Practical Statistics – part III
‘*Dealing with systematic uncertainties*’

W. Verkerke (NIKHEF)

Next subject…

- Start with basics, gradually build up to complexity of

Statistical tests with simple hypotheses for counting data — *"What do we mean with probabilities"*

↓

Statistical tests with simple hypotheses for distributions — *"p-values"*

↓

Hypothesis testing as basis for event selection — *"Optimal event selection & machine learning"*

↓

Composite hypotheses (with parameters) for distributions — *"Confidence intervals, Maximum Likelihood"*

↓

Statistical inference with nuisance parameters — *"Fitting the background"*

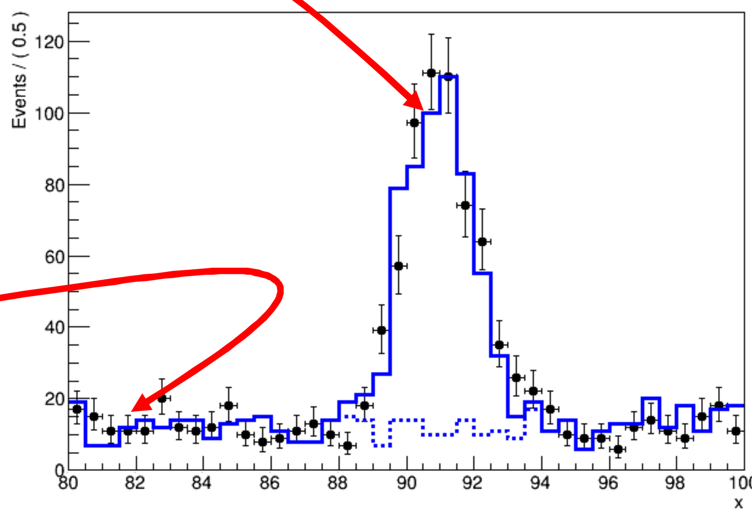
↓

Response functions and subsidiary measurements — *"Sideband fits and systematic uncertainties"*

The *ideal* experiment

- The "only thing" you need to do (as an experimental physicist) is to formulate the likelihood function for your measurement
- For an ideal experiment, where signal and background are assumed to have perfectly known properties, this is trivial

$$L(\vec{N} \mid \mu) = \prod_{bins} Poisson(N_i \mid \mu \cdot \tilde{s}_i + \tilde{b}_i)$$



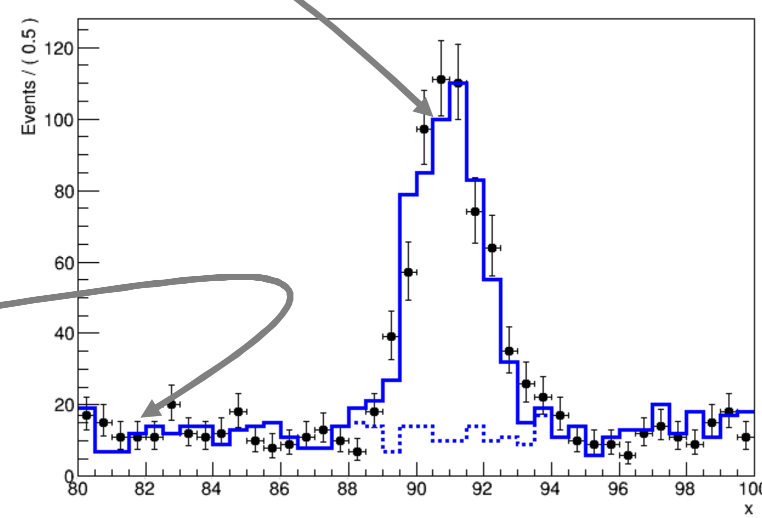
- Only a single* parameter in the likelihood:
  the physics *parameter of interest,* usually denoted as *μ*

**Unless there are of course multiple POIs…*

The imperfect experiment

- In realistic measurements many effect that we don't control exactly influence measurements of parameter of interest
- How do you model these uncertainties in the likelihood?

$$L(\vec{N} \mid \mu) = \prod_{bins} Poisson(N_i \mid \mu \cdot \tilde{s}_i + \tilde{b}_i)$$

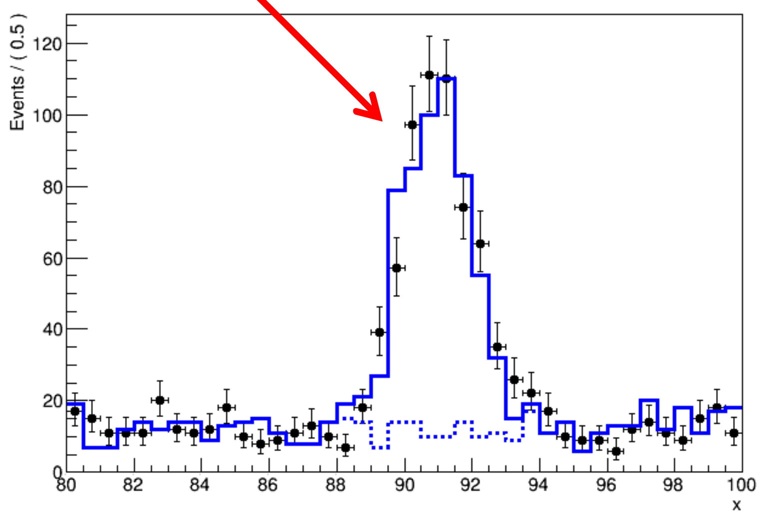


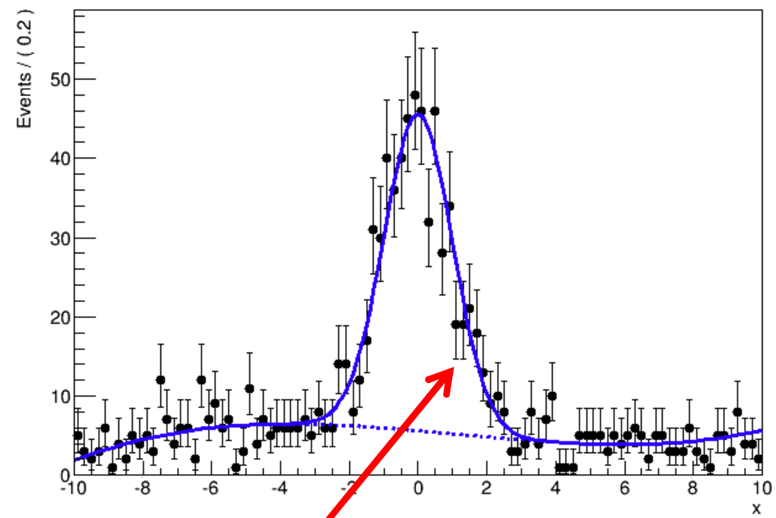
Signal and background predictions are affected by (systematic) uncertainties

Adding parameters to the model

- But parametric form of detector and theory systematic uncertainties is often, at first sight, elusive

$$L(\vec{N} \mid \mu) = \prod_{bins} Poisson(N_i \mid \mu \cdot \tilde{s}_i + \tilde{b}_i)$$

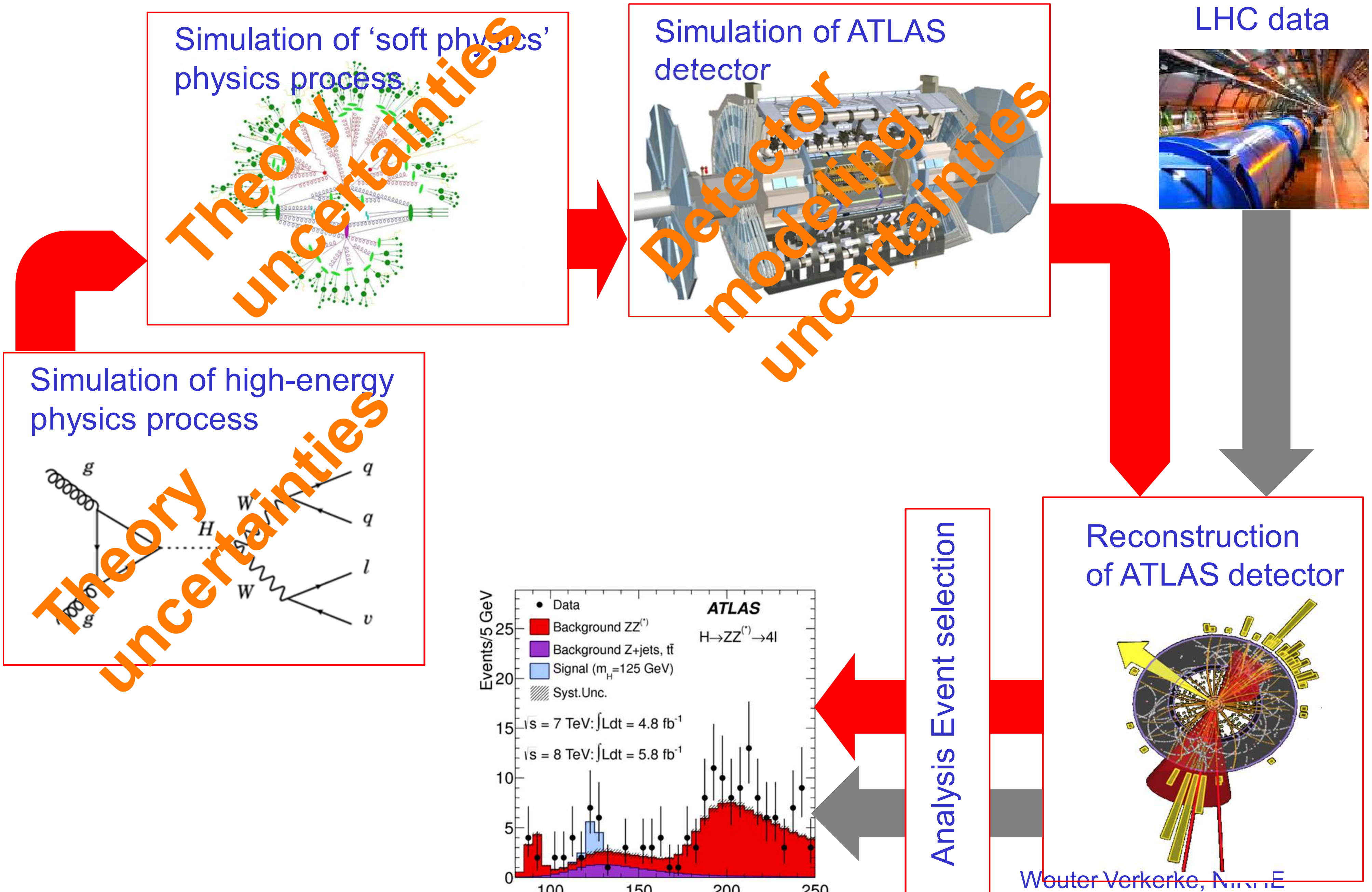




$$L(x \mid f, m, \sigma, a_0, a_1, a_2) = fG(x, m, \sigma) + (1-f)Poly(x, a_0, a_1, a_2)$$

- Ad-hoc parameterizations (like above) do not necessarily capture all uncertain degrees of freedom, provide no meaningful insight in effect of known systematic uncertainties on the analysis.

The simulation workflow and origin of uncertainties



Typical systematic uncertainties in HEP

- Detector-simulation related
  - "The Jet Energy scale uncertainty is 5%"
  - "The b-tagging efficiency uncertainty is 20% for jets with p_T<40"

- Physics/Theory related
  - The top cross-section uncertainty is 8%
  - "Vary the factorization scale by a factor 0.5 and 2.0 and consider the difference the systematic uncertainty"
  - "Evaluate the effect of using Herwig and Pythia and consider the difference the systematic uncertainty"

- MC simulation statistical uncertainty
  - Effect of (bin-by-bin) statistical uncertainties in MC samples

What can you do with *systematic* uncertainties

- As most of the typical systematic prescriptions have no immediately apparent parametric formulation in your likelihood, common approach is 'vary setting, rerun analysis, observe the difference'
- This common 'naïve' approach to assess effect of systematic uncertainties amounts to simple error propagation
- Error propagation procedure in a nutshell
  - Make nominal measurement (using your favorite statistical inference procedure)
  - Change setting in detector simulation or theory (e.g. shift Jet Calibration scale by '1 sigma' up and down) Redo measurement procedure for each shift
  - Consider propagated effect of shifted setting the systematic uncertainty

$$\mu = \underbrace{\mu_{nom} \pm \sigma_{stat}}_{\text{From statistical analysis}} \pm \underbrace{(\mu_{syst}^{up} - \mu_{syst}^{down})}_{\text{Systematic uncertainty from error propagation}} / 2 \pm \ldots$$

From statistical analysis

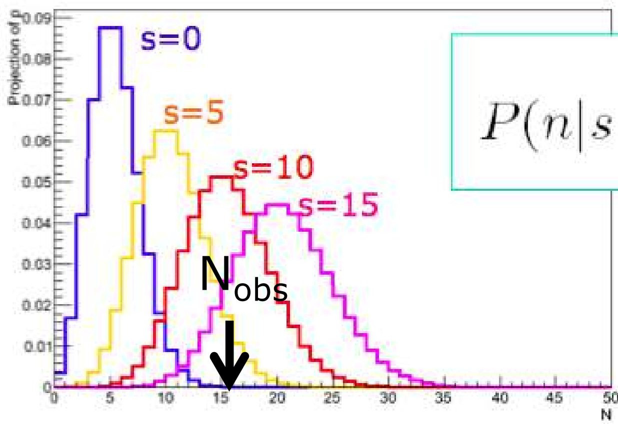
Systematic uncertainty from error propagation

Pros and cons of the 'naïve' approach

- **Pros**
  - It's easy to do
  - It results in a seemingly easy-to-interpret table of systematics
- **Cons**
  - Uncorrelated source of systematic uncertainty can have correlated effect on measurement Completely ignored
  - Magnitude of stated systematic uncertainty may be incompatible with measurement result Completely ignored
  - You lost the connection with fundamental statistical techniques (i.e. evaluation of systematic uncertainties is completely detached from statistical procedure used to estimate physics quantity of interest) No prescription to make confidence intervals, Bayesian posteriors etc in this way
  - No calibrated probabilistic statements possible (95% C.L.)
- 'Profiling' Incorporate a description of systematic uncertainties in the likelihood function that is used in statistical procedures

Everything starts with the likelihood

- **All** fundamental statistical procedures are based on the likelihood function as 'description of the measurement'



$$P(n|s+b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

NB: b is a constant in this example

Definition: the Likelihood is P(observed data|theory)

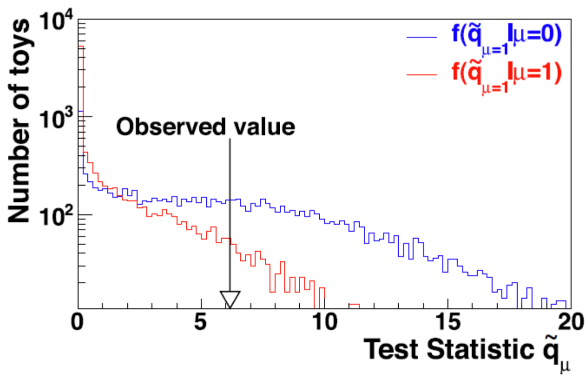
e.g. L(15|s=0)

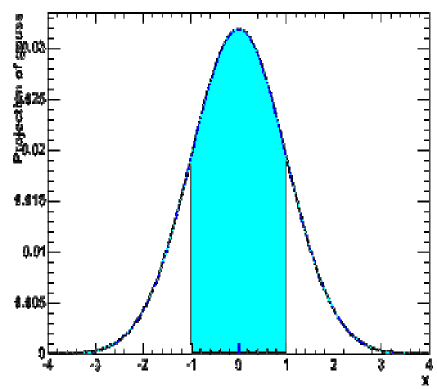
e.g. L(15|s=10)

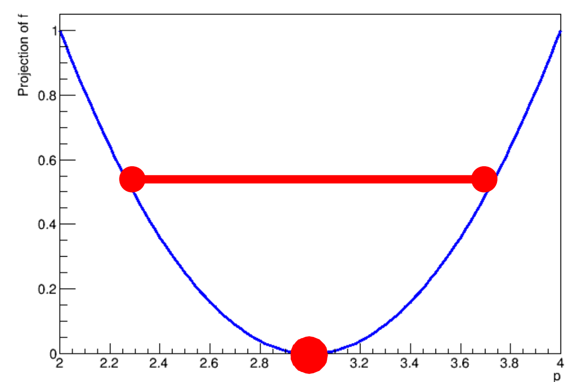
Frequentist statistics	Bayesian statistics	Maximum Likelihood
Confidence interval on s	Posterior on s	s = x ± y

Everything starts with the likelihood

Frequentist statistics	Bayesian statistics	Maximum Likelihood
$\lambda_\mu(N_{obs}) = \frac{L(N \mid \mu)}{L(N \mid \hat{\mu})}$	$P(\mu) \propto L(x \mid \mu) \cdot \pi(\mu)$	$\left.\frac{d \ln L(p)}{dp}\right\|_{p_i = \hat{p}_i} = 0$



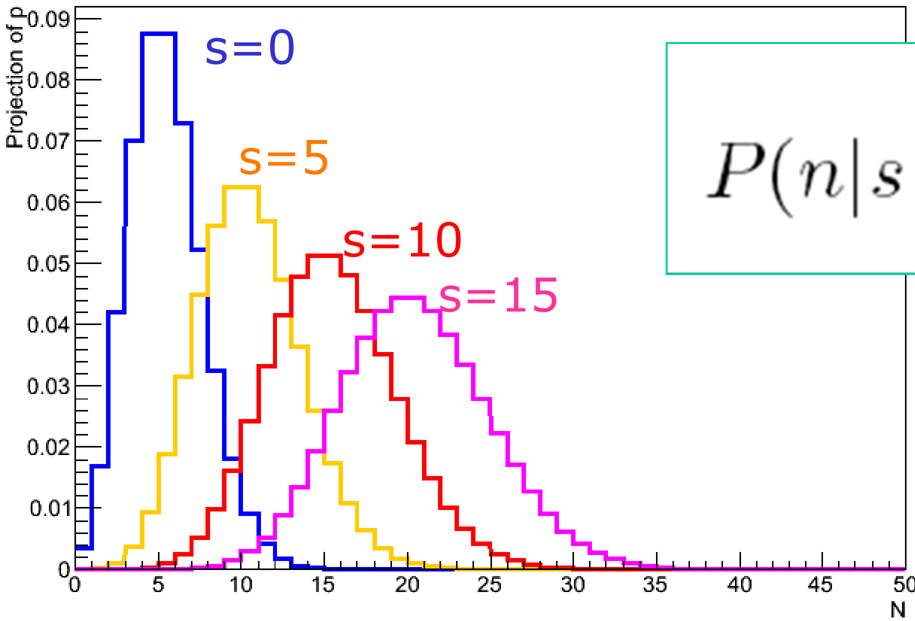




Confidence interval or p-value	Posterior on s or Bayes factor	s = x ± y

Introducing uncertainties – a non-systematic example

- The original model (with fixed *b*)



$$P(n|s+b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

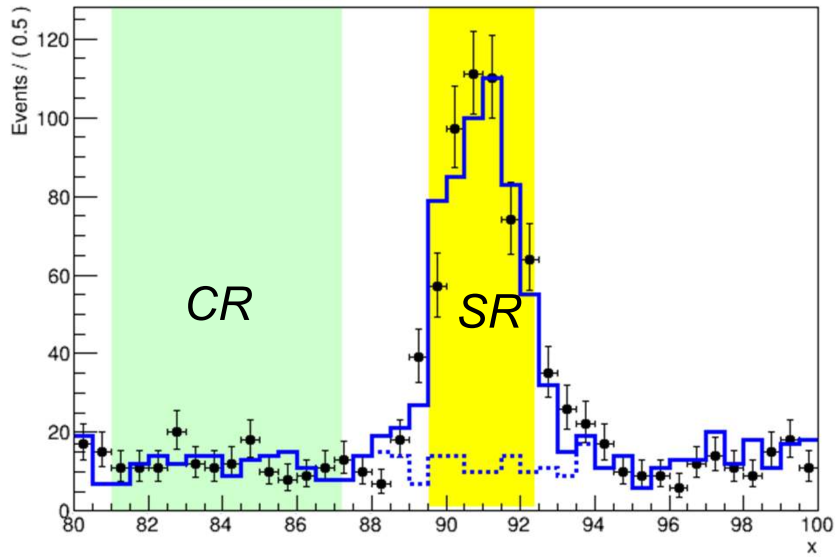
- Now consider *b* to be uncertain

L(N|s) → L(N|s,b)

- The experimental data contains insufficient to constrain both *s* and *b* → Need to add an additional measurement to constrain *b*

The sideband measurement

- Suppose your data
  in reality looks like this



Can estimate level of background in the 'signal region' from event count in a 'control region' elsewhere in phase space

$$L_{SR}(s,b) = Poisson(N_{SR} \mid s+b)$$

NB: Define parameter 'b' to represents the amount of bkg is the SR.

$$L_{CR}(b) = Poisson(N_{CR} \mid \tilde{\tau} \cdot b)$$

Scale factor τ accounts for difference in size between SR and CR

"Background uncertainty constrained from the data"

- Full likelihood of the measurement ('simultaneous fit')

$$L_{full}(s,b) = Poisson(N_{SR} \mid s+b) \cdot Poisson(N_{CR} \mid \tilde{\tau} \cdot b)$$

Generalizing the concept of the sideband measurement

- Background uncertainty from sideband clearly clearly not a 'systematic uncertainty'

$$L_{full}(s,b) = Poisson(N_{SR} \mid s+b) \cdot Poisson(N_{CR} \mid \tilde{\tau} \cdot b)$$

- Now consider scenario where *b* is not measured from a sideband, but is taken from MC simulation **with an 8% cross-section 'systematic' uncertainty**

'Measured background rate by MC simulation'

$$L_{full}(s,b) = Poisson(N_{SR} \mid s+b) \cdot \underbrace{Gauss(\tilde{b} \mid b, 0.08)}$$

'Subsidiary measurement' of background rate

  – *We can model this in the same way, because the cross-section uncertainty is also (ultimately) the result of a measurement*

  Generalize: 'sideband' → 'subsidiary measurement'

What is a systematic uncertainty?

