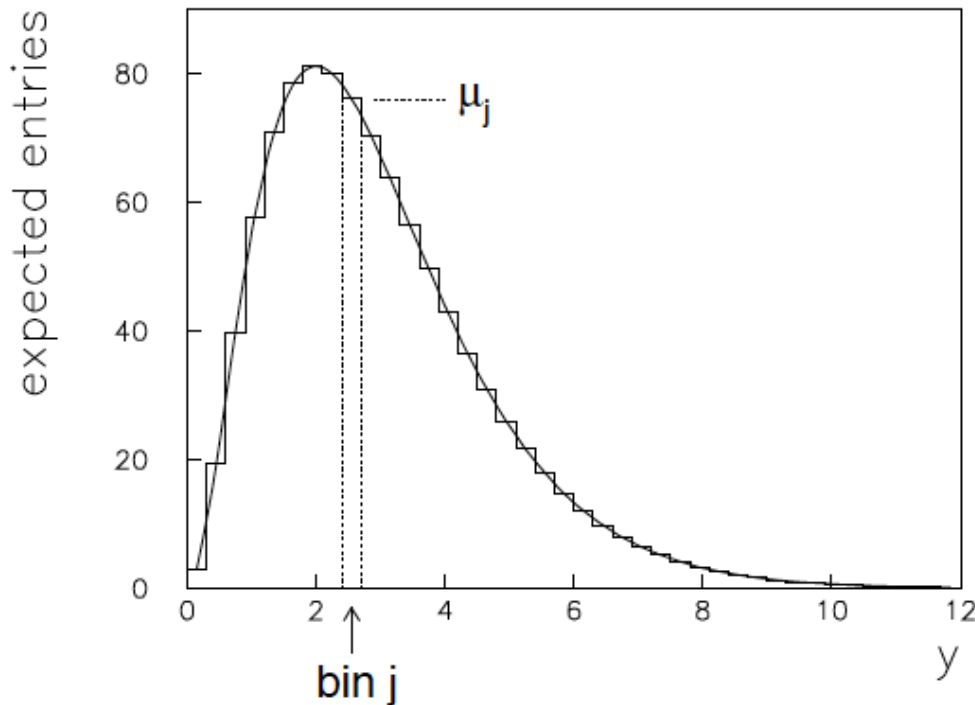
- Unfolding:
  - Formulation of the problem
  - Maximum Likelihood for unfolding
  - Regularized unfolding

# Formulation of the unfolding problem

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

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

If no parameterization available, often construct histogram:



$$p_j = \int_{\text{bin } j} f(y) dy$$

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



“true” histogram

New goal: construct estimators for the  $\mu_j$  (or  $p_j$ ).

# Migration

Effect of measurement errors:  $y$  = true value,  $x$  = observed value,  
migration of entries between bins,  
 $f(y)$  is 'smeared out', peaks broadened.

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



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

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

response matrix

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

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

# Efficiency, background

Sometimes an event goes undetected:

$$\begin{aligned}\sum_{i=1}^N R_{ij} &= \sum_{i=1}^N P(\text{observed in bin } i \mid \text{true value in bin } j) \\ &= P(\text{observed anywhere} \mid \text{true value in bin } j) \\ &= \varepsilon_j \quad \longleftarrow \text{efficiency}\end{aligned}$$

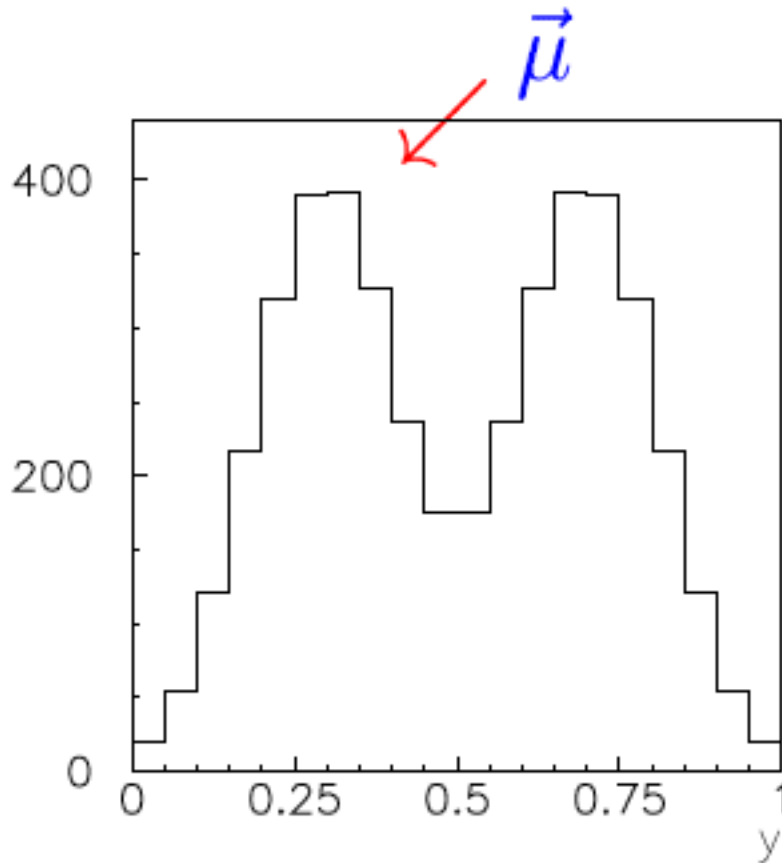
Sometimes an observed event is due to a background process:

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

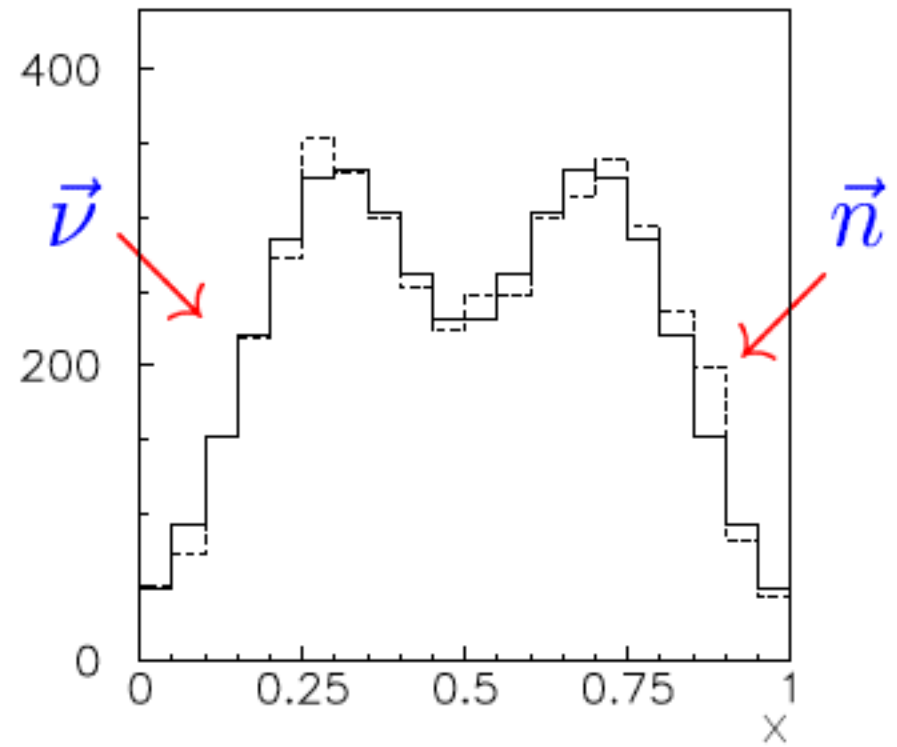
$\beta_i$  = expected number of background events in *observed* histogram.

