ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202511.00129

A Novel Approach for Signal Number Estimation in Low-Statistics Measurements

Authors: Tian, Mr. Ye, Dr. Senquan Lu, Tang, Dr. Zhi-Cheng, Dr. Hsin-Yi Chou, Dr. Feng-Ze Zhang, Chang, Prof. Yuan-Hann, Li, Prof. Zu-Hao, Lu, Senquan

Date: 2025-11-11T00:00:00+00:00

Abstract

We present CombineFit, a novel approach for estimating the signal number in low-statistics measurements. Traditional binned maximum likelihood and template fitting techniques often suffer from significant bias and increased uncertainties when the statistics of templates or target data are limited. CombineFit employs analytical functions to simultaneously fit the signal and background templates and target data by minimizing a joint likelihood function. This method is validated with toy Monte Carlo simulation by varying the number of signal/background templates and data samples and has been successfully applied in the data analysis of the Alpha Magnetic Spectrometer. With 10 events in the background template, the binned CombineFit achieved a minimal bias of 2% and an uncertainty of 5.8%, compared to TFractionFitter' s bias of 30% and uncertainty of 10%. The unbinned CombineFit further reduces the bias to 1% while maintaining the same uncertainty, whereas the RooFit with Kernel Density Estimation method yields a bias of 3.5% and an uncertainty of 11.5%. These results demonstrate that CombineFit provides a robust solution for signal number estimation under limited data statistics, offering broad applicability in the search for new physics.

Full Text

Preamble

A Novel Approach for Signal Number Estimation in Low-Statistics Measurements

Ye Tian,[1,2] Sen-Quan Lu,[1,†] Zhi-Cheng Tang,[1] Hsin-Yi Chou,[3] Feng-Ze Zhang,[1] Yuan-Hann Chang,[3] and Zu-Hao Li[1,2,‡]

[1]Key Laboratory of Particle Astrophysics, Institute of High Energy Physics, Chinese Academy of Sciences, 100049 Beijing [2]University of Chinese Academy of Sciences, 100049 Beijing [3]Institute of Physics, Academia Sinica, 115201 Taipei

We present CombineFit, a novel approach for estimating signal numbers in low-statistics measurements. Traditional binned maximum likelihood and template fitting techniques often suffer from significant bias and increased uncertainties when the statistics of templates or target data are limited. CombineFit employs analytical functions to simultaneously fit the signal and background templates and target data by minimizing a joint likelihood function. This method is validated with toy Monte Carlo simulation by varying the number of signal/background templates and data samples and has been successfully applied in the data analysis of the Alpha Magnetic Spectrometer. With 10 events in the background template, the binned CombineFit achieved a minimal bias of 2% and an uncertainty of 5.8%, compared to TFractionFitter' s bias of 30% and uncertainty of 10%. The unbinned CombineFit further reduces the bias to 1% while maintaining the same uncertainty, whereas the RooFit with Kernel Density Estimation method yields a bias of 3.5% and an uncertainty of 11.5%. These results demonstrate that CombineFit provides a robust solution for signal number estimation under limited data statistics, offering broad applicability in the search for new physics.

Keywords: Template Fitting; Statistical Analysis; RooFit; Low Statistics

Introduction

The decomposition of measured data into contributions from distinct physical processes lies at the core of statistical analysis in particle physics [**?**, **?**]. Generally, the probability density functions (PDFs) of the observable (e.g., invariant mass) from different contributions are obtained by Monte Carlo (MC) simulations, and then the template fitting technique is applied to the experimental data by minimizing the χ^2 statistics or negative log-likelihood (NLL) functions to obtain the composition of each contribution.

In some cases, Monte Carlo simulations do not reproduce the data precisely, thus the templates are obtained by applying tighter selections from data directly. For example, in the measurement of monthly cosmic antiproton fluxes with the Alpha Magnetic Spectrometer [**?**], the background mainly consists of electrons and pions, and the templates are selected using the Ring Imaging Cherenkov detector, which reduces the samples significantly. The antiproton signal, due to the nature of its production mechanism, is rare and has a low signal-to-background ratio. The low statistics of the templates or target data introduce critical challenges: the statistical fluctuations in the noisy template distort the true underlying distributions, resulting in significantly biased antiproton numbers and increasing uncertainties with traditional template fitting methods.

