# Unfolding: the Collider experience

PHYSTAT-nu 2019

Stefan Schmitt, DESY

This talk is reflecting my personal opinions only.

Examples taken from experiments at LHC and elsewhere are based on published results. The selection is following my personal biases and certainly is far from being complete and accurate.

The “big” LHC experiments all have their own experts on statistics, hopefully some of these are here to comment and to correct mistakes.

# Outline

- Introduction, definition of “unfolding” for this talk
- Unfolding methods used frequently in Collider experiments
- Comparison

# Cross section measurements

- Collider experiment, measurement of a cross section (in some fiducial volume)

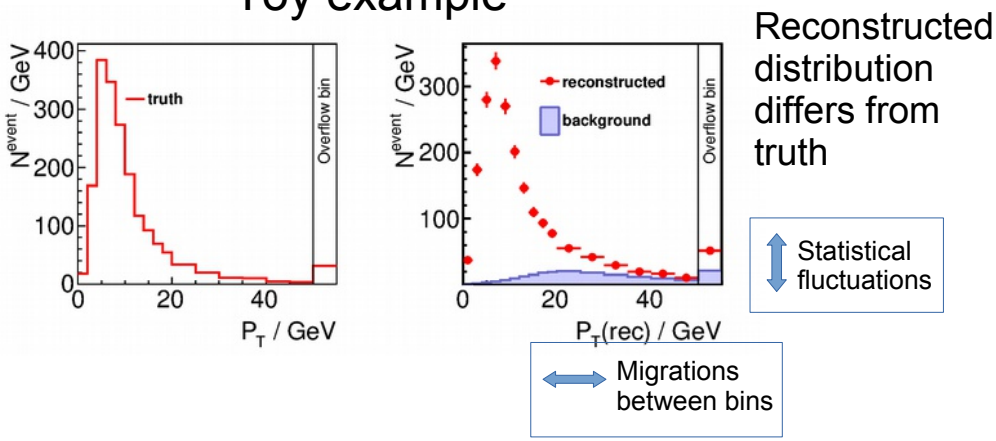
$\sigma = n / L$ [event count / integrated luminosity]

- Repeat this in several regions of phase space (bin number j=1..N)

$\sigma_j = n_j / L$

- Difficulties:
  - Statistical fluctuations: number of events $n_j$ fluctuates around Poisson parameter $\mu_i$
  - Migrations: an event belonging to truth bin j may be reconstructed in reco bin i≠j
  - Inefficiency: an event belonging to truth bin j is not reconstructed as signal
  - Background: non-signal processes may also produce a signal-like event
- In this talk, the main discussion is on migrations and statistical fluctuations.

Toy example



# Unfolding of cross sections

- Statistical model
  - Poisson expectation values are given by a folding equation, which describes migrations, inefficiencies and background

$$\mu_i = \sum_j A_{ij} x_j^{\text{truth}} + b_i$$

$A_{ij}$ : probability to find truth bin $j$ as reco bin $i$

$\epsilon_j = \sum_i A_{ij}$ : efficiency to reconstruct truth bin $j$

$b_i$ : expected number of background events

  - Observed number of events is drawn from a Poisson distribution

$$P(y_j^{\text{reco}}; \mu_j) = \exp[-\mu_j] \frac{\mu_j^{y_j^{\text{reco}}}}{(y_j^{\text{reco}})!}$$

- Unfolding: estimator of the "truth" parameters

$$\hat{x}|_{y^{reco}} : \text{estimator of the } x_j^{\text{truth}}$$

- Example: maximum-likelihood estimator

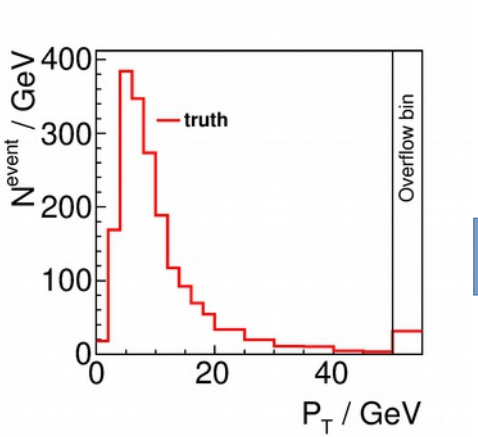
$$\frac{\partial - \ln L(\hat{x}; y^{\text{reco}})}{\partial \hat{x}_j} = 0$$

