Statistics and Likelihood

Markus Gaug
Universitat Autònoma de Barcelona and IEEC-CERES

markus.gaug@uab.cat

Questions in gamma-ray astronomy

- Is a source significantly detected?

Questions in gamma-ray astronomy

- Is a source significantly detected?
- If so, what is its flux ?
- If not, what is its upper limit ?

Questions in gamma-ray astronomy

- Is a source significantly detected?
- If so, what is its flux ?
- If not, what is its upper limit ?
- Is the source variable, periodic ?

Questions in gamma-ray astronomy

- Is a source significantly detected?
- If so, what is its flux ?
- If not, what is its upper limit ?
- Is the source variable, periodic ?

- What kind of spectrum does it have?
- What is its spectral index ?

Questions in gamma-ray astronomy

- Is a source significantly detected?
- If so, what is its flux ?
- If not, what is its upper limit ?
- Is the source variable, periodic ?

- What kind of spectrum does it have?
- What is its spectral index ?
- What is its location in the sky ?

Questions in gamma-ray astronomy

- Is a source significantly detected?
- If so, what is its flux ?
- If not, what is its upper limit ?
- Is the source variable, periodic ?

- What kind of spectrum does it have?
- What is its spectral index ?
- What is its location in the sky ?

- What are the uncertainties on these variables ?

Questions in gamma-ray astronomy

- Is a source significantly detected? **hypothesis testing**
- If so, what is its flux ? **parameter estimation**
- If not, what is its upper limit ? **parameter estimation**
- Is the source variable, periodic ? **hypothesis testing**

- What kind of spectrum does it have? **hypothesis testing**
- What is its spectral index ? **parameter estimation**
- What is its location in the sky ? **parameter estimation**

- What are the uncertainties on these variables ?

hypothesis testing / **parameter estimation**

Maximum Likelihood technique

1. Set of observed data
2. produce a model (that accurately describes the data)
3. derive the probability (density) for the data, given the model
4. treat all this as a function of the model parameters
5. maximize the likelihood with respect to these parameters

Maximum Likelihood technique

$$\boldsymbol{X} = \{x_i\} = \{x_1, x_2, \ldots, x_N\}$$

Set of observed data

$$P(\boldsymbol{x} \mid \boldsymbol{\Theta})$$

produce a model (that accurately describes the data)

derive the probability (density) for the data, given the model

$$\boldsymbol{\Theta} = \{\theta_j\} = \{\theta_1, \theta_2, \ldots, \theta_M\}$$

treat all this as a function of the model parameters

maximize the likelihood with respect to these parameters

Maximum Likelihood technique

$$\boldsymbol{X} = \{x_i\} = \{x_1, x_2, \ldots, x_N\}$$

For independent data:

$$P(x_i, x_j) = P(x_i) \cdot P(x_j \mid x_i) = P(x_i) \cdot P(x_j)$$

Set of observed data

produce a model (that accurately describes the data)

$$P(\boldsymbol{x} \mid \boldsymbol{\Theta}) = P(x_1 \mid \boldsymbol{\Theta}) \cdot P(x_2 \mid \boldsymbol{\Theta}) \cdots P(x_N \mid \boldsymbol{\Theta})$$

derive the probability (density) for the data, given the model

$$\boldsymbol{\Theta} = \{\theta_j\} = \{\theta_1, \theta_2, \ldots, \theta_M\}$$

treat all this as a function of the model parameters

maximize the likelihood with respect to these parameters

Maximum Likelihood technique

$$\boldsymbol{X} = \{x_i\} = \{x_1, x_2, \ldots, x_N\}$$

For independent data:

$$P(x_i, x_j) = P(x_i) \cdot P(x_j \mid x_i) = P(x_i) \cdot P(x_j)$$

Set of observed data

produce a model (that accurately describes the data)

$$P(\boldsymbol{x} \mid \boldsymbol{\Theta}) = P(x_1 \mid \boldsymbol{\Theta}) \cdot P(x_2 \mid \boldsymbol{\Theta}) \cdots P(x_N \mid \boldsymbol{\Theta})$$

derive the probability (density) for the data, given the model

$$\boldsymbol{\Theta} = \{\theta_j\} = \{\theta_1, \theta_2, \ldots, \theta_M\}$$

treat all this as a function of the model parameters

maximize the likelihood with respect to these parameters

$$\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\Theta}) = \prod_{i=1}^{N} P(x_i \mid \boldsymbol{\Theta})$$

