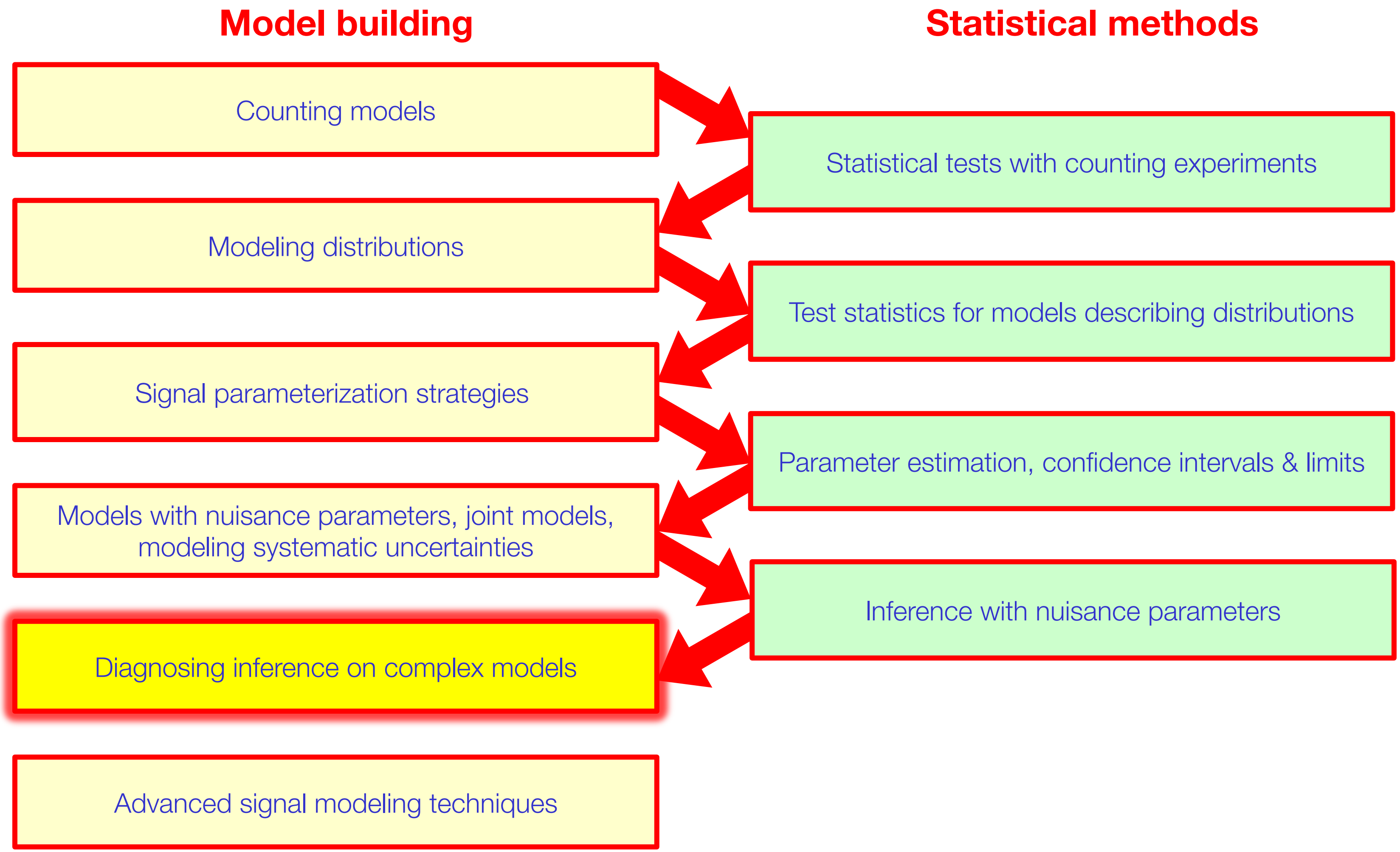
Model building 5

Diagnostics (understanding MINUIT, fit stability and convergence) and Validation (understanding your fit, overconstraining parameters, 2-point systematics etc)

Wouter Verkerke, NIKHEF

Roadmap of this course

- Start with basics, gradually build up to complexity



Model building

Statistical methods

Counting models

Statistical tests with counting experiments

Modeling distributions

Test statistics for models describing distributions

Signal parameterization strategies

Parameter estimation, confidence intervals & limits

Models with nuisance parameters, joint models, modeling systematic uncertainties

Inference with nuisance parameters

Diagnosing inference on complex models

Advanced signal modeling techniques

Minimizers and convergence of profile likelihood fits

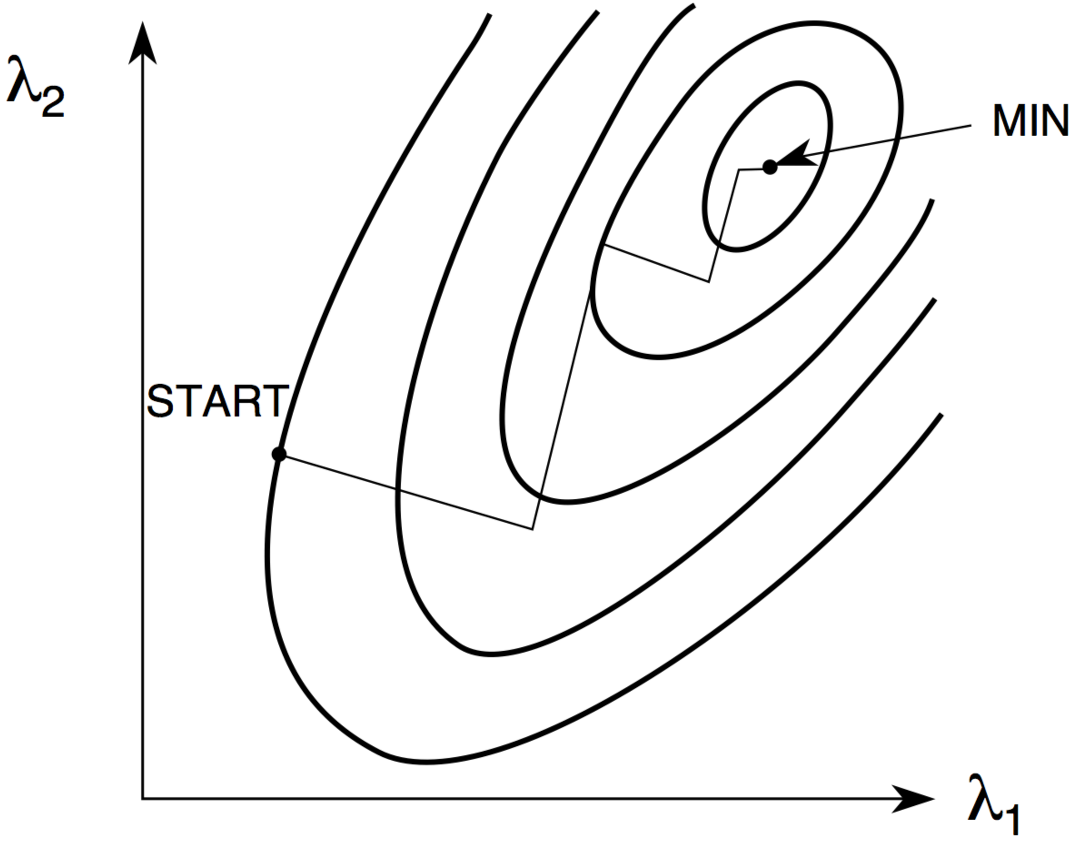
- Likelihoods with systematics modeling (‘profile likelihood fits’) tend to be more complex than ‘normal’ fits
- Sometimes these likelihood can have pathological features that frustrate the minimization process
- To help you understand I will briefly cover
  - How MINUIT works and defines ‘convergence’
  - 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 efffectively 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

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



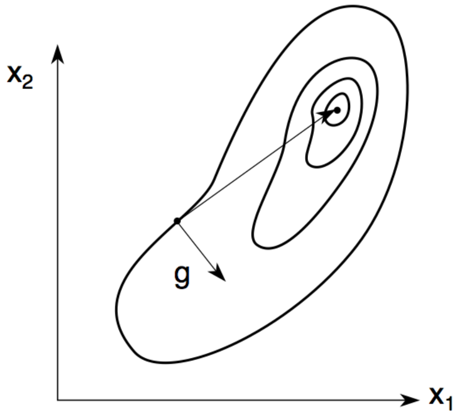
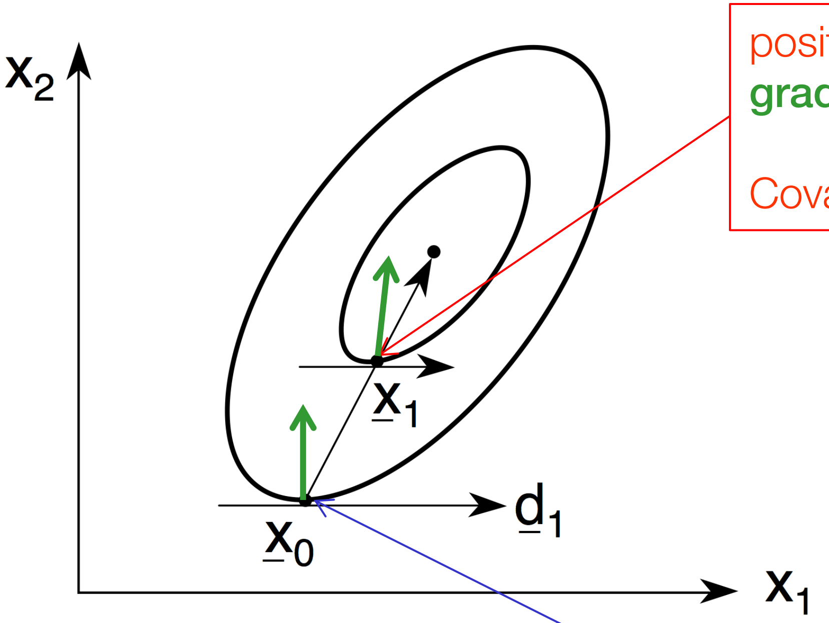


Fig. 9

Function minimization using the variable metric method

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



position: $\mathbf{x_1=x_0-V_0g_0}$

gradient: $\mathbf{g_1}$

Covariance: $\mathbf{V_1 = V_0+f(V_0,x_0,x_1,g_0,g_1)}$

Davidon-Fletcher-Power rank 2 formula

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

$$\delta = \mathbf{x}_1 - \mathbf{x} \quad \gamma = \mathbf{g}_1 - \mathbf{g}_0,$$

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

position: x_0
gradient: g_0
Covariance: $V_0 = G^{-1} = I$

NB: If function is perfectly parabolic and initial V_0 is correct, convergence in one step!

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

$$\boxed{2\cdot\mathrm{EDM}=\rho=g^{T}Vg}$$

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

Covariant metric tensor

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

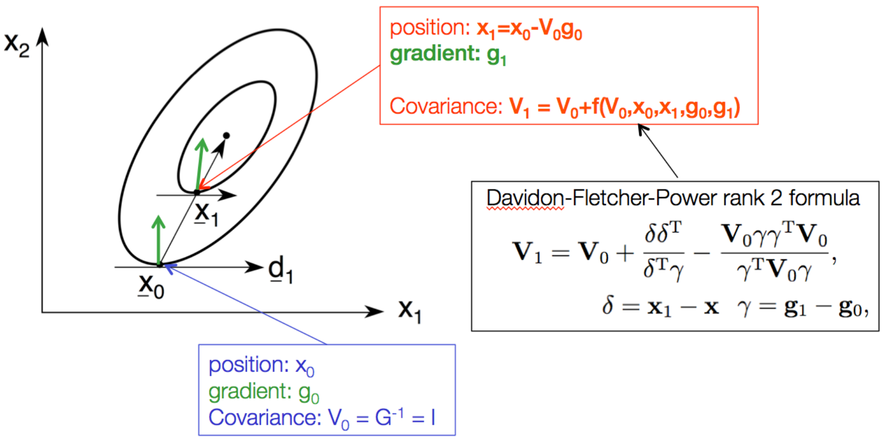
- Note that both minimization and convergence criteria depend on knowledge of covariance matrix

- There are 2 ways to calculate V
  1. From the Davidon-Fletcher-Power formula $$\mathbf{V}_1=\mathbf{V}_0+\frac{\delta\delta^{\mathrm{T}}}{\delta^{\mathrm{T}}\gamma}-\frac{\mathbf{V}_0\gamma\gamma^{\mathrm{T}}\mathbf{V}_0}{\gamma^{\mathrm{T}}\mathbf{V}_0\gamma},$$
  2. From the inversion of the Hessian matrix $$\mathbf{V}=\mathbf{G}^{-1}$$

Calculation of Hessian is expensive
(½N^2 likelihood evaluations)

Minimization convergence

- After every VariableMetric step calculate EDM = ½$g^T V g$



position: x_1=x_0-V_0g_0
gradient: g_1

Covariance: V_1 = V_0+f(V_0,x_0,x_1,g_0,g_1)

Davidon-Fletcher-Power rank 2 formula

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

$$\delta = \mathbf{x}_1 - \mathbf{x} \quad \gamma = \mathbf{g}_1 - \mathbf{g}_0,$$

position: x_0
gradient: g_0
Covariance: V_0 = G^{-1} = I

```
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 EDM<0.001

Minimization convergence

```
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 V from DFP updater formula

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

- From here on, procedure depends on ‘strategy code’
  - Code 0: terminate line search
  - Code 2: Recalculate **V** from $\mathbf{G}^{-1}$ (HESSE)
    if 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.