- Concept & definitions of 'systematic uncertainties' originates from physics, not from fundamental statistical methodology.
  - E.g. Glen Cowans (excellent) 198pp book "statistical data analysis" does not discuss systematic uncertainties at all

- A common definition is
  - "Systematic uncertainties are all uncertainties that are not directly due to the statistics of the data"

- But the notion of 'the data' is a key source of ambiguity:
  - does it include control measurements?
  - does it include measurements that were used to perform basic (energy scale) calibrations?

Typical systematic uncertainties in HEP

- Detector-simulation related
  - "The Jet Energy scale uncertainty is 5%"
  - "The b-tagging efficiency uncertainty is 20% for jets with p_T<40"

Subsidiary measurement is an actual measurement
 conceptually similar to a 'sideband' fit

- Physics/Theory related
  - The top cross-section uncertainty is 8%
  - "Vary the factorization scale by a factor 0.5 and 2.0 and consider the difference the systematic uncertainty"
  - "Evaluate the effect of using Herwig and Pythia and consider the difference the systematic uncertainty"

Subsidiary measurement unclear, but origin of prescription may well be another measurement (if yes, like sideband, if no, what is source of info?)

- MC simulation statistical uncertainty
  - Effect of (bin-by-bin) statistical uncertainties in MC samples

Subsidiary measurement is a Poisson counting experiment (but now in MC events), otherwise conceptually identical to a 'sideband fit'

Typical systematic uncertainties in HEP

- Detector-simulation related
  - "The Jet Energy scale uncertainty is 5%"
  - "The b-tagging efficiency uncertainty is 20%

Subsidiary measurement is an actual measurement conceptually to

- P

Almost all systematic uncertainties are similar in nature to 'sidebands' measurements of some form or shape

 Can always model systematics like sidebands in the Likelihood

And even when the are not the (in)direct result of some measurement (certainty theory uncertainties) we can still model them in that form

)

- MC simulation statistical uncertainty
  - Effect of (bin-by-bin) statistical uncertainties in MC samples

Subsidiary measurement is a Poisson counting experiment (but now in MC events), otherwise conceptually identical to a 'sideband fit'

Modeling a detector calibration uncertainty

$$L_{full}(s,b) = Poisson(N_{SR} \mid s+b) \cdot Gauss(\tilde{b} \mid b, 0.08)$$

- Now consider a detector uncertainty, e.g. jet energy scale calibration, which can affect the analysis acceptance in a non-trivial way (unlike the cross-section example)

$$L(N,\tilde{\alpha} \mid s,\alpha) = Poisson(N \mid s + \tilde{b}(\alpha/\tilde{\alpha})\cdot 2)) \cdot Gauss(\tilde{\alpha} \mid \alpha, \sigma_{\alpha})$$

Signal rate (our parameter of interest)

Nominal calibration

Assumed calibration

Observed event count

Nominal background expectation from MC (a constant), obtained with a=a˜

Response function for JES uncertainty (a 1% JES change results in a 2% acceptance change)

Uncertainty on nominal calibration (here 5%)

"Subsidiary measurement" Encodes 'external knowledge' on JES calibration

Modeling a detector calibration uncertainty

- Simplify expression by renormalizing "subsidiary measurement"

$$Gauss(\tilde{\alpha}\,|\,\alpha,\sigma_{\alpha})$$

Signal rate (our parameter of interest)

$$L(N\,|\,s,\alpha) = Poisson(N\,|\,s + \tilde{b}(1+0.1\alpha)) \cdot Gauss(0\,|\,\alpha,1)$$

Observed event count

Nominal background expectation from MC (a constant)

Response function for normalized JES parameter
[a unit change in α – a 5% JES change – still results in a 10% acceptance change]

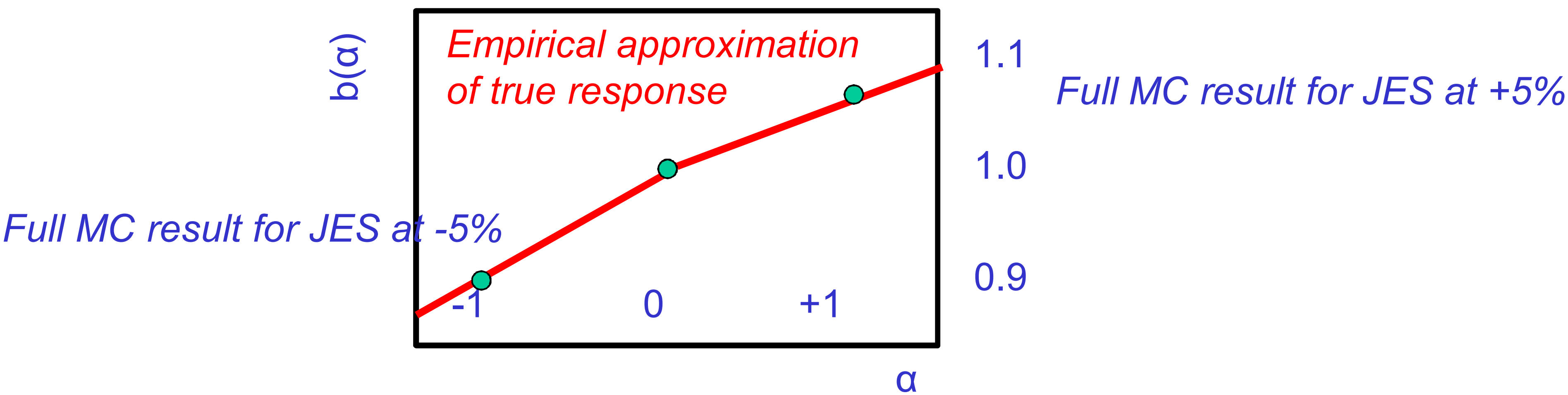
"Normalized subsidiary measurement"

The scale of parameter α is now chosen such that values ±1 corresponds to the nominal uncertainty (in this example 5%)

The response function as empirical model of full simulation

$$L(N,0|s,\alpha) = Poisson(N|s+\underbrace{b(\alpha)}) \cdot Gauss(0|\alpha,1)$$

- Note that the response function is generally not linear, but can in principle *always be determined by your full simulation chain*
  – But you cannot run your full simulation chain for any arbitrary 'systematic uncertainty variation' Too much time consuming
  – Typically, run full MC chain for nominal and ±1σ variation of systematic uncertainty, and approximate response for other values of NP with interpolation
  – For example run at nominal JES and with JES shifted up and down by ±5%



What is a systematic uncertainty?

- It is an uncertainty in the Likelihood of your physics measurement
  that is characterized deterministically, up to a set of parameters, of which the true value is unknown.

- A fully specified systematic uncertainty defines
  - 1: A set of one or more parameters
    of which the true value is unknown,
  - 2: A response model that describes the effect of those parameters on the measurement
    (sampled from full simulation, and interpolation)
  - 3: A subsidiary measurement of the parameters
    that constrains the values the parameters can take
    (implies a specific distribution: Gaussian *(default, CLT)*, Poisson *(low-stats counting)*, or otherwise)

Names and conventions – ‘profiling’ & ‘constraints’

- The full likelihood function of the form

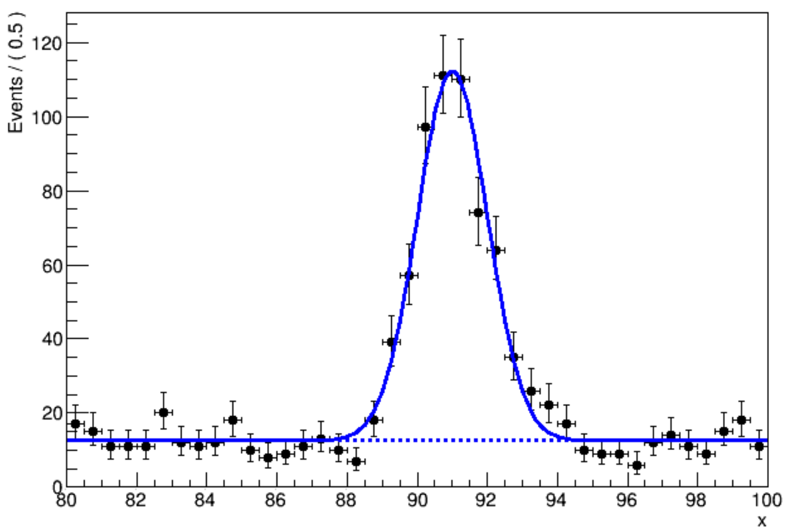
$$L(N,0\,|\,s,\alpha) = Poisson(N\,|\,s+b(\alpha))\cdot Gauss(0\,|\,\alpha,1)$$

  is usually referred to by physicists as a ‘profile likelihood’, and systematics are said to be ‘profiled’ when incorporated this way
  - Note: statisticians use the word profiling for something else
- Physicists often refer to the subsidiary measurement as a ‘constraint term’
  - This is correct in the sense that it constrains the parameter α, but this labeling commonly lead to mistaken statements (e.g. that it is a pdf for α)
  - But it is ***not*** a pdf in the NP

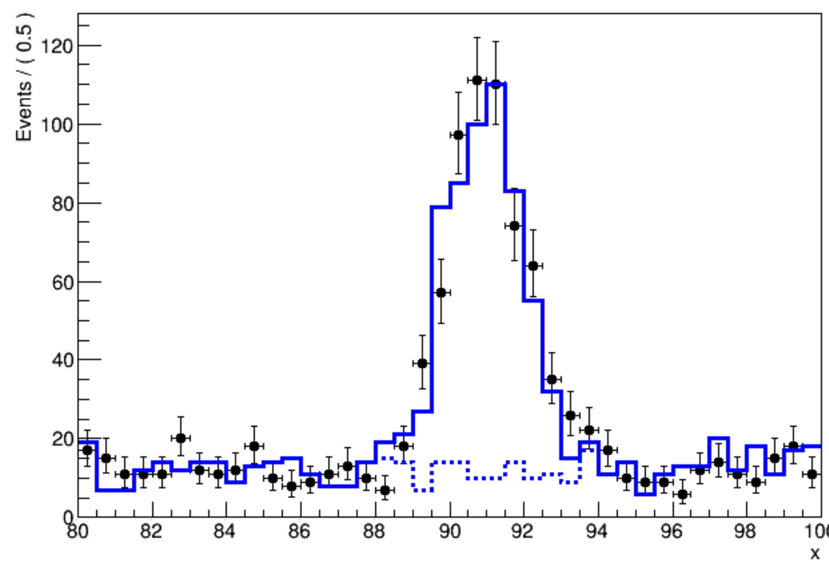
~~$Gauss(\alpha\,|\,0,1)$~~ $\qquad$ $Gauss(0\,|\,\alpha,1)$

Systematic uncertainties on shape fits

- What about systematic uncertainties on distributions?
  - So far illustrated systematics modeling with nuisance parameters in counting measurements for pedagogical reasons
- But same technique can be applied to Likelihood functions describing distributions ('shape fits')
  - Will focus on binned distributions – as these are most common at the LHC

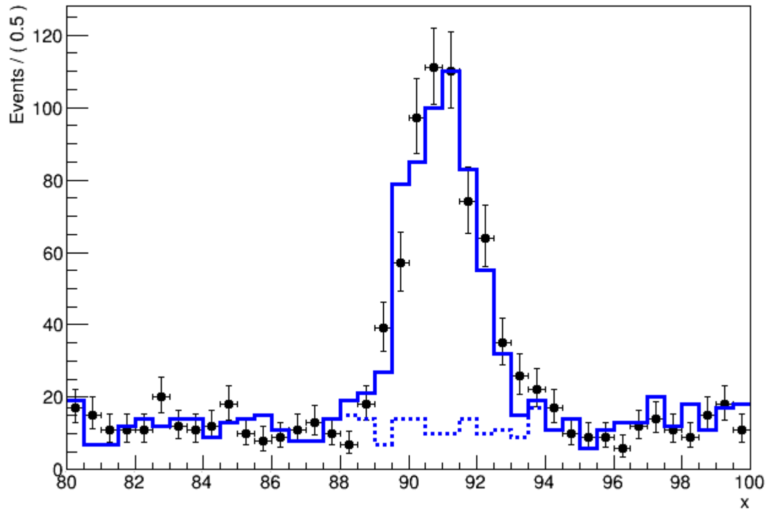


$$L(\vec{m}_{ll} \mid \mu) = \prod_i \left[\mu \cdot \mathrm{Gauss}(m_{ll}^{(i)}, 91, 1) + (1-\mu) \cdot \mathrm{Uniform}(m_{ll}^{(i)}) \right]$$



$$L(\vec{N} \mid \mu) = \prod_i Poisson(N_i \mid \mu s_i + b_i)$$

Response modeling for distributions



- For a change in the rate, response modeling of histogram-shaped distribution is straightforward: simply scale entire distribution

$$L(\vec{N}\,|\,\mu) = \prod_i Poisson(N_i\,|\,\mu\tilde{s}_i + \tilde{b}_i)$$

$$L(\vec{N}\,|\,\mu,\alpha) = \prod_i Poisson(N_i\,|\,\mu\tilde{s}_i \cdot \underbrace{(1+3.75\alpha)}_{\text{Response function for signal rate}} + \tilde{b}_i) \cdot \underbrace{Gauss(0\,|\,\alpha,1)}_{\text{Subsidiary measurement}}$$

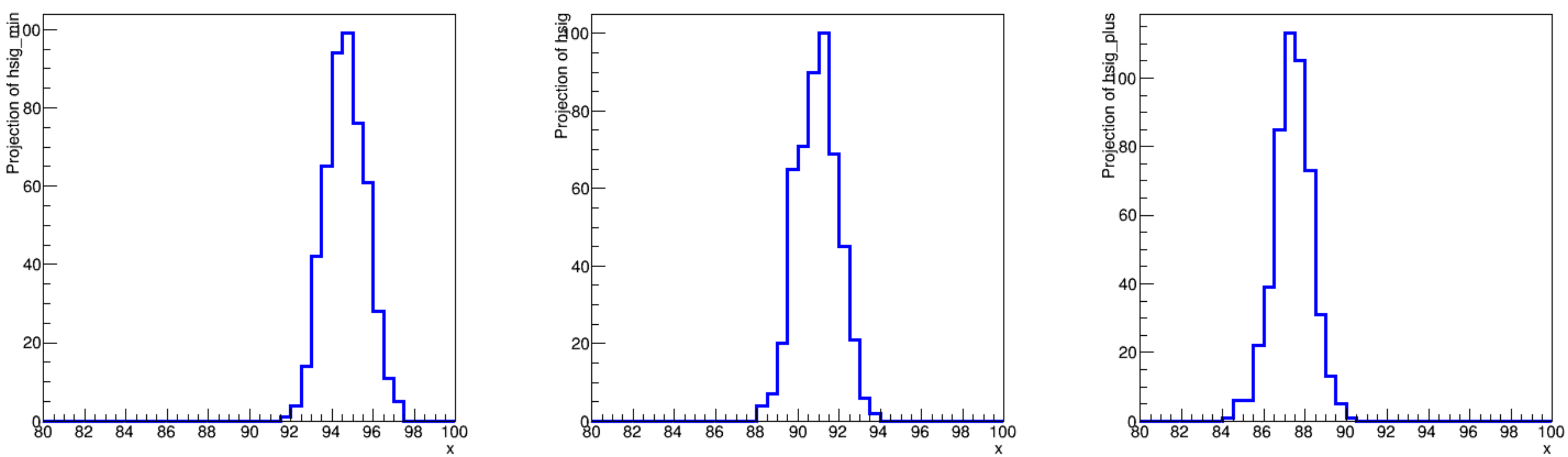
Response function for signal rate

Subsidiary measurement

- But what about a systematic uncertainty that shifts the mean, or affects the distribution in another way?

Modeling of shape systematics in the likelihood

- Effect of *any* systematic uncertainty that affects the shape of a distribution can in principle be obtained from MC simulation chain
  - Obtain histogram templates for distributions at ‘+1σ’ and ‘-1σ’ settings of systematic effect

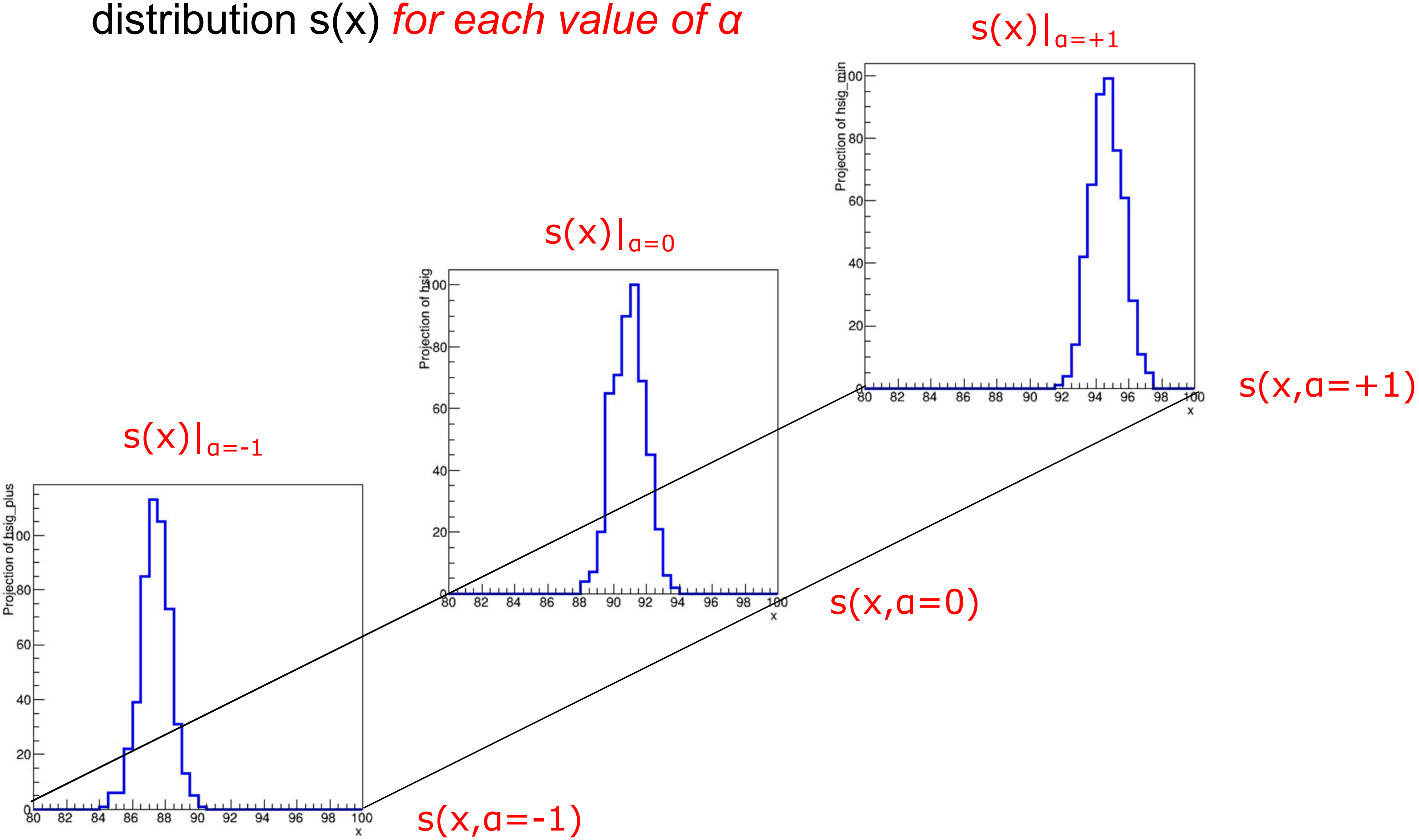


- Challenge: construct an empirical response function based on the interpolation of the shapes of these three templates.

