Performance and improvements of different unfolding methods

K. Bierwagen, U. Blumenschein, A. Quadt

2nd Institute of Physics, Georg-August-University Göttingen

May 27, 2010

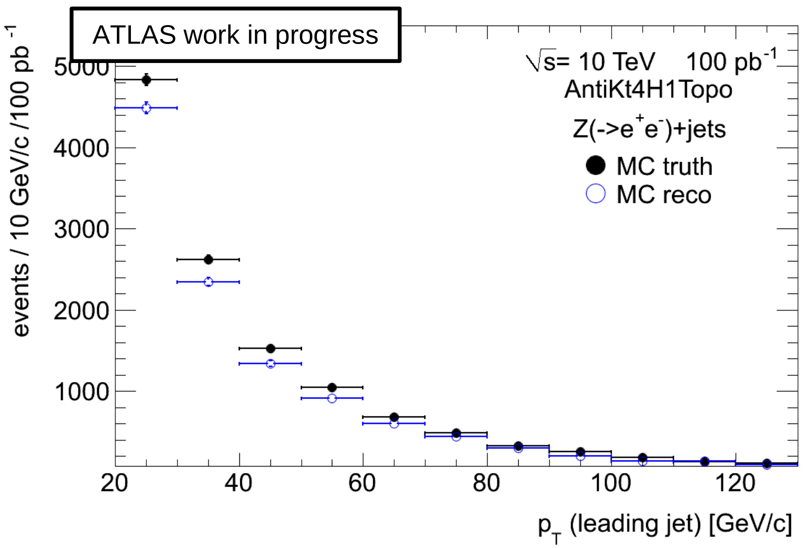
PHYSICS AT THE TERA SCALE
Helmholtz Alliance

FSP 101 ATLAS

Bundesministerium für Bildung und Forschung

Introduction

- Plan:
  - Correction for detector effects in data
- Effects:
  - Migration
  - Efficiency/acceptance
  - Resolution

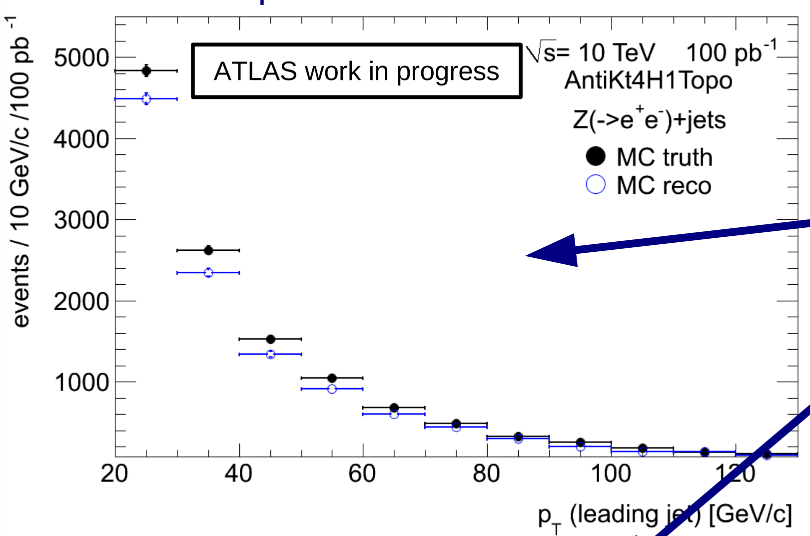


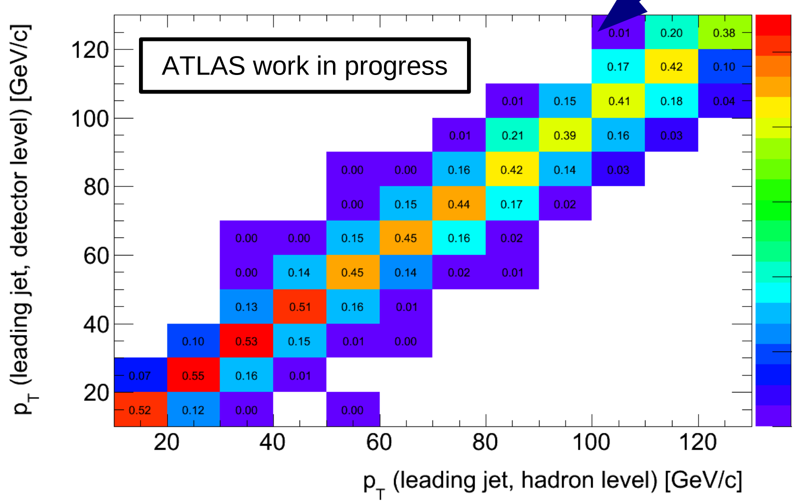
- Preparations:
  - Performance checks of available methods
  - Develope new methods
  - Improve the different methods
  - Comparison of the methods

Iterative (Bayes) Method

Iterative (Bayes) Methode

Input distributions



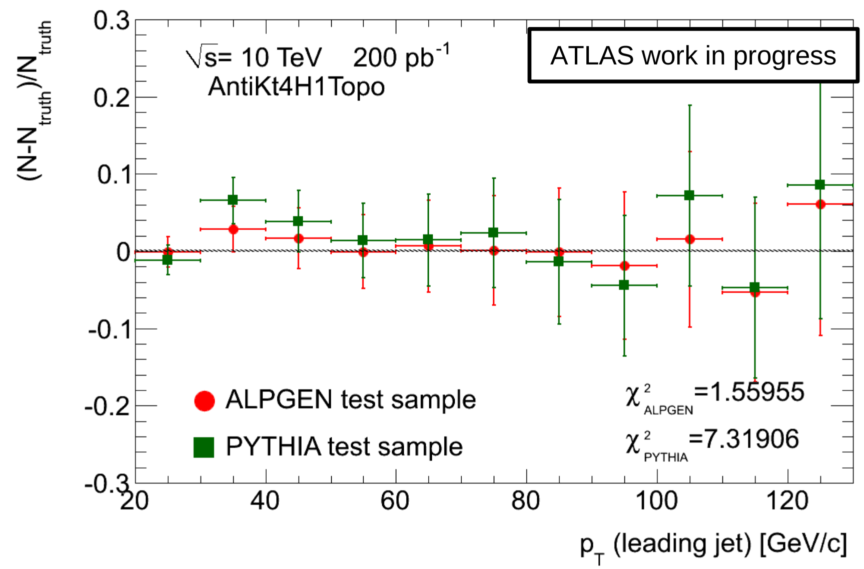


Correction for fake jets is applied

Iterative (Bayes) Methode:

- Use C++ implementation of G. D'Agostini 's paper from Marisa Sandhoff
- Training distributions (true T_i and reconstructed M_j)
- Migration matrix R_{ji} (R_{ji} fraction of events of T_i in M_j)
- Iteration step: $$\hat{\mu}_i =: \frac{1}{\varepsilon_i}\sum_{j=1}^{N}\frac{R_{ji}\,p_i}{\sum_k R_{jk}\,p_k}n_j$$

Correction aplied on two different test samples



→ Uncertainties seems to be too large
→ Check method using a simple Toy Mc

Test of C++ implementation

- Define a migration matrix

$$M_1 = \begin{pmatrix} 0 & 0.1 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.8 & 0.6 & 0.4 \end{pmatrix} \qquad M_2 = \begin{pmatrix} 0 & 0.1 & 0.8 \\ 0.2 & 0.8 & 0.2 \\ 0.8 & 0.1 & 0 \end{pmatrix} \qquad M_3 = \begin{pmatrix} 0 & 0.025 & 0.95 \\ 0.05 & 0.95 & 0.05 \\ 0.95 & 0.025 & 0 \end{pmatrix}$$

large migration   medium migration   low migration

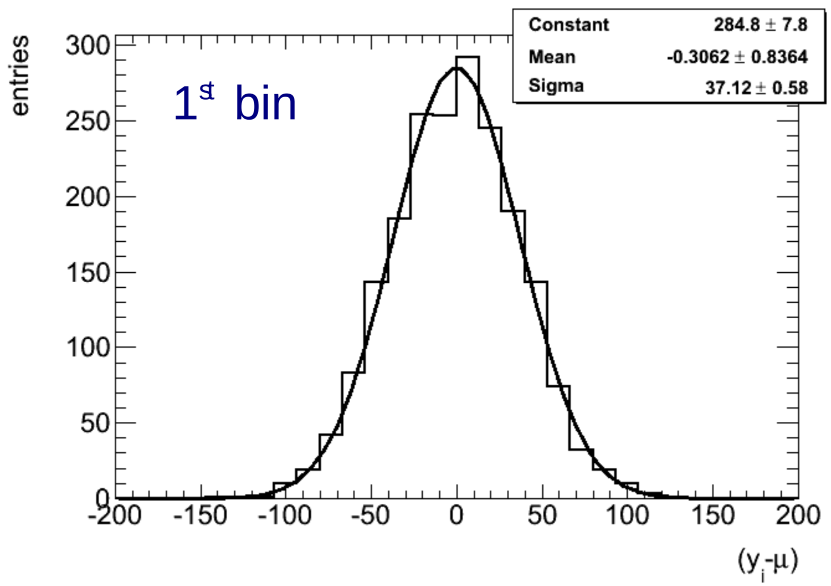
- Create a truth distribution with 3 bins with 10000 entries each
- Create randomly a test distribution
- Calculate the unfolded distribution with the program and manually