Validation of 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'

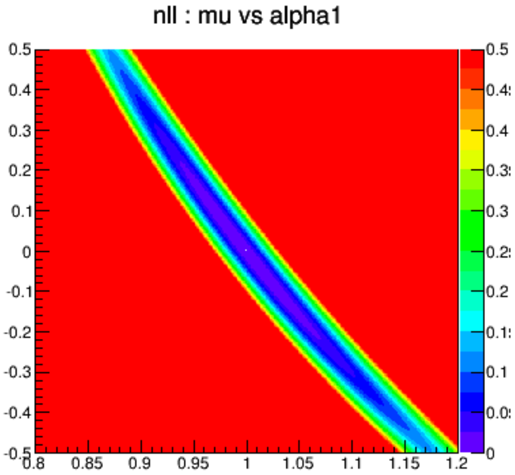
Likelihood models that cause minimization problems

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

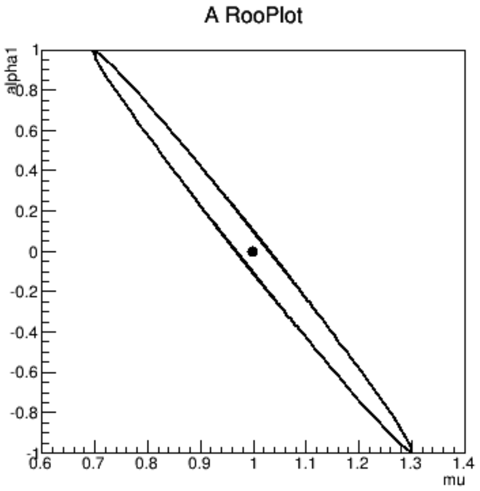
$$L_1(\mu,\alpha) = Poisson(N \mid \mu S(1+\tau\alpha))Gaussian(0 \mid \alpha,1)$$

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

Scan of –log L(μ,α)



Error ellipse from V(μ,α) HESSE



ρ=0.9945

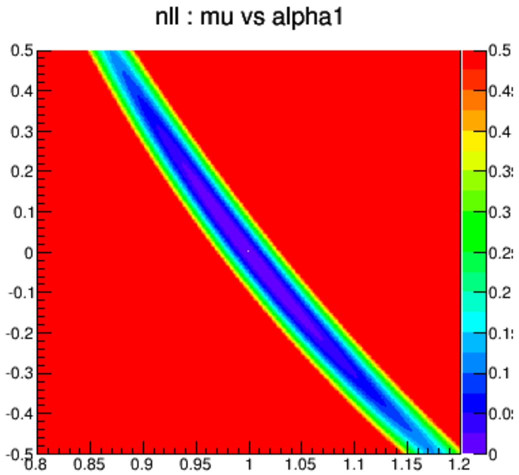
  - Strong correlations, but numerically feasible

Increasing the observed event count

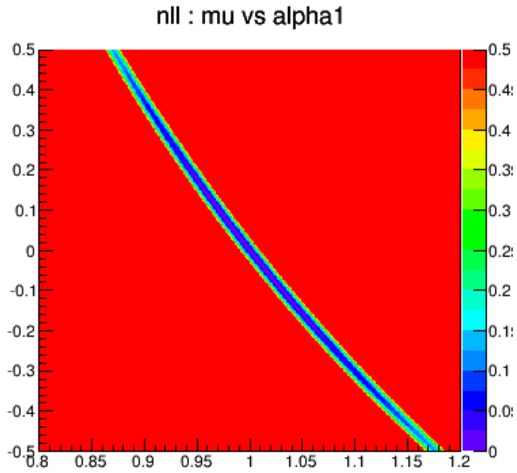
Vertical scale maximized at 0.5 units

Scan of –log L(μ,α)

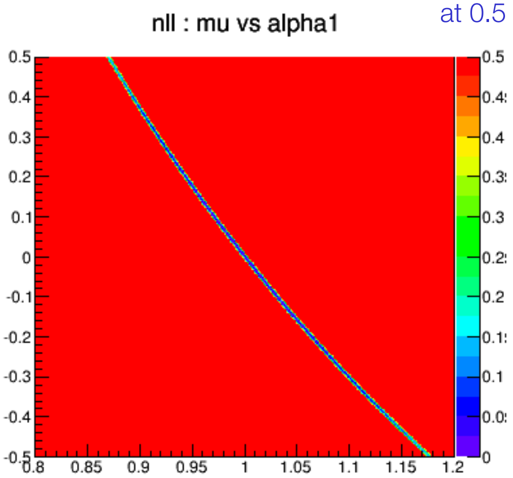
N=1000



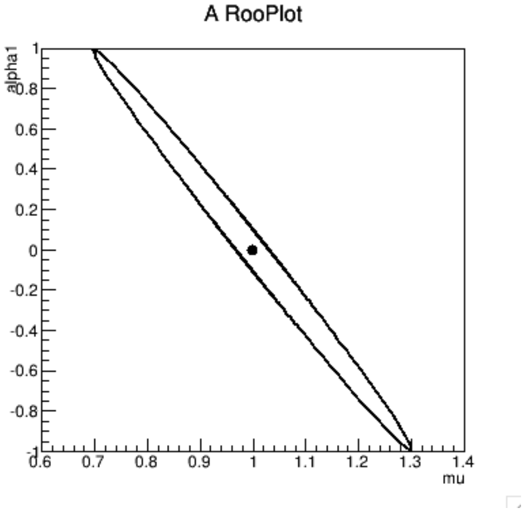
N=10.000



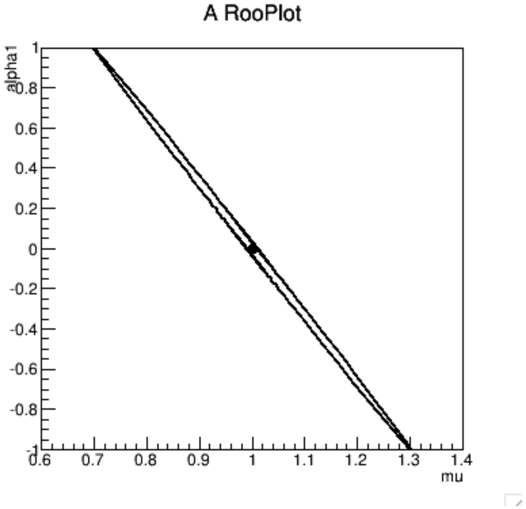
N=100.000



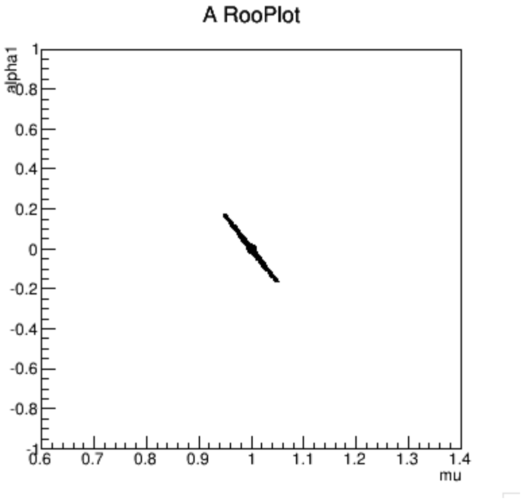
Error ellipse from V(μ,α) HESSE



ρ=-0.9945



ρ=-0.9995

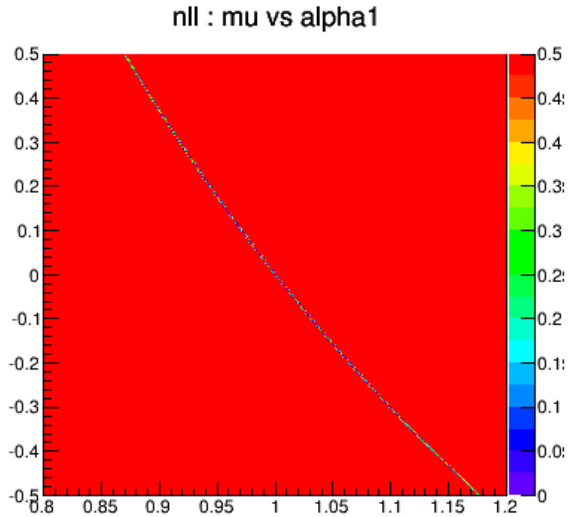


ρ=-0.98

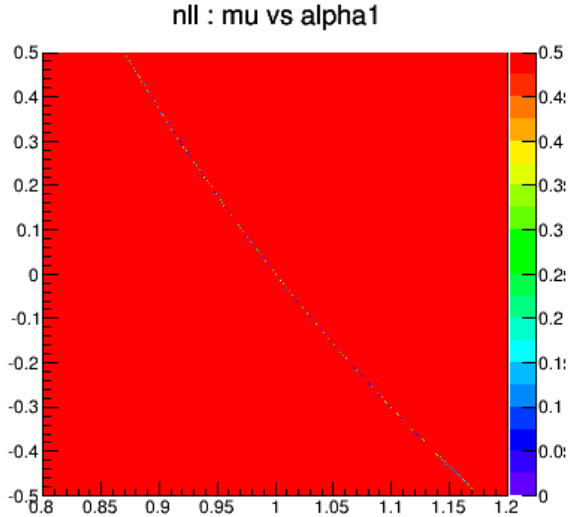
Increasing the observed event count

Scan of –log L(μ,α)

N=1.000.000

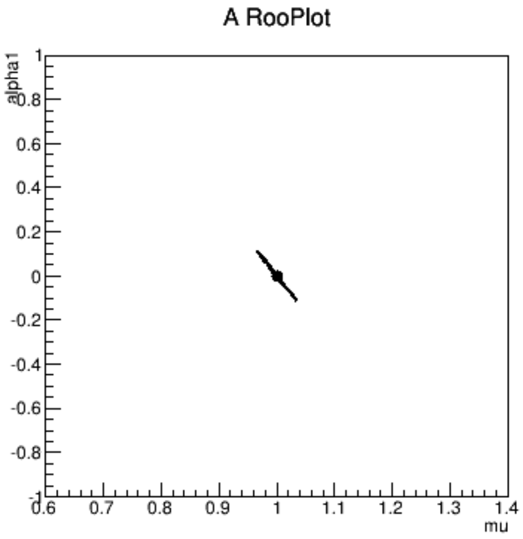


N=10.000.000

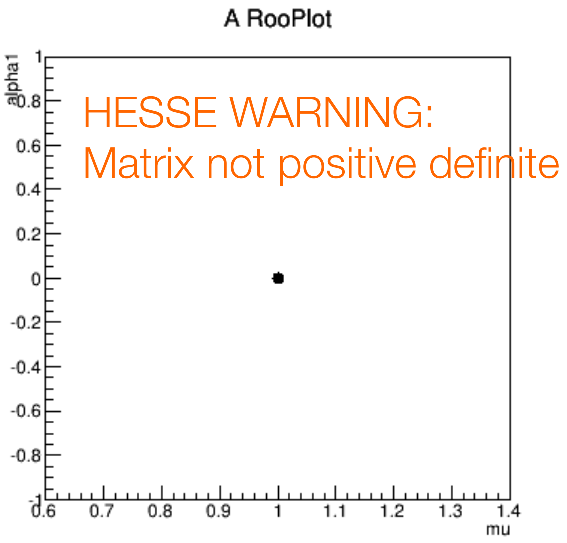


Vertical scale maximized at 0.5 units

Error ellipse from V(μ,α) HESSE



ρ=-0.9996



HESSE WARNING:
Matrix not positive definite

ρ=-0.998

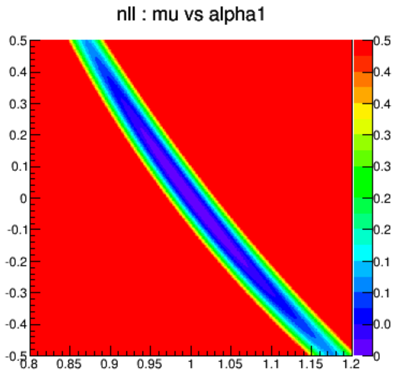
Likelihood models that cause minimization problems

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

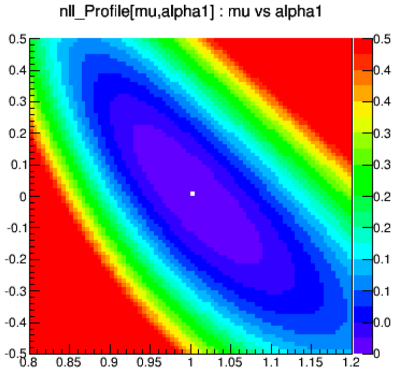
$$L_2(\mu,\alpha_1,\alpha_2) = Poisson(N \mid \mu S(1+\tau_1\alpha_1+\tau_2\alpha_2))Gauss(0 \mid \alpha_1,1)Gauss(0 \mid \alpha_2,1)$$

  - 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?

