How to treat model choice uncertainties

Liverpool HEP Seminar

Matthew Kenzie

CERN

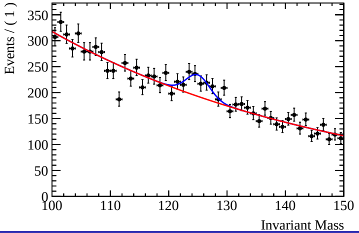
October 14, 2015

1. The model choice problem

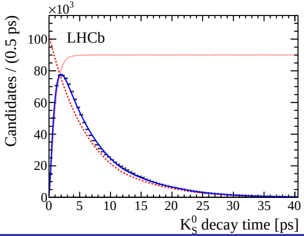
The model choice problem

- In HEP we usually have a dataset that we want to extract some physical parameter from - *parameter of interest* (POI)
  - 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 contributions
  - Acceptance effects
  - This can give a large bias on the parameter of interest (POI)

Large unknown background

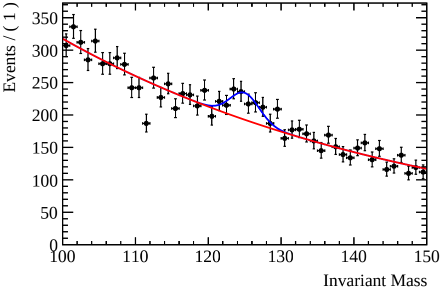


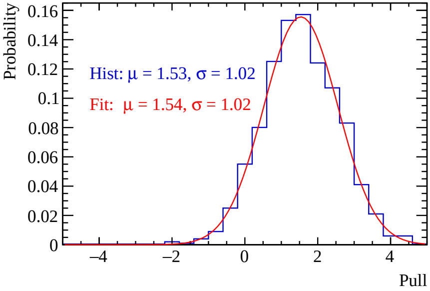
Decay time acceptance



The size of the problem

- In some cases the size of this problem can be large
- Consider the large background, small signal case
- If the *true* distribution is an exponential but I fit instead a single order polynomial
- The bias is huge
  - Measured using the pull over an ensemble of pseudoexperiments





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 and thus reduce sensitivity

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 *JINST 10 P04015* ([arXiv:1408.6865])

Handling uncertainties in background shapes: the discrete profiling method

P. D. Dauncey[a]*, M. Kenzie[b], N. Wardle[b] and G. J. Davies[a]

[a] *Department of Physics, Imperial College London, Prince Consort Road, London, SW7 2AZ, UK.*
[b] *CERN, CH-1211 Geneva 23, Switzerland.*
E-mail: P.Dauncey@imperial.ac.uk

ABSTRACT: A common problem in data analysis is that the functional form, as well as the parameter values, of the underlying model which should describe a dataset is not known *a priori*. In these cases some extra uncertainty must be assigned to the extracted parameters of interest due to lack of exact knowledge of the functional form of the model. A method for assigning an appropriate error is presented. The method is based on considering the choice of functional form as a discrete nuisance parameter which is profiled in an analogous way to continuous nuisance parameters. The bias and coverage of this method are shown to be good when applied to a realistic example.

- This method came about because of the background modelling problem in the CMS $H \to \gamma\gamma$

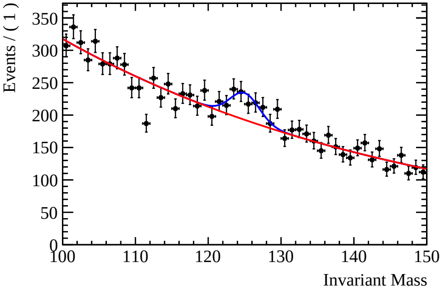
2. The envelope concept

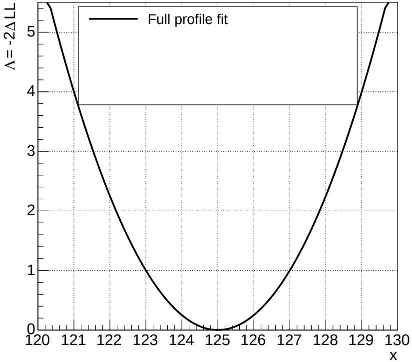
Concept of a nuisance parameter

Consider a simple situation:

- Fit a Gaussian signal and exponential background model to data with
  - one parameter of interest (observable) - 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**

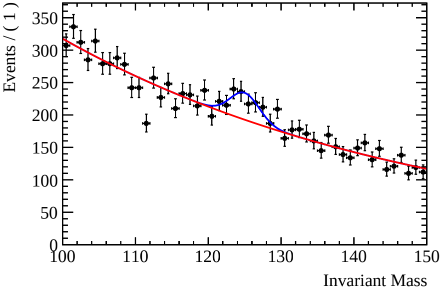


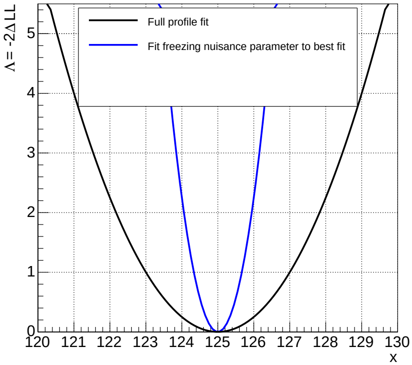


Concept of a nuisance parameter

- 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

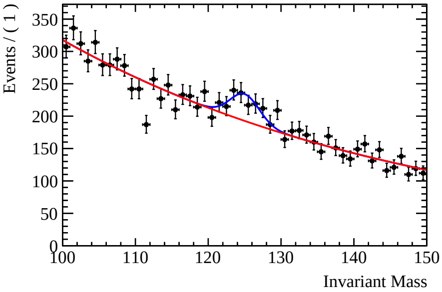


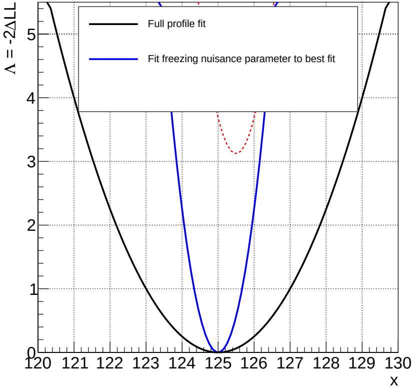


Concept of a nuisance parameter

- 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

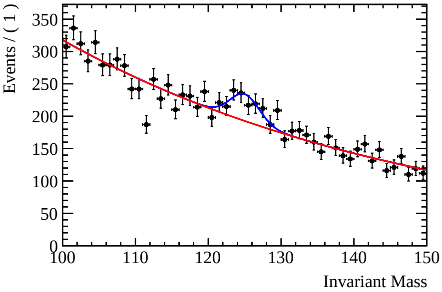


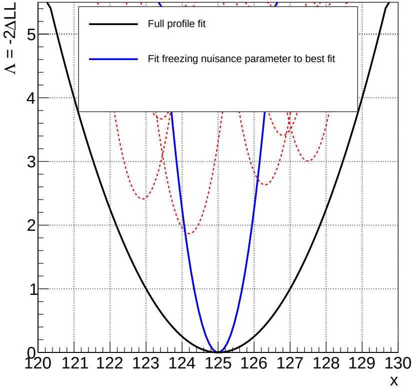


Concept of a nuisance parameter

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

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

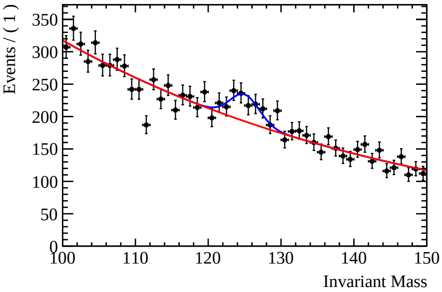


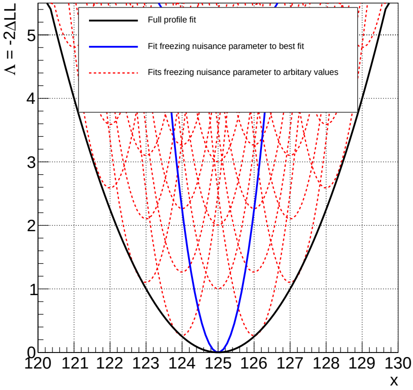


Concept of a nuisance parameter

- And even more values...