Wouter Verkerke, NIKHEF

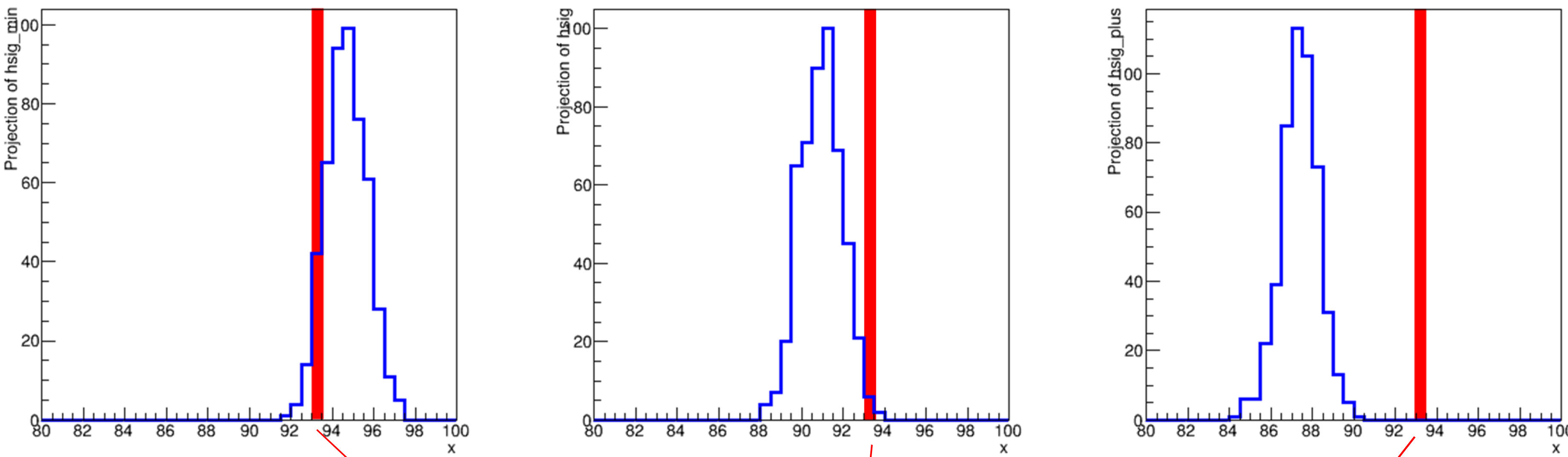
Need to interpolate between template models

- Need to define 'morphing' algorithm to define distribution s(x) *for each value of α*

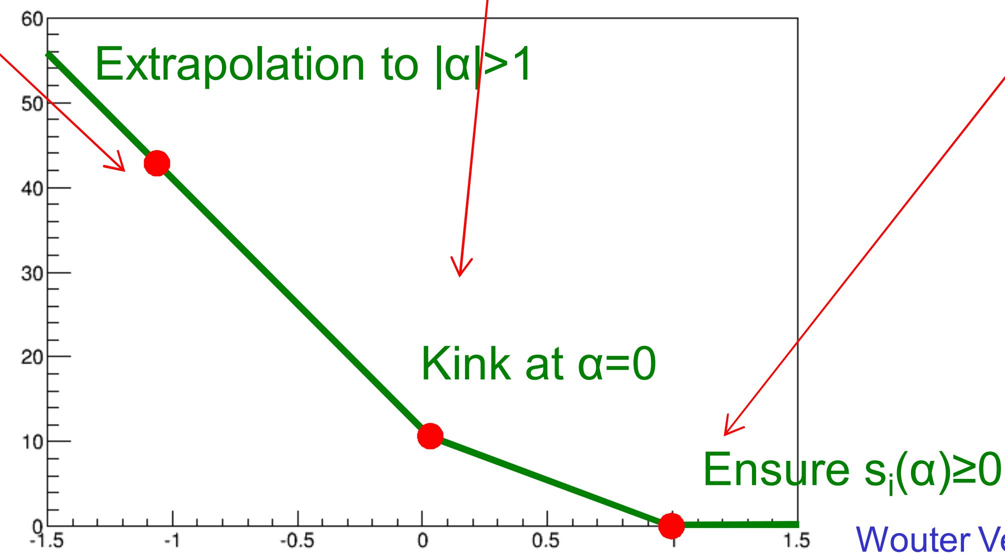


Piecewise linear interpolation

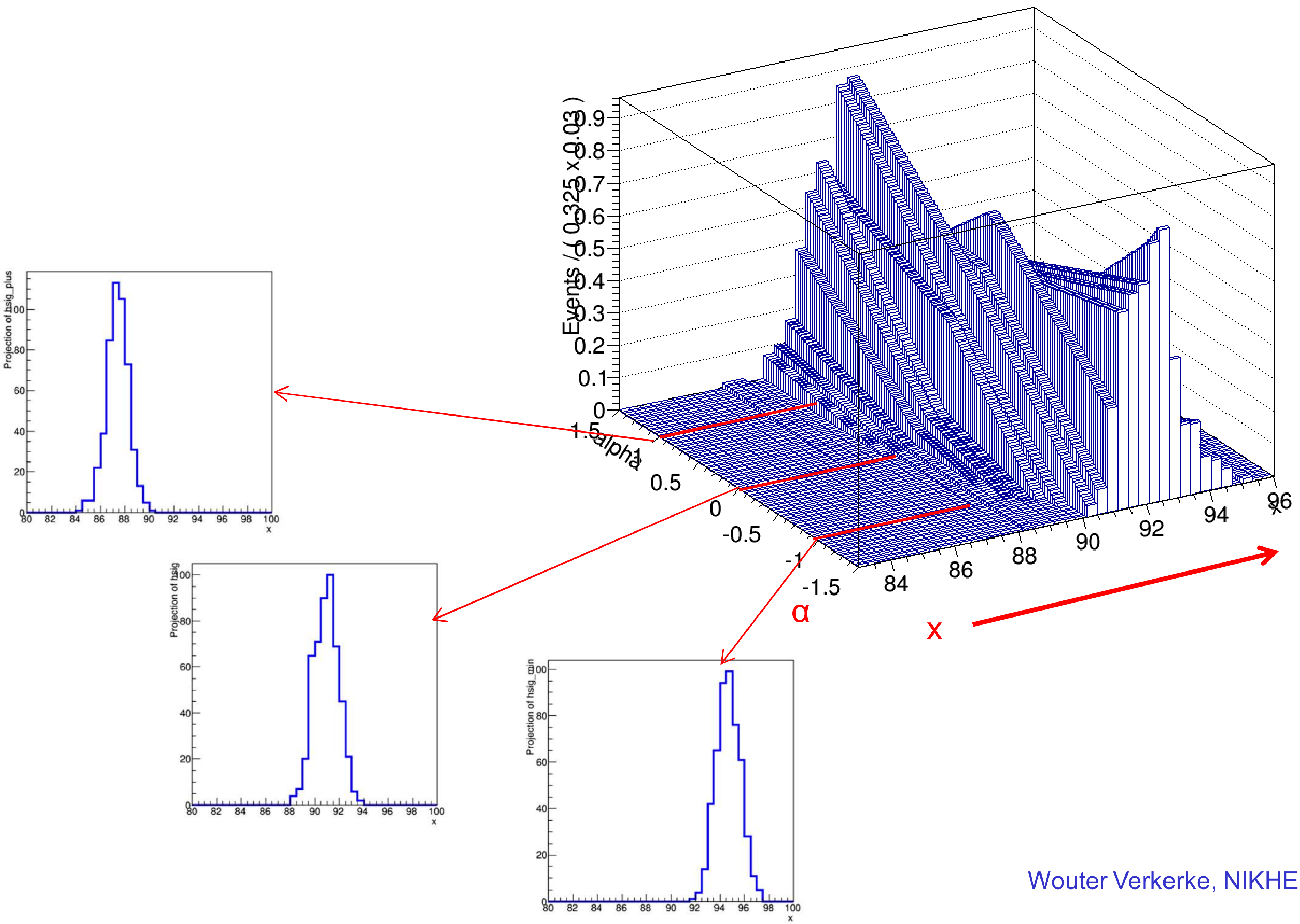
- Simplest solution is piece-wise linear interpolation for each bin



Piecewise linear interpolation response model for a one bin



Visualization of bin-by-bin linear interpolation of distribution

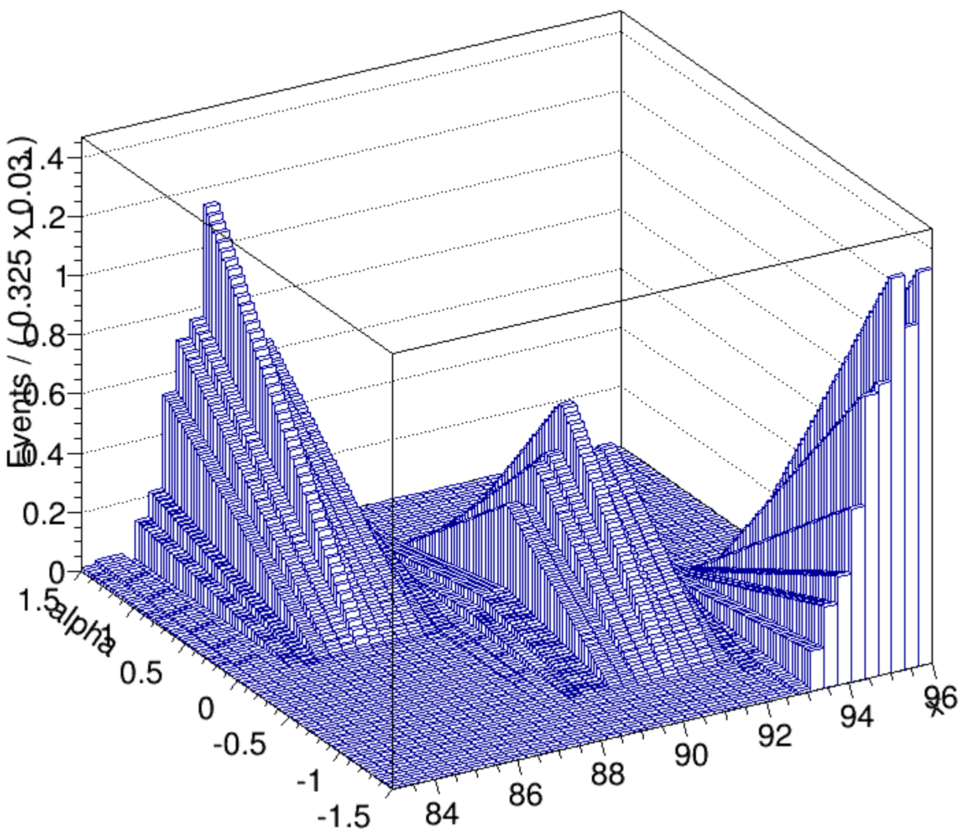


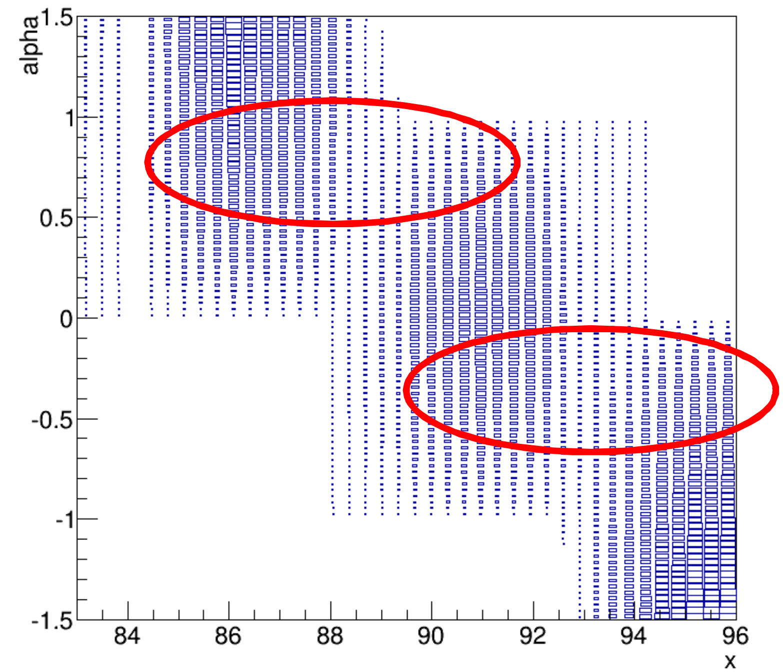
Wouter Verkerke, NIKHEF

Limitations of piece-wise linear interpolation

- Bin-by-bin interpolation looks spectacularly easy and simple, but be aware of its limitations
  - Same example, but with larger 'mean shift' between templates

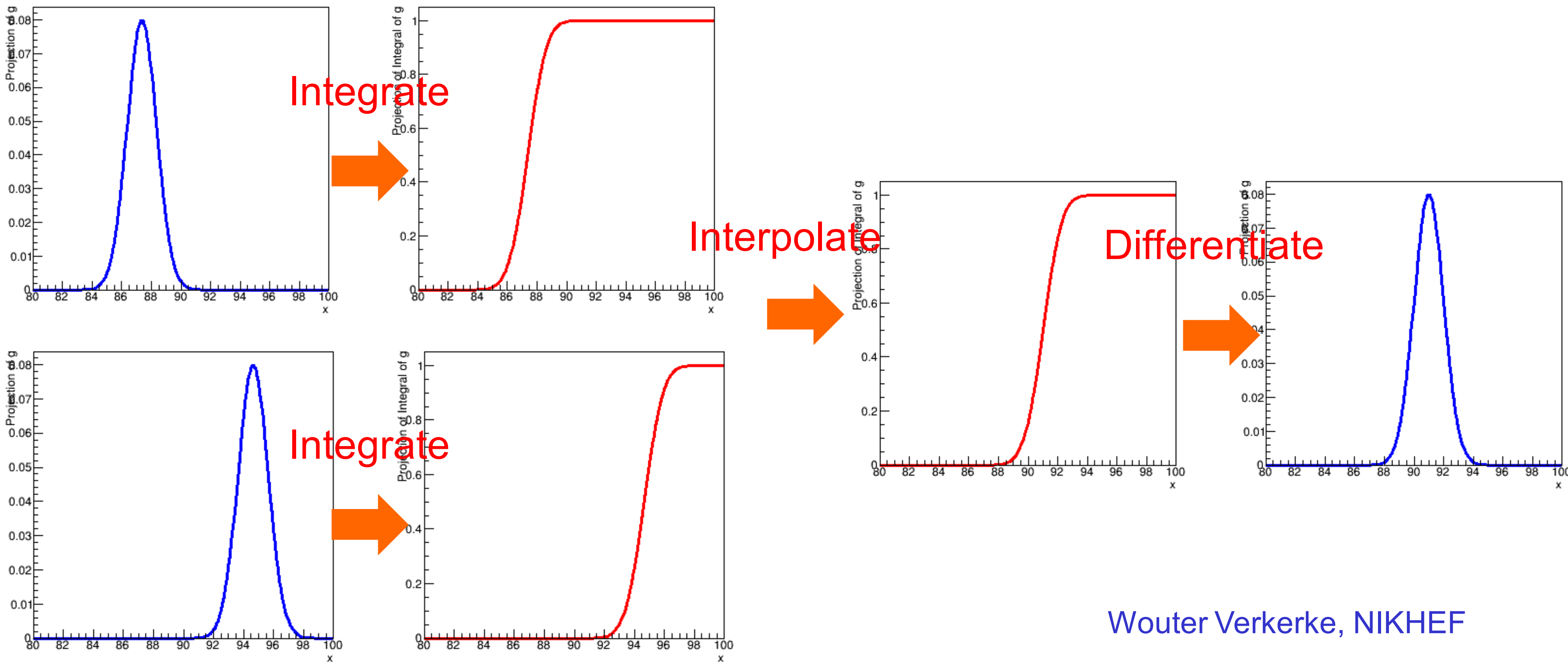
Note double peak structure around |α|=0.5





Other morphing strategies – 'horizontal morphing'

- Other template morphing strategies exist that are less prone to unintended side effects
- A 'horizontal morphing' strategy was invented by Alex read.
  - Interpolates the cumulative distribution function instead of the distribution
  - Especially suitable for shifting distributions
  - Here shown on a continuous distribution, but also works on histograms
  - Drawback: computationally expensive, algorithm only worked out for 1 NP



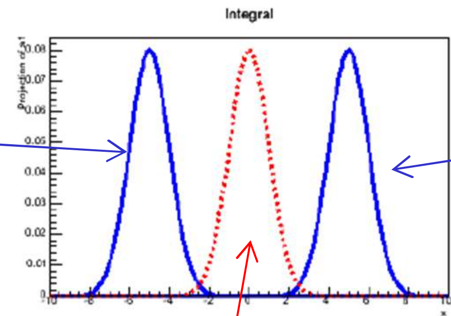
Wouter Verkerke, NIKHEF

Yet another morphing strategy – ‘Moment morphing’

M. Baak & S. Gadatsch

- Given two template model *$f_-(x)$* and *$f_+(x)$* the strategy of moment morphing considers first two moment of template models (mean and variance)

$$\mu_- = \int x \cdot f_-(x)dx$$
$$V_- = \int (x-\mu_-)^2 \cdot f_-(x)dx$$



$$\mu_+ = \int x \cdot f_+(x)dx$$
$$V_+ = \int (x-\mu_+)^2 \cdot f_+(x)dx$$

- The goal of moment morphing is to construct an interpolated function that has linearly interpolated moments

$$\mu(\alpha) = \alpha\mu_- + (1-\alpha)\mu_+$$
$$V(\alpha) = \alpha V_- + (1-\alpha)V_+ \quad [1]$$

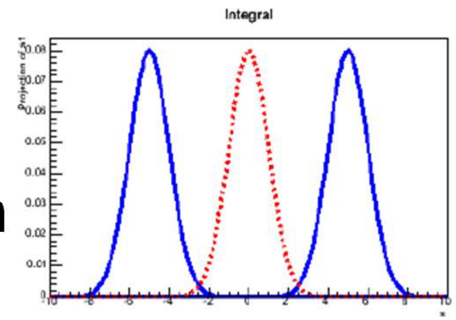
- It constructs this morphed function as combination of linearly transformed input models

$$f(x,\alpha) \rightarrow \alpha f_-(ax+b) + (1-\alpha) f_+(cx-d)$$

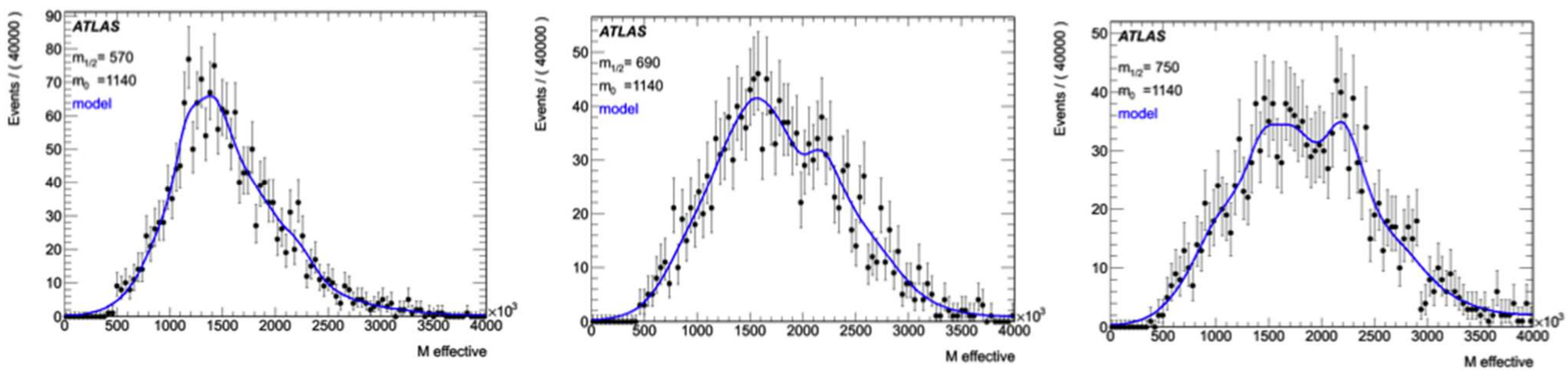
  – Where constants a,b,c,d are chosen such so that f(x,α) satisfies conditions [1]

Yet another morphing strategy – 'Moment morphing'

- For a Gaussian probability model with linearly changing mean and width, moment morphing of two Gaussian templates is the exact solution



- But also works well on 'difficult' distributions



- Good computational performance
  - Calculation of moments of templates is expensive, but just needs to be done once, otherwise very fast (just linear algebra)

$$f(x,\alpha) \rightarrow \alpha f_{-}(ax+b) + (1-\alpha) f_{+}(cx-d)$$

- Multi-dimensional interpolation strategies exist

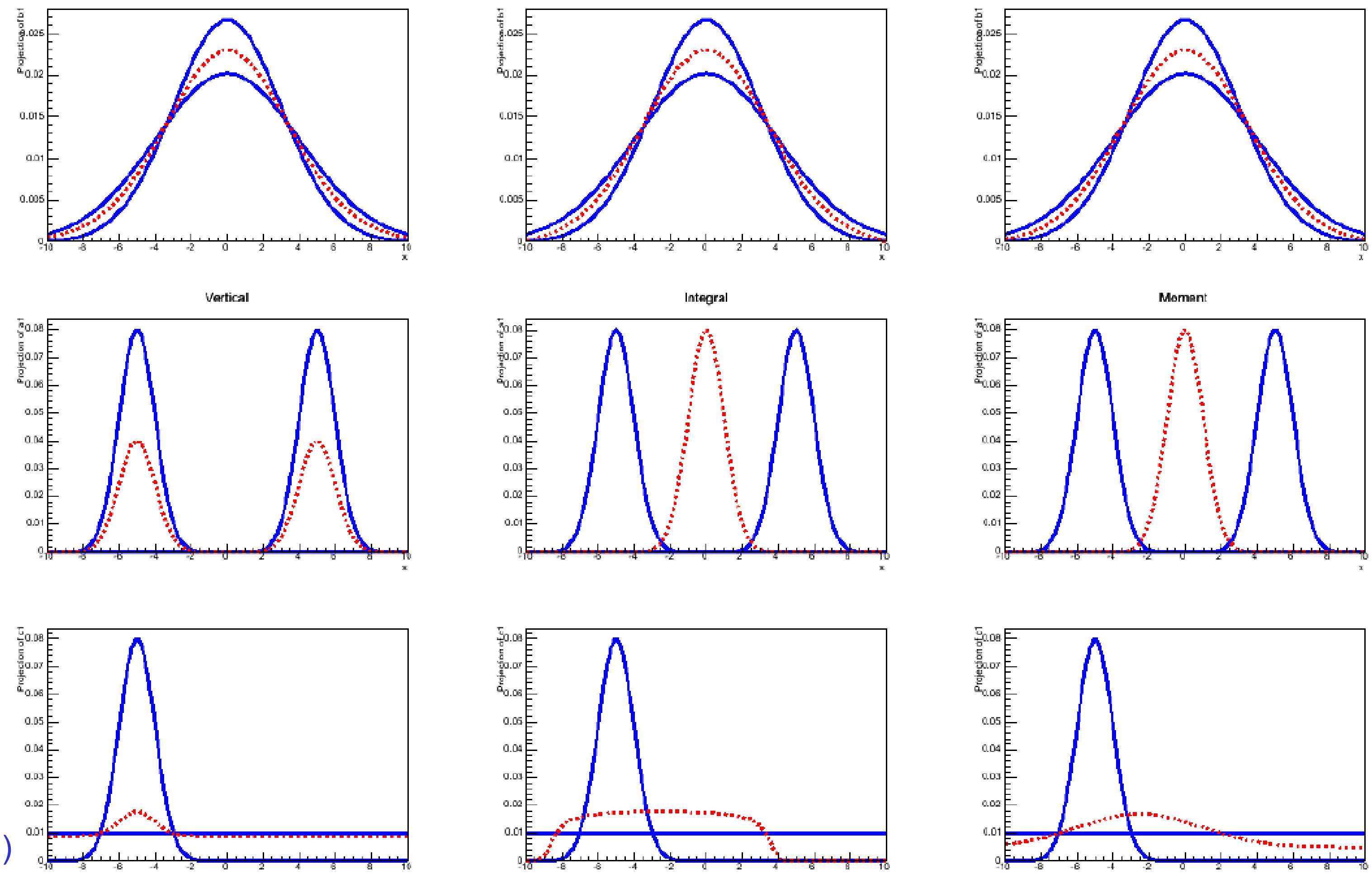
There are other morphing algorithms to choose from

