Bayesian Unfolding

K. Bierwagen, U. Blumenschein, A. Quadt

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

January 20, 2011

PHYSICS AT THE TERA SCALE
Helmholtz Alliance

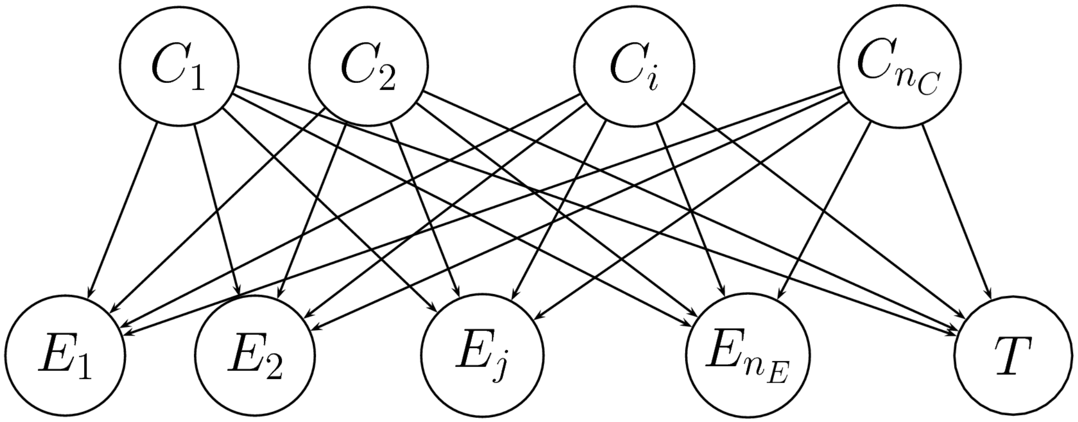
FSP 101 ATLAS

Bundesministerium für Bildung und Forschung

Introduction

Causes

Effects



G. D'Agostini

- Each cause can produce different effects
- ➔ For a given effect the exact cause is not known
- But their probabilities can be calculated assuming some knowledge about migration, efficiency and resolution
- Probabilities $P(E_j|C_i)$ are estimated from MC
- Goal: determine the probability $P(C_i|E_j)$
- Simple inversion not possible
- Bayes theorem yields solution

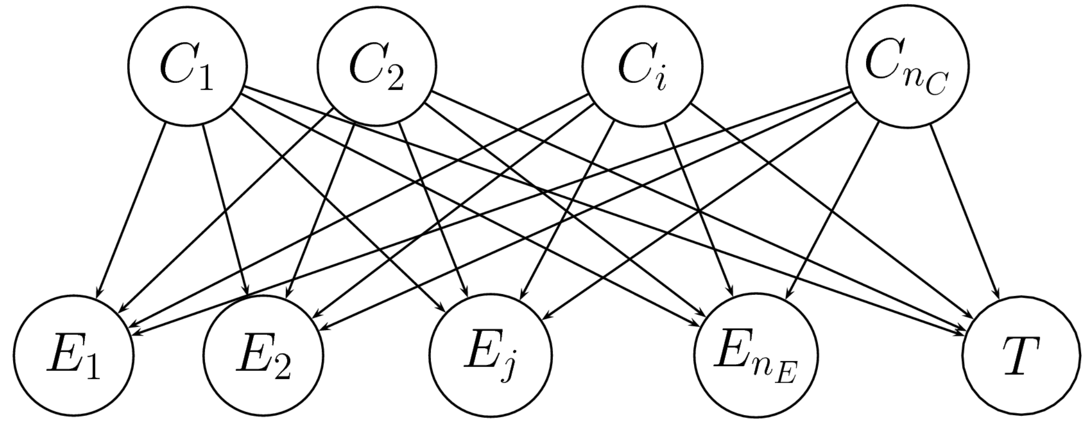
Reminder: Bayes theorem

$$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)}$$

Iterative (Bayes) Method

$n(C_i)$: #evts in cause bin i

$n(E_j)$: #evts in effect bin j



G. D'Agostini

$$P(C_i|E_j) = \frac{P(E_j|C_i) \cdot P_\circ(C_i)}{\sum_{l=1}^{n_C} P(E_j|C_l) \cdot P_\circ(C_l)}$$

$$\hat{n}(C_i) = \frac{1}{\epsilon_i} \sum_{j=1}^{n_E} n(E_j) \cdot P(C_i|E_j) \qquad \epsilon_i \neq 0$$