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

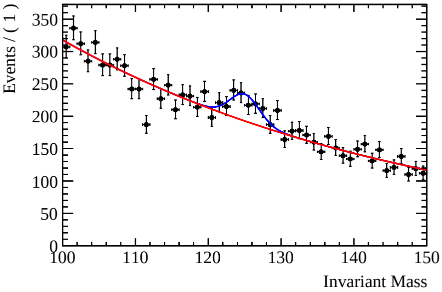


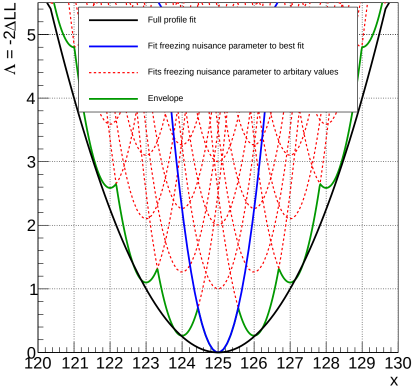


Concept of a nuisance parameter

- 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 in this way
    - For example we have ten different models (we can label them as having discrete value of a nuisance parameter $n = 1 - 10$)

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

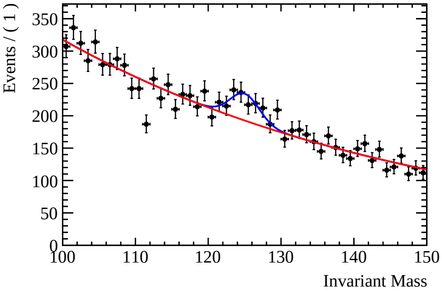


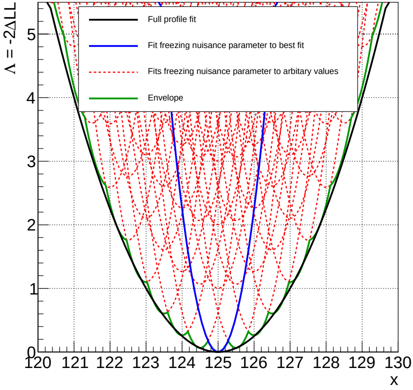


Concept of a nuisance parameter

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

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

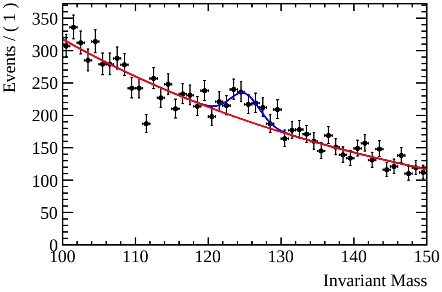


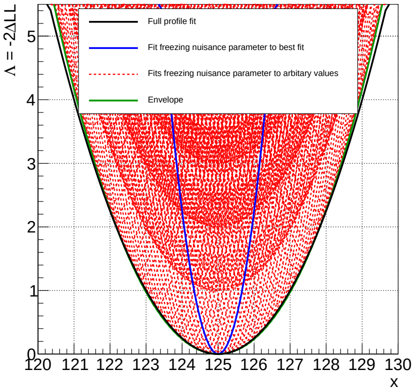


Concept of a nuisance parameter

- Clearly the more discrete values we sample the closer we get to the original
- *IMPORTANTLY - you can mix discrete nuisance parameters with continous ones*