Under low-statistics conditions, due to the low-count bins, the assumption of

Gaussian errors is no longer valid, thus the usage of traditional weighted least squares or χ^2 minimization is inadequate. To correctly account for the low statistics in each bin, one can use the likelihood functions for Poisson distribution and perform the minimization of the binned NLL to obtain the signal fraction [?]:

$$-\ln L = \sum_i \left[f_i(\lambda_s, \lambda_b, \alpha) - n_i \ln f_i(\lambda_s, \lambda_b, \alpha)\right],$$

where n_i is the number of observed events in the ith bin and f_i is the sum of predictions from signal λ_s and background λ_b with the signal fraction α.

However, such likelihood does not incorporate the fluctuations of the template distribution. Barlow and Beeston [?] found the exact likelihood by using the template expectations A_{ji} in each bin i and source j as nuisance parameters:

$$-\ln L = \sum_i (f_i - n_i \ln f_i) + \sum_{j,i} (A_{ji} - a_{ji} \ln A_{ji}),$$

where $f_i = \sum_j p_j A_{ji}$. This method is implemented as TFractionFitter in ROOT [?] and is widely used in high-energy physics experiments searching for new physics [?].

However, the large number of nuisance parameters, which scales as $\mathcal{O}(N_{\text{bins}} \times N_{\text{sources}})$, poses a significant challenge in solving the non-linear equations [?], requiring long computation time and often leading to biased results. Several works [?] propose to use only one nuisance parameter in each bin to approximate the exact likelihood, which are implemented in the iMinuit package [?]. By shrinking the number of nuisance parameters, the computation time is reduced but the resulting biases are not resolved.

With the unbinned likelihood method, the template PDFs are analytical functions or kernel densities. For N events, the joint likelihood can be written as:

$$-\ln L = -\sum_k \sum_i w_i \cdot p_i(x_k),$$

where w_i is the fraction of component i and $p_i(x_k)$ is the PDF of component i evaluated at x_k. The unbinned template fitting can reduce the bias of the fitted signal number by avoiding binning artifacts. However, the unbinned methods are sensitive to the PDF modeling, i.e., the parameters describing the template distributions or the widths of the kernel densities.

In this paper, we propose the CombineFit method, which jointly models the templates and target data through analytical likelihood optimization, aiming to significantly reduce the biases and constrain the template uncertainties.

Methodology: Likelihood for CombineFit

Generally, template fitting is performed by first constructing the signal template PDF $f_{\mathrm{sig}}(\theta_s)$ and background template PDF $f_{\mathrm{bkg}}(\theta_b)$ by minimizing:

1. Signal Template NLL

$$-\ln L_{\mathrm{sig}} = \sum_i \left[\lambda_{s,i}(\theta_s) - n_{s,i} \ln \lambda_{s,i}(\theta_s)\right],$$

where $n_{s,i}$ is the signal counts and $\lambda_{s,i}(\theta_s) = \int f_{\mathrm{sig}}(x;\theta_s)dx$ represents the number of expected signal events in the ith bin.

2. Background Template NLL

$$-\ln L_{\mathrm{bkg}} = \sum_i \left[\lambda_{b,i}(\theta_b) - n_{b,i} \ln \lambda_{b,i}(\theta_b)\right],$$

where $n_{b,i}$ is the background counts and $\lambda_{b,i}(\theta_b) = \int f_{\mathrm{bkg}}(x;\theta_b)dx$, representing the number of expected background events in the ith bin.

Then the template PDFs with θ_s and θ_b are used to fit the experimental data to determine the fraction of the signal α by minimizing the data NLL:

$$-\ln L_{\mathrm{data}} = \sum_i \left[\lambda(\alpha) - n_i \ln \lambda(\alpha)\right],$$

where

$$\lambda(\alpha) = \int \left[\alpha f_{\mathrm{sig}}(x;\theta_s) + (1-\alpha) f_{\mathrm{bkg}}(x;\theta_b)\right] dx,$$

represents the predicted number of events and n_i is the data event counts.

In this process, the minimization of NLL for the templates and data are separated. Namely, the parameters θ_s and θ_b are fixed during the fitting of α. Thus, the uncertainties of the signal and background templates are not taken into account in the fitting of the composition.