- Described by G. D'Agostini (NIM A362 (1995) 487)

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

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

  - $P(E_j|C_i)$:

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

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

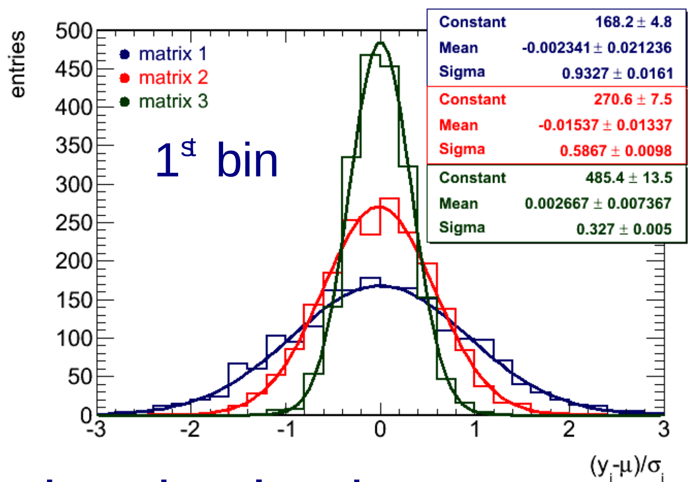
Performance checks

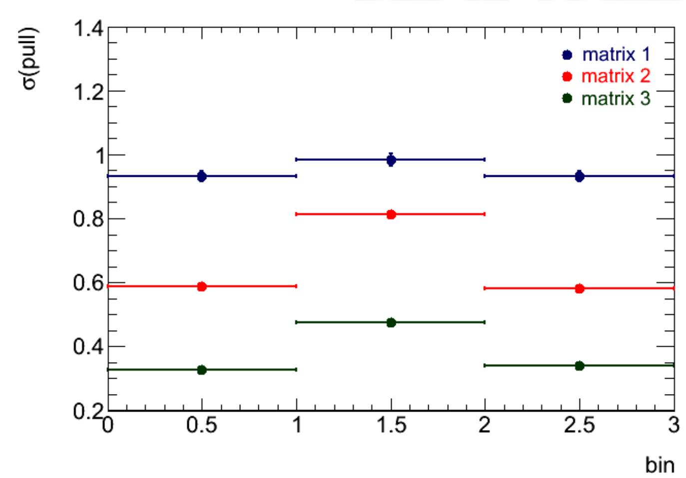
- Define a migration matrix

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

large migration — medium migration — low migration

- Use ensemble tests for performance checks
- Mean values are well described, but uncertainties are too large
- 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