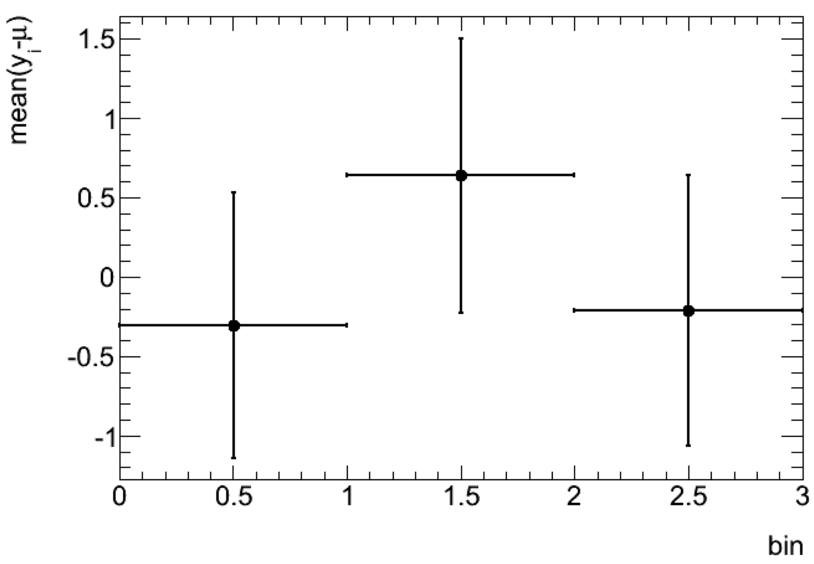
➔ Both calculations give the same result

➔ Method is correctly implemented

Method checks 1

- Define a migration matrix (M_2)
- Create a truth distribution with 3 bins with 10000 entries each
- Create randomly 2000 test distributions and calculate the unfolded distribution
- Compare unfolded distribution with the true distribution

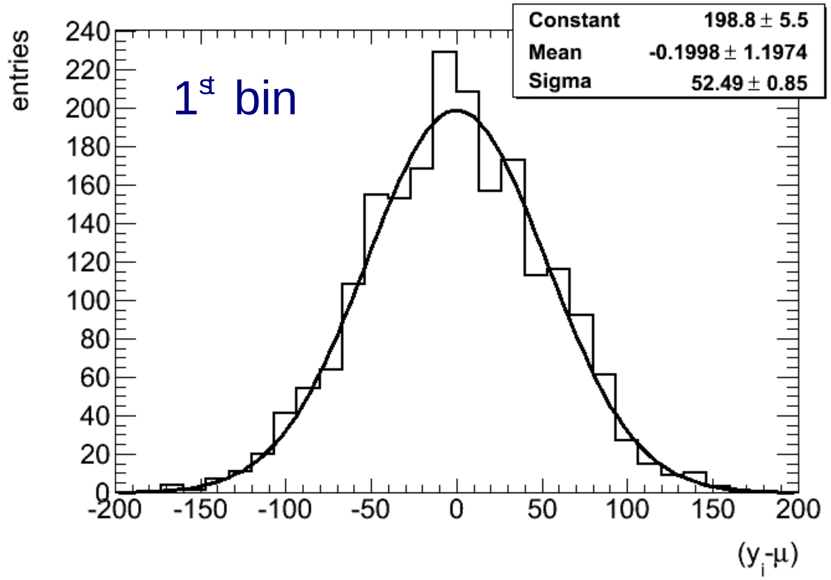


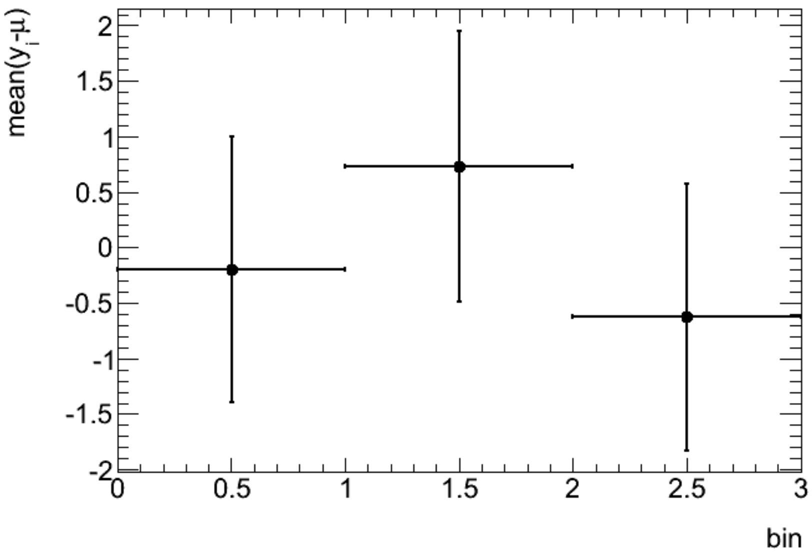


- No bias visible

Method checks 2

- Create randomly 2000 migration matrices from fixed probabilties (M_2)
- Create a truth distribution with 3 bins with 10000 entries each
- Create randomly 2000 test distributions and calculate the unfolded distribution

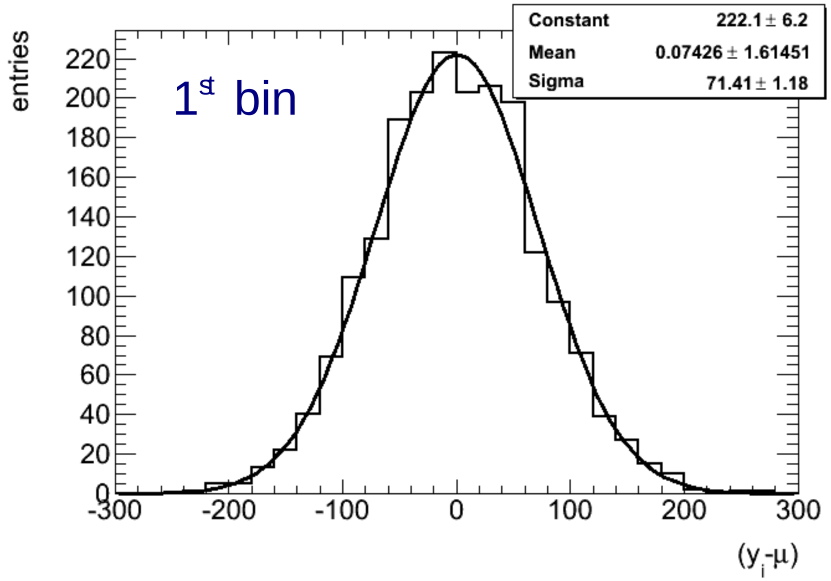


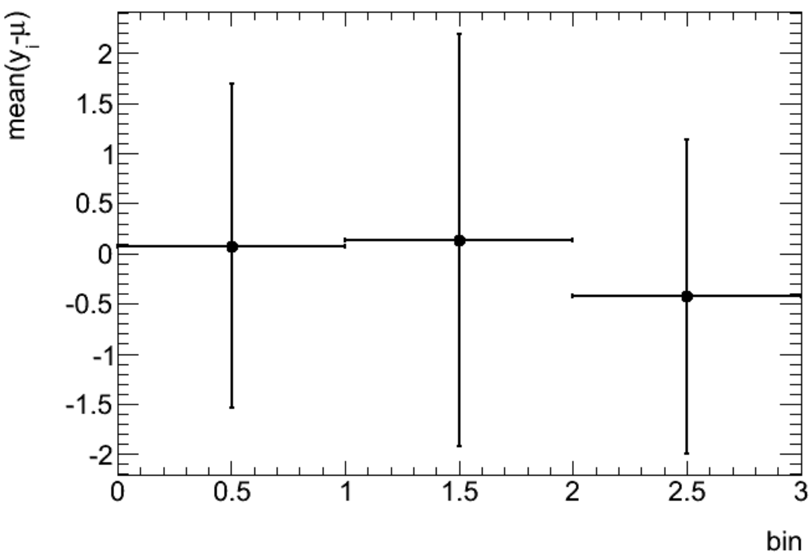


- No bias visible

Method checks 3

- Create randomly 2000 migration matrices from fixed probabilties (M_2)
- Create randomly 2000 uniformly distributed truth distributions with 30000 entries
- Create randomly 2000 test distributions and calculate the unfolded distribution





- No bias visible

Uncertainties

$$\hat{n}(C_i) = \sum_{j=1}^{n_E} M_{ij} \cdot n(E_j)$$

$$M_{ij} = \frac{P(E_j|C_i) \cdot P_o(C_i)}{\left[\sum_{l=1}^{n_E} P(E_l|C_i)\right] \cdot \left[\sum_{l=1}^{n_C} P(E_j|C_l) \cdot P_o(C_l)\right]}$$

- M_{ij} terms of the unfolding matrix M
- M is clearly not equal to the inverse of the migration matrix
- $P_0(C_i)$: initial probabilities
- $n(E_j)$: data sample
- $P(E_j|C_i)$: migration probabilities

- Sources of uncertainties:
  - $P_0(C_i)$: no uncertainty is introduced
  - $n(E_i)$: data is assumed to be mutinomial distributed

$$V_{kl}\left(\underline{n}(E)\right) = \sum_{j=1}^{n_E} M_{kj} \cdot M_{lj} \cdot n(E_j) \cdot \left(1 - \frac{n(E_j)}{\widehat{N}_{true}}\right) - \sum_{\substack{i,j=1 \\ i \neq j}}^{n_E} M_{ki} \cdot M_{lj} \cdot \frac{n(E_i) \cdot n(E_j)}{\widehat{N}_{true}}$$

