

Systematic uncertainties and profiling

Wouter Verkerke
(Nikhef/Atlas)



Wouter Verkerke, Nikhef

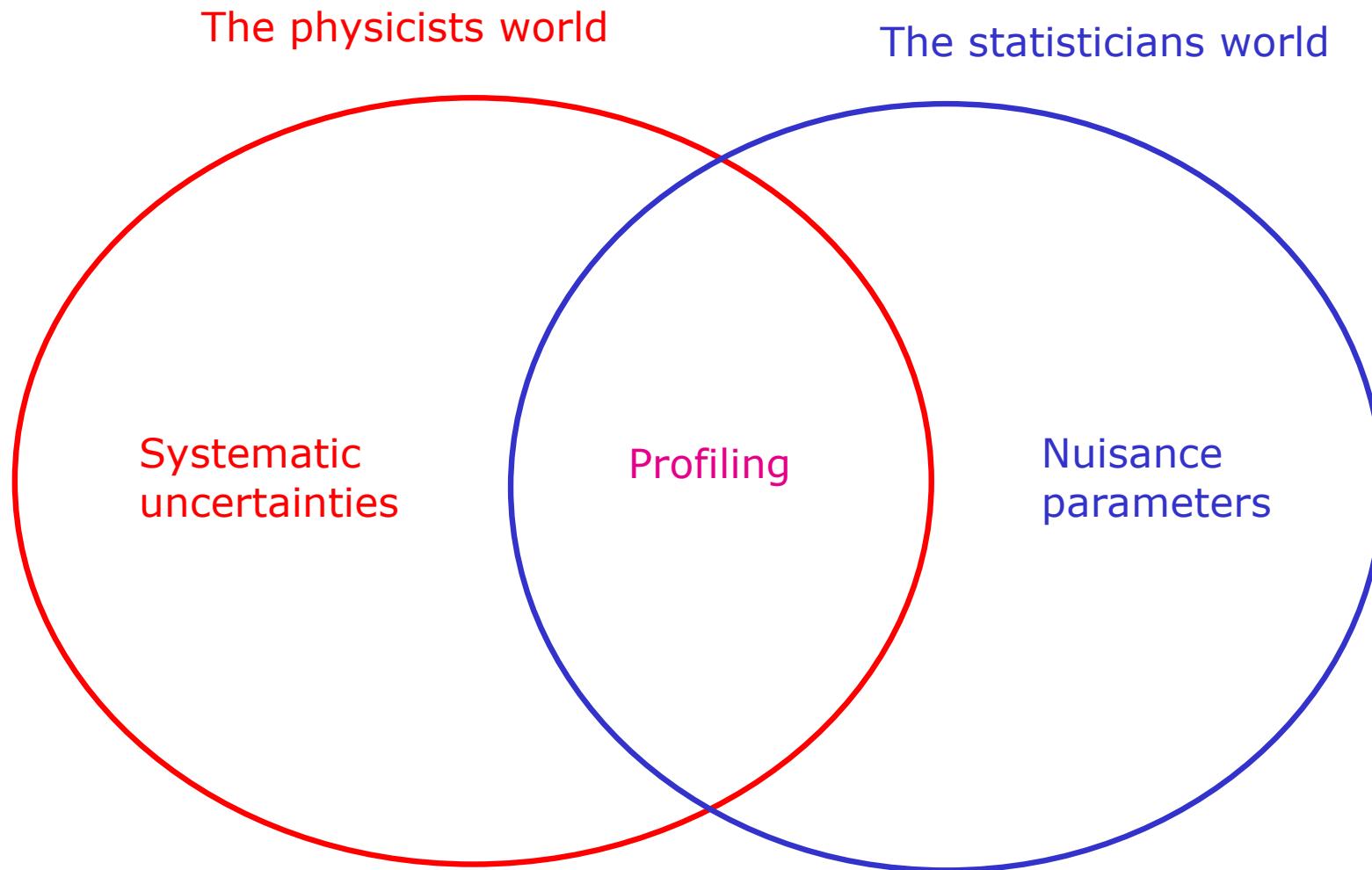
0

The scope of this course

Profiling & Systematics as part of statistical analysis

- A HEP analysis requires close integration of ‘physics concepts’ and ‘statistical concepts’
 1. Design event selection “physics”
 - Use simulated samples of signal, background to aid selection process (cuts, BDT, NN etc)
 2. Analyze (‘fit’) data in selection “statistics”
 - Measurement with statistical error, limit based on statistical uncertainty
 3. Make inventory of systematic uncertainties “physics”
 - Generally, any effect that isn’t measured constrained from your own measurement
 4. Finalize result ‘including systematics’ “statistics”
 - Variety of (empirical/fundamental) approaches to do this
 5. Interpretation “physics”
 - Better measurement, discovery etc, find mistake/sub-optimality in procedure
- Focus of this course: steps 3 and 4.
 - Practical problem: ‘physics notion’ of systematic uncertainties does not map 1-1 to a statistical procedure. Many procedures exist, some ad-hoc, some rigorous (from the statistical p.o.v.)

Profiling & Systematics as part of statistical data analysis



Outline of this course

- Outline of this course
 1. What are systematic uncertainties?
 2. The likelihood function as basis for statistical inference
 3. Incorporating systematic uncertainties in probability models
 4. Dealing with nuisance parameters in statistical inference
 5. Modeling shape systematics: template morphing
 6. Tools for modelling building
 7. Diagnostics I: Fit stability, understanding how minimizers work
 8. Diagnostics II: Result diagnostics, choice of nuisance parameters
 9. Summary

1

What are
systematic
uncertainties?

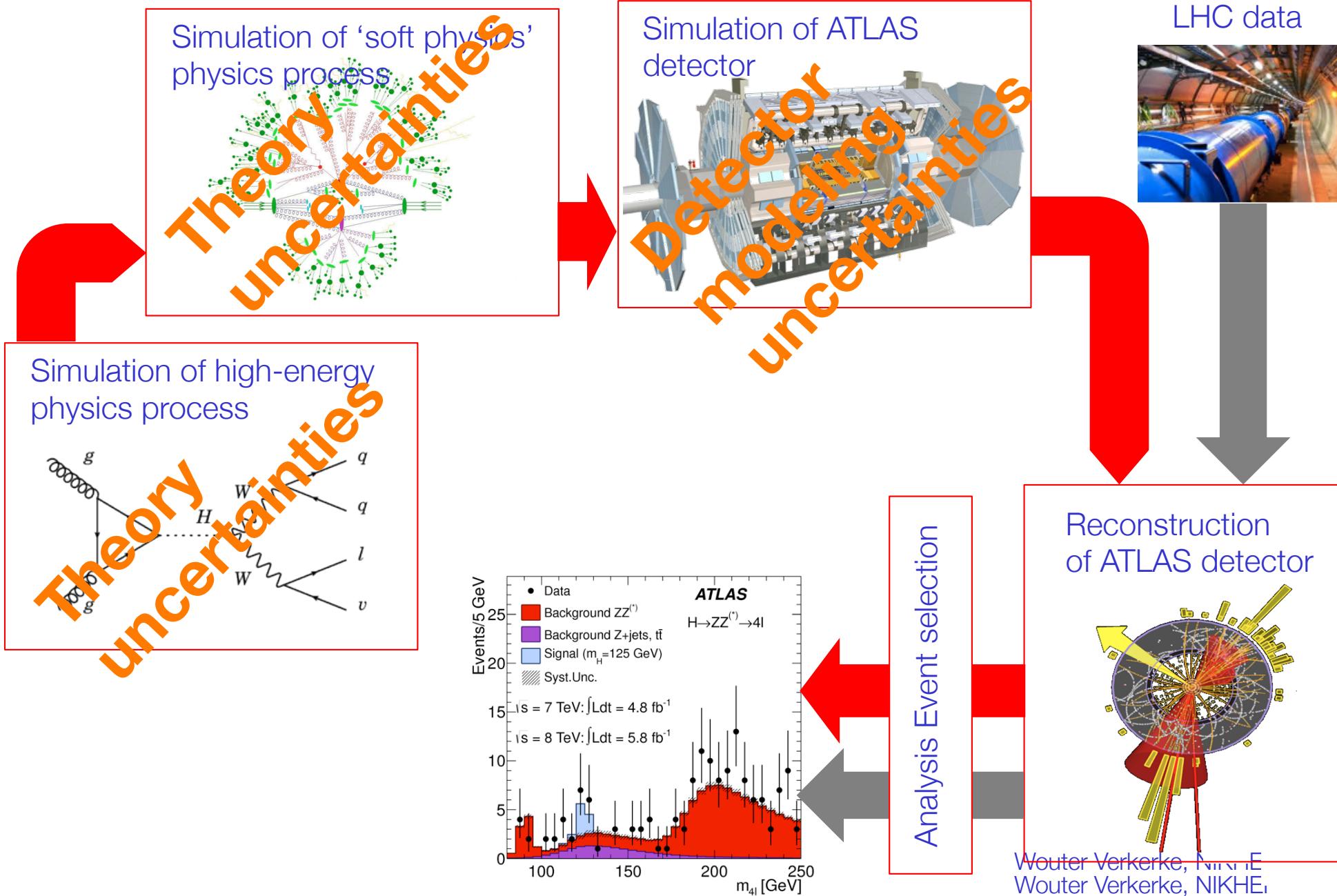
What are systematic uncertainties?

- 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?

Systematic uncertainty as a hidden measurement

- Consider 2 examples of measurements with systematic uncertainties
- Example 1: Measuring length of an object with a ruler
 - ‘Ruler calibration uncertainty’ is systematic uncertainty on length measurement
- Example 2: Counting measurement a signal in the presence of background
 - Measurement has (Poisson) statistical uncertainty.
 - Uncertainty on rate of background process introduces a systematic uncertainty on estimate of signal rate
- Is the ‘systematic uncertainty’ just a ‘hidden measurement’?
 - Ex 1: Ruler calibration could depend on temperature and uncertainty on current temperature could be dominant component of uncertainty
 - Ex 2: Background rate could be measured by a control sample

The simulation workflow and origin of uncertainties



Sources of systematic uncertainty in HEP

- Detector-simulation related uncertainty
 - Calibrations (electron, jet energy scale)
 - Efficiencies (particle ID, reconstruction)
 - Resolutions (jet energy, muon momentum)
- Theoretical uncertainties
 - Factorization/Normalization scale of MC generators
 - Choice of MC generator (ME and/or PS, e.g. Herwig vs Pythia)
- Monte Carlo Statistical uncertainties
 - Statistical uncertainty of simulated samples

Typical specifications of systematic uncertainties

- Detector-simulation related
 - “The Jet Energy scale uncertainty is 5%”
 - “The b-tagging efficiency uncertainty is 20% for jets with $p_T < 40$ ”
- Theory related
 - “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 related
 - Usually left unspecified – but quite clearly defined as a Poisson distribution with the ‘observed number of simulated events’ as mean.
 - But if MC events are weighted, it gets a bit more complicated.
- Note that specifications are often phrased as a prescription to be executed on the estimation procedure of the physics quantity of interest (‘vary and rerun...’) or can be easily cast this way.

Evaluating the effect of systematic uncertainties

- Often measurements are treated as a ‘black-box’
(e.g. as if it were a physical device that reports the measurement)
- **Inspires a ‘naive’ approach to systematic uncertainty evaluation:**
simply propagate ‘external systematic uncertainties’ into result
 - Evaluate nominal measurement (through unspecified procedure)

$$\mu_{nom} = \hat{\mu}$$

- Evaluate measurement at ‘ ± 1 sigma’ of some systematic uncertainty

$$\mu_{up} = \hat{\mu}(syst - up)$$

$$\mu_{down} = \hat{\mu}(syst - down)$$

- Calculate systematic uncertainty on measurement through numeric error propagation

$$\sigma_\mu(syst) = [\mu_{up} - \mu_{down}] / 2$$

- Repeat as needed for all systematic uncertainties,
add in quadrature for total systematic uncertainty.

$$\mu_{meas} = \mu_{nom} \pm \sigma(JES) \pm \dots$$

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
 - A maximum likelihood measurement is really nothing like a ‘device’
 - 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
 - **It's not based statistically rigorous procedures** (i.e. evaluation of systematic uncertainties is completely detached from statistical procedure used to estimate physics quantity of interest)
 - No calibrated probabilistic statements possible (95% C.L.)
 - No known good procedure for limit setting
- So what *should* we do with systematic uncertainties in statistical inference?

2

Probability models & the Likelihood

The goal of statistical inference: probabilistic statements

- (One-sided) confidence intervals (“limits”)

$$\sigma(X) < 10 \text{ pb at } 95\% C.L.$$

- Discovery of X

“Probability to obtain observed data or more extreme under hypothesis that X doesn’t exist is less than $1.1 \cdot 10^{-7}$ ”

- Measurements

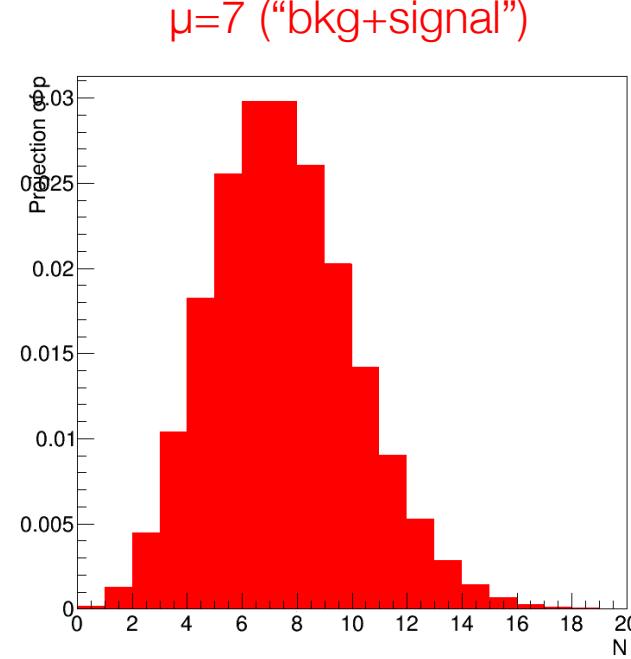
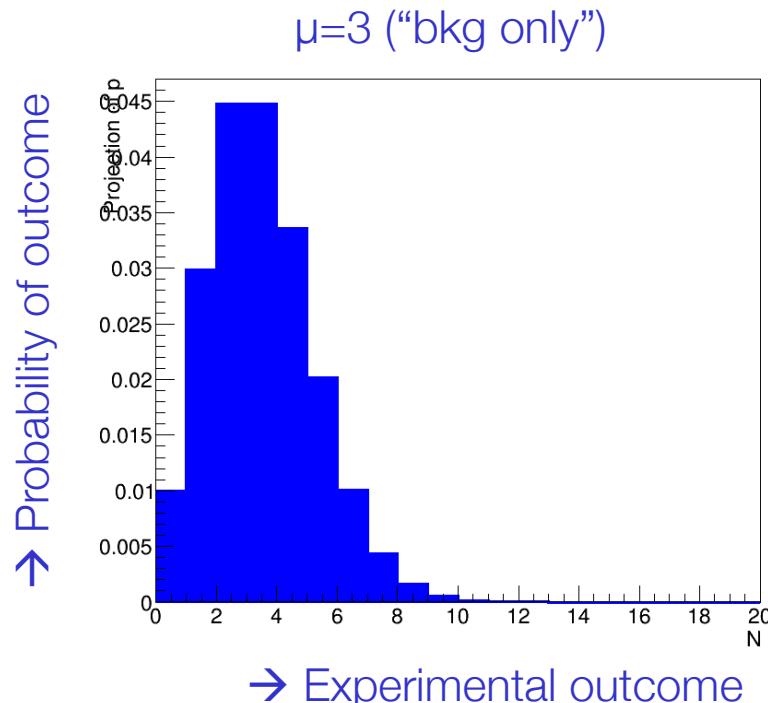
$$\sigma(X) = X \pm Y(\text{stat}) \pm Z(\text{syst}) \text{ pb}$$

- Before we discuss systematic uncertainties – review how these techniques work without systematic uncertainties

The statistical world

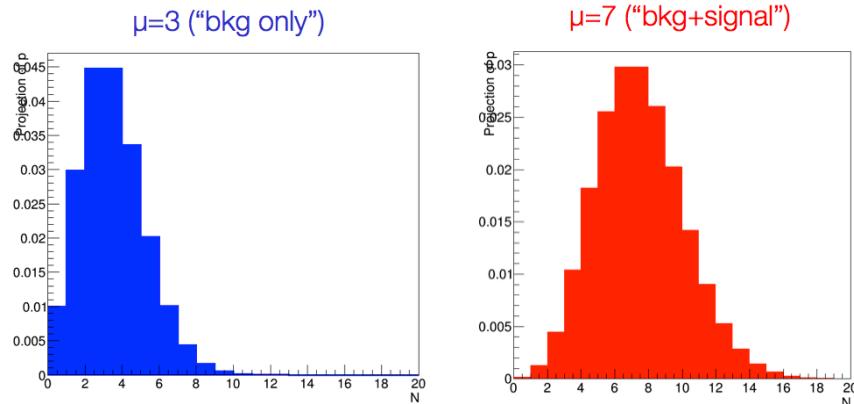
- Central concept in statistics is the '**probability model**'
- *A probability model assigns a probability to each possible experimental outcome.*
- Example: a HEP counting experiment
 - Count number of 'events' in a fixed time interval → Poisson distribution
 - Given the *expected event count*, the probability model is fully specified

$$P(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$$



Probabilities vs conditional probabilities

- Note that probability models strictly give *conditional* probabilities (with the condition being that the underlying hypothesis is true)



$$P(N) \rightarrow P(N | H_{bkg})$$

$$P(N) \rightarrow P(N | H_{sig+bkg})$$

Definition:
 $P(\text{data}|\text{hypo})$ is called
the **likelihood**

- Suppose we measure $N=7$ then can calculate

$$L(N=7 | H_{bkg}) = 2.2\%$$

$$L(N=7 | H_{sig+bkg}) = 14.9\%$$

- Data is more likely under sig+bkg hypothesis than bkg-only hypo*
- Is this what we want to know? Or do we want to know $L(H_{s+b}|N=7)$?

Inverting the conditionality on probabilities

- This conditionality inversion relation is known as **Bayes Theorem**

$$P(B|A) = P(A|B) \times P(B)/P(A)$$

Essay "Essay Towards Solving a Problem in the Doctrine of Chances" published in Philosophical Transactions of the Royal Society of London in 1764



Thomas Bayes (1702-61)

- And choosing A=data and B=theory

$$P(\text{theo}|\text{data}) = P(\text{data}|\text{theo}) \times P(\text{theo}) / P(\text{data})$$

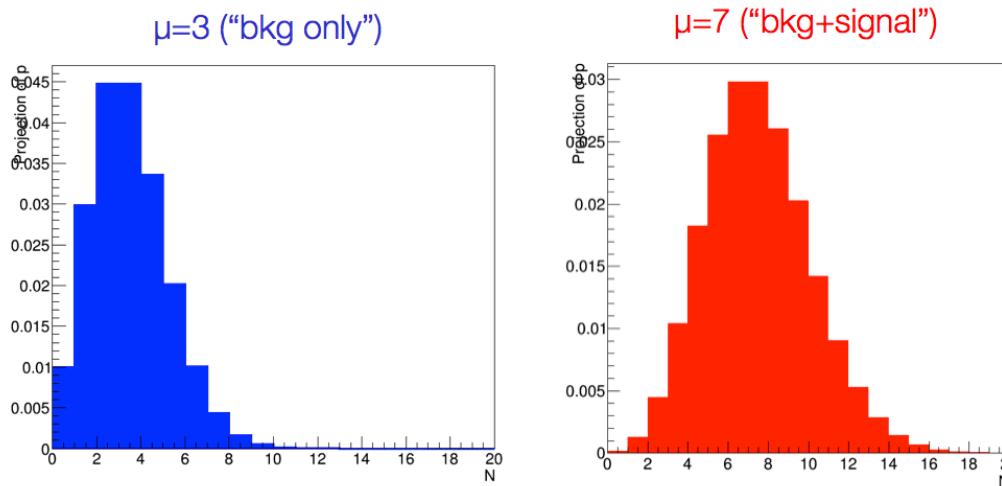
- *Return to original question:*

Do you $L(7|H_b)$ and $L(7|H_{sb})$ provide you enough information to calculate $P(H_b|7)$ and $P(H_{sb}|7)$

- **No!** → Need $P(A)$ and $P(B)$ → **Need $P(H_b)$, $P(H_{sb})$ and $P(7)$**

Summary on statistical test with simple hypotheses

- So far we considered simplest possible experiment we can do: counting experiment
- For a set of 2 or more completely specified (i.e. simple) hypotheses



- Given probability models $P(N|bkg)$, and $P(N|sig)$ we can calculate $P(N_{obs}|H_x)$ under either hypothesis
- With additional information on $P(H_i)$ we can also calculate $P(H_x|N_{obs})$

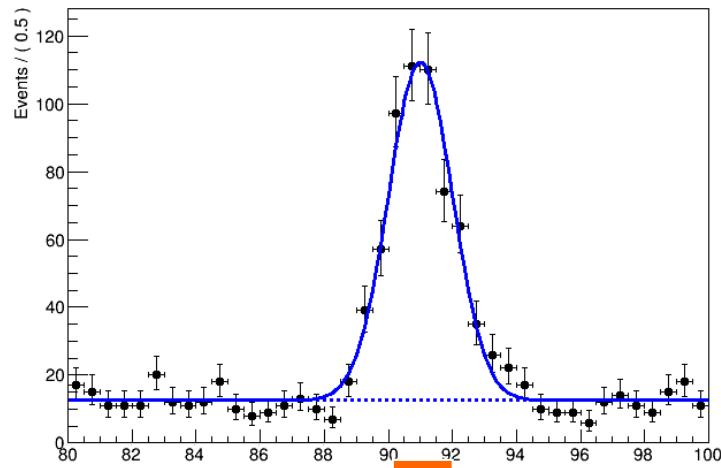
- In principle, *any potentially complex measurement (for Higgs, SUSY, top quarks) can ultimately take this a simple form.* But there is some ‘pre-work’ to get here – examining (multivariate) discriminating distributions → Now try to incorporate that

Discriminating observables & counting experiments

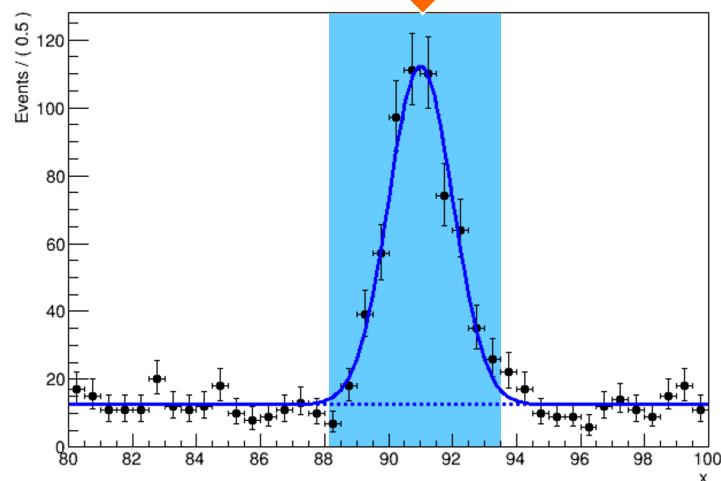
- HEP experimental data usually has many discriminating observables that carry information that can distinguish signal from background hypothesis
- In principle can use them all directly in an elaborate hypothesis test.
 - But would need to formulate **a model that describe the expected distribution of all of these → Complicated**
 - If expectations are uncertain (from simulation or theory) process of modeling becomes even more complex
- A pragmatic solution to reduce complexity is to split task in two
 - Define empirical **selection of events enriched in signal** using one or more observable properties of the event (invariant masses, distributions, angles etc)
 - Perform **statistical test** (hypothesis test, parameter estimation etc) on **sample that reduced in size** and in dimensionality of discriminating observables that are modeled
 - Most extreme reduction of dimensionality is to zero → **counting experiment**

Discriminating observables & counting experiments

- Example 1 – **Discrimination in selection stage only**



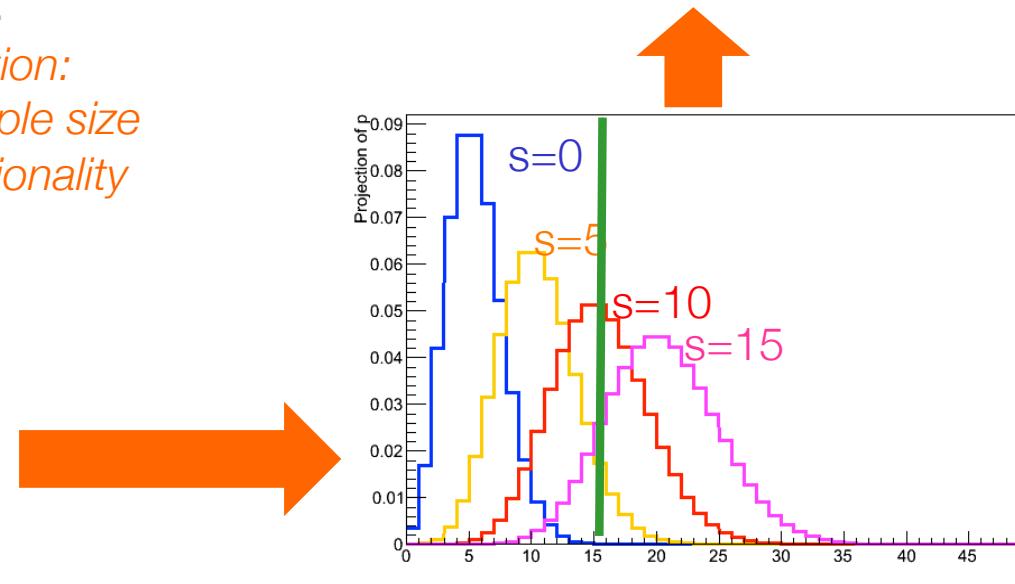
*Event selection:
reduce sample size
and dimensionality*



NB1: All discriminating power in selection step,
none in inference step. *This is a design choice!*

NB2: Selection must be tuned on a ‘figure of merit’
usually a simplified statistical inference test

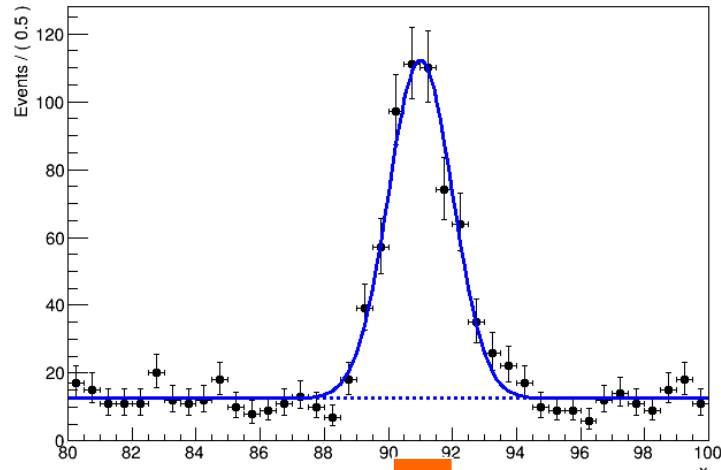
Statistical inference:
 $L(15|5) = 1.5 \cdot 10^{-4}$



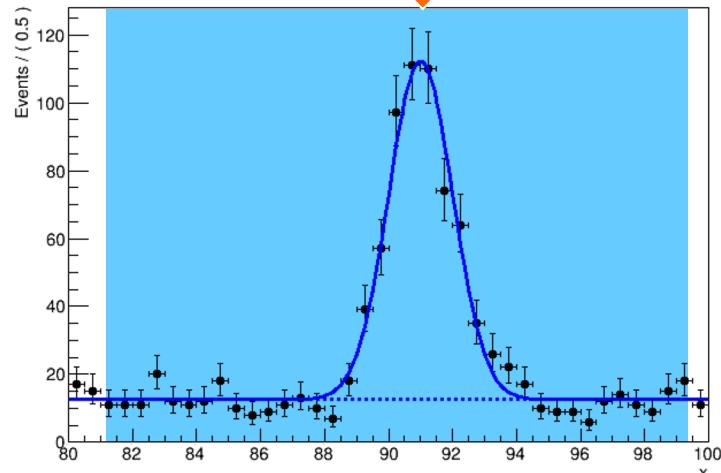
*Formulation of probability model of reduced sample:
 $\text{Poisson}(N|s+b)$*

Modeling discriminating observables

- Example 2 – **Discrimination in inference stage**



*Event selection:
reduce sample size
and dimensionality*

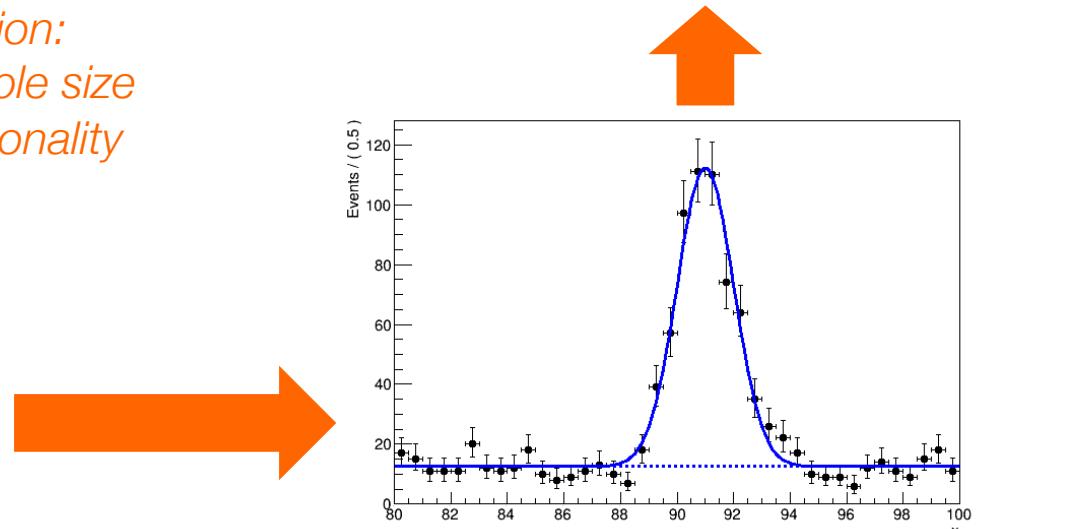


NB1: Most discrimination power in inference step.
This is again design choice!

NB2: Optimal selection less critical

NB3: Correct description of selected sample
more complex

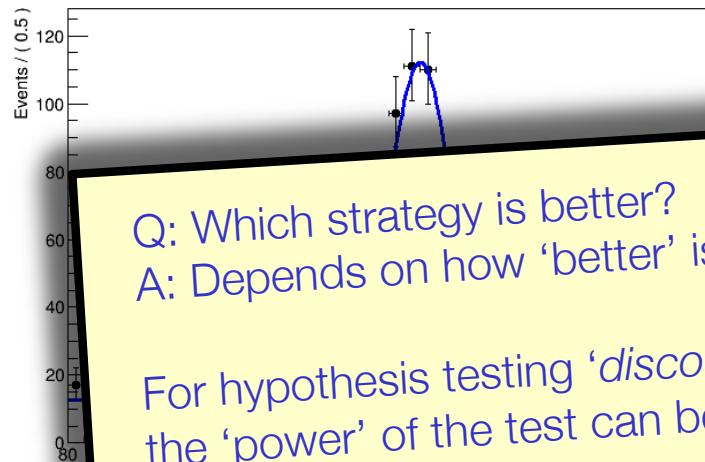
Statistical inference:
 $L(\text{data}|\text{hypo}) = \text{something}$



*Formulation of probability model of reduced sample:
 $Nbkg * \text{Uniform}(x) + Nsig * \text{Gaussian}(x)$*

Modeling discriminating observables

- Example 2 – full dataset has one discriminating observable: x

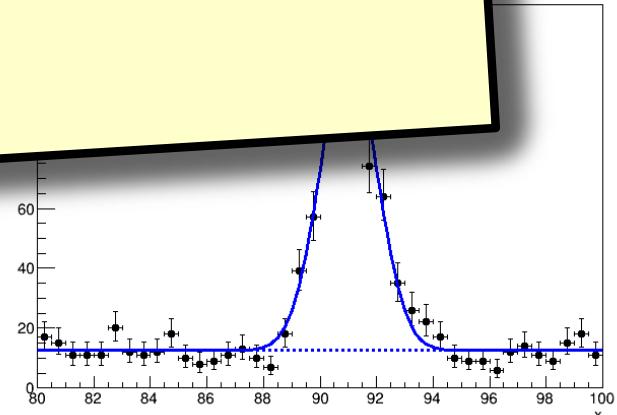
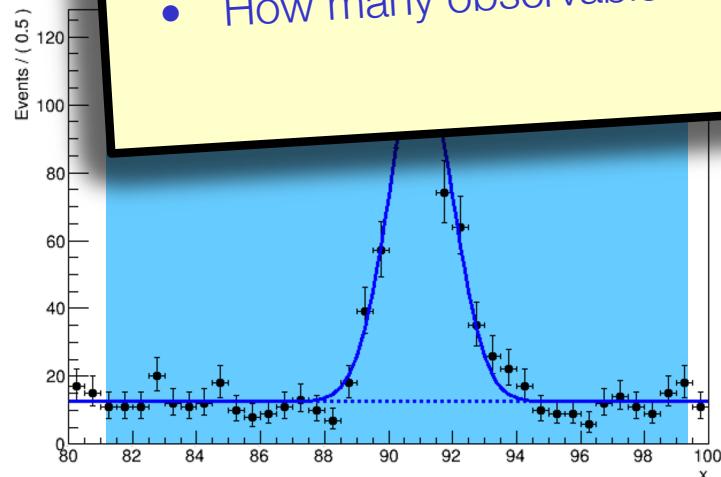


NB1: Most discrimination power in inference step.
This is again design choice!

Q: Which strategy is better?
A: Depends on how 'better' is defined?

For hypothesis testing '*discovery of a new article*'
the 'power' of the test can be the same, but doesn't need to be

Choice is real life largely dictated by practicalities
• How easy is it to formulate a description of the observables?
• How many observables are important?

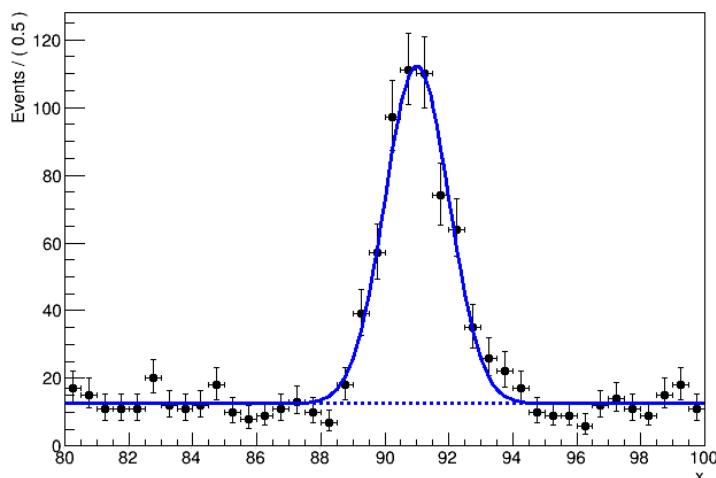


Formulation of probability model of reduced sample:
 $Nbkg * Uniform(x) + Nsig * Gaussian(x)$

PDFs with multiple process contributions

- Analogous to the counting model Poisson($N|S+B$), probability density models can describe the distribution of such hypothesis through simple addition

$$f(x) = f_{\text{sig}} \text{ Gaussian}(x) + (1-f_{\text{sig}}) \text{ Uniform}(x)$$



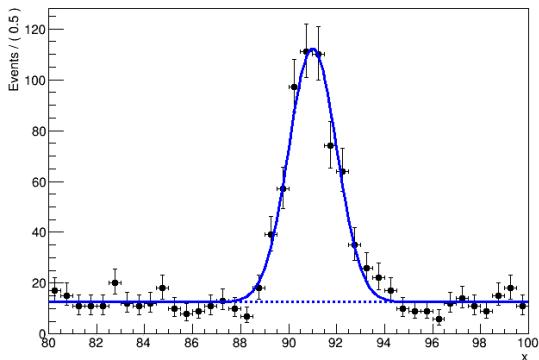
If $\text{Gaussian}(x)$ and $\text{Uniform}(x)$ are pdfs, then their sum is also a pdf, provided the sum of the coefficients is also 1

- Given a data sample $D(x)$ of N *independent identically distributed* observations of x , the Likelihood is

$$L(\vec{x}) = \prod_{i=0 \dots N} f(x_i)$$

PDFs with multiple process contributions

- Note that the Likelihood $L(x)$ of a probability density function $f(x)$ for a data sample $D(x)$ with N entries **only exploits the differential distribution in x , but not the event count N of the data**
- In many cases the event count can also distinguish the S/B hypothesis (more events expected if signal is present). If so, the probability model for the event count can be explicitly included in the Likelihood (often called ‘extended likelihood’)



$$f(x) = f_{\text{sig}} \text{Gaussian}(x) + (1-f_{\text{sig}}) \text{Uniform}(x)$$

$$P(N) = \text{Poisson}(N | N_{\text{exp}})$$

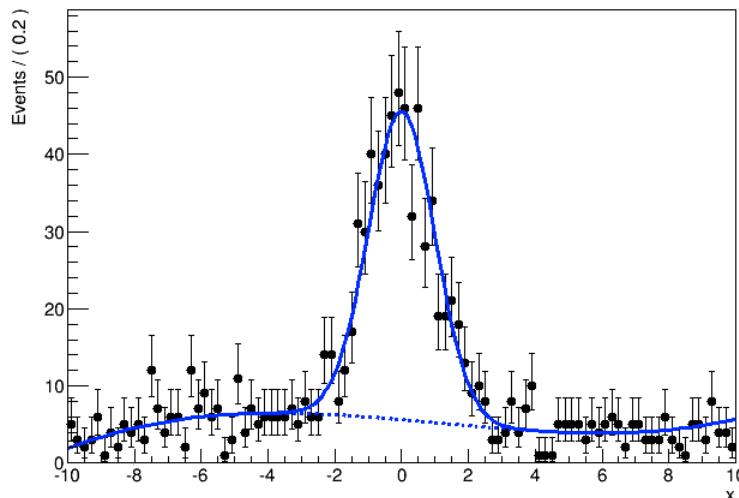
$$L(\vec{x}, N) = \prod_{i=0 \dots N} f(x_i | f_{\text{sig}}) \cdot \text{Poisson}(N | N_{\text{exp}})$$

- In the common case of a signal and background, with a respective expected event S and B , one can reparameterize $(f_{\text{sig}}, N_{\text{exp}}) \rightarrow (S, B)$

Empirical probability models

- In case no description from first principles exists for a differential distribution, empirical or simulation-based models can be deployed

Empirical models

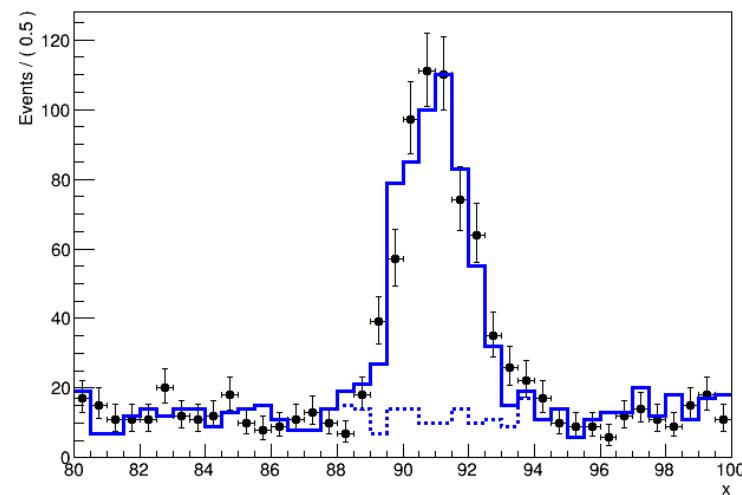


$$B(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \dots$$

Drawbacks:

- **Arbitrariness in parameterization**,
e.g. which order to choose
for a polynomial

Simulation-based models



$$B(x) = \text{histogram}$$

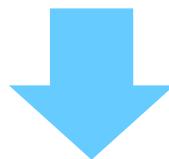
Drawbacks:

- **Quantization** of model prediction in bins
- Poor modeling in regions
with **low simulation statistics**

Working with Likelihood functions for distributions

- How do the statistical inference procedures change for Likelihoods describing *distributions*?
- Bayesian calculation of $P(\text{theo}|\text{data})$ they are *exactly the same*.
 - Simply substitute counting model with binned distribution model

$$P(H_{s+b} | \vec{N}) = \frac{L(\vec{N} | H_{s+b})P(H_{s+b})}{L(\vec{N} | H_{s+b})P(H_{s+b}) + L(\vec{N} | H_b)P(H_b)}$$

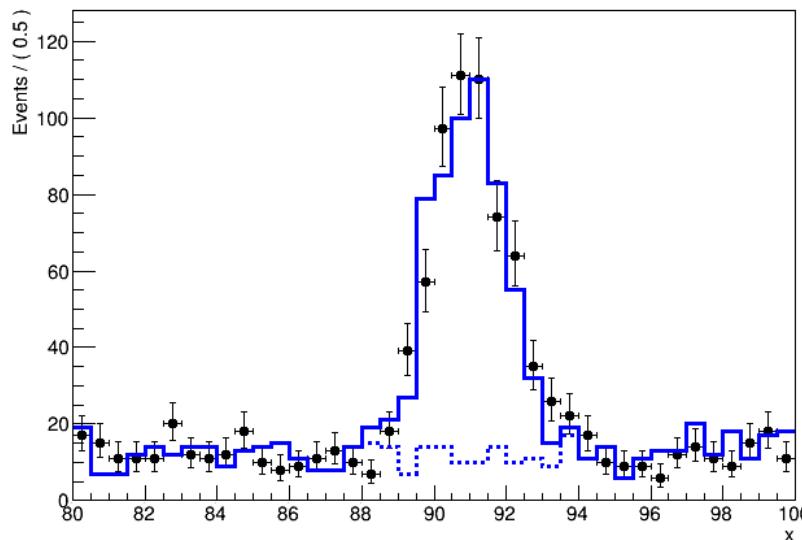


Simply fill in new Likelihood function
Calculation otherwise unchanged

$$P(H_{s+b} | \vec{N}) = \frac{\prod_i Poisson(N_i | \tilde{s}_i + \tilde{b}_i)P(H_{s+b})}{\prod_i Poisson(N_i | \tilde{s}_i + \tilde{b}_i)P(H_{s+b}) + \prod_i Poisson(N_i | \tilde{b}_i)P(H_b)}$$

Working with Likelihood functions for distributions

- Frequentist calculation of $P(\text{data}|\text{hypo})$ also unchanged, but **question arises if $P(\text{data}|\text{hypo})$ is still relevant?**



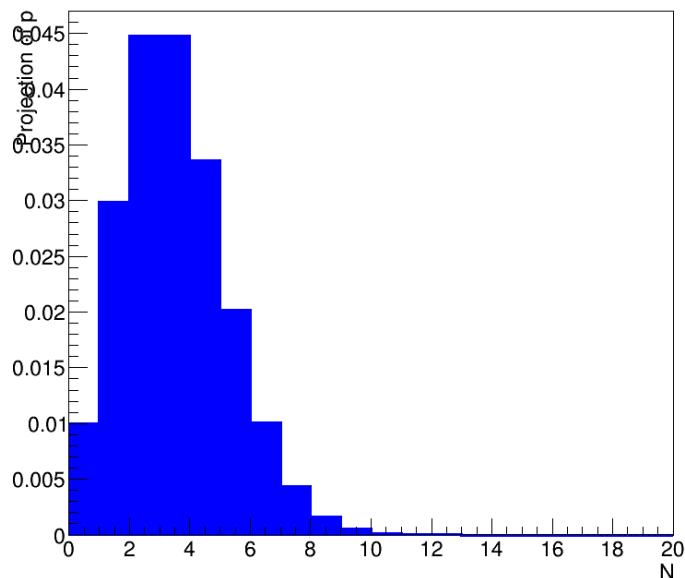
$$L(\vec{N} | H_b) = \prod_i \text{Poisson}(N_i | \tilde{b}_i)$$

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

- **$L(N|H)$ is probability to obtain exactly the histogram observed.**
- *Is that what we want to know?* Not really.. We are interested in probability to observe any ‘similar’ dataset to given dataset, or in practice dataset ‘similar or more extreme’ than observed data
- Need a way to quantify ‘similarity’ or ‘extremity’ of observed data

Working with Likelihood functions for distributions

- *Definition:* a test statistic $T(x)$ is any function of the data x
- We need a test statistic that will **classify ('order') all possible observations** in terms of ‘extremity’ (definition to be chosen by physicist)
- NB: For a counting measurement the count itself is already a useful test statistic for such an ordering (i.e. $T(x) = x$)

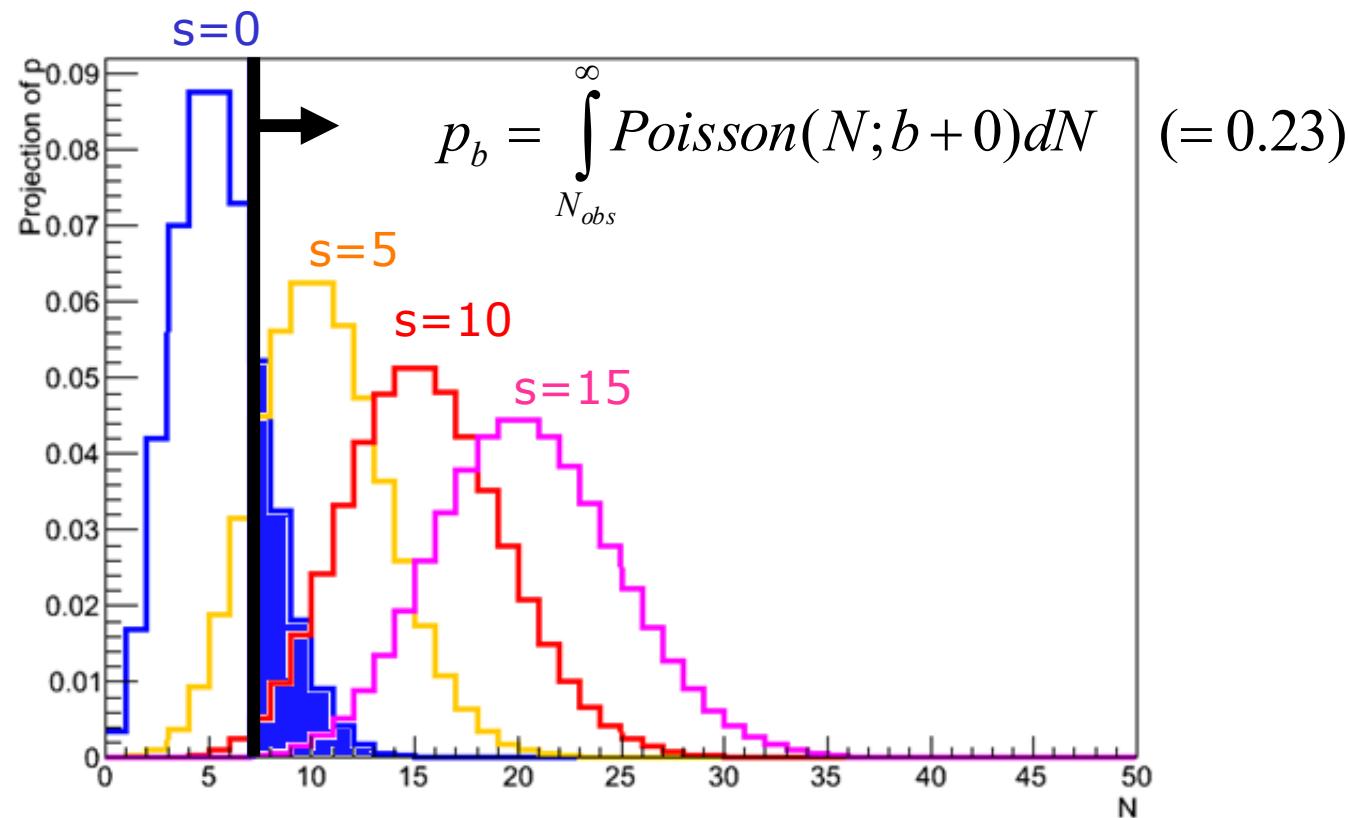


Test statistic $T(N)=N_{\text{obs}}$ orders observed events count by estimated signal yield

Low $N \rightarrow$ low estimated signal
High $N \rightarrow$ large estimated signal

P-values for counting experiments

- Now make a measurement $N=N_{\text{obs}}$ (example $N_{\text{obs}}=7$)
- **Definition: p-value:**
**probability to obtain the observed data, or more extreme
in future repeated identical experiments**
 - Example: p-value for background-only hypothesis



Ordering distributions by ‘signal-likeness’ aka ‘extremity’

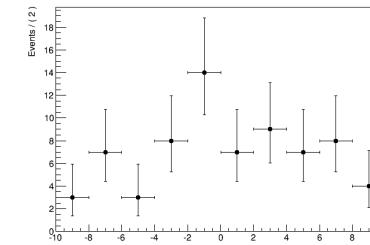
- How to define ‘extremity’ if observed data is a distribution

Observation

Counting

$$N_{\text{obs}} = 7$$

Histogram

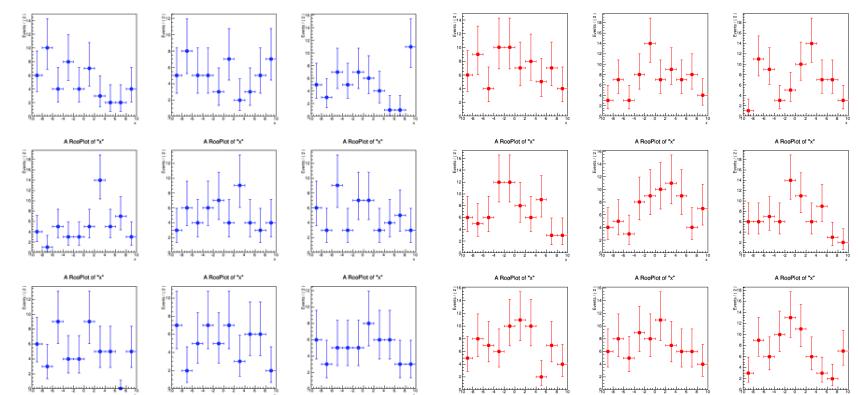
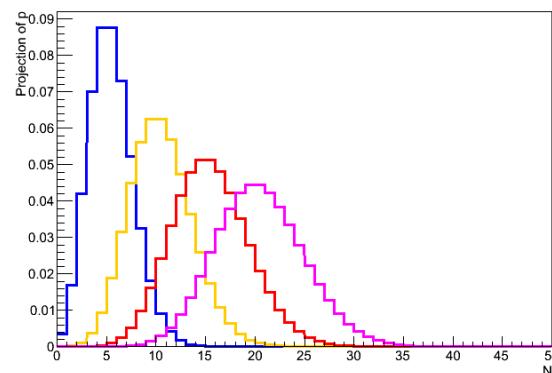


Median expected
by hypothesis

$$N_{\text{exp}}(s=0) = 5$$

$$N_{\text{exp}}(s=5) = 10$$

Predicted distribution
of observables



Which histogram is more ‘extreme’?

