Imperial College London

The discrete profiling method:

Handling uncertainties in background shapes

Nicholas Wardle

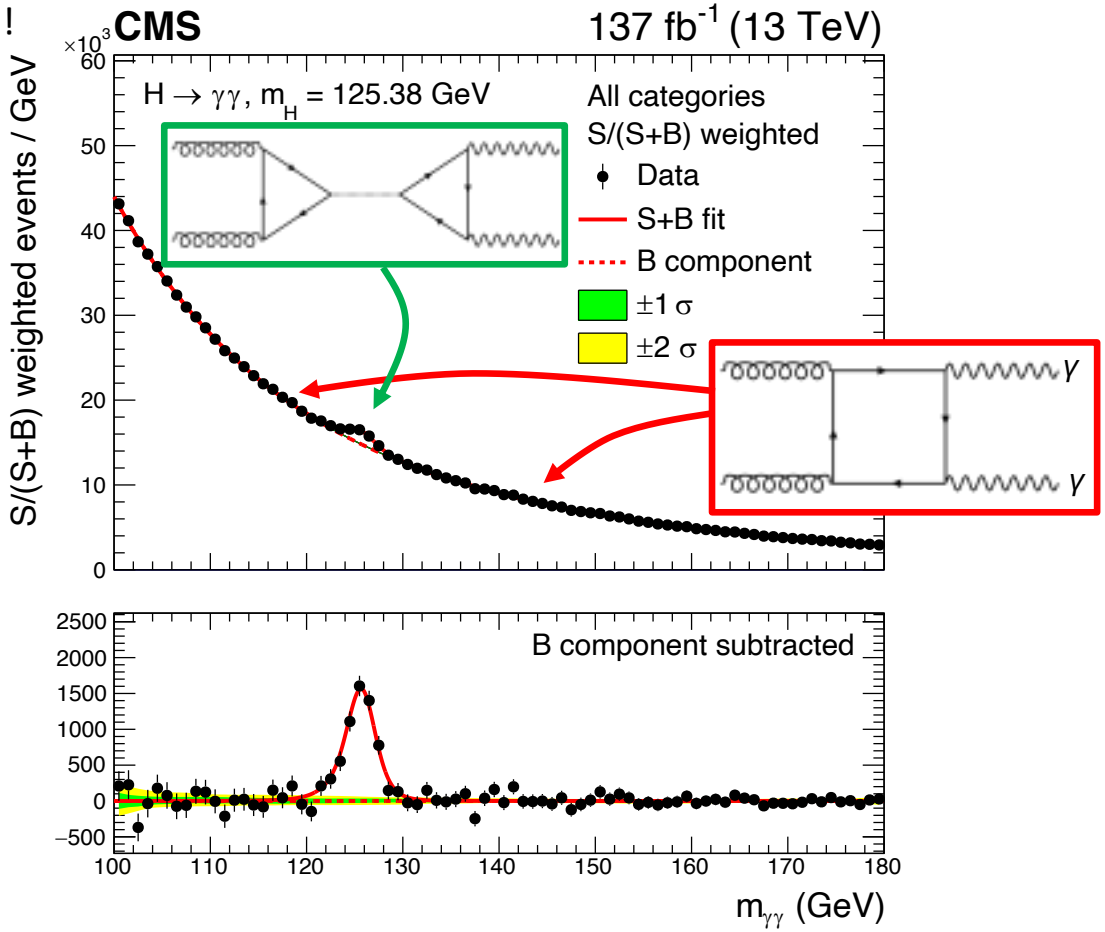
STAMPS@CMU – Feb 18 2022

The model choice problem

- Often we want to extract some physical parameter from the parameter(s) of interest (POIs) from our dataset.
  - The signal yield or branching fraction
  - Decay time
  - Mass, width, angular parameters etc.
- Usually have other parameters we don't know but also don't care about - nuisance parameters
  - Size and shape of backgrounds
  - Signal fractions etc...
- Often we don't know the true distribution of some components
  - Background compositions
  - Acceptance effects
  - Kinematic turn on's due to trigger efficiencies
- A common strategy is to fit some functional form to the data so as not to rely on simulation.
- Other strategies exist (i.e non-parameteric density estimators), not covered here.

The model choice problem

In the example of the Higgs boson decaying to photons analysis (and many other cases), we don't care about the **background** but we need to model it well to extract the **signal**!

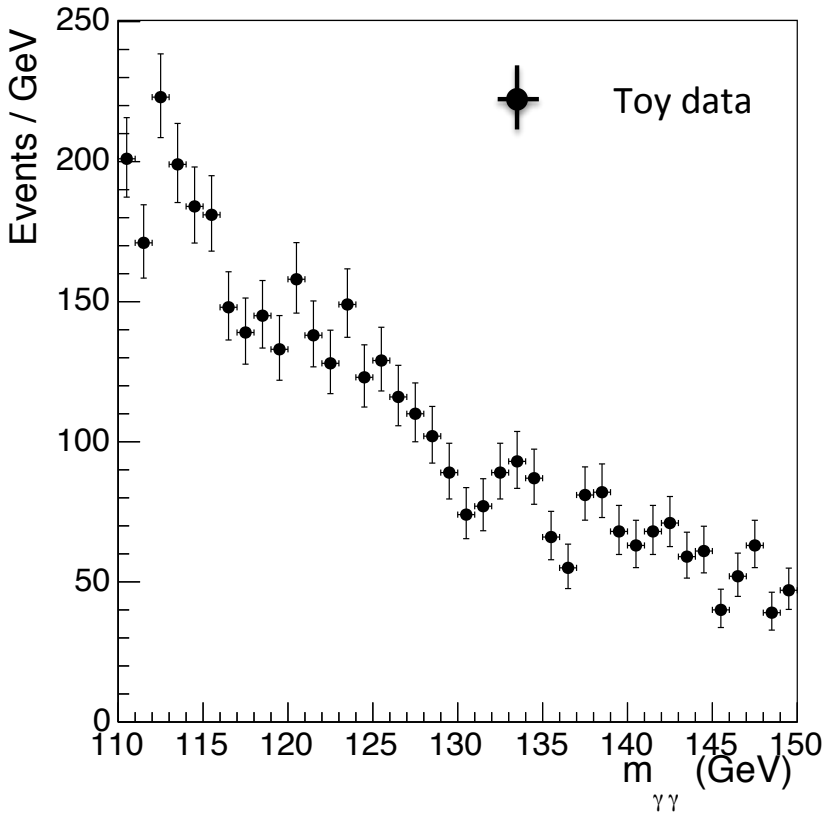


The discrete profiling method:

The model choice problem

Common problem in analysis….

Want to fit a distribution in data to model a smooth background (eg bump-hunt, H→γγ, H→μμ …). Maybe we don't know what that shape should be (insufficient MC, trigger turn on, selection bias …)

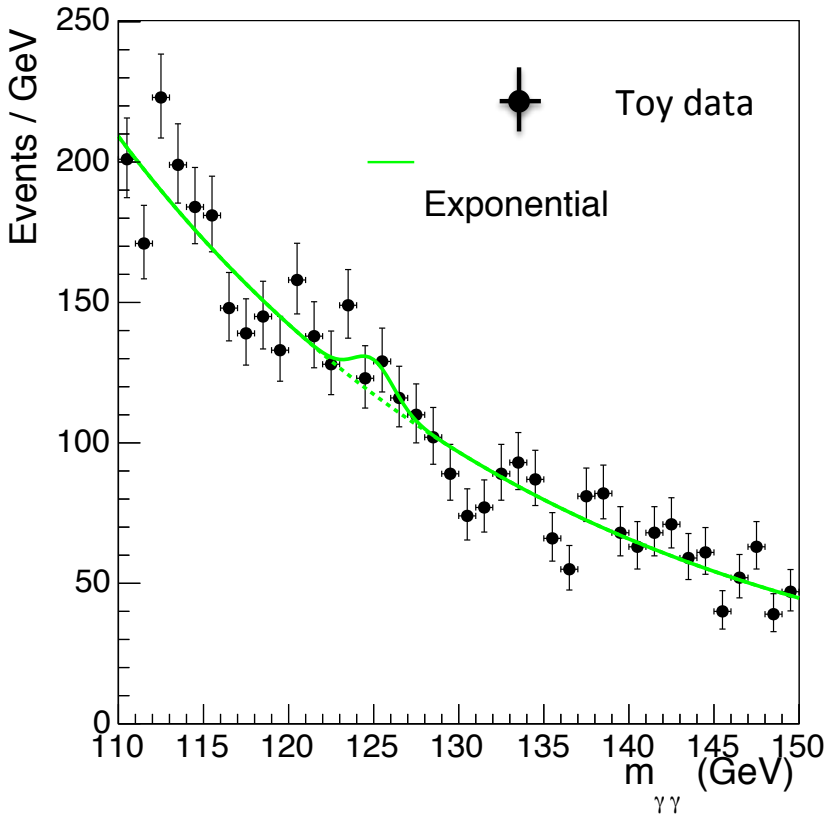


The discrete profiling method:

The model choice problem

Looks like a falling spectrum, why not try an exponential with a Gaussian for the signal?

….fits well to the data at least by eye

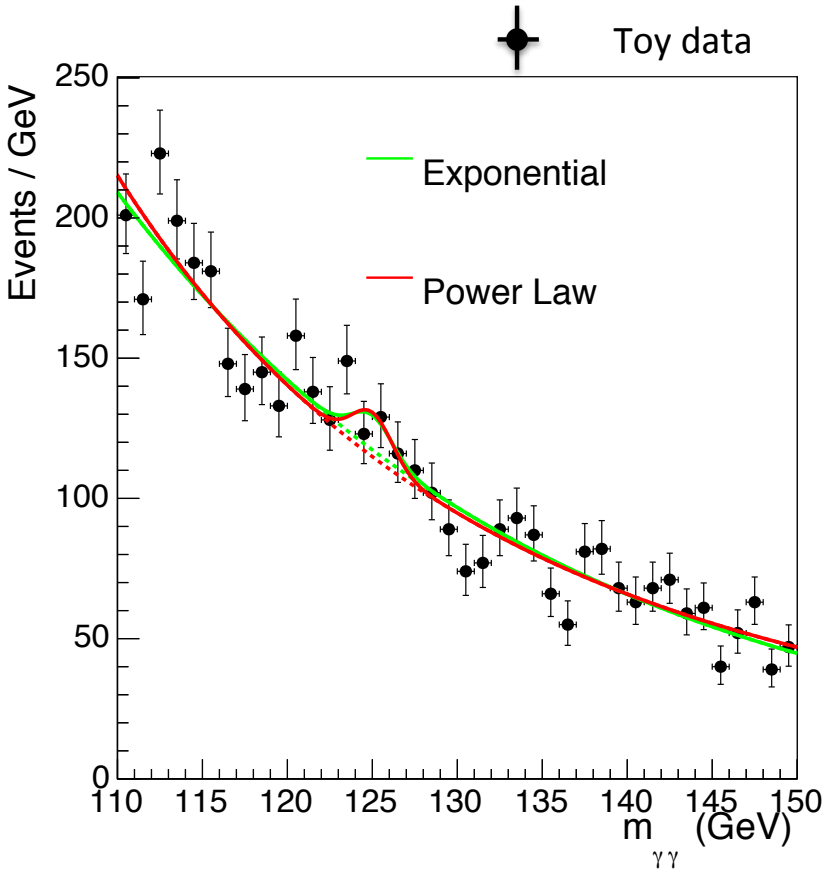


$$e^{-p_0 m}$$

The discrete profiling method:

The model choice problem

But then so does a power-law…

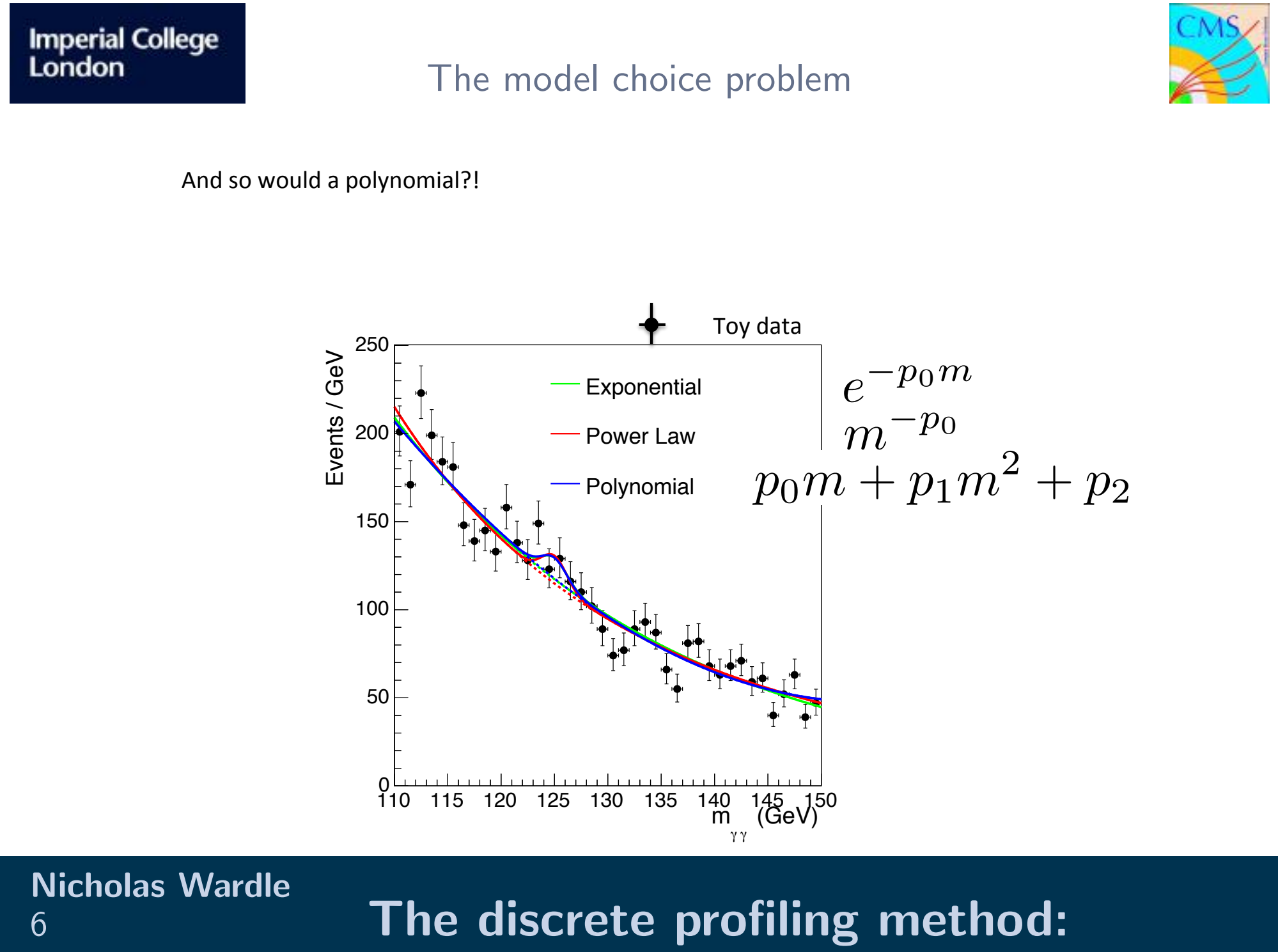


$$e^{-p_0 m}$$

$$m^{-p_0}$$

The discrete profiling method:

The model choice problem



And so would a polynomial?!

$$e^{-p_0 m}$$

$$m^{-p_0}$$

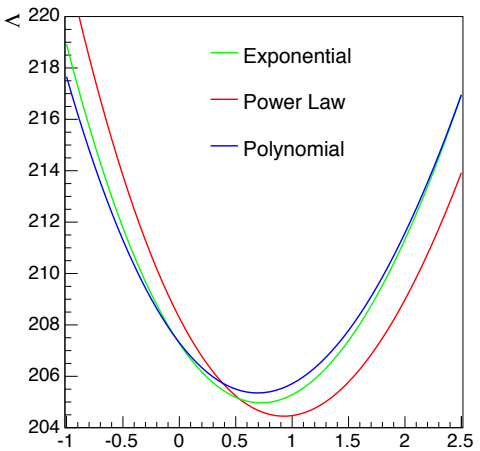
$$p_0 m + p_1 m^2 + p_2$$

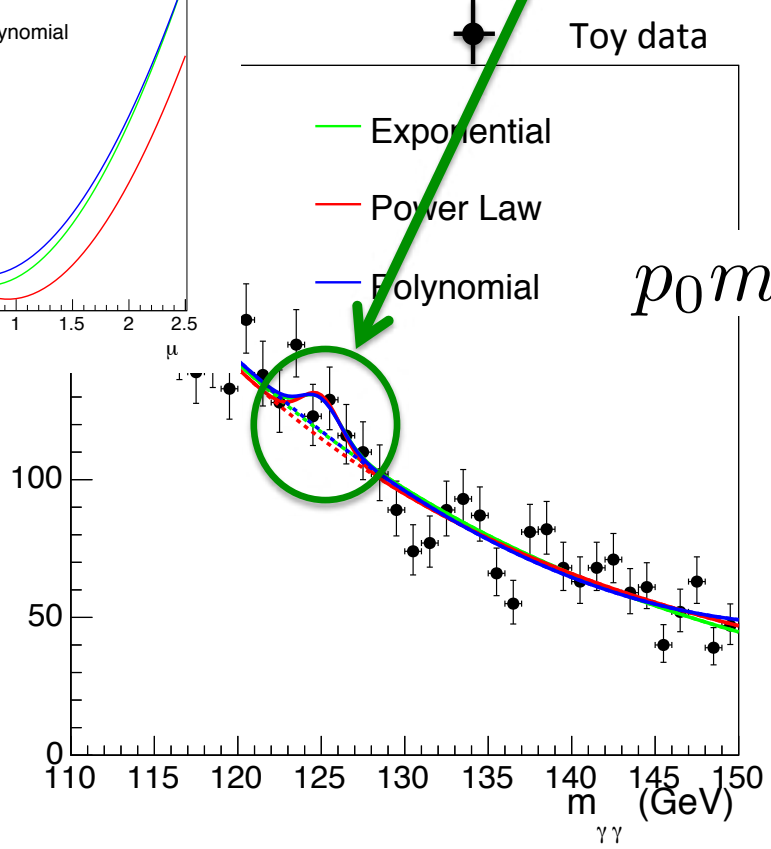
The discrete profiling method:

The model choice problem

Which one should we use (if any?)

The difference In shape can make a difference in the results depending on which one you pick





$$e^{-p_0 m}$$

$$m^{-p_0}$$

$$p_0 m + p_1 m^2 + p_2$$

The discrete profiling method:

What solutions are out there?

1. Pick your favourite model (or the one which fits best) and ignore all others
2. Look at difference in results from your favourite model with others and add as a systematic
3. Use toys to assess any difference and add this as a systematic
4. Increase freedom of the model to minimise systematic bias but increase statistical uncertainty - as in the CMS discovery paper (Phys. Lett. B 716 (2012) 30)

What we want to know is:

- How do we choose which model to use?
- How do we quote the result?
- How do we assign a systematic uncertainty from any choice we've made?

Outline

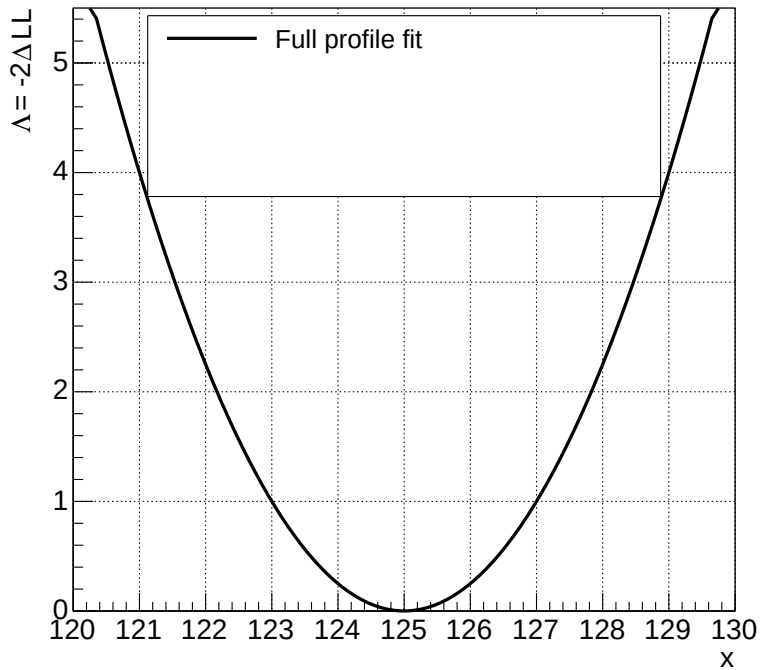
- Present here a method for treating model choice uncertainties like a discrete nuisance parameter
- It summarises the work of P. Dauncey, M. Kenzie, N. Wardle and G. Davies *JINST 10 P04015* ([arXiv:1408.6865])
- CAVEAT! - This is not a "silver bullet" solution but one method, used in CMS, to address the problem.

Concept of a nuisance parameter

Consider a simple situation:

- one parameter of interest - e.g the mass of the signal, x
- one nuisance parameter - e.g. background exponential slope, θ
- all other parameters fixed (we imagine they are known perfectly)

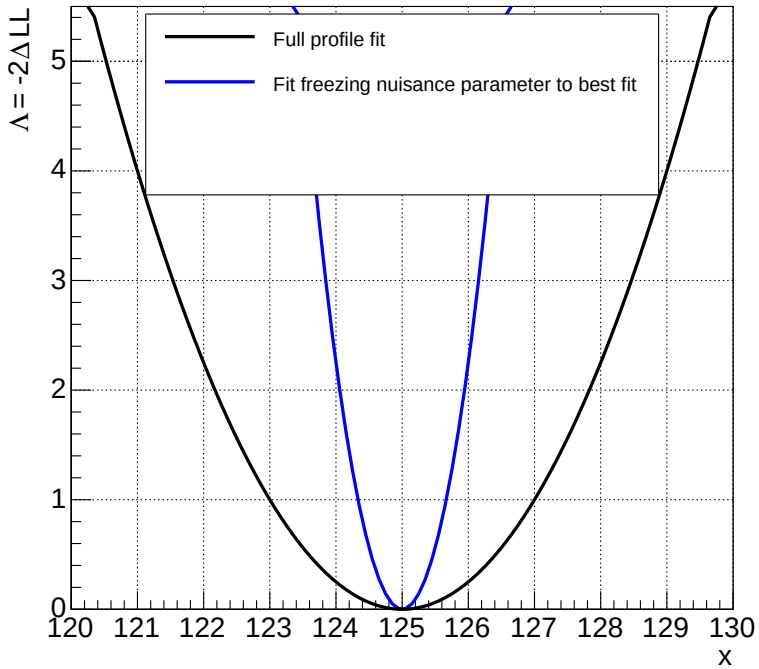
1. **Scan $\Lambda = -2LL$ of parameter x whilst profiling θ**



Concept of a nuisance parameter

- Now imagine the background parameter is perfectly known also
  - fix nuisance parameter which now has no variation
  - equivalent to the statistical only error

