ATLAS Unfolding Procedure in Run I

Florian Bernlochner, Dag Gillberg,
Bijan Haney, Sarah Heim, Andrew Pilkington, Michaela Queitsch-Maitland, Chris Meyer

LHCHXS Kick Off Meeting - June 24, 2015

Introduction

$$\sigma_i = \frac{N_i^{\text{signal}}}{\mathcal{L}_{\text{int}} \cdot \boxed{C_i}}$$

Unfolding factors are used to undo the effect of smearing due to detector resolution and efficiency. Unfolding has to be done carefully, as mis-modeling will introduce bias into the results.

- ATLAS has used both bin-by-bin correction factors and bayesian iterative unfolding factors in the Higgs cross section measurement.
  - Bin-by-bin correction factors are the main results, while bayesian iterative methods have been used as cross-checks.
- The strategy we follow is to accept a small bias (systematic error) in exchange for a large reduction in variance (statistical error).
  - Systematic errors on both the MC generator modeling and detector modeling are used in order to cover areas where bias may be present.

Unfolding Method

$$\frac{1}{C_i} = \frac{N_i^{\text{Fid}}}{N_i^{\text{Reco}}} = \frac{P_i}{\epsilon_i} \qquad P_i = \frac{N_i^{\text{Fid\&Reco}}}{N_i^{\text{Reco}}} \qquad \epsilon_i = \frac{N_i^{\text{Fid\&Reco}}}{N_i^{\text{Fid}}}$$

- Bin-by-bin correction factors, c_i, are calculated from MC simulations in order to correct for detector effects.
  - N^{Fid} is the # of truth level MC events after event selection within a fiducial volume.
  - N^{Reco} is the # of MC events after event selection with detector effects (e.g. gaps in the detector, Jet reconstruction efficiency, other smearing effects, etc.)
  - $N^{\text{Fid\&Reco}}$ are events that pass the Higgs event selection under both circumstances.
- Purity, P_i, accounts for the number of fakes in a bin.
- Efficiency, ϵ_i, accounts for poor object reconstruction and identification.

Bias

- If the model you use does not perfectly describe the real data, you will introduce a bias.

$$\delta N_i^{\mathrm{Fid}} = N_i^{\mathrm{Reco}} \times \left(\left(\frac{N_i^{\mathrm{Fid}}}{N_i^{\mathrm{Reco}}} \right)_{\mathrm{Model}} - \left(\frac{N_i^{\mathrm{Fid}}}{N_i^{\mathrm{Reco}}} \right)_{\mathrm{Reality}} \right)$$

- This bias is proportional to the off-diagonal terms of the response matrix.

$$\delta N_i^{\mathrm{Fid}} = N_i^{\mathrm{Reco}} \times \sum_{i \neq j} R_{ij}^{-1} \left(\left(\frac{N_j^{\mathrm{Reco}}}{N_i^{\mathrm{Reco}}} \right)_{\mathrm{Model}} - \left(\frac{N_j^{\mathrm{Reco}}}{N_i^{\mathrm{Reco}}} \right)_{\mathrm{Reality}} \right)$$