Vertical Morphing | Horizontal Morphing | Moment Morphing

Gaussian varying width

Gaussian varying mean

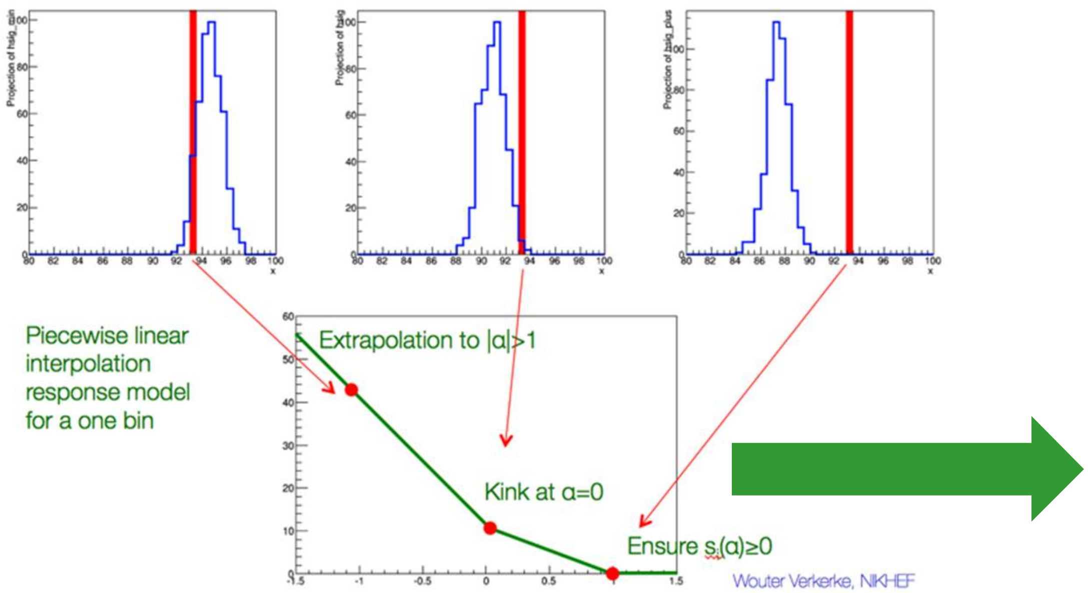
Gaussian to Uniform (this is conceptually ambigous!)



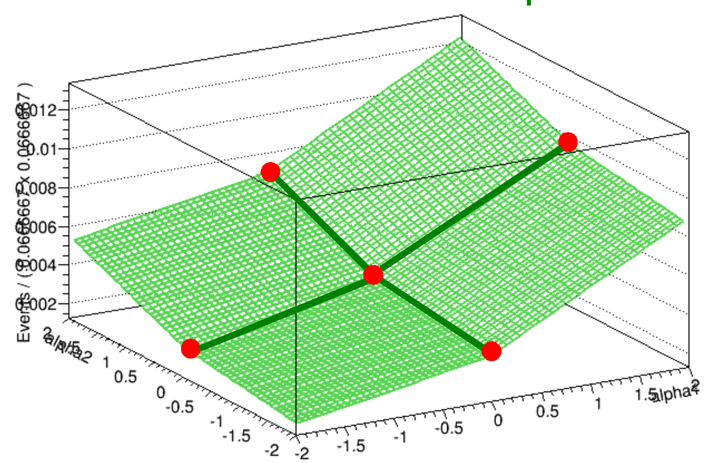
n-dimensional morphing? □ □ □

Piece-wise interpolation for >1 nuisance parameter

- Concept of piece-wise linear interpolation can be trivially extended to apply to morphing of >1 nuisance parameter.
  - Difficult to visualize effect on full distribution, but easy to understand concept at the individual bin level



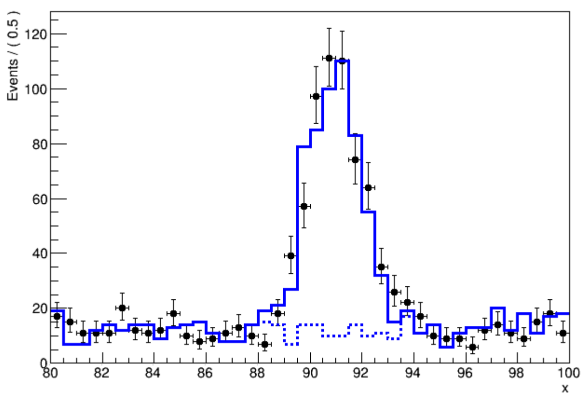
Visualization of 2D interpolation

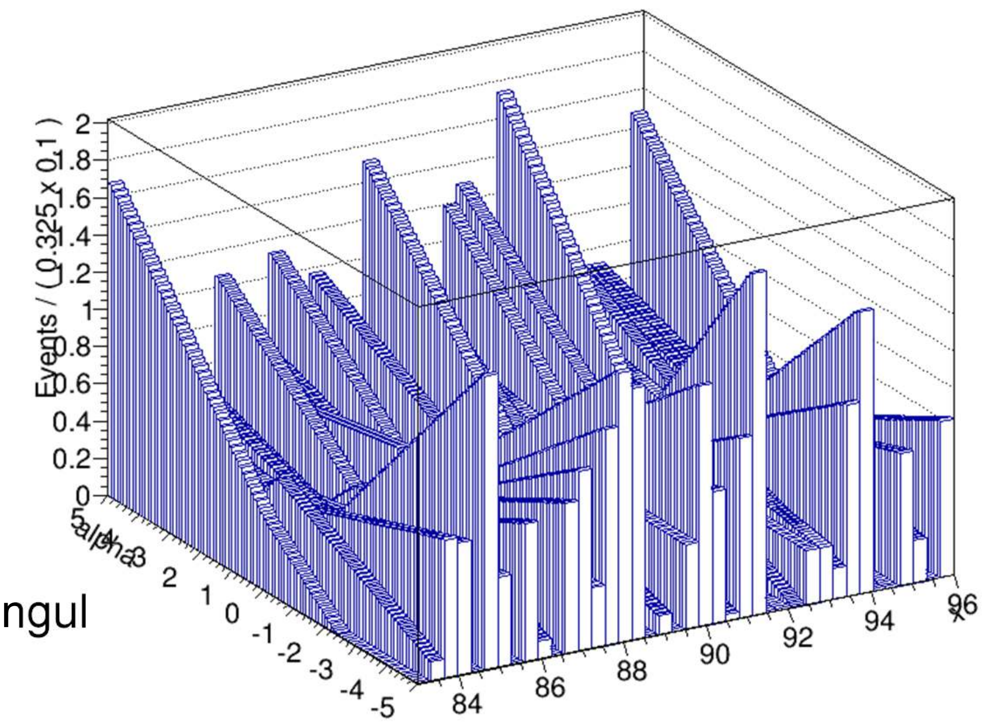


Wouter Verkerke, NIKHEF

Shape, rate or no systematic?

- Be judicious with modeling of systematic with little or no significant change in shape (w.r.t MC template statistics)
  - Example morphing of a very subtle change in the background model
  - Is this a meaningful new degree of freedom in the likelihood model?

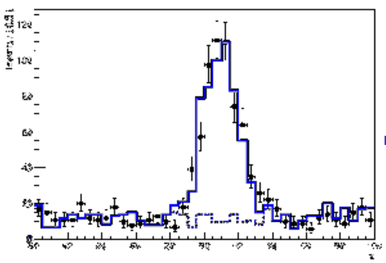


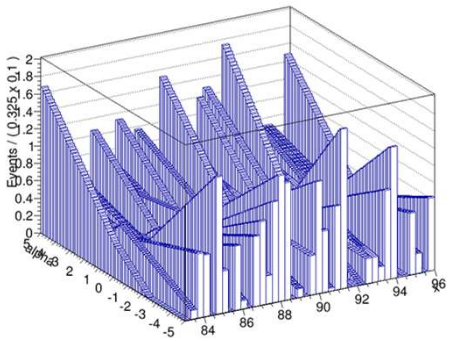


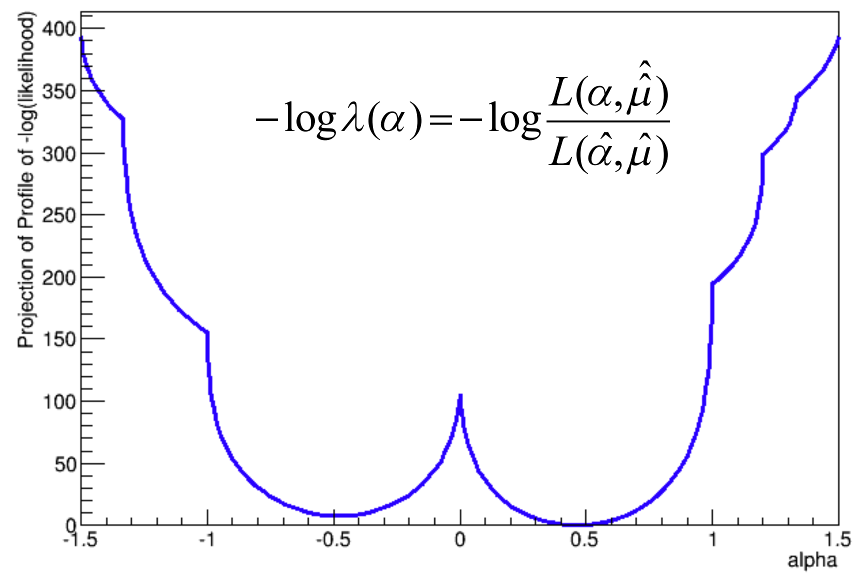
  - A χ2 or KS test between nominal and alternate template can help to decide if a shape uncertainty is meaningul
  - Most systematic uncertainties affect both rate and shape, but can make independent decision on modeling rate (which less likely to affect fit stability)

Fit stability due to insignificant shape systematics

- Shape of profile likelihood in NP α clearly raises two points



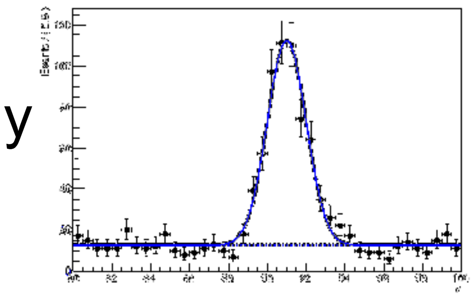




- 1) Numerical minimization process will be ‘interesting’
- 2) MC statistical effects induce strongly defined minima that are fake
  – Because for this example all three templates were sampled from the same parent distribution (a uniform distribution)

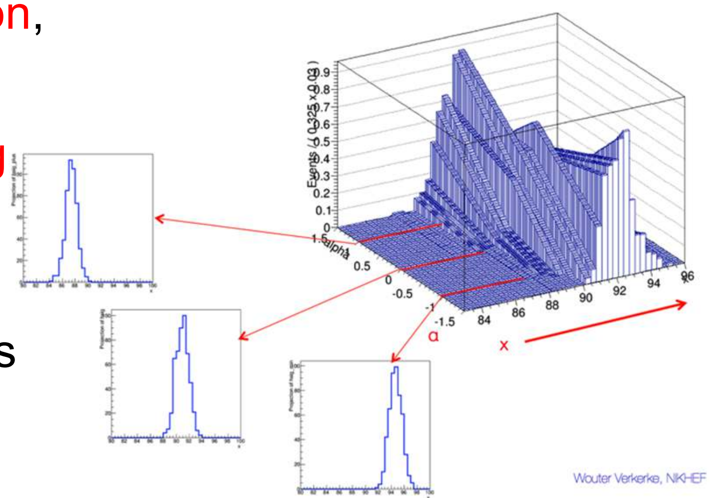
Recap on shape systematics & template morphing

- Implementation of shape systematic in likelihoods modeling distributions conceptually no different that rate systematics in counting experiments



$$L(\vec{m}_{ll}\,|\,\mu,\alpha_{LES}) = \prod_i \left[\mu\cdot \text{Gauss}(m_{ll}^{(i)}, 91\cdot(1+2\alpha_{LES}, 1) + (1-\mu)\cdot \text{Uniform}(m_{ll}^{(i)})\right]\cdot Gauss(0\,|\,\alpha_{LES}, 1)$$

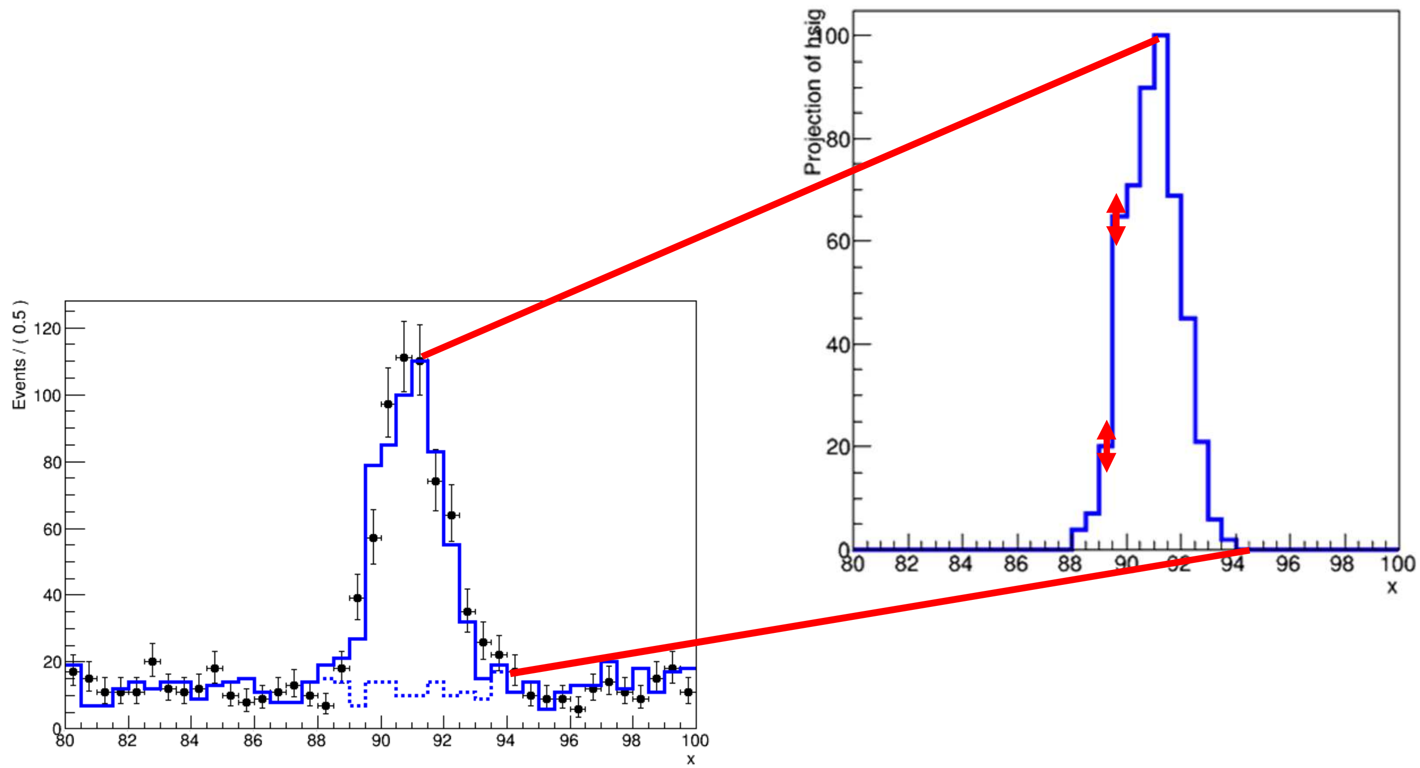
- For template modes obtained from MC simulation template provides a technical solution to implement response function
  - Simplest strategy piecewise linear interpolation, but only works well for small changes
  - Moment morphing better adapted to modeling of shifting distributions
  - Both algorithms extend to n-dimensional interpolation to model multiple systematic NPs in response function
  - Be judicious in modeling 'weak' systematics: MC systematic uncertainties will dominate likelihood



Wouter Verkerke, NIKHEF

Other uncertainties in MC shapes – finite MC statistics

- In practice, MC distributions used for template fits have finite statistics.



- Limited MC statistics represent an uncertainty on your model how to model this effect in the Likelihood?

Other uncertainties in MC shapes – finite MC statistics

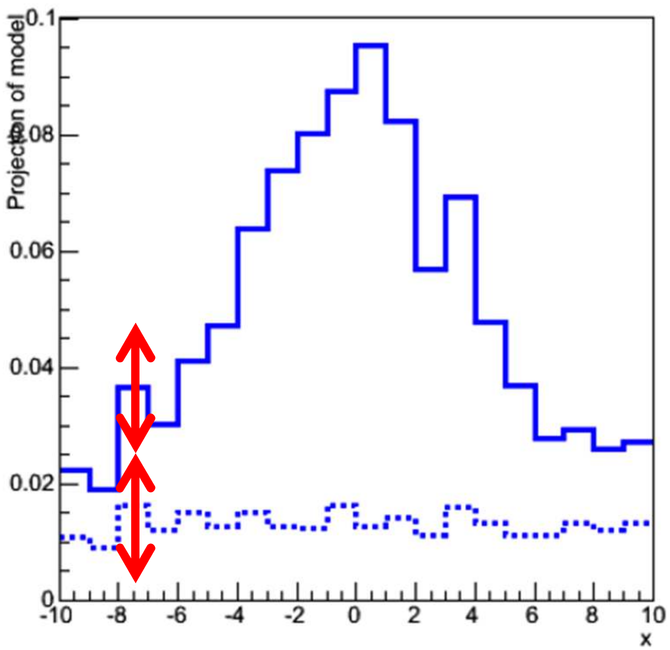
- Modeling MC uncertainties: *each MC bin has a Poisson uncertainty*
- Thus, apply usual 'systematics modeling' prescription.
- For a single bin – exactly like original counting measurement

Fixed signal, bkg MC prediction

$$L_{bin-i}(\mu) = Poisson(N_i \mid \mu \cdot \tilde{s}_i + \tilde{b}_i)$$

Signal, bkg
MC nuisance params

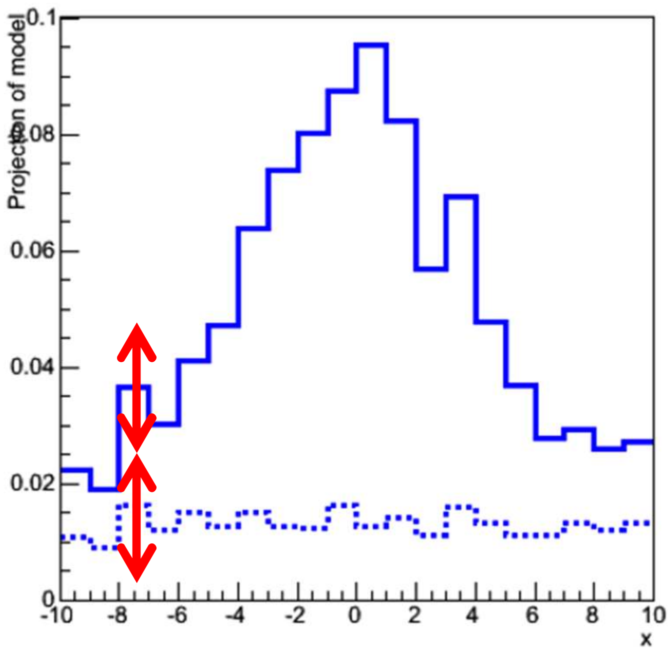
$$L_{bin-i}(\mu, s_i, b_i) = Poisson(N_i \mid \mu \cdot s_i + b_i) \cdot Poisson(N_i^{MC-s} \mid s_i) \cdot Poisson(N_i^{MC-b} \mid b_i)$$



Subsidiary measurement for signal MC
('measures' MC prediction s_i with Poisson uncertainty)

Nuisance parameters for template statistics

- Repeat for all bins



$$L(\vec{N}\,|\,\mu) = \prod_{bins} P(N_i\,|\,\mu \cdot \tilde{s}_i + \tilde{b}_i)$$

Binned likelihood with rigid template

$$L(\vec{N}\,|\,\mu,\vec{s},\vec{b}) = \prod_{bins} P(N_i\,|\,\mu \cdot s_i + b_i) \prod_{bins} P(\tilde{s}_i\,|\,s_i) \prod_{bins} P(\tilde{b}_i\,|\,b_i)$$

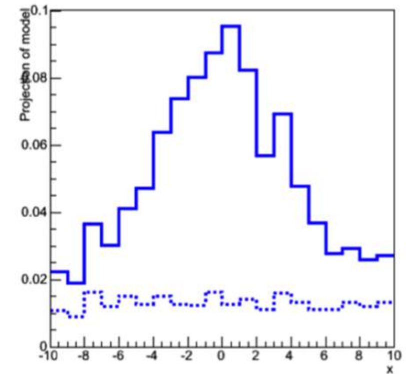
Response function w.r.t. *s, b* as parameters

2x N_{bins} subsidiary measurements of *s* ,*b* from s~,b~

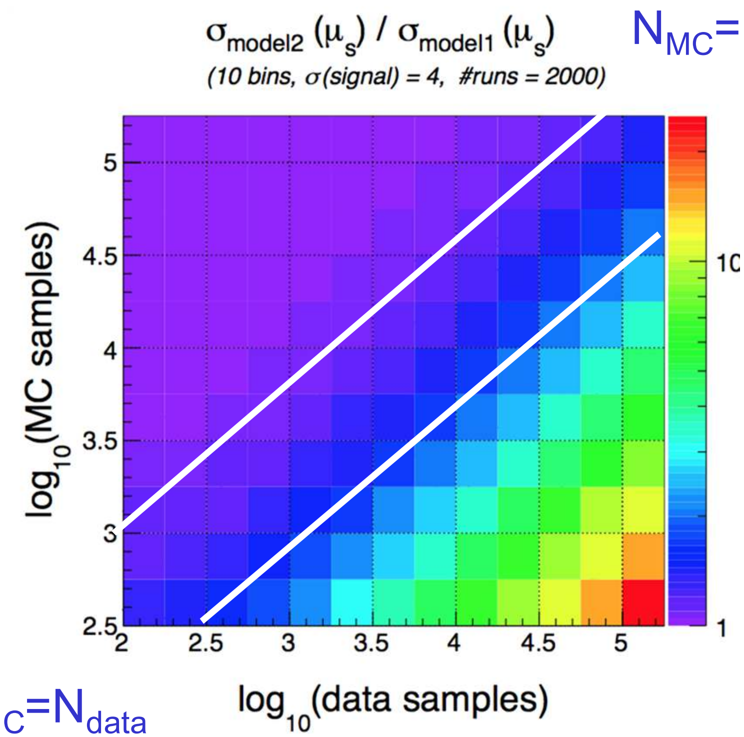
- Result: accurate model for MC statistical uncertainty, but lots of nuisance parameters (#samples x #bins)...

