Implementation of the likelihood-based ABCD method for background estimation and hypothesis testing with pyhf

Mason Proffitt

UNIVERSITY of WASHINGTON

Introduction

- An important component of any search for new physics signals is the ability to reliably estimate backgrounds and their uncertainties
- The ABCD method is a data-driven background estimation procedure that consists of dividing data into multiple regions by simple cuts
- `pyhf` is a software package that provides the functionality available in `HistFactory`, a statistical toolkit within the ROOT software framework, as a standalone Python module

ABCD method

- The ABCD method begins by forming a 2D plane from two uncorrelated variables
- Setting a cut value in each of these variables divides the plane into four regions: A, B, C, and D
- These regions are defined such that a target signal will be concentrated in region A, which is opposite to region D in both cuts
- In the simplest version of the ABCD method, the estimated number of background events in region A, n_A, is equal to $n_B n_C / n_D$
- In the general case, all regions may be non-negligibly contaminated by signal, so the likelihood-based ABCD method involves a simultaneous fit of signal and background to the observed data
- The expected number of signal events in region X is

$$n_X^{\text{signal}} = \frac{\epsilon_X}{\epsilon_A}\mu$$

  where ϵ_X is the signal efficiency of region X and μ is the signal strength, which is the parameter of interest (POI).

- The expected number of background events in each region is given by

$$n_A^{\text{bkg}} = \mu_b \qquad n_B^{\text{bkg}} = \tau_B \mu_b$$
$$n_C^{\text{bkg}} = \tau_C \mu_b \qquad n_D^{\text{bkg}} = \tau_B \tau_C \mu_b$$

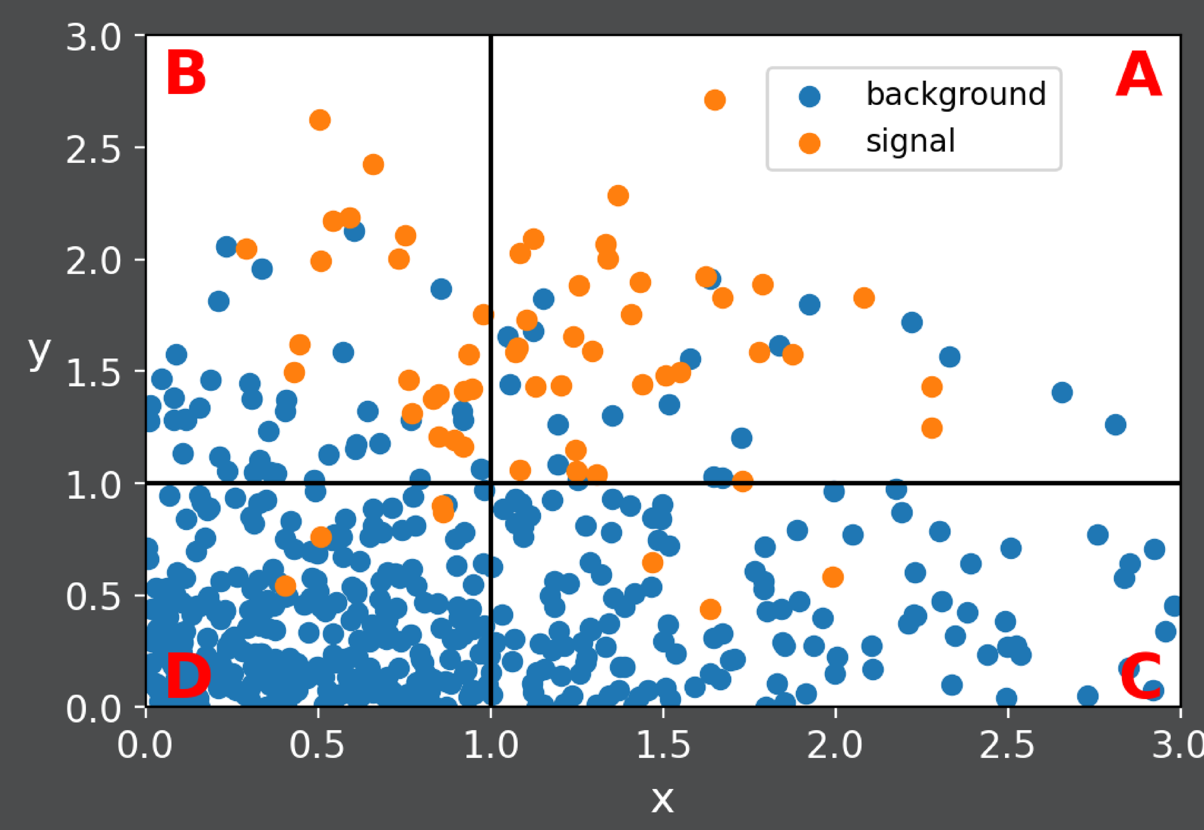
  where μ_b is a background normalization and τ_B and τ_C are nuisance parameters that enforce the relationship $n_A^{\text{bkg}} = n_B^{\text{bkg}} n_C^{\text{bkg}} / n_D^{\text{bkg}}$

