(Combined) fits of LHC data – statistical issues & systematics

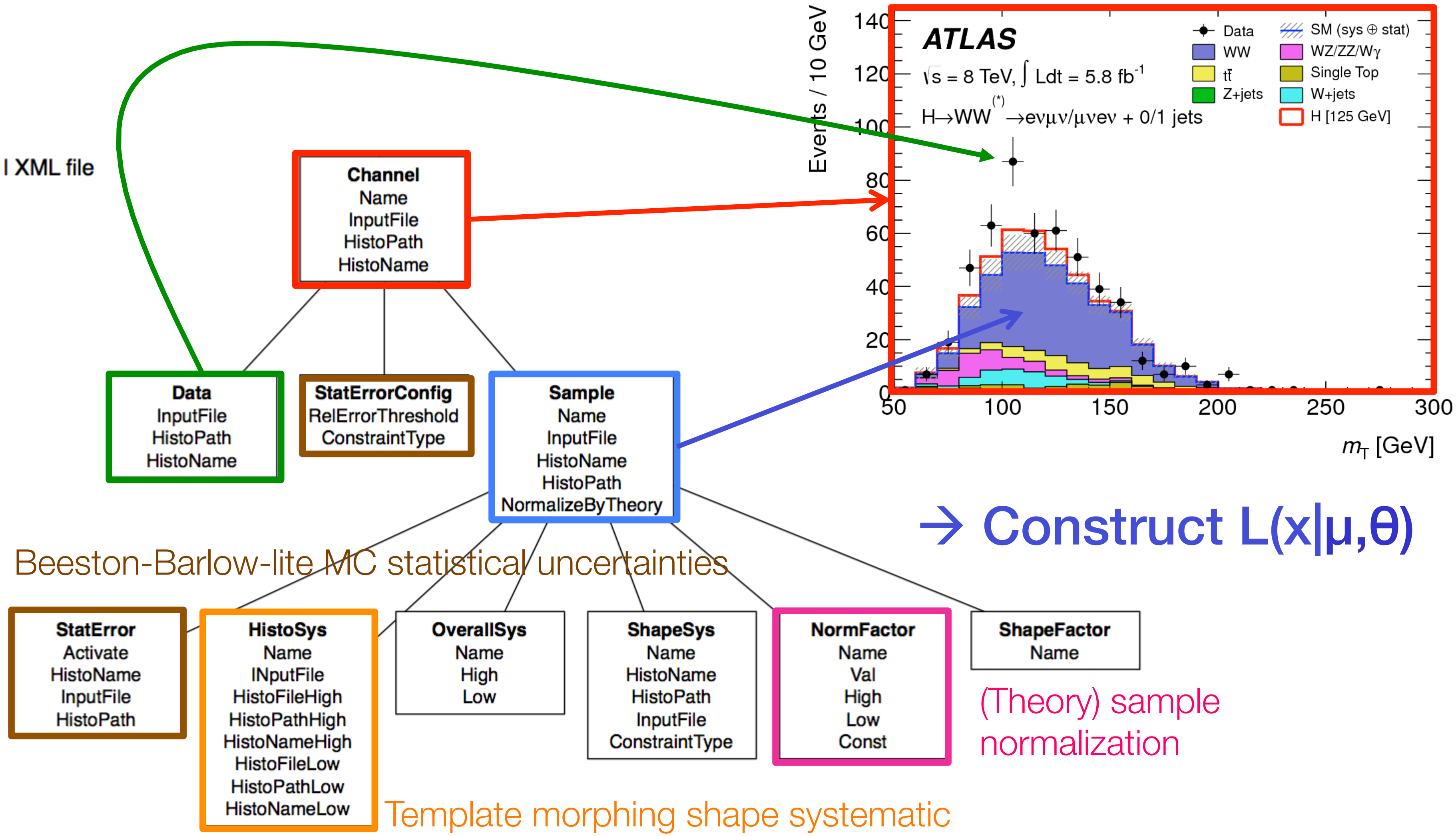
Wouter Verkerke (Nikhef/UvA)

Introduction – Understanding LHC measurements

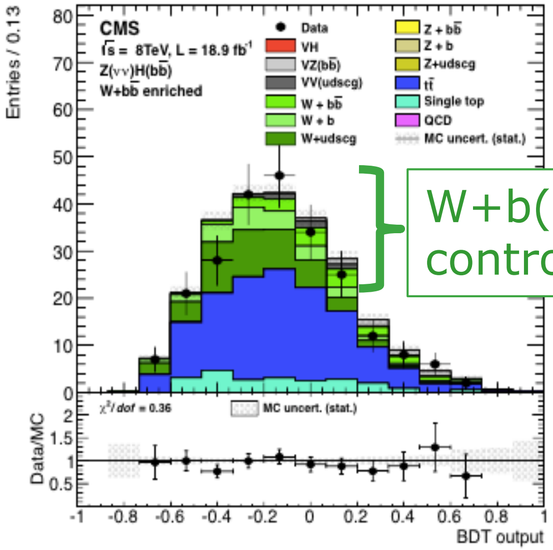
- LHC has produced an abundance of measurements that are extremely useful for further (combined) interpretation
- Many groups are doing so – both inside and outside the experiments
- In this presentation I want to show some aspects of what goes on 'under the hood' of typical published measurements
- With particular focus on statistical methods and treatment of systematic uncertainties
- Some of these issues will also be relevant (or at least good to be aware of) when interpreting these results, or using them as inputs for further combined fits

Anatomy of a ‘typical’ LHC measurement

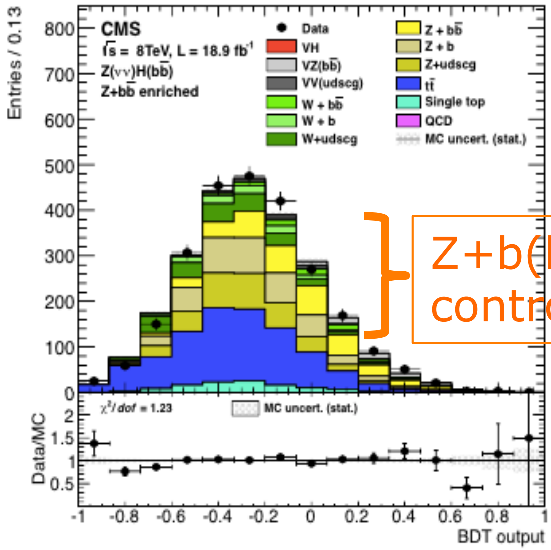
- Hierarchy of concepts for description of one measurement channel



A single measurement is often *many* measurements

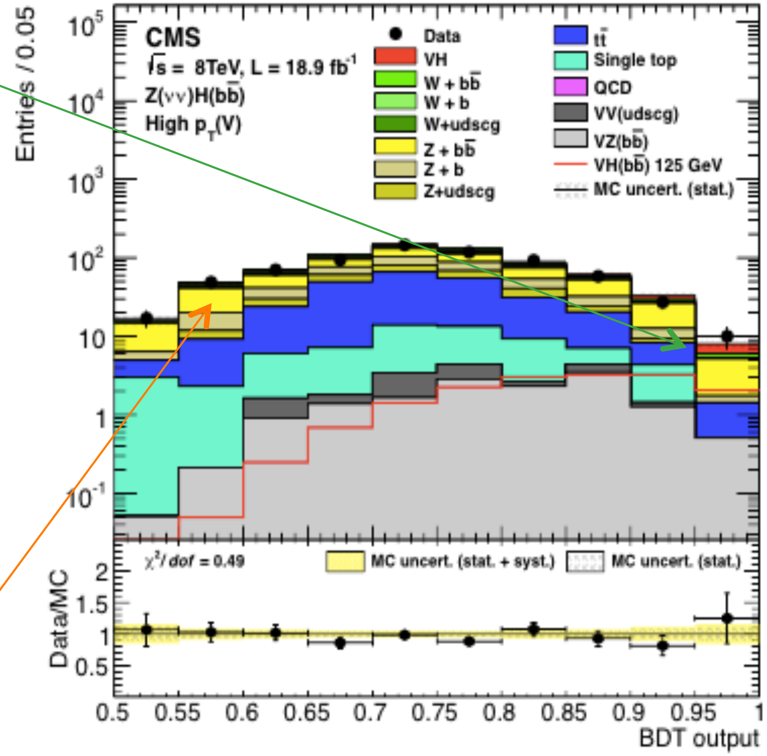


W+b(b) enriched control region



Z+b(b) enriched control region

PRD 89 (2014) 012003



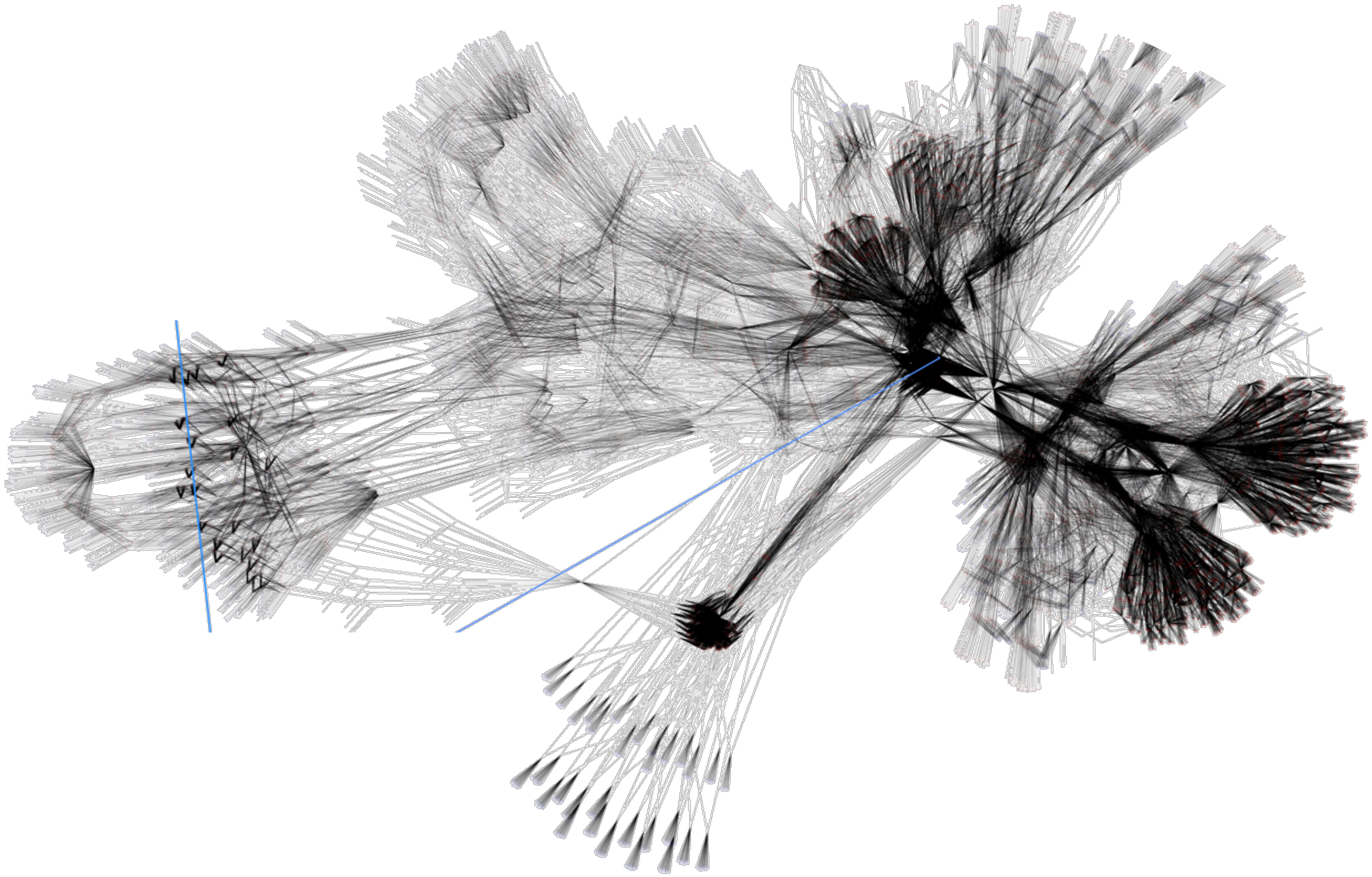
$$L_{full}(x|\mu,\theta) = L_A(x|\mu,\theta)\, L_B(x|\mu,\theta)\, L_C(x|\mu,\theta)\ \ldots$$

Information provided by measurements

- For most LHC measurements, the full likelihood model contains a lot more detail than is published
  - Typically O(10) to O(100) of nuisance parameters that quantify impact of various systematic uncertainties on each of the analysis regions of the event.
  - Not all needed to understand and interpret result, but can sometimes give extra useful information when you intend to combine measurement with other measurement → Allows for precise implementation of what NPs should be treated as correlated, or uncorrelated
- Full likelihood information of most LHC measurements is often digitally preserved (internally) in experiments
  - Allows to combine results in a later stage with full preservation of detail
  - Most famous example is Higgs boson combined measurements
  - But happens for many other measurements

A visual map of the Run-1 ATLAS Higgs combination

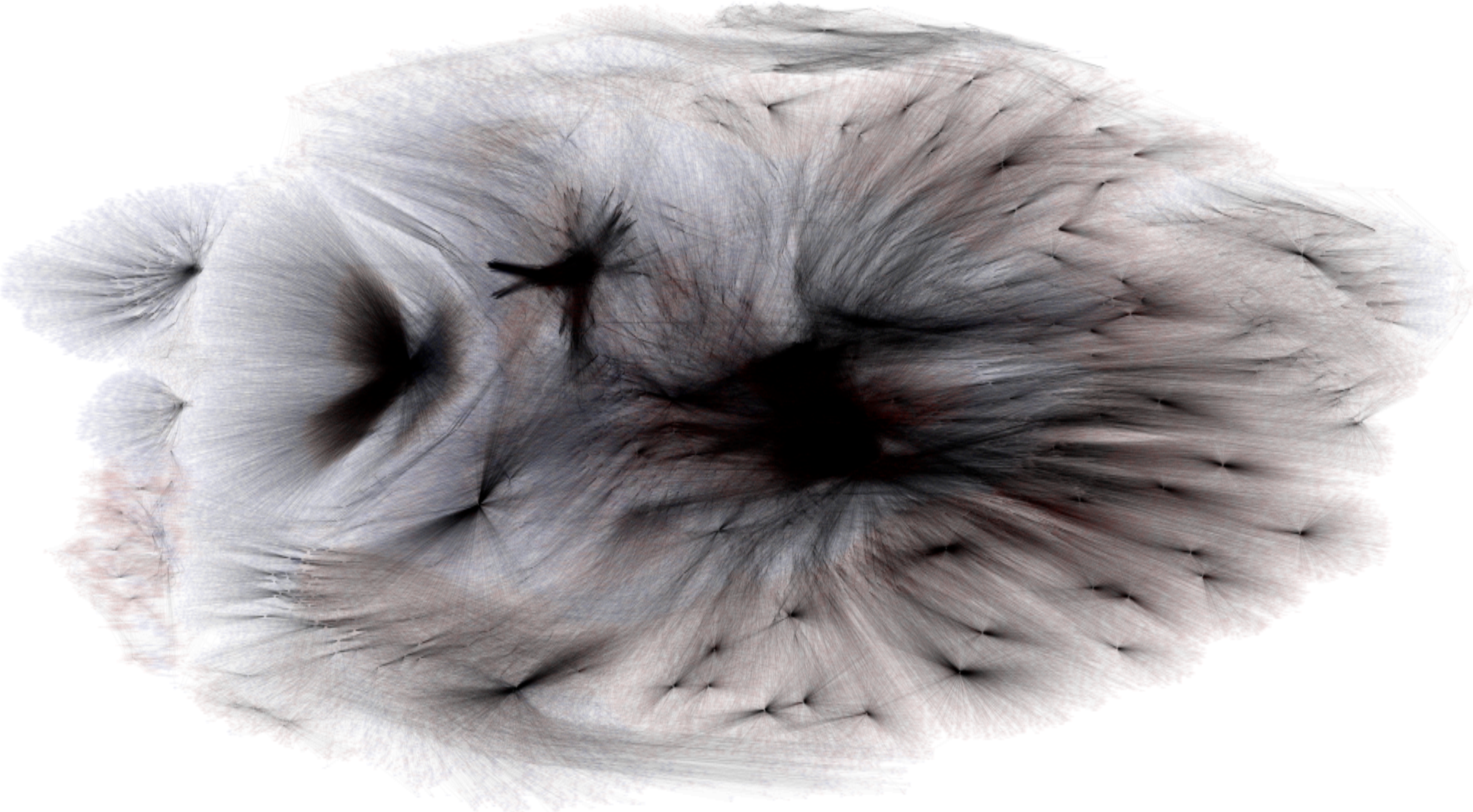
Every edge is a parameter (~1k), every node is function or pdf (~20k)



Wouter Verkerke, NIKHEF

A visual map of the Run-2 ATLAS Higgs combination

Every edge is a parameter (~2k), every node is function or pdf (~100k)



Information from LHC (combined) measurements

- Full level of detail not published, and most of it is very technical (and usually not needed after publication)
- For many measurements, details also published on HEPdata (but usually not everything)
- But detailed likelihood-level combinations allow to study effects that are otherwise not easily studied
- Will attempt to summarize some the pitfalls, points of attention, and lessons learned in this presentation
  - It is by no means exhaustive
  - It may also not all be relevant for you
  - But nevertheless some of this is useful to you I hope!

