arXiv:2104.13879v2 [physics.data-an] 14 Jun 2021

Avoiding biases in binned fits

V. V. Gligorov,[a] **S. Hageboeck,** [b,1] **T. Nanut,**[c] **A. Sciandra,**[d] **and D. Y. Tou**[a]

[a]*LPNHE,*
  Sorbonne Université, Paris Diderot Sorbonne Paris Cité, CNRS/IN2P3
[b]*CERN, 1 Esplanade des Particules, Meyrin, Switzerland*
[c]*Institute of Physics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
[d]*Santa Cruz Institute of Particle Physics, University of California at Santa Cruz, 5 High Street, Santa Cruz,*
  CA, USA

  E-mail: stephan.hageboeck@cern.ch

ABSTRACT: Binned maximum likelihood fits are an attractive option when analysing large datasets, but require care when computing likelihoods of continuous PDFs in bins. For many years the widely used statistical modelling package RooFit evaluated probabilities at the bin centre, leading to significant biases for strongly curved probability density functions. We demonstrate the biases with real-world examples, and introduce a PDF class to RooFit that removes these biases. The physics and computation performance of this new class are discussed.

---

[1]Corresponding author.

Contents

1 Introduction

RooFit [1, 2] is a C++ package for statistical modelling distributed with Root [3, 4]. RooFit's development started in the year 2000 within the BaBar collaboration. Since then, RooFit has been a reliable tool for many experiments in high-energy physics (HEP) at *B* factories and the Large Hadron Collider. With RooFit, users can define likelihood models using observables, parameters, functions and PDFs[1], which can be fitted to data, plotted or used for statistical tests.

As HEP datasets become progressively larger, it becomes increasingly important to reduce the computation time of data analysis tasks such as fits. At the same time, the higher statistics allow for measurements with higher precision. These demands of speed and accuracy can come into conflict. In this paper, we address the long-standing problem of a bias in RooFit's binned likelihood fits, while maintaining much of their speed advantage over unbinned likelihood fits.

Figure 1: Error of estimating the probability in a bin by evaluating a PDF at the bin centre.

These biases arise since in binned fits, RooFit samples probability densities at only one point in a bin, the bin centre. It is assumed that this is a decent approximation for the probability of the entire bin. When distributions are strongly curved, though, this is inaccurate. Figure 1 illustrates the error of this approximation with a Gaussian distribution. For the shown bin, the probability to the left of the bin centre is overestimated by the surface shaded in red. To the right of the bin centre,
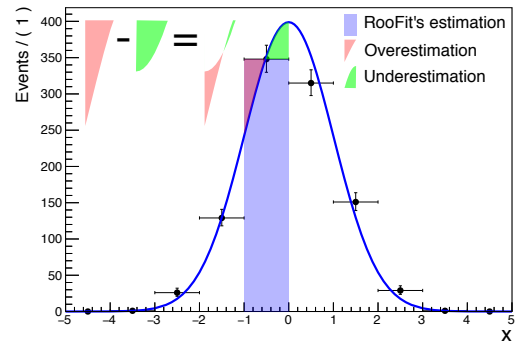
---

[1]Probability Density Functions

it is underestimated by the surface shaded in green. If the PDF was linear in the bin, these two surfaces would have equal size. However, if the PDF has a non-vanishing second derivative, the compensation is insufficient. In the marked bin, the probability is overestimated.

Although well known, removing this bias was considered too invasive for RooFit. Users were encouraged to use more bins to reduce biases, which is not always possible, though. We present here the first rigorous solution, extending RooFit with a new PDF class, which transforms a continuous into a binned PDF. This leaves likelihood and χ^2 calculation functions, as well as plotting routines, unaffected, but eliminates the bias. We demonstrate the biases with real-world examples, and discuss the physics and computation performance using this PDF class. The new PDF is available in Root since version 6.24.

2   Integrating continuous PDFs for unbiased binned fits

To eliminate biases in binned fits, the PDF class "RooBinSamplingPdf" [5, 6] was added to RooFit. It transforms continuous PDFs into binned PDFs by integrating the former in each bin and dividing the result by the bin width. The resulting PDF is constant in each bin, and evaluates to the *average* probability density in a bin (instead of the probability density at the bin centre). This allows for fitting continuous PDFs to binned data, or for adding binned and continuous PDFs to create sum models.

To integrate the original PDF, Root's IntegratorOneDim is used. Internally, this uses the adaptive Gauss-Kronrod [7] integrator with a 21-point rule from the GSL [8].[2] For smooth functions, this means that 21 times more function evaluations are required compared to using a simple binned PDF. If the adaptive algorithm has to subdivide single bins to improve the integration accuracy, the computational overhead increases to a bit more than 21-fold, but this happens only in bins where the original PDF is not smooth. Nevertheless, since binned fits are usually orders of magnitudes faster than running an unbinned fit, see section 2.1, the additional time to perform a more accurate integration is still acceptable if users get unbiased results.

Users can further set the relative precision required for the integration. They can even directly manipulate all settings of Root's IntegratorOneDim to customise accuracy and speed, e.g. switch to a 15-point Gauss-Kronrod rule.

Figure 2 shows a fit model that is a sum of three Gaussian distributions and a JohnsonSU distribution [9] — a typical signal model for analysing charm decays [10]. When RooBinSamplingPdf is used, the probability density represented by the PDF is constant in each bin, whereas it is strongly curved using the original model. The pads at the bottom show pulls[3] comparing the plotted curves evaluated at the bin centres and data points sampled from the model. Due to the over- or underestimation illustrated in figure 1, strong pulls are observed in figure 2a. Since a fitter will try to balance the overshoot in the tails of the signal model with the undershoot in the centre, fit results will be biased. Figure 2b shows that the use of RooBinSamplingPdf eliminates the pulls.

---

[2]In environments where Root is used without the GSL, Root's non-adaptive GaussIntegrator is used.