For now, assume the  $\beta_i$  are known.

# The basic ingredients



“true”



“observed”

# Summary of ingredients

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

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

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

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

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

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

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

These are related by:

$$E[\mathbf{n}] = \boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$$



# Maximum likelihood (ML) estimator from inverting the response matrix

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

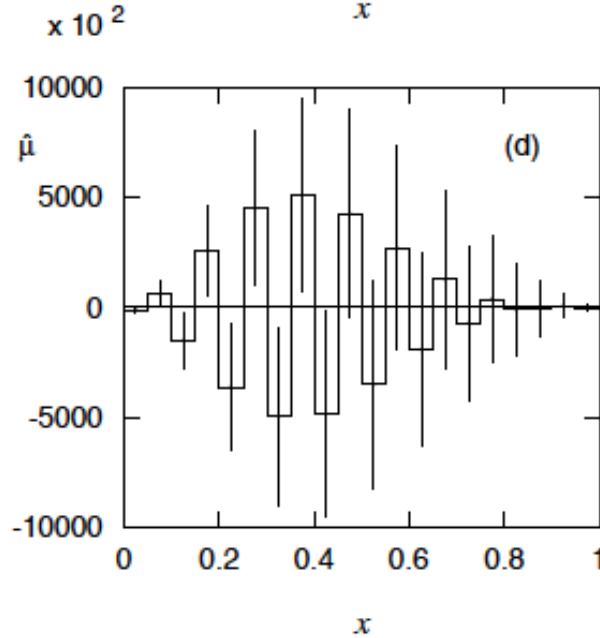
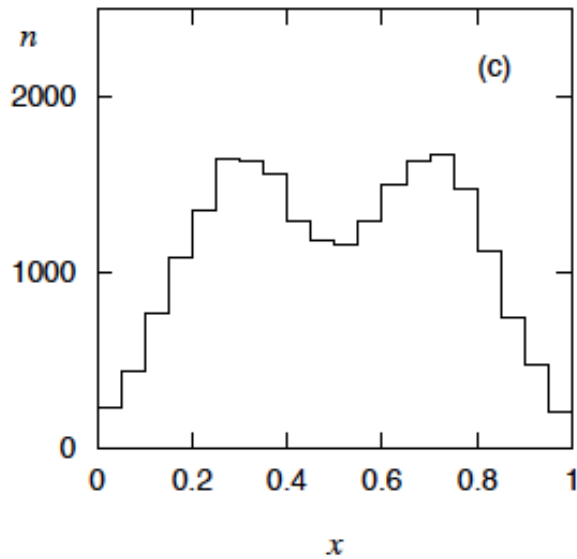
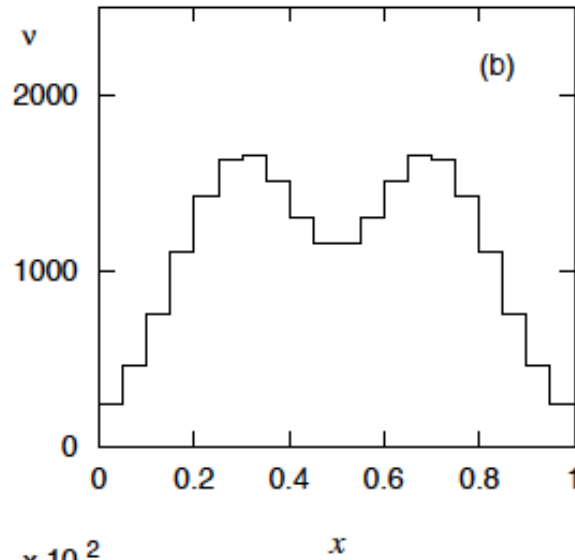
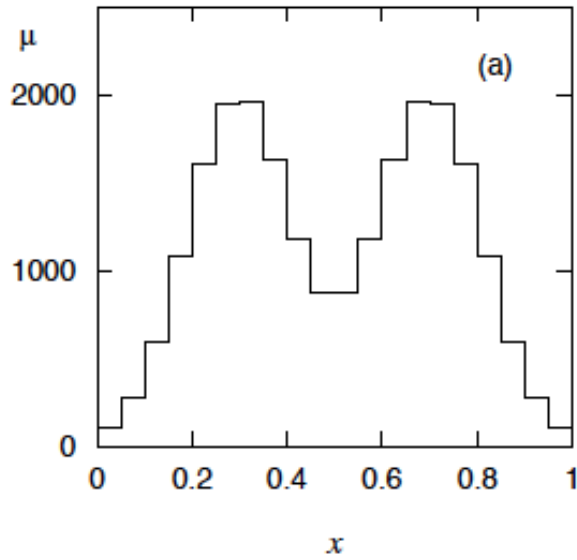
Suppose data are independent Poisson:  $P(n_i; \nu_i) = \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$

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

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

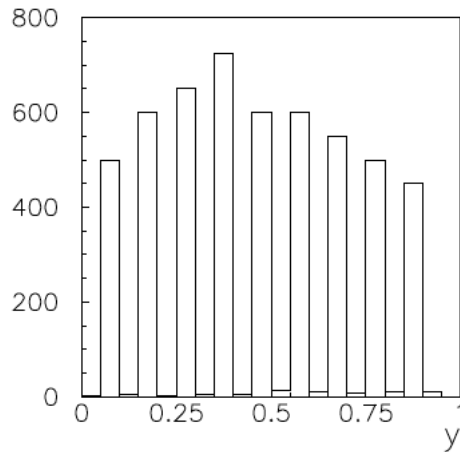
$$\longrightarrow \hat{\boldsymbol{\mu}} = R^{-1}(\mathbf{n} - \boldsymbol{\beta})$$

# Example with ML solution



Catastrophic failure???

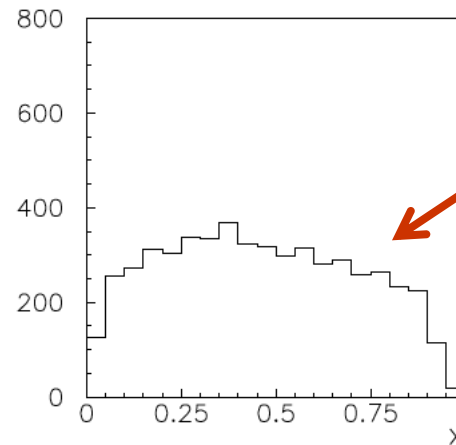
# What went wrong?



Suppose  $\vec{\mu}$  really had a lot of fine structure.

←  $\vec{\mu}$

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



→  $\vec{\nu} = R\vec{\mu}$

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

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

# Maximum likelihood solution revisited

For Poisson data the ML estimators are unbiased:

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

Their covariance is:

$$\begin{aligned} U_{ij} = \text{COV}[\hat{\mu}_i, \hat{\mu}_j] &= \sum_{k,l=1}^N (R^{-1})_{ik} (R^{-1})_{jl} \text{COV}[n_k, n_l] \\ &= \sum_{k=1}^N (R^{-1})_{ik} (R^{-1})_{jk} \nu_k \end{aligned}$$

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

## ML solution revisited (2)

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

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

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

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

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

# Correction factor method

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


$$C_i = \frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}} \quad \nu_i^{\text{MC}} \text{ and } \mu_i^{\text{MC}} \text{ from Monte Carlo simulation (no background).}$$

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

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

But the bias  $b_i = E[\hat{\mu}_i] - \mu_i$  is 
$$b_i = \left( \frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}} - \frac{\mu_i}{\nu_i^{\text{sig}}} \right)$$

Nonzero bias unless MC = Nature.