In CombineFit, instead of fitting θ_s, θ_b, and α in separate steps, we construct a combined likelihood function that includes both the data and the templates:

$$-\ln L_{\mathrm{total}} = -\ln L_{\mathrm{data}}(\alpha, \theta_s, \theta_b) - \ln L_{\mathrm{sig}}(\theta_s) - \ln L_{\mathrm{bkg}}(\theta_b).$$

By minimizing this combined NLL, the template parameters θ_s and θ_b are constrained not only by their respective template data but also by the target data; i.e., they are determined simultaneously with the signal fraction α.

This is similar to the TFractionFitter method, where the statistical uncertainties of the signal and background templates are taken into account. However, CombineFit differs from TFractionFitter in that it models the template distributions as continuous analytical functions (e.g., exponential, Gaussian mixtures), which limits the number of parameters and smooths the bin-wise fluctuations, reducing the biases and errors.

Similarly, CombineFit can also be implemented in the unbinned likelihood fit by replacing the binned Poisson likelihood functions with unbinned likelihood functions. This further reduces the biases due to binning effects.

Monte Carlo Simulation Test

The MC data generation is based on the data used in the time-variation antiproton flux analysis in the Alpha Magnetic Spectrometer (AMS-02), in which the number of antiproton signals is extracted by performing a template fit on the mass distribution of negatively charged samples [**?**]. The signal templates are constructed with the proton data sample, which is $> 10^4$ more abundant than the antiproton. The background templates are constructed with the pion+electron sample, which has a much smaller statistic (order of 10^2) since we must apply tighter event selection criteria to obtain clean samples. Since the analysis is performed on a monthly basis, the statistics of the data sample are on the order of only 4×10^2, which consists of approximately 200 background events and 200 signal events. The signal and background mass templates are modeled using a function (ExpGaussExp) characterized by a Gaussian core with exponential tails on both sides [**?**], which well represents their physical distributions. These parameterized functions are then used for Toy Monte Carlo Simulations.

1. Parameterization of Template PDFs from Data

As shown in [Figure 1: see original paper], the AMS-02 proton sample and electron+pion sample selected from flight data are fitted using the ExpGaussExp function in Eq.(1):

$$\mathrm{ExpGaussExp}(x) = \begin{cases} \frac{1}{\sigma_L} \exp\left(-\frac{(x-x_0)^2}{2\sigma_L^2}\right) \exp\left(\alpha_L \cdot \frac{x-x_0}{\sigma_L}\right), & x < x_0 - \alpha_L\sigma_L, \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right), & x_0 - \alpha_L\sigma_L \le x \le x_0 + \alpha_R\sigma_R, \\ \frac{1}{\sigma_R} \exp\left(-\frac{(x-x_0)^2}{2\sigma_R^2}\right) \exp\left(-\alpha_R \cdot \frac{x-x_0}{\sigma_R}\right), & x > x_0 + \alpha_R\sigma_R, \end{cases}$$

where: - x_0: The central location parameter of the distribution - σ_L, σ_R: The scale parameters on the left and right sides, respectively, analogous to the standard deviation, determining the spread around the central point - α_L, α_R: The transition parameters on the left and right sides, respectively, specifying where the function transitions from a Gaussian core to exponential tails

For each generated dataset, we perform the fitting to obtain the signal numbers with several methods: the CombineFit described in this paper, the TFractionFitter implemented in the ROOT framework [?], and the recommended method DA [?] that is implemented in the latest iMinuit package. The resulting PDFs are used to generate the Monte Carlo data.

2. Generation of MC Data

For each MC test, the signal templates, background templates, and target (mixture of signal and background) datasets are all independently generated using the fitted PDFs. For each test, we generate: - Signal Template: 5000 events - Background Template: Ranging from 10 to 8000 events - Target Data: 400 events per experiment, composed of a mix of 200 signal events and 200 background events

[Figure 2: see original paper] shows one example of the generated (a) background templates, (b) signal templates, and (c) target data. For each number of events configuration, a total of 10^4 rounds of experiments are simulated.

3. Decomposition of Signal Contribution

To further evaluate the robustness of our methods, we also implemented the unbinned version of CombineFit and compare with RooFit [?] results using RooAbsPdf and RooKeysPdf. The unbinned CombineFit followed the same procedure as described in the binned scenario, but with unbinned datasets for the signal, background, and target data.

