

Statistical Methods for Particle Physics

Lecture 4: unfolding

<https://indico.weizmann.ac.il/conferenceDisplay.py?confId=52>



Statistical Inference for Astro
and Particle Physics Workshop
Weizmann Institute, Rehovot
March 8-12, 2015



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

Outline for Monday – Thursday

(GC = Glen Cowan, KC = Kyle Cranmer)

Monday 9 March

GC: probability, random variables and related quantities

KC: parameter estimation, bias, variance, max likelihood

Tuesday 10 March

KC: building statistical models, nuisance parameters

GC: hypothesis tests I, p -values, multivariate methods

Wednesday 11 March

KC: hypothesis tests 2, composite hyp., Wilks', Wald's thm.

GC: asymptotics 1, Asimov data set, sensitivity

Thursday 12 March:

KC: confidence intervals, asymptotics 2

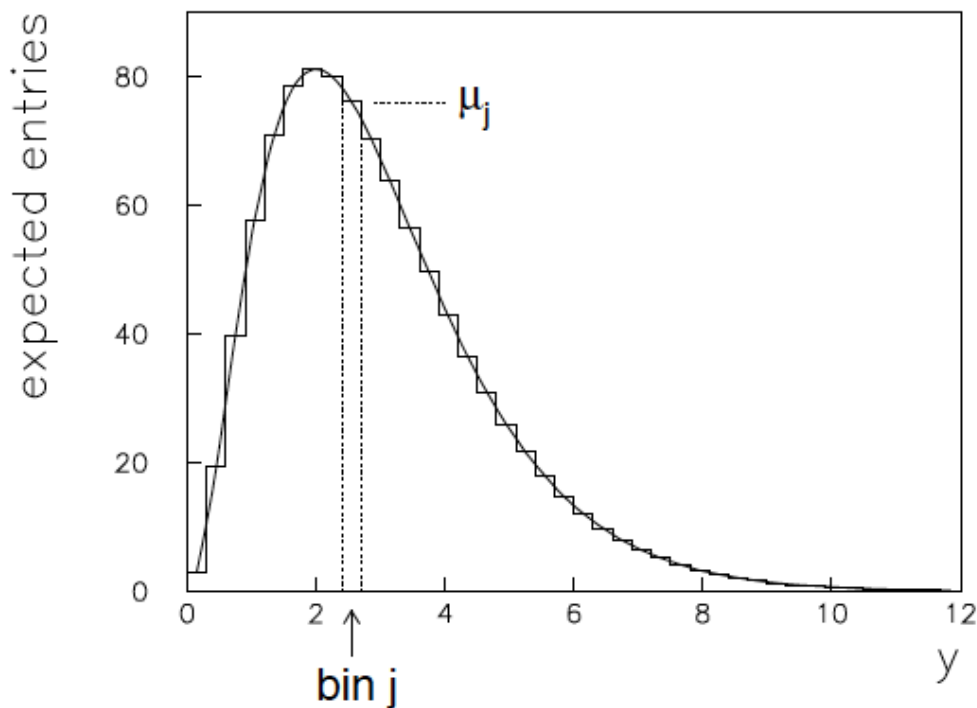
GC: unfolding

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, construct histogram:



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

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

“true” histogram

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

Migration

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

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



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

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

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

← response matrix

Note μ , ν are constants; \mathbf{n} subject to statistical fluctuations.

Efficiency, background

Sometimes an event goes undetected:

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

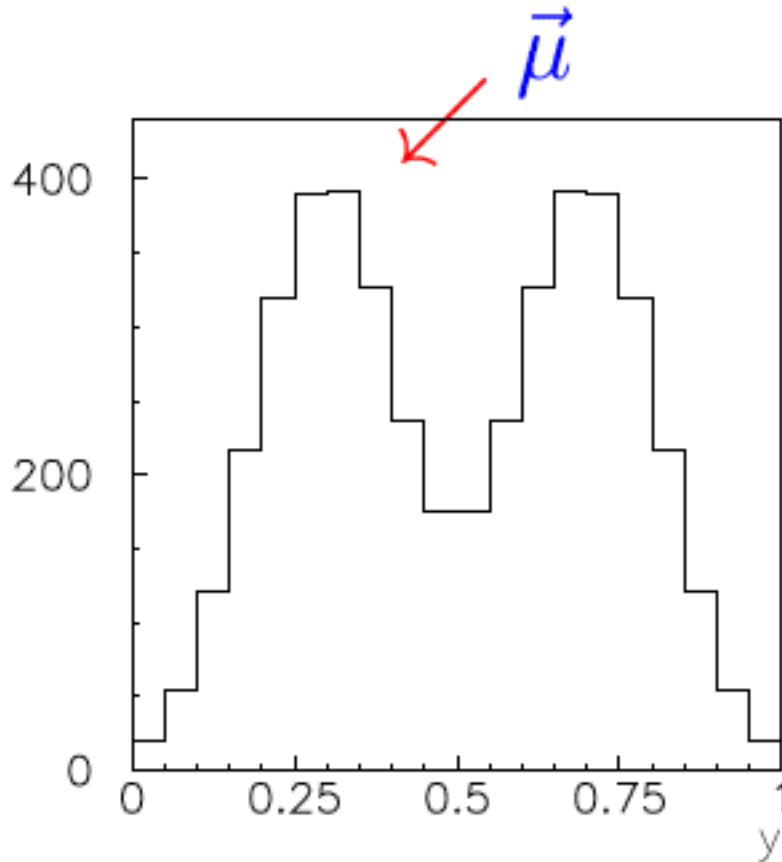
Sometimes an observed event is due to a background process:

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

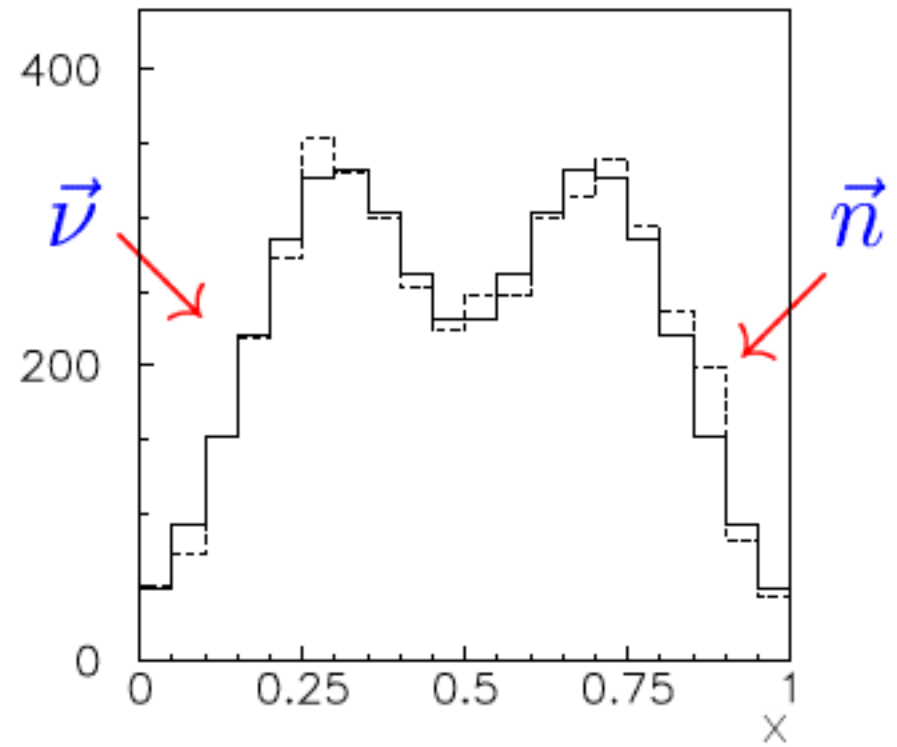
β_i = expected number of background events in *observed* histogram.

For now, assume the β_i are known.

The basic ingredients



“true”



“observed”

Summary of ingredients

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

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: $\boldsymbol{\mu} = R^{-1}(\boldsymbol{\nu} - \boldsymbol{\beta})$

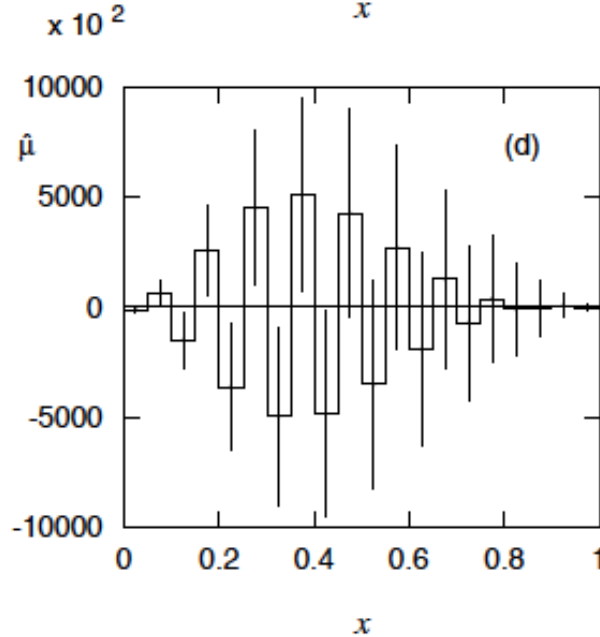
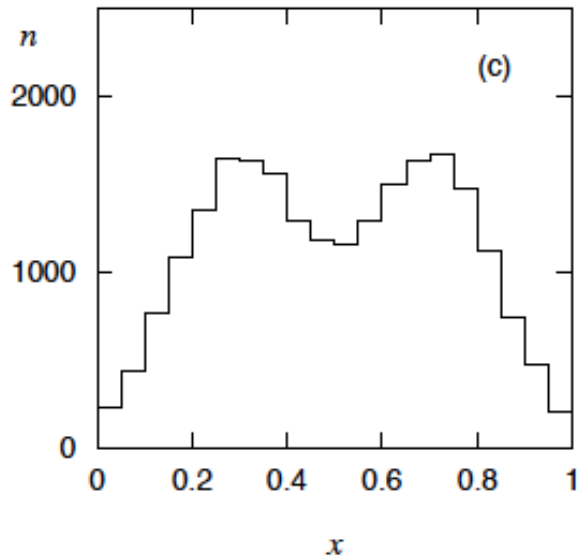
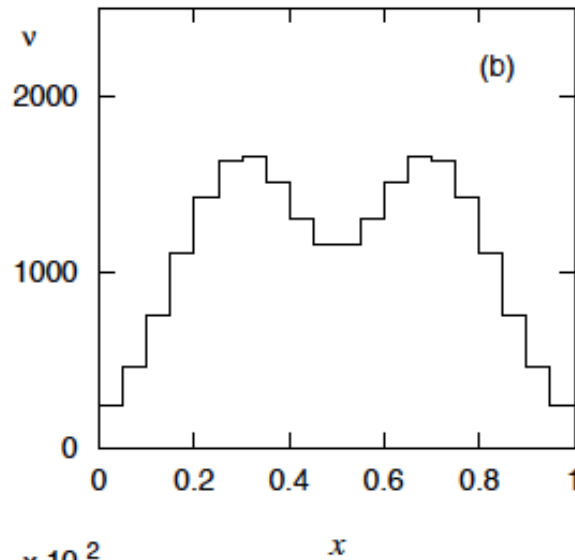
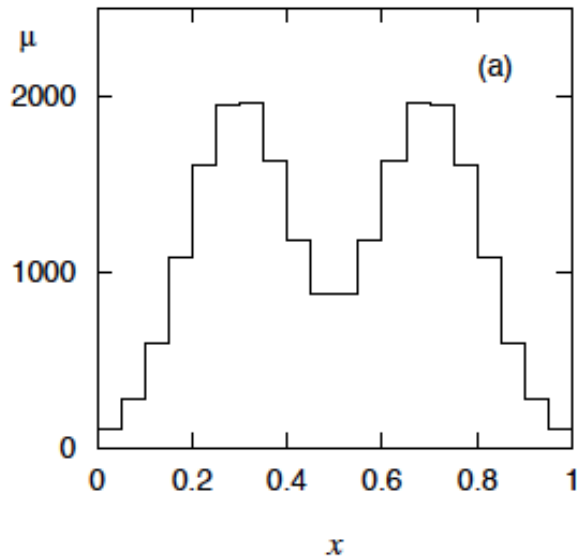
Suppose data are independent Poisson: $P(n_i; \nu_i) = \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$