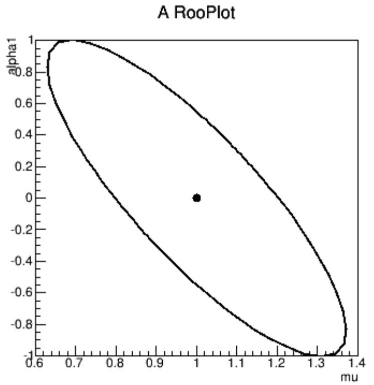
$$-\log L(\mu,\alpha_1,\underline{\hat{\alpha}_2})$$



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



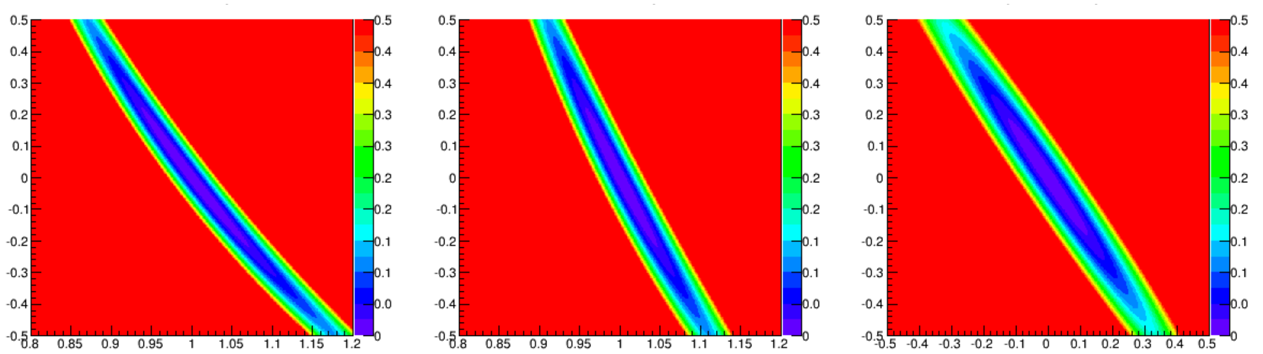
Error ellipse V(μ,α) HESSE



-logL(μ,α$_1$,α$_2$) – 1000 events

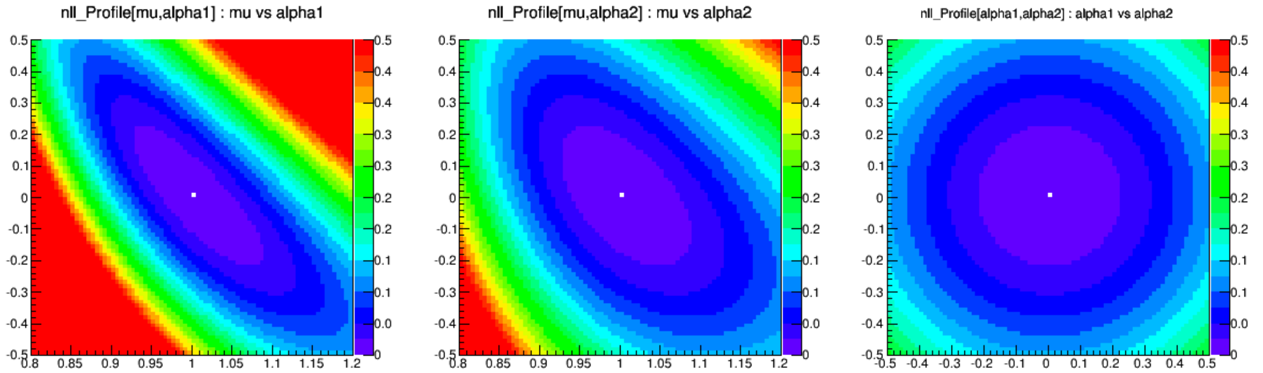
Slice in -logL

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

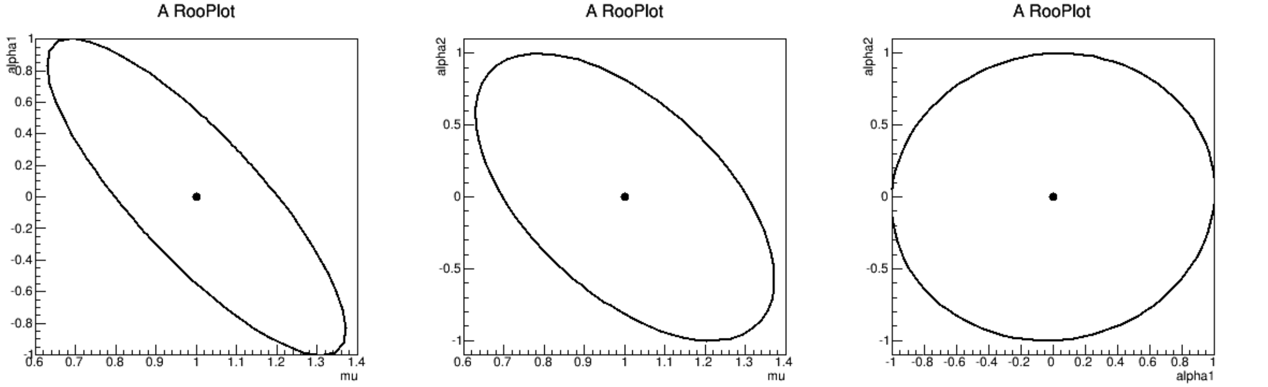


Profile likelihood

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



Error ellipse from HESSE



-logL(μ,α_1,α_2) – 10.000 events

Slice in -logL

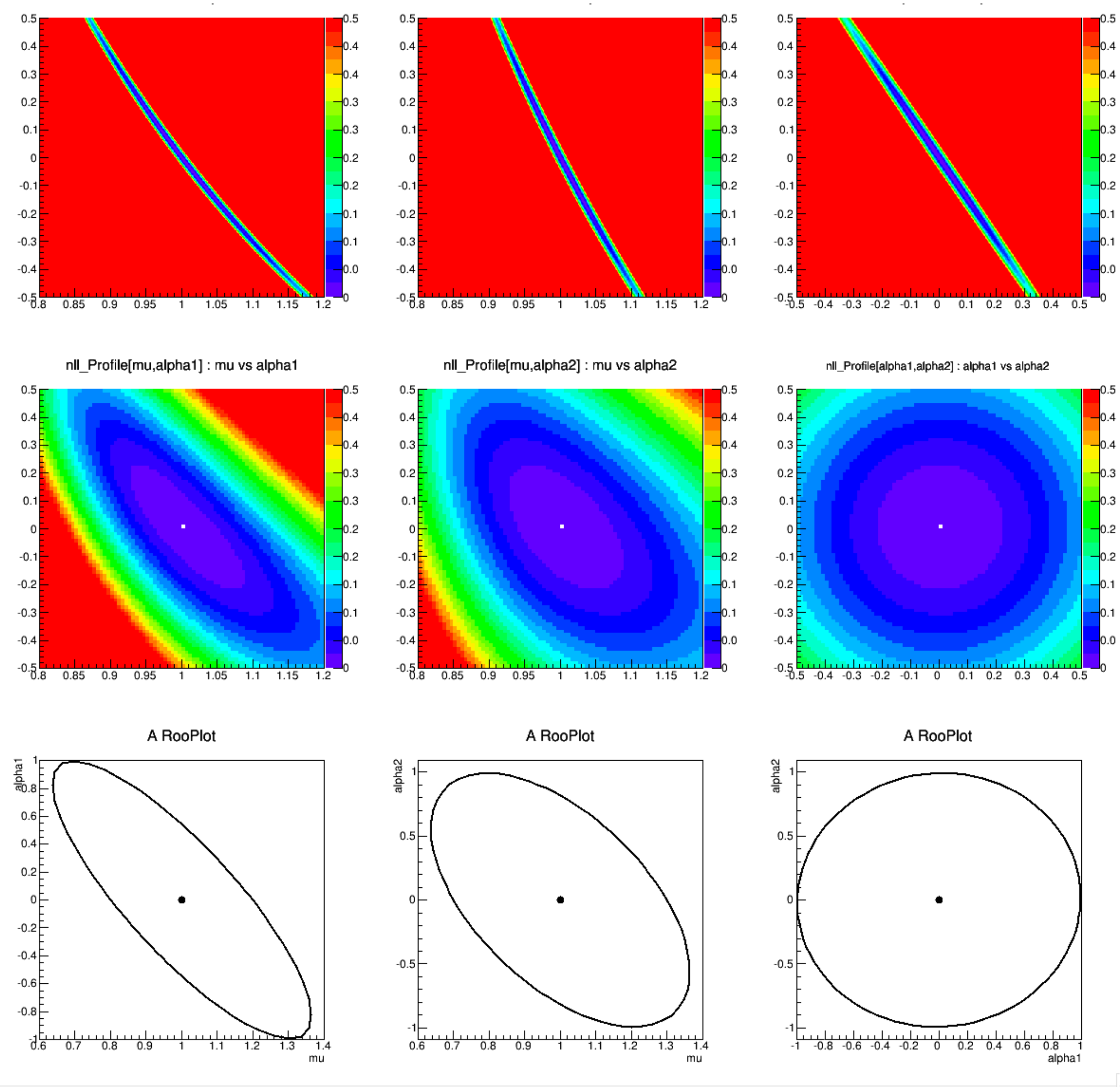
$$-\log L(a,b,\hat{c})$$

Profile likelihood

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

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

Error ellipse from HESSE



-logL(μ,α$_1$,α$_2$) – 100.000 events

Slice in -logL

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

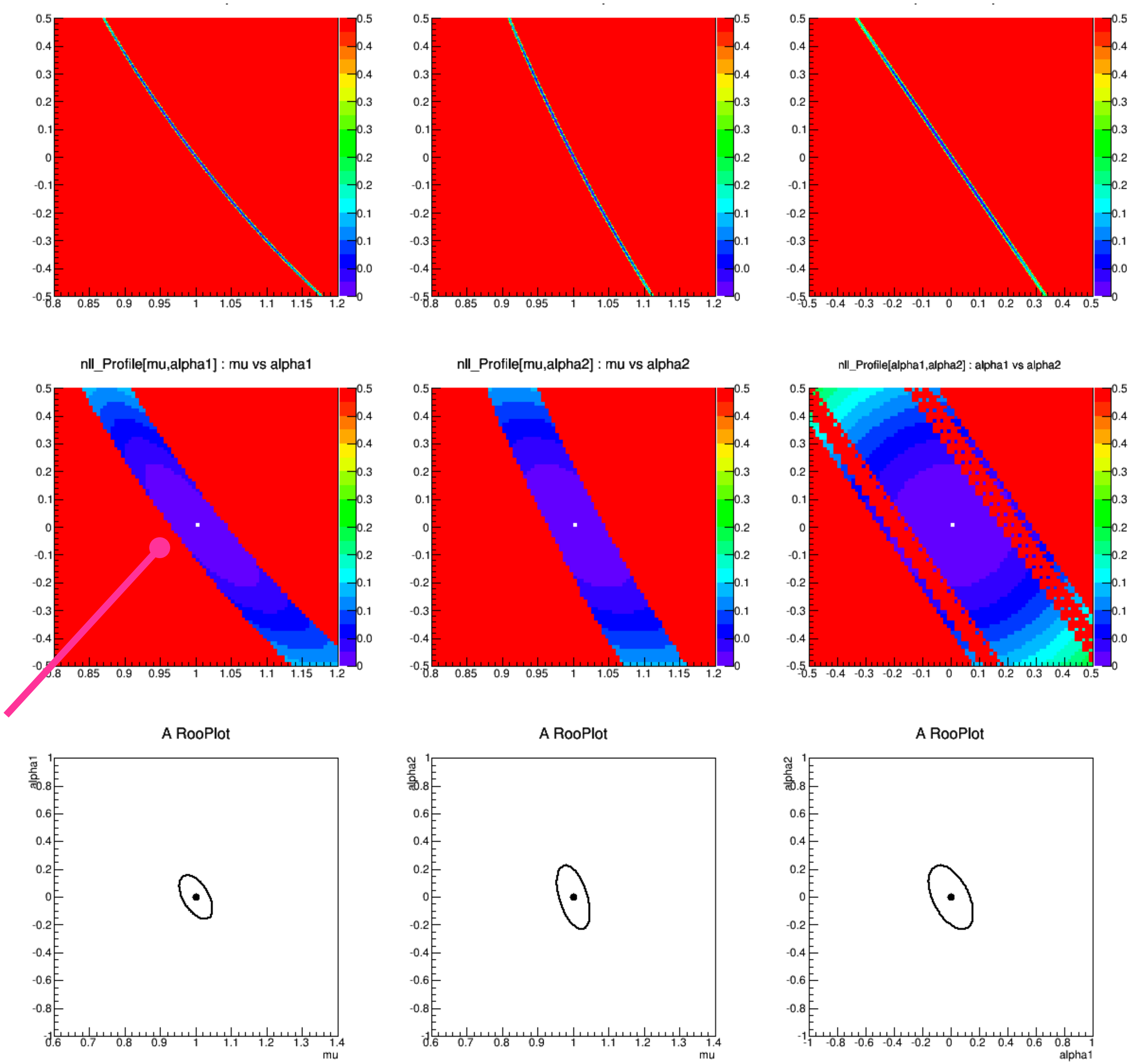
Profile likelihood

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

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



-logL(μ,α_1,α_2) – 1.000.000 events

Slice in -logL

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

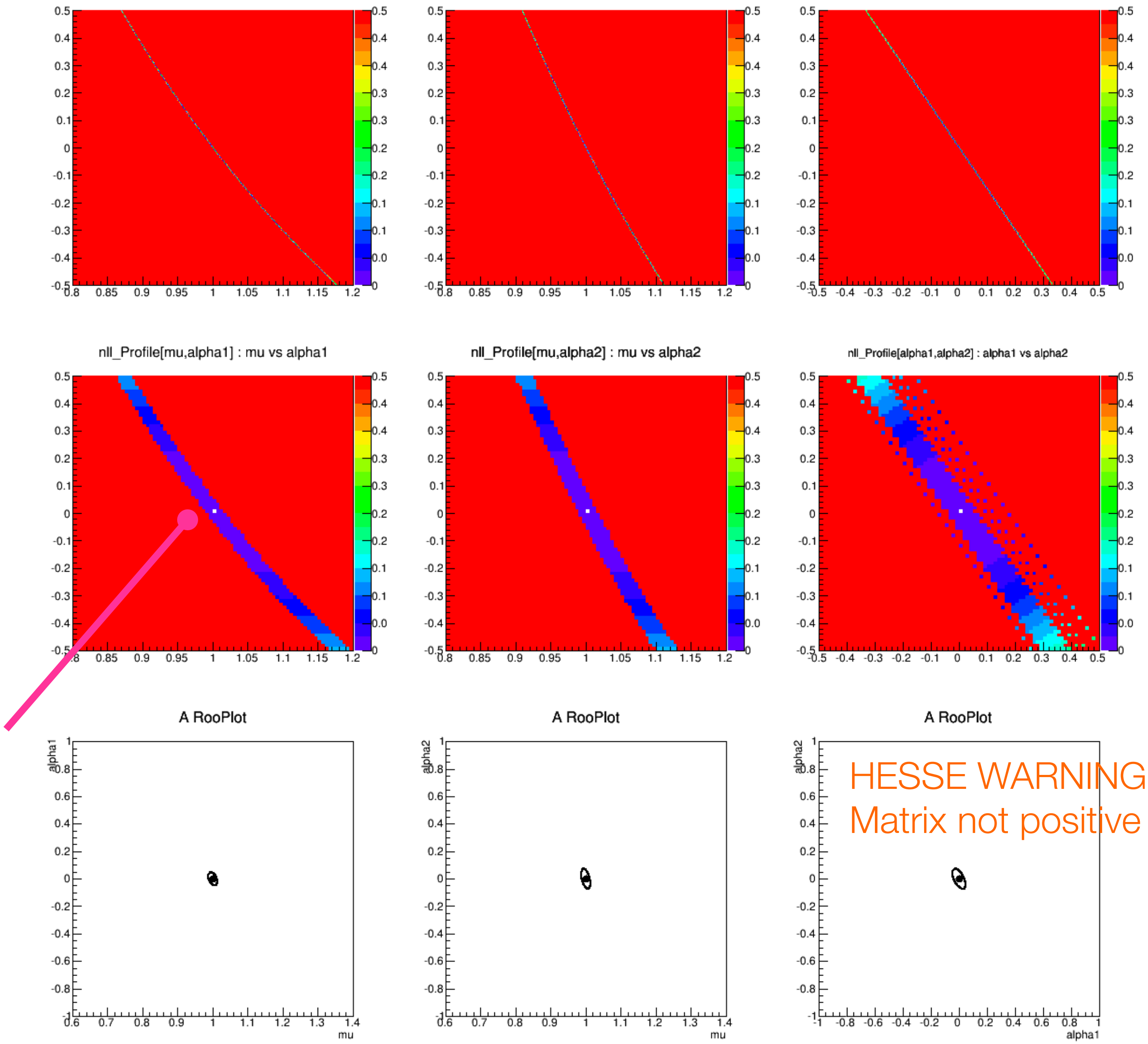
Profile likelihood

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

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



HESSE WARNING:
Matrix not positive definit

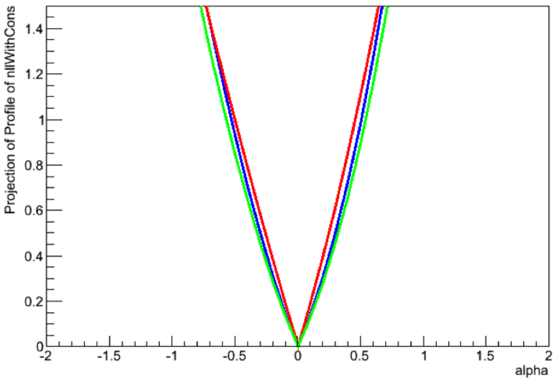
Conclusions on strong correlations

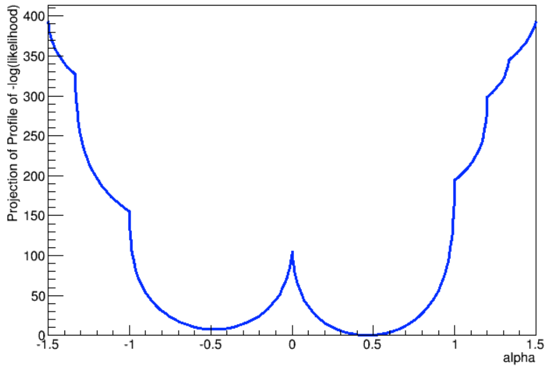
- MINUIT can handle strong correlations very well, but at some point algorithm breaks down
  - Notably HESSE will fail when ratio of weakest-to-strongest eigenvalue < 10^{-6}
- Diagnostic of the existence of strong correlations can be difficult
  - In simple models (Ex 1) this is reflected correlation 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

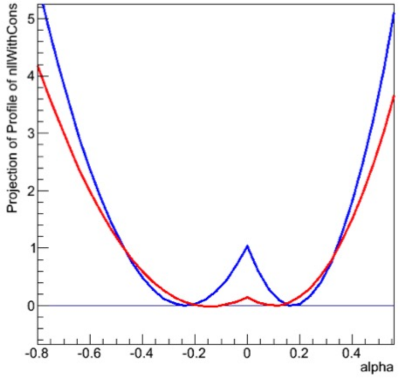
Other likelihood pathologies

- Template morphing algorithms can introduce various other pathologies in the likelihood that cause minimizers 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

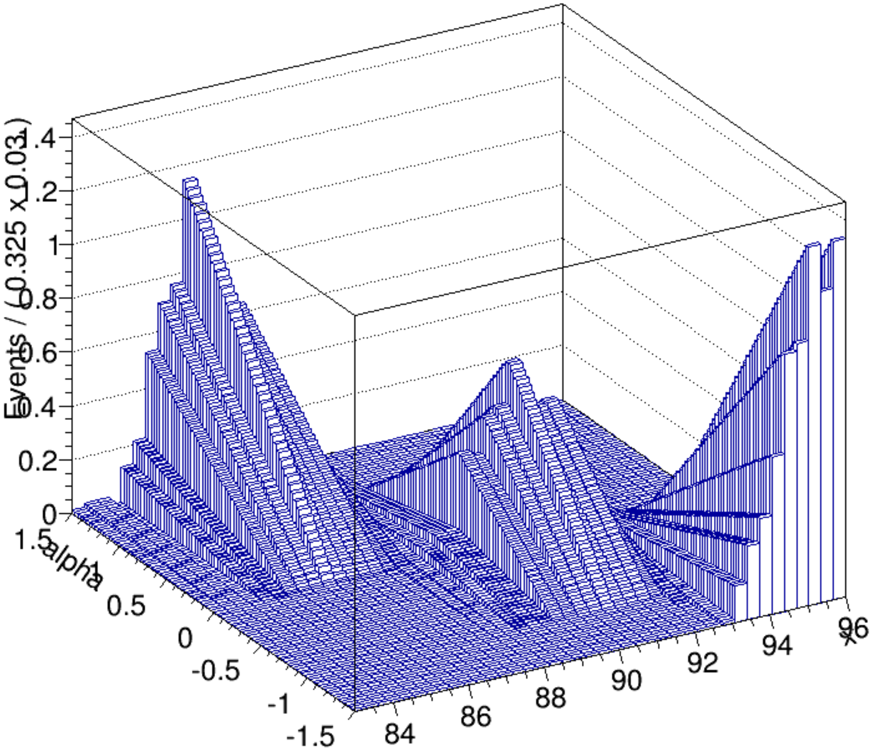




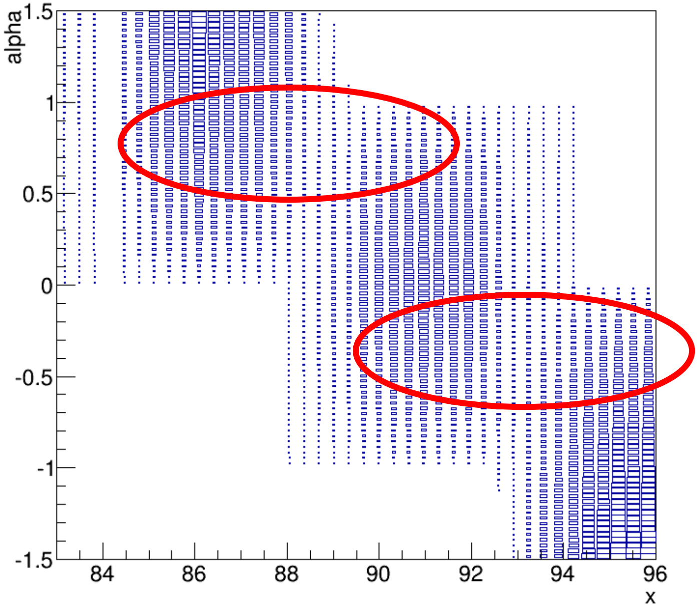


Limitations of piece-wise linear interpolation

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

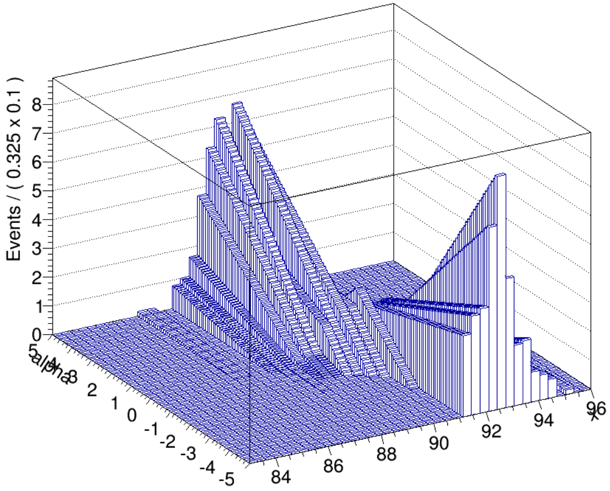


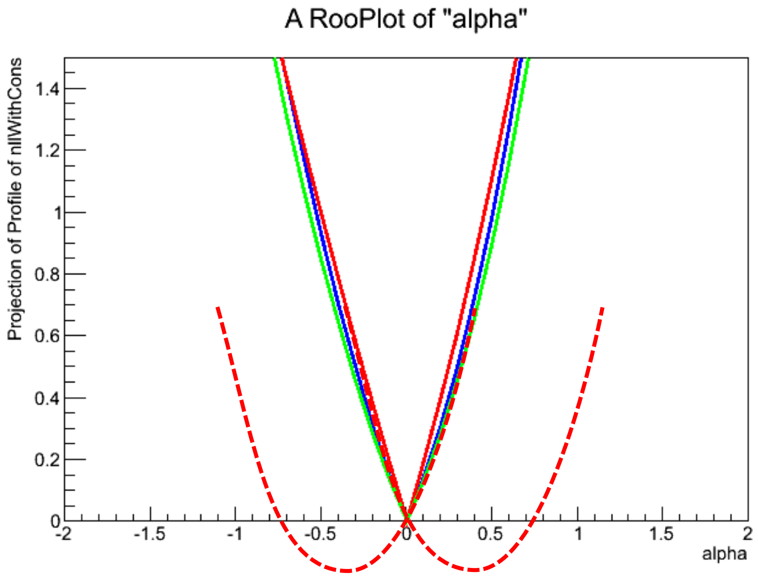
Note double peak structure around |α|=0.5