Special case: dim(x)=dim(y)

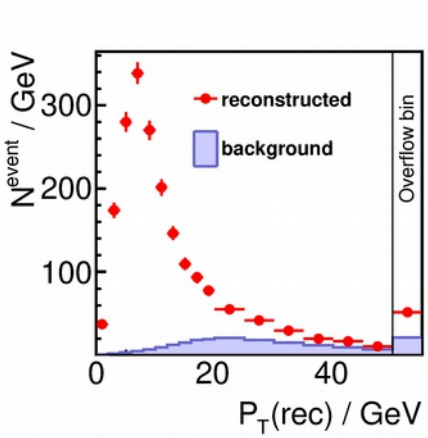
→ A is a square matrix, invert it

$$\hat{x}_j = \sum_i (A^{-1})_{ji} (y_i^{\text{reco}} - b_i)$$

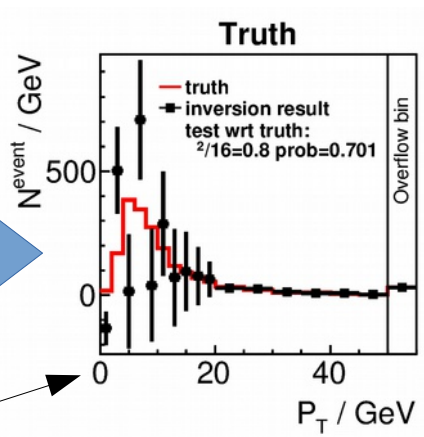
# Matrix inversion example

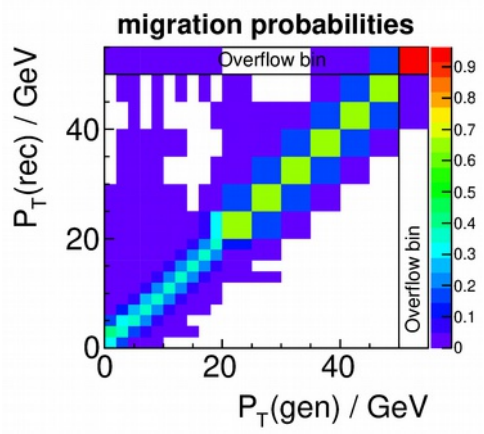



folding




Matrix inversion

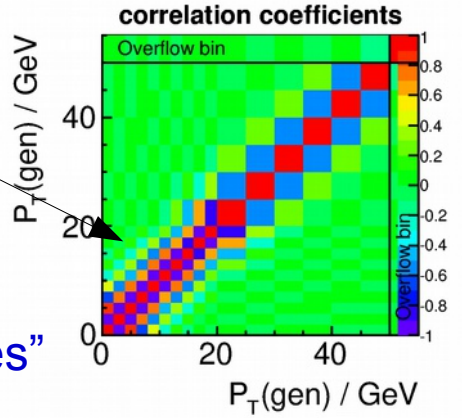



Result shows “oscillating” structures.
Large (anti-)correlations between bins.

Qualitative explaination:

Finite detector resolution σ washes out differences between bins. Unfolding “replaces” the missing information by statistical “noise”
→ statistical fluctuations are amplified



# Constructing the matrix A

- In particle physics, the matrix A is reconstructed in most cases using Monte Carlo event generators
- For each event, determine the reco bin i and the truth bin j
- Count number of events
  - In truth bin j
  - In reco bin I
  - In truth bin j and reco bin i

- The event counts are normalized to match the data luminosity
- The unfolding ingredients from simulation are