Maximum Likelihood technique

$$\boldsymbol{X} = \{x_i\} = \{x_1, x_2, \ldots, x_N\}$$

For independent data:

$$P(x_i, x_j) = P(x_i) \cdot P(x_j \mid x_i) = P(x_i) \cdot P(x_j)$$

Set of observed data

produce a model (that accurately describes the data)

$$P(\boldsymbol{x} \mid \boldsymbol{\Theta}) = P(x_1 \mid \boldsymbol{\Theta}) \cdot P(x_2 \mid \boldsymbol{\Theta}) \cdots P(x_N \mid \boldsymbol{\Theta})$$

derive the probability (density) for the data, given the model

$$\boldsymbol{\Theta} = \{\theta_j\} = \{\theta_1, \theta_2, \ldots, \theta_M\}$$

treat all this as a function of the model parameters

maximize the likelihood with respect to these parameters

easier to work with logarithm:

$$\ln\left(\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\Theta})\right) = \sum_{i=1}^{N} \ln\left(P(x_i \mid \boldsymbol{\Theta})\right)$$

Maximum Likelihood Estimation (MLE)

- Estimates of $\hat{\Theta} = \{\hat{\theta}_j\}$ can be obtained by simultaneously solving:

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta_j}\right|_{\{\hat{\theta}_k\}} = 0$$

Maximum Likelihood Estimation (MLE)

- Estimates of $\hat{\Theta} = \{\hat{\theta}_j\}$ can be obtained by simultaneously solving:

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta_j}\right|_{\{\hat{\theta}_k\}} = 0$$

- MLE has the following asymptotic properties (under certain *regularity* conditions) :
  - Consistency: $\lim_{n\to\infty}\left(\hat{\Theta}\right) = \Theta_0$
  - Asymptotic normality: $\hat{\Theta} \sim \mathcal{N}\left(\Theta_0, \left\{I\left(\Theta_0\right)\right\}^{-1}\right)$
  - Asymptotic efficiency: MLE achieves the smallest possible uncertainty (the so-called *Cramér Raó lower bound*)
  - Invariance: The MLE estimator of $f\left(\Theta_0\right)$ is $f(\hat{\Theta})$

Fisher information matrix

$$I\left(\hat{\Theta}\right) = \left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j}\right|_{\Theta=\hat{\Theta}}$$

Maximum Likelihood Estimation (MLE)

- 2^{nd} derivative of $\ln \mathcal{L}$ is related to the uncertainty of the estimate:

one-parameter case: $\ln \mathcal{L} \sim \exp\left(-\frac{\left(\Theta - \hat{\Theta}\right)^2}{2\sigma^2}\right)$ $\qquad \left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\hat{\Theta}} = -\frac{1}{\sigma^2}$

Example 1:

Independent measurements of flux of source with Gaussian uncertainties:

Model: constant flux *F* → $P(\boldsymbol{x}_i \mid F) = \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{\left(x_i - F\right)^2}{2\sigma_i^2} \right)$

$$\ln \mathcal{L} = -\sum \frac{\left(x_i - F\right)^2}{2\sigma_i^2} - \sum \ln \sigma_i - \frac{N}{2} \ln 2\pi$$

Example 1:

Independent measurements of flux of source with Gaussian uncertainties:

Model: constant flux *F* → $$P(\boldsymbol{x}_i \mid F) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\left(x_i - F\right)^2}{2\sigma_i^2} \right)$$

$$\ln \mathcal{L} = -\sum \frac{\left(x_i - F\right)^2}{2\sigma_i^2} - \sum \ln \sigma_i - \frac{N}{2} \ln 2\pi$$

Maximize MLE w.r.t. *F*: $$\frac{\partial \ln \mathcal{L}}{\partial F} = -\sum \frac{x_i - F}{\sigma_i^2} = 0 \qquad \rightarrow \hat{F} = \frac{\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$$

Example 1:

Independent measurements of flux of source with Gaussian uncertainties:

Model: constant flux *F* → $P(\boldsymbol{x}_i \mid F) = \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{(x_i - F)^2}{2\sigma_i^2} \right)$

$$\ln \mathcal{L} = -\sum \frac{(x_i - F)^2}{2\sigma_i^2} - \sum \ln \sigma_i - \frac{N}{2} \ln 2\pi$$

Maximize MLE w.r.t. *F*: $\frac{\partial \ln \mathcal{L}}{\partial F} = -\sum \frac{x_i - F}{\sigma_i^2} = 0 \quad \rightarrow \hat{F} = \frac{\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$