[3]"Pull" denotes a residual divided by the standard deviation of a quantity. In this work, this term is used both to quantify a deviation of model parameters from a theoretical value as well as for comparing a fitted distribution with data.
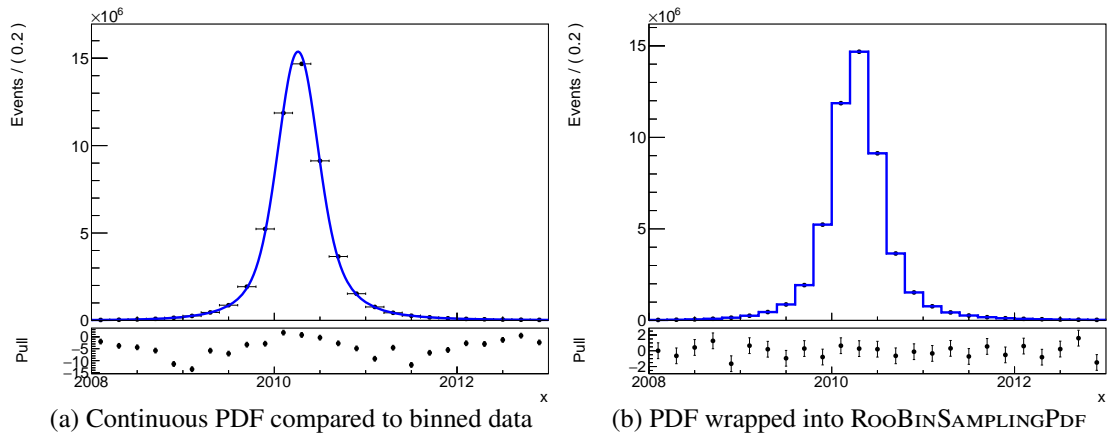
(a) Continuous PDF compared to binned data (b) PDF wrapped into RooBinSamplingPdf

Figure 2: Comparison of pulls without and with RooBinSamplingPdf [5]. The bottom pads show pulls that are computed by comparing data counts with the plotted curves evaluated at the bin centres. In (b), the *y*-axis of the pull plot is zoomed 4-fold.

Users have two ways to use this new PDF class. First, they can directly construct an instance of RooBinSamplingPdf. They pass an observable that defines the desired binning, the continuous PDF, and (if desired) the relative precision for bin integrals:

```
RooBinSamplingPdf binSampler("Name", "Plot Label",
                             binnedObservable, originalPDF,
                             /*optional integrator precision=*/ 1.E-4);
binSampler.fitTo(data);
```

Second, RooFit can be instructed to automatically construct instances of RooBinSamplingPdf before each fit:

```
pdf.fitTo(data, RooFit::IntegrateBins(<integrator precision>));
```

The first strategy is most versatile, because users can control which PDFs are transformed, and which integrators are used. Moreover, plots of RooBinSamplingPdf correctly convey the fact that a binned fit was used, and pull plots will also benefit from the correction (cf. figure 2).

The second strategy is particularly useful in simultaneous fits with multiple channels, where a mixture of binned and unbinned fits is used. When `RooFit::IntegrateBins(0.)` is passed to the fit instruction, RooFit will apply the correction to all continuous PDFs that are fit to binned data, but leave unbinned fits unchanged. Passing a non-zero value, the correction will be applied to all channels, irrespective of whether a binned or unbinned data set is used. This is important since some fitting frameworks simulate binned fits by creating an "unbinned" data set, where the coordinates of an entry correspond to a bin centre, and the weight of an entry corresponds to the bin content. Although this produces the same result as a binned fit, RooFit cannot detect that the correction discussed in this work needs to be applied, so users need a way to explicitly enable it.

Although the bias in binned fits could also have been corrected for by changing the likelihood calculation functions in RooFit, adding the new PDF class has advantages:

- It is not invasive. No changes to RooFit's fitting algorithms are required. Since the probability density of RooBinSamplingPdf is constant in each bin, RooFit's default strategy of evaluating the function in the centre and multiplying by the bin width is correct even if the original PDF is curved strongly.

- It is reusable. The same strategy that is used to correct likelihood fits can be used to correct χ^2 fits or the computation of χ^2 statistics (cf. figure 5, section 3.1).

- It allows for better plotting. Since data are usually plotted in bins, plotting a binned PDF allows for inspecting differences by eye. When plotting binned data and a continuous PDF, the viewer would have to estimate the integral of the PDF over a bin. Moreover, the new class overrides the function `RooAbsArg::plotSamplingHint()`. This function conveys the optimal points for plotting a function to RooFit's plotting routines. If this function would not be overridden, for example since the correction was implemented in the likelihood functions, RooFit would employ heuristics to estimate the best vertices for drawing a curve, which can lead to binning artifacts in the plotted curves.

- It is achieved with a limited computational overhead, which depends on the integrator settings. Independent of the integration in bins, RooFit needs to ensure that all PDFs are normalised, which means that they need to be integrated over the fit range. If no analytic integral is known, numeric integration algorithms are employed, which would typically evaluate a PDF at 64–256 points. If RooBinSamplingPdf is used, however, no further automatic normalisation needs to be performed, since the integrals in the bins can be reused to compute the integral over the fit range. The computational overhead therefore only depends on the integrator settings of RooBinSamplingPdf, but the integration step that would normally be executed automatically by RooFit is skipped.

