Generation and Analysis of Toy Particle Decays

Jonathon Howland

Motivation

Want to quantify effects of different parameters (number of events, fraction of signal events) on results such as fit likelihood

See how using different observables affects conclusions

Overall goal: For a data set with a given number of (supposed) signal events and total events, determine the certainty with which the statement that there are indeed signal events in the data set can be made.

Outline of Process

1.) Generate events using RestFrames and ROOT

Generate both signal and background events

Each event has the observables: reconstructed mass and decay angle

2.) To analyze events, we use the particles’ masses and the cosines of their decay angles (I frequently refer to them simply as decay angles)

3.) Use RooFit to create models with both signal and background as well as only background

4.) Use these models to fit datasets

Generating Events

Used RestFrames to generate simulated particle decays

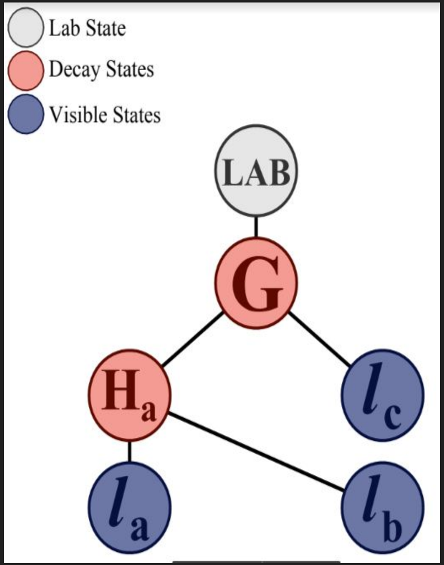
Decay trees - show the initial and final particles, as well as the path

Generator trees - show how the decay happened

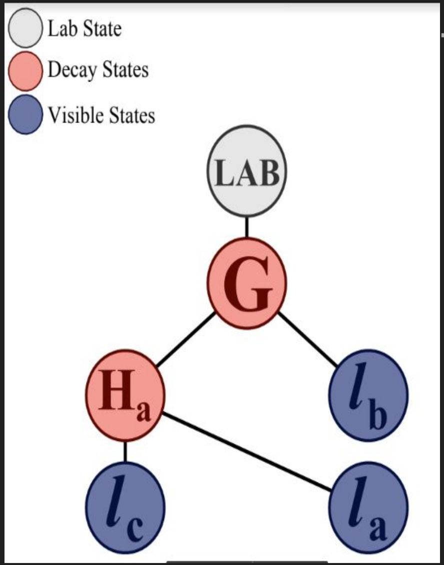
Reconstruction tree - shows how the event was analyzed

Declared the generator tree that matches the reconstruction tree as “signal” and the other two events as background

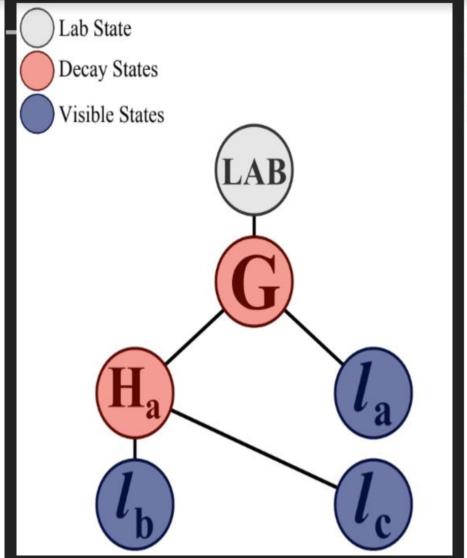
RestFrame Decay Trees



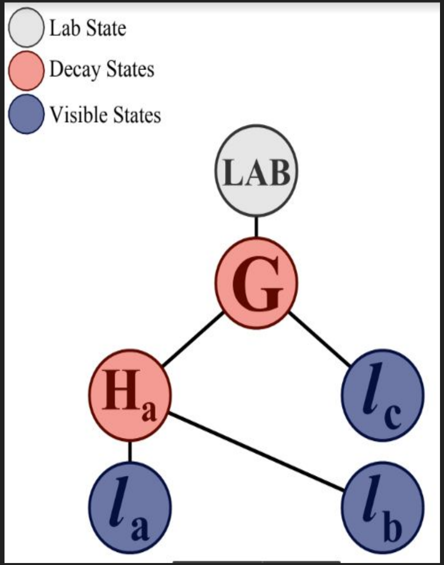
Generator Tree 1 (**signal**)



Generator Tree 2 (**background**)



Generator Tree 3 (**background**)



