



A new maximum-likelihood method for template fits

Hans Dembinski^{1,a}, Ahmed Abdelmotteleb^{2,b}

¹ Department of Physics, TU Dortmund, Otto-Hahn-Straße 4a, 44227 Dortmund, Germany

² Department of Physics, University of Warwick, Coventry CV4 7AL, UK

Received: 27 July 2022 / Accepted: 7 November 2022
© The Author(s) 2022

Abstract A common statistical problem in particle physics is to extract the number of samples which originate from a statistical process in an ensemble containing a mix of several contributing processes. The probability density function of each process is usually not exactly known. Barlow and Beeston found an exact likelihood for the problem of fitting binned templates obtained from Monte-Carlo simulation to binned data, which propagates the uncertainty of the templates into the result. Solving the exact likelihood is technically challenging, however. The original paper also did not provide a way to use weighted simulation samples with varying weights. Other papers have introduced alternative likelihoods to address these points. In this paper, a new approximate likelihood is derived from the exact Barlow–Beeston one. The new likelihood is generalized to fits of weighted templates to weighted data. The performance of the new likelihood is evaluated based on toy examples. The performance is excellent – point estimates have small bias and confidence intervals have good coverage – and is comparable to the exact Barlow–Beeston likelihood when the templates are not weighted. The new likelihood evaluates faster than the Barlow–Beeston one when the number of bins is large.

1 Introduction

Barlow and Beeston [1] were the first to describe the exact likelihood for the problem of fitting a composite model consisting of binned templates obtained from Monte-Carlo simulation to binned data. The fit estimates the yield (number of samples) which originates from each component. In this scenario, the component probability density functions (*p.d.f.s*) are not available in parameterized form, but are obtained implicitly from simulation. The shapes of the templates are

not exactly known since the simulated sample is finite. Therefore, the value of each template in each bin is a nuisance parameter, constrained through Poisson statistics by the simulated sample.

The exact likelihood has many nuisance parameters; one for each component in each bin. Since the number of bins can be large (especially when distributions are multidimensional), Barlow and Beeston provide an algorithm to estimate the nuisance parameters implicitly by solving a non-linear equation per bin for a given set of yields. Since nuisance parameters are implicitly found, only the yields remain as external parameters, which are found numerically in the usual way; for example, with the *Migrad* algorithm in the *Minuit* library [2].

This approach is elegant, but also has drawbacks. It was observed by Conway [3] that the finite accuracy of the non-linear solver may introduce discontinuities in the likelihood that confuse algorithms that use gradient-descent, like *Migrad*. This seems to be rarely an issue in practice, but it can lead to fits that fail to converge or fits that produce incorrect uncertainty estimates. Even then, solving a non-linear equation per bin numerically introduces a non-negligible computational cost.

Conway proposed a simplified treatment [3] to address these issues. The exact likelihood is replaced by a simplified one where the uncertainty in the template is captured by a multiplicative factor, which in turn is constrained by a Gaussian penalty term. This introduces only one nuisance parameter per bin instead of one per component per bin, and it allows one to estimate each nuisance parameter by solving a simple quadratic equation. The computation of the simplified likelihood does not suffer from numerical instabilities and is faster.

Conway did not derive the simplified likelihood rigorously from the exact likelihood of Barlow and Beeston. The present work grew from such a derivation, motivated by the wish to gain insight into the practical limits of Conway's approxi-

^a e-mail: hans.dembinski@tu-dortmund.de (corresponding author)

^b e-mail: ahmed.abdelmotteleb@warwick.ac.uk

mate likelihood. This exercise lead to the discovery of a new approximate likelihood that has the same computational benefits, while having better statistical properties.

The exact likelihood by Barlow and Beeston was derived under the assumption that the bin contents for the data and templates are Poisson-distributed. This assumption falls short when the bin contents in templates or data are sums of random weights, and therefore, the likelihood does not perform well when applied to such problems. Barlow and Beeston were aware of this limitation, but did not provide a full solution. Argüelles et al. [4] first described an alternative likelihood which is applicable to a weighted simulation. In their Bayesian derivation, the nuisance parameters are integrated out using a prior that is conditional on the simulation weights. The prior is based on results from Bohm and Zech [5], who showed that the distribution for a variable which is the sum of random weights can be approximately described by a modified Poisson distribution. The authors demonstrate that the marginalized likelihood works well in a frequentist approach, where point estimates for the yields are obtained by maximizing the likelihood, and limits are obtained by constructing the profile likelihood¹ and applying Wilk’s theorem [6]. The intervals have good coverage [4].

In this paper, we derive a new approximate likelihood from the Barlow–Beeston one. The likelihood is transformed so that the minimum value is asymptotically chi-square distributed, following the approach of Baker and Cousins [7]. This allows one to use the minimum value as a goodness-of-fit test statistic. Furthermore, the new approximate likelihood is generalized to handle weighted templates and weighted data, based on results from Bohm and Zech [5]. In the last part of the paper, the performance of the new likelihood is compared to alternative likelihoods in fits to two toy examples which idealize real problems in high-energy and astroparticle physics.

2 Derivation of the likelihood

We will derive the new likelihood in this section starting with a general remark by Baker and Cousins [7] who note that likelihoods for binned data can be transformed such that the minimum value doubles as an asymptotically chi-square-distributed test statistic, Q . The following monotonic transformation is applied to the likelihood \mathcal{L} , without loss of generality given here for a single bin,

$$Q(\vec{p}) = -2 \ln \left[\frac{\mathcal{L}(n; \mu(\vec{p}))}{\mathcal{L}(n; n)} \right], \tag{1}$$

¹ A profile likelihood for parameters of interest θ is obtained by maximizing with respect to all nuisance parameters ϕ , $\mathcal{L}(\theta) = \arg \max_{\phi} \mathcal{L}(\theta, \phi)$.

where n is the count of samples in the bin and μ is the model expectation, which is a function of model parameters, \vec{p} . In the constant denominator, μ is replaced by n . It is sufficient to analyse the likelihood for a single bin, since each bin is an independent sample so that $\mathcal{L}_{\text{tot}} = \prod_i \mathcal{L}(n_i; \mu_i(\vec{p}))$ and $Q_{\text{tot}} = \sum_i Q(n_i; \mu_i(\vec{p}))$.

If the model is correct, Q follows a chi-square distribution with n_{dof} degrees of freedom in the asymptotic limit of infinite sample size, where n_{dof} is the difference between the number of bins and the number of fitted parameters. In the following, we will refer to Q as the “likelihood”, although it is in fact a monotonic function of a likelihood.

In the likelihoods that are derived here, the nuisance parameters from the templates add a balance of zero to n_{dof} , since each bin with a simulated count has one corresponding nuisance parameter. Therefore, n_{dof} of the total likelihood is still given by the difference of the number of bins and the number of yields.