Simulated truth: $x_j^{\text{truth}} = N_j^{\text{sim,truth}}$

Simulated observation: $\mu_i = N_i^{\text{sim,reco}}$

Probability matrix: $A_{ij} = \frac{N_{ij}}{N_j^{\text{sim,truth}}}$

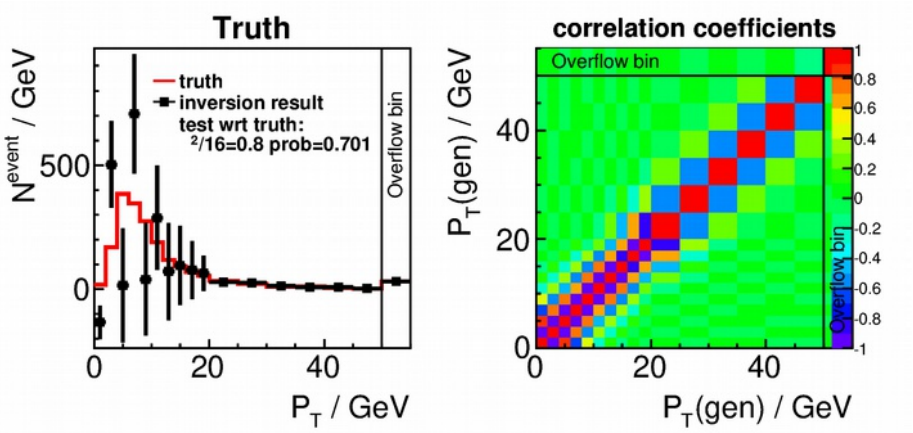
Background from "fakes": $b_i^{\text{fakes}} = N_i^{\text{sim,reco}} - \sum_j N_{ij}$

Often, a “reweighting” is applied to improve the simulation: a weighting function w=f(x) is chosen such that the predicted ($N_i^{\text{sim,reco}}$) agree better with the data ($y_i$) → improved unfolding input.

Possible problem: the data are used twice (to determine f(x) and for the unfolding. This may give incorrect uncertainties if the unfolding method chosen produces a notable bias to $N_j^{\text{sim,truth}}$

# Regularisation

- Example of matrix inversion: unfolding results shows oscillation patterns and large (anti-)correlations




- Regularisation:

  Put in prior knowledge, for example:

  - Cross section is strictly positive
  - Result is expected to be “smooth”
  - Prefer estimators with small correlation coefficients

## Examples

- Bin-by-bin
  - Simple estimator with zero correlation coefficients but potentially large bias
- Tikhonov regularisation
  - Add prior knowledge as extra “measurements”, with tunable weight parameter τ
  - Zero τ → unbiased result
  - Large τ → bias
- Iterative methods with given start value
  - Ensures that result is positive
  - Small number of iterations: strong bias to start value
  - Many iterations → reduced bias

# Unfolding methods overview

- Methods used frequently in Collider experiments
  - Square matrices: $N_{truth}=dim(x)=dim(y)=N_{reco}$
    - Bin-by-bin
  - More general case: $N_{truth}=dim(x) \leq dim(y)=N_{reco}$
    - Non-regularized maximum-likelihood fit (RooFit)
    - Least-square with Tikhonov regularisation (TUnfold, TSVDunfold)
    - EM iterations with early stopping, "D'Agostini"

Note1: certain implementations of SVD and EM iterations only work with square matrices
Note2: RooUnfold provides a common interface to many unfolding algorithms but also has comes with certain limitations

# Bin-by-bin corrections

- Main assumption: matrix of probabilities A is approximately diagonal
- Observed data are "corrected" using the simulation

$$\hat{x}_i^{\mathrm{BBB}} = f_i\, y_i^{\mathrm{data}} \text{ , where: } f_i = \frac{N_i^{\mathrm{sim,gen}}}{N_i^{\mathrm{sim,reco}}}$$

- This estimator by definition is (statistically) uncorrelated between bins
- However, it introduces a bias to the simulation, depending on the bin "purity" (fraction of reconstructed events which originate from that truth bin)

- Approximate formula to estimate the bias of this method:

$$\text{Purity: } P_i = \frac{A_{ii} N_i^{\mathrm{sim,gen}}}{N_i^{\mathrm{sim,reco}}} = \frac{n_i^{\mathrm{sim,gen}\wedge\mathrm{reco}}}{n_j^{\mathrm{sim,reco}}}$$

$$\text{BBB expectation: } \langle \hat{x}_i^{\mathrm{BBB}} \rangle \sim (1-P_i)\, x_i^{\mathrm{sim,gen}} + P_i\, x_i^{\mathrm{truth}}$$

- **WARNING**: Can be quite problematic for testing models

Imagine: simulation was NLO at the time of publication, purity was 50%
→ unfolded cross section is half-way between truth and NLO
→ goodness of NNLO theory can not be tested with these data!