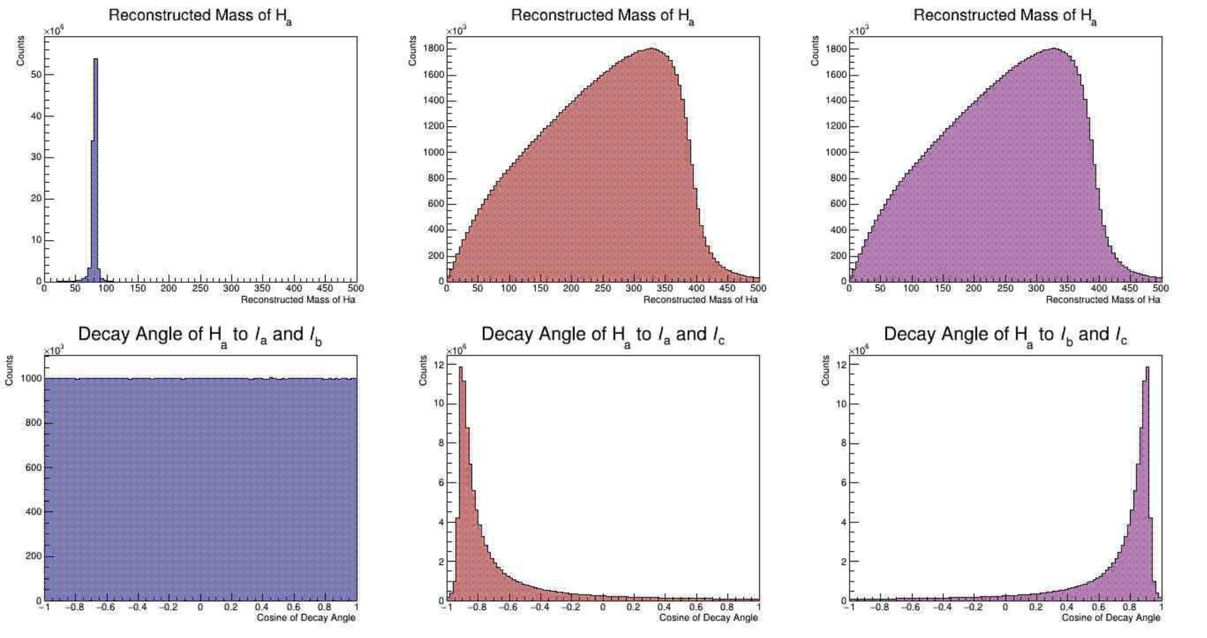
Reconstruction Tree

1D Histograms

Leftmost set of graphs shows decays from generator tree 1 (signal)

Other two sets of graphs show background (generator trees 2 and 3)

Mass and decay angle are of H_a in the decay trees

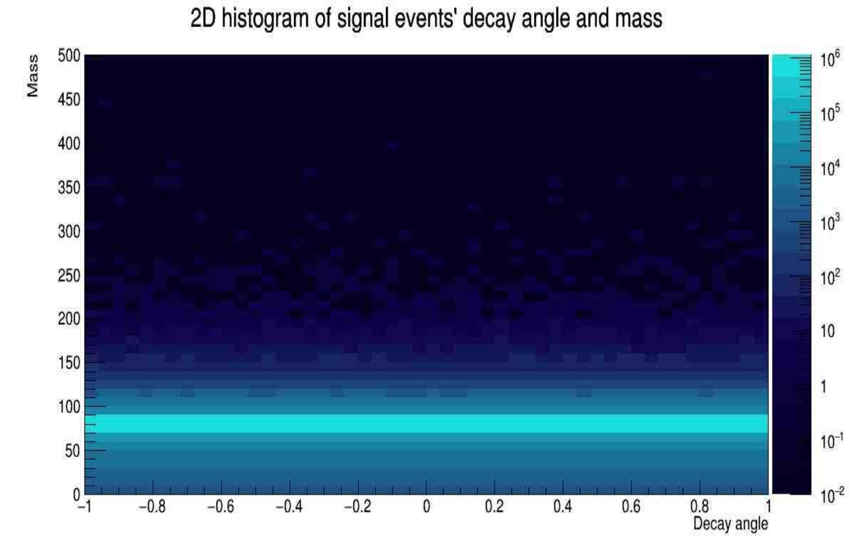


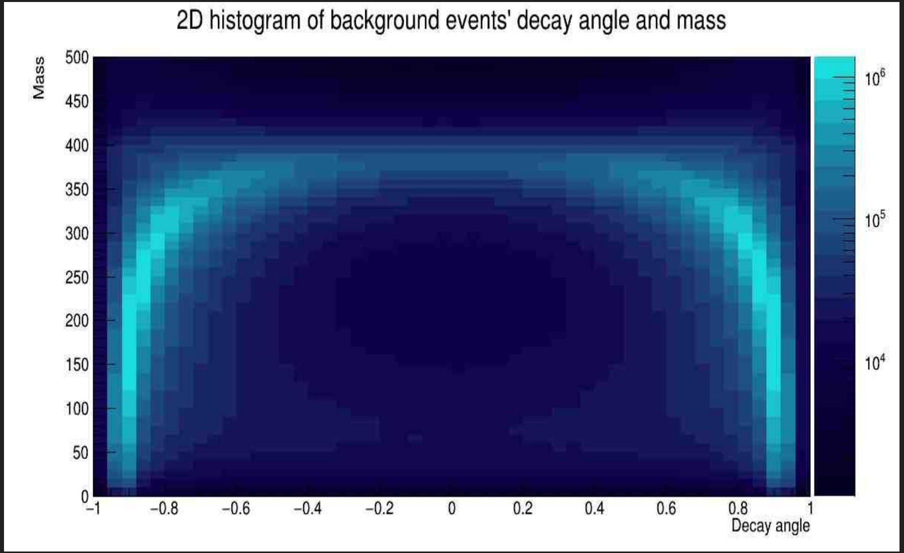
Signal

Background

2D Histograms

Histograms showing the 2D distribution of decay angle and mass with a log(Z) color scale.