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

## Reality check on the statistical errors

Suppose for some bin  $i$  we have:

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

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

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

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

# Discussion of correction factor method

As with all unfolding methods, we get a reduction in statistical error in exchange for a bias..

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

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

For a more careful analysis the other regularized unfolding methods are usually preferred.



# Regularized unfolding

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

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

Out of these estimators, choose the ‘smoothest’, by maximizing

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

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

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

## Regularized unfolding (2)

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

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

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

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

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

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

a prescription for setting  $\alpha$ .

# Tikhonov regularization

Take measure of smoothness = mean square of  $k$ th derivative,

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

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

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

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

→ setting derivatives of  $\varphi$  equal to zero gives linear equations.

Solution using Singular Value Decomposition (SVD).

# SVD implementation of Tikhonov unfolding

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

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

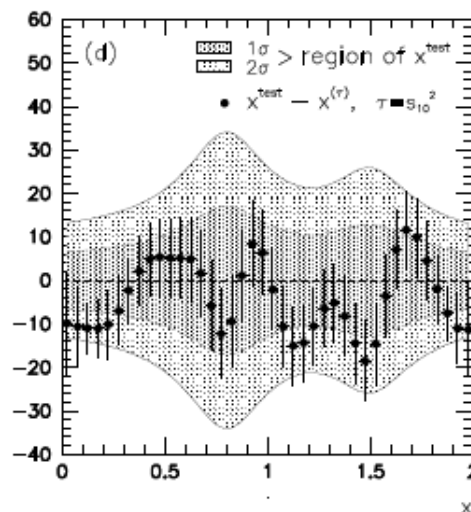
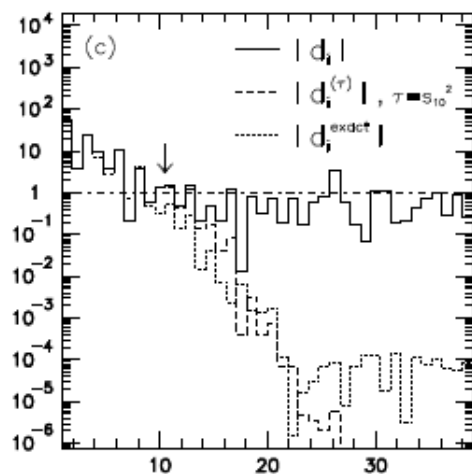
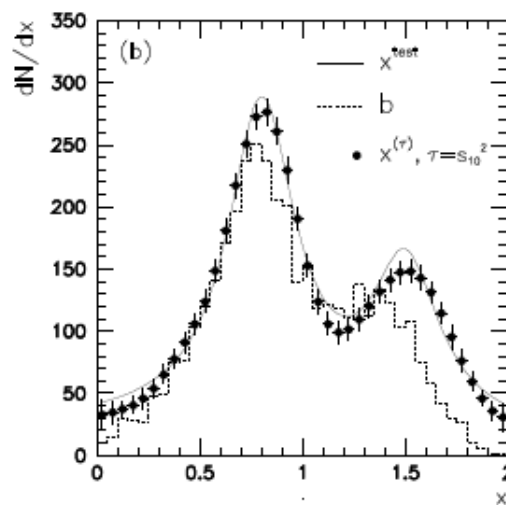
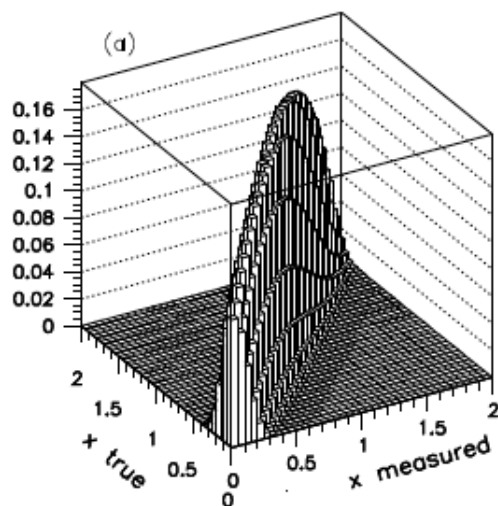
Numerical implementation using Singular Value Decomposition.

Recommendations for setting regularization parameter  $\tau$ :

Transform variables so errors  $\sim \text{Gauss}(0,1)$ ;  
number of transformed values significantly different  
from zero gives prescription for  $\tau$ ;  
or base choice of  $\tau$  on unfolding of test distributions.