2. **Draw minimum “envelope”**
    - green line

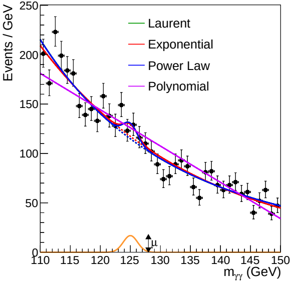




3. An example case

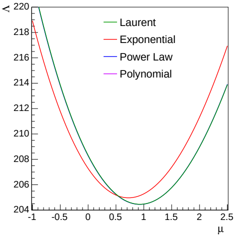
A more realistic example

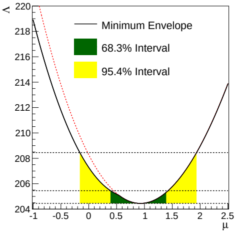
- A small signal component
- Some realistic (and one unrealistic) background models
- Do a profile scan for each model and take the envelope
  - Choices which are very similar have no effect (**Laurent** and **Power Law**)
  - Choices which are bad have no effect (**Polynomial**)
  - Choices which compete increase the uncertainty (**Exponential**)
- Uncertainty is increased if models are different
- *NOTE:* No explicit model choice has to be made
  - We don't actually care what model "is the best"



Result:

- A best fit value ✔
- A confidence interval ✔
- A systematic from the model choice ✔

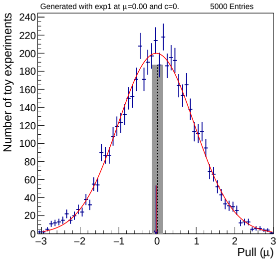




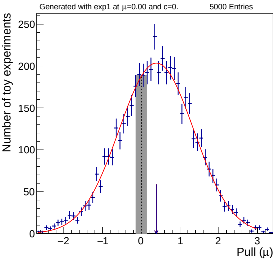
Bias and Coverage properites

- Generate toy MC from various background hypotheses and then refit to asses the bias (using the pull) and the coverage
- For example generate with exponential background distribution:
- Grey band shows 14% of statistical uncertainty

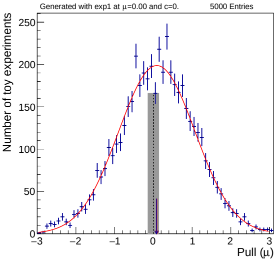
Fit back with exponential



Fit back with power law



Fit back with envelope

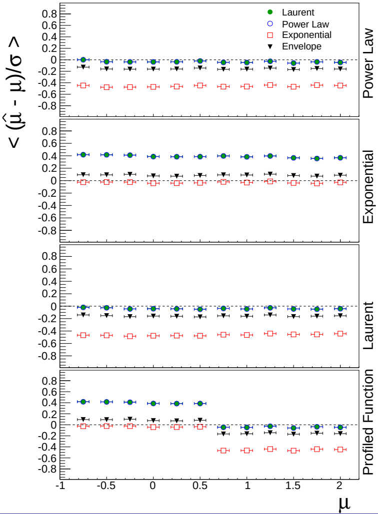


Bias and Coverage properites

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

Bias:

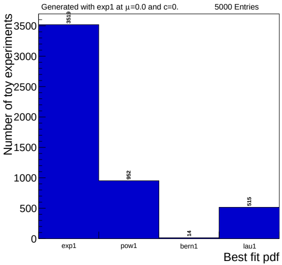
- When you generate and fit back with *the same* (or similar) background function the bias is neglible (**green points** in top panel, **red points** in second panel)
- When you generate and fit back with *different* functions the bias is large (**red points** in top panel, **green points** in second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



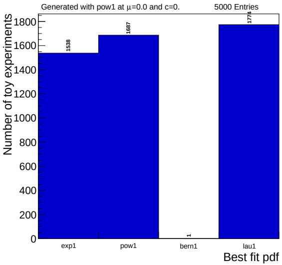
Which PDF fits best?