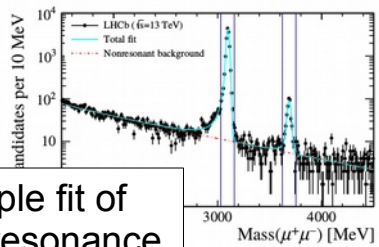
Important message: when using BBB despite of purities not close to unity, always publish both the purities and the (truth) prediction used for unfolding

# Example of a bin-by-bin analysis

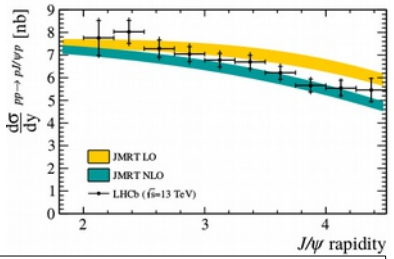
- Bin-by-bin analyses have been used frequently by past experiments
- Nowadays, bin-by-bin corrections often are still used when measuring kinematic distributions of resonance decays
- Reason: tracker gives precision measurement of these kinematic variables

  → matrix A is almost diagonal, corrections for migrations are very small

- Example: central exclusive production of J/psi @13 TeV, LHCb [arXiv:1806.04079]
  - Signals are extracted in fits to the $\pi^+\pi^-$ mass distribution
  - Migrations between analysis bins are not corrected for




Example fit of J/psi resonance




Fit is repeated in rapidity bins, which are reconstructed precisely

A bin migration uncertainty has been estimated using simulation to relate the reconstructed and true rapidity bin. The difference is smaller than 0.06% in all bins and so is considered negligible.

# Template fits

- Template: shape of a reconstructed distribution predicted by simulation or a data control sample
- Template fit: determine normalisation factors of all templates

$$y_i^{\text{data}} \sim \sum A_{ij} \hat{x}_j$$

$A_{ij}$ : template $j$ distribution

$\hat{x}_j$ : template $j$ normalisation, fit parameter

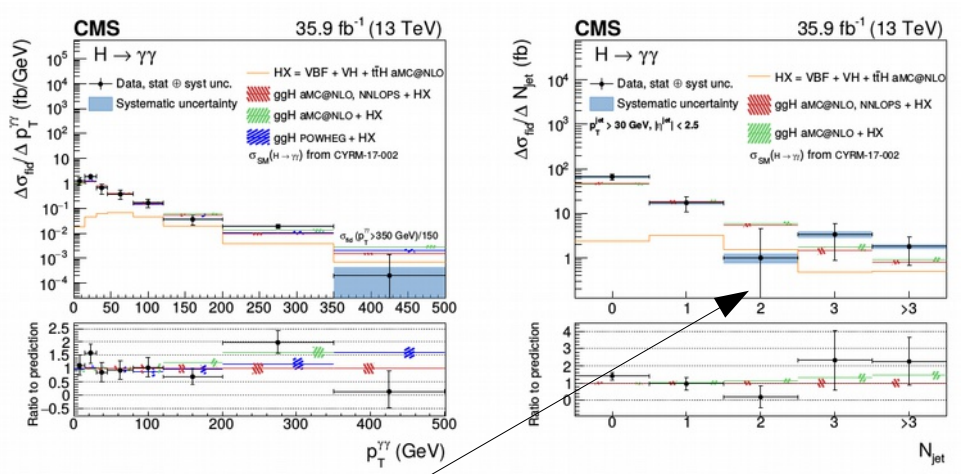
- "Template fit" and "unfolding" are both dealing with the same problem: to decompose an observed distribution into its sources

- The distinction between "tempate fit" and "unfolding" is made in particle physics approximately along these lines:
  - "Template fits" typically are done using Poission likelihoods and without regularisation. They often include "control" distributions to discriminate signal and background
  - Unfolding usually involves regularisation and is often based on a least-square fits. In many case, only in the distribution of interest is probed, whereas background is subtracted and not included in the fit.

# Template fit example

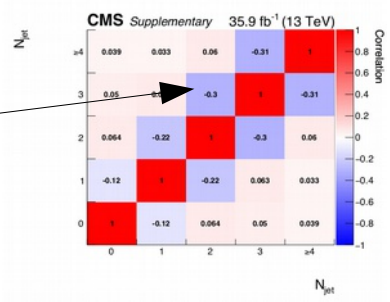
- Template fits often are done in background-dominated analyses
- In many cases RooFit is used as tool – designed to work with many distributions and many bins (signal and control regions)
- RooFit also can handle non-linear nuisance parameters besides the signal and background normalisation factors

- Example: H→γγ differential cross sections [CMS, arXiv:1807.03825]




"Oscillating" behavior is typical for non-regularized unfolding, together with negative correlation coefficients

[Plot of correlation coefficients taken from ICHEP2018 talk by V. Tavolaro, backup slides]