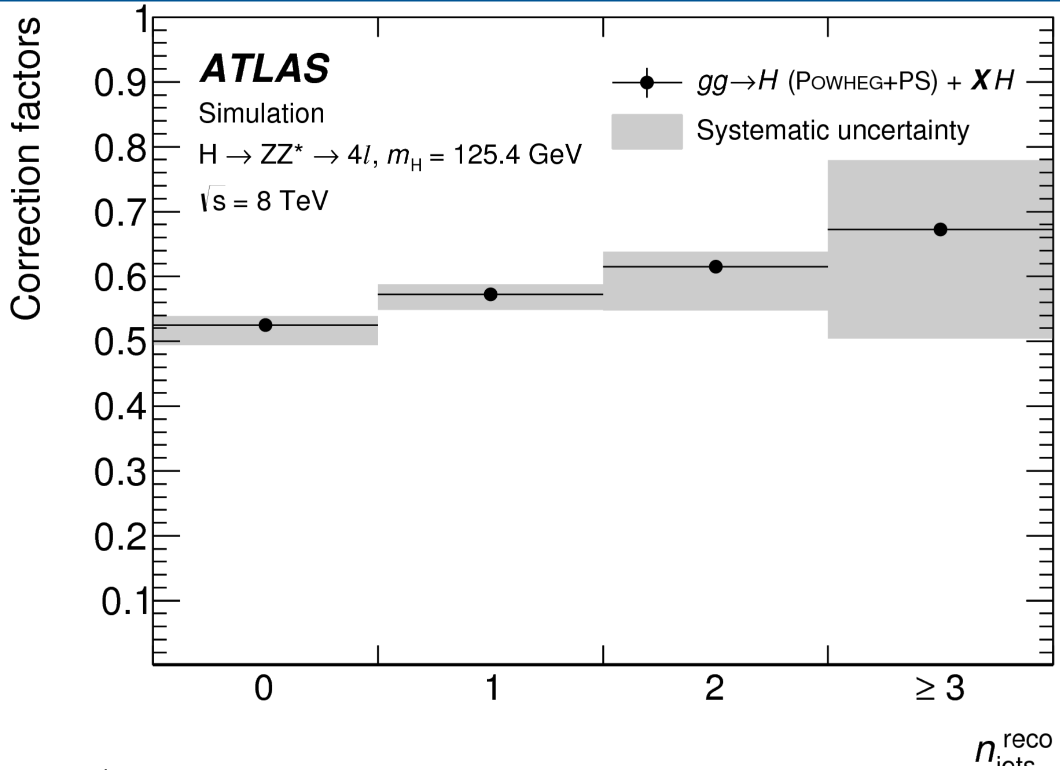
- Therefore, the size of this bias goes to zero as the response matrix becomes diagonal
  - In other words, the bin-to-bin migrations should be small to use this method.
- For variables with non-negligible off-diagonal elements (e.g. jet variables), systematic uncertainties are assigned to cover possible variations in the signal model.