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

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

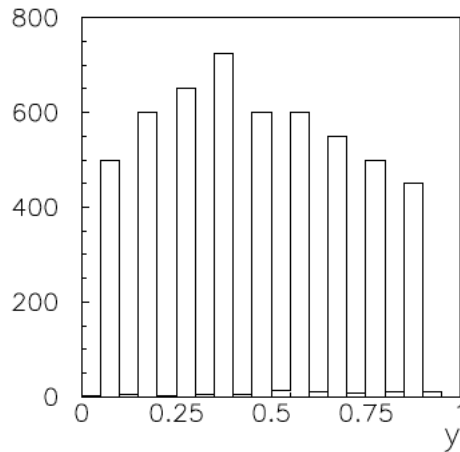
$$\longrightarrow \hat{\boldsymbol{\mu}} = R^{-1}(\mathbf{n} - \boldsymbol{\beta})$$

Example with ML solution



Catastrophic failure???

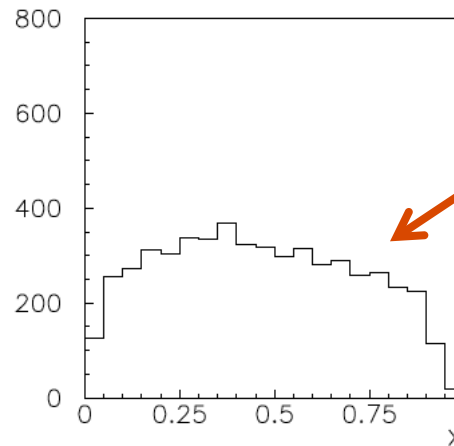
What went wrong?



Suppose μ really had a lot of fine structure.

← $\vec{\mu}$

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



← $\vec{v} = R\vec{\mu}$

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

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

ML solution revisited

For Poisson data the ML estimators are unbiased:

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

Their covariance is:

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

(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:

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

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

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; here the bias is difficult to quantify (difficult also for many other unfolding methods).

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

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

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

Regularized unfolding

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

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

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),

α = 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 λ 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 α .

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 φ equal to zero gives linear equations.

Solution using Singular Value Decomposition (SVD).

SVD implementation of Tikhonov unfolding

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

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

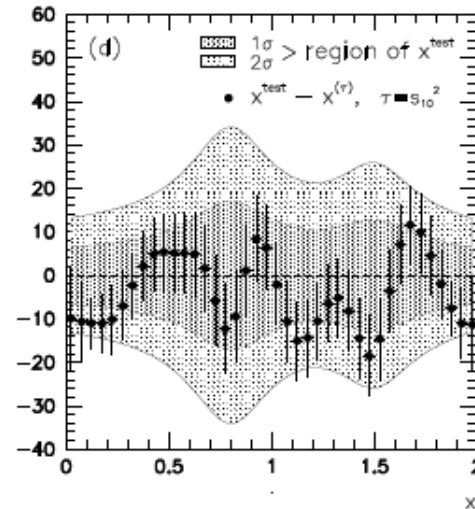
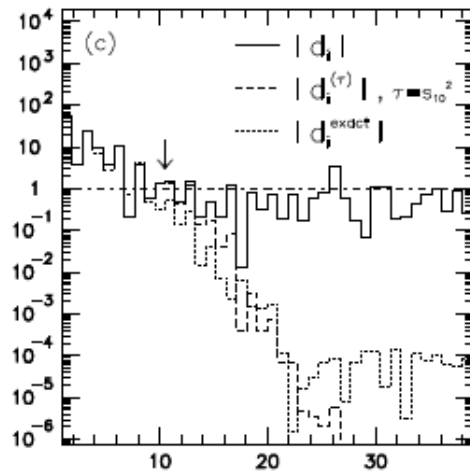
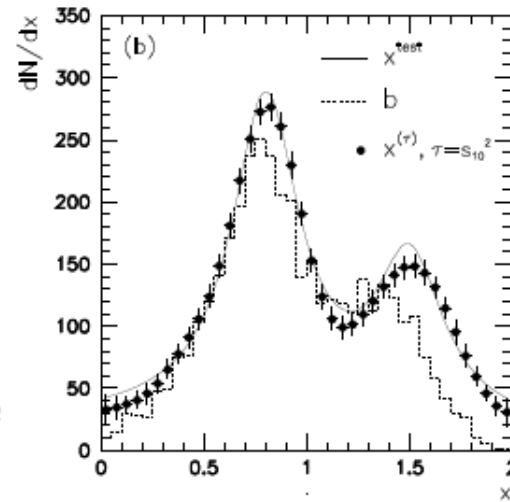
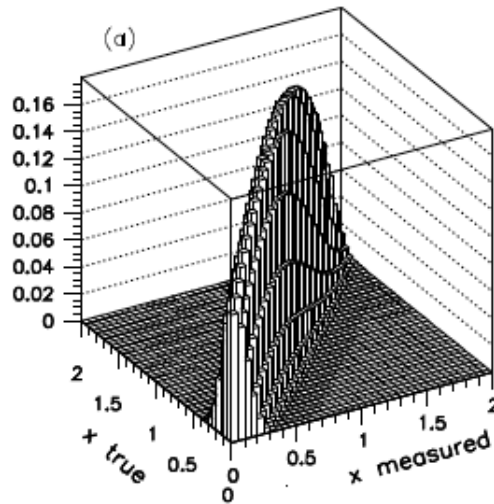
Numerical implementation using Singular Value Decomposition.