# SVD example

A. Höcker, V. Kartvelishvili, NIM A**372** (1996) 469.



# Regularization function based on entropy

Shannon entropy of a set of probabilities is

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

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

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

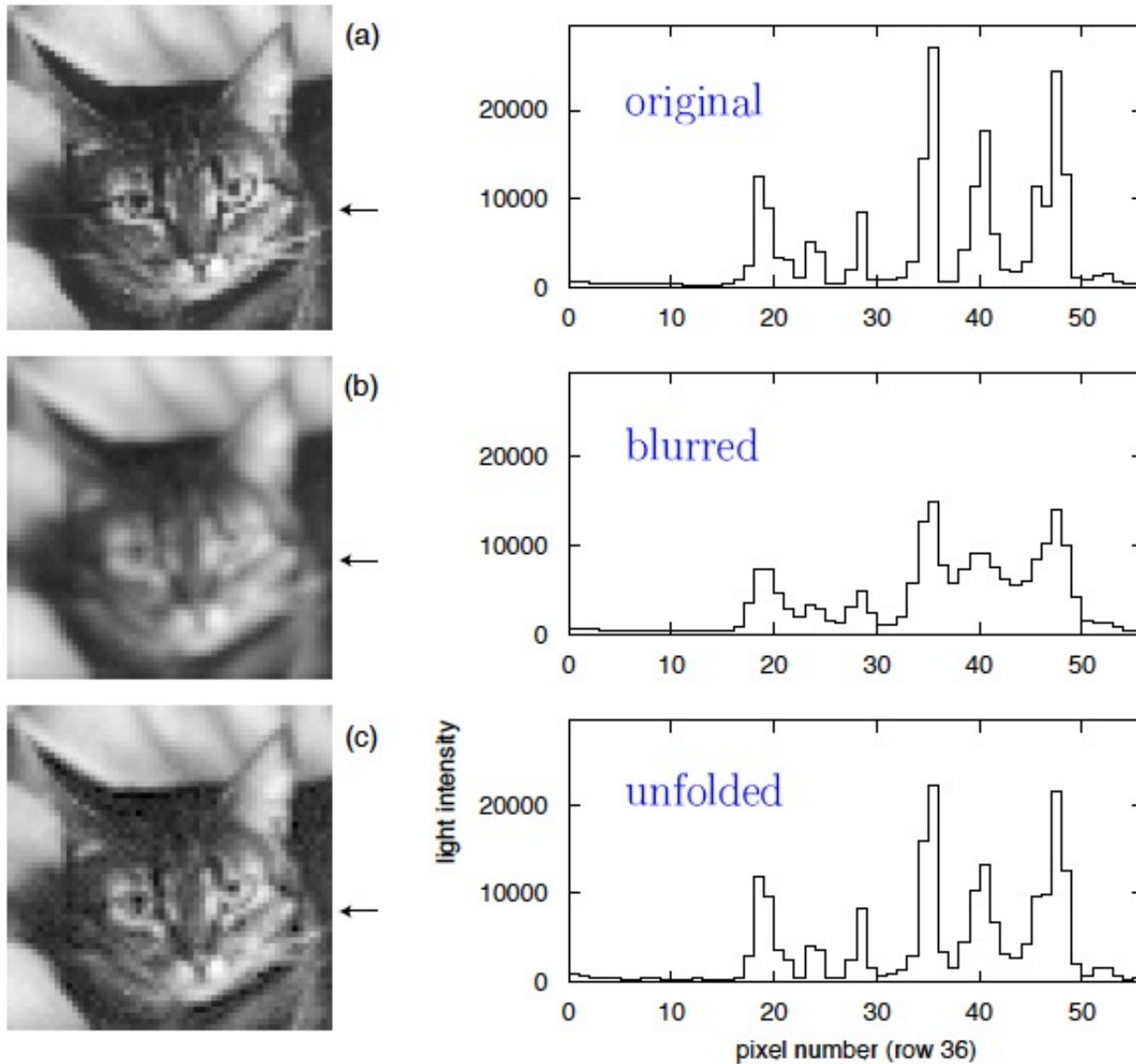
Use entropy as regularization function,

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

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

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

# Example of entropy-based unfolding



## Estimating bias and variance

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

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

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

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

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

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



## Choosing the regularization parameter

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

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

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

Minimize mean squared error,

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

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

## Choosing the regularization parameter (2)

Or look at changes in  $\chi^2$  from unregularized (ML) solution,

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

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

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

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

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

# Choosing the regularization parameter (3)

Or look at the average coverage probability of the 68.3% confidence intervals for individual bins

$$P_{\text{cov}} = \frac{1}{M} \sum_{i=1}^M P(\hat{\mu}_i - \sigma_{\hat{\mu}_i} < \mu_i < \hat{\mu}_i + \sigma_{\hat{\mu}_i} | \mu_i)$$

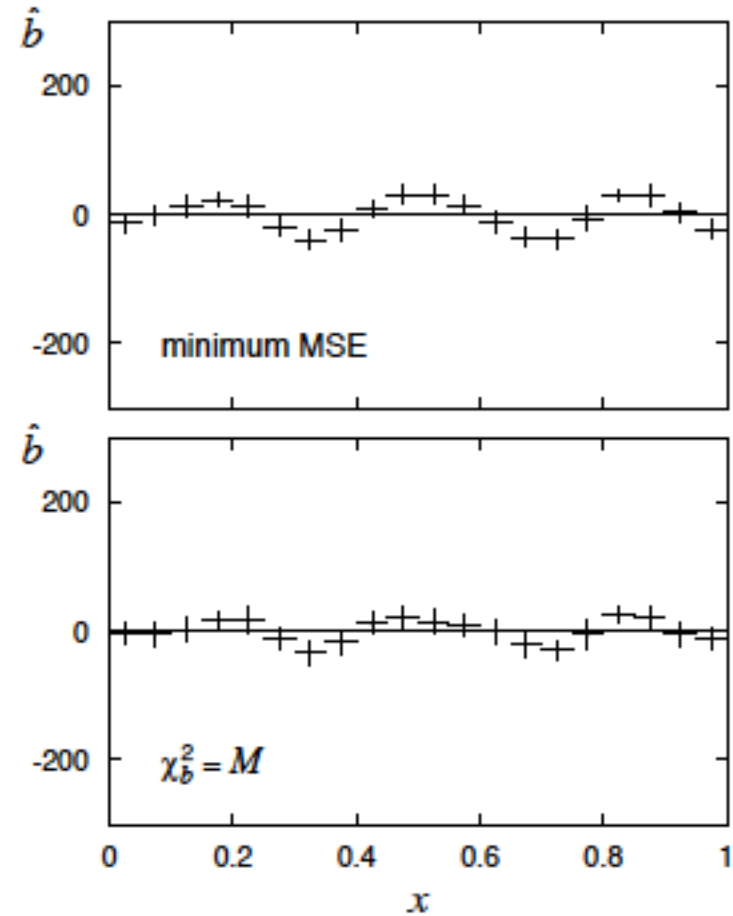
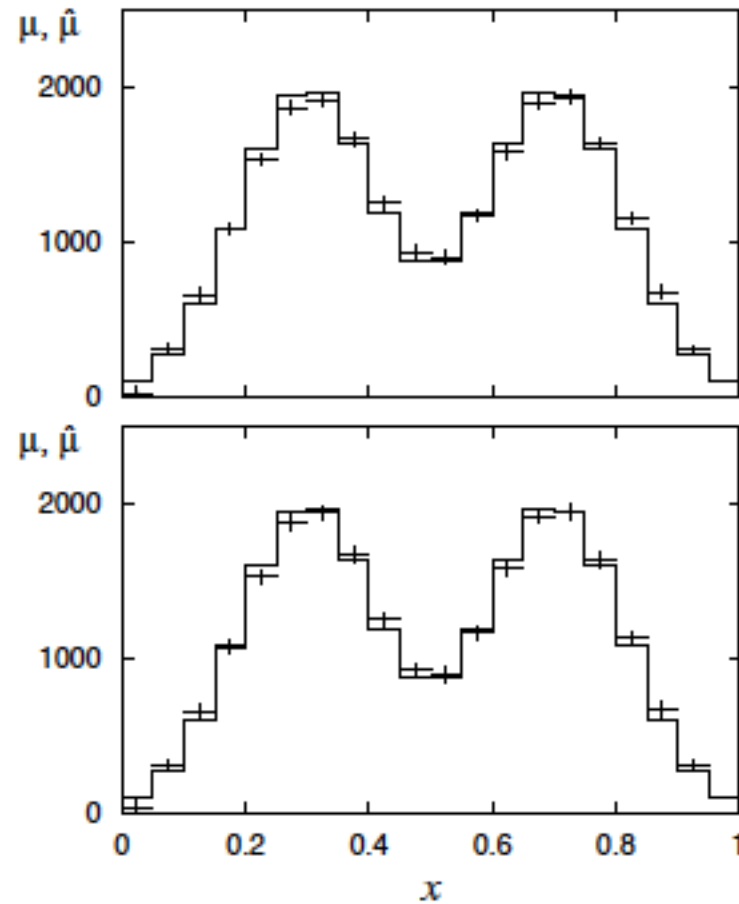
In the limit of no regularization, the coverage probability should converge to its nominal value  $P_{\text{nom}} = 0.683$ .

So choose the regularization parameter that gives a minimum  $P_{\text{cov}} = P_{\text{nom}} - \varepsilon$  defined by a tolerance parameter  $\varepsilon$ .

