Improved DATA MODELLING with Goodness-of-Fit & Likelihood-ratio tests

Terascale Statistics School, DESY, Feb 23, 2018

Olaf Behnke (DESY)

Content

Goodness-of-Fit tests for

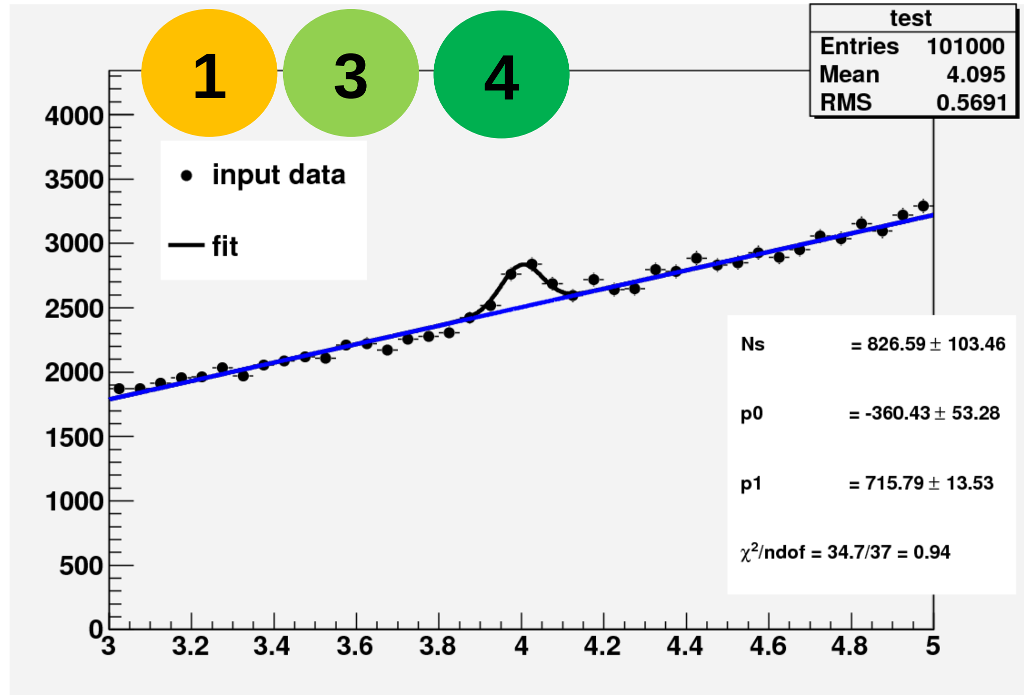