Creating and Using Numeric PDFs

Used the data shown in the histograms to create numeric PDFs

- 2 observables - mass and decay angle from Reconstruction Tree
  - **For each process:**
    - Created a 1D PDF for each observable, and a 2D PDF for each pair of decay angle and mass
- Created models of experimental observations by adding PDFs together
- Generated more datasets by sampling from PDFs

Building Models and Fitting

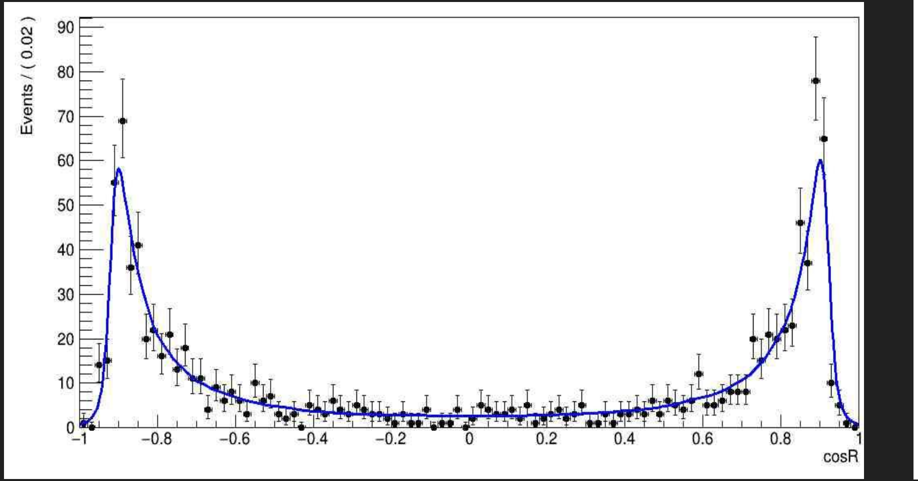
Created models for decay angle only, mass only, and both together

> Added PDFs together using RooFit, with normalizations corresponding to the number of events from each PDF

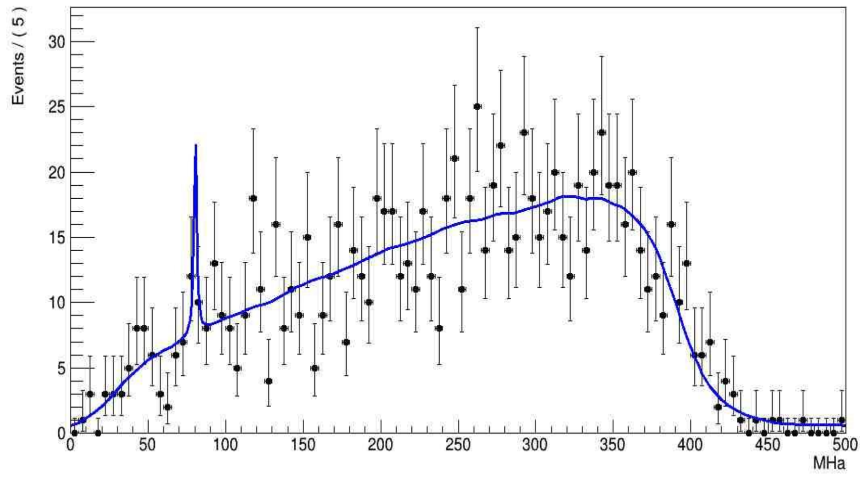
Fit new datasets (created from PDFs) with models, with normalizations as floating parameters

Goal of fit: determine how many events came from each distribution

1D Fits (10 signal and 990 background events)



