FNSPE CTU in Prague

Unfolding

Decin 2018

Roman Lavička

May 2, 2018, Decin

Supervisor: Guillermo Contreras

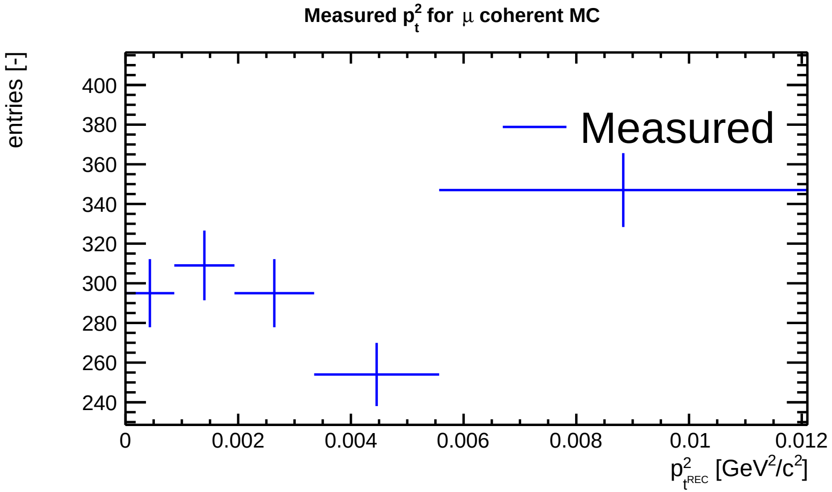
This work has been partially supported by the grant 18-07880S of the Czech Science Foundation (GACR).

Content

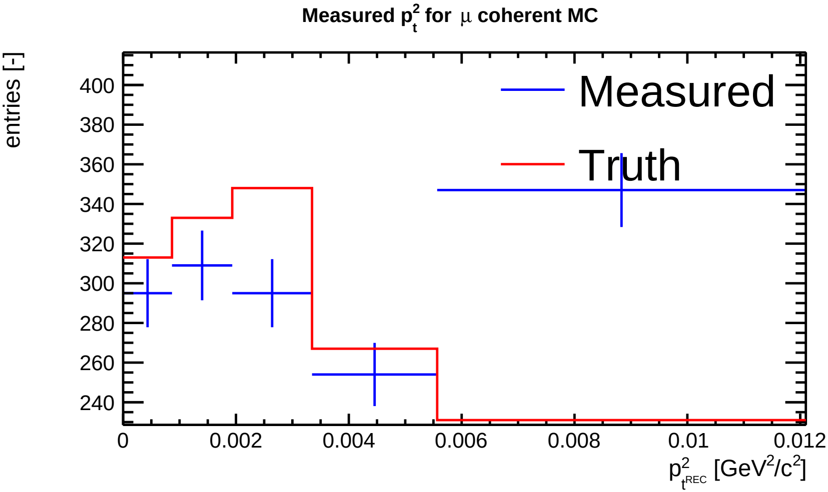
Measurement

Analysis

What you have measured

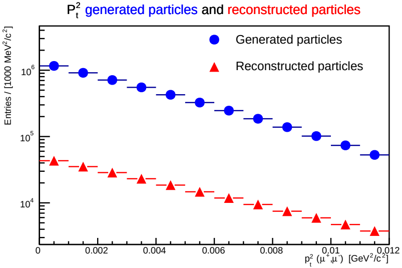


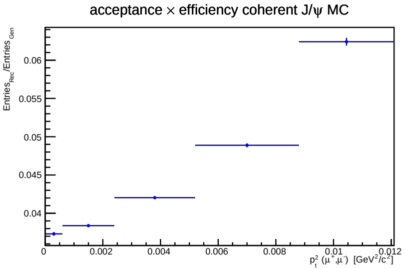
What really happened



How and why to correct

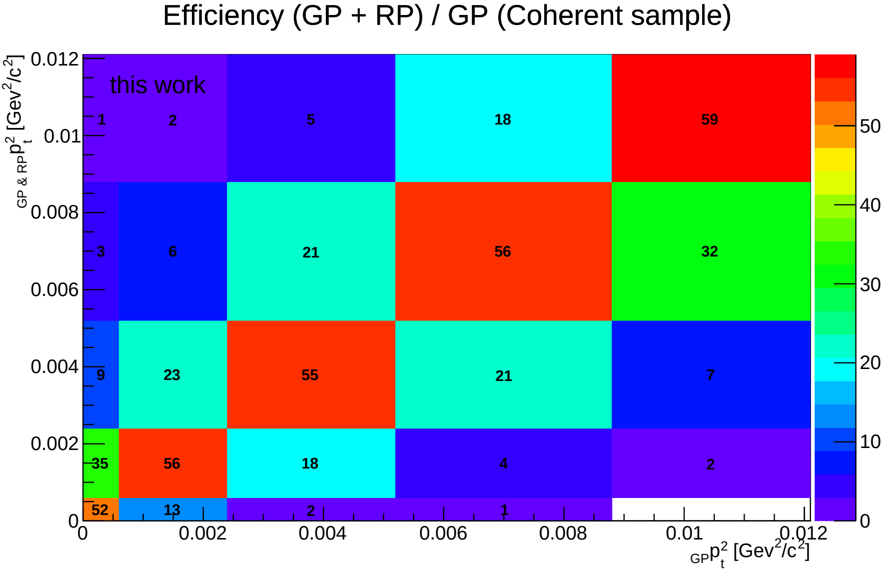
- You want to publish the original distribution, because
  - you want to compare with theory,
  - you want to compare between experiments.
- Acceptance × efficiency is usually preferred...





- ...but is not always the best solution.