Implementation

- The likelihood-based ABCD method has been implented using `pyhf`, with the code available in a public GitHub repository [2]
- The likelihood function is implicitly specified by a `Model` object, which consists of `channels`
- Each of the four ABCD regions is represented by a `channel`, which in turn contains two `samples`, `signal` and `background`, representing the expected yield of each in the corresponding region
- The profile likelihood is calculated via `pyhf` library functions such as `fixed_poi_fit` and `twice_nll`
- An `ABCD` class is used to conveniently package these functions

Example

- To demonstrate this implementation, toy signal and background distributions were randomly generated for a simplified analysis
- The background follows an exponential distribution in the variables x and y, while the signal is sampled from a 2D normal distribution with its mean inside region A
- 500 background events and 60 signal events were used to emulate observed results, as shown below:



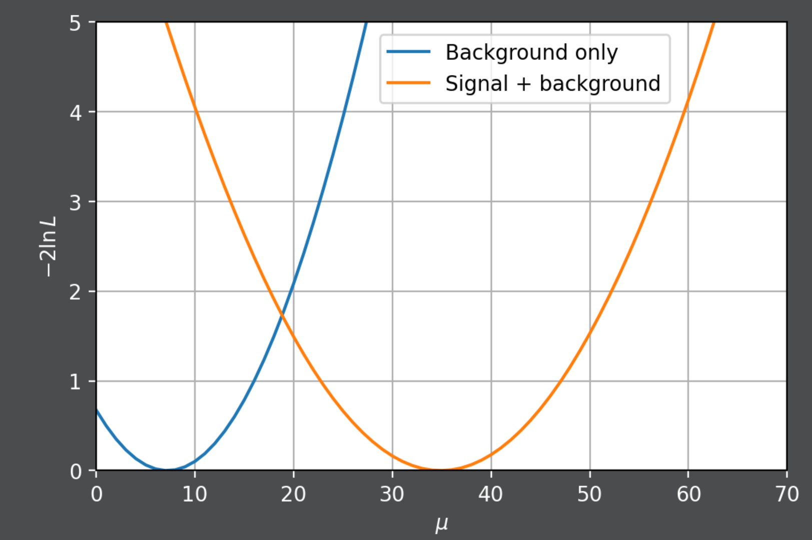
- A much larger sample (100,000) of signal events was used to evaluate the signal efficiencies of each region, as would be done by Monte Carlo signal samples in a real analysis

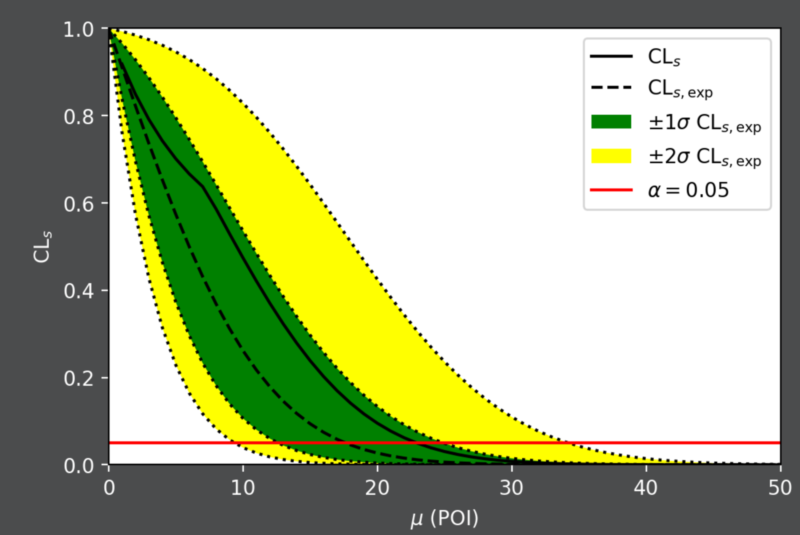
Results

- This implementation was tested for the example under two different ground truth scenarios: background only and signal + background
- The region A results are shown below and in a GitHub repository [3]:

	Background only	Signal + background
Expected events (prefit)	22 ± 4	33 ± 5
p-value (prefit)	0.21	0.0017
Observed events	27	59
Expected events (postfit)	24 ± 3	46 ± 5
p-value (postfit)	0.21	0.0029

- The "background only" test results in p-values that indicate consistency with the null hypothesis as expected
- The small p-values for the "signal + background" test shows that there is a statistically significant excess of events in region A
- Shown below are the negative log-likelihood curves of the signal strength for the two tests and CL_s [4] for the "background only" test





- Based on the results above, a 95% CL upper limit of 23 can be set on the signal strength μ in the "background only" test

Conclusion

- An implementation of the likelihood-based ABCD method was developed using `pyhf`
- This was demonstrated in an example analysis of toy signal and background distributions generated by Monte Carlo
- This work can be used to provide background estimation and hypothesis testing for physics analyses that use the ABCD method

References

[1] Lukas Heinrich, Matthew Feickert, & Giordon Stark. (2021). pyhf: v0.6.3 (0.6.3). Zenodo. `https://doi.org/10.5281/zenodo.5426790`

[2] Mason Proffitt. (2021). `https://github.com/masonproffitt/abcd-pyhf`

[3] Mason Proffitt. (2021). `https://github.com/masonproffitt/abcd-pyhf-examples`

[4] Read, A. L. (2002). "Presentation of search results: The CL(s) technique". *Journal of Physics G: Nuclear and Particle Physics.* **28** (10): 2693–2704.