The Likelihood Ratio as a test statistic

- Given two hypothesis H_b and H_{s+b} the ratio of likelihoods is a useful test statistic

$$\lambda(\vec{N}) = \frac{L(\vec{N} | H_{s+b})}{L(\vec{N} | H_b)}$$

- Intuitive picture:

→ If data is likely under H_b ,
 $L(N|H_b)$ is **large**,
 $L(N|H_{s+b})$ is smaller

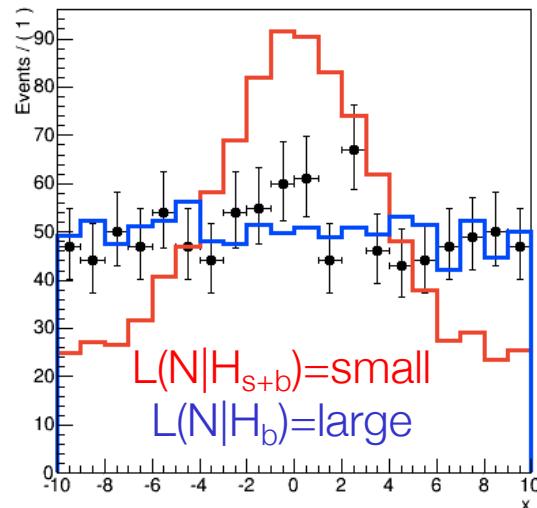
→ If data is likely under H_{s+b}
 $L(N|H_{s+b})$ is **large**,
 $L(N|H_b)$ is smaller

$$\lambda(\vec{N}) = \frac{\text{small}}{\text{large}} = \text{small}$$

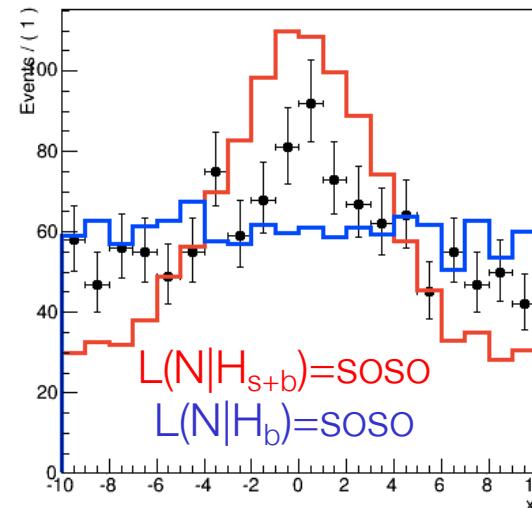
$$\lambda(\vec{N}) = \frac{\text{large}}{\text{small}} = \text{large}$$

Visualizing the Likelihood Ratio as ordering principle

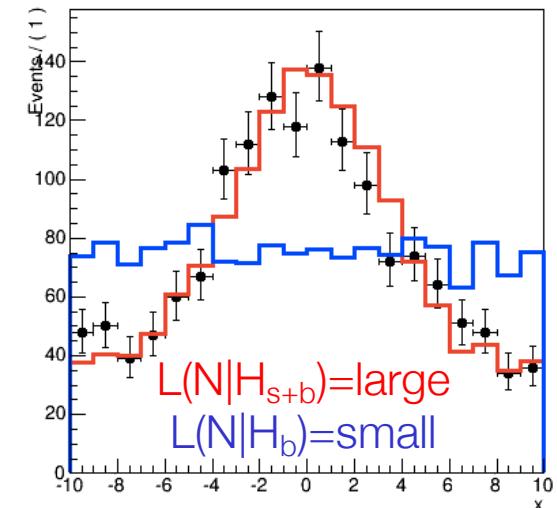
- The Likelihood ratio as ordering principle



$\lambda(N)=0.0005$



$\lambda(N)=0.47$



$\lambda(N)=5000$

- Frequentist solution to ‘relevance of $P(\text{data}|\text{theory})$ ’ is to order all observed data samples using a (Likelihood Ratio) test statistic
 - Probability to observe ‘similar data or more extreme’ then amounts to calculating ‘probability to observe test statistic $\lambda(N)$ as large or larger than the observed test statistic $\lambda(N_{\text{obs}})$

A different Likelihood ratio for composite hypothesis testing

- On *composite hypotheses*, where both null and alternate hypothesis map to values of μ , we can define an alternative likelihood-ratio test statistics that has better properties

$$\lambda(\vec{N}) = \frac{L(\vec{N} | H_0)}{L(\vec{N} | H_1)} \rightarrow \lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N} | \mu)}{L(\vec{N} | \hat{\mu})}$$

'simple hypothesis' 'composite hypothesis'

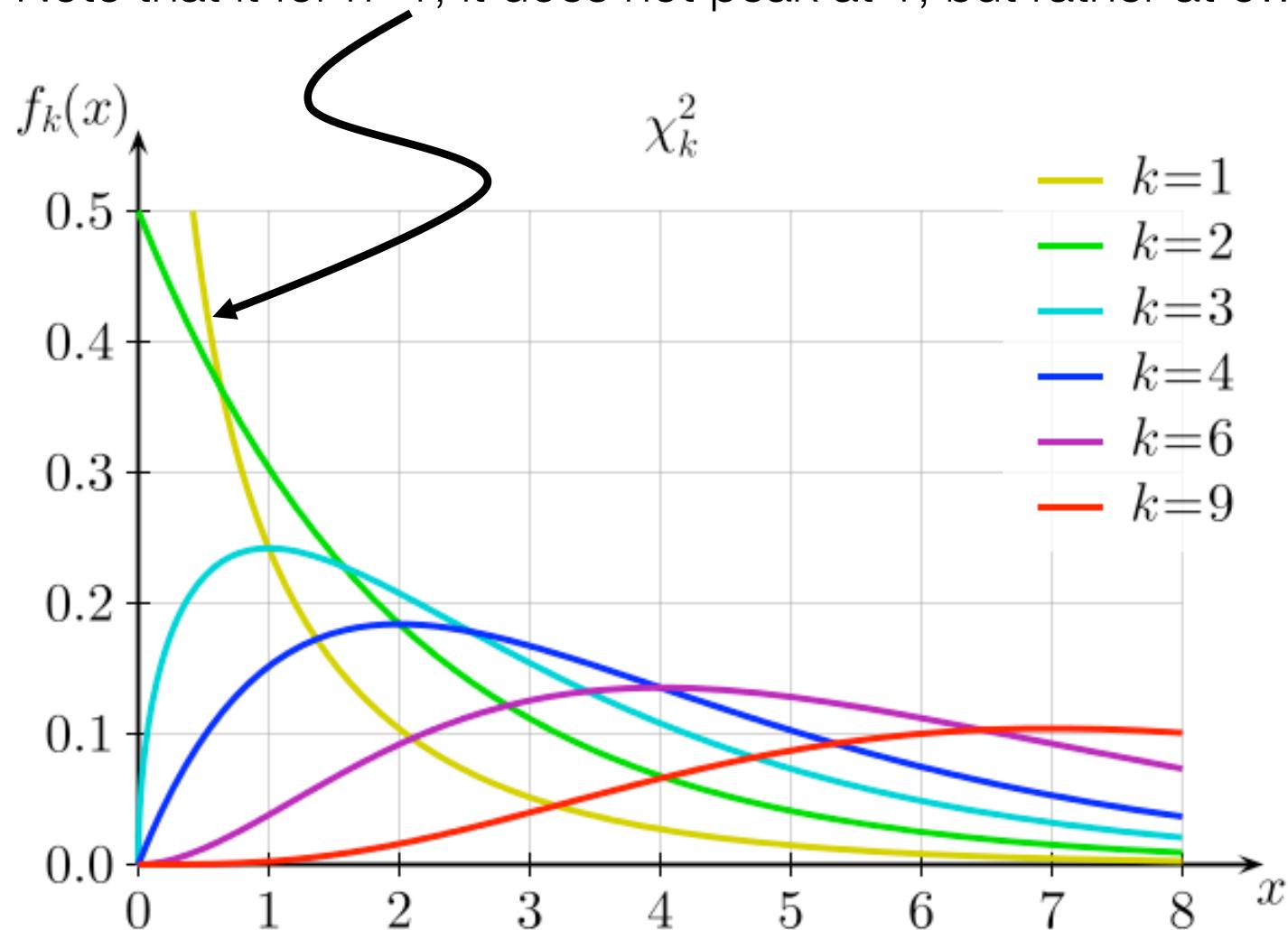
Hypothesis μ that is being tested

'Best-fit value'

- **Advantage: distribution of new λ_μ has known asymptotic form**
 - **Wilks theorem:** distribution of $-\log(\lambda_\mu)$ is asymptotically distribution as a χ^2 with N_{param} degrees of freedom*
- *Some regularity conditions apply
- → Asymptotically, we can *directly* calculate p-value from λ_μ^{obs}

What does a χ^2 distribution look like for n=1?

- Note that for n=1, it does not peak at 1, but rather at 0...



Composite hypothesis testing in the asymptotic regime

- For ‘histogram example’: what is p-value of null-hypothesis

‘likelihood assuming zero signal strength’

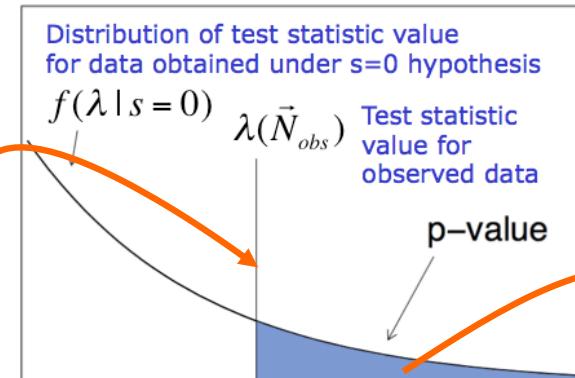
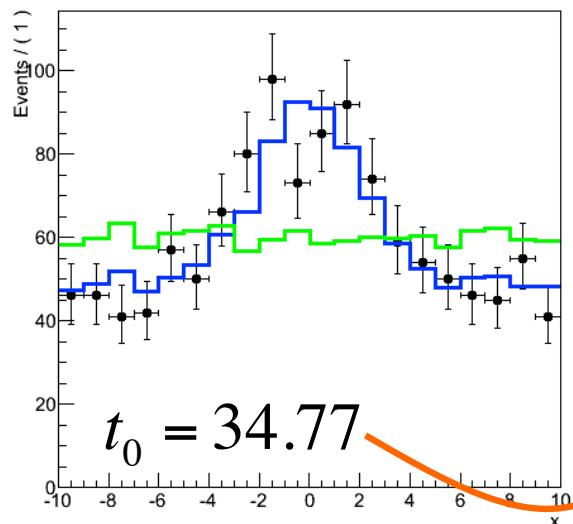
$$t_0 = -2 \ln \frac{L(\text{data} | \mu = 0)}{L(\text{data} | \hat{\mu})}$$

$\hat{\mu}$ is best fit value of μ

‘likelihood of best fit’

$-\log \mu$

On signal-like data t_0 is large



Wilks: $f(\lambda|0) \rightarrow \chi^2$ distribution

$$\begin{aligned} \text{P-value} &= \text{TMath::Prob}(34.77, 1) \\ &= 3.7 \times 10^{-9} \end{aligned}$$

Composite hypothesis testing in the asymptotic regime

- For ‘histogram example’: what is p-value of null-hypothesis

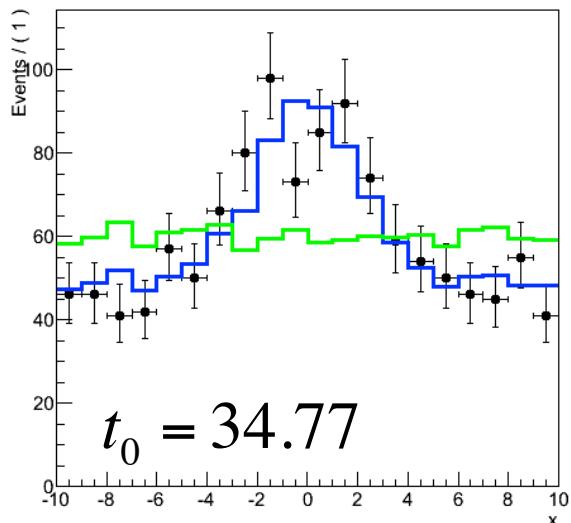
‘likelihood assuming zero signal strength’

$$t_0 = -2 \ln \frac{L(\text{data} | \mu = 0)}{L(\text{data} | \hat{\mu})}$$

$\hat{\mu}$ is best fit
value of μ

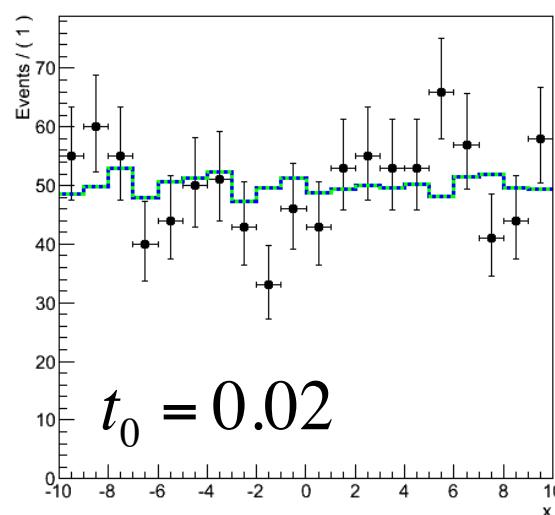
‘likelihood of best fit’

On signal-like data t_0 is large



$$\begin{aligned}\text{P-value} &= \text{TMath::Prob}(34.77, 1) \\ &= 3.7 \times 10^{-9}\end{aligned}$$

On background-like data t_0 is small



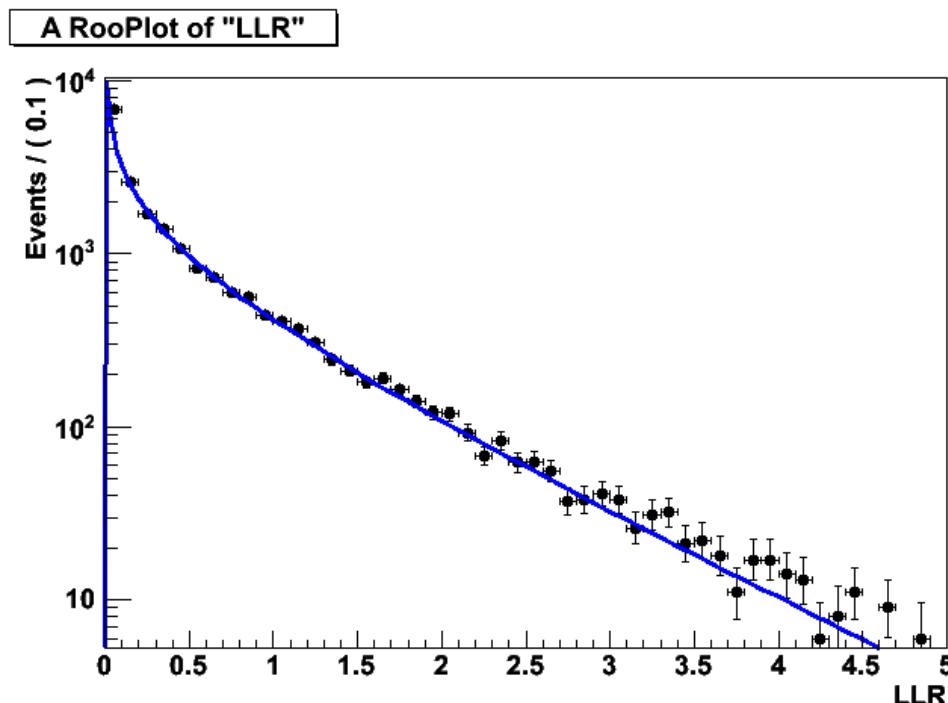
Use
Wilks
Theorem

$$\begin{aligned}\text{P-value} &= \text{TMath::Prob}(0.02, 1) \\ &= 0.88\end{aligned}$$

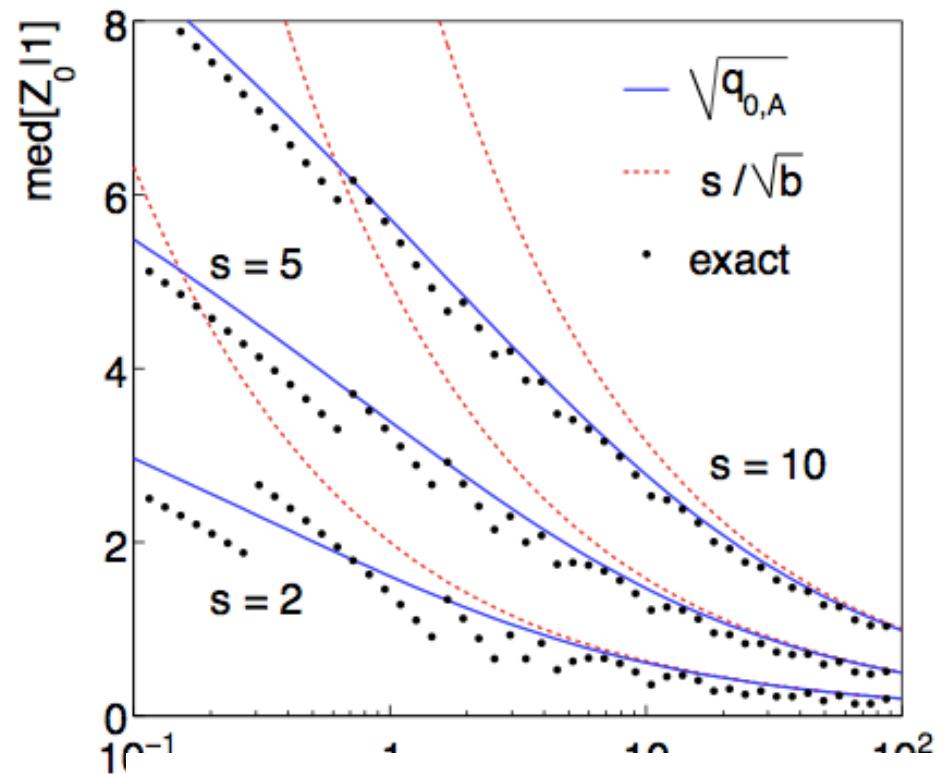
How quickly does $f(\lambda_\mu|\mu)$ converge to its asymptotic form

- Pretty quickly –

Here is an example of likelihood function for 10-bin distribution with 200 events



Here is an example for event counting at various s,b



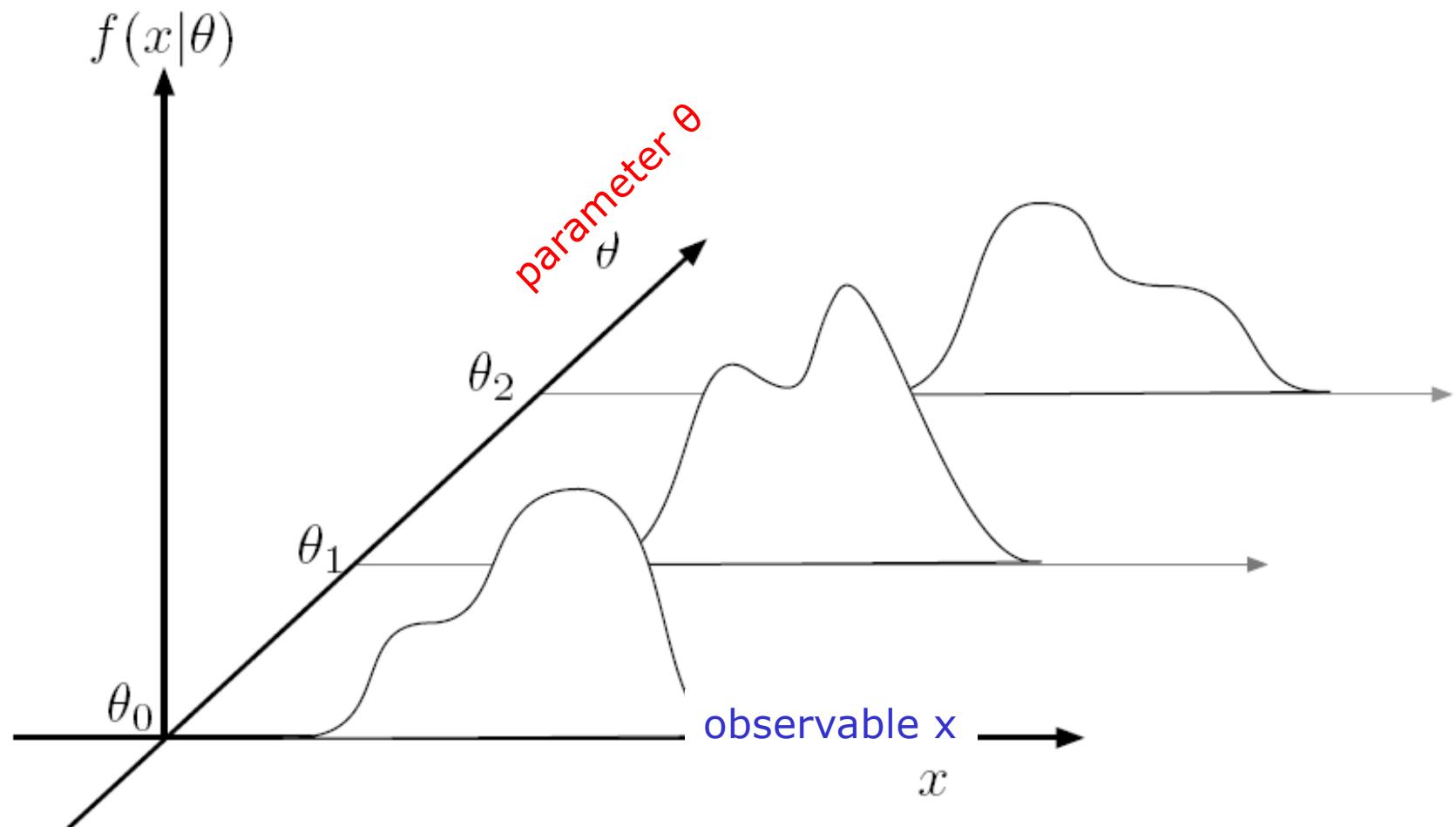
$$\sqrt{q_{0,A}} = \sqrt{2((s+b)\ln(1+s/b) - s)} .$$

From hypothesis testing to confidence intervals

- Next step for composite hypothesis is to go from p-values for a hypothesis defined by fixed value of μ to *an interval statement on μ*
- Definition: A interval on μ at X% confidence level is defined such that the true value of μ is contained X% of the time in the interval.
 - Note that the output is *not* a probabilistic statement on the true s value
 - The true μ is fixed but unknown – each observation will result in an estimated interval $[\mu_-, \mu_+]$. X% of those intervals will contain the true value of μ
 - Coverage = guarantee that probabilistic statements is true (i.e. repeated future experiments do reproduce results in X% of cases)
- Definition of confidence intervals does not make any assumption on shape of interval
 - Can choose one-sided intervals ('limits'), two-sided intervals ('measurements'), or even disjoint intervals ('complicated measurements')

Exact confidence intervals – the Neyman construction

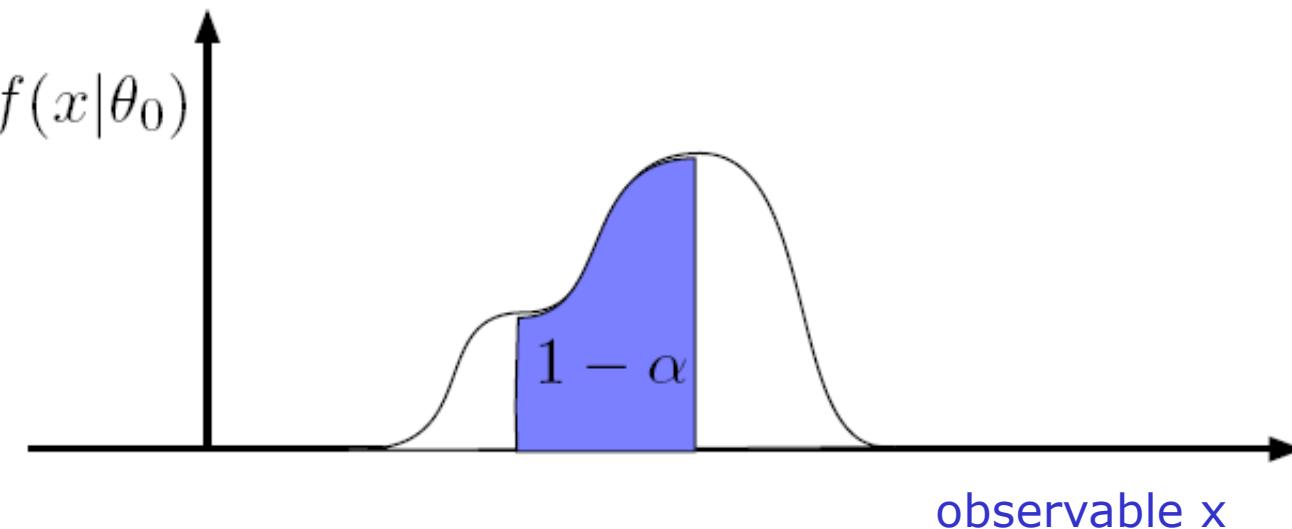
- Simplest experiment: one measurement (x), one theory parameter (θ)
- For each value of **parameter θ** , determine distribution in **observable x**



How to construct a Neyman Confidence Interval

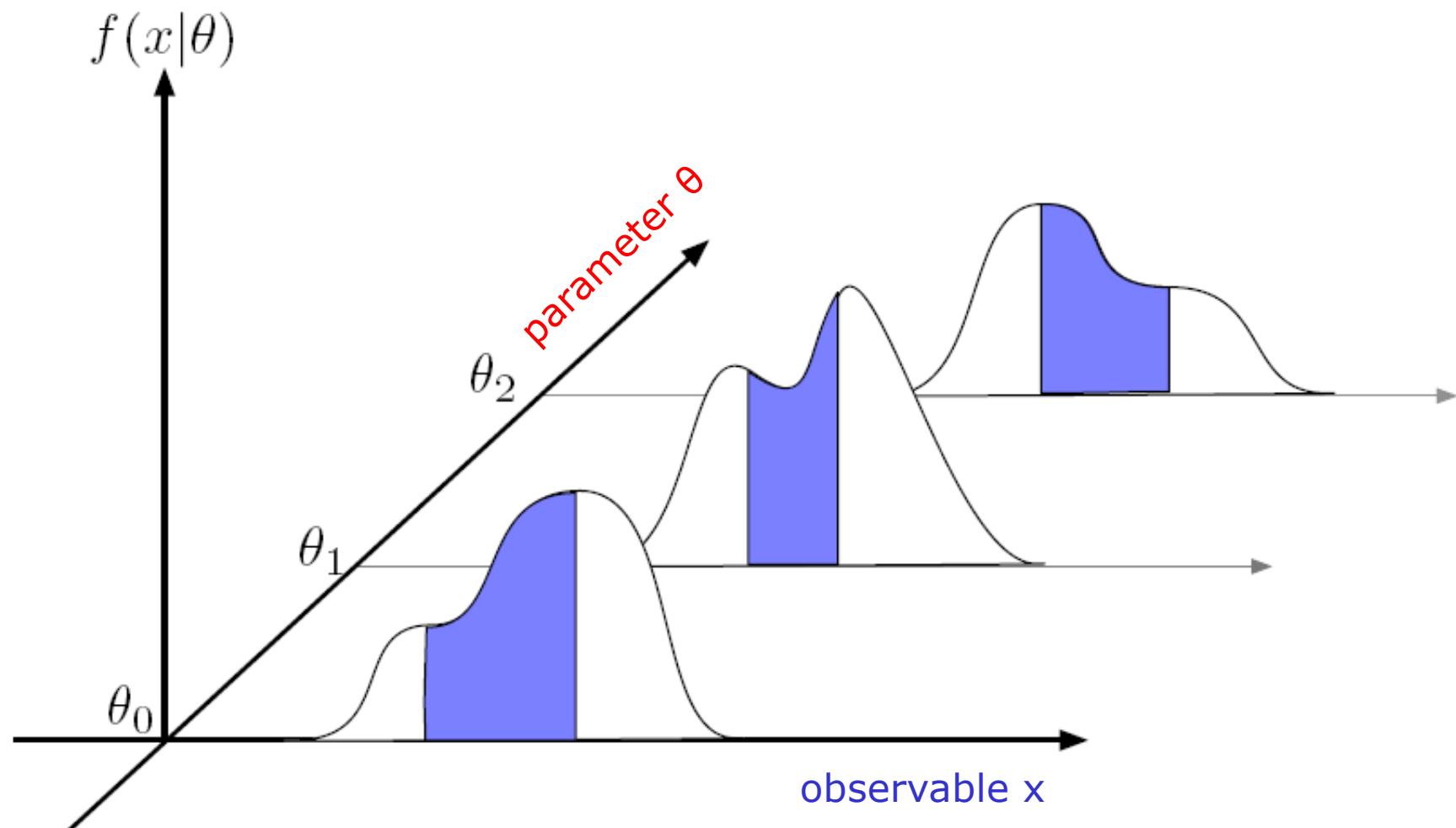
- Focus on a slice in θ
 - For a $1-\alpha\%$ confidence Interval, define *acceptance interval* that contains $100\%- \alpha\%$ of the distribution

pdf for observable x
given a parameter value θ_0



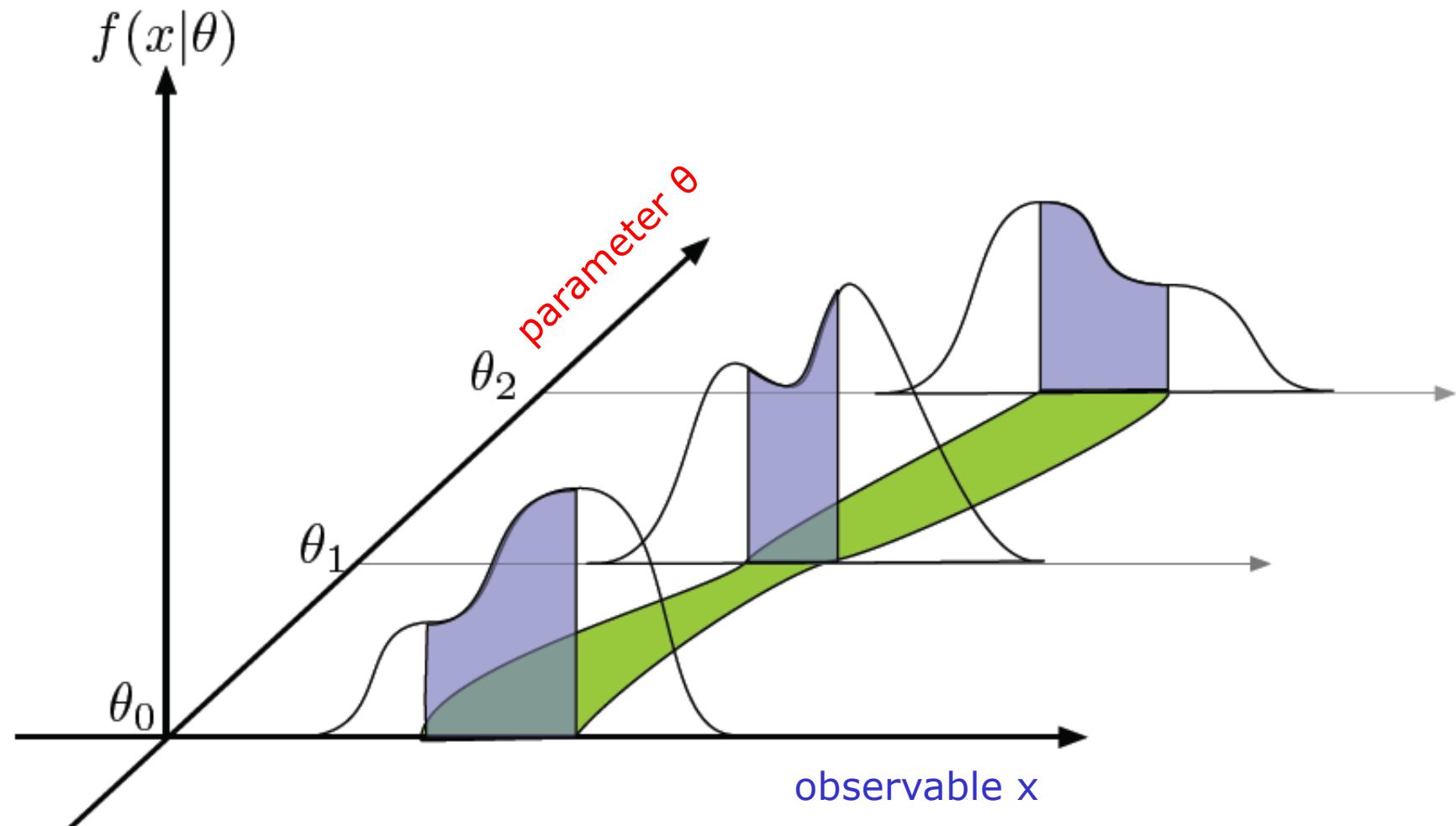
How to construct a Neyman Confidence Interval

- Now make an acceptance interval in observable x for each value of parameter θ



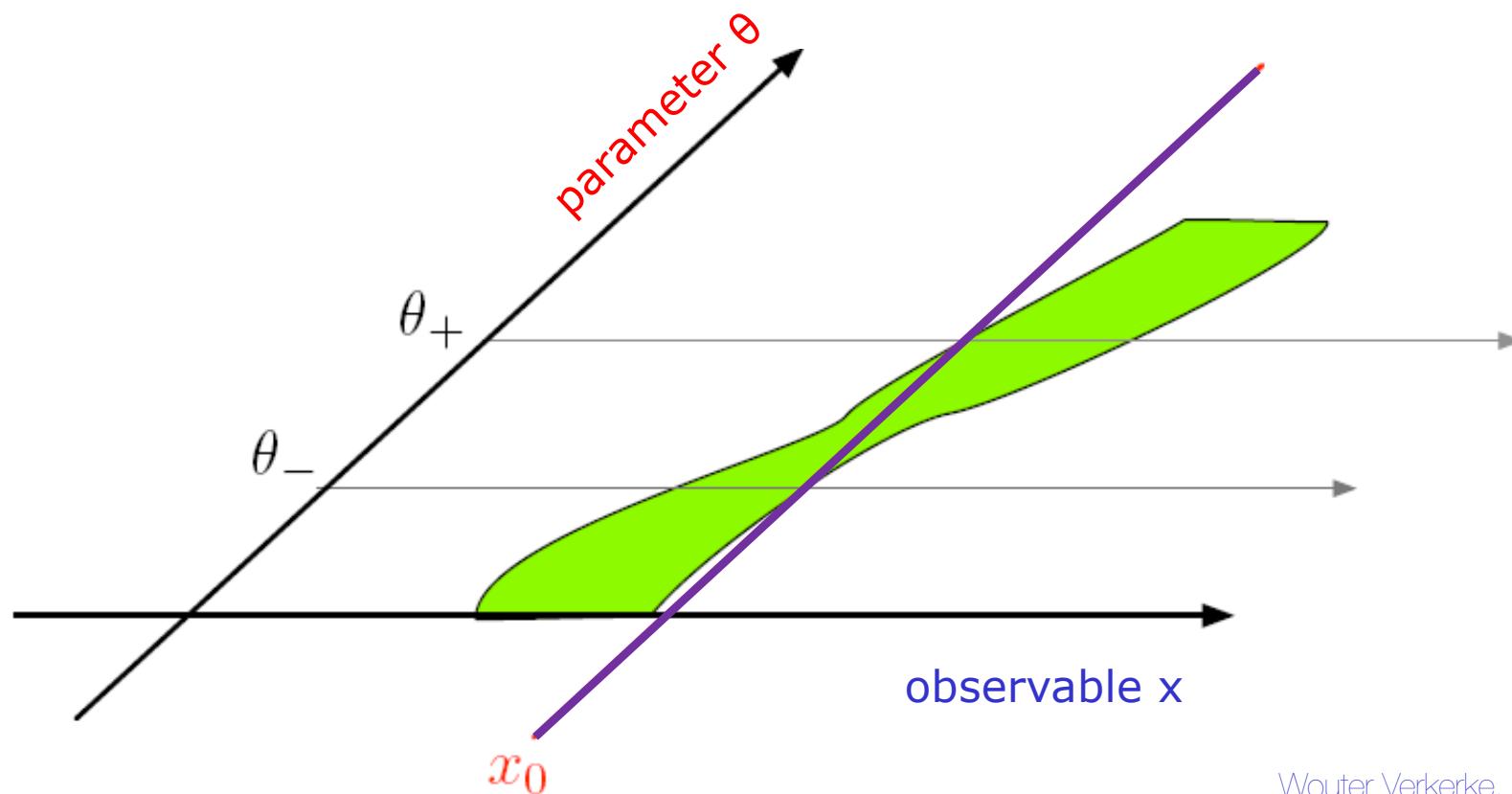
How to construct a Neyman Confidence Interval

- This makes the confidence belt



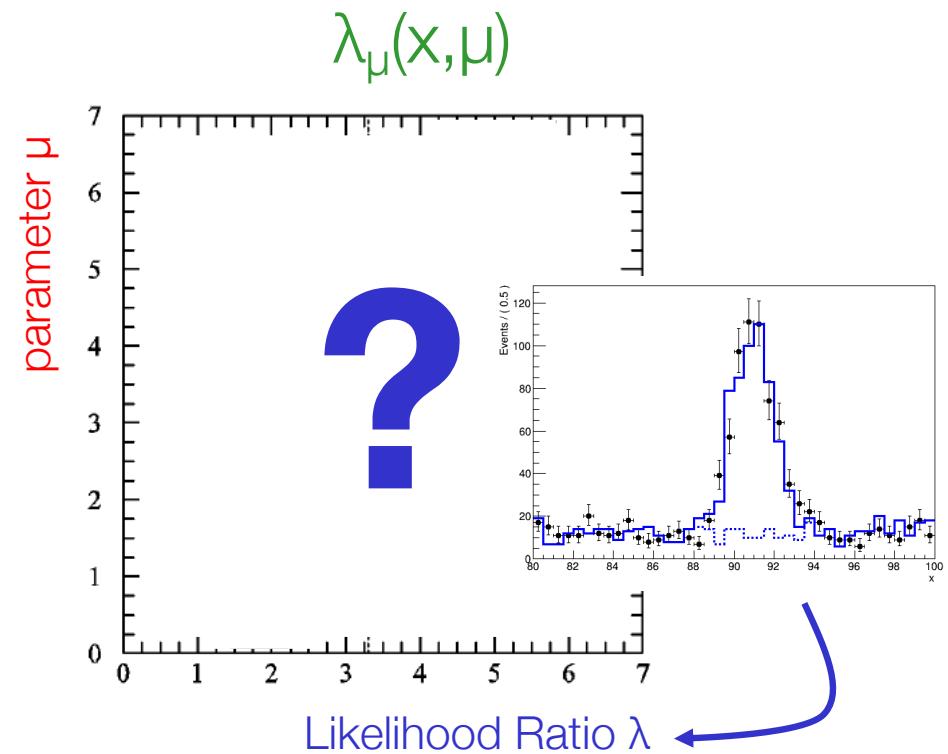
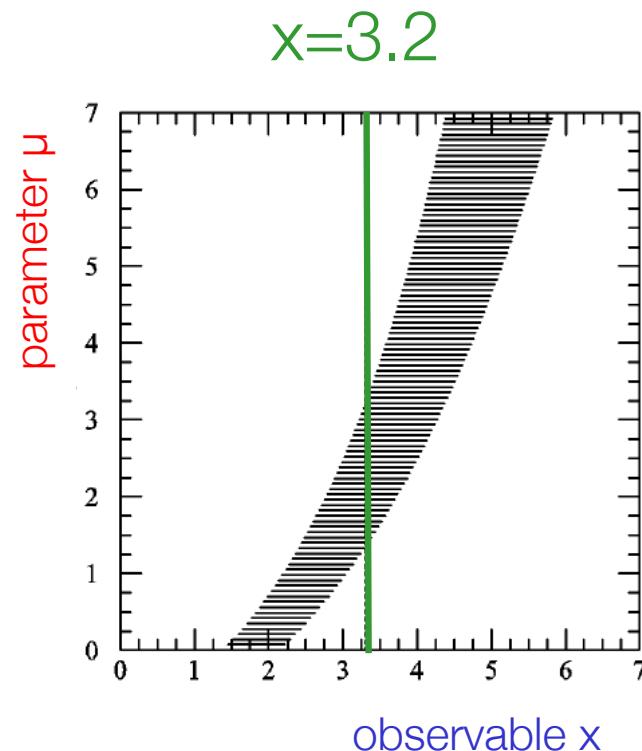
How to construct a Neyman Confidence Interval

- The confidence belt can be constructed *in advance of any measurement*, it is a property of the model, not the data
- Given a measurement x_0 , a confidence interval $[\theta_-, \theta_+]$ can be constructed as follows
- The interval $[\theta_-, \theta_+]$ has a 68% probability to cover the true value



Confidence intervals using the Likelihood Ratio test statistic

- Neyman Construction on Poisson counting looks like ‘textbook’ belt.
- In practice we’ll use the **Likelihood Ratio test statistic** to summarize the measurement of a (multivariate) distribution for the purpose of hypothesis testing.
- Procedure to construct belt with LR is identical:
obtain distribution of λ for every value of μ to construct confidence belt



The asymptotic distribution of the likelihood ratio test statistic

- Given the likelihood ratio

$$t_\mu = -2 \log \lambda_\mu(x) = -2 \log \frac{L(x|\mu)}{L(x|\hat{\mu})}$$

Q: What do we know about asymptotic distribution of $\lambda(\mu)$?

- A: Wilks theorem \rightarrow Asymptotic form of $f(t|\mu)$ is a χ^2 distribution

$$f(t_\mu|\mu) = \chi^2(t_\mu, n)$$

Where

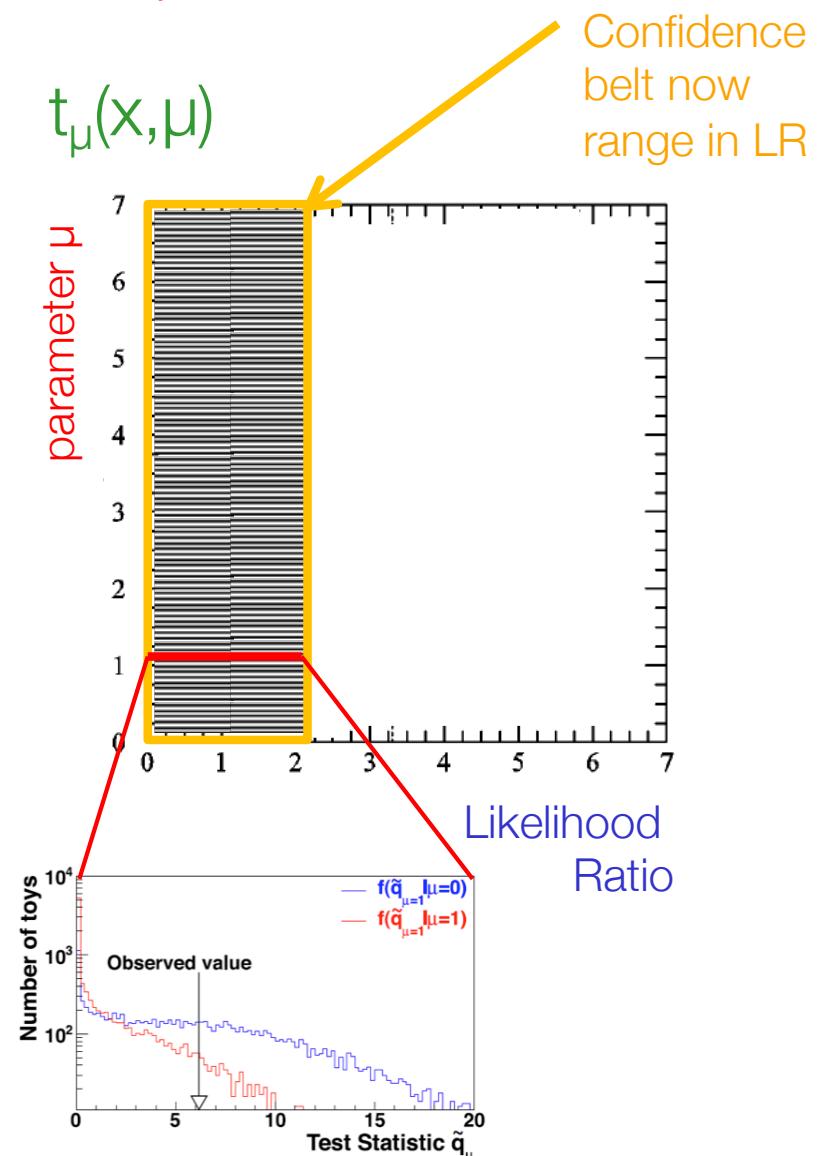
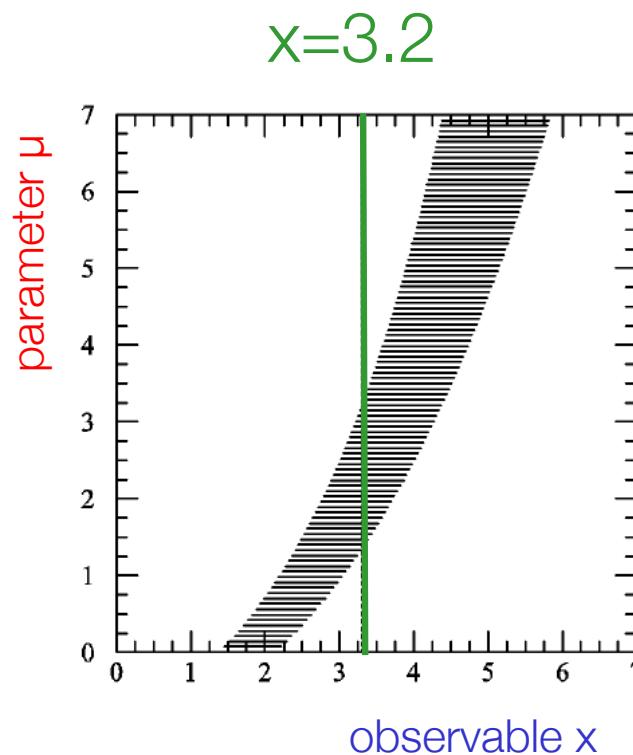
μ is the hypothesis being tested and

n is the number of parameters (here 1: μ)

- Note that $f(t_\mu|\mu)$ is independent of μ !**
 \rightarrow Distribution of t_μ is the same for every ‘horizontal slice’ of the belt

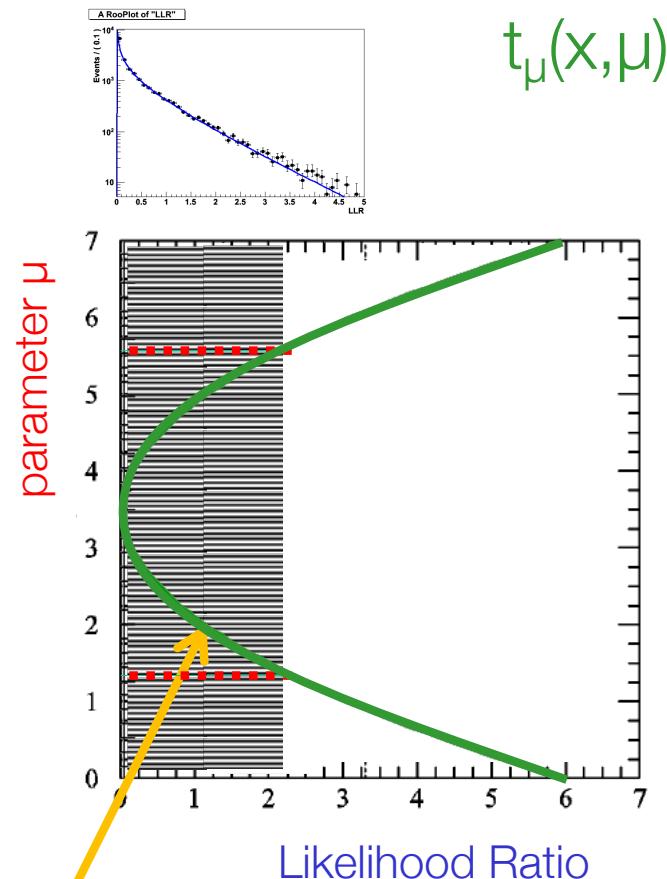
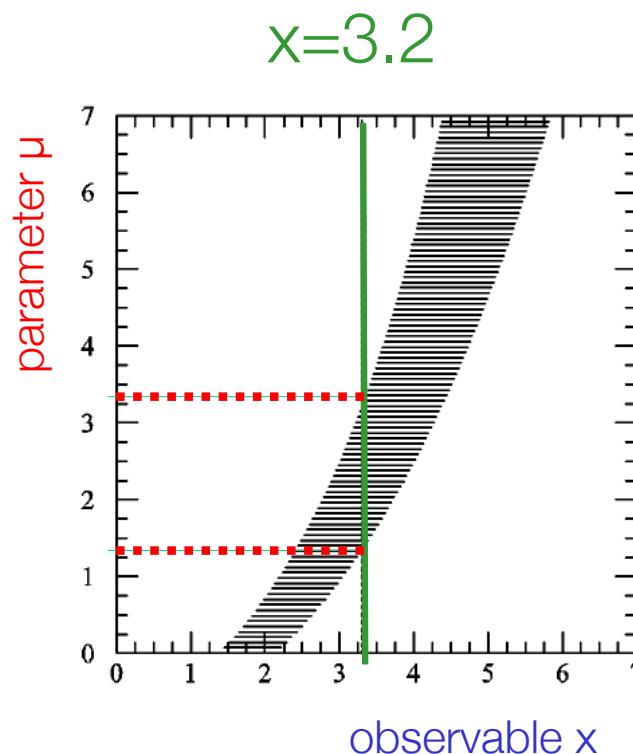
Confidence intervals using the Likelihood Ratio test statistic

- Procedure to construct belt with LR is identical:
obtain distribution of λ for every value of μ to construct belt



What does the observed data look like with a LR?