1 Checking data modelling

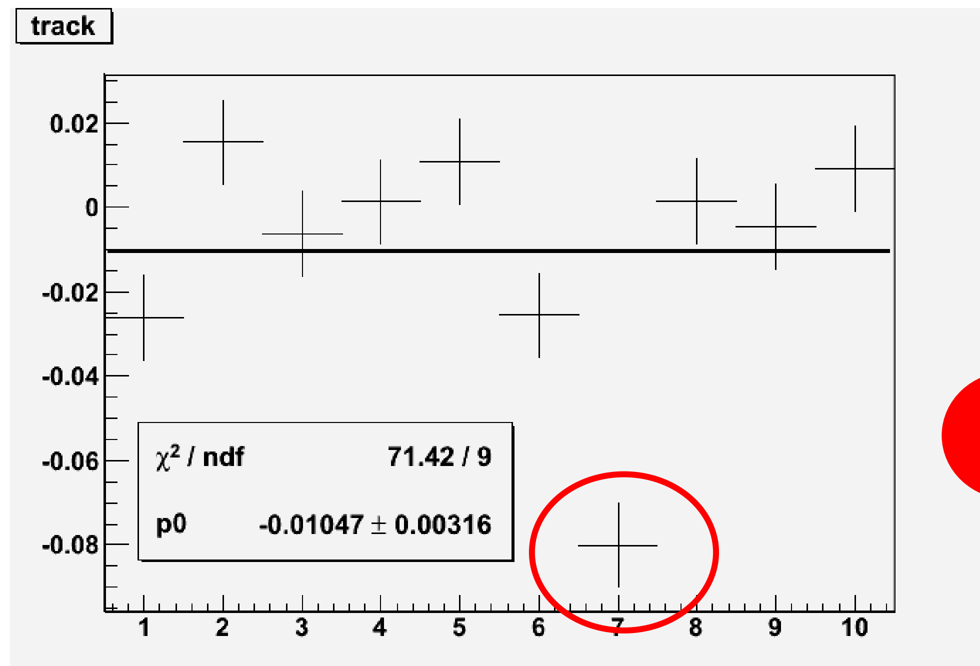
2 Outlier rejection

Likelihood ratio tests for background

3 Optimal parametrisation

4 Shape systematics (discrete profiling)