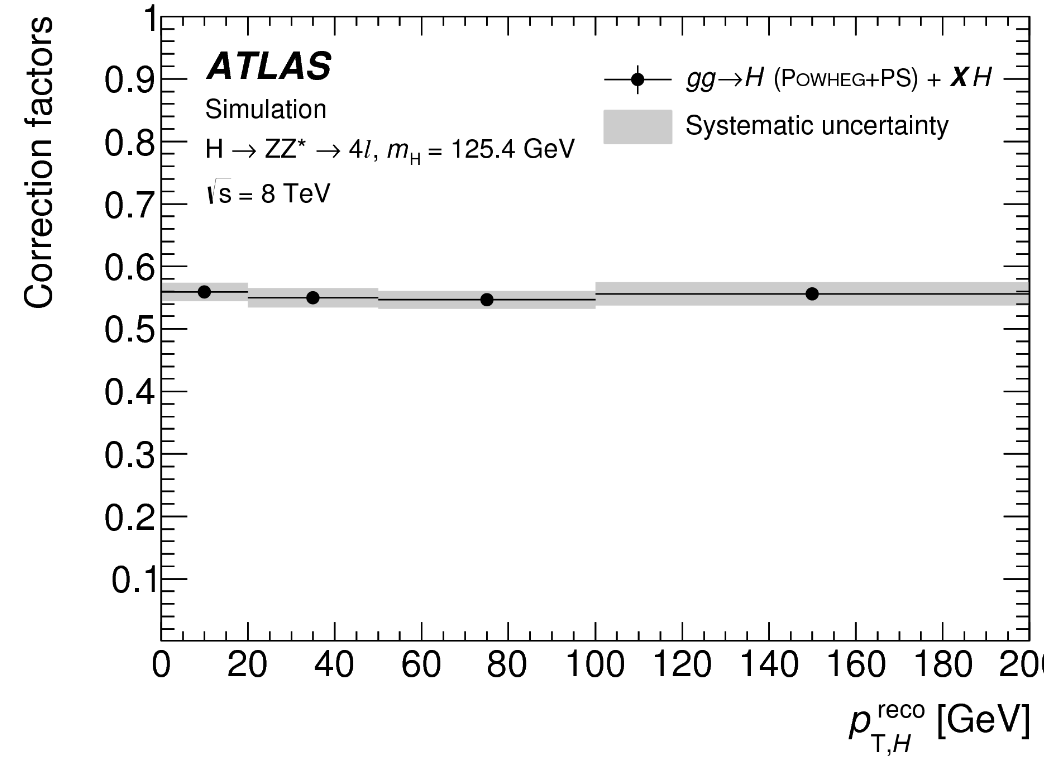
Uncertainty

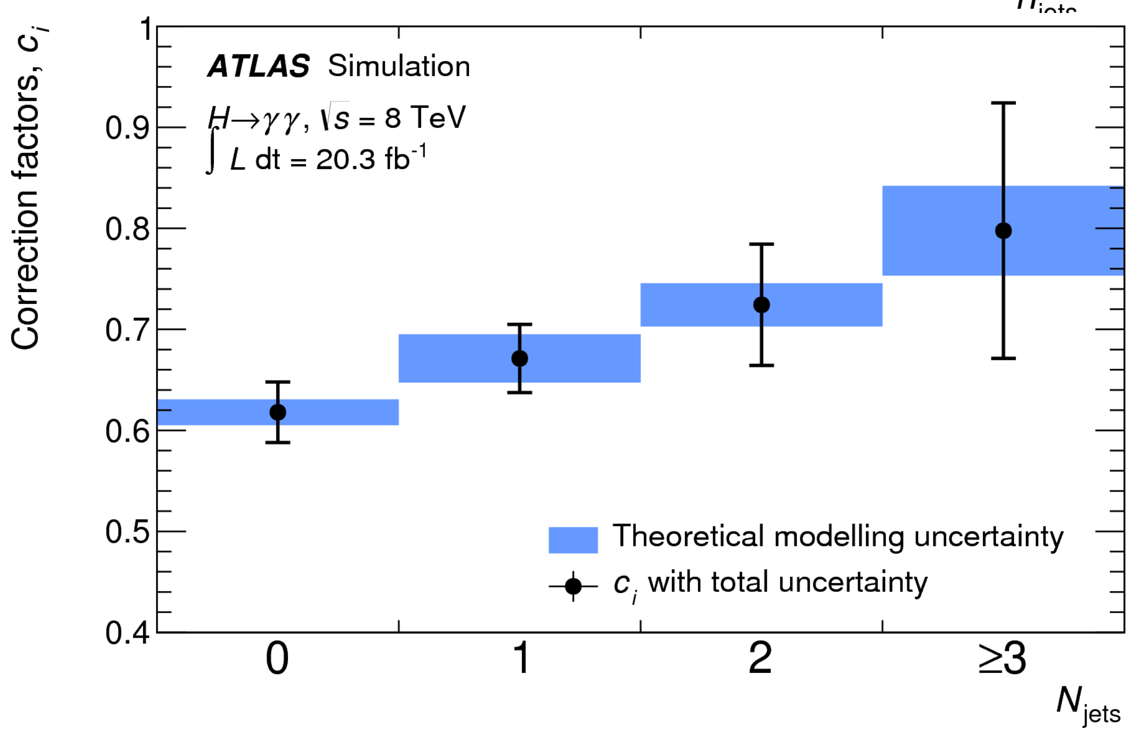
There is uncertainty in both N^{Reco} and N^{fid} because the generators themselves may not match reality. There is extra uncertainty specifically in N^{Reco} because the MC smearing may not match reality.