The effect of template statistics

- When is it important to model the effect of template statistics in the likelihood
  - Roughly speaking the effect of template statistics becomes important when $N_{templ} < 10\text{x}\ N_{data}$ (from Beeston & Barlow)



- Measurement of effect of template statistics in previously shown toy likelihood model, where POI is the signal yield

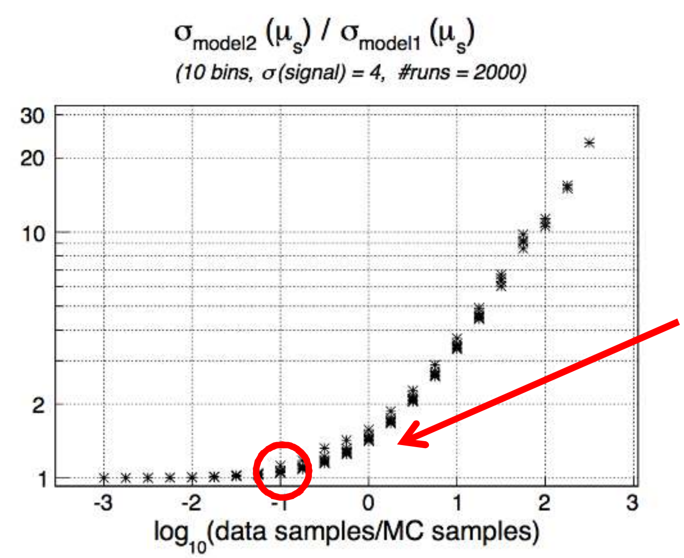


$N_{MC}=10N_{data}$

$N_{MC}=N_{data}$

'model 1 – plain template likelihood'

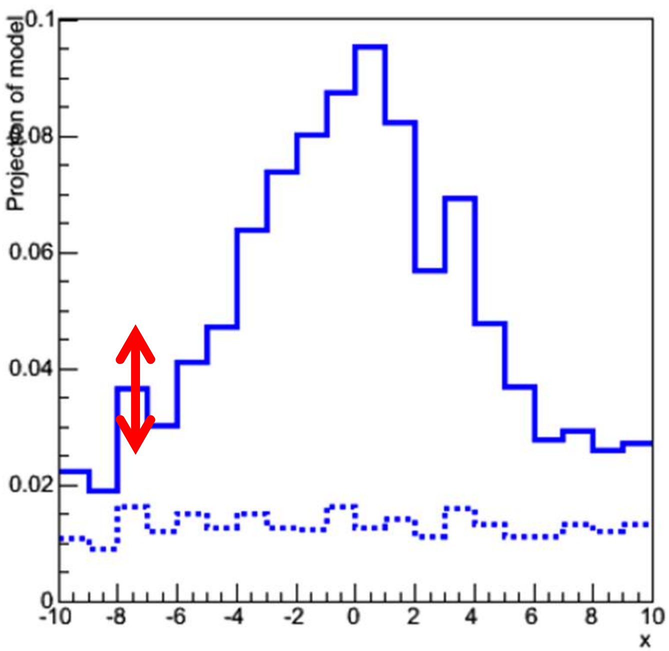
'model 2 – Beeston-Barlow likelihood'



Note that even at $N_{MC}=10N_{data}$ uncertainty on POI can be underestimated by 10% without BB

Reducing the number NPs – Beeston-Barlow ‘lite’

- Another approach that is being used is called ‘BB’ – lite
- Premise: effect of statistical fluctuations on sum of templates is dominant → Use one NP per bin instead of one NP per component per bin



‘Beeston-Barlow’

$$L(\vec{N}|\vec{s},\vec{b}) = \prod_{bins} P(N_i | s_i + b_i) \prod_{bins} P(\tilde{s}_i | s_i) \prod_{bins} P(\tilde{b}_i | b_i)$$

‘Beeston-Barlow lite ’

$$L(\vec{N}|\vec{n}) = \prod_{bins} P(N_i | n_i) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | n_i)$$

Response function w.r.t. *n* as parameters

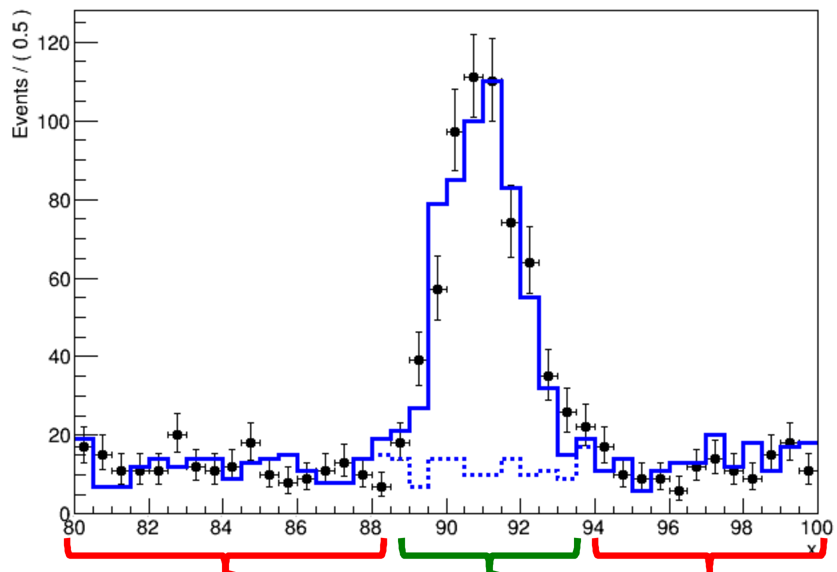
Subsidiary measurements of *n* from s~+b~

$$L(\vec{N}|\vec{\gamma}) = \prod_{bins} P(N_i | \gamma_i(\tilde{s}_i + \tilde{b}_i)) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | \gamma_i(\tilde{s}_i + \tilde{b}_i))$$

Normalized NP lite model (nominal value of all γ is 1)

Pruning complexity – MC statistical for selected bins

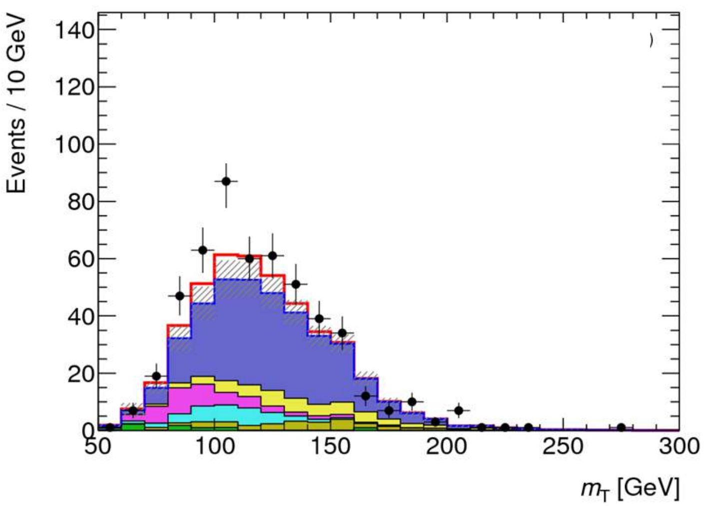
- Can also make decision to model MC statistical uncertainty on a bin-by-bin basis
  - No modeling for high statistics bins
  - Explicit modeling for low-statistics bins

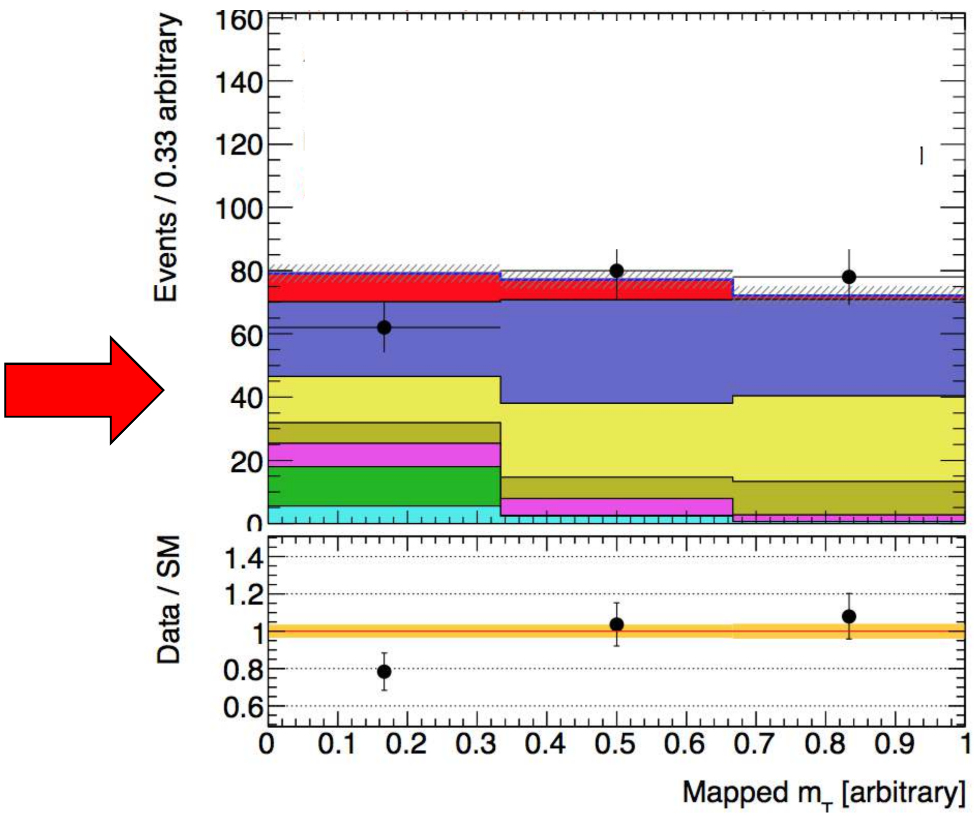


$$L(\vec{N}|\vec{\gamma}) = \prod_{bins} P(N_i \,|\, \gamma_i(s_i + b_i)) \prod_{low-stats\ bins} P(s_i + b_i \,|\, \gamma_i(s_i + b_i)) \prod_{hi-stats\ bins} \delta(\gamma_i)$$

Adapting binning to event density

- Effect of template statistics can also be controlled by rebinning data such all bins contain expected and observed events
  - For example choose binning such that expected background has a uniform distribution (as signals are usually small and/or uncertain they matter less)





Intermezzo – Software tools for likelihood modeling

- Techniques shown to model systematic uncertainties can lead to complex likelihood functions
  - Realistic analyses describe many distributions, with O(10) to O(100) systematic uncertainties
- Not trivial to write these by hand – but various tools have been developed in the past decade to implement these techniques
- Will highlight a few key feature of RooFit/RooStats tool suite
  - Joined software development project ROOT/ATLAS/CMS over past 6 years
  - My personal favorite

Modular software design

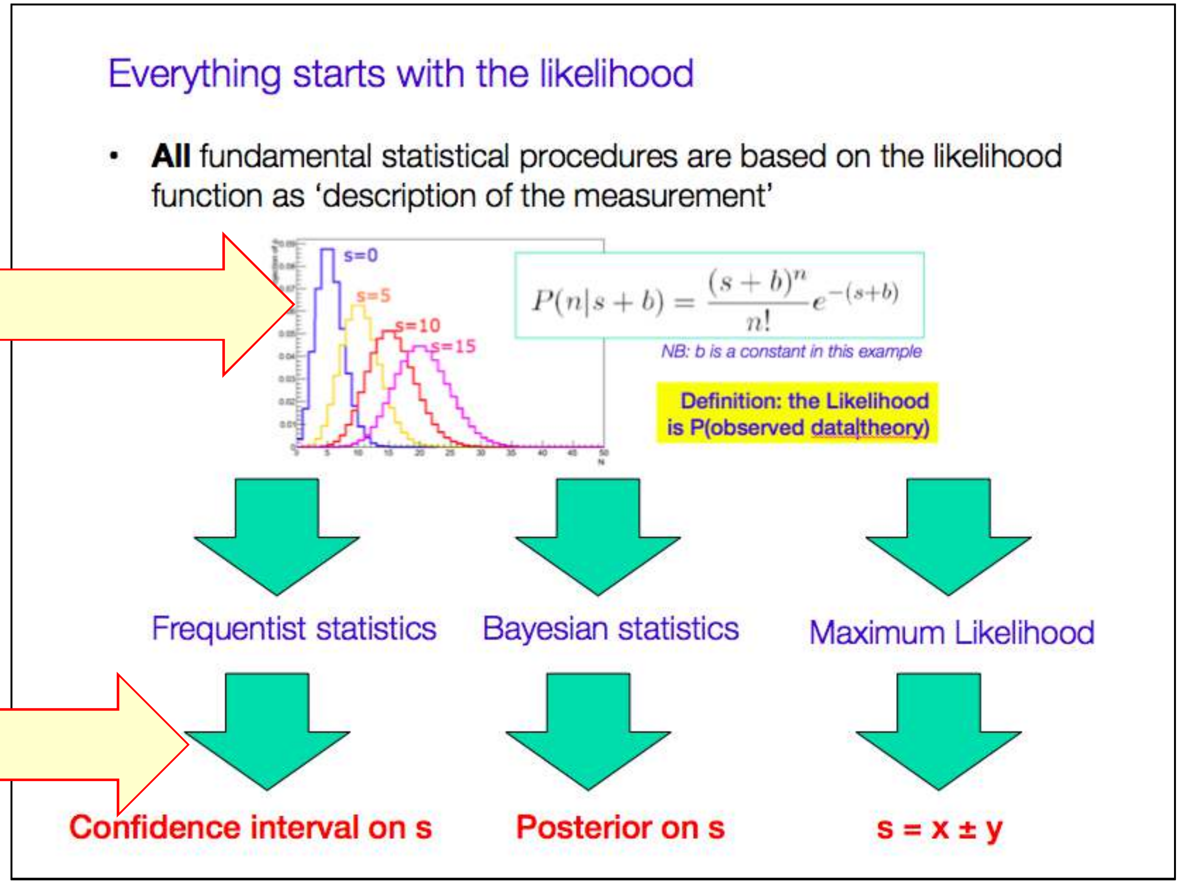
RooFit/HistFactory
Language for building probability models

Comprises datasets, likelihoods, minimization, toy data generation, visualization and persistence

(RooFit Workspace)

RooStats

Suite of statistical tests operating on RooFit probability models



RooFit core design philosophy

- Mathematical objects are represented as C++ objects

Mathematical concept		RooFit class
variable	x	`RooRealVar`
function	$f(x)$	`RooAbsReal`
PDF	$f(x)$	`RooAbsPdf`
space point	x	`RooArgSet`
integral	$\int_{x_{\min}}^{x_{\max}} f(x)dx$	`RooRealIntegral`
list of space points		`RooAbsData`

RooFit core design philosophy

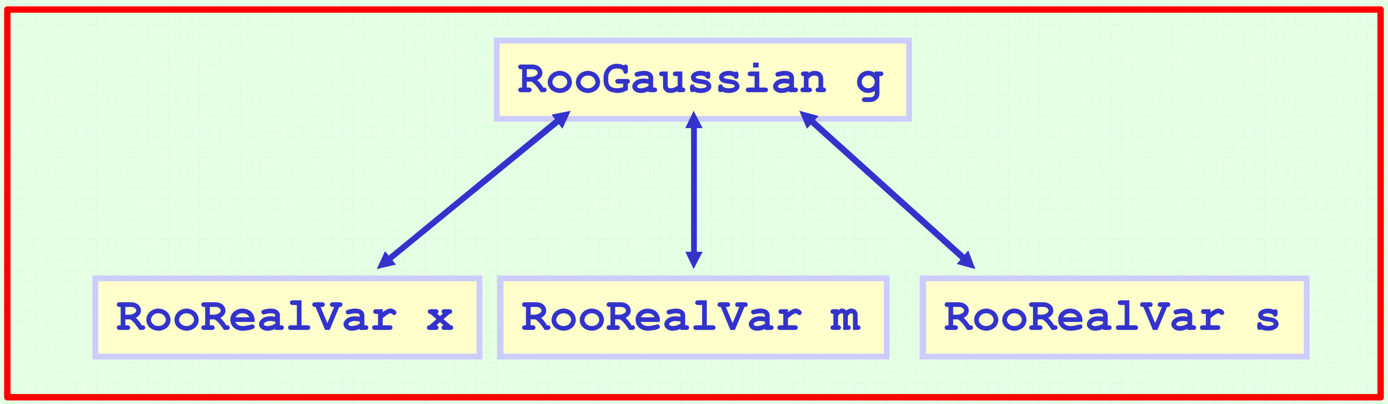
- Instead of **‘double Likelihood(double paramVec[])’**, a flexible modular structure of ‘programmed’ functions

Math

Gauss(x,μ,σ)

RooFit diagram

RooWorkspace (keeps all parts together)



RooFit code

```
RooRealVar x("x","x",-10,10) ;
RooRealVar m("m","y",0,-10,10) ;
RooRealVar s("s","z",3,0.1,10) ;
RooGaussian g("g","g",x,m,s) ;
RooWorkspace w("w") ;
w.import(g) ;
```

RooFit core design philosophy - Workspace

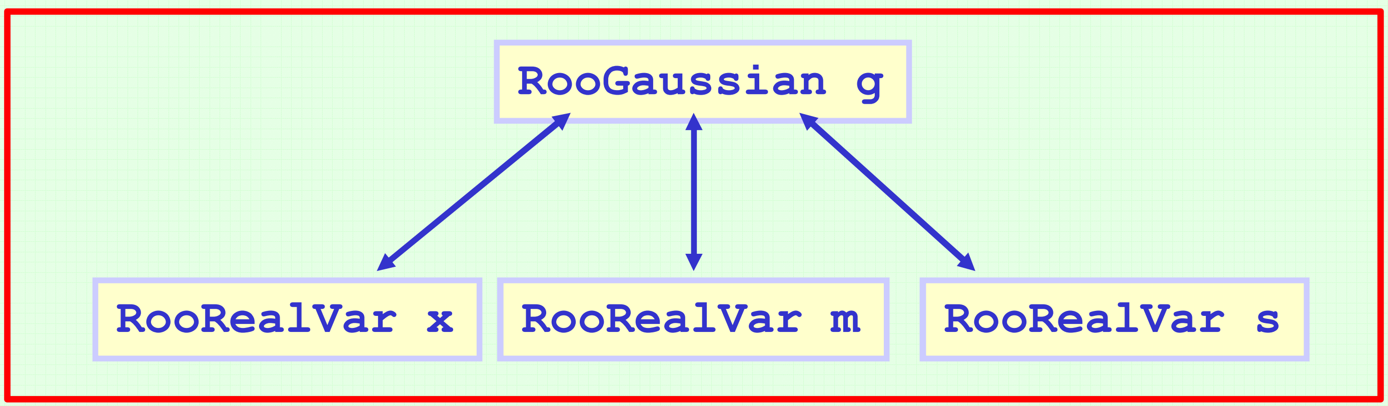
- Alternatively, a simple math-like ‘factory language’ can quickly populates a workspace with the same objects