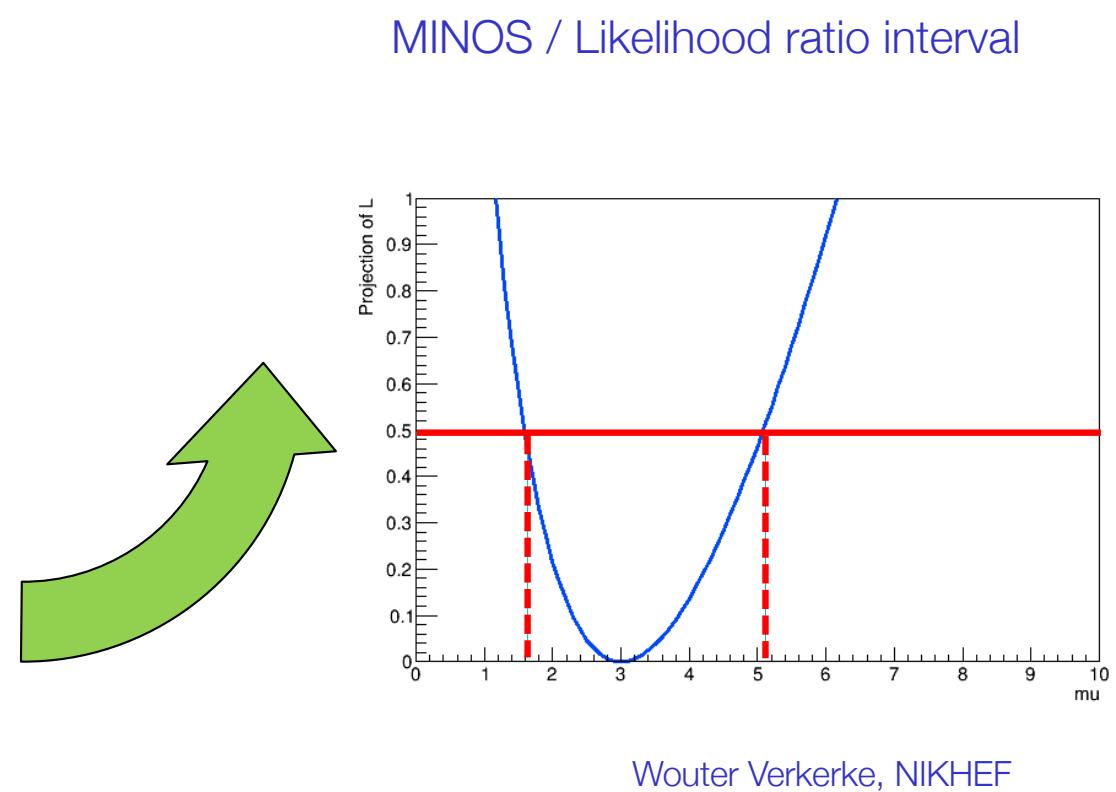
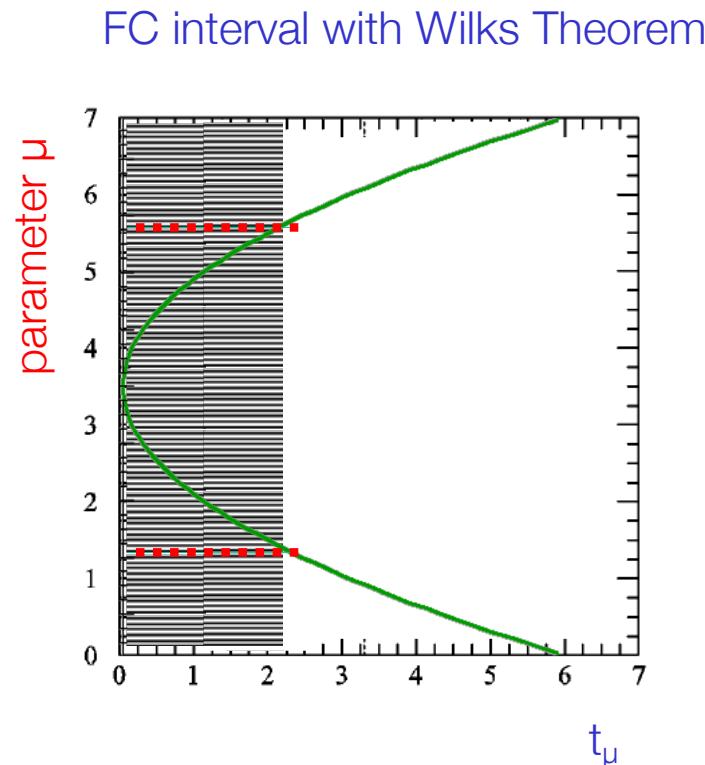
- Note that while belt is (asymptotically) independent of parameter μ , observed quantity now is dependent of the assumed μ



Measurement = $t_\mu(x_{\text{obs}}, \mu)$
is now a function of μ

Connection with likelihood ratio intervals

- If you assume the asymptotic distribution for t_μ ,
 - Then the confidence belt is exactly a box
 - And the constructed confidence interval can be simplified to finding the range in μ where $t_\mu = \frac{1}{2} \cdot Z^2$
- This is exactly the MINOS error



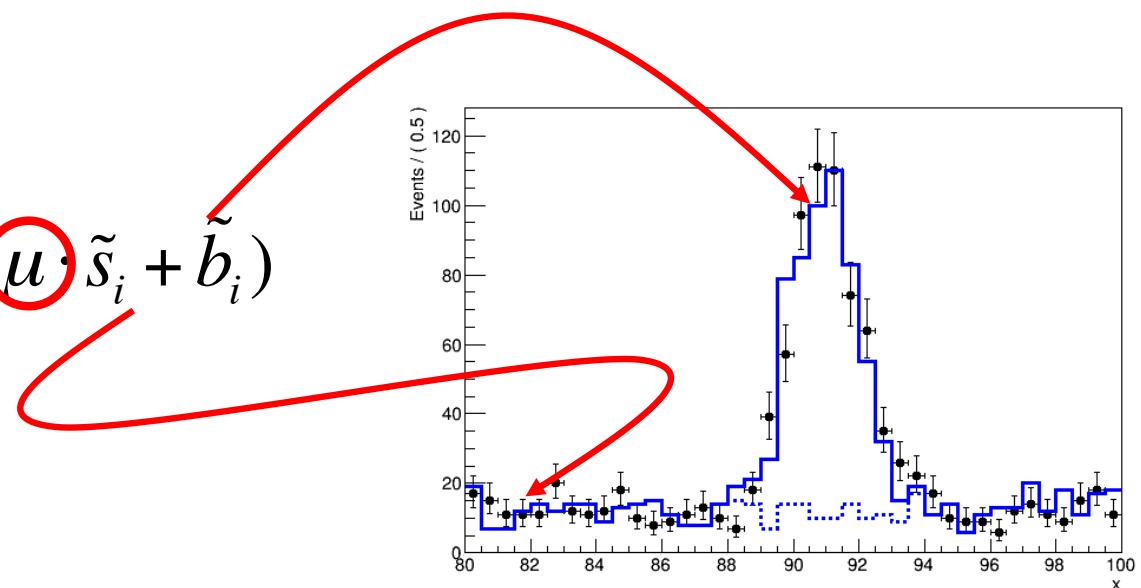
3

Incorporating systematic uncertainties in probability models

So far we've only considered 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

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

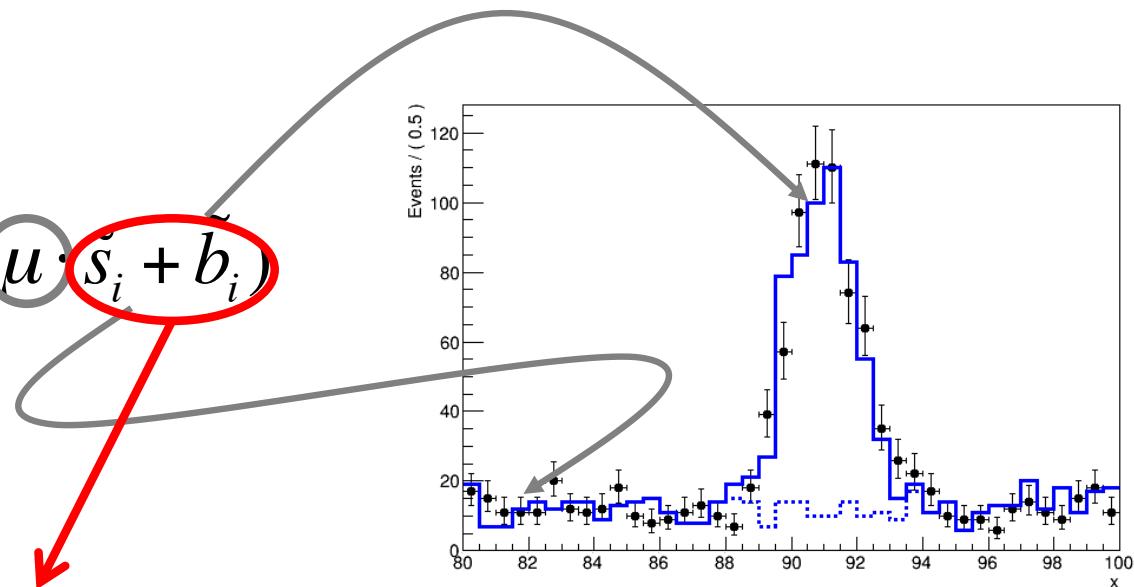


- So far only considered a single parameter in the likelihood: the physics *parameter of interest*, usually denoted as μ

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?

$$L(\vec{N} \mid \mu) = \prod_{\text{bins}} \text{Poisson}(N_i \mid \mu \cdot s_i + b_i)$$



*Signal and background predictions
are affected by (systematic) uncertainties*

Modeling systematic uncertainties in the likelihood

- What is a systematic uncertainty? It consists of
 - 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.
 - 3: A distribution of possible values for the parameters
 - In practice these (response) models are often only formulated implicitly, but modeling of systematic uncertainties in the likelihood requires an explicit model
- Example of ‘typical’ systematic uncertainty prescription

“The Jet Energy Scale Uncertainty is 5%”

- Note that example does not meet definition standards above
 - Specification specifies variance of the distribution unknown parameter, *but not the distribution itself* (is it Gaussian, Poisson, something else)
 - *Response model left unspecified*

Formulating a response model

- Why does the statement

“the JES uncertainty is X%”

not a formulate a response model, while an additional statement

*“If the JES is off by +X%, the energy of every jet
in the event is increased by X%”*

does constitute a response model?

- The first statement doesn't specify any correlation between jets with different kinematics
 - Can low pT jets be miscalibrated by -4% and high pT jets be calibrated by +5%?
 - Or must all jets be miscalibrated by exactly the same amount?
- The former interpretation would require 2 (or more) model parameters to capture the effect of the miscalibration of the simulation, the latter only one.
- Once the response model is defined, the effect of a systematic uncertainty is deterministically described, up to an (a set of) unknown strength parameter(s).

Formulating a response model

- Note that the construction of a response model for a systematic uncertainty is no different from choosing a model to describe your physics of interest
 - You define a probability model that deterministically describes the consequences of the underlying hypothesis, up to set of (*a priori*) unknown model parameter
- Will (for now) assume that for our example measurement the example systematic uncertainty – the Jet Energy Scale – can be correctly described **with a single parameter that coherently moves the calibration of all jets in the event.**
 - The correctness of such an assumption we'll revisit later (but note that this is a *physics* argument)

Modeling the strength parameter

- What do we know about distribution of the corresponding strength parameter?
 - The $\text{sqrt}(\text{variance})$ of the distribution was specified to be 5%
- But a variance does not completely specify a distribution
 - Does the JES measurement follow a Gaussian distribution?
 - Does the JES measurement follow a Poisson distribution?
 - Or, a ‘block-shaped’ distribution, or anything else?
- Not specified by “JES is 5%” prescription
 - Often not a difficult issue as detector-related uncertainties, as these since they are based on (calibration) measurements (and/or central limit theorem applies) → Gaussian or Poisson distribution
 - For theory uncertainties this can be tricky, what distribution to assume for ‘renormalization scale uncertainty’? Will come back to this later

Formalizing systematic uncertainties

- The original systematic uncertainty prescription

“the JES uncertainty is 5%”

- The formalized prescription for use in statistical analysis

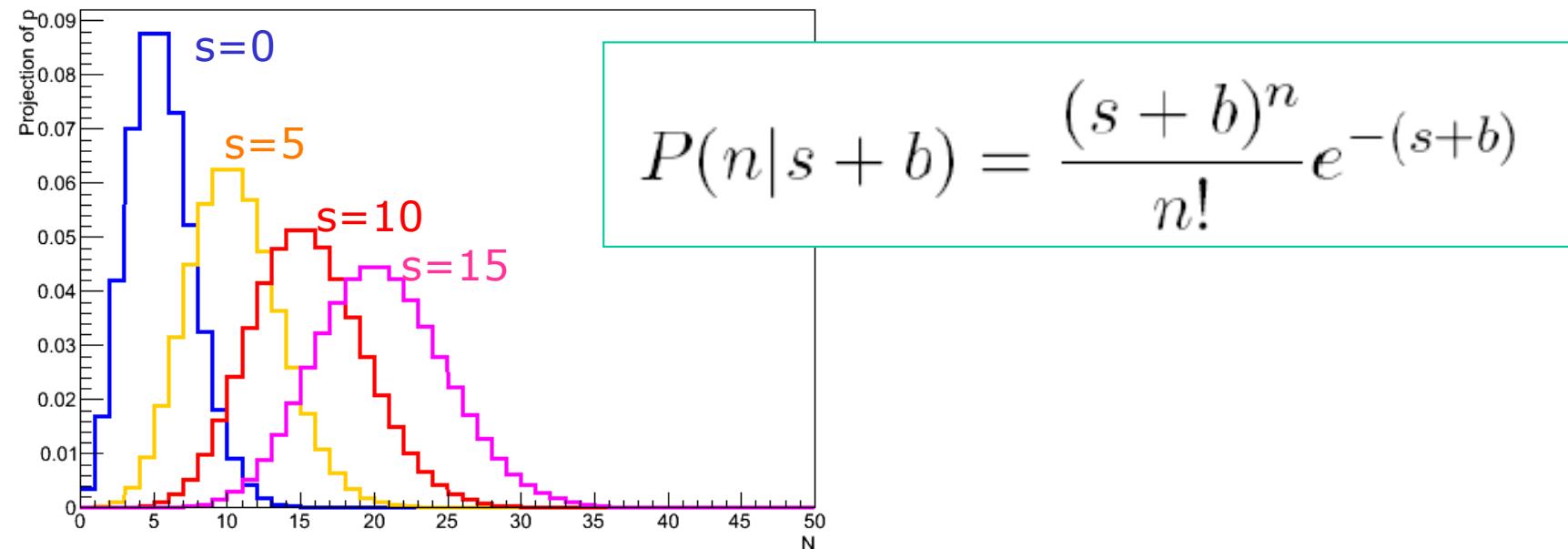
“There is a calibration parameter in the likelihood
of which the true value is unknown

The distribution of this parameter is a Gaussian
distribution with a 5% width

The effect of changing the calibration by 1%
is that energy of all jets in the event is
coherently increased by 1% ”

Introducing uncertainties – a non-systematic example

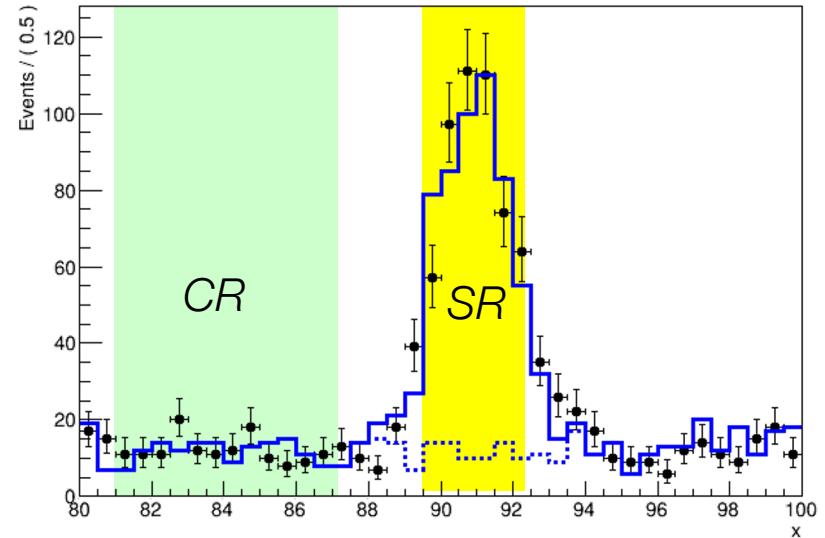
- The original model (with fixed b)



- Now consider b to be uncertain
- $$\mathcal{L}(N|s) \rightarrow \mathcal{L}(N|s,b)$$
- 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

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

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

NB: Define parameter ‘b’ to represents the amount of bkg in the SR.

Scale factor $\tilde{\tau}$ accounts for difference in size between SR and CR

“Background uncertainty constrained from the data”

- Full likelihood of the measurement (‘simultaneous fit’)

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

Generalizing the concept of the sideband measurement

- Background uncertainty from sideband clearly clearly not a ‘systematic uncertainty’

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

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

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

‘Measured background rate by MC simulation’

‘Subsidiary measurement’
of background rate

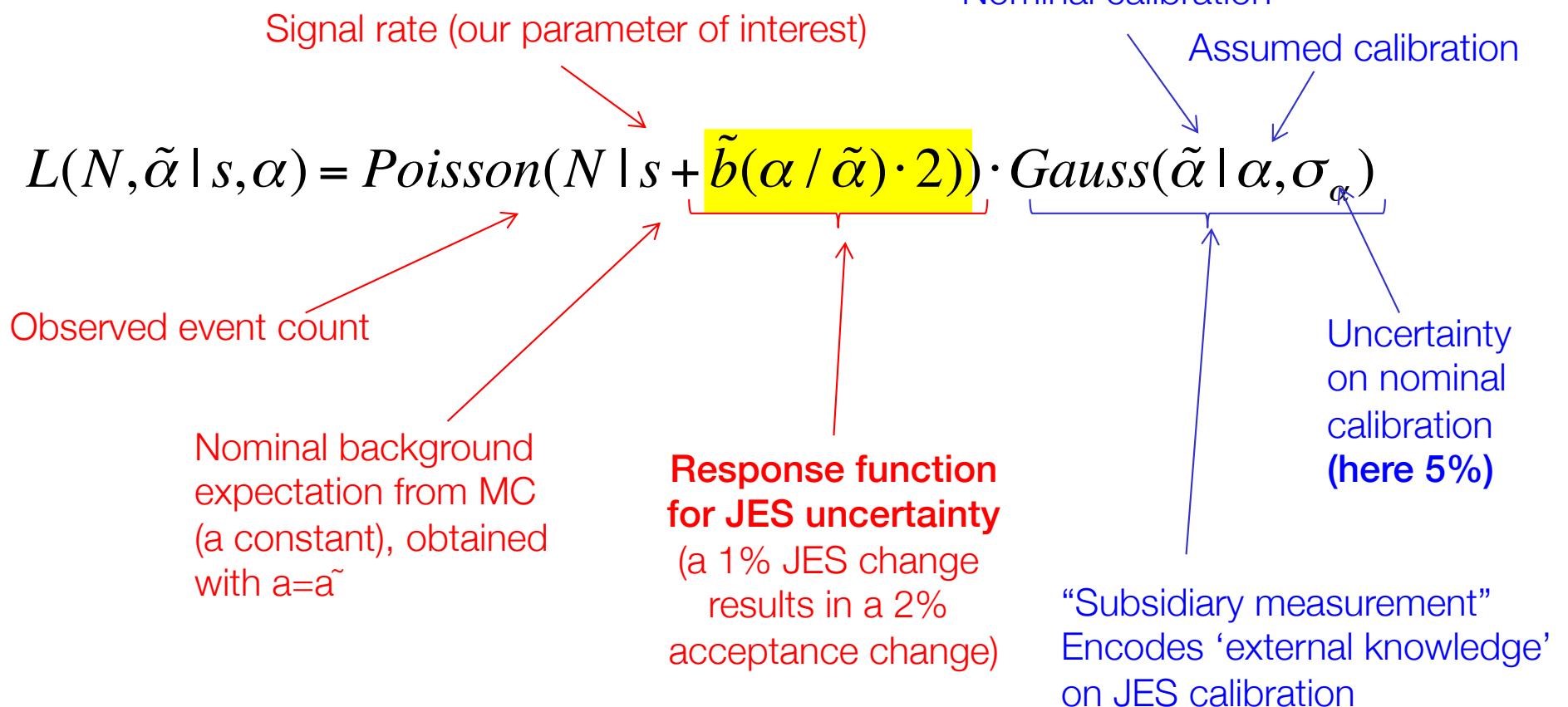
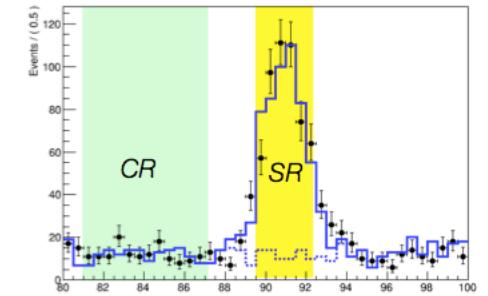
- We can model this in the same way, because the cross-section uncertainty is also (ultimately) the result of a measurement

Generalize: ‘sideband’ → ‘subsidiary measurement’

Modeling a detector calibration uncertainty

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

- 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)



Modeling a detector calibration uncertainty

- Simplify expression by renormalizing “subsidiary measurement”

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

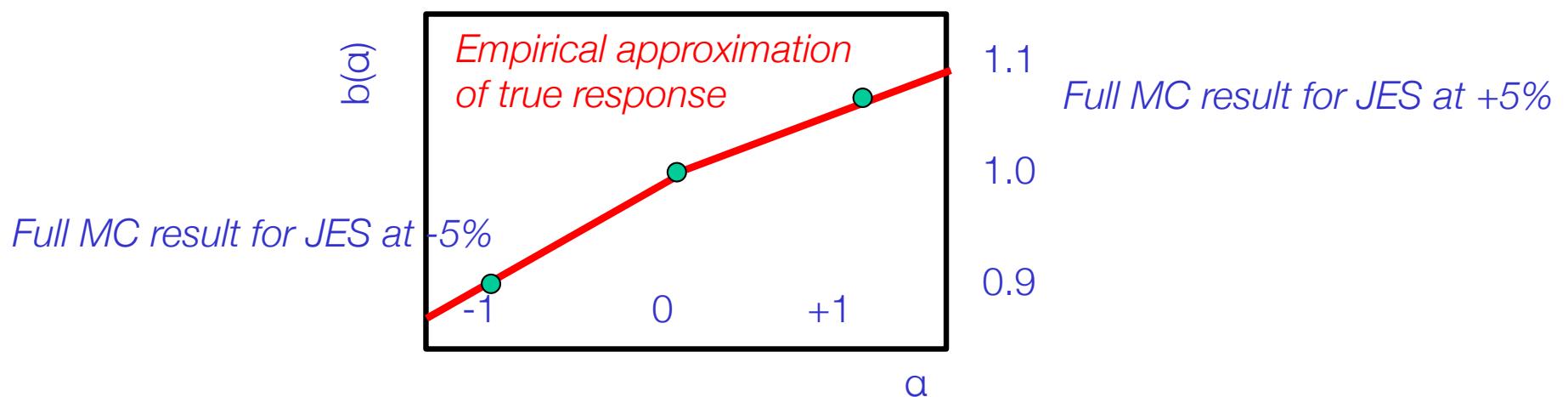
Diagram illustrating the components of the likelihood function:

- Observed event count** (red arrow pointing to the first term s)
- Nominal background expectation from MC (a constant)** (red arrow pointing to the term \tilde{b})
- Signal rate (our parameter of interest)** (red arrow pointing to the term $1 + 0.1\alpha$)
- Response function for normalized JES parameter** (red arrow pointing to the coefficient 0.1α)
[a unit change in α – a 5% JES change – still results in a 10% acceptance change]
- Gauss($\tilde{\alpha}|\alpha, \sigma_\alpha$)** (grey arrow pointing down to the right)
- “Normalized subsidiary measurement”** (blue text)
- The scale of parameter α is now chosen such that values ± 1 corresponds to the nominal uncertainty (in this example 5%)** (blue text)

The response function as empirical model of full simulation

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

- 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 $\pm 1\sigma$ 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 $\pm 5\%$



Names and conventions – ‘profiling’ & ‘constraints’

- The full likelihood function of the form

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

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

The diagram shows two mathematical expressions side-by-side. On the left, there is a crossed-out term $Gauss(\alpha \mid 0, 1)$, indicated by a large red 'X' drawn through it. On the right, there is a term $Gauss(0 \mid \alpha, 1)$. A blue arrow points from the text 'But it is *not* a pdf in the NP' above to the '0' in the expression $Gauss(0 \mid \alpha, 1)$.

Names and conventions

- The ‘subsidiary measurement’ as simplified form of the ‘full calibration measurement’ also illustrates another important point
 - The full likelihood is simply a *joint likelihood of a physics measurement and a calibration measurement* where both terms are treated on equal footing in the statistical procedure
 - In a perfect world, not bound by technical modelling constraints you would use this likelihood

$$L(N, \vec{y} | s, \alpha) = \text{Poisson}(N | s + b(1 + 0.1\alpha)) \cdot L_{JES}(\vec{y} | \alpha, \vec{\theta})$$

where L_{JES} is the full calibration measurement as performed by the Jet calibration group, based on a dataset y , and which may have other parameters θ specific to the calibration measurement.

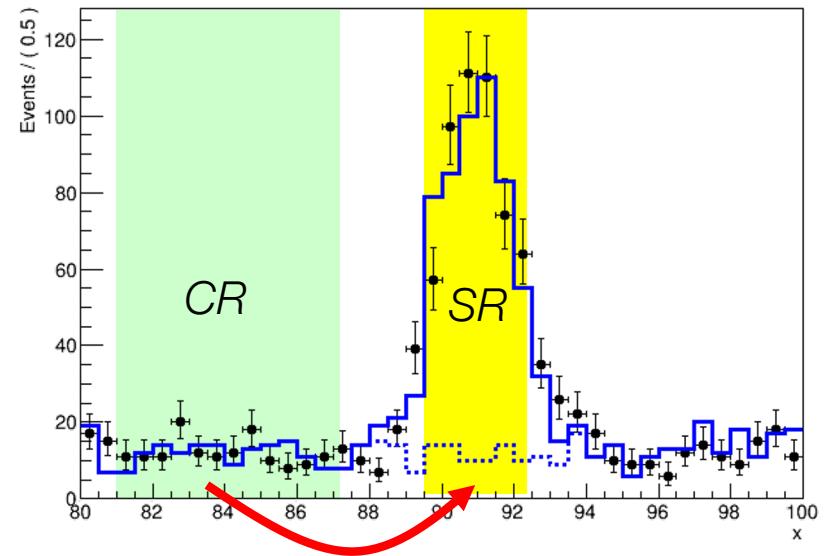
- Since we are bound by technical constraints, we substitute L_{JES} with simplified (Gaussian) form, but the statistical treatment and interpretation remains the same

The sideband measurement *with* a systematic uncertainty

- The extrapolation from the sideband (CR) to the SR always assumes a model (that may carry an uncertainty)
 - The factor τ may depend on theory or detector factors that are uncertainty.
 - One should account for these!

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

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



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

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

$$L(N, N_{ctl}, 0 | s, b, \alpha_{JES}) = \text{Poisson}(N | s + b) \cdot \text{Poisson}(N_{ctl} | \tau(1 + X\alpha_{JES}) \cdot b) \cdot \text{Gauss}(0 | \alpha_{JES}, 1)$$

JES response model for ratio b_{SR}/b_{CR}

Subsidiary measurement of JES response parameter

MC statistical uncertainties as systematic uncertainty

- In original JES uncertainty example, the MC statistical uncertainty was ignored (since 100Mevt were available)
- What should you do if MC statistical uncertainties cannot be ignored?
- Follow same procedure again as before:
 - Define response function (this is trivial for MC statistics: it is the luminosity ratio of the MC sample and the data sample)
 - Define distribution for the ‘subsidiary measurement’ – This is a Poisson distribution – since MC simulation is also a Poisson process
 - Construct full likelihood (‘profile likelihood’)

$$L(N, N_{MC} | s, b) = \text{Poisson}(N | s + b) \cdot \text{Poisson}(N_{MC} | \tau \cdot b)$$

↑
Constant factor $\tau = L(\text{MC})/L(\text{data})$

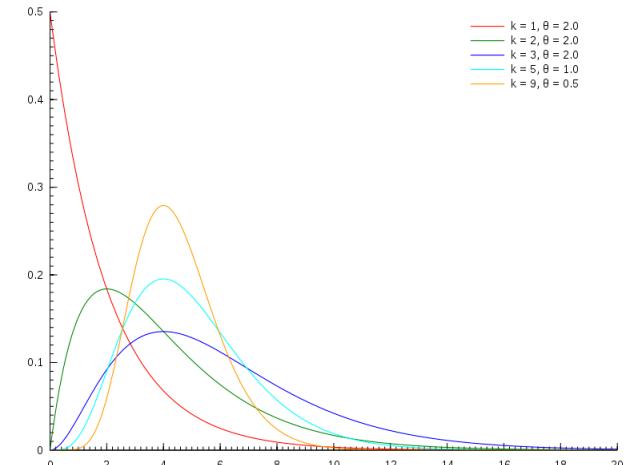
- Note uncanny similarity to full likelihood of a sideband measurement!

$$L(N, N_{ctl} | s, b) = \text{Poisson}(N | s + b) \cdot \text{Poisson}(N_{ctl} | \tau \cdot b)$$

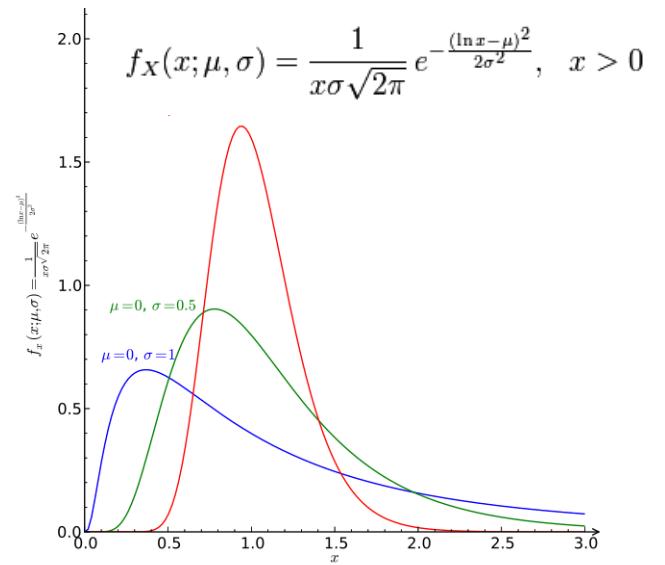
Overview of common subsidiary measurement shapes

- Gaussian $G(x|\mu,\sigma)$
 - ‘Default’, motivated by Central Limit Theorem
(asymp dist for sum of random variables)
- (Rescaled) Poisson $P(N|\mu\tau)$ →

- Obvious choice for any subsidiary measurement that is effectively a counting experiment
- NB: For a Poisson model the distribution in μ is a Gamma distribution (posterior of Poisson)
- Scale factor τ allows to choose variance independently of mean (e.g. to account for side-band size ratio, data/mc lumi ratio)



- LogNormal $LN(x|\mu,\sigma)$ →
- Asymptotic distribution for product of random variables
- Appealing property for many applications is that it naturally truncates at $x=0$



Specific issues with theory uncertainties

- Modeling of **theoretical** syst. uncertainties follows familiar pattern
 - Define response
 - Define distribution for the ‘subsidiary measurement’
 - Construct full likelihood
- But **distribution of subsidiary theory measurement** can be a thorny issue
 - For detector simulation uncertainties, subsidiary measurement usually based on actual measurement → Central Limit Theorem → convergence to Gaussian distribution when measurement is based on many events
 - This argument does not always apply to theoretical uncertainties, as there may be no underlying measurement
- Example: (N)LO scale uncertainties in Matrix Element calculations
 - Typical prescription “vary to 0.5x nominal and 2x nominal and consider the difference” makes no statement on distribution
 - Yet proper statistical treatment of such an uncertainty (i.e. modeling in the likelihood) demands a specified distribution
 - Not clear what to do. You can ask theory expert, but not clear if has a well-motivated choice of distribution...
 - In any case if choice of distribution turns out not to matter too much, you just pick one.

Specific issue with 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 CTEQ and MRST parton density functions and take the difference as systematic uncertainty.
 - Evaluate measurement with Herwig and Pythia showering Monte Carlos and take the difference as systematic uncertainty
- I call these ‘2-point systematics’.
 - You have the technical means to evaluate two known different configurations, but reasons for underlying difference are not clearly identified.

Specific issue with theory uncertainties

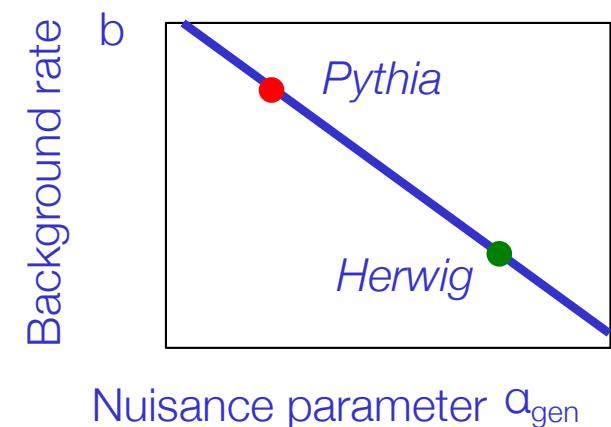
- It is difficult to define rigorous statistical procedures to deal with such 2-point uncertainties. So you need to decide
- If their estimated effect is small, you can pragmatically ignore these lack of proper knowledge and ‘just do something reasonable’ to model these effects in a likelihood
- If their 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.
 - Hypothetical example:
 - * You measure an inclusive cross-section.
 - * But Pythia-Herwig effect is largest uncertainty, originates from the visible-to-inclusive acceptance factor.
 - * Does it make sense to publish the inclusive cross-section, or is it better to publish visible cross-section in some well-defined fiducial range?
 - * Your measurement can then contribute to further discussion and validation of various showering MC packages.

Specific issues with theory uncertainties

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

$$L(N|s,\alpha) = \text{Poisson}(N|s + b \cdot f(\alpha)) \cdot \text{SomePdf}(0|\alpha)$$

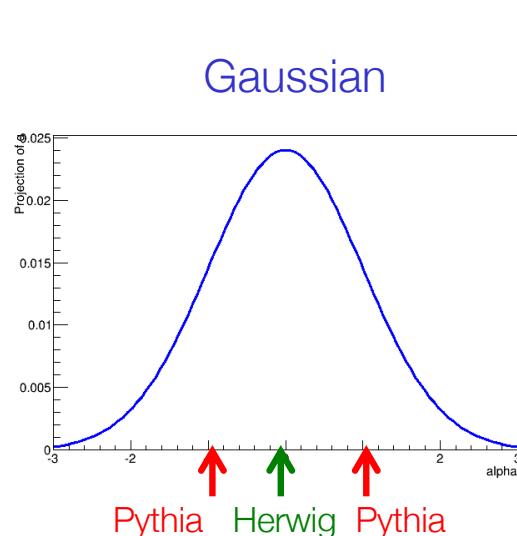
- Since underlying concept of systematic uncertainty not defined, the only option is to *define its meaning terms in terms of response in the physics measurement*
 - Straightforward for a counting measurement, much more challenging for a distribution
- Example
 - Estimate of bkg with Herwig = 8, with Pythia = 12
 - In the likelihood choose $b=8$ and then define $f(a) = |1+4*a|$, so that $f(0)$ results in ‘Herwig ($b.f=8$)’ and $f(\pm 1)$ results in ‘Pythia ($b.f=12$)’
 - For lack of a better word you could call a now the ‘Herwigness of fragmentation w.r.t its effect on my background estimate’
- A thorny question remains: What is the subsidiary measurement for a ?
 - This should reflect your current knowledge on a .



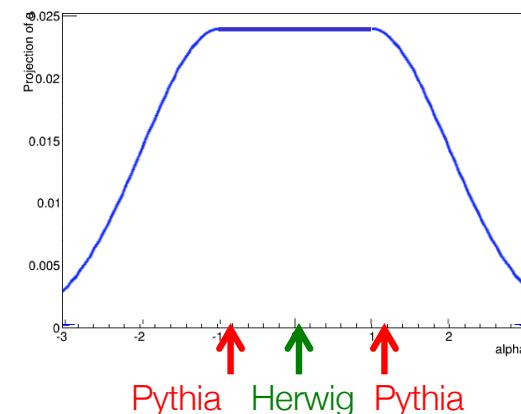
Specific issues with theory uncertainties

- Subsidiary measurement of a theoretical 2-point uncertainty effectively quantifies your ‘prior belief’ in models
 - Formally staying in concepts of frequentist statistics here: likelihood of subsidiary measurement $L(x|a)$ is strictly $P(\text{data}|\text{theory})$, but you ‘data’ here is not really data but something that quantifies your belief since you have no data on this problem.
 - I realize this sounds very much like “you have no idea what you’re doing”, but to some extent this is precisely the problem with 2-point systematics – you really don’t know (or decided not to care about) the underlying physics issues.

- Some options and their effects

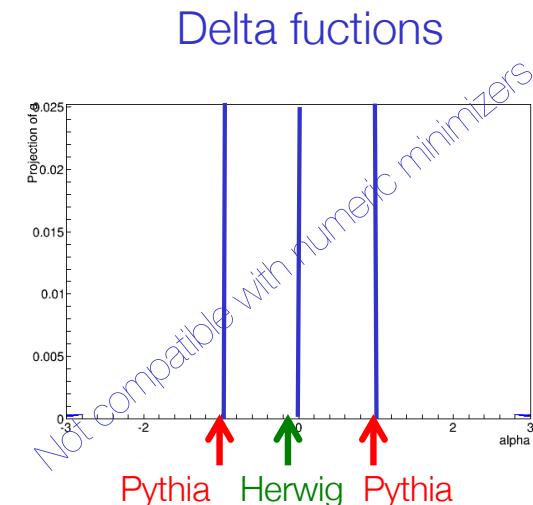


Box with
Gaussian wings



Prefers Herwig at 1σ

All predictions ‘between’
Herwig and Pythia equally
probable



Only ‘pure’ Herwig
and Pythia exist

Modeling multiple systematic uncertainties

- Introduction of multiple systematic uncertainties presents no special issues
- Example JES uncertainty plus generator ISR uncertainty

$$L(N, 0 \mid s, \alpha_{JES}, \alpha_{ISR}) = P(N \mid s + b(1 + 0.1\alpha_{JES} + 0.05\alpha_{ISR})) \cdot G(0 \mid \alpha_{JES}, 1) \cdot G(0 \mid \alpha_{ISR}, 1)$$

Joint response function
for both systematics One subsidiary
measurement for each
source of uncertainty

- A brief note on correlations
 - Word “correlations” often used sloppily – proper way is to think of correlations of parameter estimators. Likelihood defines parameters α_{JES} , α_{ISR} .
The (ML) estimates of these are denoted $\hat{\alpha}_{JES}, \hat{\alpha}_{ISR}$
 - The ML estimators of $\hat{\alpha}_{JES}, \hat{\alpha}_{ISR}$ using the Likelihood of the subsidiary measurements are uncorrelated (since the product factorize in this example)
 - The ML estimators of $\hat{\alpha}_{JES}, \hat{\alpha}_{ISR}$ using the full Likelihood may be correlated.
This is due to physics modeling effects encoded in the joint response function

Modeling systematic uncertainties in multiple channels

- Systematic effects that affect multiple measurements should be modeled coherently.
 - Example – Likelihood of two Poisson counting measurements

$$L(N_A, N_B | s, \alpha_{JES}) = P(N_A | s \cdot f_A + b_A(1 + 0.1\alpha_{JES})) \cdot P(N_B | s \cdot f_B + b_B(1 - 0.3\alpha_{JES})) \cdot G(0 | \alpha_{JES}, 1) \cdot$$

JES response function for channel A JES response function for channel B JES subsidiary measurement

- Effect of changing JES parameter α_{JES} coherently affects both measurement.
- Magnitude and sign effect does not need to be same, this is dictated by the physics of the measurement

Summary on likelihood modeling of systematic uncertainties

- To describe a systematic uncertainty in a likelihood model you need
 - A response model that deterministically describes the effect underlying the uncertainty (e.g. a change in calibration). Such a model has one or more parameters that control the strength of the effect
 - The ‘external knowledge’ on the strength of the effect is modeled as Likelihood representing the ‘subsidiary measurement’ through which this knowledge was obtained
 - Conceptually this is identical to including the likelihood of the actual calibration measurement in the likelihood of the physics analysis
 - In practice a simplified form of the measurement is included, but you must choose an explicit distribution that best represents the original measurement. For systematic uncertainties that relate to external measurements (calibrations), this is often a Gaussian or Poisson distribution
- Modeling prescription can easily be repeated to extend describe effect of multiple uncertainties in multiple simultaneous measurement
 - Conceptually it is not more complicated, but technically it can get tedious. We have good tools for this → will discuss these later

Summary on likelihood modeling of systematic uncertainties

- Often the process of modeling uncertainties in the likelihood requires **information that is traditionally not provided** as part of a systematic uncertainty prescription
- **This is good thing** – your evaluation of these uncertainties otherwise relies on tacit assumptions on these. **Discuss modeling assumptions you make with the prescription ‘provider’**
- You may also learn that your measurement is strongly affect by something you don't know (e.g. distribution of a theory uncertainty). **This is also a good thing**. This is a genuine physics problem, that you might have otherwise overlooked
- Theory uncertainty modeling can pose difficult questions
 - Usually discovered 3 days before approval deadline, tendency is to ‘be conservative’ and not think much about problem. ‘Conservative’ solution tend to be ‘naïve error propagation’ → problem gets hidden behind unspecified assumptions of that method.

4

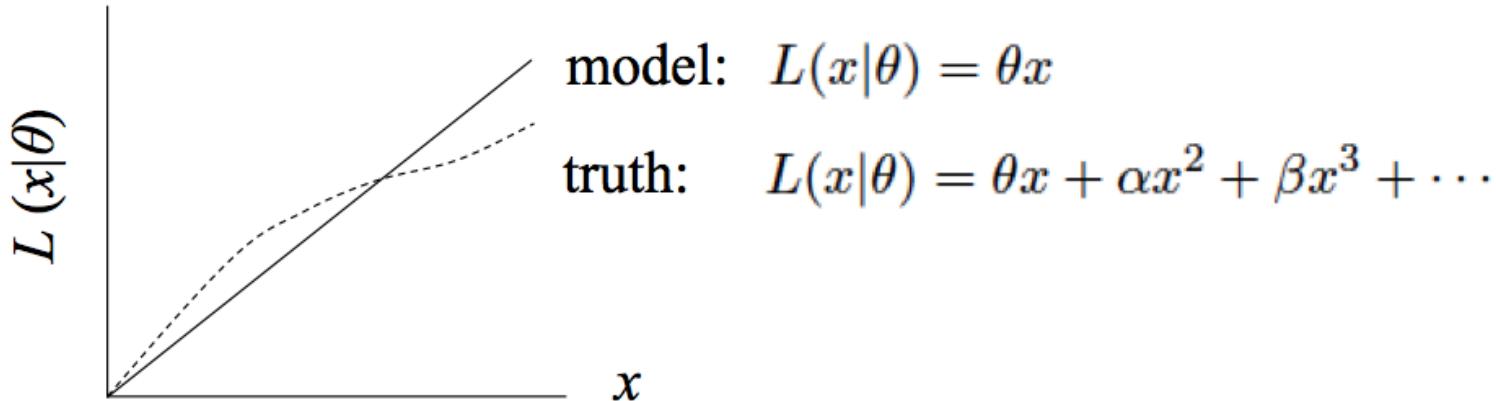
Dealing with nuisance parameters in statistical inference

Dealing with nuisance parameters

- Modeling of systematic uncertainties has introduced many extra parameters in likelihood model that we're not really interested in
 - How do we deal with these in statistical inference?
- Semantic definition
 - **Parameter(s) of Interest** – The (physics) parameter you are interested in. This result goes in your paper. Usually there is one, but sometimes more
 - **Nuisance parameters** – Any other parameter of your model
- The goal of practical statistical inference is to make a statement about the POI that accounts for the uncertainties in the NPs so that the NPs themselves don't need to be reported
 - Procedure to accomplish this differs somewhat for various techniques parameter/variance estimation, hypothesis testing, confidence intervals, Bayesian posteriors
- This section is purely on statistical methods – at no point it is important that a NP models a systematic uncertainty.

The statisticians view on nuisance parameters

- In general, our model of the data is not perfect



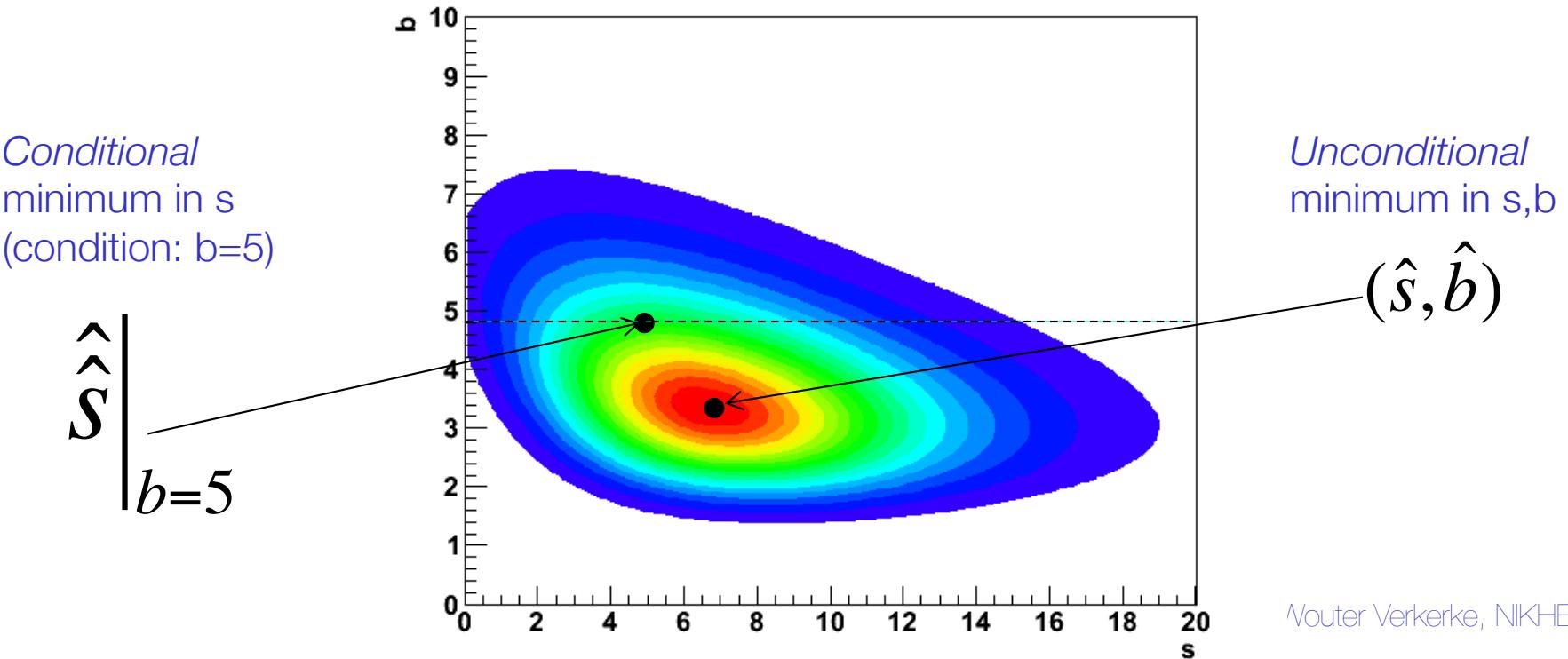
- Can improve modeling by including additional adjustable parameters
- Goal: some point in the parameter space of the enlarged model should be “true”
- Presence of nuisance parameters decreases the sensitivity of the analysis of the parameter(s) of interest

Treatment of nuisance parameters in parameter estimation

- In POI parameter estimation, the effect of NPs incorporated through *unconditional minimization*
 - I.e. minimize Likelihood w.r.t all parameter simultaneously.
- Simple example with 2-bin Poisson counting experiment

$$L(s) = \text{Poisson}(10|s+5)$$

$$L(s,b) = \text{Poisson}(10|s+b)\text{Poisson}(10|3\cdot b)$$



Treatment of nuisance parameters in variance estimation

- Maximum likelihood estimator of parameter variance is based on 2nd derivative of Likelihood
 - For multi-parameter problems this 2nd derivative is generalized by the **Hessian Matrix** of partial second derivatives

$$\hat{\sigma}(p)^2 = \hat{V}(p) = \left(\frac{d^2 \ln L}{d^2 p} \right)^{-1} \quad \rightarrow \quad \hat{\sigma}(p_i)^2 = \hat{V}(p_{ii}) = (H^{-1})_{ii}$$

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- For multi-parameter likelihoods estimate of covariance V_{ij} of pair of 2 parameters in addition to variance of individual parameters
 - Usually re-expressed in terms dimensionless correlation coefficients ρ

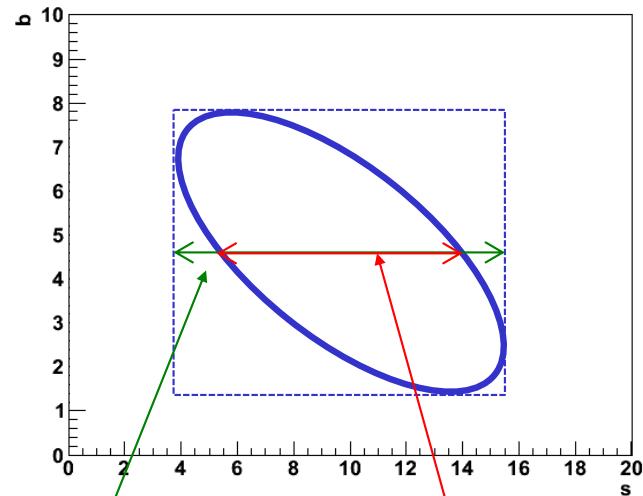
$$V_{ij} = \rho_{ij} \sqrt{V_{ii} V_{jj}}$$

Treatment of nuisance parameters in variance estimation

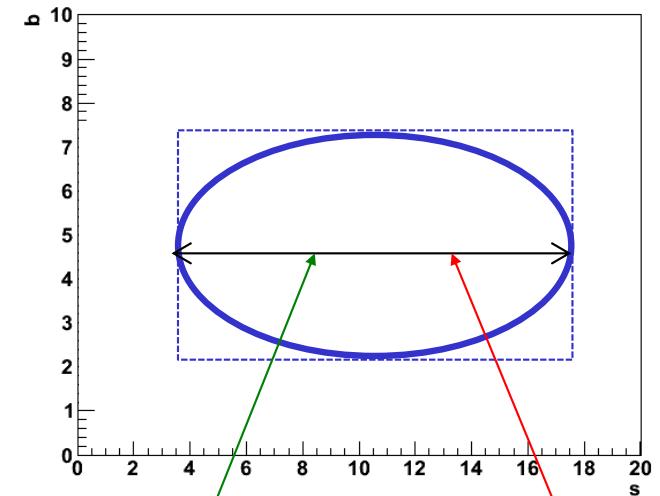
- Effect of NPs on variance estimates visualized

Scenario 1
Estimators of
POI and NP correlated
i.e. $\rho(s,b) \neq 0$

Scenario 2
Estimators of
POI and NP correlated
i.e. $\rho(s,b) = 0$



$$\hat{V}(s) \text{ from } \begin{bmatrix} \frac{\partial^2 L}{\partial s^2} & \frac{\partial^2 L}{\partial s \partial b} \\ \frac{\partial^2 L}{\partial s \partial b} & \frac{\partial^2 L}{\partial b^2} \end{bmatrix}^{-1} \quad \hat{V}(s) \text{ from } \left[\frac{\partial^2 L}{\partial s^2} \right]_{b=\hat{b}}^{-1}$$



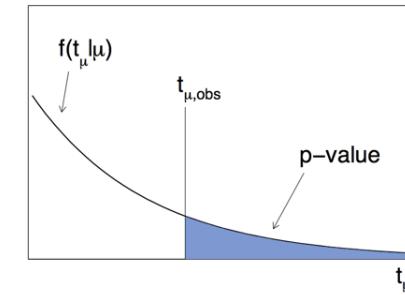
$$\hat{V}(s) \text{ from } \begin{bmatrix} \frac{\partial^2 L}{\partial s^2} & \frac{\partial^2 L}{\partial s \partial b} \\ \frac{\partial^2 L}{\partial s \partial b} & \frac{\partial^2 L}{\partial b^2} \end{bmatrix}^{-1} \quad \hat{V}(s) \text{ from } \left[\frac{\partial^2 L}{\partial s^2} \right]_{b=\hat{b}}^{-1}$$

Uncertainty on background increases uncertainty on signal

Treatment of NPs in hypothesis testing and conf. intervals

- We've covered frequentist hypothesis testing and interval calculation using likelihood ratios based on a likelihood with a single parameter (of interest) $L(\mu)$
 - Result is p-value on hypothesis with given μ value, or
 - Result is a confidence interval $[\mu_-, \mu_+]$ with values of μ for which p-value is at or above a certain level (the confidence level)
- How do you do this with a likelihood $L(\mu, \theta)$ where θ is a nuisance parameter?
 - With a test statistics q_μ , we calculate p-value for hypothesis θ as

$$p_\mu = \int_{q_{\mu, \text{obs}}}^{\infty} f(q_\mu | \mu, \theta) dq_\mu$$



- But what values of θ do we use for $f(q_\mu | \mu, \theta)$?
Fundamentally, we want to reject μ only if $p < \alpha$ for all θ
→ Exact confidence interval

Hypothesis testing & conf. intervals with nuisance parameters

- The goal is that the parameter of interest should be covered at the stated confidence **for every value of the nuisance parameter**
- if there is *any value* of the nuisance parameter which makes the data consistent with the parameter of interest, that value of the POI should be considered:
 - e.g. don't claim discovery if any background scenario is compatible with data
- But: technically very challenging and significant problems with over-coverage
 - Example: *how broadly should 'any background scenario' be defined?* Should we include background scenarios that are clearly incompatible with the observed data?