2.1 Barlow–Beeston likelihood

Barlow and Beeston first described an exact likelihood for Poisson-distributed binned data and templates [1]. To make this paper self-contained, we briefly summarize their derivation. The log-likelihood for a bin that contains a Poisson distributed count n with expectation μ is

$$\ln \mathcal{L} = n \ln \mu - \mu - \ln n!. \tag{2}$$

The last term is constant with respect to changes in μ and can be omitted. In a template fit, μ is the sum over templates that are normalized and scaled with the respective yield y_k of that component,

$$\mu = \sum_k \frac{y_k \xi_k}{M_k}, \tag{3}$$

where ξ_k is the expected contribution of component k to this bin, and M_k is the sum over all ξ_k from different bins. The key insight here is that ξ_k cannot be identified with the count a_k obtained from a particular simulation run. Instead, a_k is a random realization from a Poisson distribution with expectation ξ_k . Just like μ is constrained by n , ξ_k is constrained by a_k . We can write down another log-likelihood for each template k ,

$$\ln \mathcal{L}_k = a_k \ln \xi_k - \xi_k - \ln a_k!, \tag{4}$$

where the last term can be omitted again. The (pseudo)random process that generates a_k is independent of the real process that generates n . Therefore, the likelihoods can be added to form the total likelihood for the inference problem,

$$\ln \mathcal{L} + \sum_k \ln \mathcal{L}_k = n \ln \mu - \mu + \sum_k a_k \ln \xi_k - \xi_k, \tag{5}$$

now without constant terms. This is the exact likelihood for this statistical problem, if n and a_k are Poisson distributed. Estimates for the y_k are obtained by maximizing Eq. 5 over y_k and ξ_k . The ξ_k are nuisance parameters. They are not of interest, but the likelihood has to be maximized with respect to the ξ_k , too. The number of nuisance parameters can be large; evaluating to $K \times N$ for K components and N bins. Barlow and Beeston show how these can be found numerically by solving only N decoupled non-linear equations; for details we refer the reader to the original paper [1].

The transformation in Eq. 1 applied to the exact likelihood gives

$$Q = 2(\mu - n - n(\ln \mu - \ln n)) + 2 \sum_k \xi_k - a_k - a_k(\ln \xi_k - \ln a_k). \tag{6}$$

This can be written more succinctly using the Cash statistic $C(k; \lambda) = 2(\lambda - k - k(\ln \lambda - \ln k))$ [8];

$$Q(\vec{y}, \vec{\xi}) = C(n; \mu(\vec{y}, \vec{\xi})) + \sum_k C(a_k; \xi_k). \tag{7}$$

2.2 Conway’s approximation

Conway proposed an approximate likelihood [3] as an alternative to Eq. 5. The likelihood was not rigorously derived in the original publication, therefore we will attempt to do that here.

Without loss of generality, the template amplitudes in Eq. 7 can be parameterized as $\xi_k = a_k \beta_k$, since the extremum of the likelihood is also invariant to monotonic transformations of the parameters. The central approximation is to set $\beta \approx \beta_k$; the component factors are replaced by a single factor,

$$\mu = \sum_k \frac{y_k \xi_k}{M_k} = \sum_k \frac{y_k \beta_k a_k}{M_k} \approx \beta \underbrace{\sum_k \frac{y_k a_k}{M_k}}_{\mu_0}. \tag{8}$$

This approximation is valid in the limit where μ is dominated by a single component. Applied to Eq. 7, one gets

$$Q \approx C(n; \beta \mu_0) + \sum_k C(a_k; \beta a_k). \tag{9}$$

The second term can be simplified,

$$\begin{aligned} & \sum_k C(a_k; \beta a_k) \\ &= 2 \sum_k \beta a_k - a_k - a_k(\ln(\beta a_k) - \ln a_k) \\ &= 2(\beta a - a - a \ln \beta) = 2a(\beta - 1 - \ln \beta), \end{aligned}$$

with $a = \sum_k a_k$. One obtains the intermediate result

$$Q = C(n; \beta \mu_0) + 2a(\beta - 1 - \ln \beta). \tag{10}$$

To obtain Conway’s likelihood, the logarithm $\ln \beta$ in the second term is approximated further by a Taylor series around $\beta = 1$ to second order,

$$\begin{aligned} & 2a(\beta - 1 - \ln \beta) \\ & \approx 2a \left(\beta - 1 - (\beta - 1) + \frac{1}{2}(\beta - 1)^2 \right) \\ & = a(\beta - 1)^2. \end{aligned}$$

This approximation is valid in the asymptotic limit $a \rightarrow \infty$ which implies that β can deviate only slightly from 1. In practice, the simulation sample is often smaller than the data sample, such that a is often smaller than n . The approximation is not expected to work well in these cases.

In Conway’s original formula [3], $(\beta - 1)^2$ is divided by the estimated variance V_β of β , not multiplied by a . This largely overcomes the limitation introduced by setting $\beta_k \approx \beta$. Conway’s idea is to treat the second term like a Gaussian penalty for the nuisance parameter β , which means we need to divide by the expected variance V_β of β .

We demonstrate that $V_\beta \approx 1/a$ when a single component is dominant, while in general, V_β is larger. Recall the definition $\beta = \mu/\mu_0$, where μ_0 is considered constant. One finds via error propagation,

$$V_\beta = \frac{V_\mu}{\mu_0^2} \quad \text{with} \quad V_\mu = \sum_k \frac{y_k^2 V_{\xi,k}}{M_k^2}. \tag{11}$$

With the plug-in estimate $V_{\xi,k} = \xi_k \approx a_k$, one obtains

$$V_\beta = \frac{\sum_k \frac{y_k^2}{M_k^2} a_k}{\left(\sum_k \frac{y_k}{M_k} a_k \right)^2}. \tag{12}$$

In the limit where only one of the components is dominant (the same limit in which the central approximation $\beta_k \approx \beta$ is valid), one gets

$$V_\beta \approx \frac{\frac{y_k^2}{M_k^2} a}{\left(\frac{y_k}{M_k} a \right)^2} = \frac{1}{a}. \tag{13}$$

When more than one component is dominant, $V_\beta > 1/a$.

We thus obtain Conway’s likelihood [3],

$$Q(\vec{y}, \beta) = C(n; \beta \mu_0(\vec{y})) + \frac{(\beta - 1)^2}{V_\beta}. \tag{14}$$

A value for the nuisance parameter β is obtained by solving the score function $\partial Q/\partial \beta = 0$ which leads to a quadratic

equation for β that has only one valid solution ($\beta > 0$),

$$\beta = \frac{1 - V_\beta \mu_0}{2} + \sqrt{n V_\beta + \left(\frac{1 - V_\beta \mu_0}{2}\right)^2}. \quad (15)$$

In summary, Conway's likelihood is only expected to perform well in the limit of large simulation samples.

2.3 New approximation

Starting again from Eq. 10, we note that the likelihood treats data and simulation symmetrically,

$$\begin{aligned} Q &= C(n; \beta \mu_0) + 2a((\beta - 1) - \ln \beta) \\ &= C(n; \beta \mu_0) + C(a; \beta a). \end{aligned} \quad (16)$$

One can solve the score function $\partial Q / \partial \beta = 0$ directly, the solution is

$$\beta = \frac{n + a}{\mu_0 + a}. \quad (17)$$

This formula for β is even simpler than the one derived by Conway, and can be easily interpreted. The nuisance parameter β balances the pulls of the observed value n in data and the observed value in the simulation a . If $a \ll n$, $\beta \mu_0 \rightarrow n$, irrespective of the actual value of μ_0 . In other words, the bin provides no information for the component yields y_k , which are constrained only through the potential tension between μ_0 and n . For $a \rightarrow \infty$, $\beta \rightarrow 1$ and $Q \rightarrow C(n; \mu_0)$; we recover the basic Poisson-likelihood.

