Statistics for CAKE

Ata Sattari

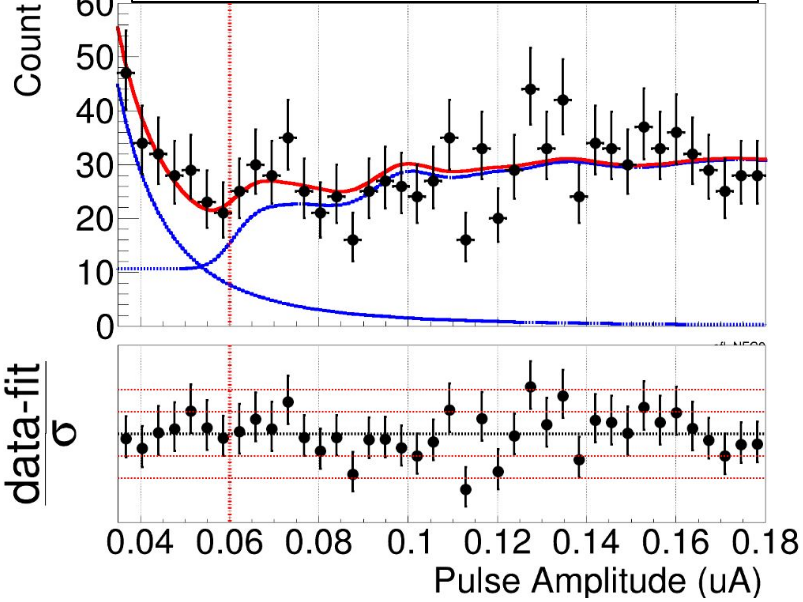
UNIVERSITY OF TORONTO

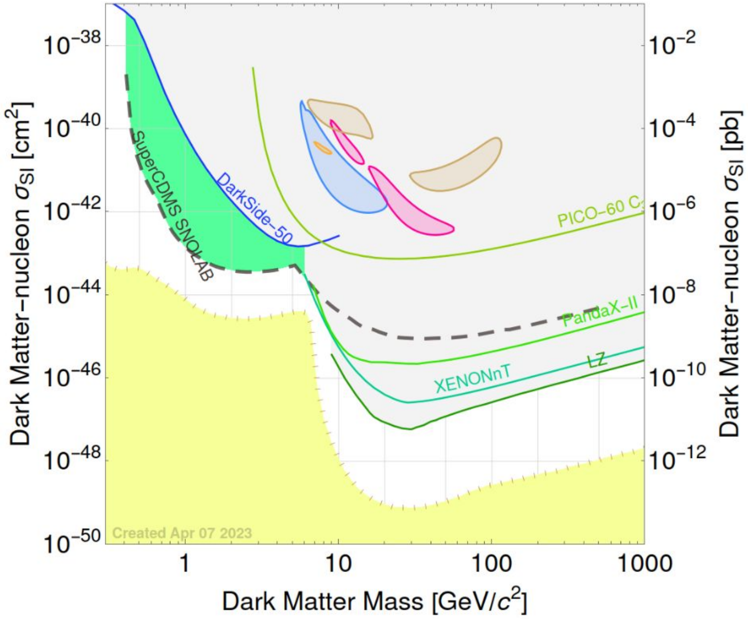
SUPER CDMS Cryogenic Dark Matter Search

Why are we here?

1. Discuss how to fit a model to data.
2. Learn how to perform statistical tests. (Upper limits, ….)

F(x|θ) where θ is unknown.



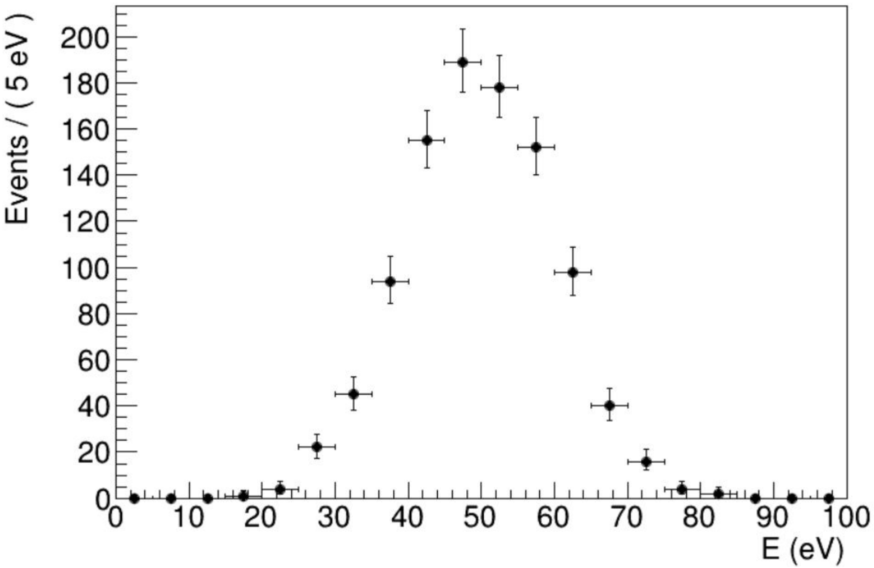


General statistics

Likelihood definition

What is the probability to observe this data?

PDF: Gaussian(E|μ,σ)



Likelihood definition

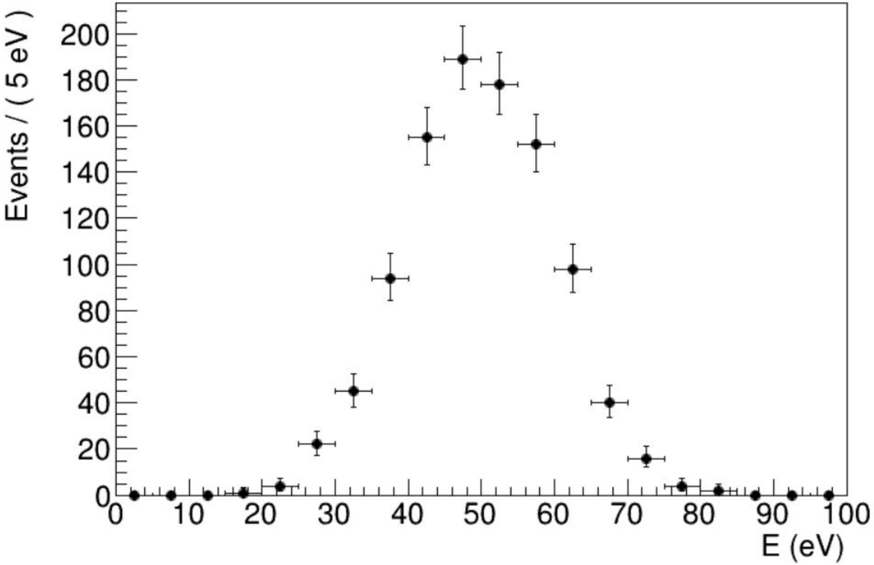
What is the probability to observe this data?

Multiply probabilities to observe events:

$$L(data|\vec{\theta}) = \prod_{Events} \underline{PDF(E_i|\vec{\theta})\cancel{dE}}$$

Shape only fit
(Unbinned likelihood)

PDF: Gaussian(E|μ,σ)



Likelihood definition

What is the probability to observe this data?

PDF: Gaussian(E|μ,σ)

Multiply p

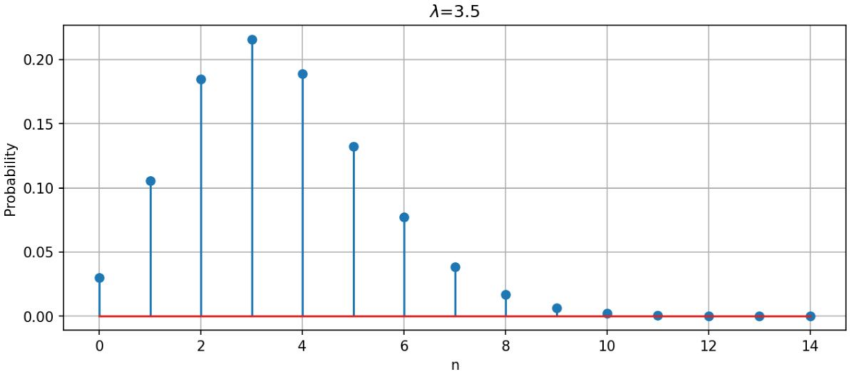
$L(data|\vec{\theta}) =$

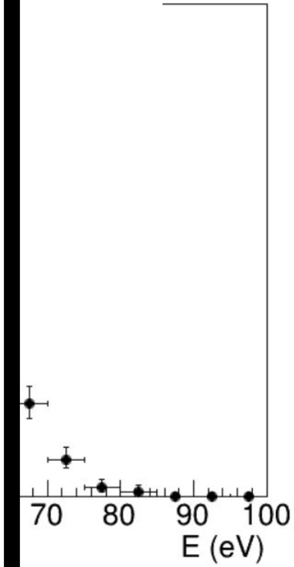
What is the probability to see n events when we expect λ?

n : Integer

λ : Positive real (average number of event)

$$Poisson(n, \lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$





Likelihood definition - Extended unbinned likelihood

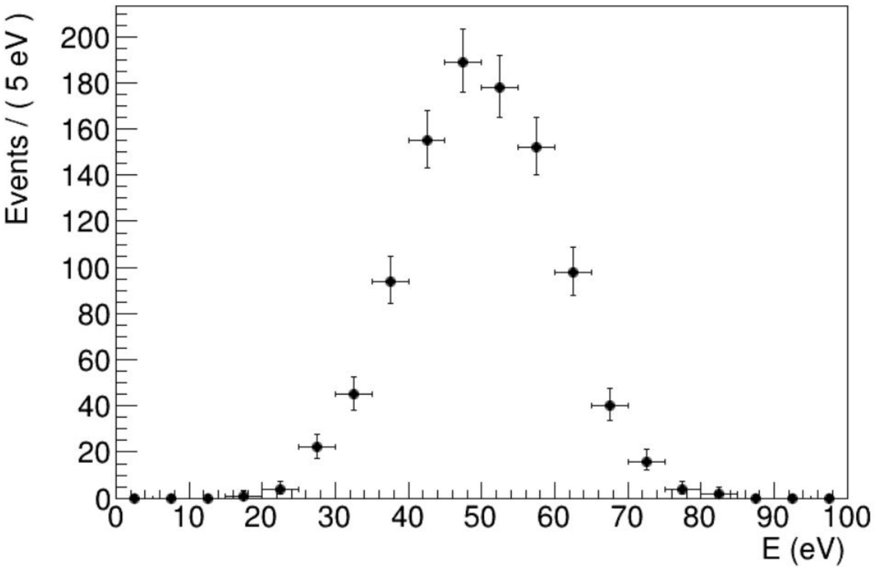
What is the probability to observe this data?

Multiply probabilities to observe events:

$$L(data|\vec{\theta}) = \prod_{Events} PDF(E_i|\vec{\theta})\cancel{dE} \cdot Poisson(N_{total}|\lambda(\vec{\theta}))$$

Shape and event count fit
(Extended unbinned likelihood)

PDF: Gaussian(E|μ,σ)



Likelihood definition

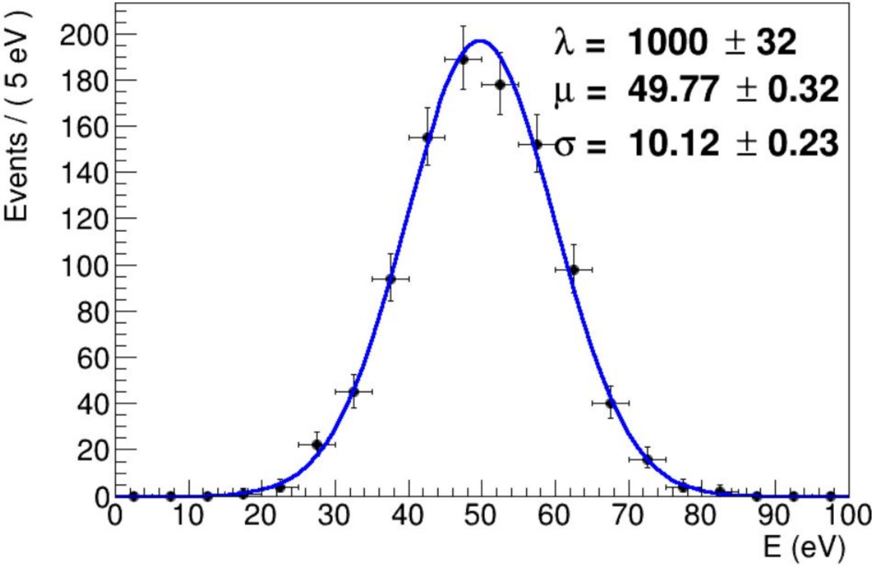
What is the probability to observe this data?