2.1 Time complexity of likelihood calculations

An unbinned fit in RooFit estimates the parameters of a model by maximising the negative logarithmic likelihood, $\mathcal{L}$, defined as

$$-\log \mathcal{L} = -\sum_{i=1}^{N} \log f(x_i \mid \boldsymbol{\theta}), \tag{2.1}$$

where $f(x_i \mid \boldsymbol{\theta})$ is the PDF of the model, and the likelihood is a product over N observations x_i. The term $f(x_i \mid \boldsymbol{\theta})$ is the probability to observe an event x_i, given the model f and parameters $\boldsymbol{\theta}$. In the case where the PDF normalisation is a parameter of interest, the extended negative log likelihood is used

$$-\log \mathcal{L} = -\mu - N \log \mu + \log N! - \sum_{i=1}^{N} \log f(x_i \mid \boldsymbol{\theta}), \tag{2.2}$$

where μ is the expected number of events predicted by the model. The additional terms have a negligible impact on the time of likelihood computations.[4] From equations (2.1) and (2.2), it is

---

[4]The constant terms are ignored by RooFit, as they are not needed to find the minimum, the other terms are only evaluated once.

evident that the time to compute a log likelihood scales with the number of events in the data set.

$$T\,[\text{unbinned}] = N_{\text{event}} \cdot T[\log f(x_i \mid \boldsymbol{\theta})] = O(N_{\text{event}}). \tag{2.3}$$

Analogously, the likelihood to fit binned data

$$\mathcal{L} = \prod_{i=1}^{N_{\text{bin}}} \frac{\mu_i^{n_i}(\boldsymbol{\theta})}{n_i!} \cdot \exp\left(\mu_i(\boldsymbol{\theta})\right) \tag{2.4}$$

scales as

$$T\,[\text{binned}] = N_{\text{bin}} \cdot T[\log \mu_i(\boldsymbol{\theta})] = O(N_{\text{bin}}). \tag{2.5}$$

where n_i is the number of events observed in bin i, and $\mu_i(\boldsymbol{\theta})$ the expected number of events in that bin. Computing μ takes roughly the same time as computing one event likelihood in the unbinned case, because it requires evaluating the model in the bin: $f(x_{\text{centre}} \mid \boldsymbol{\theta})$.

By default, RooFit minimises negative log likelihoods using the MIGRAD algorithm of the MINUIT package [11], and covariance matrices are estimated using the HESSE algorithm. Since both MIGRAD and HESSE rely on calculating finite differences, their time complexities scale with the complexity of the likelihood calculations. The speed up when using binned fits is therefore proportional to $N_{\text{event}}/N_{\text{bin}}$. Despite the fact that $T[\log \mu_i(\boldsymbol{\theta}); \text{bin sampling}] \approx 20 \cdot T[\log f(x_i \mid \boldsymbol{\theta})]$ with our proposed solution and default integrator settings, since 21 points have to be evaluated to integrate each bin, binned fits are usually invoked for large data sets where $N_{\text{bin}} \ll N_{\text{event}}$. Hence, the solution presented in this paper still offers orders of magnitude speed up compared to unbinned fits, while eliminating unwanted biases.

3   Validation with realistic pseudo experiments

In order to demonstrate the real-world impact of these biases and the performance of our method, we perform pseudo experiments using examples representative of LHCb [12] and ATLAS [13] collaboration analyses.

3.1   LHCb charm example

We demonstrate and evaluate the performance of the default RooFit binned fit method and the RooBinSamplingPdf using the signal model and data set conditions (yield, number of bins) from [14]. The PDF models the D^{*+} peak of the decay $D^{*+} \rightarrow D^0(\rightarrow K^-\pi^+)\pi^+$ and is comprised of two Gaussian functions and a JohnsonSU distribution. The number of floating parameters is 10. While the total signal yield is 520×10^6 events, the fit is performed separately in 20 equipopulous bins of decay time, yielding 26×10^6 events to be fitted in one sample. The number of bins is 160.

For each of the two methods, we generate 10^4 pseudo experiments using the PDF values from one of the fits in reference [14]. The generated data set is then fitted using the PDF with floating parameters. Example fits for each method are shown in figure 3. As in figure 2, the bias is clearly seen when using the default fit method, while not so when using RooBinSamplingPdf. For each pseudo experiment and each fit parameter, we compute the difference between the fitted and generated value, divided by the error on the fitted value. While this quantity is not the exact
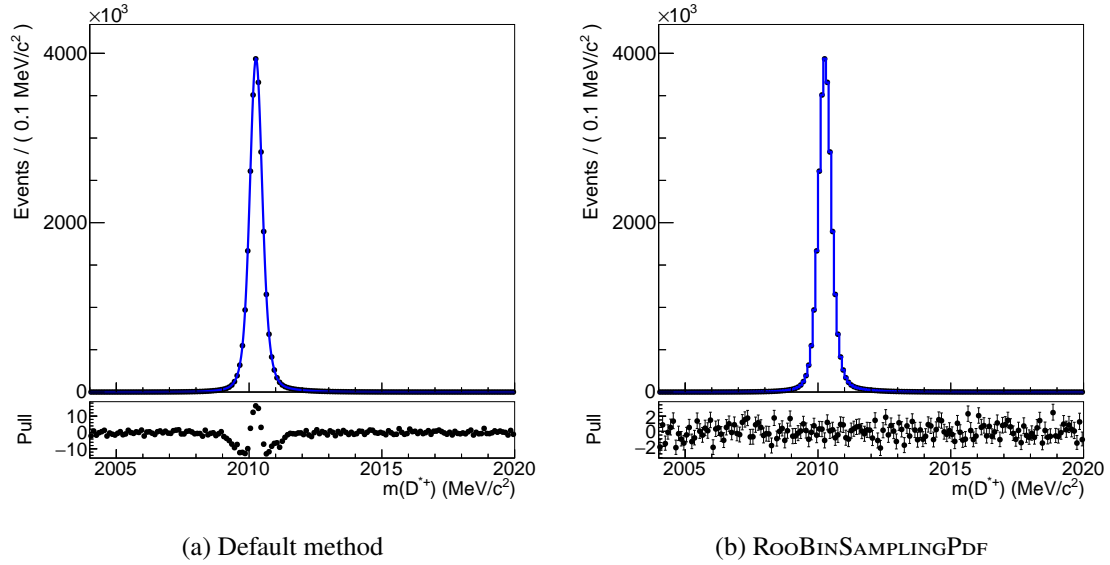
(a) Default method

(b) RooBinSamplingPdf

Figure 3: Example of a fit to one pseudo experiment for both methods. The bottom pads show pulls that are computed by comparing data counts with the plotted curves evaluated at the bin centres.

equivalent of the statistical pull, as we use the generated value and not the expected one, the resulting distribution is still expected to follow a Gaussian with mean at 0 and standard deviation 1.