To summarize, Eq. 16 is derived from the exact likelihood without using a truncated Taylor series, in contrast to Conway's likelihood. Real and simulated counts are described with Poisson statistics, same as the exact Barlow–Beeston likelihood. Maximum-likelihood estimation based on Poisson statistics generally performs better than a likelihood based on a Gaussian approximation for Poisson-distributed data [1,4]. The new likelihood is therefore expected to perform better in fits where the simulated sample is small.

Bins with $a = 0$ should be excluded from the calculation, since they do not contribute anything to the yields. Alternatively, one can replace $a = 0$ in numerical calculations with a tiny number like 10^{-100} , which has the same effect.

Equation 16 is still limited by the original approximation $\beta \approx \beta_k$. It does not perform well when more than one component dominates in a bin. A mixture of components has a larger variance than βa , which is the variance of the Poisson distribution associated to the term $C(a; \beta a)$. One can obtain a better approximation by replacing C with a Poisson-like distribution that allows for a larger variance, analogue to the final step in the derivation of Conway's likelihood. This is discussed next, after the likelihood is generalized for weighted samples.

2.4 Weighted samples

In practice, the data and the simulation samples may be weighted. Simulations are often weighted to reduce discrepancies between the simulated and the real experiment, or as a form of importance sampling. Data may be weighted to correct losses from finite detection efficiency. In both cases, a count n is replaced by a sum of weights, $n = \sum_i w_i$. Barlow and Beeston discussed the impact of weighted simulated samples in their paper [1] and showed how the exact likelihood can be adapted when the distribution of weights is extremely narrow, but do not provide a solution for the general case.

Bohm and Zech [5] derived the scaled Poisson distribution (SPD), an approximate probability distribution for a sum of independent weights. The SPD becomes exact in the limit that all weights are equal, and is a good approximation to the correct distribution otherwise. The basic idea is to use a Poisson distribution that is scaled such that its variance is equal to $V_n = \sum_i w_i^2$, which differs in general from $n = \sum_i w_i$ [5]. This can be achieved by multiplying n and its model prediction, μ , with the scale factor $t = n / V_n$.

Argüelles, Schneider, and Yuan were the first to use the SPD in the context of template fitting [4]. In contrast to the frequentist approach in this paper, they derived a Bayesian marginalized likelihood, in which the nuisance parameter μ is removed by integrating over a prior $p(\mu)$,

$$\mathcal{L}_G = \int_0^\infty \frac{\mu^n e^{-\mu}}{n!} p(\mu) d\mu. \quad (18)$$

The prior $p(\mu)$ in this equation is conditional on the simulation outcome. It is the posterior obtained by applying Bayes' theorem on the simulated sample,

$$p(\mu; \mu_0, V_\mu) = \frac{1}{N_p} \mathcal{L}(\mu_0; \mu, V_\mu) q(\mu), \quad (19)$$

where $\mathcal{L}(\mu_0; \mu, V_\mu)$ is the likelihood to observe μ_0 , given that μ and V_μ are the true values, μ_0 is the expected bin content given components yield \vec{y} and the simulation weights \vec{w} , V_μ is the estimated variance of μ around μ_0 , N_p is a normalization constant, and $q(\mu)$ is a subjective prior for μ . The authors take $q(\mu)$ to be uniform in their main result, but also consider other priors, see Ref. [4] for details. The variables μ_0 and V_μ are calculated as

$$\begin{aligned} \mu_0 &= \sum_{k,i} \frac{y_k w_{k,i}}{M_k} = \sum_k \frac{y_k m_k}{M_k} \\ V_\mu &= \sum_{k,i} \left(\frac{y_k w_{k,i}}{M_k} \right)^2 = \sum_k \frac{y_k^2 V_{m,k}}{M_k^2}, \end{aligned} \quad (20)$$

where m_k and $V_{m,k}$ are the sum of weights and the sum of weights squared for component k , respectively, and M_k is the sum of m_k over all bins. The authors identify $\mathcal{L}(\mu_0; \mu, V_\mu)$

with the SPD, which takes the following form,

$$\mathcal{L}(\mu_0; \mu, V_\mu) = \frac{e^{-s\mu}(s\mu)^{s\mu_0}}{\Gamma(s\mu_0 + 1)}, \tag{21}$$

where $\Gamma(x)$ is the gamma function and $s = \mu_0/V_\mu$. Inserting Eq. 21 into Eq. 18 leads to an integral that can be solved analytically [4]. The final marginalized likelihood is

$$\mathcal{L}_{\text{ASY}} = \frac{s^{s\mu_0+1}\Gamma(n + s\mu_0 + 1)}{n!(s + 1)^{n+s\mu_0+1}\Gamma(s\mu_0 + 1)}. \tag{22}$$

The authors demonstrate that the marginalized likelihood \mathcal{L}_{ASY} has good statistical properties even when used in a frequentist approach (point estimation via maximization over parameter space, interval computation via likelihood profiling). It cannot be transformed into a chi-square distributed test statistic with the result from Baker and Cousins [7], since the form of Eq. 22 is not compatible.

We take a different approach and generalize the likelihood in Eq. 16 to weighted samples by introducing the generalized Cash statistic for weighted samples,

$$C(n; \mu) \rightarrow C(sn; s\mu), \tag{23}$$

which is obtained by constructing the SPD likelihood and then applying the Baker–Cousins transform. We exploit the symmetry of Eq. 16 and apply the generalization to both terms, which allows us to handle both weighted data and weighted simulation. By doing so, we also lift the limitation that Eq. 16 is only a good approximation for a single dominant component. Since the SPD is fundamentally a generalized Poisson distribution with larger variance, we can now account for the larger variance that a mix of components has over a single component. The final form of the new likelihood is

$$Q = C(tn; \beta t\mu_0) + C(s\mu_0; \beta s\mu_0), \tag{24}$$

with

$$n = \sum_i w'_i, \quad V_n = \sum_i w'^2_i, \quad t = \frac{n}{V_n},$$

$$\mu_0 = \sum_k \frac{y_k m_k}{M_k}, \quad V_\mu = \sum_k \frac{y^2_k V_{m,k}}{M^2_k}, \quad s = \frac{\mu_0}{V_\mu},$$

where w'_i are data weights, while m_k and $V_{m,k}$ are the sum of weights and sum of weights squared of the simulated component k respectively. The nuisance parameter β becomes

$$\beta = \frac{tn + s\mu_0}{t\mu_0 + s\mu_0}. \tag{25}$$

The final form of the new likelihood in Eq. 24 is no longer identical to Eq. 16 even if all weights are unity. In this case, we have

$$s = \sum_k \frac{y_k a_k}{M_k} / \sum_k \frac{y^2_k a_k}{M^2_k}, \tag{26}$$

where a_k is the count in template k , compared to Eq. 11. If only a single component k is dominant, we recover $s\mu_0 \approx a$.