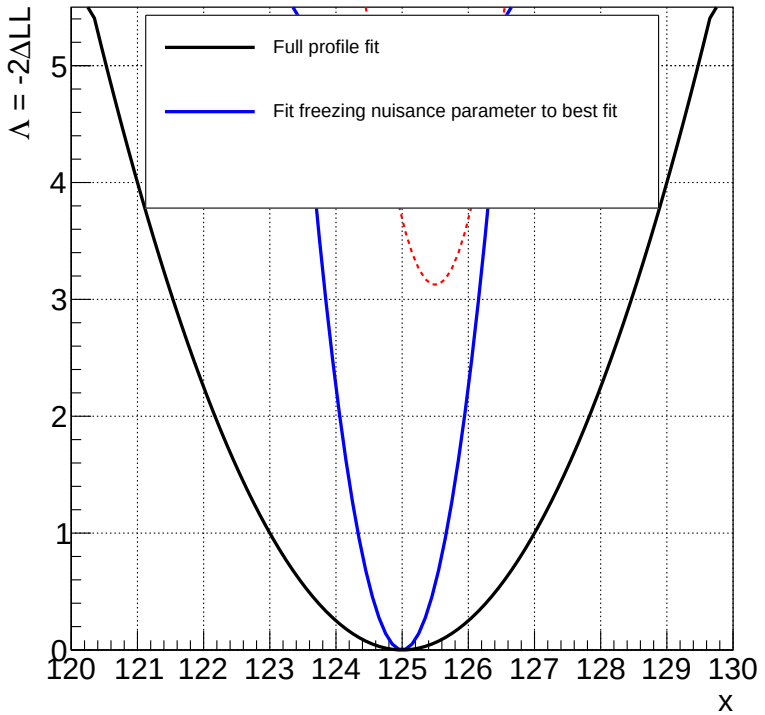
2. **Fix θ to it's best fit value**
   - blue line



Concept of a nuisance parameter

- What about if we fix the background parameter to some other value?
  - this gives some other curve
  - not necessarily near the minimum

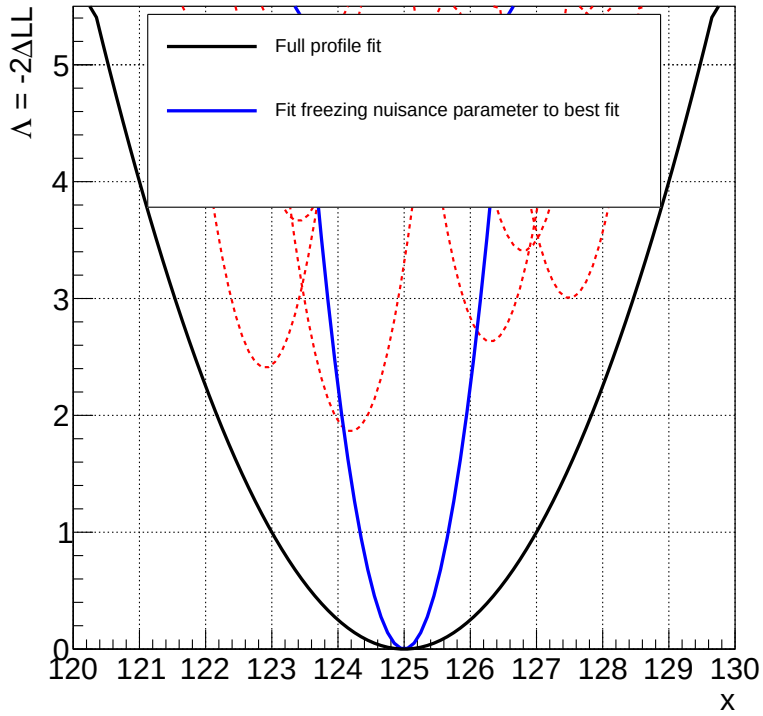
3. **Fix θ to a random value**
   - red dashed line



Concept of a nuisance parameter

- Can do this for a few different values of the background parameter

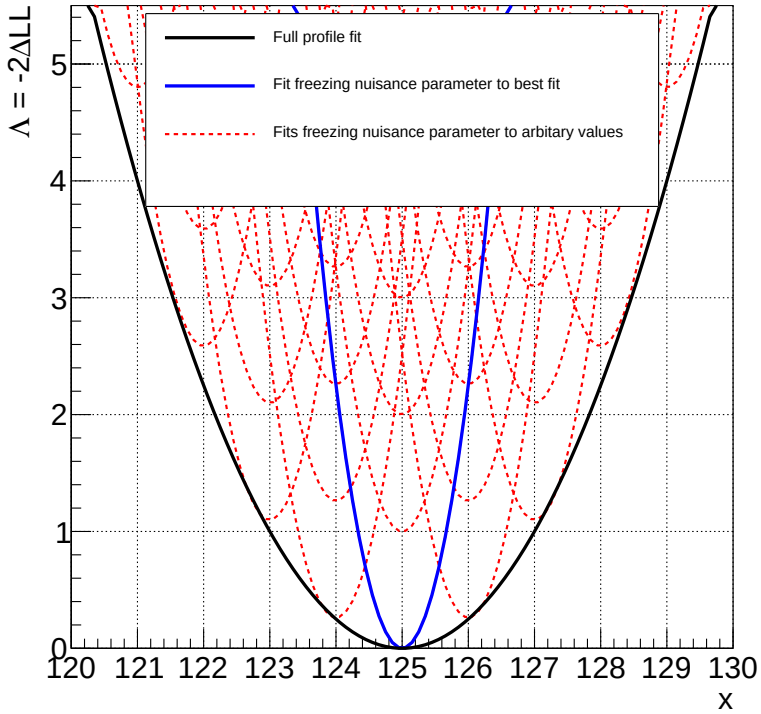
2. **Fix θ to a few random values**
   - red dashed lines



Concept of a nuisance parameter

- And even more values...

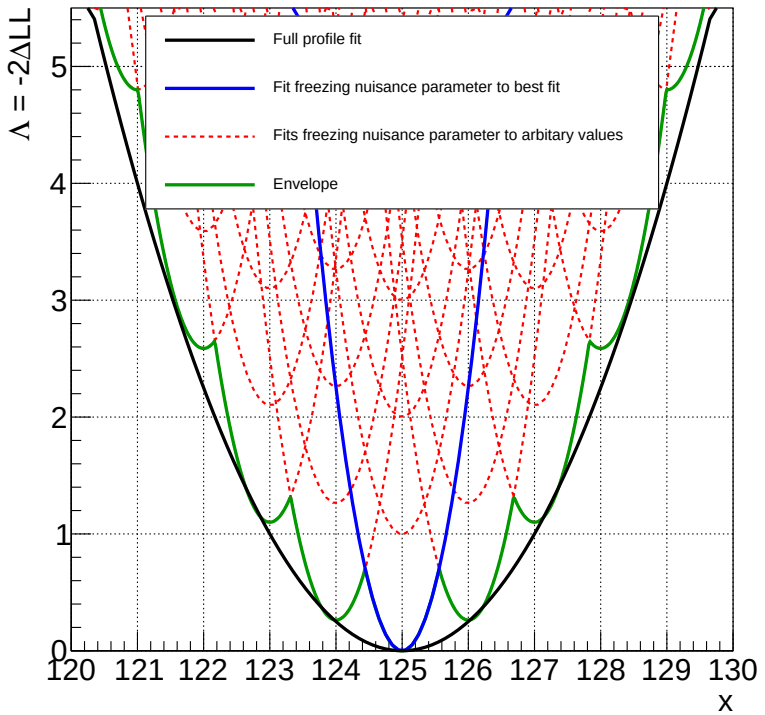
2. **Fix θ to a few random values**
   - red dashed lines



Concept of a nuisance parameter

- If you draw the minimum contour around all of the red dashed lines you begin to recover the original curve
  - In this case it doesn't matter because θ is a continuous nuisance parameter
  - But if we have a parameter that can ONLY take discrete values then we can make a profile likelihood.

2. **Draw minimum "envelope"**
   - green line

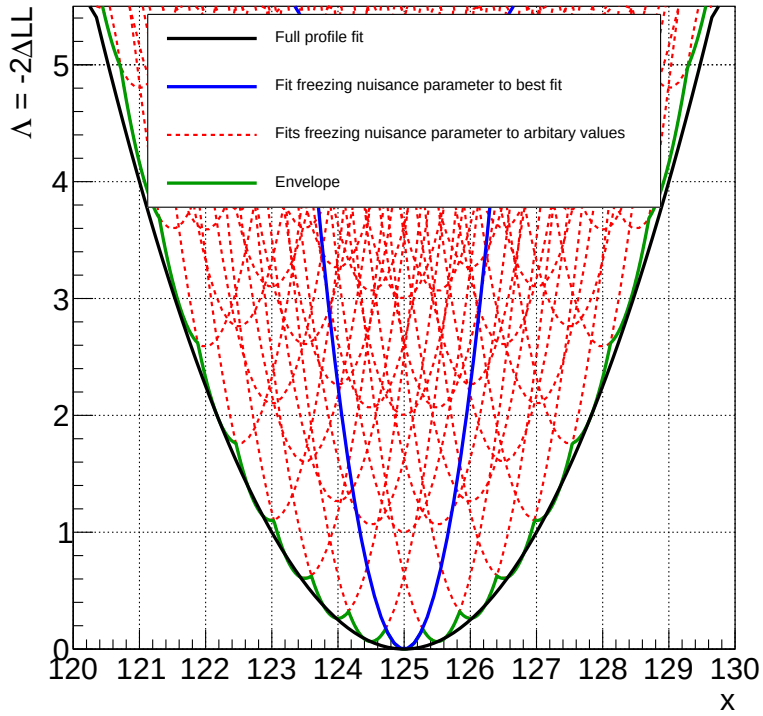


The discrete profiling method:

Concept of a nuisance parameter

- Clearly the more discrete values we sample the closer we get to the original

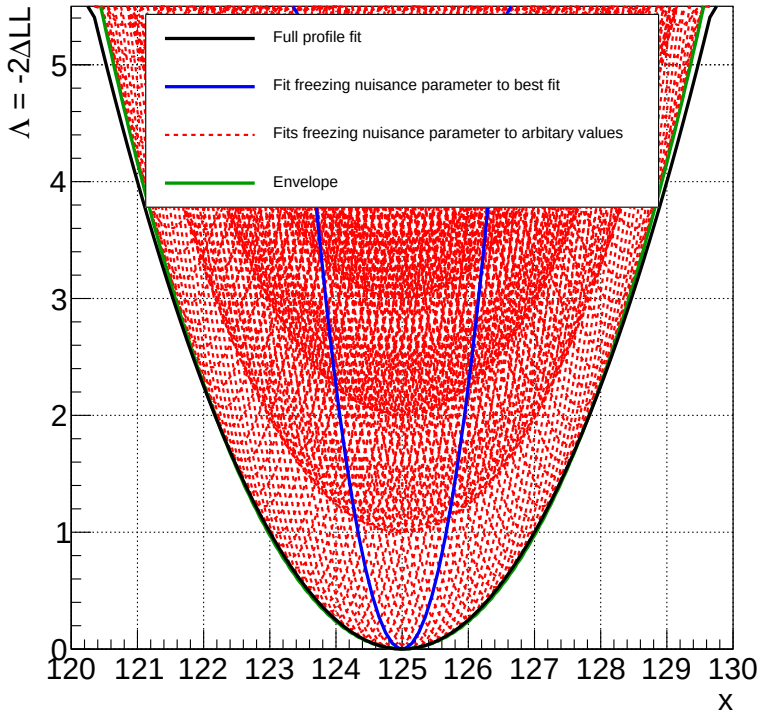
2. **Draw minimum "envelope"**
   - green line



Concept of a nuisance parameter

- Clearly the more discrete values we sample the closer we get to the original
- In principle one can mix discrete nuisance parameters with continous ones

2. **Draw minimum "envelope"**
   - green line



The discrete profiling method:

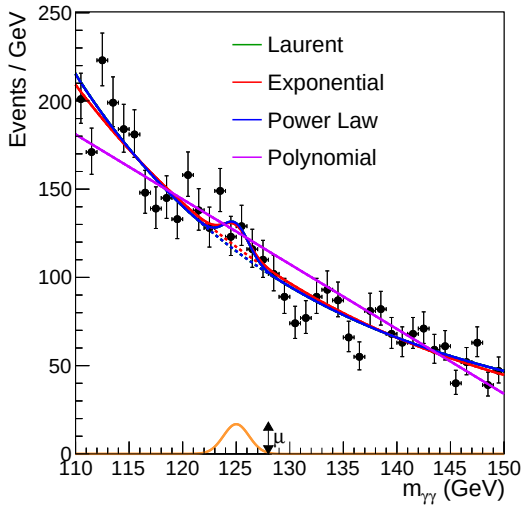
A more realistic example

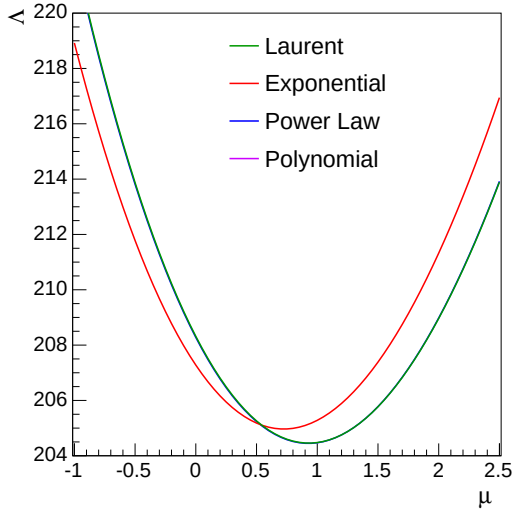
Choose signal strength (μ) as POI. Define a binned log-likelihood ratio[1] as ...

$$\Lambda = 2\sum_i \nu_i - n_i + n_i \ln\left(\frac{n_i}{\nu_i}\right)$$

- The expectation in each bin, ν_i is given by a background model plus $\mu\times$ signal model
- If the background parameters (θ) are free parameters, we let $\nu_i \to \nu_i(\theta)$ and profile them.
- As a function of μ, we now calculate the profiled likelihood to obtaine a profiled likelihood curve for each choice of background model

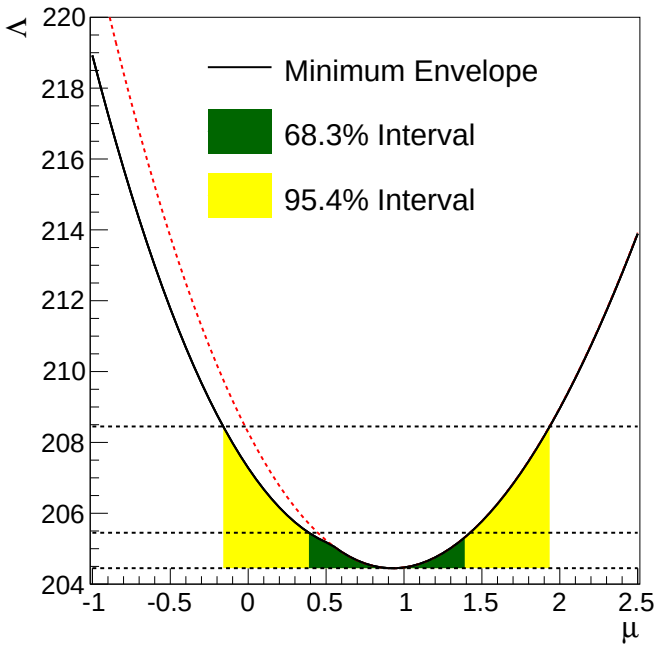
- Choices which are similar shouldn't effect our result (**Laurent** and **Power Law**)
- Choices which are poor should have little impact (**Polynomial**)
- Choices which seem equally valid but disagree should increase our uncertainty (**Exponential**)





[1] Procedure in general does not rely on a binned likelihood

- Take the minimum of all the curves as a function of μ (the envelope). If more than one function contributes then that envelope is wider than any of the individual curves $\rightarrow$ parameter uncertainty is increased
- No explicit model choice has to be made and we don't actually care what model "is the best", since this choice is dynamic in the scan



Result:

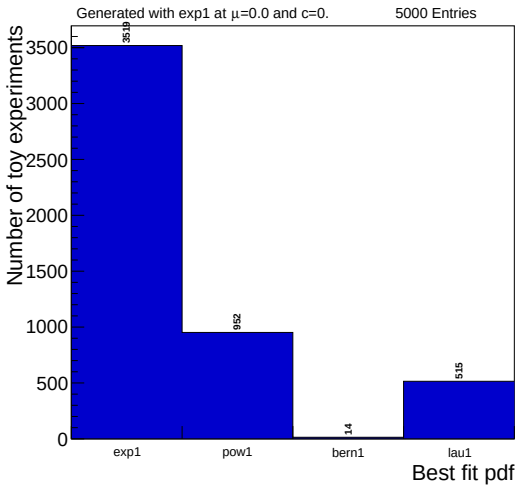
- A best fit value ($\hat{\mu}$) ✔
- A confidence interval ($\Delta\Lambda \leq 1$)✔
- A systematic from the model choice ✔

The discrete profiling method:

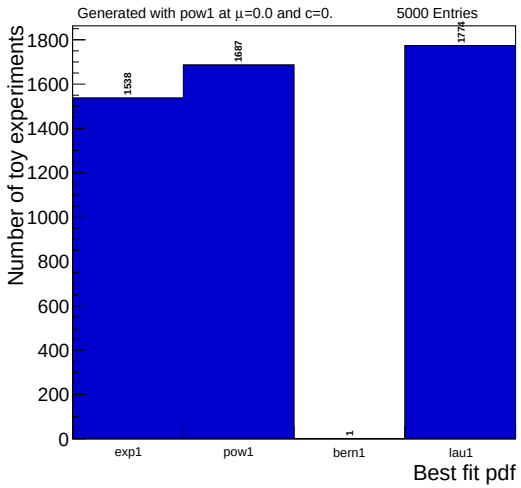
Which PDF fits best?

- Can assess toys to see which PDF minimises the envelope

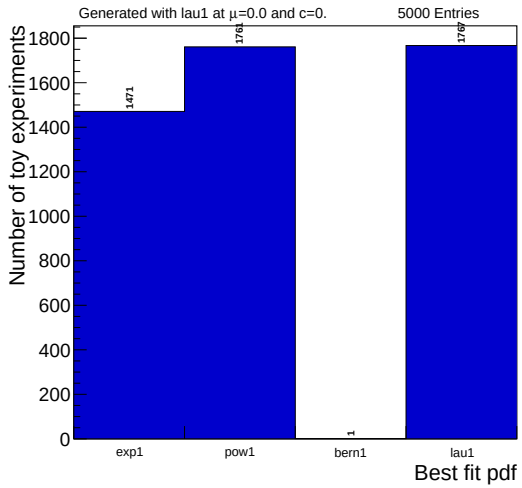
Toy - exponential



Toy - power law



Toy - Laurent



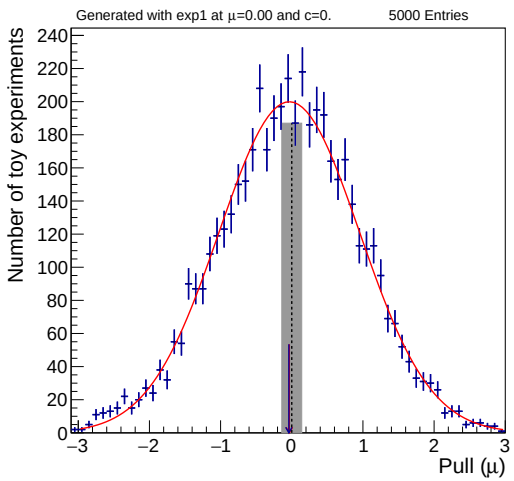
Bias and Coverage properites

How well does the method (including the procedure of obtaining a confidence interval) behave?

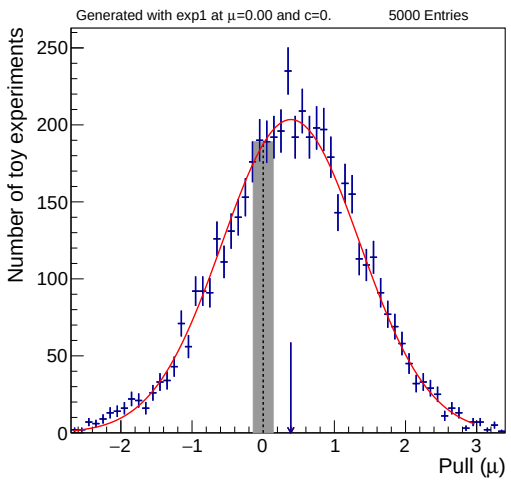
- Generate toy MC from each background hypotheses and then refit to asses the bias (using the pull) and the coverage
- For example generate with exponential background distribution:

Example: generate toys from the exponential function and try fitting with the other functions including fitting back with the envelope procedure, define the pull as $(\hat{\mu} - \mu_{gen})/\sigma_{\mu}$.

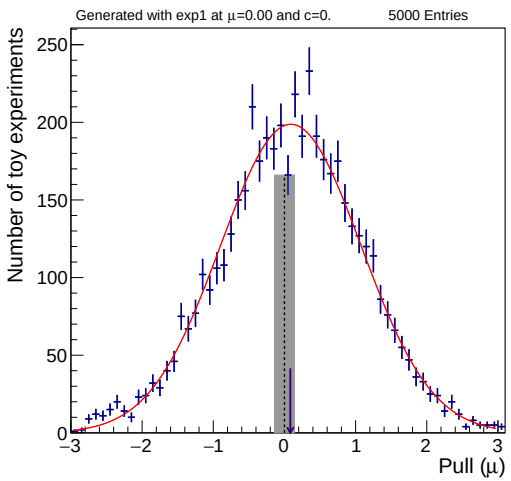
Fit back with exponential



Fit back with power law



Fit back with envelope

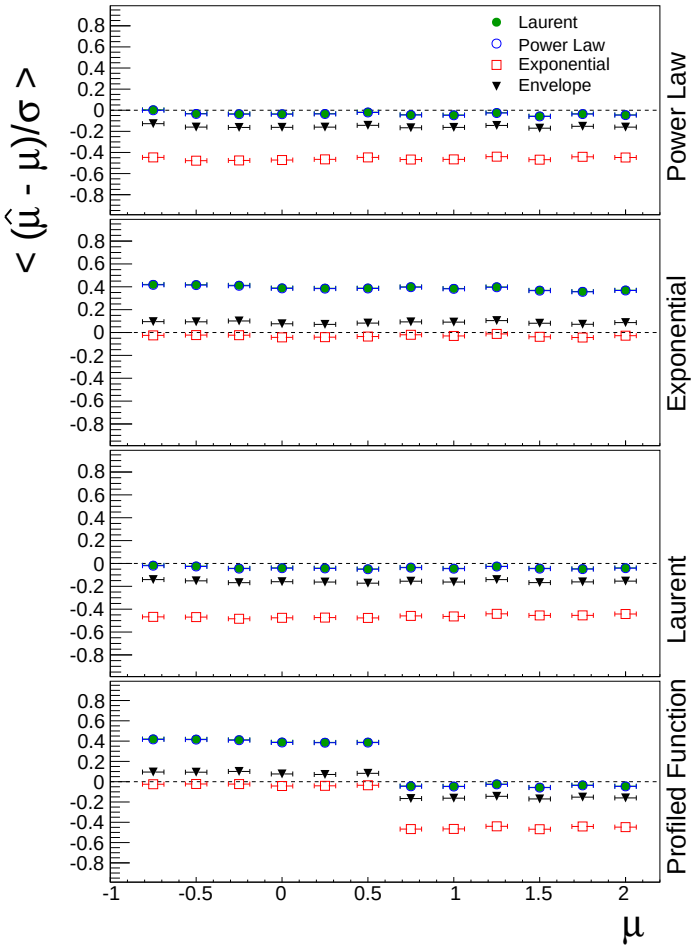


Bias and Coverage properites

- Generate toy MC from various background hypotheses, as a function of the signal size, and then refit to asses the bias