Example of over-coverage

- The 1958 thought expt of David R. Cox focused the issue:
 - Your procedure for weighing an object consists of flipping a coin to decide whether to use a weighing machine with a 10% error or one with a 1% error; and then measuring the weight.
- Then “surely” the error you quote for your measurement should reflect which weighing machine you actually used, and not the average error of the “whole space” of all measurements!
- But this is not how the classical frequentist confidence interval works!
 - Suppose weight=100, coin='1% error' Can you exclude weight=90 at 95% C.L?
 - No: because for 'coin=10% error' weight=90 cannot be excluded at 95% C.L.
- Solution: conditioning on observed data will make result more relevant (at expense of exact frequentist coverage)
 - Restricting whole space of probabilities to 'coin=1% error' only if that is observed allows to exclude weight=90 at 95% C.L.

The profile likelihood construction as compromise

- For LHC the following prescription is used:

Given $L(\mu, \theta)$
POI ↗
↖ NPs

perform hypothesis test for each value of μ (the POI),

using values of nuisance parameter(s) θ that best fit the data under the hypothesis μ

- Introduce the following notation

$$\hat{\theta}(\mu) \quad \begin{array}{l} \text{M.L. estimate of } \theta \text{ for a given value of } \mu \\ (\text{i.e. a conditional ML estimate}) \end{array}$$

- The resulting confidence interval will have exact coverage for the points $(\mu, \hat{\theta}(\mu))$
 - Elsewhere it may overcover or undercover (but this can be checked)

The profile likelihood ratio

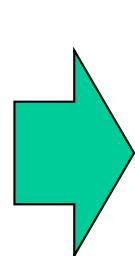
- With this prescription we can construct the **profile likelihood ratio** as test statistic

Likelihood for given μ

$$\lambda(\mu) = \frac{L(\mu)}{L(\hat{\mu})}$$

Maximum Likelihood

Maximum Likelihood for given μ

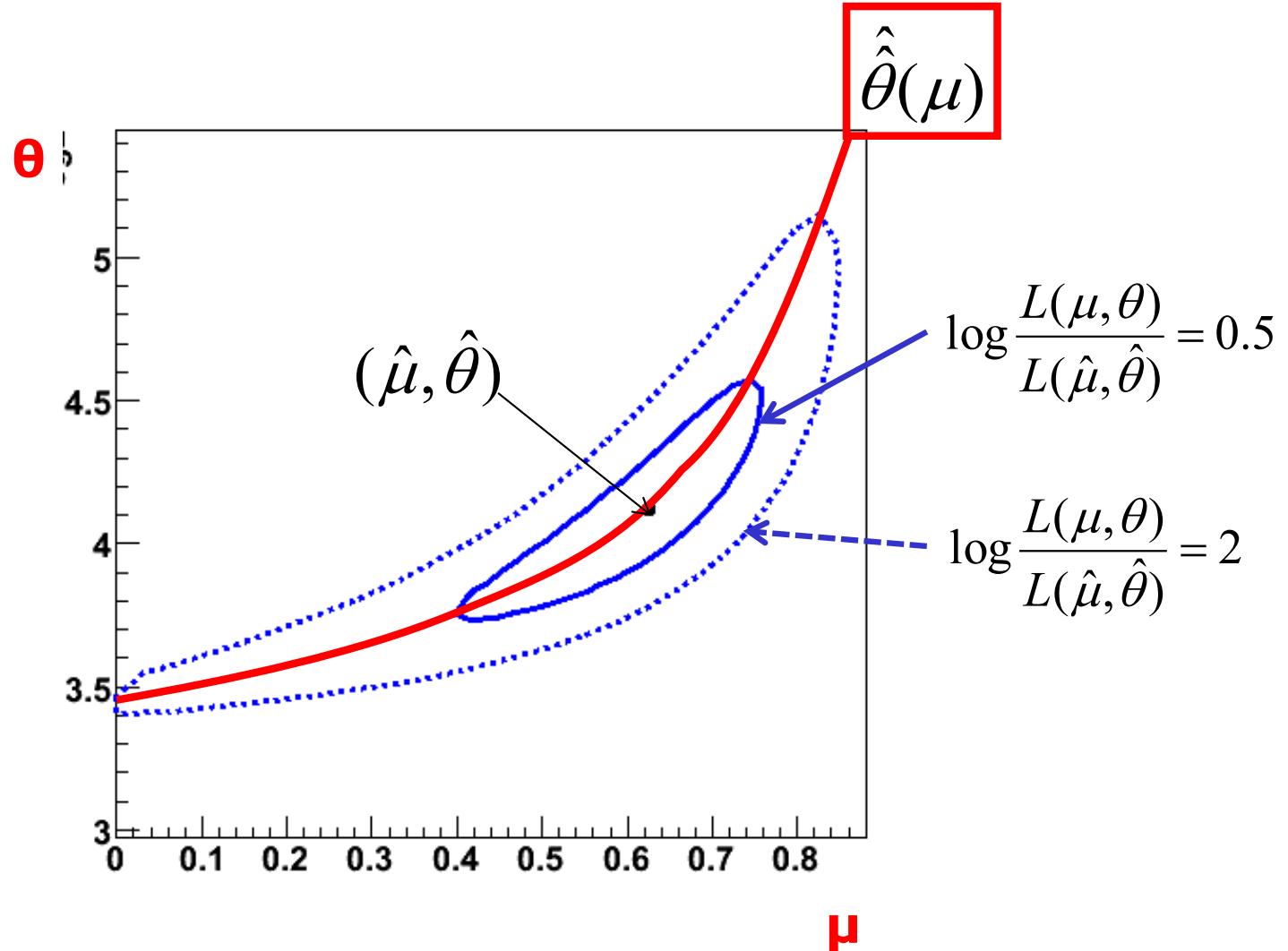


$$\lambda(\mu) = \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})}$$

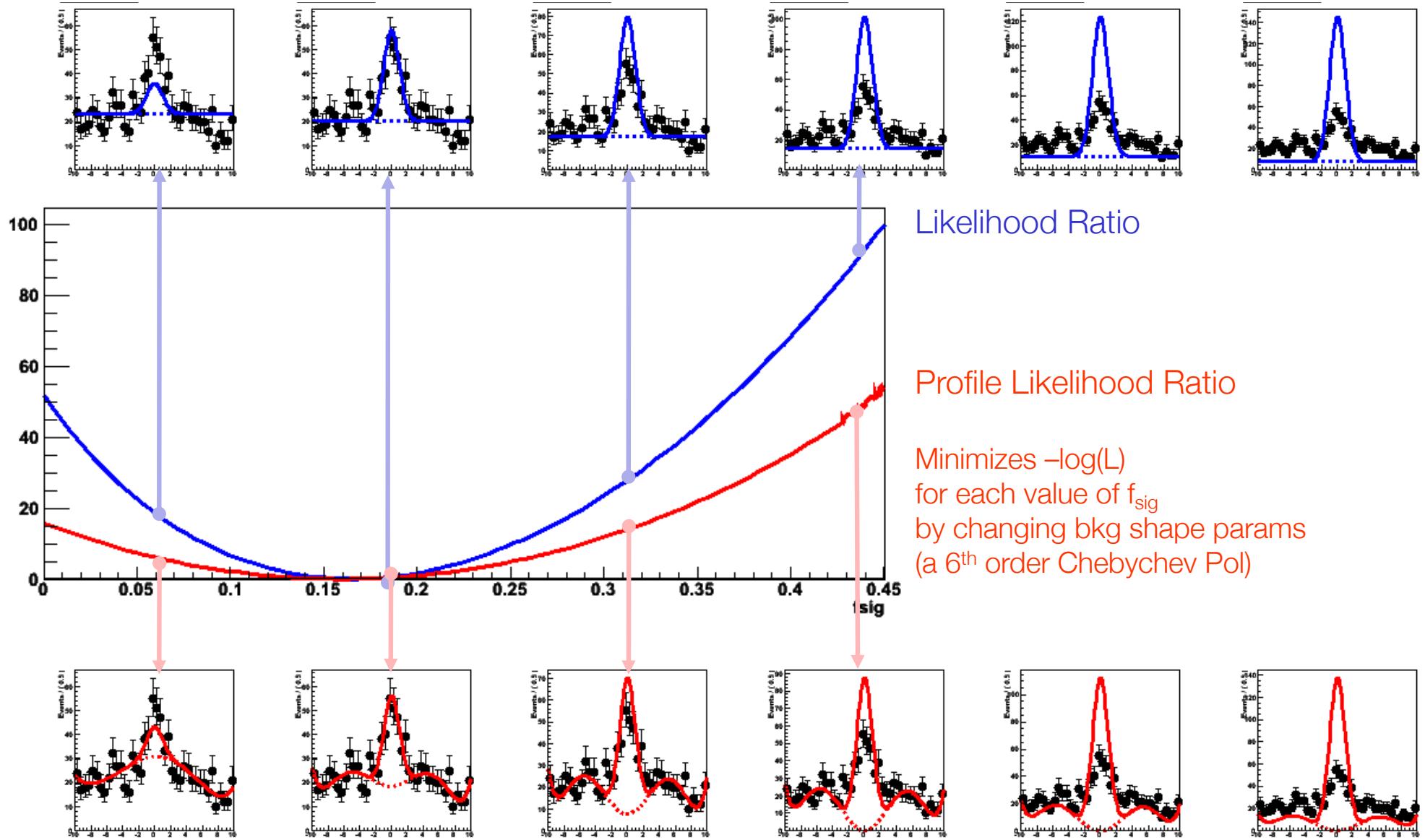
Maximum Likelihood

- NB: value profile likelihood ratio does *not* depend on θ

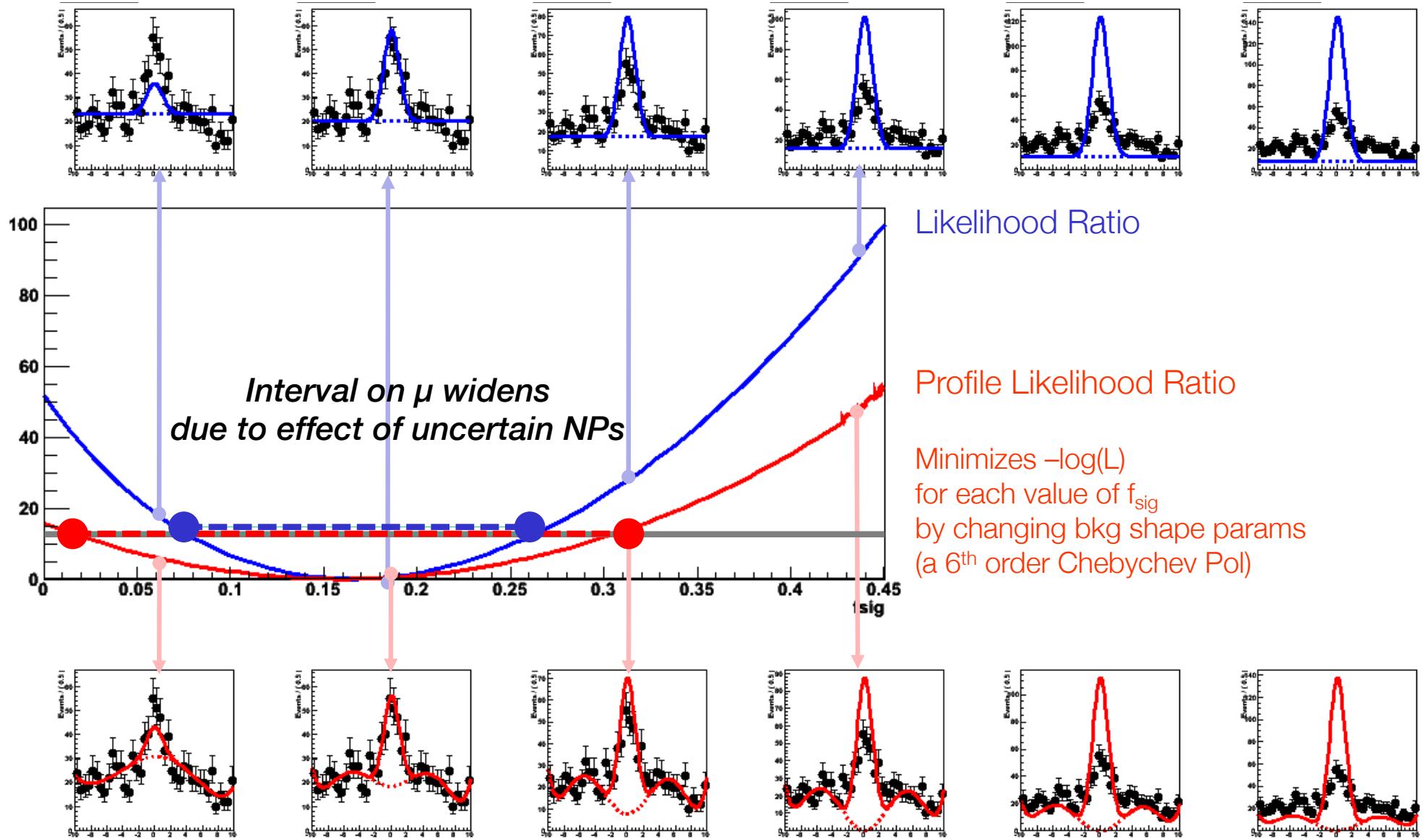
Profiling illustration with one nuisance parameter



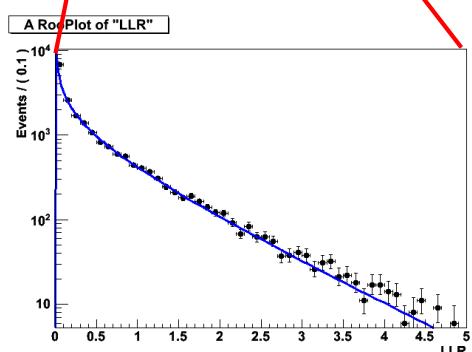
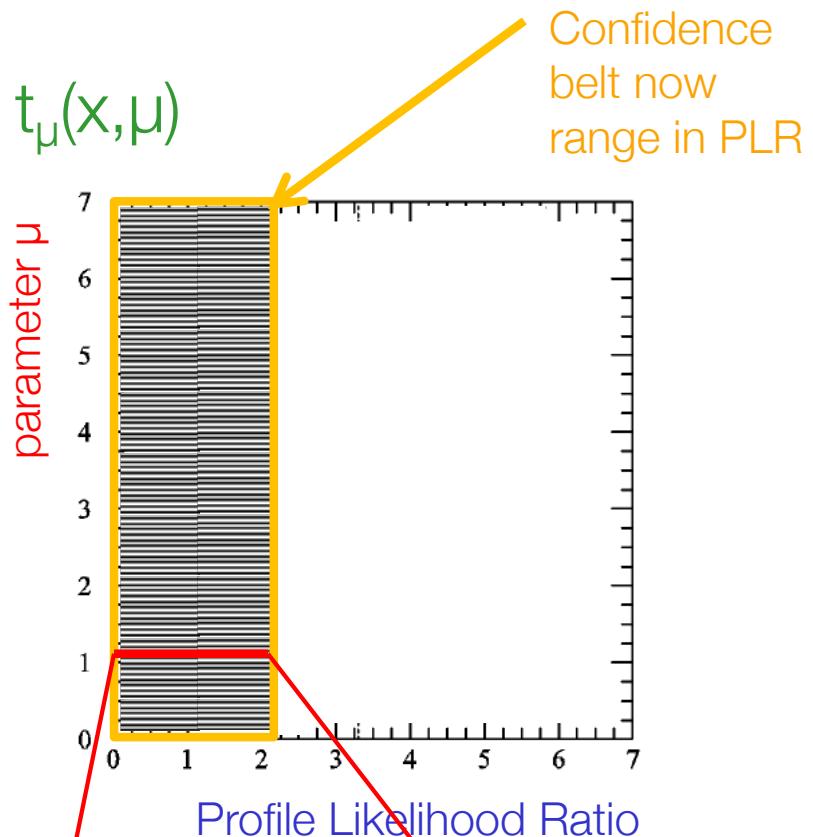
Profile scan of a Gaussian plus Polynomial probability model



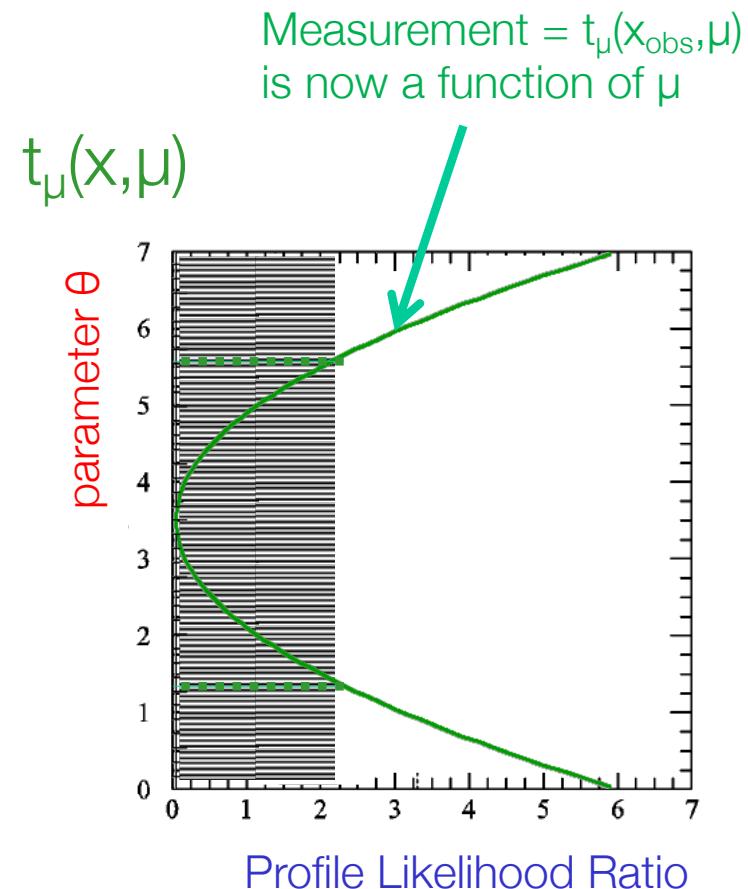
Profile scan of a Gaussian plus Polynomial probability model



PLR Confidence interval vs MINOS



Asymptotically,
distribution is identical
for all μ



*NB: asymptotically, distribution
is also independent of true
values of θ*

$$f(t_\mu; \Lambda) = \frac{1}{2\sqrt{t_\mu}} \frac{1}{\sqrt{2\pi}} \left[\exp\left(-\frac{1}{2} (\sqrt{t_\mu} + \sqrt{\Lambda})^2\right) + \exp\left(-\frac{1}{2} (\sqrt{t_\mu} - \sqrt{\Lambda})^2\right) \right]$$

$$\Lambda = \frac{(\mu - \mu')^2}{\sigma^2} .$$

Summary of statistical treatment of nuisance parameters

- All of the statistical techniques mentioned in section 2 have an associated technique to propagate the effect of the NPs on the POI
 - Parameter estimation → Joint unconditional estimation
 - Variance estimation → Replace d^2L/dp^2 with Hessian matrix
 - Hypothesis tests & confidence intervals → Use profile likelihood ratio
 - Bayesian credible intervals → Integration ('Marginalization')
- Be sure to use the right procedure with the right method
 - Anytime you integrate a Likelihood you are a Bayesian
 - If you sample something chances are you performing either a (Bayesian) Monte Carlo integral, or are doing glorified error propagation
- Answers can differ substantially between methods!
 - This is not always a problem, but can also be a consequence of a difference in the problem statement

Systematic uncertainties and profiling

Wouter Verkerke
(Nikhef/Atlas)



Wouter Verkerke, Nikhef

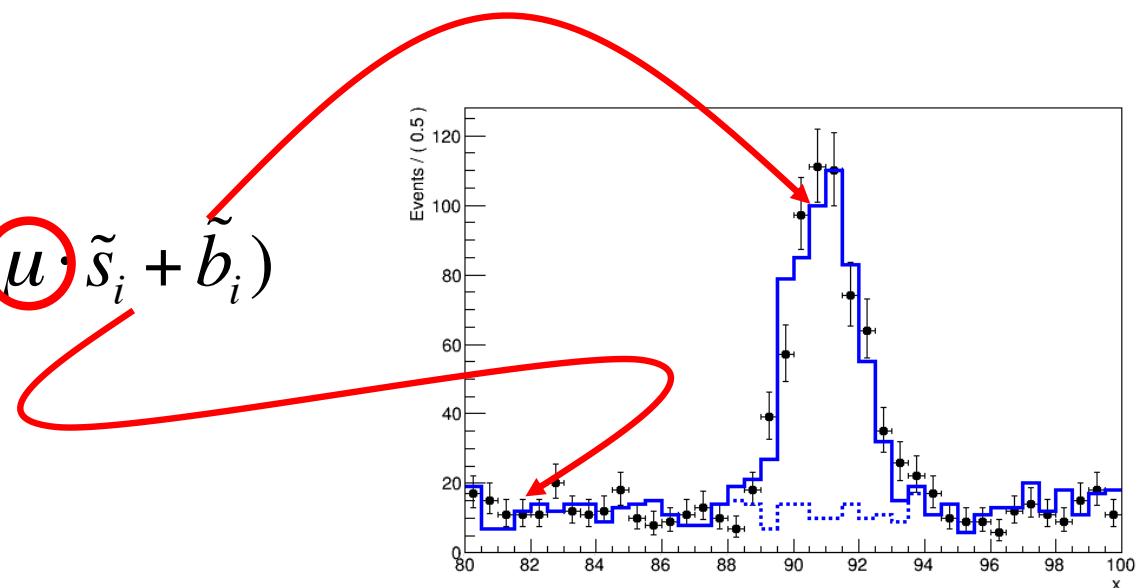
Outline of this course

- Outline of this course
 1. What are systematic uncertainties?
 2. The likelihood function as basis for statistical inference
 3. Incorporating systematic uncertainties in probability models
 4. Dealing with nuisance parameters in statistical inference
- 5. **Modeling shape systematics: template morphing**
- 6. Tools for modelling building
- 7. **Diagnostics I: Fit stability, understanding how minimizers work**
- 8. **Diagnostics II: Result diagnostics, choice of nuisance parameters**
- 9. Summary

So far we've only considered 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

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

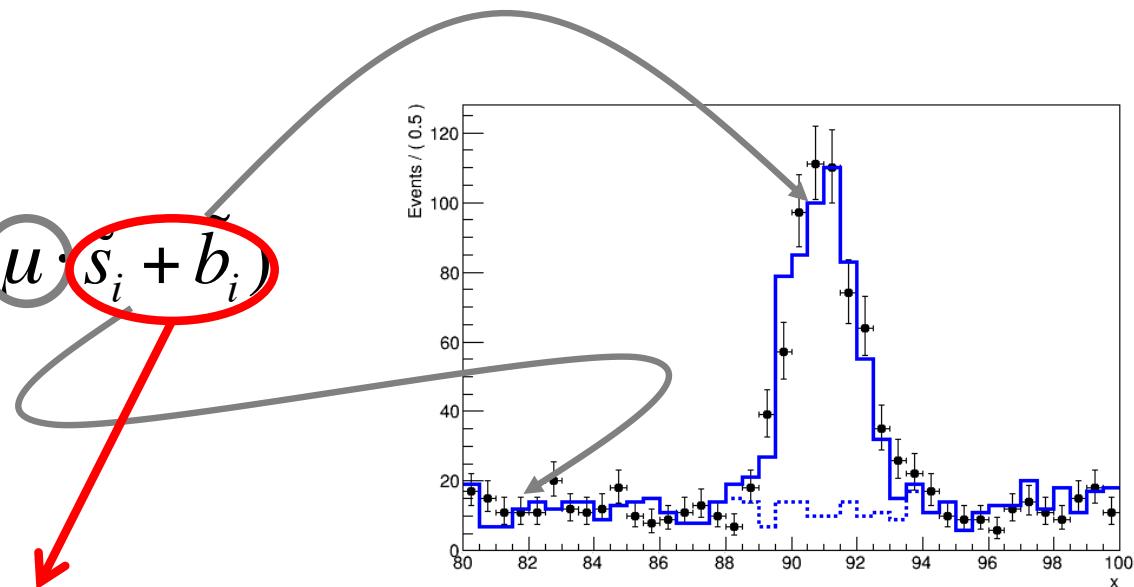


- So far only considered a single parameter in the likelihood: the physics *parameter of interest*, usually denoted as μ

The imperfect experiment

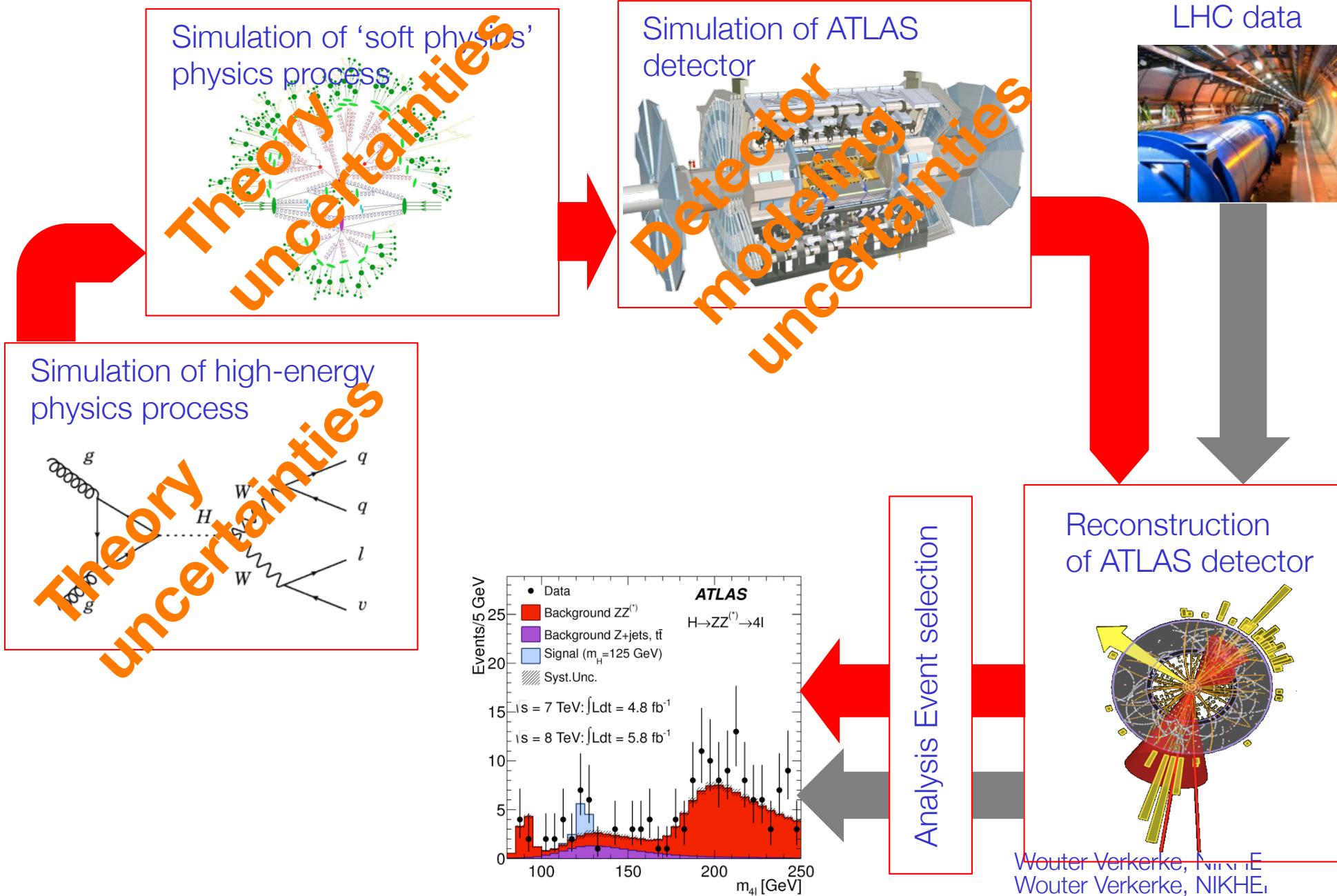
- In realistic measurements many effect that we don't control exactly influence measurements of parameter of interest
- How do you model these uncertainties in the likelihood?

$$L(\vec{N} \mid \mu) = \prod_{\text{bins}} \text{Poisson}(N_i \mid \mu \cdot s_i + b_i)$$



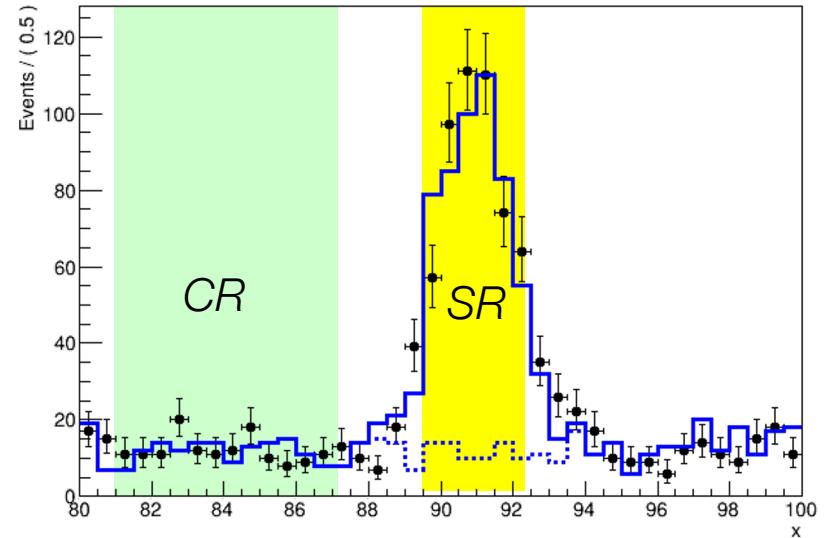
*Signal and background predictions
are affected by (systematic) uncertainties*

The simulation workflow and origin of uncertainties



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

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

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

NB: Define parameter ‘b’ to represents the amount of bkg in the SR.

Scale factor $\tilde{\tau}$ accounts for difference in size between SR and CR

“Background uncertainty constrained from the data”

- Full likelihood of the measurement (‘simultaneous fit’)

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

Generalizing the concept of the sideband measurement

- Background uncertainty from sideband clearly clearly not a ‘systematic uncertainty’

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

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

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

‘Measured background rate by MC simulation’

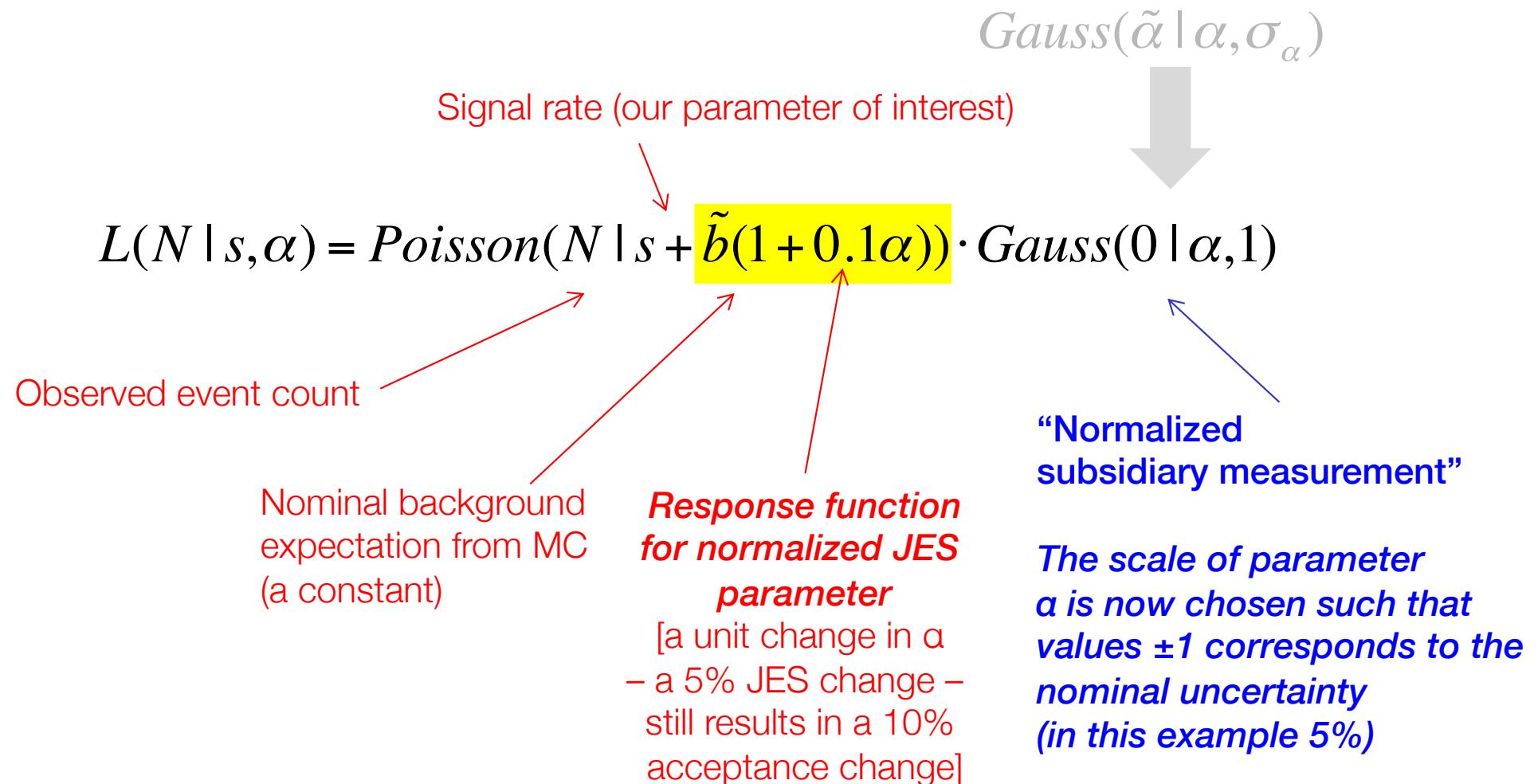
‘Subsidiary measurement’
of background rate

- *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’

Modeling a detector calibration uncertainty

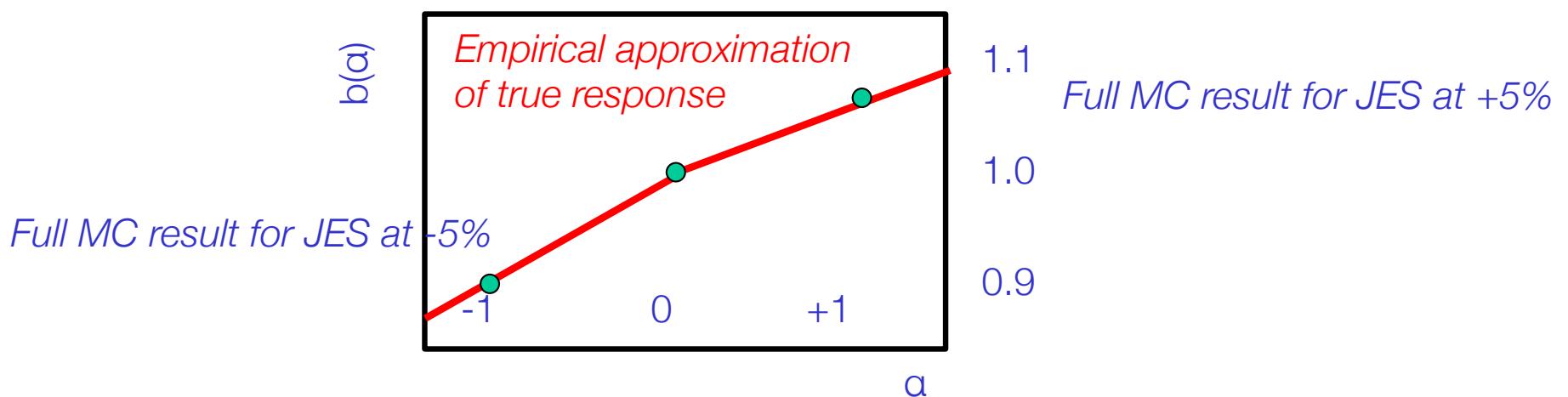
- Simplify expression by renormalizing “subsidiary measurement”



The response function as empirical model of full simulation

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

- 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 $\pm 1\sigma$ 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 $\pm 5\%$



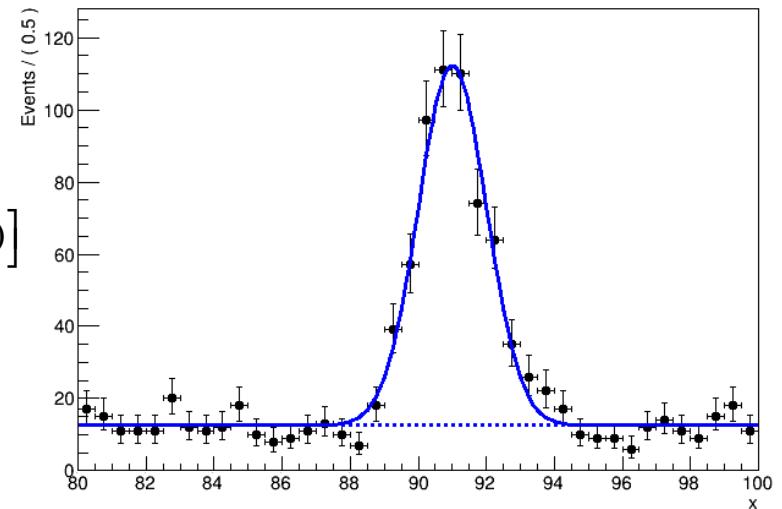
5

Modeling shape systematics: template morphing

Introducing response functions for shape uncertainties

- Modeling of systematic uncertainties in **Likelihoods describing distributions** follows the same procedure as for counting models
 - Example: Likelihood modeling distribution in a di-lepton invariant mass. POI is the signal strength μ

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

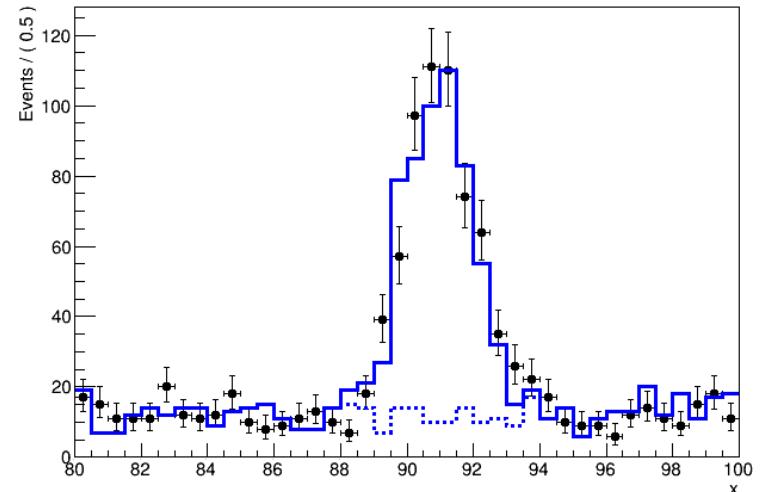


- Consider a lepton energy scale systematic uncertainty that affects this measurement
 - The LES has been measured with a 1% precision
 - The effect of LES on m_{ll} has been determined to a 2% shift for 1% LES change

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

Response modeling for distributions

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



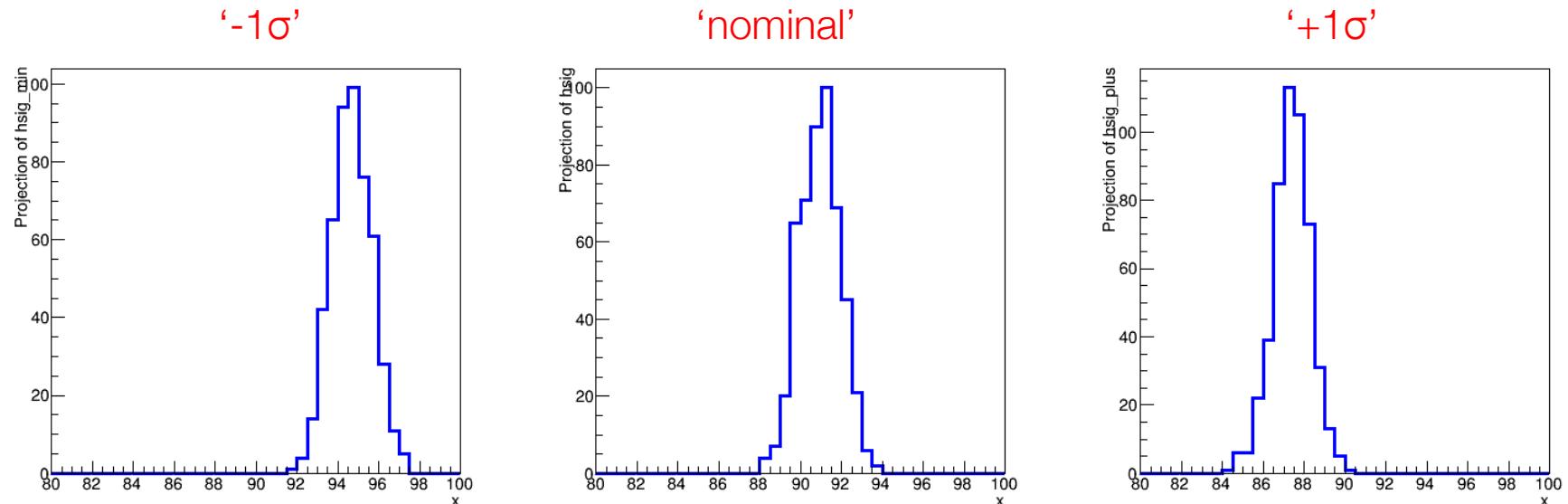
$$L(\vec{N} | \mu) = \prod_i Poisson(N_i | \mu \tilde{s}_i + \tilde{b}_i)$$

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

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

Modeling of shape systematics in the likelihood

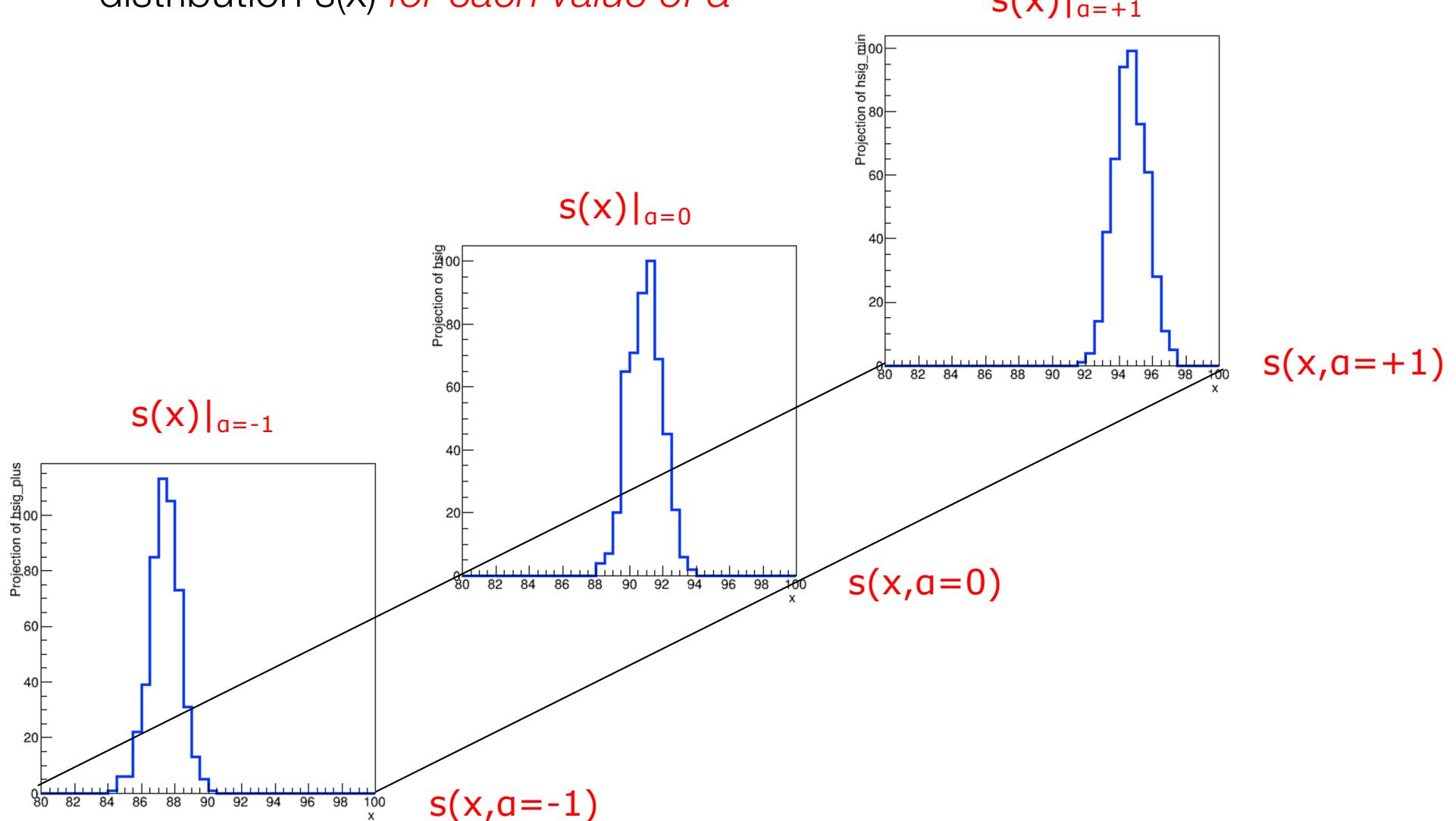
- Effect of *any* systematic uncertainty that affects the shape of a distribution can in principle be obtained from MC simulation chain
 - Obtain histogram templates for distributions at ' $+1\sigma$ ' and ' -1σ ' settings of systematic effect



- Now construct a response function based on the shape of these three templates.