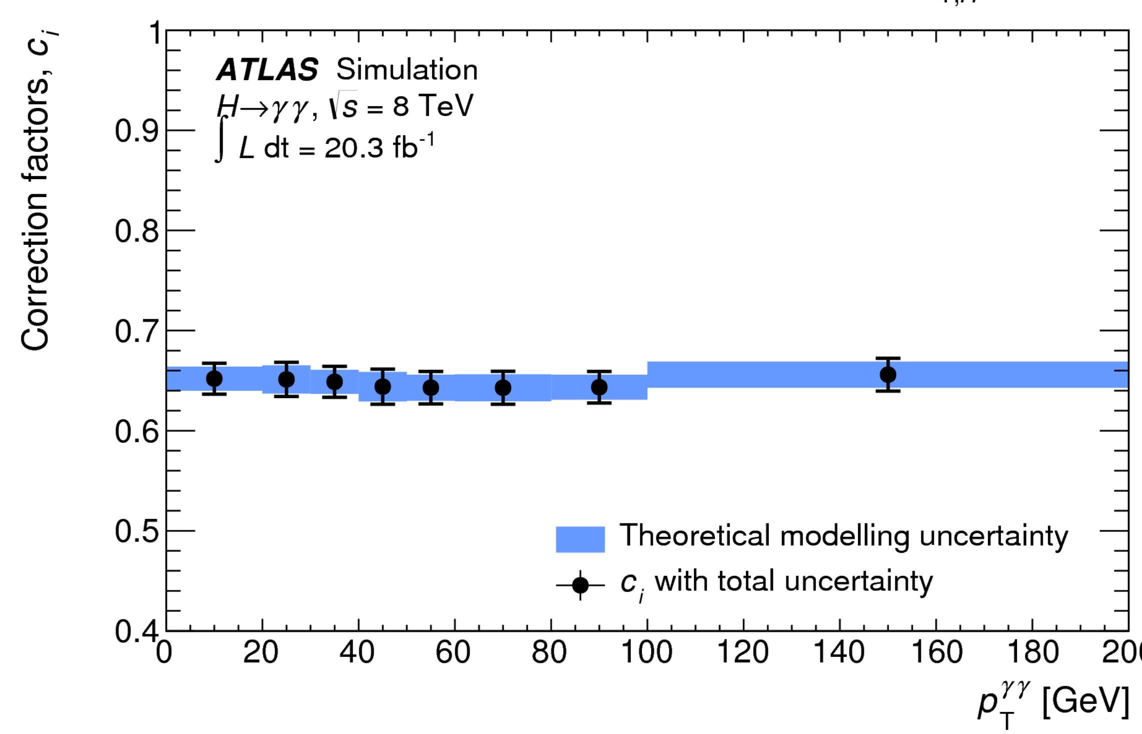
- Generator Modeling and Uncertainty
  - Alternative MC generators were used and their envelope was taken as an uncertainty.
    - eigenvector variations of the baseline CT10 PDF.
    - central values of alternative MSTW2008NLO and NNPDF2.3 PDFs.
  - Signal composition of the production modes was varied.
    - VBF+WH+ZH production XS were doubled and halved.
    - ttH production XS was multiplied x5 and x0.
  - Varying the renormalization and factorization scales by double and a half.
  - Reweighing was applied to the MC to make it more closely reflect the observed distribution of data.
    - The unfolded data distributions of p_T and |y| were compared to fiducial MC predictions. Reweighing functions from data/MC were used to correct the fiducial p_T and |y| spectrum.
    - Data tend to have harder Higgs p_T, and more forward |y|.

Effects on Correction factor









Jet variables had higher systematic uncertainties than Higgs variables from bigger off-diagonal terms in the response matrix.

