Bayesian Unfolding

Katharina Bierwagen, Ulla Blumenschein, Arnulf Quadt
2nd Institute of Physics, Georg-August-University Göttingen

Abstract
I give a short introduction to Bayesian Unfolding and describe my work on the improvement of the uncertainty calculation for this method.

1 Introduction

Bayesian Unfolding has been used since 1994 and was introduced by G. D'Agostini (see Ref. [1]). For our application area the method shows problems in the evaluation of the uncertainties. Therefore, performance studies using ensemble testing are shown for the Bayesian Unfolding Method. An improved uncertainty calculation has been developed to providing a better error calculation.

2 Bayes Method

The procedure of Bayesian Unfolding can be explained using a picture of causes C and effects E, in which causes correspond to the true values before smearing and effects to the values after smearing. Each cause can produce different effects, but for a given effect, as is the case in a measurement, the exact cause is not known. However, the probability for an effect produced from a defined cause $P(E_j|C_i)$ can be estimated assuming some knowledge about the migration, efficiency and resolution. This is usually achieved by using Monte Carlo. Now the goal is to estimate the probability $P(C_i|E_j)$ that different causes C_i were responsible for the observed effect E_j. A simple inversion cannot be used to solve this problem, but Bayes theorem yields a solution.

$$P(C_i|E_j) = \frac{P(E_j|C_i) \cdot P_0(C_i)}{\sum_{l=1}^{n_C} P(E_j|C_l) \cdot P_0(C_l)}, \tag{1}$$

$$\hat{n}(C_i) = \frac{1}{\epsilon_i} \sum_{j=1}^{n_E} n(E_j) \cdot P(C_i|E_j) \qquad \epsilon_i \neq 0, \tag{2}$$

with $\hat{n}(C_i)$ the expected number of events in the cause bin i, $n(E_j)$ the number of events in the effect bin j, $P_0(C_i)$ the initial probabilities and ϵ_i the efficiency that the cause i has an effect. This formula can be rewritten in terms of the unfolding matrix M

$$\hat{n}(C_i) = \sum_{j=1}^{n_E} M_{ij} \cdot n(E_j), \tag{3}$$

$$M_{ij} = \frac{P(E_j|C_i) \cdot P_0(C_i)}{\left[\sum_{l=1}^{n_E} P(E_l|C_i)\right] \cdot \left[\sum_{l=1}^{n_C} P(E_j|C_l) \cdot P_0(C_l)\right]}, \tag{4}$$

which is clearly not equal to the inverse of the migration matrix. For the calculation of the covariance matrix of $\hat{n}(C_i)$ two different sources are taken into account, an uncertainty on the distribution of the effects $n(E_j)$

$$\begin{aligned} V_{kl}(\underline{n}(E)) = & \sum_{j=1}^{n_E} M_{kj} \cdot M_{lj} \cdot n(E_j) \cdot \left(1 - \frac{n(E_j)}{\hat{N}_{true}}\right) \\ & - \sum_{\substack{i,j=1 \\ i \neq j}}^{n_E} M_{ki} \cdot M_{lj} \cdot \frac{n(E_i) \cdot n(E_j)}{\hat{N}_{true}}, \end{aligned} \tag{5}$$

with the true number of events $\hat{N}_{true}$ and an uncertainty on the migration probabilities $P(E_j|C_i)$

$$V_{kl}(M) = \sum_{i,j=1}^{n_E} n(E_i) \cdot n(E_j) \cdot Cov(M_{ki} M_{lj}). \tag{6}$$

The total uncertainty is calculated from the sum of both covariance matrices

$$V_{kl} = V_{kl}(\underline{n}(E)) + V_{kl}(M). \tag{7}$$

In the following this method with its uncertainties is checked using a toy Monte Carlo.

2.1 The Toy Monte Carlo

This section describes a simple toy Monte Carlo, which is used in the following to check the method.