Multiply probabilities to observe events:

$$L(data|\vec{\theta}) = \prod_{Events} PDF(E_i|\vec{\theta})\cancel{dE} \cdot Poisson(N_{total}|\lambda(\vec{\theta}))$$

Shape and event count fit
(Extended unbinned likelihood)

PDF: Gaussian(E|μ,σ)



Choose θ to maximize L(data|θ).

Likelihood definition

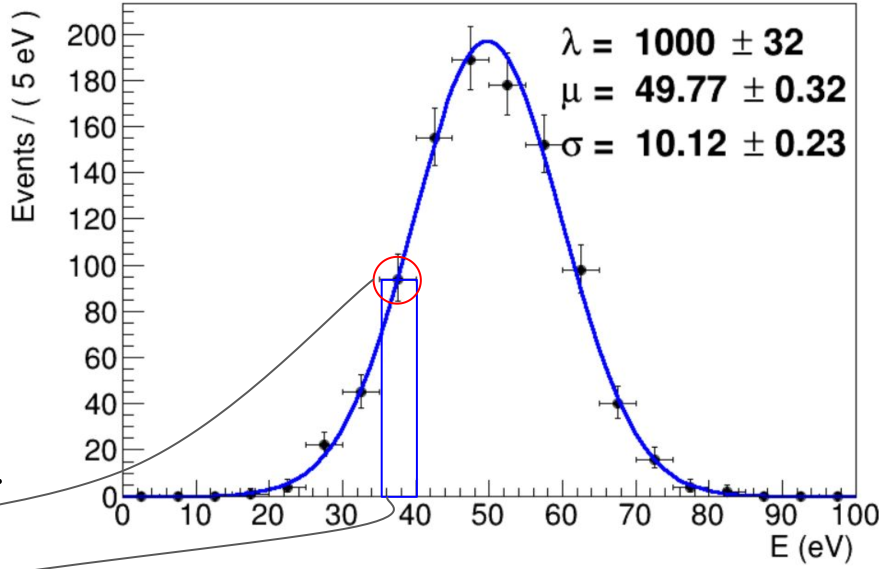
What is the probability to observe this data?

Multiply probabilities to observe events:

$$L(data|\vec{\theta}) = \prod_{Events} PDF(E_i|\vec{\theta})\cancel{dE} \cdot Poisson(N_{total}|\lambda(\vec{\theta}))$$

Shape and event count fit
(Extended unbinned likelihood)

PDF: Gaussian(E|μ,σ)



Alternatively, we can multiply bin probabilities.

$$L(data|\vec{\theta}) = \prod_{Bins} Poisson(n_i|\lambda_i(\vec{\theta}))$$

(Extended binned likelihood)

Choose θ to maximize L(data|θ).

Likelihood definition

What is the … E|μ,σ)

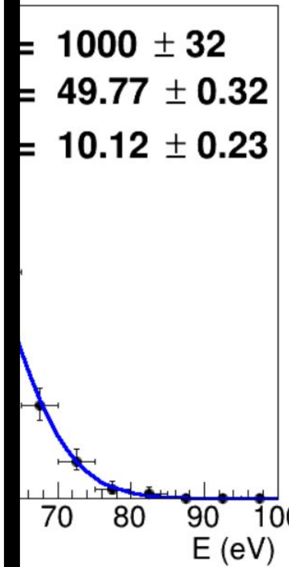
Multiply pro…

$L(data|\vec{\theta}) =$ …

Alternativel…

$L(data|\vec{\theta}) = \prod$ …

(Exte…



… L(data|θ).

Features of binned likelihood:

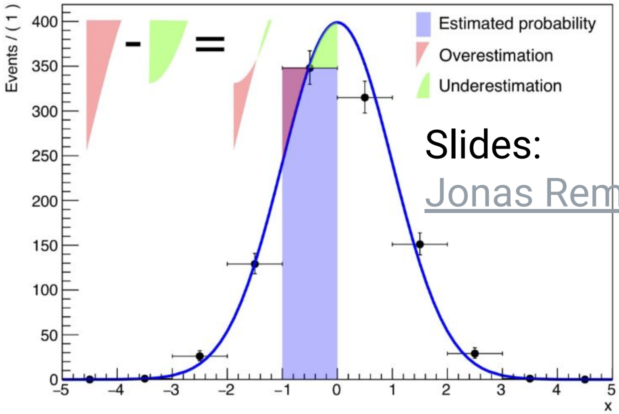
+ Less time consuming minimization. (100 events per bin ?)
+ Numerically more stable.

- May be biased based on the bin size. (Integral estimations)

Typically:

Unbinned for low statistic and analytical PDFs.

Binned for high statistic and models from MC simulation.



Slides: [Jonas Remsber](#)

Illustration of bias in binned fits when not integrating PDF over bins

Parameter estimation from likelihood

Change θ to maximize $L(data|\theta)$.

θ_{max} gives the best fit (point estimate).

Parameter estimation from likelihood

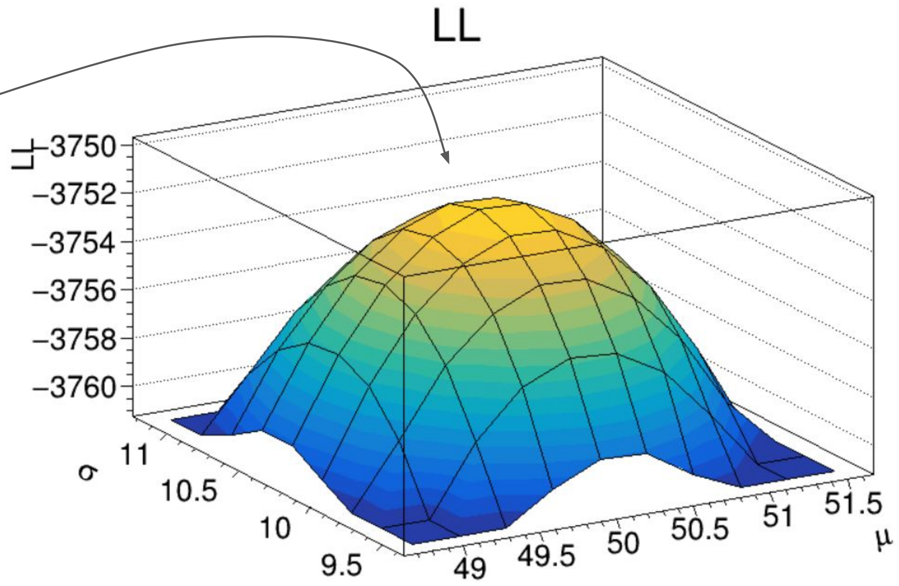
Change θ to maximize L(data|θ).

θ_{max} gives the best fit (point estimate).

How to do this?

1- For numerical stability use Log(L).

$$\prod_i \rightarrow \sum_i$$



Parameter estimation from likelihood

Change θ to maximize L(data|θ).

θ_{max} gives the best fit (point estimate).

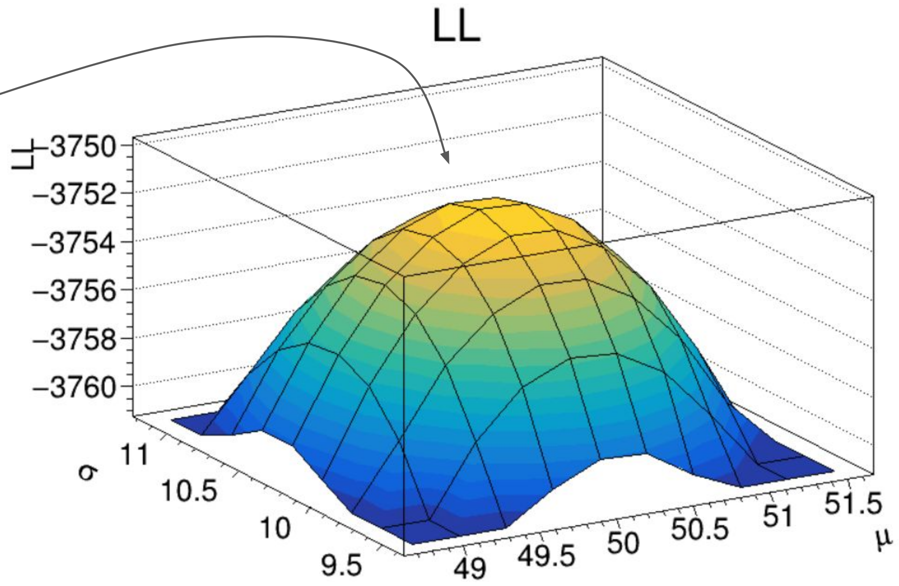
How to do this?

1- For numerical stability use Log(L).

$$\prod_i \rightarrow \sum_i$$

2- Impossible to evaluate LL on a grid.

2 parameters → 200^2 evaluations!
200 points each → (What about 100s of parameters?)



Parameter estimation from likelihood

Change θ to maximize L(data|θ).

θ_{max} gives the best fit (point estimate).

How to do this?

1- For numerical stability use Log(L).

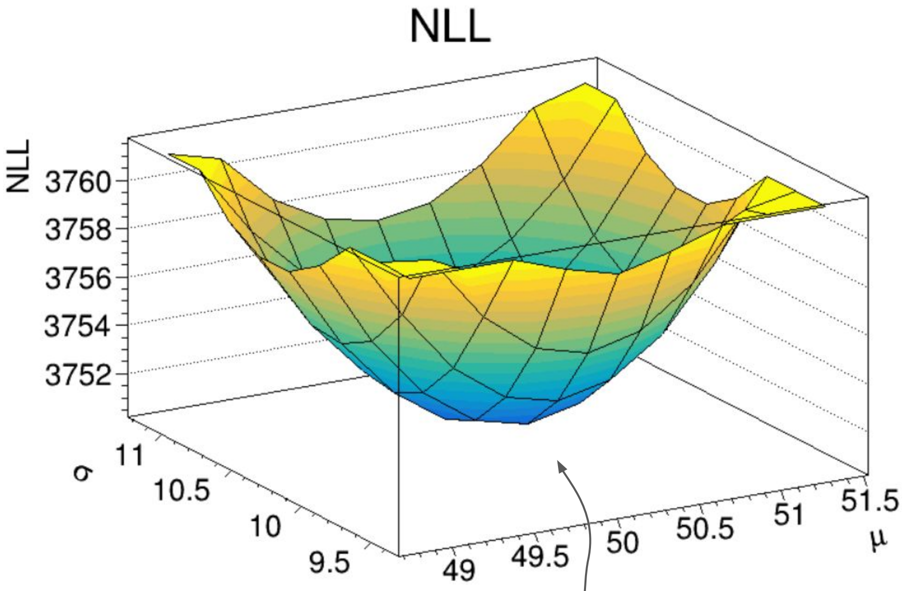
$$\prod_i \rightarrow \sum_i$$

2- Impossible to evaluate LL on a grid.

2 parameters
200 points each → 200^2 evaluations!
(What about 100s of parameters?)

3- Flip LL (negative LL) and use derivatives to go down.

ROOT minimizes the NLL (θ_{min}).



Frequentist confidence (uncertainty)

What is uncertainty on θ_{min}?

Frequentist confidence (uncertainty)

What is uncertainty on θ_{min}?

Imagine there are many replicas of the data to fit.

$$
\begin{cases}
L(\text{data}_1|\vec{\theta}) \rightarrow \vec{\theta}^{1}_{min} \\
L(\text{data}_2|\vec{\theta}) \rightarrow \vec{\theta}^{2}_{min} \\
\vdots
\end{cases}
$$

Frequentist confidence (Neyman construction)

What is uncertainty on θ_{min}?

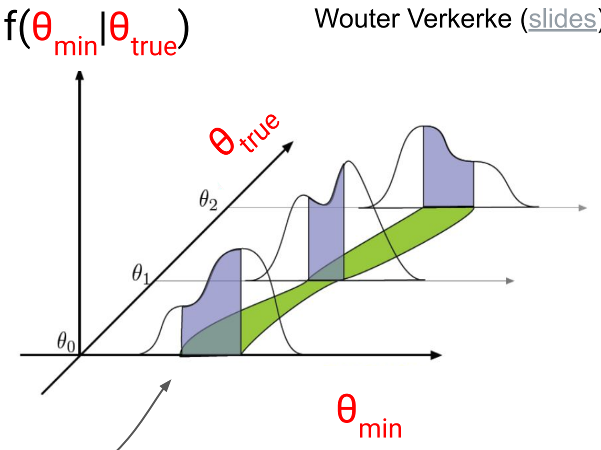
Imagine there are many replicas of the data to fit.

$$\begin{cases} L(\text{data}_1|\vec{\theta}) \rightarrow \vec{\theta}^{1}_{min} \\ L(\text{data}_2|\vec{\theta}) \rightarrow \vec{\theta}^{2}_{min} \\ \quad\quad\vdots \end{cases}$$

θ_{true} can be anything and is unknown.