## Likelihood function from paper

all the analysis categories. The complete likelihood is given in Eq. (1):

$$\mathcal{L}(\text{data}|\Delta\vec{\sigma}^{\text{fid}}, \vec{n}_{\text{bkg}}, \vec{\theta}_{\text{S}}, \vec{\theta}_{\text{B}}) =$$

$$\prod_{i=1}^{n_{\text{cat}}} \prod_{j=1}^{n_{\text{b}}} \prod_{l=1}^{n_{m_{\gamma\gamma}}} \left( \frac{\sum_{k=1}^{n_b} \Delta\sigma_k^{\text{fid}} K_k^{ij}(\vec{\theta}_{\text{S}}) S_k^{ij}(m_{\gamma\gamma}^l|\vec{\theta}_{\text{S}}) L + n_{\text{OOA}}^{ij} S_{\text{OOA}}^{ij}(m_{\gamma\gamma}^l|\vec{\theta}_{\text{S}}) + n_{\text{bkg}}^{ij} B^{ij}(m_{\gamma\gamma}^l|\vec{\theta}_{\text{B}})}{n_{\text{sig}}^{ij} + n_{\text{bkg}}^{ij}} \right)^{n_{\text{ev}}^{lij}} \quad (1)$$

$$\text{Pois}(n_{\text{ev}}^{ij}|n_{\text{sig}}^{ij} + n_{\text{bkg}}^{ij}) \text{Pdf}(\vec{\theta}_{\text{S}}) \text{Pdf}(\vec{\theta}_{\text{B}}),$$

# Least-square & Tikhonov regularisation

- Least-square fit: normal-distributed data, approximate for event counts
- Minimize $\chi^2$=-log(L) with regularisation term added

$$\chi^2=(Ax-y)^T V_y^{-1}(Ax-y)+\tau^2(L(x-x_B))^T(L(x-x_B))$$

$V_y$ : covariance of y, $A$ : matrix of probabilities
$y$ : observations, $x$ : unfolding parameters

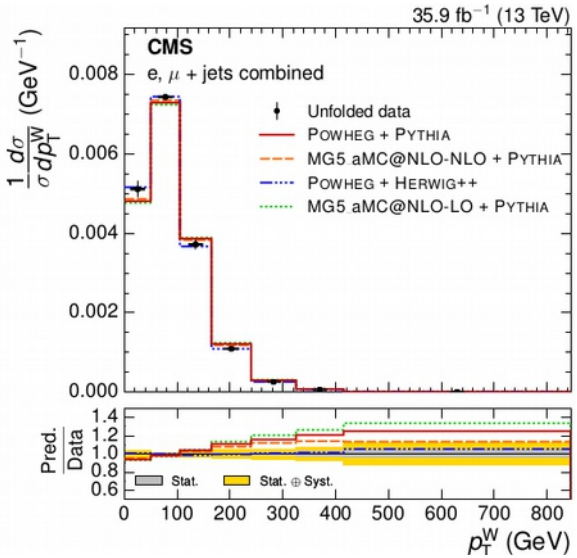
$x_B$ : bias (simulation truth)
$L$ : regularisation pattern (unity matrix or curvature matrix)
$\tau$ : regularisation strength

[to include background, replace $y_i$ by $(y_i-b_i)$ ]

- For zero τ: standard least-square fit (in analogy to Poisson log-likelihood on previous slides)
- For non-zero τ: penalty for large differences between x and $x_B$
  →damps the oscillations
- Equations can be solved easily, question is on the choice of τ
  - TUnfold: test unfolding for many τ, select distinct point for final result (L-curve kink, minimum correlation, ...)
  - TSVDUnfold and related methods: τ is the result of an analysis of eigenvalues (of the matrix $A^TV^{-1}A$ or the matrix $V^{-½}A$ or similar)

# Two Examples using TUnfold

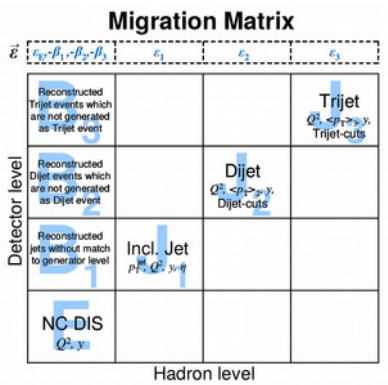
- Differential ttbar (single ℓ) cross sections [CMS, arXiv:1803.03991]