Conway’s likelihood in Eq. 14 is generalized to weighted samples in an analogous way. The variances $V_{\xi,k}$ in Eq. 11 are replaced by $V_{m,k}$, and weighted data is handled in the same way as in Eq. 24 by replacing the Cash statistic. The original Barlow–Beeston likelihood in Eq. 7 can also be generalized with Eq. 23, but we do not attempt this here. The marginalized likelihood in Eq. 22 cannot treat weighted data in its current form, only weighted simulation. We also do not attempt to generalize it further here.

3 Toy study

The properties of the estimated yields, obtained with the likelihoods presented here, are studied in two toy examples. The yields are estimated by minimizing the likelihoods given by Eq. 7 (Barlow–Beeston), Eq. 14 (Conway), Eq. 22 (Argüelles–Schneider–Yuan), and Eq. 24 (this work). We are interested in the bias of the estimated signal yield, the bias of the estimated uncertainty of the signal yield, and the coverage probability of intervals obtained by likelihood profiling. Biases should be small, and the coverage probability should be equal to the expected confidence level.

The minimization is performed with the `Minuit2` library [2] as implemented in `iminuit` [9]. In the case of Eq. 7, the nuisance parameters are found by the Barlow–Beeston algorithm from the reference implementation `TFractionFitter` in the `ROOT` framework [10]. Two-sided limits are generated with the `Minos` algorithm, which computes the profile likelihood and applies Wilk’s theorem [6] to construct an interval.

In both toy scenarios, samples are drawn from two overlapping components. The parameters of these components are listed in Table 1. In case A, a normally distributed signal is mixed with a comparably flat exponentially distributed background. In this example, most bins are dominated by a single component. It idealizes the common problem in high-energy physics where a narrow resonance peaks in the mass distribution over a smooth background. Two normally distributed components are mixed for case B. Most bins are not clearly dominated by one component. Moreover, the lowest and highest bins at the tails of the distribution have a low density. It idealizes a problem in cosmic ray physics, in which the distribution of the depth of shower maximum from many cosmic-ray induced air showers is analysed to extract yields of different elemental groups of primary cosmic rays (see e.g. Ref. [11]).

In both scenarios, the expected signal yield in the data sample is 250 and the expected background yield is 750. The expected size of each component in the simulation sample is N_{mc} ; we run sets of experiments with $N_{\text{mc}} \in$

Table 1 Parameters of the distributions used in the toy simulation. The signal and background distributions are truncated to the interval $x \in [0, 2]$

	Signal	Background
Case A	Normal, $\mu = 1, \sigma = 0.1$	Exponential, $\mu = 2$
Case B	Normal, $\mu = 1.2, \sigma = 0.4$	Normal, $\mu = 0.9, \sigma = 0.4$
Weights	Uniform in the interval $[0, 10]$	

{100, 1000, 10000}. When a new toy experiment is simulated, the sample sizes are randomly drawn from a Poisson distribution. Generated samples are sorted into histograms with equidistant bins over the interval $x \in [0, 2]$; 15 bins in case A and 8 bins in case B. The number of bins in case B is reduced to increase the density in the lowest and highest bin.

An example for each case and each expected size N_{mc} is shown in Fig. 1. In the examples with $N_{\text{mc}} = 10000$, the uncertainty of the template is negligible compared to the data sample. In the examples with $N_{\text{mc}} = 1000$, the uncertainties in the simulated sample are comparable to the data sample, and they dominate the total uncertainty per bin for $N_{\text{mc}} = 100$. The templates also show strong fluctuations then. We expect all likelihoods to perform well for $N_{\text{mc}} = 10000$ and expect the largest differences for $N_{\text{mc}} = 100$.

Template fits are performed based on these inputs. The value $\Delta Q = -2 \ln(\mathcal{L}/\mathcal{L}_{\text{max}})$ as a function of the signal yield is shown in Fig. 2 for the different likelihoods in one particular toy experiment, where \mathcal{L} is the profile likelihood with respect to all other parameters. The new likelihood has a profile close to the Barlow–Beeston one, while the others differ. Fits are also performed with a weighted simulation. In this variation, a count k drawn from the Poisson-distribution as previously described is replaced by a sum over k random weights drawn from a uniform distribution in the interval $[0, 10]$. This sum of weights has a variance that is larger than the original Poisson distribution by a factor of 6 to 7.

Independent samples are generated and fitted 2000 times for each combination. To judge the performance, the pull distribution of the estimated signal yield is computed, where the pull is defined as $z = (\hat{s} - s)/\hat{V}_s^{1/2}$, with true signal yield s , estimate \hat{s} , and estimated variance \hat{V}_s (obtained by the Hesse routine in `Minuit2`) for \hat{s} . The performance is indicated by the degree of agreement of the mean of z with the value 0 and the standard deviation of z with the value 1.

The results are shown in Figs. 3 and 4. As expected, all likelihoods produce compatible results for $N_{\text{mc}} = 10000$, since the uncertainty contributed by the finite size of the simulation sample is negligible. At smaller values of N_{mc} , we see that the Barlow–Beeston likelihood underestimates the uncertainty of the signal yield if the simulated sample is weighted; the width of the pull distribution is larger than one. This was also expected. In the strongly mixed case B, some outliers are produced for $N_{\text{mc}} = 100$, and signal amplitudes are fairly biased. The bias and outliers originate from sam-

ples where the simulation has zero or only one entry in some bins. A bin with zero simulated entries cannot be used by the fit and is discarded. Bins which only contribute to the fit when the simulation templates do not fluctuate downwards introduce a selection effect that causes a bias over an ensemble of toy experiments. This is more likely to happen in case B, and even more so when the simulation is weighted, where the tails of the normal distributions only have a few entries. When bins in the simulation have only one entry, this sometimes causes a fit to converge to a result more than six standard deviations away from the true value, as the fit attempts to reconcile a tiny simulated expectation with a comparably large data value. The Argüelles–Schneider–Yuan likelihood produces the smallest number of outliers among all compared likelihoods. In practice, the bias and outliers can be avoided by restricting the fit to bins which are sufficiently populated in the simulation or by making bins wider. Apart from these special cases, we generally observe a non-negligible bias for the yield estimates obtained by the Argüelles–Schneider–Yuan likelihood, while the bias for the other likelihoods is small or negligible.

To evaluate the coverage of two-sided intervals produced by likelihood profiling, we generate intervals for a fine grid of confidence level values in each toy experiment. The coverage probability is the fraction of intervals that contain the true value for that confidence level. Ideally, coverage probability should be equal to the confidence level. The results are shown in Fig. 5. Coverage is generally excellent except for $N_{\text{mc}} = 100$, where the intervals for case A are too wide, and too narrow for case B. Even then, the new likelihood is competitive, with the coverage consistently being the closest or among the closest to the expected confidence level. Intervals generated from the Barlow–Beeston likelihood are generally too narrow when the simulation is weighted, as expected (Fig. 5).