Math

$$\mathrm{Gauss}(x,\mu,\sigma)$$

RooFit diagram

RooWorkspace

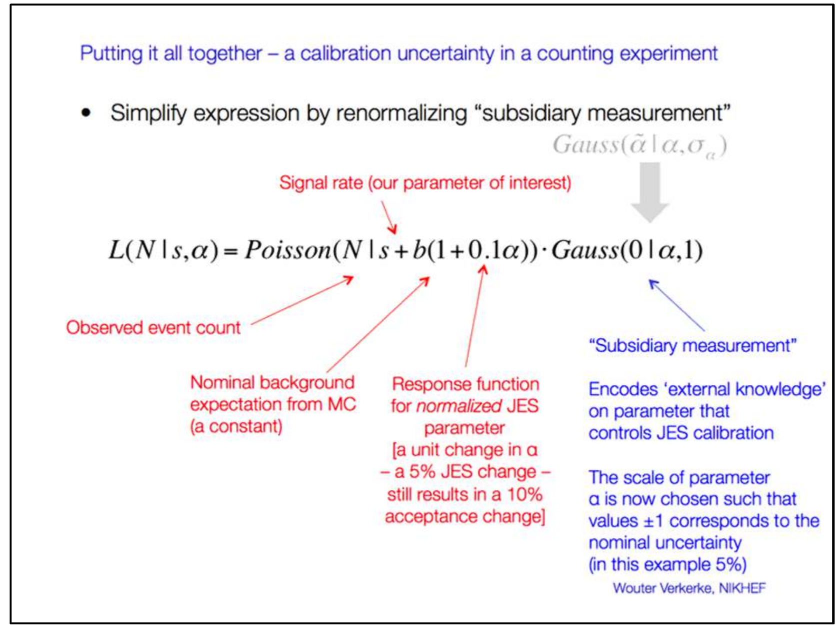


RooFit code

```
RooWorkspace w(“w”) ;
w.factory(“Gaussian::g(x[-10,10],m[0],s[5])”) ;
```

Example 1: counting expt

- Will now demonstrate how to construct a model for a counting experiment with a systematic uncertainty

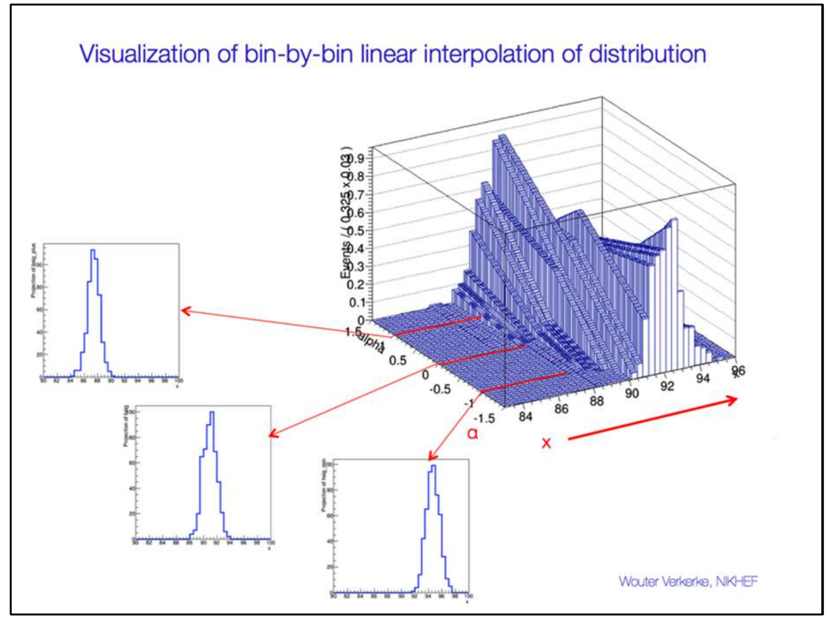


$$L(N\,|\,s,\alpha) = Poisson(N\,|\,s+b(1+0.1\alpha))\cdot Gauss(0\,|\,\alpha,1)$$

```
// Subsidiary measurement of alpha
w.faxtory("Gaussian::subs(0,alpha[-5,5],1)") ;

// Response function mu(alpha)
w.factory("expr::mu('s+b(1+0.1*alpha)',s[20],b[20],alpha)") ;

// Main measurement
w.factory("Poisson::p(N[0,10000],mu)");

// Complete model Physics*Subsidiary
w.factory("PROD::model(p,subs)") ;
```

Example 2 : binned L with syst

- Example of template morphing systematic in a binned likelihood



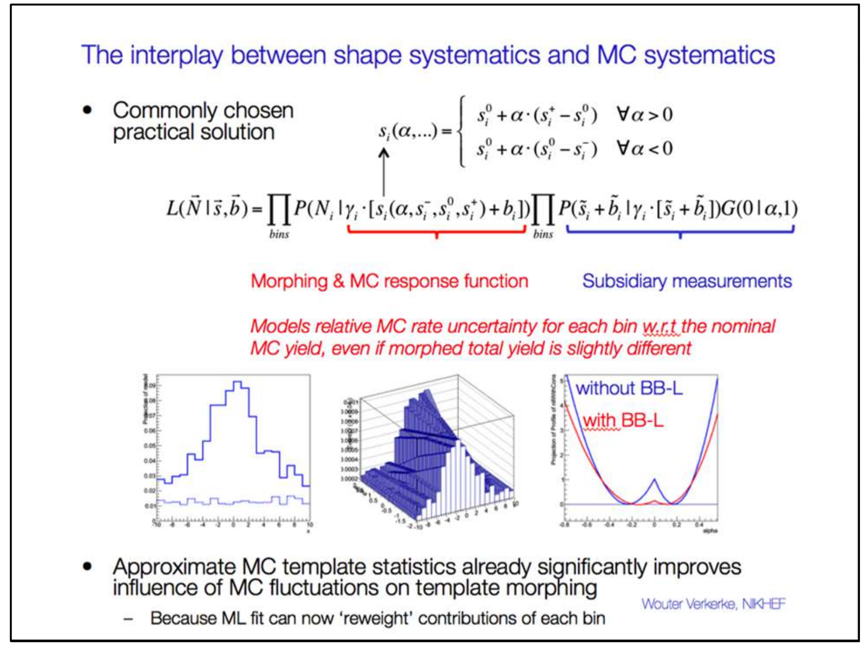
$$s_i(\alpha,...) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$

$$L(\vec{N}\,|\,\alpha,\vec{s}^-,\vec{s}^0,\vec{s}^+) = \prod_{bins} P(N_i\,|\,s_i(\alpha,s_i^-,s_i^0,s_i^+)) \cdot G(0\,|\,\alpha,1)$$

```
// Import template histograms in workspace
 w.import(hs_0,hs_p,hs_m) ;

 // Construct template models from histograms
 w.factory("HistFunc::s_0(x[80,100],hs_0)") ;
 w.factory("HistFunc::s_p(x,hs_p)") ;
 w.factory("HistFunc::s_m(x,hs_m)") ;

 // Construct morphing model
 w.factory("PiecewiseInterpolation::sig(s_0,s_,m,s_p,alpha[-5,5])") ;

 // Construct full model
 w.factory("PROD::model(ASUM(sig,bkg,f[0,1]),Gaussian(0,alpha,1))") ;
```

Example 3 – BB-lite + morphing

- Template morphing model with Beeston-Barlow-lite MC statistical uncertainties



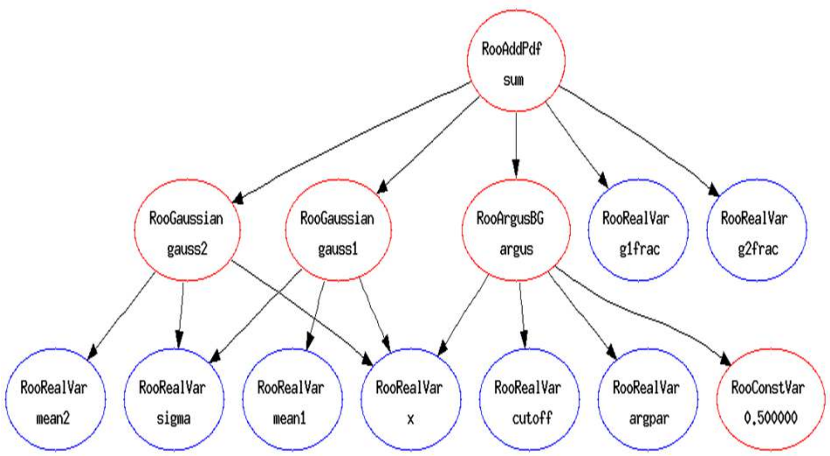
$$s_i(\alpha,...)=\begin{cases} s_i^0+\alpha\cdot(s_i^+-s_i^0) & \forall\alpha>0 \\ s_i^0+\alpha\cdot(s_i^0-s_i^-) & \forall\alpha<0 \end{cases}$$

$$L(\vec{N}|\vec{s},\vec{b})=\prod_{bins}P(N_i|\gamma_i\cdot[s_i(\alpha,s_i^-,s_i^0,s_i^+)+b_i])\prod_{bins}P(\tilde{s}_i+\tilde{b}_i|\gamma_i\cdot[\tilde{s}_i+\tilde{b}_i])G(0|\alpha,1)$$

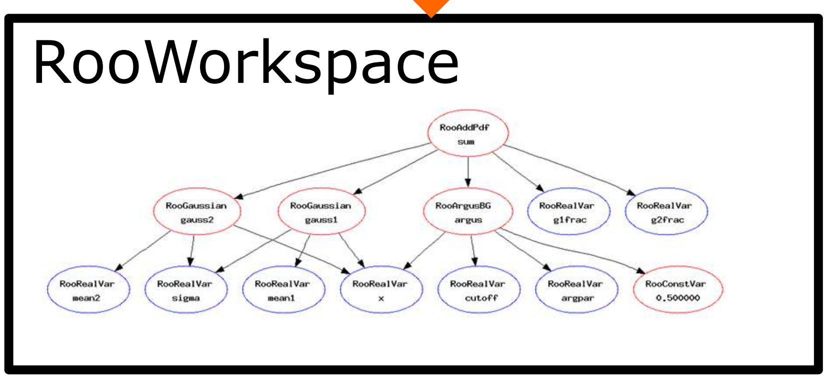
```
// Import template histograms in workspace
w.import(hs_0,hs_p,hs_m,hb) ;

// Construct parametric template morphing signal model
w.factory("ParamHistFunc::s_p(hs_p)") ;
w.factory("HistFunc::s_m(x,hs_m)") ;
w.factory("HistFunc::s_0(x[80,100],hs_0)") ;
w.factory("PiecewiseInterpolation::sig(s_0,s_,m,s_p,alpha[-5,5])") ;

// Construct parametric background model (sharing gamma's with s_p)
w.factory("ParamHistFunc::bkg(hb,s_p)") ;

// Construct full model with BB-lite MC stats modeling
w.factory("PROD::model(ASUM(sig,bkg,f[0,1]),
           HistConstraint({s_0,bkg}),Gaussian(0,alpha,1))") ;
```

The workspace – *the portable likelihood function*

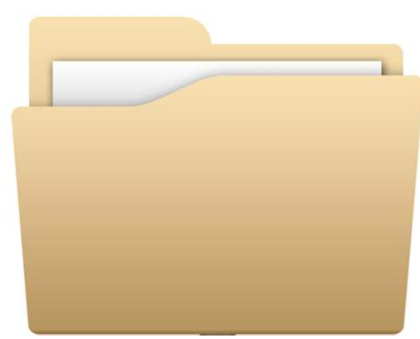
- **The workspace concept has revolutionized the way people share and combine analysis**
  - **Completely factorizes process of building and using likelihood functions**
  - You can give somebody an analytical likelihood of a (potentially very complex) physics analysis in a way to the easy-to-use, provides introspection, and is easy to modify.




```
RooWorkspace w("w") ;
w.import(sum) ;
w.writeToFile("model.root") ;
```

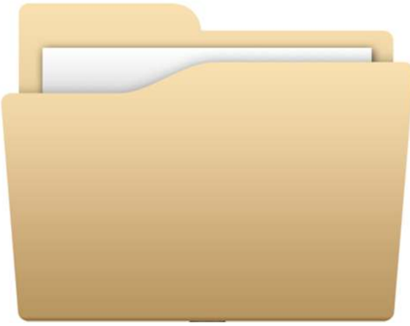


model.root

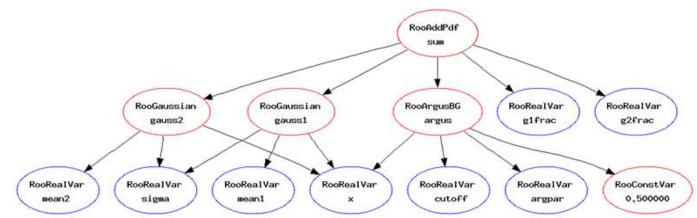


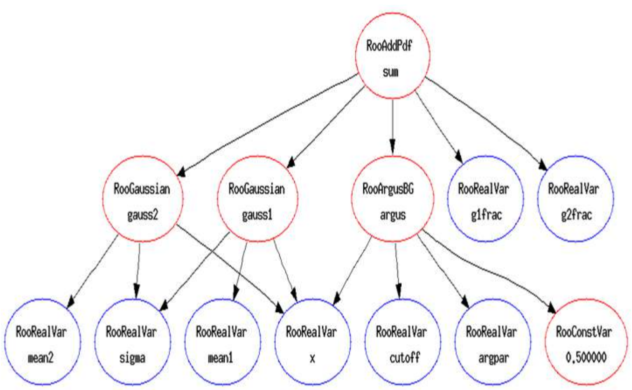
Wouter Verkerke, NIKHEF

Using a workspace

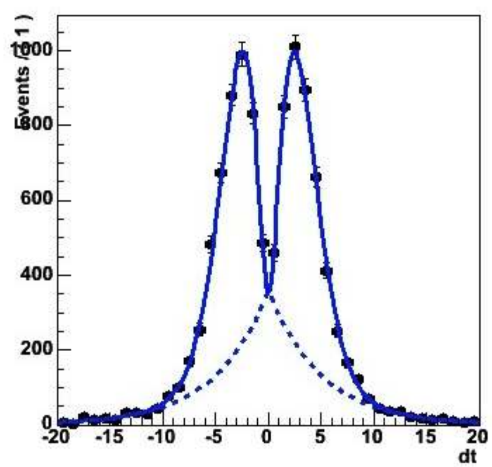


RooWorkspace



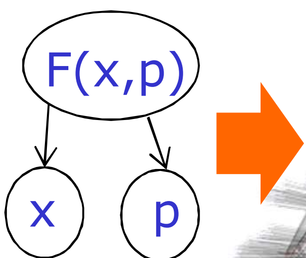


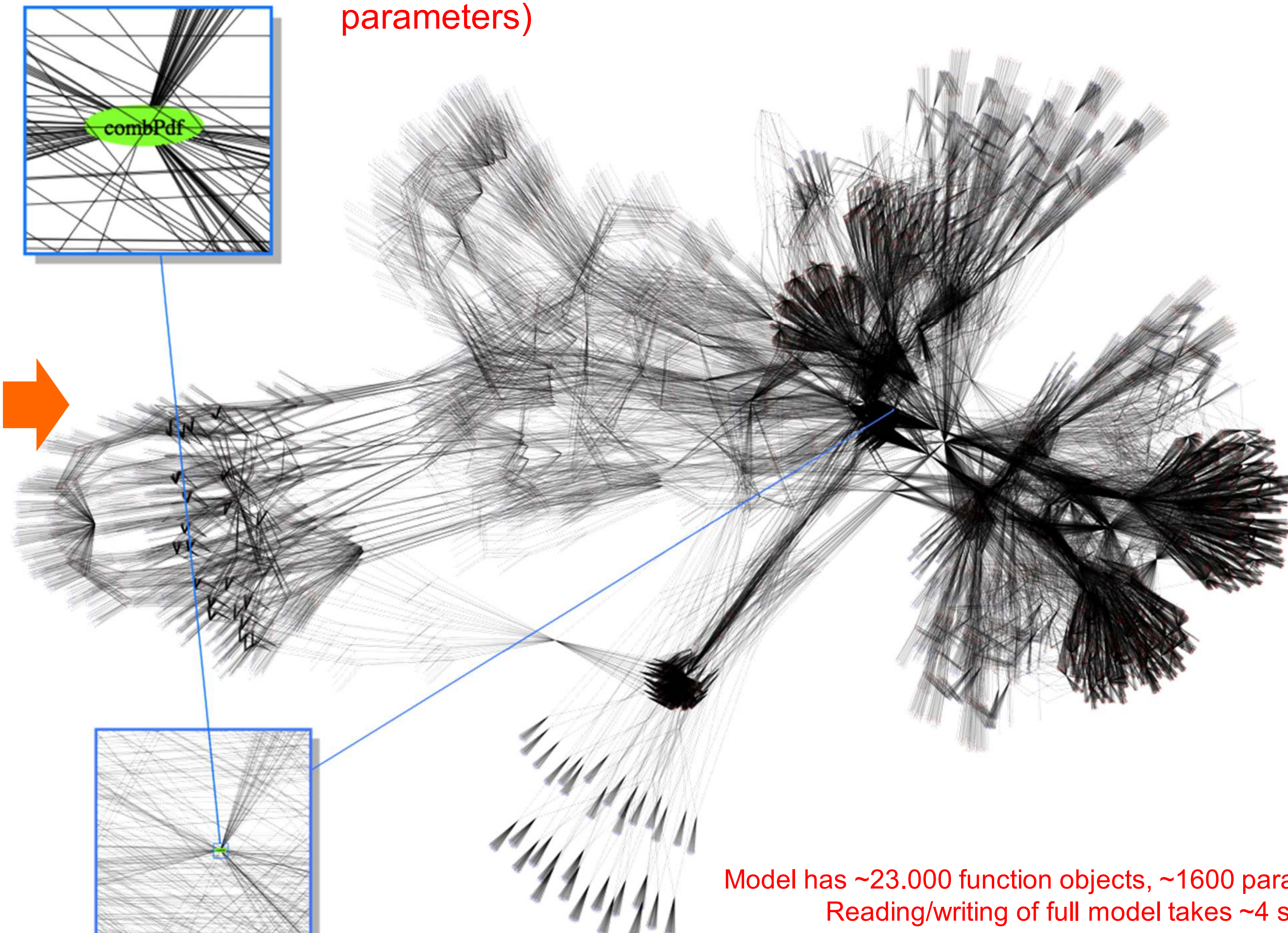

```
// Resurrect model and data
TFile f("model.root") ;
RooWorkspace* w = f.Get("w") ;
RooAbsPdf* model = w->pdf("sum") ;
RooAbsData* data = w->data("xxx") ;

// Use model and data
model->fitTo(*data) ;

RooPlot* frame =
        w->var("dt")->frame() ;
data->plotOn(frame) ;
model->plotOn(frame) ;
```



The full ATLAS Higgs combination in a single workspace…

Atlas Higgs combination model (23.000 functions, 1600 parameters)





Model has ~23.000 function objects, ~1600 parameters
Reading/writing of full model takes ~4 seconds
ROOT file with workspace is ~6 Mb

Being a good physicist – Understand your model!

- Full (profile) likelihood treats physics and subsidiary measurement on equal footing

$$L(N,0\,|\,s,\alpha) = \underbrace{Poisson(N\,|\,s+b(1+0.1\alpha))}_{\text{Physics measurement}} \cdot \underbrace{Gauss(0\,|\,\alpha,1)}_{\text{Subsidiary measurement}}$$

Physics measurement → "measures s"

Subsidiary measurement → "measures α"

- Our mental picture:

"dependence on α weakens inference on s"

- Is this picture (always) correct?