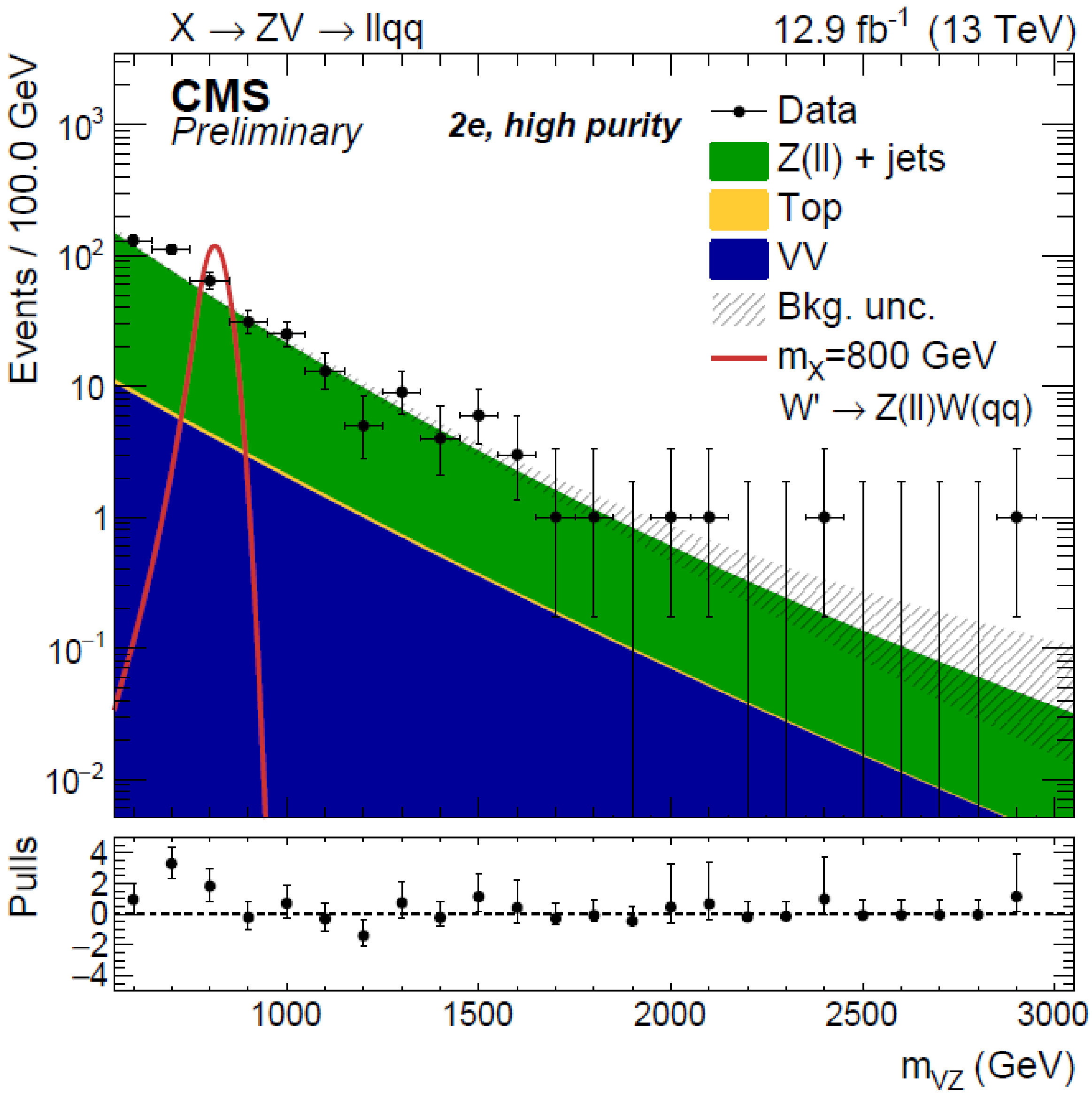




2

1 GOF tests for *checking data modelling*

Searches with Likelihood ratio



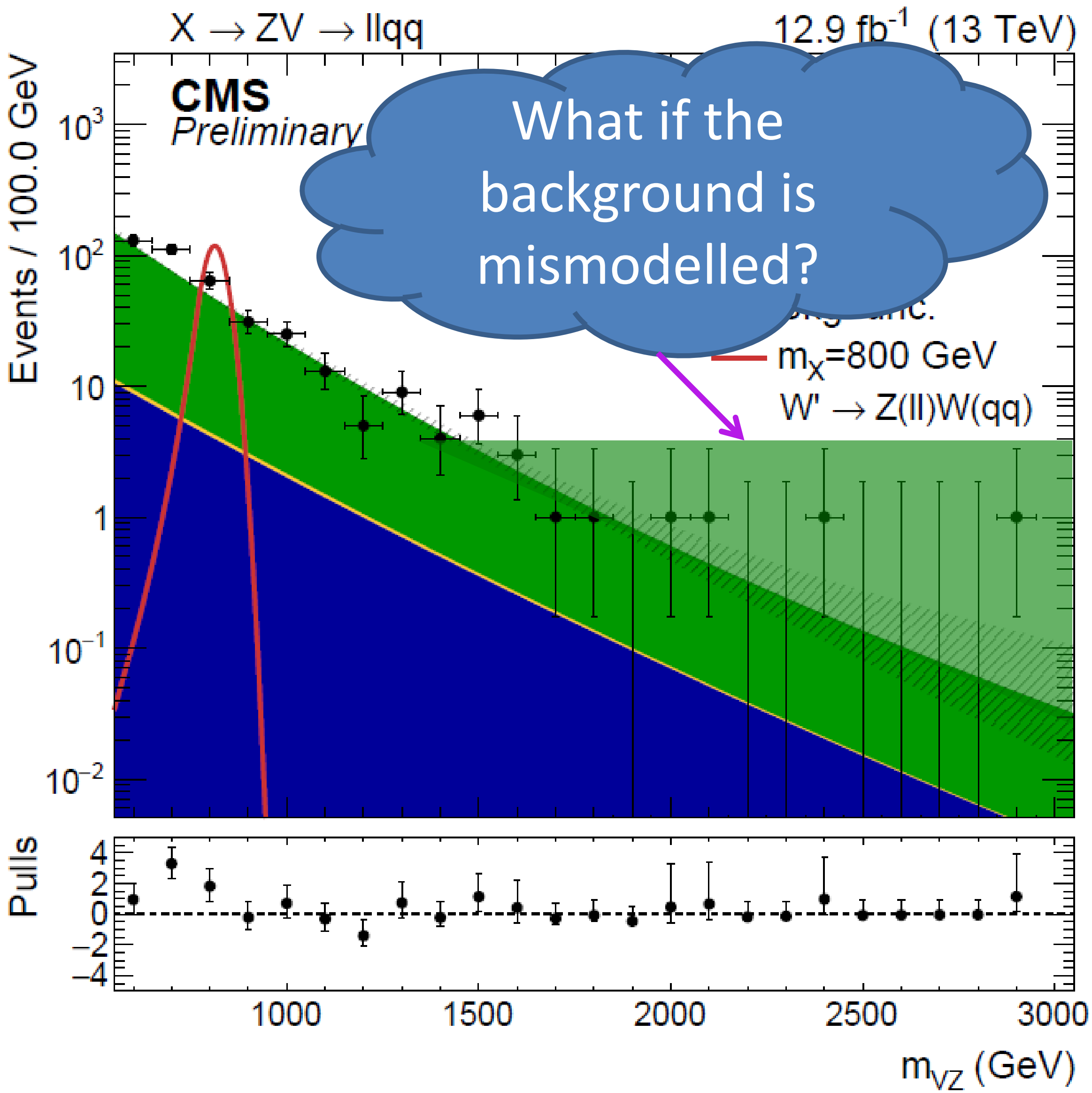
$$L(\mu) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)}$$