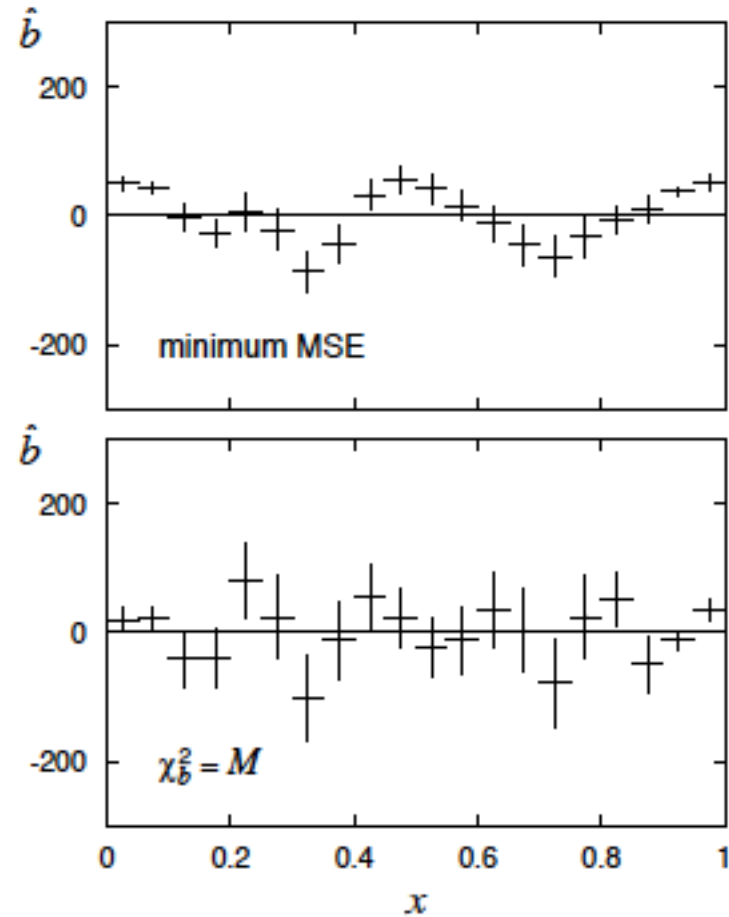
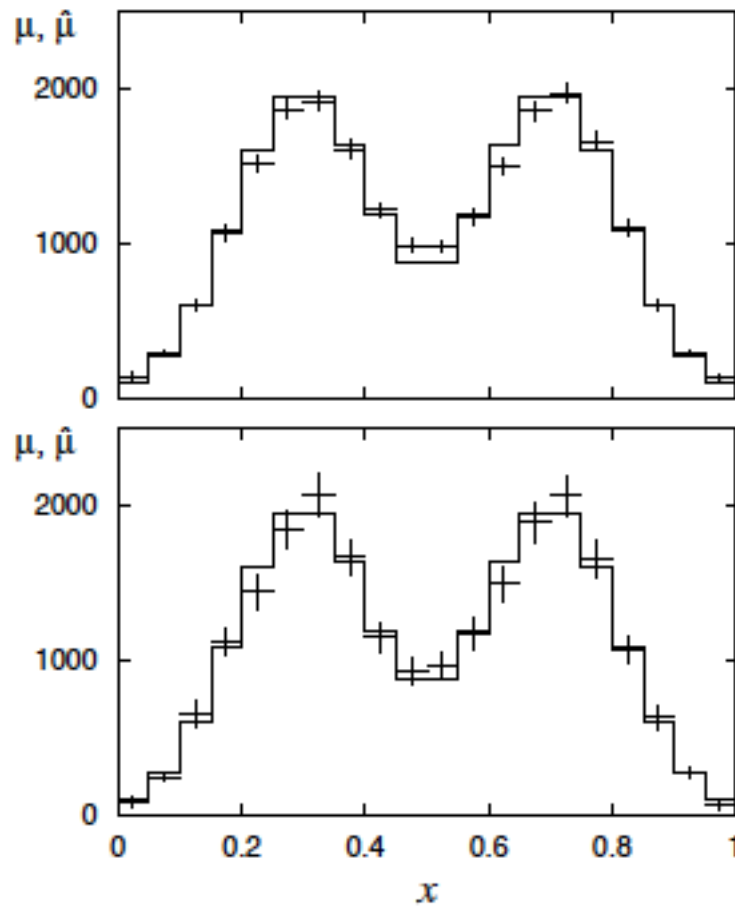
M. Kuusela and V. M. Panaretos, “Statistical unfolding of elementary particle spectra: Empirical Bayes estimation and bias-corrected uncertainty quantification,” *Ann. Appl. Stat.*, vol. 9, pp. 1671–1705, 2015.

Pim Jordi Verschuuren, *Novel Unfolding Methods and Measurements of  $t\bar{t}$  Differential Cross Sections with SMEFT interpretation using the ATLAS detector at the LHC* (PhD thesis), Royal Holloway, University of London, 2021.

# Some examples with Tikhonov regularization



# Some examples with entropy regularization



# Stat. and sys. errors of unfolded solution

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

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

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

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

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

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

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

# Folding

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

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

$$\nu = R\mu + \beta$$

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

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

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

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

# Versus unfolding

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

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

where

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

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

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



# ML solution again

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

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

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

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

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

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

# Summary/discussion on unfolding

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

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

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

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

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

# Some references on unfolding

Lydia Brenner, et al., *Comparison of unfolding methods using RooFitUnfold*, International Journal of Modern Physics A, Vol. 35, No. 24, 2050145 (2020); e-print: arXiv:1910.14654.

P.A. Zyla et al., (PDG), Prog. Theor. Exp. Phys. 2020, 083C01 (2020) and 2021 update; (Sec. 40.2.5 on unfolding).

G. Cowan, Statistical Data Analysis (1998) (Ch. 11).

O. Behnke *et al.*, editors, *Data analysis in high energy physics*, Wiley-VCH, Weinheim, (2013) (Ch. 6 by Volker Blobel).

S. Schmitt, *Data Unfolding Methods in High Energy Physics*, EPJ Web of Conferences **137**, 11008 (2017), e-print arXiv:1611.01927.

G. Cowan, A Survey of Unfolding Methods in Particle Physics, in M. Whalley and L. Lyons (eds.), *Advanced Statistical Techniques in Particle Physics (Proceedings)* Durham, UK, March 18-22, 2002, Conf. Proc. C 0203181 (2002) 248-257.

# Finally...

Estimation of parameters is usually the “easy” part of statistics:

Frequentist: maximize the likelihood.

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

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

But there are many potential stumbling blocks:

