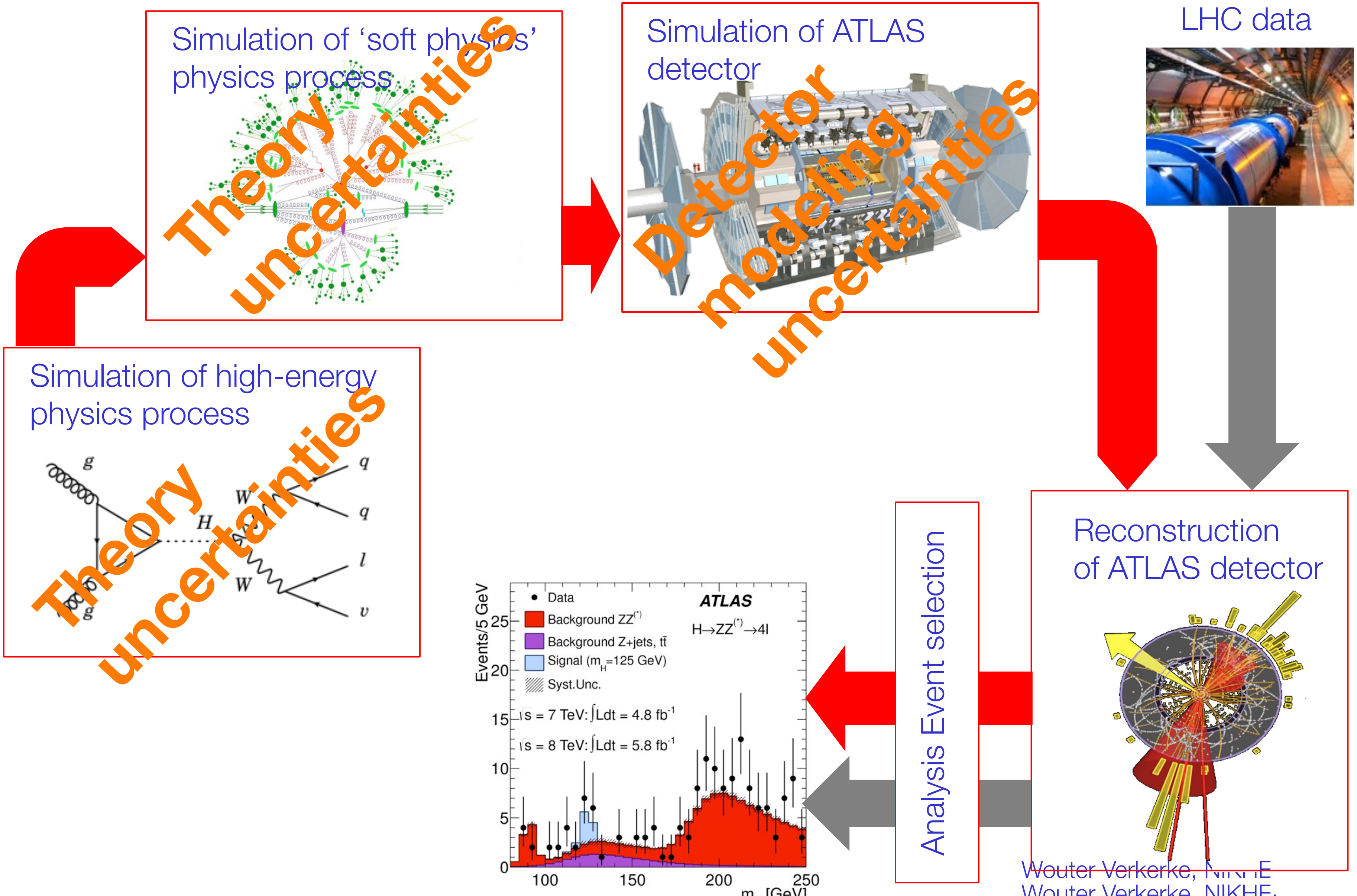
Statistics

W. Verkerke

[Recap]

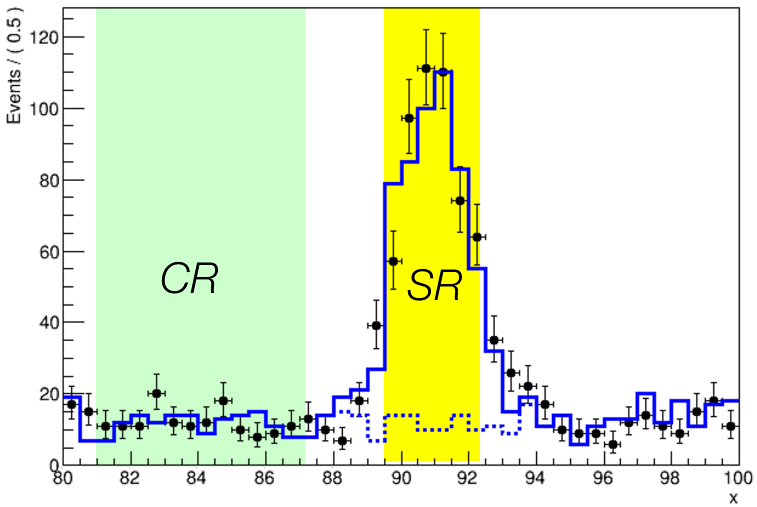
The simulation workflow and origin of uncertainties



Wouter Verkerke, NIKHEF

The sideband measurement

- Suppose your data in reality looks like this ➔



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

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

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

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

"Background uncertainty constrained from the data"

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

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

Generalizing the concept of the sideband measurement

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

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

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

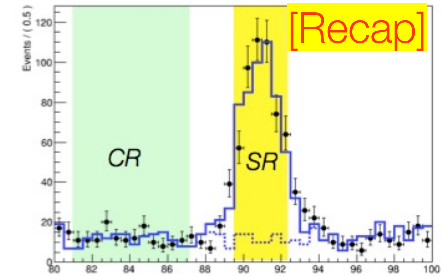
'Measured background rate by MC simulation'

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

'**Subsidiary measurement**' of background rate

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

Generalize: 'sideband' → 'subsidiary measurement'



Modeling a detector calibration uncertainty

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

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

Nominal calibration

Signal rate (our parameter of interest)

Assumed calibration

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

Observed event count

Uncertainty on nominal calibration **(here 5%)**

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

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

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

Modeling a detector calibration uncertainty

- Simplify expression by renormalizing "subsidiary measurement"

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

Signal rate (our parameter of interest)

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

Observed event count

Nominal background expectation from MC (a constant)

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

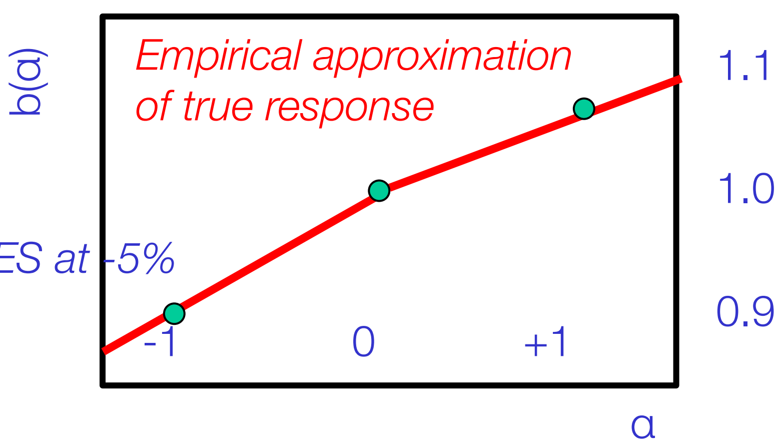
"Normalized subsidiary measurement"