We thus fit the obtained distribution with a Gaussian function and study the mean and standard deviation. The results are shown in figure 4. It is clear that when using the default method, large biases are present in almost all fitted parameters. Contrary, for the RooBinSamplingPdf, the fitted parameters show no discernible bias. The plots of the standard deviation show that the error estimation is reliable for all parameters when using RooBinSamplingPdf, and all but one for the default method.

As an additional cross-check, we examine the distribution of χ^2 values obtained from the fits of the pseudo experiments. The number of degrees of freedom is the number of bins minus the number of free parameters in the fit, i.e. 150, so the obtained distribution is compared to the theoretical χ^2 distribution for 150 degrees of freedom. The results are shown in figure 5. We can observe that when using RooBinSamplingPdf, the obtained distribution matches reasonably well the theoretical one, while for the default binned fit method, it is very different.

In the above described studies, the number of bins was taken as that reported in reference [14]. Figure 6 shows plots of the mean and width of the Gaussian distribution, fitted to the ratio of the difference between the generated and fitted values and the error of the fitted value (analogue to figure 4), for different binnings using RooBinSamplingPdf. It can be concluded that a different choice of binning does not affect the performance of the RooBinSamplingPdf.
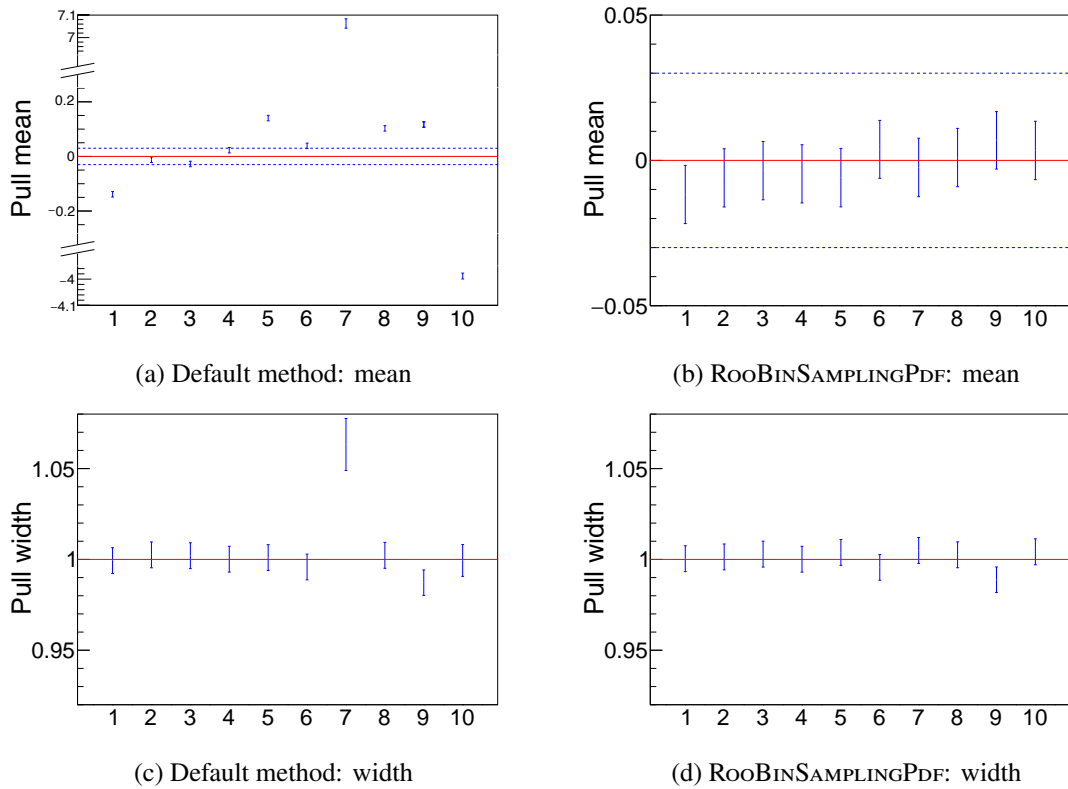
Figure 4: Mean (top) and width (bottom) of the pull of various fit parameters (*x*-axis), with respect to the generated and fitted values, for both methods. The blue dashed lines in (a) and (b) are at the same values to serve as visualisation aid.
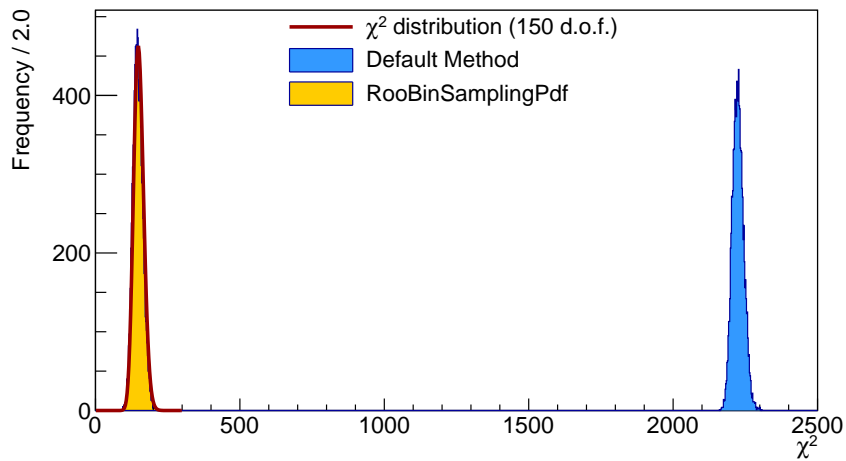