Combined measurement strategies inside (or between) LHC results

- Most combined results follow one of 3 classes
  1. Full-detail likelihood-based combination
  2. Likelihood combination with simplified parametrized likelihood formulations of existing results (assuming Gaussian/Linear modeling)
  3. BLUE (variants) of combination
- Hybrid combinations are also performed (e.g. Gaussian/Linear likelihood constructed from a covariance-formulated result with a full likelihood)
  - In many cases the difficult problems do not so much arise from the precise formulation of the likelihood function (many measurements are Gaussian/ Linear and hard-coding that assumption does not hurt n such cases)
  - But the difficult problems are in the formulation of systematic uncertainties and their (partially) correlated implementation of effects in combined measurements

Unfolding measurements

- A special class of LHC measurements – from the statistical treatment perspective – are unfolding measurements
- Typical ‘unfolding’ LHC measurements follows multi-step approach
  1. Modeling of signal and background in (2 or more)-D distribution
  2. Subtraction of estimated background component from data
  3. Unfolding of bkg-subtracted data from reco to particle or parton space
  4. Publication of results in covariance form
- While detailed information of systematic effects (through NP parametrization) is available in 1st step – many details are lost in the subsequent processing → typically less detailed information available than for full likelihood-based measurement.
  - This limitation is not fundamental. New unfolding tools can preserve information → transport uncertainties from reco-level through unfolding matrix to parton level. But fairly new, and therefore not much used yet

Formulating uncertainties – Hessian vs PLL approach

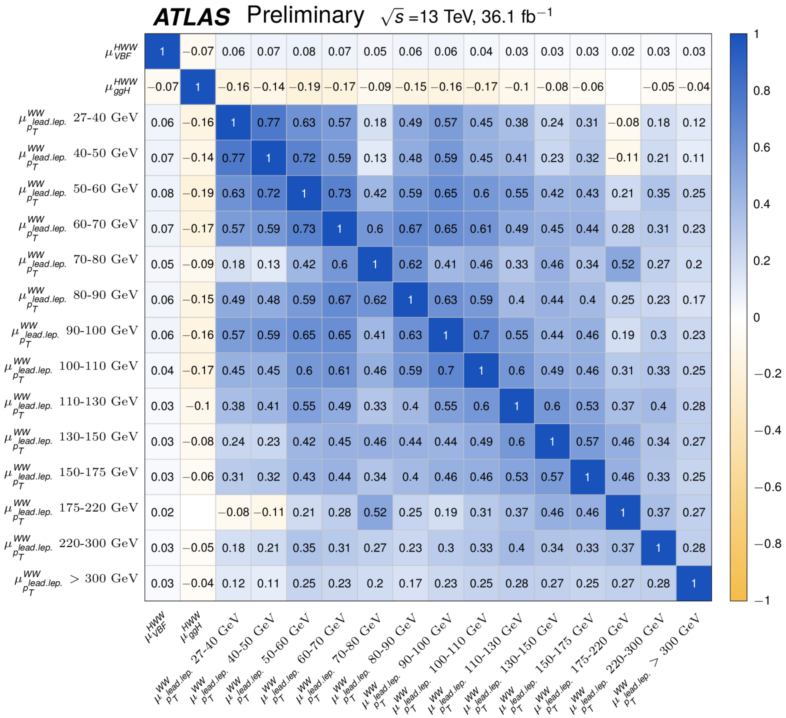
- In the regime of Gaussian uncertainties with a linear response, Hessian and Profile Likelihood formulation are equivalent
  - Almost all measurements at LHC are in this regime
  - Searches and low-statistics measurements not necessarily
- Nevertheless – majority of LHC measurements are expressed in PLL formalism
  - Main exception are unfolded results (as mentioned before)
- As lot of the ‘tricky’ aspects of systematics are in the response function and not in the subsidiary measurement, it is often natural to work with the full likelihood in the PLL formalism

A experimentalist guide to combining results

- Golden Rule: **Avoid (almost) intractable statistical problems through thorough advance preparation of the combination strategy**
  - Issues we *really* try to avoid include
- Overlapping event selections
  - *Every event should only appear at most once in the likelihood*
  - Allowing events to be used more than once leads to underestimation of statistical uncertainties
  - This is a really hard problem that requires significant study of event selection of measurements with dedicated samples
  - It often also requires non-trivial adjustments to existing analysis to explicitly remove such overlap
- Systematic uncertainties with an undefined degree of correlation
  - For example: what is correlation in flavor-tagging efficiency algorithm at 70% efficiency operating point and at 80% operating?
  - Could in principle be studied and resolved – but very significant amount of work and flavor tagging groups in experiments will refuse to take on such tasks
  - But much more intractable variants exist for theoretical uncertainties. For example Stewart-Tackman vs JetVeto procedures, or hadronization uncertainties evaluated with different MC generators.
  - General strategy: 1) coordinate in advance to avoid. 2) if that failed, retroactive change analyses to common strategy.

A non-trivial combination example

- Joint interpretation of pp→H→WW and pp→WW measurements
- Different input formulations:
  - Higgs measurement in full likelihood form, SM measurement in covariance form
  - Solution: express SM measurement as Gaussian likelihood
  - Harmonize/match nuisance parameters
- Overlapping samples:
  - Non-resonant WW production used in H→WW sample as control sample
  - Sololution: remove HWW control sample. reformulate HWW likelihood to use corresponding information from SM



Combined effective field theory interpretation of $H \rightarrow WW^*$ and WW measurements using ATLAS data

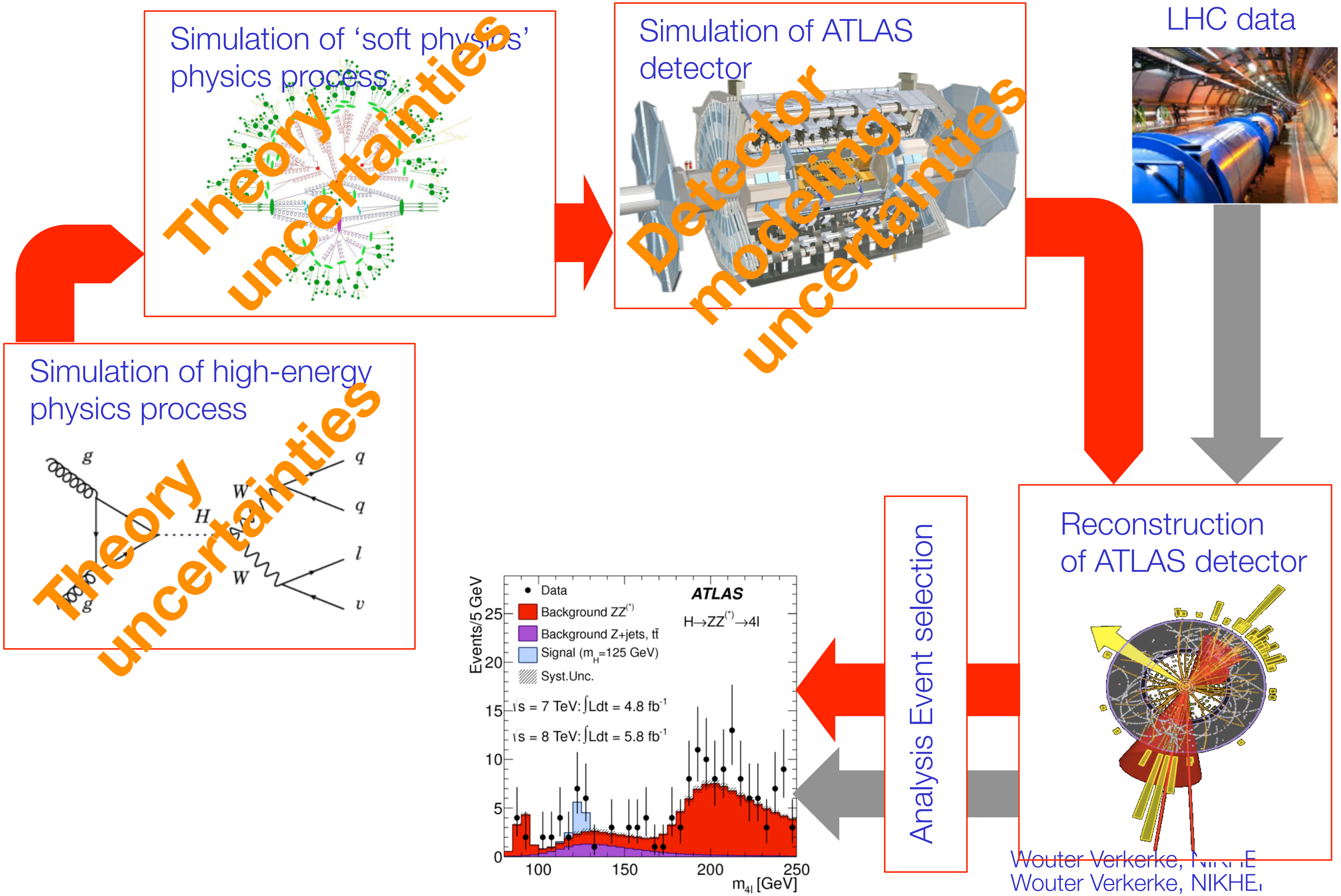
ATL-PHYS-PUB-2021-010

Points of attention in combining measurements

- After having avoided the ‘intractable’ problems through proper preparation, still plenty of issues left
- Most of these are in the domain of formulation of systematic uncertainties in input measurements
- Even if definitions are harmonized – still plenty of room for surprising issues

- In the remainder of this presentation will focus mostly on definition, implementation and effect of systematic uncertainty representations.

The simulation workflow and origin of uncertainties



Wouter Verkerke, NIKHEF

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.

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*

Formalizing systematic uncertainties

- The original systematic uncertainty prescription

“the JES uncertainty is 5%”

- The formalized prescription for use in statistical analysis

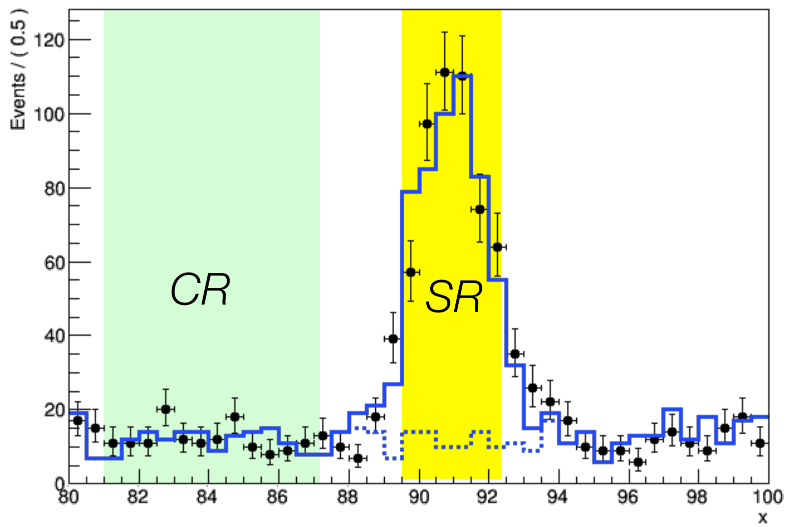
“There is a calibration parameter in the likelihood
of which the true value is unknown

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% ”

The sideband measurement

- Suppose your data in 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'

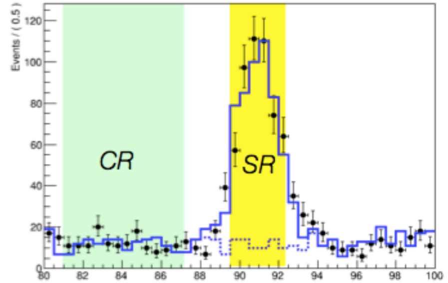
$$L_{full}(s,b) = Poisson(N_{SR} \mid s+b) \cdot Gauss(\tilde{b} \mid b, 0.08)$$

'**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)

$$L(N,\tilde{\alpha} \mid s,\alpha) = Poisson(N \mid s + \tilde{b}(1+0.02\alpha/\tilde{\alpha})) \cdot Gauss(\tilde{\alpha} \mid \alpha, \sigma_\alpha)$$

Signal rate (our parameter of interest)

Observed event count

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

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

Nominal calibration

Assumed calibration

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

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

Modeling a detector calibration uncertainty

- Simplify expression by renormalizing "subsidiary measurement"

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

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]

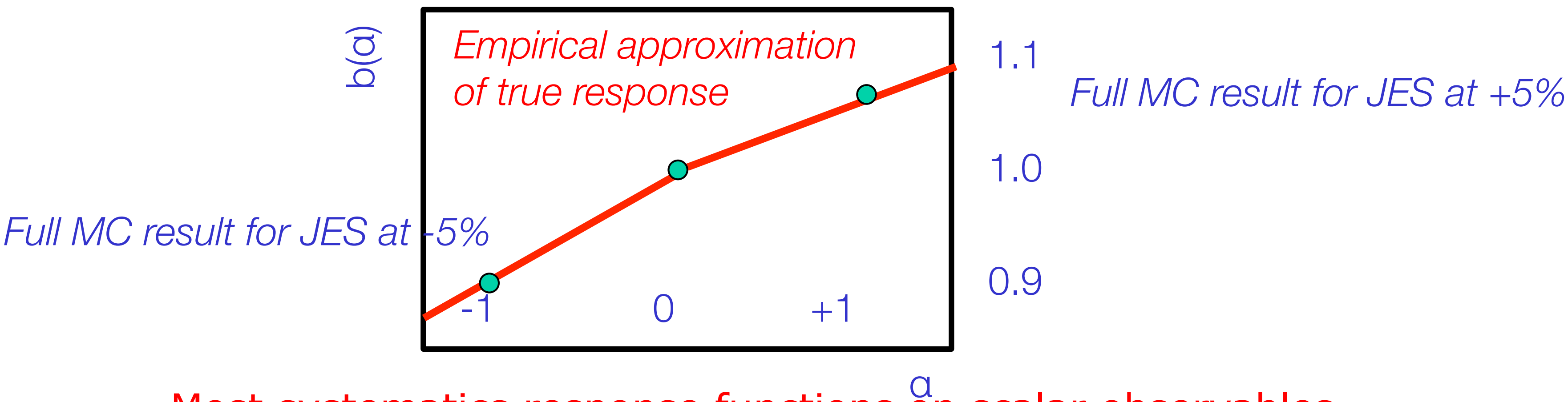
"Normalized subsidiary measurement"

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

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

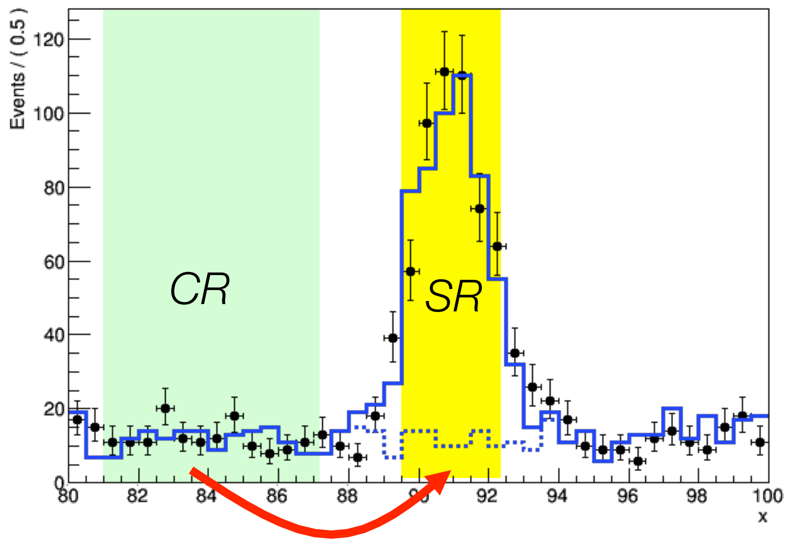
- Note that the response function is generally not linear, but can in principle *always be determined by your full simulation chain*
  - But you cannot run your full simulation chain for any arbitrary 'systematic uncertainty variation' → Too much time consuming
  - Typically, run full MC chain for nominal 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%



Most systematics response functions on scalar observables are implemented as piece-wise linear functions

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!