Recommendations for setting regularization parameter τ :

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

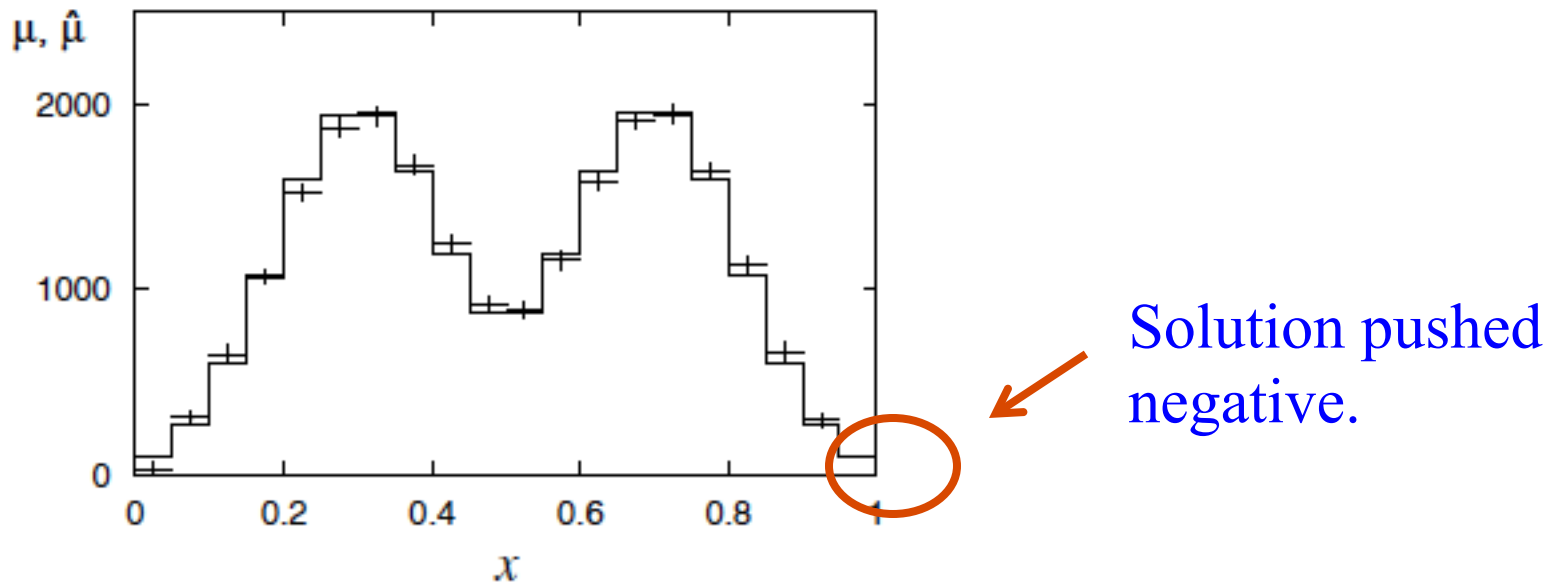
SVD example

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



Edge effects

Regularized unfolding can lead to “edge effects”. E.g. in Tikhonov regularization with Gaussian data, solution can go negative:



Important e.g. if New Physics would appear as a longer tail of a distribution.