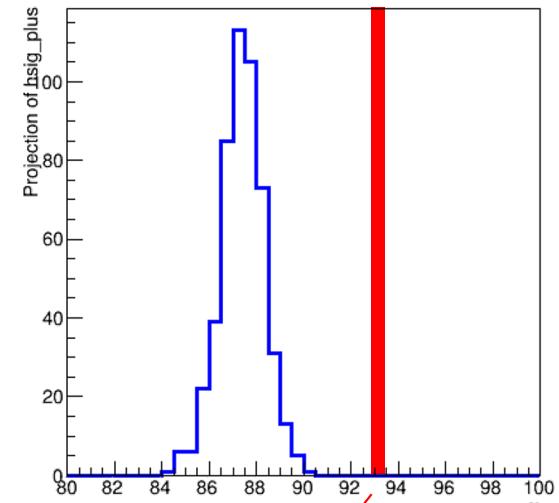
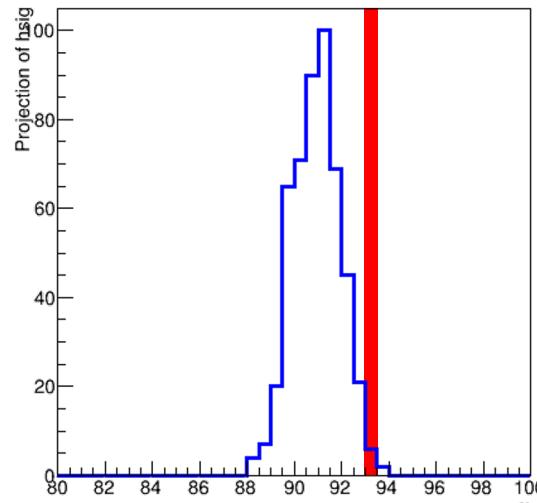
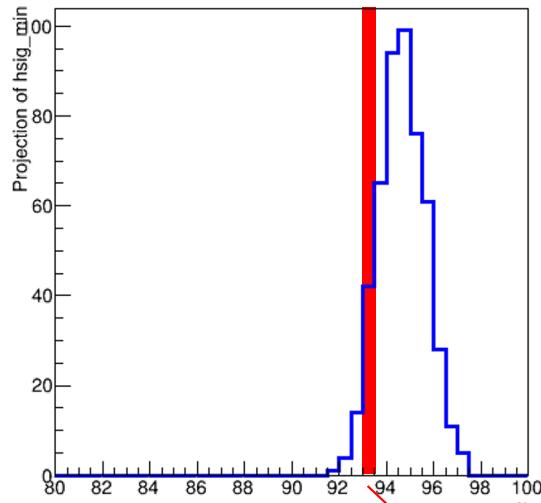
Need to interpolate between template models

- Need to define ‘morphing’ algorithm to define distribution $s(x)$ *for each value of a*

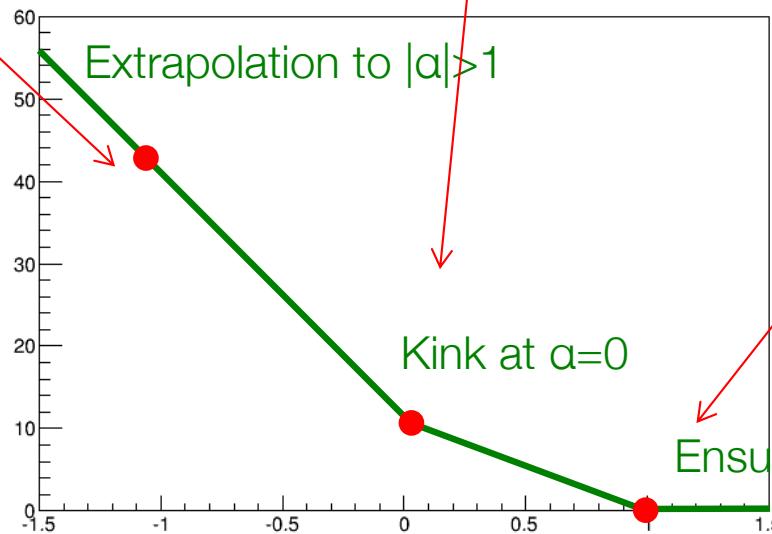


Piecewise linear interpolation

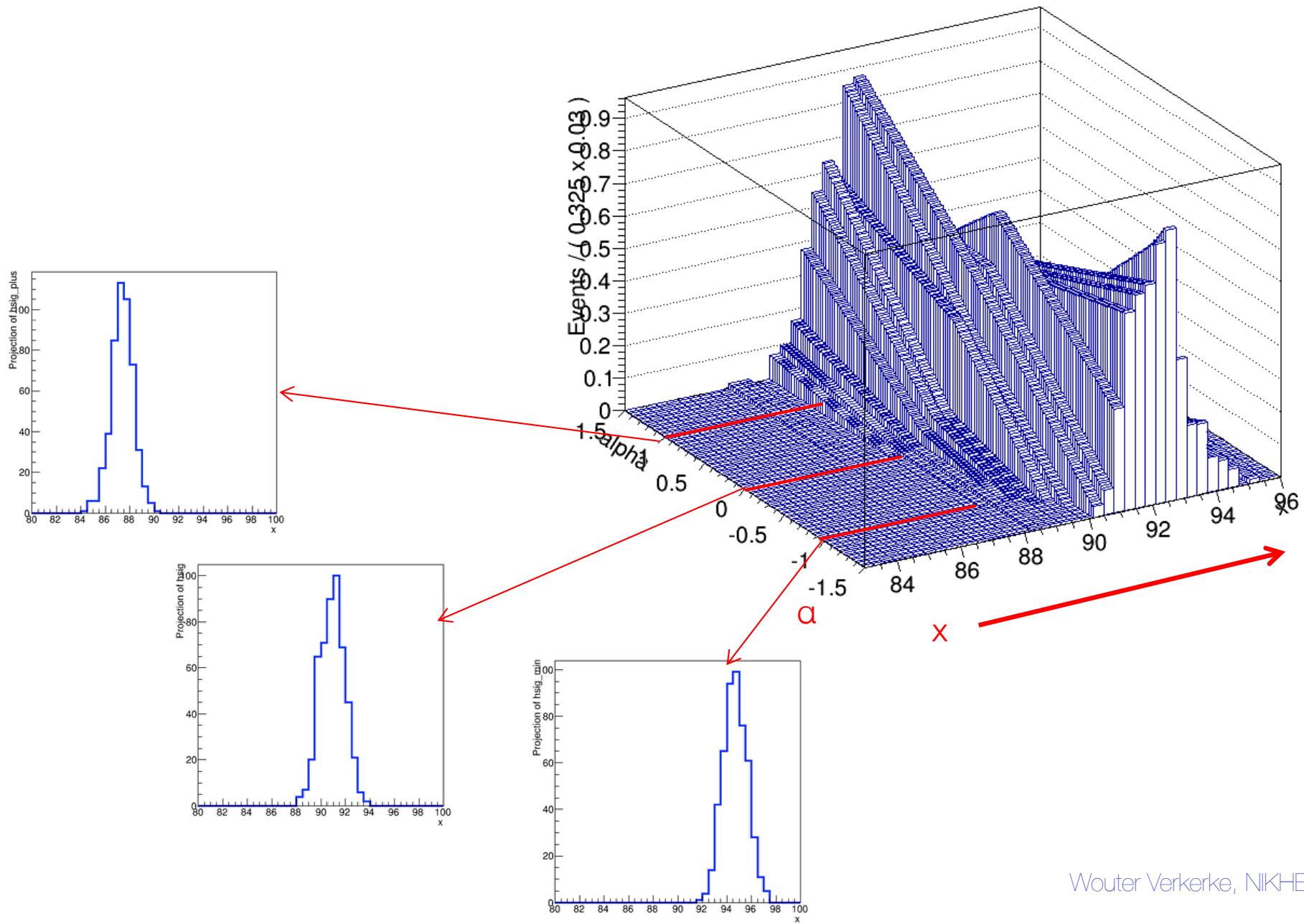
- Simplest solution is piece-wise linear interpolation for each bin



Piecewise linear
interpolation
response model
for a one bin



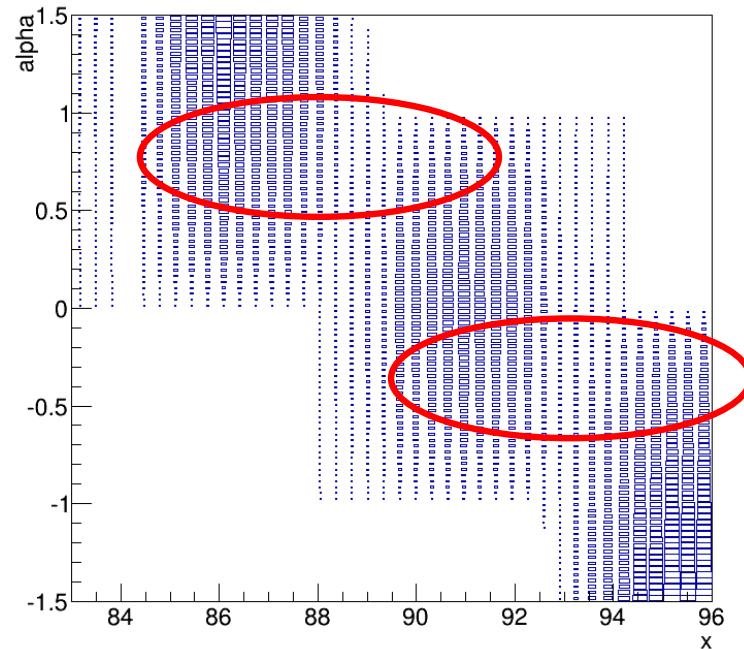
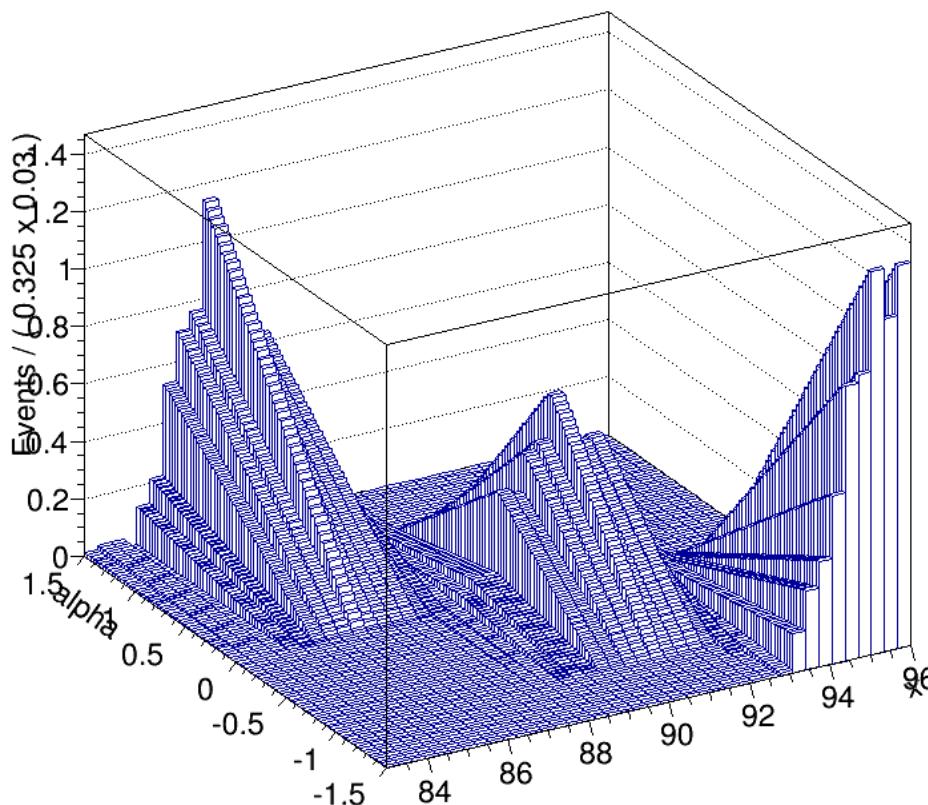
Visualization of bin-by-bin linear interpolation of distribution



Limitations of piece-wise linear interpolation

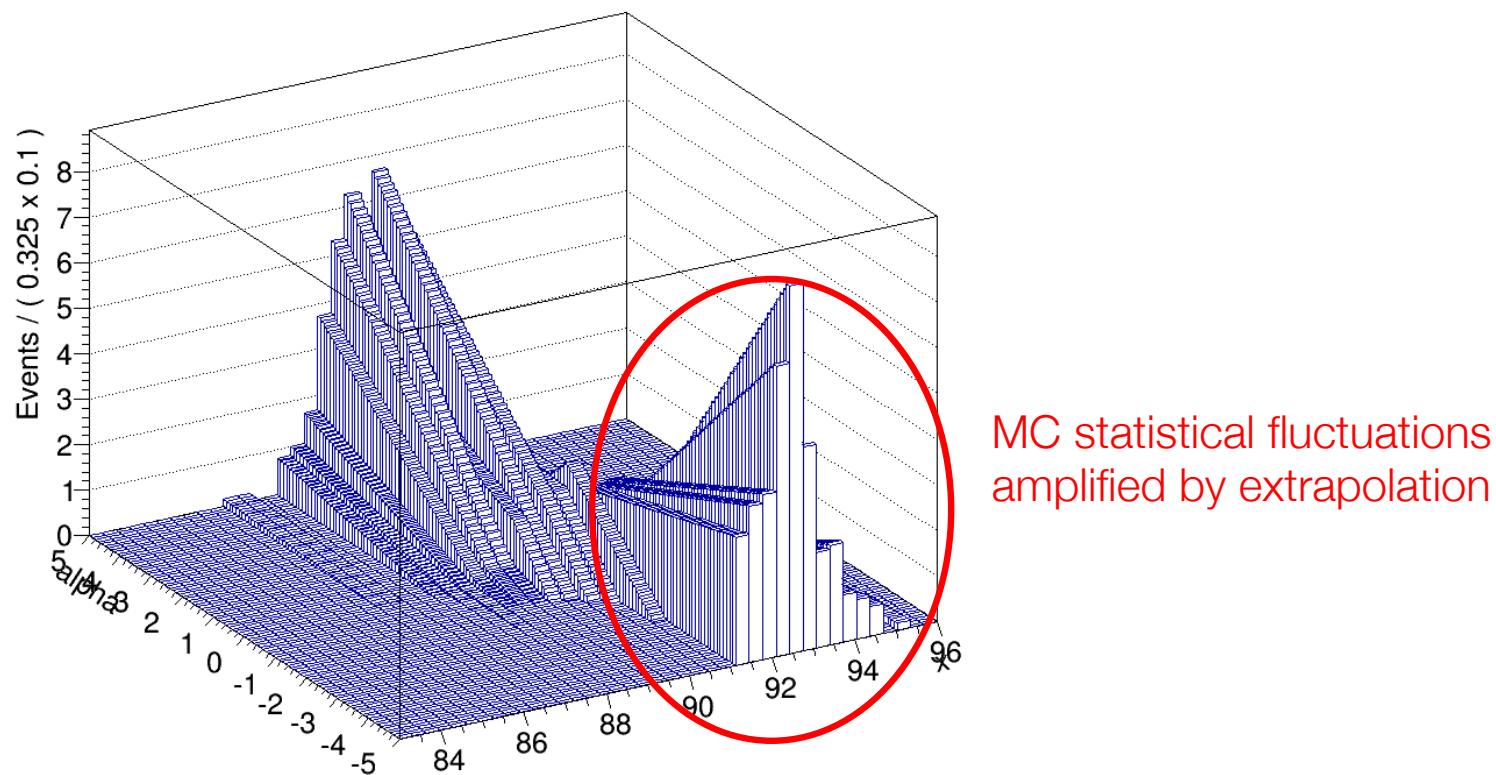
- Bin-by-bin interpolation looks spectacularly easy and simple, but be aware of its limitations
 - Same example, but with larger ‘mean shift’ between templates

Note double peak structure around $|alpha|=0.5$



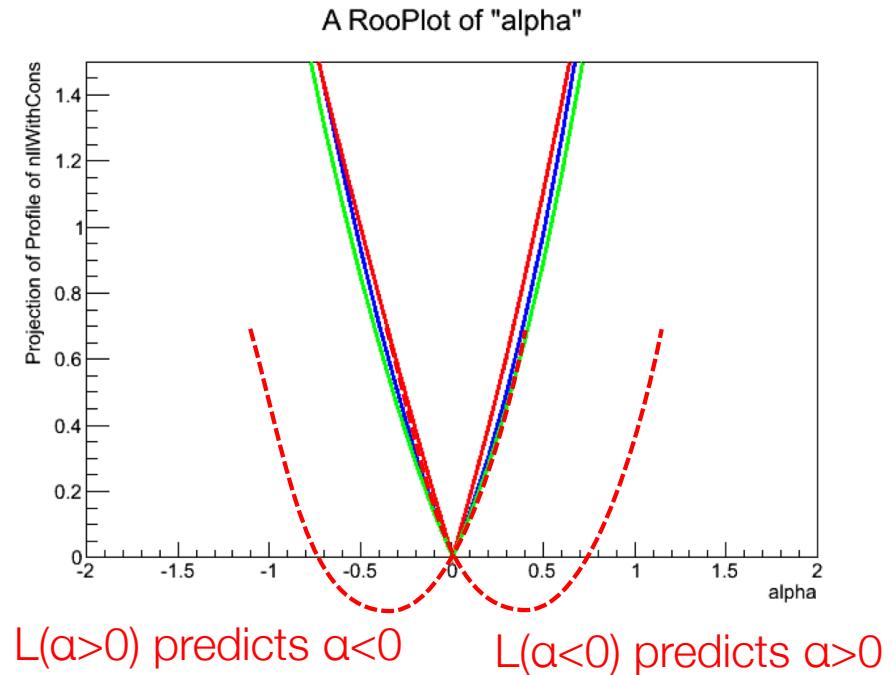
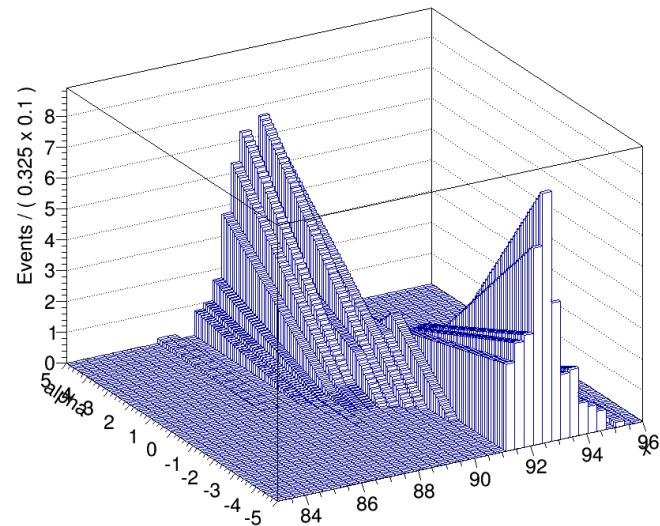
Limitations of piece-wise linear interpolation

- Also be aware of extrapolation effects
 - Nuisance parameters associated to systematic uncertainties can be pulled well beyond ‘ 1σ ’, especially in high-significance hypothesis testing
 - Original example (with modest shift), **but now visualized up to $|a|=5$**



Non-linear interpolation options

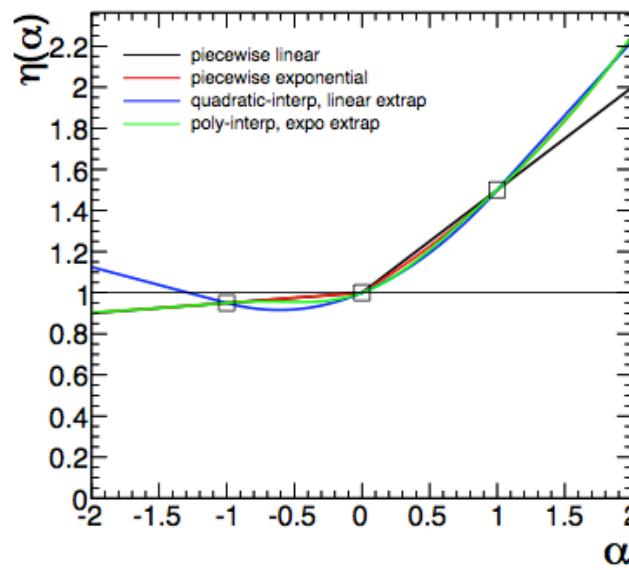
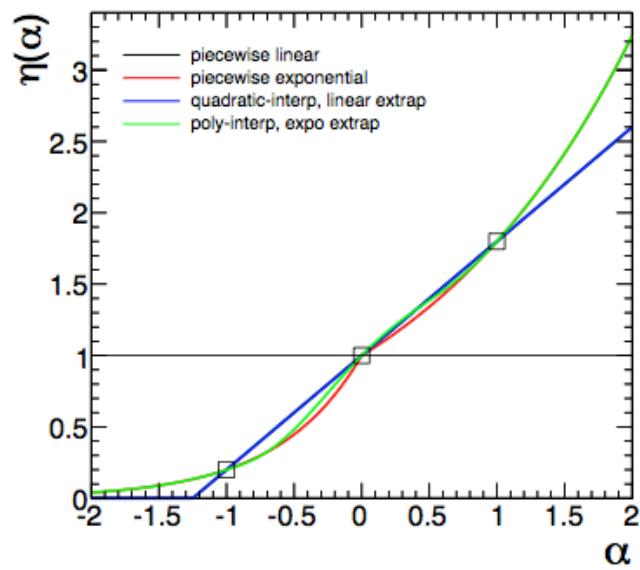
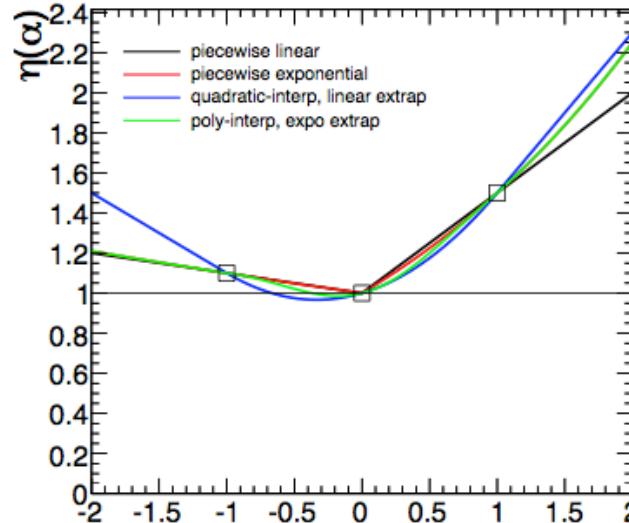
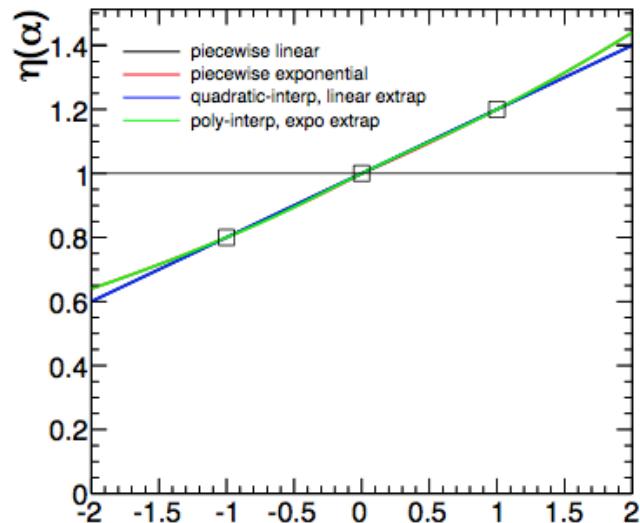
- Piece-wise linear interpolation leads to kink in response functions that may result in pathological likelihood functions



- A variety of other interpolation options exist that improve this
 - Parabolic interpolation/linear extrapolation (but causes shift of minimum)
 - Polynomial interpolation [orders 1,2,4,6]/linear extrapolation (order 1 term allows for asymmetric modeling of templates)

Non-linear interpolation options

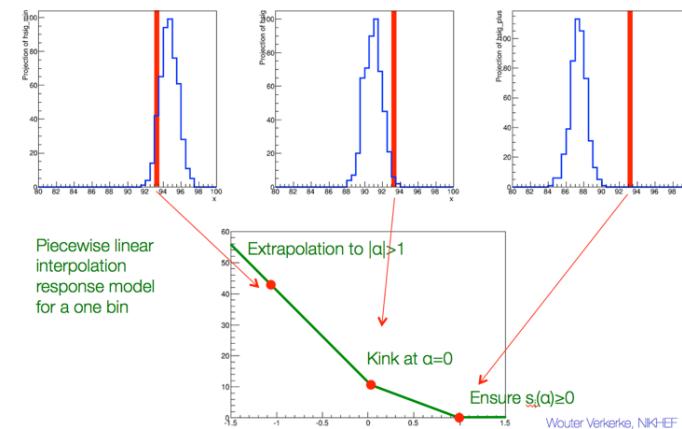
- Comparison of common interpolation options



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
 - One-parameter interpolation

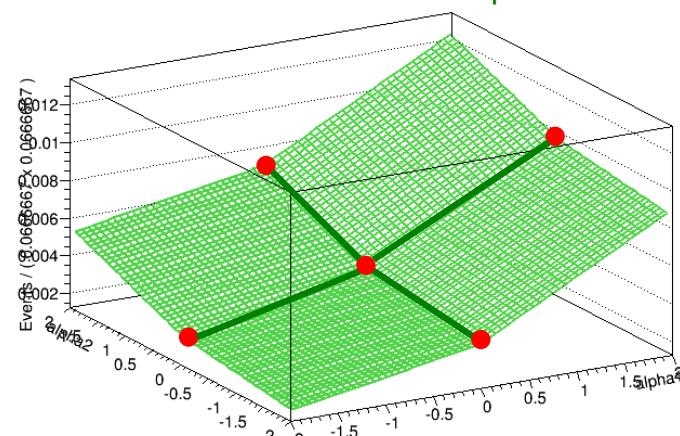
$$s_i(\alpha) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$



- N-parameter interpolation

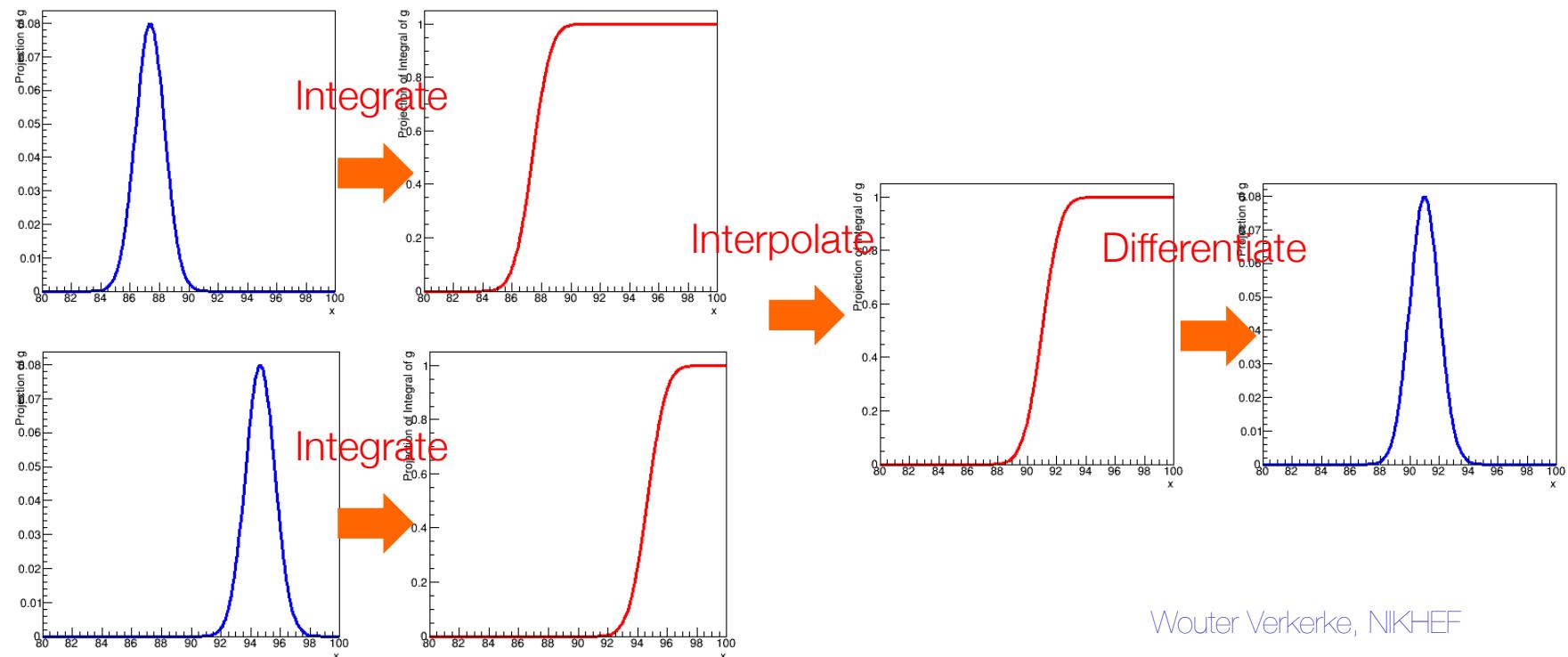
$$s_i(\vec{\alpha}) = \begin{cases} s_i^0 + \sum_j \alpha_j \cdot (s_i^{+,j} - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \sum_j \alpha_j \cdot (s_i^0 - s_i^{-,j}) & \forall \alpha < 0 \end{cases}$$

Visualization of 2D interpolation



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

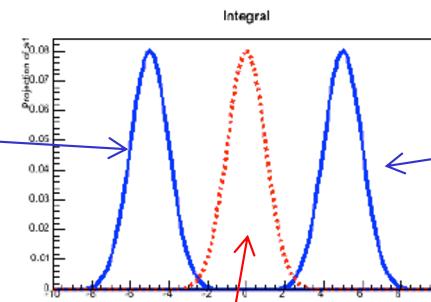


Yet another morphing strategy – ‘Moment morphing’

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

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

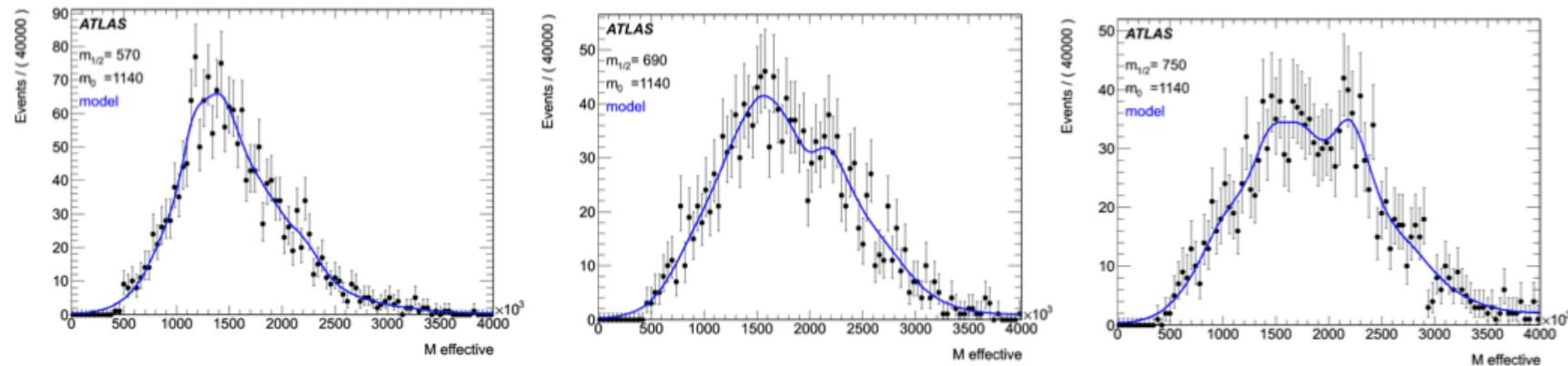
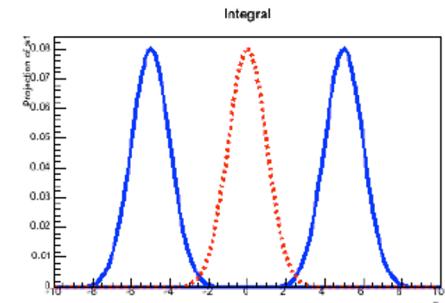
- It constructs this morphed function as combination of linearly transformed input models

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

- Where constants a, b, c, d are chosen such so that $f(x, \alpha)$ satisfies conditions [1]

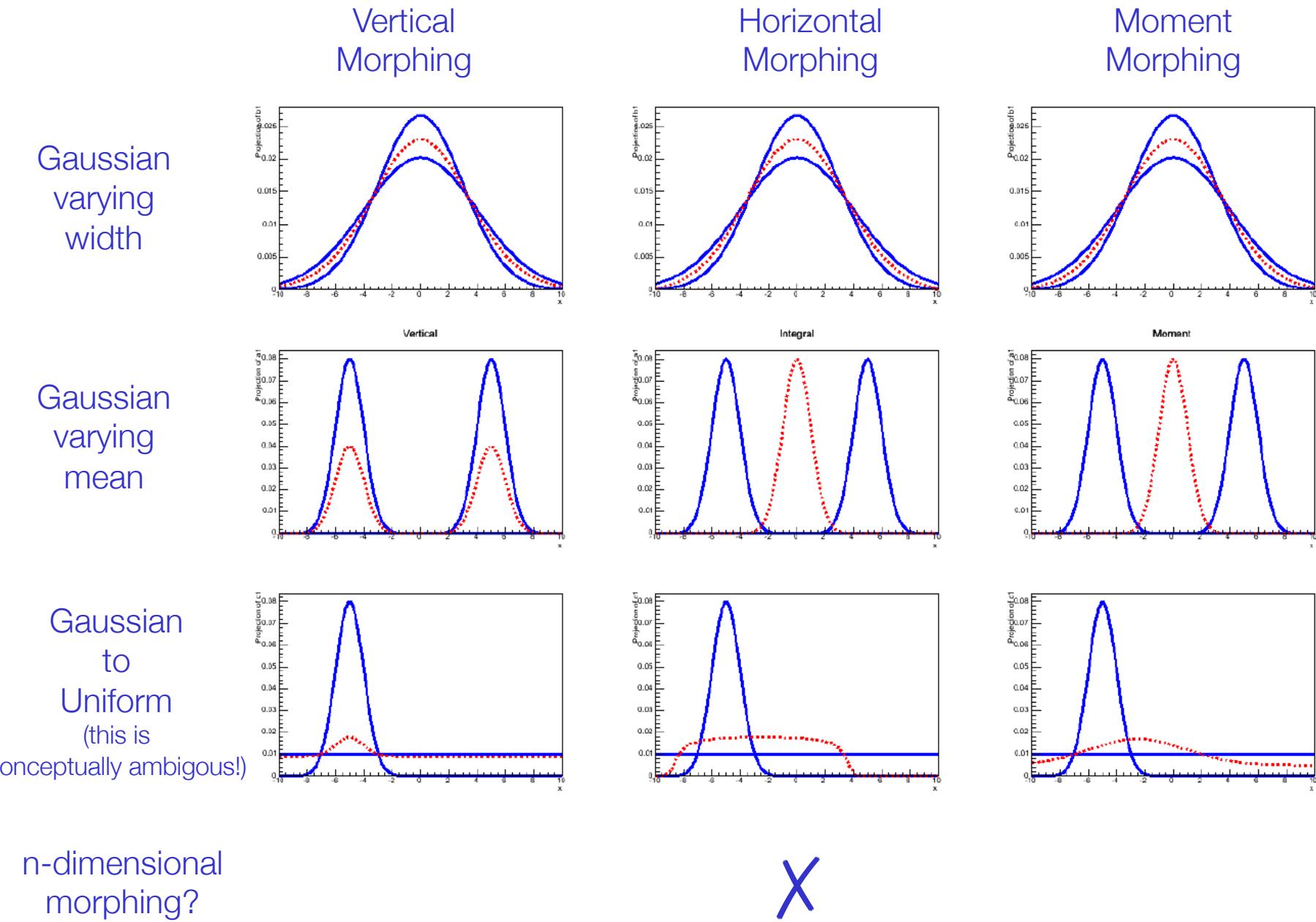
Yet another morphing strategy – ‘Moment morphing’

- 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



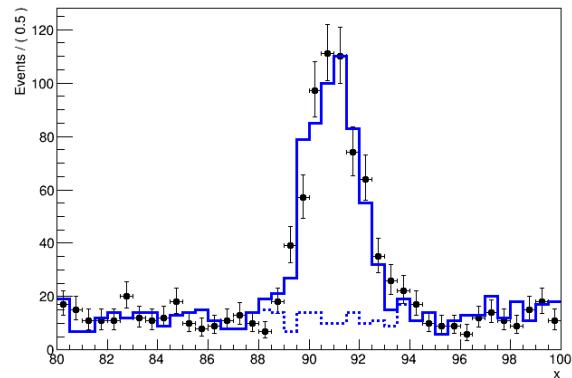
- Good computational performance
 - Calculation of moments of templates is expensive, but just needs to be done once, otherwise very fast (just linear algebra)
$$f(x, \alpha) \rightarrow \alpha f_{-}(ax + b) + (1 - \alpha)f_{+}(cx - d)$$
- Multi-dimensional interpolation strategies exist

Comparison of morphing algorithms

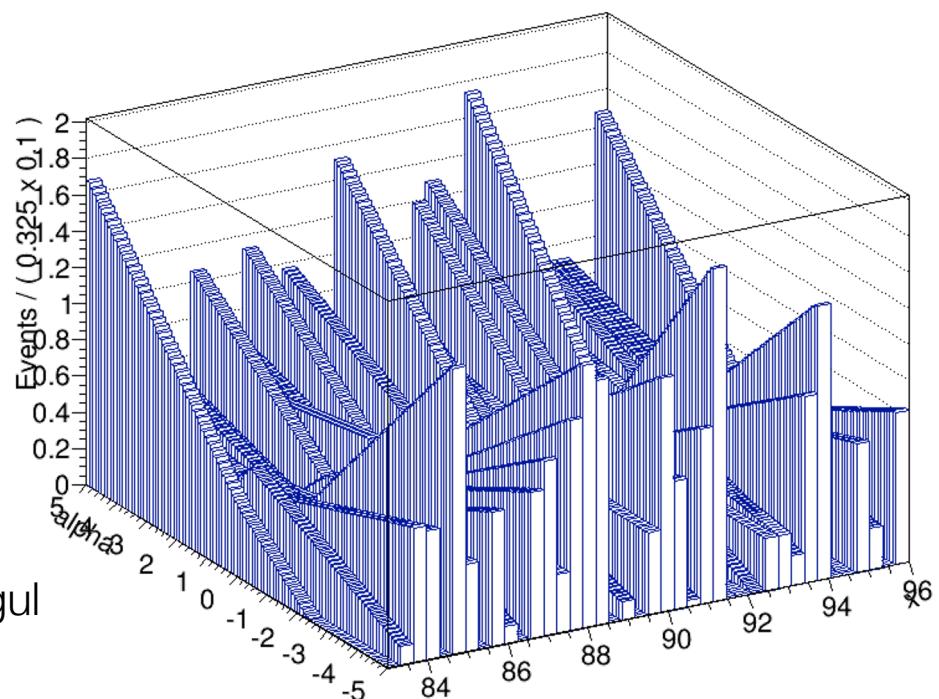


Shape, rate or no systematic?

- 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?

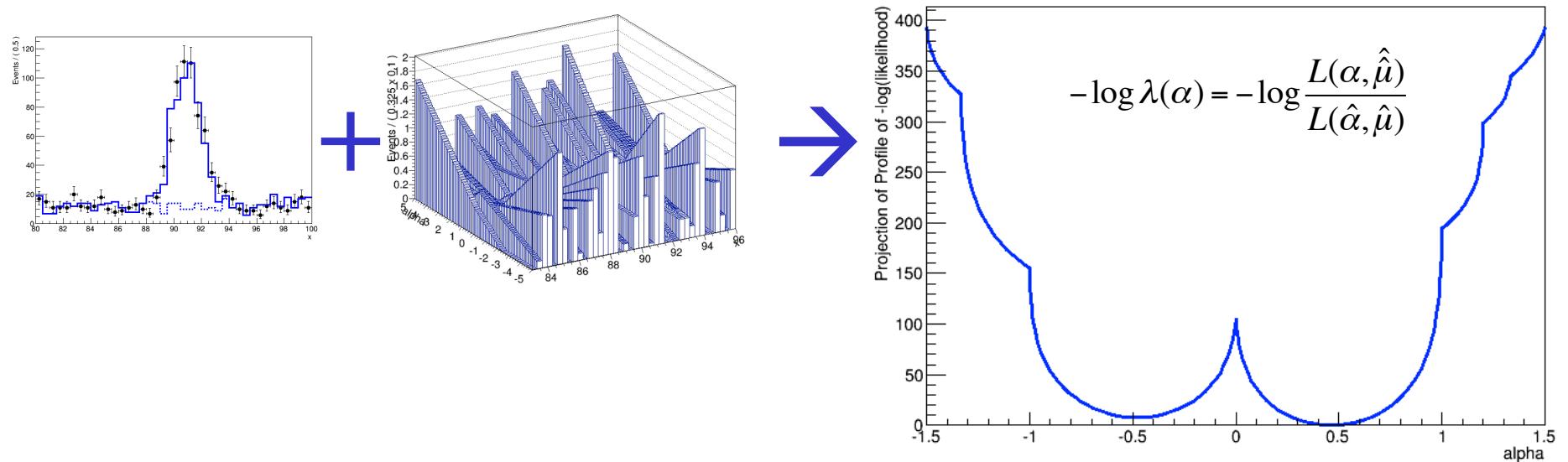


- A χ^2 or KS test between nominal and alternate template can help to decide if a shape uncertainty is meaningful
- Most systematic uncertainties affect both rate and shape, but can make independent decision on modeling rate (which less likely to affect fit stability)



Fit stability due to insignificant shape systematics

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



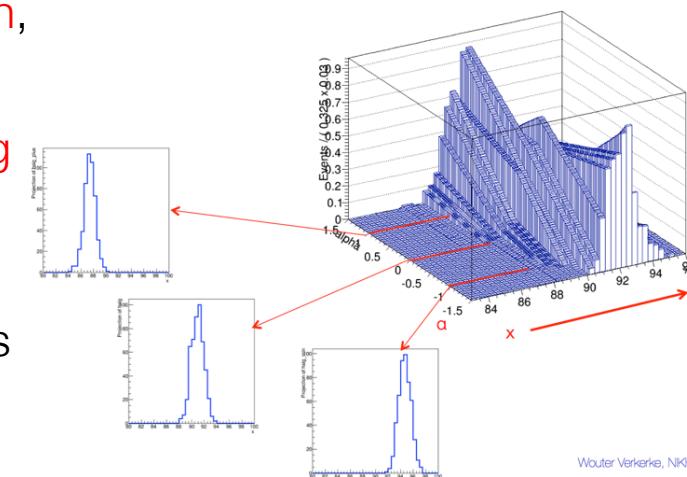
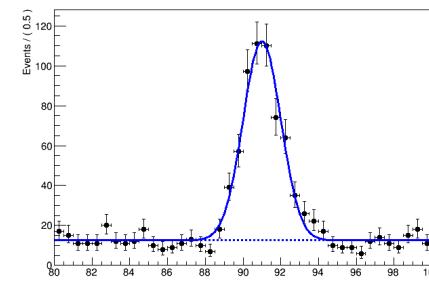
- 1) Numerical minimization process will be ‘interesting’
- 2) MC statistical effects induce strongly defined minima that are fake
 - Because for this example all three templates were sampled from the same parent distribution (a uniform distribution)

Recap on shape systematics & template morphing

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

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

- 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



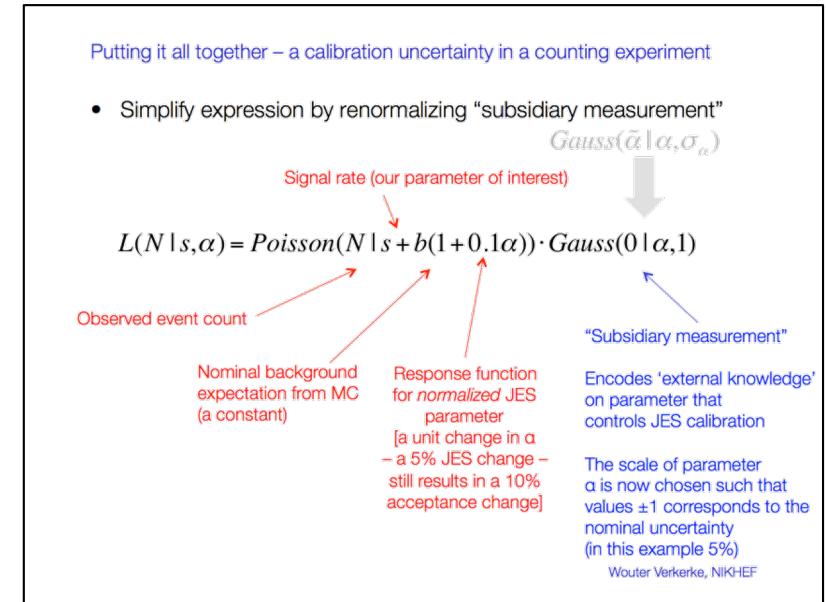
Wouter Verkerke, Nikhef

Wouter Verkerke, Nikhef

Example 1: counting expt

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

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



```
// Subsidiary measurement of alpha
w.factory("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: unbinned L with syst.

- Will now demonstrate how to code complete example of the unbinned profile likelihood of Section 5:

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

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

// Response function m(alpha)
w.factory("expr::m_a(\"m*(1+2alpha)\",m[91,80,100],alpha)");

// Signal model
w.factory("Gaussian::sig(x[80,100],m_a,s[1])");

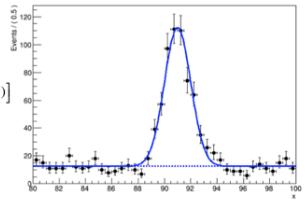
// Complete model Physics(signal plus background)*Subsidiary
w.factory("PROD::model(SUM(mu[0,1]*sig,Uniform::bkg(x)),subs)");
```

Introducing shape systematic uncertainties

- Modeling of systematic uncertainties in Likelihood describing distributions follows the same procedure as for counting models

- Example: Likelihood modeling distribution in a di-lepton invariant mass. POI is the signal strength μ

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



- Consider a lepton energy scale systematic uncertainty that affects this measurement

- The LES has been measured with a 1% precision
- The effect of LES on m_{ll} has been determined to a 2% shift for 1% LES change

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

Response function Subsidiary measurement

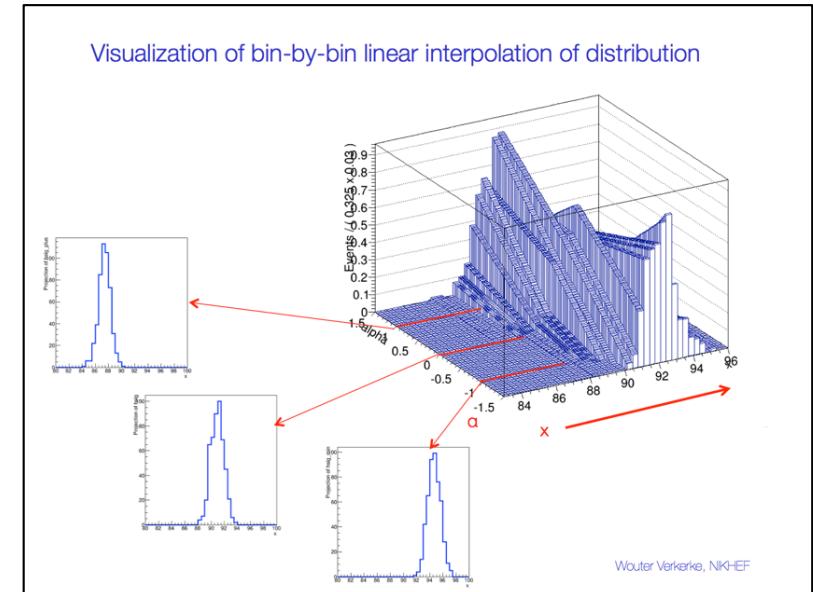
Wouter Verkerke, Nikhef

Example 3 : binned L with syst

- Example of template morphing systematic in a binned likelihood

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

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



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

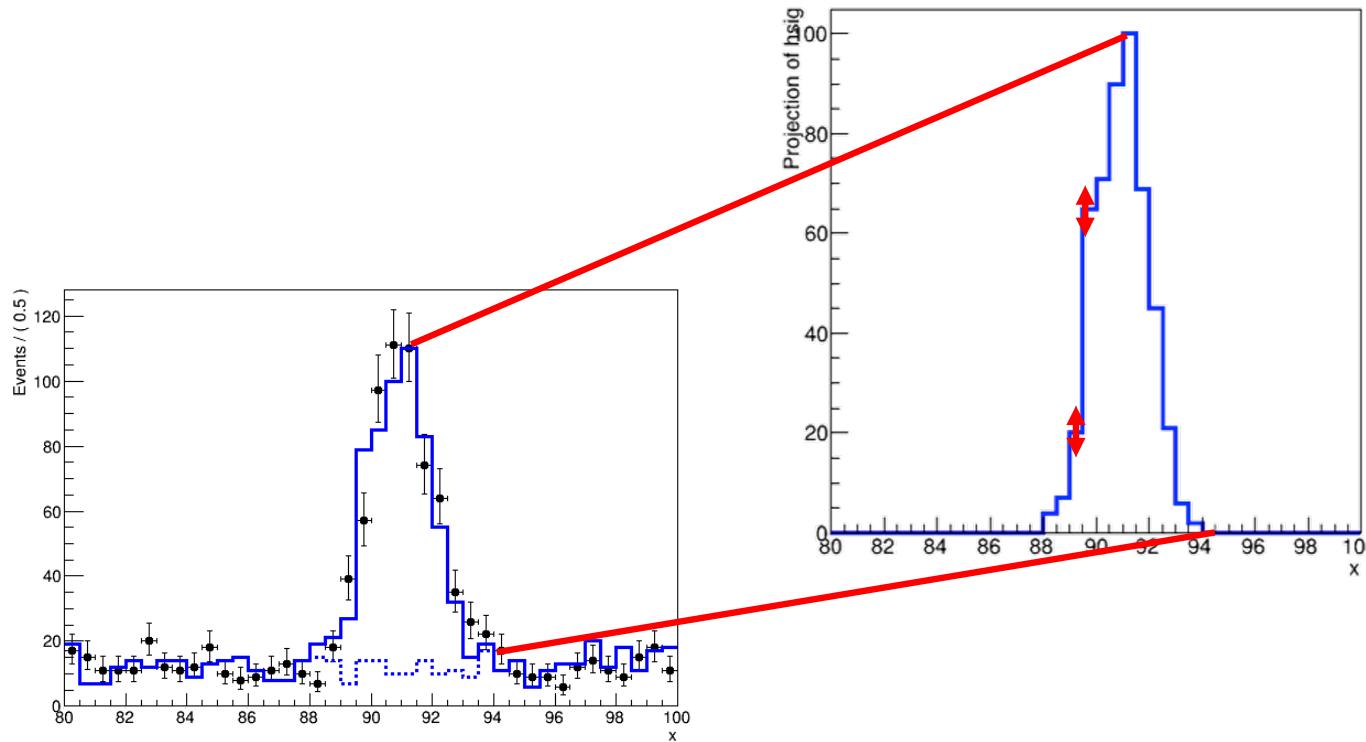
// Construct template models from histograms
w.factory("HistFunc::s_0(x[80,100],hs_0)" );
w.factory("HistFunc::s_p(x,hs_p)" );
w.factory("HistFunc::s_m(x,hs_m)" );

