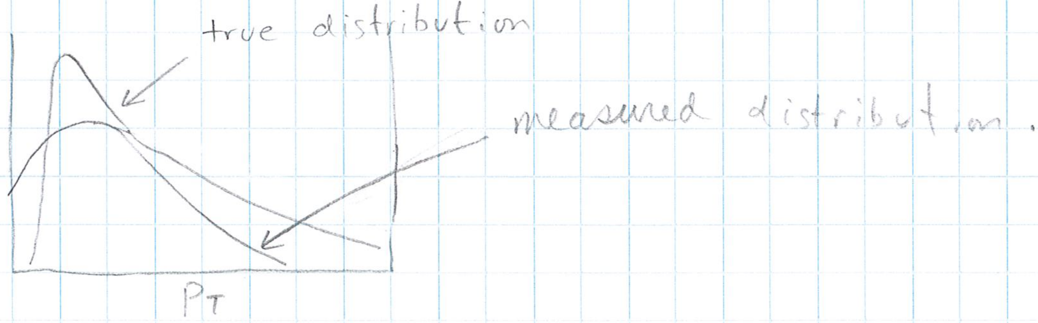
Non parametric models.

Frequently, we may not have a functional form that accurately models the data from an experiment. Examples; an underlying distribution can be modified by detector response which includes

- inefficiency
- bias
- resolution



Efficiencies are frequently calculated using Monte Carlo simulations

$$\varepsilon_i = \frac{N_i^{reco}}{N_i^{Gen}}$$

The "true" distribution is recovered by

$$\frac{dN_i^{true}}{dp_T} = \frac{1}{\varepsilon_i}\frac{dN_i^{meas}}{dp_T}$$

This procedure is frequently applied but has some the drawback that it does not account for migration between bins.

In fact, when Monte Carlo simulations model detector response, an event generated in bin i can be reconstructed in bin j due to migration from resolution or bias.

$$A_{ij} = \frac{N_i^{reco}}{N_j^{gen}}$$

Then the observed number of events in reconstructed bin i is expected to be

$$y_i = \sum_{j=1}^{m} A_{ij} x_j + b_i$$

where x_j is the "true" distribution and b_i is the expected background in bin i.

The general statement of the problem is how best to determine x_i from measurements y_i given estimates for b_i.

Bin-by-bin corrections.

If A_{ij} is diagonal then

$$x_i = \frac{(y_i - b_i)}{A_{ii}}$$

where $A_{ii} = \varepsilon_i$ is the efficiency for reconstructing an event in bin i.

What if A_{ij} is not diagonal.

Naive approach:

$$x_i = \sum_{j=1}^{m} (A^{-1})_{ij} (y_j - b_j)$$

This generally does not work ...

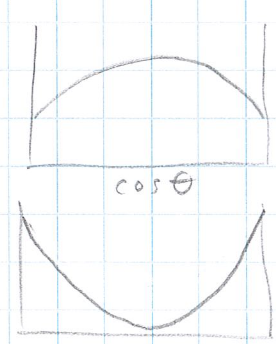
1. There is no requirement that A is a square matrix
2. A could have rows that are all zero (implying lack of information)
3. A may not be invertable.

Template Fits.

Example with angular distribution.

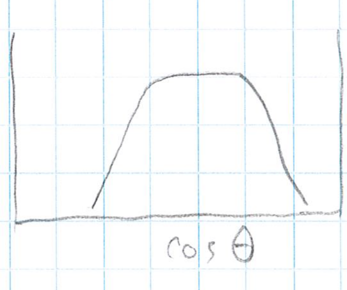
$$\frac{dN_T}{d(\cos\theta)} \propto 1 + \cos^2\theta$$

$$\frac{dN_L}{d(\cos\theta)} \propto 1 - \cos^2\theta$$

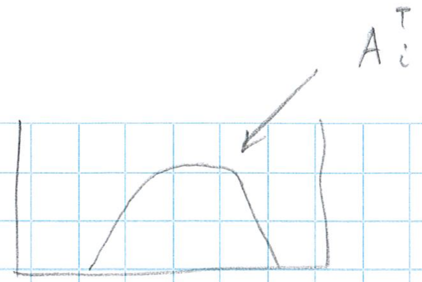


But, the detector may not have uniform acceptance for all cos θ.