Understanding your model – what constrains your NP

- The answer is no – not always! Your physics measurement may in some circumstances constrain α *better* than your subsidiary measurement.
- Doesn't happen in Poisson counting example
  - Physics likelihood has no information to distinguish effect of s from effect of α

$$L(N,0\,|\,s,\alpha) = \underbrace{Poisson(N\,|\,s+b(1+0.1\alpha))}_{\text{Physics measurement}} \cdot \underbrace{Gauss(0\,|\,\alpha,1)}_{\text{Subsidiary measurement}}$$

- But if physics measurement is based on a distribution or comprises multiple distributions this is well possible

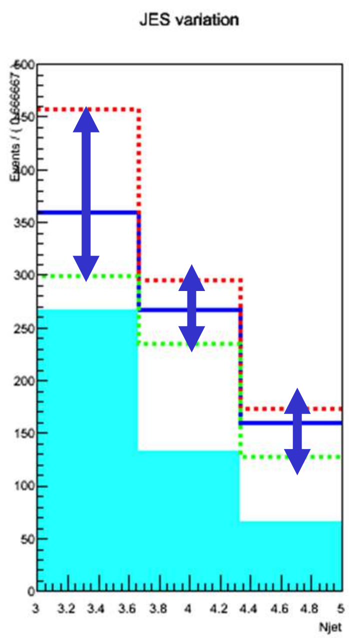
Understanding your model – what constrains your NP

- A case study – measuring jet multiplicity (3j,4j,5j)

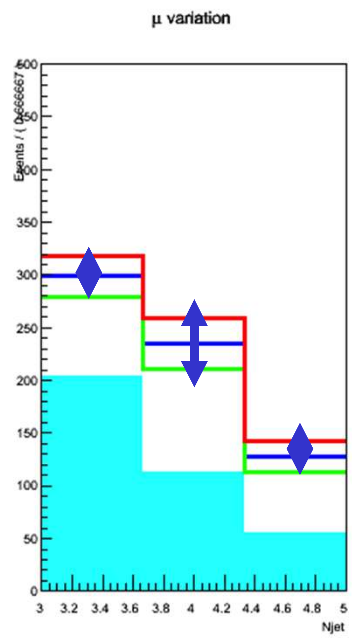
$$L(\vec{N}\,|\,\mu,\alpha_{JES}) = \prod_{i=3,4,5} Poisson(N_i\,|\,(\mu\cdot \vec{s}_i\cdot + \vec{b}_i)\cdot r_s(\alpha_{JES})))\cdot Gauss(0\,|\,\alpha_{JES},1)$$

- Signal mildly peaks in 4j bin, sits on top of a falling background

Effect of changing α_{JES}

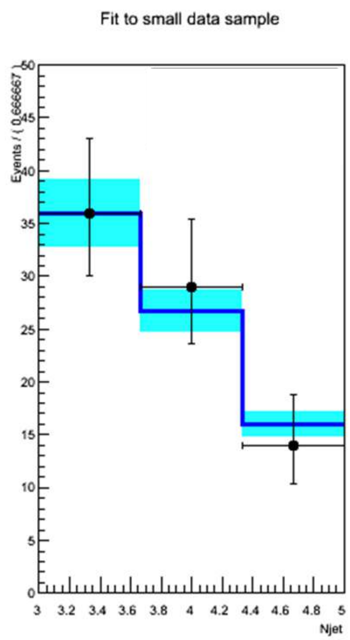


Effect of changing μ

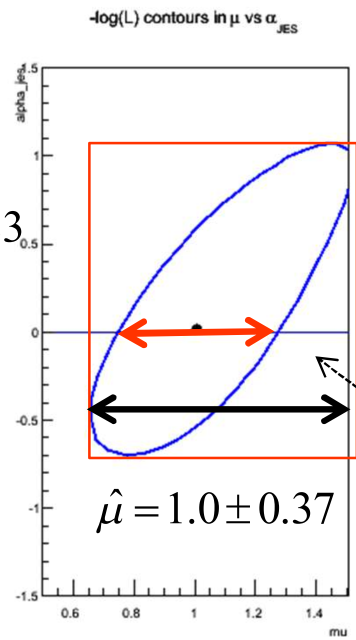


Understanding your model – what constrains your NP

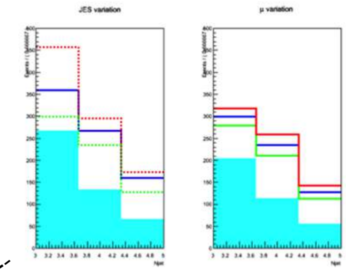
- Now measure (μ,α) from data – 80 events



$\hat{\alpha} = 0.01 \pm 0.83$



$\hat{\mu} = 1.0 \pm 0.37$



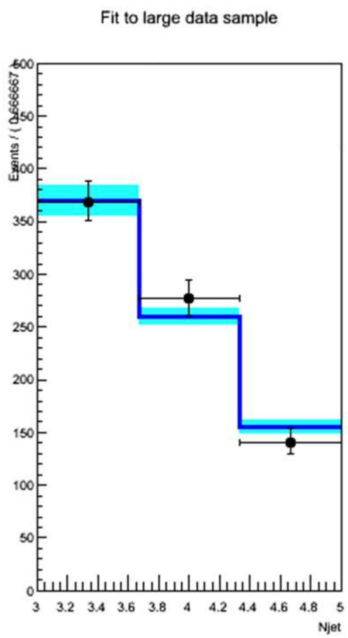
Estimators of μ, α correlated due to similar response in physics measurement

Uncertainty on μ with/without effect of JES

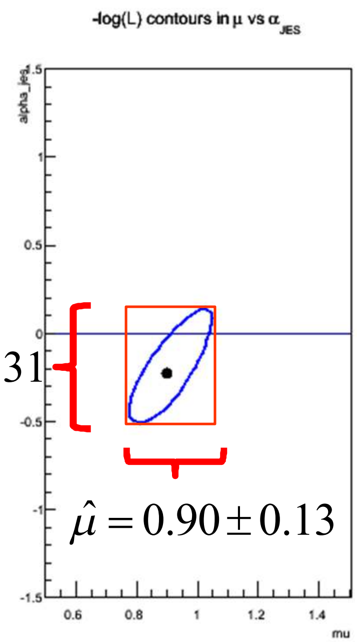
- Is this fit OK?
  - Effect of JES uncertainty propagated in to μ via response modeling in likelihood. Increases total uncertainty by about a factor of 2
  - Estimated uncertainty on α is not precisely 1, as one would expect from unit Gaussian subsidiary measurement…

Understanding your model – what constrains your NP

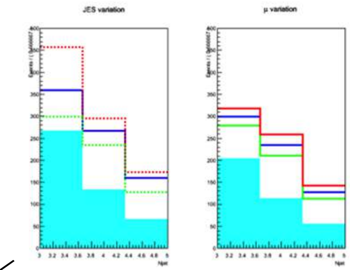
- The next year – 10x more data (800 events) repeat measurement with same model



$\hat{\alpha} = -0.23 \pm 0.31$



$\hat{\mu} = 0.90 \pm 0.13$



Estimators of μ, α correlated due to similar response in physics measurement

- Is this fit OK?
  - Uncertainty of JES NP *much reduced* w.r.t. subsidiary meas. ($\alpha = 0 \pm 1$)
  - Because the physics likelihood can measure it better than the subsidiary measurement (the effect of μ, α are sufficiently distinct that both can be constrained at high precision)

Understanding your model – what constrains your NP

- Is it OK if the physics measurement constrains NP associated with a systematic uncertainty better than the designated subsidiary measurement?
  - From the statisticians point of view: no problem, simply a product of two likelihood that are treated on equal footing 'simultaneous measurement'
  - From physicists point of view? Measurement is only valid is model is valid.
- Is the probability model of the physics measurement valid?

$$L(\vec{N} \mid \mu, \alpha_{JES}) = \prod_{i=3,4,5} Poisson(N_i \mid (\mu \cdot \vec{s}_i \cdot + \vec{b}_i) \cdot r_s(\alpha_{JES}))) \cdot Gauss(0 \mid \alpha_{JES}, 1)$$

- Reasons for concern
  - Incomplete modeling of systematic uncertainties,
  - Or more generally, model insufficiently detailed

Understanding your model – what constrains your NP

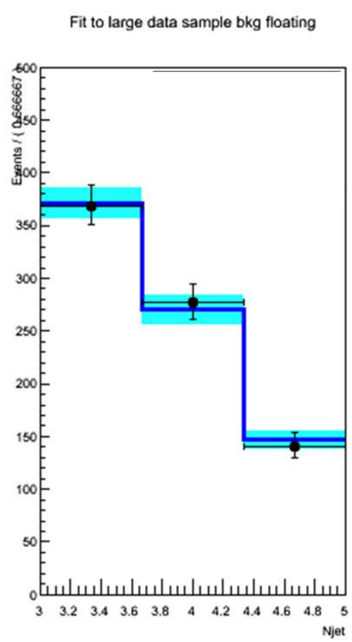
- What did we overlook in the example model?
  - The background rate has no uncertainty!
- Insert modeling of background uncertainty

$$L(\vec{N}\,|\,\mu,\alpha_{JES},\alpha_{bkg}) = \prod_{i=3,4,5} Poisson(N_i\,|\,(\mu\cdot\vec{s}_i\cdot+\vec{b}_i\cdot r_b(\alpha_{bkg}))\cdot r_s(\alpha_{JES})))\cdot Gauss(0\,|\,\alpha_{JES},1)\cdot Gauss(0\,|\,\alpha_{bkg},1)$$

Background rate response function

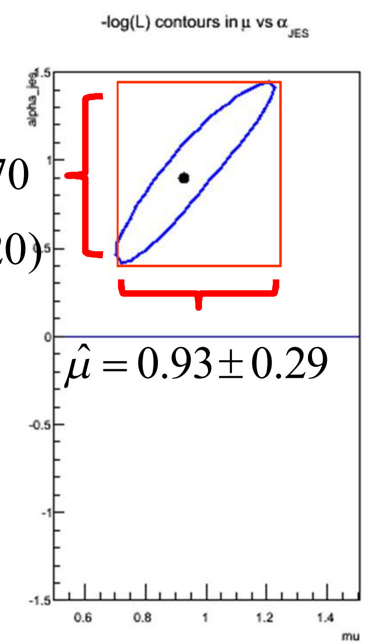
Background rate subsidiary measurement

- With improved model accuracy estimated uncertainty on both α_{JES}, μ goes up again…
  - Inference weakened by new degree of freedom α_{bkg}
  - NB α_{JES} estimate still deviates a bit from normal distribution estimate…



$$\hat{\alpha}_{JES} = 0.90 \pm 0.70$$

$$(\hat{\alpha}_{bkg} = 1.36 \pm 0.20)$$



$$\hat{\mu} = 0.93 \pm 0.29$$

Understanding your model – what constrains your NP

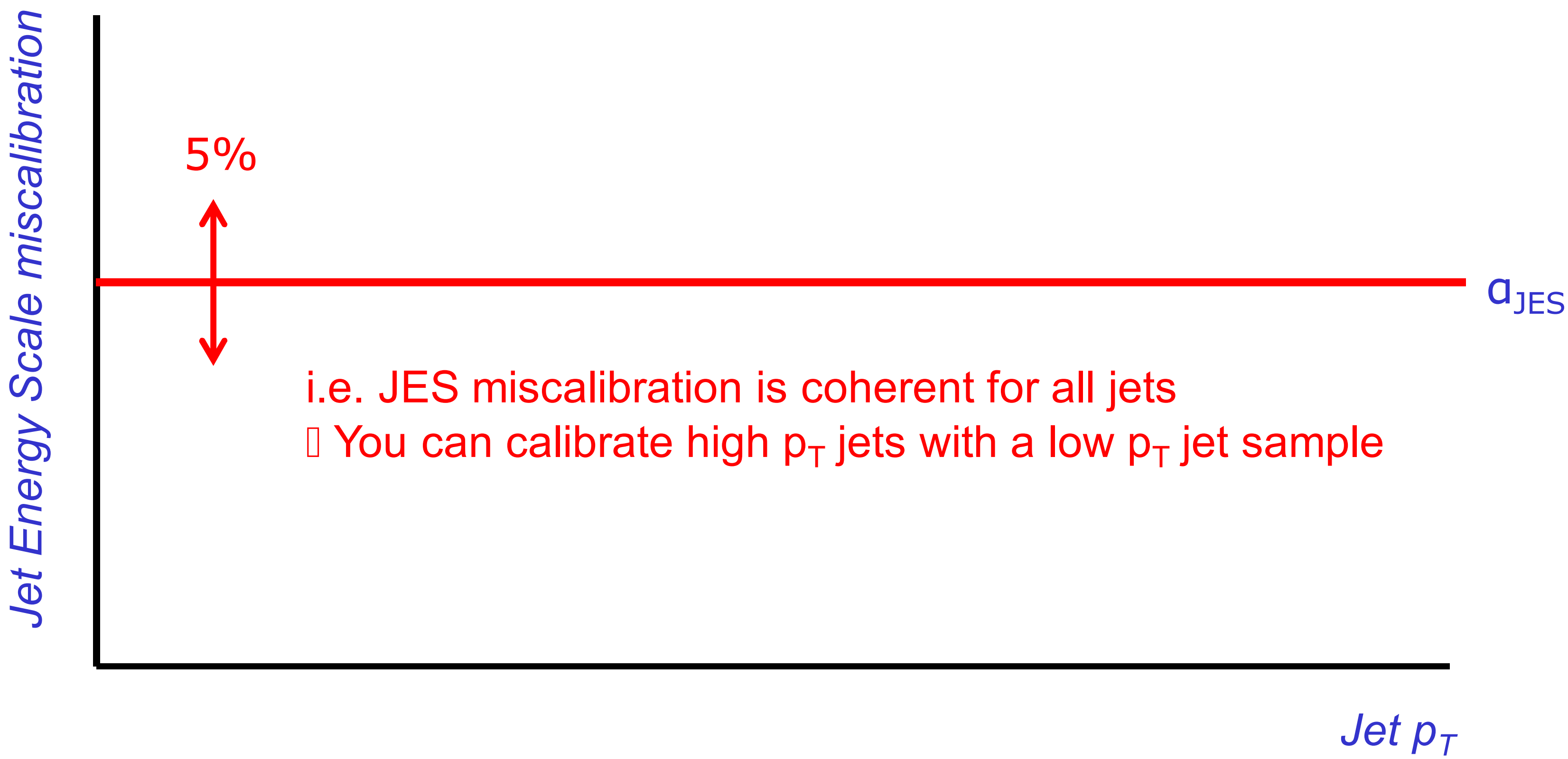
- Lesson learned: if probability model of a physics measurement is insufficiently detailed (i.e. flexible) you can *underestimate* uncertainties
- Normalized subsidiary measurement provide an excellent diagnostic tool
  - Whenever estimates of a NP associated with unit Gaussian subsidiary measurement deviate from $\alpha = 0 \pm 1$then physics measurement is constraining or biases this NP.
- Is 'over-constraining' of systematics NPs always bad?
  - No, sometimes there are good arguments why a physics measurement can measure a systematic uncertainty better than a dedicated calibration measurement (that is represented by the subsidiary measurement)
  - Example: in sample of reconstructed hadronic top quarks t→bW(qq), the pair of light jets should always have m(jj)=mW. For this special sample of jets it will possible to calibrate the JES better than with generic calibration measurement

Commonly heard arguments in discussion on over-constraining

- Overconstraining of a certain systematic is OK "because this is what the data tell us"
  - It is what the data tells you *under the hypothesis that your model is correct*. The problem is usually in the latter condition
- "The parameter α_{JES} should not be interpreted as Jet Energy Scale uncertainty provided by the jet calibration group"
  - A systematic uncertainty is always combination of response prescription and one or more nuisance parameters uncertainties.
  - If you implement the response prescription of the systematic, then the NP in your model really is the same as the prescriptions uncertainty
- "My estimate of $\alpha_{JES} = 0 \pm 0.4$ doesn't mean that the 'real' Jet Energy Scale systematic is reduced from 5% to 2%
  - It certainly means that in your analysis a 2% JES uncertainty is propagated to the POI instead of the "official" 5%.
  - One can argue that the 5% shouldn't apply because your sample is special and can be calibrated better by a clever model, but this is a physics argument that should be documented with evidence for that (e.g. argument JES in t→bW(qq) decays)

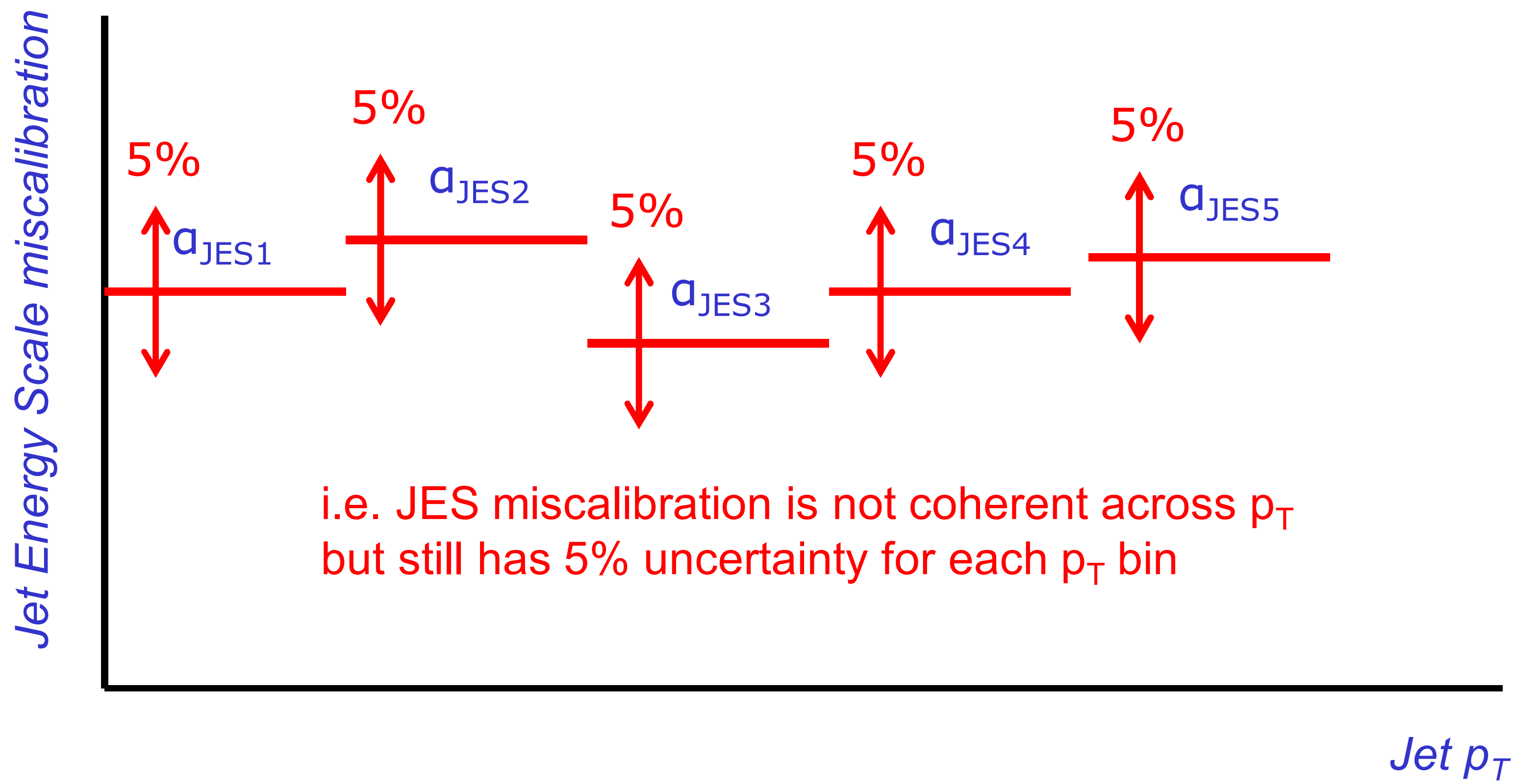
Dealing with over-constraining – introducing more NPs

- Some systematic uncertainties are not captured well by one nuisance parameter.
- Written prescription often not clear on *number* of nuisance parameters:
- Does "*the JES uncertainty is 5% for all jets*" mean one NP



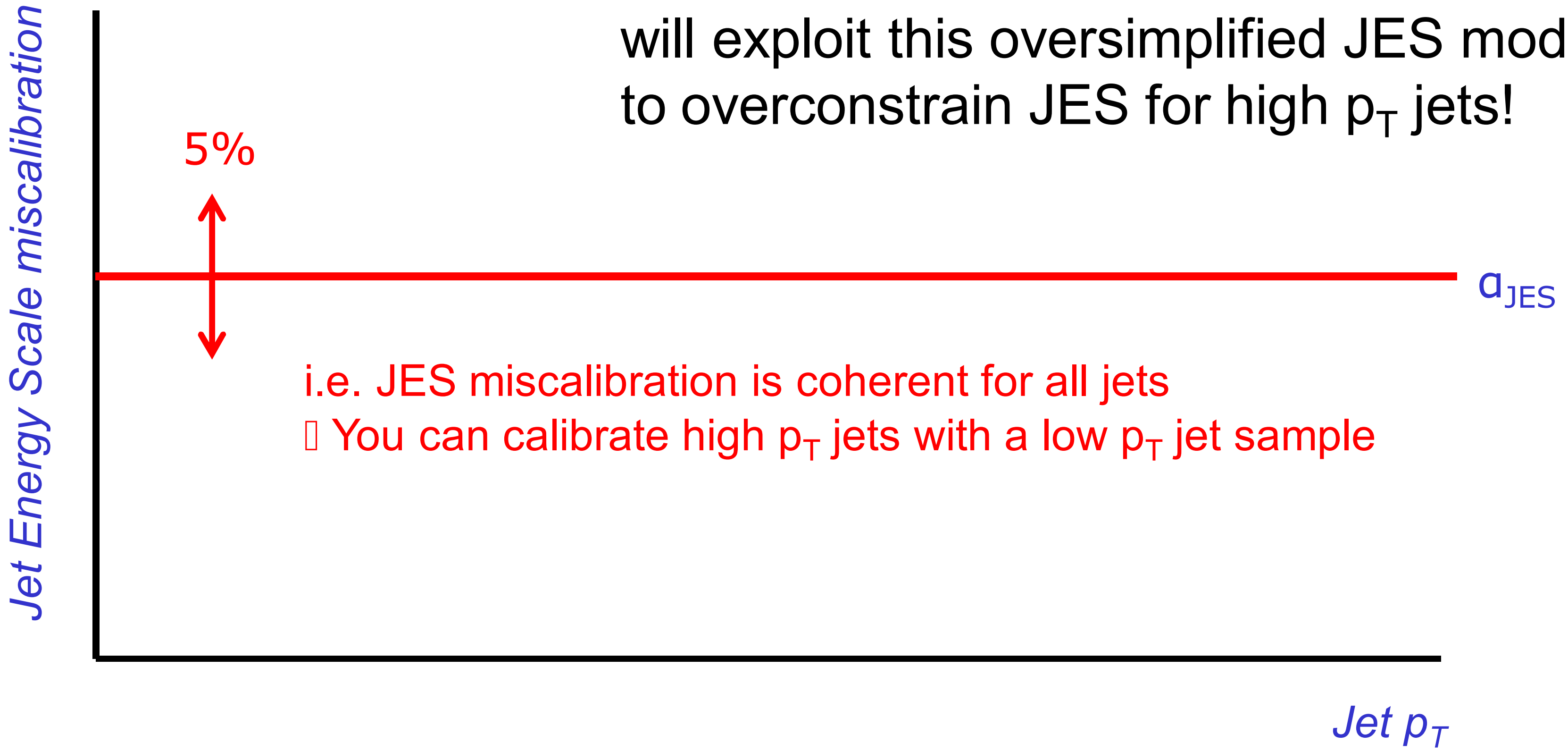
Dealing with over-constraining – introducing more NPs

- Some systematic uncertainties are not captured well by one nuisance parameter.
- Written prescription often not clear on *number* of nuisance parameters:
- Or does "*the JES uncertainty is 5% for all jets*" mean 5 NPs?



Dealing with over-constraining – introducing more NPs

- Some systematic uncertainties are not captured well by one nuisance parameter.
- Written prescription often not clear on *number* of nuisance parameters:
- If you assume one NP – chances are that your physics Likelihood will exploit this oversimplified JES model to overconstrain JES for high p_T jets!