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

The response function as empirical model of full simulation

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

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



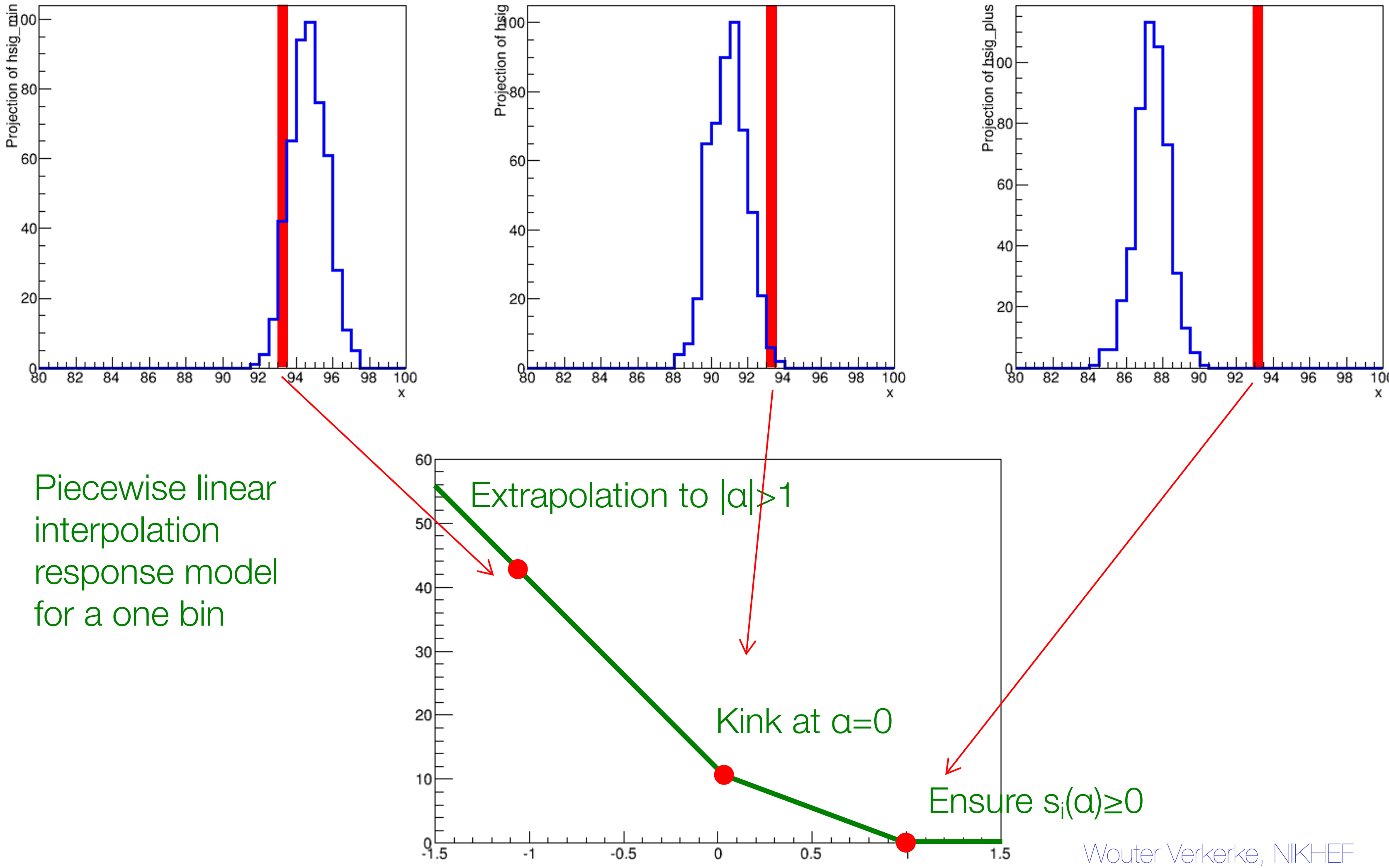
What is a systematic uncertainty?