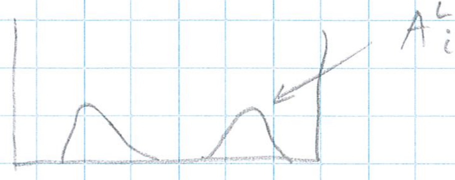
A uniform distribution might be reconstructed as



Transverse template



Longitudinal template



Fit the data:

$$y_i = f A_i^T + (1-f) A_i^L$$

$$\chi^2 = \sum \frac{(m_i - y_i)^2}{y_i}$$

minimize χ^2 wrt α.

Some drawbacks: What if $y_i = 0$?

Then we expect $m_i = 0$. Otherwise the source of the events must be background or mismodeling of the detector response.

Unfolding.

We need to impose our prior knowledge that the "true" distribution is expected to be relatively smooth.

TUnfold algorithm

$$L_1 = (\vec{m} - A\vec{x})^T V^{-1} (\vec{m} - A\vec{x})$$

This is just a χ^2 function. Minimize with respect to parameters $\vec{x}$.

Require that if m_i, $i = 1, \ldots, m$
and x_i, $i = 1, \ldots, n$
then $n < m$.

This has the same statistical fluctuations inherent in the matrix inversion case.

In fact, $\frac{\partial L_1}{\partial \vec{x}} = -2 A^T V^{-1} (\vec{m} - A\vec{x}) = 0$

$$A^T V^{-1} A \vec{x} = A^T V^{-1} \vec{m}$$

$$\vec{x} = (A^T V^{-1} A)^{-1} A^T V^{-1} \vec{m}$$

Introduce a regularization procedure to damp wild departures from a smooth function

$$L_2 = \tau^2 (\vec{x} - f_b \vec{x}_0)^T L^T L (\vec{x} - f_b \vec{x}_0)$$

Typical cases:

Bias vector $f_b \vec{x}_0 = 0$.
L = identity matrix.

Then $L_2 = \tau^2 |\vec{x}|^2$

This penalizes the minimization when $|\vec{x}|$ becomes very large.

If $f_b = 1$ then deviations from $\vec{x}_0$ are suppressed:

$$L_2 = \tau^2 |\vec{x} - \vec{x}_0|^2 .$$

L could also reflect the curvature of $\vec{x}$.

We can also require that observed number of events is constrained:

$$L_3 = \lambda (Y - \vec{e}^T \vec{x})$$

$$e_j = \sum_i A_{ij}$$

e_j is the efficiency for reconstructing true bin j in any of the reconstructed bins.

Still a bit arbitrary ... how best to select τ and $\vec{x}_0$?

When $L = L_1 + L_2$

$$\vec{x} = E\left(A^T V^{-1} \vec{m} + \tau^2 L^T L f_b \vec{x}_0\right)$$

$$E = \left(A^T V^{-1} A + \tau^2 L^T L\right)^{-1}$$

When $L = L_1 + L_2 + L_3$

$$\vec{x}' = \vec{x} + \frac{\lambda}{2} E \vec{e}$$

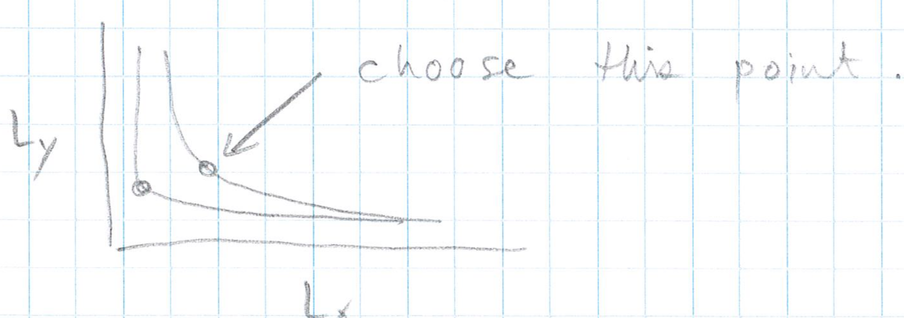
$$\lambda = \frac{Y - e^T \vec{x}}{\vec{e}^T E \vec{e}}$$