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

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

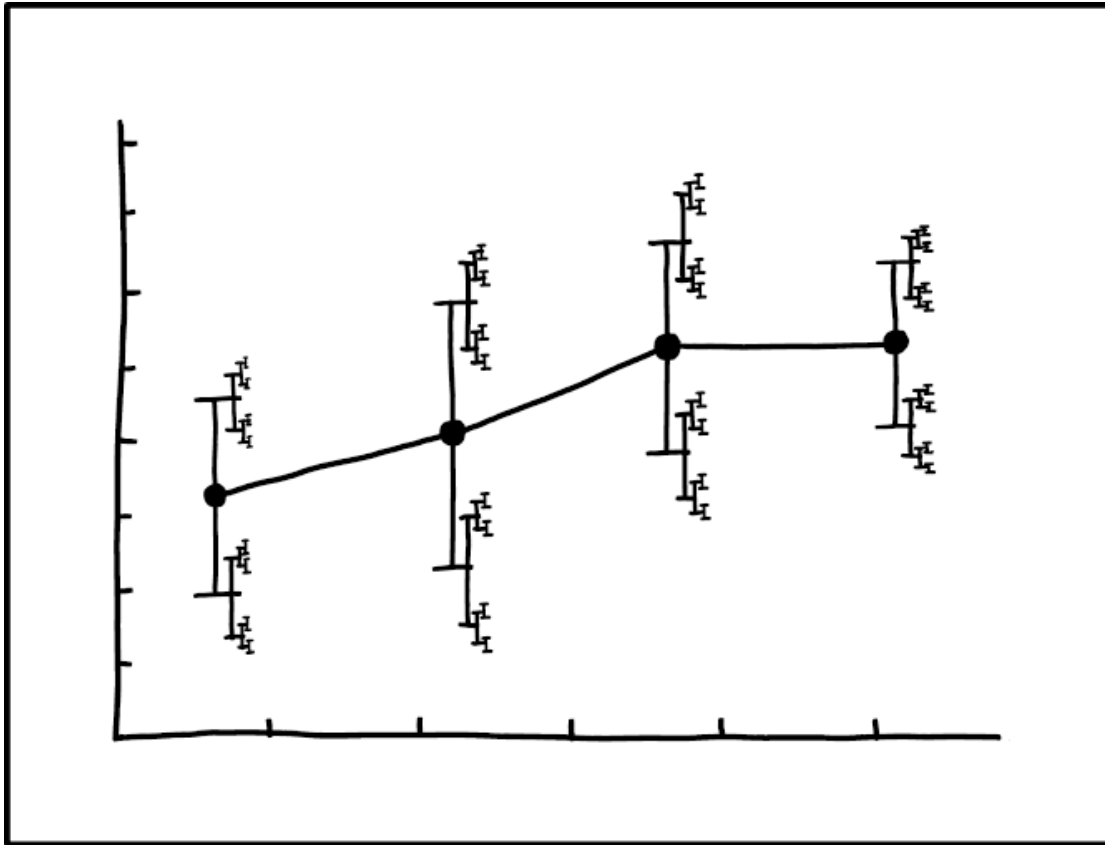
choice of prior for Bayesian approach;

unexpected behaviour in LS averages with correlations,

unknown errors (“errors on errors”)

# Extra Slides

# 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 (→ 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

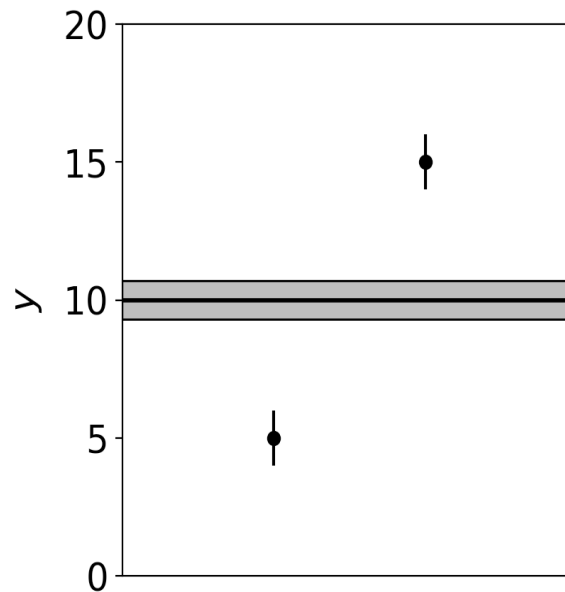
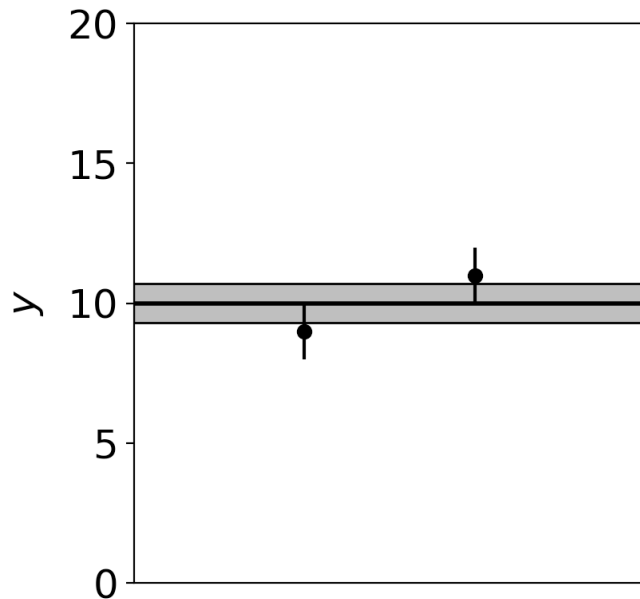
- Any analysis where assigned systematic error is uncertain

# Motivation (2)

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

Least squares average of  $9 \pm 1$  and  $11 \pm 1$  is  $10 \pm 0.71$

Least squares average of  $5 \pm 1$  and  $15 \pm 1$  is  $10 \pm 0.71$



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



# Formulation of the problem

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

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

$\boldsymbol{\theta}$  = nuisance parameters

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

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

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  $\alpha$  and  $\beta$  so that they give desired relative uncertainty  $r$  in  $\sigma_u$ .

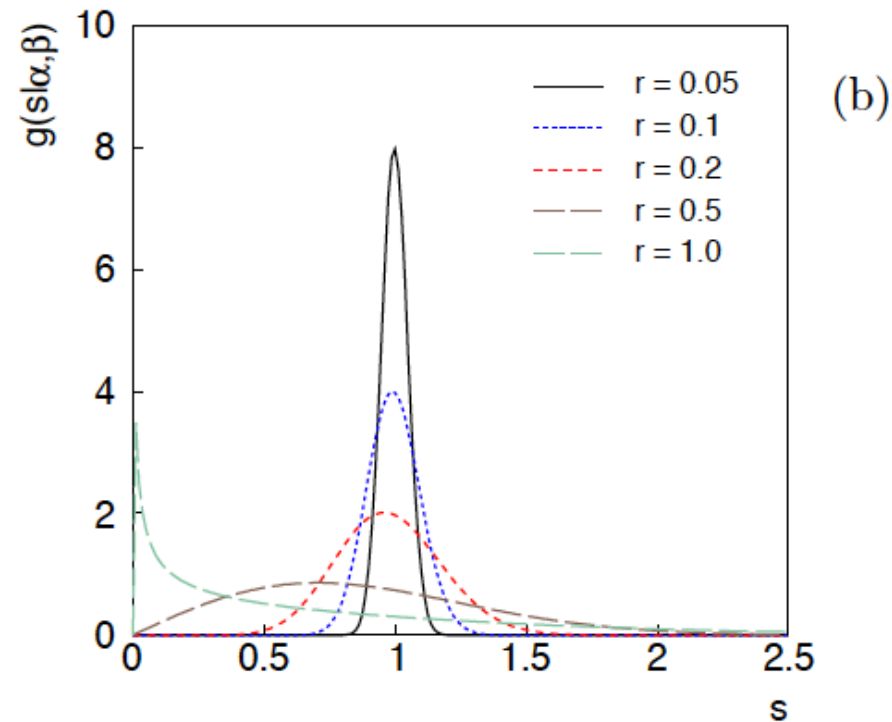
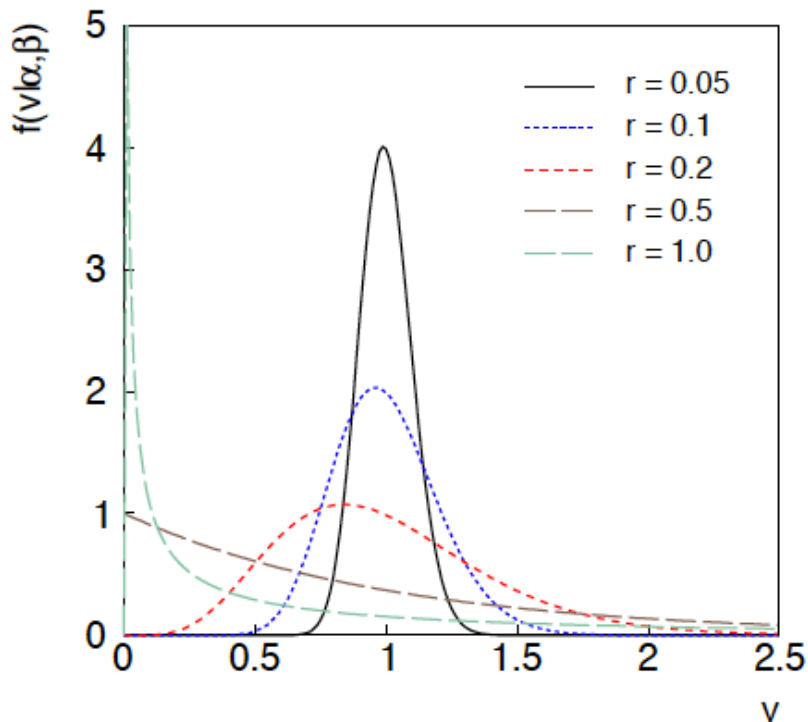
Other "bell-shaped" models tried; qualitatively similar results.