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

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

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

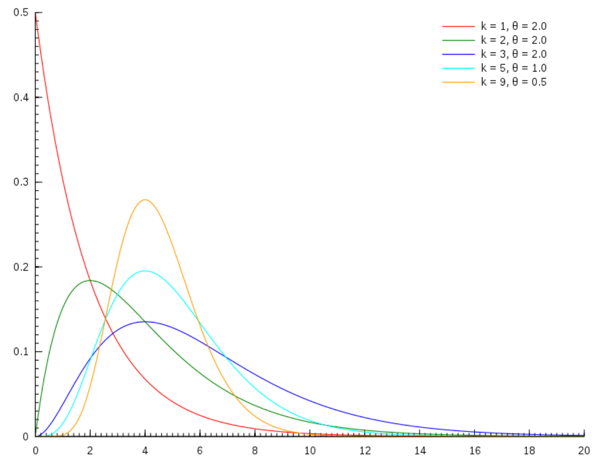
$$L(N, N_{ctl}, 0 \mid s, b, \alpha_{JES}) = Poisson(N \mid s+b) \cdot Poisson(N_{ctl} \mid \underbrace{\tau(1+X\alpha_{JES})}_{\text{JES response model for ratio } b_{SR}/b_{CR}} \cdot b) \cdot \underbrace{Gauss(0 \mid \alpha_{JES}, 1)}_{\text{Subsidiary measurement of JES response parameter}}$$

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

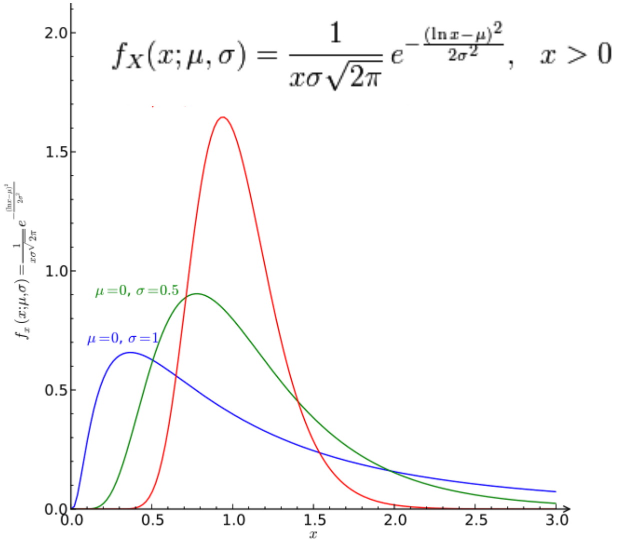
Subsidiary measurement of JES response parameter

Overview of common subsidiary measurement shapes

- Gaussian G(x|μ,σ)
  - 'Default', motivated by Central Limit Theorem (asymp dist for sum of random variables)
- (Rescaled) Poisson P(N|μτ) →
  - 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|μ,σ) →
  - Asymptotic distribution for product of random variables
  - Appealing property for many applications is that it naturally truncates at x=0



Additive vs Multiplicative systematic uncertainties

- Additive vs Multiplicative effect of systematic effects is expressed in the response function.

- Additive response function with Gaussian subsidiary

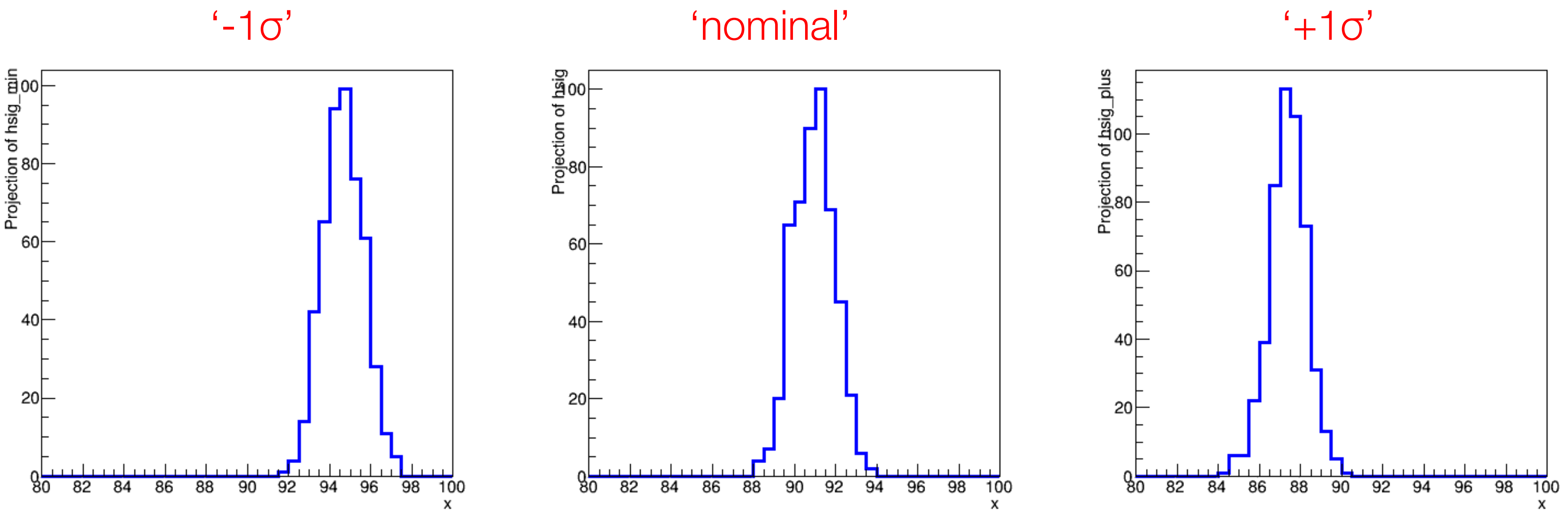
$$L(N \mid s,\alpha) = Poisson(N \mid s + \tilde{b}(1+0.1\alpha)) \cdot Gauss(0 \mid \alpha,1)$$

- Multiplicative response function with LogNormal subsidiary
  - LogNormal cannot go negative, but asymptotically Gaussian away from 0

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

Modeling of shape systematics in the likelihood

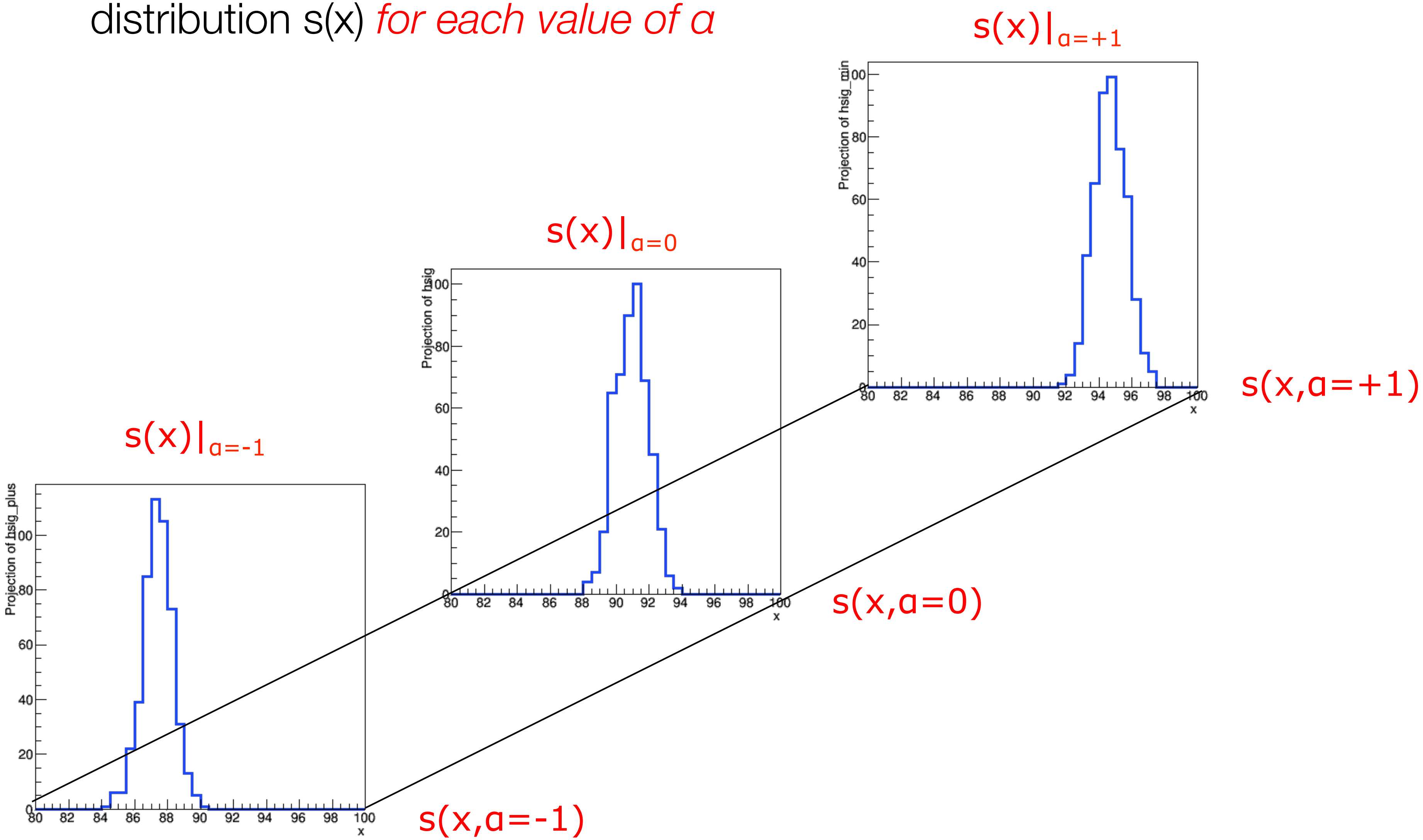
- Effect of *any* systematic uncertainty that affects the shape of a distribution can in principle be obtained from MC simulation chain
  - Obtain histogram templates for distributions at '+1σ' and '-1σ' settings of systematic effect



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

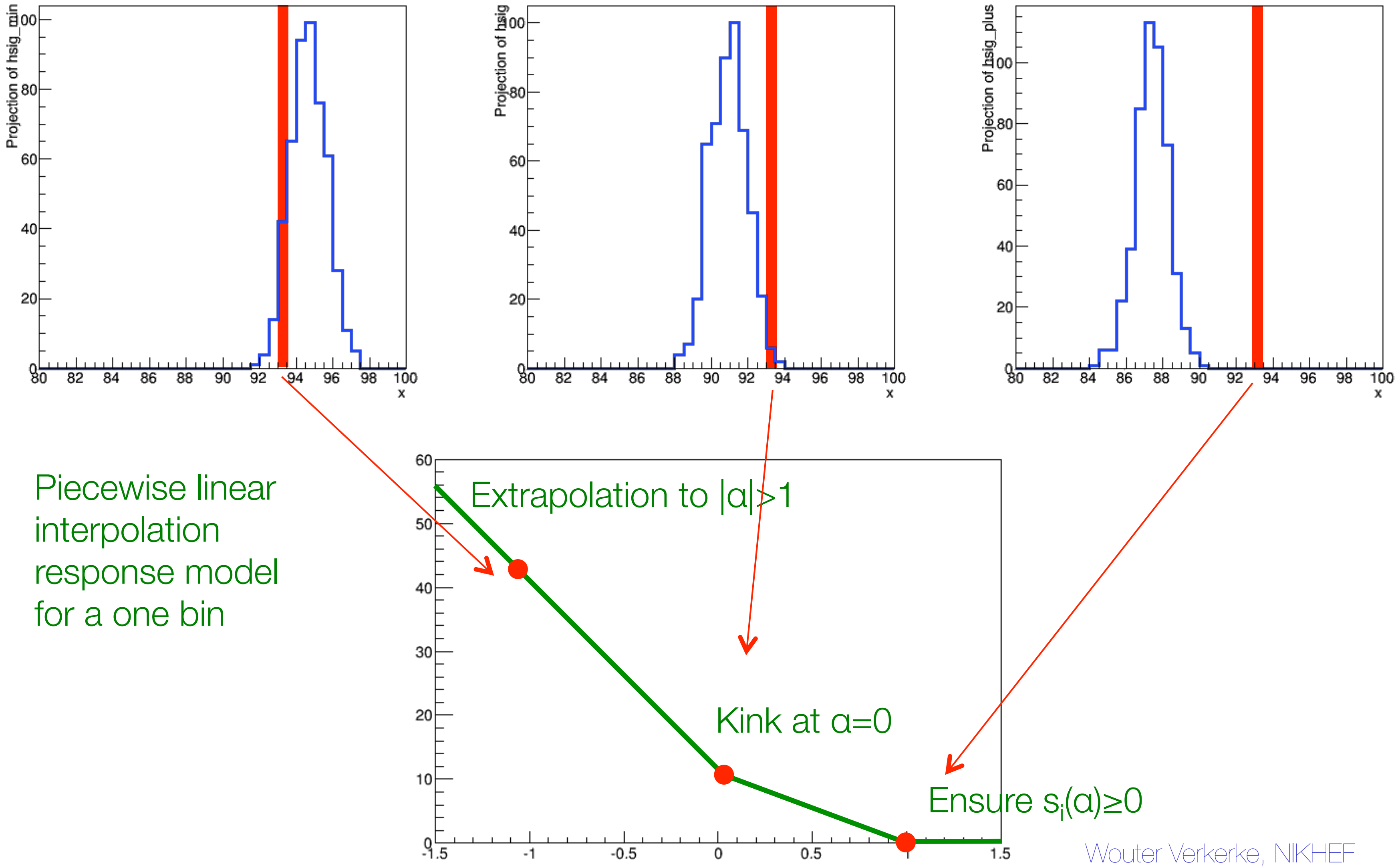
Need to interpolate between template models