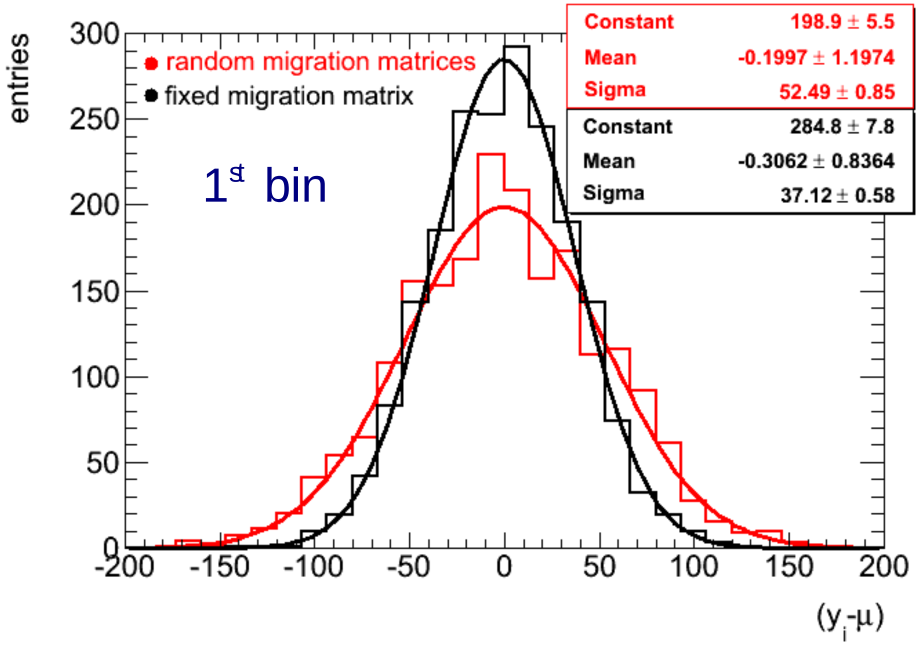
  - $P(E_j|C_i)$:

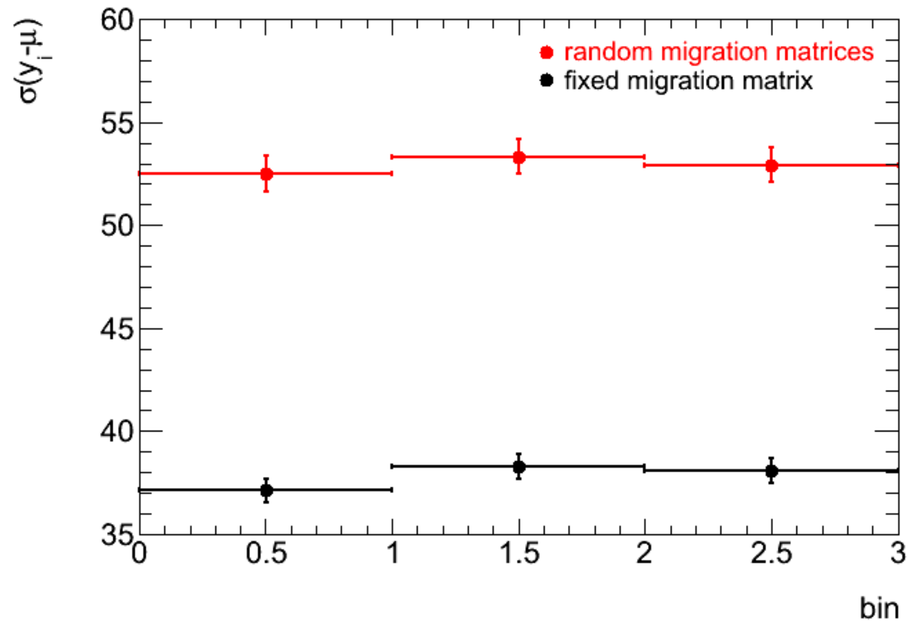
$$V_{kl}\left(\mathbf{M}\right) = \sum_{i,j=1}^{n_E} n(E_i) \cdot n(E_j) \cdot Cov(M_{ki}, M_{lj})$$

  - Total uncertainty: $V_{kl} = V_{kl}\left(\underline{n}(E)\right) + V_{kl}\left(\mathbf{M}\right)$

Uncertainties

- Absolute uncertainties from ensemble tests with and w/o fluctuations in the matrix (M_2)

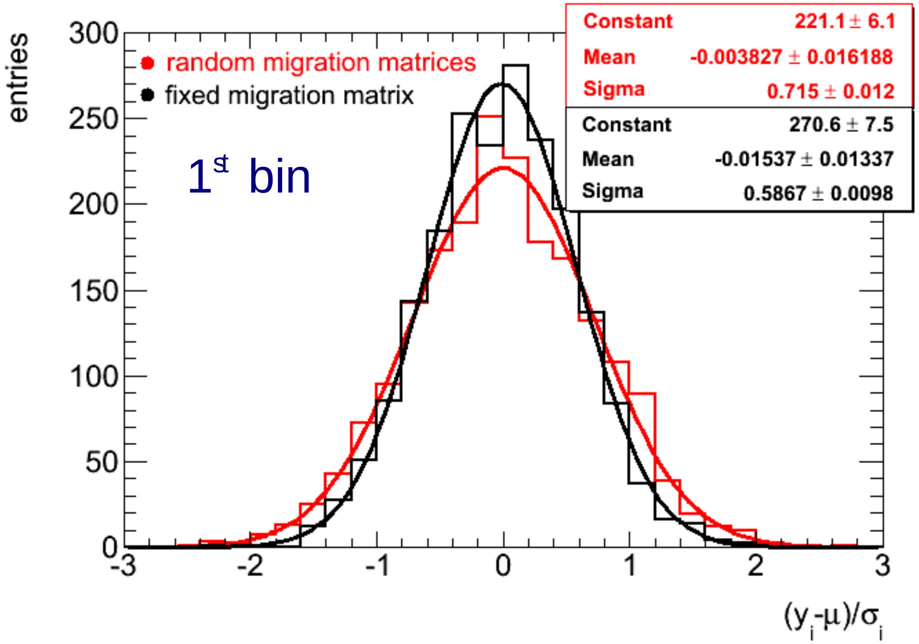


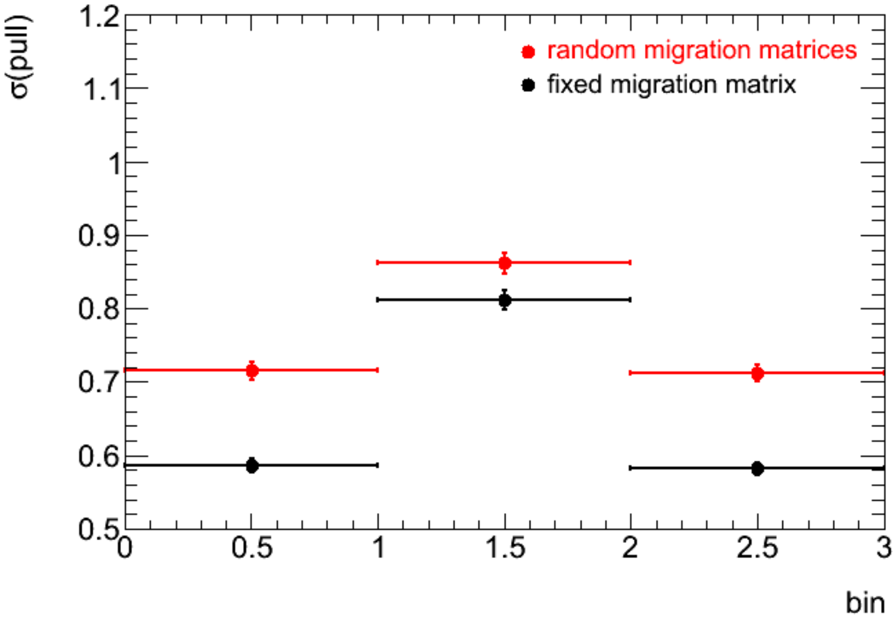


- Fluctuations in the matrix increases the uncertainty by a factor of 1.4 in this case

Uncertainties

- Create pull distribution → comparison between uncertainties from ensemble tests and the program (M_2)





- Uncertainties given from the program are too large (σ(pull)<1.0)

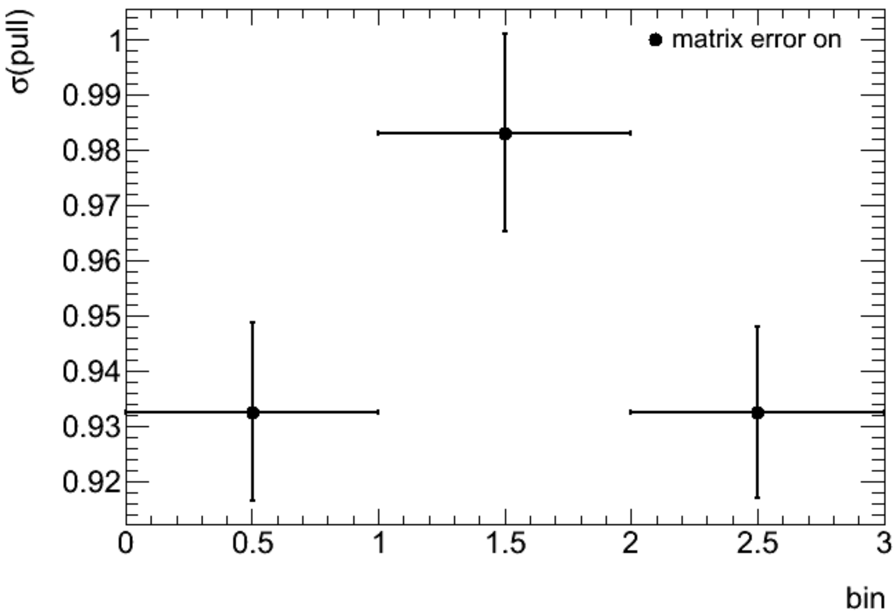
➔ Seems that fluctuations in data are not treated correctly