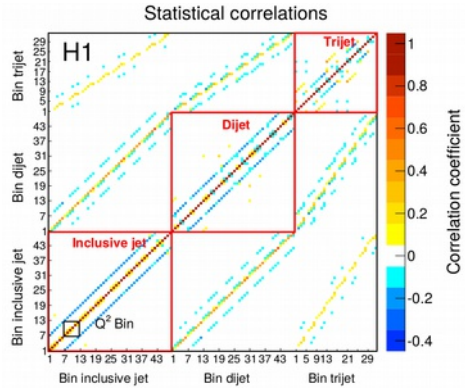



Parameter τ is chosen to minimize the statistical correlations

trix, based on a least-squares fit with Tikhonov regularization, implemented in the TUNFOLD software framework [53]. Regularization dampens nonphysical fluctuations in the unfolded $t\bar{t}$ yields, and the regularization parameter is chosen by minimizing the average global statistical correlation between the bins of each variable. The typical regularization parameters are found to be of order $10^{-4} - 10^{-3}$, and significantly lower for the $|\eta^{\ell}|$ variable.

- Jet cross sections in ep collisions [H1, arXiv:1611.03421 & 1406.4709]




In this rather complex example, the data covariance includes non-diagonal elements (jets emerging from the same event are correlated). The matrix A has dimension of order 12000x3300. Only 10% (~320) of the unfolded bins are used for cross-sections, the other bins are nuisance parameters for control regions.

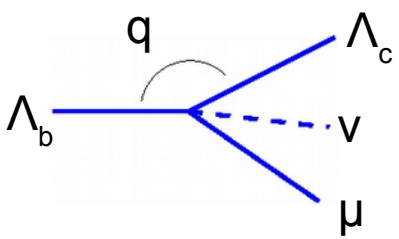
# Example using SVD

- Measurement of the $\Lambda_b \to \Lambda_c \mu \nu$ differential decay rate [LHCb arXiv:1709.01920]
- The decay rate is unfolded as a function of the invariant w

$w = v(\Lambda_b)\, v(\Lambda_c)$, with velocity four-vector $v = p/m$

- w is related to the momentum transfer $q^2$

$$q^2 = m^2_{\Lambda_b} + m^2_{\Lambda_c} - 2\, w\, m_{\Lambda_b} m_{\Lambda_c}$$



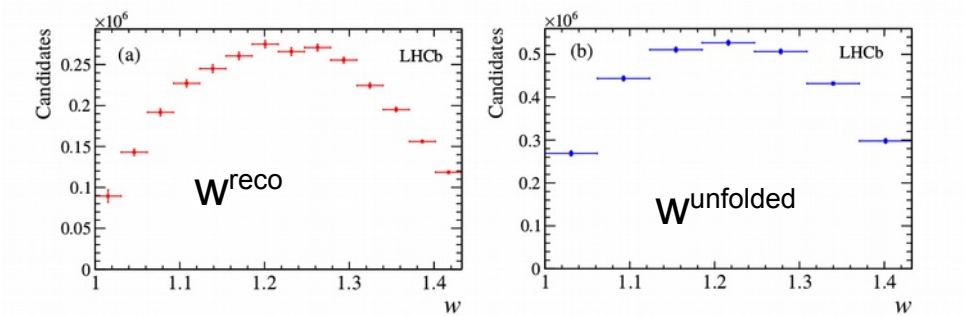




Figure 3: The spectra (a) $dN_{\rm meas}/dw$ before unfolding and (b) $dN_{\rm u}/dw$ after unfolding, for the decay $\Lambda_b^0 \to \Lambda_c^+ \mu^- \overline{\nu}_\mu$. The latter spectrum is then corrected for acceptance and reconstruction efficiency and fitted to the IW function $\xi_B(w)$ with the procedure discussed in the text.

- The SVD algorithm used here depends on a single parameter, which is the ordered Eigenvector index. Choice: k=4
- 14 bins are reconstructed, 7 unfolded

corrected spectrum. The SVD method includes a regularization procedure that depends upon a parameter $k$ [41], ranging between unity and the number of degrees of freedom, in our case 14. Simulation studies demonstrate that $k = 4$ is optimal in our case. Variations associated with different choices of $k$ have been studied and are included in the systematic uncertainties. We have performed closure tests with different simulation models of the $\Lambda_b^0 \to \Lambda_c^+ \mu^- \overline{\nu}_\mu$ dynamics, and verified that this unfolding procedure does not bias the reconstructed distribution. The spectra before and after unfolding are shown in Fig. 3.