Migrations for a **5** bins configuration - 2015 measurement



Unfolding

Migration treatment

Unfolding - theory

$$f_{meas}(y) = \int R(x, y) f_{true}(x) dx,$$

$f_{meas}(y)$ - measured distribution
$f_{true}(x)$ - true distribution
$R(x, y)$ - response function

- Response function contains:
    - detector acceptance,
    - efficiencies,
    - bin migrations.
- Unfolding = inversion of response matrix and its application on measured distribution (in discrete notation).

Unfolding methods

- Need to be solved:

$$\vec{x} = \mathbf{R}^{-1}\vec{y},$$

  - $\vec{x}$ is true vector, $\mathbf{R}$ is response matrix, $\vec{y}$ is measured vector
- Easy task - just invert the response matrix.
- Mathematically correct, but does not work.
- Large variance, which can not be reduced.
- Explained in back-up or read Cowen: Statistical Data and Analysis

Unfolding methods

- You need to somehow modify the response matrix.
- In general, you add bias and smaller variance.
- Theoretical prediction in form of Monte Carlo.
- Without regularisation:
  - Response matrix inversion. (large variance)
  - Method of correction factors. (large bias)
- With regularisation:
  - TUnfold (likelihood or χ^2 test).
  - Singular Value Decomposition (SVD)
- Iterative:
  - Bayes by D'Agostiny (iteration).

Available tools for unfolding

- RooUnfold.
  - Interface to wrap several methods.
  - Bayesian, SVD, bin-by-bin, TUnfold interface, inversion without regularization.
  - User friendly, less control.
  - Disadvantage for TUnfold - true vector used for regularization fixed to MC Truth distribution
- TUnfold.
  - ROOT classes TUnfold and TUnfoldDensity.
  - Example macros available.
  - Several methods to decide the strength of regularisation.
  - Several methods to perform regularisation itself.
- The best method/tool doesn't exists.
- A choice of the method depends on your distribution.

Unfolding

Iterative method

Bayes method (1/2)

x_i - true event with momentum in bin i
y_j - measured event with momentum in bin j
P(x_i) - probability of x_i
P(y_j,x_i) - probability of measurement of event y_j caused by event x_i

- From Bayes theorem:

$$P(x_i \mid y_j) = \frac{P(y_i \mid x_j)P(x_j)}{\sum_l^{nbins} P(y_i \mid x_l)P(x_l)}, \tag{1}$$

Bayes method (2/2)

x_i - true event with momentum in bin i

y_j - measured event with momentum in bin j

P(x_i) - probability of x_i

P(y_j,x_i) - probability of measurement of event y_j caused by event x_i

- Procedure:
  1. Choose first P(x_j) from MC,
  2. Do the Bayes estimation and get no. of true events,

$$n(x_j) = \sum_{l}^{nbins} n(y_l) P(x_j \mid y_l),$$

  3. Do χ^2 test,
  4. If condition not fulfilled, repeat with

$$P(x_j) = \frac{n(x_j)}{\sum_l^{nbins} n(y_l)}.$$

Implementation

Iterative method

Prerequisites

- RooUnfold package available in ROOT
- Data distribution
- Monte Carlo, where:
    - you have true distribution,
    - you have smeared distribution (true after application of detector effects)
    - you can connect both distributions for each event

Steps

- RooUnfold package available in ROOT
- Data distribution
- Monte Carlo, where:
  - you have true distribution,
  - you have smeared distribution (true after application of detector effects)
  - you can connect both distributions for each event

- machinery has 3 steps:
  - training of the method
  - testing of the method
  - application of the method

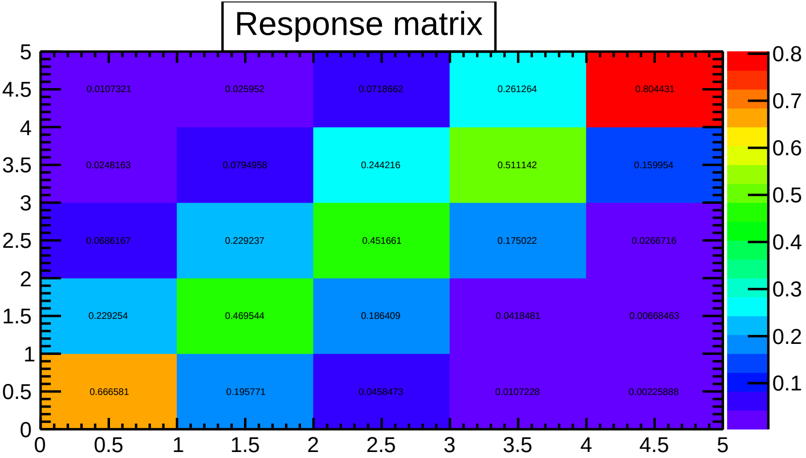
Training

```
\\ Object, which holds the response matrix
\\ uses input histograms to set response matrix dimensions
RooUnfoldResponse response (histoSmeared,histoTrue);
  \\ Loop over smeared MC events
  for (iev(0);iev<treeInput->GetEntries();iev++){
    treeInput->GetEntry(iev);\\ Get event from tree
    \\ Creates response matrix
    if (event_was_reconstructed) {
      \\ deals with migration
      response.Fill (varSmeared, varTrue);
    }
    else {
      \\ deals with reconstruction efficiency
      response.Miss (varTrue);
    }
  }
```

Look what you have

```
\\TMatrixD object, which can be drawn on canvas
  TMatrixD mtx = response.Mresponse();
  TPaveText *label = new TPaveText(0.35,0.9,0.65,1.0,"brNDC");
  label->AddText("Response␣matrix");
  TCanvas *canvas = new TCanvas("canvas","canvas",1600,900);
  canvas->cd();
    mtx.Draw("colzTEXT");
    label->Draw("same");
```