// 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))" );
```

Other uncertainties in MC shapes – finite MC statistics

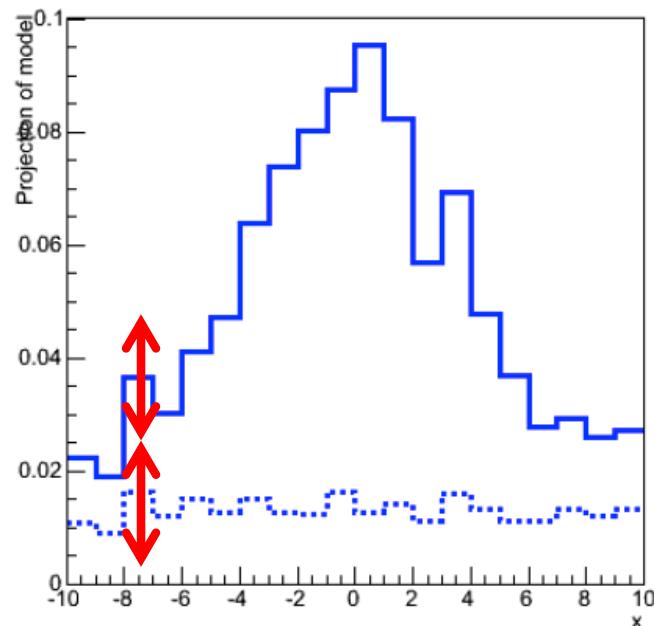
- In practice, MC distributions used for template fits have finite statistics.



- 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



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

Fixed signal, bkg MC prediction

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

Signal, bkg MC nuisance params

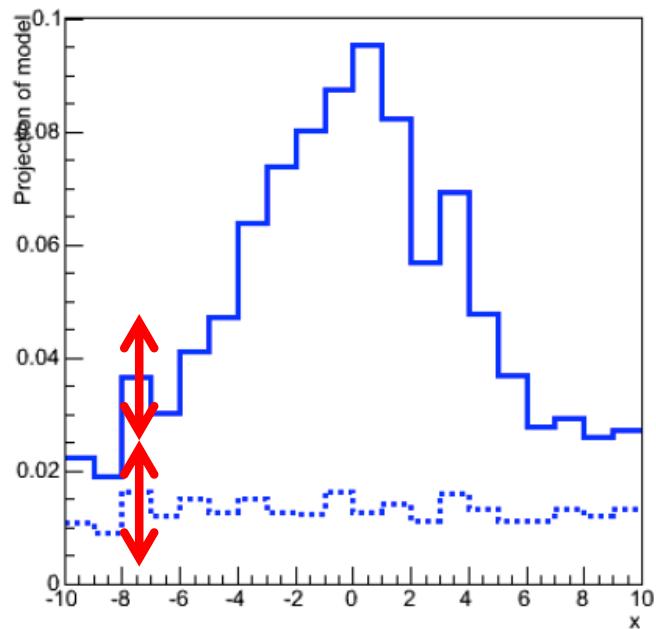
$$L_{bin-i}(\mu, s_i, b_i) = \text{Poisson}(N_i | \mu \cdot s_i + b_i)$$

$\cdot \text{Poisson}(N_i^{MC-s} | s_i)$

$\cdot \text{Poisson}(N_i^{MC-b} | b_i)$

Nuisance parameters for template statistics

- Repeat for all bins



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

Binned likelihood
with rigid template

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

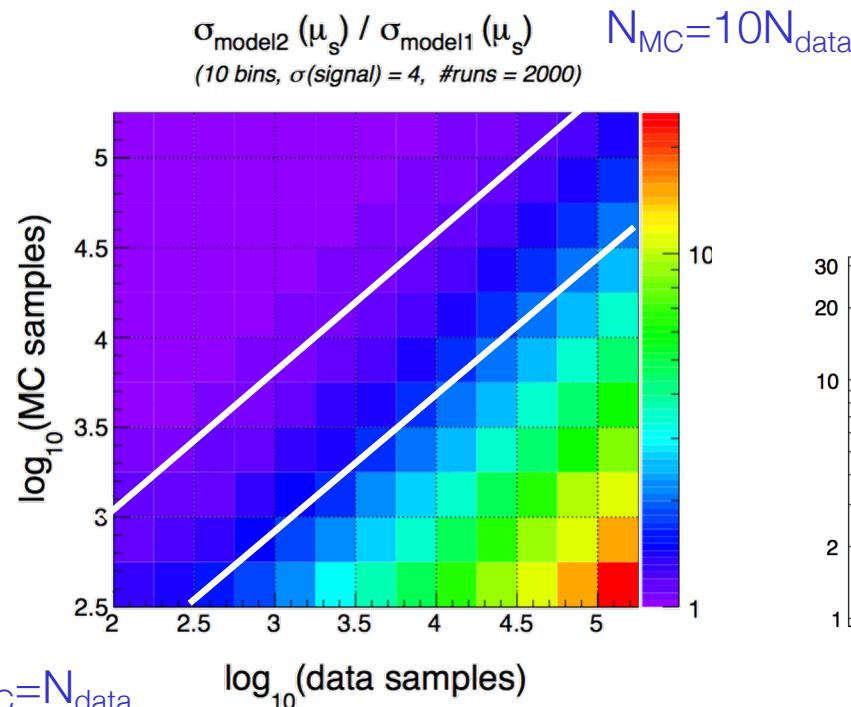
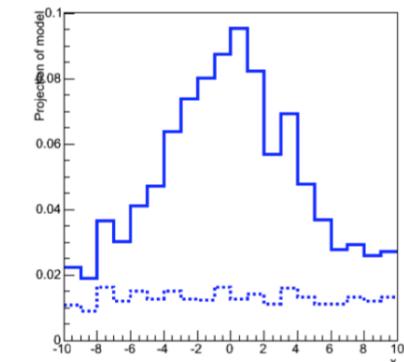
Response function
w.r.t. s, b as parameters

$2 \times N_{\text{bins}}$ subsidiary
measurements
of s, b from $s \sim, b \sim$

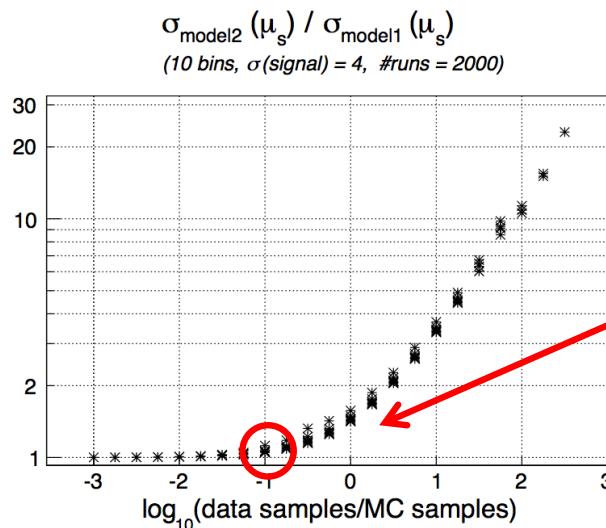
- Result: accurate model for MC statistical uncertainty, but lots of nuisance parameters (#samples \times #bins)...

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 $N_{\text{templ}} < 10 \times N_{\text{data}}$ (from Beeston & Barlow)
- Measurement of effect of template statistics in previously shown toy likelihood model, where POI is the signal yield



'model 1 – plain template likelihood'
'model 2 – Beeston-Barlow likelihood'



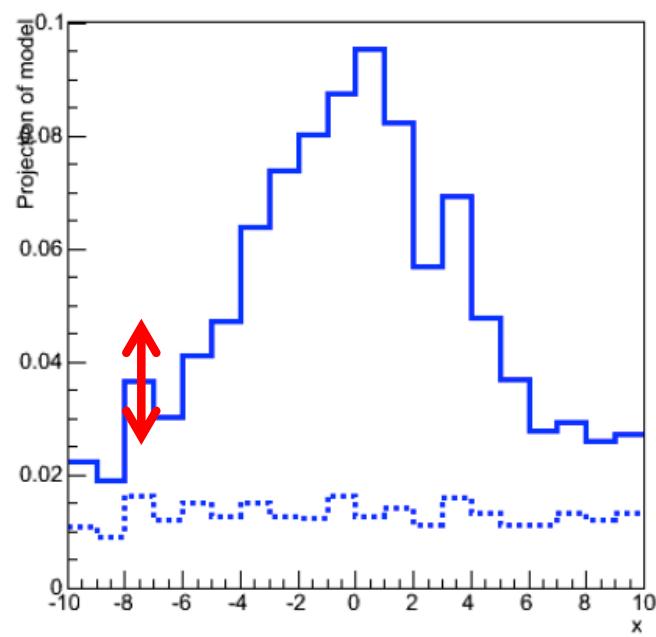
Note that even at $N_{\text{MC}} = 10N_{\text{data}}$ uncertainty on POI can be underestimated by 10% without BB

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

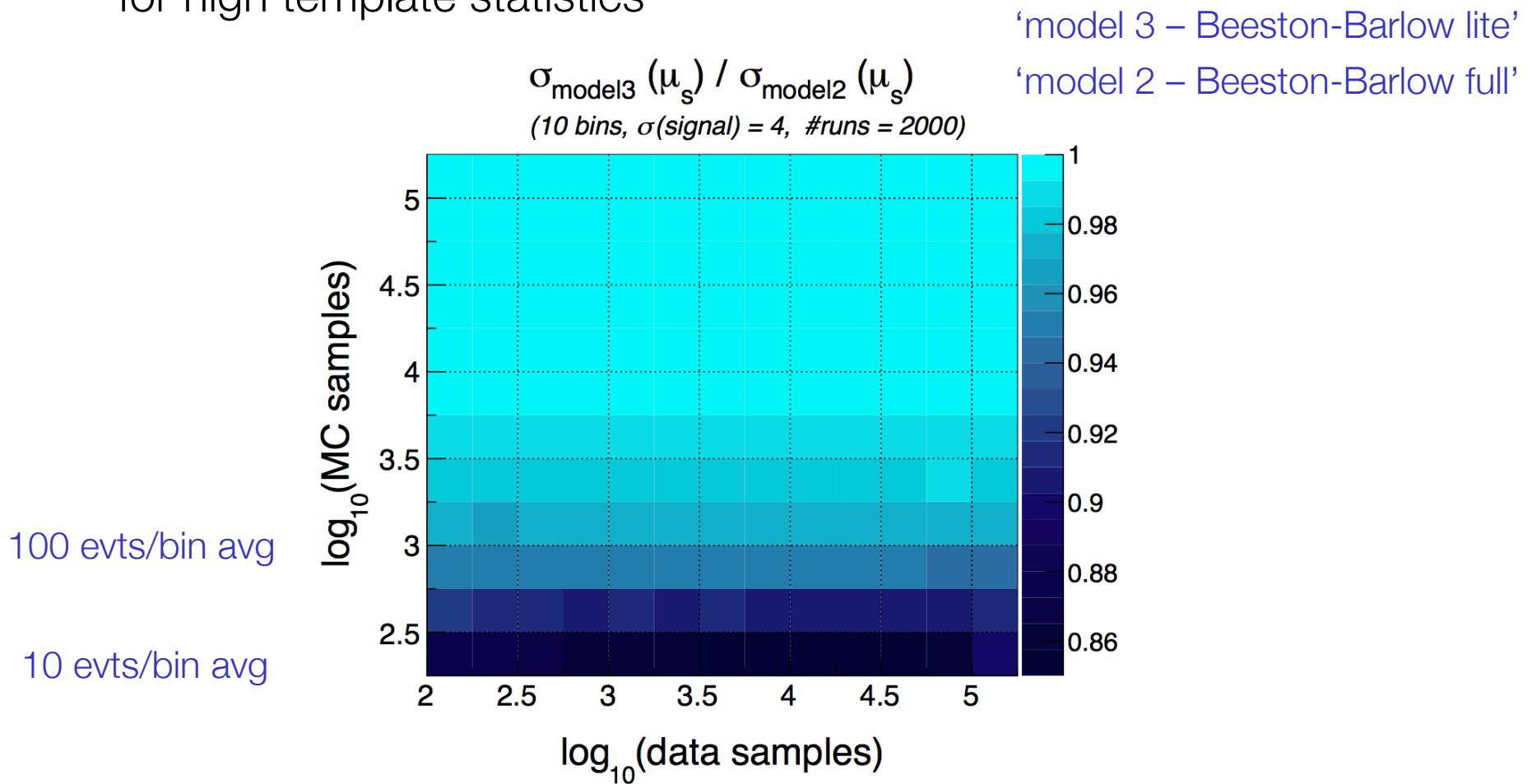
Subsidiary measurements
of n from $s+b$

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

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

The accuracy of the BB-lite approximation

- The Beeston-Barlow ‘lite’ approximation is quite good for high template statistics



- Deviation at low template statistics large due to imperfect modeling of template bins with zero content

The interplay between shape systematics and MC systematics

- Best practice for template morphing models is to also include effect of MC systematics
- Note that for every ‘morphing systematic’ there is a set of two templates that have their own (independent) MC statistical uncertainties.
 - A completely accurate model should model MC stat uncertainties of all templates

$$L(\vec{N} \mid \alpha, \vec{s}^-, \vec{s}^0, \vec{s}^+) = \prod_{bins} P(N_i \mid s_i(\alpha, s_i^-, s_i^0, s_i^+)) \prod_{bins} P(\tilde{s}_i^- \mid s_i^-) \prod_{bins} P(\tilde{s}_i^0 \mid s_i^0) \prod_{bins} P(\tilde{s}_i^+ \mid s_i^+)$$
$$s_i(\alpha, \dots) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$



Morphing response function Subsidiary measurements

- But has severe practical problems
 - Can only be done in ‘full’ Beeston-Barlow model, not in ‘lite’ mode, enormous number of NP models with only a handful of shape systematics...

The interplay between shape systematics and MC systematics

- Commonly chosen practical solution

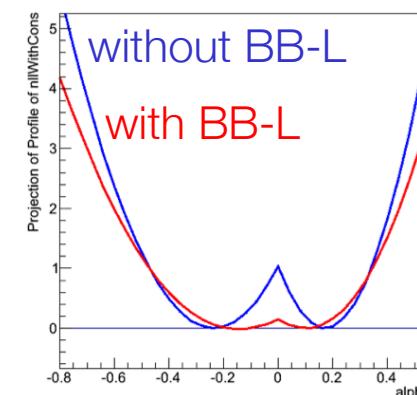
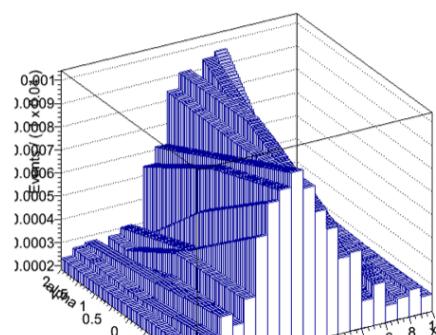
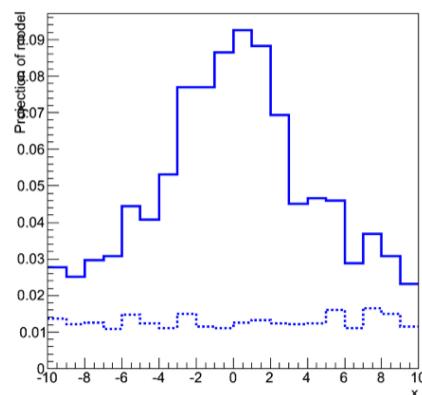
$$L(\vec{N} \mid \vec{s}, \vec{b}) = \prod_{bins} P(N_i \mid \gamma_i \cdot [s_i(\alpha, s_i^-, s_i^0, s_i^+) + b_i]) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i \mid \gamma_i \cdot [\tilde{s}_i + \tilde{b}_i]) G(0 \mid \alpha, 1)$$

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

Morphing & MC response function

Subsidiary measurements

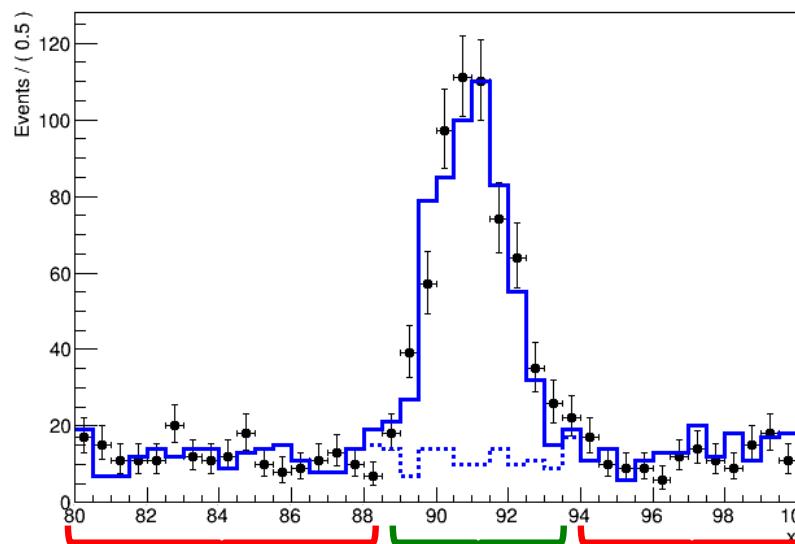
Models relative MC rate uncertainty for each bin w.r.t the nominal MC yield, even if morphed total yield is slightly different



- Approximate MC template statistics already significantly improves influence of MC fluctuations on template morphing
 - Because ML fit can now ‘reweight’ contributions of each bin

Pruning complexity – MC statistical for selected bins

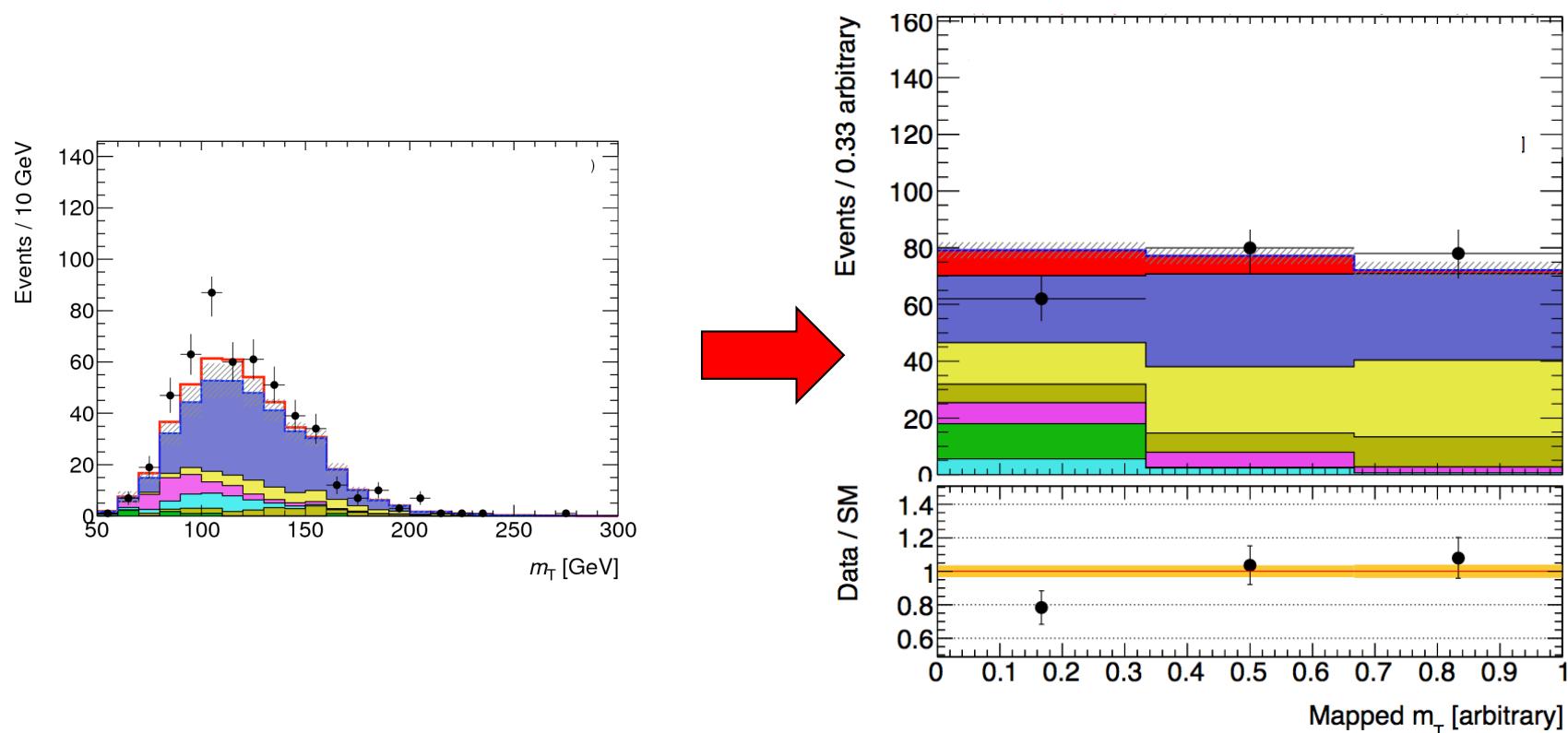
- Can also make decision to model MC statistical uncertainty on a bin-by-bin basis
 - No modeling for high statistics bins
 - Explicit modeling for low-statistics bins



$$L(\vec{N} \mid \vec{\gamma}) = \prod_{bins} P(N_i \mid \gamma_i(\tilde{s}_i + \tilde{b}_i)) \prod_{low-stats\ bins} P(\tilde{s}_i + \tilde{b}_i \mid \gamma_i(\tilde{s}_i + \tilde{b}_i)) \prod_{hi-stats\ bins} \delta(\gamma_i)$$

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)



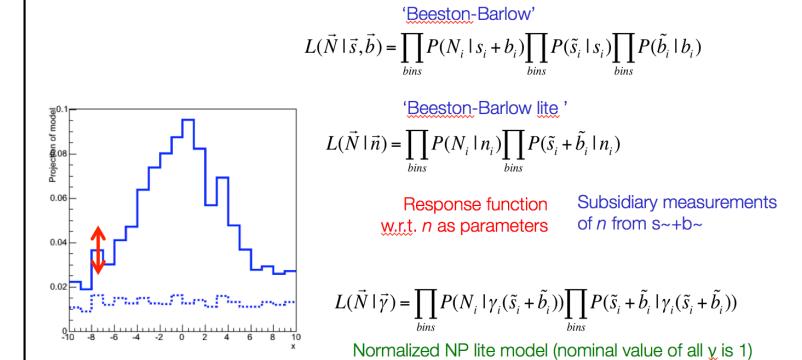
Example 4 – Beeston-Barlow light

- Beeston-Barlow-(lite) modeling of MC statistical uncertainties

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


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



```
// Import template histogram in workspace
w.import(hs) ;

// Construct parametric template models from histograms
// implicitly creates vector of gamma parameters
w.factory("ParamHistFunc::s(hs)" ) ;

// Product of subsidiary measurement
w.factory("HistConstraint::subs(s)" ) ;

// Construct full model
w.factory("PROD::model(s,subs)" ) ;
```

Example 5 – BB-lite + morphing

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

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

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

```
// 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 interplay between shape systematics and MC systematics

- Commonly chosen practical solution

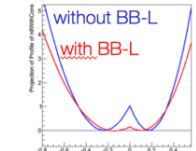
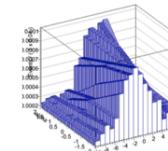
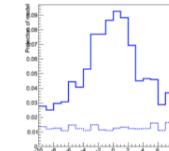
$$s_i(\alpha, \dots) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$

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

Morphing & MC response function

Subsidiary measurements

Models relative MC rate uncertainty for each bin w.r.t. the nominal MC yield, even if morphed total yield is slightly different



- Approximate MC template statistics already significantly improves influence of MC fluctuations on template morphing
 - Because ML fit can now 'reweight' contributions of each bin

Wouter Verkerke, Nikhef

Summary on template morphing and template statistics

- Template morphing allows to model arbitrary responses of shape systematics in template models
 - Various techniques exist, choose carefully, be wary of MC statistical effects that can dominate systematic effect
- Modeling of MC statistical uncertainties important if $N_{MC} < 10 \times N_{data}$
 - Full Beeston-Barlow likelihood most accurate, but leads to enormous number of Nuisance parameters
 - Beeston-Barlow-lite procedures gives very comparable answers if template statistics are sufficient and results in less NPs
 - Modeling of MC statistical uncertainties improves stability of template morphing algorithms

6

Modeling tools: RooFit, RooStats & HistFactory

Coding probability models and likelihood functions

- Discussion on implementation of systematic uncertainties in likelihood models has lead to very complex probability models
- All statistical techniques discussed in Section 2,4 require numeric minimization of likelihood functions. See problem in three parts
 1. Construction of probability models and likelihood functions (always needed)
 2. Minimization of likelihood functions (for parameter estimation, variance estimate, likelihood-ratio intervals)
 3. Construction of test statistics and calculation of their distributions, construction of Neyman constructions on test statistics (p-values, confidence intervals) and calculation (MC(MC)) integrals over Likelihood (Bayesian credible intervals, Bayes factors)
- For step 2 (minimization) the HEP industry standard is MINUIT
- For steps 1, 3 good tools have been developed in the past years, will discuss these now.

RooFit, RooStats and HistFactory

RooFit

Language for building probability models

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

W. Verkerke & D. Kirkby
(exists since 1999)

Will briefly sketch workings of RooStats

Will cover RooFit and HistFactory in a bit more detail since they relate to model building – the key topic of this course

HistFactory

Language to simplify construction of RooFit models of a particular type: binned likelihood template (morphing) models

K. Cranmer, A. Shibata, G. Lewis, L. Moneta, W. Verkerke
(exists since 2010)

RooStats

Suite of statistical tests operating on RooFit probability models

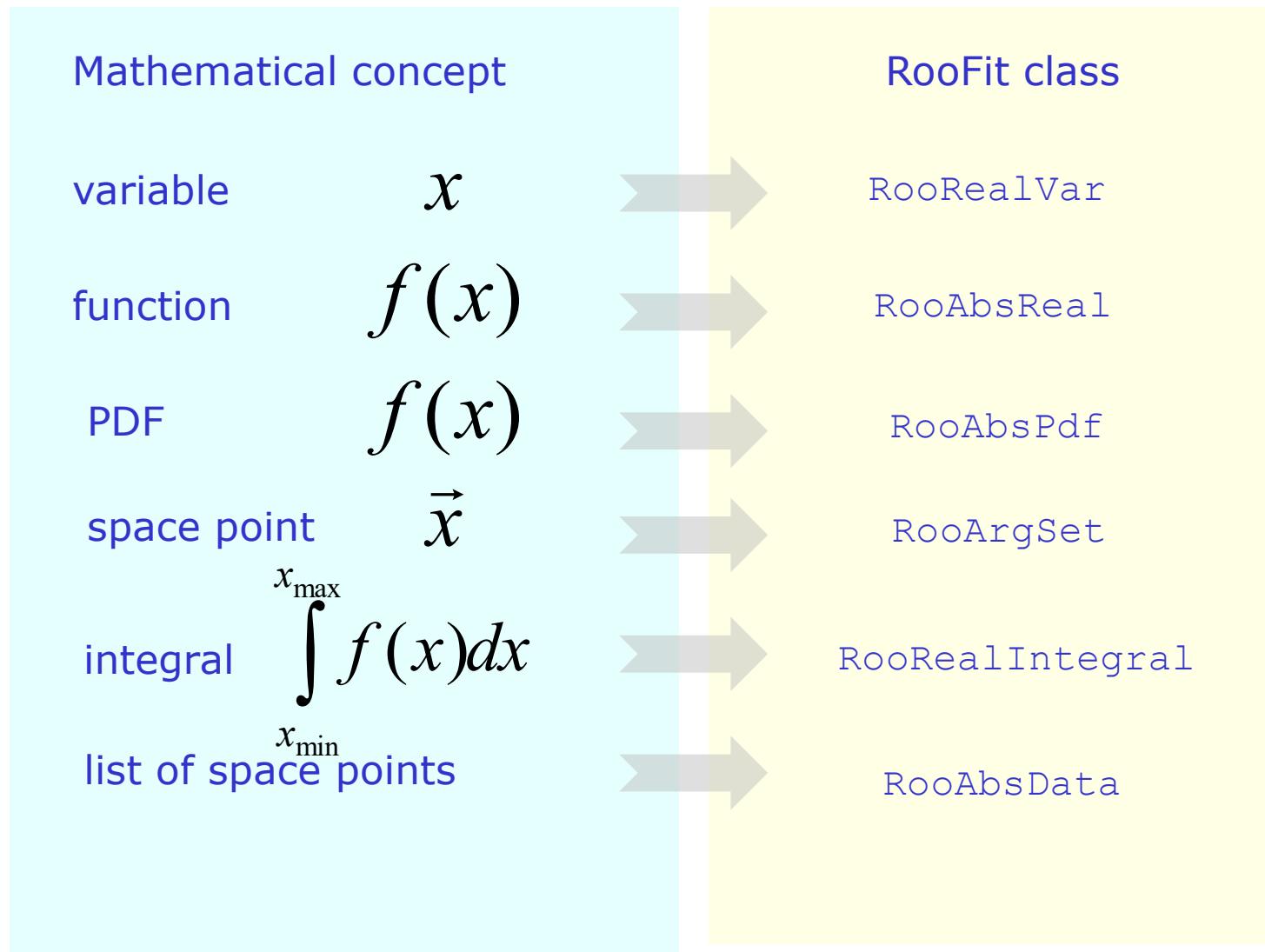
K. Cranmer, G. Schott,
L. Moneta, W. Verkerke
(exists since 2008)

HistFitter
is a tool that configures HistFactory and RooStats in a consistent way

Wouter Verkerke, NIKHEF

RooFit core design philosophy

- Mathematical objects are represented as C++ objects



RooFit core design philosophy - Workspace

- The workspace serves a container class for all objects created

Math	$\text{Gauss}(x,\mu,\sigma)$
RooFit diagram	<pre> graph TD g[RooGaussian g] <--> x[RooRealVar x] g --> y[RooRealVar y] g --> z[RooRealVar z] </pre>
RooFit code	<pre> 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) ; </pre>

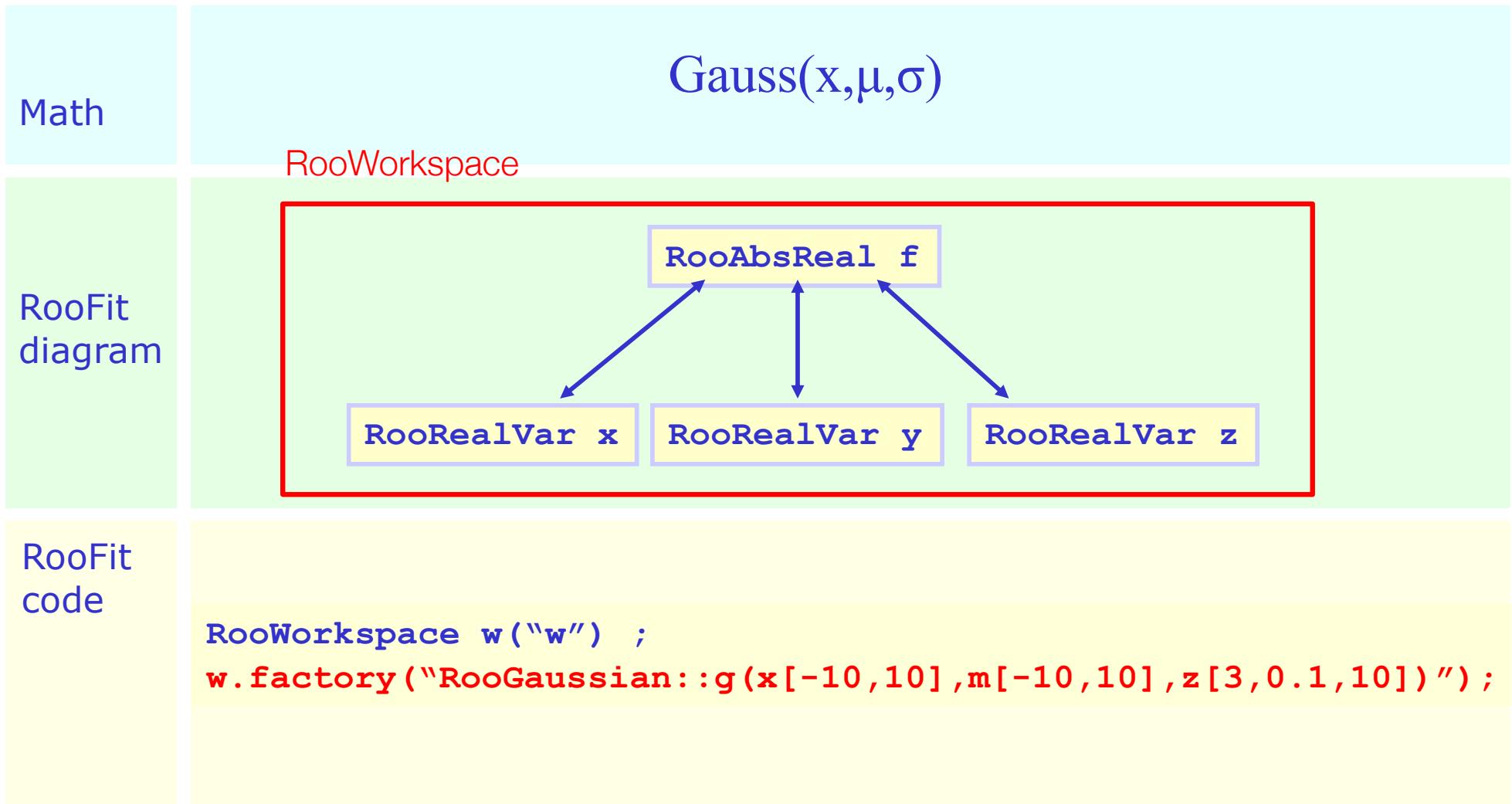
RooFit core design philosophy - Workspace

- The workspace serves a container class for all objects created

Math	$\text{Gauss}(x,\mu,\sigma)$
RooFit diagram	<p>RooWorkspace</p> <pre> graph TD x[RooRealVar x] --> g[RooGaussian g] y[RooRealVar y] --> g z[RooRealVar z] --> g </pre> <p>A diagram illustrating the RooFit workspace structure. A red box labeled "RooWorkspace" contains three blue boxes: "RooRealVar x", "RooRealVar y", and "RooRealVar z". Above the workspace, the mathematical expression $\text{Gauss}(x,\mu,\sigma)$ is shown. Arrows point from each of the three workspace variables to the Gaussian object.</p>
RooFit code	<pre> 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) ; </pre>

Populating a workspace the easy way – “the factory”

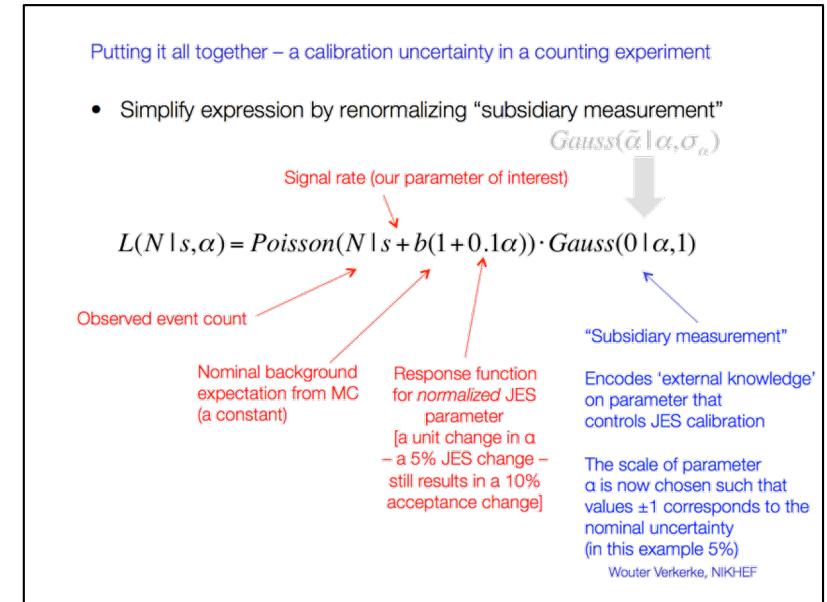
- The **factory** allows to fill a workspace with pdfs and variables using a simplified scripting language



Example 1: counting expt

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

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



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

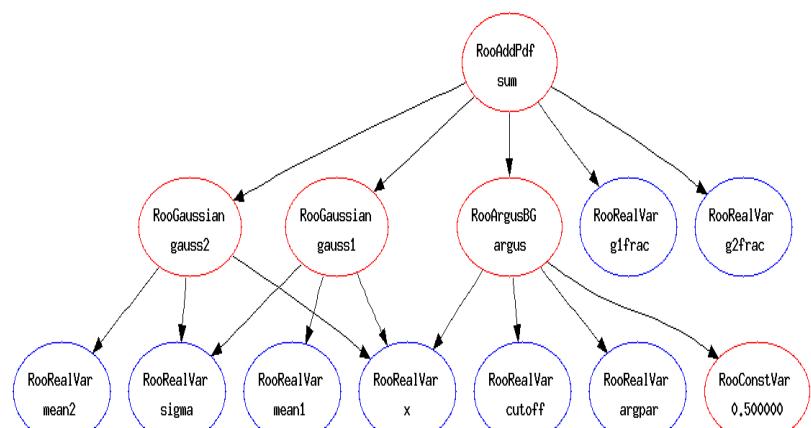
// Response function mu(alpha)
w.factory("expr::mu('s+b(1+0.1*alpha)',s[20],b[20],alpha)" );

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

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

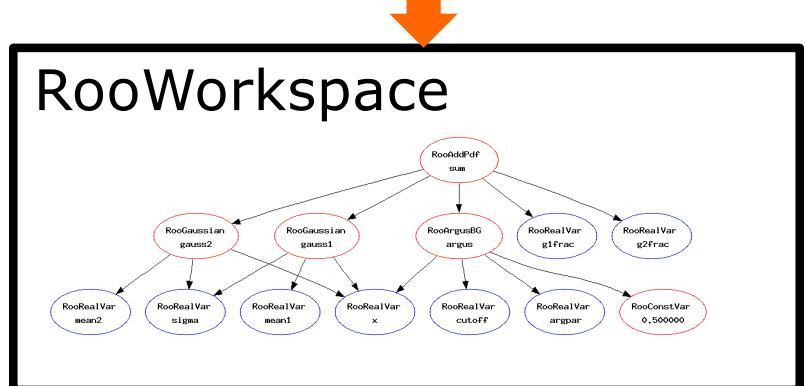
The workspace

- 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.



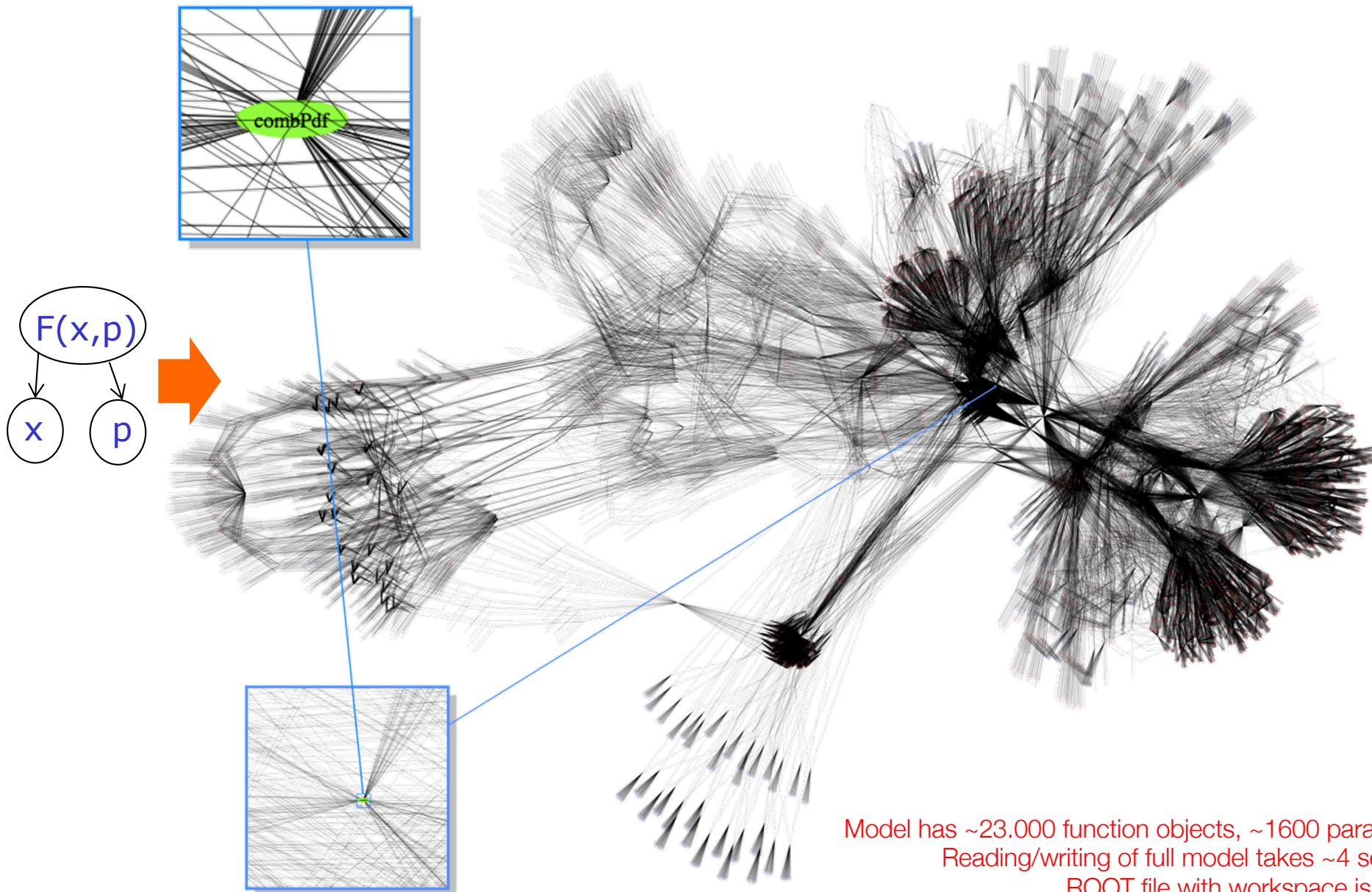
```
RooWorkspace w("w") ;  
w.import(sum) ;  
w.writeToFile("model.root") ;
```

model.root



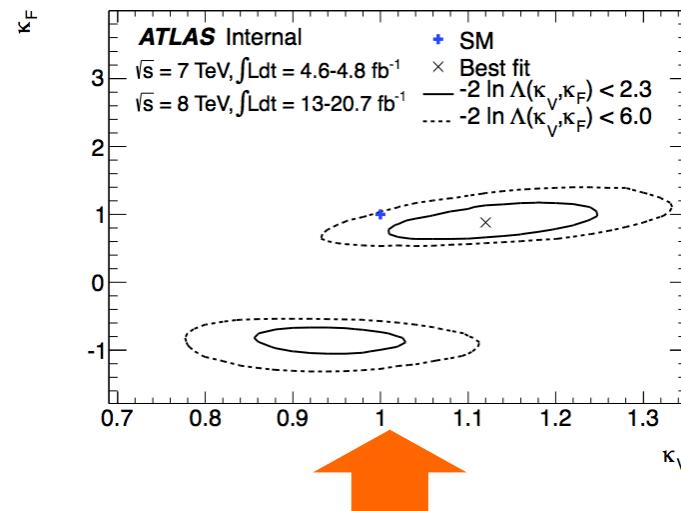
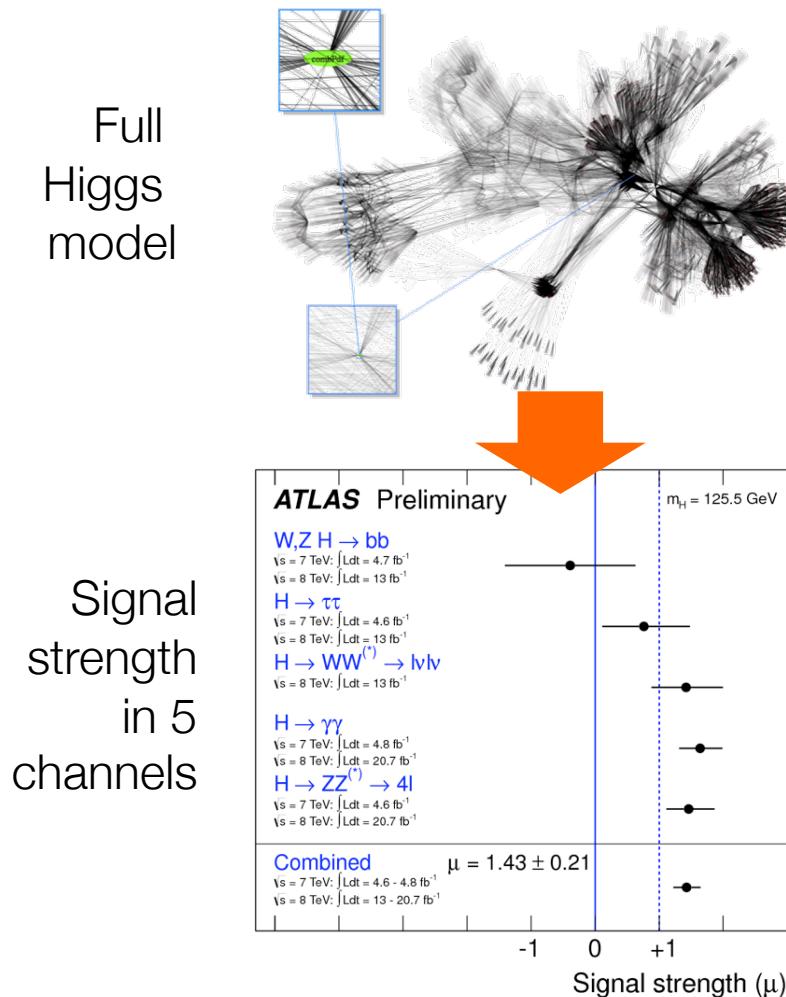
The full ATLAS Higgs combination in a single workspace...

Atlas Higgs combination model (23.000 functions, 1600 parameters)



Collaborative analyses with workspaces

- Workspaces allow to share and modify very complex analyses with very little *technical* knowledge required
- Example: Higgs coupling fits



Confidence intervals on Higgs fermion, ν-boson couplings

Reparam in terms of fermion, ν-boson scale factors

Collaborative analyses with workspaces

- How can you reparametrize existing Higgs likelihoods *in practice*?
- Write functions expressions corresponding to new parameterization

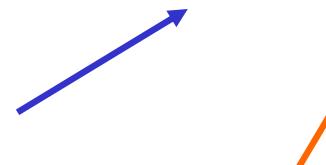
$$\sigma(gg \rightarrow H) * \text{BR}(H \rightarrow \gamma\gamma) \sim \frac{\kappa_F^2 \cdot \kappa_\gamma^2(\kappa_F, \kappa_V)}{0.75 \cdot \kappa_F^2 + 0.25 \cdot \kappa_V^2}$$

```
RooFormulaVar mu_gg_func("mu_gg_func",
                           "(KF2*Kg2)/(0.75*KF2+0.25*KV2)" ,
                           KF2,Kg2,KV2) ;
```

- Edit existing model

```
w.import(mu_gg_func) ;
w.factory("EDIT::newmodel(model,mu_gg=mu_gg_func)") ;
```

*Top node of **modified**
Higgs combination pdf*



*Top node of **original**
Higgs combination pdf*



Modification prescription
replace parameter *mu_gg*
with function *mu_gg_func*
everywhere

7

Diagnostics I: MINUIT, Fit stability & convergence

MINUIT and convergence of profile likelihood fits

- Likelihoods with systematics modeling ('profile likelihood fits') tend to be more complex than 'normal' fits
- Sometimes these likelihood can have pathological features that frustrate the minimization process
- To help you understand I will briefly cover
 - How MINUIT works and defines 'convergence'
 - Typical problems that occur in profile likelihood models and how these affect MINUIT

MINUIT in a nutshell

- MINUIT is a function minimization and analysis packages written by Fred James
 - Original FORTRAN version more than 40 years old!
 - Currently two versions in C++ in ROOT: TMinuit and Minuit2. Former is a ‘machine translated version’ from FORTRAN, latter hand-ported version under the supervision of Fred James
 - I recommend to always use Minuit2 – performance has been exhaustively validated against the original minuit and you get much more useful diagnostic information out of it.
- Three analysis routines implement main functionality
 - **MIGRAD**: Function minimization using the *variable metric method* developed by Fletcher Davidon and Powell. (This is effectively equivalent to the ‘industry standard’ method of Broyden, Fletcher, Goldfarb and Shanno ‘BFGS’)
 - **HESSE**: Error analysis: Calculates Hessian matrix of 2nd derivatives and inverts this into the covariance matrix
 - **MINOS**: Calculates intervals based on the profile likelihood ratio

Function minimization using the variable metric method

- MINUIT does *not* implement a simple ‘steepest descent’ method as plain gradient often does not point well in direction of minimum

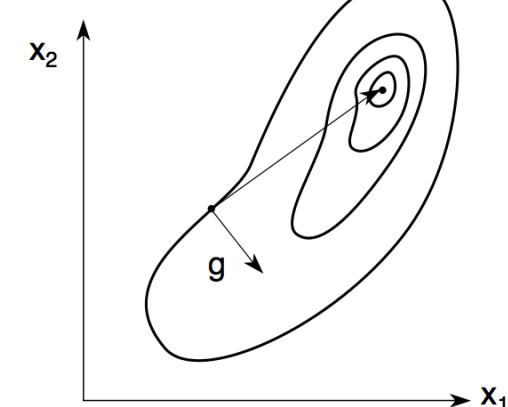
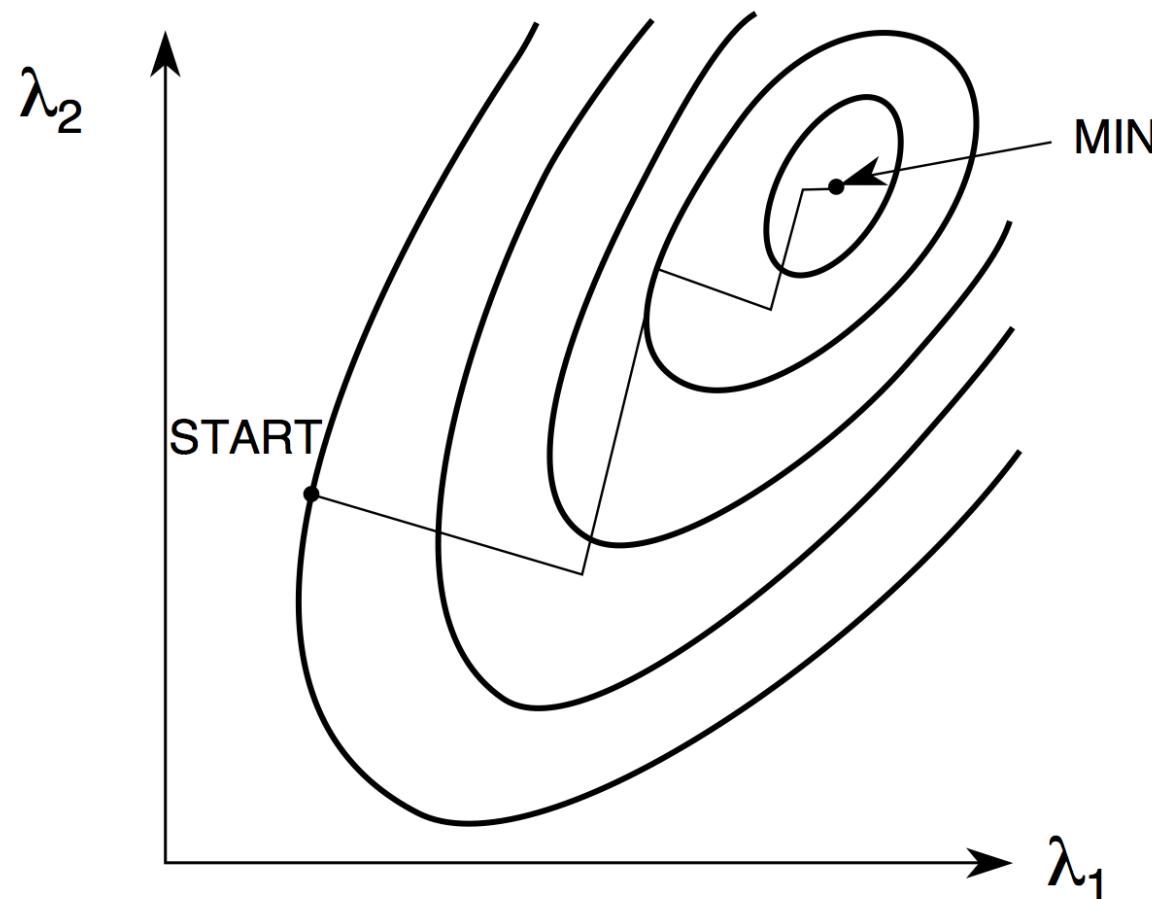
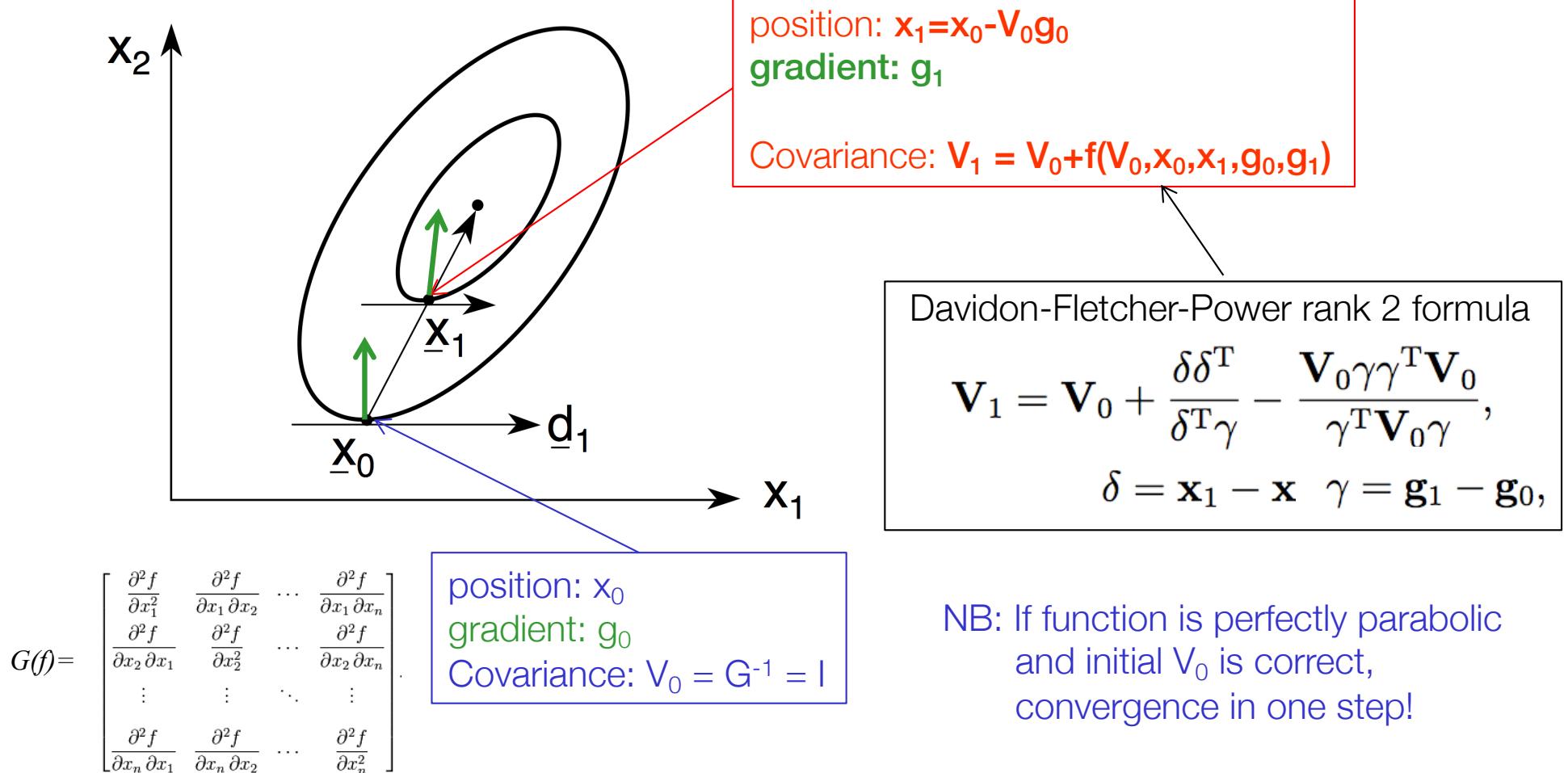


Fig. 9

Function minimization using the variable metric method

- Instead concept of ‘conjugate gradients’ that exploit knowledge of covariance information



Function minimization using the variable metric method

- Convergence criteria is based on ‘estimated distance to minimum’
 - EDM ‘estimated vertical distance to minimum’ assuming parabolic function

$$2 \cdot \text{EDM} = \rho = g^T V g$$

- NB: Derives from general distance metric in non-Euclidian space

$$\Delta s^2 = \Delta x^T A \Delta x$$

Covariant metric tensor

- Note that both minimization and convergence criteria depend on knowledge of covariance matrix
- There are 2 ways to calculate V

1. From the Davidon-Fletcher-Power formula

$$\mathbf{V}_1 = \mathbf{V}_0 + \frac{\delta \delta^T}{\delta^T \gamma} - \frac{\mathbf{V}_0 \gamma \gamma^T \mathbf{V}_0}{\gamma^T \mathbf{V}_0 \gamma},$$

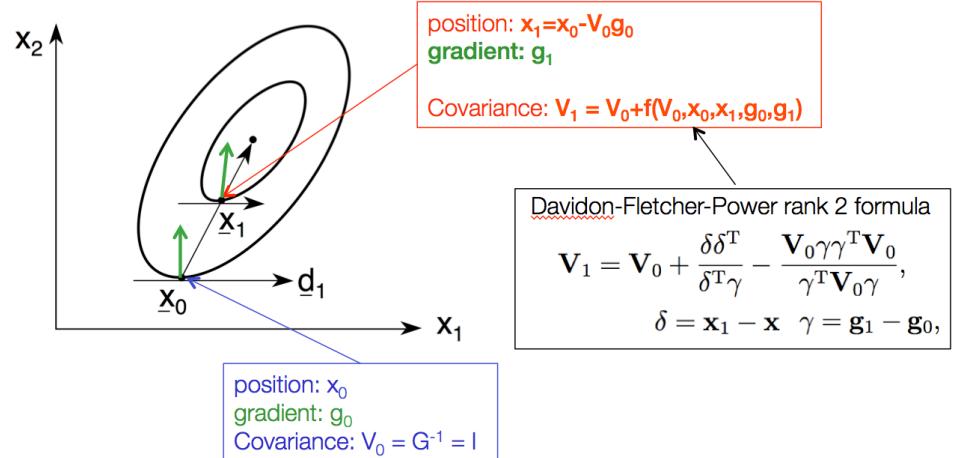
2. From the inversion of the Hessian matrix

$$\mathbf{V} = \mathbf{G}^{-1}$$

Calculation of Hessian is expensive
($\frac{1}{2}N^2$ likelihood evaluations)

MINUIT convergence

- After every VariableMetric step calculate $\text{EDM} = \frac{1}{2}\mathbf{g}^T\mathbf{V}\mathbf{g}$



```

VariableMetric: start iterating until Edm is < 0.001
VariableMetric: Initial state   - FCN = -289.1204081677 Edm =      46.0713 NCalls =    1826
VariableMetric: Iteration # 1 - FCN = -299.3073097602 Edm =      9.18415 NCalls =   2226
VariableMetric: Iteration # 2 - FCN = -304.9468725143 Edm =      2.22698 NCalls =   2624
VariableMetric: Iteration # 3 - FCN = -306.3323972775 Edm =      1.43793 NCalls =   3016
VariableMetric: Iteration # 4 - FCN = -307.199970017 Edm =      0.615574 NCalls =   3410
VariableMetric: Iteration # 5 - FCN = -307.6493784582 Edm =      0.352904 NCalls =   3804
VariableMetric: Iteration # 6 - FCN = -307.8960954798 Edm =      0.0749124 NCalls =   4196
VariableMetric: Iteration # 7 - FCN = -307.9549184882 Edm =      0.0498047 NCalls =   4588
VariableMetric: Iteration # 8 - FCN = -308.0068371877 Edm =      0.03473 NCalls =   4980
VariableMetric: Iteration # 9 - FCN = -308.0564661263 Edm =      0.0266955 NCalls =   5372
VariableMetric: Iteration # 10 - FCN = -308.1092267909 Edm =      0.038622 NCalls =   5764
VariableMetric: Iteration # 11 - FCN = -308.1547659161 Edm =      0.0290921 NCalls =   6156
VariableMetric: Iteration # 12 - FCN = -308.1870210082 Edm =      0.00827767 NCalls =   6548
VariableMetric: Iteration # 13 - FCN = -308.2008924182 Edm =      0.0034224 NCalls =   6940
VariableMetric: Iteration # 14 - FCN = -308.2064790118 Edm =      0.00151676 NCalls =   7332
VariableMetric: Iteration # 15 - FCN = -308.2090105175 Edm =      0.00106118 NCalls =   7724
VariableMetric: Iteration # 16 - FCN = -308.2106535849 Edm =      0.000634155 NCalls =   8116

```

- Terminate VM procedure when $\text{EDM} < 0.001$

MINUIT converge

```
VariableMetric: Iteration # 12 - FCN = -308.1870210082 Edm = 0.00827767 NCalls = 6548
VariableMetric: Iteration # 13 - FCN = -308.2008924182 Edm = 0.0034224 NCalls = 6940
VariableMetric: Iteration # 14 - FCN = -308.2064790118 Edm = 0.00151676 NCalls = 7332
VariableMetric: Iteration # 15 - FCN = -308.2090105175 Edm = 0.00106118 NCalls = 7724
VariableMetric: Iteration # 16 - FCN = -308.2106535849 Edm = 0.000634155 NCalls = 8116
```

- (Terminate VM procedure when EDM<0.001)
 - Note that EDM up to here was calculated with \mathbf{V} from DFP updater formula

$$\mathbf{V}_1 = \mathbf{V}_0 + \frac{\delta\delta^T}{\delta^T\gamma} - \frac{\mathbf{V}_0\gamma\gamma^T\mathbf{V}_0}{\gamma^T\mathbf{V}_0\gamma},$$

- From here on, procedure depends on ‘strategy code’
 - Code 0: terminate line search
 - Code 2: Recalculate \mathbf{V} from \mathbf{G}^{-1} (HESSE)
if $\text{EDM(HESSE)} > 0.001$ restart line search, else terminate
 - Code 1: If accuracy of \mathbf{V}_n from DFP better than 5% terminate,
else follow Code 2 procedure
- Strategy 1 is the default.