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

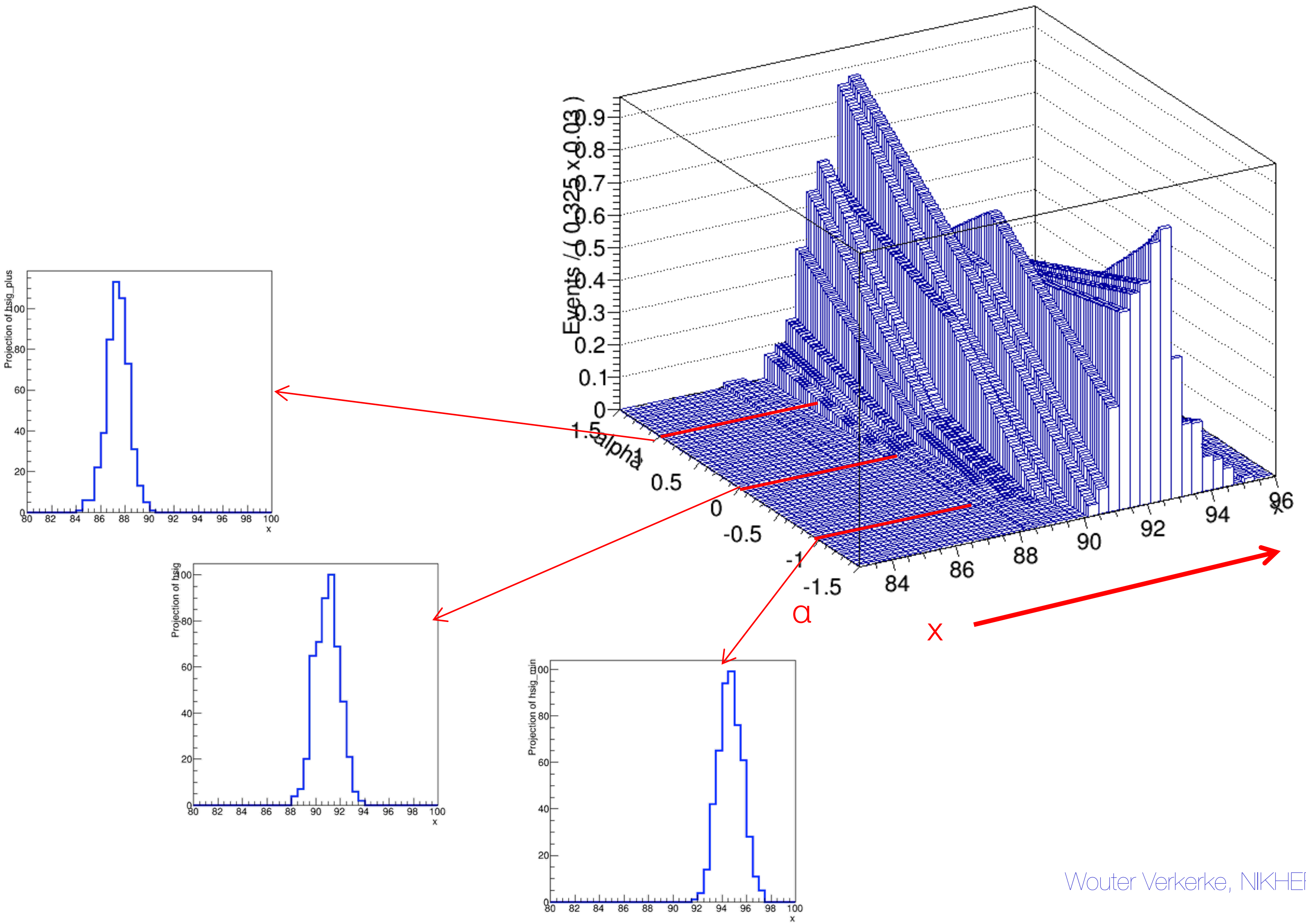
Piecewise linear interpolation

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

Piecewise linear interpolation response model for a one bin

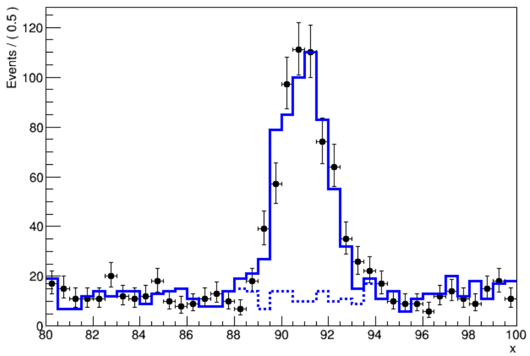


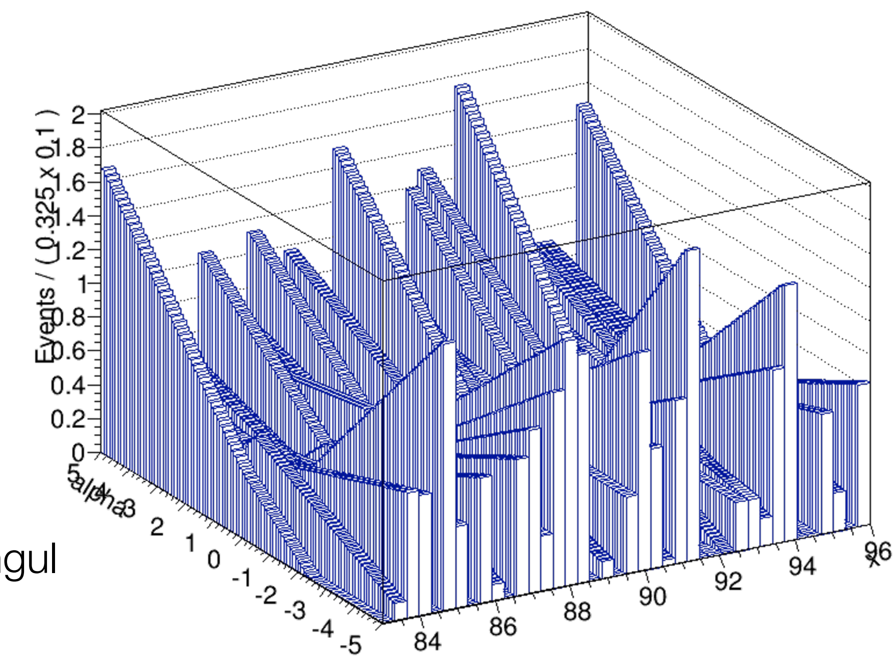
Visualization of bin-by-bin linear interpolation of distribution



Shape, rate or no systematic?

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

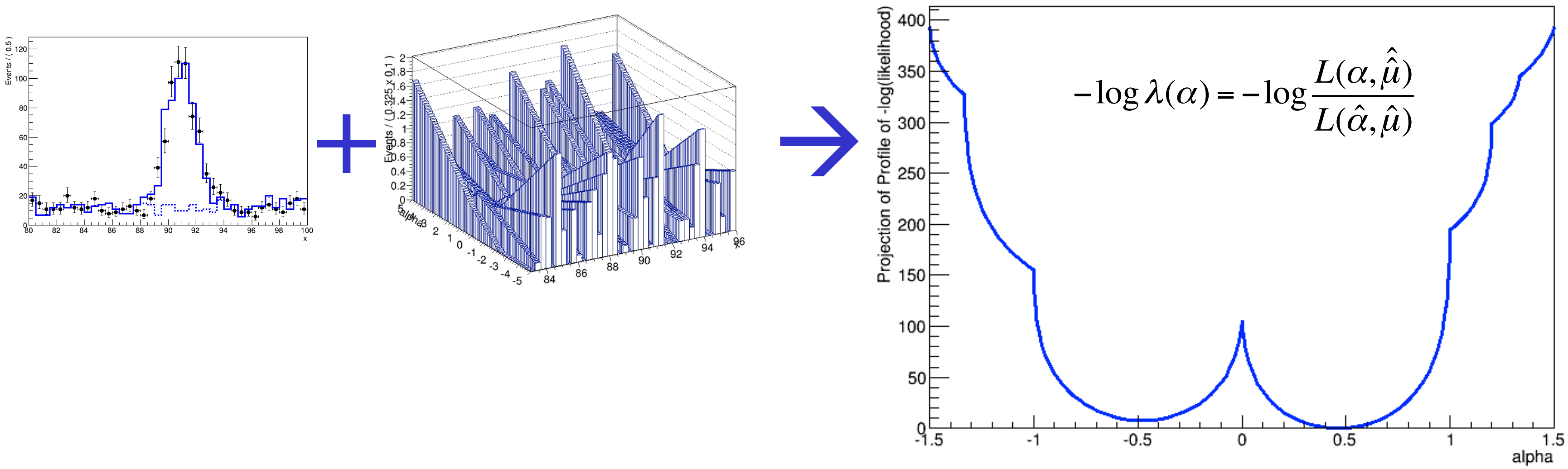




  - A χ2 or KS test between nominal and alternate template can help to decide if a shape uncertainty is meaningul
  - Most systematic uncertainties affect both rate and shape, but can make independent decision on modeling rate (which less likely to affect fit stability)

Fit stability due to insignificant shape systematics

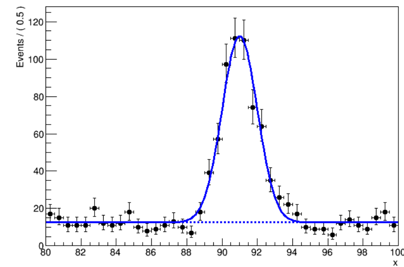
- Shape of profile likelihood in NP α clearly raises two points



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

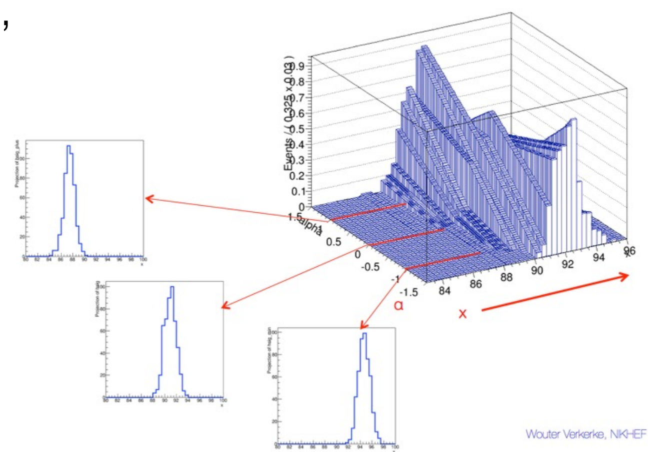
Recap on shape systematics & template morphing