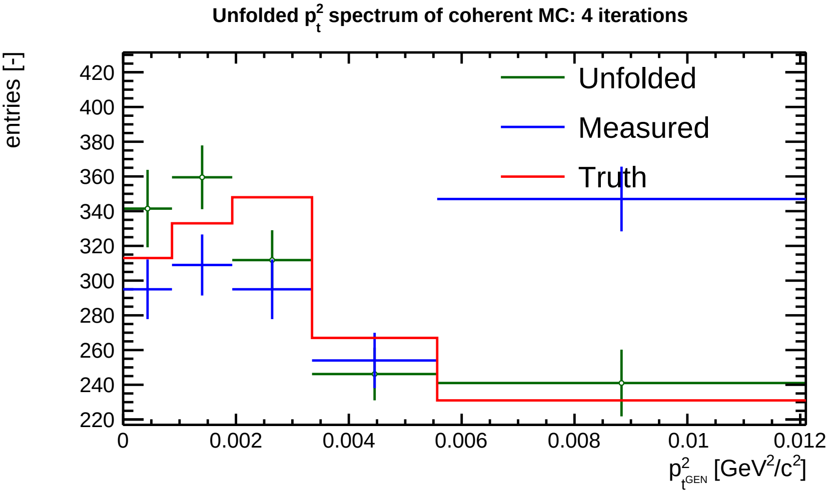
Look what you have



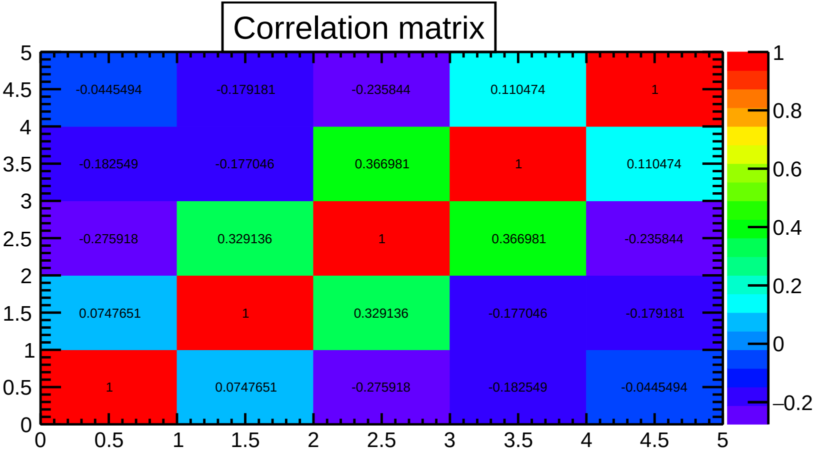
Testing

```
\\ Prepare MC histo for testing
TH1F* histoTesting = histoSmeared->Reduce("n_evts_as_data");
TH1F* histoTestTrue = histoTrue->Reduce("n_evts_as_data");
\\ Loop over number of iterations
for (iter(0);iter<ten_is_ussually_enough;iter++){
  \\ Object, which performs Bayes unfolding
  \\ Needs response matrix, smeared distribution
  \\ and no. of iterations for unfolding
  RooUnfoldBayes unfold (&response, histoTesting, iter);
  \\Get your results
 (TH1F*) unfold.Hreco(); //Unfolded spectrum
 (TH1F*) unfold.Hmeasured(); //Inserted measured spectrum
(TMatrixD) unfold.UnfoldingMatrix();  //Unfolding matrix
(TMatrixD) unfold.Ereco(); //Covariance mtx of unfold. spec.
}
```

Example - Bayes method



Example - Bayes method



Bayes unfolding

Finding iteration limit

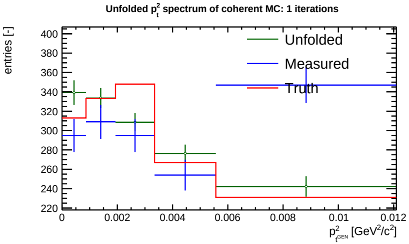
RooUnfold results

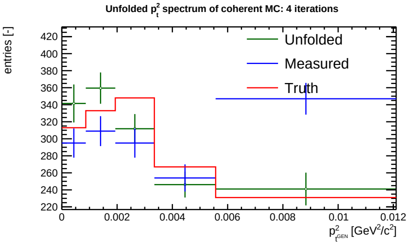
According to RooUnfold tutorial:

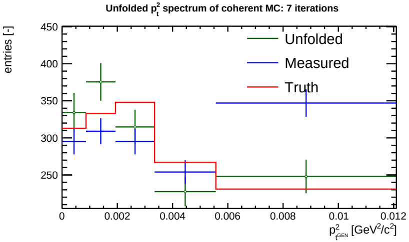
The optimal regularisation parameter can be selected by finding the largest value up to which the errors remain reasonable (ie. do not become much larger than previous values). This will give the smallest systematic errors (reconstructed distribution least biased by the training truth), without too-large statistical errors.

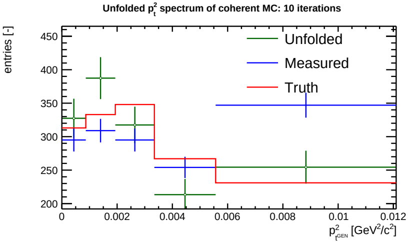
- Bayes - 4-5 iterations optimal.
- Investigate in more detail (Let's talk numbers).