HESSE Convergence

- For smooth functions covariance estimates from HESSE are generally more accurate than those from Davidon-Fletcher-Powell but matrix inversion step is vulnerable to singularity issues
- Singularities detected with eigenvalue analysis of Hessian matrix G before matrix inversion
 - If ‘smallest eigenvalue’/‘largest eigenvalue’ $< 10^{-6}$ then matrix is declared ‘not positive definite’
 - Note that happens for both negative *and* small eigenvalues
 - In that case an ‘ad-hoc’ term is added to the diagonal of the Hessian matrix to force it positive definite so that it can be inverted
- The ‘adjusted’ V from HESSE is then used to calculate the EDM
 - EDM estimate less reliable in this case, may cause MINUIT to endlessly go back to VariableMetric line search and eventually give up ‘maximum number of calls exceeded’

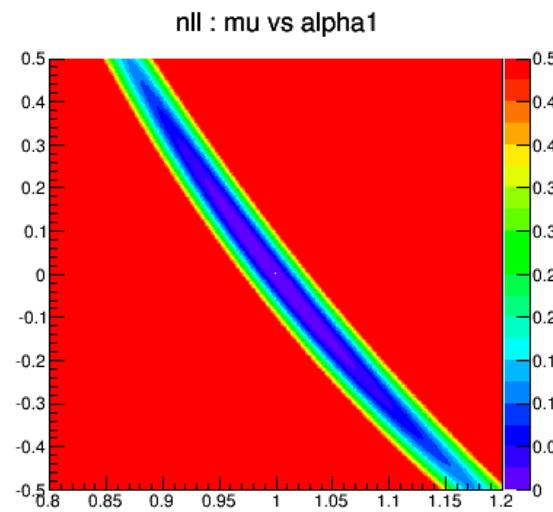
Likelihood models that cause MINUIT problems

- Example 1 – Strong correlations
 - Consider this simple likelihood model with one NP

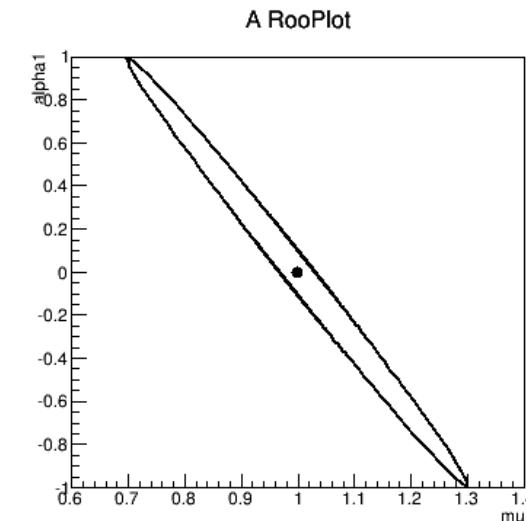
$$L_1(\mu, \alpha) = \text{Poisson}(N | \mu S(1 + \tau\alpha)) \text{Gaussian}(0 | \alpha, 1)$$

- What does the likelihood look like, e.g. for N=1000?

Scan of $-\log L(\mu, \alpha)$



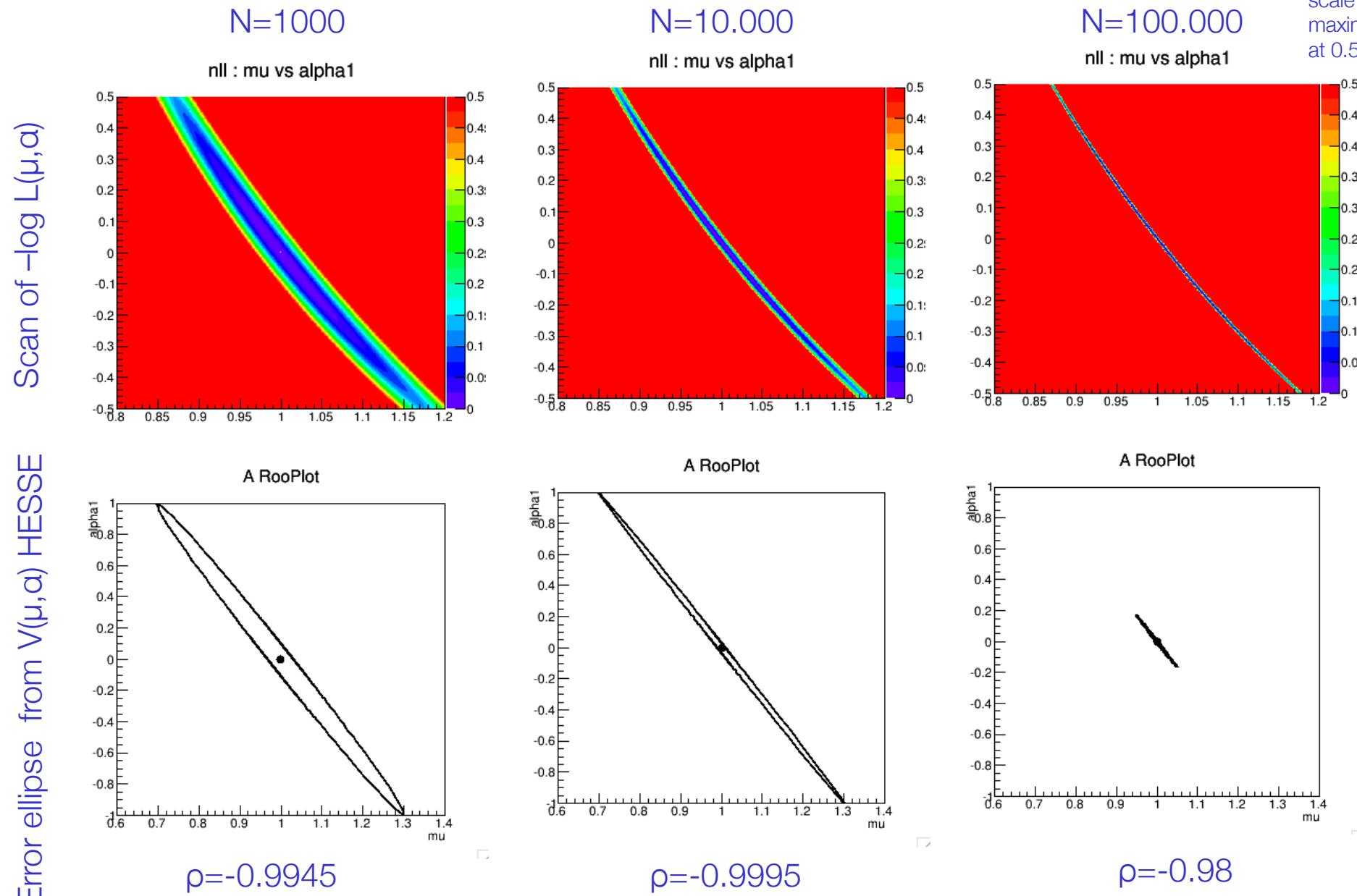
Error ellipse from $V(\mu, \alpha)$ HESSE



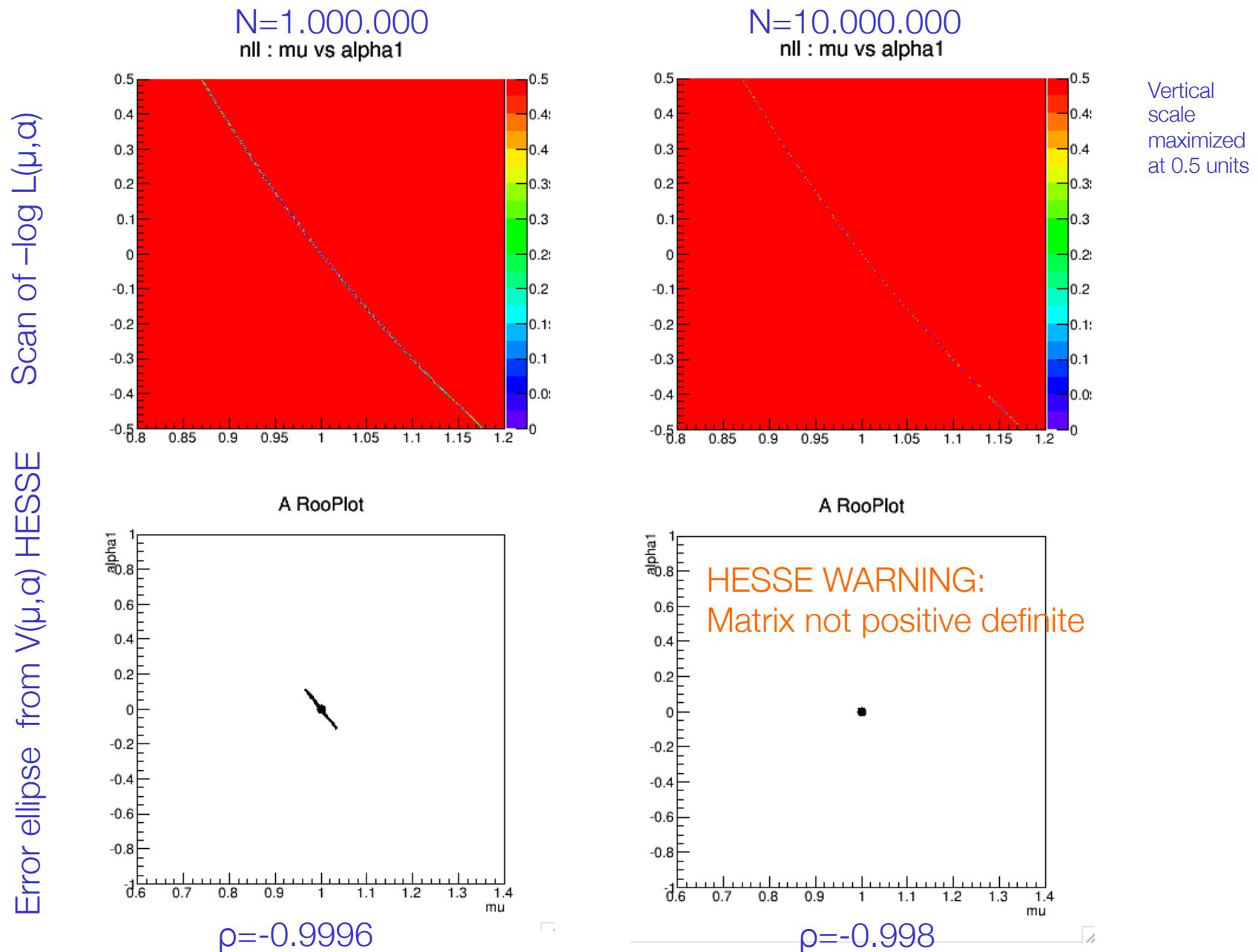
$\rho = 0.9945$

- Strong correlations, but numerically feasible

Increasing the observed event count



Increasing the observed event count



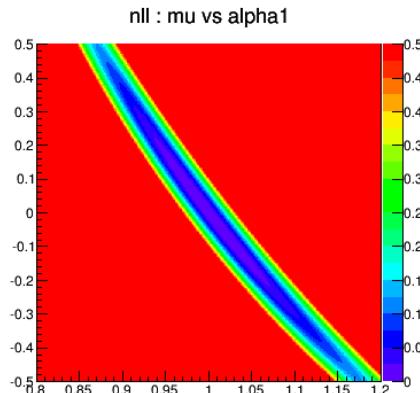
Likelihood models that cause MINUIT problems

- Example 2 – **Hidden** strong correlations
 - Consider this trivial extension of the previous example with 2 NPs

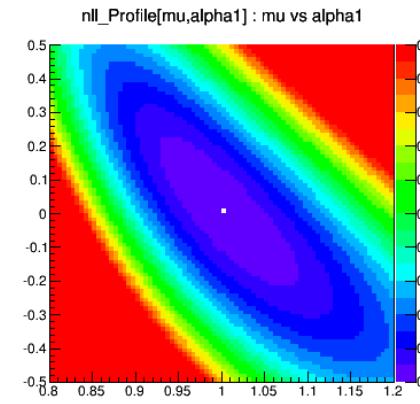
$$L_2(\mu, \alpha_1, \alpha_2) = \text{Poisson}(N \mid \mu S(1 + \tau_1 \alpha_1 + \tau_2 \alpha_2)) \text{Gauss}(0 \mid \alpha_1, 1) \text{Gauss}(0 \mid \alpha_2, 1)$$

- Underlying scenario: two (independent) sources of systematic uncertainty that have a similar effect on the physics measurement
- What does (profile) likelihood look like for various S ?

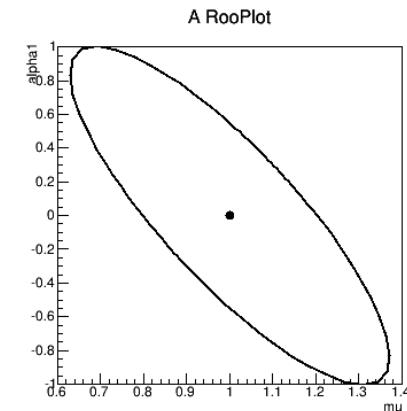
$-\log L(\mu, \alpha_1, \hat{\alpha}_2)$



$-\log L(\mu, \alpha_1, \hat{\alpha}_2(\alpha_1, \mu))$



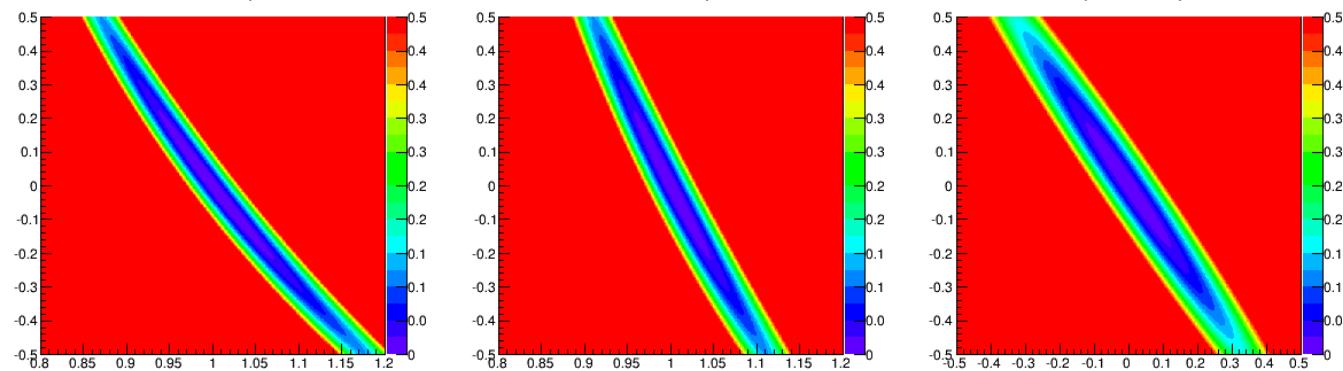
Error ellipse $V(\mu, \alpha)$ HESSE



$-\log L(\mu, \alpha_1, \alpha_2) - 1000$ events

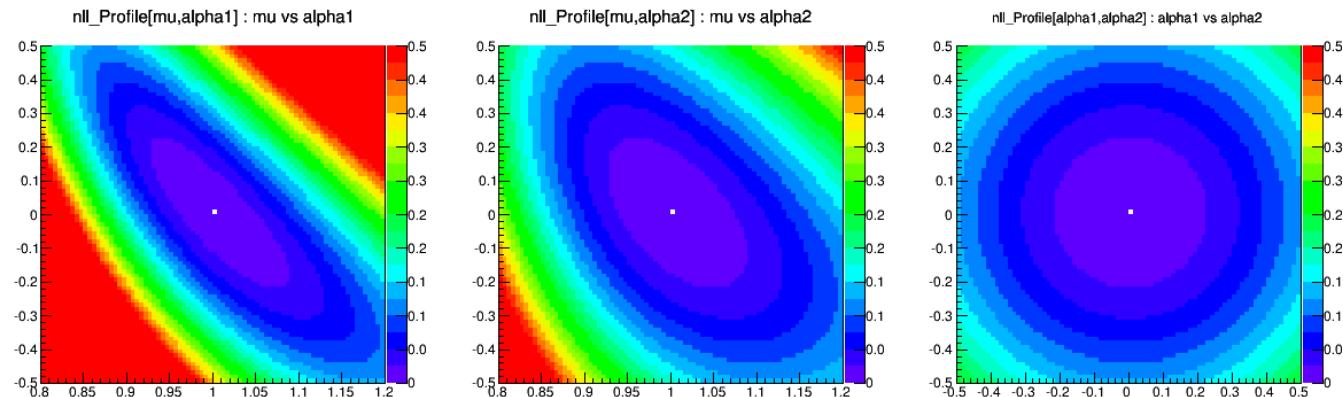
Slice in $-\log L$

$-\log L(a, b, \hat{c})$

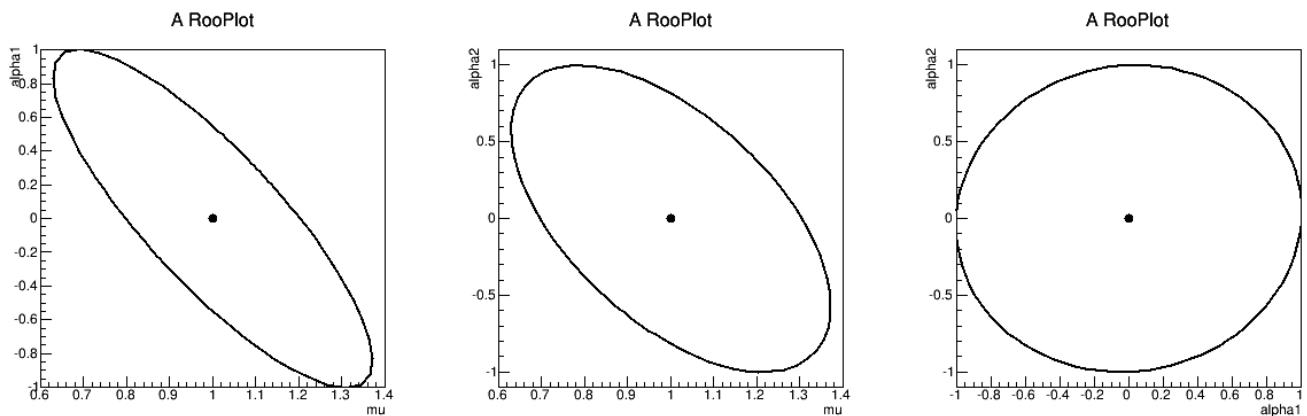


Profile likelihood

$-\log L(a, b, \hat{c}(a, b))$



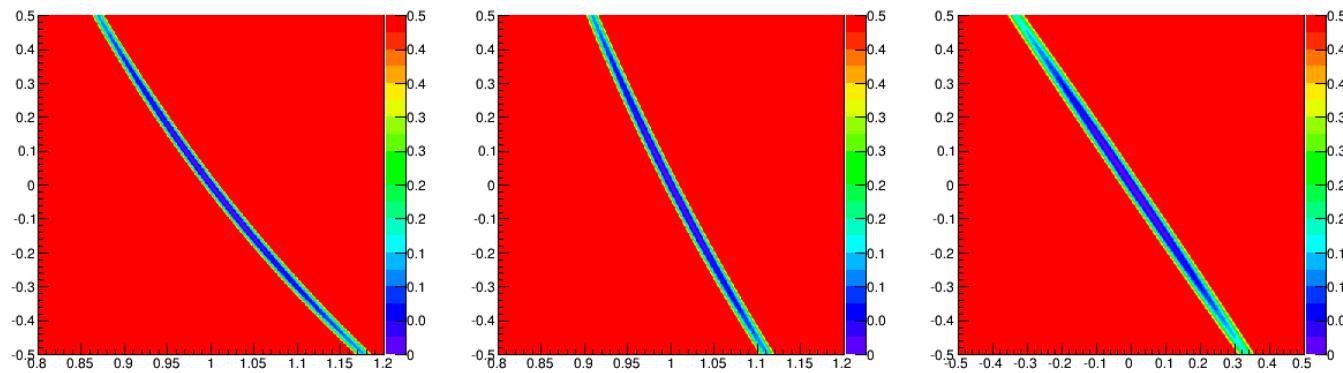
Error ellipse
from HESSE



$-\log L(\mu, \alpha_1, \alpha_2) - 10.000$ events

Slice in $-\log L$

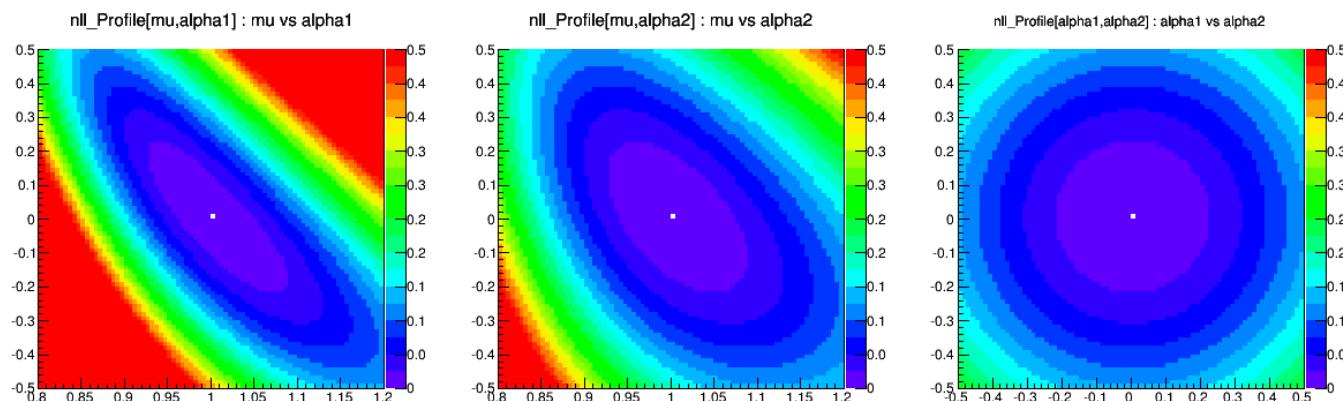
$-\log L(a, b, \hat{c})$



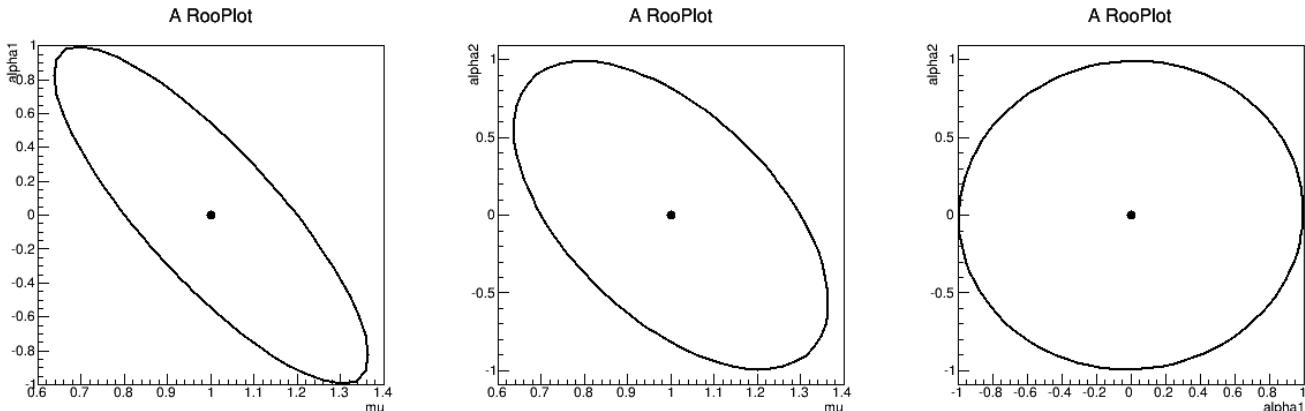
Profile likelihood

$-\log L(a, b, \hat{c}(a, b))$

Note that PLL
contours don't
change between 1K
and 10k!



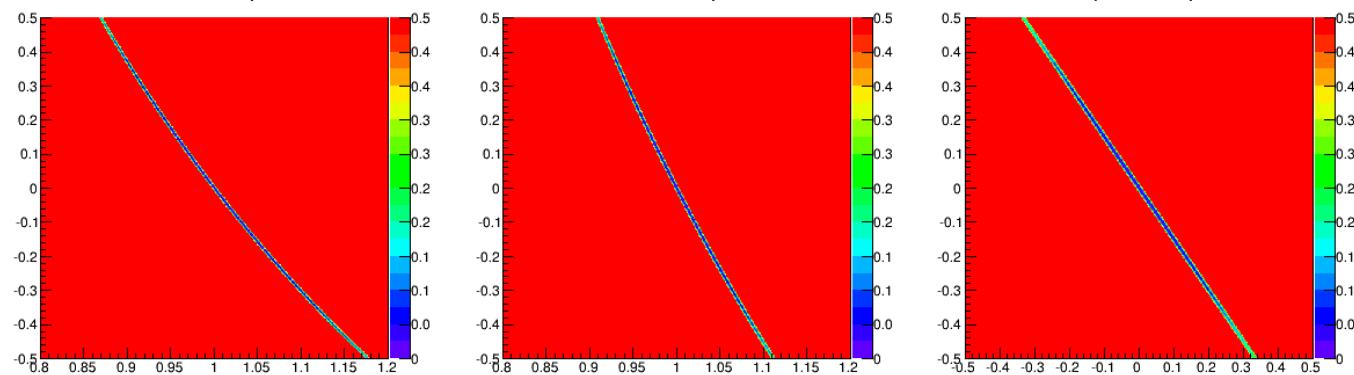
Error ellipse
from HESSE



$-\log L(\mu, \alpha_1, \alpha_2) - 100.000$ events

Slice in $-\log L$

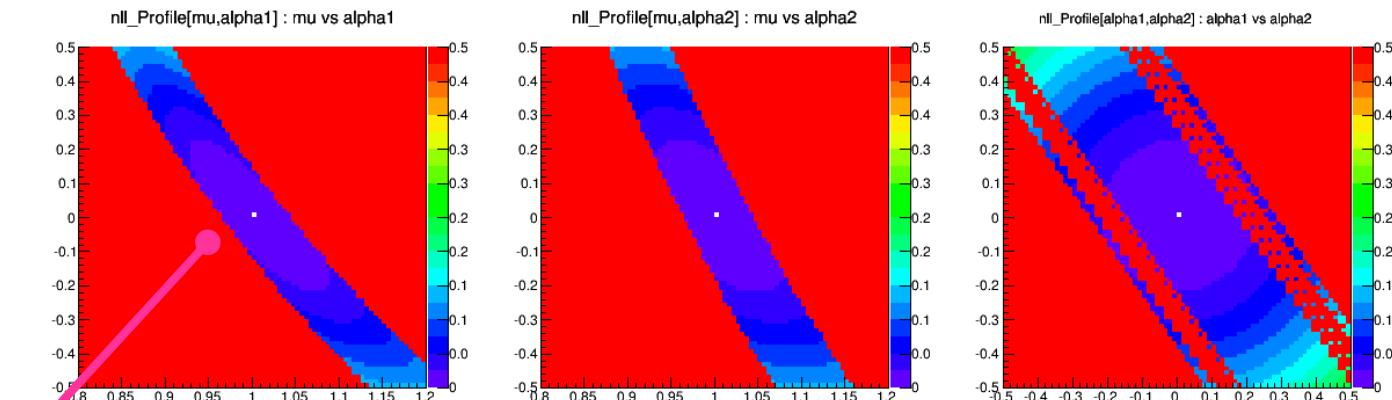
$-\log L(a, b, \hat{c})$



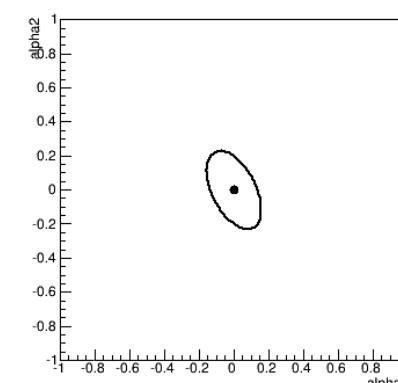
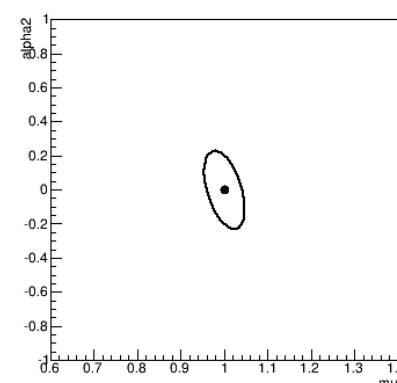
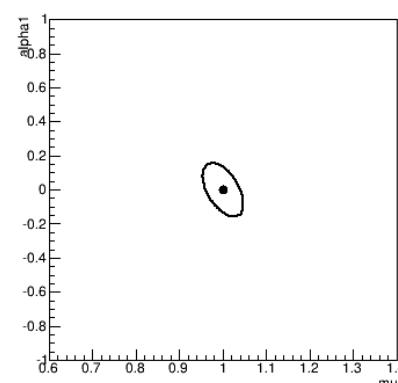
Profile likelihood

$-\log L(a, b, \hat{c}(a, b))$

Note that PLL
contours don't
change between 10K
and 100k close to min.!
(but onset of fit failures
further away...)



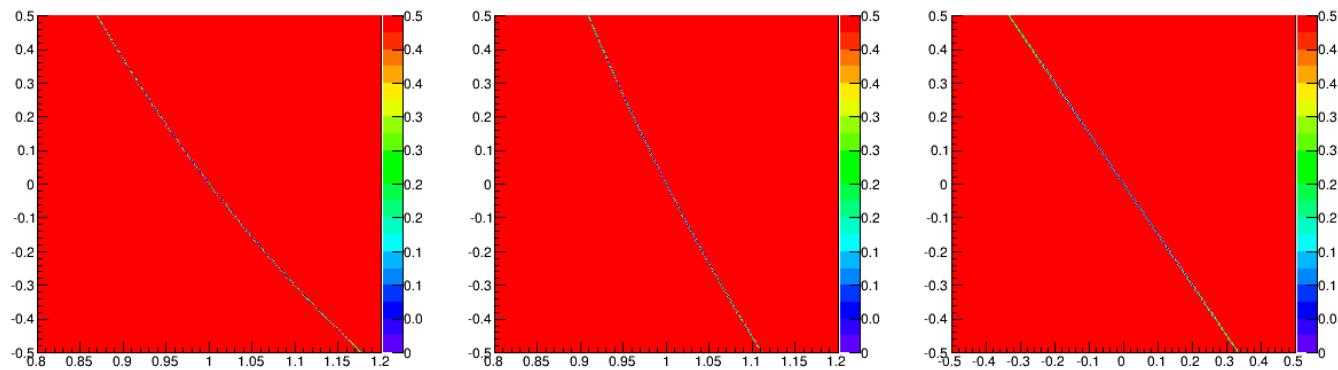
Error ellipse
from HESSE



$-\log L(\mu, \alpha_1, \alpha_2) - 1.000.000$ events

Slice in $-\log L$

$-\log L(a, b, \hat{c})$

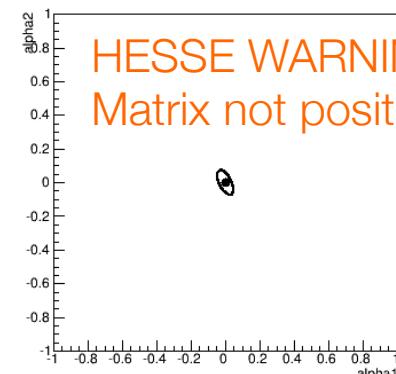
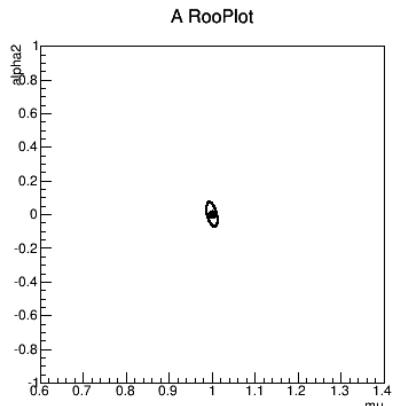
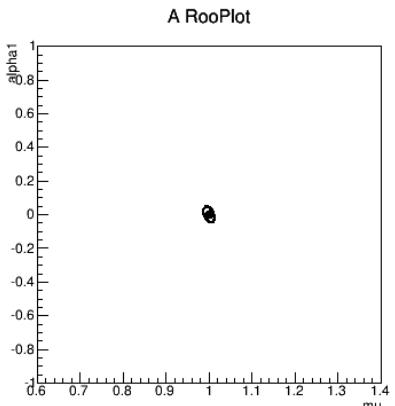
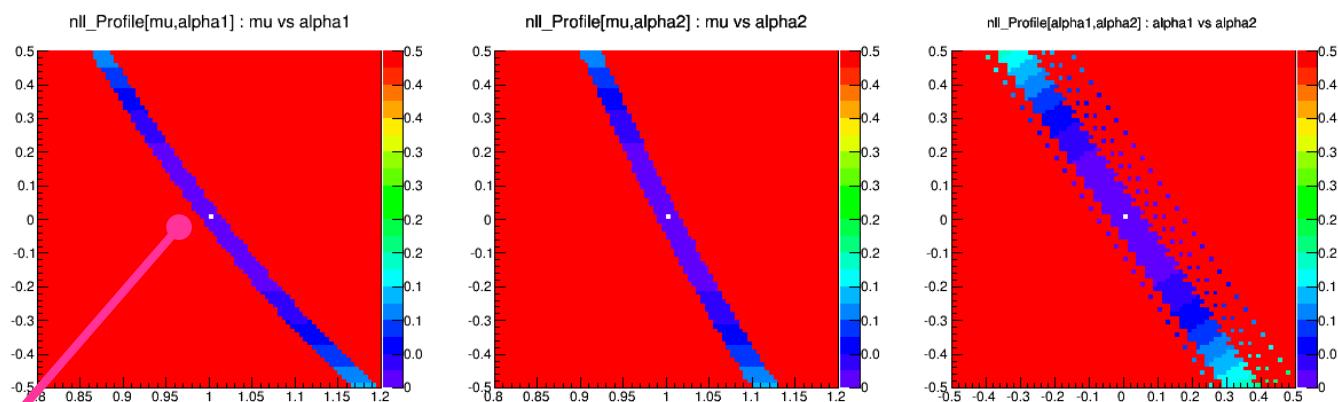


Profile likelihood

$-\log L(a, b, \hat{c}(a, b))$

Note that PLL contours don't change between 100K and 1M close to min.
(but further increase of fit failures further away...)

Error ellipse
from HESSE

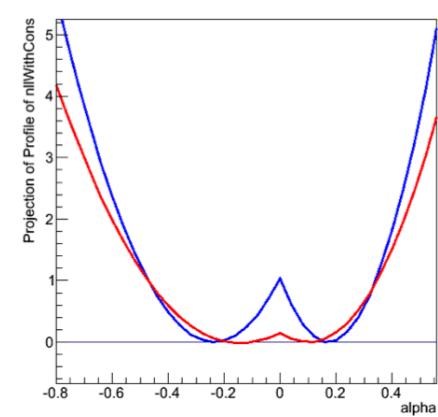
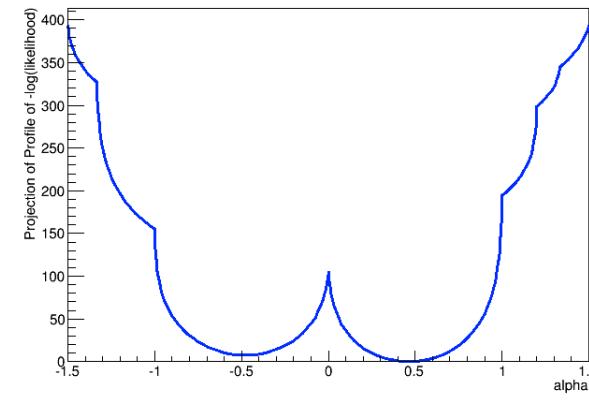
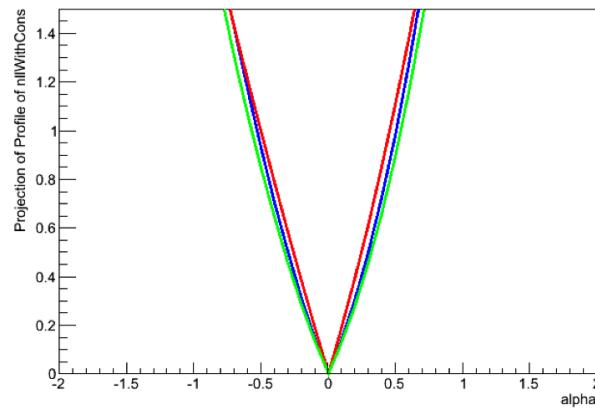


Conclusions on strong correlations

- MINUIT can handle strong correlations very well, but at some point algorithm breaks down
 - Notably HESSE will fail when ratio of weakest-to-strongest eigenvalue $< 10^{-6}$
- Diagnostic of the existence of strong correlations can be difficult
 - In simple models (Ex 1) this is reflected correlation coefficients
 - In more complex models (Ex 2) this may not show at all in the correlation coefficients because strong ‘N-point correlations’ may still project out to modest 2-point correlations (i.e. the usual Pearson correlation coefficients)
 - Better diagnostic tools is eigenvalues of Hessian matrix before inversion, but not (yet) available in Minuit2 [I am discussing this with ROOT team]
- Solution: consider to simplify model:
 - If two NPs represent conceptually distinct systematic uncertainties, but their effect on the likelihood is virtually identical, then there is effectively a redundant degree of freedom. You can eliminate one

Other likelihood pathologies

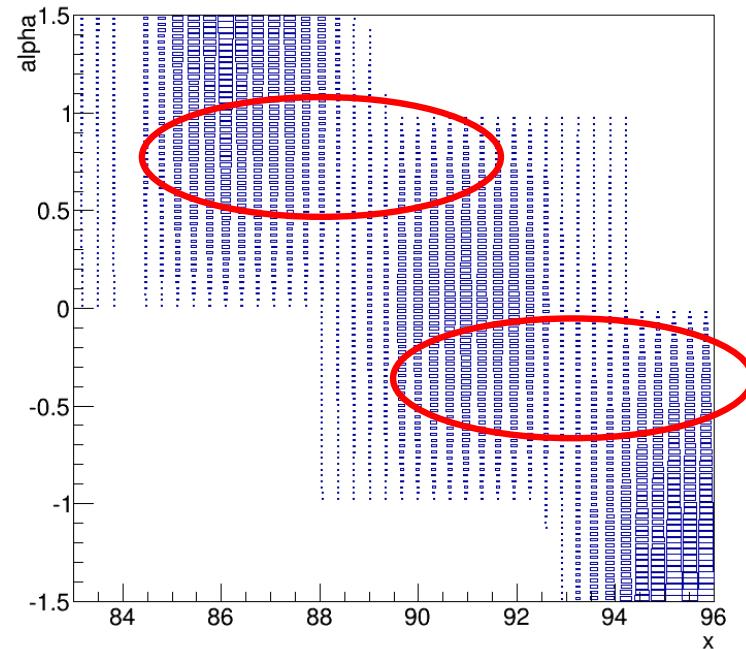
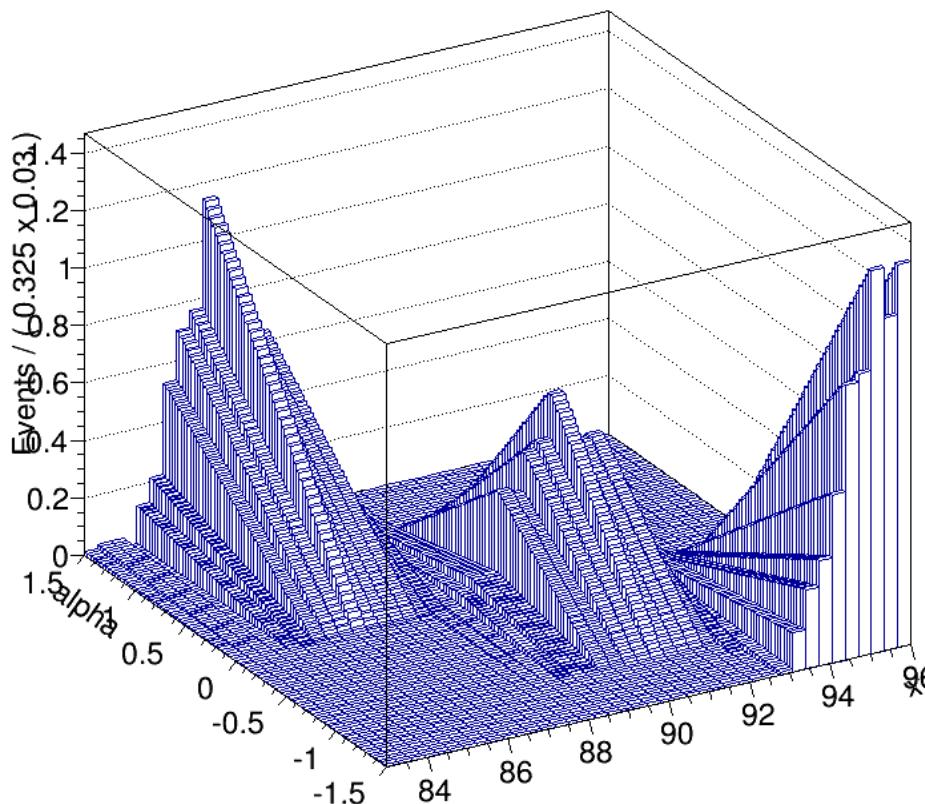
- Template morphing algorithms can introduce various other pathologies in the likelihood that cause MINUIT to fail
 - We've already seen some of them
- Kinks & Multiple minima
 - Caused by (among others) template morphing with piece-wise linear interpolation and morphing of (low-statistics) template distributions where MC statistical effects are larger than systematic effect



Limitations of piece-wise linear interpolation

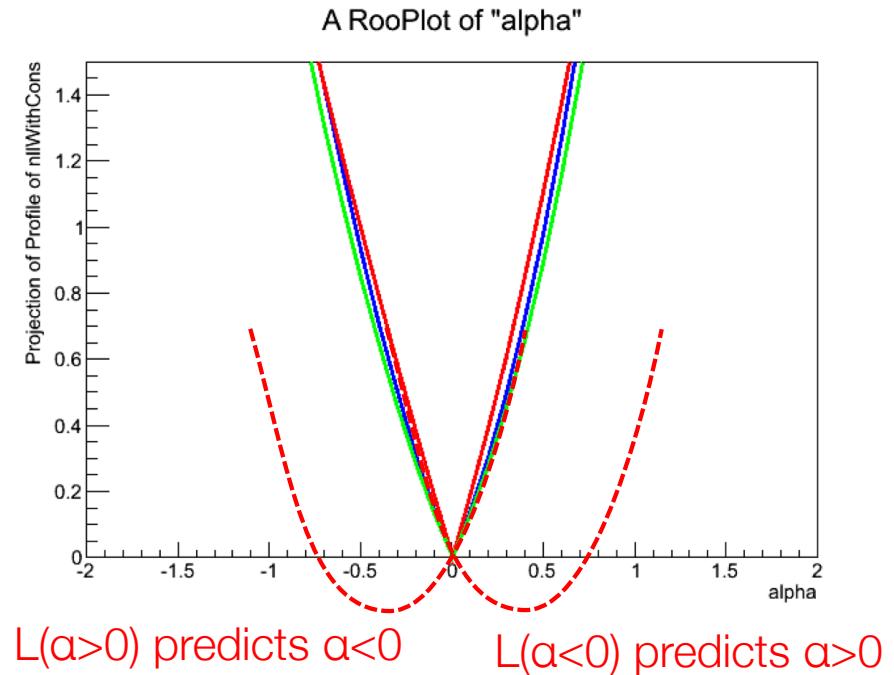
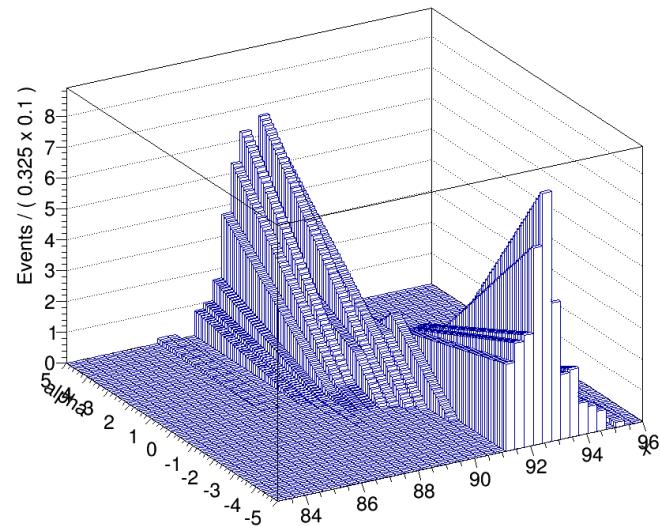
- Bin-by-bin interpolation looks spectacularly easy and simple, but be aware of its limitations
 - Same example, but with larger ‘mean shift’ between templates

Note double peak structure around $|alpha|=0.5$



Non-linear interpolation options

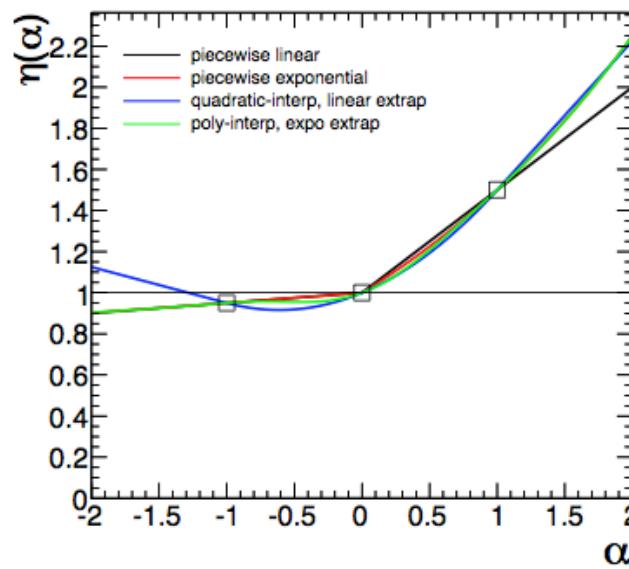
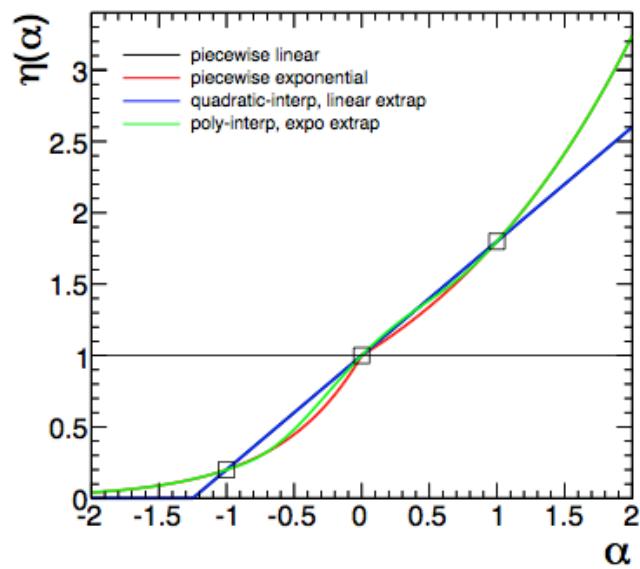
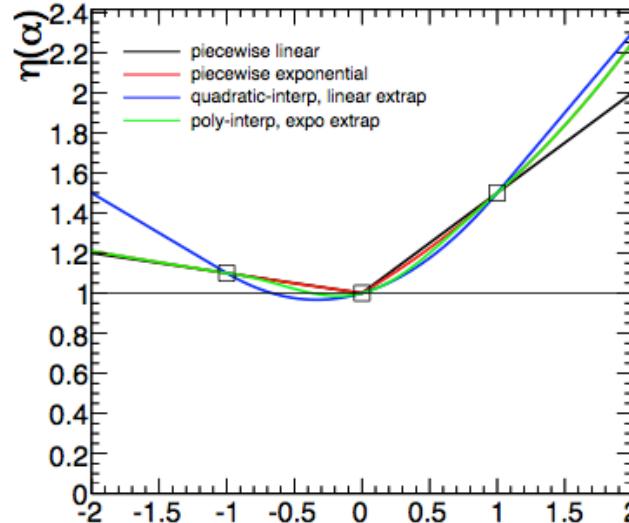
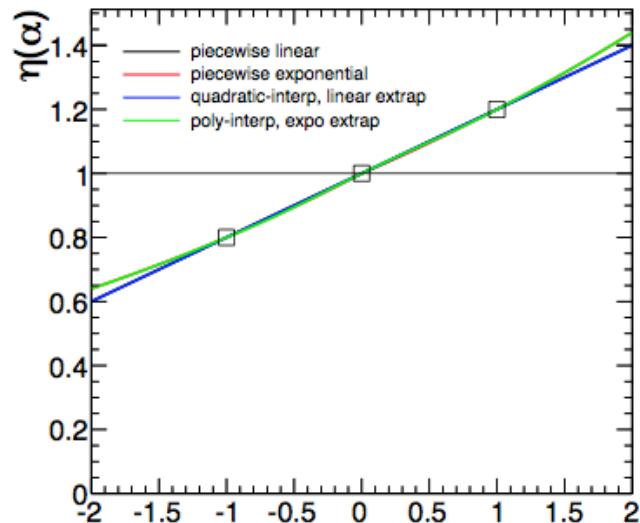
- Piece-wise linear interpolation leads to kink in response functions that may result in pathological likelihood functions