The runtime of a fit is dominated by the time required to evaluate the likelihood. This work was also motivated by the desire to have a likelihood that can be evaluated without solving non-linear equations numerically, to speed up the calculation. It is therefore interesting to compare likelihood evaluation times. These times are specific to the problem, computing platform, and the implementation (Fig. 6). The Barlow–Beeston likelihood that we use is implemented in C++, while the approximate likelihoods discussed in this work are implemented in Python. Code execution in Python

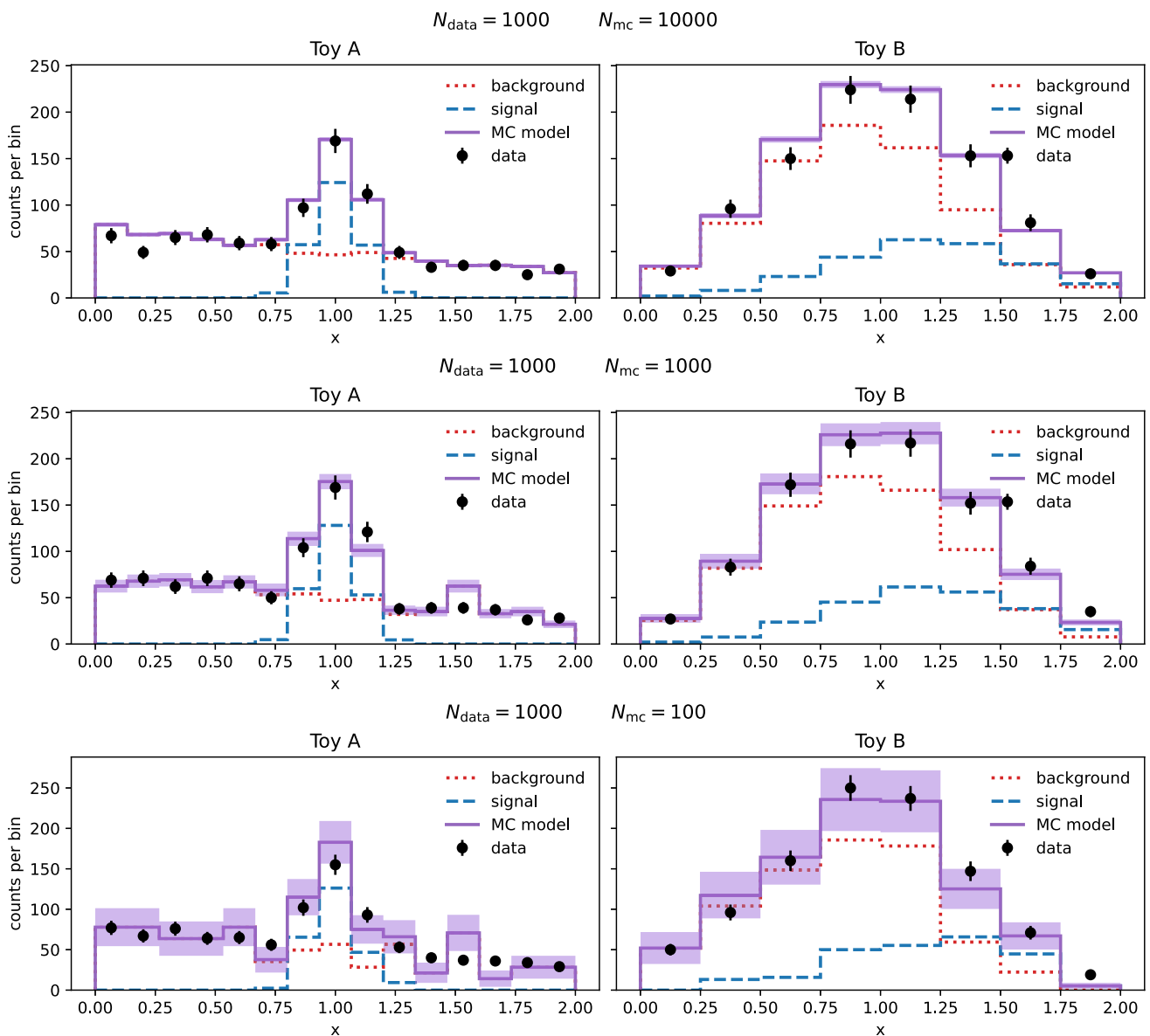


Fig. 1 Examples from the toy simulations. The examples from case A (B) are shown on the left (right). From top to bottom, the expected size of the simulated samples N_{mc} decreases by a factor of 10, starting at

10000. The data sample is shown with data points and error bars, the simulated sample is shown with a solid line and an error band. Dashed (signal) and dotted (background) lines show the two components

is orders of magnitude slower than in C++, but the use of the Numpy library [12] can make the performance of Python implementations competitive with C++. We measure the likelihood evaluation time for the toy experiment A, while varying the number of bins (Fig. 6). The Barlow–Beeston likelihood falls behind the other likelihoods when the number of bins exceeds about 200. This is not a large number, when the working distributions are multidimensional. The evaluation time for the new likelihood is among the best, only Conway’s likelihood is a little faster when the number of bins exceeds 2000.

4 Conclusions

A new approximate likelihood was derived for a template fit from the exact likelihood described by Barlow and Beeston. The new likelihood treats data and simulation symmetrically as Poisson or SPD distributed. It was generalized to describe weighted data and/or weighted templates, and to correctly take into account the increased fluctuations that weighted histograms have when the weights vary in size. This goes beyond the capabilities of the Barlow–Beeston likelihood. Argüelles, Schneider, and Yuan previously derived a marginalized likelihood using a Bayesian approach for weighted simulated