Figure 5: Comparison of the theoretical χ^2 distribution with NDOF = 150 and the empirical distributions obtained from fits to pseudo experiments for both methods. While the standard method results in very high values of χ^2, the values with RooBinSamplingPdf follow the theoretical distribution.
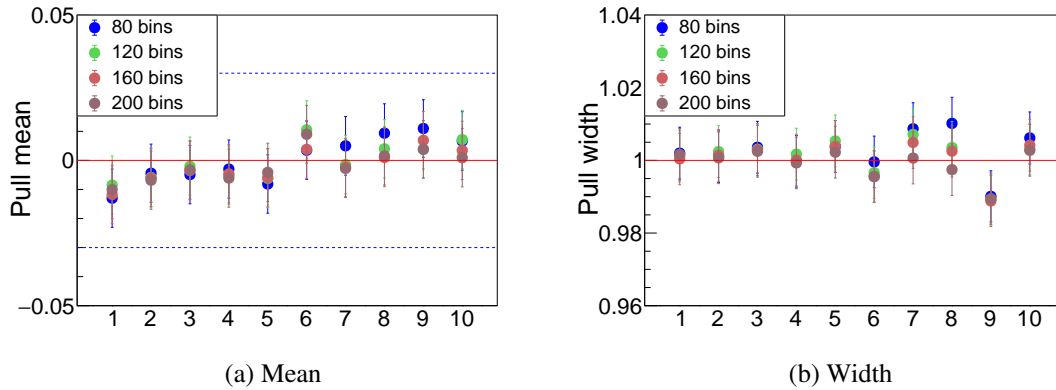
Figure 6: Mean and width of the pulls of various fit parameters (*x*-axis), with respect to the generated and fitted values, for different choices of binning using RooBinSamplingPdf.
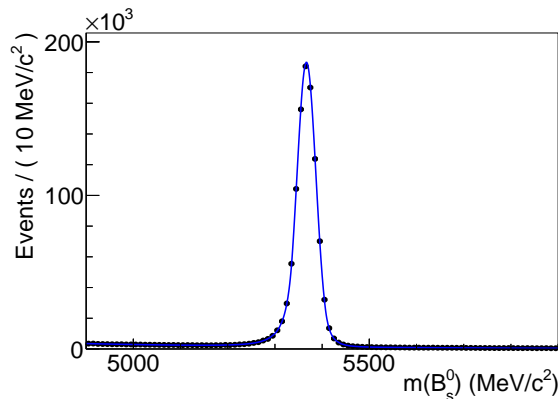

Figure 7: An example pseudo experiment generated by the baseline mode. The data set (black) is generated by a model (blue) which consists of a DSCB signal and a combinatorial background.

3.2 LHCb beauty example

In this section, we explore in more detail the effects of the bias with regard to signal width, number of signal events and number of bins, using a model with a signal PDF similar to typical beauty meson decays at LHCb and a combinatorial background model.

The signal is a Double Sided Crystal Ball (DSCB) function [15], with a Gaussian core of mean $\mu_{\text{gauss}} = 5366.79$ MeV (corresponding to the B_s^0 rest mass) and width $\sigma_{\text{gauss}} = 20$ MeV. The combinatorial background is an exponential function with a slope of $b_{\text{slope}} = -2 \times 10^{-3}$ MeV^{-1}. The baseline model has 1×10^6 signal events and 1.5×10^5 background events within a mass window of 4900–5900 MeV. A plot of the typical data set generated by this model is shown in figure 7.

We study the bias as a function of signal width, the number of signal events, and the number of bins. At a given bin width, broader signal distributions are expected to be less biased than narrower ones. Similarly, fits to the same signal distribution with narrower bin widths are expected to be less biased than fits with wider bins.
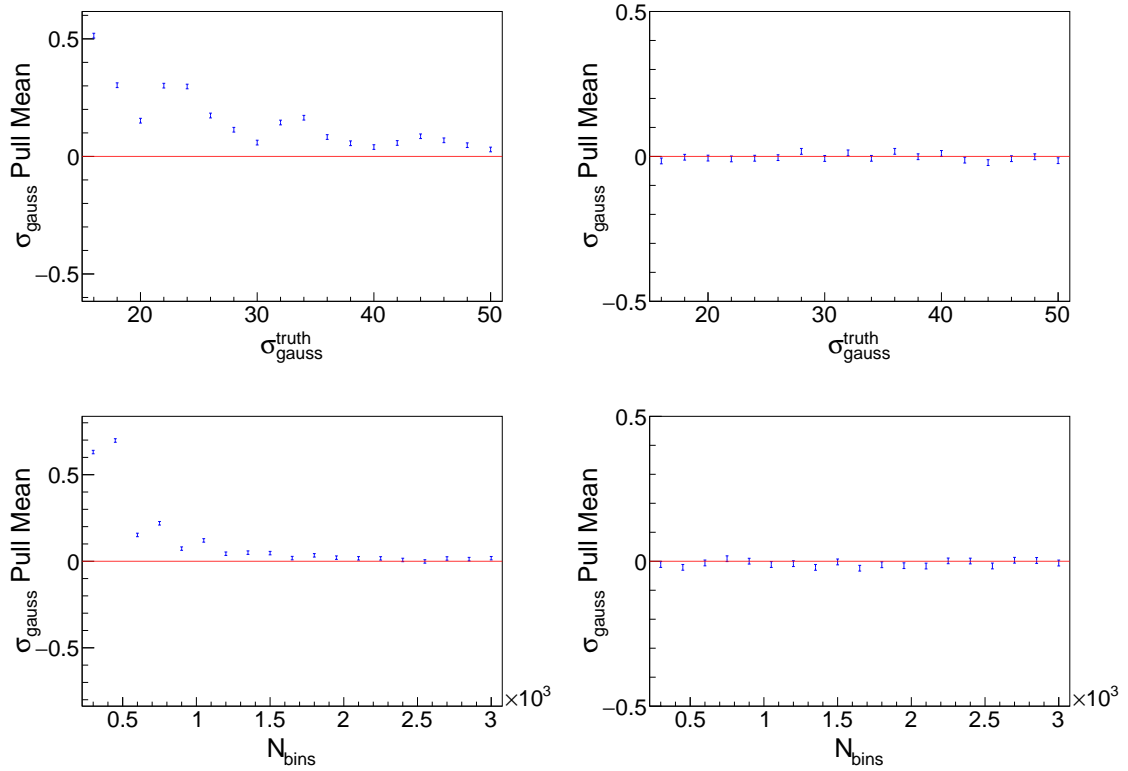
Figure 8: Mean pull of σ_{gauss} in dependence of $\sigma_{\text{gauss}}^{\text{truth}}$ (top) and N_{bins} (bottom) with the default method (left) and RooBinSamplingPdf (right).
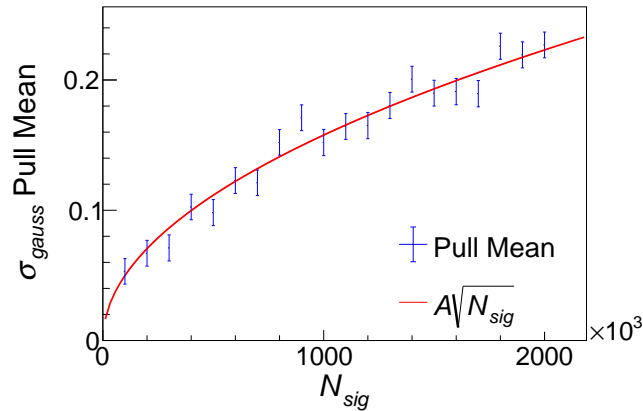