Non-linear interpolation options

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





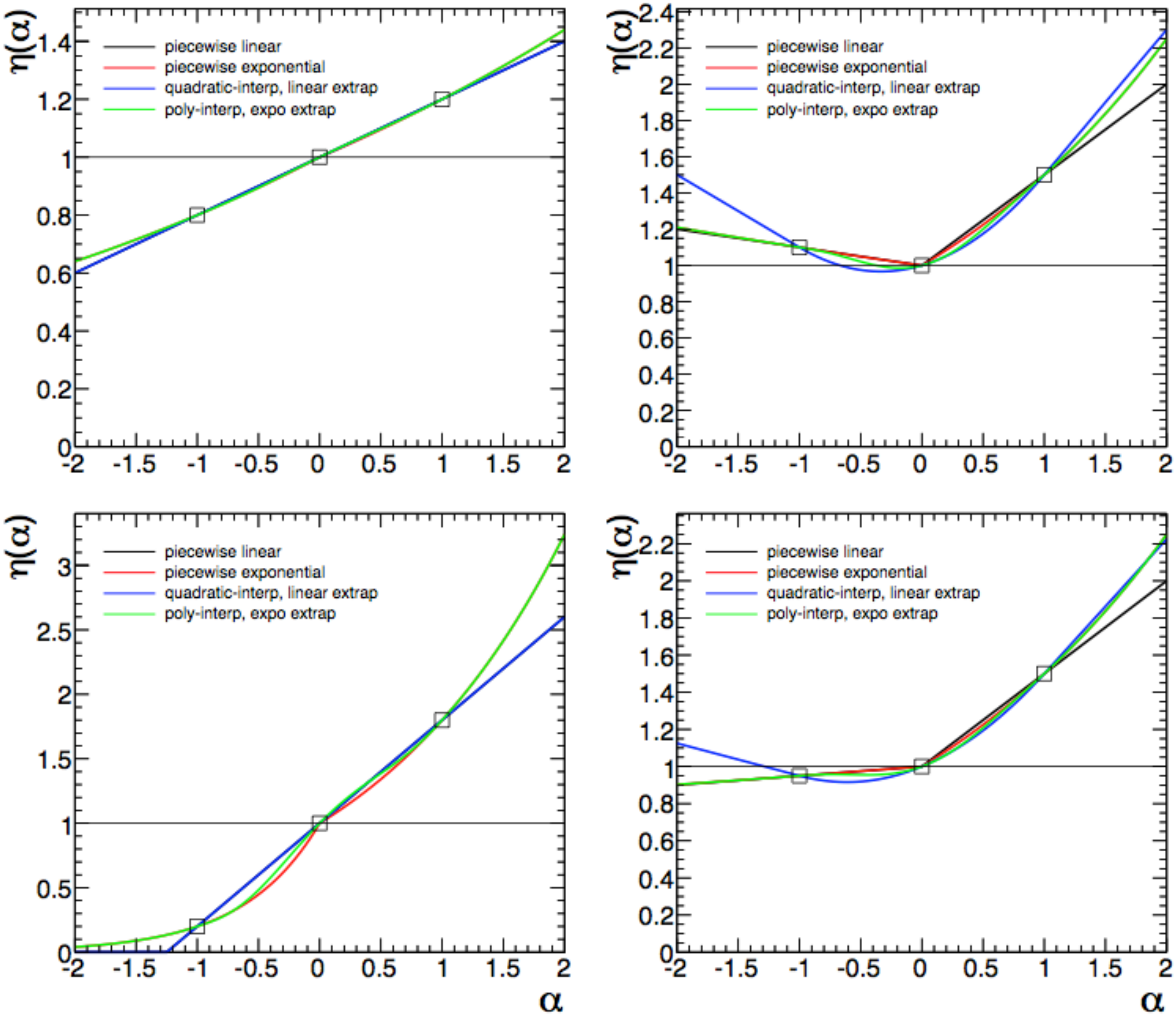
L(α>0) predicts α<0

L(α<0) predicts α>0

- 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

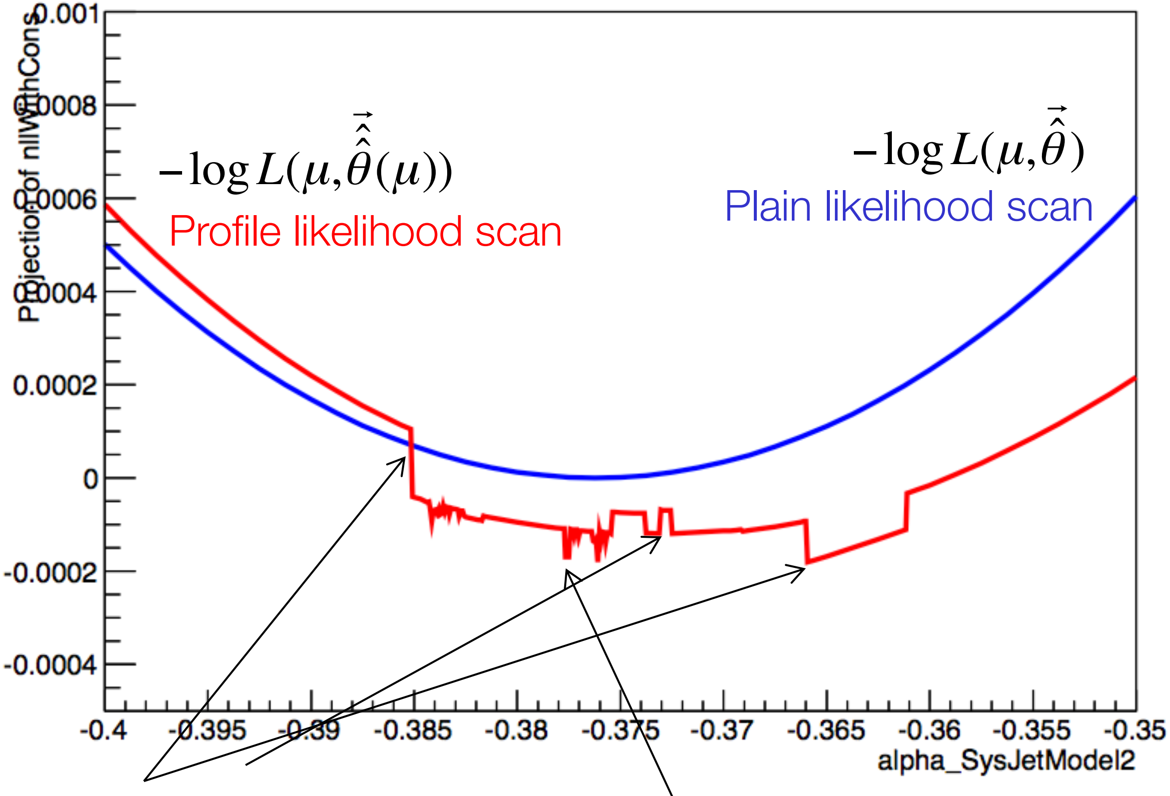
- Comparison of common interpolation options



Wouter Verkerke, NIKHEF

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



Jump to another minimum solution
in one of the profiled θ parameters

Jitter/noise

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)