How does θ_{min} distribute for a θ_{true}?

Wouter Verkerke (slides)



Frequentist confidence (Neyman construction)

What is uncertainty on θ_{min}?

Imagine there are many replicas of the data to fit.

$$\begin{cases} L(\text{data}_1|\vec{\theta}) \rightarrow \vec{\theta}^1_{min} \\ L(\text{data}_2|\vec{\theta}) \rightarrow \vec{\theta}^2_{min} \\ \vdots \end{cases}$$

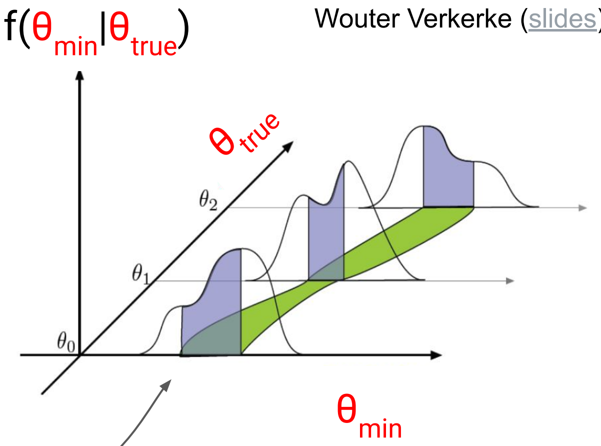
θ_{true} can be anything and is unknown.

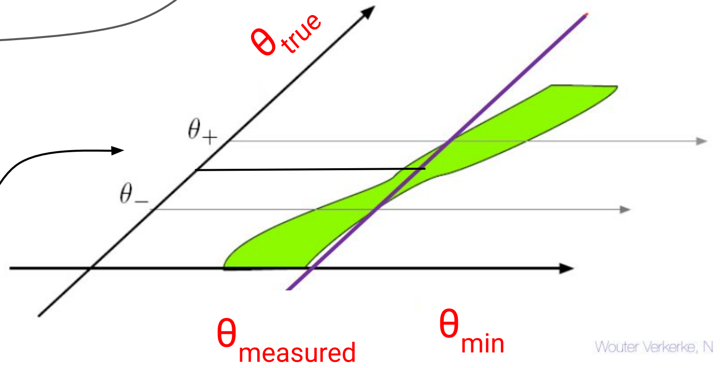
How does θ_{min} distribute for a θ_{true}?

The confidence (uncertainty) range shows how often θ_{true} is within $\theta_{measured}$.

$1\ \sigma \sim 68\%$	$2\ \sigma \sim 95\%$

Wouter Verkerke (slides)





Frequentist confidence (Neyman construction)

What

Imag ... t.

θ_{true} c

How

The c

how c

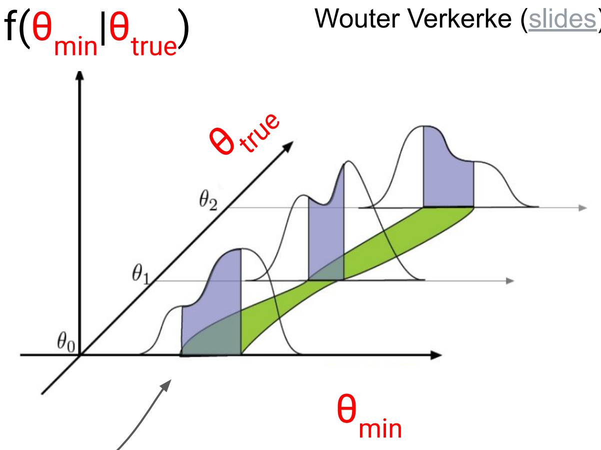
Neyman construction is intuitive but has issues.

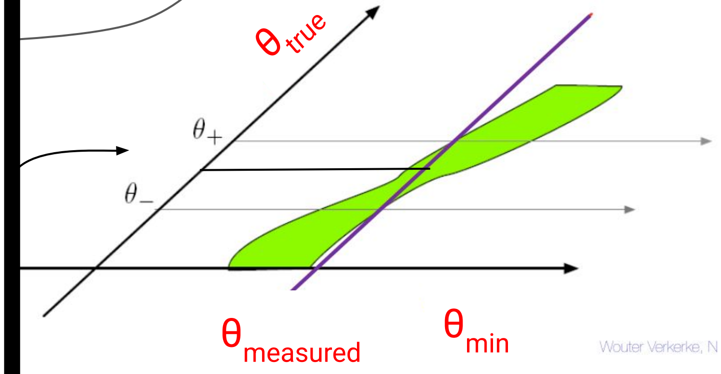
For instance finding $f(\theta_{min}|\theta_{true})$.

What to do?

Use likelihood ratio test

Wouter Verkerke (slides)



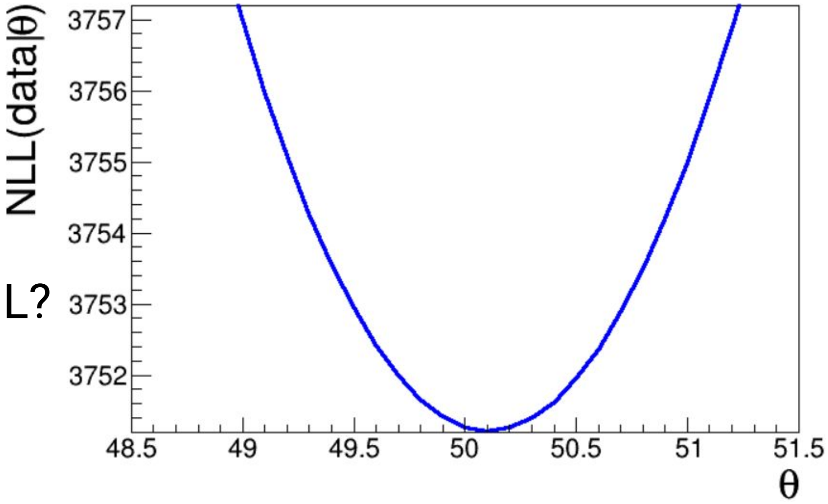


Confidence interval from likelihood ratio

Value of NLL(data|θ):

- θ_{min} gives the minimum.
- As θ deviates NLL increases.

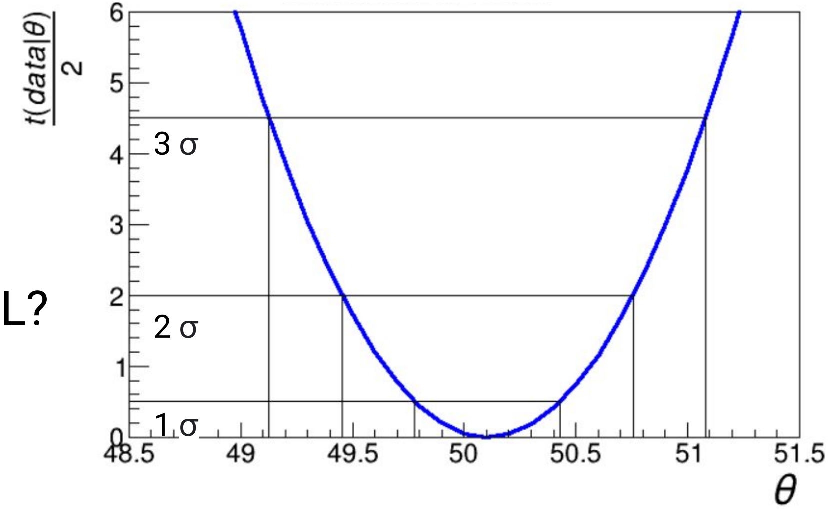
How much θ changes for a significant change in L?



Likelihood ratio test to measure the confidence interval

Value of NLL(data|θ):

- θ_{min} gives the minimum.
- As θ deviates NLL increases.



How much θ changes for a significant change in L?

Decide with the Likelihood Ratio (LR) test:

$$t(data|\theta) = -2\log \frac{L(data|\theta)}{L(data|\hat{\theta})}\bigg|_{\hat{\theta}:\text{Best fit}} = \left(\frac{\theta-\hat{\theta}}{\sigma}\right)^2\bigg|_{data\to\infty}$$

Papers sometime show Hessian errors. (An approximation!)

Likelihood ratio test to measure the confidence interval

Value

- θ
- A

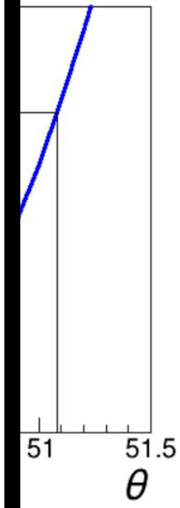
How r

Decid

$t(data$

ssian

o!)

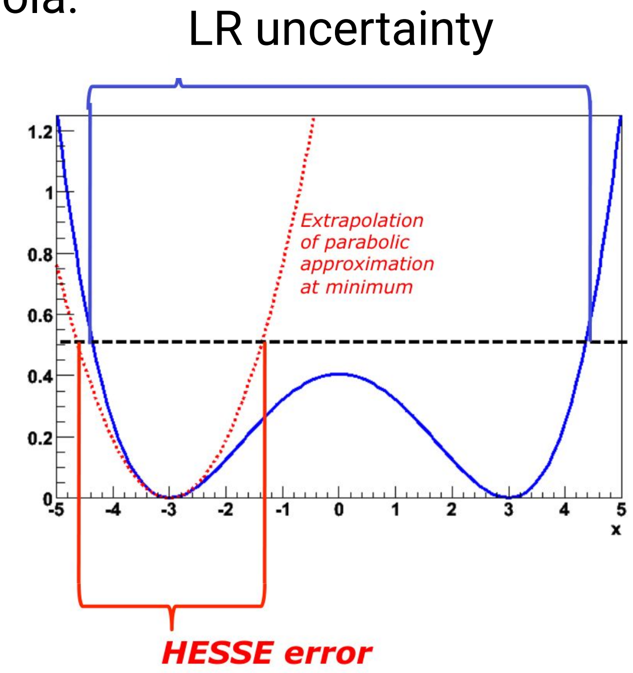


At lower statistic LR may not give a parabola.

$$-2\log \frac{L(data|\theta)}{L(data|\hat{\theta})}\Big|_{\hat{\theta}:\text{Best fit}} \neq \left(\frac{\theta-\hat{\theta}}{\sigma}\right)^2\Big|_{data\to\infty}$$

Still the threshold of 1 for t(data|θ) gives a good 68% coverage.

Why?



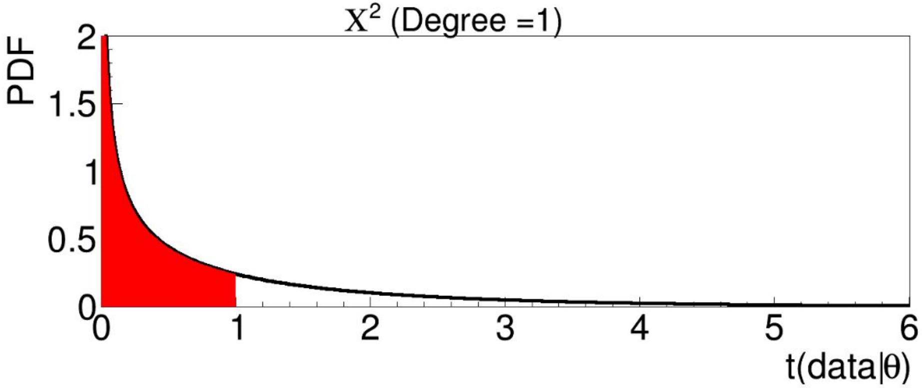
Distribution of $t(data|\theta)$

What is the distribution of $t(data|\theta)$?

Distribution of t(data|θ)

What is the distribution of t(data|θ)?

t(~~data|~~θ) follows a chi^2 distribution, independent of θ!