Figure 9: The bias of σ_{gauss} against $N^{\text{truth}}(\text{signal})$ (black) is well modelled by $A\sqrt{N(\text{signal})}$ (red).

The results are shown in figure 8. As we can see, the pull bias in σ_{gauss} increases as $\sigma_{\text{gauss}}^{\text{truth}}$ decreases and as N_{bins} decreases. The σ_{gauss} bias is plotted as a function of $N^{\text{truth}}(\text{signal})$ in figure 9. Interestingly, the bias is well modelled as $A\sqrt{N(\text{signal})}$. This proves that the absolute bias remains unchanged with respect to $N^{\text{truth}}(\text{signal})$, but it is the $\frac{1}{\sqrt{N(\text{signal})}}$ uncertainty scaling that amplifies the pull bias in σ_{gauss}.
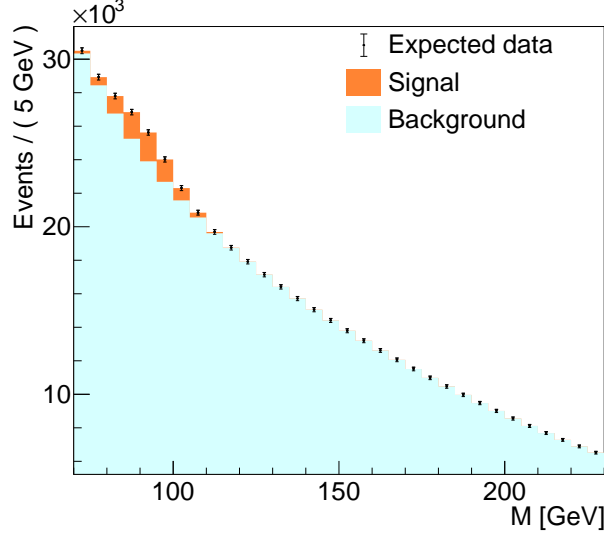
Figure 10: The default signal-plus-background expected distribution fitted with the RooBinSam-plingPdf integration. Expected data points are shown as black points, whereas the resulting fitted model consists of the signal (orange) stacked on top of the background (cyan).

3.3 ATLAS example

The adoption of RooBinSamplingPdf functionalities finds extensive application in analyses carried out by the ATLAS Collaboration. In fact, some ATLAS analyses rely on analytic models and binned likelihood fits, see for example $H \rightarrow \mu\mu$ [16]. The absence of RooBinSamplingPdf forced many physics analyses to rely on very small bin widths, as the bias shown in figure 2a becomes negligible with bin size orders of magnitude smaller than the resolution in the characteristic structures of the considered fit observable.

However, analyses based on a mixture of pure analytic models and template PDFs from simulation may not be able to accommodate a full parameterisation of the fit model. Recent ATLAS studies of the inclusive boosted $H \rightarrow b\bar{b}$ production at high Higgs-boson transverse momentum [17] exploit the new class, object of this paper. The preliminary version of the work discussed in Ref. [18] was completed before the presented RooFit developments and is, therefore, taken as a natural field of application for the RooBinSamplingPdf class. This example supplements studies shown in sections 3.1 and 3.2, as the signal-to-background ratio is much smaller and an analytic model is exploited to describe the background only, inducing an indirect bias in the signal extraction.

The model consists of a linear sum of the "background" and "signal" models. The background is described by means of a polynomial exponential PDF model of the form:

$$f_N\left(x \,\middle|\, \vec{\theta}\right) = \exp\left(\sum_{i=1}^{N} \theta_i x^i\right),$$

where $N = 4$, $x = (M - 150)/80$, with M the fit observable, and θ_i are fit parameters. A fit range of [70,230] is taken for M, resulting in [-1,1] for x. The following model parameter values are taken: $\vec{\theta} = (-0.7, -0.05, -0.1, 0.05)$. The signal shape is fixed and modelled via a Gaussian centred at

$M = 91$ with a standard deviation of $\sigma = 8.5$. The default signal fraction (f_{sig}), total number of events (N_{ev}) and of bins (N_{bins}) are 0.015, 5×10^5 and 32, respectively. Figure 10 shows the resulting falling background spectrum with a "Z-like" resonance on top.

The bias in the extracted fit parameters is dependent on the global and local statistical accuracy and is, therefore, studied as a function of the available statistics over the whole fit range (N_{ev}) and in each fit bin (N_{bins}). For each value of N_{ev} and N_{bins} a thousand data sets are generated through Poisson-sampling of the expected distribution, and fitted with both the default method and the RooBinSamplingPdf.