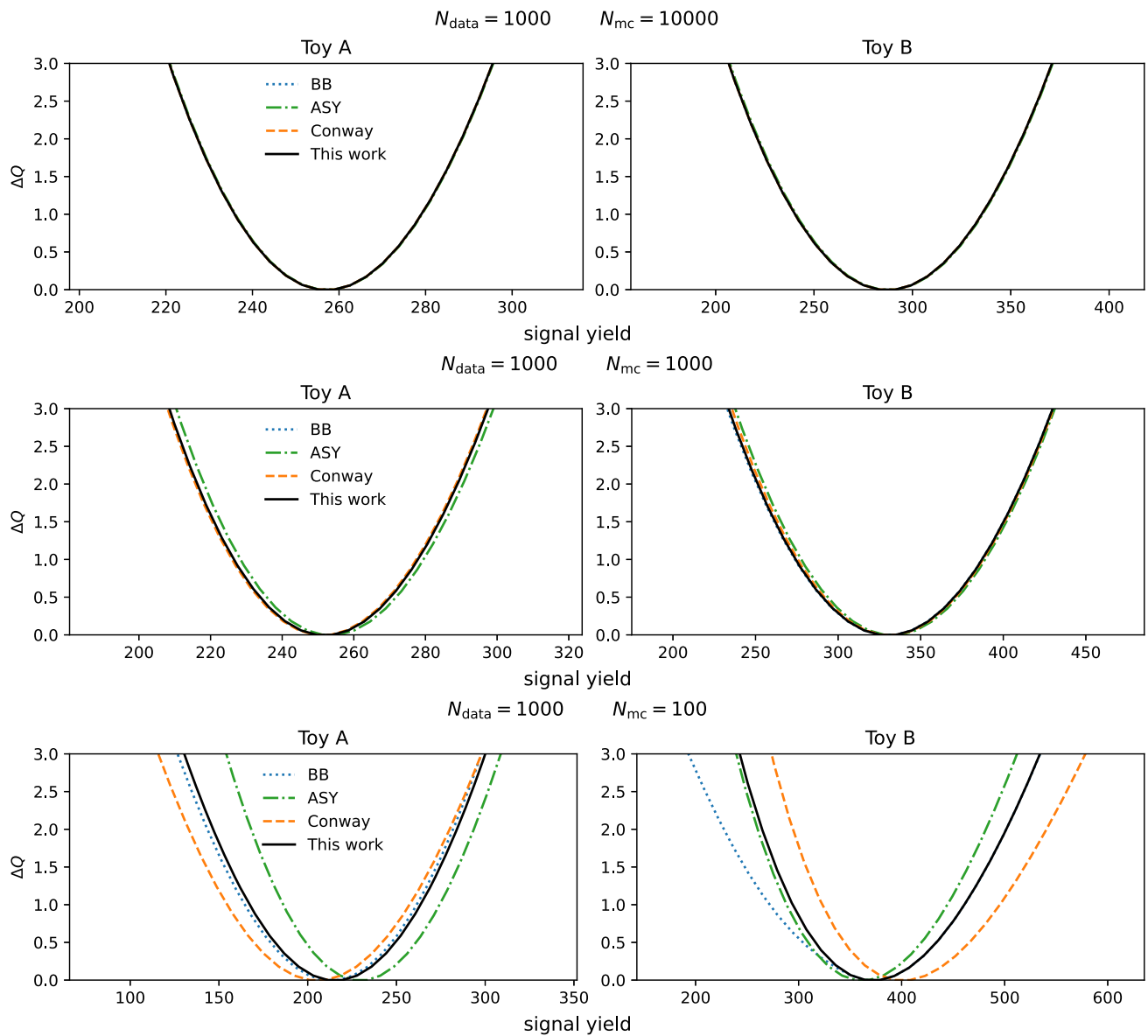


Fig. 2 Profiled ΔQ values (the minimum value of each profile is subtracted) as a function of the signal yield for the likelihoods from Barlow–Beeston (BB), Argüelles–Schneider–Yuan (ASY), Conway, and the one derived in this work. Data and templates samples are those from Fig. 1

samples, while a frequentist approach was used here. The new likelihood can treat both weighted data and weighted simulated samples. The new likelihood was transformed such that the minimum value is asymptotically chi-square distributed following the approach described by Baker and Cousins; enabling the minimum value to serve as a goodness-of-fit test statistic. This cannot be replicated for the marginalized Argüelles–Schneider–Yuan likelihood.

The Barlow–Beeston likelihood, the Argüelles–Schneider–Yuan likelihood, Conway’s likelihood, and the new likelihood were compared in an ensemble study based on two toy examples. The bias of point estimates and uncertainty estimates was studied by analysing pull distributions. The cov-

erage of two-sided intervals obtained from likelihood profiling was measured. Normal and weighted simulation samples were tested. The new likelihood performs well in all tested scenarios. Performance is comparable to the Barlow–Beeston likelihood when the simulation is not weighted. The signal yield in the tested examples is less biased compared to the Argüelles–Schneider–Yuan likelihood. Fits with the Argüelles–Schneider–Yuan likelihood are more stable and produced the fewest outliers. The new likelihood produces some outliers, because it is sensitive to bins with few simulation entries. To avoid outliers, bins should be made wide enough so that the template bins are sufficiently populated. The coverage of intervals extracted from the new likelihood

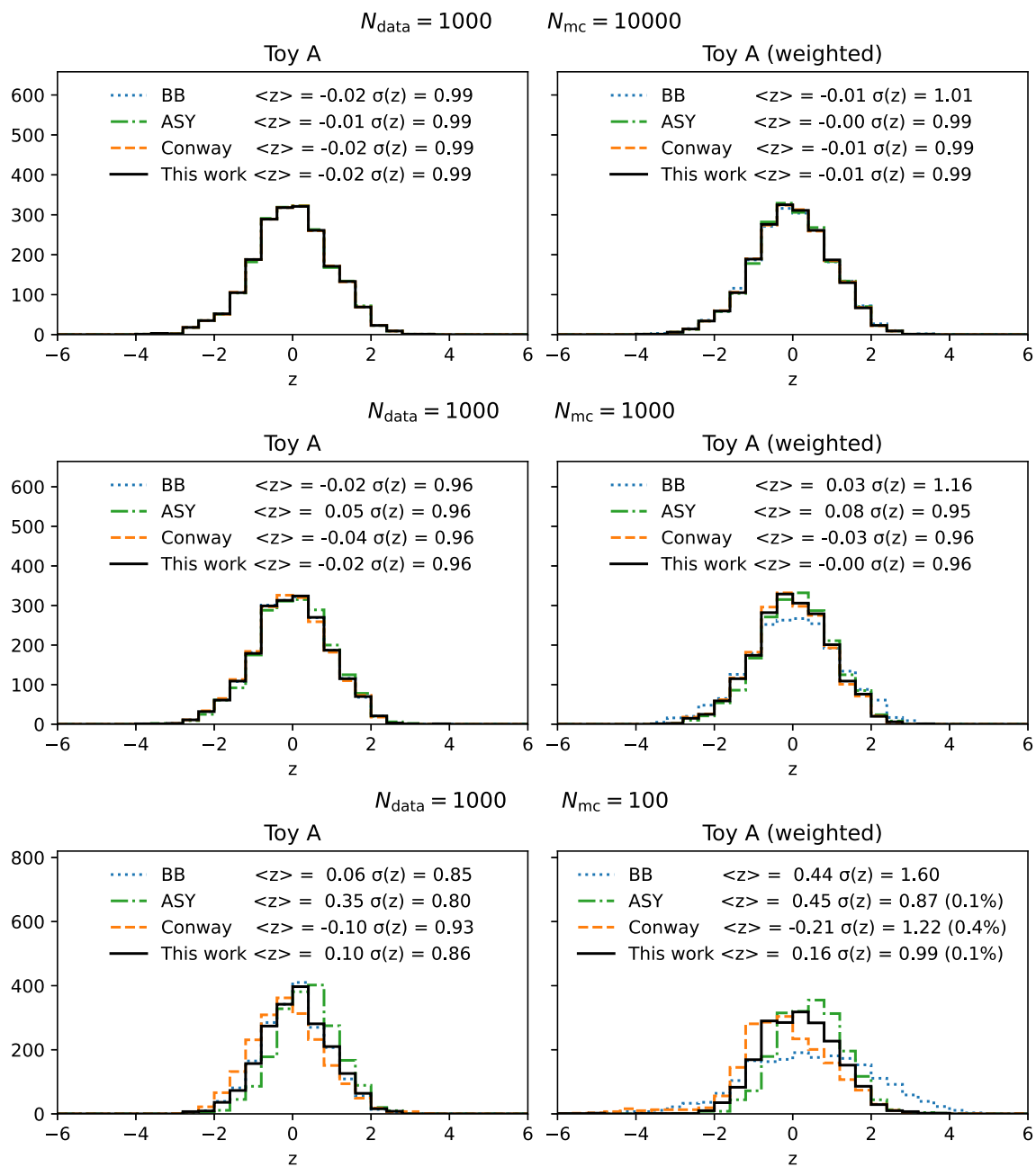


Fig. 3 Pull distributions for toy example A for the estimates obtained from maximizing the likelihoods from Barlow–Beeston (BB), Argüelles–Schneider–Yuan (ASY), Conway, and the one derived in this work. Shown on the left-hand side are fits with Poisson-distributed tem-

plates, shown on the right-hand side are fits with weighted templates as described in the text. Outliers with $|z| > 6$ are not included in the calculation of mean and standard deviation in the legend. The fraction of outliers is included in the legend if it is larger than zero

is accurate or close to the best performers in challenging scenarios.