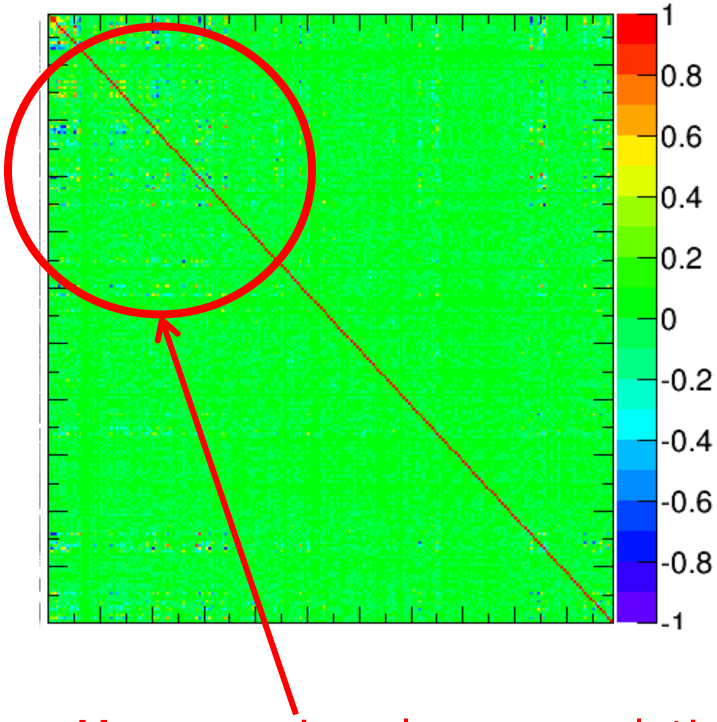
```
VariableMetric: start iterating until Edm is < 0.001
VariableMetric: Initial state   - FCN =  -289.1204081677 Edm =      46.0713 NCalls =   1826
VariableMetric: Iteration #   1 - FCN =  -299.3073097602 Edm =      9.18415 NCalls =   2226
VariableMetric: Iteration #   2 - FCN =  -304.9468725143 Edm =      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: <u>simplify model</u>: eliminate nuisance parameters that suffer from dominant MC statistical effects (causing multiple minima, kinks etc…)

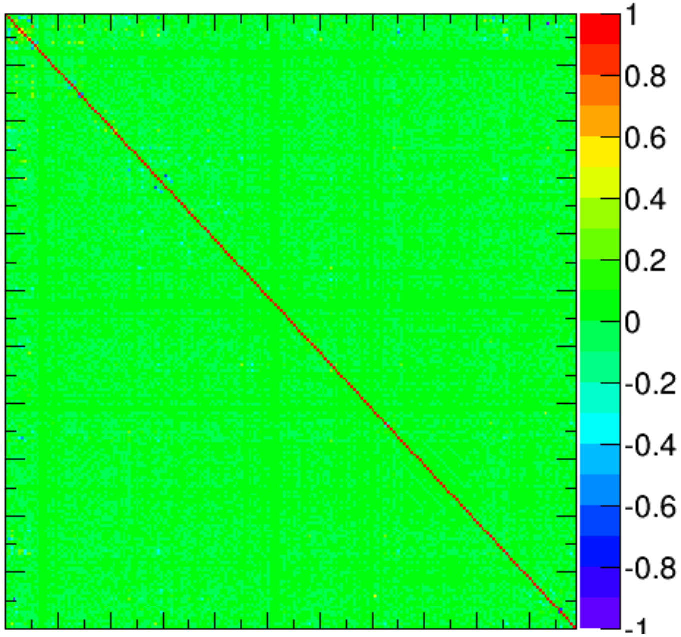
Other likelihood pathologies

- 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

V from Davidon-Fletcher-Powell



V from inversion of Hessian



Many spurious large correlations

- 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

Being a good physicist – **Understand your model!**

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

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

⬇ ⬇

- Our mental picture: “measures s” “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

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

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

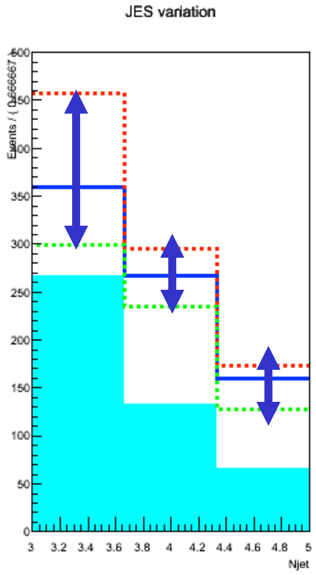
Understanding your model – what constrains your NP

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

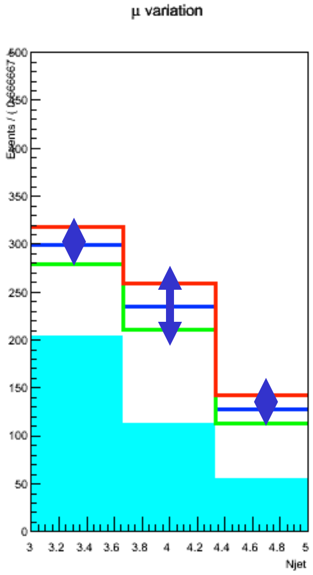
$$L(\vec{N}\,|\,\mu,\alpha_{JES}) = \prod_{i=3,4,5} Poisson(N_i\,|\,(\mu\cdot\tilde{s}_i\cdot+\tilde{b}_i)\cdot r_s(\alpha_{JES})))\cdot Gauss(0\,|\,\alpha_{JES},1)$$

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

Effect of changing α_{JES}

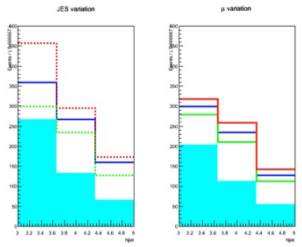


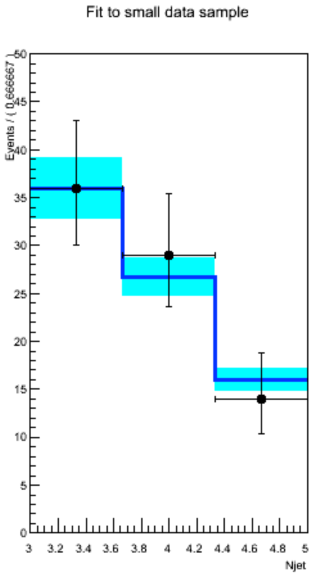
Effect of changing μ



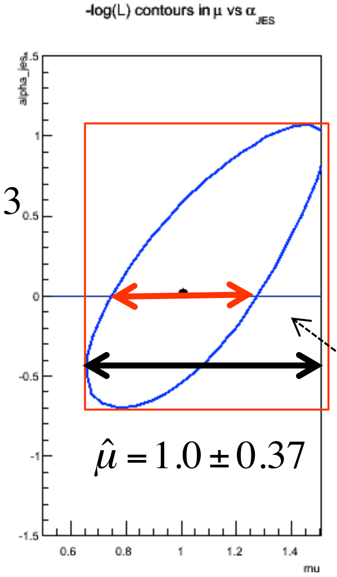
Understanding your model – what constrains your NP

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





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



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

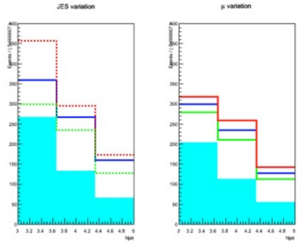
Estimators of μ, α correlated due to similar response in physics measurement

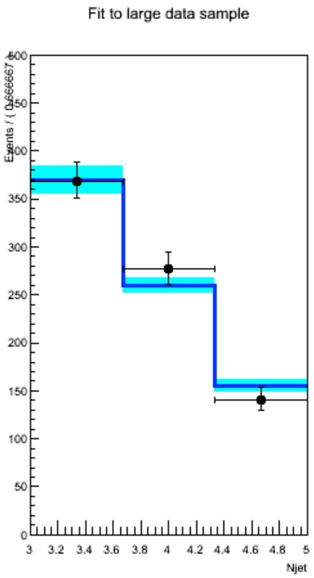
Uncertainty on μ with/without effect of JES

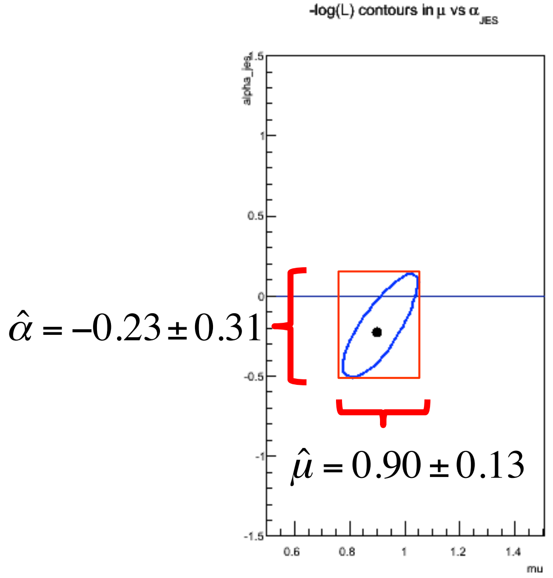
- Is this fit OK?
  - Effect of JES uncertainty propagated in to μ via response modeling in likelihood. Increases total uncertainty by about a factor of 2
  - Estimated uncertainty on α is not precisely 1, as one would expect from unit Gaussian subsidiary measurement…

Understanding your model – what constrains your NP

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







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

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

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

- Is this fit OK?
  - Uncertainty of JES NP *much reduced* w.r.t. subsidiary meas. (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?

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

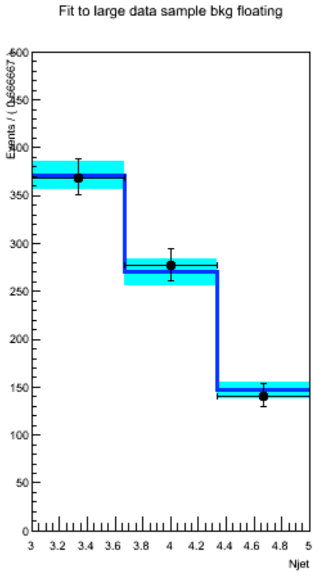
- Reasons for concern
  - Incomplete modeling of systematic uncertainties,
  - Or more generally, model insufficiently detailed

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

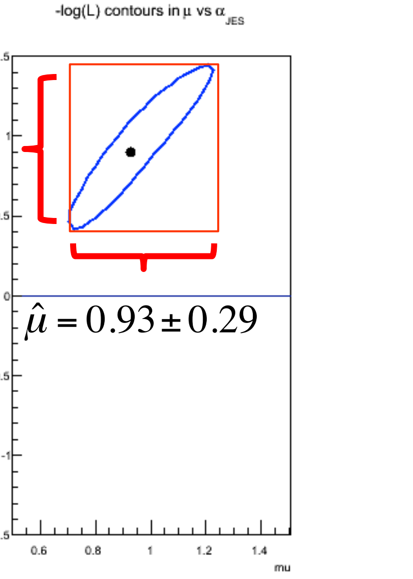
$$L(\vec{N}\,|\,\mu,\alpha_{JES},\alpha_{bkg}) = \prod_{i=3,4,5} Poisson(N_i\,|\,(\mu\cdot\tilde{s}_i\cdot+\tilde{b}_i\cdot \underbrace{r_b(\alpha_{bkg})}_{\text{Background rate response function}})\cdot r_s(\alpha_{JES})))\cdot Gauss(0\,|\,\alpha_{JES},1)\cdot \underbrace{Gauss(0\,|\,\alpha_{bkg},1)}_{\text{Background rate subsidiary measurement}}$$

- 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