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



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

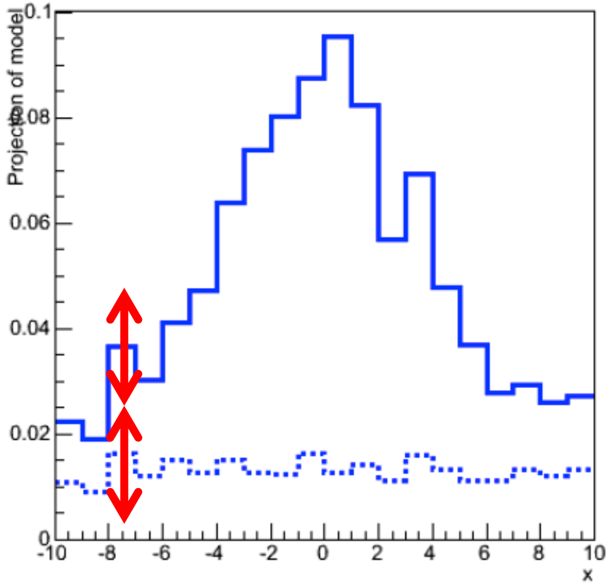
- For template modes obtained from MC simulation template provides a technical solution to implement response function
  - Simplest strategy piecewise linear interpolation, but only works well for small changes
  - Moment morphing better adapted to modeling of shifting distributions
  - Both algorithms 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

Other uncertainties in MC shapes – finite MC statistics

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



Fixed signal, bkg MC prediction

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

Signal, bkg
MC nuisance params

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

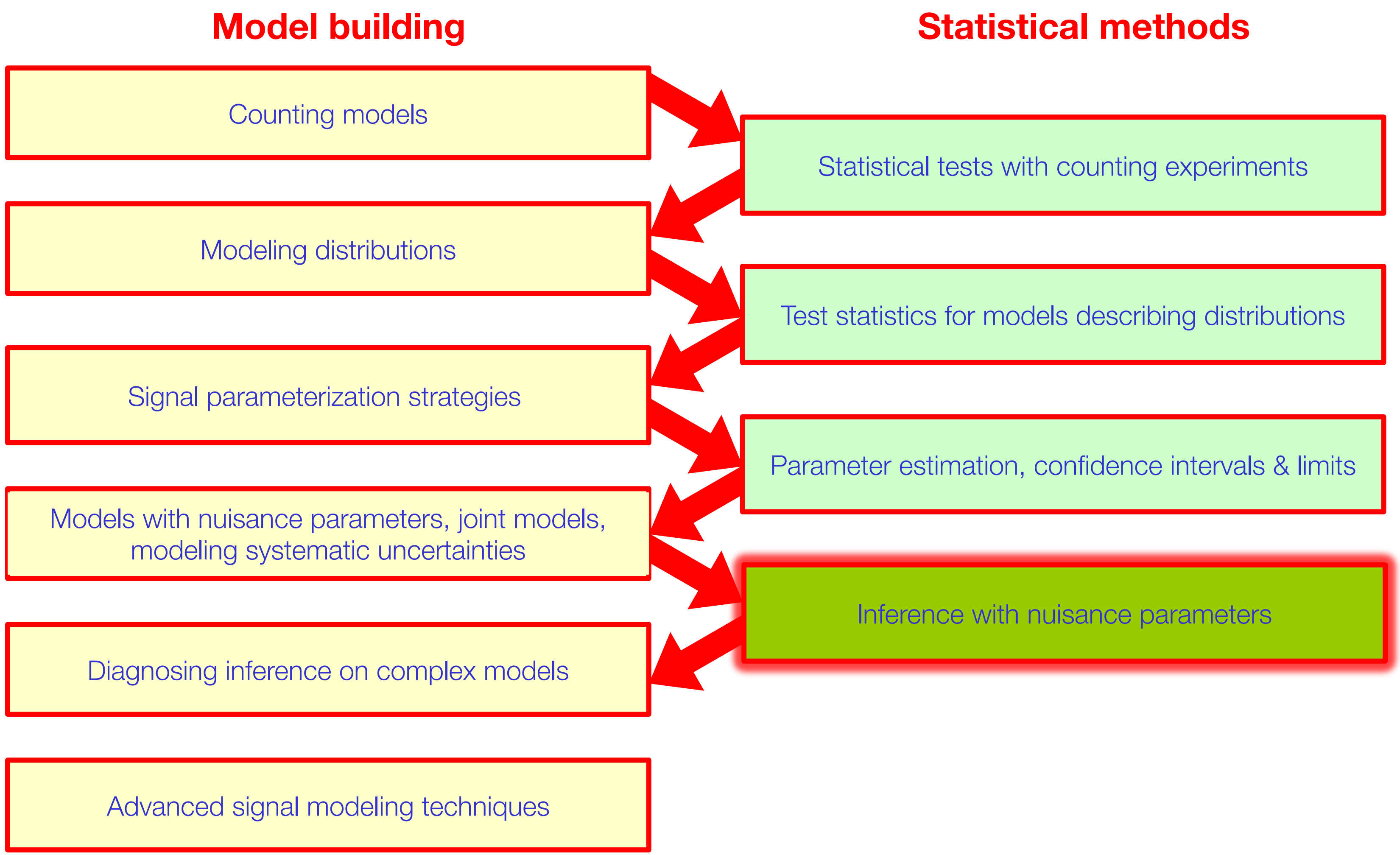
$$\cdot Poisson(N_i^{MC-s} \mid s_i)$$

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

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

Roadmap of this course

- Start with basics, gradually build up to complexity



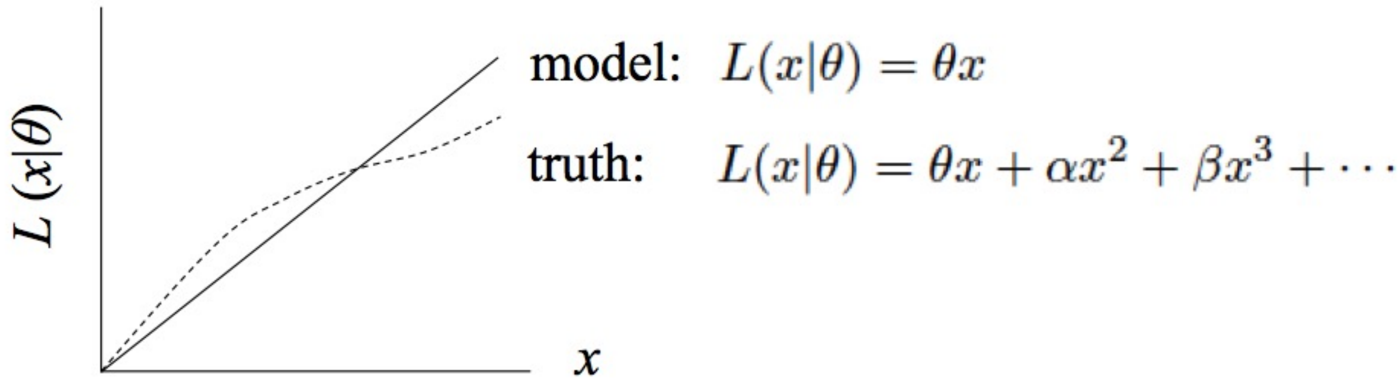
Statistical methods 4

Parameters of interest vs nuisance parameters, dealing with nuisance parameters in inference methods

Wouter Verkerke, NIKHEF

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 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

