Discrete Profiling (The Envelope Method)

Experimental Perspective

Matt Kenzie

University of Cambridge

PHYSTAT informal review

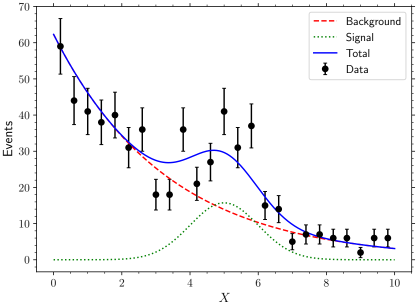
Wednesday 4$^{\text{th}}$ December

The physics problem

- Dataset, X_i, comes from some underlying distribution which is a composite of
  - Background - usually flat(ish) or smoothly falling
  - Signal - usually peaking
- Normally interested in the **properties of the signal**

$$p(X;\mu,\vec{\theta}) = \mu \underbrace{s(X;m_0,\sigma)}_{\text{signal}} + N_b \underbrace{b(X;\lambda)}_{\text{background}}$$

  - Signal strength, μ, or peak position, m_0
- Don't care about the background parameters, λ, nor its *parameterisation*, $b(X)$
  - Although *we do care* about their contribution to signal parameter uncertainties



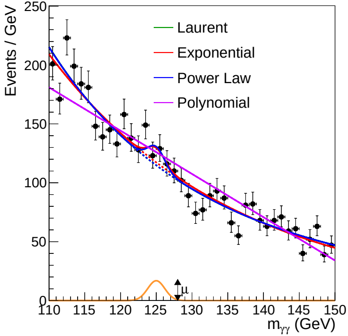
- Our models contain *parameters of interest* (POIs)
- And often contain several *nuisance parameters*
- Normally we *profile* over them in likelihood fits
- **BUT** what if we don't know the underlying p.d.f (*i.e. functional form*) of the model or part of the model?
- This is what we tried to address in our paper [JINST 10 P04015]
  - The specific application was the search for the Higgs boson

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

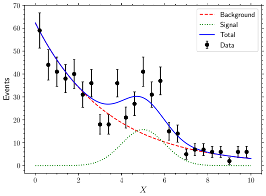
ABSTRACT: A common problem in data analysis is that the functional form, as well as the parameter values, of the underlying model which should describe a dataset is not known *a priori*. In these cases some extra uncertainty must be assigned to the extracted parameters of interest due to lack of exact 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.



Conceptualisation of a nuisance parameter

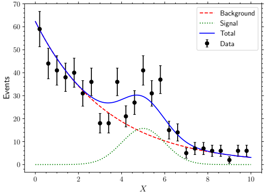
Using the example above

- Smoothly falling background, $b(X;\lambda) = \lambda e^{-\lambda X}$
  - *Nuisance parameter, λ*
- Peaking signal, $s(X; m_0, \sigma) = \mathcal{N}(m_0, \sigma)$
  - *Parameter of interest, μ*



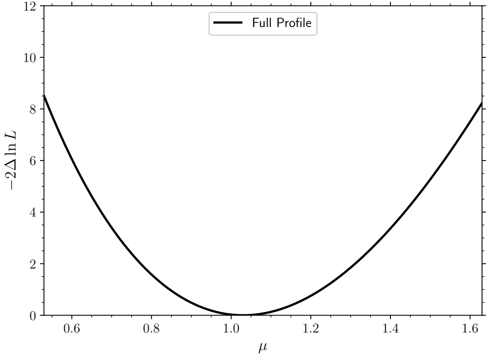
Using the example above

- Smoothly falling background, $b(X;\lambda) = \lambda e^{-\lambda X}$
  - *Nuisance parameter, λ*
- Peaking signal, $s(X;m_0,\sigma) = \mathcal{N}(m_0,\sigma)$
  - *Parameter of interest, μ*