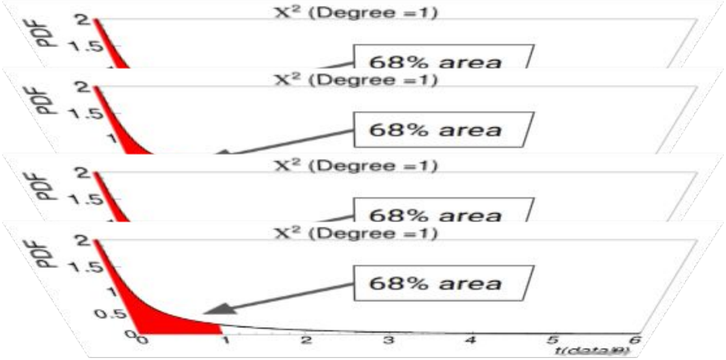
Distribution of t(data|θ)

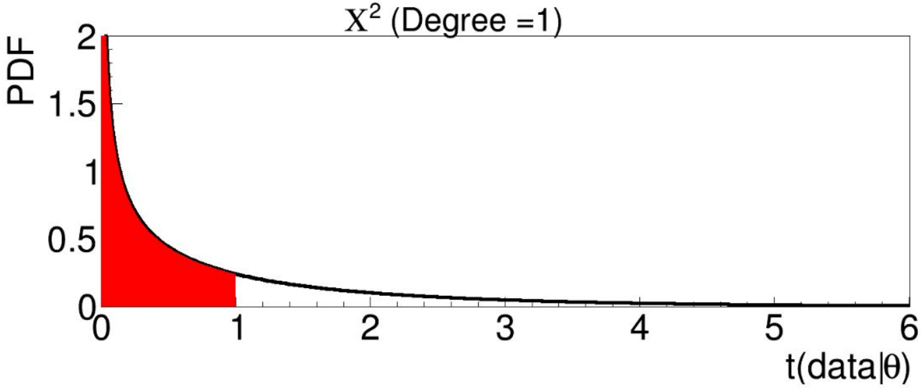
What is the distribution of t(data|θ)?

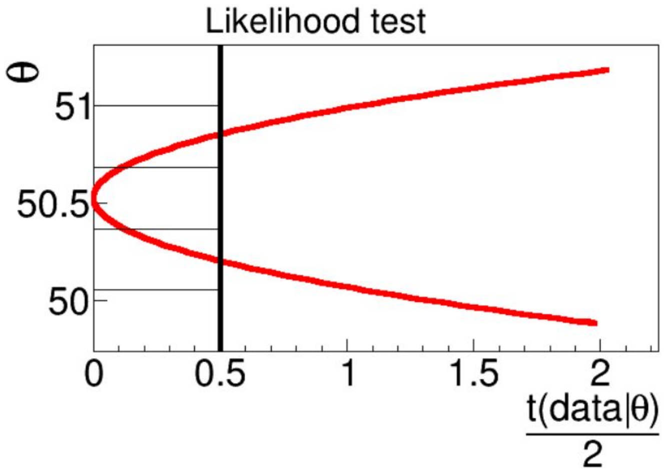
t(~~data|~~θ) follows a chi^2 distribution, independent of θ!

You find a value for t at each θ.

A value above 1 is an unlikely outcome.





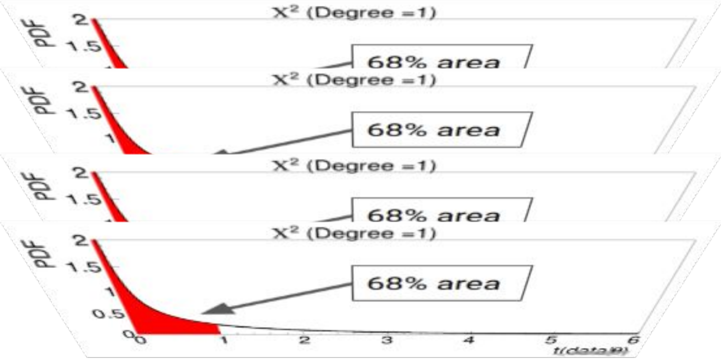


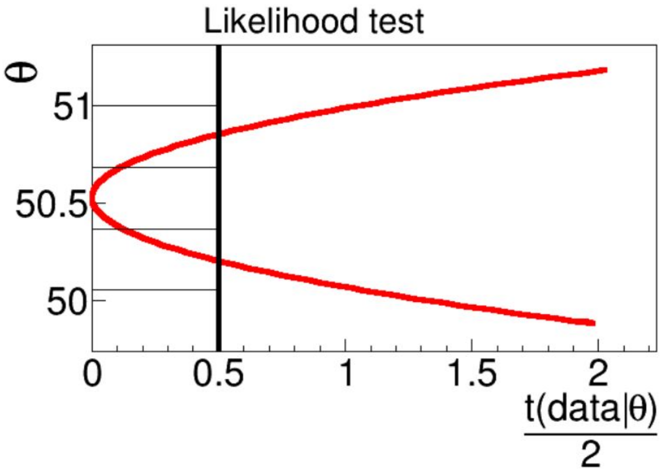
Distribution of t(data|θ)

It is always good to do more validation, especially with small data.

More on this later!

What if the likelihood has multiple parameters?





Nuisance parameters

Definition: parameters that will impact our results/change our PDF but are not the parameter of interest we care about fitting.

Typically they are measured or computed separately and have some uncertainty associated with them.

Examples: resolution, efficiency, calibration constants, isotope activity…

Two ways of including these in your likelihood, depending on your school of thought.

Notation in next few slides assumes μ is the POI and θ the nuisance parameter/s.

Nuisance parameters

Definition: parameters that will impact our results/change our PDF but are not the parameter of intere

Typically they are n uncertainty associated with the

Examples: resoluti ty…

Two ways of incluc chool of thought.

Notation in next fe e parameter/s.

We want to measure the parameter of interest. (fraction of DM events in data)

We also have NPs that control the shape of the fit model.

How to propagate the uncertainty of NPs into the measurement?

Profile likelihood ratio

Nuisance parameters

Frequentist:
Construct your likelihood with two data sets so we include the probability of observing data_x given μ and θ: $P(\text{data}_x|\mu,\theta)$ and the probability of observing our other "nuisance" data_y (e.g., calibration data set) given θ $P(\text{data}_y|\theta)$

Total likelihood is then: $\mathcal{L}(\mu,\theta) = P(\text{data}_x|\mu,\theta)P(\text{data}_y|\theta)$

Bayesian:
Adjust your priors given the information you have on θ (nominally some mean value θ_0 and uncertainty $\Delta\theta$) given your previous measurements.
Typically takes Poisson or Gaussian form:

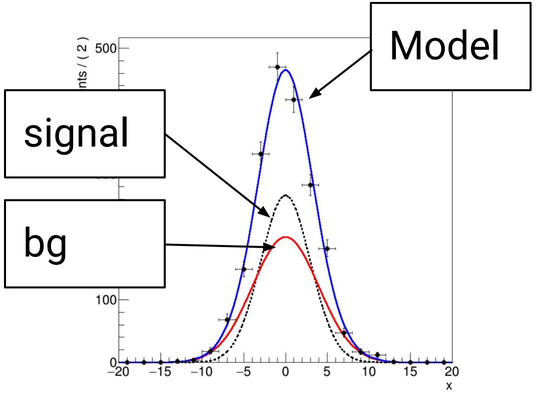
$$\text{pdf}(\mu,\theta) = P(\text{data}_x|\mu,\theta)\text{Gauss}(\theta|\theta_0,\Delta\theta)$$

From likelihood to profile likelihood (PL)

Model: Gaussian-1 (signal) + Gaussian-2 (bg).

$\mu_{Interest}$: Ratio of sig/bg θ_{NP}: Background width.

Signal width is constant



We want to measure $\mu_{Interest}$.

From likelihood to profile likelihood (PL)

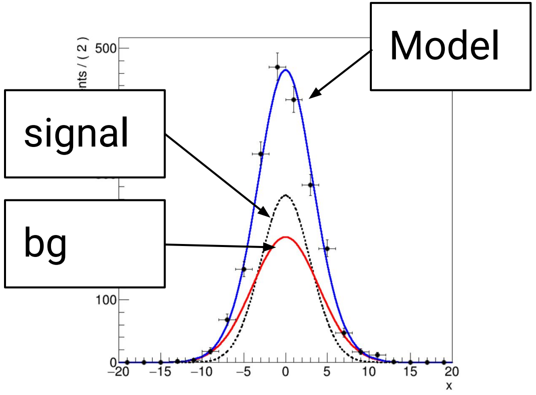
Model: Gaussian-1 (signal) + Gaussian-2 (bg).

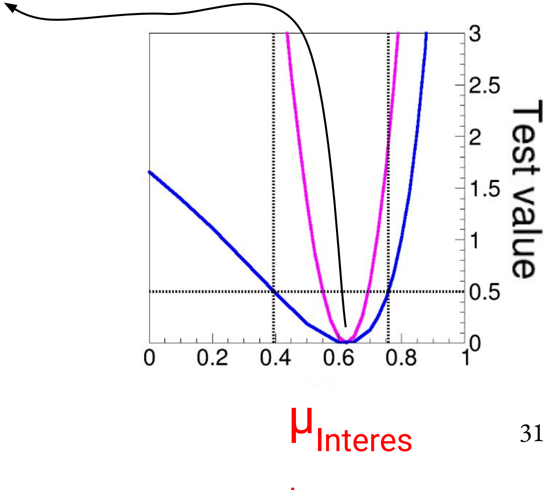
$\mu_{Interest}$: Ratio of sig/bg θ_{NP}: Background width.

Signal width is constant

We want to measure $\mu_{Interest}$.

LR quantifies how fast the model becomes inconsistent with data.





From likelihood to profile likelihood (PL)

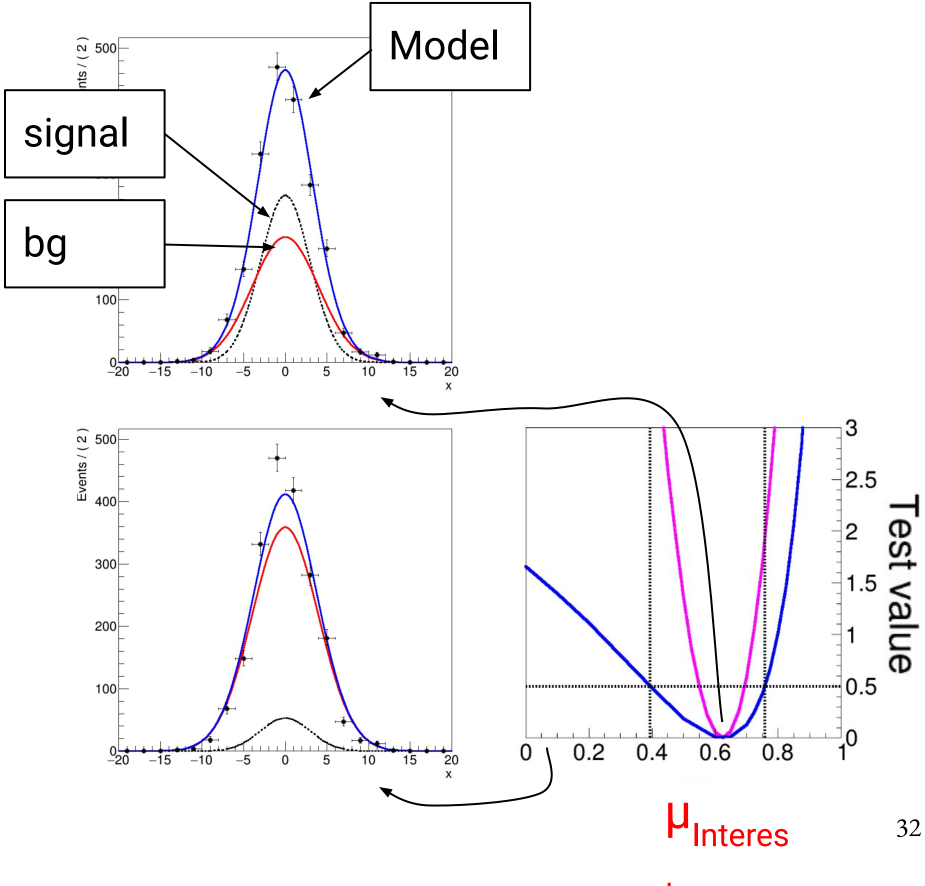
Model: Gaussian-1 (signal) + Gaussian-2 (bg).

μ_{Interest}: Ratio of sig/bg θ_{NP}: Background width.

Signal width is constant

We want to measure μ_{Interest}.

LR quantifies how fast the model becomes inconsistent with data.