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

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

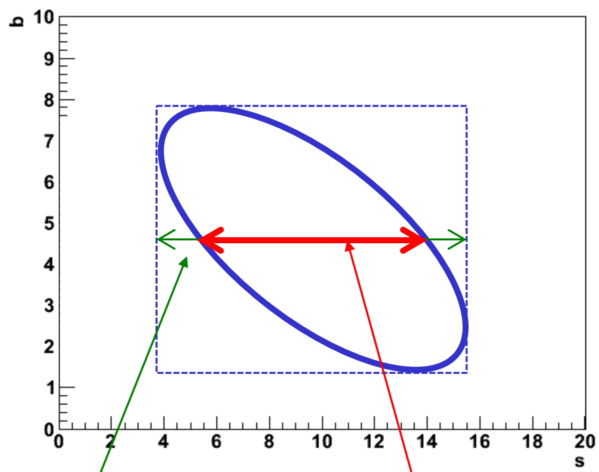
- For multi-parameter likelihoods estimate of covariance V_{ij} of pair of 2 parameters in addition to variance of individual parameters
  - Usually re-expressed in terms dimensionless correlation coefficients ρ

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

Treatment of nuisance parameters in variance estimation

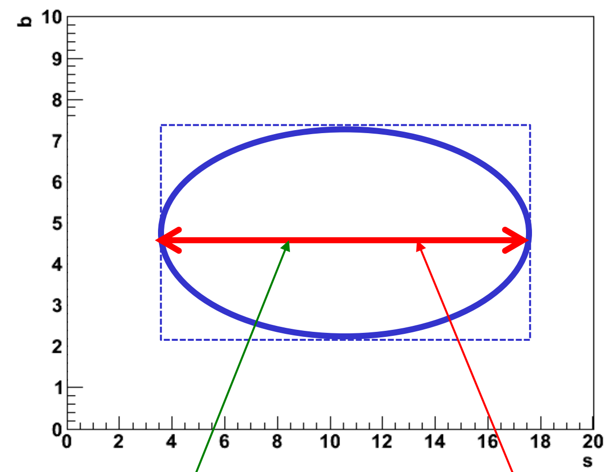
- Effect of NPs on variance estimates visualized

Scenario 1
Estimators of
POI and NP correlated
i.e. ρ(s,b)≠0



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

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



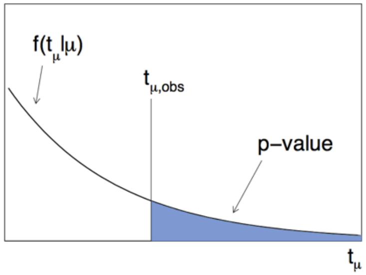
$$\hat{V}(s) \; from \begin{bmatrix} \frac{\partial^2 L}{\partial s^2} & \frac{\partial^2 L}{\partial s \partial b} \\ \frac{\partial^2 L}{\partial s \partial b} & \frac{\partial^2 L}{\partial b^2} \end{bmatrix}^{-1} \qquad \hat{V}(s) \; from \left[\frac{\partial^2 L}{\partial s^2} \right]^{-1}_{b=\hat{b}}$$

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(μ)
  - 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(μ,θ) where θ is a nuisance parameter?
  - With a test statistics q_μ, we calculate p-value for hypothesis θ as

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



- But what values of θ do we use for $f(q_\mu|\mu,\theta)$?
  Fundamentally, we want to reject μ only if p<α 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(μ,θ) — NPs: θ; POI: μ

  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

  $$\hat{\hat{\theta}}(\mu)$$

  M.L. estimate of θ for a given value of μ (i.e. a conditional ML estimate)

- The resulting confidence interval will have exact coverage for the points $(\mu, \hat{\hat{\theta}}(\mu))$
  - 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