Inspect the profiled $-2\Delta \ln L(\mu)$

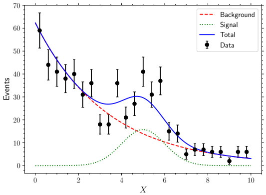
- **with λ floating**



Conceptualisation of a nuisance parameter

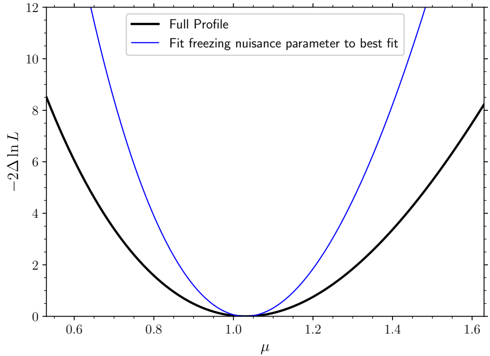
Using the example above

- Smoothly falling background, $b(X;\lambda) = \lambda e^{-\lambda X}$
    - *Nuisance parameter, λ*
- Peaking signal, $s(X; m_0, \sigma) = \mathcal{N}(m_0, \sigma)$
    - *Parameter of interest, μ*



Inspect the profiled $-2\Delta \ln L(\mu)$

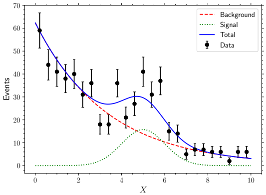
- **with λ floating**
- with λ fixed to its best fit value



Conceptualisation of a nuisance parameter

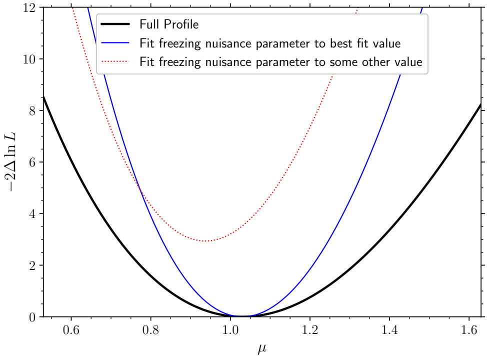
Using the example above

- Smoothly falling background, $b(X; \lambda) = \lambda e^{-\lambda X}$
  - *Nuisance parameter, λ*
- Peaking signal, $s(X; m_0, \sigma) = \mathcal{N}(m_0, \sigma)$
  - *Parameter of interest, μ*



Inspect the profiled $-2\Delta \ln L(\mu)$

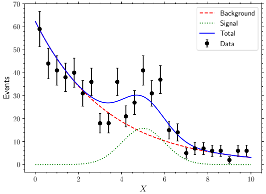
- **with λ floating**
- with λ fixed to its best fit value
- with λ fixed to other values



Conceptualisation of a nuisance parameter

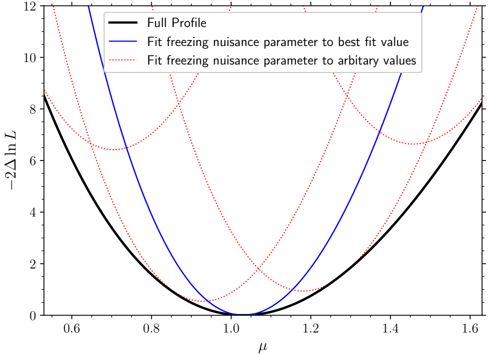
Using the example above

- Smoothly falling background, $b(X;\lambda) = \lambda e^{-\lambda X}$
  - *Nuisance parameter, λ*
- Peaking signal, $s(X; m_0, \sigma) = \mathcal{N}(m_0, \sigma)$
  - *Parameter of interest, μ*



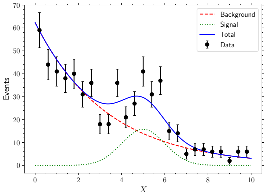
Inspect the profiled $-2\Delta \ln L(\mu)$

- **with λ floating**
- with λ fixed to its best fit value
- with λ fixed to other values