Bias:

- When you generate and fit back with the same (or similar) background function the bias is neglible (**green points** in top panel, **red points** in second panel)
- When you generate and fit back with different functions the bias is large (**red points** in top panel, **green points** in second panel)
- Using the profile envelope (**black points**) you find a small bias **for all cases**



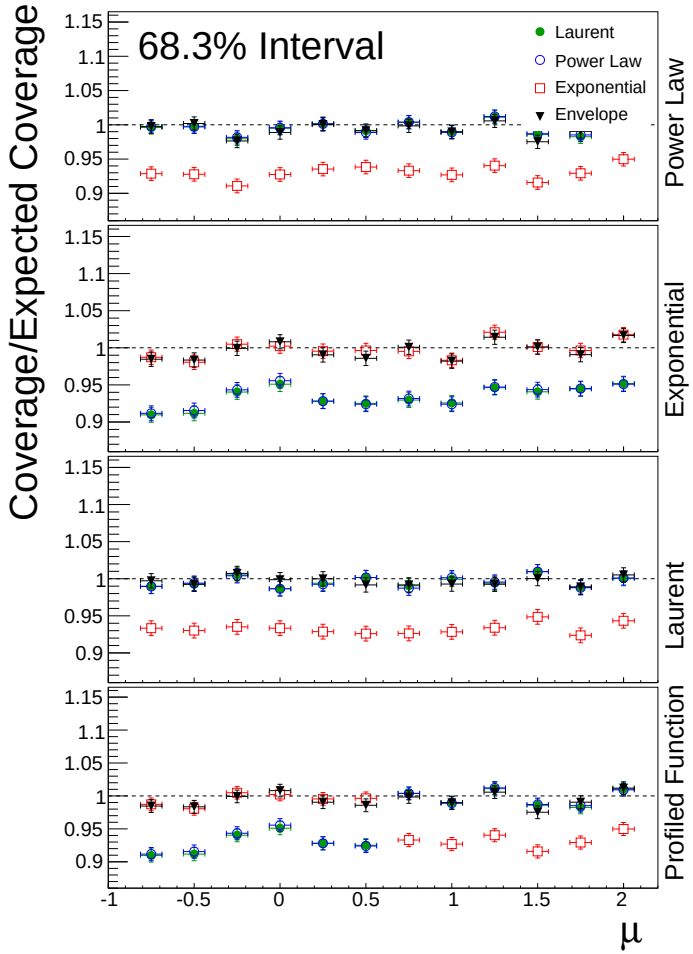
The discrete profiling method:

Bias and Coverage properites

- Generate toy MC from various background hypotheses, as a function of the signal size, and then refit to asses the coverage

Coverage:

- When you generate and fit back with the same (or similar) background function the coverage is good (**green points** in top panel, **red points** in second panel)
- When you generate and fit back with different functions there can be under-coverage (**red points** in top panel, **green points** in second panel)
- Using the profile envelope (**black points**) you recover good coverage **for all cases**

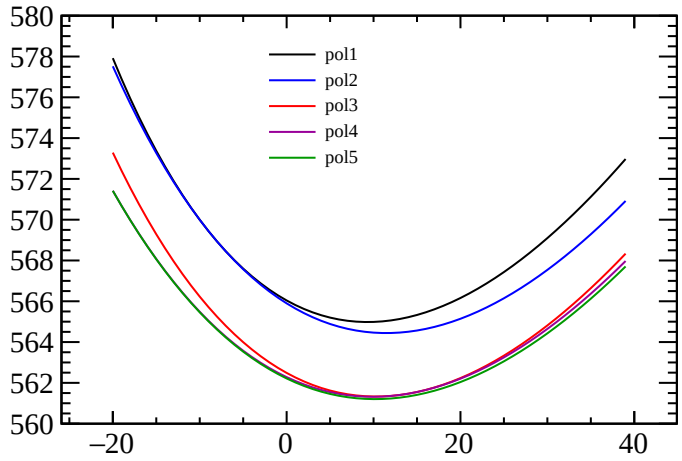


The discrete profiling method:

Different number of parameters?

- How do we compare models with different numbers of parameters?
- The value of Λ is simply an indicator of relatively how well the data agrees with a particular point in the model space
  - It does not account for degrees of freedom used to make that agreement!
- Consequently using Λ without any penalty would always result in choosing the most flexible model(s) available
- There is also no natural mechanism for ignoring higher and higher order functions when calculating Λ

- We aim to correct the Λ for this but it is not obvious by how one should do this (we tried a few in the paper)

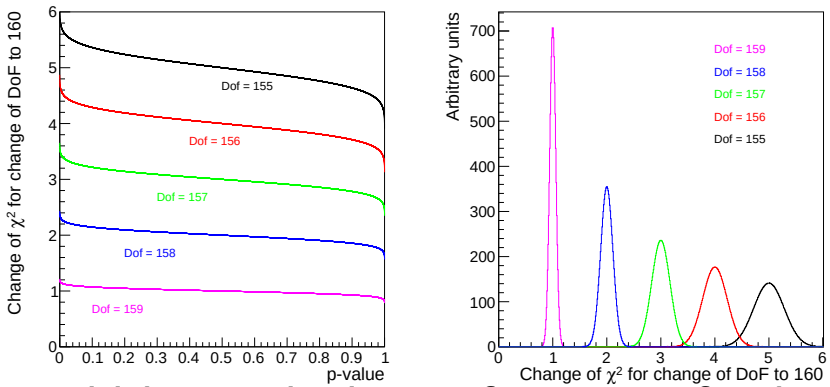


The p-value correction

- For binned fits, in the high statistics limit then $\Lambda \approx$ distributed as a χ^2 with degrees of freedom : $n_{bins} - n_{pars}$.
- Convert this to a p-value using $p = 1 - F(\Lambda, n_{bins} - n_{pars})$
- Now find Λ' which would have given the same p-value but with degrees of freedom ($n_{pars} = 0$) i.e the one which satisfies,

$$p = 1 - F(\Lambda', n_{bins}) \tag{1}$$

- The "correction" to the log-likelihood is just $\Lambda' - \Lambda$. It depends on number of bins, number of parameters and the value of p.[2]



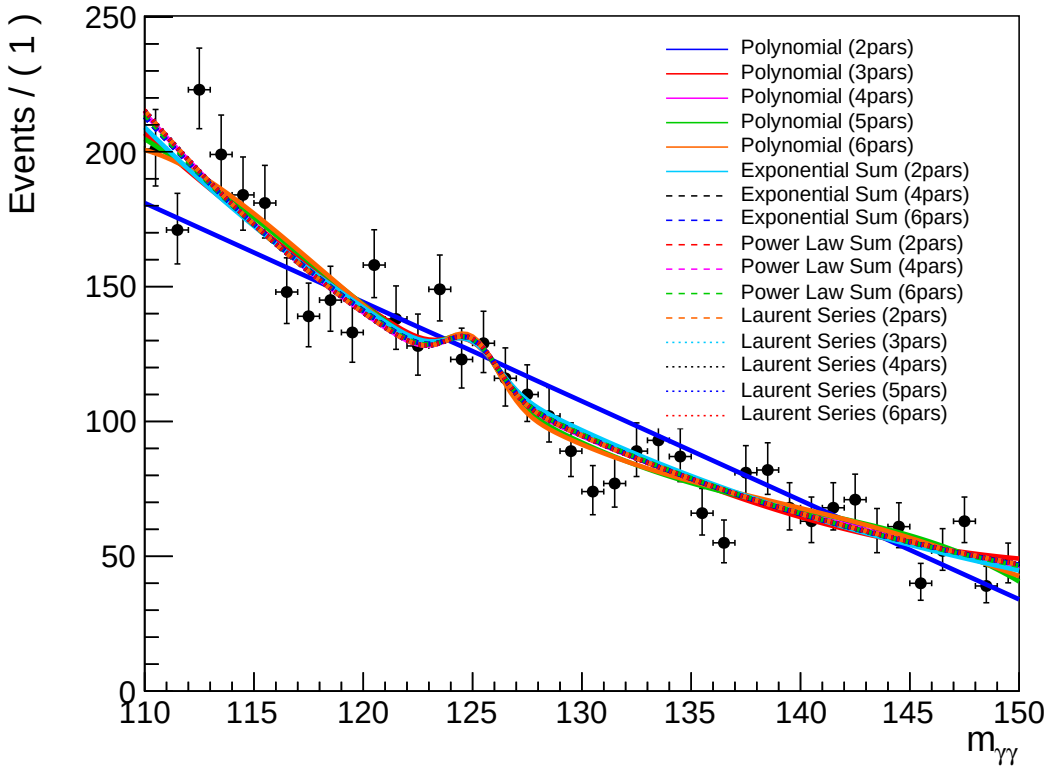
- The correction could be applied as a function of μ but in a wide range of p-values, this correction yields

$$\Lambda' - \Lambda \approx n_{\text{pars}} \quad \textit{so} \quad \Lambda_{\text{corrected}} \approx \Lambda + n_{\text{pars}} \tag{2}$$

[2] `TMath::ChisquareQuantile(1-p,160) - TMath::ChisquareQuantile(1-p,160-N)`

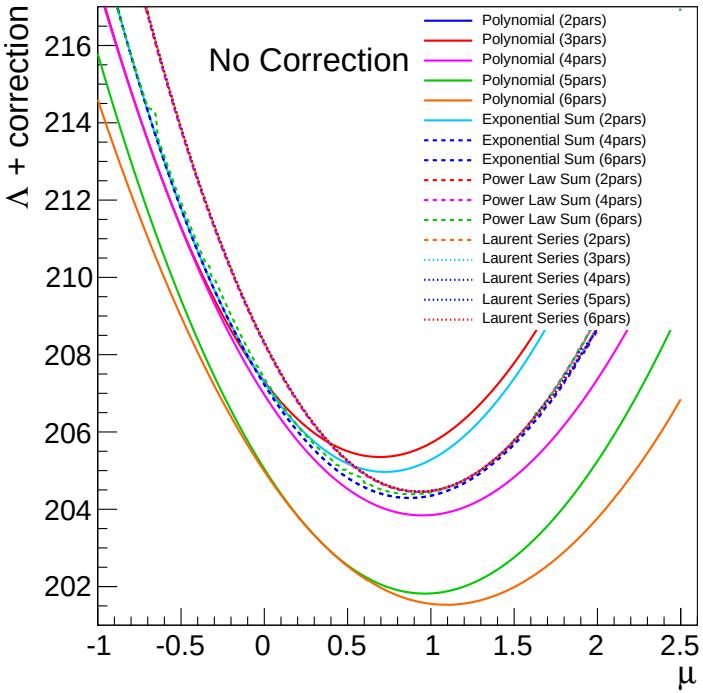
Back to the example spectrum

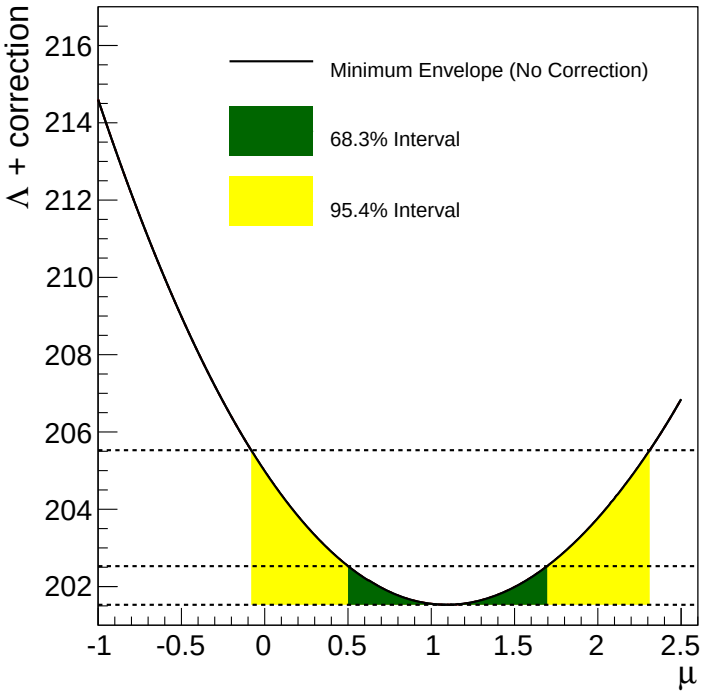
- Take the same dataset and now try many functions (of different orders)
- Scan the likelihoods as before now applying the correction