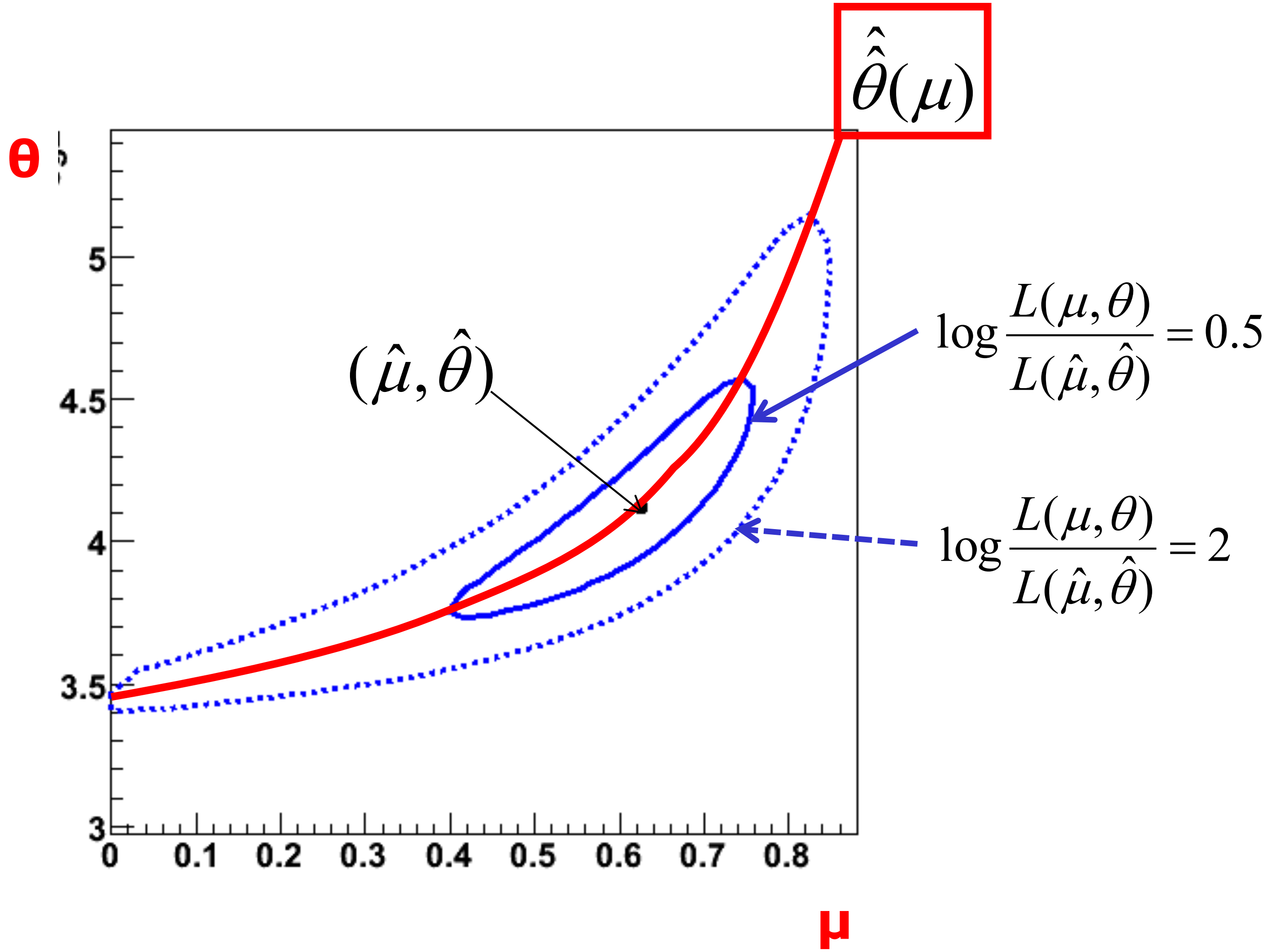
Maximum Likelihood for given μ

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

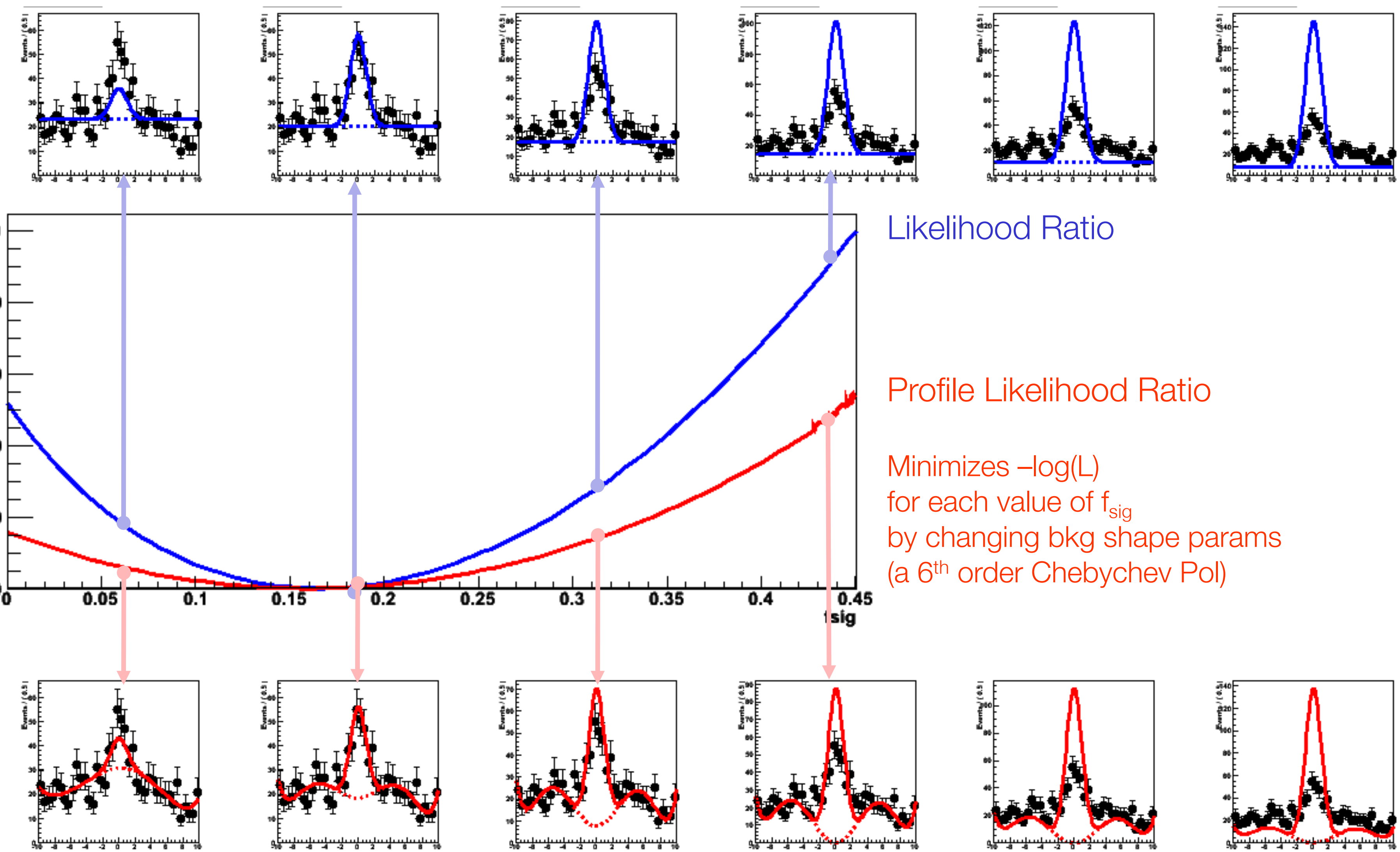
Maximum Likelihood

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

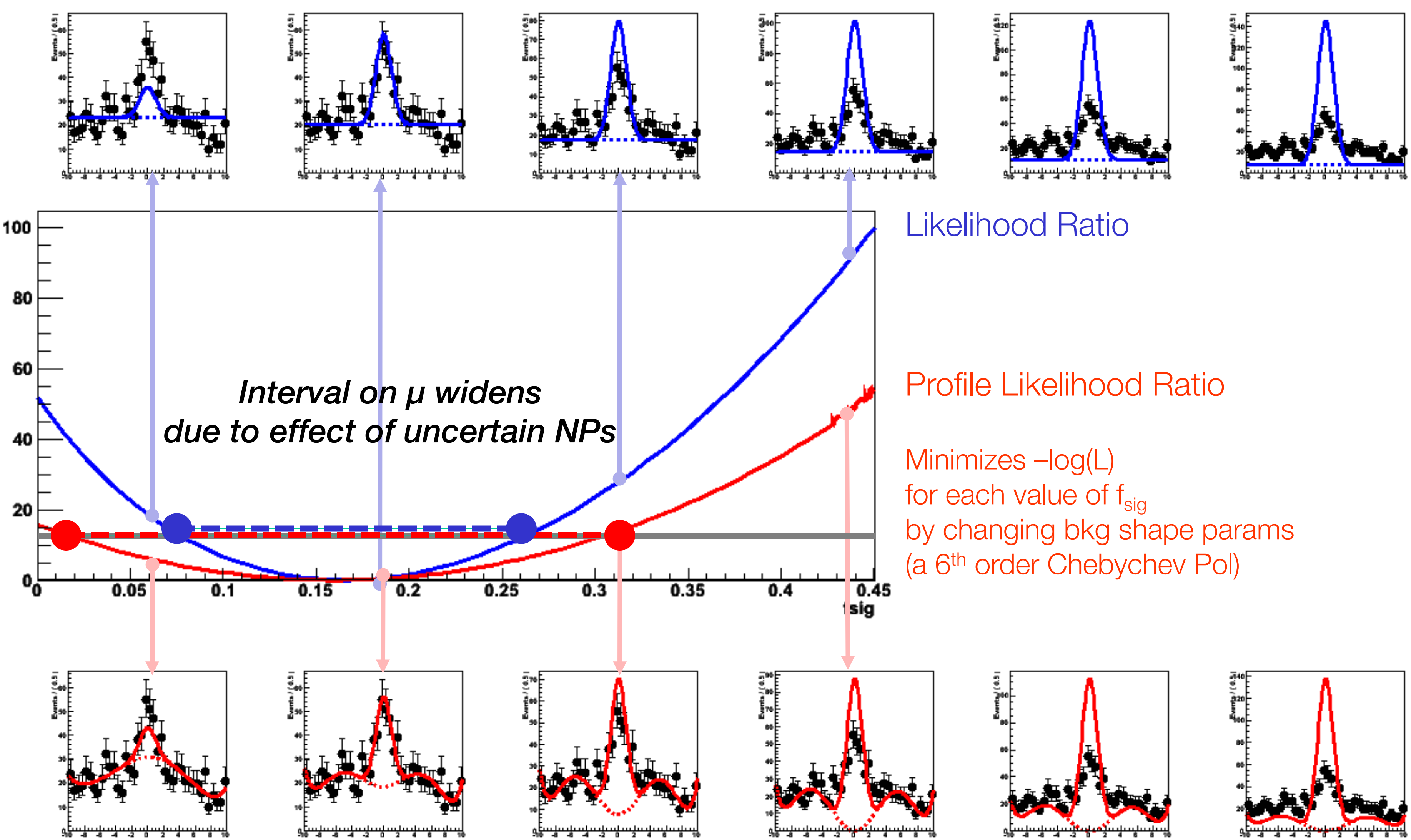
Profiling illustration with one nuisance parameter