# EM Iterative method

- Basic idea: instead of maximizing the likelihood using gradients (Minuit), iteratively improve the maximum
  (EM=Expectation / maximisation)
- Few iterations → close to start value
- Infinite iterations → exact solution
- Iteration prescription:

$$x_j^{(N+1)} = x_j^{(N)} / \epsilon_j \sum_i \frac{A_{ij}\, y_i}{\sum_k A_{ik}\, x_k^{(N)}}$$

- Typically one would apply such a method for a very large number of bins and sparse matrices, where exact inversion of A is not practical (image processing: $10^5$ pixels but most of the $10^{10}$ elements of A a zero)

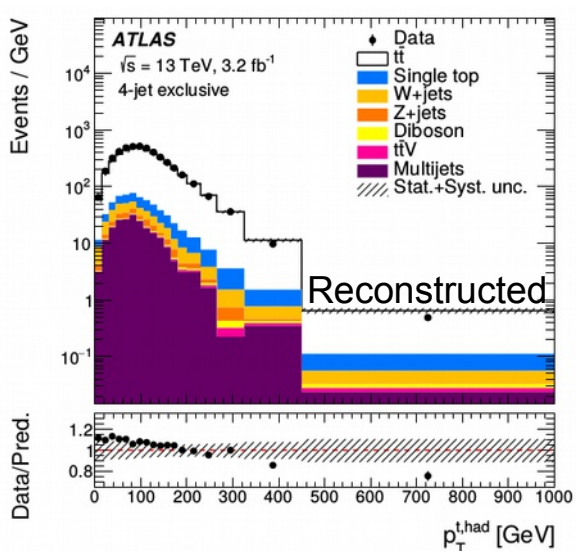
  → the method was introduced for tomography image processing by Shepp & Vardi [IEEE trans.med.im. MI-1 (1982) 113]

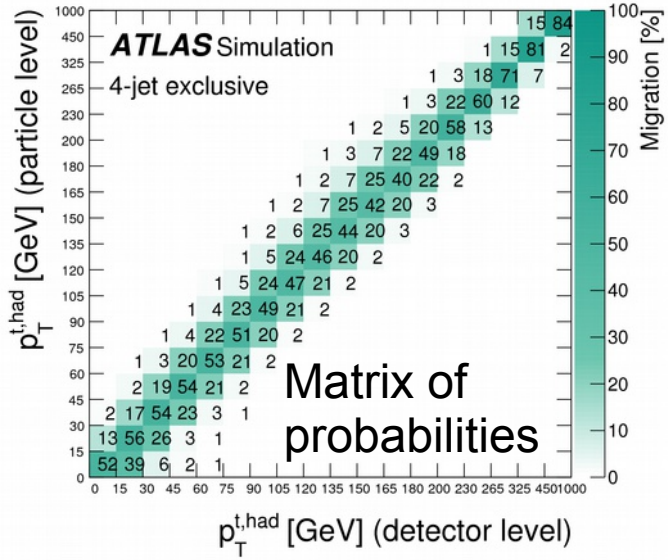
[background $b_i$ often is subtracted in the enumerator, but rather should be added in the denominator Ax→Ax+b]

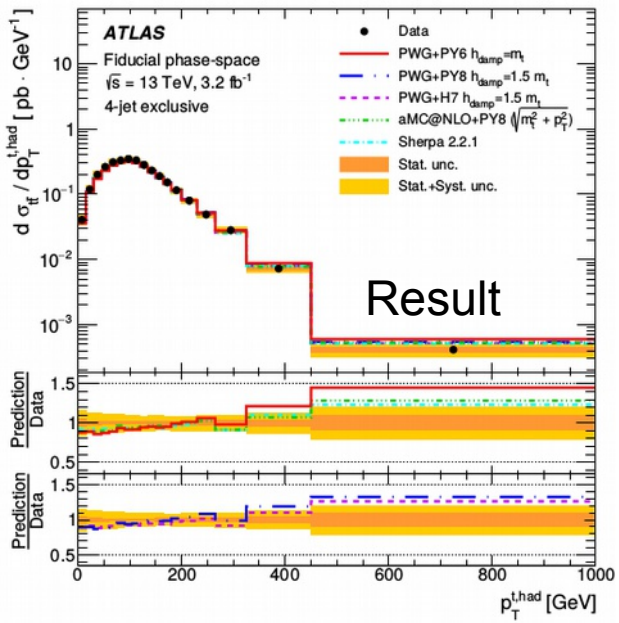
- For HEP, the method was introduced by Mülthei/Schorr [NIMA 257 (1987) 371]
- Reinvented as "Iterative Bayesian unfolding" by D'Agostini [NIMA 362 (1995) 487]
  Since then, used in many HEP analyses (because of its apparent simplicity?)