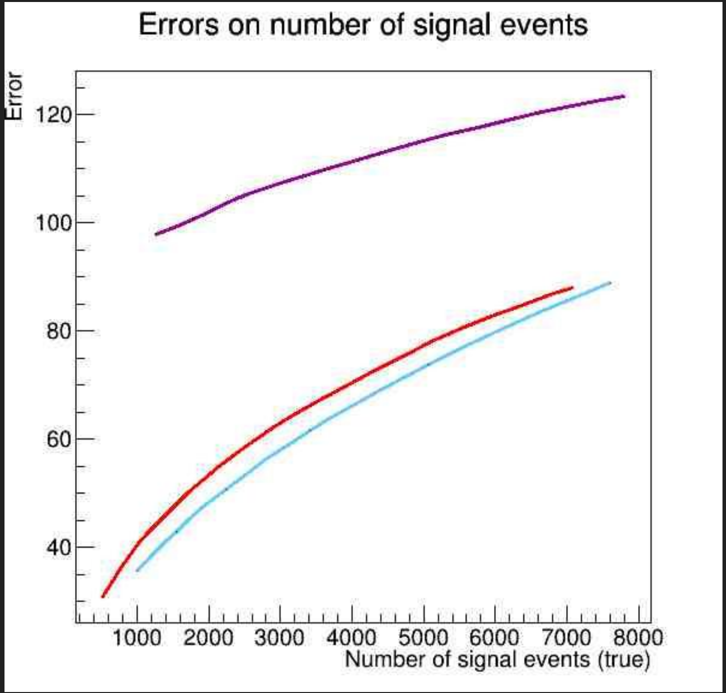
Fit with cosine of decay angle

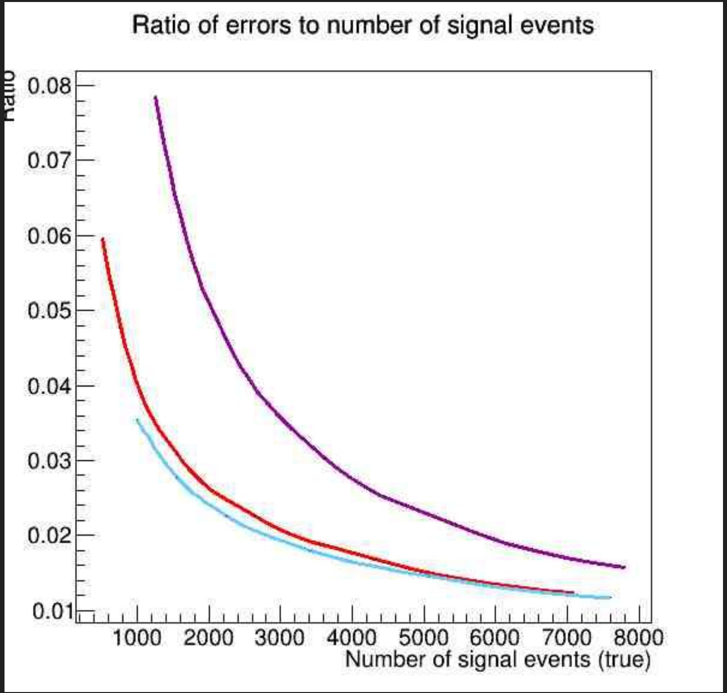


Fit with mass

Mass, Decay Angle, or Both?

Legend
Using only cos
Using only mass
Using both cos and mass





Likelihood

A measure of how well a model describes data

$$L = \frac{e^{-\sum_c^{components} N_c}}{N_{events}} \prod_e^{events} \sum_c^{components} N_c P_c(\vec{x})$$

Here, the N_c's are the floating parameters (numbers of each component), while the P_c's are the probabilities of the vectors of observables.

The summation is over all components (3, in this example) and the product is over all events.

The additional exponential piece is because this is an “extended” likelihood, since all three N_c's are floating in the fit

Can use ratios of likelihoods of two models to decide which model better suits the data

Log Likelihood Ratio (LLR) and Sigmas

Taking the logarithm changes the product in the likelihood to a summation

LLR = log(Likelihood Ratio)

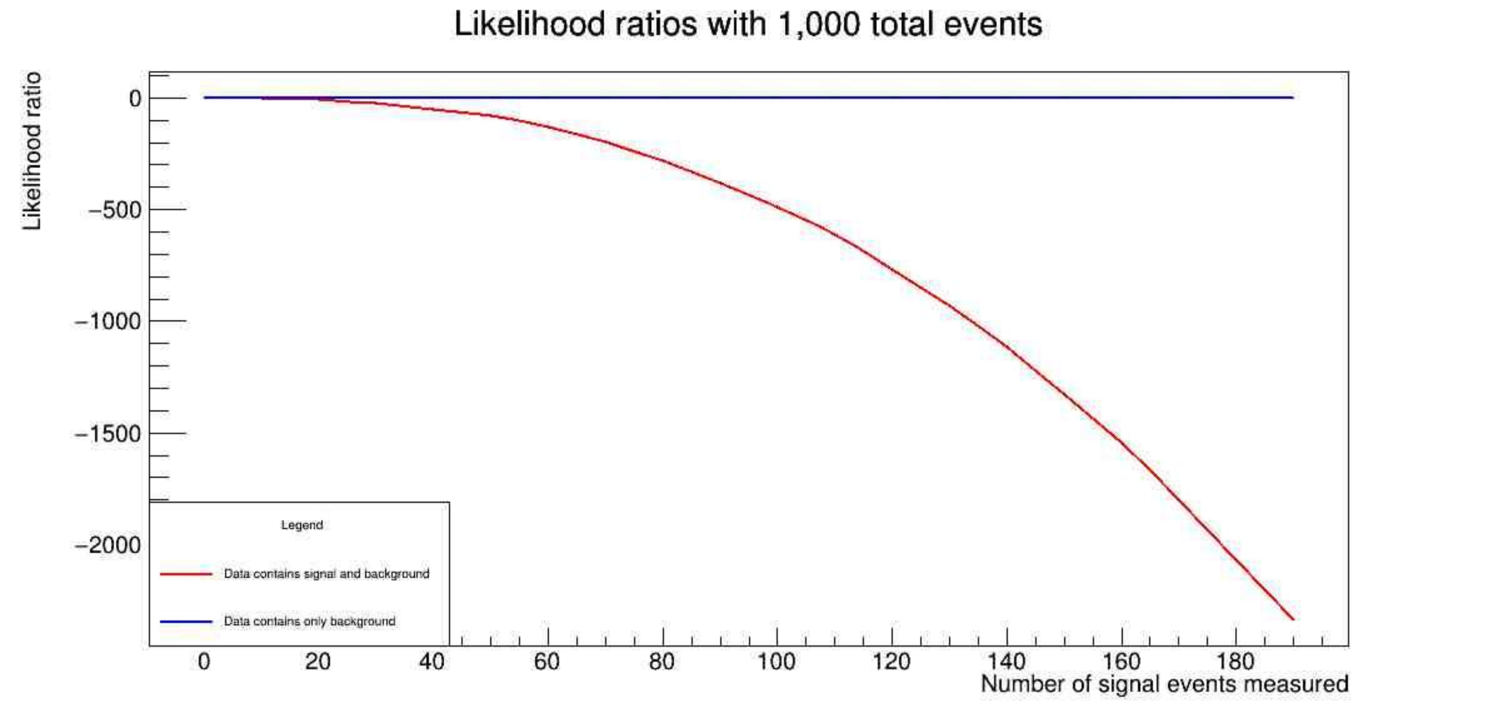
Can convert log likelihood ratios to sigmas