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



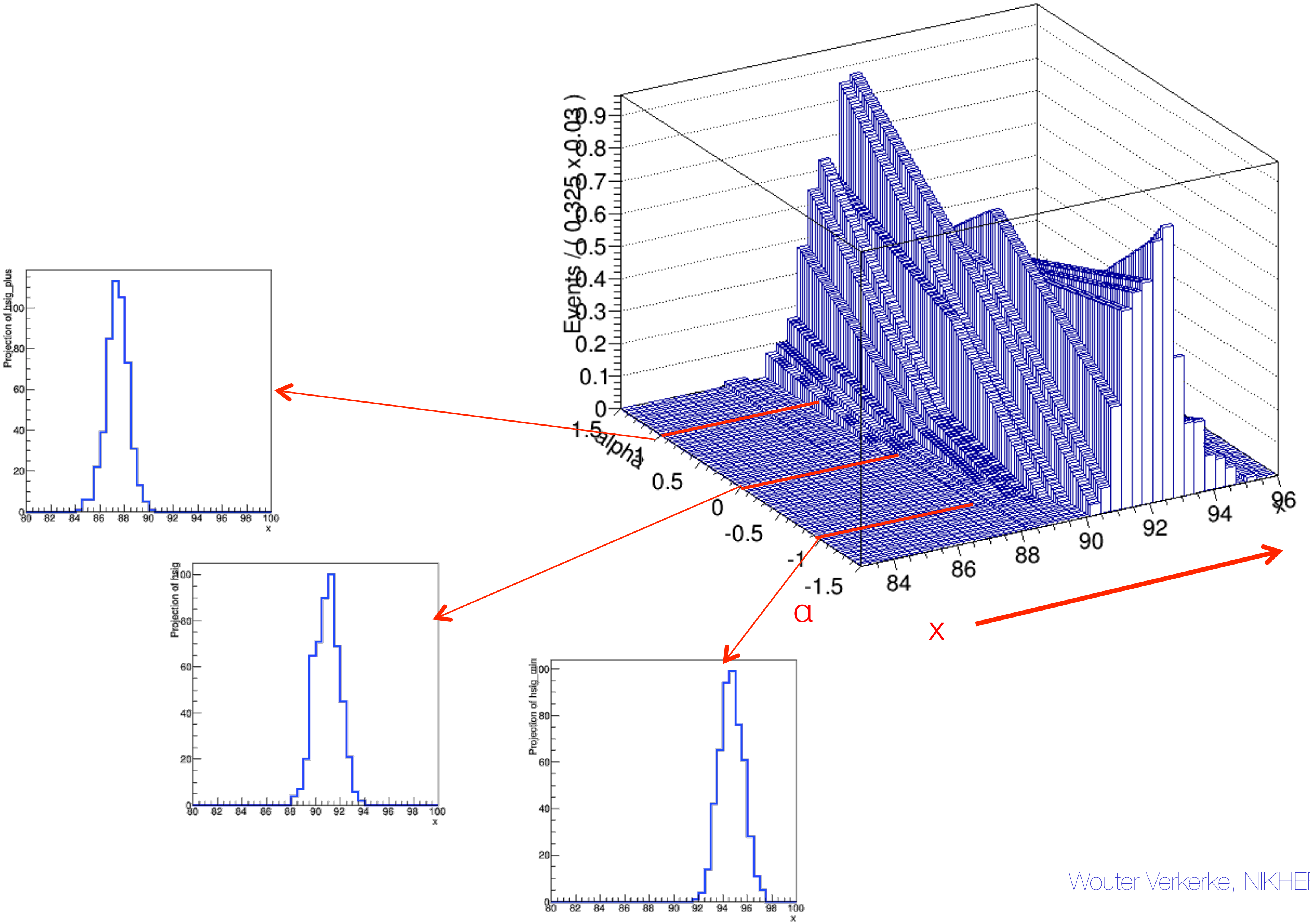
Piecewise linear interpolation

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



Piecewise linear interpolation response model for a one bin

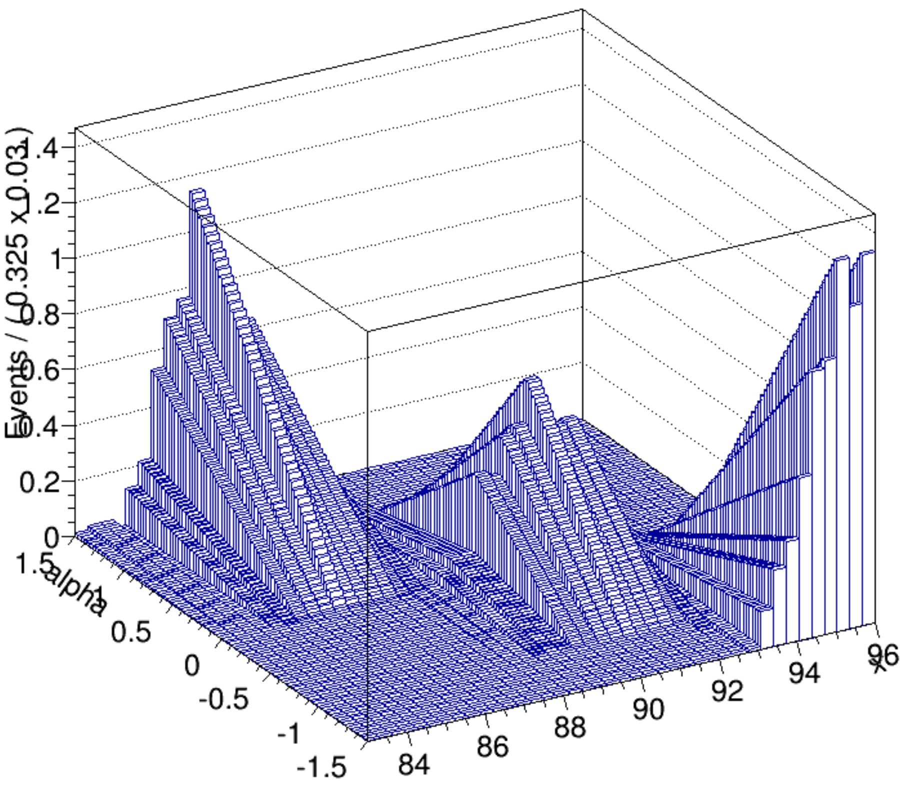
Visualization of bin-by-bin linear interpolation of distribution

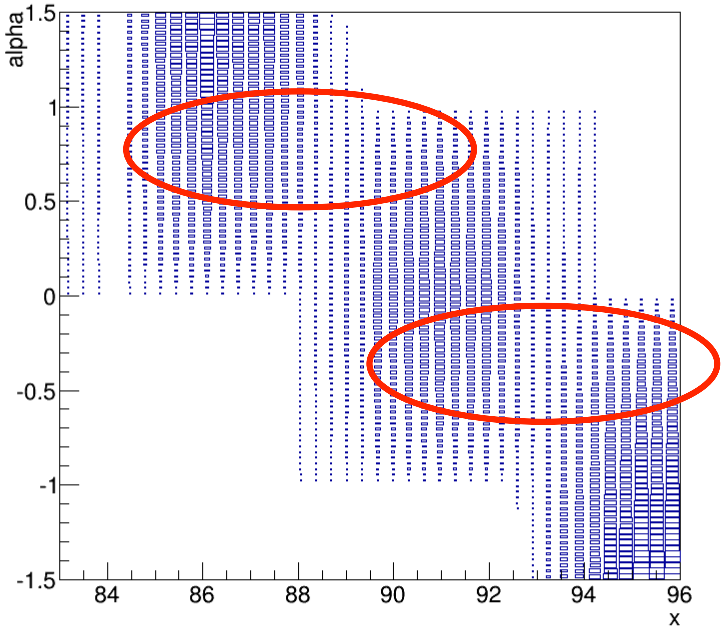


Limitations of piece-wise linear interpolation

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

Note double peak structure around |α|=0.5

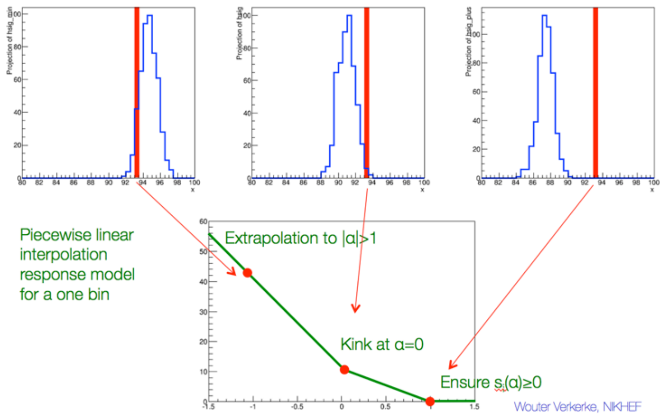




Piece-wise interpolation for >1 nuisance parameter

- 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

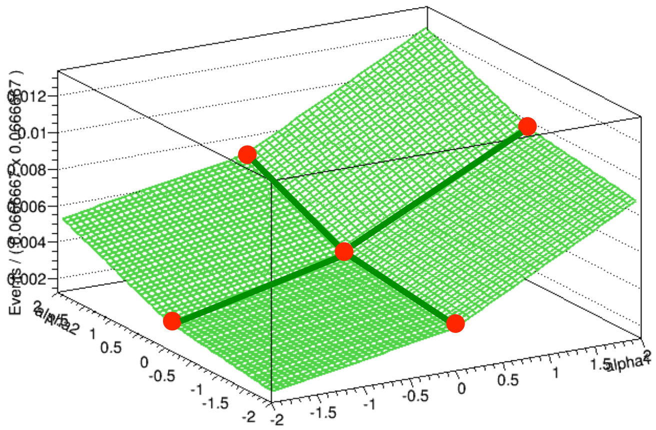
$$s_i(\alpha) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$



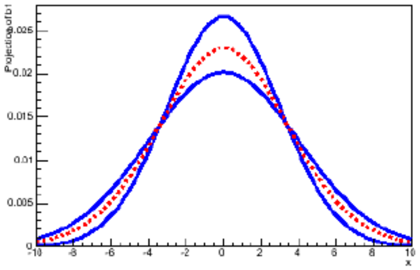
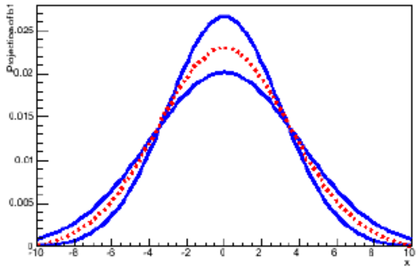
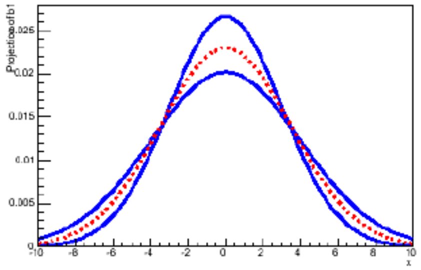
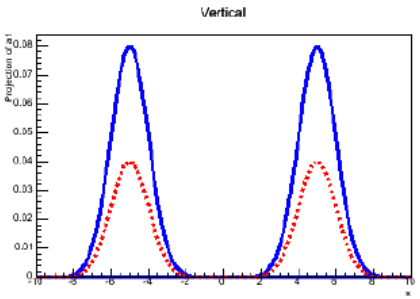
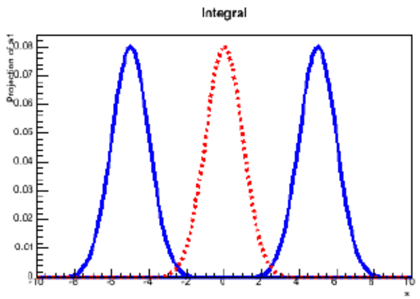
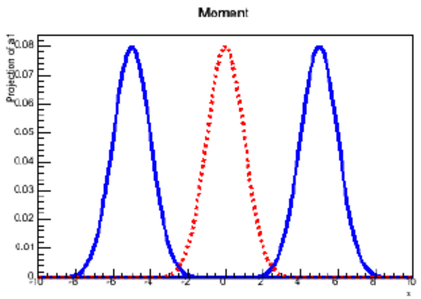
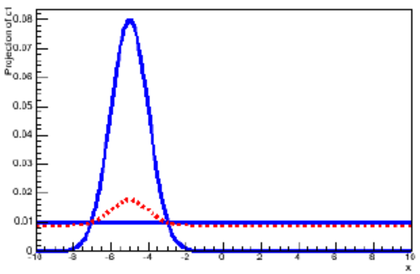
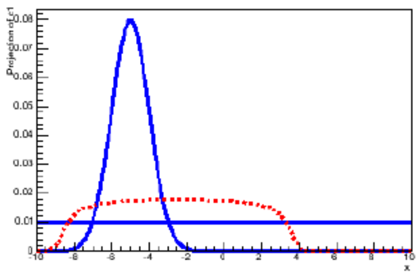
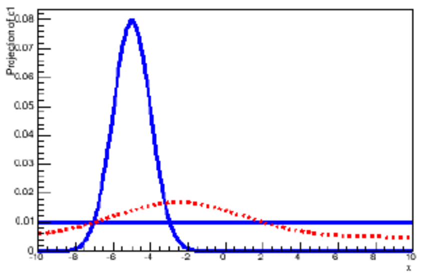
  - N-parameter interpolation

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

Visualization of 2D interpolation



Comparison of morphing algorithms

	Vertical Morphing	Horizontal Morphing	Moment Morphing
Gaussian varying width			
Gaussian varying mean			
Gaussian to Uniform (this is conceptually ambiguous!)			
n-dimensional morphing?	✔	✗	✔

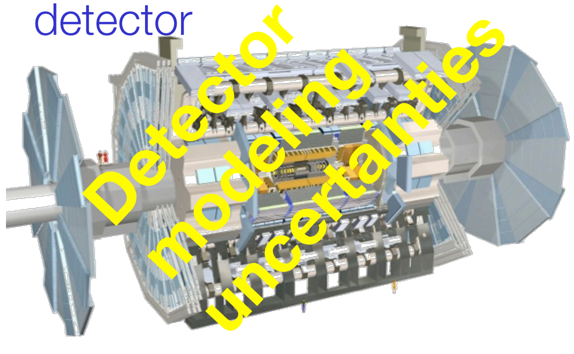
Systematic uncertainties – what exactly is uncertain?

- Quite a bit of technical machinery involved implementing shape systematics in measurements, but good tools for this and technical aspects generally very well under control
- More difficult conceptual question with many systematic uncertainties is: **'what exactly is uncertain'?**
- ***For many systematic uncertainties this is a quite difficult question to answer.***
  - *For experimental systematics it does not only relate to e.g. the design of the detector, but also how calibration measurements were done.* E.g. when energy calibration of the calorimeter are performed with dijet events, energy balance is explicitly used and will introduce a correlation structure
  - For theory systematics these questions can be even more difficult to answer, as many formulations are simply proxies for concepts without well-defined physical meaning at the observable level (e.g. scale uncertainties as proxy for MHOU)

Understanding of systematic uncertainty models

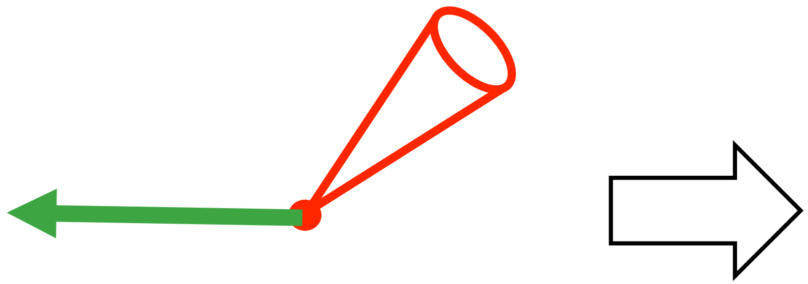
- **"Best case"**
  - Well-defined **parametric model** that define uncertain degrees of freedom in a particular aspect of the simulation
  - Well-defined **uncertainty estimates** on the set of parameters

Simulation of ATLAS detector

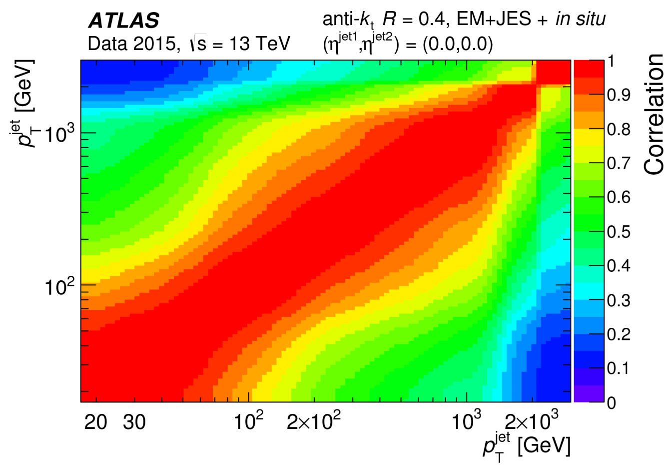


- Example – detector calibration uncertainties (e.g JES)

Measurement on calibration data (e.g. jet-γ balance)



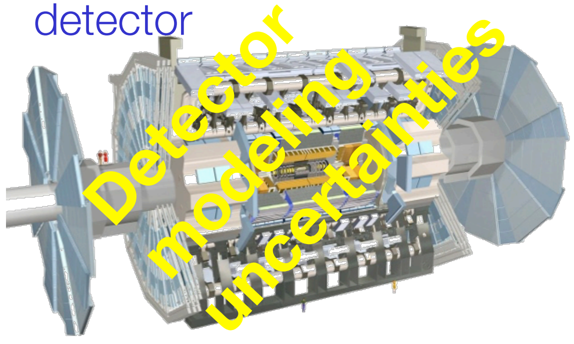
Calibration with parameterization and correlation structure motivated by underlying measurement



Understanding of systematic uncertainty models

- **"Best case"**
  - Well-defined **parametric model** that define uncertain degrees of freedom in a particular aspect of the simulation
  - Well-defined **uncertainty estimates** on the set of parameters

Simulation of ATLAS detector



- Formulation of systematic uncertainty in L of physics measurement

Likelihood describing physics measurement

Likelihood describing calibration measurement

$$L_{joint}(x_{phys}, x_{calib} | \mu, \theta_c) = L_{phys}(x_{phys} | \mu, \theta_c)\ L_{calib}(x_{calib} | \theta_c)$$

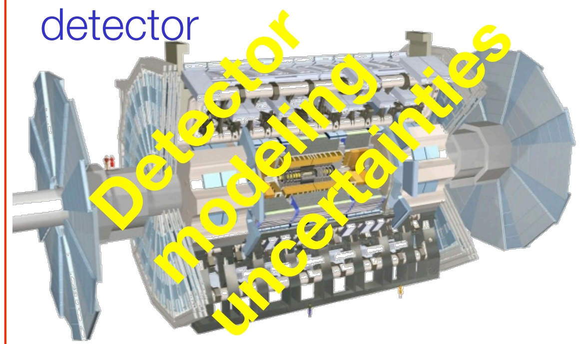
Physics parameter of interest

Uncertain Nuisance Parameters that capture systematic uncertainty

Understanding of systematic uncertainty models

- **"Best case"**
  - Well-defined **parametric model** that define uncertain degrees of freedom in a particular aspect of the simulation
  - Well-defined **uncertainty estimates** on the set of parameters

Simulation of ATLAS detector



- Formulation of systematic uncertainty in L of physics measurement

Likelihood describing physics measurement

Simplified (Gaussian) approximation of calibration measurement

$$L_{joint}(x_{phys}|\mu,\theta) = L_{phys}(x_{phys}|\mu,\theta)\, L_{calib\text{-}simplified}(0|\theta)$$

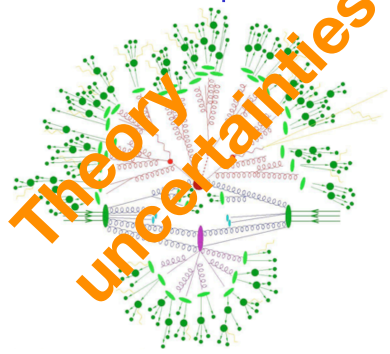
Coordinate transformation in nuisance parameters

θ_c=nom-1σ → θ=-1
θ_c=nom → θ=0
θ_c=nom+1σ → θ=+1

Understanding of systematic uncertainty models

- **"Bad case"**
  - Well-defined **parametric model** that define uncertain degrees of freedom in a particular aspect of the simulation
  - No well-motivated **uncertainty estimates** on the set of parameters