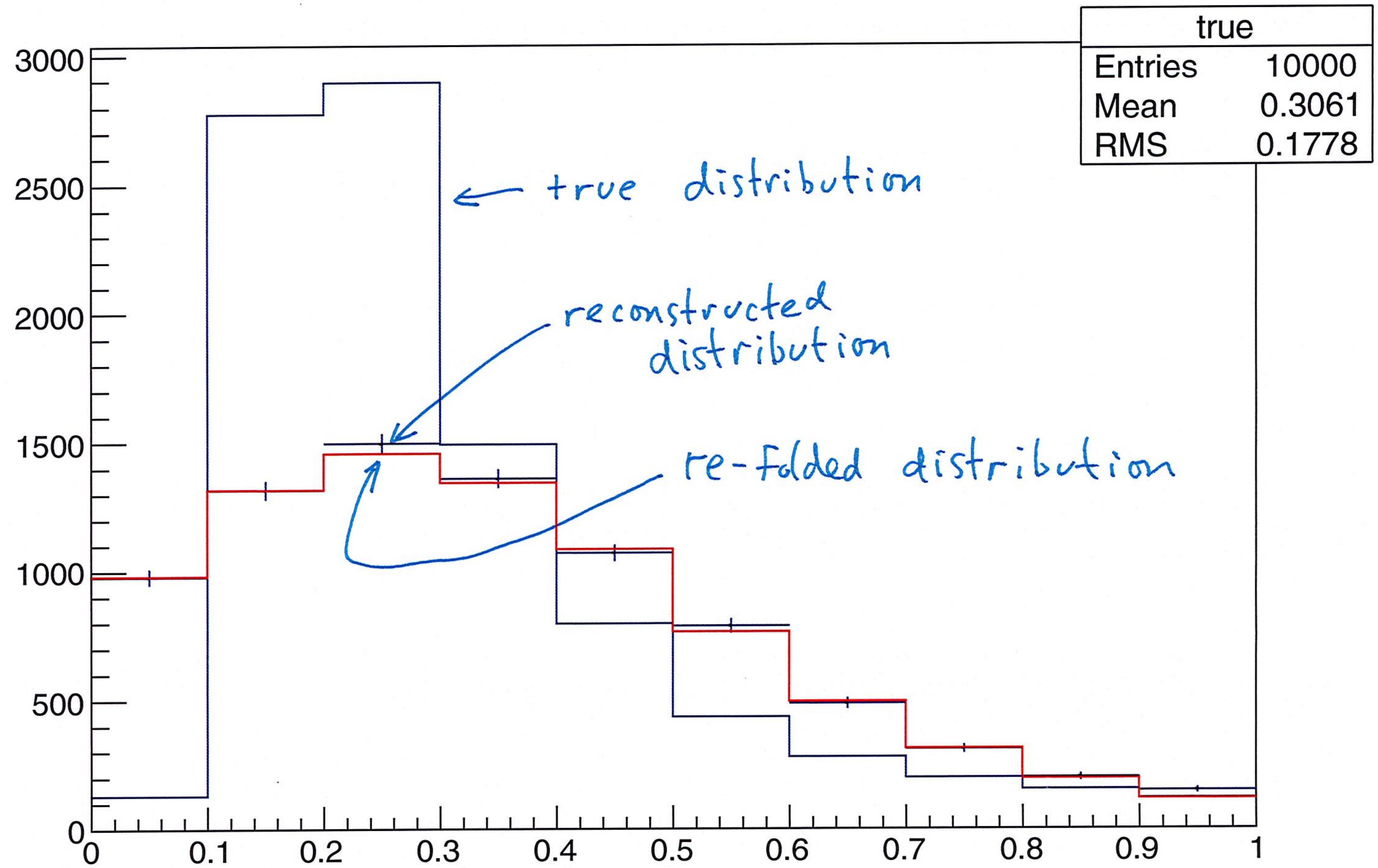
Still, how best to choose τ ?