Example case with higher order functions

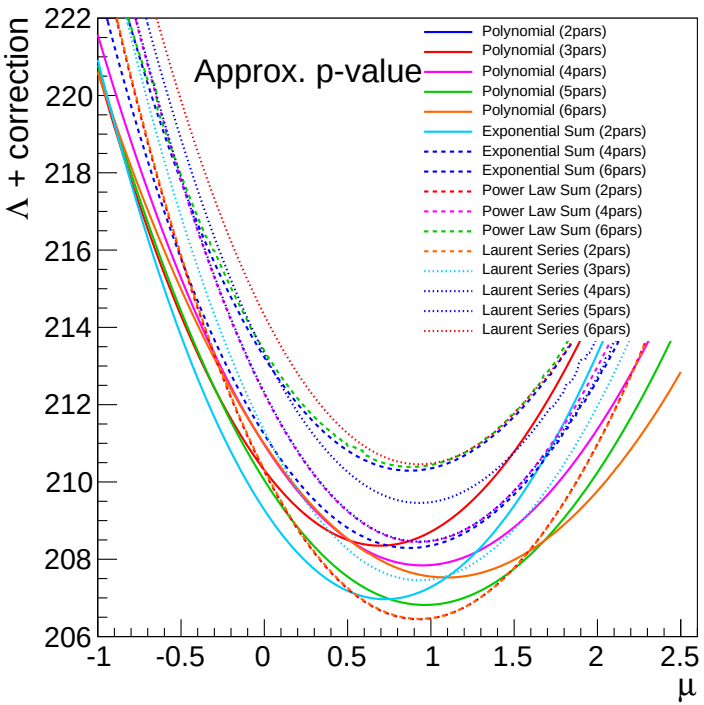
- Profile same dataset with many functions (of different orders)
- **With no correction**
  - Best Fit: 6th order polynomial (highest order tried)

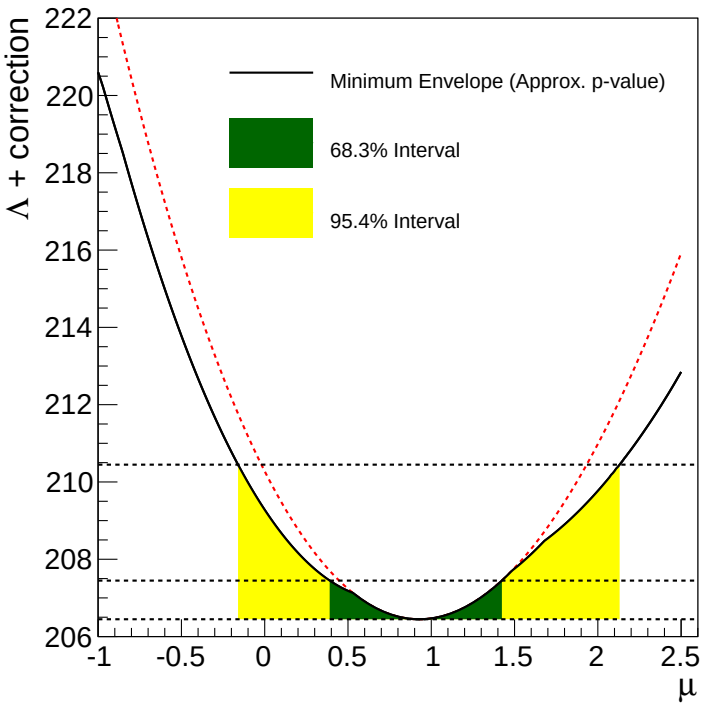




Example case with higher order functions

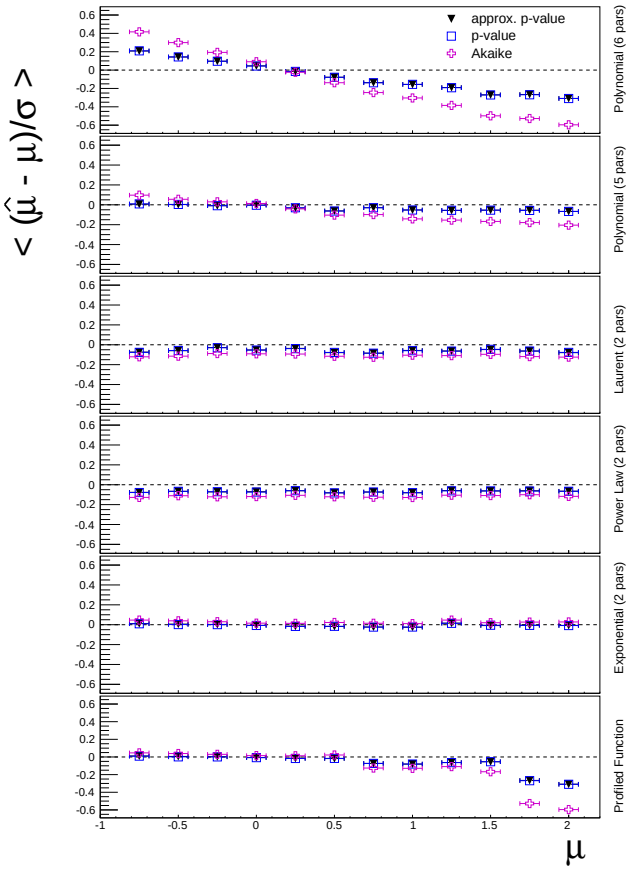
- Profile same dataset with many functions (of different orders)
- **With approx. *p*-value correction** (Λ +1 per dof)
  - Best Fit: 2 parameter power law

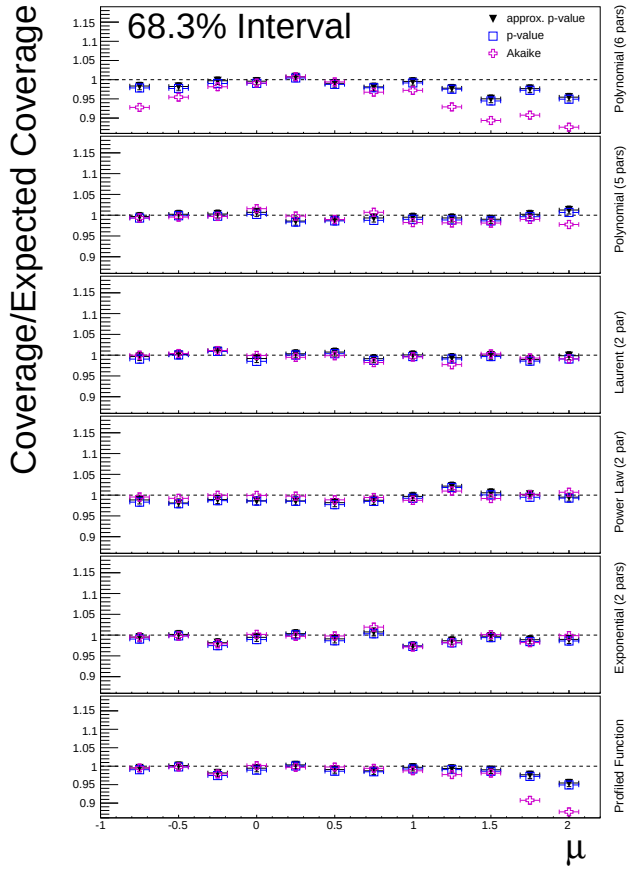




Bias and coverage for many order functions

- Now comparing envelope of all functions with different correction schemes





Other forms of correction

- Using the p-value argument suggests:

$$\Lambda_{\text{corr}} = -2 \ln \mathcal{L} + N_{\text{par}} \tag{3}$$

- There are other forms of likelihood correction out there
- Aikaike information criterion (AIC):

$$\Lambda_{\text{corr}} = -2 \ln \mathcal{L} + 2N_{\text{par}} \tag{4}$$

- Bayesian information criterion (BIC):

$$\Lambda_{\text{corr}} = -2 \ln \mathcal{L} + N_{\text{par}} \ln(n) \tag{5}$$

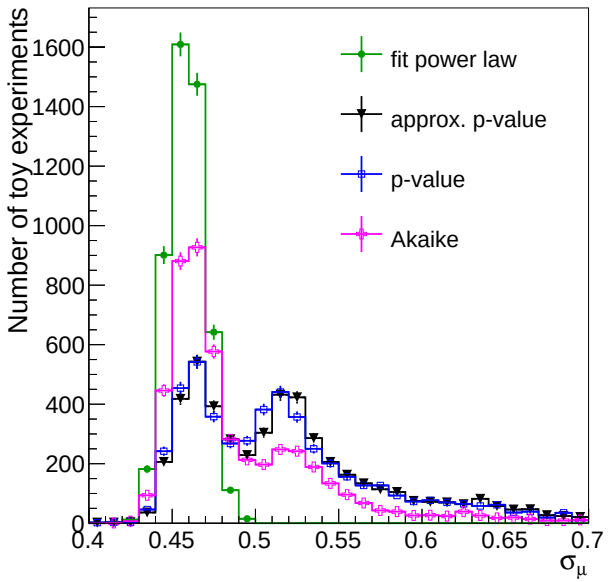
- In general they take the form:

$$\Lambda_{\text{corr}} = -2 \ln \mathcal{L} + cN_{\text{par}} \tag{6}$$

where c is some "correction value" to be determined

What happens to the error?

- Over a set of pseudoexperiments the error when using the envelope increases
- This quantifies the systematic uncertainty contribution from the model choice
- The size of this systematic is smaller depending on the choice of c

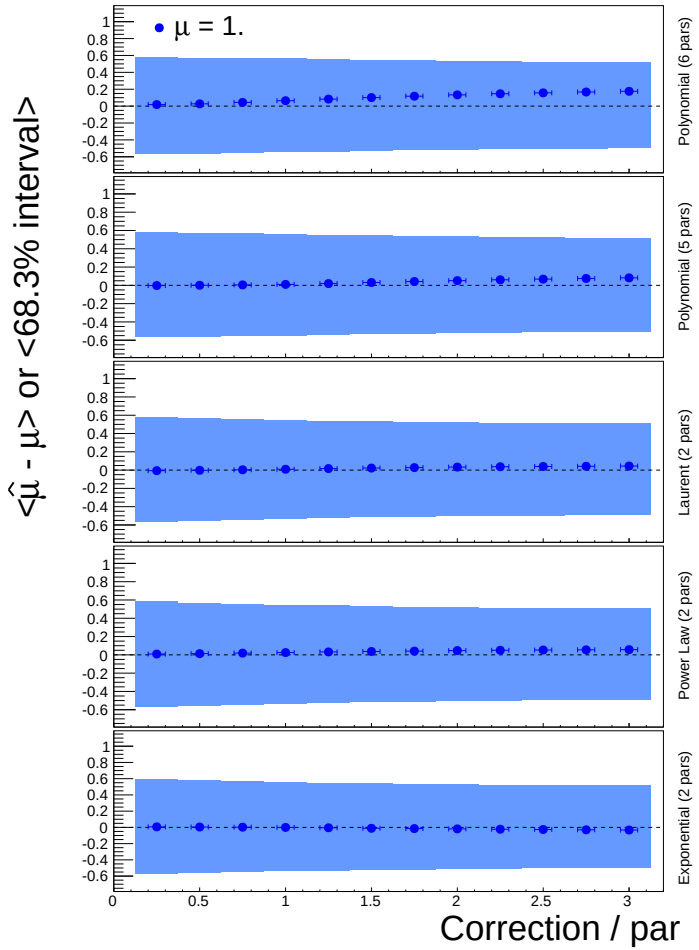


What correction to use?