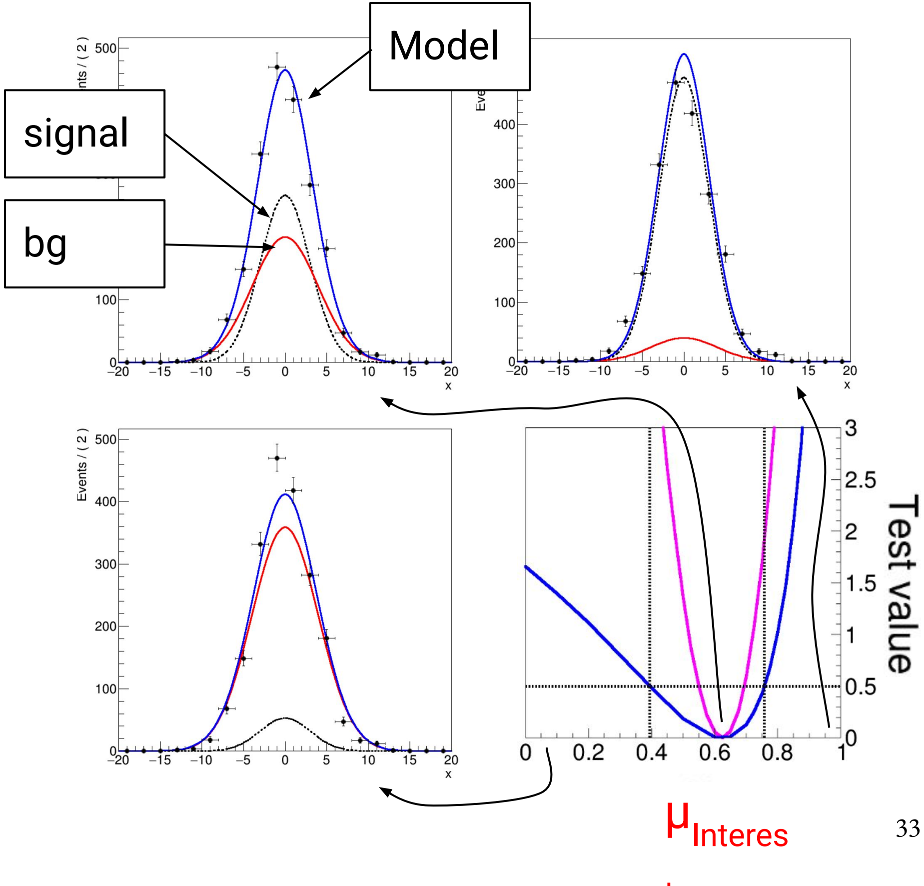


From likelihood to profile likelihood (PL)

Model: Gaussian-1 (signal) + Gaussian-2 (bg).

μ_{Interest}: Ratio of sig/bg $\quad\theta_{\mathrm{NP}}$: Background width.

Signal width is constant



We want to measure μ_{Interest}.

LR quantifies how fast the model becomes inconsistent with data.

θ_{NP} can compensate for the change in μ_{Interest}.

For small signal (μ_{Interest}), make bg (θ_{NP}) pointier to keep the model consistent (smaller LR).

From likelihood to profile likelihood (PL)

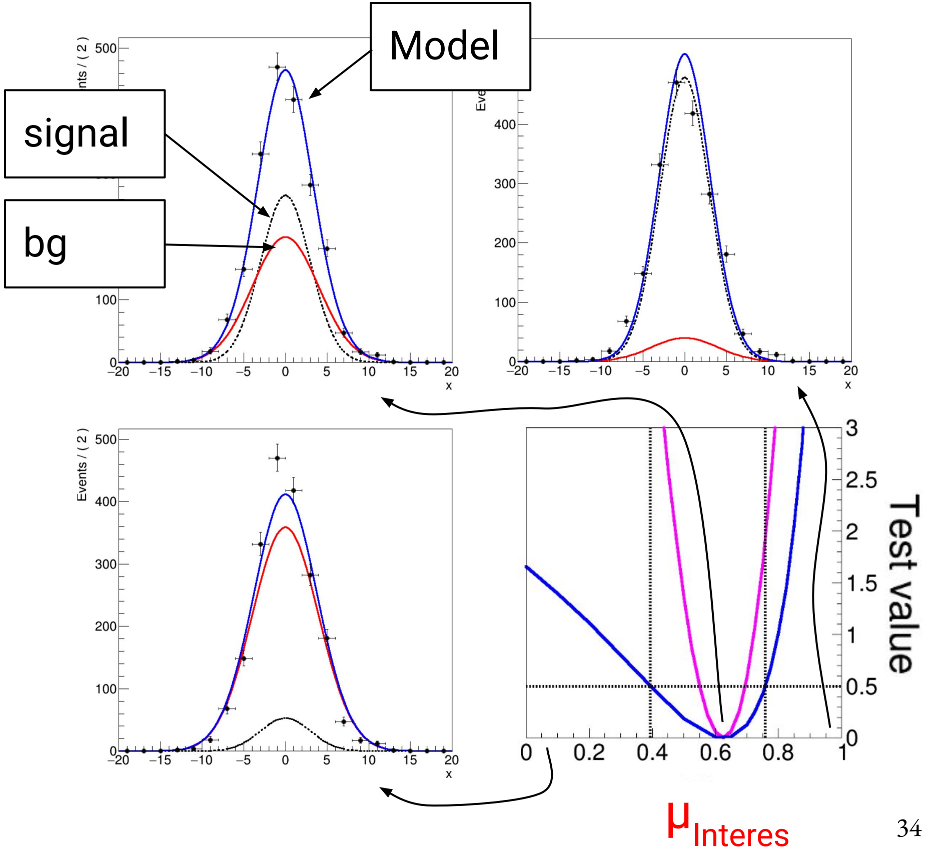
Model: Gaussian-1 (signal) + Gaussian-2 (bg).

This is the effect of NPs.

NPs have uncertainties. Let them float as a function of $\mu_{Interest}$ to keep the model consistent with data.

This is called Profile likelihood.

PL grows slower, uncertainty becomes wider.



From likelihood to profile likelihood (PL)

PL propagates the uncertainty of NPs.

Best fit of θ for given μ

$$Log\left(\frac{L(\mu)}{L(\hat{\mu})}\right) \rightarrow Log\left(\frac{L(\mu, \boxed{\hat{\hat{\theta}}(\mu)})}{L\boxed{(\hat{\mu}, \hat{\theta})}}\right)$$

Global best fit values

From likelihood to profile likelihood (PL)

PL propagates the uncertainty of NPs.

Best fit of θ for given μ

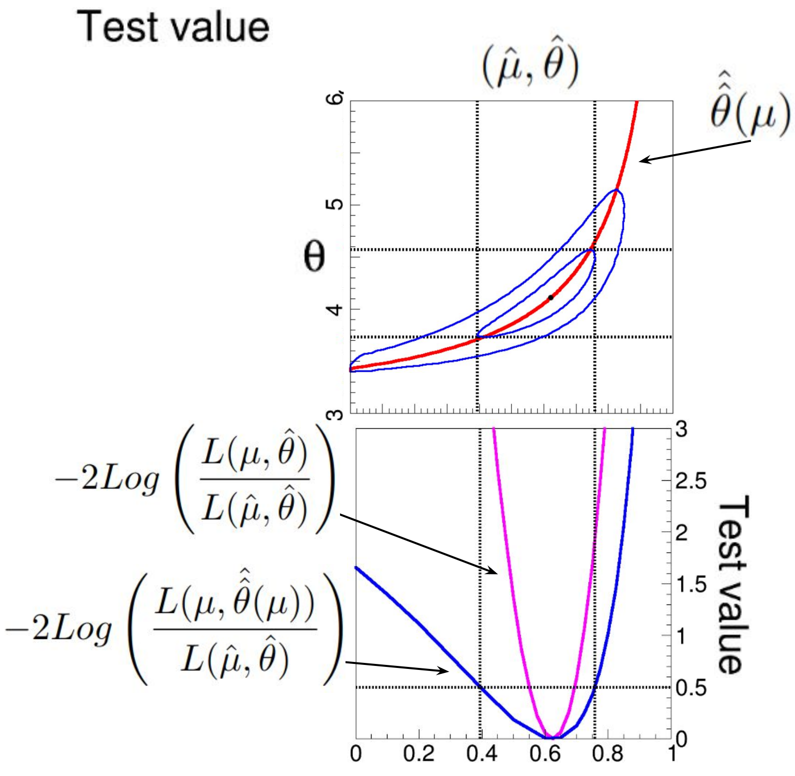
$$Log\left(\frac{L(\mu)}{L(\hat{\mu})}\right) \rightarrow Log\left(\frac{L(\mu,\boxed{\hat{\hat{\theta}}(\mu)})}{\boxed{L(\hat{\mu},\hat{\theta})}}\right)$$

Global best fit values

At each μ_{Interest} repeat the fit with floating θ_{NP}.

Changes in θ_{NP} compensates for the change in μ_{Interest}.

Test reaches 0.5 slower -> Wider uncertainty.



μ_{Interes}

From likelihood to profile likelihood (PL)

PL propagates the uncertainty of NPs.

Best fit of θ for given μ

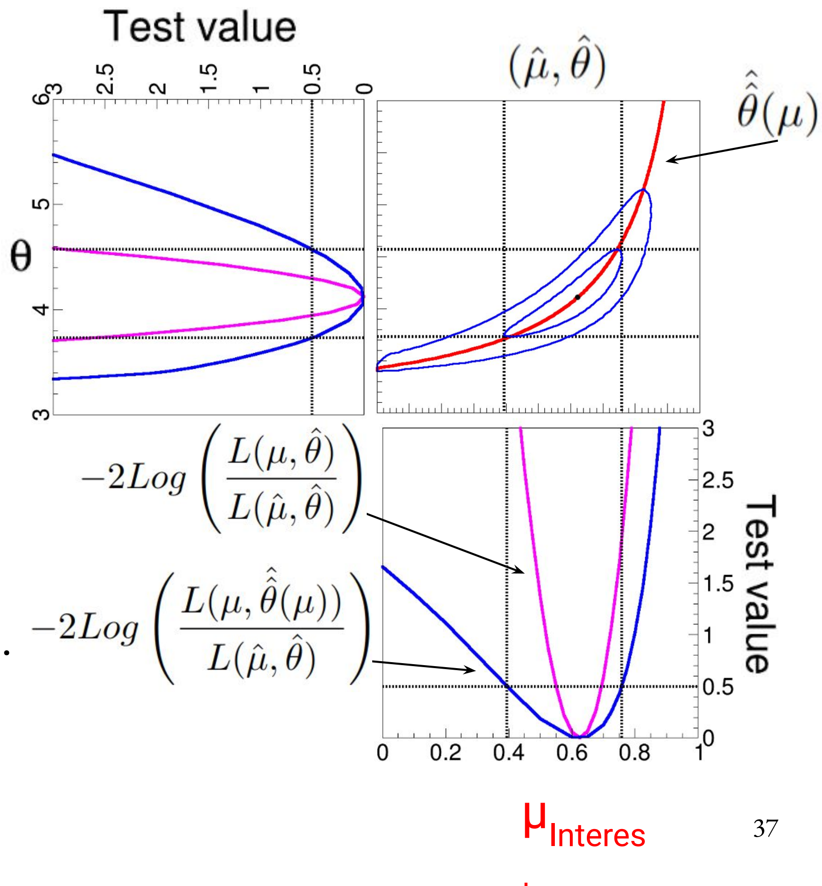
$$Log\left(\frac{L(\mu)}{L(\hat{\mu})}\right) \rightarrow Log\left(\frac{L(\mu, \boxed{\hat{\hat{\theta}}(\mu)})}{L\boxed{(\hat{\mu}, \hat{\theta})}}\right)$$

Global best fit values

At each $\mu_{Interest}$ repeat the fit with floating θ_{NP}.

Changes in θ_{NP} compensates for the change in $\mu_{Interest}$.

Test reaches 0.5 slower -> Wider uncertainty.