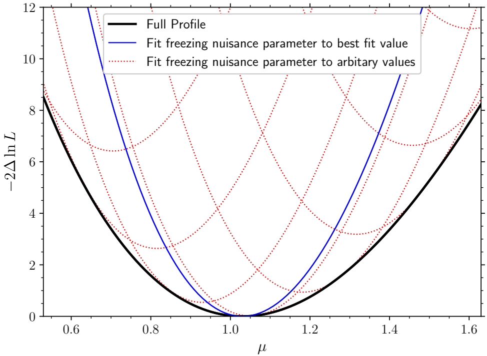
Using the example above

- Smoothly falling background, $b(X;\lambda) = \lambda e^{-\lambda X}$
  - *Nuisance parameter, λ*
- Peaking signal, $s(X;m_0,\sigma) = \mathcal{N}(m_0,\sigma)$
  - *Parameter of interest, μ*



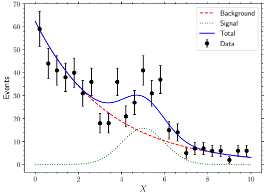
Inspect the profiled $-2\Delta \ln L(\mu)$

- **with λ floating**
- with λ fixed to its best fit value
- with λ fixed to other values



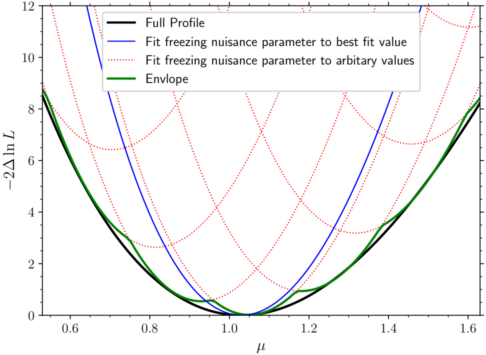
Using the example above

- Smoothly falling background, $b(X;\lambda) = \lambda e^{-\lambda X}$
  - *Nuisance parameter, λ*
- Peaking signal, $s(X; m_0, \sigma) = \mathcal{N}(m_0, \sigma)$
  - *Parameter of interest, μ*



Inspect the profiled $-2\Delta \ln L(\mu)$

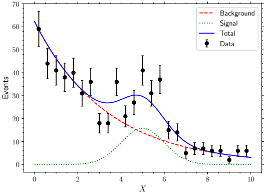
- **with λ floating**
- with λ fixed to its best fit value
- with λ fixed to other values
- draw the minimum "envelope"



Conceptualisation of a nuisance parameter

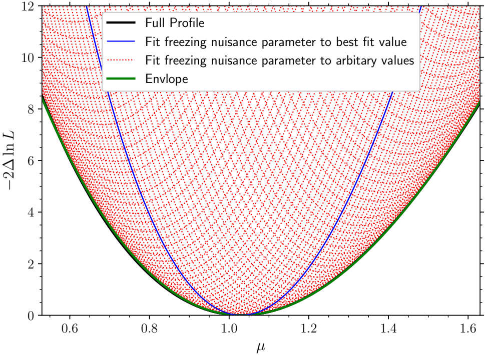
Using the example above

- Smoothly falling background, $b(X;\lambda) = \lambda e^{-\lambda X}$
  - *Nuisance parameter, λ*
- Peaking signal, $s(X;m_0,\sigma) = \mathcal{N}(m_0,\sigma)$
  - *Parameter of interest, μ*



Inspect the profiled $-2\Delta \ln L(\mu)$

- **with λ floating**
- with λ fixed to its best fit value
- with λ fixed to other values
- draw the minimum "envelope"
- eventually the "envelope" $\rightarrow$ the full profile

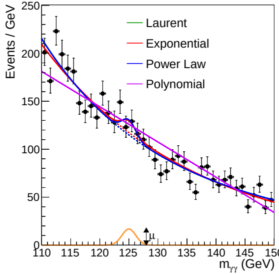


Module choice as a discrete nuisance parameter

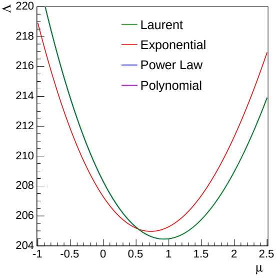
- The choice of underlying model can be treated as a discrete nuisance parameter in this way
- Profile over all of them and find the *minimum envelope*
- Gives me *freedom* over several choices and allows me to
  - Pick the model that "fits best" (as it will maximise the likelihood)
  - Compute an uncertainty related to the *model choice*
- Question for the statisticians: Is the space of model choices infinite / is this imagined nuisance parameter really discrete valued?