The pull for the fit parameters with respect to the generated and fitted values, expected to be centred around $\mu = 0$ with a standard deviation of $\sigma = 1$, is fitted with a Gaussian function. The dependence of the bias in all of the fit parameters is shown in figures 11 and 12 for six values of N_{bins} and five values of N_{ev}, respectively. It is evident that a default binned fit induces significant biases, unless a large N_{bins} or a small N_{ev} is assumed. In fact, expected biases smoothly decrease as a function of the increasing (decreasing) N_{bins} (N_{ev}). Sometimes, as for example looking at f_{sig} for $N_{\text{bins}} = 16$, the estimate of the fit-parameter standard deviation may be biased, too. On the contrary, when exploiting the RooBinSamplingPdf, the estimate of fit-parameter central values and standard deviations appears robust on a large ensemble of N_{bins} and N_{ev} values. Small deviations with very small N_{bins} or very large N_{ev}, see for instance $N_{\text{ev}} = 50 \times 10^6$ in figure 12, and the RooBinSamplingPdf may be recovered by requiring higher integration precision.

4 Conclusions and future work

Currently, all PDFs are integrated numerically, even if analytic integrals are known to RooFit. This can be changed should the need arise. Furthermore, the method is restricted to one-dimensional PDFs. Similarly, this can be extended to multi-dimensional PDFs if necessary.
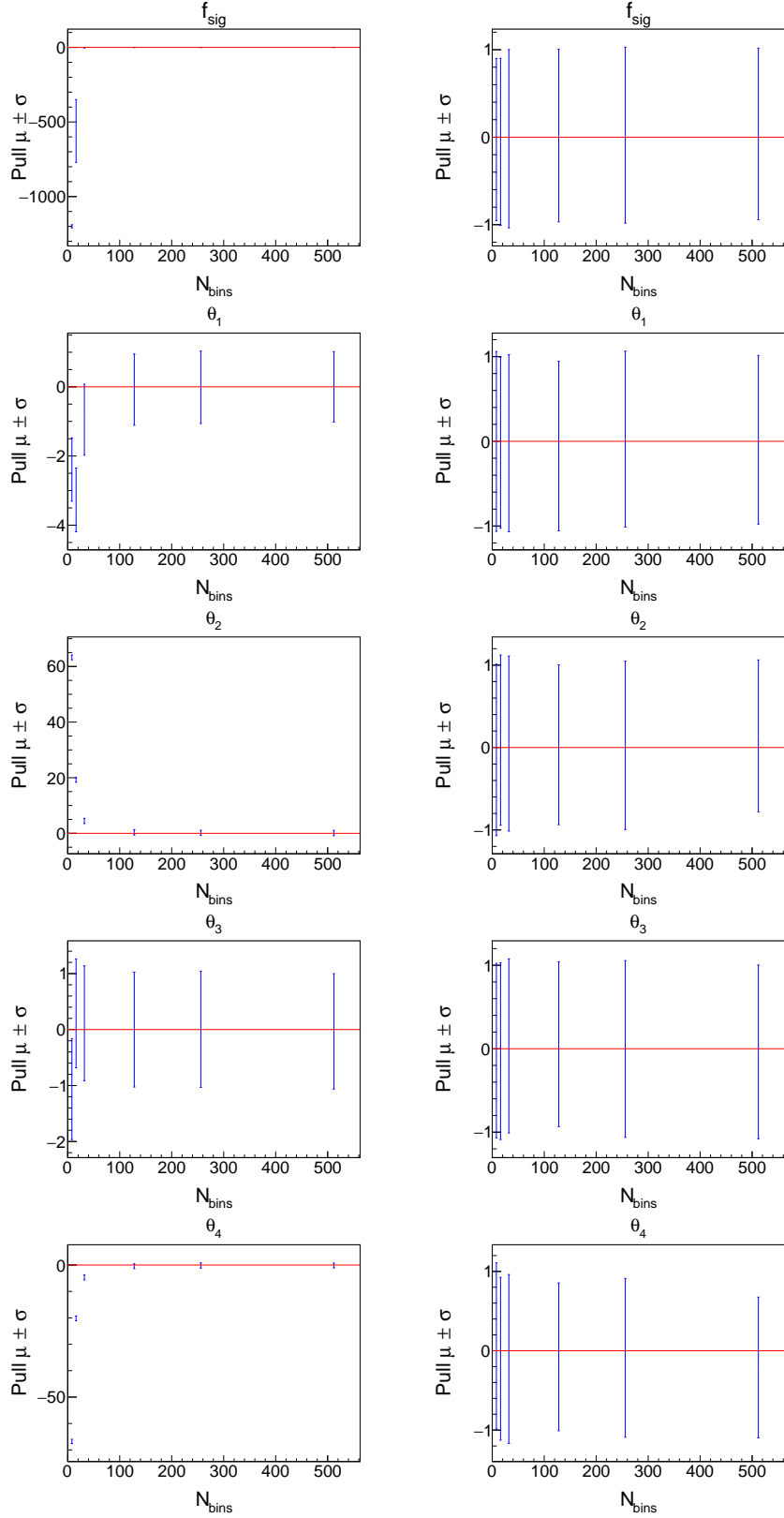
Acknowledgments