Estimate uncertainty of *F*: $\frac{1}{\sigma_F^2} = -\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\bigg|_{\hat{\Theta}} = \sum \frac{1}{\sigma_i^2} \quad \rightarrow \sigma_F = \frac{1}{\sqrt{\sum \frac{1}{\sigma_i^2}}}$

Example 2:

Counting experiment (e.g. gamma-rays): Detector detected *n* events

Model: Possonian process with mean of λ: → $P(n|\lambda) = \frac{e^{-\lambda} \cdot \lambda^n}{n!}$

$$\ln \mathcal{L} = -n \ln \lambda - \lambda - \ln n!$$

Example 2:

Counting experiment (e.g. gamma-rays): Detector detected *n* events

Model: Possonian process with mean of λ: → $P(n|\lambda) = \frac{e^{-\lambda} \cdot \lambda^n}{n!}$

$$\ln \mathcal{L} = -n \ln \lambda - \lambda - \ln n!$$

Maximize MLE w.r.t. *λ*: $\frac{\partial \ln \mathcal{L}}{\partial F} = \frac{n}{\lambda} - 1 = 0 \qquad \rightarrow \hat{\lambda} = n$

Example 2:

Counting experiment (e.g. gamma-rays): Detector detected *n* events

Model: Possonian process with mean of λ: → $P(n\,|\,\lambda) = \frac{e^{-\lambda} \cdot \lambda^{n}}{n!}$

$$\ln \mathcal{L} = -n \ln \lambda - \lambda - \ln n!$$

Maximize MLE w.r.t. *λ*: $\frac{\partial \ln \mathcal{L}}{\partial \lambda} = \frac{n}{\lambda} - 1 = 0 \qquad \rightarrow \hat{\lambda} = n$

Estimate uncertainty of *λ*: $\frac{1}{\sigma_{\lambda}^{2}} = -\left.\frac{\partial^{2} \ln \mathcal{L}}{\partial \lambda^{2}}\right|_{\hat{\lambda}} = \frac{n}{\lambda^{2}} \qquad \rightarrow \sigma_{\lambda} = \sqrt{n}$

Hypothesis Testing

For a model with *N* parameters and the sample size $n \rightarrow \infty$:

$$2\left(\ln \mathcal{L}\left(\hat{\Theta}\right) - \ln \mathcal{L}\left(\Theta_0\right)\right) \sim \chi^2\left(N\right)$$

Wilk's theorem

Hypothesis Testing

For a model with *N* parameters and the sample size $n \to \infty$:

$$2\left(\ln \mathcal{L}\left(\hat{\Theta}\right) - \ln \mathcal{L}\left(\Theta_0\right)\right) \sim \chi^2\left(N\right)$$

Wilk's theorem

Caveats:

– The model must describe the data correctly !!