Iterative Unfolding

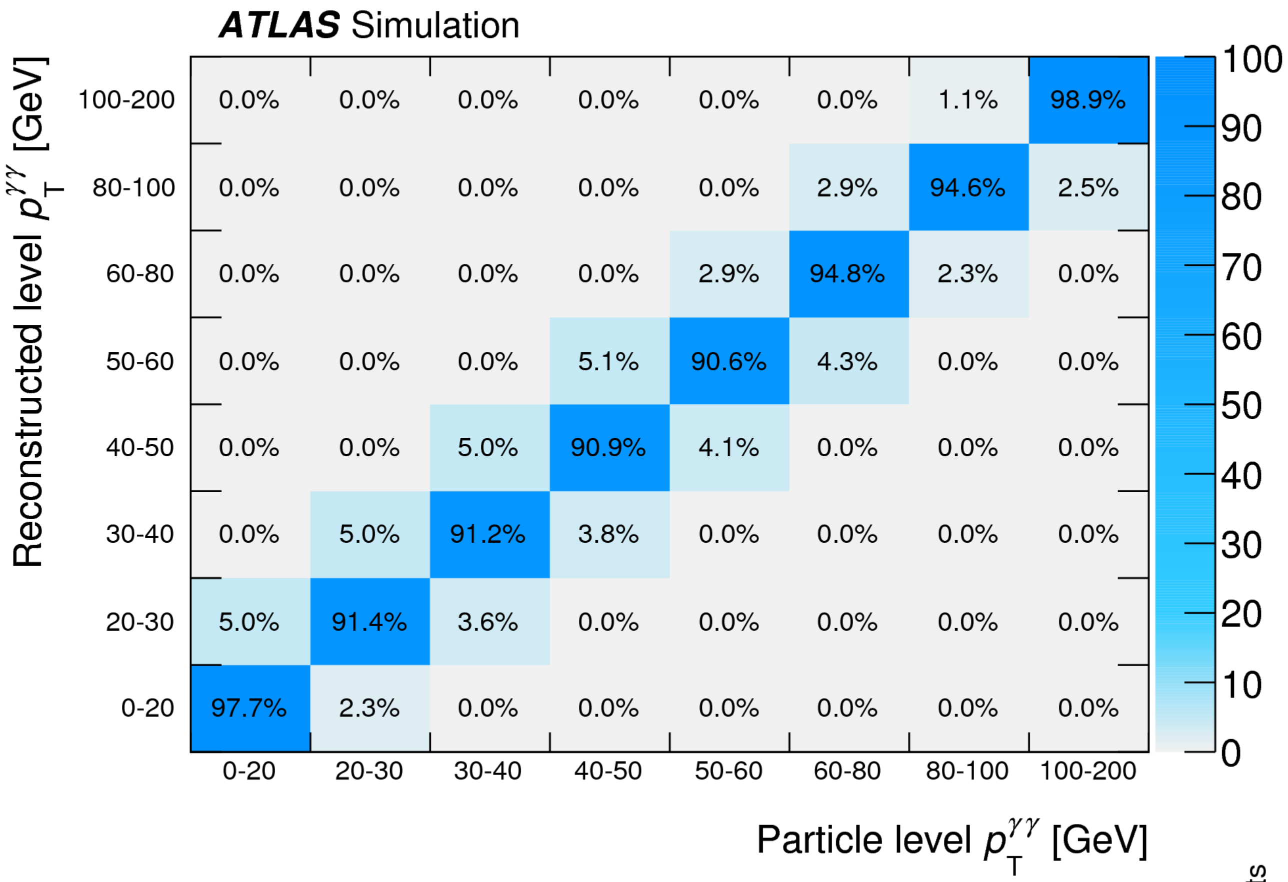
- Another method is to use iterative Bayesian unfolding to converge of an unfolding factor.
  - This was used a cross check to the bin-by-bin correction factors. Where the bayesian method converged, the central values in bin-by-bin were also the same.
- ‘C’ = causes and ‘E’ = effects. In our case, ‘C’ correspond to the fiducial values before smearing, and ‘E’ correspond to the values after detector reconstruction.
- P(E|C) is essentially our response matrix, and what we want to know is P(C|E), to get the true value of the event from what we observed.
- P(C|E) can be used to figure out n(C), which can be fed back into P(C|E) iteratively.

$$\hat{n}(C_i) = \frac{1}{\varepsilon_i} \sum_j n(E_j) P(C_i|E_j) \qquad P(C_i|E_j) = \frac{P(E_j|C_i) P_0(C_i)}{\sum_k P(E_j|C_k) P_0(C_k)}$$