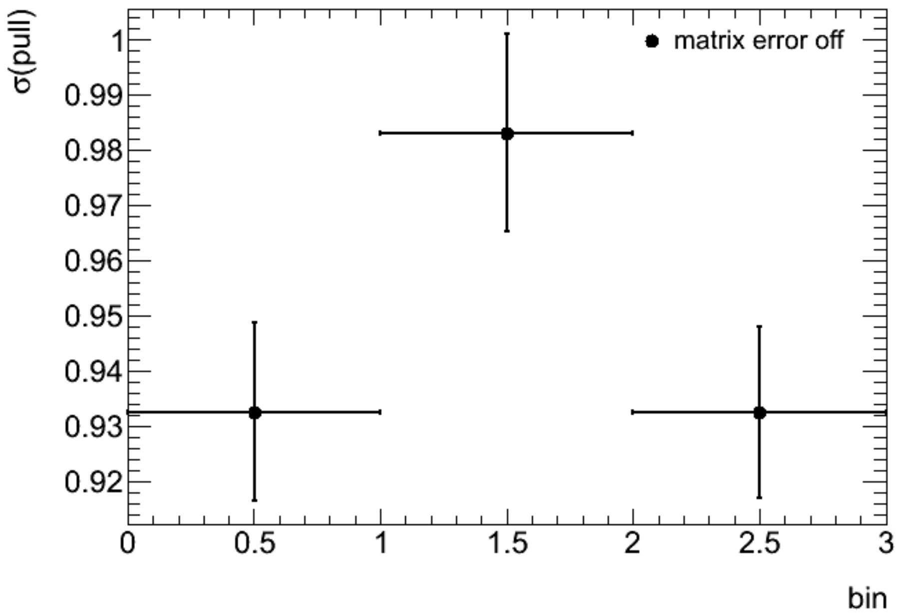
- Pull distributions with and w/o fluctuations in the matrix are not equal

➔ Seems that fluctuations in the matrix are not treated correctly

Uncertainties

- Infinite statistics in the migration matrix $\rightarrow$ contribution close to 0 to the total uncertainty
- Compare the pull distributions for the fixed migration matrix with uncertainty on and off on the migration matrix

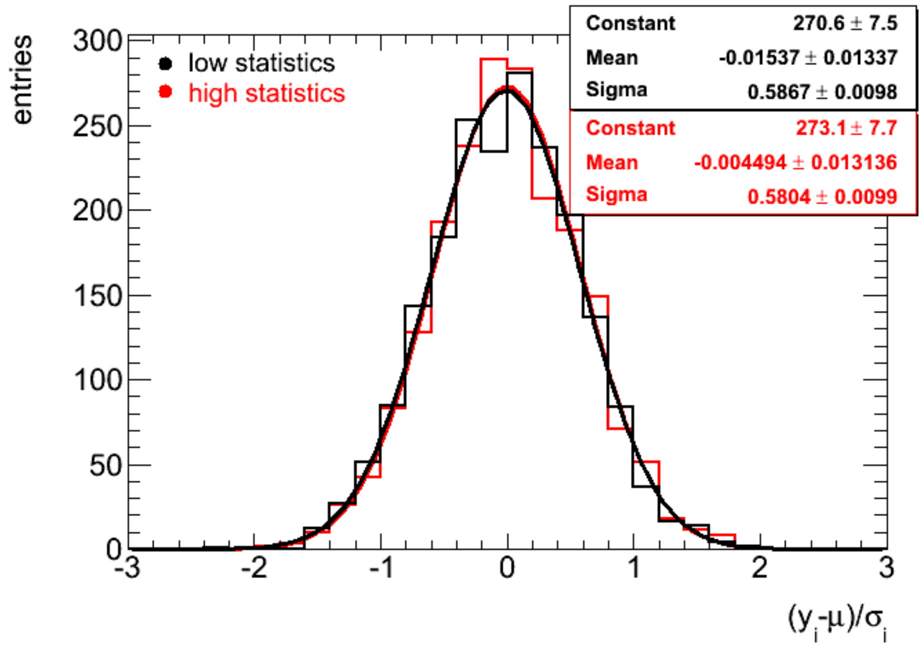


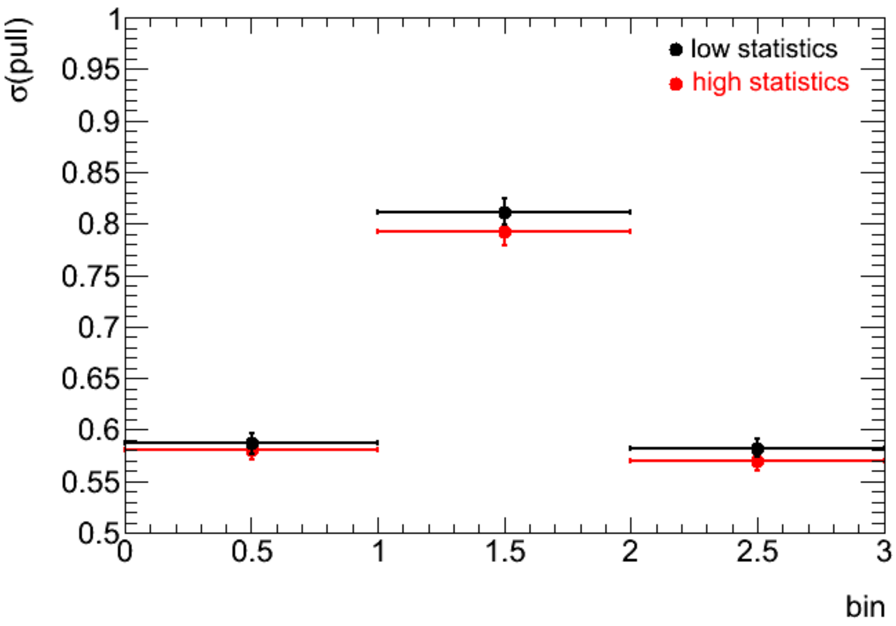


- Calculation of the uncertainty on the migration matrix for infinite statistics seems to work correctly

Uncertainties

- Comparison between uncertainties from ensemble tests and the program for high and low statistics in data with infinite statistics in the training

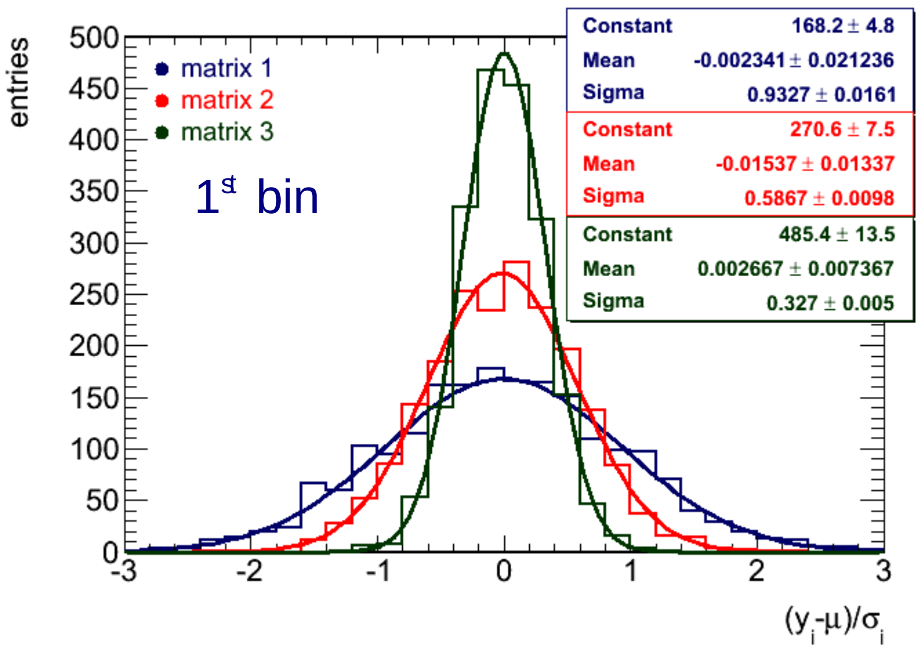


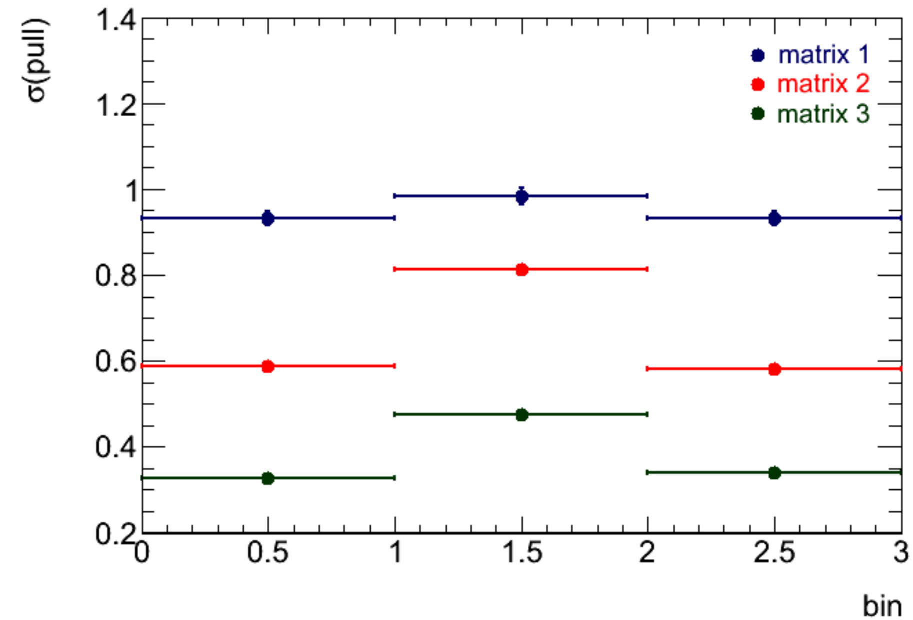


- As expected the influence of the amount of statistics in data is very small

Uncertainties

- Comparison between uncertainties from ensemble tests and the program for 3 different migration matrices





- Less migration leads to an over estimation of the uncertainties

➔ Migration effect is not treated correctly in the error calculation

➔ Assumptions for the error calculation have to be checked

Uncertainties

- Problem: Program assumes a multinomial distribution for the data
- Multinomial distribution:

$$var = np_j \cdot (1 - p_j)$$
$$cov = -np_i\, p_j$$

- But each bin is multinomial distributed
- The sum of multinomial distributions is only a multinomial distribution if all distributions are the same