Example - Bayes method

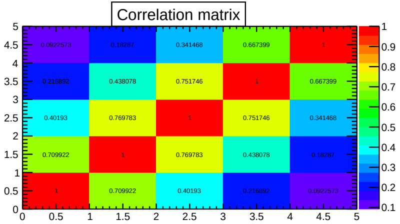

Unfolded p_t^2 spectrum of coherent MC: 1 iterations


Unfolded p_t^2 spectrum of coherent MC: 4 iterations


Unfolded p_t^2 spectrum of coherent MC: 7 iterations

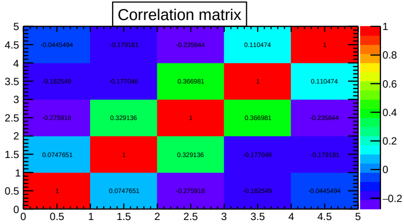

Unfolded p_t^2 spectrum of coherent MC: 10 iterations

Example - Bayes method



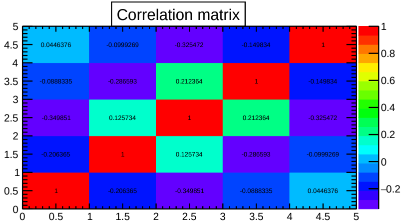
Correlation matrix

0.0922573	0.18287	0.341468	0.667399	1
0.216892	0.438078	0.751746	1	0.667399
0.40193	0.769783	1	0.751746	0.341468
0.709922	1	0.769783	0.438078	0.18287
1	0.709922	0.40193	0.216892	0.0922573



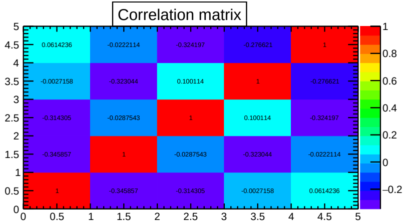
Correlation matrix

-0.0445494	-0.179181	-0.235844	0.110474	1
-0.182549	-0.177046	0.366981	1	0.110474
-0.275918	0.329136	1	0.366981	-0.235844
0.0747651	1	0.329136	-0.177046	-0.179181
1	0.0747651	-0.275918	-0.182549	-0.0445494



Correlation matrix

0.0446376	-0.0999269	-0.325472	-0.149834	1
-0.0888335	-0.286593	0.212364	1	-0.149834
-0.349851	0.125734	1	0.212364	-0.325472
-0.206365	1	0.125734	-0.286593	-0.0999269
1	-0.206365	-0.349851	-0.0888335	0.0446376



Correlation matrix

0.0614236	-0.0222114	-0.324197	-0.276621	1
-0.0027158	-0.323044	0.100114	1	-0.276621
-0.314305	-0.0287543	1	0.100114	-0.324197
-0.345857	1	-0.0287543	-0.323044	-0.0222114
1	-0.345857	-0.314305	-0.0027158	0.0614236

Bin no. comparison - constant bin difference distribution

Iteration:	1	2	3	4	5	6	7	8
Number of bins: 5								
Aver. Sys. Err. [%]	1.62	1.35	1.13	0.95	0.81	0.69	0.60	0.54
Aver. Stat. Err [%]	4.01	5.27	6.28	7.11	7.82	8.46	9.03	9.55
Number of bins: 6								
Aver. Sys. Err. [%]	1.38	1.13	0.92	0.79	0.72	0.66	0.60	0.56
Aver. Stat. Err [%]	4.09	5.44	6.53	7.45	8.25	8.97	9.62	10.23
Number of bins: 7								
Aver. Sys. Err. [%]	1.66	1.47	1.30	1.15	1.02	0.91	0.81	0.72
Aver. Stat. Err [%]	4.14	5.55	6.70	7.68	8.54	9.32	10.03	10.70
Number of bins: 8								
Aver. Sys. Err. [%]	1.47	1.30	1.18	1.08	0.99	0.91	0.84	0.77
Aver. Stat. Err [%]	4.19	5.65	6.86	7.88	8.79	9.62	10.38	11.09
Number of bins: 9								
Aver. Sys. Err. [%]	1.62	1.43	1.27	1.12	1.00	0.90	0.81	0.75
Aver. Stat. Err [%]	4.23	5.73	6.97	8.04	8.98	9.85	10.65	11.39

Fit results - constant bin difference distribution

Iteration:	1	2	3	4	5	6	7	8
No. bins: 5								
Norm. fit err. [%]	4.07	5.60	6.56	7.12	7.44	7.63	7.74	7.81
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.70 ± 0.05	6.69 ± 0.08	6.69 ± 0.10	6.68 ± 0.11	6.67 ± 0.12	6.67 ± 0.12	6.67 ± 0.12	6.67 ± 0.12
χ^2/ndf	8.68	2.62	1.57	1.23	1.09	1.02	0.98	0.96
No. bins: 6								
Norm. fit err. [%]	4.62	6.22	7.14	7.63	7.89	8.03	8.11	8.15
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.50 ± 0.06	6.58 ± 0.09	6.63 ± 0.11	6.66 ± 0.12	6.67 ± 0.12	6.68 ± 0.12	6.69 ± 0.13	6.69 ± 0.13
χ^2/ndf	19.93	5.85	3.00	1.97	1.49	1.22	1.06	0.96
No. bins: 7								
Norm. fit err. [%]	7.48	7.82	8.22	8.40	8.46	8.48	8.49	8.48
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.12 ± 0.10	6.39 ± 0.11	6.53 ± 0.12	6.60 ± 0.13	6.64 ± 0.13	6.67 ± 0.13	6.68 ± 0.13	6.69 ± 0.13
χ^2/ndf	110.04	36.45	16.35	8.82	5.34	3.51	2.46	1.82
No. bins: 8								
Norm. fit err. [%]	5.03	6.43	7.22	7.62	7.82	7.92	7.98	8.01
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.54 ± 0.07	6.65 ± 0.10	6.70 ± 0.11	6.73 ± 0.12	6.75 ± 0.12	6.76 ± 0.12	6.77 ± 0.13	6.77 ± 0.13
χ^2/ndf	29.50	8.51	4.23	2.73	2.06	1.71	1.51	1.38
No. bins: 9								
Norm. fit err. [%]	8.19	7.83	8.04	8.14	8.18	8.18	8.17	8.16
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.19 ± 0.11	6.45 ± 0.11	6.58 ± 0.12	6.66 ± 0.12	6.70 ± 0.13	6.73 ± 0.13	6.75 ± 0.13	6.76 ± 0.13
χ^2/ndf	100.54	39.67	20.03	11.99	8.04	5.82	4.47	3.59

Long story short a.k.a. summary

- Once you find your best number of iterations, simply use RooUnfoldBayes with your data.
- You can easily try SVD within RooUnfold.
  - You can compare both methods.
  - Use RooUnfoldSVD instead RooUnfoldBayes.