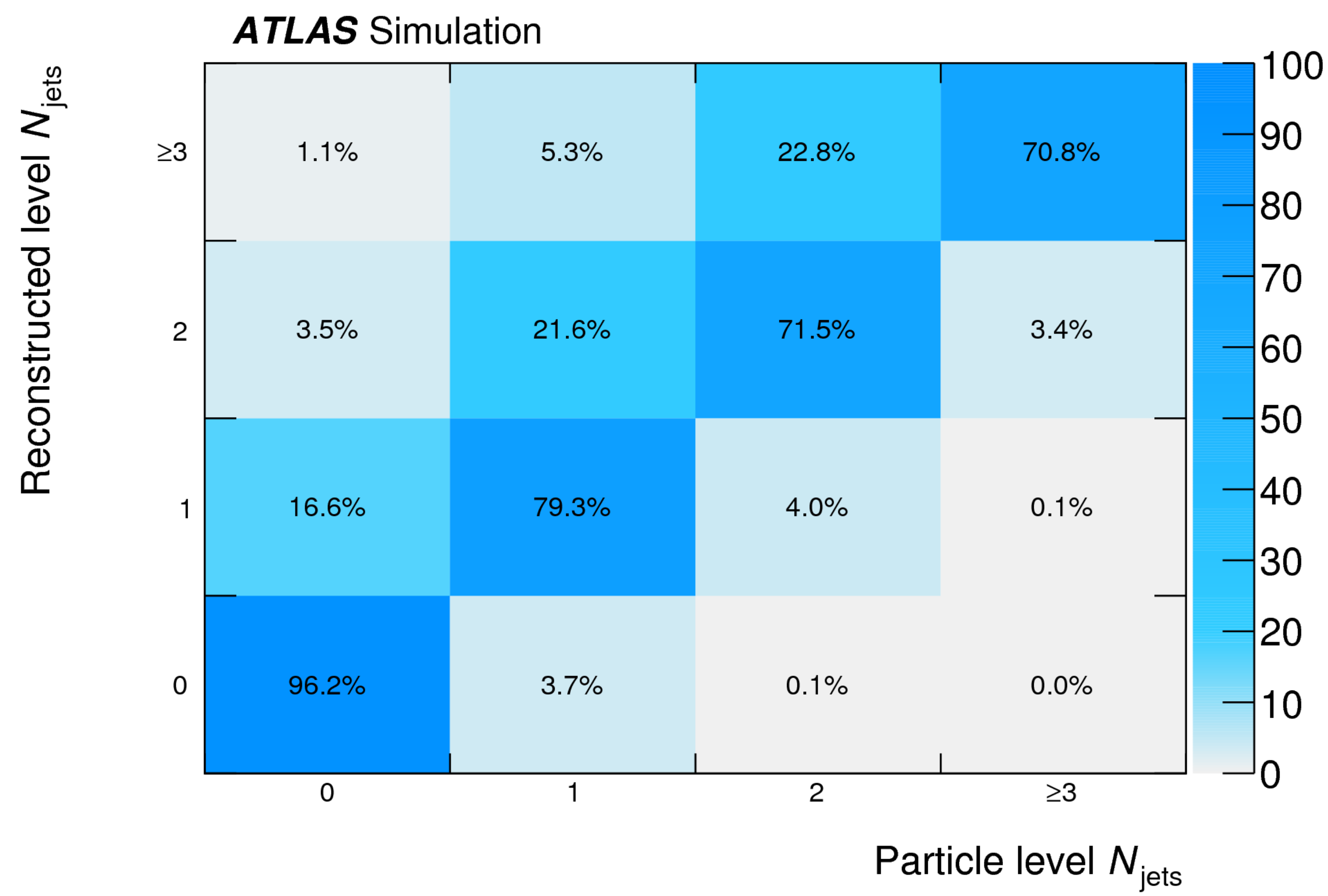
n(C) = expected #of events in the bin.
n(E) = observed number of events in the bin.
ϵ_i = efficiency for the bin.
P(C) = Priors and iterated priors.

$$P(C_i) = P(C_i|n(E)) = \frac{n(C_i)}{\sum_j n(E_j) P(C_i|E_j)}$$

Response Matrices



- Higgs variables have little migration.



- Jet variables have more migration.

Using the Response matrix

$$N_i^{\mathrm{sig}} = \mathcal{L}_{\mathrm{int}} \cdot \sigma_i \cdot C_i$$

$$N_i^{\mathrm{sig}} = \mathcal{L}_{\mathrm{int}} \cdot \epsilon_i \sum_j \epsilon_j' \sigma_j M_{ij}$$

$$\epsilon_i' = \frac{N_i^{\mathrm{Reco}} \cap N_i^{\mathrm{Fid}}}{N_i^{\mathrm{Fid}}} \qquad \epsilon_i = \frac{N_i^{\mathrm{Reco}}}{N_i^{\mathrm{Reco}} \cap N_i^{\mathrm{Fid}}}$$

- To create more model independence, one can also separate the correction factor out into its response matrix M_{ij}, and two transfer terms, ϵ_i and ϵ_i', taken from Monte Carlo.
- The transfer term ϵ_i' takes out MC fiducial events that are not also MC reconstructed events, while ϵ_i puts in events that are MC reconstructed events but not MC fiducial.
- The response matrix captures the resolution effects of the detector while making less assumptions about how the truth distribution looks like.

Conclusion

- Bin-by-bin is not perfect. There are reasons why one should and should not use it.
  - Bin-by-bin is much more sensitive to biased MC modeling.
  - But the total statistical power remains the same.
  - Bias in the MC model is taken into account in the systematic uncertainties.
  - There is a trade-off between variance and bias.
  - Using the migration matrix would make the unfolding less model dependent and reduce the systematic errors.

“Unfolding is a complicated business and one is well advised to ask in each problem if it can be avoided.”

Glen Cowan

source: https://www.ippp.dur.ac.uk/old/statistics/proceedings/cowan.ps