- Can assess toys to see which PDF minimises the envelope

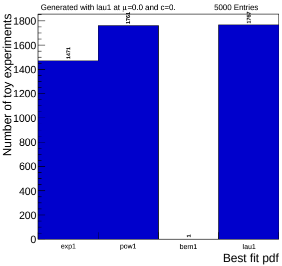
Generated with exponential



Generated with power law



Generated with Laurent

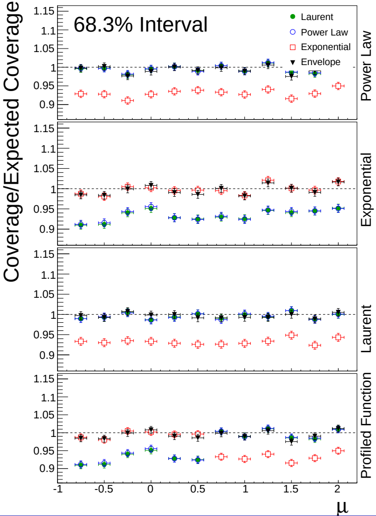


Bias and Coverage properites

- 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**

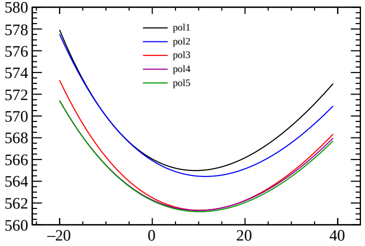


4. Different degrees of freedom

Hang on a minute...

- How do we compare models with different numbers of parameters?
  - In the combinatorial background case a single exponential and an 8th order polynomial are surely not on equal footing?
- The value of $\Lambda = -2LL$ is simply a measure of how well the data agrees with a particular probability distribution
  - It does not account for degrees of freedom
- Consequently using Λ *without any penalty* would alway result in choosing the highest order model(s) available [i]
- There is also no *natural* mechanism for ignoring higher and higher order functions [ii]

- The answers is to correct the Λ for this
  - It is not obvious by how much one should do this
  - There are several possibilities:
    1. Approximate *p*-value correction
    2. Exact *p*-value correction
    3. Aikaike information criteria (AIC)
    4. Bayesian information criteria (BIC)



[i] At least for nested families such as polynomials

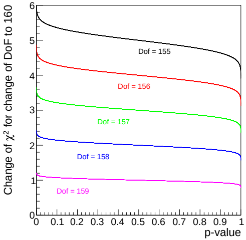
[ii] Fisher-test is however a possibility (although arbitrary)

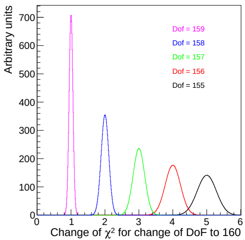
The p-value correction

- For binned fits, in the high statistics limit then $\Lambda \approx \chi^2$ and has corresponding $p(\chi^2, n_{bins} - n_{pars})$
- Can now find χ'^2 namely that which would have given the same p-value but with different degrees of freedom ($n_{pars} = 0$) and consequently,

$$\Lambda_{\rm corr} = \chi'^2 = \Lambda + (\chi'^2 - \chi^2) \quad (1)$$

- Correction depends on number of bins, number of parameters and quality of original fit [iii]





- Can be applied for specific p-value but also should note that on average:

$$\chi'^2 - \chi^2 \approx N_{\rm par} \quad \textit{so} \quad \Lambda_{\rm corr} \approx \Lambda + N_{\rm par} \quad (2)$$

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

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):

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

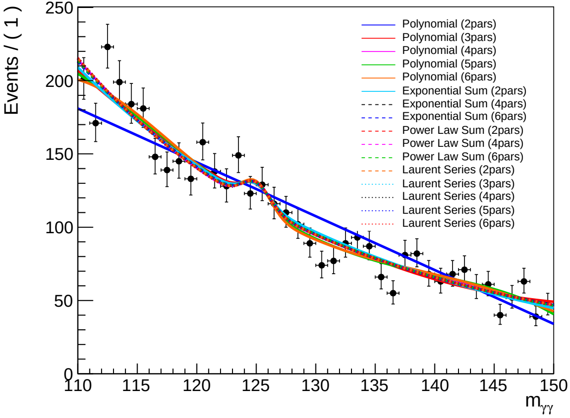
- In general they take the form:

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

where c is some "correction value" to be determined

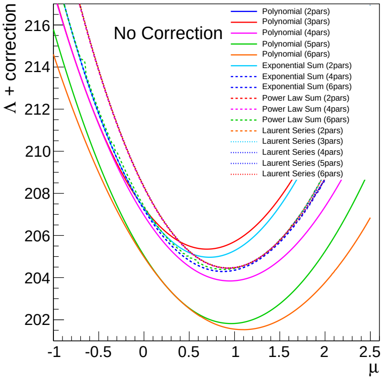
Back to the example case with higher order functions