- Some general observations:
  - Results do not really depend on binning scheme.
    - D'Agostini claims that a uniform distribution is also a good first guess.
  - Bayes and SVD are comparable (in my case)
- It is save to use this method as a black box.
  - You can try to understand it deeply, but at the end, you will use it as a black box anyway.

References

- Unfolding in general
  - Cowen: Statistical Data and Analysis
  - Conf. Proc. C **0203181** (2002) 248.
  - arXiv:1505.04768v3
  - Redmer Alexander Bertens - ALICE analysis tutorial.
- Bayes
  - doi:10.1016/0168-9002(95)00274-X
  - arXiv:1201.4612
  - arXiv:1010.0632
- SVD
  - arXiv:hep-ph/9509307
- Methods implementations
  - RooUnfold
  - TUnfold

BACK UP

Conclusion III

- Here, average correlation drops to 0.35.
- Goes against the first conclusion that less iterations is better.
- Question. What would be the best mix of correlation+stat.error+sys.error?
- Because systematical errors from unfolding are correlated, we also investigated the response of fit parameters errors on different number of bins and iterations - all fits includes covariance matrix of unfolding results.
- Next, we show results of this fits and fitting function taken from STARlight.

STARlight function

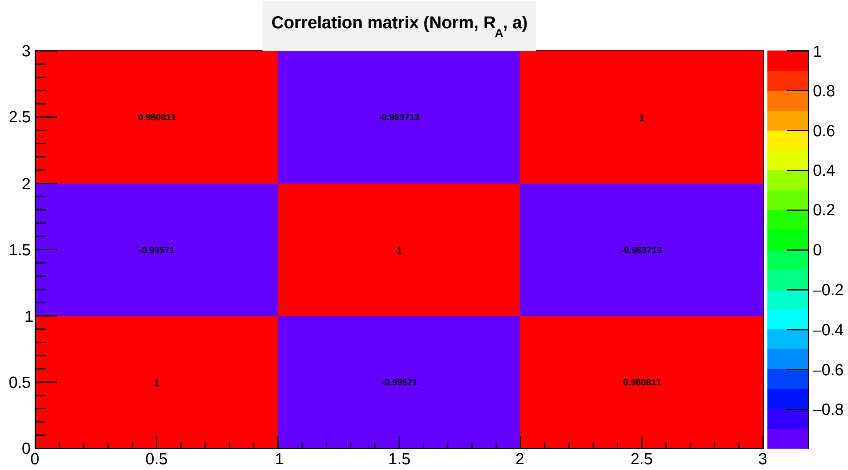
- STARlight production is based on Vector Meson Dominance Model
- as a form factor is used a convolution of Yukawa potential and hard sphere

$$\frac{d\sigma(\gamma A \rightarrow VA)}{dt} = NORM|F(t)|^2$$

$$F(q=\sqrt{|t|}) = \frac{4\pi\rho_0}{Aq^3}[\sin(qR_A) - qR_A\cos(qR_A)][\frac{1}{1+a^2q^2}]$$

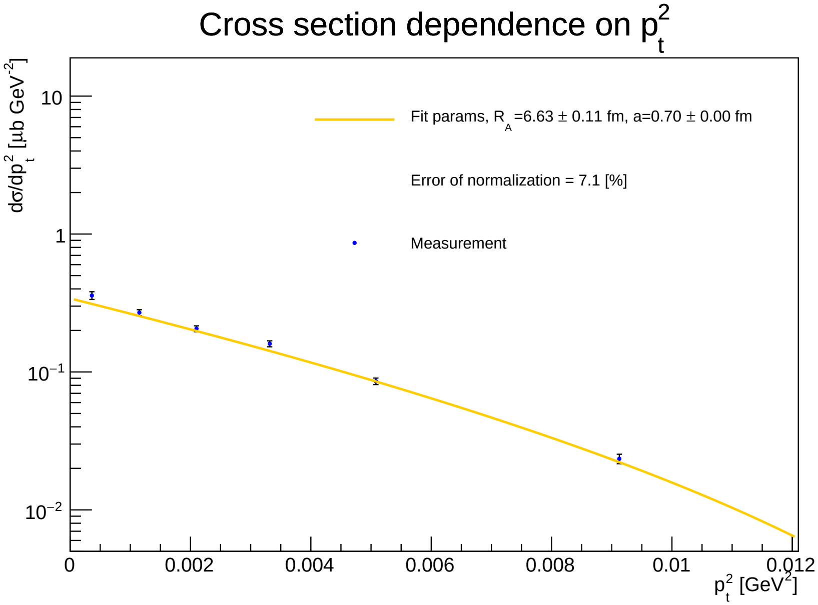
- R_A is the transversal nucleus radius, a is the range of Yukawa potential, A is the atomic number and ρ_0 is the nuclear density of a hard sphere
- input numbers are $R_A = 6.62$ fm and $a = 0.7$

Example: Covariance mtx of fit with 6 Bins, 3 iterations



- a is strongly correlated with $R_A \rightarrow a$ is fixed.