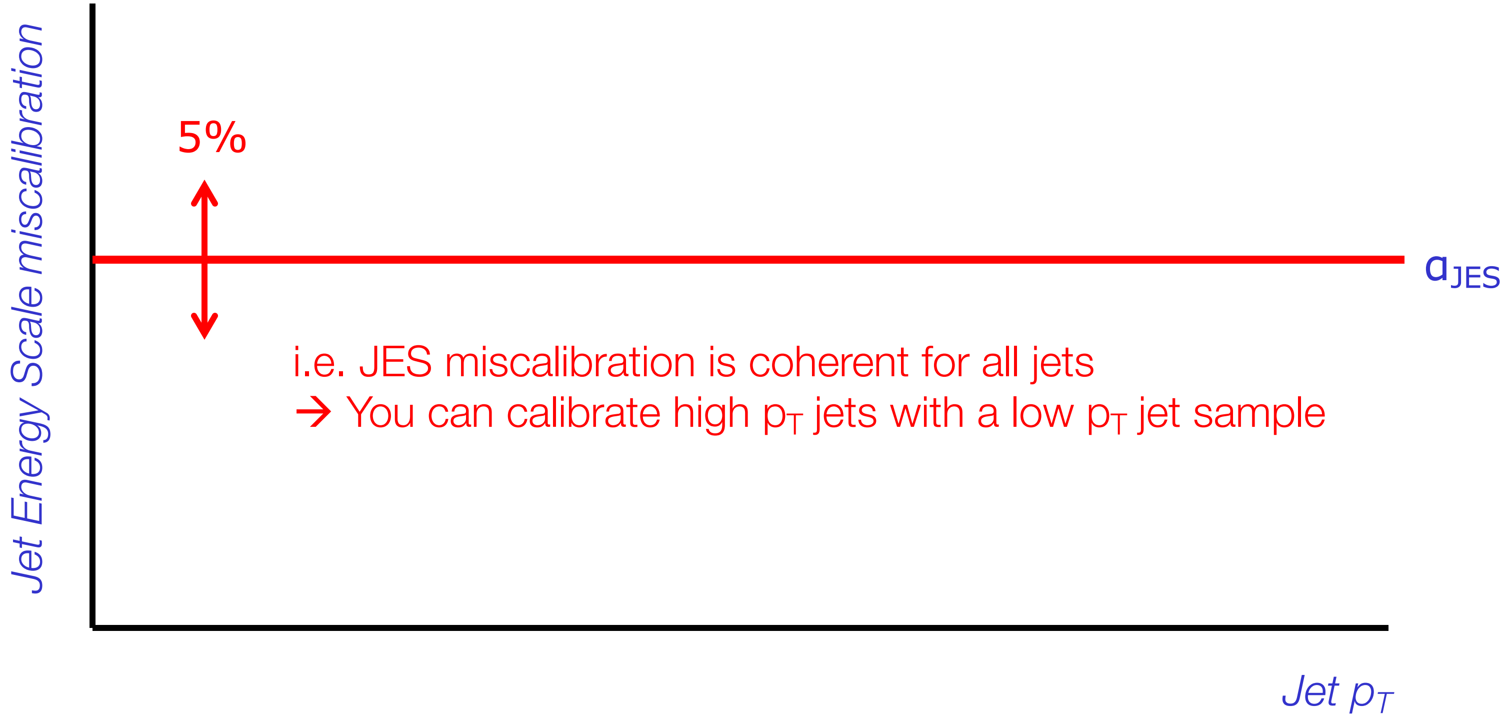
- Lesson learned: if probability model of a physics measurement is insufficiently detailed (i.e. flexible) you can *underestimate* uncertainties
- Normalized subsidiary measurement provide an excellent diagnostic tool
  - Whenever estimates of a NP associated with unit Gaussian subsidiary measurement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

Commonly heard arguments in discussion on over-constraining

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

Dealing with over-constraining – introducing more NPs

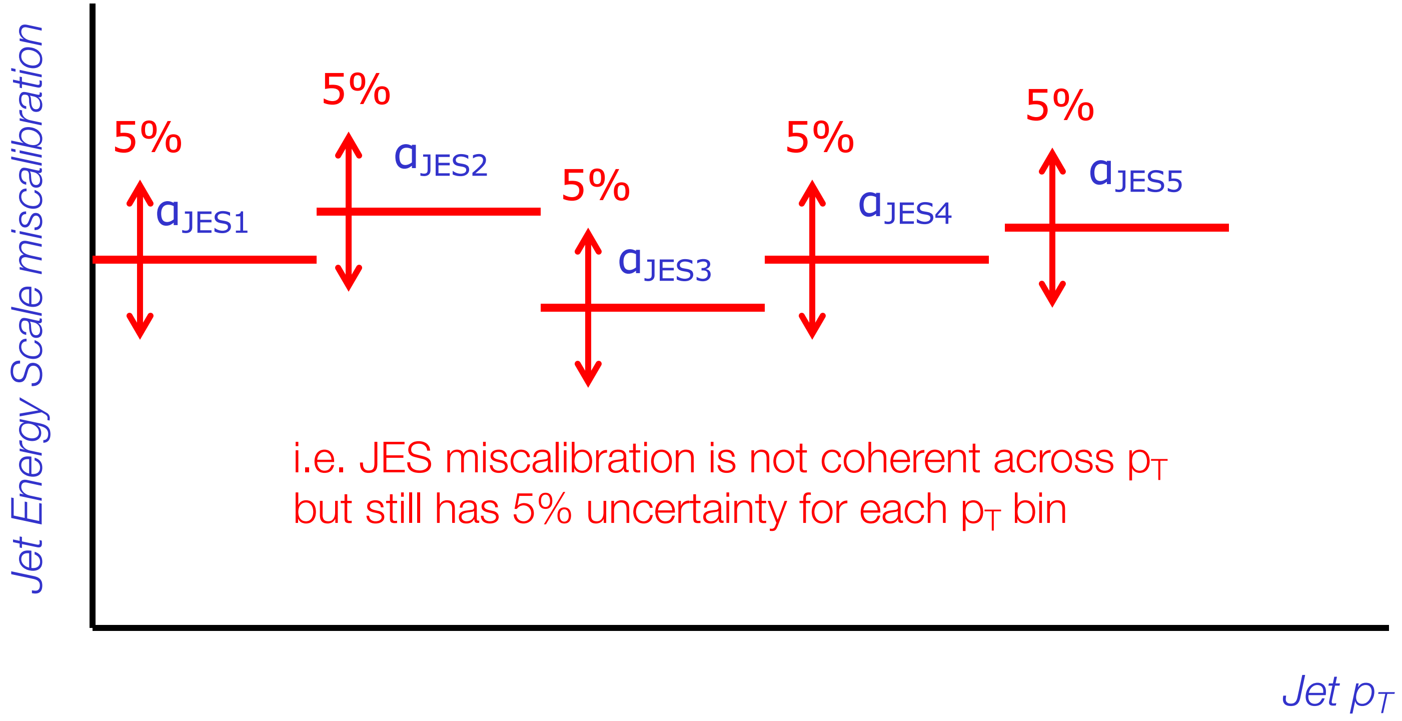
- Some systematic uncertainties are not captured well by one nuisance parameter.
- Written prescription often not clear on *number* of nuisance parameters:
- Does "*the JES uncertainty is 5% for all jets*" mean one NP



i.e. JES miscalibration is coherent for all jets
→ You can calibrate high p_T jets with a low p_T jet sample

Dealing with over-constraining – introducing more NPs

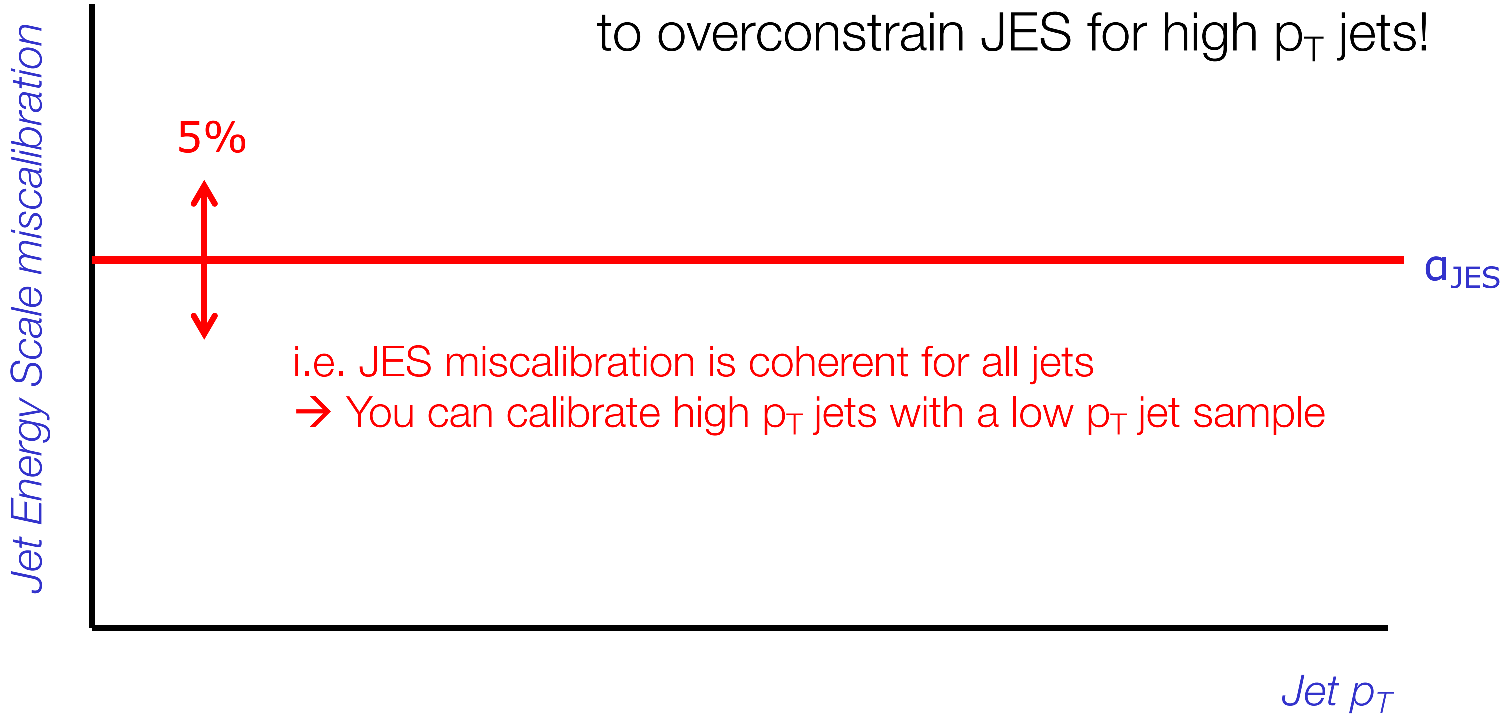
- Some systematic uncertainties are not captured well by one nuisance parameter.
- Written prescription often not clear on *number* of nuisance parameters:
- Or does "*the JES uncertainty is 5% for all jets*" mean 5 NPs?



i.e. JES miscalibration is not coherent across p_T but still has 5% uncertainty for each p_T bin

Dealing with over-constraining – introducing more NPs

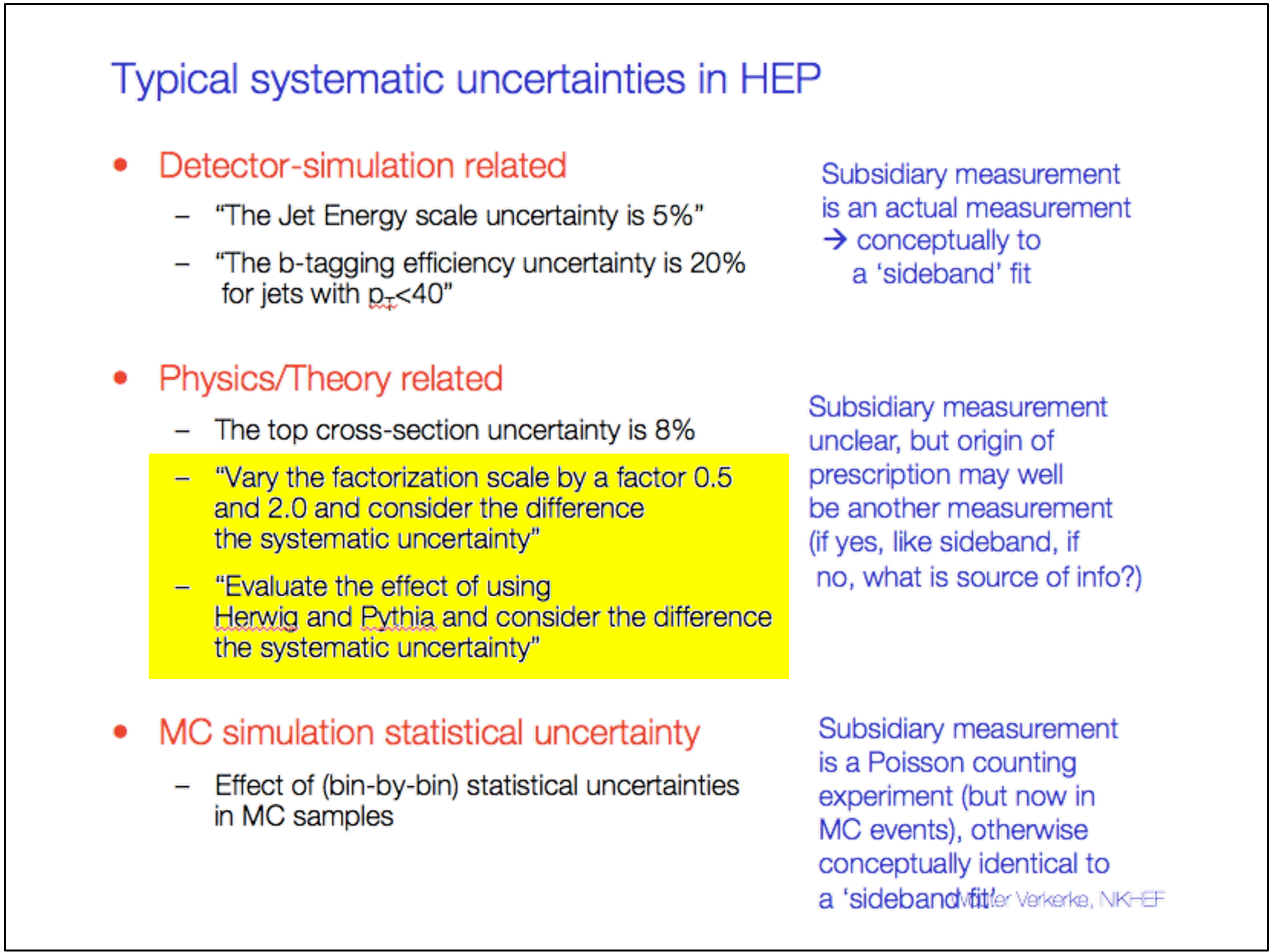
- Some systematic uncertainties are not captured well by one nuisance parameter.
- Written prescription often not clear on *number* of nuisance parameters:
- If you assume one NP – chances are that your physics Likelihood will exploit this oversimplified JES model to overconstrain JES for high p_T jets!



i.e. JES miscalibration is coherent for all jets
→ You can calibrate high p_T jets with a low p_T jet sample

Modeling theory uncertainties

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



Typical systematic uncertainties in HEP

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

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

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

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

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

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

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

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

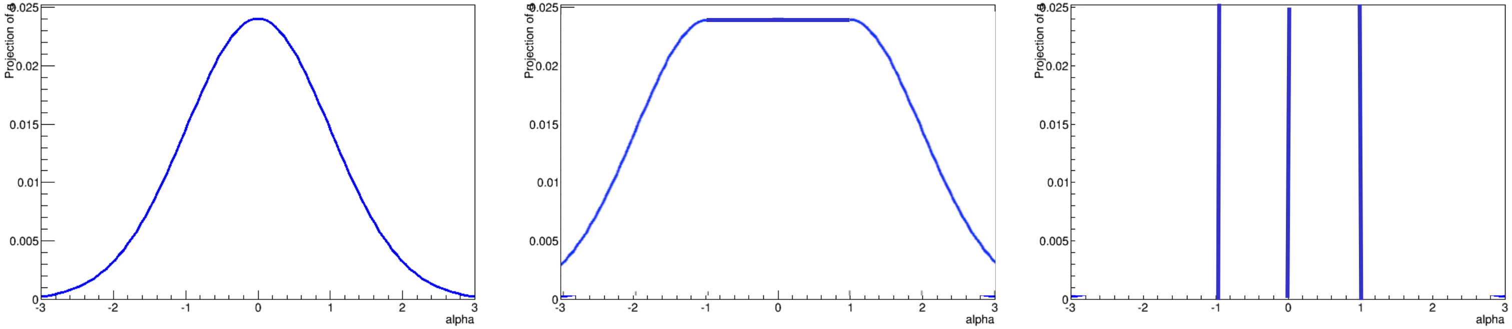
“Vary Factorization Scale by x0.5 and x” → F(α) is probably not Gaussian
So what distribution was meant?

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 – Factorization scale uncertainty

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

  - One parameter: the factorization scale → Clearly described and connected to the underlying theory model
  - You can ask yourself if there are additional uncertainties in the theory model (renormalization scale etc), this a valid, but distinct issue.

- Difficult example – Hadronization/Fragmentation model
  - Source uncertainty**: you run different showering MC generators (e.g. HERWIG and PYTHIA)** and you observe you get different results from your physics analysis
  - **How do you model this in the likelihood?**

Modeling theory uncertainties

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

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

Specific issue with theory uncertainties

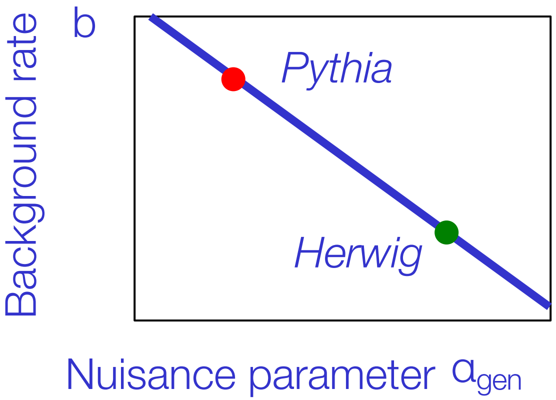
- It is difficult to define rigorous statistical procedures to deal with such 2-point uncertainties. So you need to decide