Results

A total of 10^4 rounds of MC simulations for each set of event number configuration were performed, and different template fitting methods were applied to obtain the signal numbers. The distribution of 10^4 fitted signal numbers for each method was obtained and follows a Gaussian distribution. A Gaussian fit is performed, and the difference between the Gaussian mean and the number of generated signal events N_{gen} represents the bias, δ, while the Gaussian σ represents the uncertainty of the method. The root mean squared error represents the total error relative to the number of generated signal events: $\sqrt{\delta^2 + \sigma^2}$.

1. Binned Scenario

As shown in Figure 3: see original paper, the bias in the signal numbers as a function of the statistics of the background template is presented for different methods. As expected, with all methods, the bias decreases as the number of events in the background template increases. The results show that TFractionFitter and DA methods exhibit large bias under low-template statistics (30%), while the CombineFit method resulted in minimal bias that’ s around 2% with

10 events in the background template and drops to 0.5% with 100 events in the background template, while the other methods still have more than 2% biases.

Figure 3: see original paper shows the statistical uncertainties of the methods. With 10 events in the background template, CombineFit shows an uncertainty of 5.8%, compared to about 9% for DA and 10% for TFractionFitter. With more than 100 events in the background template, all methods show less than 5% uncertainties.

Figure 3: see original paper shows the total error of the methods. With less than 100 events in the background templates, since CombineFit has both smaller bias and sigma, the total error is also smaller. As the number of events increases, all methods perform similarly.

2. Unbinned Scenario

As shown in Figure 4: see original paper, similar to the binned scenario, as the number of background template events increases, the estimated signal yields converge toward the true signal value, resulting in diminishing biases. Particularly, the results with unbinned RooFit function fitting show larger biases than the binned CombineFit method, demonstrating that the separated minimization between the templates and the target data causes larger biases. The unbinned CombineFit method further reduced the bias to below 1% with 10 events in the background template and 0.25% bias with 100 events in the background template. The RooFit KDE method shows faster decreasing biases; however, it exhibits a systematic shift to negative biases at larger numbers of events, demonstrating the limitation of this method.

Figure 4: see original paper shows the statistical uncertainties of these methods. With 10 events in the background template, the unbinned RooFit method showed larger uncertainties of 12% compared to the binned methods. The unbinned and binned versions of CombineFit have similar uncertainties. With more than 100 events, all methods show below 5% uncertainties, consistent with the binned methods.

Figure 4: see original paper shows the total error of the unbinned methods. With 10 events in the background template, since the unbinned CombineFit has smaller bias, the total error is also smaller. With more than 100 events, all methods perform similarly.

3. Robustness Across Configurations

Tables 1-4 demonstrate CombineFit' s robust performance under varied signal/background composition in the target data and different numbers of events in the templates.

and show the comparison of the total errors between different methods for $N_b : N_s = 200 : 200$, $N_b : N_s = 50 : 200$, or $N_b : N_s = 200 : 50$ in the target data, with different numbers of events in the background or signal templates.

In , the numbers of signal and background are the same in the target data, and both template statistics are varied. In , only the template corresponding to the component with fewer events in the target data is varied. When one template statistic is varied, the other is fixed at 8000 to independently check their contributions.

As shown in the tables, in low-statistics scenarios, the CombineFit methods result in the smallest errors compared to the other methods. In most cases, the binned CombineFit method shows smaller errors compared to the unbinned RooFit methods.

and show the comparison of errors between CombineFit and DA when the numbers of events in both the signal and background templates are small, for $N_b : N_s = 200 : 200$ in the target data. In the binned scenario (), CombineFit consistently outperforms DA. In the unbinned scenario (), unbinned CombineFit shows smaller errors compared to RooFit(func) when both templates have limited statistics.

In most cases, CombineFit shows the smallest total errors across different configurations.

Conclusions

A new approach for signal number estimation in low-statistics measurements, CombineFit, that uses analytical functions to simultaneously fit the templates and data by combining the likelihood of data and every component is proposed and tested. CombineFit improves fitting stability and accuracy compared to traditional methods by reducing bias from \$ \$30% to <2% and lowering errors by 35% (with 10 events in the background template), and further reduces the bias to 1% in the unbinned scenario while maintaining accuracy across different configurations.

This method has been successfully applied in the time-variation analysis of cosmic antiproton flux with AMS, providing unique information for understanding particle transport in the Solar System [**?**]. This method can also be applied to other experiments in search of new physics.