➔ The columns of the migration matrix has to be equal to get the correct estimate for the uncertainty

➔ Not the typical case in data analysis

- Example 1 (large migration):

$$\begin{pmatrix} 0 & 0.1 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.8 & 0.6 & 0.4 \end{pmatrix} \longrightarrow \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.16 & -0.16 \\ 0 & -0.16 & 0.16 \end{pmatrix} + \begin{pmatrix} 0.09 & -0.03 & -0.06 \\ -0.03 & 0.21 & -0.18 \\ -0.06 & -0.18 & 0.24 \end{pmatrix} + \begin{pmatrix} 0.09 & -0.05 & -0.04 \\ -0.05 & 0.25 & -0.2 \\ -0.04 & -0.2 & 0.24 \end{pmatrix} = \begin{pmatrix} 0.18 & -0.08 & -0.1 \\ -0.08 & 0.62 & -0.54 \\ -0.1 & -0.54 & 0.64 \end{pmatrix}$$

program

$$\frac{1}{3}\cdot\begin{pmatrix} 0.2 \\ 1.0 \\ 1.8 \end{pmatrix}$$

Calculate covariance matrix

$$\begin{pmatrix} 0.187 & -0.07 & -0.14 \\ -0.07 & 0.7 & -0.6 \\ -0.12 & -0.6 & 0.72 \end{pmatrix}$$

- As mentioned before both calculations give different results

- Example 2 (low migration):

$$\begin{pmatrix} 0 & 0.1 & 0.8 \\ 0.2 & 0.8 & 0.2 \\ 0.8 & 0.1 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.16 & -0.16 \\ 0 & -0.16 & 0.16 \end{pmatrix} + \begin{pmatrix} 0.09 & -0.08 & -0.01 \\ -0.08 & 0.16 & -0.08 \\ -0.01 & -0.08 & 0.09 \end{pmatrix} + \begin{pmatrix} 0.16 & -0.16 & 0 \\ -0.16 & 0.16 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0.25 & -0.24 & -0.01 \\ -0.24 & 0.48 & -0.24 \\ -0.01 & -0.24 & 0.25 \end{pmatrix}$$

program

$$\frac{1}{3} \cdot \begin{pmatrix} 0.9 \\ 1.2 \\ 0.9 \end{pmatrix}$$

Calculate covariance matrix

$$\begin{pmatrix} 0.63 & -0.36 & -0.27 \\ -0.36 & 0.72 & -0.36 \\ -0.27 & -0.36 & 0.63 \end{pmatrix}$$

- As mentioned before both calculations give different results
- Differences become larger for less migration in the migration matrix

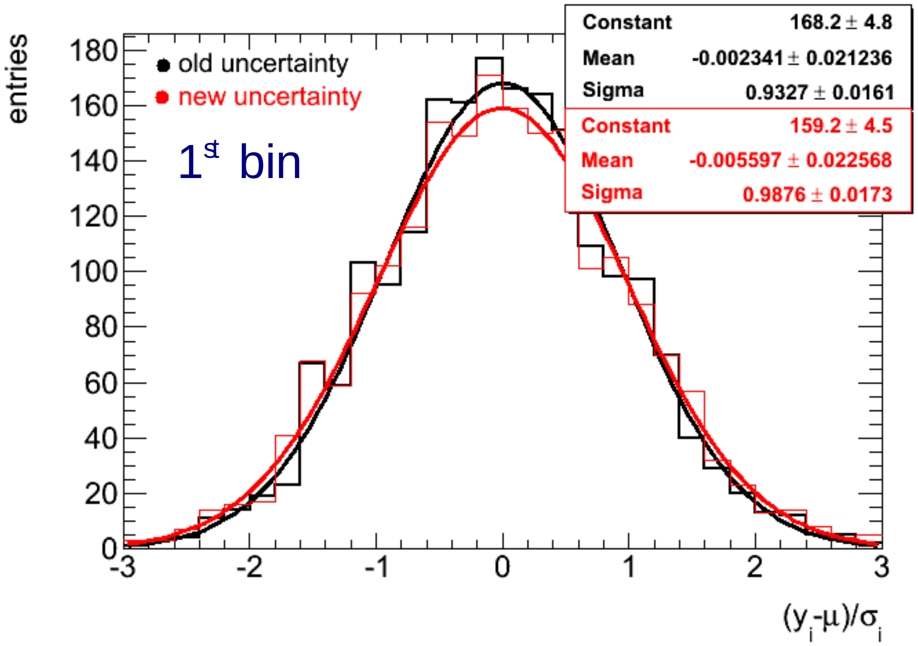
Uncertainties

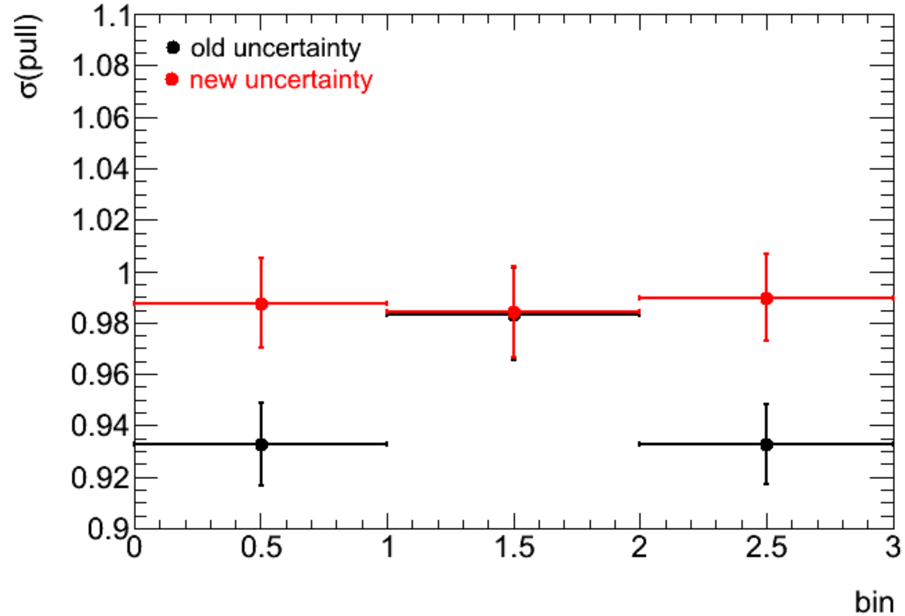
- Implement the new uncertainty calculation for the data into the program
- Assumption: The data sample is a realization of a sum of multinomial distributions

$$V_{kl}(\underline{n}(E)) = \sum_{j=1}^{n_E} M_{kj} \cdot M_{lj} \cdot \sum_{r=1}^{n_E} \hat{n}(C_r) \cdot P(E_j|C_r) \cdot \left(1 - P(E_j|C_r)\right) - \sum_{\substack{i,j=1 \\ i \neq j}}^{n_E} M_{ki} \cdot M_{lj} \cdot \sum_{r=1}^{n_E} \hat{n}(C_r) \cdot P(E_i|C_r) \cdot P(E_j|C_r)$$

Uncertainties

- Comparison of the pull distributions for the old and the new uncertainty calculation for a migration matrix with large migration (M_1)

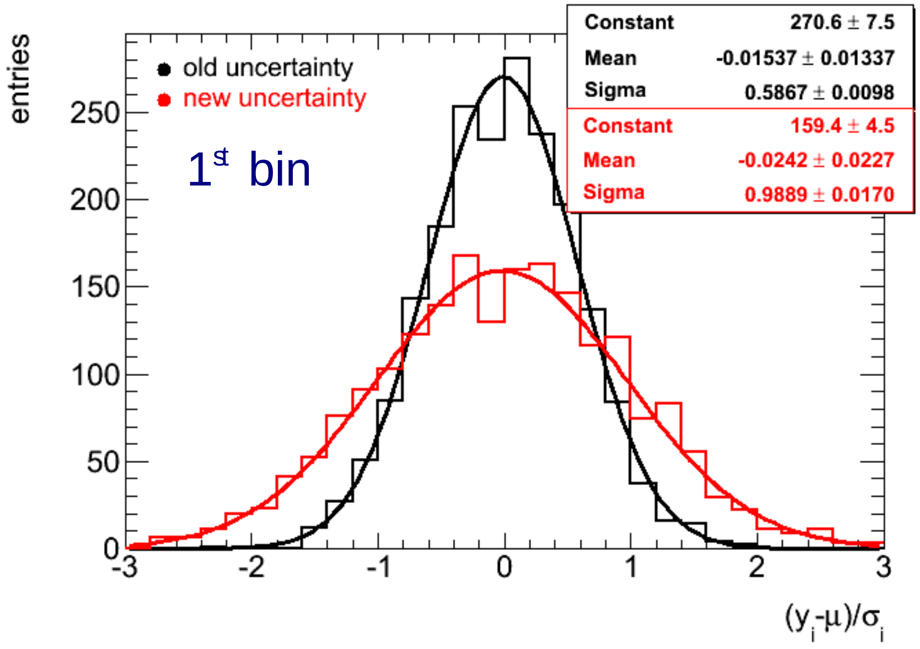


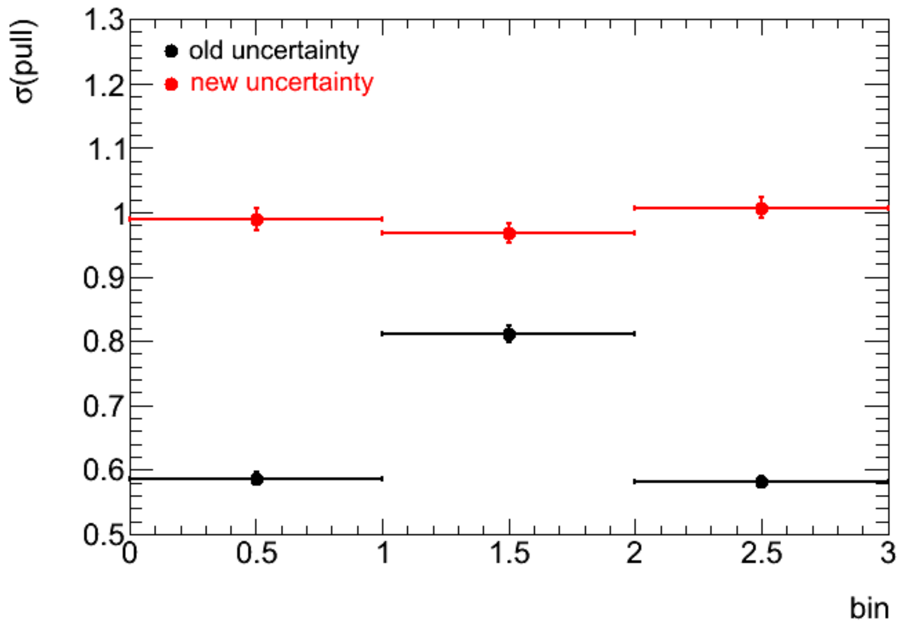


➔ As expected for large migration only a small improvement of the uncertainty calculation is visible