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

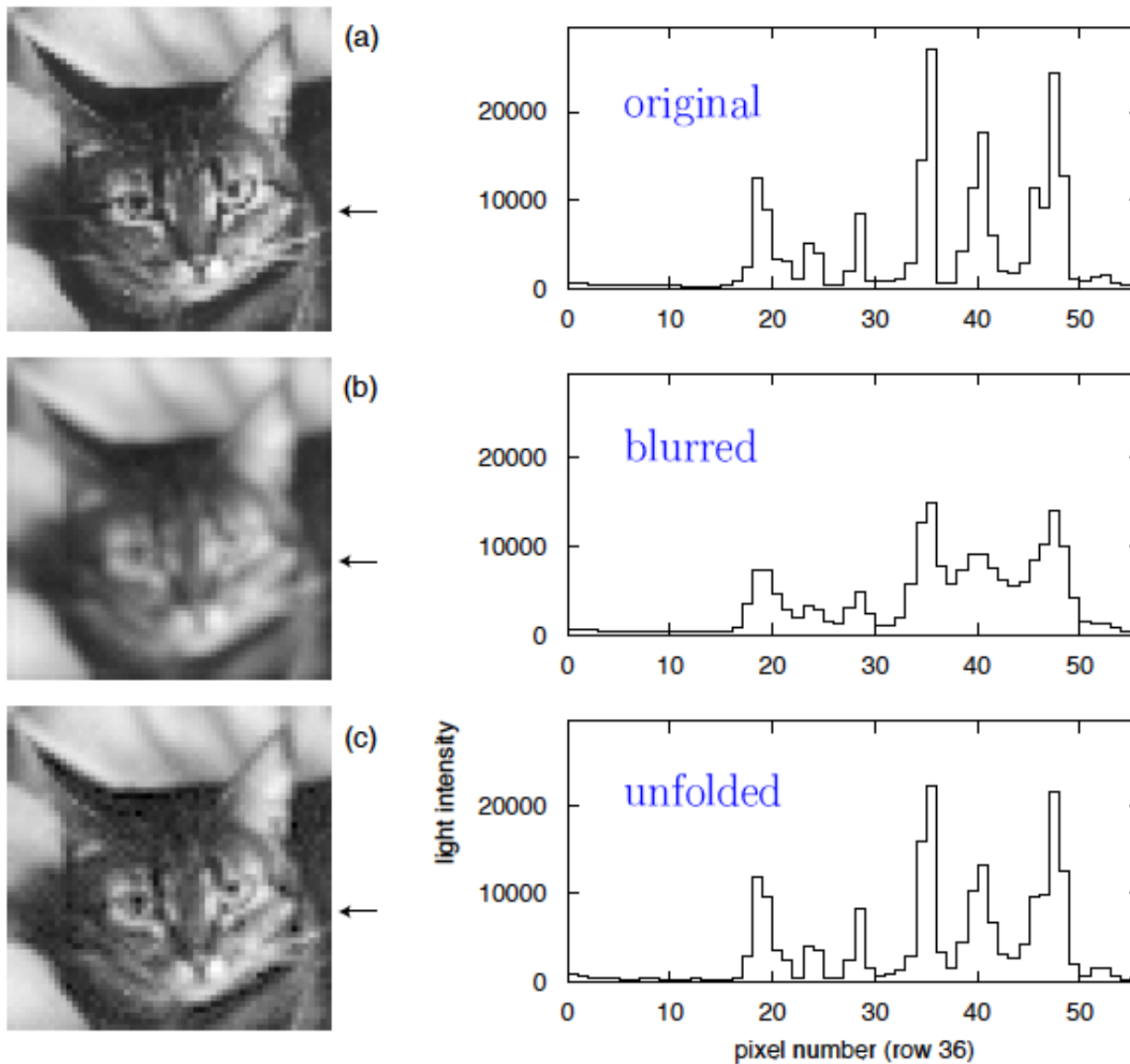
Use entropy as regularization function,

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

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

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}_{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 χ^2 from unregularized (ML) solution,

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

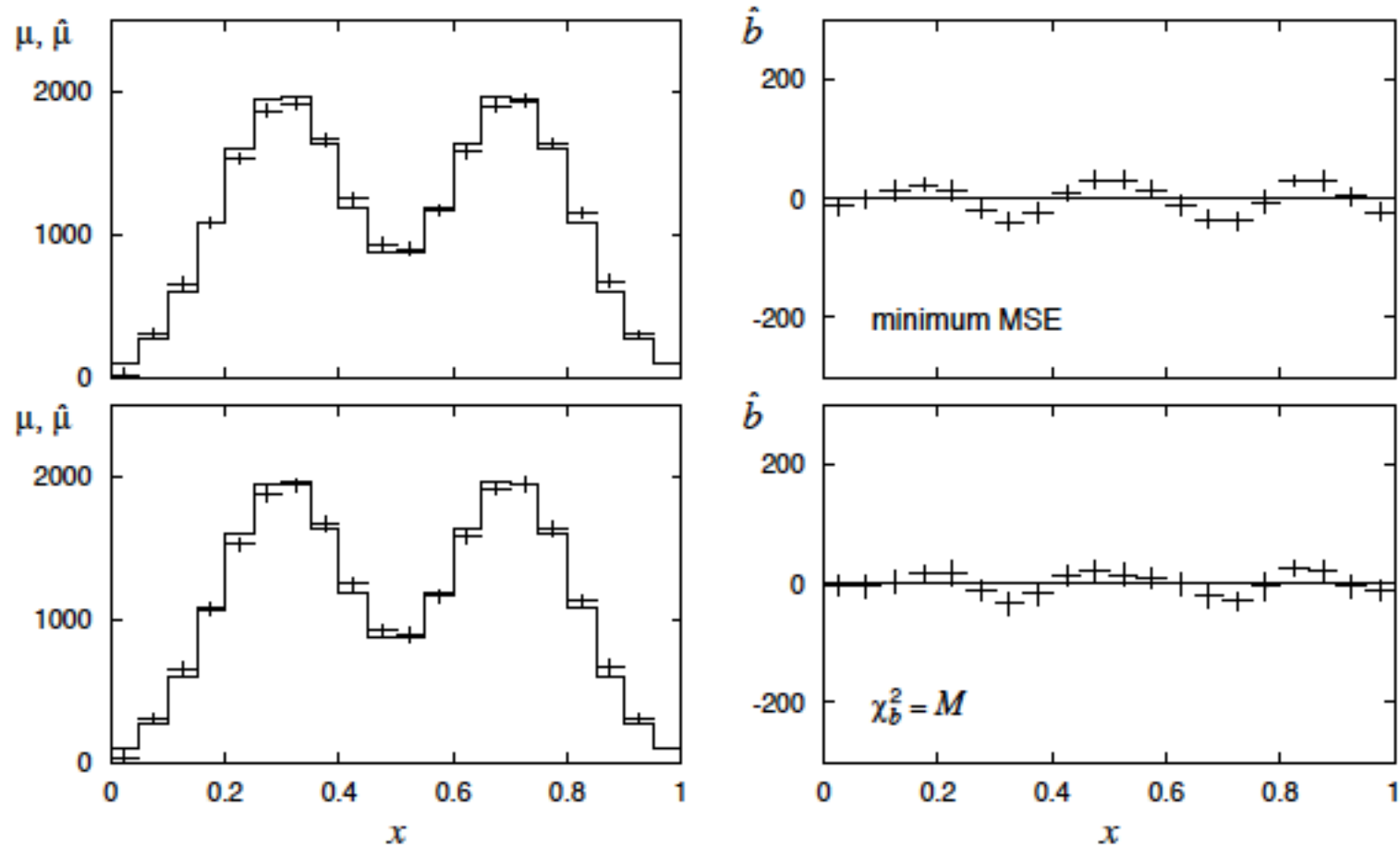
Or require that bias be consistent with zero to within its own error,