Gamma pdf leads to important mathematical simplifications.

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

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

$$r_i \equiv \frac{1}{2} \frac{\sigma_{v_i}}{E[v_i]} = \frac{1}{2} \frac{\sigma_{v_i}}{\sigma_{u_i}^2} \approx \frac{\sigma_{s_i}}{E[s_i]} \quad \leftarrow \text{relative "error on error"}$$



# Likelihood for Gamma Variance Model

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

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

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

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

Treated like data:  $y_1, \dots, y_L$  (the primary measurements)  
 $u_1, \dots, u_N$  (estimates of nuisance par.)  
 $v_1, \dots, v_N$  (estimates of variances of estimates of NP)

Adjustable parameters:  $\mu_1, \dots, \mu_M$  (parameters of interest)  
 $\theta_1, \dots, \theta_N$  (nuisance parameters)  
 $\sigma_{u,1}, \dots, \sigma_{u,N}$  (sys. errors = std. dev. of of NP estimates)

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

# Profiling over systematic errors

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

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

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

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

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  $z_i \equiv \frac{u_i - \theta_i}{\sqrt{v_i}}$

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

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

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

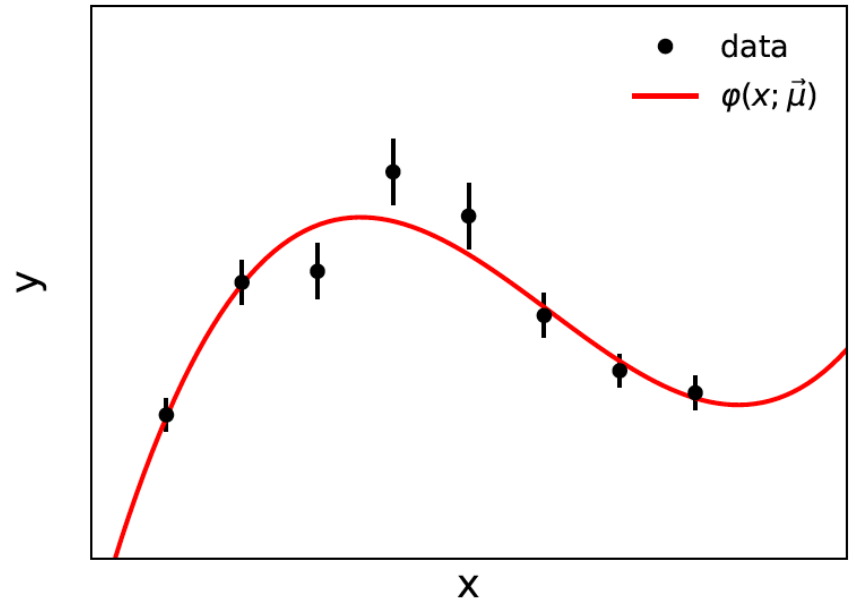
$$L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \prod_{i=1}^N \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\sqrt{\nu_i\pi}\Gamma(\nu_i/2)} \left(1 + \frac{z_i^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}}$$

# Curve fitting, averages

Suppose independent  
 $y_i \sim \text{Gauss}$ ,  $i = 1, \dots, N$ , 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  
 $u_i \sim \text{Gauss}(\theta_i, \sigma_{u,i})$ , 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 $\theta_i$ with known $\sigma_{u,i}$

Profiling over the bias parameters  $\theta_i$  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  $\theta_i$  but profiling over  $\sigma_{u,i}^2$  gives

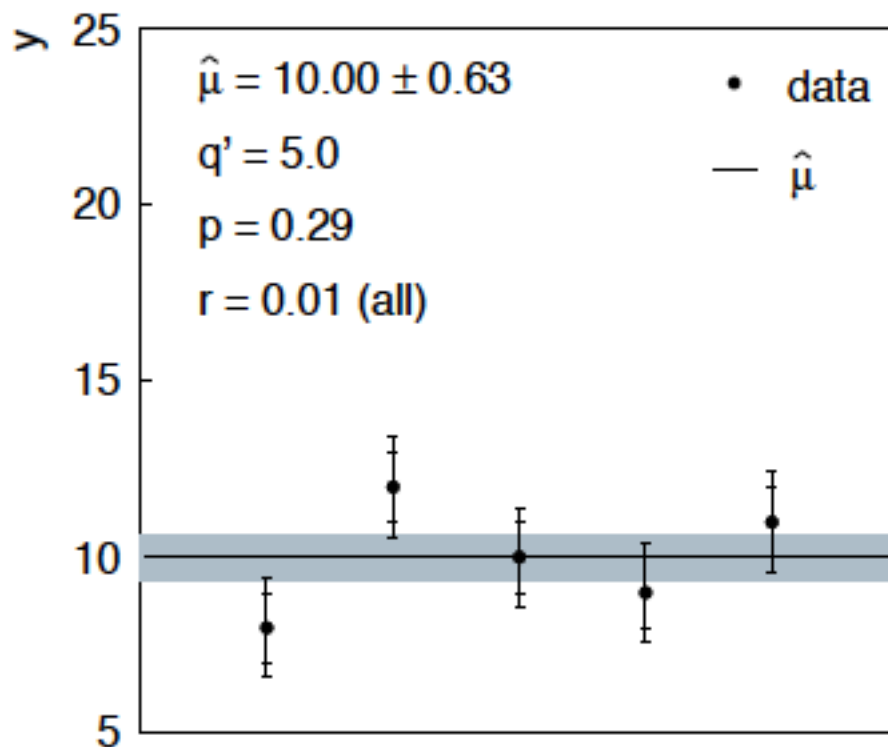
$$-2 \ln L'(\boldsymbol{\mu}, \boldsymbol{\theta}) = \sum_{i=1}^N \left[ \frac{(y_i - \varphi(x_i; \boldsymbol{\mu}) - \theta_i)^2}{\sigma_{y_i}^2} + \left(1 + \frac{1}{2r_i^2}\right) \ln \left(1 + 2r_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right) \right]$$