The new likelihood has a significantly smaller computational cost than the Barlow–Beeston one when the number of bins in the distributions is large. This benefits the runtime of a fit, which is usually dominated by the evaluation time of the likelihood.

In summary, the new likelihood can accommodate weighted data and weighted simulation sample. It shows excellent performance in all tests we made, is fast to evaluate, and has no particular weaknesses, apart from a sensitivity to bins with low simulation samples, which can be overcome. Its theoretical derivation suggests that it should perform well in general.

The new likelihood developed in this paper, Conway’s likelihood (the generalized form that can handle weighted

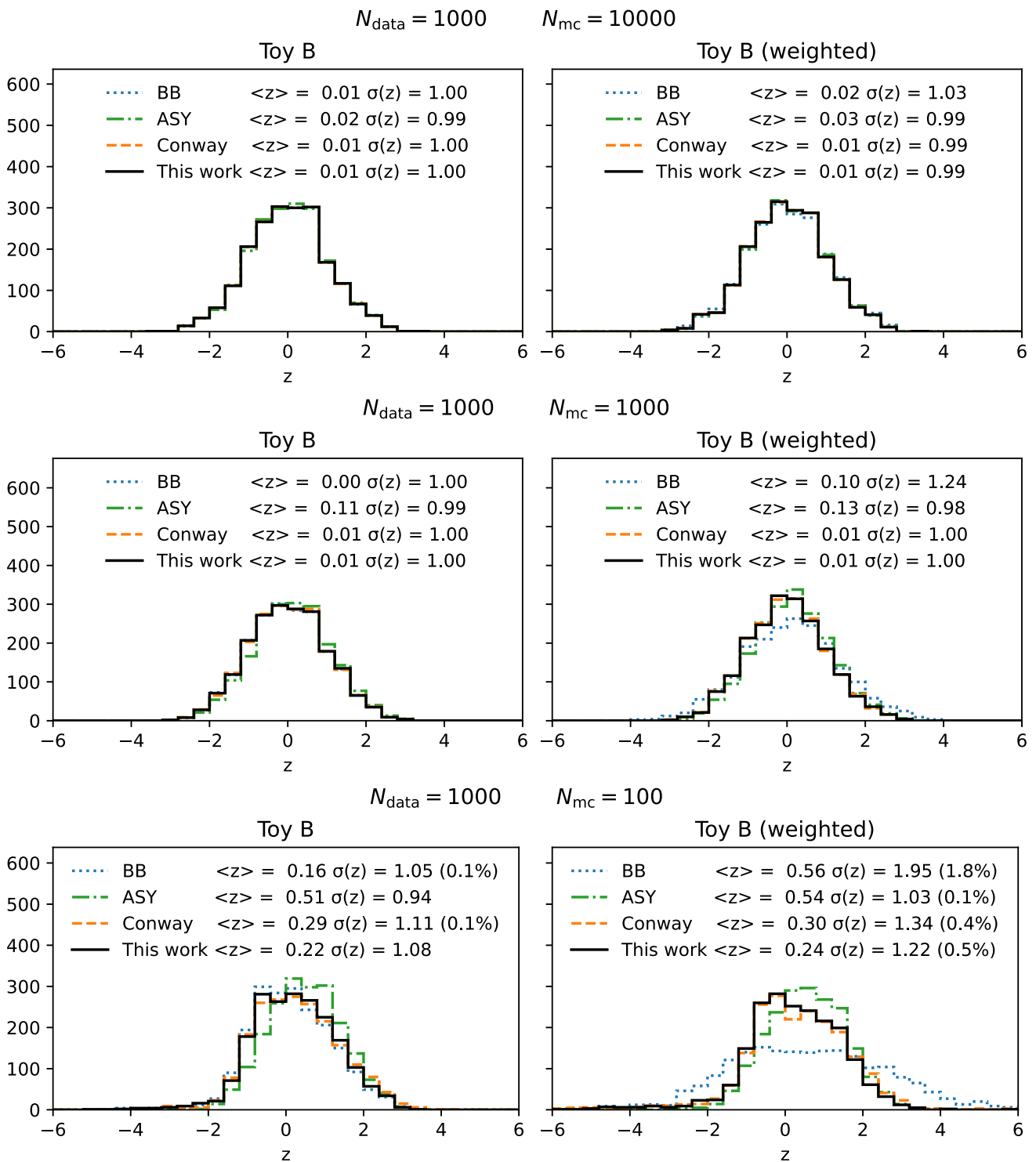


Fig. 4 Plots for toy example B analogous to the ones in Fig. 3

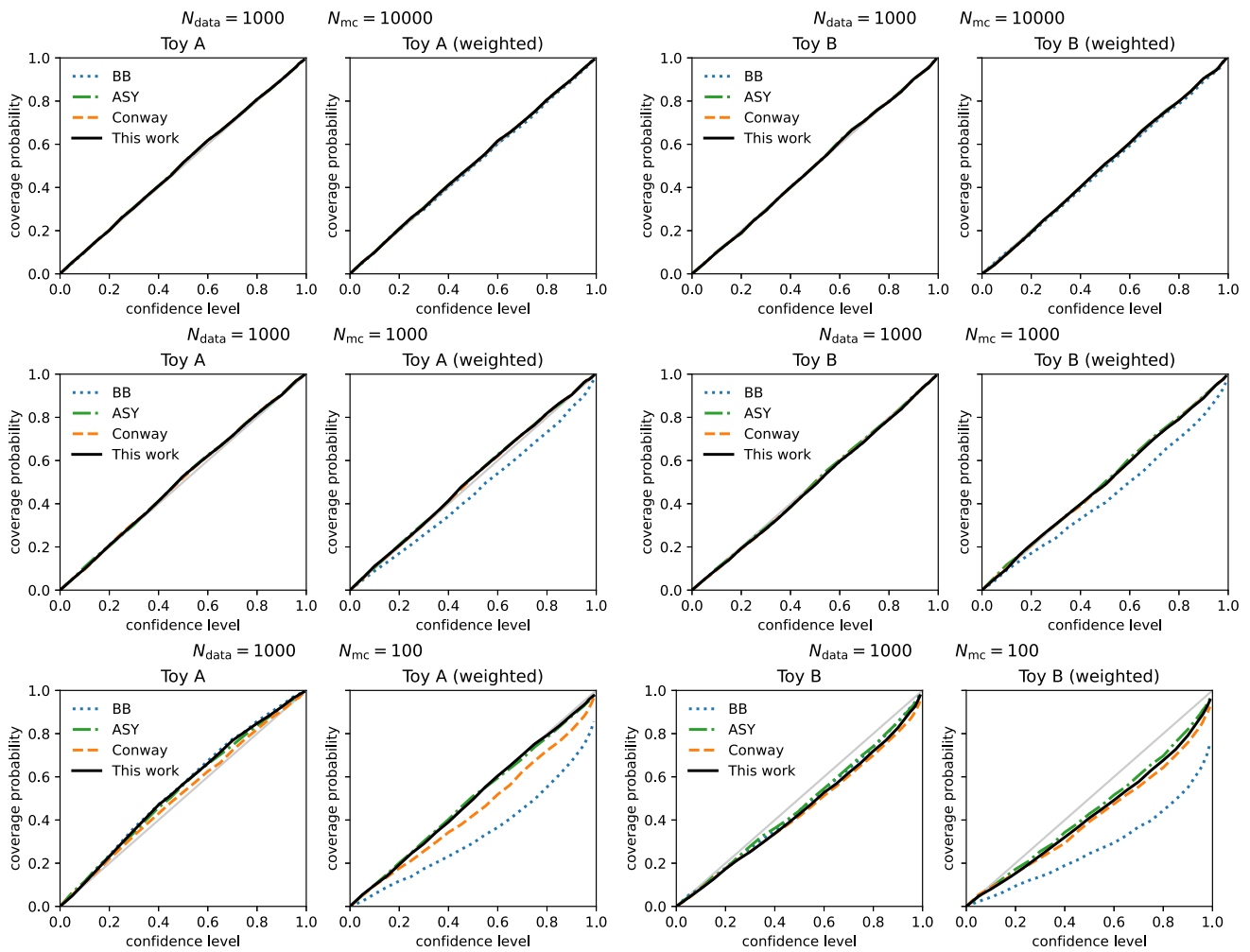


Fig. 5 Coverage probability as a function of the expected confidence level for the two toy ensembles with unweighted and weighted simulation samples. Outliers are included when the coverage probability is calculated

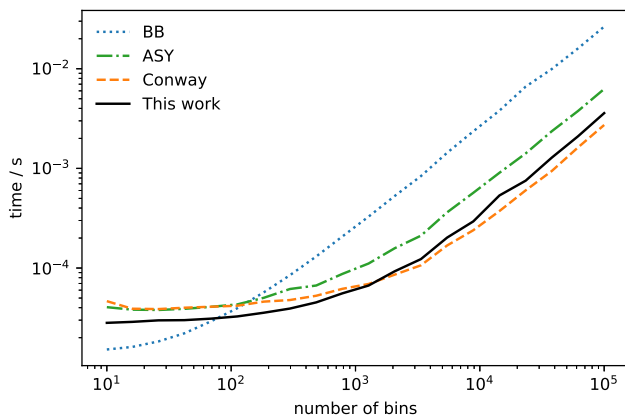


Fig. 6 Time required to evaluate the likelihoods described in the text; smaller is better. The likelihoods were evaluated on the toy example A. The number of bins in the data and template histograms was varied

samples), and the Argüelles–Schneider–Yuan likelihood are available in the `iminuit` library [9].

Acknowledgements We thank Roger Barlow, Carlos Argüelles, Austin Schneider, and Tianlu Yuan for valuable discussions in preparation of this work.

Funding HD acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project no. 449728698.

Data Availability Statement The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Code availability The likelihoods for template fits are available on the iminuit website: <https://iminuit.readthedocs.io/en/stable/reference.html#iminuit.cost.Template>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Funded by SCOAP³. SCOAP³ supports the goals of the International Year of Basic Sciences for Sustainable Development.

References

1. R.J. Barlow, C. Beeston, Fitting using finite Monte Carlo samples. *Comput. Phys. Commun.* **77**, 219–228 (1993). [https://doi.org/10.1016/0010-4655\(93\)90005-W](https://doi.org/10.1016/0010-4655(93)90005-W)