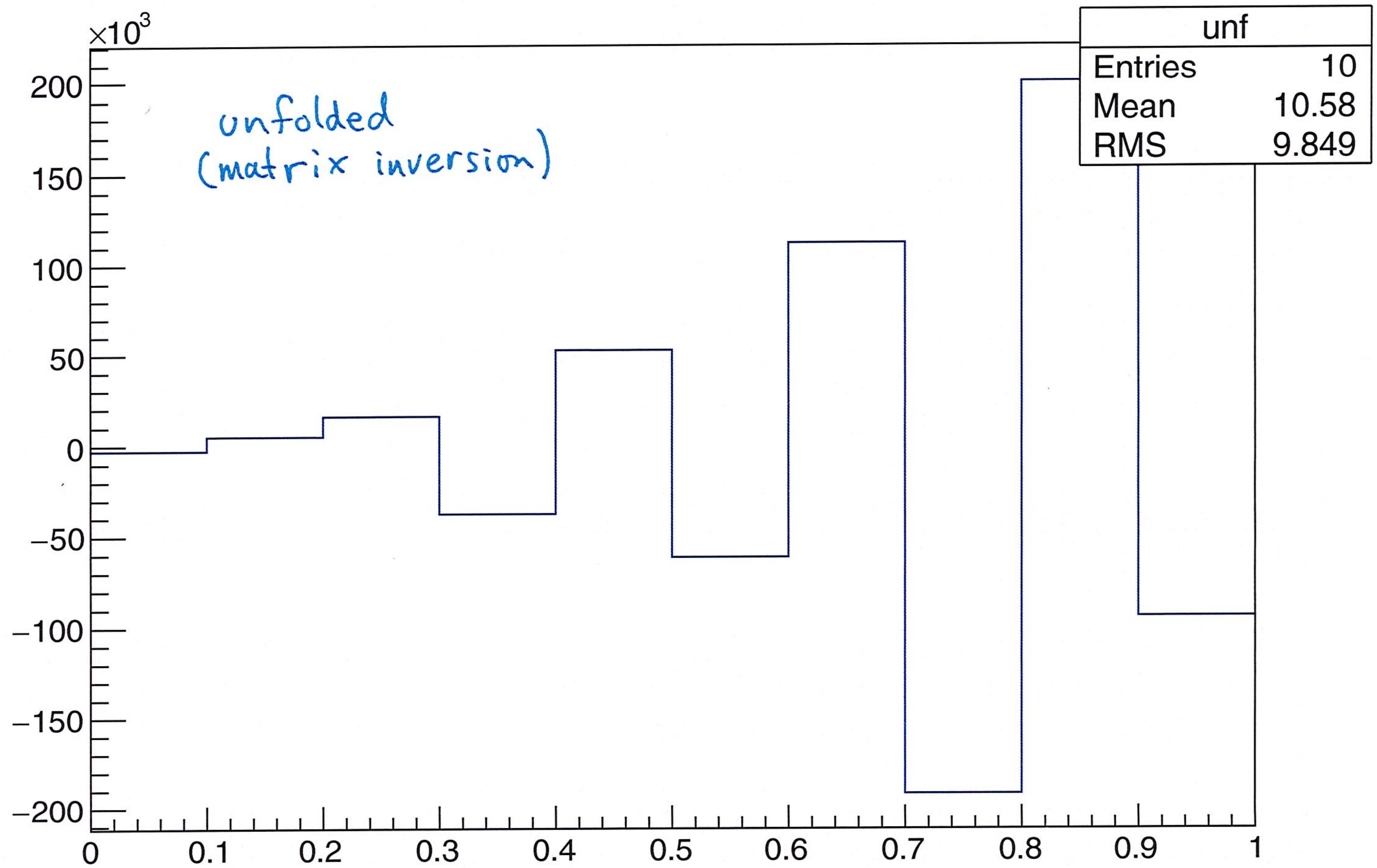
Find the point that is most sensitive to changes in τ.

$$L_x = \log L_1$$

$$L_y = \log \frac{L_2}{\tau^2}$$




true
Entries 10000
Mean 0.3061
RMS 0.1778
true distribution
reconstructed distribution
re-folded distribution
3000
2500
2000
1500
1000
500
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1



unf
Entries 10
Mean 10.58
RMS 9.849
×10^3
200
150
100
50
0
−50
−100
−150
−200
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
unfolded
(matrix inversion)