- If their estimated effect is small, you can pragmatically ignore these lack of proper knowledge and 'just do something reasonable' to model these effects in a likelihood
- If their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

Specific issues with theory uncertainties

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

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

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

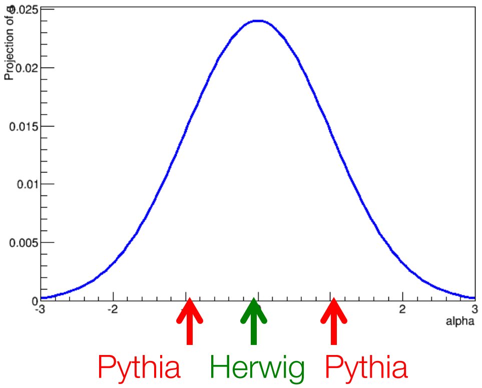


- A thorny question remains:
  What is the subsidiary measurement for α?
  This should reflect you current knowledge on α.

Specific issues with theory uncertainties

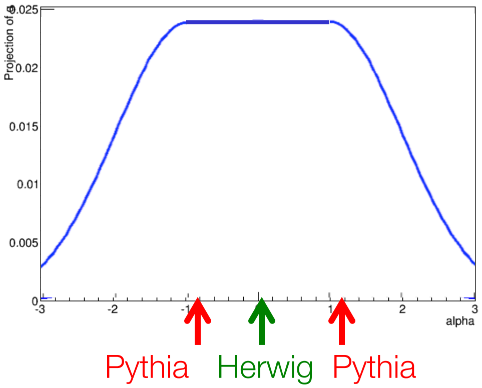
- Subsidiary measurement of a theoretical 2-point uncertainty effectively quantifies the ‘knowledge’ on these models
  - *Extra difficult to make meaningful statement about this*, since meaning of parameter is not well embedded in underlying theory model
  - But again, all procedures need to assume some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

Gaussian



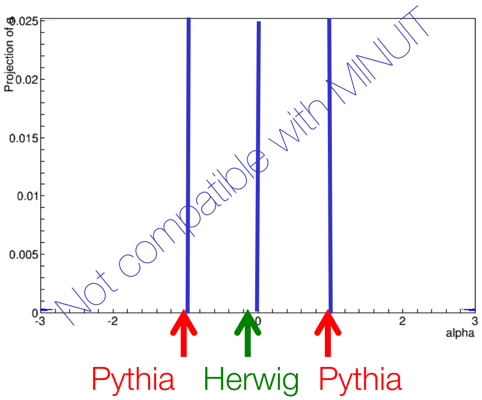
Prefers Herwig at 1σ

Box with
Gaussian wings



All predictions ‘between’ Herwig and Pythia equally probable

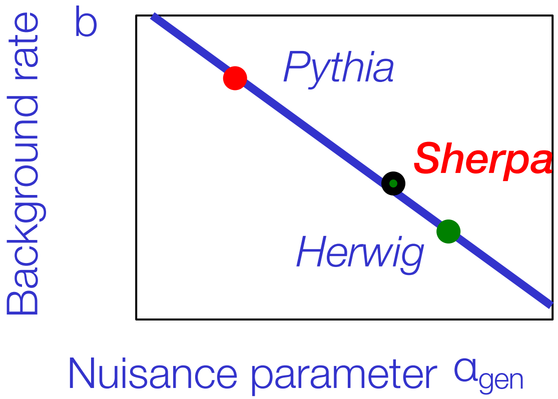
Delta fuctions



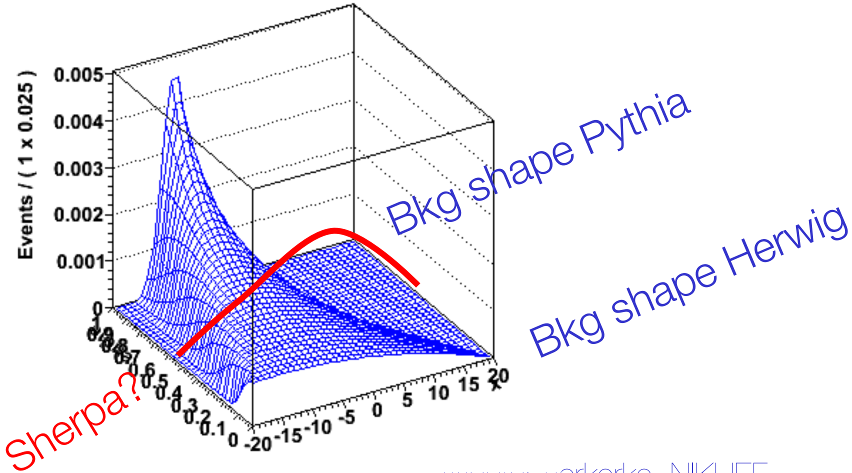
Only ‘pure’ Herwig and Pythia exist

Two-point systematics on non-counting measurements

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

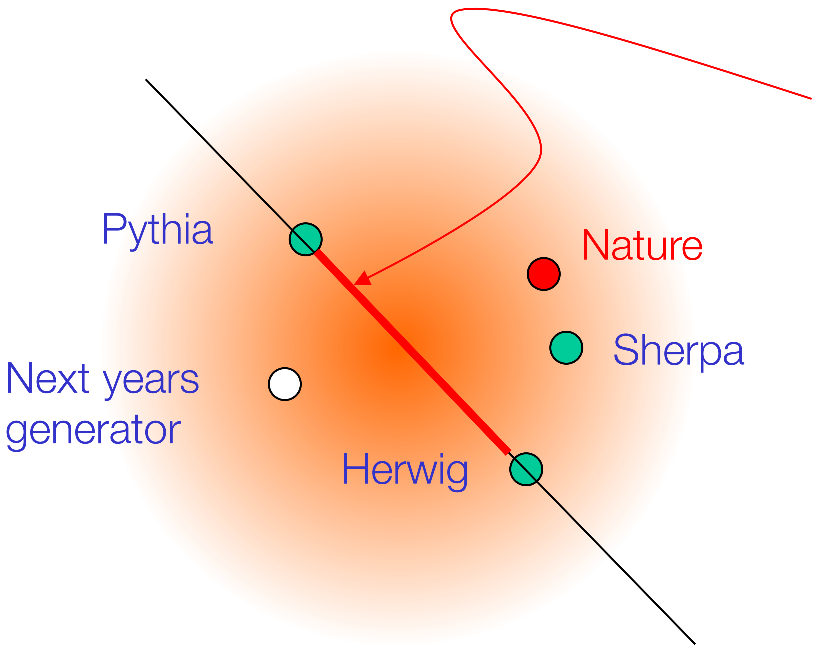


- This is not generally true for distributions. A shape interpolation between 'pythia' and 'herwig' does not necessarily describe shape of 'sherpa' (or of Nature!)
  - Fundamental modeling problem!
  - You may need more parameters…



Dealing with ‘two-point’ uncertainties

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



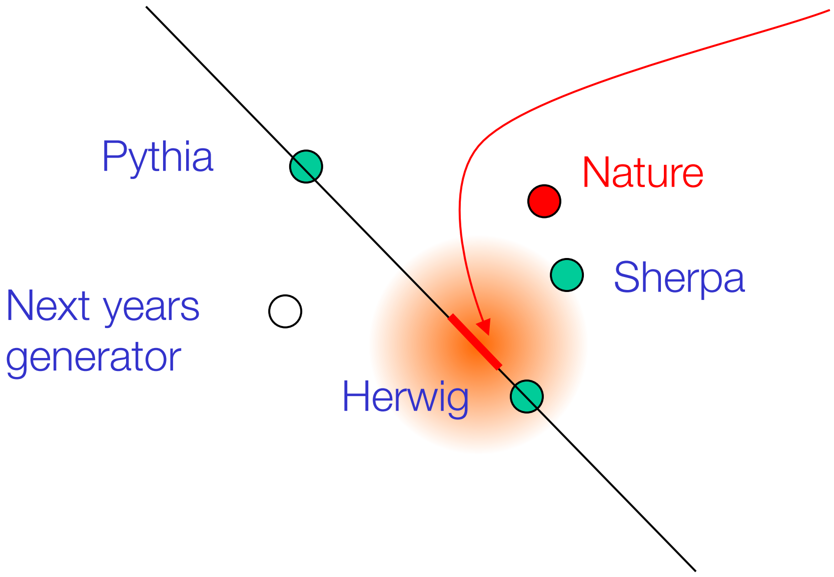
Modeled uncertainty (1 dimension) assuming ‘nature is on line’

Effectively captured uncertainty

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

Dealing with ‘two-point’ uncertainties

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

Summary

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

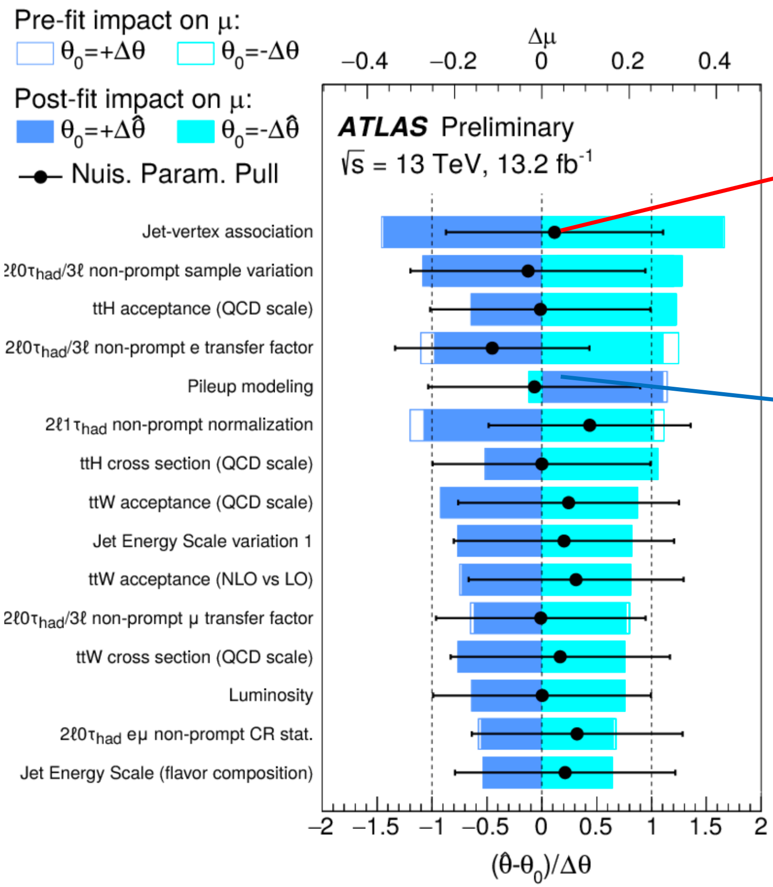
Summary

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

Fit diagnostics – NP ranking/impact plots

Does the fit constrain (reduce) the systematic uncertainty from the data, based on the choice of NP model, w.r.t. the input specifications?

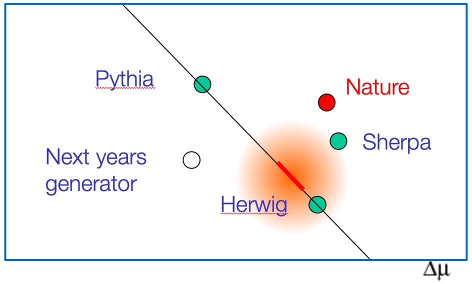
→ **Diagnostics are crucial!**

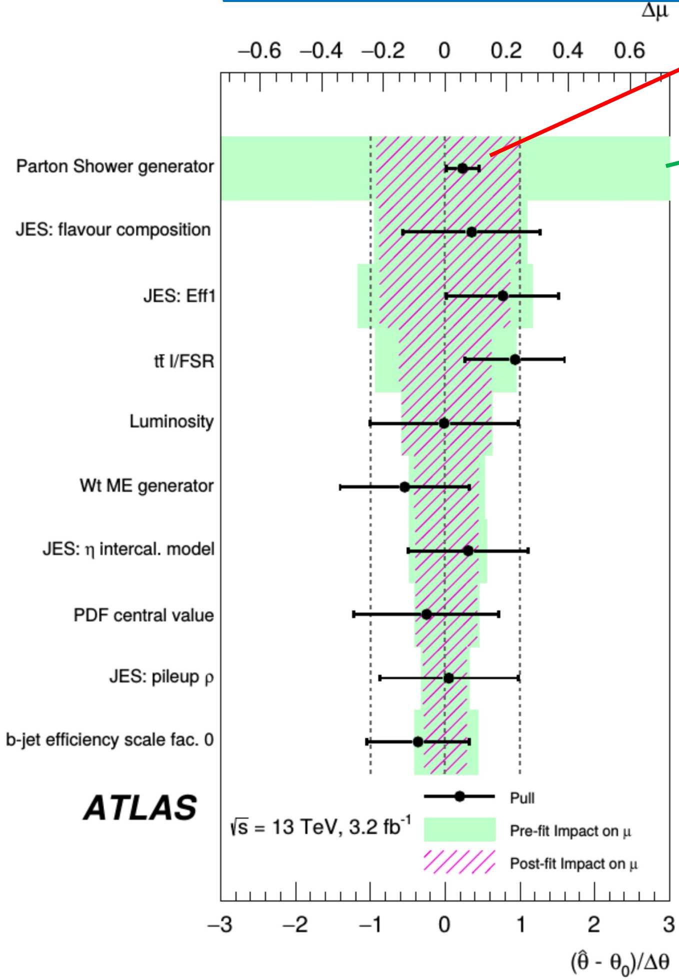


Physics data biases / constrains systematic uncertainty if not 0 +/- 1

Impact quantifies correlation with POI. Small impact → NP is (almost) irrelevant for this analysis

Fit diagnostics – NP ranking/impact plots





Physics data biases / constrains systematic uncertainty if not 0 +/- 1

Impact quantifies correlation with POI. Small impact → NP is (almost) irrelevant for this analysis

NP bias or constraint can be due to

1) Statistical fluctuation in data or template (common)
2) Invalid (over)somplified NP model (common)
3) Genuine physics information (not common)

If impact large: always investigate and fix as needed
If impact is small, may ignore, use your judgement

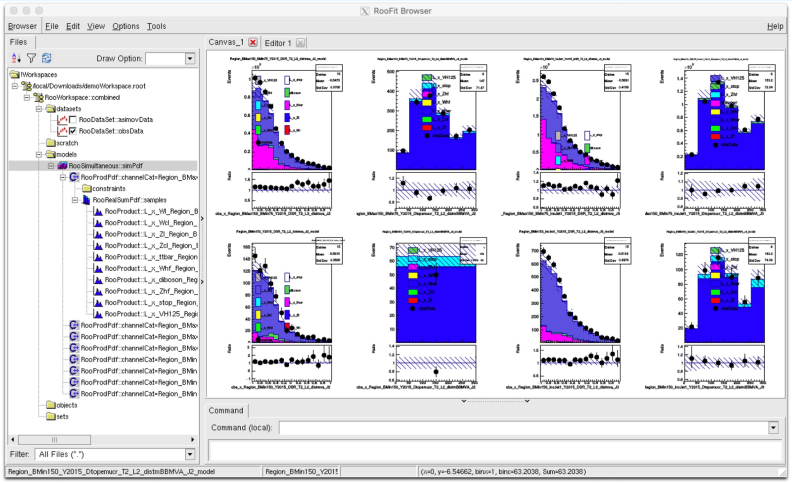
Instructive to look both at *expected* and *observed* NP rankings

- Expected has no data fluctuations (Asimov)
- Additional pulls/constraints in ‘observed’ NP rankings have origin in data

Visualization of model predictions in observable space useful diagnostic!

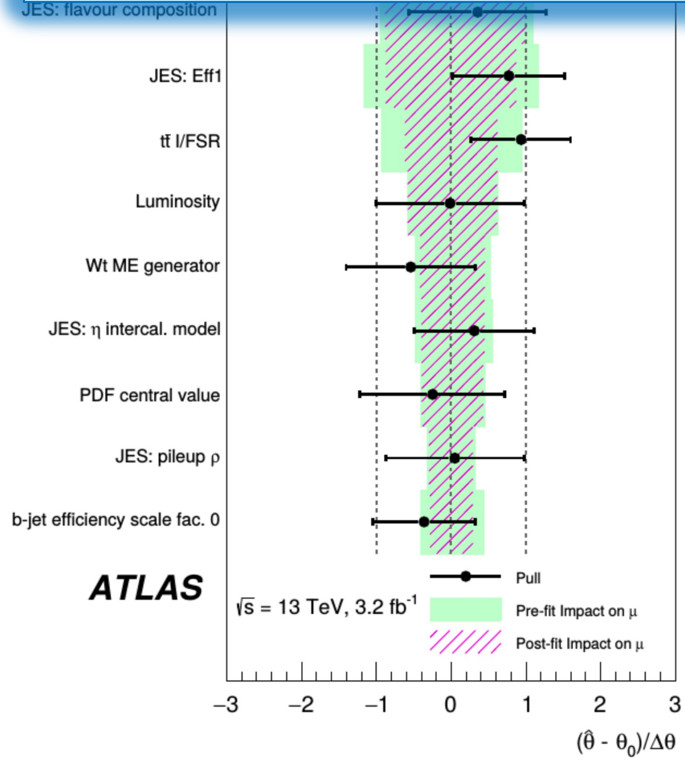
- Localize fluctuations in templates that constrain/pull fits
- Observe magnitude of model change with variation of NPs within uncertainty