From likelihood to profile likelihood (PL)

PL propagates the un

Best fit of θ

$$Log\left(\frac{L(\mu)}{L(\hat{\mu})}\right) \to L$$

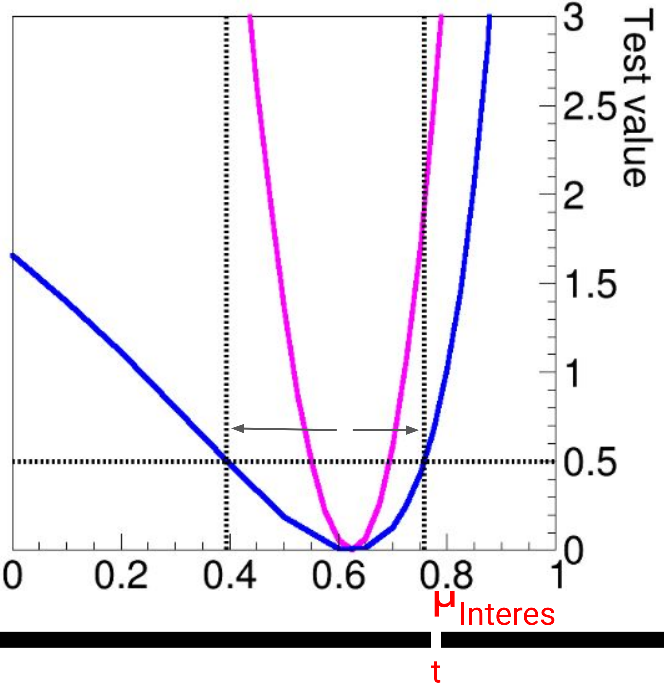
Global best

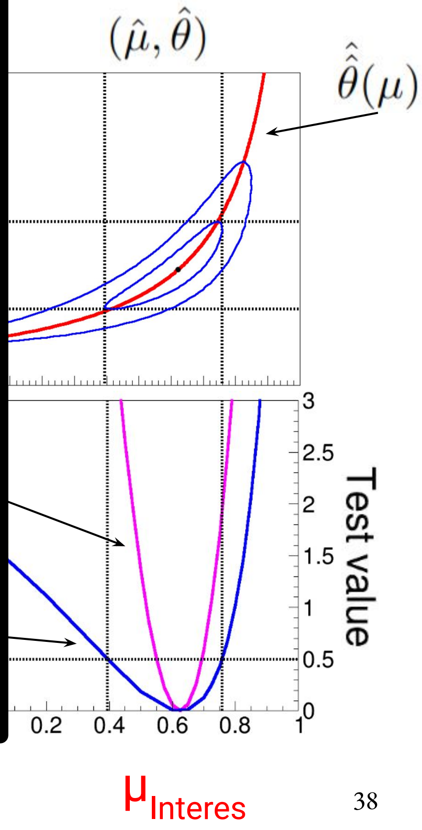
At each $\mu_{Interest}$ repeat th

Changes in θ_{NP} compen

Test reaches 0.5 slowe

Why is PL uncertainty asymmetric?



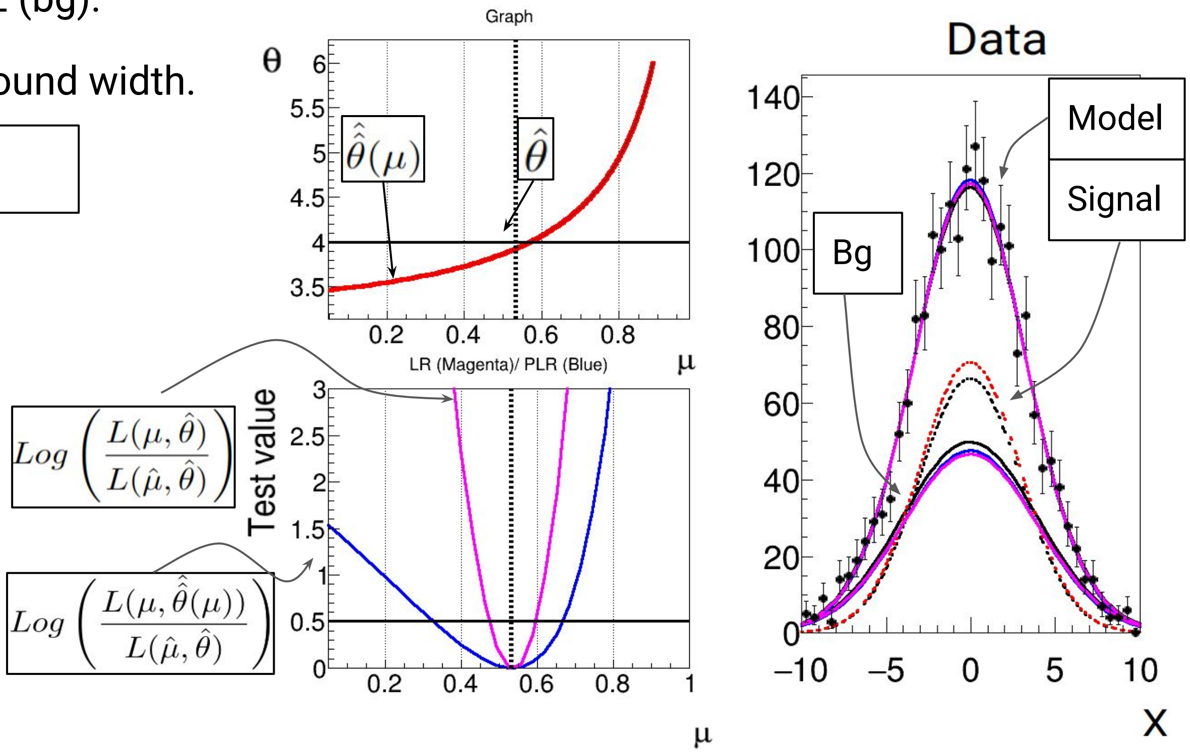


Profile likelihood is not voodoo magic

Model: Gaussian-1 (signal) + Gaussian-2 (bg).

μ_{Interest}: Ratio of sig/bg θ_{NP}: Background width.

signal width is constant



Profile likelihood is not voodoo magic

Model: Gaussian-1 (signal) + Gaussian-2 (bg).

$\mu_{Interest}$: Ratio of sig/bg θ_{NP}: Background width.

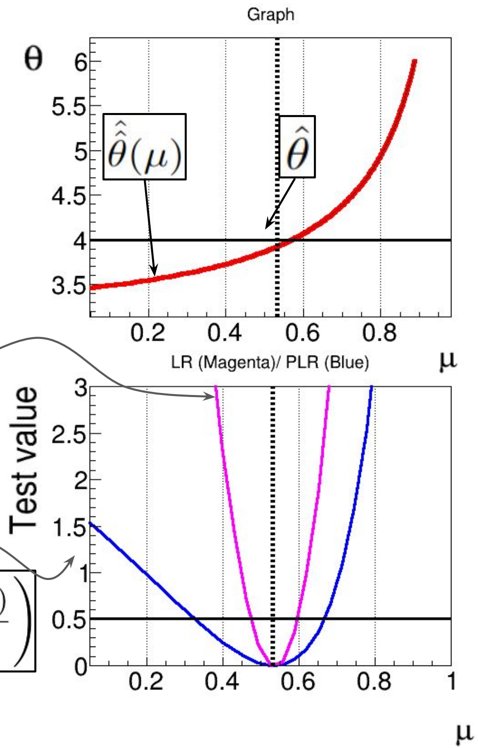
signal width is constant

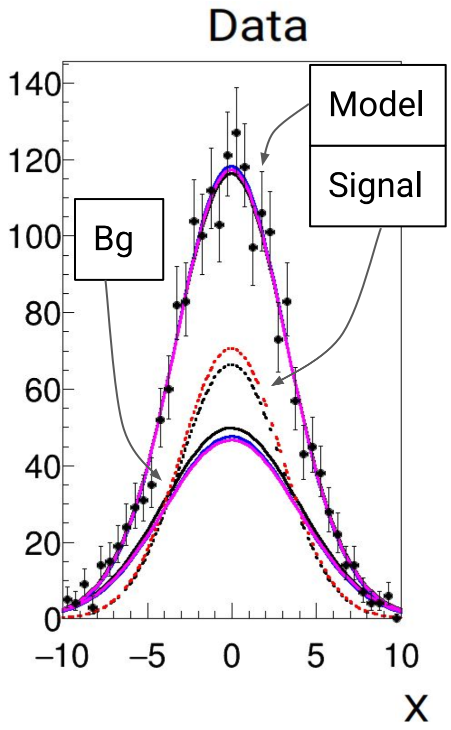
Low μ:

- Weak signal.
- bg becomes narrower to model the bulk.
- It compensates the signal
- PLR grows slower than LR.

$$Log\left(\frac{L(\mu,\hat{\theta})}{L(\hat{\mu},\hat{\theta})}\right)$$

$$Log\left(\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})}\right)$$





Profile likelihood is not voodoo magic

Model: Gaussian-1 (signal) + Gaussian-2 (bg).

$\mu_{Interest}$: Ratio of sig/bg θ_{NP}: Background width.

signal width is constant

Low μ:

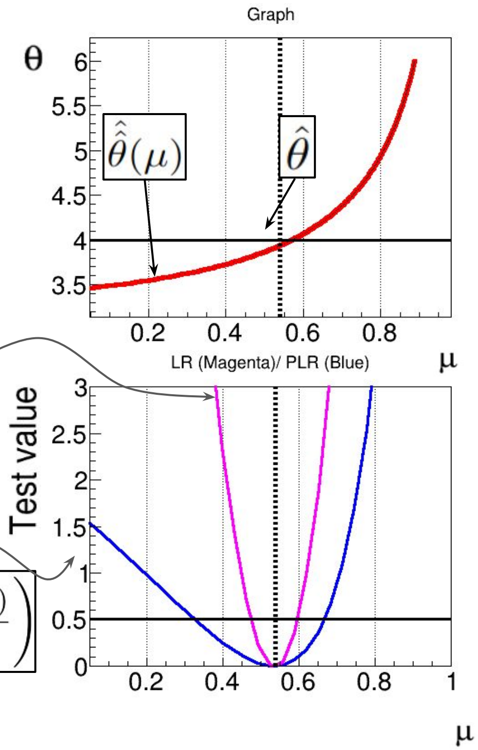
- Weak signal.
- bg becomes narrower to model the bulk.
- It compensates the signal
- PLR grows slower than LR.

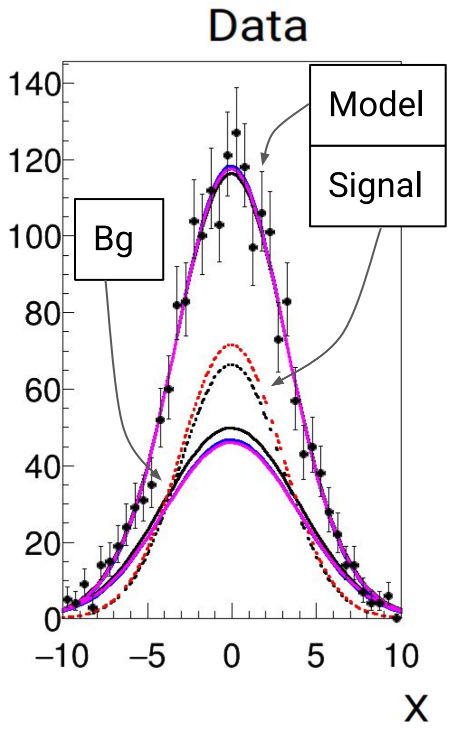
$$Log\left(\frac{L(\mu,\hat{\theta})}{L(\hat{\mu},\hat{\theta})}\right)$$

High μ:

- bg is depleted.
- Tail cannot be modeled anymore.
- Model becomes inconsistent faster.
- PLR and LR increase parallel

$$Log\left(\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})}\right)$$





Profile likelihood is not voodoo magic

Model: Gaussian-1 (signal) + Gaussian-2 (bg)

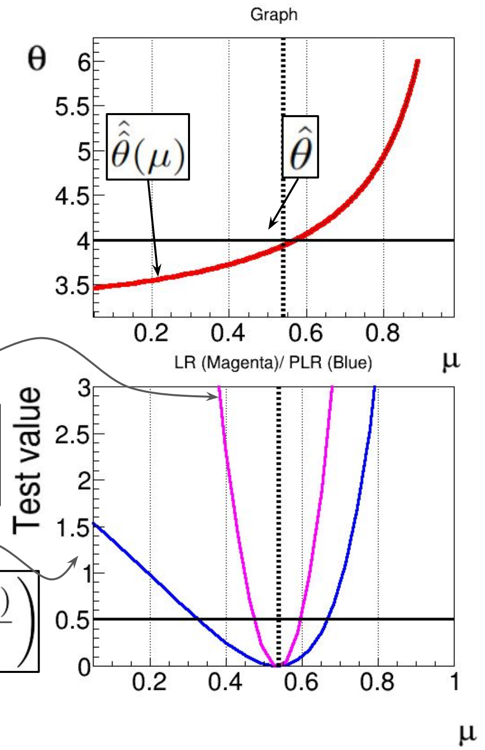
μ

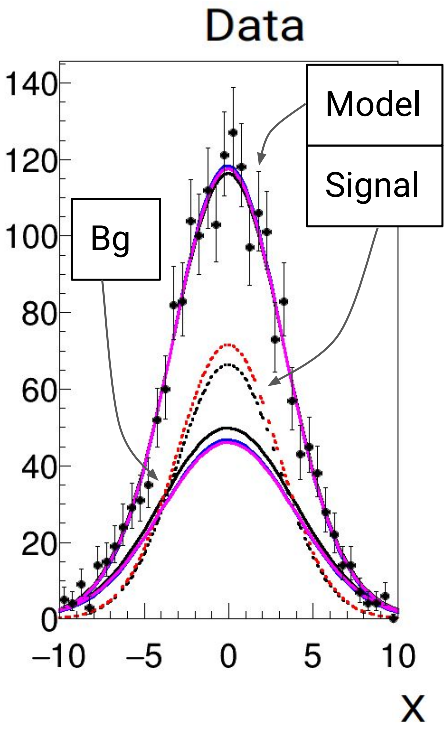
Low

High

Bottom line:

1. NPs increase the uncertainty if there is correlation.
2. Try to understand strong correlations and the PLR response.
3. Use PLR to propagate uncertainties.