Hypothesis Testing

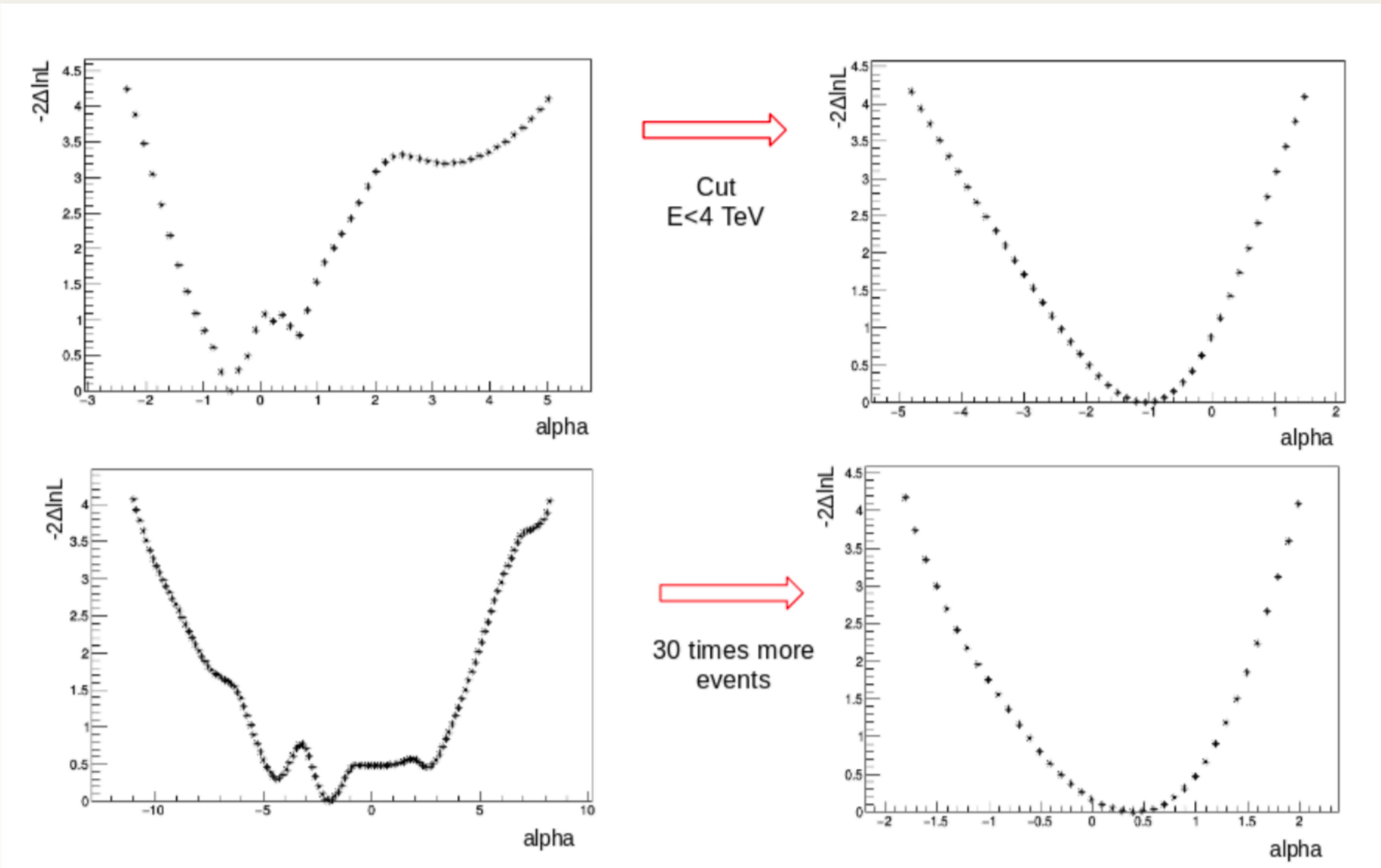
For a model with *N* parameters and the sample size $n \to \infty$:

$$TS = 2\left(\ln \mathcal{L}\left(\hat{\Theta}\right) - \ln \mathcal{L}\left(\Theta_0\right)\right) \sim \chi^2\left(N\right)$$ Wilk's theorem

Caveats:

– The model must describe the data correctly !!
– If the MLE behaves asymptotically, it is well-behaved (i.e. Wilk's theorem applies), otherwise not!
– Sometimes, n can be large, but asymptotic behaviour not yet reached because of a high weight given only to a small sub-sample of few events.

Hypothesis Testing



→ ∞ :

s theorem

from:
Leyre Nogués,
PhD thesis,
Univ. de Zaragoza,
2018

Hypothesis Testing

Normally, we do not know Θ_0 (that's why we take a measurement!)

Hypothesis Testing

Normally, we do not know Θ_0 (that's why we take a measurement!)

BUT:

We make an assumption about the model (*the null hypothesis*), in which the parameters have some *presumed "true" values*.

Hypothesis Testing

Normally, we do not know Θ_0 (that's why we take a measurement!)

BUT:

We make an assumption about the model (*the null hypothesis*), in which the parameters have some *presumed "true" values*.

Nobody can tell if the *null hypothesis is right !*

(except in MC simulated data samples)

Hypothesis Testing

Normally, we do not know Θ_0 (that's why we take a measurement!)

BUT:

We make an assumption about the model (*the null hypothesis*), in which the parameters have some *presumed "true" values*.

Compute $\ln \mathcal{L}(\Theta_0)$ for the null hypothesis (instead of the true values)

Hypothesis Testing

Normally, we do not know Θ_0 (that's why we take a measurement!)

BUT:

We make an assumption about the model (*the null hypothesis*), in which the parameters have some *presumed "true" values*.

Compute $\ln \mathcal{L}(\Theta_0)$ for the null hypothesis (instead of the true values)

Hope to show that $2\left(\ln \mathcal{L}(\hat{\Theta}) - \ln \mathcal{L}(\Theta_0)\right)$ is so large that it is improbable from χ^2

Hypothesis Testing

Normally, we do not know Θ_0 (that's why we take a measurement!)

BUT:

We make an assumption about the model (*the null hypothesis*), in which the parameters have some *presumed "true" values*.