Modeling theory uncertainties

- Modeling of systematic uncertainties originating from theory sources can pose some extra & thorny problems

Typical systematic uncertainties in HEP

- Detector-simulation related
  - "The Jet Energy scale uncertainty is 5%"
  - "The b-tagging efficiency uncertainty is 20% for jets with p_T<40"

Subsidiary measurement is an actual measurement → conceptually to a 'sideband' fit

- Physics/Theory related
  - The top cross-section uncertainty is 8%
  - "Vary the factorization scale by a factor 0.5 and 2.0 and consider the difference the systematic uncertainty"
  - "Evaluate the effect of using Herwig and Pythia and consider the difference the systematic uncertainty"

Subsidiary measurement unclear, but origin of prescription may well be another measurement (if yes, like sideband, if no, what is source of info?)

- MC simulation statistical uncertainty
  - Effect of (bin-by-bin) statistical uncertainties in MC samples

Subsidiary measurement is a Poisson counting experiment (but now in MC events), otherwise conceptually identical to a 'sideband' fit

Wouter Verkerke, NIKHEF

Modeling theory uncertainties

- Difficulties are not in the modeling procedure, but in quantifying what precisely we know
- Difficulty 1 – What is distribution of the subsidiary measurement?
- Easy example – Top cross-section uncertainty

$$L_{full}(s,\sigma_{tt}) = Poisson(N_{SR} \mid s + \varepsilon_{tt} \cdot \sigma_{tt}) \cdot Gauss(\tilde{\sigma}_{tt} \mid \sigma_{tt}, 0.08)$$

“XS Uncertainty is 8%” Gaussian subsidiary with 8% uncertainty
(because XS uncertainty is ultimately from a measurement)

- Difficult example – Factorization scale uncertainty

$$L_{full}(s,\sigma_{tt}) = Poisson(N_{SR} \mid s + b(\alpha_{FS})) \cdot F(\tilde{\alpha}_{FS} \mid \alpha_{FS})$$

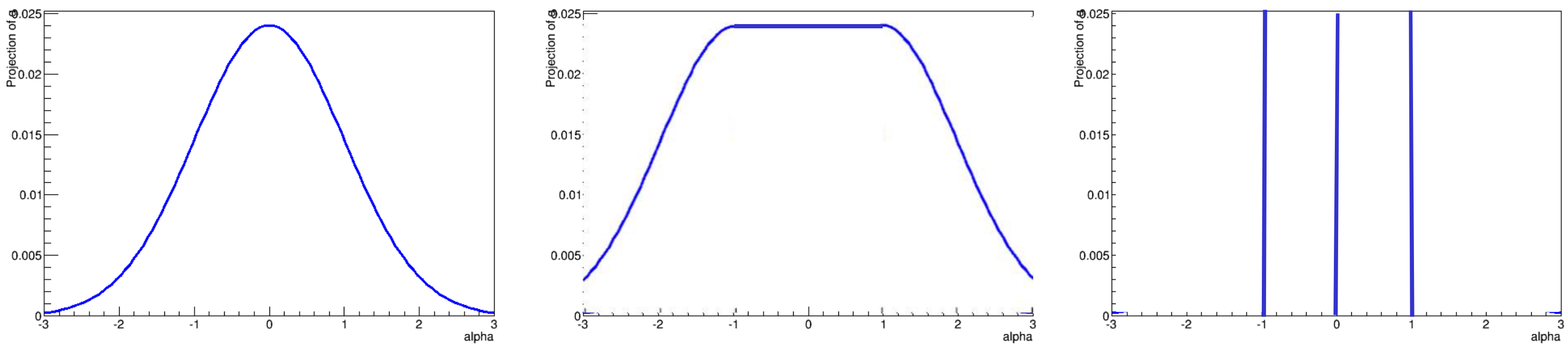
“Vary Factorization Scale by x0.5 and x” F(α) is probably not Gaussian
So what distribution was meant?

Modeling theory uncertainties

- Difficult example – Factorization scale uncertainty

$$L_{full}(s,\sigma_{tt}) = Poisson(N_{SR}\,|\,s+b(\alpha_{FS}))\cdot F(\tilde{\alpha}_{FS}\,|\,\alpha_{FS})$$

“Vary Factorization Scale by x0.5 and x” F(α) is probably not Gaussian
So what distribution was meant?



- Difficult arises from imprecision in original prescription.
  - NB: Issue is *physics* question, not a statistical procedure question. Answer will also need to be motivated with physics arguments
- Note that you *always* assume some distribution (even if you do error propagation) Profiling approach requires you to write it out explicitly. This is *good*!

Modeling theory uncertainties

- Difficulty 2 – What are the *parameters* of the systematic model?
- Easy example – Factorization scale uncertainty

$$L_{full}(s,\sigma_{tt}) = Poisson(N_{SR} \mid s + b(\alpha_{FS})) \cdot F(\tilde{\alpha}_{FS} \mid \alpha_{FS})$$

  – One parameter: the factorization scale Clearly described and connected to the underlying theory model

  – You can ask yourself if there are additional uncertainties in the theory model (renormalization scale etc), this a valid, but distinct issue.

- Difficult example – Hadronization/Fragmentation model

  – Source uncertainty: you run different showering MC generators (e.g. HERWIG and PYTHIA) and you observe you get different results from your physics analysis

  – **How do you model this in the likelihood?**

Modeling theory uncertainties

- Worst type of ‘theory’ uncertainty are prescriptions that result in an observable difference that cannot be ascribed to clearly identifiable effects. Examples of such systematic prescriptions
  – Evaluate measurement with Herwig and Pythia showering Monte Carlos and take the difference as systematic uncertainty
  – Evaluate measurement with CTEQ and MRST parton density functions and take the difference as systematic uncertainty.

- I call these ‘2-point systematics’.
  – You have the technical means to evaluate (typically) two known different configurations, but reasons for underlying difference are not clearly identified.

Specific issue with theory uncertainties

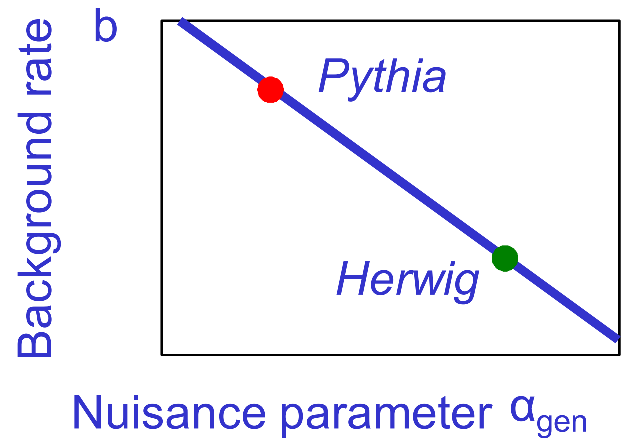
- It is difficult to define rigorous statistical procedures to deal with such 2-point uncertainties. So you need to decide
- If their estimated effect is small, you can pragmatically ignore these lack of proper knowledge and ‘just do something reasonable’ to model these effects in a likelihood
- If their estimated effect is large, your leading uncertainty is related to an effect that largely ununderstood effect. This is bad for physics reasons!
  - You should go back to the drawing board and design a new measurement that is less sensitive to these issues.
  - E.g. If your inclusive cross-section uncertainty is dominated by full fiducial acceptance uncertainty due to Herwig/Pythia issue, shouldn’t you rather be publishing the fiducial cross-section?

Specific issues with theory uncertainties

- Pragmatic solutions to likelihood modeling of '2-point systematics'
- Final solution will need to follow usual pattern

$$L(N \mid s, \alpha) = Poisson(N \mid s + b(\alpha)) \cdot SomePdf(0 \mid \alpha)$$

- Defining an (empirical) response function *b(α)* is the easy part

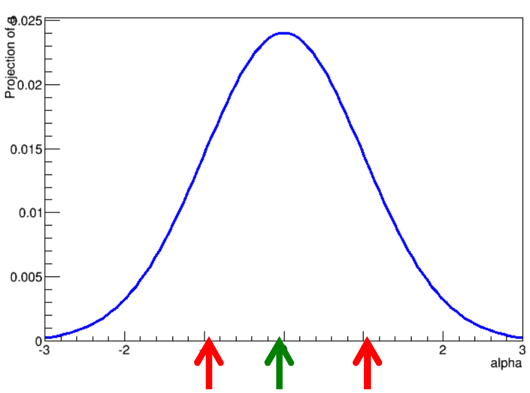


- A thorny question remains:
  What is the subsidiary measurement for α?
  This should reflect you current knowledge on α.

Specific issues with theory uncertainties

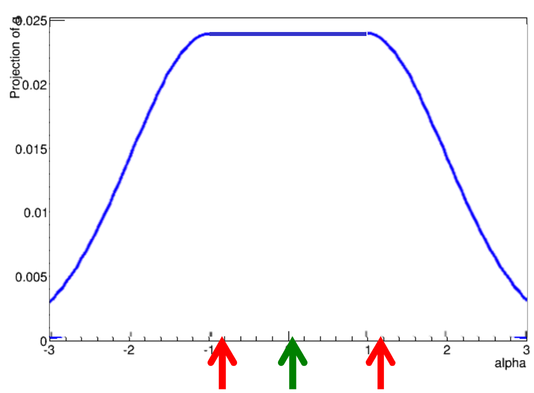
- Subsidiary measurement of a theoretical 2-point uncertainty effectively quantifies the ‘knowledge’ on these models
  - *Extra difficult to make meaningful statement about this*, since meaning of parameter is not well embedded in underlying theory model
  - But again, all procedures need to assume some distribution… Profiling requires you to spell it out
- Some options and their effects

Gaussian



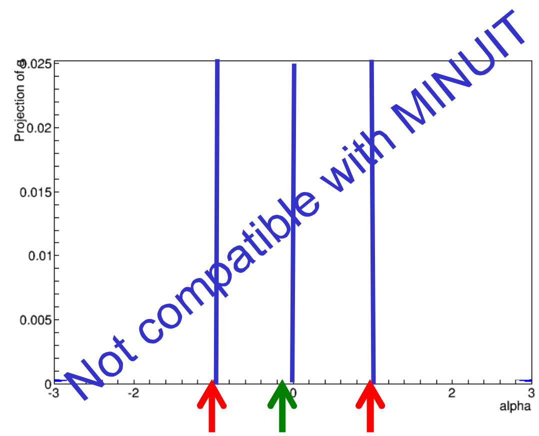
Prefers Herwig at 1σ

Box with
Gaussian wings



All predictions ‘between’ Herwig and Pythia equally probable

Delta fuctions

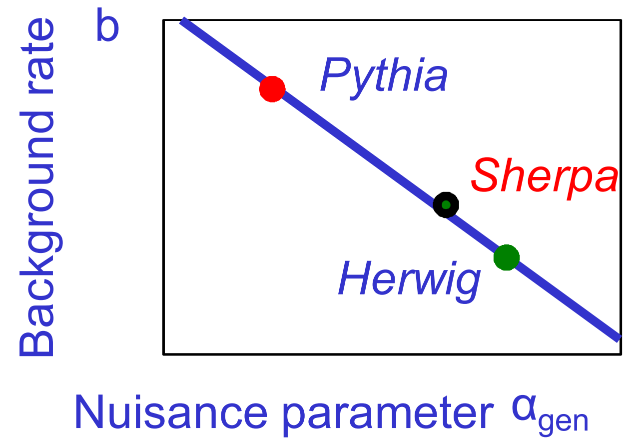


Only ‘pure’ Herwig and Pythia exist

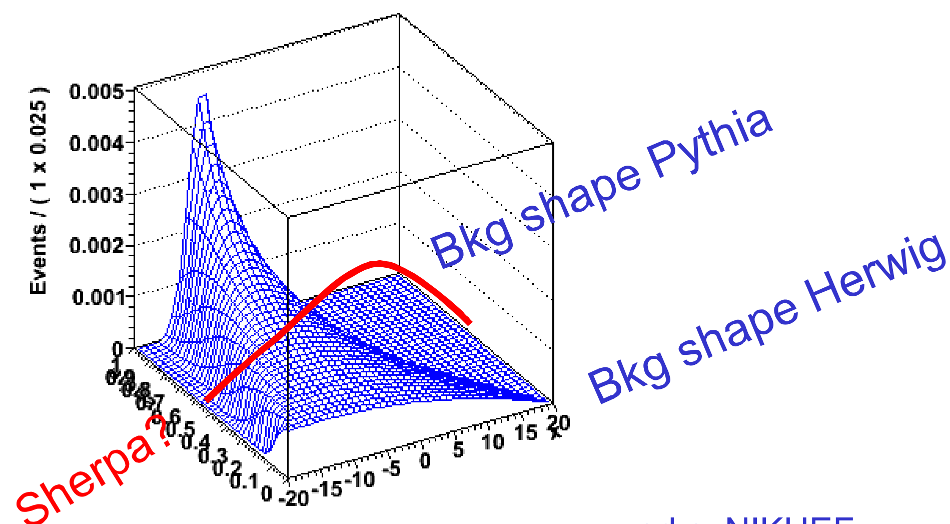
Wouter Verkerke, NIKHEF

Two-point systematics on non-counting measurements

- In a counting experiment you can argue that for every conceivable background rate there exists a value of the NP that corresponds to that rate
  - Even if 'SHERPA' was never used to construct the model, you can still represent its outcome

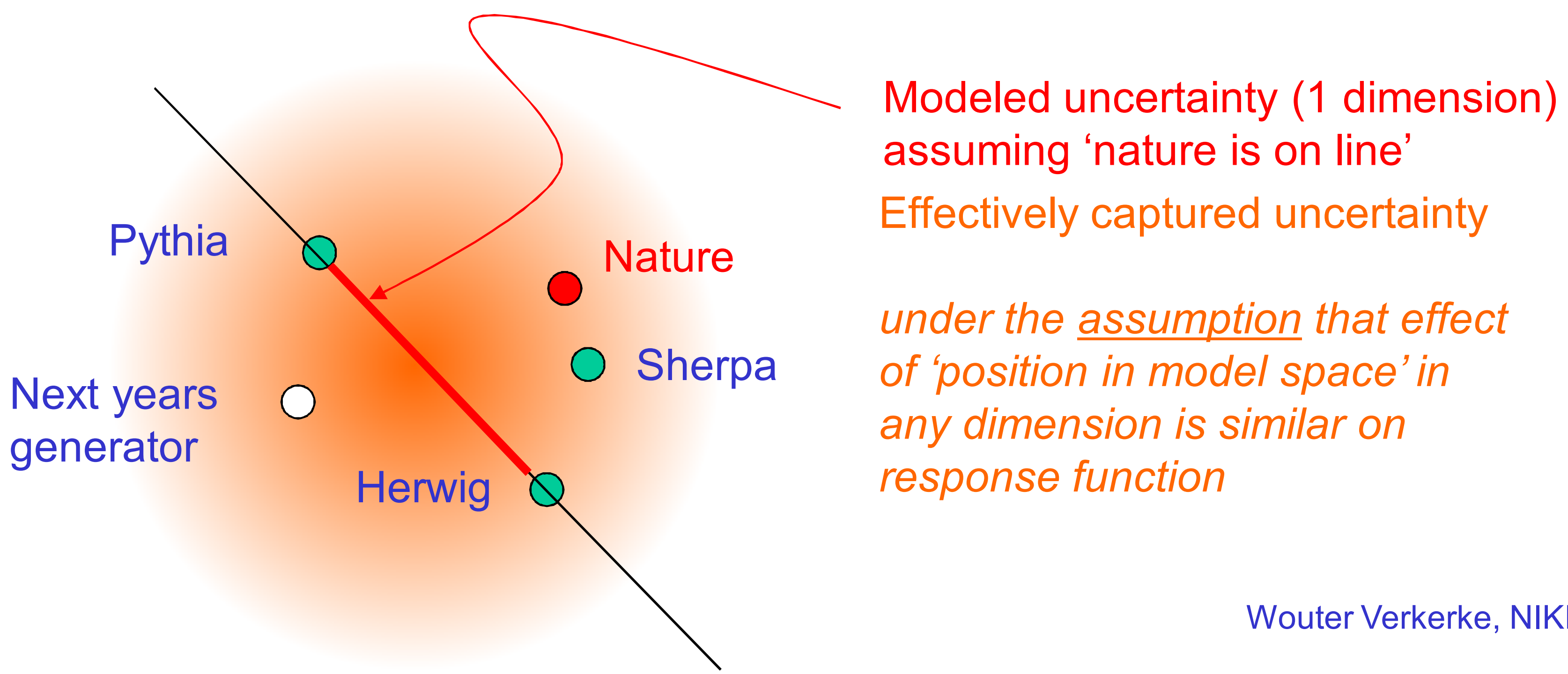


- This is not generally true for distributions. A shape interpolation between 'pythia' and 'herwig' does not necessarily describe shape of 'sherpa' (or of Nature!)
  - Fundamental modeling problem!
  - You may need more parameters…



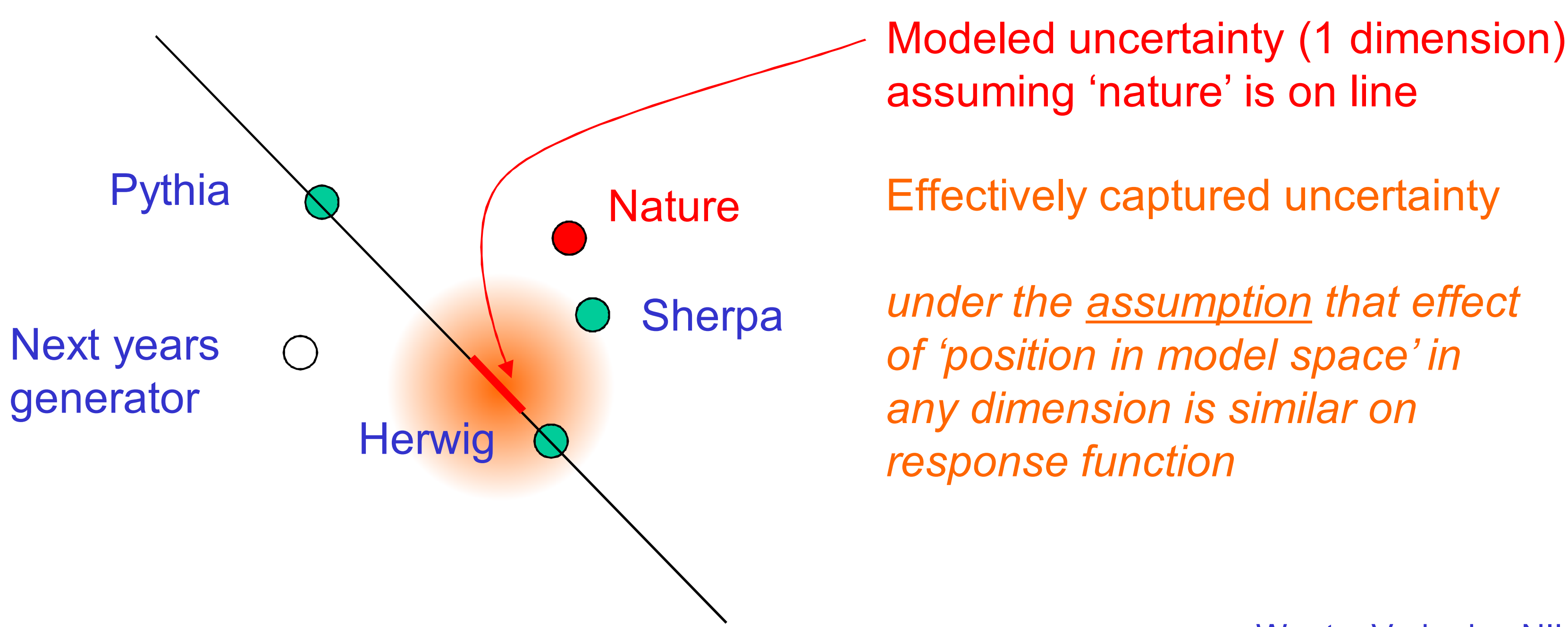
Dealing with 'two-point' uncertainties

- *Key issue: How many d.o.f. does you systematic uncertainty have?*
- Especially important in the discussion to what extent a two-point response function can be over-constrained.
  - A result α_{2p} = *0.5* ± *1* has 'reasonable' odds to cover the 'true generator' assuming all generators are normally scattered in an imaginary 'generator space'



Dealing with 'two-point' uncertainties

- *Key issue: How many d.o.f. does you systematic uncertainty have?*
- Especially important in the discussion to what extent a two-point response function can be over-constrained.
  - Does a hypothetical overconstrained result α_{2p} = *0.1 ± 0.2 'reasonably'* cover the generator model space?



Summary

- The key challenge for experimental physicist is to construct the likelihood function describing his analysis/experiment
- 'Profiling' is a technique allows to effectively incorporate all model uncertainties that are traditionally thought of as 'systematic uncertainties'
  - By empirically parametrizing the response of the full simulation chain
- Profiling enable used of all fundamental statistical inference techniques (frequentist/Bayesian), which start with the likelihood
  - A 'profile likelihood' allows execution of fundamental statistical techniques without cutting corners
  - Confidence intervals with guaranteed coverage, Bayesian posteriors, etc

Summary

- Profile likelihood implements and diagnoses many analysis issues that are missed by naïve approaches to systematic uncertainties (e.g. error prop)
  - "Posterior correlation" – Effect of correlations between systematics introduced by features of the physics measurement
  - "Overconstraining" – Either input magnitude was too conservative, or response model for systematic uncertainty was too simple (you'd like to know in either case)
  - "Imprecisely specified systematics" – Profiling requires physicist to explicit spell out precise model that is used
- But is important to run diagnostics on a profile likelihood model
  - Default interpretation in case of overconstraining is 'input uncertainty too conservative', which may lead to underestimated uncertainties if simplistic response model was the real problem
- 'Profiling' is the best way we know to incorporate systematic uncertainties is probability models