When one hypothesis is a subset of the other

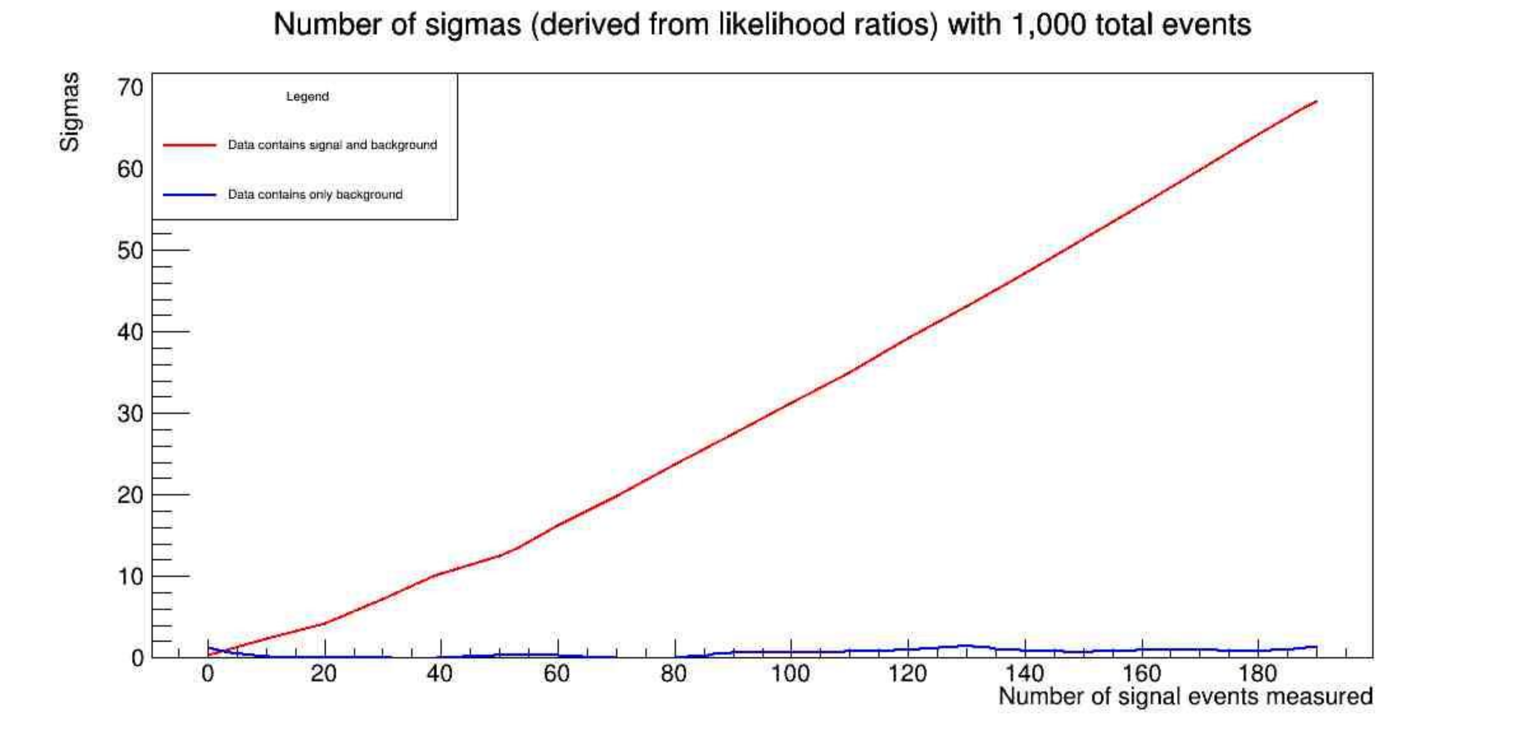
One has an additional parameter (in this case, signal)