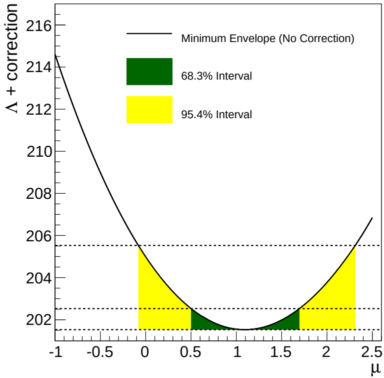
- Take the same dataset and now try many functions (of different orders)
- Scan the likelihoods as before now applying a correction, *c*, for different degrees of freedom



Example case with higher order functions

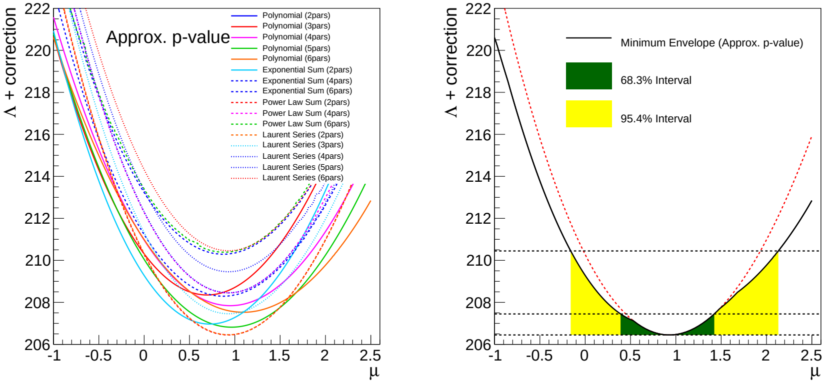
- Profile same dataset with many functions (of different orders)
- **With no correction,** $c = 0$
  - Best Fit: 6th order polynomial (highest order tried)





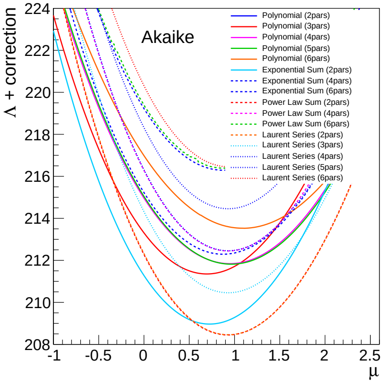
Example case with higher order functions

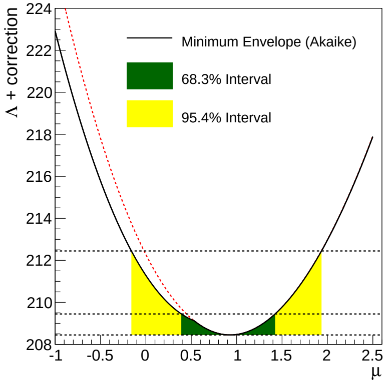
- Profile same dataset with many functions (of different orders)
- **With approx. *p*-value correction, $c = 1$** (Λ +1 per dof)
  - <u>Best Fit:</u> 2 parameter power law



Example case with higher order functions

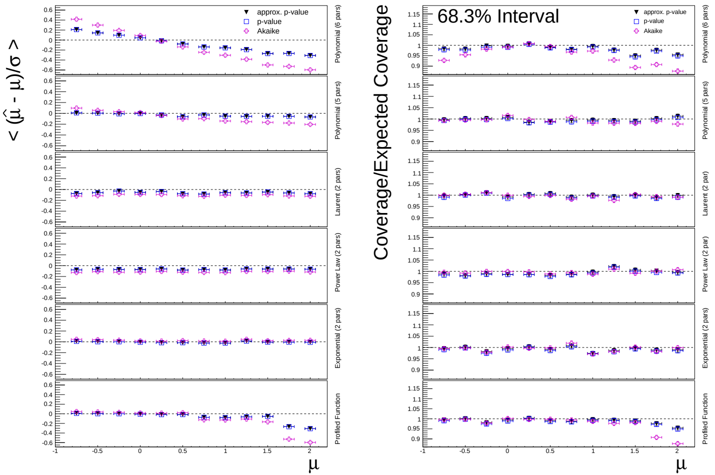
- Profile same dataset with many functions (of different orders)
- **With Aikaike correction, $c = 2$** (Λ +2 per dof)
  - <u>Best Fit:</u> 2 parameter power law