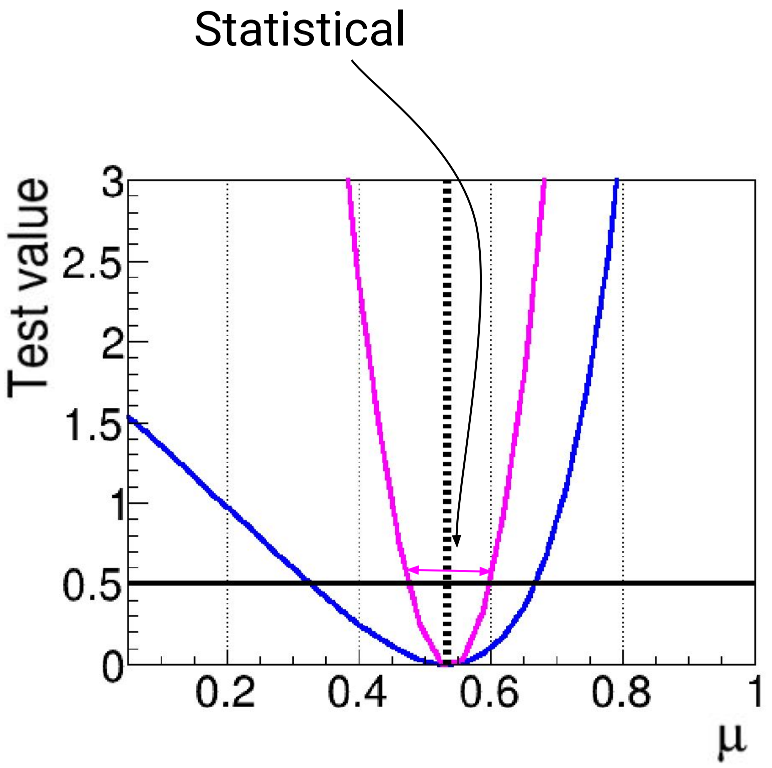




Measuring statistical and systematic uncertainties

For fits with many floating parameters (5, 100, 1000, …):

1. Fix all parameter to the best fit. $(\hat{\mu}, \hat{\theta})$
2. Do LR on μ to find the Statistical uncertainty.



Measuring statistical and systematic uncertainties

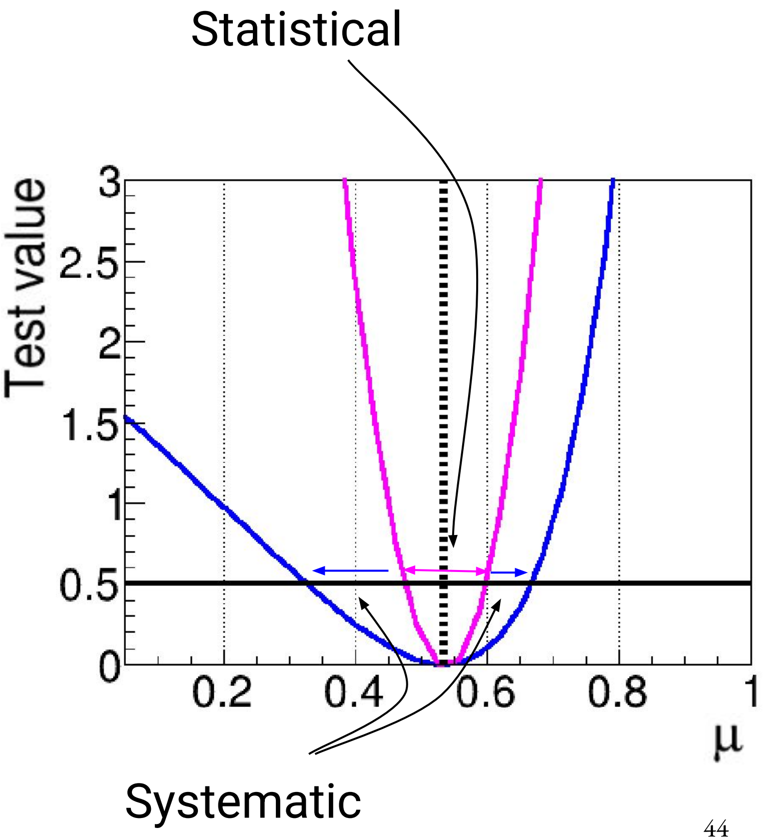
For fits with many floating parameters (5, 100, 1000, …):

1. Fix all parameter to the best fit. $(\hat{\mu}, \hat{\theta})$
2. Do LR on μ to find the Statistical uncertainty.
3. Do PLR with one NP at a time for systematics.

Statistical uncertainty is dominant?

Deep breath, your life is easy …

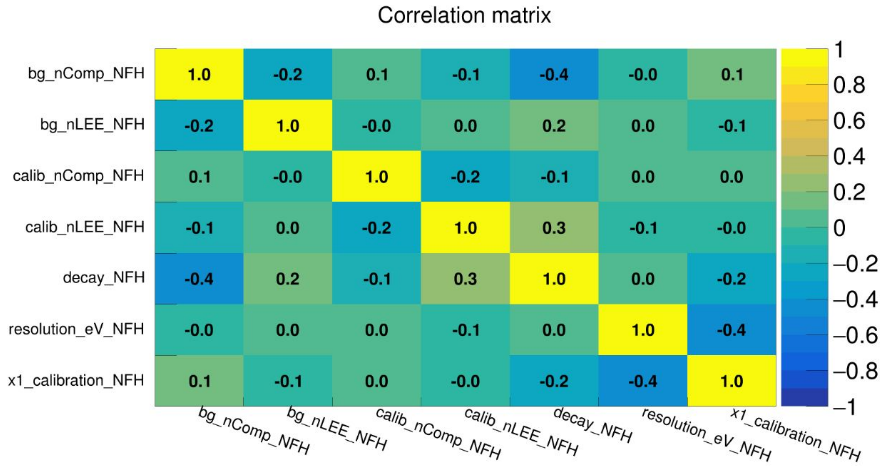
Assuming you considered all systematics ;)

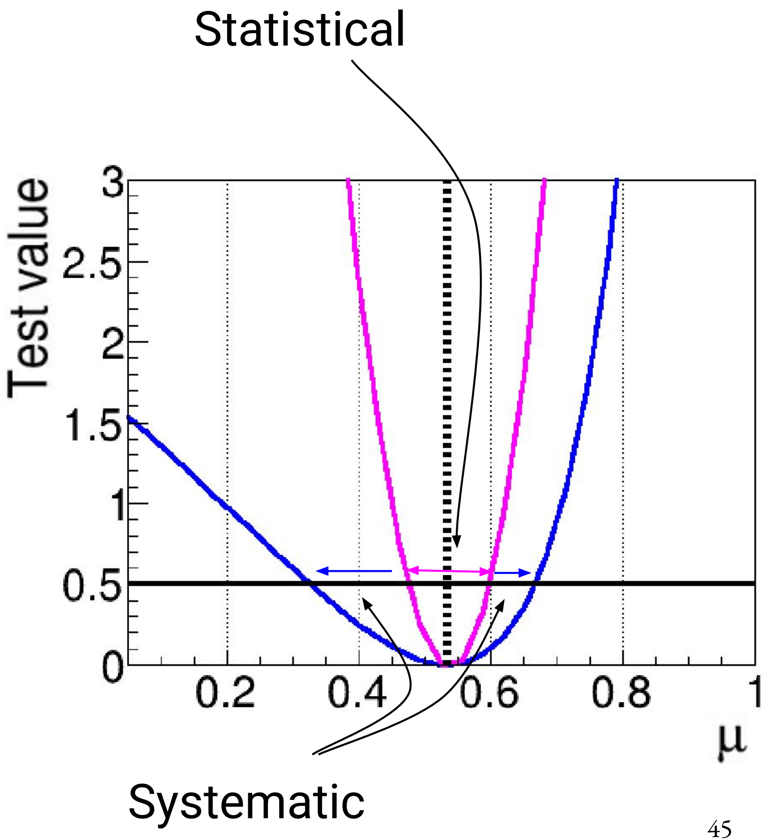


Measuring statistical and systematic uncertainties

For fits with many floating parameters (5, 100, 1000, …):

1. Fix all parameter to the best fit. $(\hat{\mu}, \hat{\theta})$
2. Do LR on μ to find the Statistical uncertainty.
3. Do PLR with one NP at a time for systematics.
4. Prioritize to understand strong correlations.
5. Check ranking plots. (More on this later)





Measuring statistical and systematic uncertainties

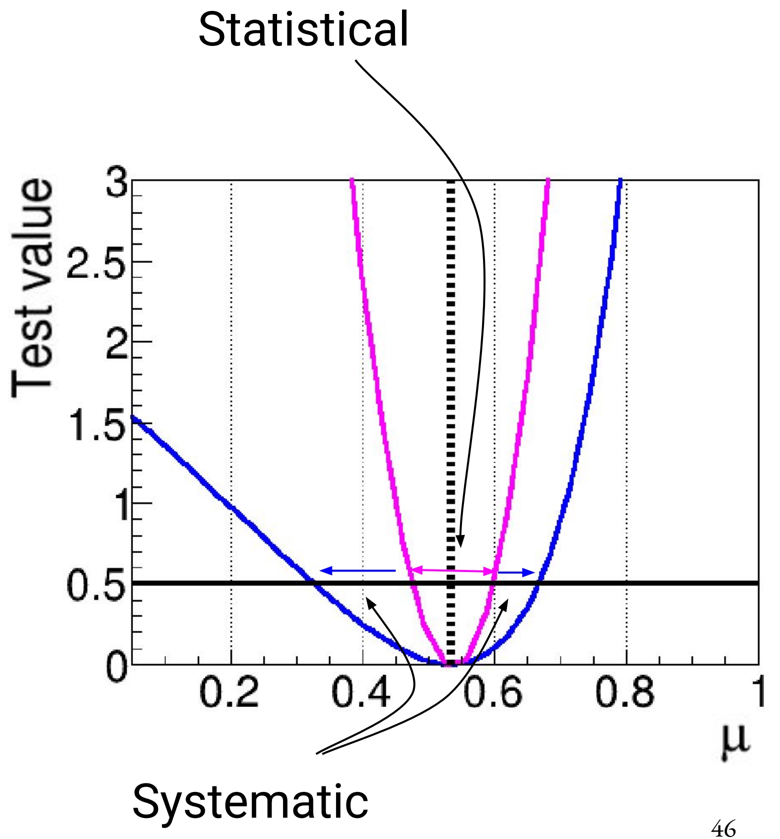
For fits with many floating parameters (5, 100, 1000, ...):

1. Fix
2. Do
3. Do
4. Pri
5. Ch

Up to this point we learned about:

1. Likelihood function.
2. Point estimate (best fit).
3. Frequentist confidence.
4. PLR and treatment of NPs.
5. Stat and Sys uncertainties.

How does ROOT do these things?