$$\sigma = \sqrt{-2 \times LLR}$$

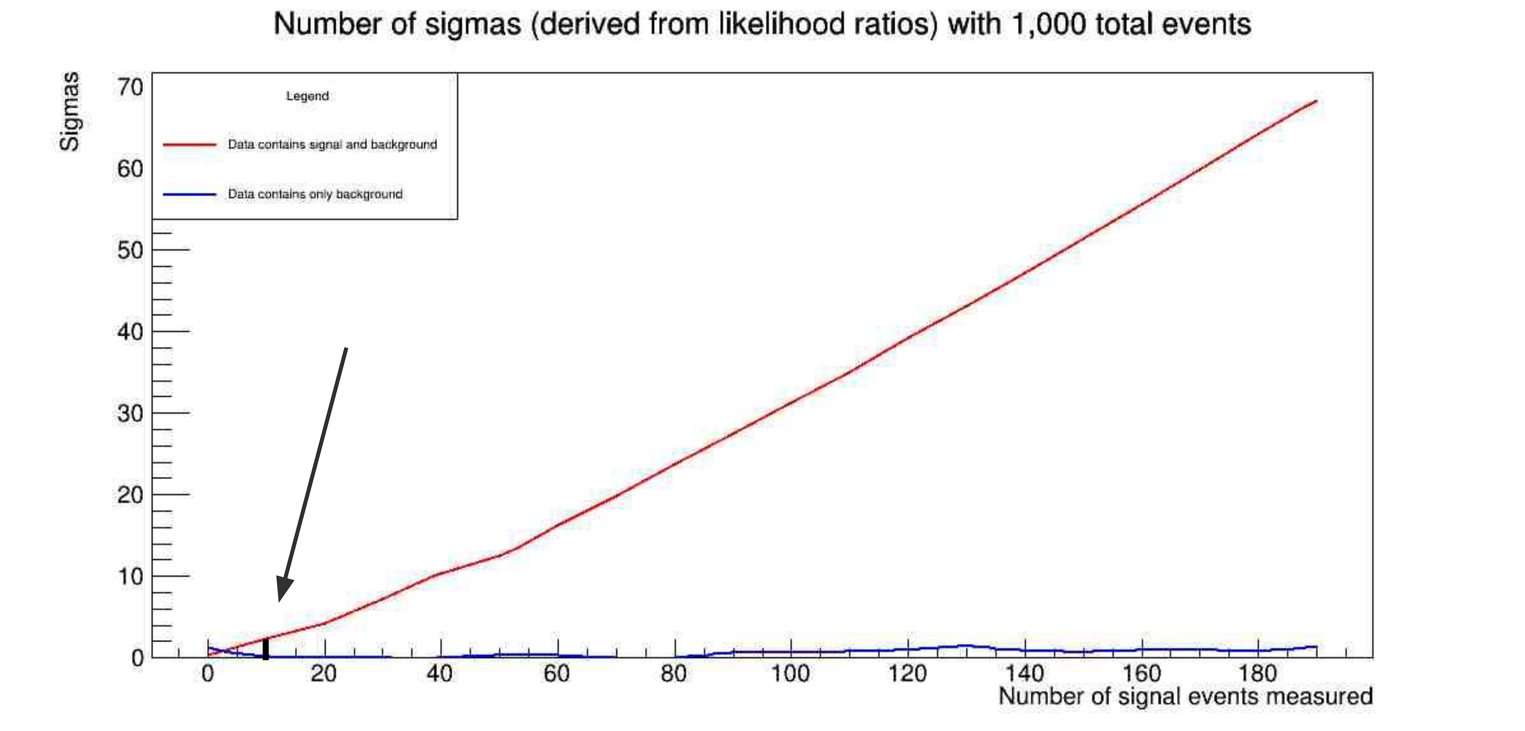
Signal vs Background Only



Signal vs Background Only (Sigmas)



Signal vs Background Only (Sigmas)



An Interesting Region

Find an area with low enough sigmas to be interesting

Run multiple simulations using that configuration

- Generate independent toy datasets for each simulation

The point on the previous line is one fit

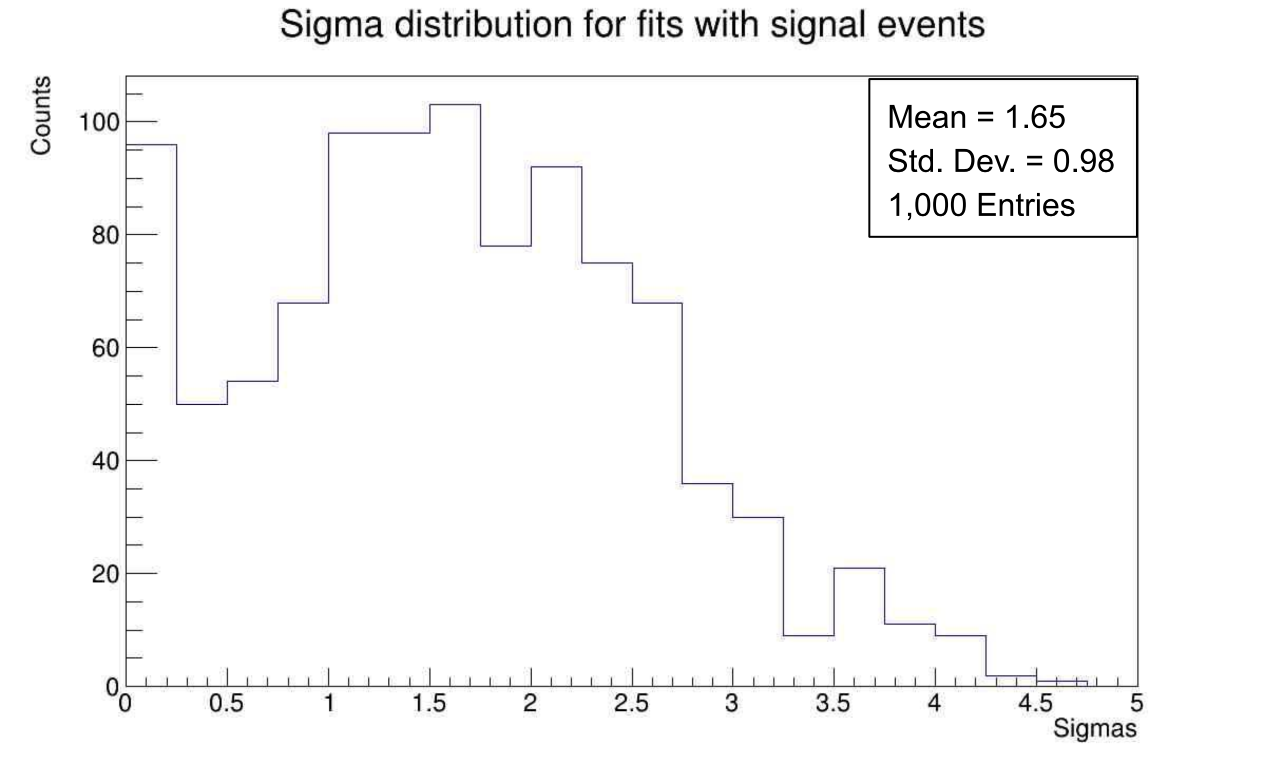
- Now, do that fit many times with different toy datasets