Example: Fit with 6 Bins, 3 iterations



Fit results - constant bin difference distribution

Iteration:	1	2	3	4	5	6	7	8
No. bins: 5								
Norm. fit err. [%]	4.07	5.60	6.56	7.12	7.44	7.63	7.74	7.81
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.70 ± 0.05	6.69 ± 0.08	6.69 ± 0.10	6.68 ± 0.11	6.67 ± 0.12	6.67 ± 0.12	6.67 ± 0.12	6.67 ± 0.12
χ^2/ndf	8.68	2.62	1.57	1.23	1.09	1.02	0.98	0.96
No. bins: 6								
Norm. fit err. [%]	4.62	6.22	7.14	7.63	7.89	8.03	8.11	8.15
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.50 ± 0.06	6.58 ± 0.09	6.63 ± 0.11	6.66 ± 0.12	6.67 ± 0.12	6.68 ± 0.12	6.69 ± 0.13	6.69 ± 0.13
χ^2/ndf	19.93	5.85	3.00	1.97	1.49	1.22	1.06	0.96
No. bins: 7								
Norm. fit err. [%]	7.48	7.82	8.22	8.40	8.46	8.48	8.49	8.48
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.12 ± 0.10	6.39 ± 0.11	6.53 ± 0.12	6.60 ± 0.13	6.64 ± 0.13	6.67 ± 0.13	6.68 ± 0.13	6.69 ± 0.13
χ^2/ndf	110.04	36.45	16.35	8.82	5.34	3.51	2.46	1.82
No. bins: 8								
Norm. fit err. [%]	5.03	6.43	7.22	7.62	7.82	7.92	7.98	8.01
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.54 ± 0.07	6.65 ± 0.10	6.70 ± 0.11	6.73 ± 0.12	6.75 ± 0.12	6.76 ± 0.12	6.77 ± 0.13	6.77 ± 0.13
χ^2/ndf	29.50	8.51	4.23	2.73	2.06	1.71	1.51	1.38
No. bins: 9								
Norm. fit err. [%]	8.19	7.83	8.04	8.14	8.18	8.18	8.17	8.16
a [fm]	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00	0.70 ± 0.00
R_A [fm]	6.19 ± 0.11	6.45 ± 0.11	6.58 ± 0.12	6.66 ± 0.12	6.70 ± 0.13	6.73 ± 0.13	6.75 ± 0.13	6.76 ± 0.13
χ^2/ndf	100.54	39.67	20.03	11.99	8.04	5.82	4.47	3.59

Conclusion IV

- Fit still very bad for low number of iterations
- When fit OK, relative error increases with no. of iterations - expected!
- Fit parameters ussually correct within error.
- a (or R_A) has to be fixed, otherwise the minimum finding algorithm hits a plato.
- Now, we have enough information to decide, what number of iterations we will use. What will be the strategy?
- Next step is to apply on real data. Let's hope they will be available soon.

Unfolding

Inversion failure

A trivial problem?

The solution seems simple: multiply the measurement by the inverted response matrix

$$\mathbf{x}_i = \mathbf{R}_{ij}\mathbf{y}_j \qquad \longrightarrow \qquad \mathbf{x}_i \approx \mathbf{n}_i \pm \sigma_i \qquad \longrightarrow \qquad \mathbf{R}_{ij}^{-1}\mathbf{n}_i = \mathbf{R}_{ij}^{-1}\mathbf{R}_{ij}\mathbf{y}_j = \mathbf{y}_j$$

let's give it a go ... (provided the matrix *has* an inverse)

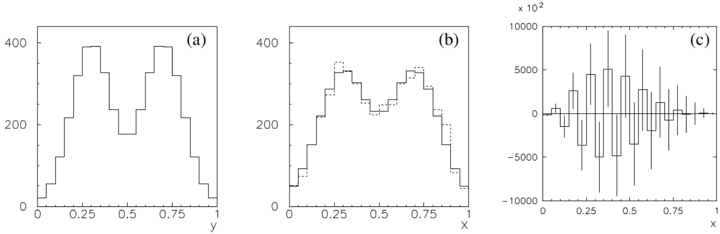


fig. (a) true distribution **y**

fig. (b) measurement **n** (full) and true distribution (**x**, dashed)

fig. (c) distribution **y** from inversion ... o dear ! ...

To understand what goes wrong, we have to look at the response matrix

... where inversion *doesn't* work ... [2/2]

In practice, the bin size of a measurement is small compared to experimental resolution

- Result: off-diagonal elements in the response matrix
- At equal efficiency ϵ, the problem becomes e.g.

$$\underbrace{\begin{pmatrix}0.8\\1\\0.9\end{pmatrix}}_{\epsilon} \qquad \underbrace{\begin{pmatrix}1\\2\\3\end{pmatrix}}_{\mathbf{n}} = \underbrace{\begin{pmatrix}0.5 & 0.3 & 0\\0.3 & 0.4 & 0.3\\0 & 0.4 & 0.5\end{pmatrix}}_{\mathbf{R}} \cdot \underbrace{\begin{pmatrix}?\\?\\?\end{pmatrix}}_{\mathbf{y}}$$

... the inverted response looks quite different now ...

$$\underbrace{\begin{pmatrix}-16 & 30 & -18\\30 & -50 & 30\\-24 & 40 & -22\end{pmatrix}}_{\mathbf{R}^{-1}} \cdot \underbrace{\begin{pmatrix}1\\2\\3\end{pmatrix}}_{\mathbf{n}} = \underbrace{\begin{pmatrix}-10 \pm 293\\20 \pm 382\\-10 \pm 338\end{pmatrix}}_{\mathbf{y}}$$

and **y** is probably not the solution we're looking for! Inversion gives

- a 'mathematically' correct (unbiased) estimator of **y** ...
- ... at the cost of wildly oscillating y with catastrophically large variances