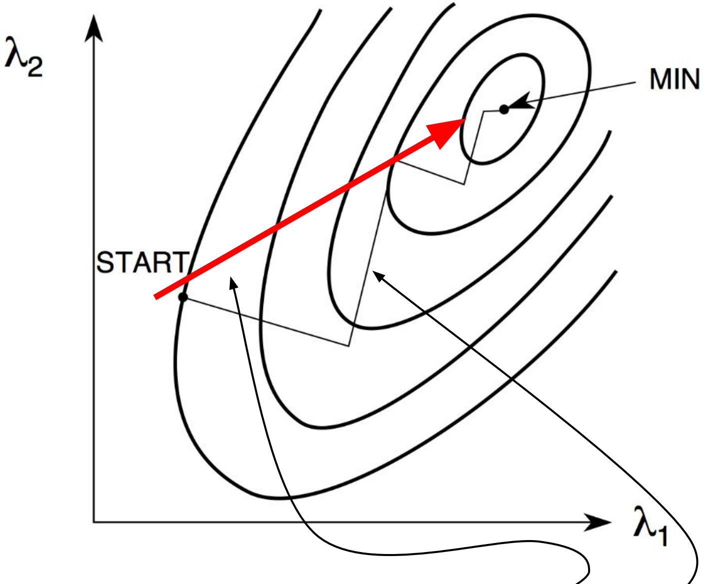
PLR diagnosis - (1) Understanding the minimizer (MINUIT)

MINUIT is the minimization package behind RooFit

It checks the gradient, and its change to find the direction to min.

$$\mathrm{f}(\vec{\theta_0} + \delta\vec{\theta}) = f(\vec{\theta}) + \Delta f^T \theta + \frac{1}{2}\delta\vec{\theta}^T H \delta\vec{\theta} + \cancel{\cdots}$$

$$\nabla \mathrm{f}(\vec{\theta_0} + \delta\vec{\theta}) = \nabla \mathrm{f}(\vec{\theta_0}) + H\delta\vec{\theta} + \cancel{\cdots}$$



Red: MINUIT
Black: Gradient descent

PLR diagnosis - (1) Understanding the minimizer (MINUIT)

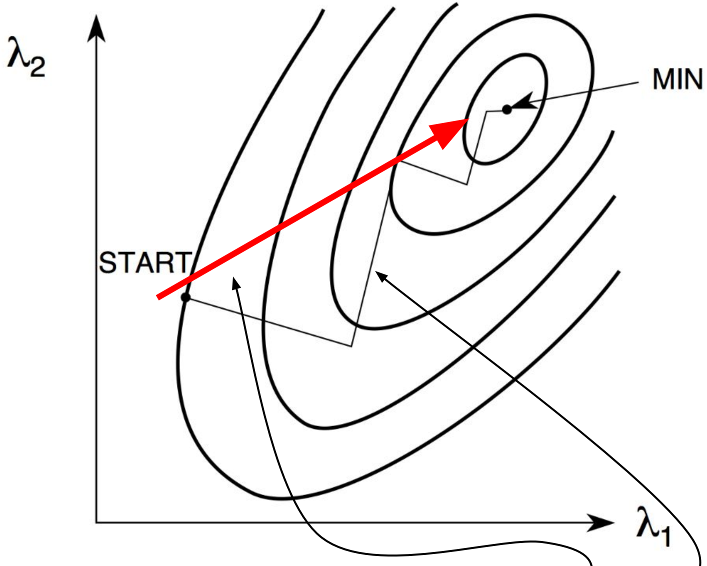
MINUIT is the minimization package behind RooFit

It checks the gradient, and its change to find the direction to min.

$$\mathrm{f}(\vec{\theta_0} + \delta\vec{\theta}) = f(\vec{\theta}) + \Delta f^T \theta + \frac{1}{2}\delta\vec{\theta}^T \boxed{H} \delta\vec{\theta} + \cancel{\cdots}$$

$$\nabla \mathrm{f}(\vec{\theta_0} + \delta\vec{\theta}) = \nabla \mathrm{f}(\vec{\theta_0}) + \boxed{H} \delta\vec{\theta} + \cancel{\cdots}$$

- Intuition: Take a big step if gradient is stationary.
- Benefit: Fast convergence.
- Tradeoff: Requires the Hessian (second derivative matrix).



Red: MINUIT
Black: Gradient descent

PLR diagnosis - (1) Understanding the minimizer (MINUIT)

MINUIT is the minimization package behind RooFit

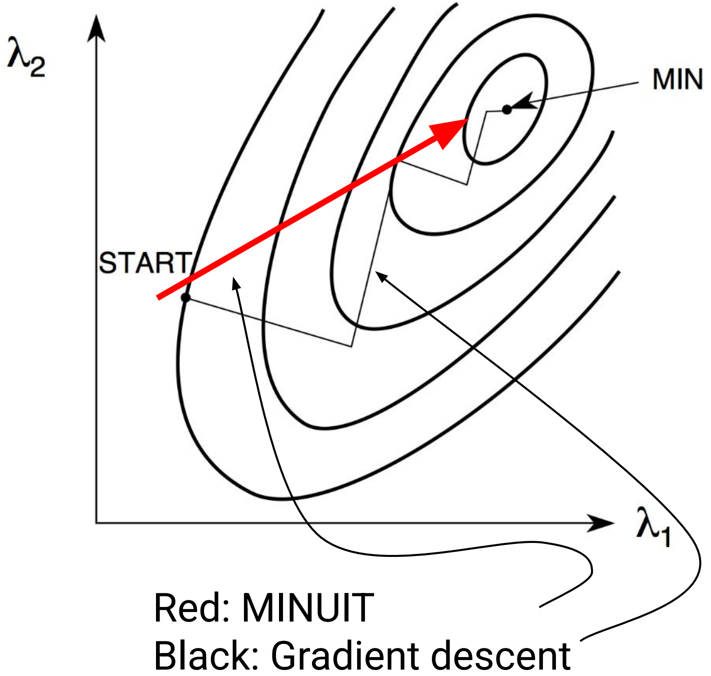
It checks t nin.

When does a fit converge with RooFit?

What diagnostic RooFit returns?

- Intui
- Bene
- Trad x).

MINUIT al



Check " $\mathrm{EDM} = \frac{1}{2} \cdot \nabla f^T H^{-1} \nabla f < 0.001$" per step.

PLR diagnosis - (2) Convergence and output (MINUIT)

1- Initial minimization to EDM< 0.001 using approximate calculation of the Hessian (H).

2- Depending on the “strategy code”:

- 0 -> claim convergence.
- 2 -> Find exact H, and continue the minimization if EDM>0.001
- 1 (default) -> If approximate and exact H are close terminate. Continue with strategy 2 otherwise.

PLR diagnosis - (2) Convergence and output (MINUIT)

1- Initial minimization to EDM< 0.001 using approximate calculation of the Hessian (H).

2- Depending on the “strategy code”:

- 0 -> claim convergence.
- 2 -> Find exact H, and continue the minimization if EDM>0.001
- 1 (default) -> If approximate and exact H are close terminate. Continue with strategy 2 otherwise.

$$\mathrm{EDM} = \frac{1}{2} \cdot \nabla f^T H^{-1} \nabla f < 0.001$$

Stop after this

```
Info in <Minuit2>: VariableMetricBuilder Start iterating until Edm is < 0.001 with call limit = 1500
Info in <Minuit2>: VariableMetricBuilder    0 - FCN =       -2118.332081 Edm =        33.51993972 NCalls =      61
Info in <Minuit2>: VariableMetricBuilder Start iterating...
Info in <Minuit2>: VariableMetricBuilder Initial State:
  Parameter:     [   0.007160101288     -0.161471224    -0.9272951793]
  Gradient:      [     -119.5636239     -549.1843092  1.189489143e-05]
  InvHessian:
[[   2.8564027e-05               0               0]
 [              0   0.00022092437               0]
 [              0               0   0.00011111111]]]
  Edm: 33.5199
```

NLL value

Current EDM

Ncalls used

More info: [TMinuit reference](#) - [MINUIT manual](#)

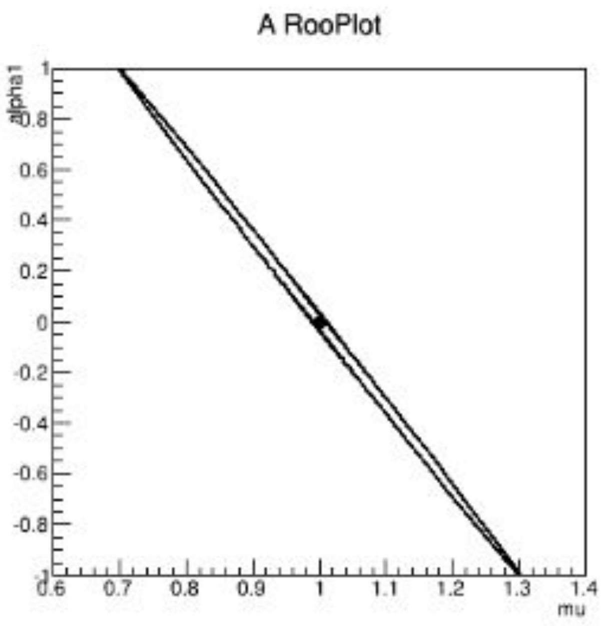
PLR diagnosis - (3) what may go wrong

Extreme correlations

- Hessian and inverse Hessian calculation problems.
- Global minimum becomes like a valley instead of a point.
- Not well-defined minimum or uncertainty.
- HESSE fails when ratio of weakest-to-strongest eigenvalue < 10-6

Output:

```
Warning in <Minuit2>: VariableMetricBuilder Matrix not pos.def, gdel = 0.506583 > 0
Warning in <Minuit2>: MnPosDef non-positive diagonal element in covariance matrix[ 9 ] = -0.0231748
Warning in <Minuit2>: MnPosDef non-positive diagonal element in covariance matrix[ 10 ] = -0.270261
Warning in <Minuit2>: MnPosDef non-positive diagonal element in covariance matrix[ 12 ] = -21.0648
Warning in <Minuit2>: MnPosDef non-positive diagonal element in covariance matrix[ 16 ] = -0.0247647
Warning in <Minuit2>: MnPosDef Added to diagonal of Error matrix a value 21.5648
Warning in <Minuit2>: MnPosDef Matrix forced pos-def by adding to diagonal 1.38155
```



$\rho=-0.9995$

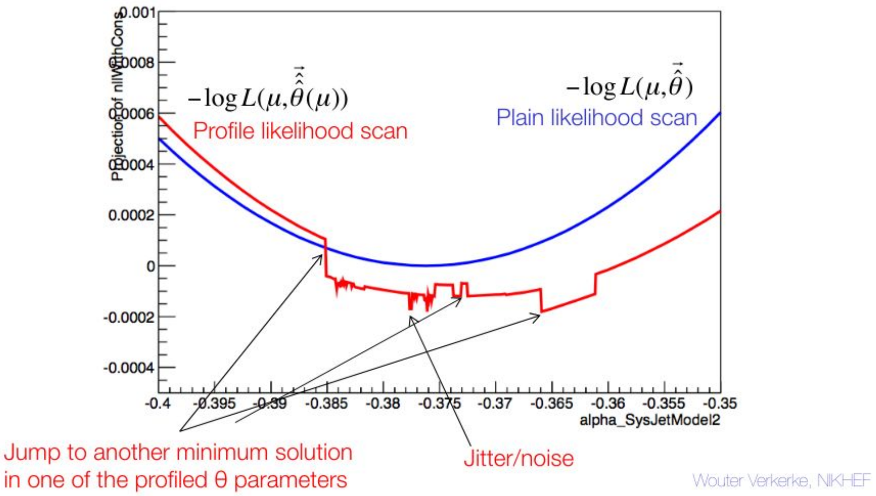
Solution:

- Reparameterize or simplify the model.

PLR diagnosis - (3) what may go wrong

Numerical instabilities

- Numerical integral.
- Numerical smearing.
- Very high statistic and complex model.
- Multiple minimums close the global minimum.



Solution:

- Increasing the numerical integral precession.
- Define Analytical integrals for PDF.
- Case by case debugging.
- Also talk to people with more experience.

Back ups

Useful references

- Asymptotic formulae for likelihood-based tests of new physics (Cowan, Cranmer, Gross, Vitells) https://arxiv.org/pdf/1007.1727.pdf
- PDG stats chapter
- Advances statistics (Verkerke) https://www.precision.hep.phy.cam.ac.uk/wp-content/people/mitov/lectures/GraduateLectures/Advanced-Statistics-Verkerke.pdf
- Dealing with uncertainty/errors: https://www.nikhef.nl/~ivov/Talks/2013_03_21_DESY_PoissonError.pdf
- RooFit tutorials - available in c++ and python (in regular files and notebook format): https://root.cern/doc/master/group__tutorial__roofit.html
- RooStats tutorials - available in c++ and python (in regular files and notebook format): https://root.cern/doc/master/group__tutorial__roostats.html
- Diagnostics guide from ATLAS: https://statisticalmethods.web.cern.ch/StatisticalMethods/recommendations/rec_diagnostics_checks/#investigating-simple-fits
- Introductory ROOT/RooFit/RooStats tutorials: https://root.cern/get_started/courses/