# Example analysis with iterative method

- ATLAS ttbar+jet [arXiv:1082.06572] http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/TOPQ-2017-01/
- Cross section is measured differential in transverse momenta




Background is subtracted from data

For each observable, the unfolding procedure starts from the number of events at reconstruction level in bin $j$ of the distribution ($N^j_{\text{reco}}$), after subtracting the background events estimated as described in Section 5 ($N^j_{\text{bg}}$). Next, the acceptance correction $f^j_{\text{acc}}$ is defined as the ratio of the nu

the elements of each row add up to unity (within rounding). The number of bins is optimised for maximum information extraction under stable unfolding conditions. This is achieved by requiring that closure and stress tests are satisfied without introducing any bias. The unfolding is performed using four iterations to balance the unfolding stability with respect to the previous iteration (below 0.1%) and the growth of the statistical uncertainty. The effect of varying the number of iterations by one was found to be negligible.

Choice of number of iterations: 4

# Choice of number of iterations

- Measurement shown on previous slide: 4 iterations. Motivated by:

  "unfolding stability wrt previous iteration"

- Similar criteria are applied in many other HEP analyses using iterative unfolding

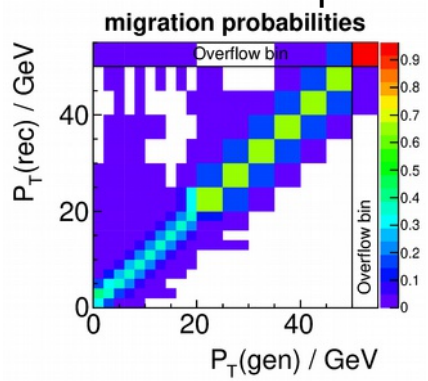
  In my opinion there is a problem with this criterion, because it depends critically on the start value:

  - Simulation out of the box ↔ larger number of iterations
  - Simulation tuned to data ↔ smaller number of iterations

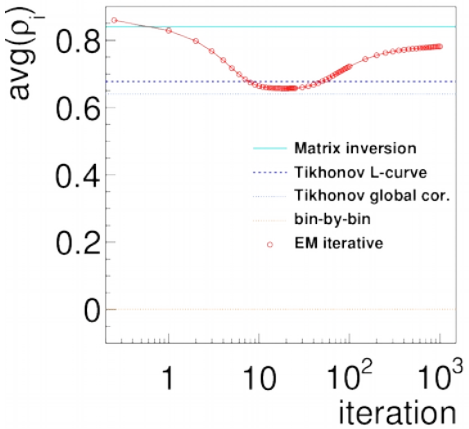
  But the stat. uncertainties and correlations grow with the number of iterations → if too small this may result in underestimated data statistical uncertainties

- Proposal: use a different objective to decide # iterations, selecting on properties of the covariance matrix (e.g. correlations)

Example: probability matrix from matrix inversion example




Global correlations are similar for Tihkonov and EM with N=20



# Summary

- This talk: present four unfolding algorithms which are used frequently in HEP analyses. Each has its own advantages and drawbacks

| Algorithm | Main advantage | Technical difficulties | Main disadvantage | Root Tools |
|---|---|---|---|---|
| Likelihood fit | unbiased | Often deadling with many bins | not regularized → correlations and oscillations between bins | RooFit |
| Least-square +Tikhonov | small and well controlled bias | Choice of τ, Binning | Using least-square and not Poisson → statistical bias | TUnfold TSVDUnfold |
| Bin-by-bin | simple to use | none | Large bias to simulation | RooUnfold |
| Truncated EM iterations | simple to use | Choose number of iterations | Bias is difficult to quantify | RooUnfold |

- Some ideas:
  - Tikhonov regularisation in Poisson Likelihood fits (RooFit with regularized unfolding?)
  - Try to use similar objectives to choose the regularisation strength for Tikhonov and iterative method, to enable direct comparisons. Example: minimum global correlstion coefficients.