Bias and coverage for many order functions

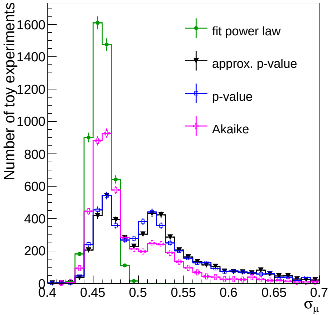
- Now comparing envelope of all functions with different correction schemes



5. How large a correction?

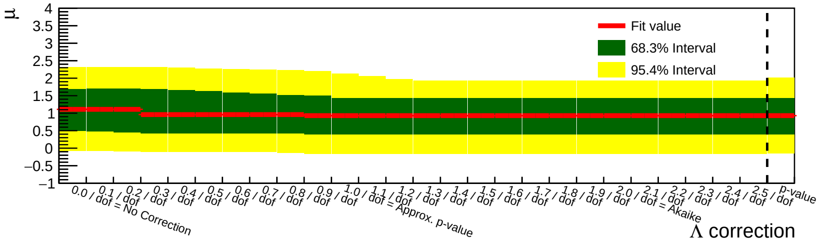
What happens to the error?

- 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



Central value and error dependence on the correction

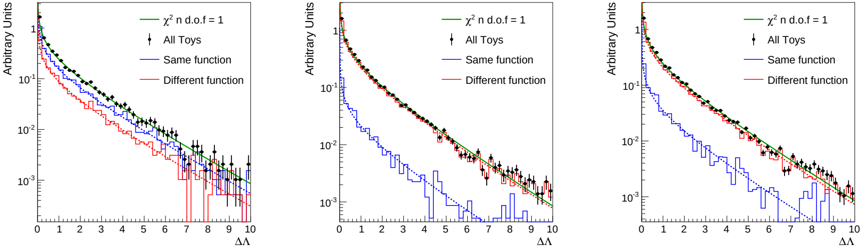
- As a function of the correction value the uncertainty (and central value) can change
- At lower values of c you have a large statistical uncertainty
  - In principle for this example if $c = 0$ the statistical error is infinite
- At larger values of c you have a potentially large bias



How reasonable is it to quote an uncertainty like this?

- Difference in Λ between the true and fitted values of μ follows a χ^2 distribution.

$c = 0$ $c = 1$ $c = 2$

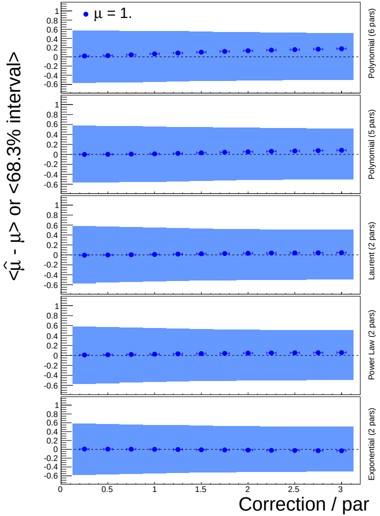


What correction to use?