Figure 11: Mean and standard deviation of the pull as a function of N_{bins} from a Gaussian fit for various fit parameters with respect to the generated and fitted values. Fits are performed with the default (left) and RooBinSamplingPdf (right) methods. A thousand data sets are generated and fitted for each value of N_{bins}.
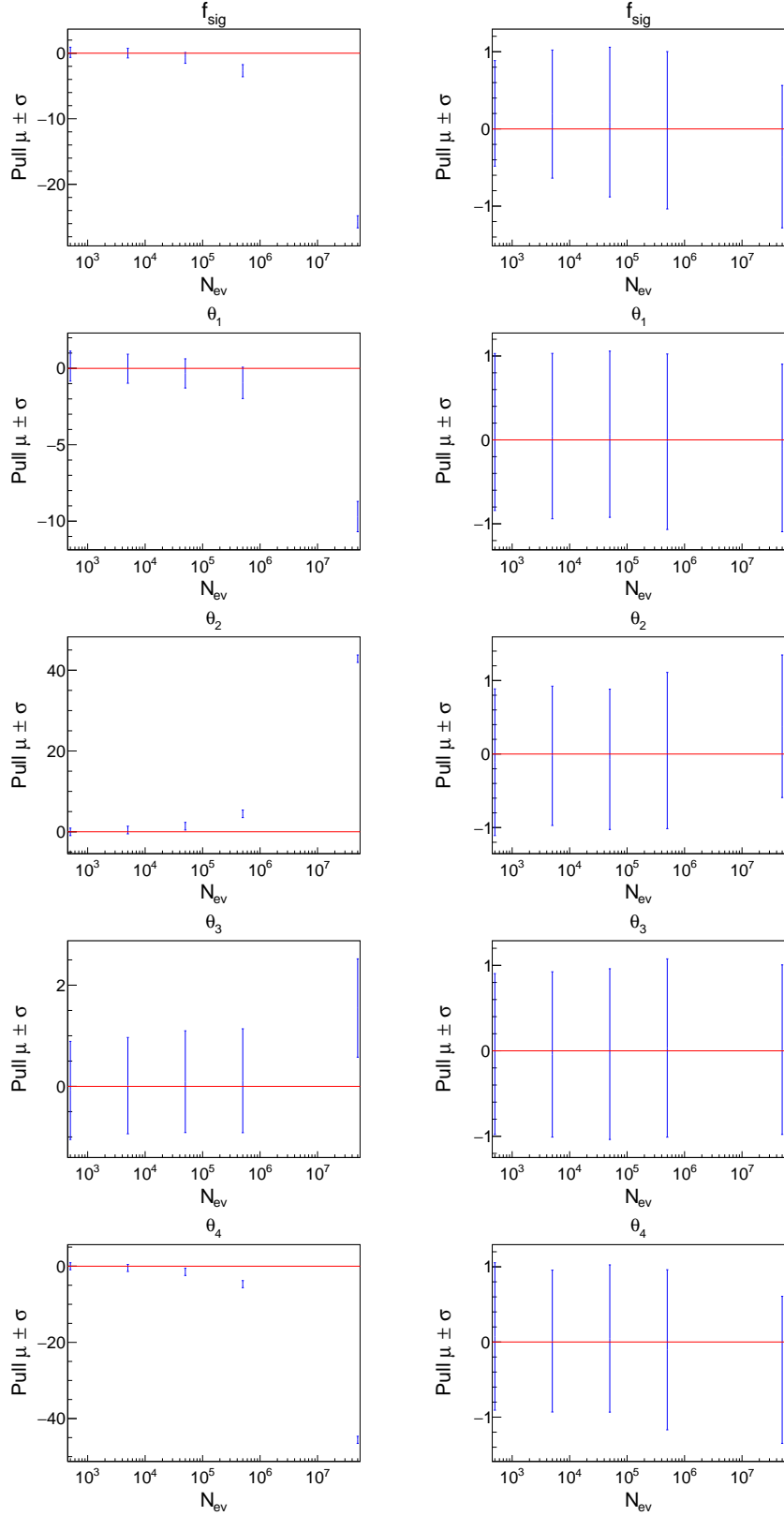
Figure 12: Mean and standard deviation of the pull as a function of N_{ev} from a Gaussian fit for various fit parameters with respect to the generated and fitted values. Fits are performed with the default (left) and RooBinSamplingPdf (right) methods. A thousand data sets are generated and fitted for each value of N_{ev}. The x-axis is shown in logarithmic scale.

5 Bibliography

[1] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling*, *eConf* **C0303241** (2003) MOLT007 [`physics/0306116`].

[2] ROOT Team, *RooFit Manual*, 2021. root.cern/manual/roofit.

[3] R. Brun and F. Rademakers, *ROOT: An object oriented data analysis framework*, *Nucl. Instrum. Methods Phys. Res.* **A389** (1997) 81.

[4] ROOT Team, *ROOT Manual*, 2021. root.cern/manual.

[5] S. Hageboeck, *What the new RooFit can do for your analysis*, in *Proceedings of 40th International Conference on High Energy physics — PoS(ICHEP2020)*, vol. 390, p. 910, 2021, DOI [`2012.02746`].

[6] S. Hageboeck and A. Sciandra, *RooBinSamplingPdf Class Reference*, 2020. root.cern/doc/master/classRooBinSamplingPdf.html.

[7] A. Kronrod, *Nodes and Weights of Quadrature Formulas: Sixteen-place Tables*. Consultants Bureau, 1966.

[8] B. Gough, *GNU Scientific Library Reference Manual - Third Edition*. Network Theory Ltd., 3rd ed., 2009.

[9] N. L. Johnson, *Systems of frequency curves generated by methods of translation*, *Biometrika* **36** (1949) 149.

[10] LHCb Collaboration, *Observation of CP Violation in Charm Decays*, *Phys. Rev. Lett.* **122** (2019) 211803 [`1903.08726`].

[11] F. James, "MINUIT Function Minimization and Error Analysis: Reference Manual Version 94.1." CERN-D-506, 1994.

[12] LHCb Collaboration, *The LHCb Detector at the LHC*, *JINST* **3** (2008) S08005.

[13] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003.

[14] LHCb Collaboration, *Search for time-dependent CP violation in $D^0 \rightarrow K^+K^-$ and $D^0 \rightarrow \pi^+\pi^-$ decays*, *In preparation* (2021) .

[15] T. Skwarnicki, *A study of the radiative CASCADE transitions between the Upsilon-Prime and Upsilon resonances*, Ph.D. thesis, Cracow, INP, 1986.

[16] ATLAS Collaboration, *A search for the dimuon decay of the Standard Model Higgs boson with the ATLAS detector*, *Phys. Lett. B* **812** (2021) 135980 [`2007.07830`].

[17] ATLAS Collaboration, "Study of Higgs-boson production with large transverse momentum using the $H \rightarrow b\bar{b}$ decay with the ATLAS detector." ATLAS-CONF-2021-010.

[18] ATLAS Collaboration, "Search for boosted resonances decaying to two b-quarks and produced in association with a jet at $\sqrt{s}$ = 13 TeV with the ATLAS detector." ATLAS-CONF-2018-052.