An Interesting Region

On average, 10 signal events and 990 background events in each toy dataset

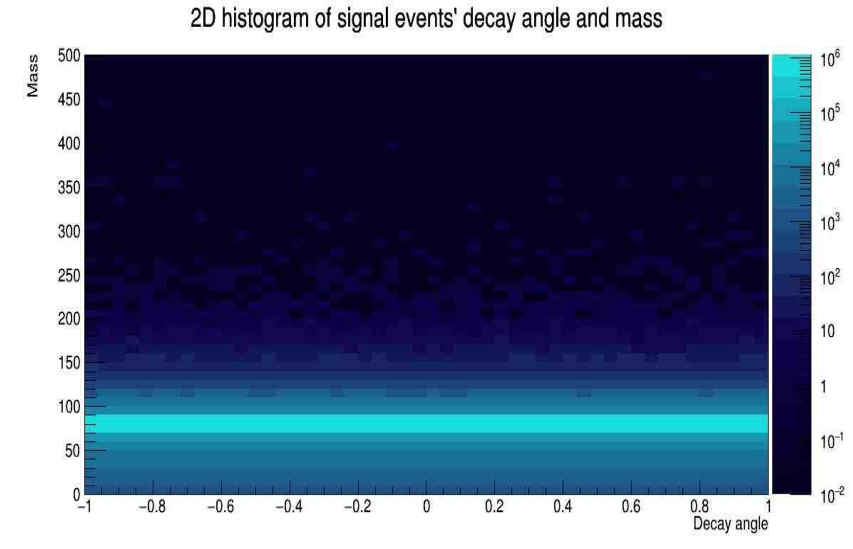
1,000 independent toy datasets

Fit each toy dataset with S+B model, and a B-only model

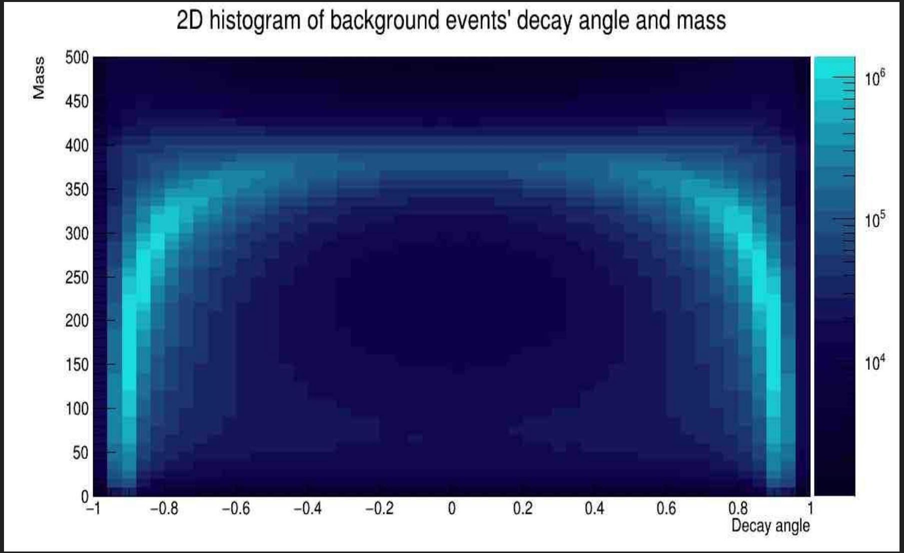


Explaining the Previous Distribution

Histograms showing the 2D distribution of decay angle and mass with a log(Z) color scale.



2D histogram of signal events' decay angle and mass



2D histogram of background events' decay angle and mass

Distributions of Distributions (of Sigmas)