Simulation of parton showers



- Canonical example – Uncertainty on b-quark mass
  - Evaluation strategy clear – propagate impact of uncertain NP representing m_b
  - **But no clean equivalent of 'calibration measurement' that constrains magnitude of m_b** → Magnitude of uncertainty not precisely defined

$$L_{joint}(x_{phys}|\mu,\theta) = L_{phys}(x_{phys}|\mu,\theta)\ L_{calib\text{-}simplified}(0|\theta)$$

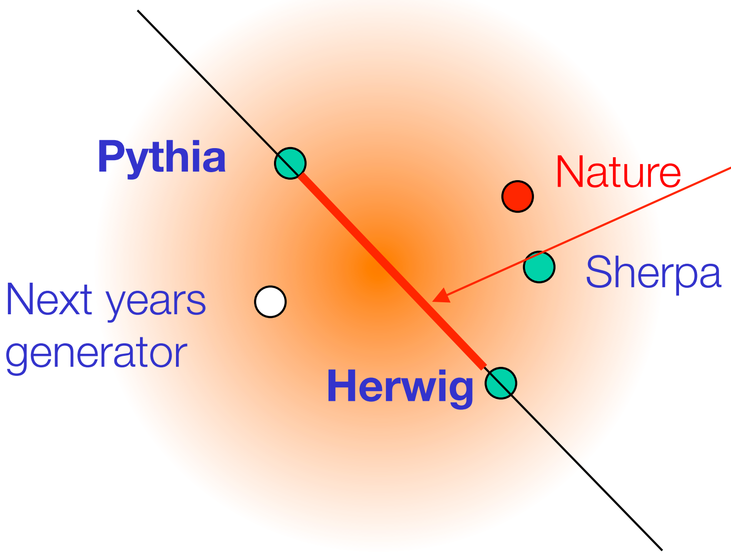
(Mostly an issue if this uncertainty dominates total measurement uncertainty)

Understanding of systematic uncertainty models

- **"Ugly case"**
  - No well defined **parametric model** that define uncertain degrees of freedom
  - No well-motivated **uncertainty estimates** on the set of parameters

Simulation of parton showers

- Canonical example – Fragmentation/Hadronization
  - We have several MC generators that give different answers that sample 'points in the theory space' that are indicative of the uncertainty



Dimensionality of problem unclear

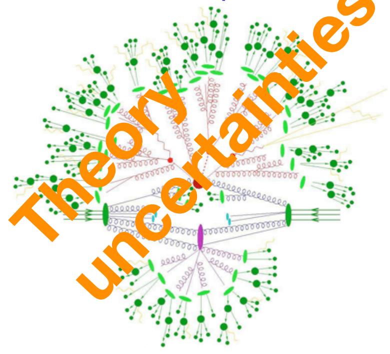
Can construct 1-parameter model with 'brute force' with morphing interpolation

'2-point systematic'

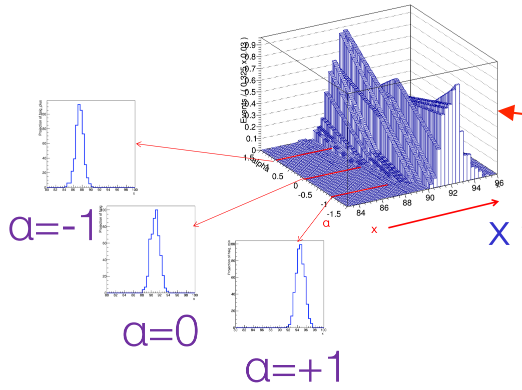
Parameterizing systematics uncertainties in the Likelihood

- **"Ugly case"**
  - No well defined **parametric model** that define uncertain degrees of freedom
  - No well-motivated **uncertainty estimates** on the set of parameters

Simulation of parton showers



- Canonical example – Fragmentation/Hadronization
  - We have several MC generators that give different answers that sample 'points in the theory space' that are indicative of the uncertainty



Nuisance parameter α shifts peak in simulated distribution of x Approximated with linear morphing between templates sampled at α=-1,0,1

Simplified Gaussian subsidiary measurement of parameter α. α is defined as the 'pull' of the original NP

$$L_{joint}(x_{phys}|\mu,\alpha) = L_{phys}(x_{phys}|\mu,\alpha)\ \mathrm{Gauss}(0|\alpha,1)$$

Build good response model for systematic uncertainties

- A lot of work has happened in the past 10 years for experimental uncertainties
- Detailed response models reflect detailed knowledge of calibrations. Can also provide a variable level of detail (e.g. a counting measurement would never need more than 1 NP for any systematic)
- Theory uncertainties remain a challenging topic (but more on that later)
- Getting the level of detail right is crucial*. It is surprisingly easy to underestimate the impact of systematic effects through improper modeling (either too simple, or too elaborate)*

Understanding your fit – beware unintended effects

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

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

- Our mental picture:

Physics measurement → "measures s"

Subsidiary measurement → "measures α"

"dependence on α weakens inference on s"

- **Is this picture (always) correct?**

Understanding your model – what constrains your NP

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

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

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

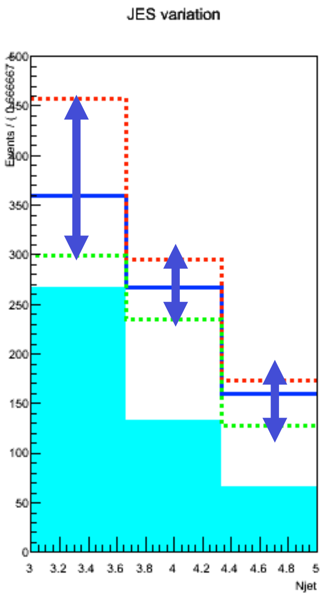
Understanding your model – what constrains your NP

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

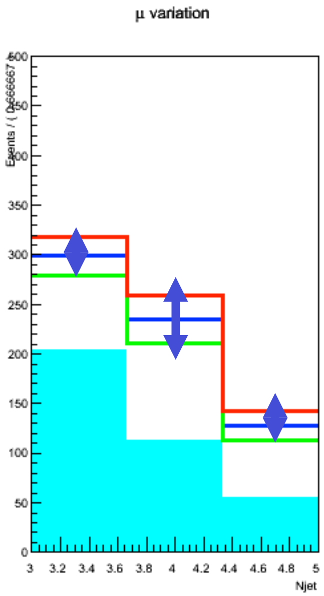
$$L(\vec{N} \mid \mu, \alpha_{JES}) = \prod_{i=3,4,5} Poisson(N_i \mid (\mu \cdot \tilde{s}_i \cdot + \tilde{b}_i) \cdot r_s(\alpha_{JES}))) \cdot Gauss(0 \mid \alpha_{JES}, 1)$$

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

Effect of changing α_{JES}

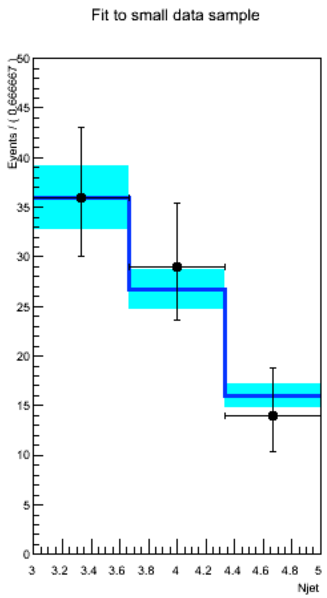


Effect of changing μ

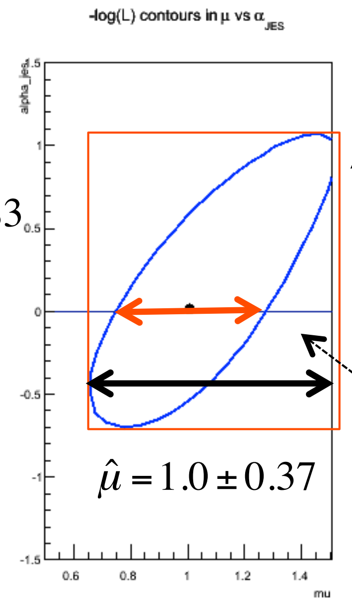


Understanding your model – what constrains your NP

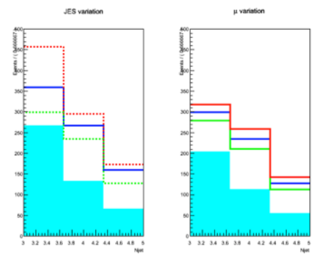
- Now measure (μ,α) from data – 80 events



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



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



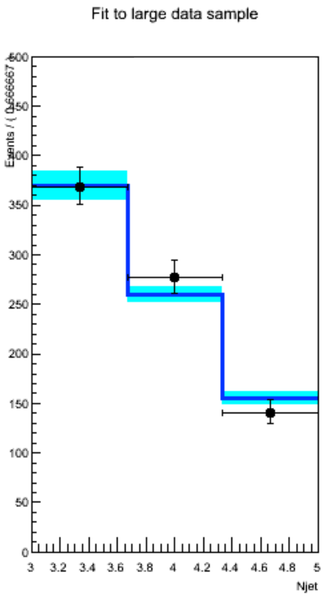
Estimators of μ, α correlated due to similar response in physics measurement

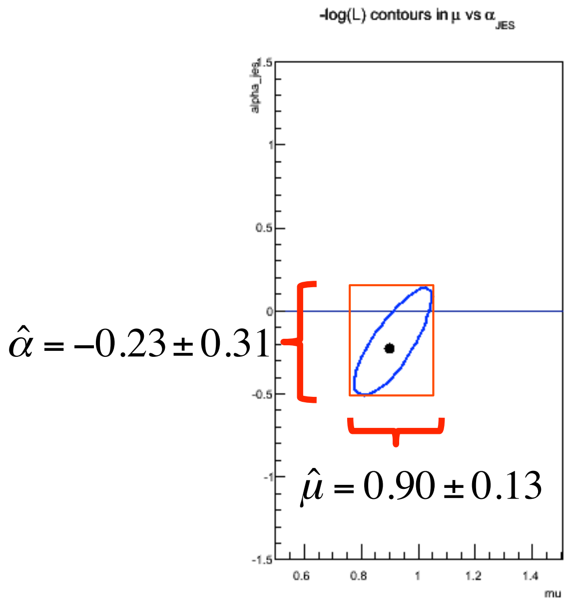
Uncertainty on μ with/without effect of JES

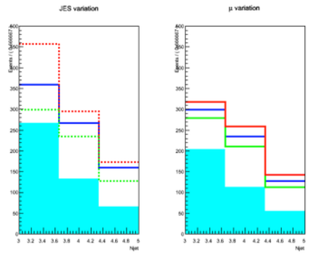
- Is this fit OK?
  - Effect of JES uncertainty propagated in to μ via response modeling in likelihood. Increases total uncertainty by about a factor of 2
  - Estimated uncertainty on α is not precisely 1, as one would expect from unit Gaussian subsidiary measurement…

Understanding your model – what constrains your NP

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







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. (α = 0 ± 1)
  - Because the physics likelihood can measure it better than the subsidiary measurement (the effect of μ, α are sufficiently distinct that both can be constrained at high precision)

Understanding your model – what constrains your NP

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

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

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

Understanding your model – what constrains your NP

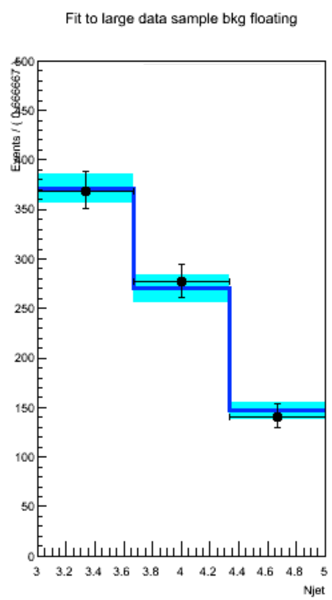
- What did we overlook in the example model?
  - The background rate has no uncertainty!
- Insert modeling of background uncertainty

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

Background rate response function

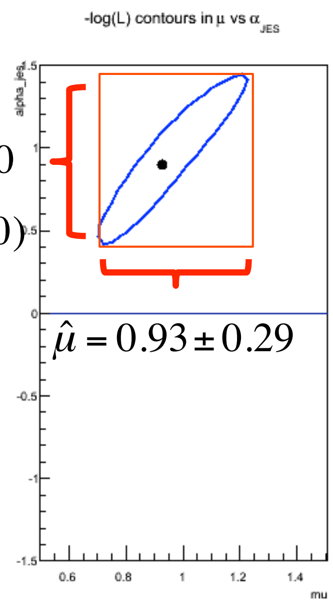
Background rate subsidiary measurement

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



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

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



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

Understanding your model – what constrains your NP

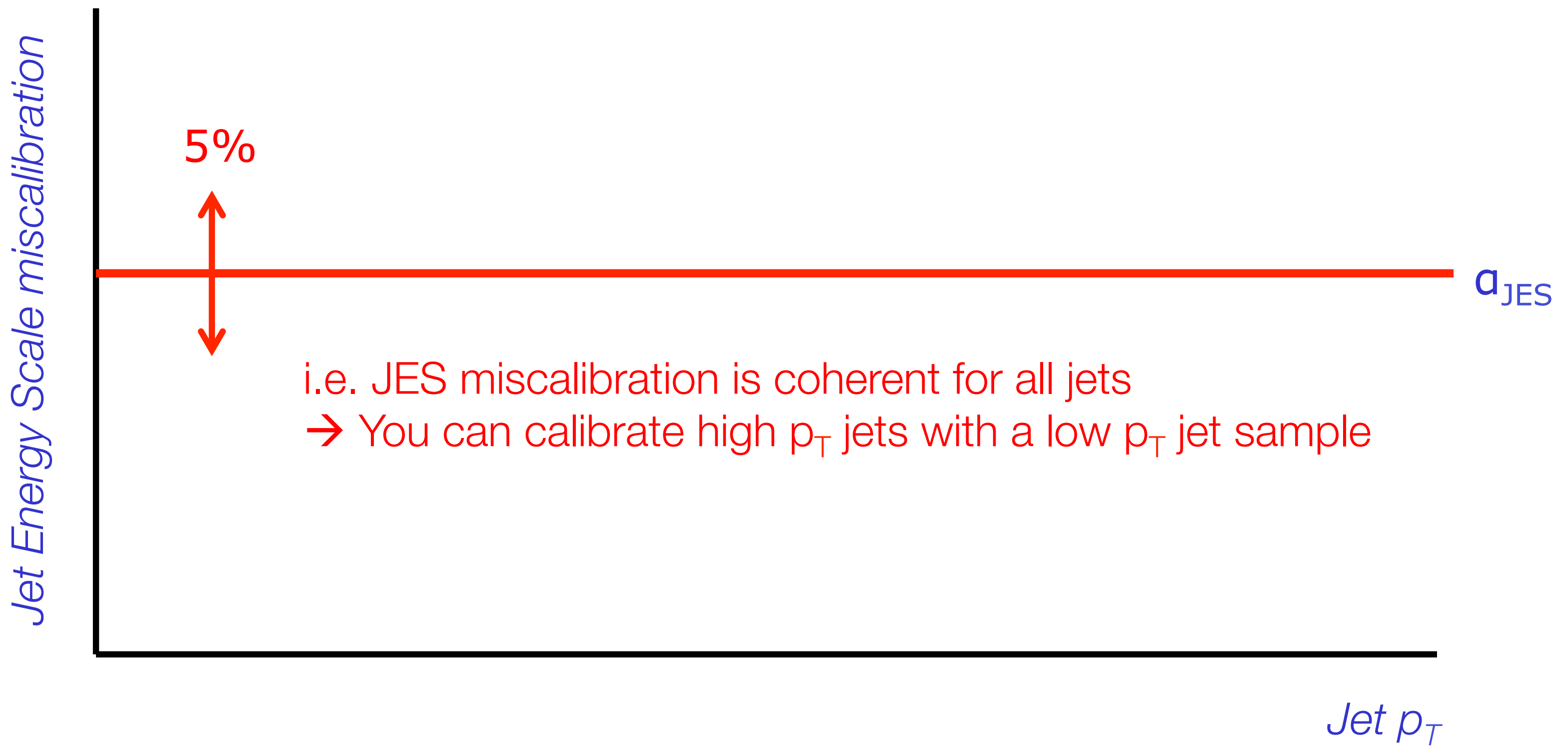
- Lesson learned: if probability model of a physics measurement is insufficiently detailed (i.e. flexible) you can *underestimate* uncertainties
- Normalized subsidiary measurement provide an excellent diagnostic tool
  - Whenever estimates of a NP associated with unit Gaussian subsidiary measurement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