Compute $\ln \mathcal{L}\left(\Theta_0\right)$ for the null hypothesis (instead of the true values)

Hope to show that $2\left(\ln \mathcal{L}\left(\hat{\Theta}\right) - \ln \mathcal{L}\left(\Theta_0\right)\right)$ is so large that it is improbable from χ^2 → hence *reject* the *null hypothesis*

Hypothesis Testing

Normally, we do not know Θ_0 (that's why we take a measurement!)

BUT:

We make an assumption about the model (*the null hypothesis*), in which the parameters have some *presumed "true" values*.

Compute $\ln\mathcal{L}\left(\Theta_0\right)$ for the null hypothesis (instead of the true values)

Hope to show that $TS = 2\left(\ln\mathcal{L}\left(\hat{\Theta}\right) - \ln\mathcal{L}\left(\Theta_0\right)\right)$ is so large that it is improbable from χ^2 → hence *reject* the *null hypothesis with* $\sqrt{TS}$ *signficance*

Profile likelihood and treatment of nuisance parameters

- Often we are either concerned only with the one parameter (of interest) λ, and treat the rest of other free (nuisance) parameters $\boldsymbol{\nu}$ separately: $\boldsymbol{\Theta} = \{\lambda, \boldsymbol{\nu}\}$

- Produce "profile log-likelihood" curve, a function of only one parameter (at a time), maximized over all others.

- Wilk's theorem say that this "profile log-likelihood" curve should behave as a

$$TS = 2\left(\ln \mathcal{L}\left(\lambda, \hat{\hat{\nu}}\left(\lambda\right)\right) - \ln \mathcal{L}\left(\hat{\lambda}, \hat{\nu}\right)\right) \sim \chi^2(1)$$

Set of parameters that maximize the likelihood simultaneously

Profile likelihood and treatment of nuisance parameters

- Often we are either concerned only with the one parameter (of interest) λ, and treat the rest of other free (nuisance) parameters ***ν*** separately: $\boldsymbol{\Theta} = \left\{ \lambda, \boldsymbol{\nu} \right\}$

- Produce "profile log-likelihood" curve, a function of only one parameter (at a time), maximized over all others.

- Wilk's theorem say that this "profile log-likelihood" curve should behave as a

$$TS = 2\left(\ln \mathcal{L}\left(\lambda, \hat{\hat{\nu}}\left(\lambda\right)\right) - \ln \mathcal{L}\left(\hat{\lambda}, \hat{\nu}\right)\right)$$

Given value of the parameter of interest to be tested

Profile likelihood and treatment of nuisance parameters

- Often we are either concerned only with the one parameter (of interest) λ, and treat the rest of other free (nuisance) parameters $\boldsymbol{\nu}$ separately: $\Theta = \{\lambda, \boldsymbol{\nu}\}$

- Produce "profile log-likelihood" curve, a function of only one parameter (at a time), maximized over all others.

- Wilk's theorem say that this "profile log-likelihood" curve should behave as a

$$TS = 2\left(\ln \mathcal{L}\left(\lambda, \hat{\hat{\nu}}(\lambda)\right) - \ln \mathcal{L}\left(\hat{\lambda}, \hat{\nu}\right)\right)$$

The set of nuisance parameters that
maximize the likelihood (simultaneously) for the given λ

Profile likelihood and treatment of nuisance parameters

- Often we are either concerned only with the one parameter (of interest) λ, and treat the rest of other free (nuisance) parameters ***ν*** separately: $\boldsymbol{\Theta} = \{\lambda, \boldsymbol{\nu}\}$

Caveat:
only true for (any) fixed set of nuisance parameters!

- Wilk's theorem say that this "profile log-likelihood" curve should behave as a

$$TS = 2\left(\ln \mathcal{L}\left(\lambda, \hat{\hat{\nu}}(\lambda)\right) - \ln \mathcal{L}\left(\hat{\lambda}, \hat{\nu}\right)\right)$$

The set of nuisance parameters that maximize the likelihood (simultaneously) for the given λ

Confidence intervals

- Find two values of λ where *TS* decreases by 1 w.r.t. its maximum:
  - yields a 2-sided 1σ confidence interval (68% probability)
  - is usually asymmetric !

- *Normally* can derive any confidence interval of N-σ, where *TS* decreases by N^2 w.r.t. its maximum.