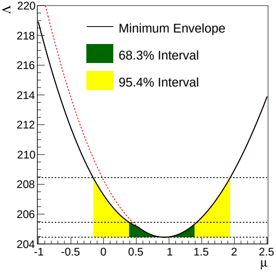
A (slightly) more realistic example

- The example from our paper is inspired by the Higgs search
- Small *signal* on a large smoothly falling *background*
- A few realistic (and one unrealistic) background models
  - Choices which are similar overlap (**Laurent** and **Power Law**)
  - Choices which are bad have no effect (**Polynomial**)
  - Choices which compete increase the uncertainty (**Exponential**)

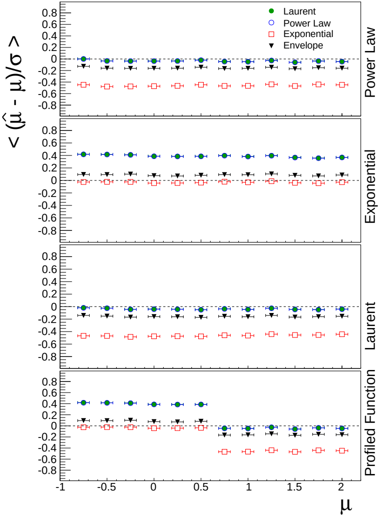


- Uncertainty is increased if models are different
- No explicit model choice has to be made

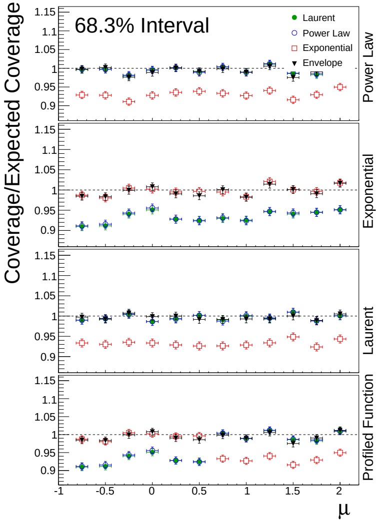




- Generate samples from different background hypotheses and refit



Small biases

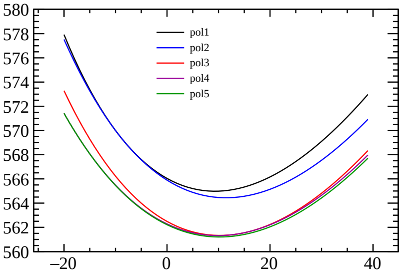


Good coverage

Ah but wait...

- All seems hunky dory but what about models with *different numbers of parameters*
- The likelihood only measures agreement of data with model
  - Does not account for *degrees of freedom*
- Without any kind of *regularisation* would always choose the model with the *most freedom* [i]
- No *natural mechanism* for ignoring higher order functions [ii]
  - Question for the statisicians: when can we stop adding functions to try?

- Our solution is to *correct (regularise) the likelihood*
  - Not obvious by how much
  - 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 like polynomials
[ii] Maybe something like a Fisher test?

What correction term?