Profile scan of a Gaussian plus Polynomial probability model



Profile scan of a Gaussian plus Polynomial probability model



Likelihood Ratio

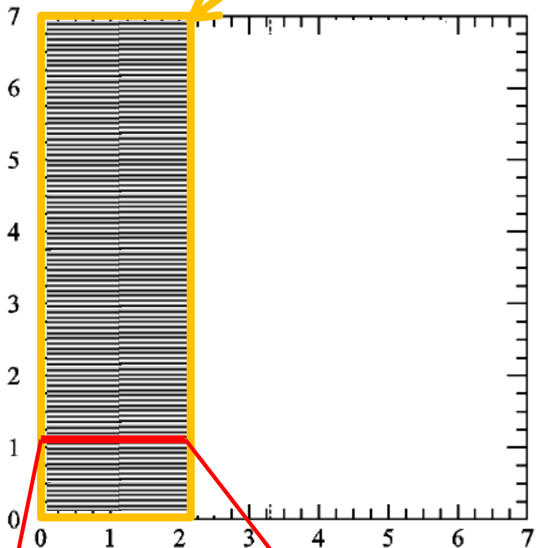
Profile Likelihood Ratio

Minimizes –log(L)
for each value of f_{sig}
by changing bkg shape params
(a 6th order Chebychev Pol)

PLR Confidence interval vs MINOS

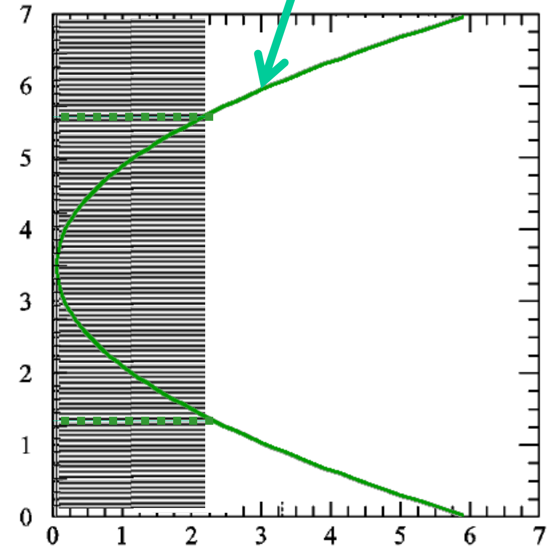
Confidence belt now range in PLR

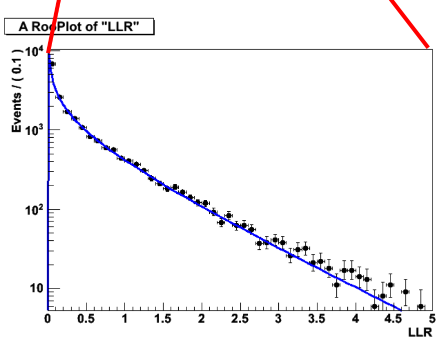
$t_\mu(x,\mu)$



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

$t_\mu(x,\mu)$

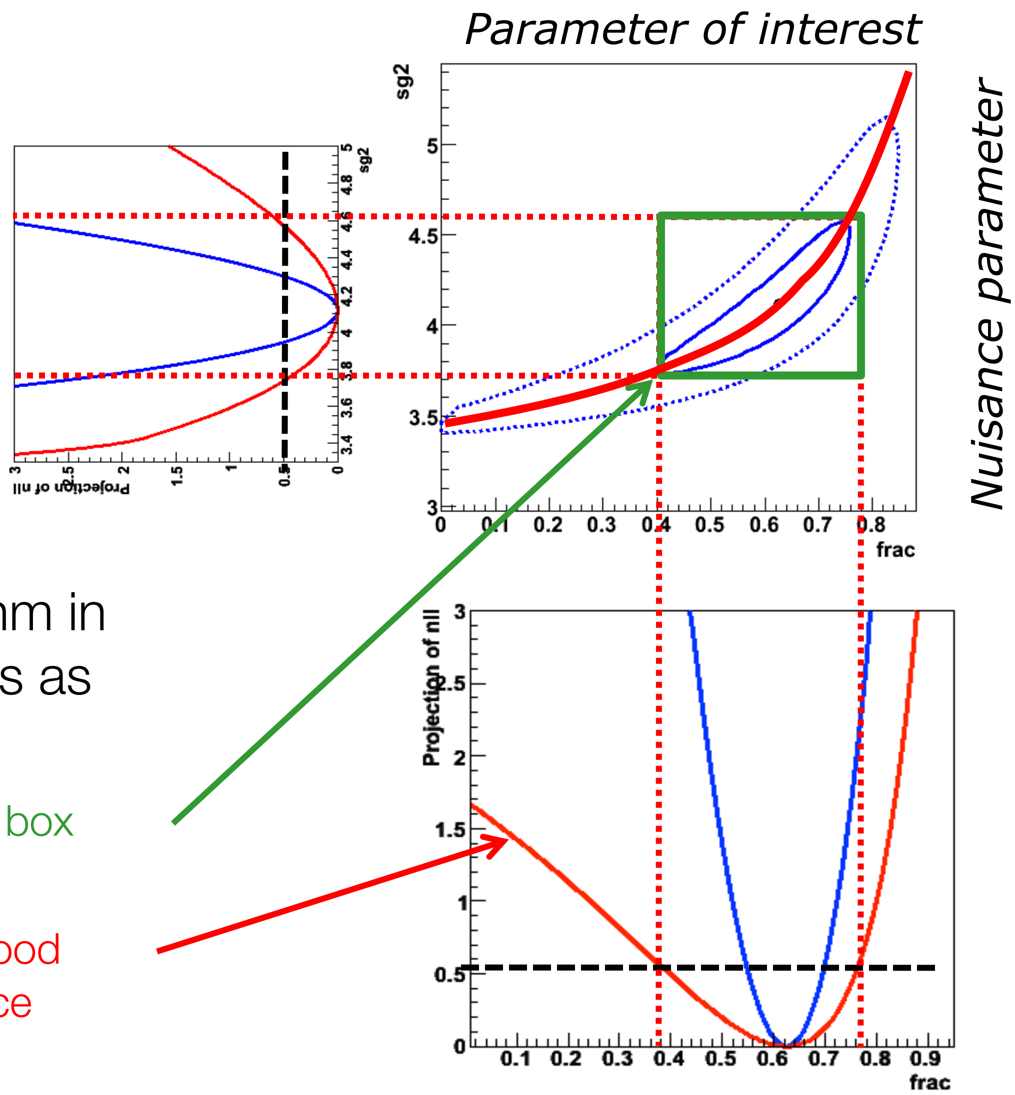




Asymptotically, distribution is identical for all μ

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

Link between MINOS errors and profile likelihood



- Note that MINOS algorithm in MINUIT gives same errors as Profile Likelihood Ratio
  - MINOS errors is bounding box around $\lambda(s)$ contour
  - Profile Likelihood = Likelihood minimized w.r.t. all nuisance parameters

NB: Similar to graphical interpretation of variance estimators, but those always assume an elliptical contour from a perfectly parabolic likelihood