- If Wilk's theorem holds (i.e. the likelihood is *well-behaved*), the range of parameters enclosed by the $(TS - N^2)$ contains the true parameter λ in a part of cases which correspond to an integrated normal distribution in a range of $(\mu - N\sigma, \mu + N\sigma)$, e.g. if the same experiment was repeated many times.

Confidence intervals

- Finding points where *TS* decreases by 1 w.r.t. its maximum:
  - yields a 2-sided 1σ confidence interval (68%)
  - is usually asymmetric !
- Normally can derive any confidence interval of N-σ, where *TS* decreases by N^2 w.r.t. its maximum.
- If Wilk's theorem holds (i.e. the likelihood is *well-behaved*), the range of parameters enclosed by the $(TS - N^2)$ contains the true parameter λ in a part of cases which correspond to an integrated normal distribution in a range of (μ-$N\sigma$, μ+$N\sigma$), e.g. if the same experiment was repeated many times.
- If the previous relation hold, the likelihood is said to have the correct coverage.

Confidence intervals

- Finding points where *TS* decreases by 1 w.r.t. its maximum:
  - yields a 2-sided 1σ confidence interval (68%)
  - is usually asymmetric !

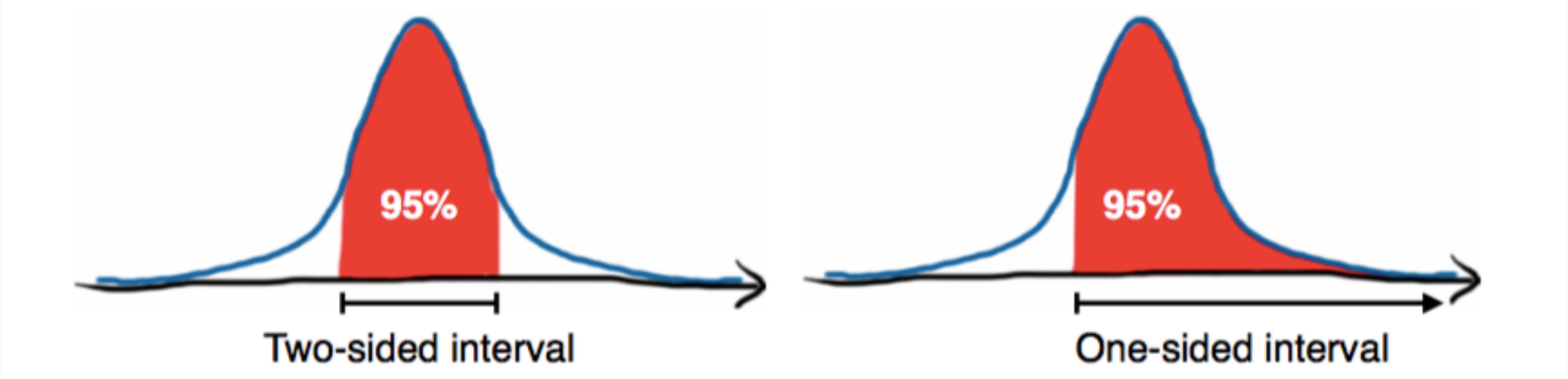
The relation below can (and should!) be checked with MC simulations !

- If Wilk's theorem holds (i.e. the likelihood is *well-behaved*), the range of parameters enclosed by the $(TS - N^2)$ contains the true parameter λ in a part of cases which correspond to an integrated normal distribution in a range of (μ-$N\sigma$, μ+$N\sigma$), e.g. if the same experiment was repeated many times.

- If the previous relation hold, the likelihood is said to have the correct coverage.

Confidence limits

(see Rolke et al., NIM A, 551, 493 (2005))



In two-sided interval search for two points where

$$2\left(\ln\mathcal{L}\left(\lambda,\hat{\hat{\nu}}(\lambda)\right)-\ln\mathcal{L}\left(\hat{\lambda},\hat{\nu}\right)\right)=N$$

For one-sided interval, we need to find single such a point for which $\int_{0.5}^{x}\mathcal{N}(0,1)=(1-CL)/2$

E.g. for *CL=0.95* we search for $2\left(\ln\mathcal{L}\left(\lambda,\hat{\hat{\nu}}(\lambda)\right)-\ln\mathcal{L}\left(\hat{\lambda},\hat{\nu}\right)\right)=2.71$

Good practices

- Define all the parameters of an analysis before looking at the data.
  – Data selection “cuts”
  – Thresholds for claiming detection.

Good practices

- Define all the parameters of an analysis before looking at the data.
  - Data selection "cuts"
  - Thresholds for claiming detection.