Commonly heard arguments in discussion on over-constraining

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

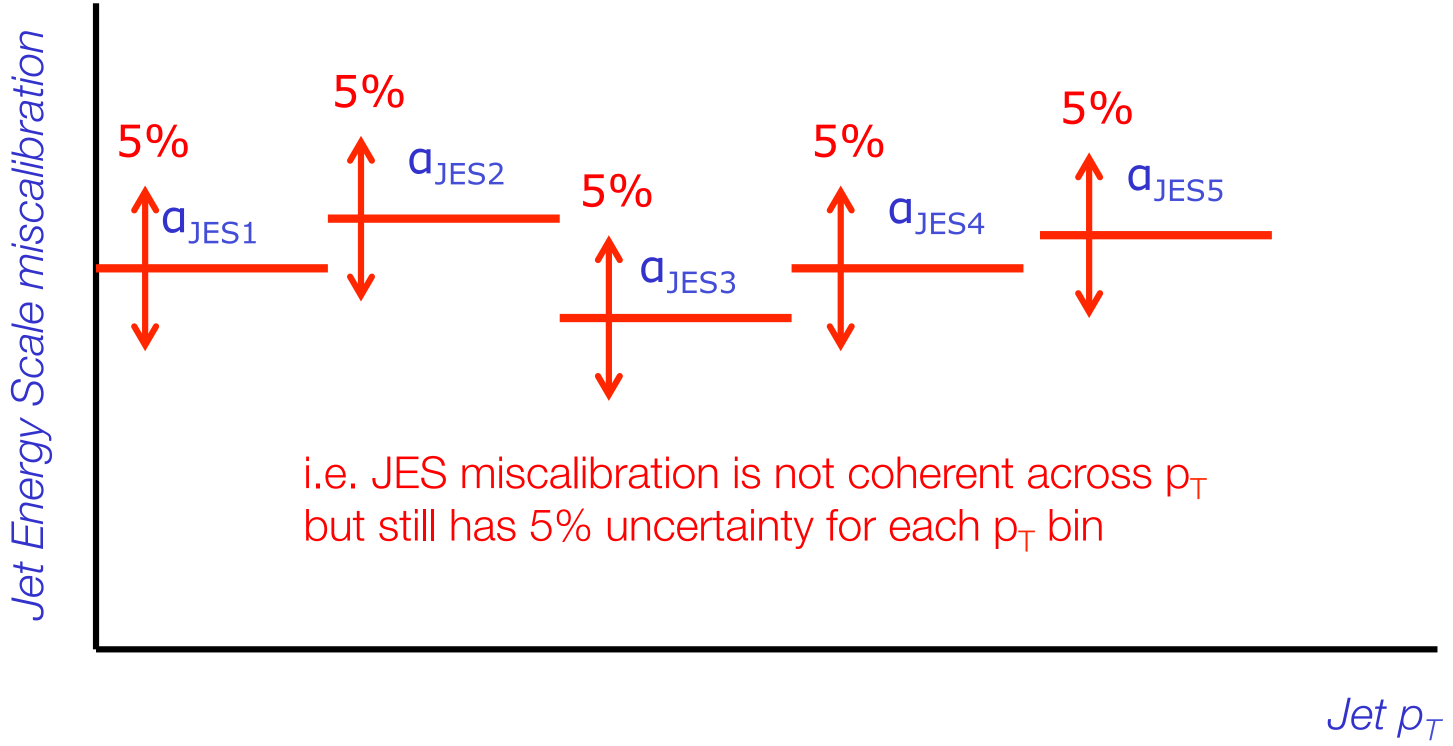
Dealing with over-constraining – introducing more NPs

- 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



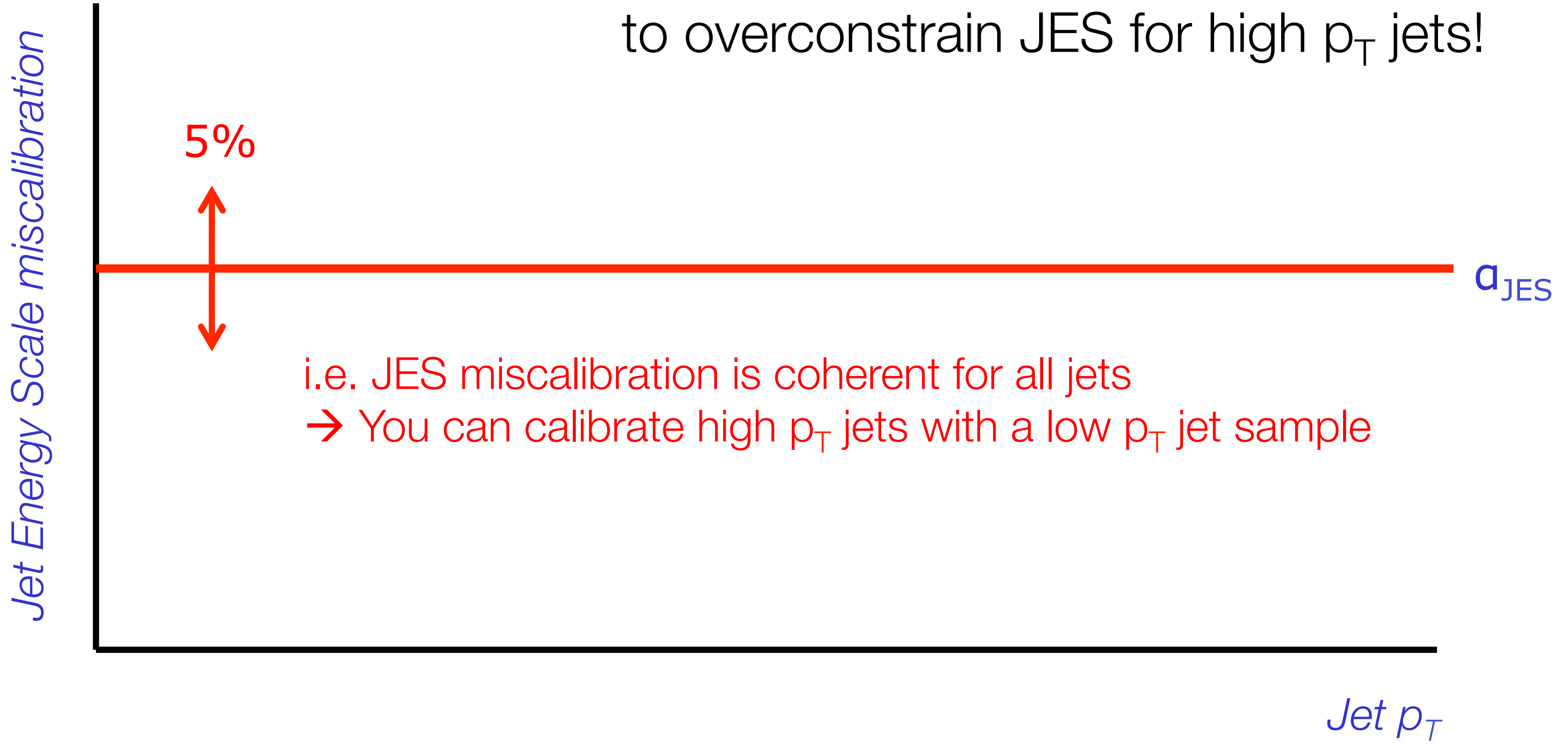
Dealing with over-constraining – introducing more NPs

- 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?

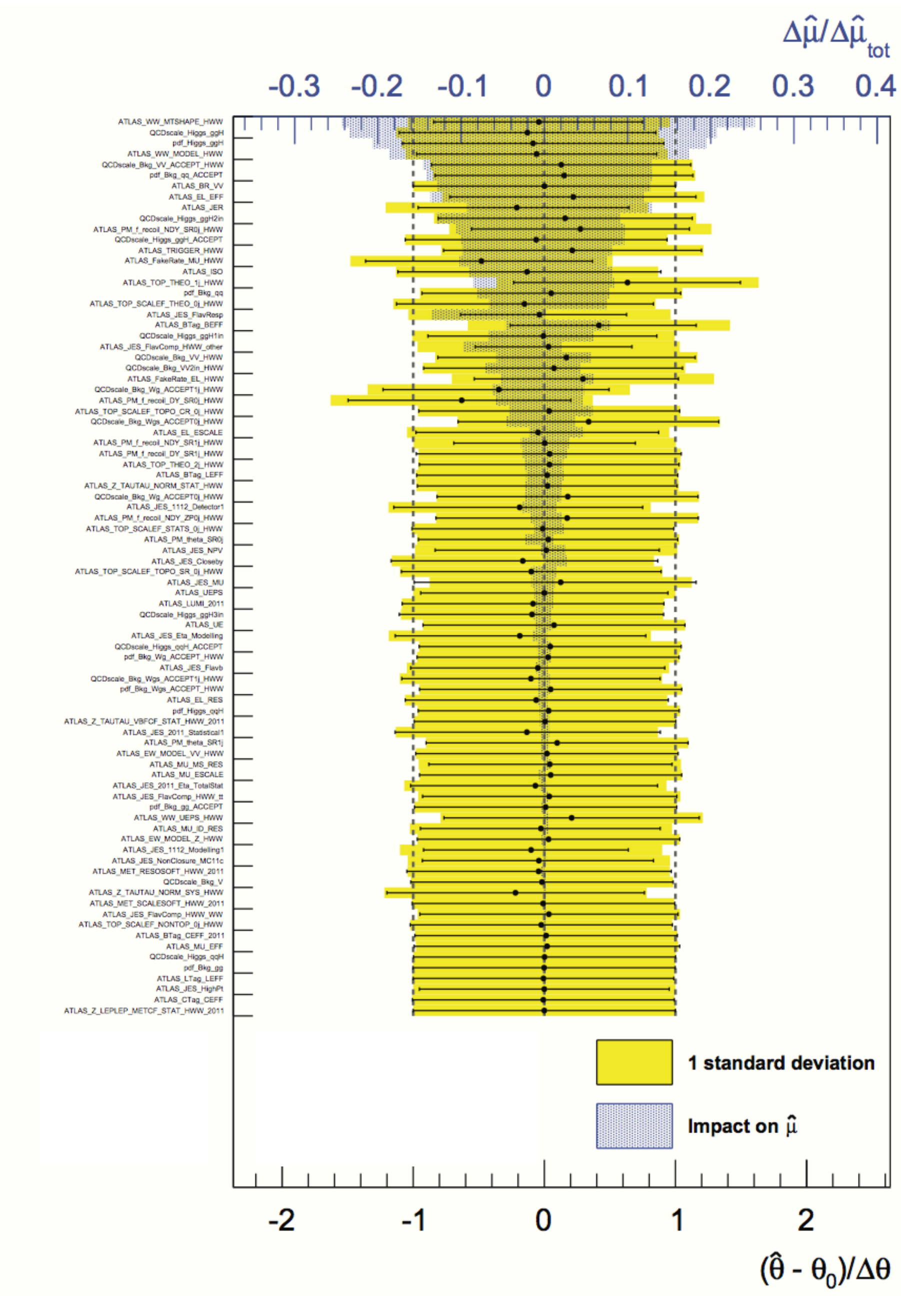


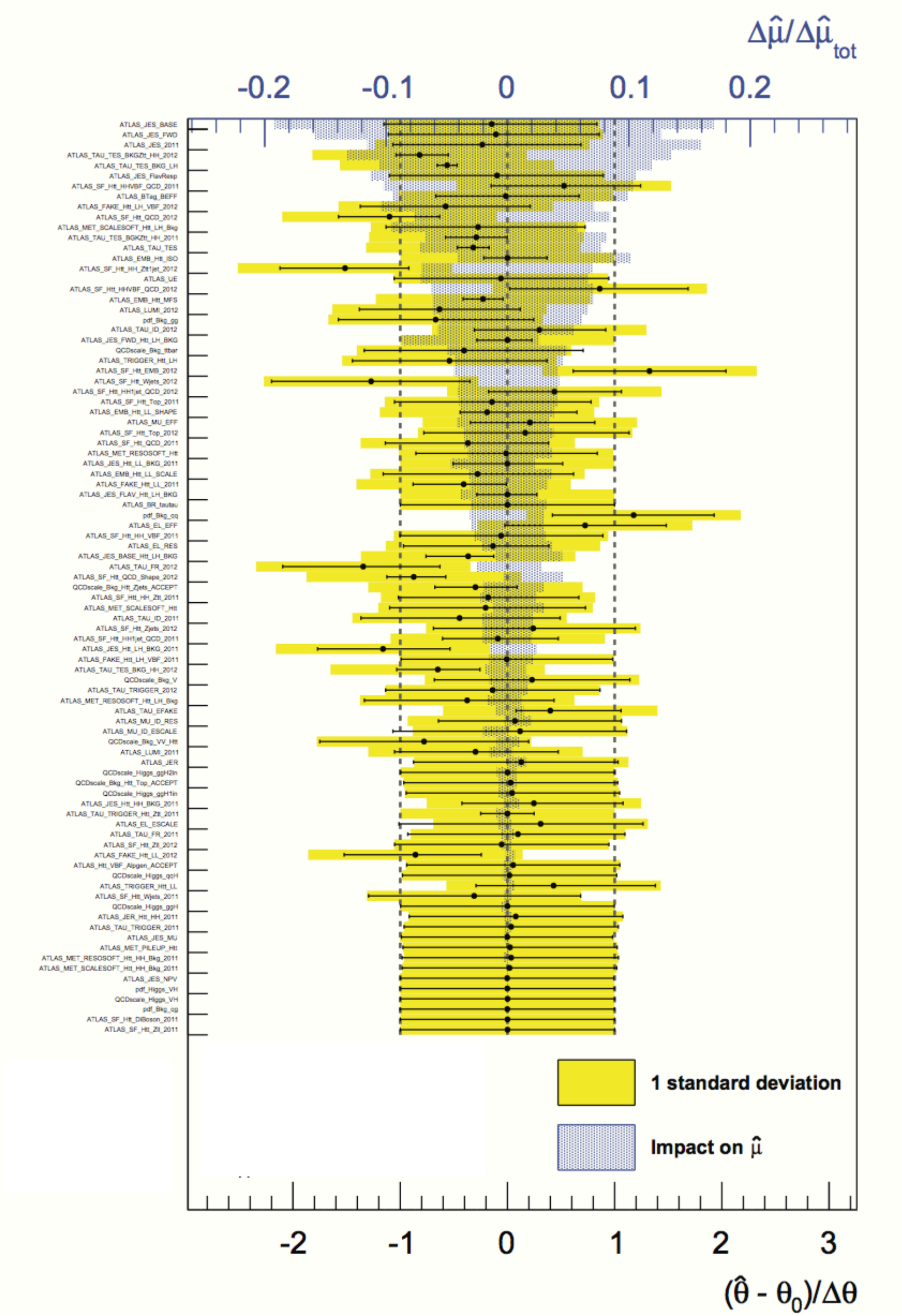
Dealing with over-constraining – introducing more NPs

- 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!



Example of likelihood modeling diagnostics





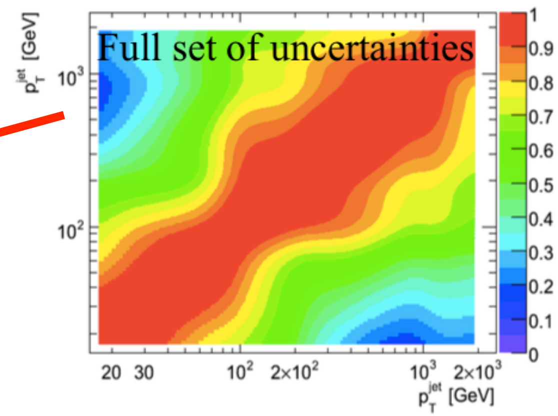
How many NPs you need to capture a systematic uncertainty?

- Detector systematics (calibrations, efficiencies) are complex entities, mapping det. performance measurements with variable resolution of the detector phase space → **Need >>1 parameter**

$$L_{full}(x\,|\,\mu,\alpha_1,\ldots,\alpha_n) = L_{physics}(x\,|\,\mu,\alpha_1,\ldots,\alpha_n)\cdot L_{subsidiary}(0\,|\,\alpha_1,\ldots,\alpha_n)$$

Some α_i parameters can also be *correlated* by subsidiary calibration measurement
(typical for in-situ calibration measurements)

$$L_{subsidiary}(0\,|\,\alpha_1,\ldots,\alpha_n) = \exp(-0.5\cdot\vec{\alpha}^T\cdot V^{-1}\cdot\vec{\alpha})$$

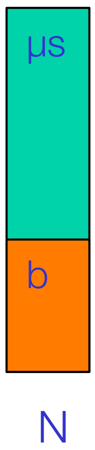


How many NPs you need to capture a systematic uncertainty?