2. F. James, M. Roos, Minuit: a system for function minimization and analysis of the parameter errors and correlations. *Comput. Phys. Commun.* **10**, 343–367 (1975). [https://doi.org/10.1016/0010-4655\(75\)90039-9](https://doi.org/10.1016/0010-4655(75)90039-9)
3. J.S. Conway, Incorporating nuisance parameters in likelihoods for multisource spectra, in *PHYSTAT 2011* (2011), pp. 115–120. <https://doi.org/10.5170/CERN-2011-006.115>
4. C.A. Argüelles, A. Schneider, T. Yuan, A binned likelihood for stochastic models. *J. High Energy Phys.* **2019**(6), 1–18 (2019). [https://doi.org/10.1007/JHEP06\(2019\)030](https://doi.org/10.1007/JHEP06(2019)030)
5. G. Bohm, G. Zech, Statistics of weighted Poisson events and its applications. *Nucl. Instrum. Methods A* **748**, 1–6 (2014). <https://doi.org/10.1016/j.nima.2014.02.021>. [arXiv:1309.1287](https://arxiv.org/abs/1309.1287) [physics.data-an]
6. S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**(1), 60–62 (1938). <https://doi.org/10.1214/aoms/1177732360>
7. S. Baker, R.D. Cousins, Clarification of the use of Chi square and likelihood functions in fits to histograms. *Nucl. Instrum. Methods* **221**, 437–442 (1984). [https://doi.org/10.1016/0167-5087\(84\)90016-4](https://doi.org/10.1016/0167-5087(84)90016-4)
8. W. Cash, Parameter estimation in astronomy through application of the likelihood ratio. *Astrophys. J.* **228**, 939–947 (1979). <https://doi.org/10.1086/156922>
9. H. Dembinski, P. Ongmongkolkul, C. Deil, H. Schreiner, M. Feickert, Andrew, C. Burr, J. Watson, F. Rost, A. Pearce, L. Geiger, B.M. Wiedemann, C. Gohlke, Gonzalo, J. Drotleff, J. Eschle, L. Neste, M.E. Gorelli, M. Baak, O. Zapata, odidev: scikit-hep/iminuit: v2.11.2. Zenodo (2022). <https://doi.org/10.5281/zenodo.6389982>
10. R. Brun, F. Rademakers, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne. *Nucl. Inst. Meth. Phys. Res. A* **389**, 81–86 (1997). <https://doi.org/10.5281/zenodo.3895860>
11. A. Aab et al., Depth of maximum of air-shower profiles at the Pierre Auger Observatory. II. Composition implications. *Phys. Rev. D* **90**(12), 122006 (2014). <https://doi.org/10.1103/PhysRevD.90.122006>. [arXiv:1409.5083](https://arxiv.org/abs/1409.5083) [astro-ph.HE]
12. C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy. *Nature* **585**(7825), 357–362 (2020). <https://doi.org/10.1038/s41586-020-2649-2>