Uncertainties

- Comparison of the pull distributions for the old and the new uncertainty calculation for a migration matrix with low migration (M_2)

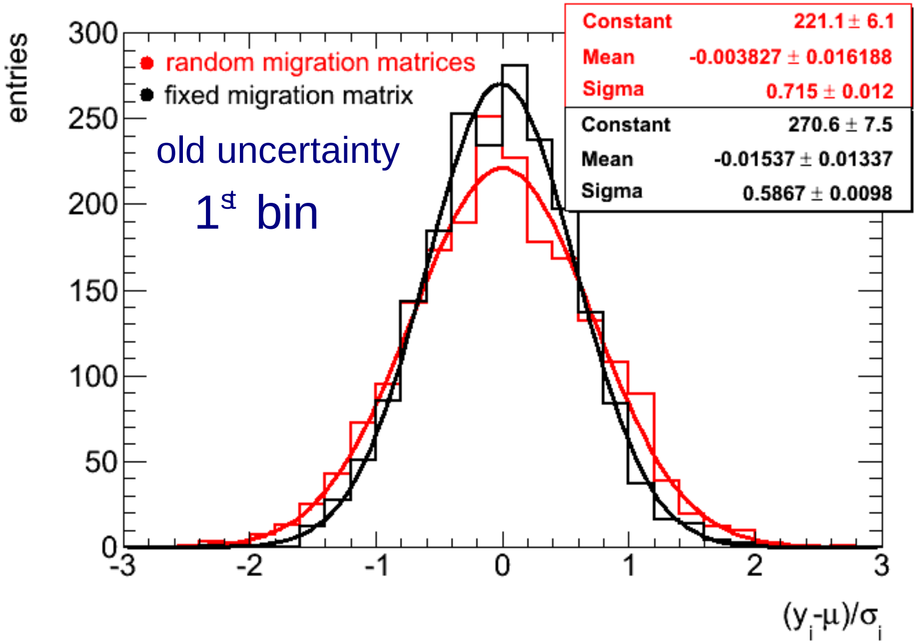


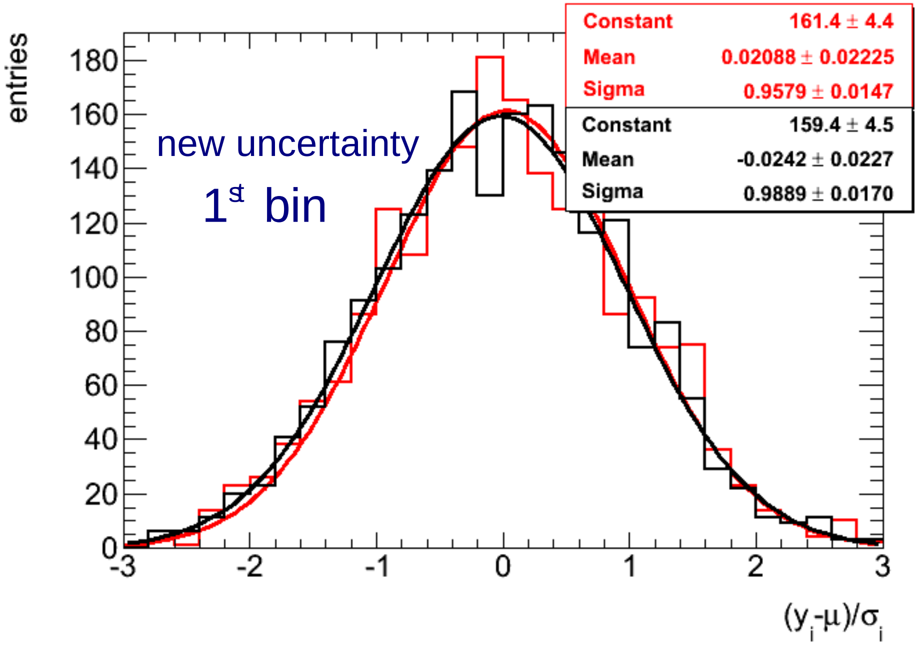


➔ For low migration in the migration matrix a clear improvement of the uncertainty calculation is visible

Uncertainties

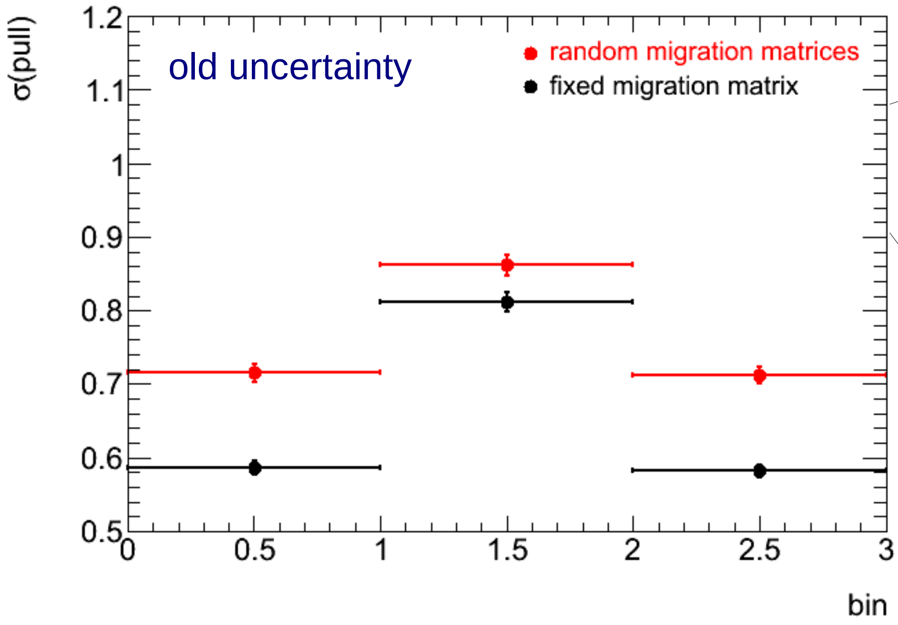
- Comparison between uncertainties from ensemble tests and the program with and w/o fluctuations in the migration matrix (M_2) for the new error calculation

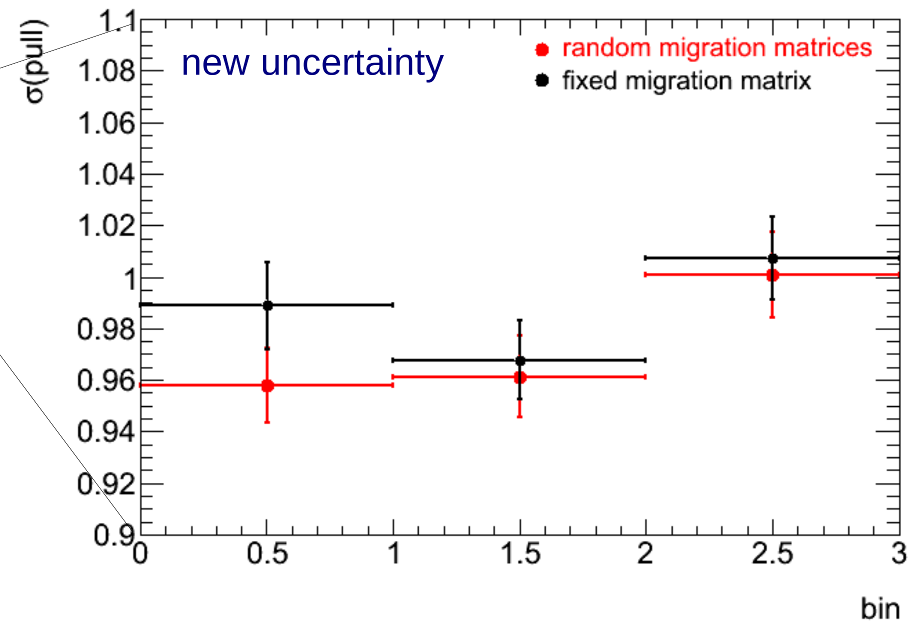




Uncertainties

- Comparison between uncertainties from ensemble tests and the program with and w/o fluctuations in the migration matrix (M_2) for the new error calculation