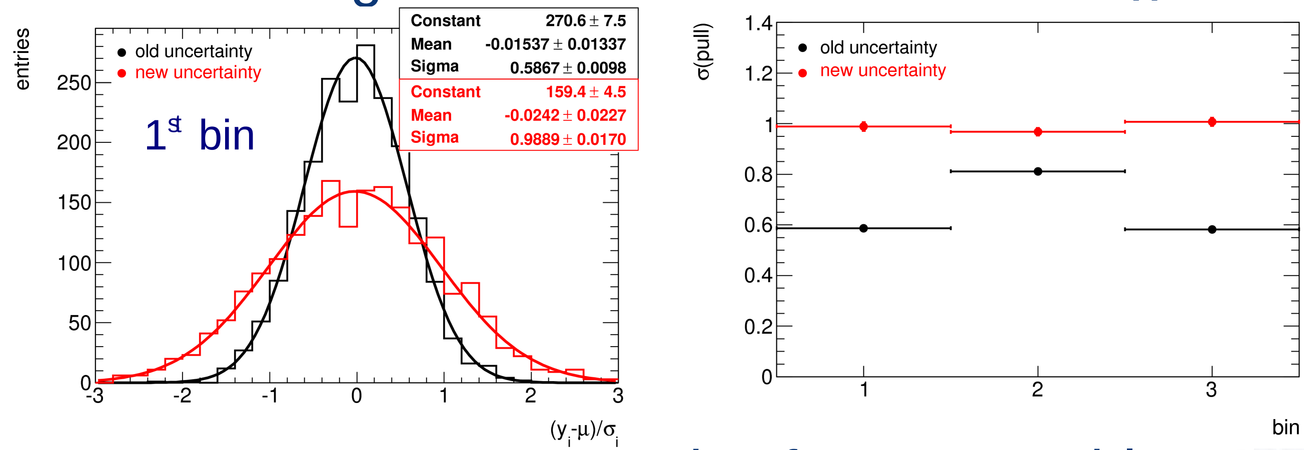
➔ Not the typical case in data analysis

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

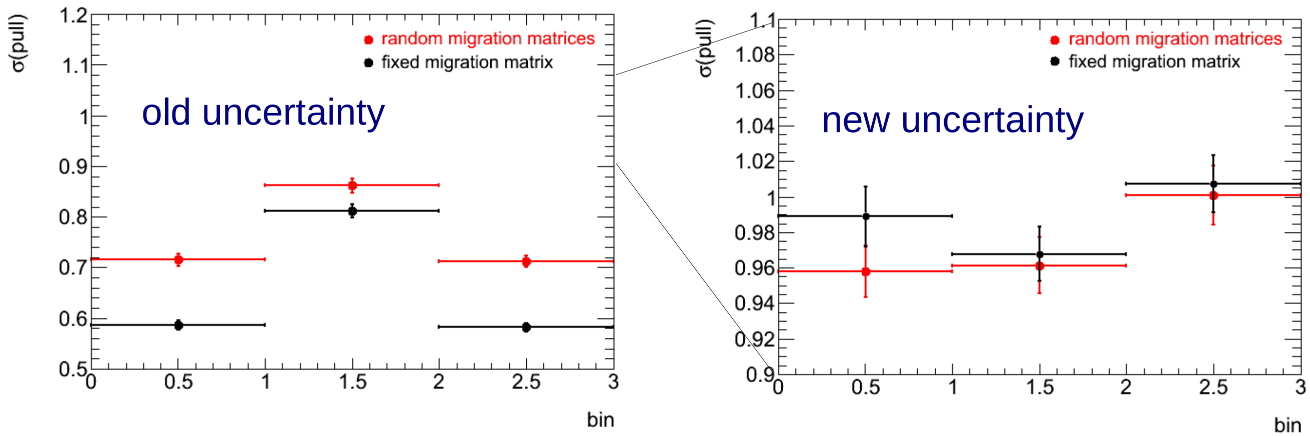
$$V_{kl}(\underline{n}(E)) = \sum_{j=1}^{n_E} M_{kj} \cdot M_{lj} \cdot \sum_{r=1}^{n_E} \hat{n}(C_r) \cdot P(E_j|C_r) \cdot \left(1 - P(E_j|C_r)\right)$$
$$- \sum_{\substack{i,j=1 \\ i \neq j}}^{n_E} M_{ki} \cdot M_{lj} \cdot \sum_{r=1}^{n_E} \hat{n}(C_r) \cdot P(E_i|C_r) \cdot P(E_j|C_r)$$

Uncertainties

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



- 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

Improved iterative (Bayes) Method

- Described by G. D'Agostini (arxiv:1010.0632 (2010))
- Based on the previous method
- But uncertainties are treated differently:
  - Quantities are described by probability density functions
  - Uncertainty propagation is done by sampling
- Results will be compared soon with my improvements for the old method

Conclusion & Outlook

- Iterative (Bayes) Method:
  - Performance of this method is checked
  - New uncertainty calculation shows a clear improvement
- Improved iterative (Bayes) Method:
  - Code exist in R implemented by G. D'Agostini and in C++ implemented by J. Therhaag
- Outlook:
  - Compare improved iterative (Bayes) method with the results of my improvements on the old method