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

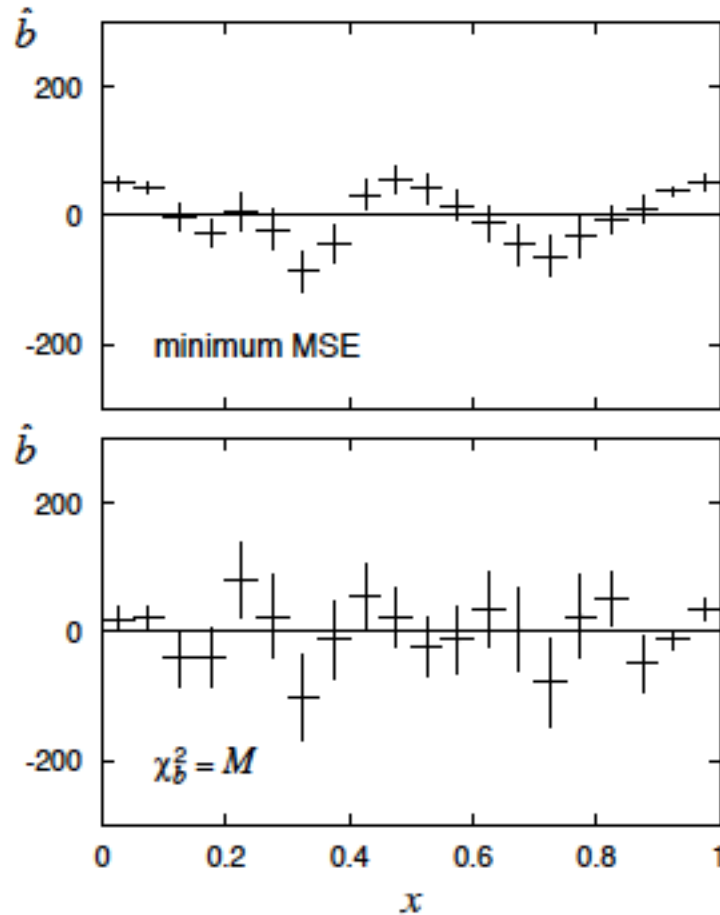
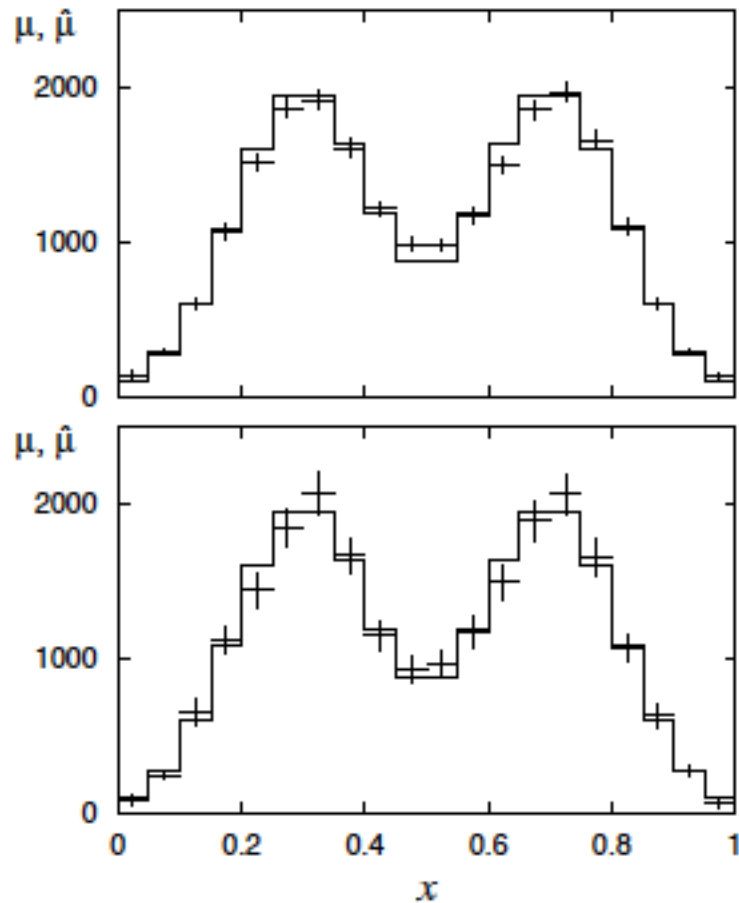
i.e. if bias significantly different from zero, we would subtract it;

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

Some examples with Tikhonov regularization



Some examples with entropy regularization



Iterative unfolding (“Bayesian”)

G. D’Agostini, NIM A **362** (1995) 487; see also arXiv:1010.0632.

Goal is to estimate probabilities: $\mathbf{p} = (p_1, \dots, p_M)$


For initial guess take e.g. $p_i = 1/M$

Initial estimators for μ are $\hat{\mu}_0 = n_{\text{tot}} \mathbf{p}_0$,

Update according to the rule

$$\hat{\mu}_i = \frac{1}{\varepsilon_i} \sum_{j=1}^N P(\text{true value in bin } i \mid \text{found in bin } j) n_j$$

uses Bayes’ theorem here


$$= \frac{1}{\varepsilon_i} \sum_{j=1}^N \left(\frac{R_{ij} p_i}{\sum_k R_{jk} p_k} \right) n_j$$

Continue until solution stable using e.g. χ^2 test with previous iteration.