- As we have seen the corrected likelihood takes the form,

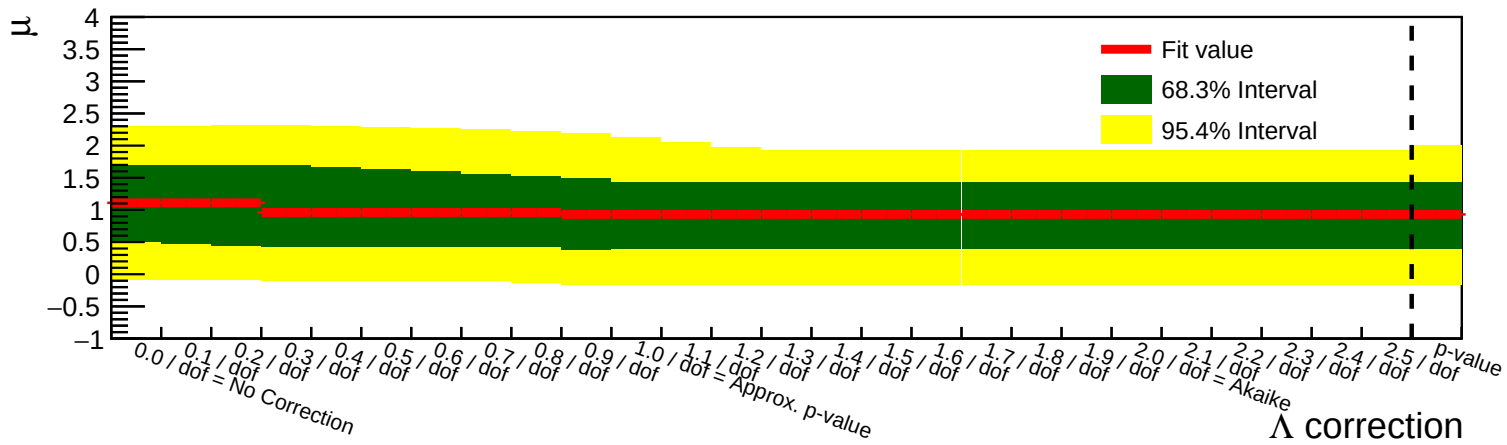
$$\Lambda_{\text{corr}} = -2 \ln \mathcal{L} + cN_{\text{par}}$$

- The coverage is largely independent of the choice of c
  - Within reason the choice for the value of c can be motivated by other considerations
  - This will depend on the application and the size of the dataset available
- Ends up being a trade off between:
  - the size of the correction (eventual bias)
  - statistical precision
- Depends on specific analysis and individual preference



Central value and error dependence on the correction

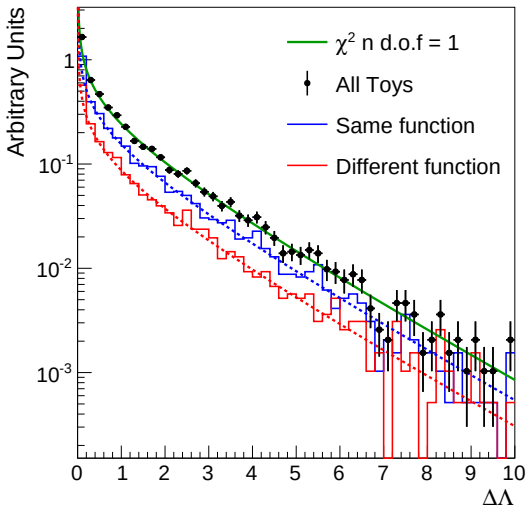
- As a function of the correction value the uncertainty (and central value) can change
- At lower values of c you have a large statistical uncertainty
  - In principle for this example if $c = 0$ the statistical error is infinite
- At larger values of c you have a potentially large bias



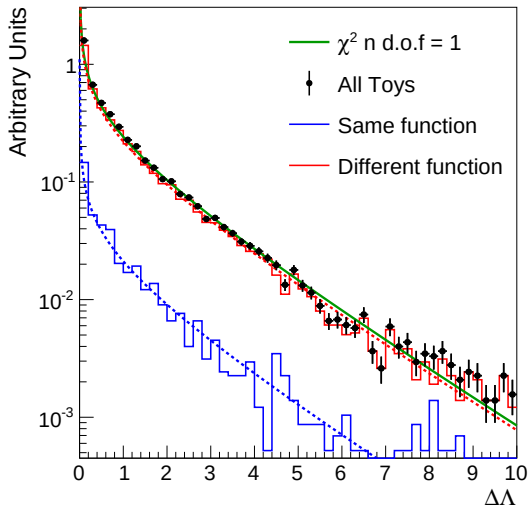
Does the uncertainty make sense?

- Difference in 2×Log-likelihood between the true and fitted values of μ in general follows a χ^2 distribution with 1 d.o.f (Wilks'theorem.)
- This appears to be the case in this example! Also true whether or not the same function which was used to generate is fitted back or the discrete profiling picks another function

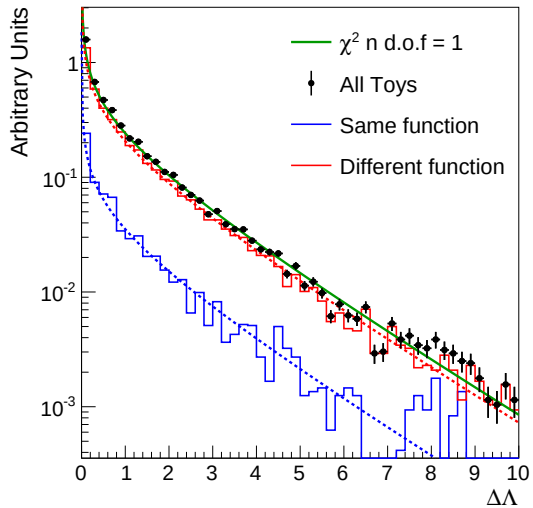
$c = 0$



$c = 1$



$c = 2$

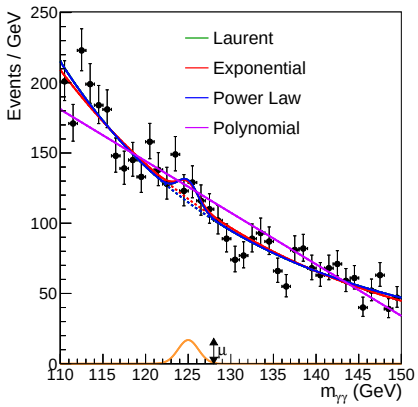


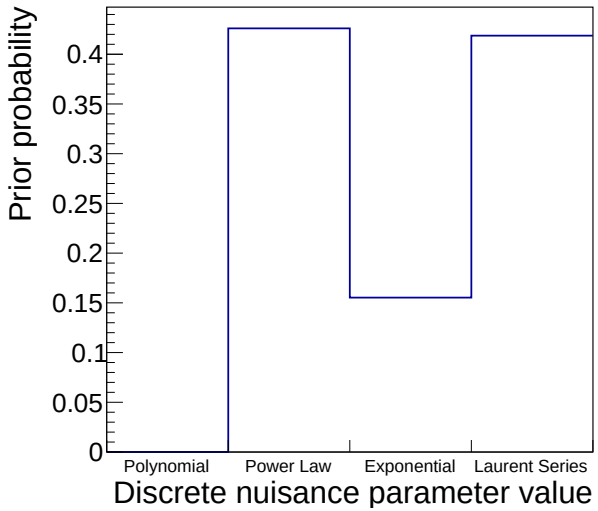
Open questions for (any?) method like this

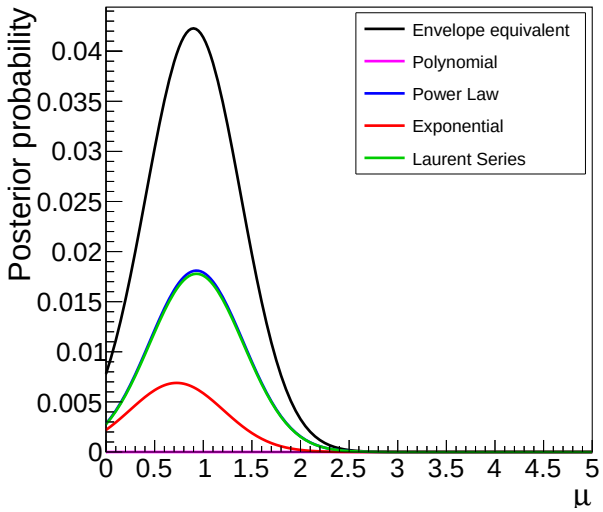
- Is there an analytical proof / better motivation of a correction for Λ to use?
- How can we use the method to set Bayesian credible intervals rather than frequentist confidence intervals?
  - What, if any, prior should be used
- How do you decide how many models to include in the envelope if the choice is infinitely many?
  - Fisher test used to find a reasonable range of n_{pars} in each model family.
- How should one assess how many "model" choices is appropriate?
  - Are there other ways of sampling more of the "model phase space"?
  - Is there a way to get a "complete set".
- What restrictions should be placed on this set?
  - Should avoid hiding the signal – Restrict higher derviatives/inflections?.

Bayesian formalism?

- So far the method discussed has been in a frequentist formalism
- In Bayesian context, the "discrete" profiling equates to adding up posterior PDFs each with a weight $\propto e^{-cN}$
- Would make an interesting study for a student?