- From *Wilks' theorem*, as $N \to \infty$, then $-2\Delta \ln L \to \chi^2$ with $p(\chi^2, n_{\text{bins}} - n_{\text{pars}})$
- Find χ'^2 which would have given same p-value but with different degrees of freedom

$$-2\Delta \ln L_{\text{corr}} = \chi'^2 = -2\Delta \ln L + (\chi'^2 - \chi^2)$$

- On average $\chi'^2 - \chi^2 \approx N_{\text{par}}$ and therefore *p-value* correction

$$-2\Delta \ln L_{\text{corr}} = -2\Delta \ln L + N_{\text{par}}$$

- Other options are available
- Aikaike information criterion (AIC):

$$-2\Delta \ln L_{\text{corr}} = -2\Delta \ln L + 2N_{\text{par}}$$

- Bayesian information criterion (BIC):

$$-2\Delta \ln L_{\text{corr}} = -2\Delta \ln L + N_{\text{par}} \ln(n)$$

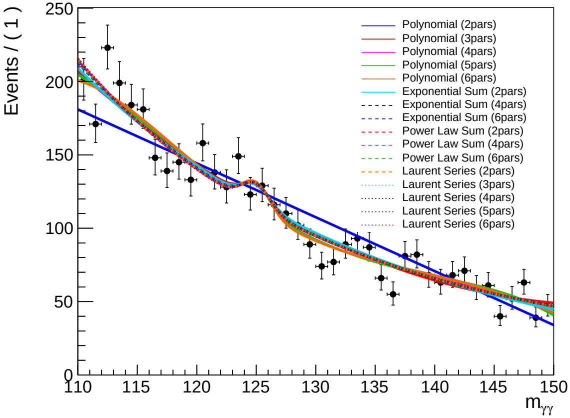
- In general the correction takes the form

$$-2\Delta \ln L_{\text{corr}} = -2\Delta \ln L + cN_{\text{par}}$$

where c is some "correction value" to be determined **by the user** based on the use case and desired *bias / variance* trade-off

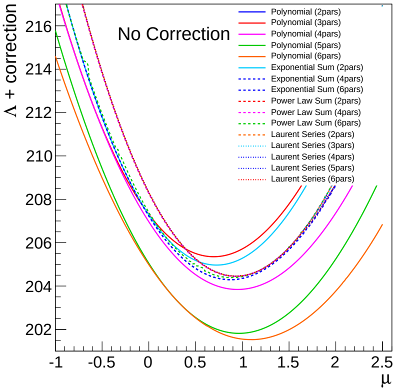
The example case with higher order functions

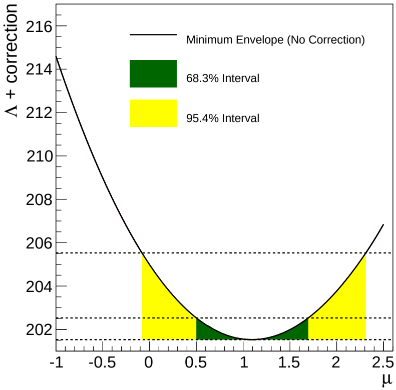
- Take the same dataset and try many functions (of different orders)
- Profile the likelihood as before investigating different corrections



The example case with higher order functions

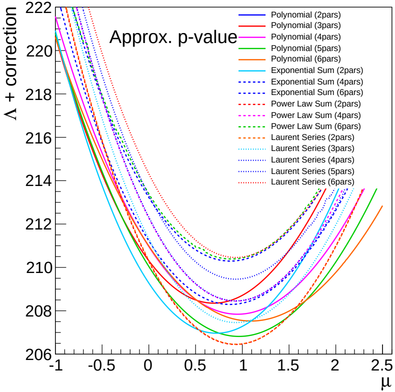
- **With no correction**, $c = 0$
- <u>Best Fit:</u> 6th order polynomial (highest order tried)