Profiled values of  $\theta_i$  from solution to cubic equations

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

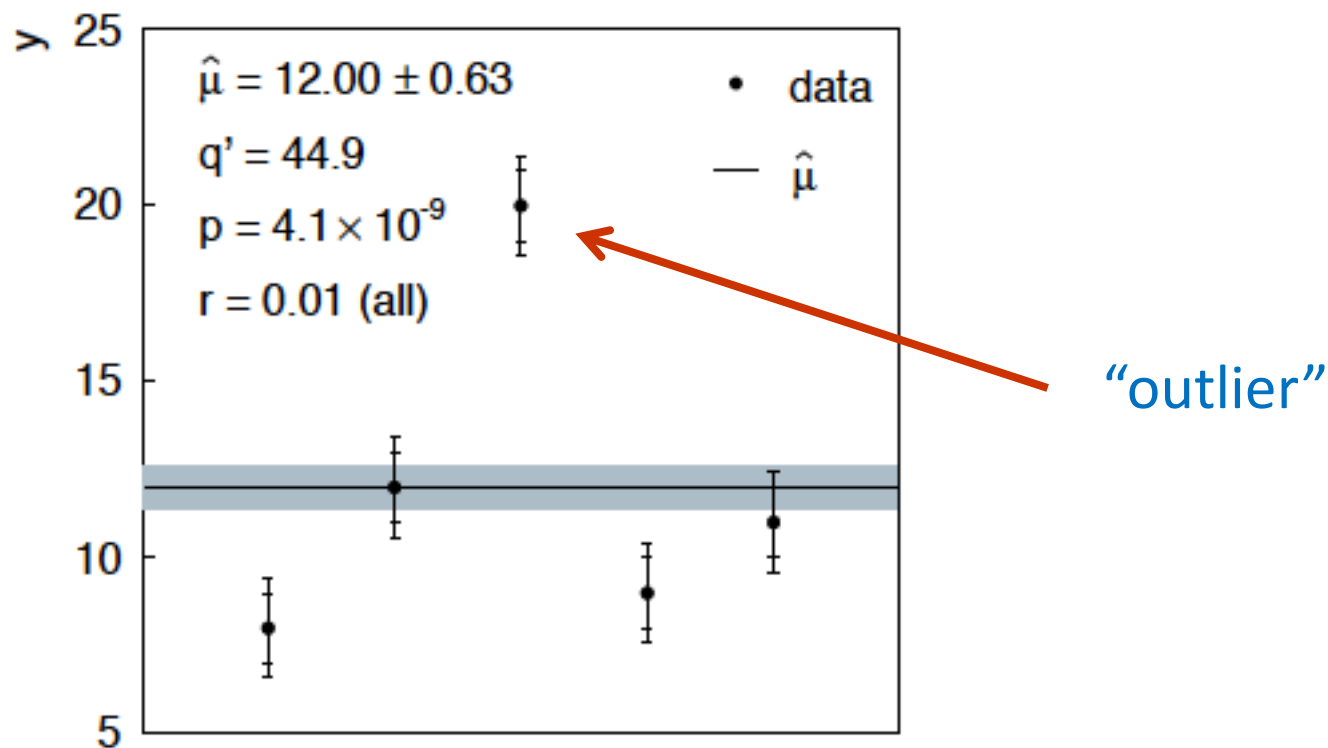
# Sensitivity of average to outliers

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)

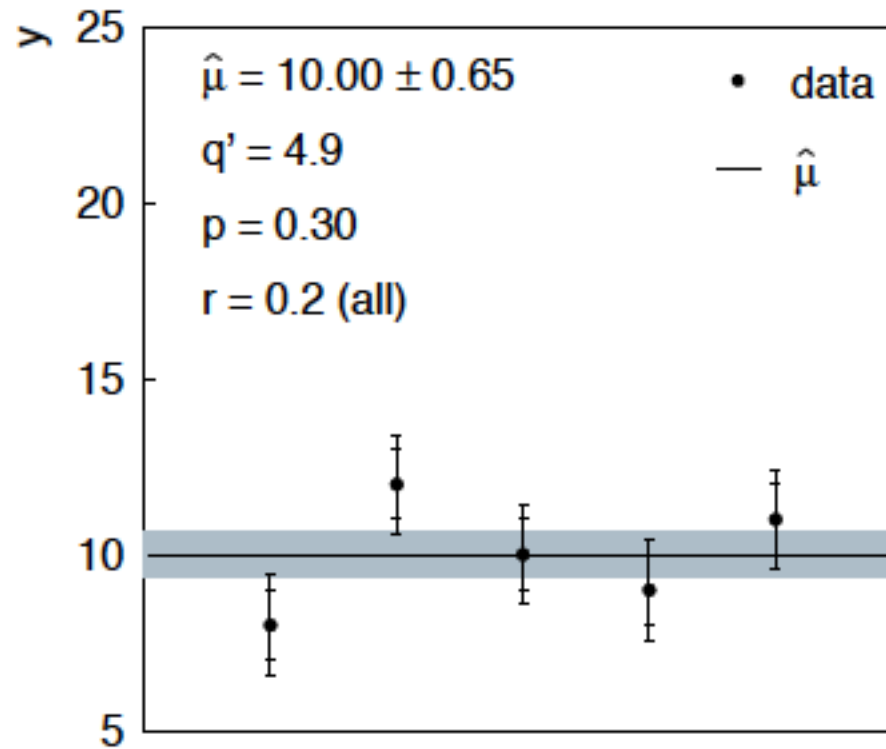
Now suppose the measurement at 10 was actually at 20:



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

# Average with all $r = 0.2$

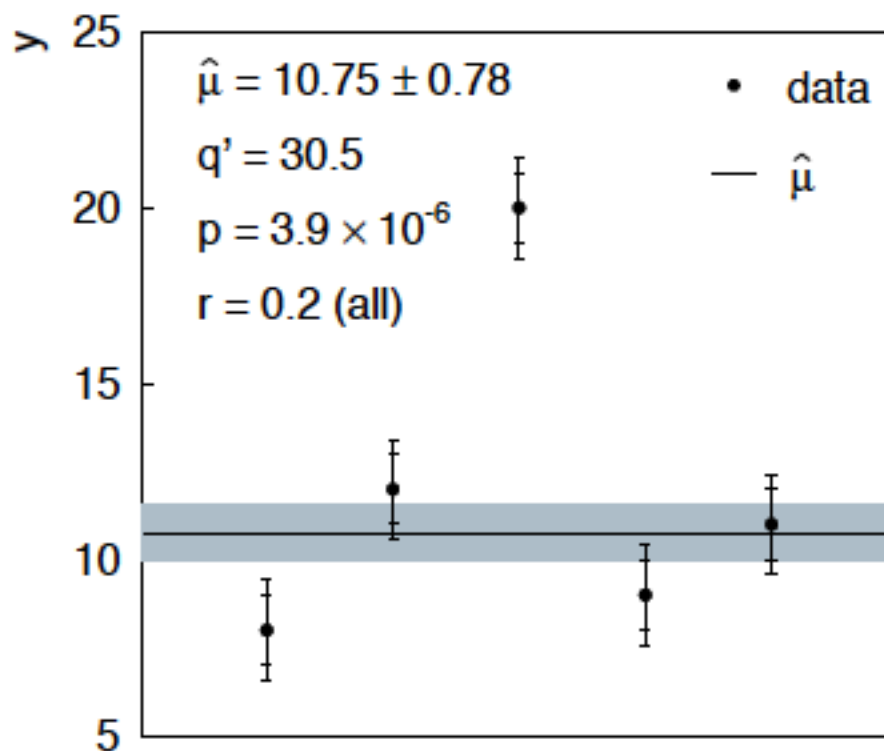
If we assign to each measurement  $r = 0.2$ ,



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

# 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.,

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

But this gives

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

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

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:

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