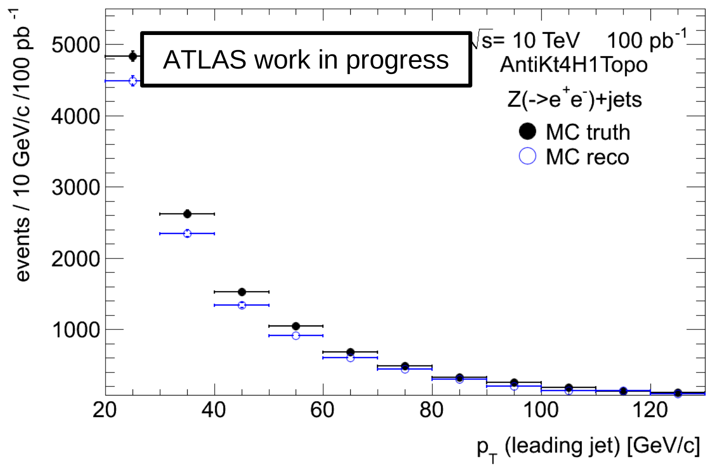


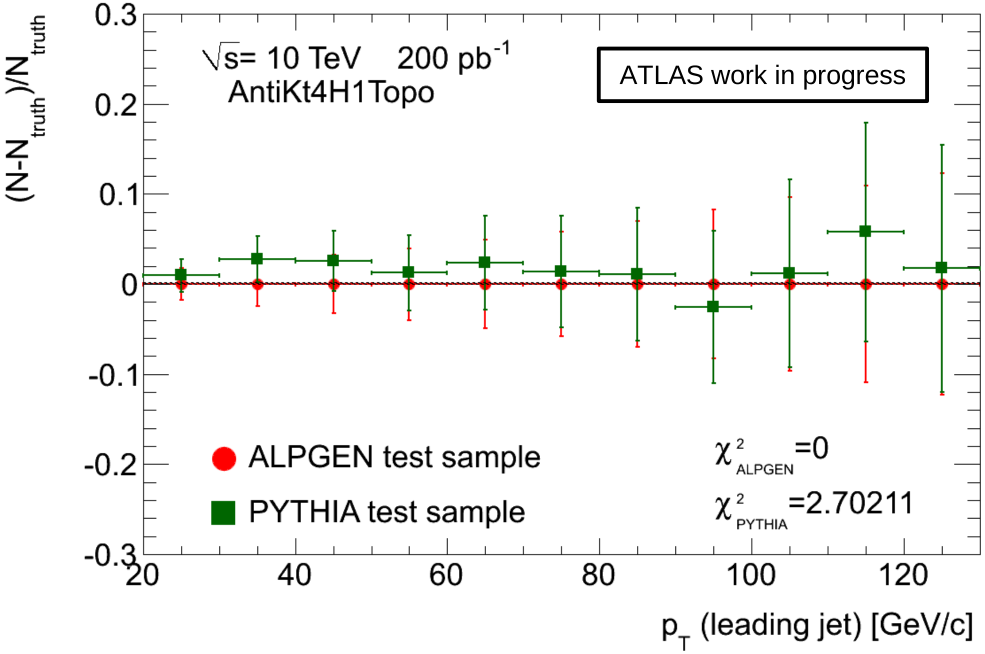
- The new uncertainty calculation shows a clear improvement
- Seems that also the problem with fluctuations in the matrix is solved

Bin-by-Bin Method

Bin-by-Bin Method:

- Assumes that migration between the bins is negligable
- ➔ Migration matrix is diagonal
- Only needs the reconstructed and the truth distribution as input
- No correction for fake jets is needed

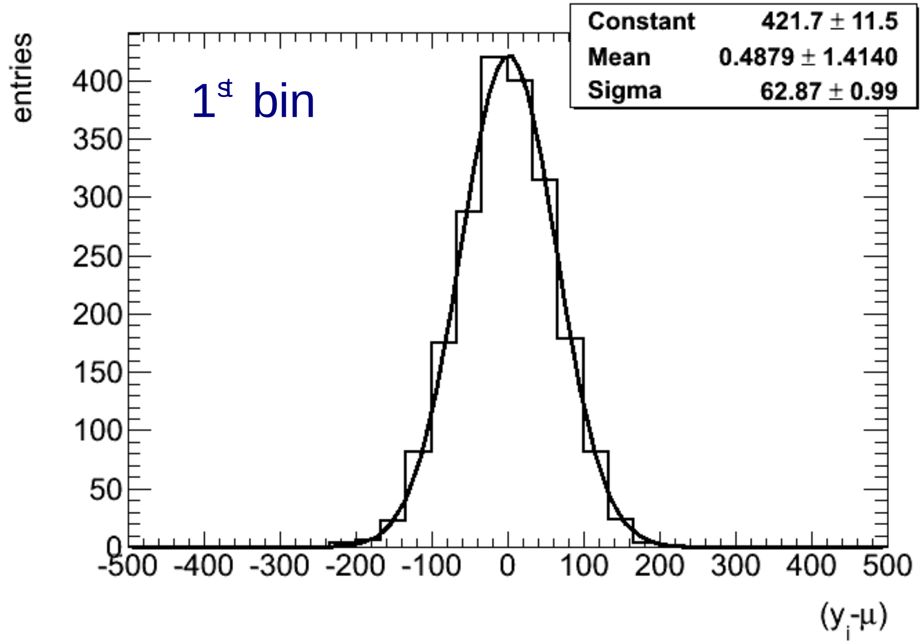


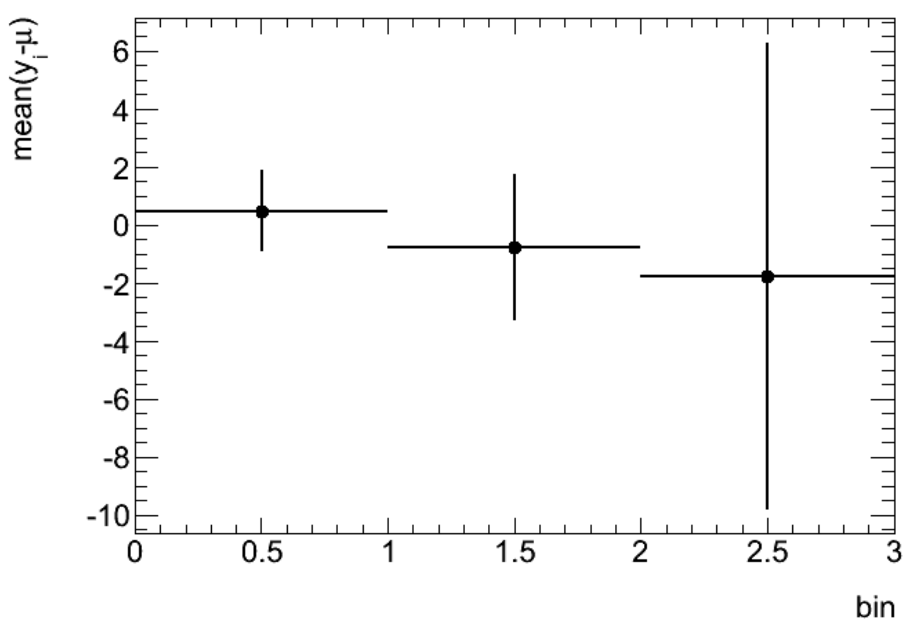


- ➔ Uncertainties seems to be too large
- ➔ Check method using a simple Toy Mc

Method checks

- Create a truth distribution with 3 bins with 10000 entries each
- Create randomly 2000 training distributions from fixed probabilties
- Calculate for each bin a correction factor
- Create randomly 2000 test distributions and calculate the unfolded distribution

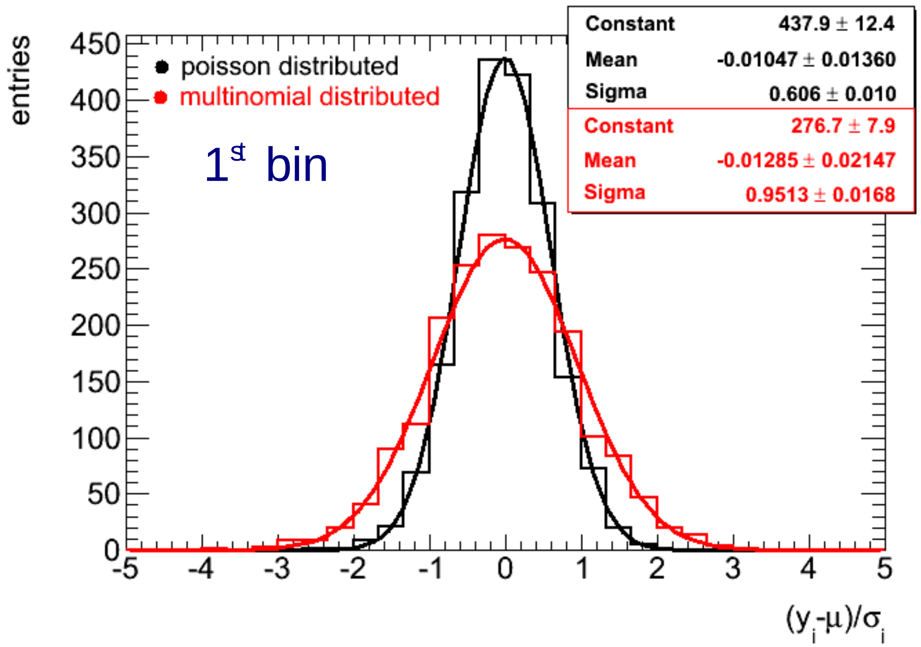


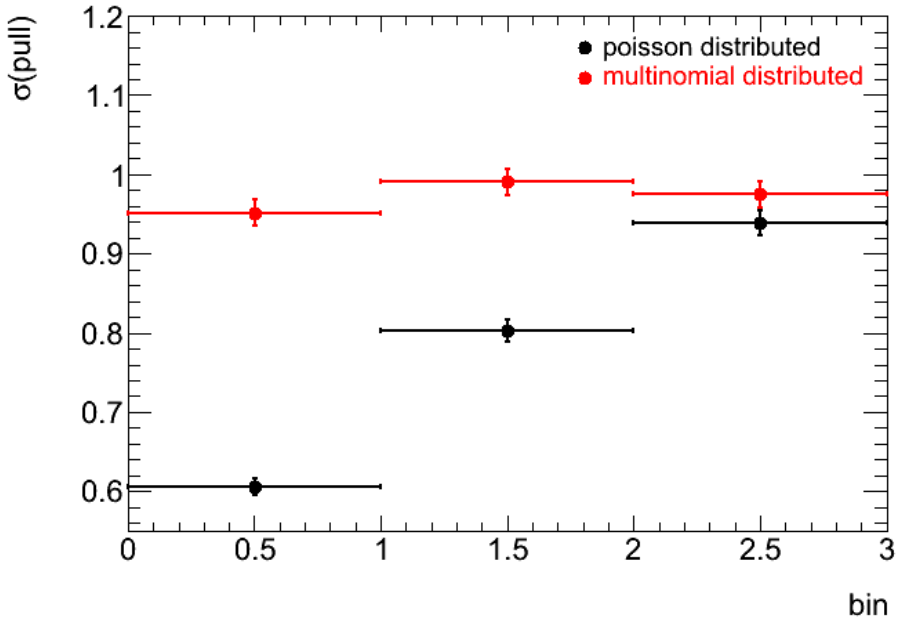


- No bias visible

Uncertainties

- Uncertainties for the Bin-by-Bin method can be calculated assuming a multinomial distribution or poisson distribution for the data
- Create pull distribution → comparison between uncertainties from ensemble tests and the program with infinite statistics in the training