Estimating systematic uncertainty

We know that unfolding introduces a bias, but quantifying this (including correlations) can be difficult.

Suppose a model predicts a spectrum

$$f(y; \theta) \sim 1/y^\theta \rightarrow \mu(\theta)$$

A priori suppose e.g. $\theta \approx 8 \pm 2$. More precisely, assign prior $\pi(\theta)$. Propagate this into a systematic covariance for the unfolded spectrum:

$$U_{ij}^{(\theta)} = \int (\hat{\mu}_i - \mu_i(\theta))(\hat{\mu}_j - \mu_j(\theta)) \pi(\theta) d\theta \quad \text{(Typically large positive correlations.)}$$

Often in practice, one doesn't have $\pi(\theta)$ but rather a few models that differ in spectrum. Not obvious how to convert this into a meaningful covariance for the unfolded distribution.

Stat. and sys. errors of unfolded solution

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

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

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

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

To test a new theory depending on parameters θ , 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_{sys} itself has large uncertainty.

Folding

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

Given the response matrix R and expected background β , 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 μ compute likelihood

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

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_{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

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.

Extra slides

Some unfolding references

H. Prosper and L. Lyons (eds.), Proceedings of PHYSTAT 2011, CERN-2011-006; many contributions on unfolding from p 225.

G. Cowan, *A Survey Of Unfolding Methods For Particle Physics*, www.ippp.dur.ac.uk/old/Workshops/02/statistics/proceedings/cowan.pdf

G. Cowan, *Statistical Data Analysis*, OUP (1998), Chapter 11.

V. Blobel, *An Unfolding Method for High Energy Physics Experiments*, arXiv:hep-ex/0208022v1

G. D'Agostini, *Improved iterative Bayesian unfolding*, arXiv:1010.0632

G. Choudalakis, *Fully Bayesian Unfolding*, arXiv:1201.4612

Tim Adye, *Unfolding algorithms and tests using RooUnfold*, arXiv:1105.1160v1

Dependence of response matrix on physics model

Often to good approximation, the response matrix is independent of the physics model, i.e.,

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

is by construction independent of how often the model puts an event in bin j .

But it can still depend on the underlying model. E.g. if one model generates jets very forward, this has different p_T resolution than another model that puts jets mainly in the barrel.

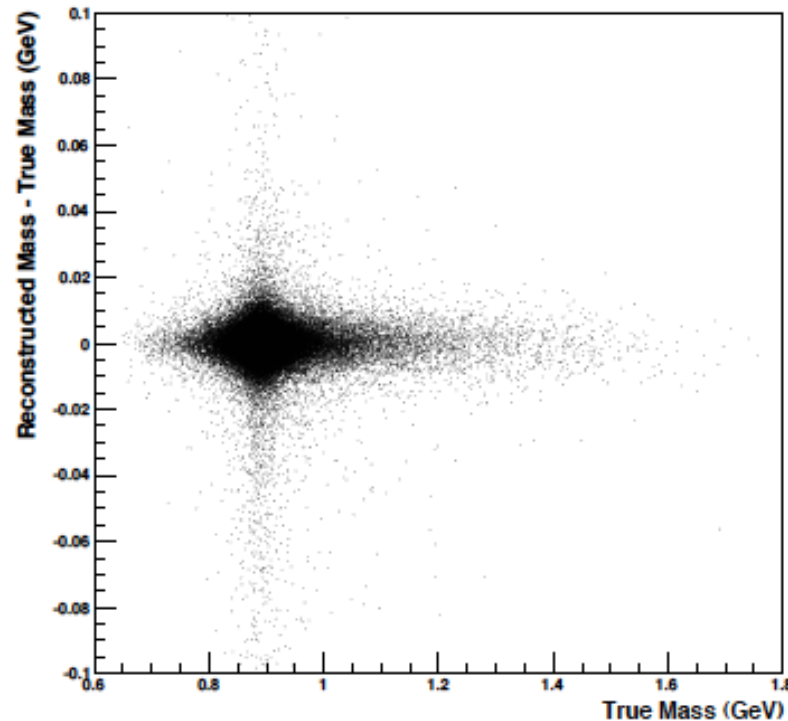
Same problem regardless of whether folding or unfolding.

Can improve situation by measuring both the variable of interest (say, p_T), as well as another variable that correlates with the resolution. Response “matrix” now has 4 indices:

$$R_{ijkl} = P(\text{observed in cell } ij \mid \text{true value in cell } kl)$$

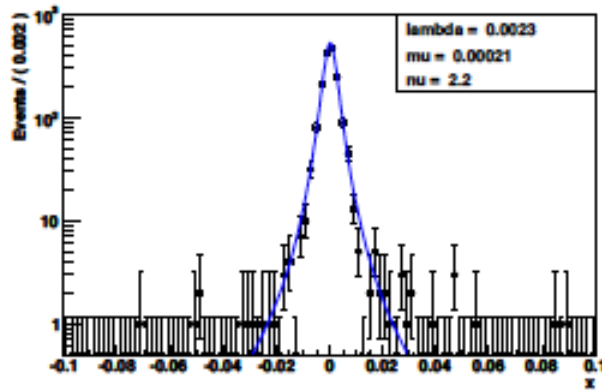
Determining the response matrix

Usually determine response matrix from MC, some smoothing often useful. E.g. first make scatterplot of measured versus true:

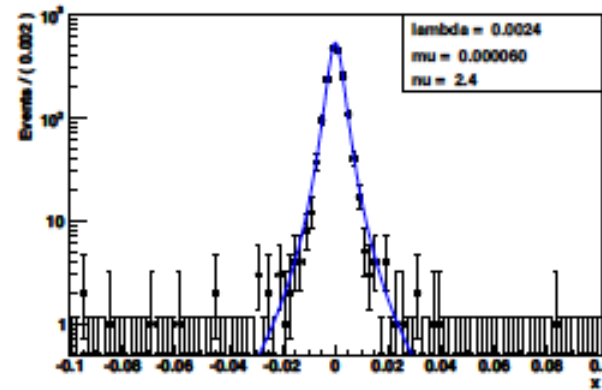


Parameterizing the response matrix

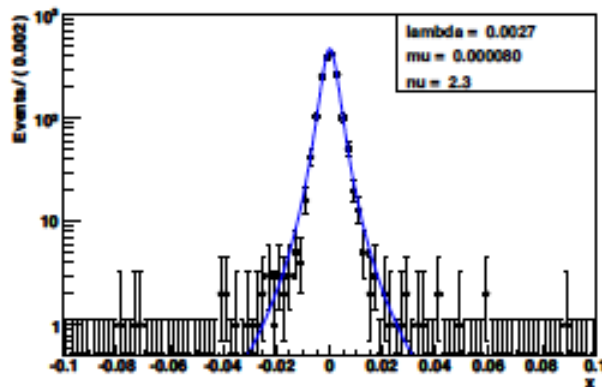
Then cut the scatterplot into vertical slices, and fit each with an appropriate function (here, Student's t distribution):



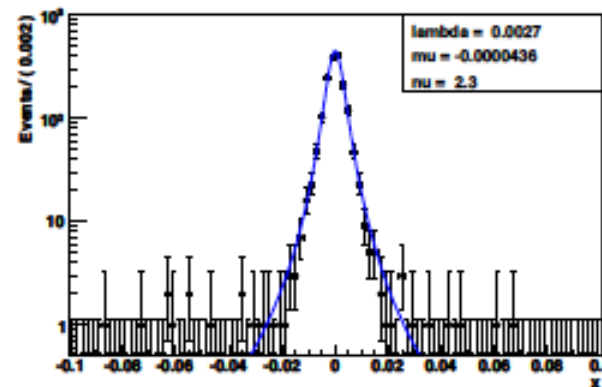
(a) 0.600 - 0.804 GeV



(b) 0.804 - 0.829 GeV



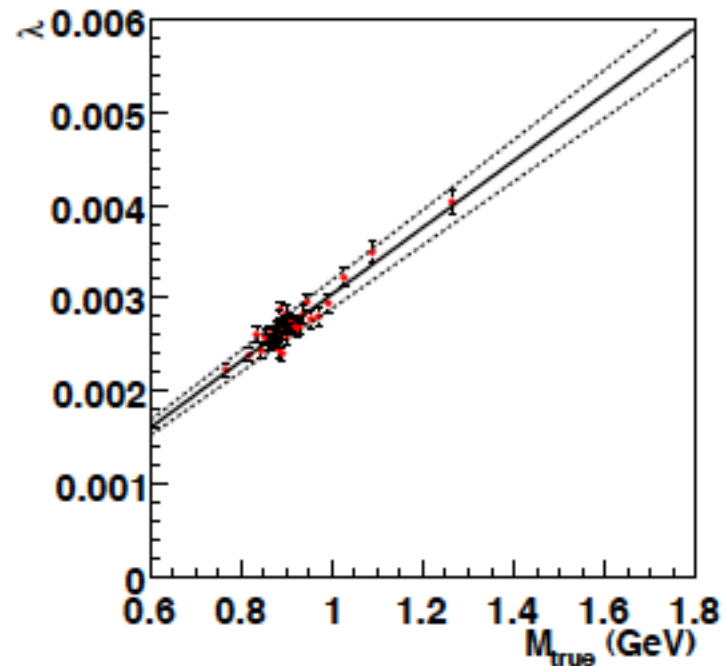
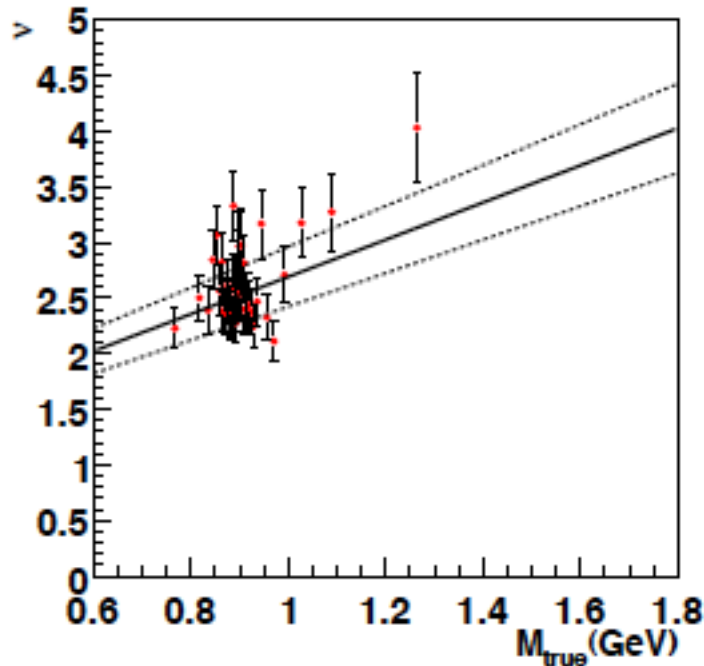
(c) 0.887 - 0.889 GeV



(d) 0.889 - 0.891 GeV

Parameterizing the response matrix (2)

Then fit these parameters as a function of the true variable:



These can then be treated as nuisance parameters and the likelihood function extended to include the constraints on them from knowledge of the detector (e.g., from the fits above or subsidiary measurements).