Three different migration matrices with different migration (large migration (S_1), medium migration (S_2) and low migration (S_3)) for 3 bins without efficiency losses are defined

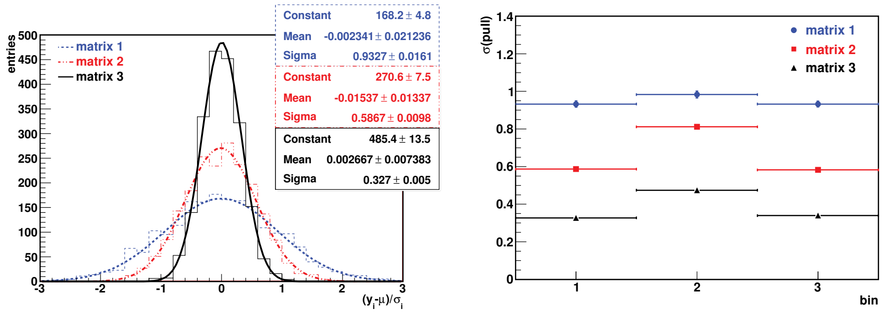
$$S_1 = \begin{pmatrix} 0 & 0.1 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.8 & 0.6 & 0.4 \end{pmatrix}, \; S_2 = \begin{pmatrix} 0 & 0.1 & 0.8 \\ 0.2 & 0.8 & 0.2 \\ 0.8 & 0.1 & 0 \end{pmatrix}, \; S_3 = \begin{pmatrix} 0 & 0.025 & 0.95 \\ 0.05 & 0.95 & 0.05 \\ 0.95 & 0.025 & 0 \end{pmatrix}.$$

For the unfolding, two different sources of uncertainties are present: a finite amount of data and a finite number of events for the creation of the migration matrix. Therefore, three different cases were considered: create randomly 2000 test distributions to simulate the finite amount of data events, create randomly 2000 migration matrices from fixed probabilities on top of the test distributions to simulate statistical fluctuations in the migration matrix due to a finite statistics in Monte Carlo and finally create randomly 2000 uniformly distributed true distributions in addition.

2.2 Performance checks

For the performance checks of the method, a C++ implementation of Ref. [1] is used.

In order to quantify if the absolute values and the uncertainties are correct ensemble testing is used. The width of the Gaussian distribution from ensemble testing is compared to the uncertainties calculated by the program using pull distributions: pull $= (y_i - \mu)/\sigma_i$ with the unfolded values y_i, the truth value μ and the calculated uncertainties σ_i. If the uncertainties are correctly estimated, the width of the pull distributions is expected to be compatible with 1. These tests show that the mean values are well described as expected, but the uncertainties are too large.



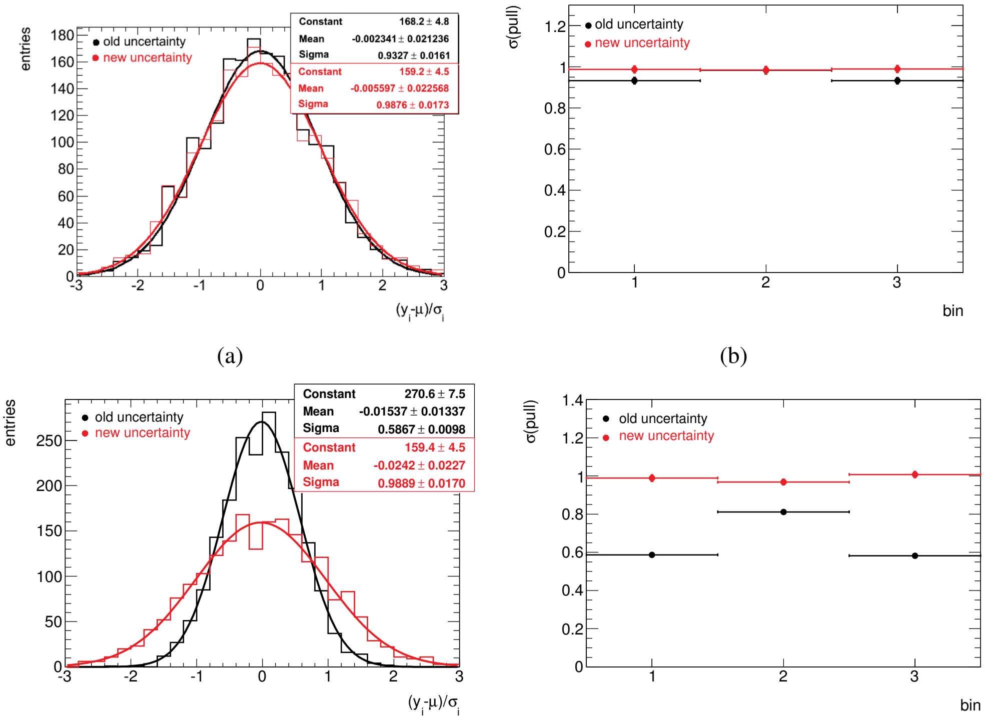
(a) (b)

Fig. 1: Comparison between the deviations from ensemble tests and the uncertainties calculated by the program for 3 different migration matrices.