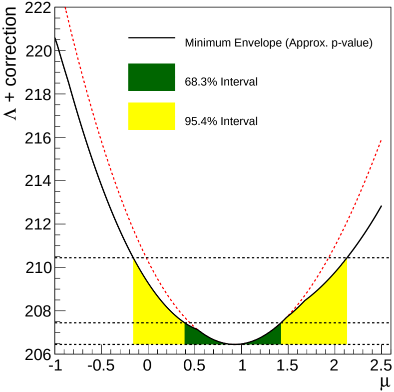




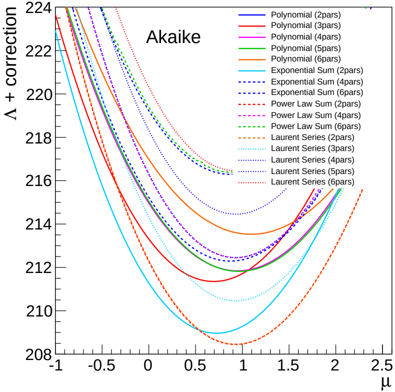
The example case with higher order functions

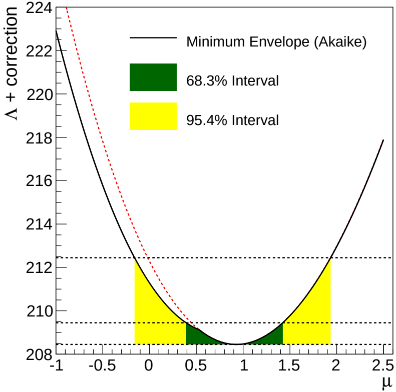
- **With p-value correction**, $c = 1$ ($\Lambda + 1/$d.o.f)
- Best Fit: 2 parameter power law





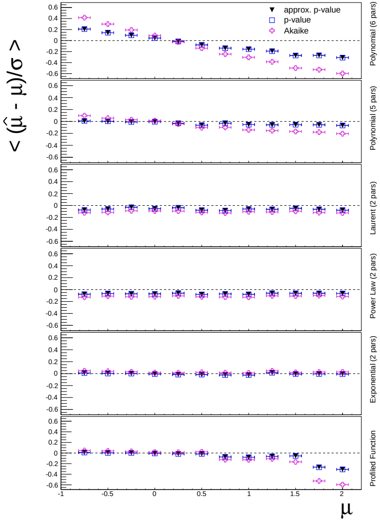
- **With Aikaike correction**, $c = 2$ $(\Lambda + 2/\mathrm{d.o.f})$
- Best Fit: 2 parameter power law

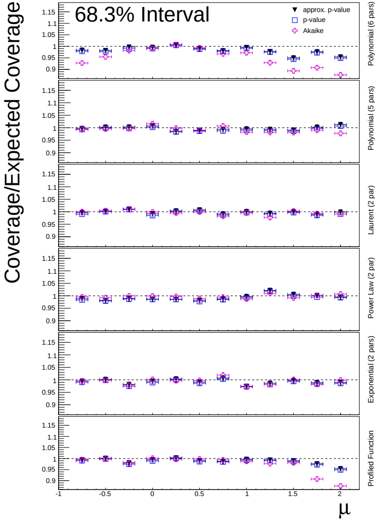




Bias and coverage properties

- Generate samples from different background hypotheses and refit
- Bias and coverage properties of AIC considerably worse in this case

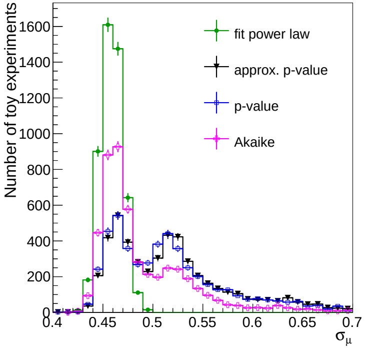




What happens to the error?

- For ensembles of samples 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
- **BUT** at lower values of c the statistical uncertainty is larger
  - In principle if **every** function is allowed it is *infinite*
- **ON THE OTHER HAND** at large values of c the bias gets larger

- So the user has a choice *bias or variance*?
- Question for the statisicians: which correction should we use?



Other questions

- Studies with *mixed functions*
  - With two functions e.g. e^{-px} and x^{-p} does it make sense to try $fe^{-p_1 x} + (1-f)x^{-p_2}$?
  - 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?
- 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 prior should be used?
- How do you decide *how many orders* to include in the envelope if the choice is infinitely many?
  - Fisher test?

BACK UP