Unfolding

Regularization

Regularizations

μ - true histogram with M bins, $S(\mu)$ - regularization function,
τ - strength of regularization, $\log L(\mu) = -\chi^2(\mu)/2$

$$\Phi(\mu) = \log L(\mu) + \tau S(\mu),$$

- Tikhonov regularization

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

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

- Entropy-based regularization

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

Strength of regularizations τ

- The key task.
- Represents the trade-off between bias and variance.
- TUnfold offers 2 principles how to find τ:
  - L-curve scan ($\log \chi^2 vs. \log \tau$)
  - Global correlation coefficient scan ($\overline{\rho} vs. \log \tau$)

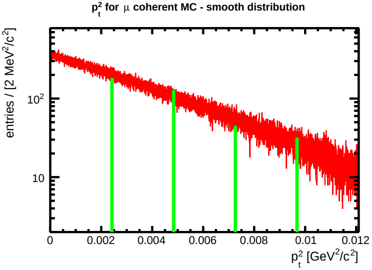
tau limits:

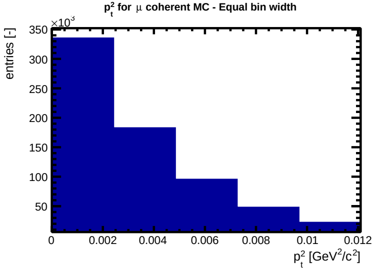
It is desirable to have a small τ parameter, otherwise the result becomes largely biased. On the other hand it should be strong enough to damp large oscillations.

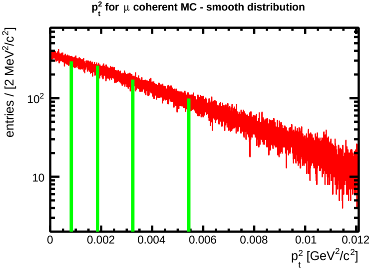
Tests

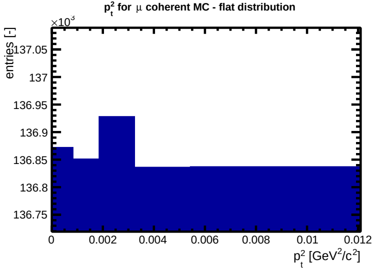
Tested binnings 1/2

Equal bin width and flat distributions



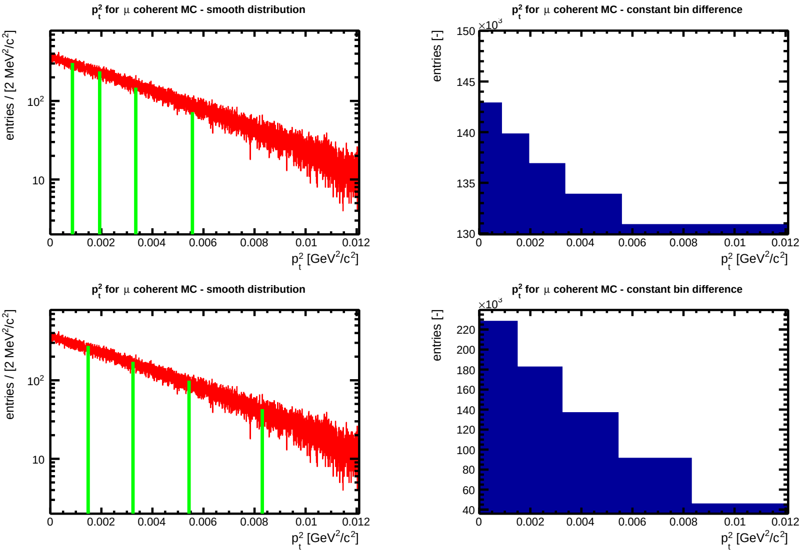






Tested binnings 2/2

Constant bin difference without and with respect to overflow bin



TUnfold results

- Tikhonov regularization.
  - Can be modified.
  - (derivative, curvature) = (k=1,k=2)
  - Change the number of adjacent bins used for 1 calculation.
- Global correlation coefficient scan used to find regularization parameter.
  - Can be modified.
  - Minimalisation of global average/maximum.
  - Average used.

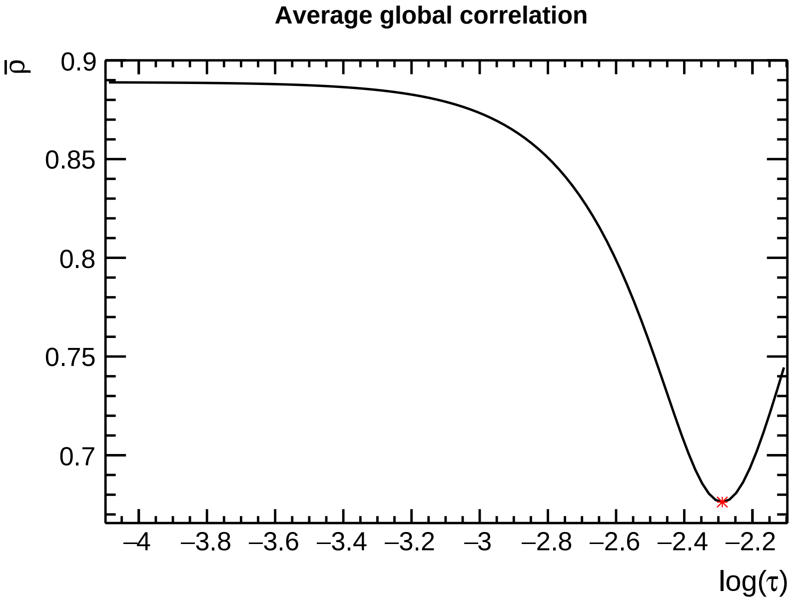
Average global correlation coefficient method