- It is tempting to adjust the analysis procedure to enhance some small signal, **BUT THIS WILL DESTROY (artificially enhance) ANY DETECTION SIGNIFICANCE!**

Good practices

- Define all the parameters of an analysis before looking at the data.
  - Data selection “cuts”
  - Thresholds for claiming detection.

- It is tempting to adjust the analysis procedure to enhance some small signal, **BUT THIS WILL DESTROY (artificially enhance) ANY DETECTION SIGNIFICANCE!**

- Best practice is to do a blind analysis.
- Use MC or test (Crab Nebula) data to refine analysis in advance.

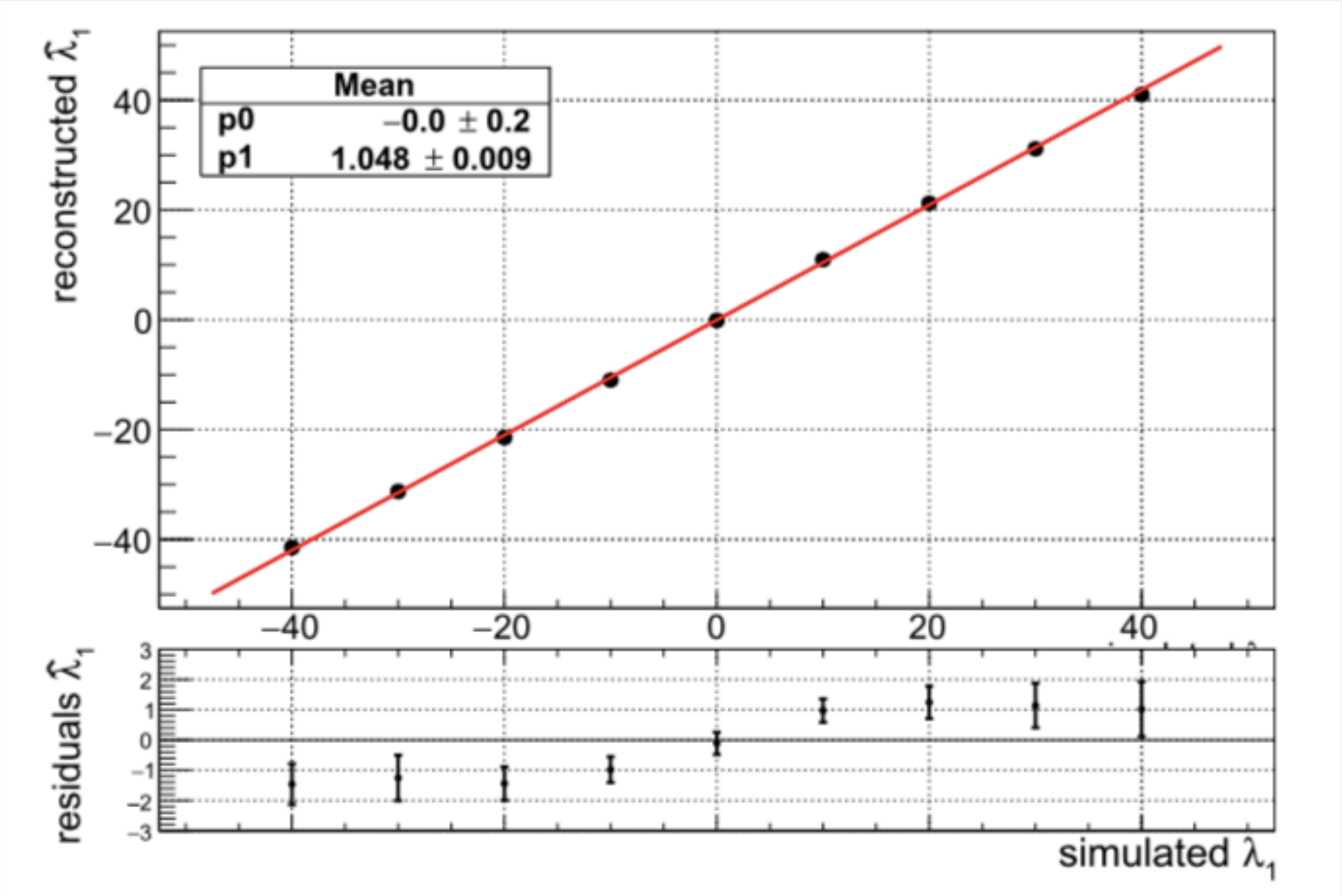
Good practices

- Define all the parameters of an analysis before looking at the data.
  - Data selection "cuts"
  - Thresholds for claiming detection.