Summary on NPs in confidence intervals

- Exact confidence intervals are difficult with nuisance parameters
  - Interval should cover for any value of nuisance parameters
  - Technically difficult and significant over-coverage common
- LHC solution Profile Likelihood ratio → Guaranteed coverage at *measured* values of nuisance parameters only
  - Technically replace likelihood ratio with profile likelihood ratio
  - Computationally more intensive (need to minimize likelihood w.r.t all nuisance parameters for each evaluation of the test statistic), but still very tractable
- Asymptotically confidence intervals constructed with profile likelihood ratio test statistics correspond to (MINOS) likelihood ratio intervals
  - As distribution of profile likelihood becomes asymptotically independent of θ, coverage for all values of θ restored

Dealing with nuisance parameters in Bayesian intervals

- Elimination of nuisance parameters in Bayesian interval: Integrate over the full subspace of all nuisance parameters;

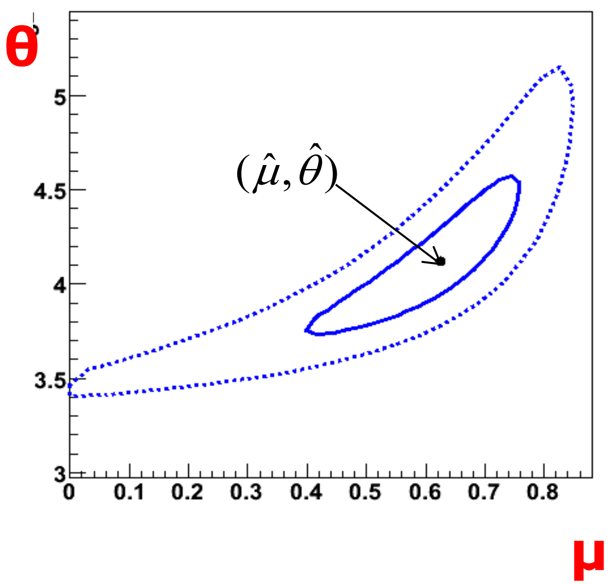
$$P(\mu \mid x) \propto L(x \mid \mu) \cdot \pi(\mu)$$

$$\Downarrow$$

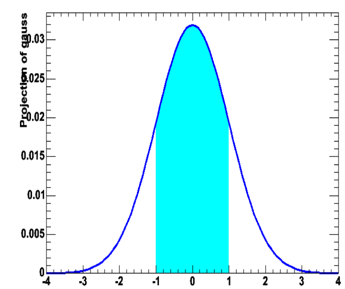
$$P(\mu \mid x) \propto \int \left(L(x \mid \mu, \vec{\theta}) \pi(\mu) \pi(\vec{\theta}) \right) d\vec{\theta}$$

- You are left with posterior pdf for μ

$\int$



$\times \; \pi(\mu, \theta) =$



Credible interval: area that integrates X% of posterior

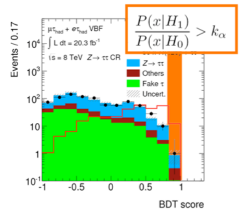
Computational aspects of dealing with nuisance parameters

- Dealing with many nuisance parameters is computationally intensive in both Bayesian and (LHC) Frequentist approach
- Profile Likelihood approach
  - Computational challenge = ***Minimization*** *of likelihood w.r.t. all nuisance parameters for every point in the profile likelihood curve*
  - Minimization can be a difficult problem, e.g. if there are strong correlations, or multiple minima
- Bayesian approach
  - Computational challenge = ***Integration*** *of posterior density of all nuisance parameters*
  - Requires sampling of very potentially very large space.
  - Markov Chain MC and importance sampling techniques can help, but still very CPU consuming

Nuisance parameters also impact event selection optimization!

Choosing the 'best' high-signal region

- A common scenario for searches in a low-statistics regime is to perform a simplified analysis
  1. Train MVA to obtain discriminant D
  2. Apply a cut on D
  3. Perform only a counting analysis



- And a common question is then – what is the 'optimal cut on D'?
  - NB: the question arise due to choice for simplified
    If a *probability density model* is used for the analy
    'the full range of the discriminant'
  - To answer question a 'figure of merit' (FOM) mus
    the optimality of the selection. The ideal FOM for
    expected signal significance.

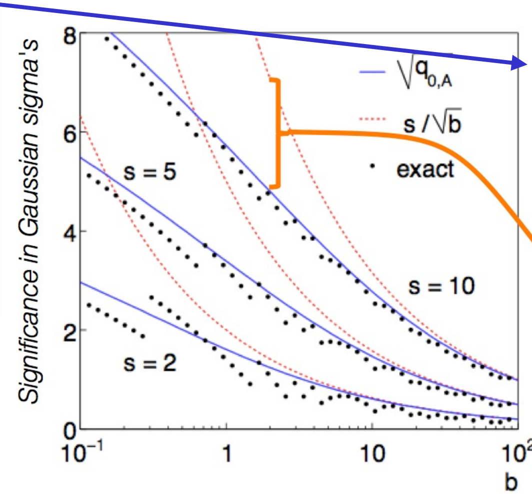
If the estimate of the background rate B is uncertain then

Figure of Merit $\sqrt{q_{0,A}}$ (and also $S/\sqrt{B}$)

overestimate counting model significance. Effect depends both on B and σ(B) → *can also effect location of optimum*

Choosing the 'best' high-signal region

- The estimated significance assuming a Poisson process modeled by Poisson(N|S+B) is $\sqrt{2\left((s+b)\ln(1+s/b)-s\right)}$.
- E.g. for 'discovery FOM' $s/\sqrt{b}$ illustration of approximation for s=2,5,10 and b in range [0.01-100] *shows significant deviations of $s/\sqrt{b}$ from actual significance at low b*



$$\sqrt{q_{0,A}} = \sqrt{2\left((s+b)\ln(1+s/b)-s\right)} \; .$$
$$= \frac{s}{\sqrt{b}}\left(1+\mathcal{O}(s/b)\right) \; .$$

Wouter Verkerke, NIKHEF

Nuisance parameters also impact event selection optimization!

Choosing the 'best' high-signal region

- A common scenario for searches in a low-statistics regime is to perform a simplified analysis

$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$

Can improve counting model significance estimate used as Figure of Merit *by including background uncertainty* (if known and sizable)

Approximate counting probability model with B uncertainty as

$\mathrm{Poisson}(N_{on}|\mu S+B)$ **$\mathrm{Poisson}(N_{off}|\tau B)$**

NB: Assumes Poisson (not Gaussian) model for B uncertainty.
For x% fractional uncertainty on B choose

$N_{off}=1/x^2$ and $\tau=N_{off}/B_{nom}$ → $\hat{B}=B_{nom}$, $\sigma(\hat{B})=x\%$

Signal significance for this model is analytically known in terms of the 'Incomplete Beta funtion'

→ Easy to use implementation in ROOT (returns significance Z)

```
RooStats::NumberCountingUtils::BinomialObsZ(Double_t nObs,
                                  Double_t bExp, Double_t fracBUnc) ;
```

isson process modeled

f approximation for
s significant deviations of

$\sqrt{2\left((s+b)\ln(1+s/b)-s\right)}$.

$\frac{s}{\sqrt{b}}(1+\mathcal{O}(s/b))$.

b

Summary of statistical treatment of nuisance parameters

- Each statistical method has an associated technique to propagate the effect of uncertain NPs on the estimate of 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 are minimizing the likelihood you are usually a Frequentist
  - 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
- Don't forget large nuisance parameters in your event selection optimization