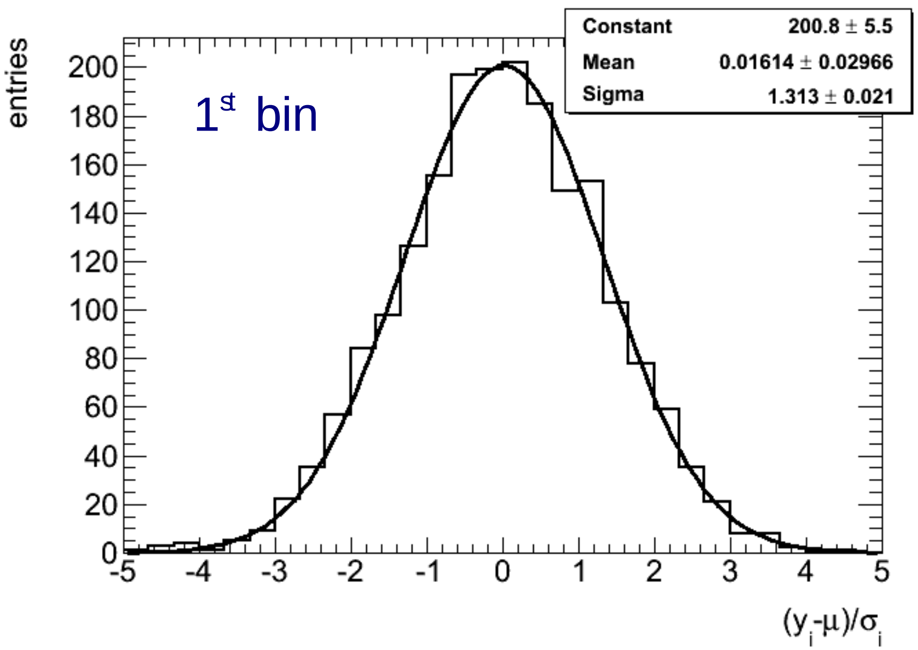


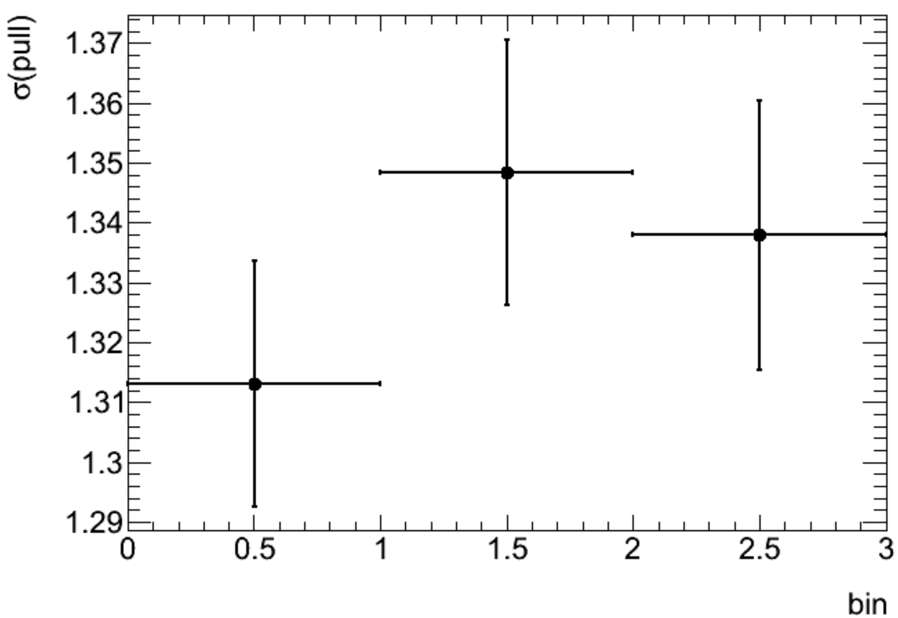


➔ The multinomial distribution gives a better and a stable estimation of the uncertainties due to fluctuations in data

Uncertainties

- Create pull distribution → comparison between uncertainties from ensemble tests and the program with finite statistics in the training assuming a multinomial distribution for the data





➔ Uncertainties given from the program are too small (σ(pull)>1.0) with finite statistics in the training

➔ Have to introduce an additional uncertainty on the correction factor

Other Methods

SVD

- SVD describes a change of Basis with a diagonal response matrix A

$$A=USV^T$$

U and V are orthogonal and S is a diagonal matrix with non-negative diagonal elements

$$S_{ij}=0 \text{ for } i\neq j,\ S_{ii}\equiv s_i\geq 0$$

s_i are called singular values of A

- Some singular values are significantly smaller than others

➔ The system is difficult to solve

➔ The small singular values are set to zero to solve the system

- **Problem:** Due to the cut on the singular values the unfolded distributions becomes periodic, not a good method if it is known that the function is smooth

Tikhonov regularisation

$$Y = A \cdot X$$

- The migration matrix A itself is not invertable
- But $(\alpha I + A^T A)$ is invertable

$$X = (\alpha I + A^T A)^{-1} A^T Y$$

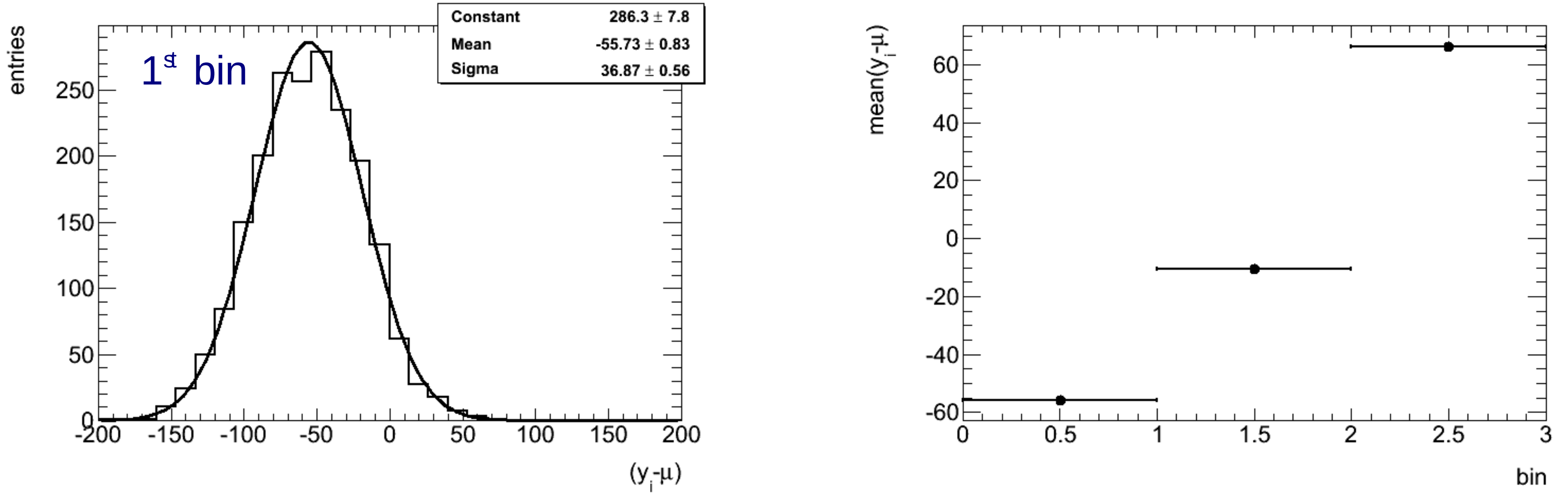
- For $\alpha \to 0$ the system converges to the initial system
- → Tikhonov regularisation is a reweighting of the singular values
- → Smoother result
- α can be estimated using cross validation

Conclusion & Outlook

- Iterative (Bayes) Method:
  - Performance of this method is checked
  - New uncertainty calculation shows a clear improvement
  - Code exists in C++, but is not yet user friendly enough
- Bin-by-Bin Method:
  - Performace of this method is checked
  - Code exists also in C++
- Next steps:
  - Include efficiency loss
  - Ensemble tests for Tikhonov regularisation and SVD
  - Look at physics distributions
  - Make the code public and document it
  - Compare the different methods

Method checks 2

- Create randomly a migration matrix from fixed probabilties (M_2)
- Create a truth distribution with 3 bins with 10000 entries each
- Create randomly 2000 test distributions and calculate the unfolded distribution



- ➔ Create randomly a migration matrix results in fluctuations in the matrix for finite statistics
- ➔ Introduce a bias
- ➔ Fluctuations have to be taken into account
- ➔ Have to create randomly different migration matrices