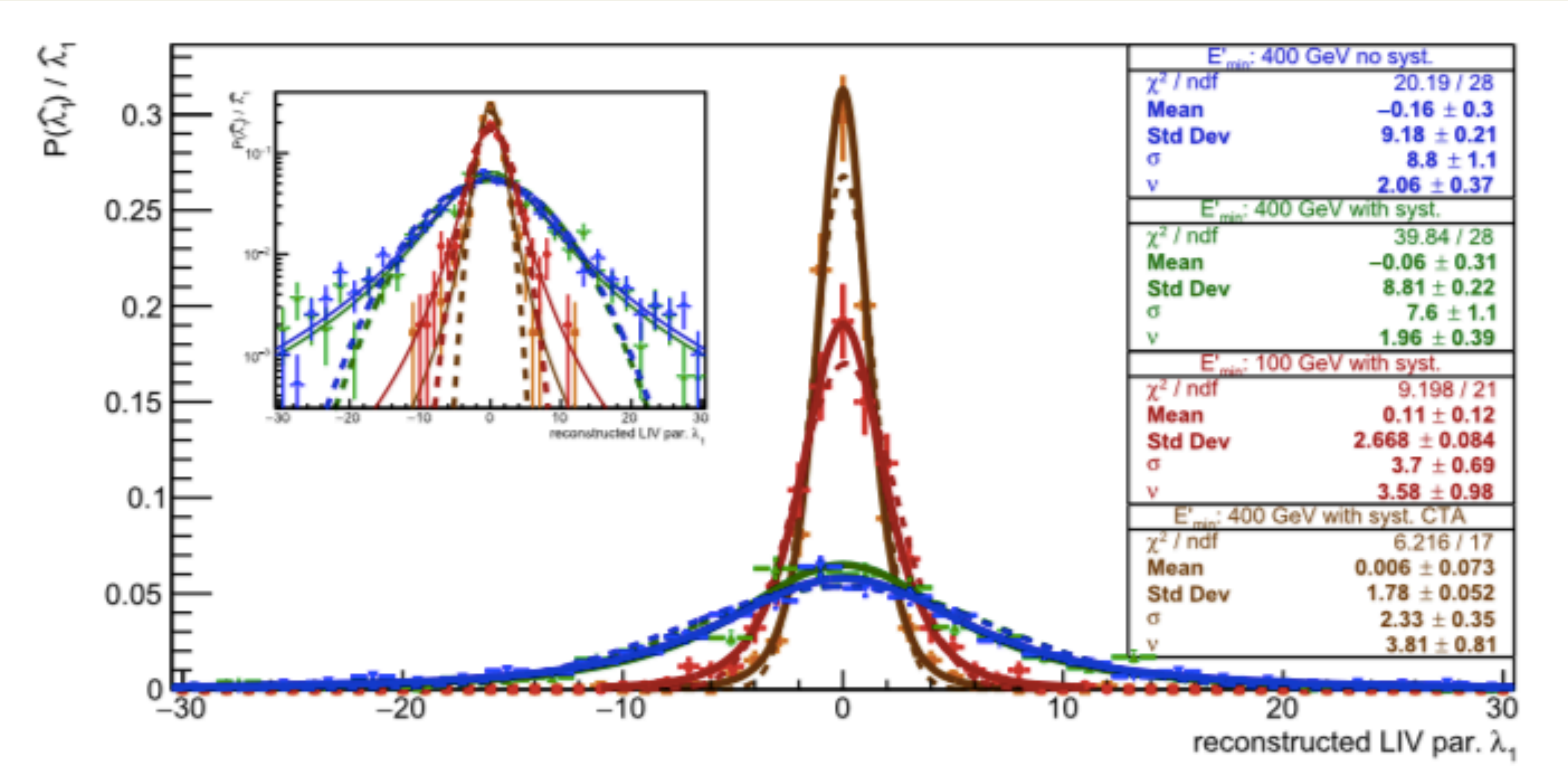
- It is tempting to adjust the analysis procedure to enhance some small signal, **BUT THIS WILL DESTROY (artificially enhance) ANY DETECTION SIGNIFICANCE!**
- Best practice is to do a blind analysis.
- Use MC or test (Crab Nebula) data to refine analysis in advance.

- Use extensive MC simulations to test the behaviour of your (profile) likelihood, particularly whether it converges correctly and whether it has the desired coverage.

Good practices



Good practices



And now, let's move to the real stuff…