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

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

- 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



6. Use cases

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

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

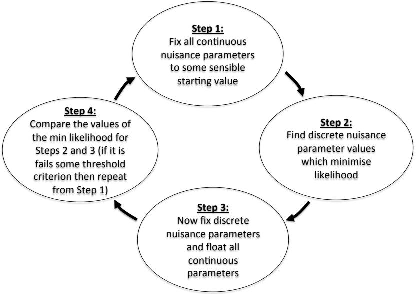
- For CMS $H \to \gamma\gamma = 16^{25} \approx 10^{30}$ combinations
- For any reasonable practical use this has to be reduced

Technical implementation

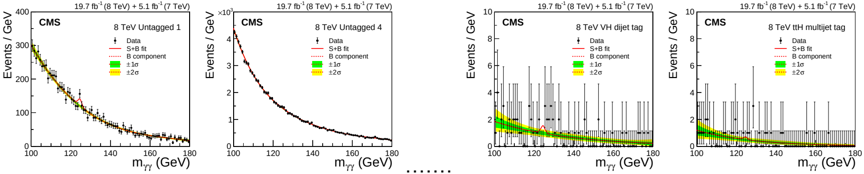
- These studies were developed and performed in `RooFit`
  - Specialised class written: `RooMultiPdf`
  - Not in `RooFit` public release yet
  - Private version being used by both CMS and ATLAS
- **How to reduce numbers of combinations** (given 10^{30} minimisations is impractical for Higgs combination)
  - 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

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

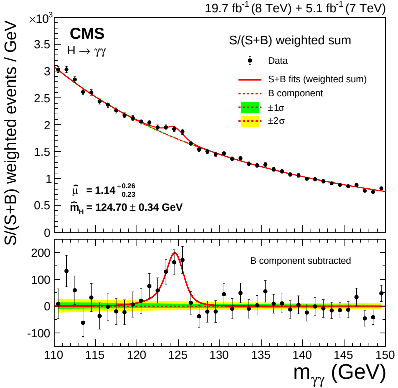
for N_I iterations

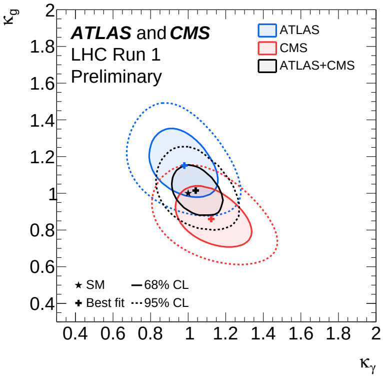


Use in Higgs analyses



....

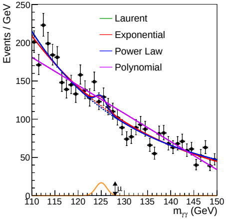


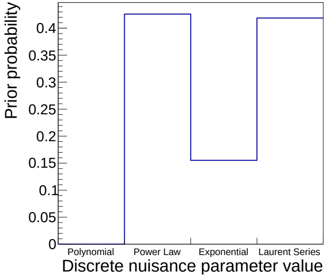


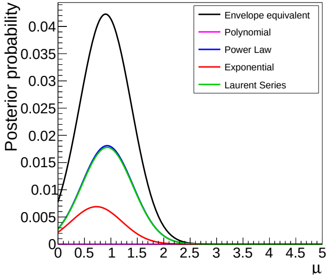
7. The Bayesian way

Bayesian formalism

- So far the method discussed has been in a frequentist formalism
- Work ongoing to publish a Bayesian equivalent
- The "discrete" profiling equates to adding up posterior PDFs each with a weight $\sim e^{-\chi^2}$







8. Extensions and Open Questions

Extensions and Open Questions

- Studies with mixed functions
  - Given a comparison of two functions of the form e^{-px} and x^{-p} does it make sense to try $fe^{-p_1 x} + (1-f)x^{-p_2}$?
  - This is then 3 free parameters not 1. Does the correction handle this appropriately?
- Is there an analytical proof of which correction to use?
- How should one assess how many "model" choices is appropriate?
- Are there other ways of sampling more of the "*model phase space*" cheaply?
- Can one "*interpolate*" gaps in the discontinuous profiles?
- What happens in very non-Gaussian situations?
- Are there fairer ways of generating MC from mixed model hypotheses?
  - How does one generate an "*Asimov*" toy from a composite model?
- 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

9. Summary

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 a real life case
  - Small bias and good coverage shown under several scenarios
  - Lead to improvements in technical implementation and recommendations for use
- Similar studies are highly recommended for each use case
- Several possible extensions and open questions

Thanks for your attention!