```
'ex16.C'
```



```
!Workspaces
  /local/Downloads/demoWorkspace.root
    RooWorkspace::combined
      datasets
        □ RooDataSet::asimovData
        ☑ RooDataSet::obsData
      scratch
      models
        RooSimultaneous::simPdf
          RooProdPdf::channelCat=Region_BMax
            constraints
            RooRealSumPdf::samples
              RooProduct::L_x_Wl_Region_B
              RooProduct::L_x_Wcl_Region_
              RooProduct::L_x_Zl_Region_B
              RooProduct::L_x_Zcl_Region_I
              RooProduct::L_x_ttbar_Region
              RooProduct::L_x_Whf_Region_
              RooProduct::L_x_diboson_Reg
              RooProduct::L_x_Zhf_Region_
              RooProduct::L_x_stop_Region_
              RooProduct::L_x_VH125_Regic
          RooProdPdf::channelCat=Region_BMax
          RooProdPdf::channelCat=Region_BMax
          RooProdPdf::channelCat=Region_BMax
          RooProdPdf::channelCat=Region_BMin
          RooProdPdf::channelCat=Region_BMin
          RooProdPdf::channelCat=Region_BMin
          RooProdPdf::channelCat=Region_BMin
      objects
      sets
```

```
Region_BMin150_Y2015_Dtopemucr_T2_L2_distmBBMVA_J2_model | Region_BMin150_Y2015 | (x=0, y=-6.54662, binx=1, binc=63.2038, Sum=63.2038)
```



NP bias or constraint can be due to

1) Statistical fluctuation in data or template (common)
2) Invalid (over)somplified NP model (common)
3) Genuine physics information (not common)

If impact large: always investigate and fix as needed
If impact is small, may ignore, use your judgement

Instructive to look both at *expected* and *observed* NP rankings

- Expected has no data fluctuations (Asimov)
- Additional pulls/constraints in 'observed' NP rankings have origin in data