Fig. 1 shows the pull distributions for the three different migration matrices. With decreasing migration in the migration matrix, the pull distributions become broader. If the uncertainty calculation is correct and for a correct treatment of the migration effect, the pull distributions are expected to be standard Gaussians (mean zero and unit variance). As shown, less migration leads to an overestimate of the uncertainties, so the migration effect is not treated correctly in the uncertainty calculation. This problem seems to come from the fact that the program assumes a multinomial[1] distribution for the data, but each bin is multinomially distributed and the sum of multinomial distributions is only a multinomial distribution if all distributions are equal. In order to fulfil this requirement, the columns of the migration matrix have to be equal to get the correct estimate for the uncertainty from the program, which is not the typical case in data analysis.

Due to the fact, that the data sample is a sum of multinomial distributions, the formula for the calculation of the covariance matrix V_{kl} for the data sample $\underline{n}(E)$ is changed from Eqn. 5 to

$$\begin{aligned} V_{kl}(\underline{n}(E)) = \sum_{j=1}^{n_E} M_{kj}\cdot M_{lj}\cdot \sum_{r=1}^{n_C} \hat{n}(C_r)\cdot P(E_j|C_r)\cdot(1-P(E_j|C_r)) \\ - \sum_{\substack{i,j=1\\ i\neq j}}^{n_E} M_{ki}\cdot M_{lj}\cdot \sum_{r=1}^{n_C} \hat{n}(C_r)P(E_i|C_r)\cdot P(E_j|C_r). \end{aligned} \qquad (8)$$



(a)

(b)

(c)

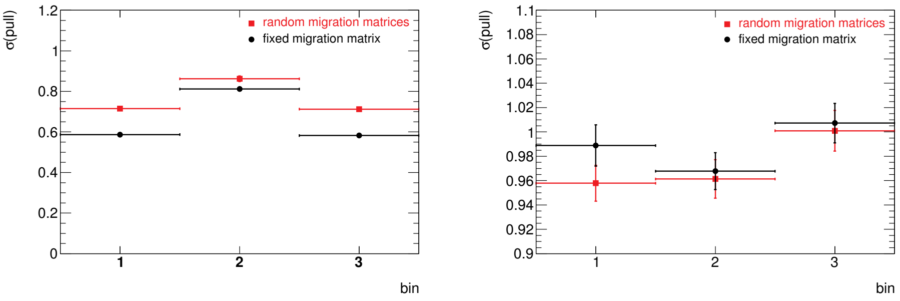
(d)

Fig. 2: Comparison of the pull distributions (a,c) and the width of the pull distributions (b,d) for the old and the new uncertainty calculation for large and medium migration.

Fig. 2 shows the comparison of the pull distributions between the old and the new uncertainty calculation for the migration matrix with large migration and with medium migration for the first bin and the width of the pull distributions for each of the three bins. As expected, for large migration,

[1] A multinomial distribution is a generalization of the binomial distribution with more than two possible outcomes.

only a small improvement due to the new uncertainty calculation is visible, whereas for the migration matrix with medium migration a clear improvement is visible. For both cases with the new uncertainty calculation the width of the pull distribution is now compatible with unity. Furthermore, the differences between the pull distributions for a randomly generated migration matrices and a fixed migration matrix vanishes using the new uncertainty calculation (Fig 3).



(a)

(b)

Fig. 3: Comparison of the width of the pull distributions for the old uncertainty calculation (a) and the new uncertainty calculation (b).

The new uncertainty calculation shows a clear improvement and solves all previously mentioned problems. Meanwhile, there is an improved Bayesian Unfolding method available, which was also described by G. D'Agostini in Ref. [2] and which is based on the previous method but the uncertainties are treated differently. In this method the quantities are described by probability density functions and the error propagation is done by sampling. The next step is to compare this method with my improvements for the old method. This will be pursued in the near future.

3 Summary

This article describes performance studies using ensemble testing for Bayesian Unfolding. This study is motivated by the fact that the uncertainties for this method seem to be too large compared to the fluctuations. In order to solve this problem, an improved uncertainty calculation is presented which shows a good performance.

References

[1] G. D'Agostini. A multidimensional unfolding method based on Bayes' theorem. *Nuclear Instruments and Methods in Physics Research A*, 362:487–498, February 1995.

[2] G. D'Agostini. Improved iterative Bayesian unfolding. *ArXiv e-prints*, October 2010.