References

[1] F. James, *Statistical Methods in Experimental Physics* (2nd Edition), World Scientific Publishing Company, 2006. ISBN: 9789813101845. Available at: https://books.google.com.hk/books?id=c0A8DQAAQBAJ.

[2] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370-384, 1972. Available at: http://www.jstor.org/stable/2344614.

[3] M. Aguilar, G. Ambrosi, H. Anderson, et al. (AMS Collaboration), "Antiprotons and Elementary Particles over a Solar Cycle: Results from the Alpha Mag-

netic Spectrometer" , *Physical Review Letters*, vol. 134, 051002, 2025. Available at: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.134.051002

[4] S. Baker and R. Cousins, "Clarification of the use of CHI-square and likelihood functions in fits to histograms", *Nuclear Inst. and Methods in Physics Research*, 221, (1984) 437-442. Available at: https://www.sciencedirect.com/science/article/pii/0167508784900164

[5] R. Barlow and C. Beeston, "Fitting using finite Monte Carlo samples" , *Computer Physics Communications* 77 (1993) 219–228. Available at: https://www.sciencedirect.com/science/article/pii/001046559390005W

[6] R. Brun and F. Rademakers, "ROOT —An object oriented data analysis framework" , *Nuclear Inst. and Methods in Physics Research*, A (1997) 389. Available at: https://doi.org/10.1016/S0168-9002(97)00048-X

[7] M. Aguilar, L. Ali Cavasonza, G. Ambrosi, et al. (AMS Collaboration), "The Alpha Magnetic Spectrometer (AMS) on the International Space Station: Part II —Results from the First Seven Years" , *Physics Reports*, vol. 894, pp. 1-116, 2021. Available at: https://www.sciencedirect.com/science/article/pii/S0370157320303434.

[8] G. Aad, B. Abbott, K. Abeling, et al. (ATLAS Collaboration), "Precise measurements of W- and Z-boson transverse momentum spectra with the ATLAS detector using pp collisions at $\sqrt{s} = 5.02$ TeV and 13 TeV" , *Eur.Phys.J.C* 84, 1126 (2024). Available at: https://doi.org/10.1140/epjc/s10052-024-13414-0

[9] G. Aad et al. (ATLAS Collaboration), "Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment", *Phys. Rev. D* 101 (2020) no.1, 012002. Available at: https://doi.org/10.1103/PhysRevD.101.012002

[10] A. M. Sirunyan et al. (CMS Collaboration), "Combined measurements of Higgs boson couplings in proton–proton collisions at $\sqrt{s} = 13$ TeV" , *Eur. Phys. J. C* 79, no.5, 421 (2019). Available at: https://link.springer.com/article/10.1140/epjc/s10052-019-6909-y

[11] M. Aguilar et al. (AMS Collaboration), "Antiproton Flux, Antiproton-to-Proton Flux Ratio, and Properties of Elementary Particle Fluxes in Primary Cosmic Rays Measured with the Alpha Magnetic Spectrometer on the International Space Station" , *Phys. Rev. Lett.* 117, 091103 (2016). Available at: https://doi.org/10.1103/PhysRevLett.117.091103

[12] J.S. Conway, "Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra", arXiv:1103.0354. Available at: http://arxiv.org/abs/1103.0354

[13] H. Dembinski, M. Schmelling and R. Waldi, "A new maximum-likelihood method for template fits" , *Eur. Phys. J. C* (2022) 82:1043. Available at: https://doi.org/10.1140/epjc/s10052-022-11019-z

[14] C. A. Argüelles, A. Schneider and T. Yuan, "A binned likelihood for stochastic models" , *J. High Energ. Phys.* (2019) 2019:30. Available at: https://doi.org/10.1007/JHEP06(2019)030

[15] F. James and M. Roos, "Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations" , *Comput. Phys. Commun.* (1975) 10:343–367. Available at: https://doi.org/10.1016/0010-4655(75)90039-9

[16] S. Das, "A simple alternative to the Crystal Ball function", arXiv:1603.08591v1. Available at: https://doi.org/10.48550/arXiv.1603.08591

[17] W. Verkerke and D. Kirkby, "The RooFit toolkit for data modeling" , arXiv:physics/0306116. Available at: https://arxiv.org/abs/physics/0306116

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.