Varied either fraction of signal events or number of total events

- Kept the other constant

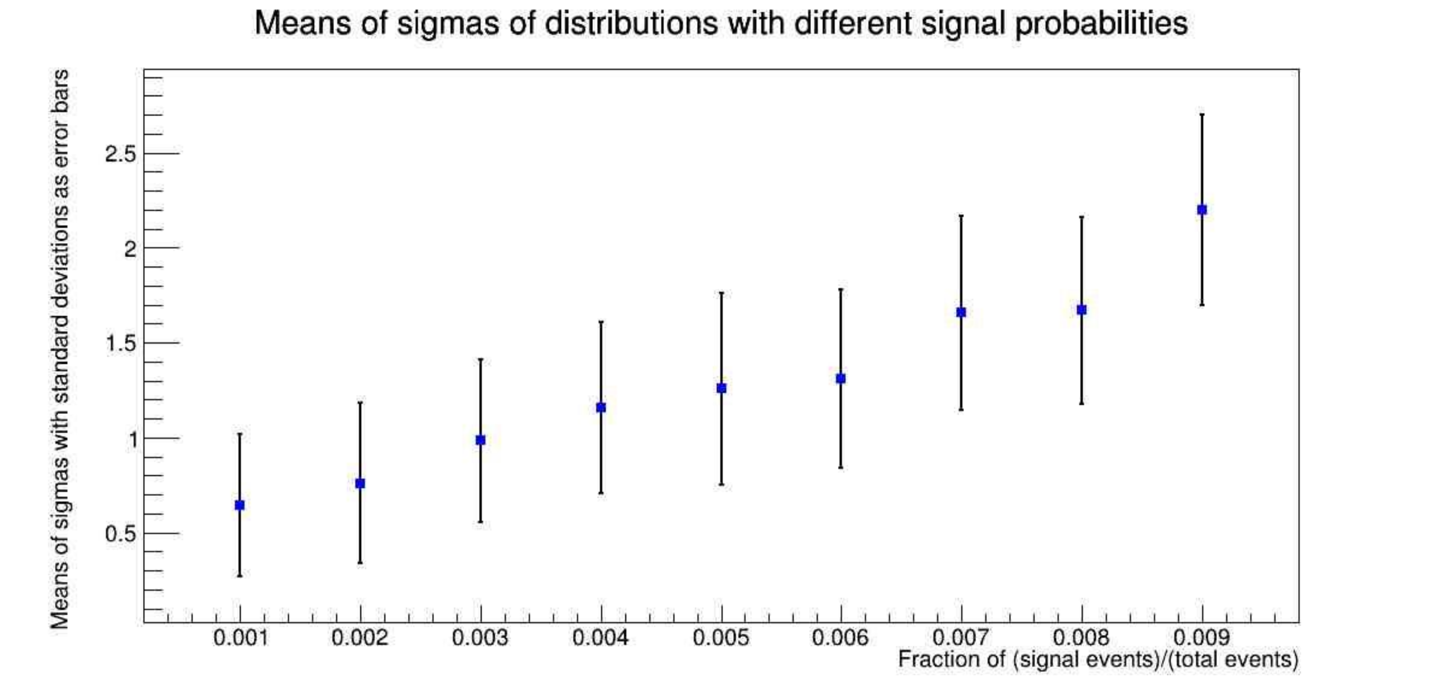
Ran 100 fits at each point

- Similar to previous histogram

Plotted results

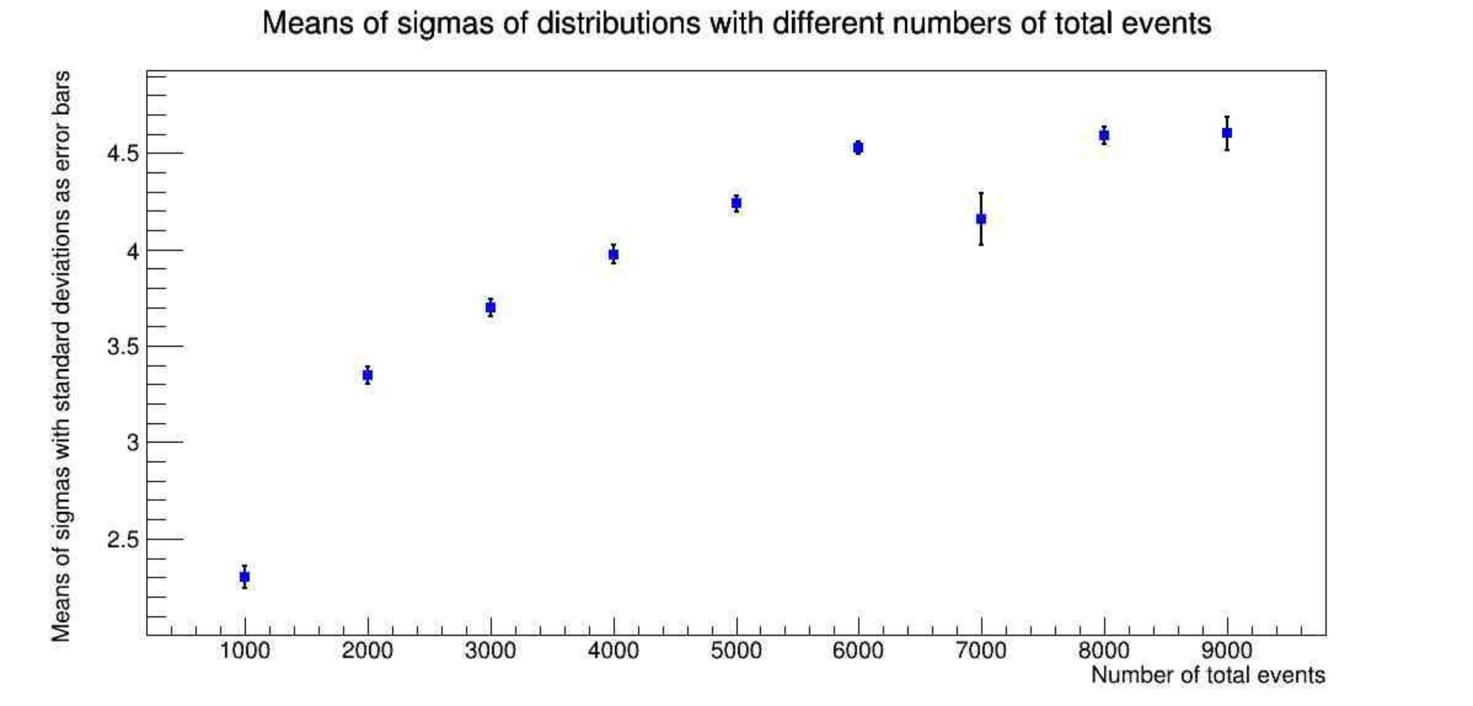
Distributions (N = 1,000)

Means of sigmas increase ~linearly with probability of signal events



Distributions (p = 0.01)

Means of sigmas increase ~as the square root of total events



Conclusion

Using two variables results in a better fit than using one variable

Fits with both had ~½ the error of fits with only decay angle, and were slightly better than with only mass

Distinctly different mass distributions lead to very high sigma values even with very low numbers of signal events

Roughly 5 sigmas with around 20 signal and 980 background events

Mean sigma values increase as either the fraction of signal events or the total number of events increase

Increases ~linearly with fraction and ~as the square root of total number

Conclusion

Same principles as in experiments with real data instead on toy data

Real data is more complicated

Models here are exactly the same shapes as datasets

No experimental errors due to equipment or lost data

Makes it easier to analyze the process due to less complication

References

ROOT - root.cern.ch

RestFrames - restframes.com