- A variety of other interpolation options exist that improve this
 - Parabolic interpolation/linear extrapolation (but causes shift of minimum)
 - Polynomial interpolation [orders 1,2,4,6]/linear extrapolation (order 1 term allows for asymmetric modeling of templates)

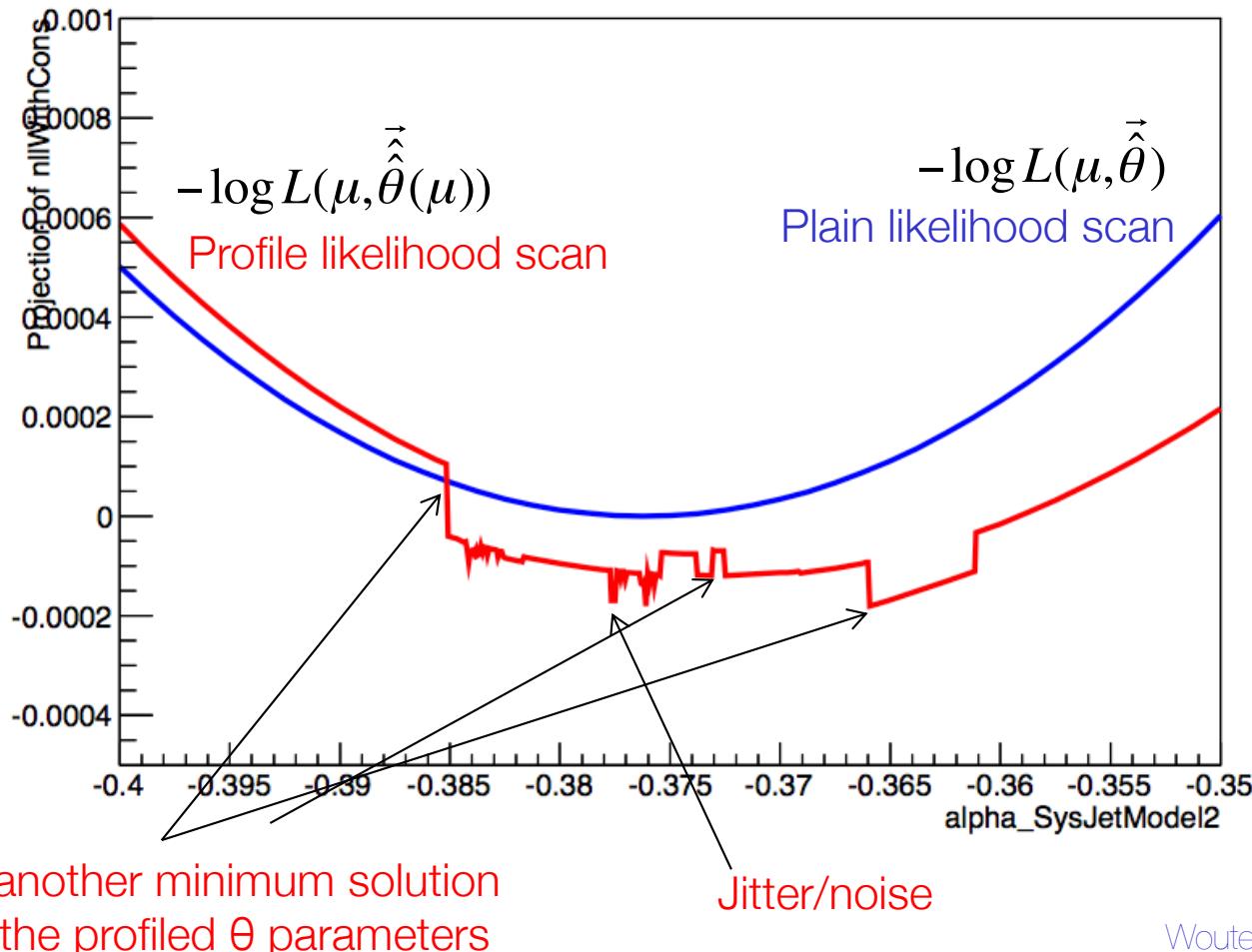
Non-linear interpolation options

- Comparison of common interpolation options



Other likelihood pathologies

- Effects of likelihood pathologies
 - Numerical noise and ‘jumping’ of profile likelihoods
 - Example NP (profile) likelihood scan of an ATLAS Higgs trial model



Other likelihood pathologies

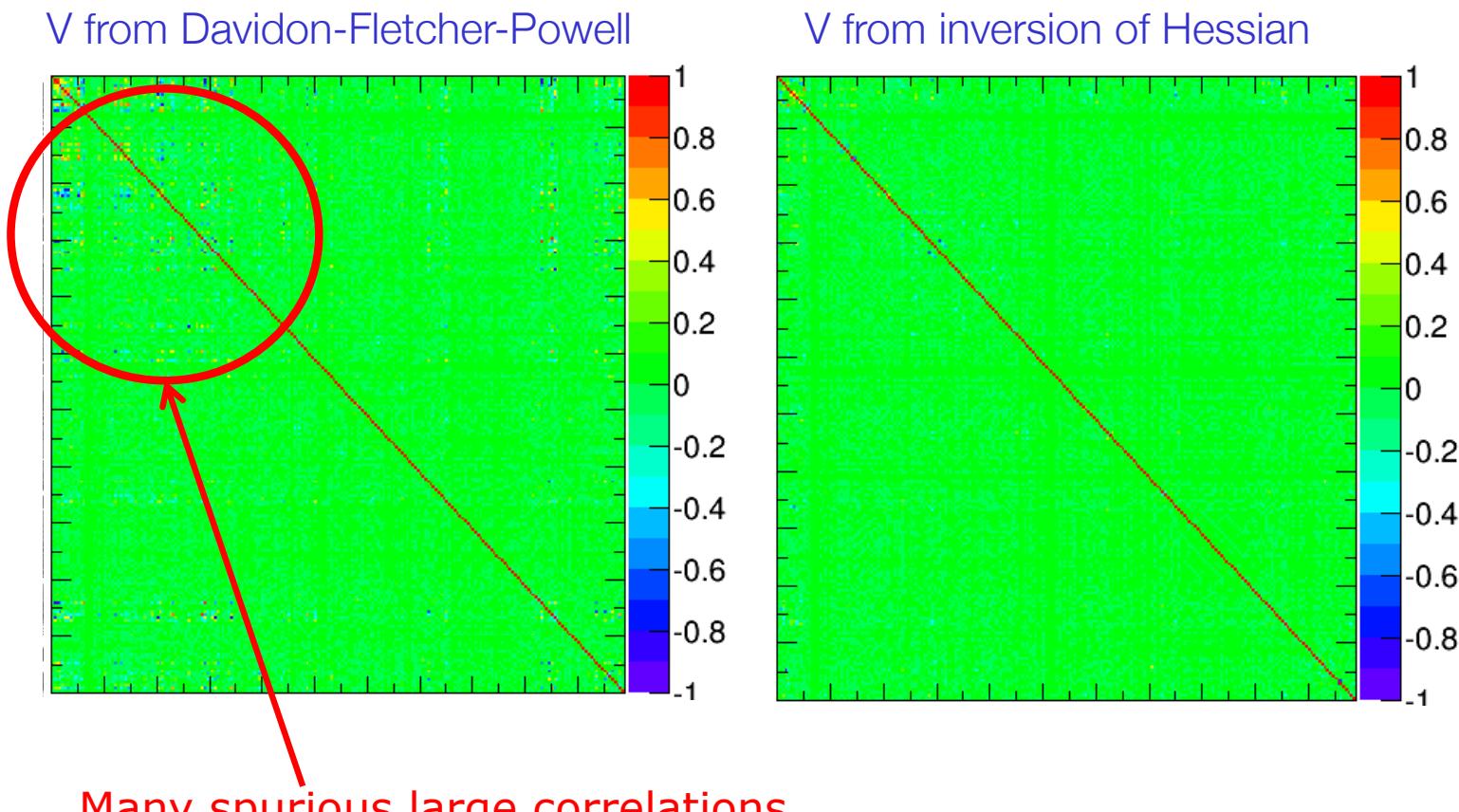
- Another effect of likelihood pathologies is that calculation of derivatives and notably the Hessian from either FDP or HESSE matrix become inaccurate
 - Slows down minimization
 - Can blow up EDM calculation → no convergence
- Red flags: EDM estimates that don't decrease ~monotonically
 - Only possible in Minuit2 (Minuit1 does not report EDM per step)

```
VariableMetric: start iterating until Edm is < 0.001
VariableMetric: Initial state - FCN = -289.1204081677 Edm = 46.0713 NCalls = 1826
VariableMetric: Iteration # 1 - FCN = -299.3073097602 Edm = 9.18415 NCalls = 2226
VariableMetric: Iteration # 2 - FCN = -304.9468725143 Edm = 2.22698 NCalls = 2624
VariableMetric: Iteration # 3 - FCN = -306.3323972775 Edm = 1.43793 NCalls = 3016
VariableMetric: Iteration # 4 - FCN = -307.199970017 Edm = 0.615574 NCalls = 3410
VariableMetric: Iteration # 5 - FCN = -307.6493784582 Edm = 0.352904 NCalls = 3804
VariableMetric: Iteration # 6 - FCN = -307.8960954798 Edm = 0.0749124 NCalls = 4196
VariableMetric: Iteration # 7 - FCN = -307.9549184882 Edm = 0.298047 NCalls = 4588
VariableMetric: Iteration # 8 - FCN = -308.0068371877 Edm = 3.40473 NCalls = 4980
```

- Solutions: simplify model: eliminate nuisance parameters that suffer from dominant MC statistical effects (causing multiple minima, kinks etc...)

Other likelihood pathologies

- Note that pathologies can affect calculation of V via iterative DFP updating and Hessian inversion differently
- A real-life example of complex likelihood fit where DFP estimate is strongly affected by likelihood pathologies



- But other likelihood pathologies can affect Hessian inversion more

Summary

- A variety of pathological features in likelihood models can interfere with minimization
 - Strong correlations
 - Kinks
 - Multiple minima
 - ‘Forbidden regions’ where likelihood is not defined
- Problems affect various steps of the minimization process
 - Understanding these effects requires basic understanding of the minimization algorithms and strategies
- Solutions usually involve simplifications of models

8

Diagnostics II: Overconstraining & choices in modeling parametrization

Understanding profile likelihood fits

- The previous section discussed technical diagnostics of profile likelihood fits
 - “Why doesn’t my fit converge”?
- The next level of diagnostics is on the interpretation level
 - “Do my fit results make sense”?
 - “What part of the likelihood model is measuring what?”

Being a good physicist – Understand your model!

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

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



Physics measurement Subsidiary measurement

- Our mental picture:

“measures s”



“measures a”

“dependence on a
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 a *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 a

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

Physics measurement Subsidiary measurement

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

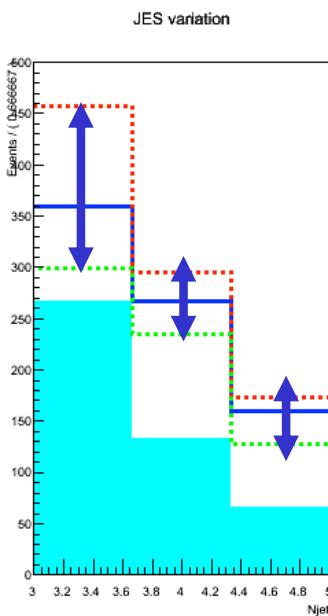
Understanding your model – what constrains your NP

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

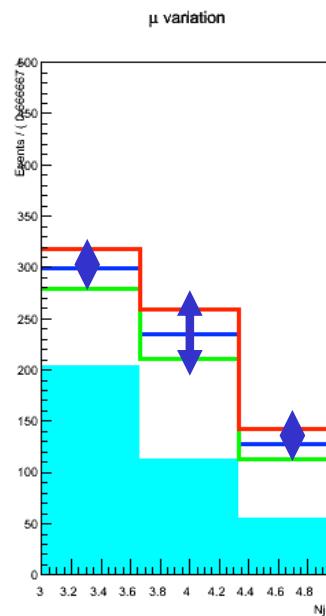
$$L(\vec{N} \mid \mu, \alpha_{JES}) = \prod_{i=3,4,5} Poisson(N_i \mid (\mu \cdot \tilde{s}_i \cdot + \tilde{b}_i) \cdot r_s(\alpha_{JES})) \cdot Gauss(0 \mid \alpha_{JES}, 1)$$

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

Effect of changing α_{JES}

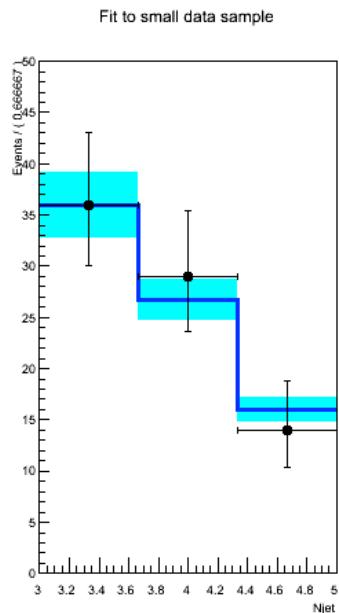


Effect of changing μ

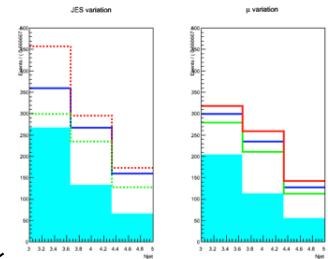
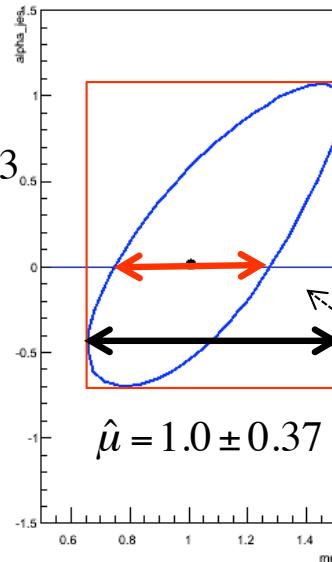


Understanding your model – what constrains your NP

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



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



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

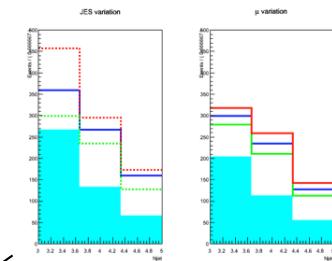
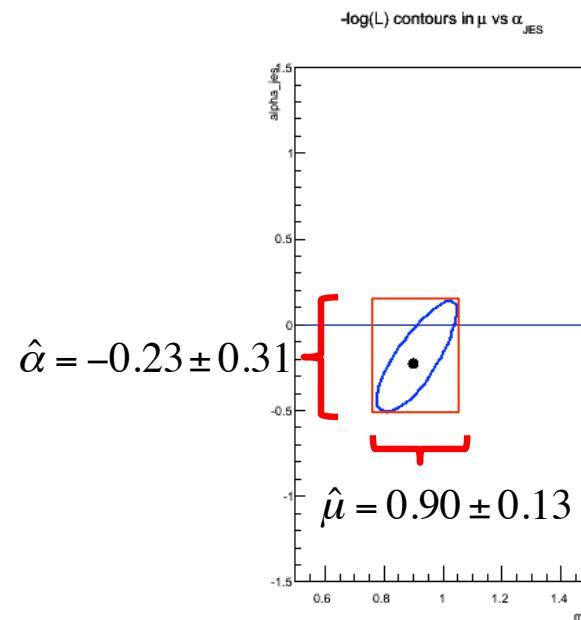
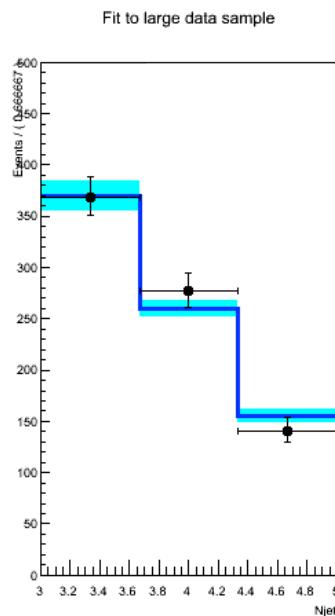
Uncertainty on μ with/without effect of JES

- 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



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 if model is valid.
- Is the probability model of the physics measurement valid?

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

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

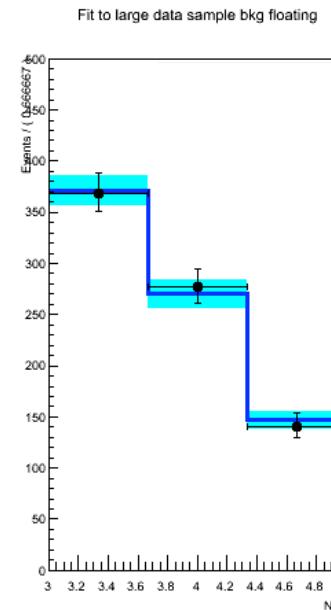
Understanding your model – what constrains your NP

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

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

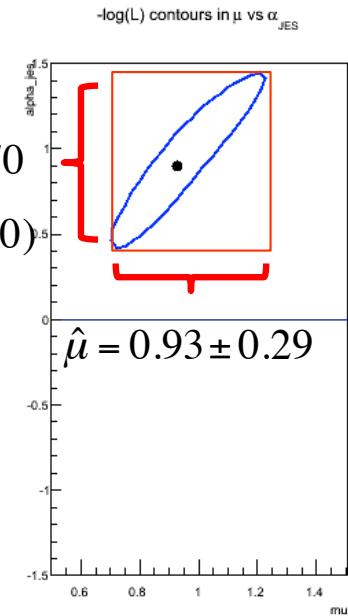
Background rate response function
Background rate subsidiary measurement

- 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...



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

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



Understanding your model – what constrains your NP

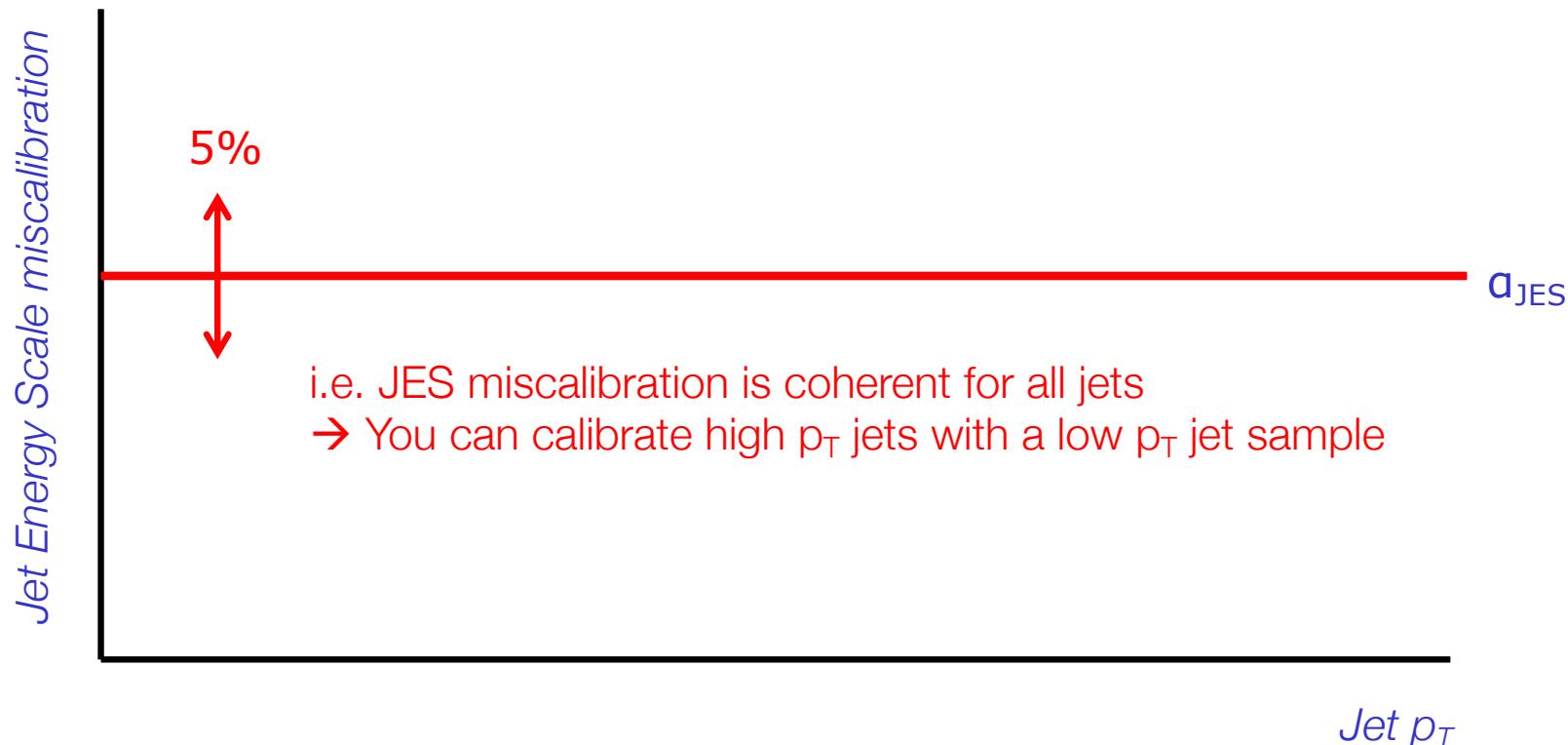
- Lesson learned: if probability model of a physics measurement is insufficiently detailed (i.e. flexible) you can *underestimate* uncertainties
- Normalized subsidiary measurement provide an excellent diagnostic tool
 - Whenever estimates of a NP associated with unit Gaussian subsidiary measurement deviate from $a = 0 \pm 1$ 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 $t \rightarrow bW(qq)$, the pair of light jets should always have $m(jj)=mW$. For this special sample of jets it will be 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 a_{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 $a_{\text{JES}} = 0 \pm 0.4$ 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 $t \rightarrow bW(qq)$ decays)

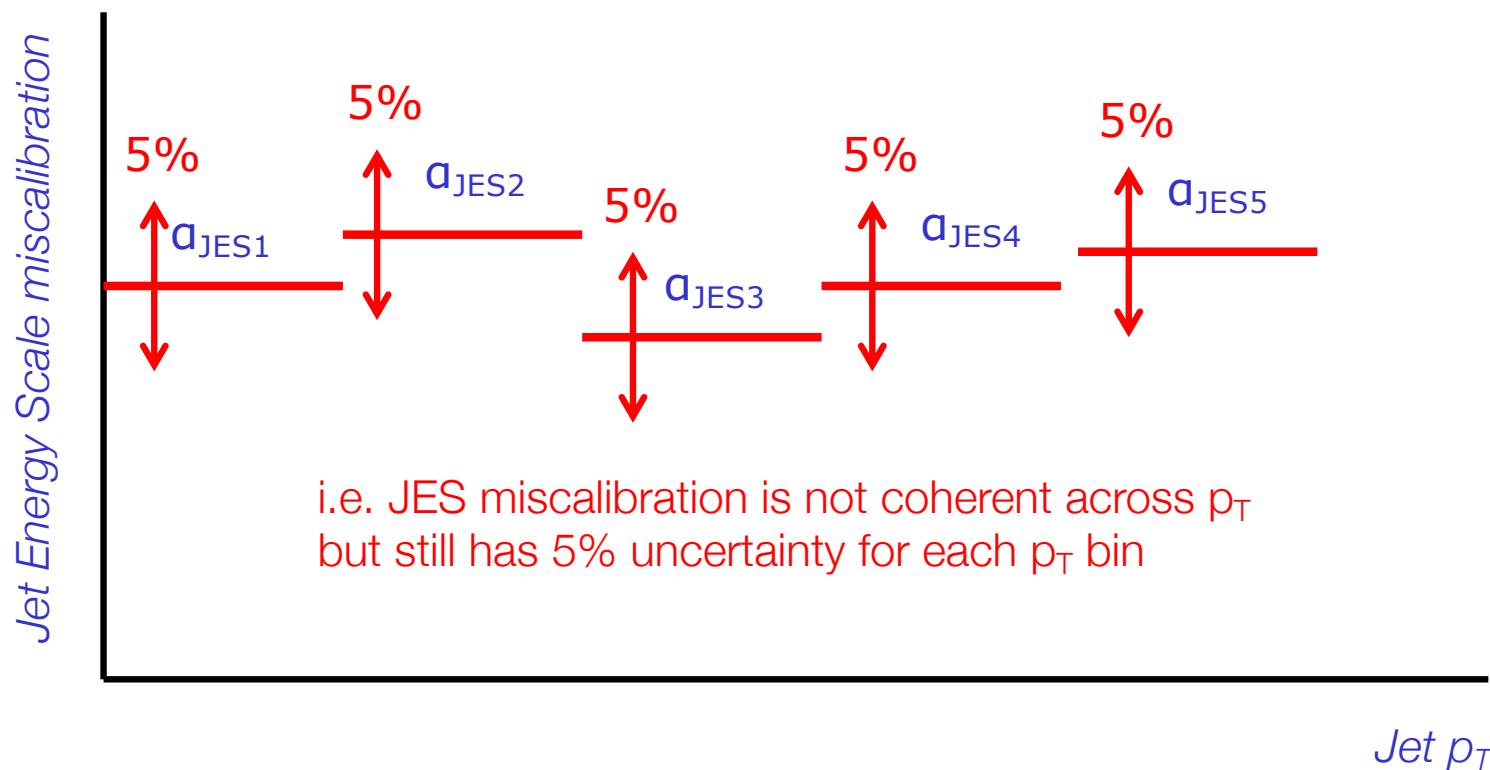
Dealing with over-constraining – introducing more NPs

- 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



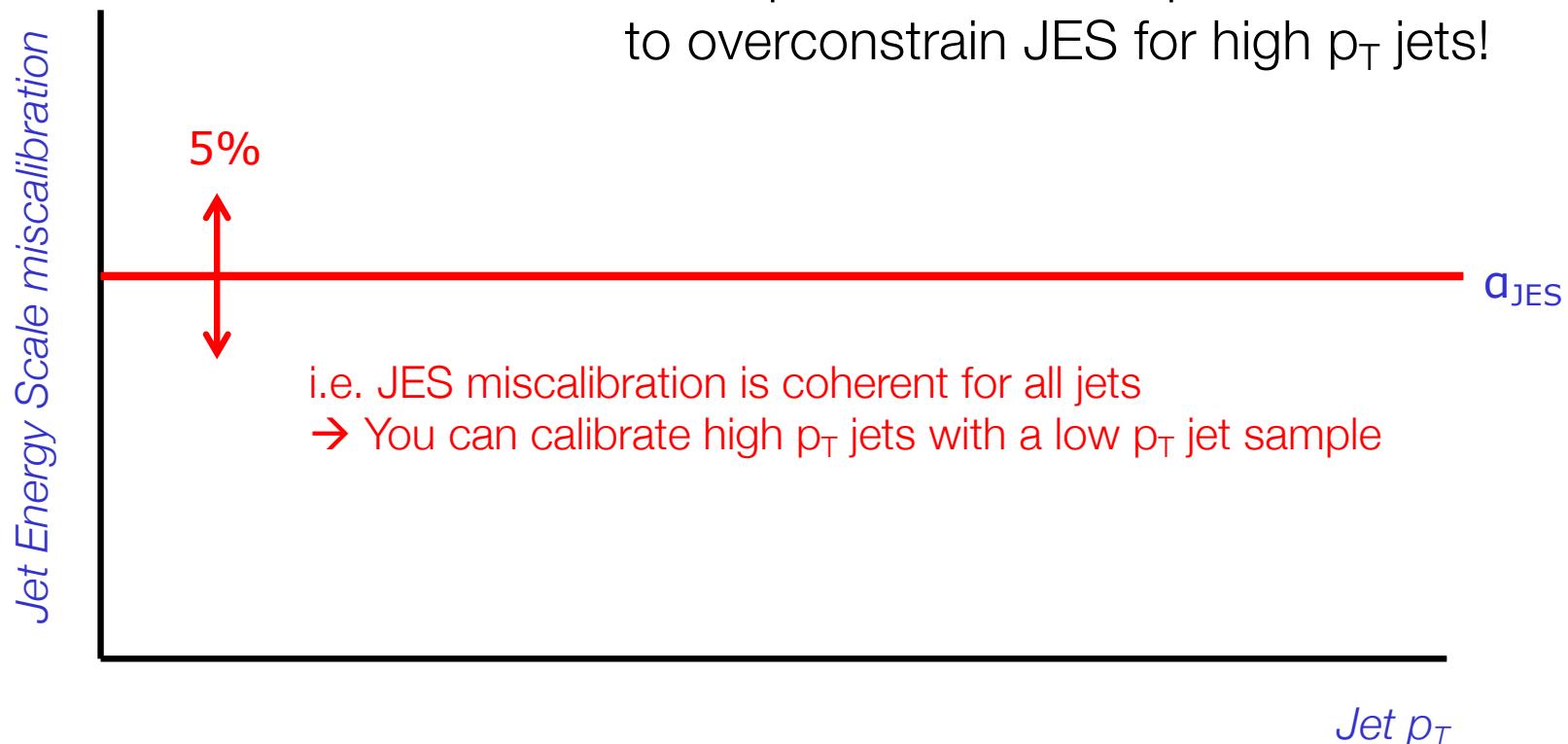
Dealing with over-constraining – introducing more NPs

- 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 $p_T < 40$ ”
 - 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
- Subsidiary measurement is an actual measurement
→ conceptually to a ‘sideband’ fit
- Subsidiary measurement unclear, but origin of prescription may well be another measurement (if yes, like sideband, if no, what is source of info?)
- Subsidiary measurement is a Poisson counting experiment (but now in MC events), otherwise conceptually identical to a ‘sideband’ fit

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

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

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

- **Difficult example** – Factorization scale uncertainty

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

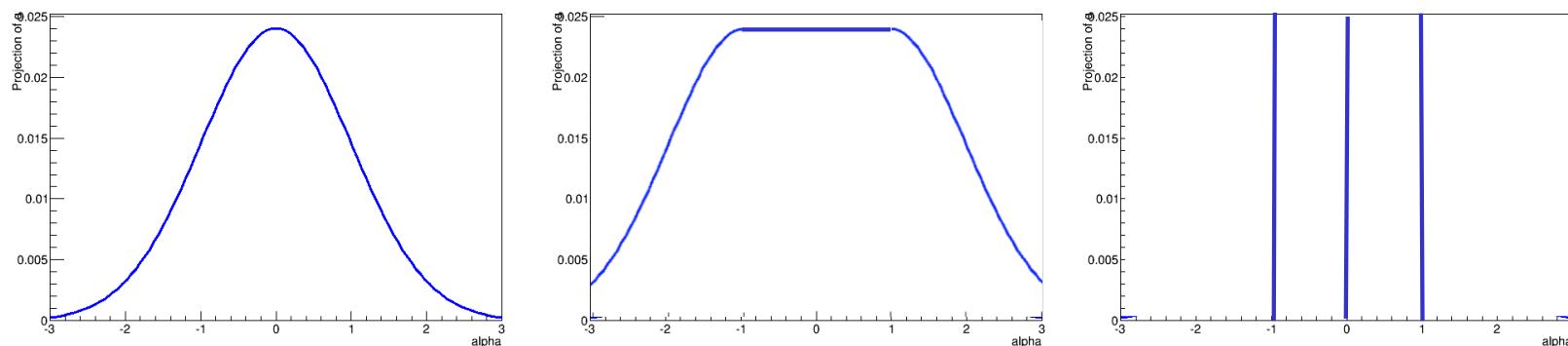
“Vary Factorization Scale by x0.5 and x” → $F(a)$ is probably not Gaussian
So what distribution was meant?

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” → $F(a)$ 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** – b-quark mass uncertainty

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

- One parameter: the quark mass → Clearly described and connected to the underlying theory model
- **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

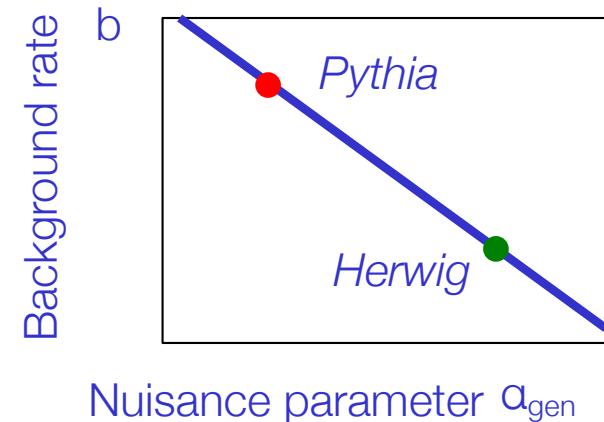
- It is difficult to define rigorous statistical procedures to deal with such 2-point uncertainties. So you need to decide
- If their estimated effect is small, you can pragmatically ignore these lack of proper knowledge and ‘just do something reasonable’ to model these effects in a likelihood
- If their 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 \rightarrow 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

$$L(N|s, \alpha) = \text{Poisson}(N|s + b(\alpha)) \cdot \text{SomePdf}(0|\alpha)$$

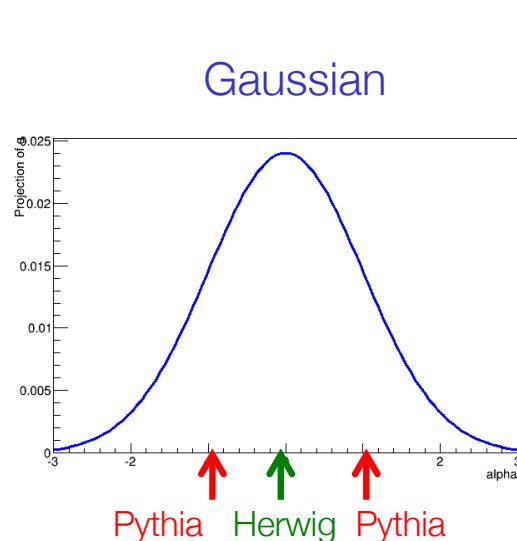
- Defining an (empirical) response function $b(a)$ is the easy part



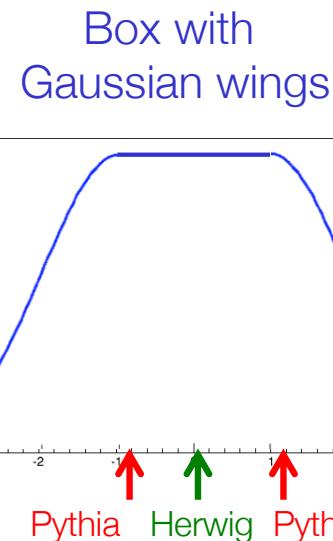
- A thorny question remains:
What is the subsidiary measurement for a ?
This should reflect your current knowledge on a .

Specific issues with theory uncertainties

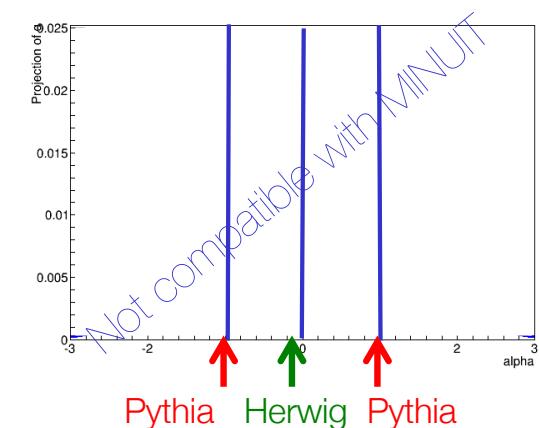
- Subsidiary measurement of a theoretical 2-point uncertainty effectively quantifies the ‘knowledge’ on these models
 - *Extra difficult to make meaningful statement about this*, since meaning of parameter is not well embedded in underlying theory model
 - But again, all procedures need to assume some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



Prefers Herwig at 1σ



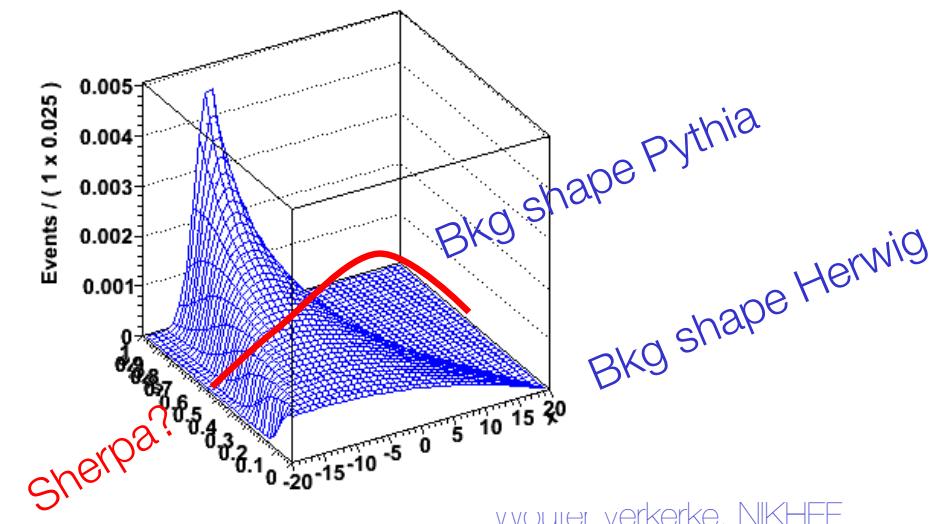
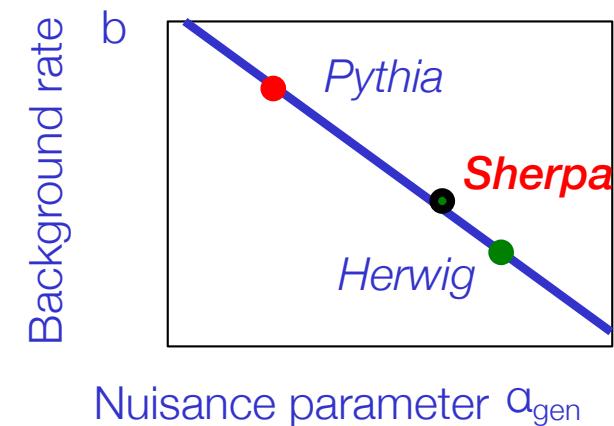
All predictions ‘between’
Herwig and Pythia equally
probable



Only ‘pure’ Herwig
and Pythia exist

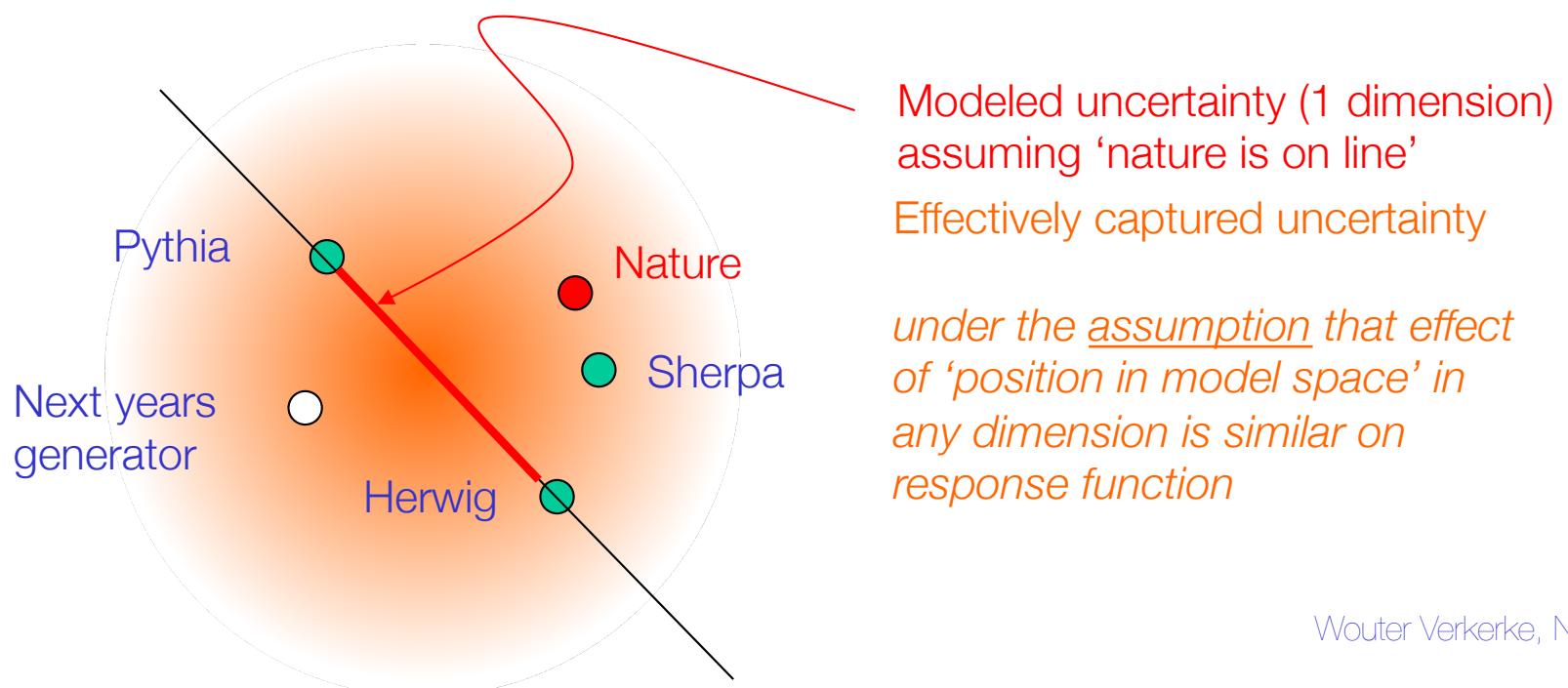
Two-point systematics on non-counting measurements

- 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...



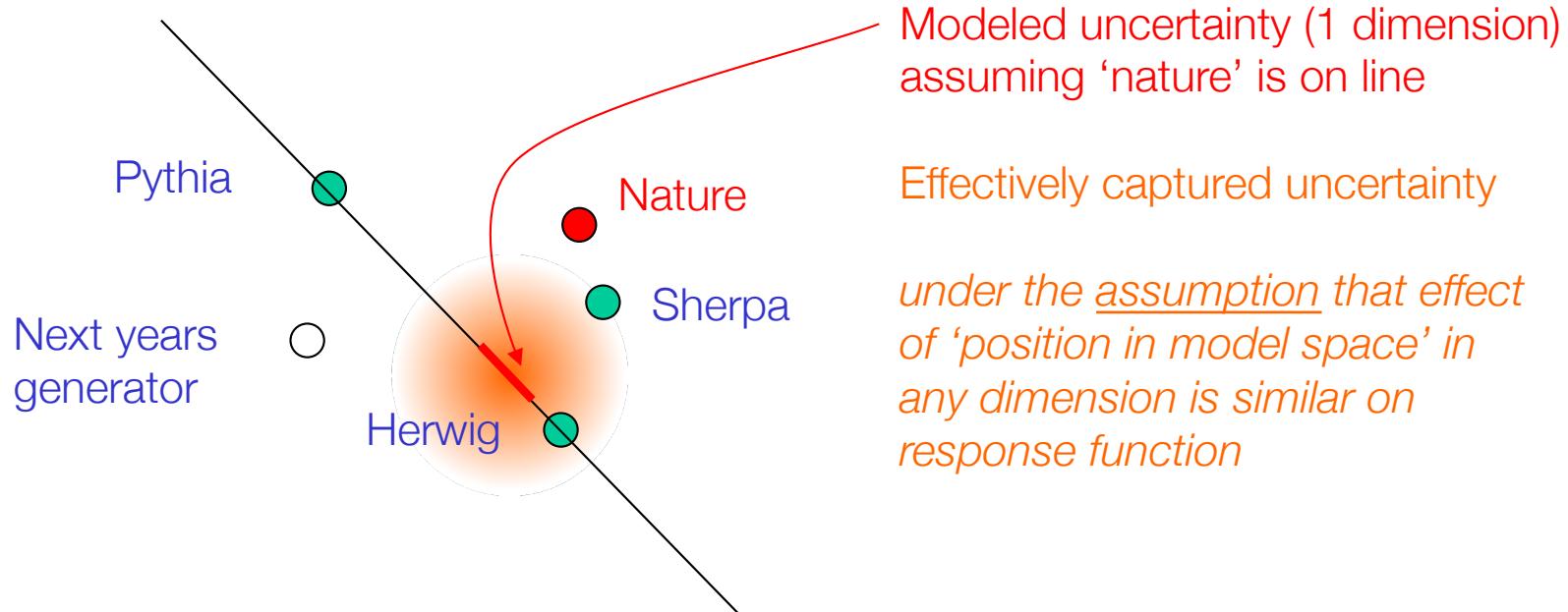
Dealing with ‘two-point’ uncertainties

- Key issue: *How many d.o.f. does your systematic uncertainty have?*
- Especially important in the discussion to what extent a two-point response function can be over-constrained.
 - A result $a_{2p} = 0.5 \pm 1$ 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 your 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 $a_{2p} = 0.1 \pm 0.2$ ‘reasonably’ cover the generator model space?



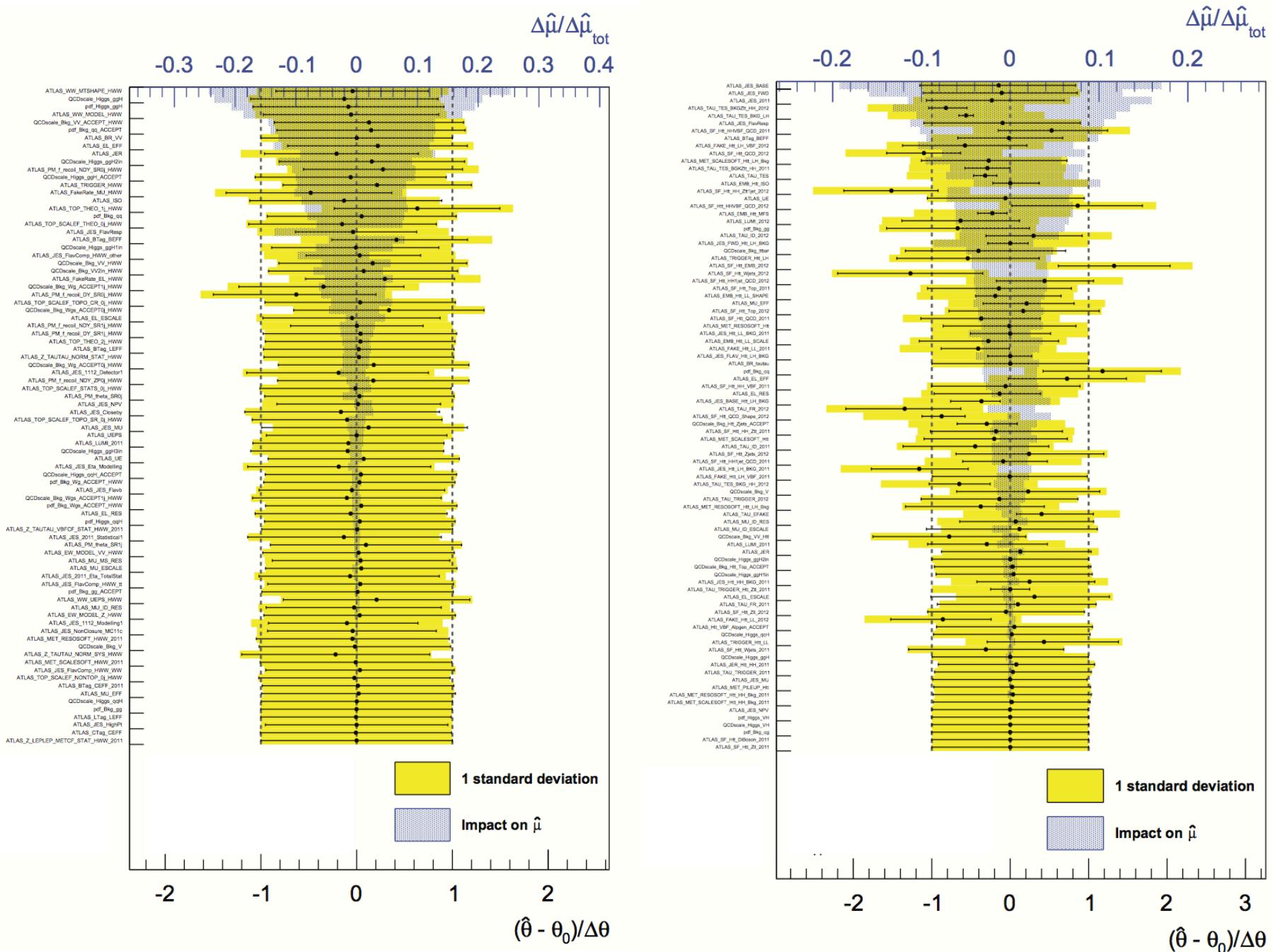
Summary

- 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 explicitly spell out precise model that is used
- **But it 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 in probability models

Example of likelihood modeling diagnostics



Summary

- Diagnostics over NP overconstraining provide powerful insight into your analysis model
 - An overconstrained NP indicates an externally provided systematic is inconsistent with physics measurement
 - This may point to either an incorrect response modeling of that uncertainty, to result in a genuinely better estimate of the uncertainty
 - Solution not always clear-cut, but you should be at least aware of it.
 - Note that over-constraining always points to an underlying physics issue (lack of knowledge, simplistic modeling) → Treat it as a physics analysis problem, not as a statistics problem
- Diagnostic power of profile likelihood models highlights one of the major shortcomings of the ‘naïve’ strategy of error propagation (as discussed in Section 1)
 - Physics measurement can entangle in non-trivial ways with systematic uncertainties

9

Summary & conclusions

Summary

- Modelling of systematic uncertainties in the likelihood ('profiling') is the best we know to incorporate systematic uncertainties in rigorous statistical procedures
 - Profiling requires more a 'exact' specification of what a systematic uncertainty means than traditional prescriptions → this is good thing, it makes you think about (otherwise hidden) assumption
 - It's important to involve the 'author' of uncertainty prescription in this process, as flawed assumptions can be exploited by statistical methods to arrive at unwarranted conclusions
 - Systematic uncertainties that have conceptual fuzziness ('pythia-vs-herwig') are difficult to capture in the likelihood, but this is a reflection of an underlying physics problem
 - Good software tools exist to simplify the process of likelihood modeling
 - It's important to carefully diagnose your profile likelihood models for both technical and interpretational problems ('over-constraining')