- The need for detailed (detector) systematic modeling also depends on complexity of the physics analysis
→ **Sometimes 1 parameter enough, sometimes all *n* needed**

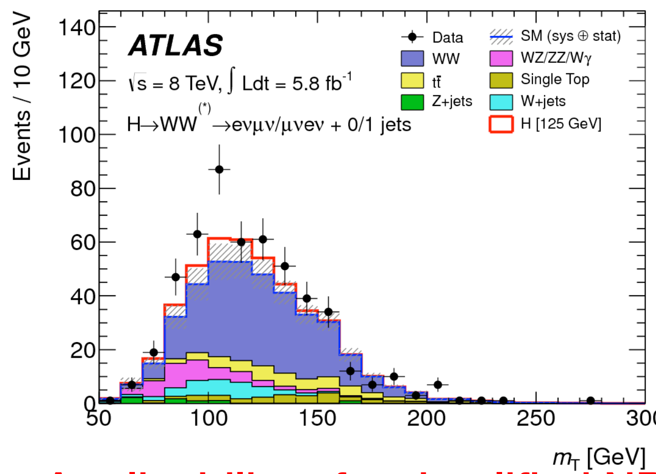
$$L_{full}(x\,|\,\mu,\alpha_1,\ldots,\alpha_n) = L_{physics}(x\,|\,\mu,\alpha_1,\ldots,\alpha_n)\cdot L_{subsidiary}(0\,|\,\alpha_1,\ldots,\alpha_n)$$

Extreme case 1:
Counting measurement in a single bin



1 parameter in the likelihood model describing total effect likely sufficient

Extreme case 2:
Shape fit to multiple distributions

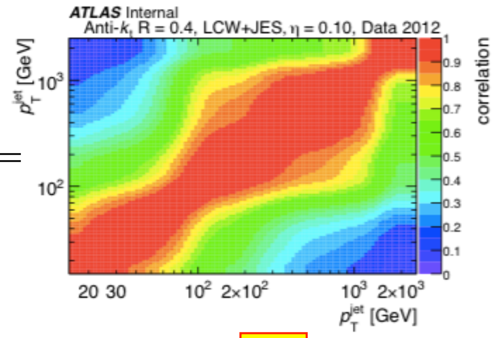


Applicability of a simplified NP model in the likelihood not obvious

1: Reduction – **Eigenvalue decomposition**

- Large NP sets in CP groups usually arise from in-situ calibrations. Common feature of in-situ calibration uncertainties is that nuisance parameters that are strongly correlated.
  - *A priori* all of these NPs are important, making reduction of detail difficult
  - Common solution implemented by performance groups is EigenVector decomposition of these NPs.

$V=$



- Correlation issues solved with an eigenvector decomposition constructs a *rotated set of NPs* that are largely uncorrelated and can be ranked in importance using the eigenvalues
  - EV decomposition doesn't reduce #NPs in its own, but simplifies subsequent merging or pruning of NPs

$$V' = \begin{pmatrix} V_{11} & & & \\ & V_{22} & & \\ & & V_{33} & \\ & & & V_{44} \end{pmatrix}$$

- CP-provided solution for NP reduction: **merging** combine weakest *n* NPs into a single NP
  - Makes ad-hoc assumption that weak modes can be treated as fully correlated
  - Good: With uniform prescription can reduce full set of N nuisance parameters to anywhere between N-1 and 1 parameters
  - But, interpretation of rotated NPs difficult, correlation with other analyses (or even CMS) becomes more complicated (but solutions are foreseen)

$$V'' = \begin{pmatrix} V_{11} & \\ & V_{mm} \end{pmatrix}$$

Modeling theory uncertainties

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

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

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

- Difficult example – Factorization scale uncertainty

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

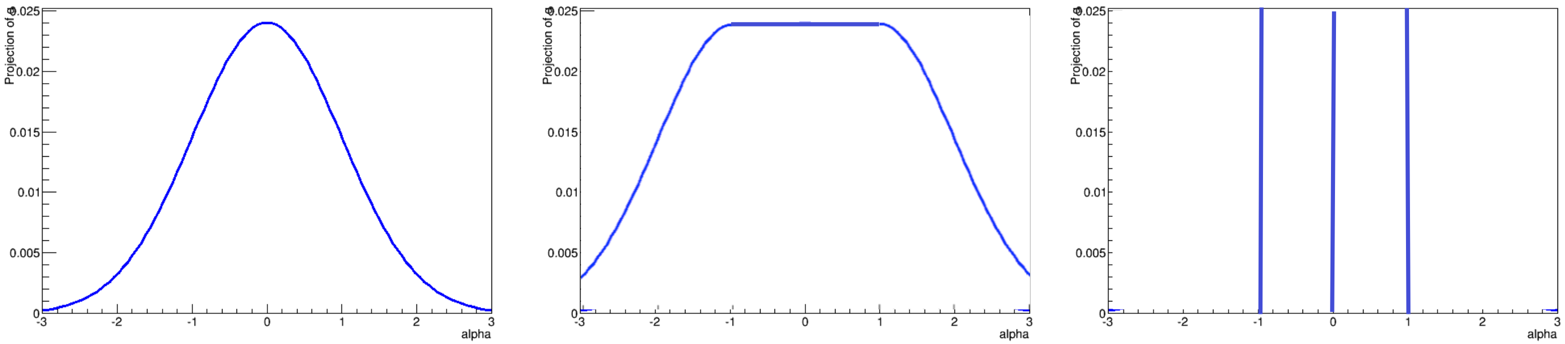
“Vary Factorization Scale by x0.5 and x” → F(α) is probably not Gaussian
So what distribution was meant?

Modeling theory uncertainties

- Difficult example – Factorization scale uncertainty

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

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



- 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

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

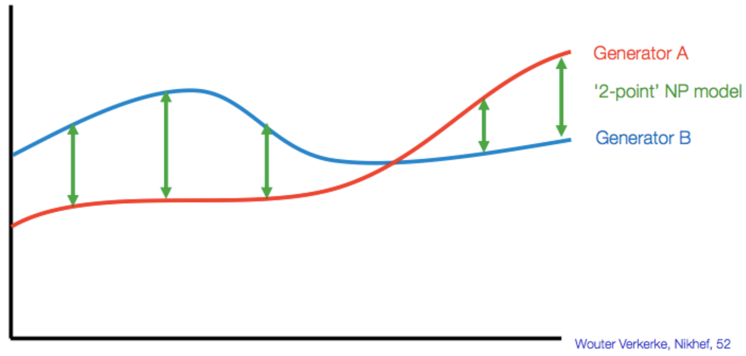
  – One parameter: the quark mass→ Clearly described and connected to the underlying theory model

- 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?**

Type of modeling uncertainties for theory systematics

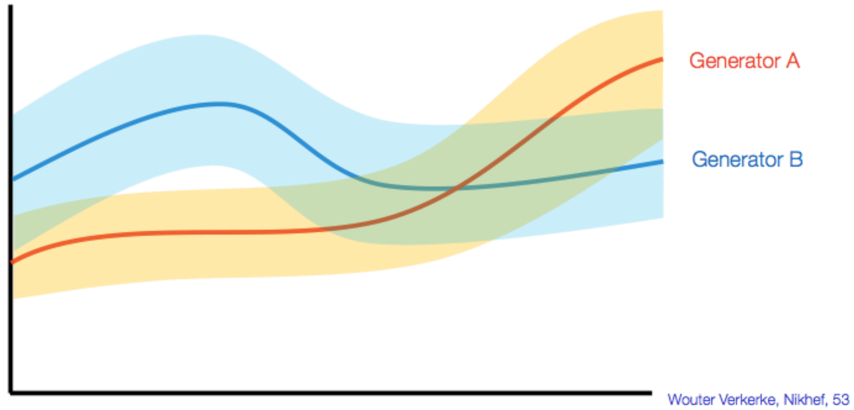
1. ‘Two-point uncertainties’

- Two concrete predictions
- Disagree within stat. uncertainty
- No well-defined intrinsic uncertainty
- Example: Herwig vs Pythia



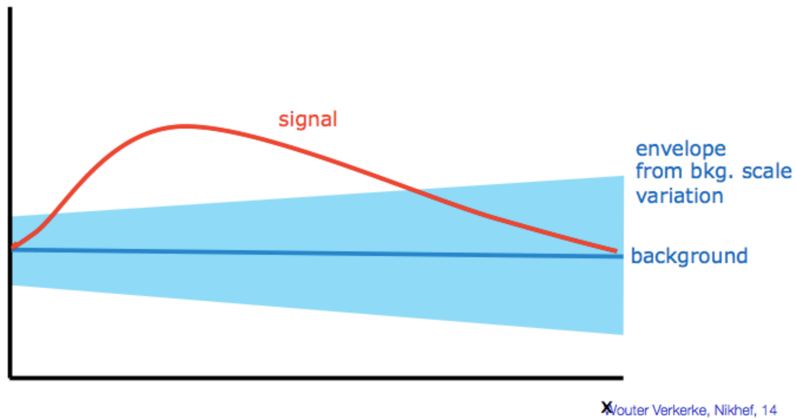
2. ‘Two-band uncertainties’

- Two concrete predictions
- Well-defined intrinsic uncertainty
- Disagree within stated uncertainty
- Example: PDF sets



3. ‘Envelope uncertainties’

- Uncertainty band from heuristic prescription
- No knowledge of correlation structure inside band
- Example: Scale Unc / MHOU

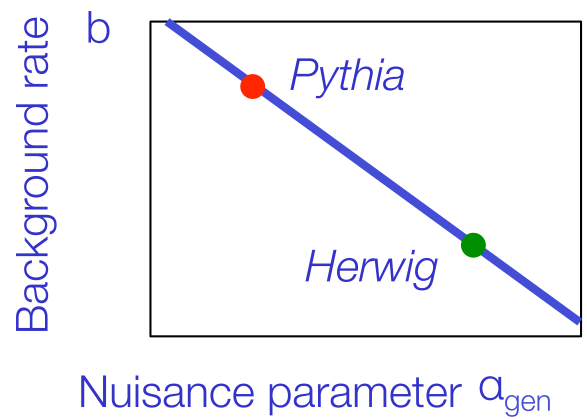


Two-point uncertainties

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

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

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

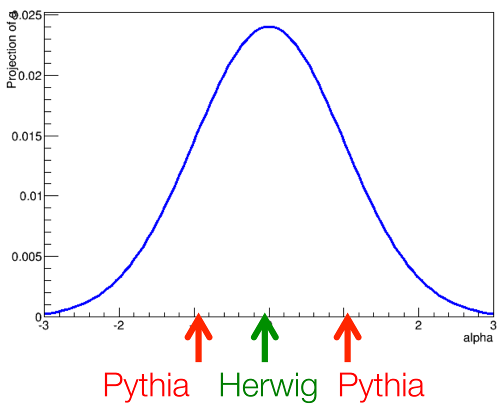


- A thorny question remains:
  What is the subsidiary measurement for α?
  This should reflect you current knowledge on α.

Two-point uncertainties

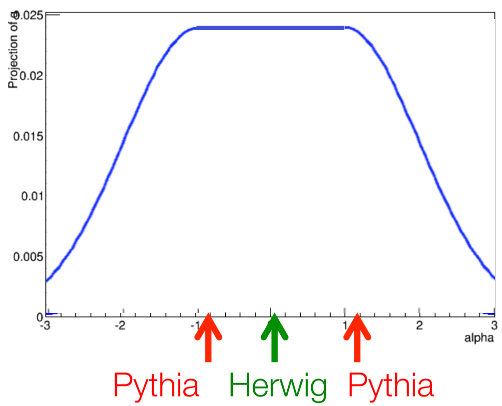
- Subsidiary measurement of a theoretical 2-point uncertainty effectively quantifies the ‘knowledge’ on these models
  - *Extra difficult to make meaningful statement about this*, since meaning of parameter is not well embedded in underlying theory model
  - But again, all procedures need to assume some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

Gaussian



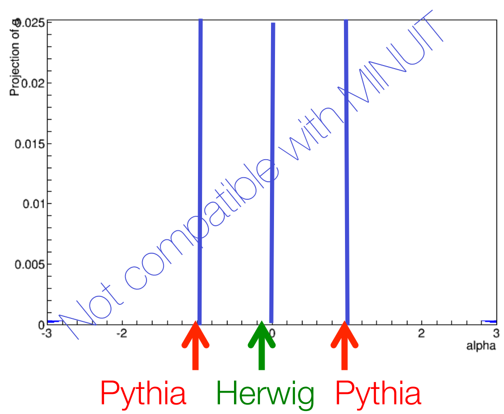
Prefers Herwig at 1σ

Box with
Gaussian wings



All predictions ‘between’ Herwig and Pythia equally probable

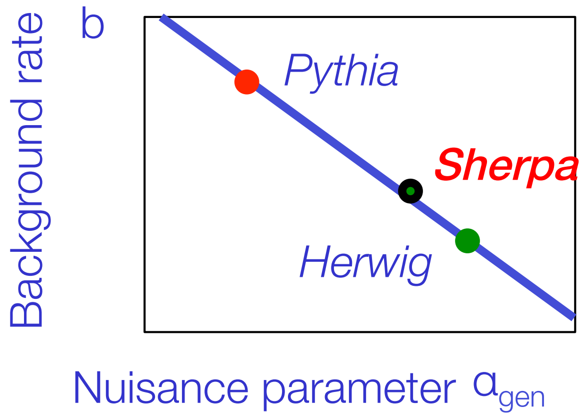
Delta fuctions



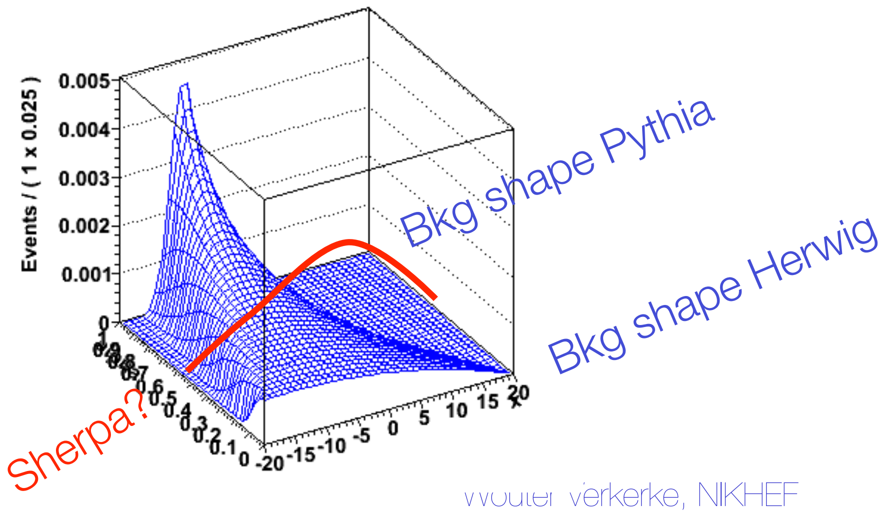
Only ‘pure’ Herwig and Pythia exist

Two-point uncertainties

- 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

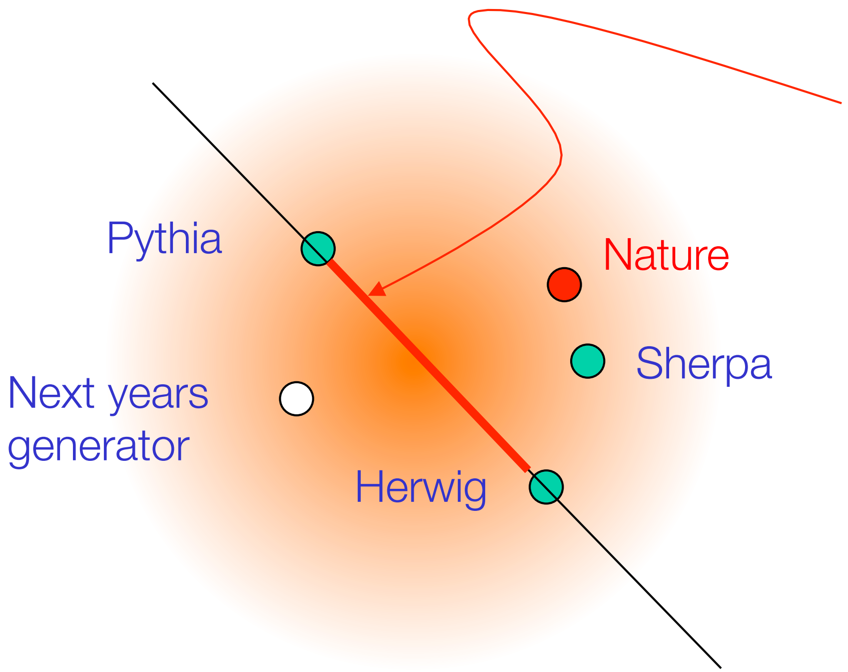


- This is not generally true for distributions. A shape interpolation between ‘pythia’ and ‘herwig’ does not necessarily describe shape of ‘sherpa’ (or of Nature!)
  - Fundamental modeling problem!
  - You may need more parameters…