ρ - global correlation coefficient, V - covariance matrix

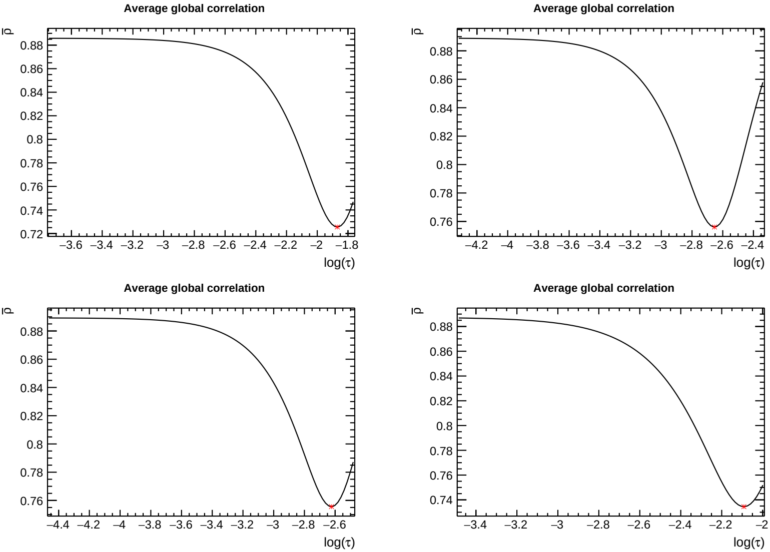
$$\rho_i = \sqrt{1 - \frac{1}{V_{ii} V_{ii}^{-1}}},$$

- $V_{ii} V_{ii}^{-1} = \sigma^2$ - variance of the bin.
- τ (strength of regularization) finding algorithm.
    - Take some τ and unfold.
    - Take covariance matrix of the result.
    - Average ρ over all bins.
    - Plot the point (τ vs. ρ)
    - Repeat for another τ.
    - Once enough points, fit and find minimum.

TUnfold - regularization parameter finding example



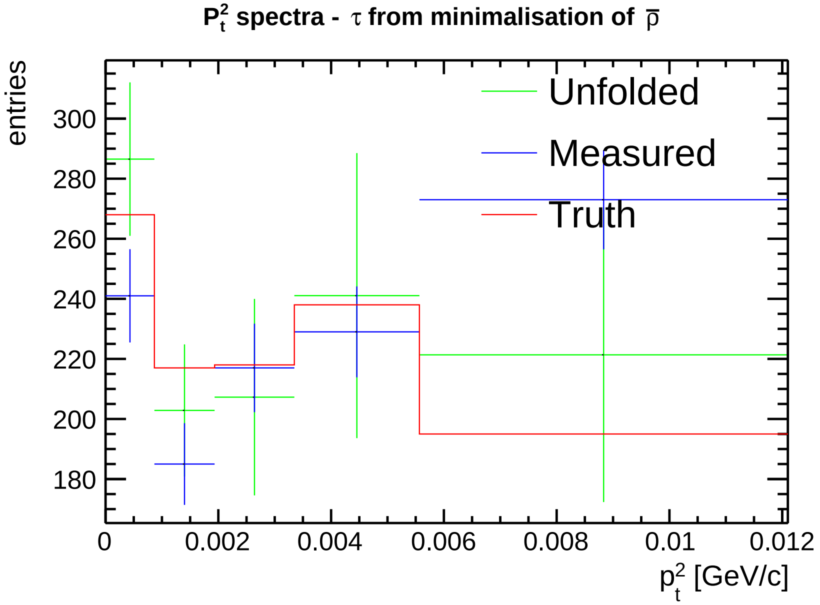
TUnfold Curvature - comparison of τ



τ comparison

- Derivative regularization fails for non-flat distributions.
- The least bias is for flat distribution.
- Curvature regularization also good for constant bin difference.
- Equal bin width largely biased.

TUnfold Curvature - constant bin difference

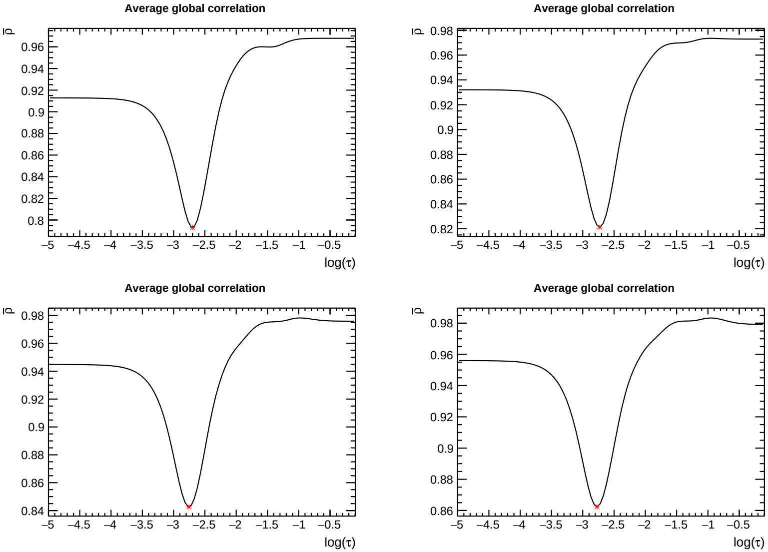


Different bin numbers

τ comparison of different bin number

- All plots for constant bin difference.
- Testing τ for different number of bins.
- Smaller variance yields smaller ρ.
- In this scope make sense that more bins (larger variance) yields larger average ρ (see next slides).
- Testing τ for different number of bins.
- Structure of the next plots:
  - Next slide:
    - Left top - 6 bins.
    - Right top - 7 bins.
    - Left bottom - 8 bins.
    - Right bottom - 9 bins.

TUnfold Curvature - comparison of τ



τ comparison - summary

- Higher τ = higher bias.
- With more bins used, result is less biased.
- However, the difference is within 0.1 %

- TUnfold summary:
    - Depends a lot on distribution.
    - Little sensitivity on number of bins.
    - In our case, variance is not suppressed enough.

Template title

- Template item.