The discrete profiling method:

Summary

- Demonstrated a new method for treating model choices as discrete nuisance paramters
  - "Profile" the choice and take the "envelope"
  - Choice of correction open to user
  - Choice of which models to include open to user
- The method in a toy example shows small bias and good coverage
- The method has been used in real data analyses (at CMS)
  - Small bias and good coverage found in those scenarios
- Similar studies are highly recommended for each use case
- Several possible extensions and open questions

Imperial College London

CMS

BACKUP SLIDES

Higgs to two photons at CMS

- This is what the technique was developed for
- 25 analysis categories all with different signal to background, resolution and background shapes
- Perform a simultaneous fit across all 25 for signal size
- Profile between 4-16 background functions in each category
- Order of 50 additional continuous nuisance parameters in this fit also
  - Many of which are correlated across categories
- Without nuisance parameter correlation then number of combinations goes like

$$N_c = \sum_i^c n_i \tag{7}$$

  for c categories with n_i functions in each.
- With correlated nuisances then every combination is required which goes like

$$N_c = \prod_i^c n_i \tag{8}$$

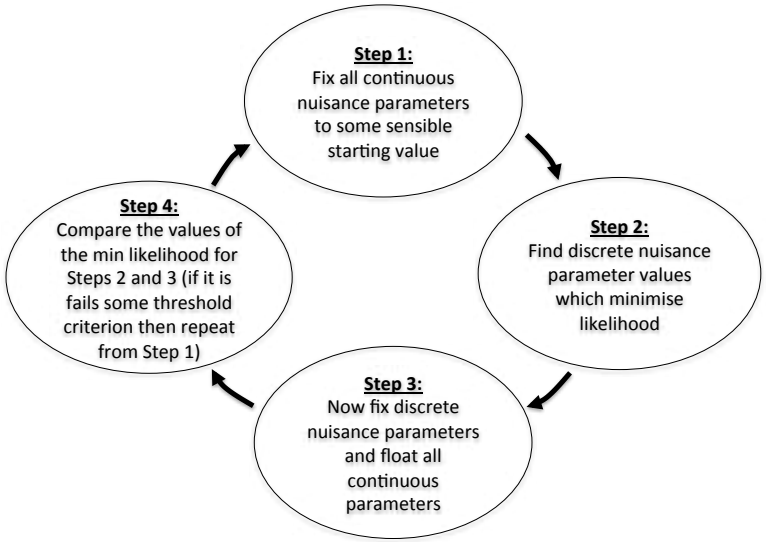
- For CMS $H \to \gamma\gamma = 16^{25} \approx 10^{30}$ combinations
- For any reasonable practical use this has to be reduced

Technical implementation

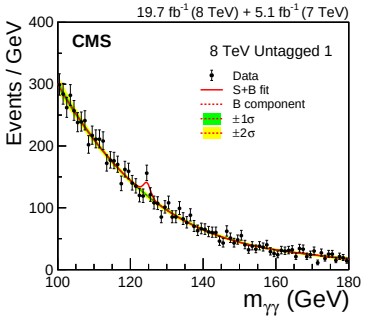
- These studies were developed and performed in `RooFit`
  - Specialised class written: `RooMultiPdf`
  - Not in `RooFit` public release yet
  - Private version being used by both CMS and ATLAS
- **How to reduce numbers of combinations** (given 10^{30} minimisations is impractical for Higgs combination)
  - Run continuous and discrete parts of minimisations separately in iterative procedure
  - Have found that in the $H \to \gamma\gamma$ case the true likelihood is found after $\approx 3-4$ iterations
  - Now number of minimisations goes like

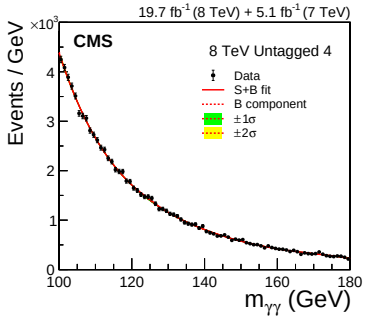
$$N_c = N_I \sum_i^c n_i \qquad (9)$$

for N_I iterations

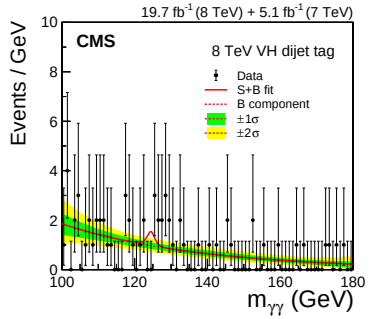


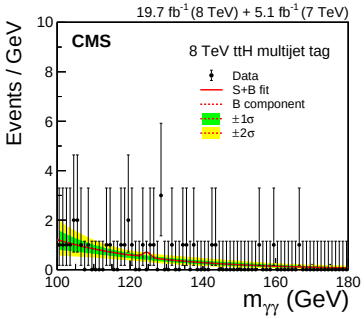
Use in Higgs analyses

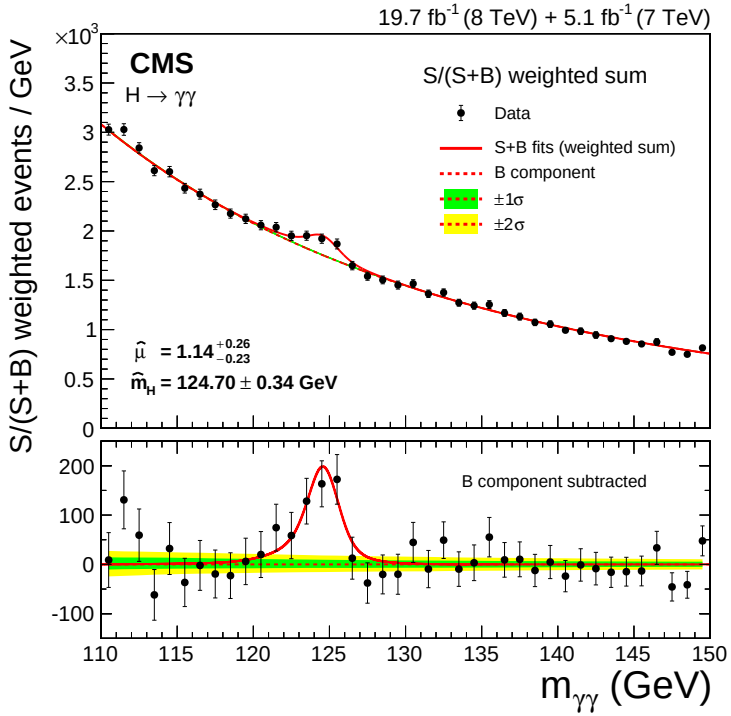




.....







The discrete profiling method:

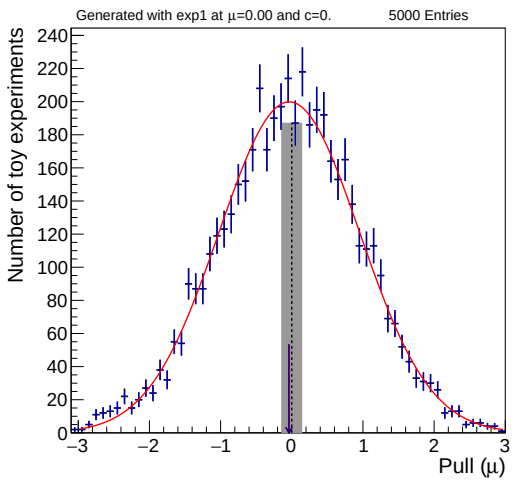
Bias and Coverage properites

How well does the method (including the procedure of obtaining a confidence interval) behave?

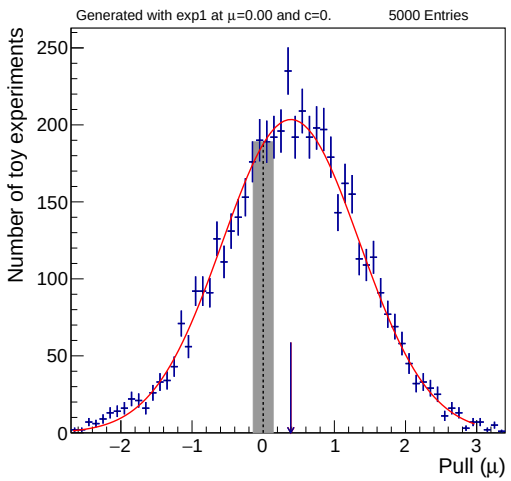
- Generate toy MC from each background hypotheses and then refit to asses the bias (using the pull) and the coverage
- For example generate with exponential background distribution:

Example: generate toys from the exponential function and try fitting with the other functions including fitting back with the envelope procedure, define the pull as $(\hat{\mu} - \mu_{gen})/\sigma_{\mu}$.

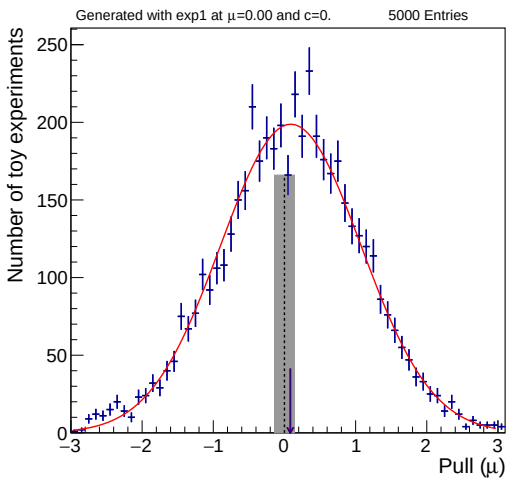
Fit back with exponential



Fit back with power law

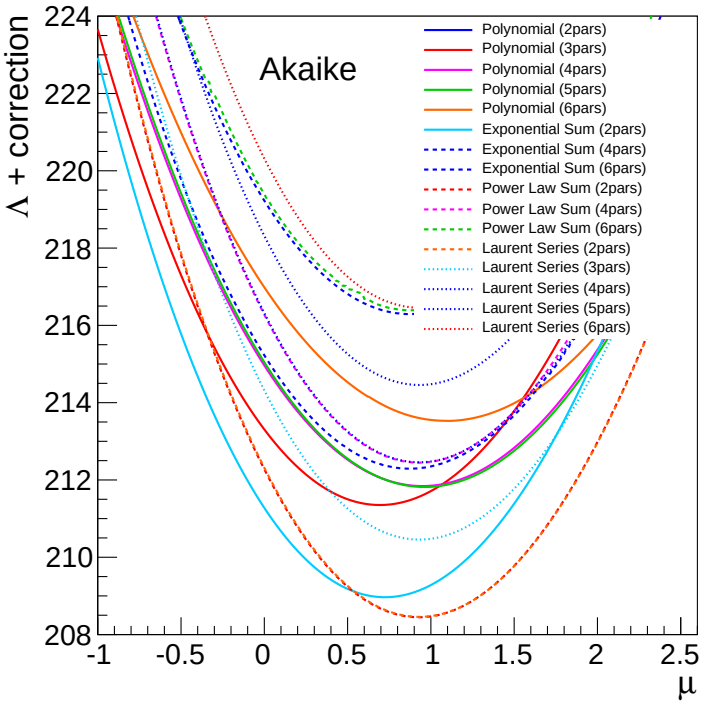


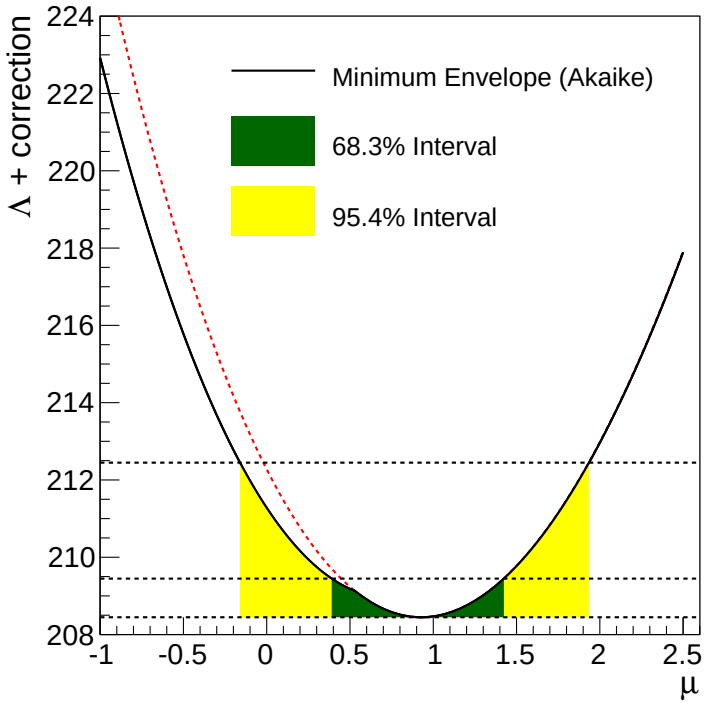
Fit back with envelope



Example case with higher order functions

- Profile same dataset with many functions (of different orders)
- **With Aikaike correction,** $c = 2$ (Λ +2 per dof)
  - Best Fit: 2 parameter power law





Does the uncertainty make sense?

- Difference in 2×Log-likelihood between the true and fitted values of μ in general follows a χ^2 distribution with 1 d.o.f (Wilks'theorem.)
- This appears to be the case in this example! Also true whether or not the same function which was used to generate is fitted back or the discrete profiling picks another function