Two-point uncertainties

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



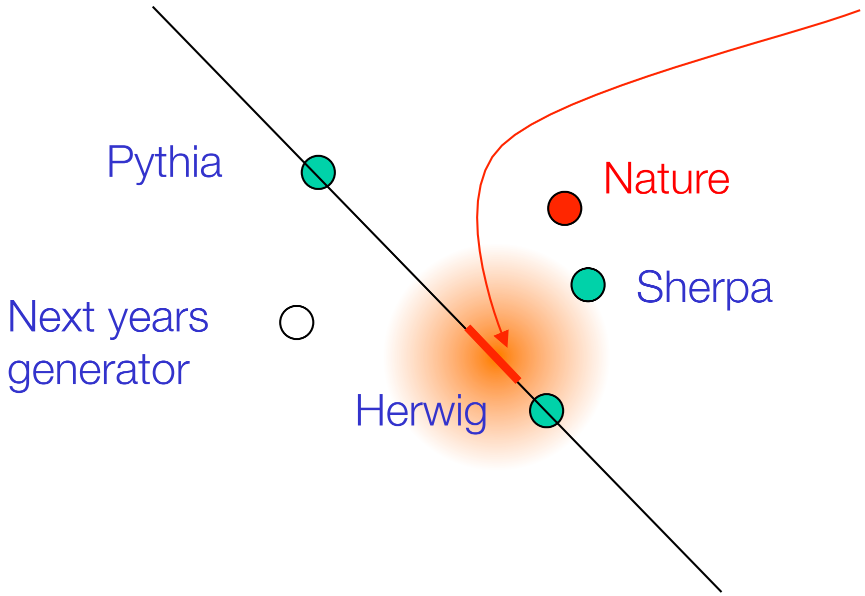
Modeled uncertainty (1 dimension) assuming ‘nature is on line’

Effectively captured uncertainty

under the <u>assumption</u> that effect of ‘position in model space’ in any dimension is similar on response function

Two-point uncertainties

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



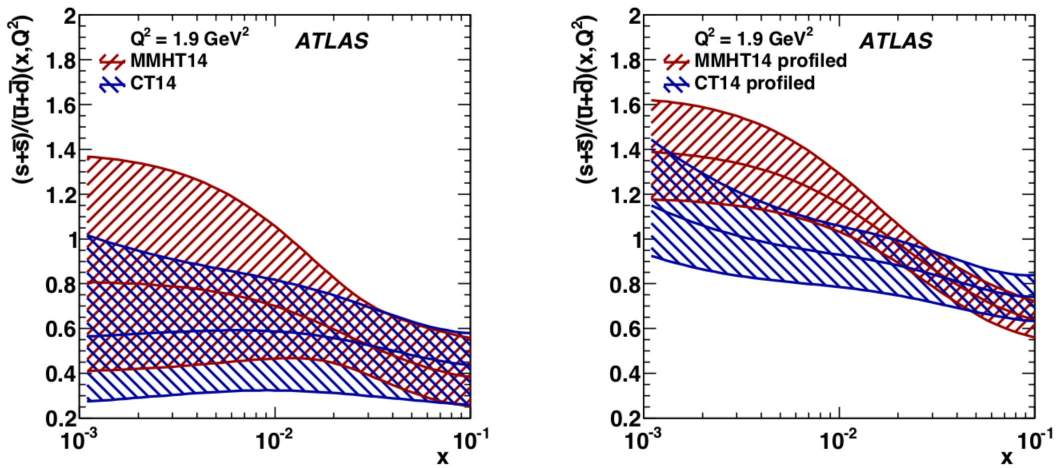
Modeled uncertainty (1 dimension) assuming 'nature' is on line

Effectively captured uncertainty

under the assumption that effect of 'position in model space' in any dimension is similar on response function

Two-band uncertainties – i.e. PDF uncertainties

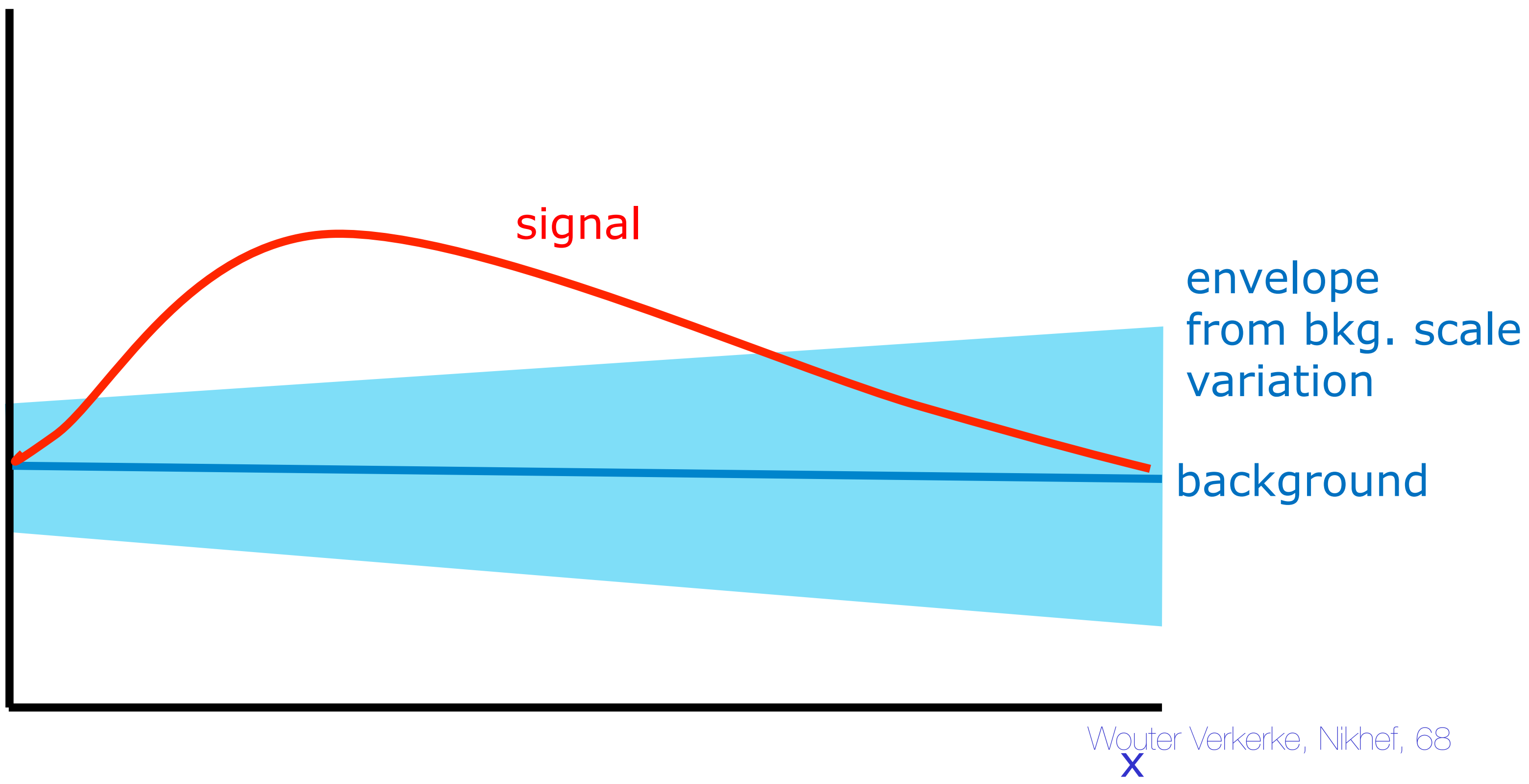
- Scenario A: physics measurement cannot constrain PDF
  - *If PDF sets disagree within their error bands* – usually add the difference between sets as a 2-point systematic in addition to the intrinsic uncertainty of the chosen 'nominal' PDF set. (In the spirit of 'being conservative')
  - Of course issue of arbitrariness – which PDF is taken as the 'reasonable alternative'. We have no real good answers/solutions for this (but generally tend to be conservative)
- Scenario B: physics measurement *can* constrain PDF
  - We choose to allow physics measurement to constrain PDF uncertainty further.
  - Assumes that parametrization chosen by PDF authors is flexible enough
  - In particular – if 2 constrained PDF sets agree post-fit – can argue that this serves as validation of the constrained parametrization → No need for additional systematic



Verkerke, NIKHEF

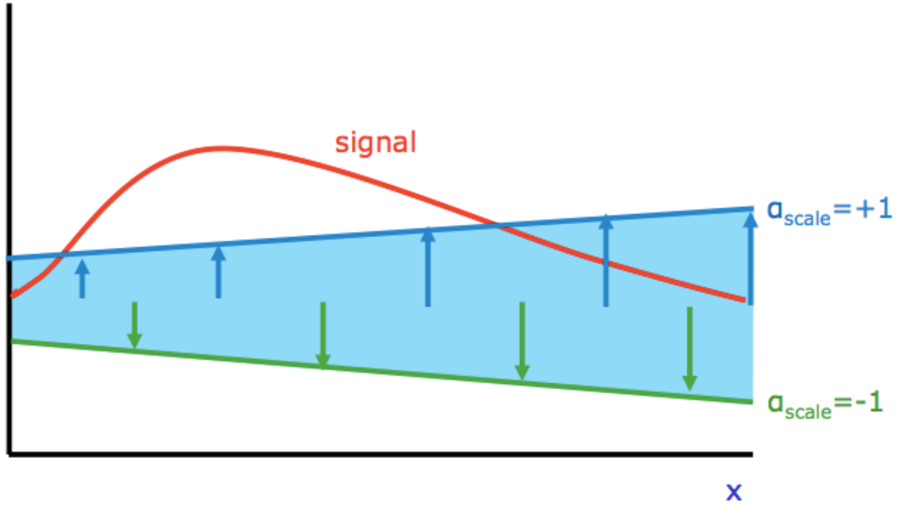
Envelope uncertainties

- Input information (most cases) an ‘envelope’ defined by scale variations
  - ‘7-point procedure’ set the size of an envelope
  - **In most cases - no known/usable information on correlation structure → constraining uncertainty using NP model assumptions effectively never allowed**



Envelope uncertainties

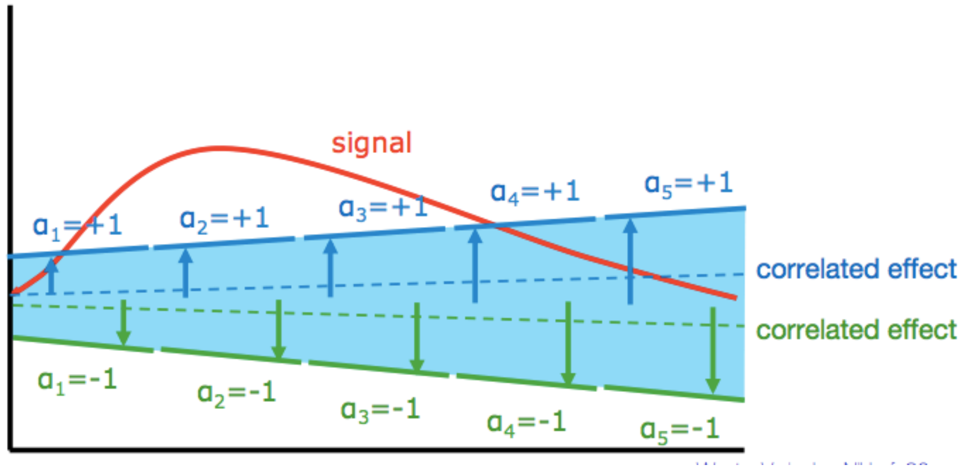
- For measurement with 1 POI the NP model that propagates the ‘full’ uncertainty is always the model that is maximally correlated with the POI → Parametrization can always be analytically derived
  - Most common example: POI is a global yield parameter



- Main headache here is *what to do if physics measurement has power to constrain NP propagates MHOU uncertainty*?
  - A constrained NP could (potentially) significantly constrain the propagated uncertainty.

Envelope uncertainties

- Commonly proposed solution: split NP into 2 or more NPs up to point where constraint goes away. (each NP representing a region in the phase space)

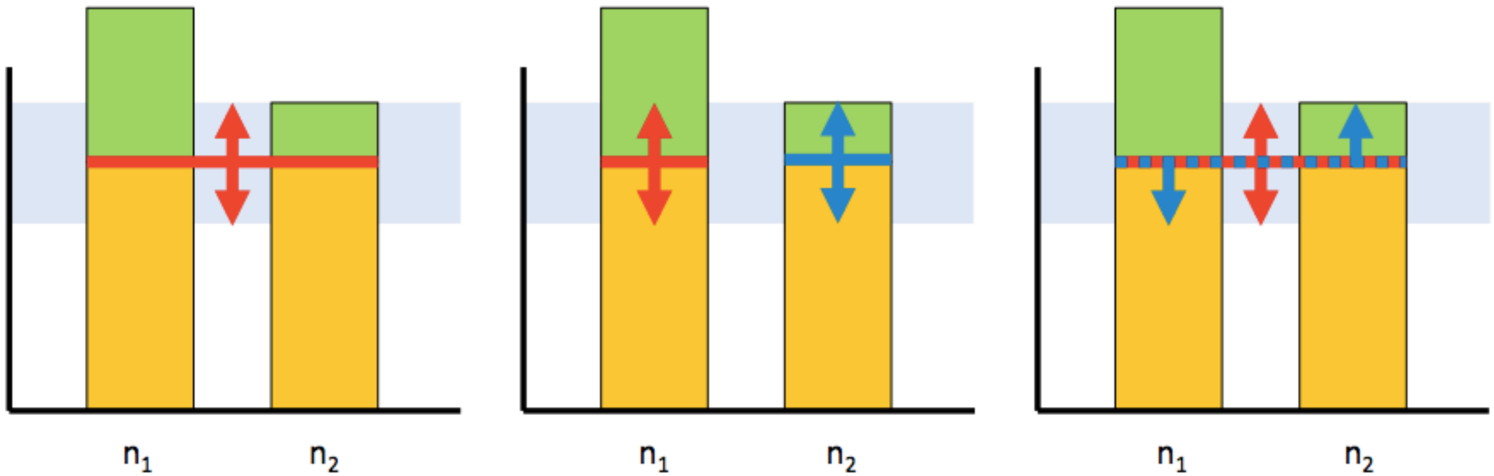


- But not necessarily a good idea: de-correlating reduces propagated uncertainty to POI up to a factor √N for N regions
  - Assumption of decorrelation reduced uncertainty – but is not necessarily a warranted assumption
  - Cure might be worse than original problem (constraint of single NP)….
  - Also: in many cases it is observed to be impossible to avoid constraints, even with aggressive regional splitting

Envelope uncertainties

- There *are* good technical solutions for NP models that neither overconstrain neither inappropriately assume decorrelation
- For example for a 2-region measurement

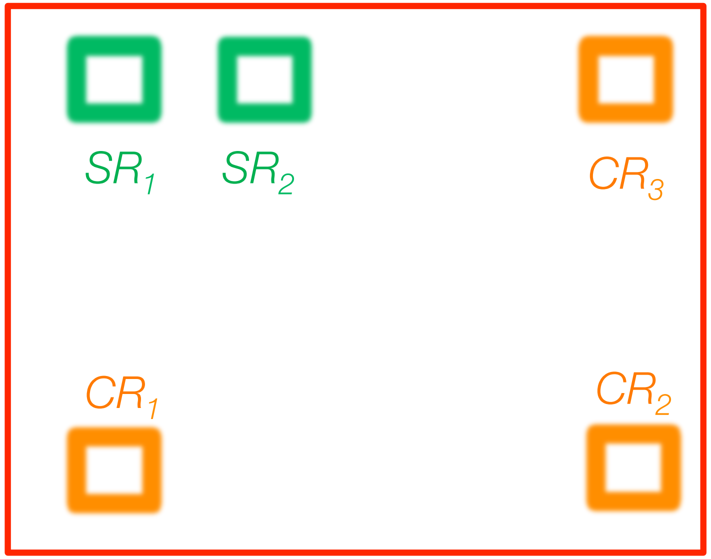
Single correlated NP (prone to overconstraining)	Two uncorrelated NPs (propagating √2-reduced unc)	Rotated NP model (corr/anti-corr NP) no constr, no red.
$\sigma(n_1+n_2) = 10\%$	$\sigma(n_1+n_2) = 7\%$	$\sigma(n_1+n_2) = 10\%$



Can be generalized to N regions

Joint fits tend to exacerbate NP modeling issues

- Practical experience is that often constraining power of physics measurements does not come from the individual likelihoods that measure a single region, but more often from the joint inference of multiple regions ('lever arm')



- I.e. issues of constraining of NPs (and corresponding unwanted reduction of propagated theory systematics) often arises first when making combinations → Watch carefully, good diagnostics vital.

Wrapping up – modeling systematics in joint fits

- Experimental systematics often have NP models that are reasonably robust
  - Often EV-decomposed with multiple parameters.
  - Number of used parameters in individual measurements tuned to avoid constraints
  - Combination of measurements usually do not lead to constraints on experimental systematics (*but it can happen and should be checked*)
  - If measurement use NP models with different level of detail (e.g. 1 vs 5 NPs – correlation should be done per EV (i.e. correlate EV_0 and not total uncertainty)
  - Intractable problems possible (e.g. using both 70% and 80% efficiency points of flavor tagging in same combination). Typically avoided with a priori coordination in LHC 'internal' combinations. But be careful with 'home grown' combinations

Wrapping up – modeling systematics in joint fits

- Theory systematics often more difficult to parametrize safely
  - Explicit validation/tuning/reparametrization in measurements common to avoid both constraints and unwarranted decorrelation assumptions (which would both result in underestimated uncertainties)
  - Tuned NP models for individual measurements often not robust against new problems that arise combinations → proceed with care:
  - Overconstraining might happen,
  - Newly introduced correlation assumptions could strongly reduce uncertainties (e.g. ratio of measurements with a single correlated NP)


Questions?