H0: μ=0, H1: μ>0

Neyman-Pearson:
L(H1)/L(H0)
→ max. power test
→ Use for discovery

CMS-PAS-B2G-16-022

Searches with Likelihood ratio



$$L(\mu)=\prod_{j=1}^{N}\frac{(\mu s_j+b_j)^{n_j}}{n_j!}e^{-(\mu s_j+b_j)}$$

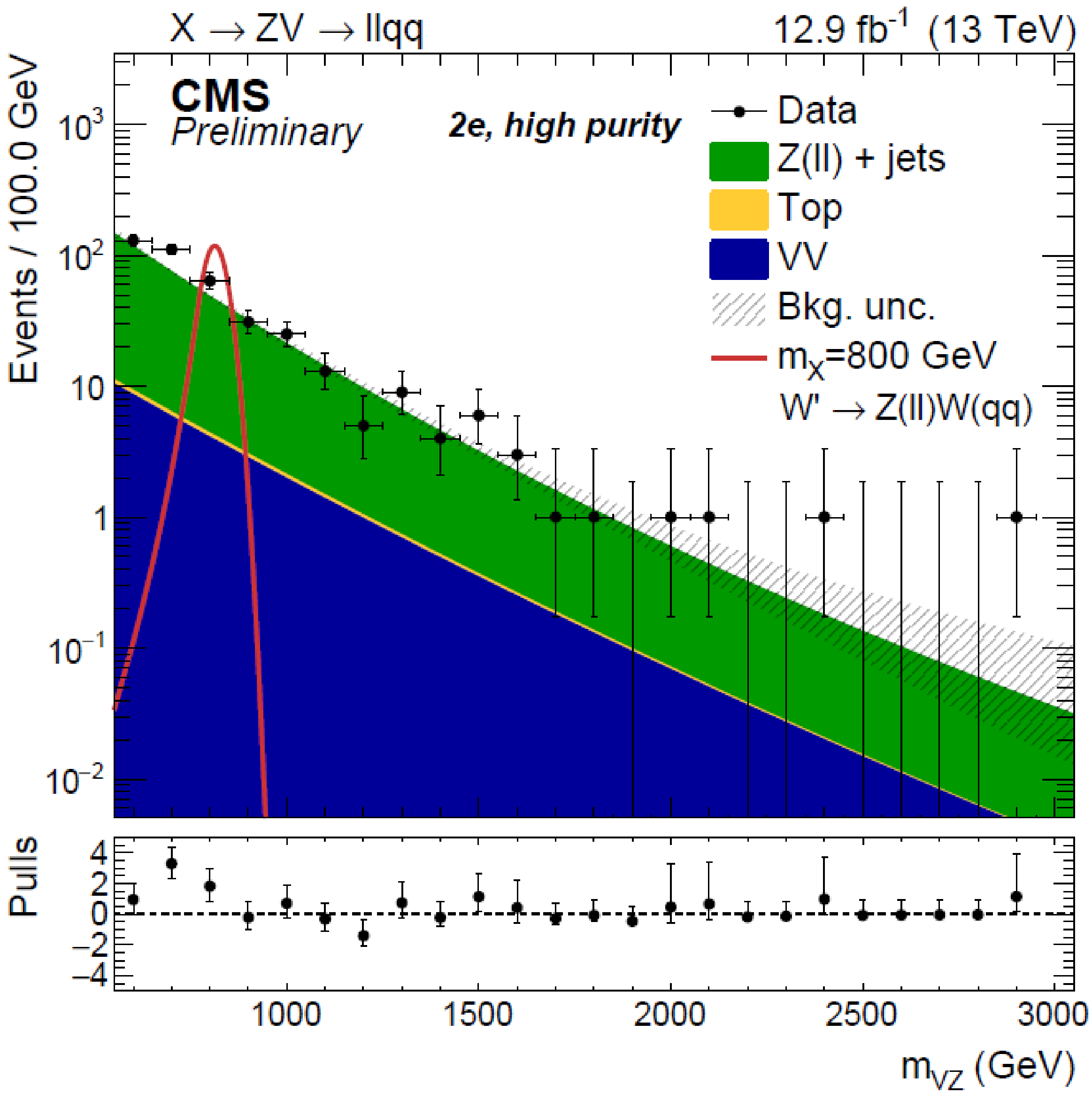
H0: $\mu=0$, H1: $\mu>0$

Neyman-Pearson:
L(H1)/L(H0)
→ max. power test
→ Use for discovery

⇒ Don't trust a damn thing !!

CMS-PAS-B2G-16-022

Searches with Likelihood ratio



$$L(\mu) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)}$$

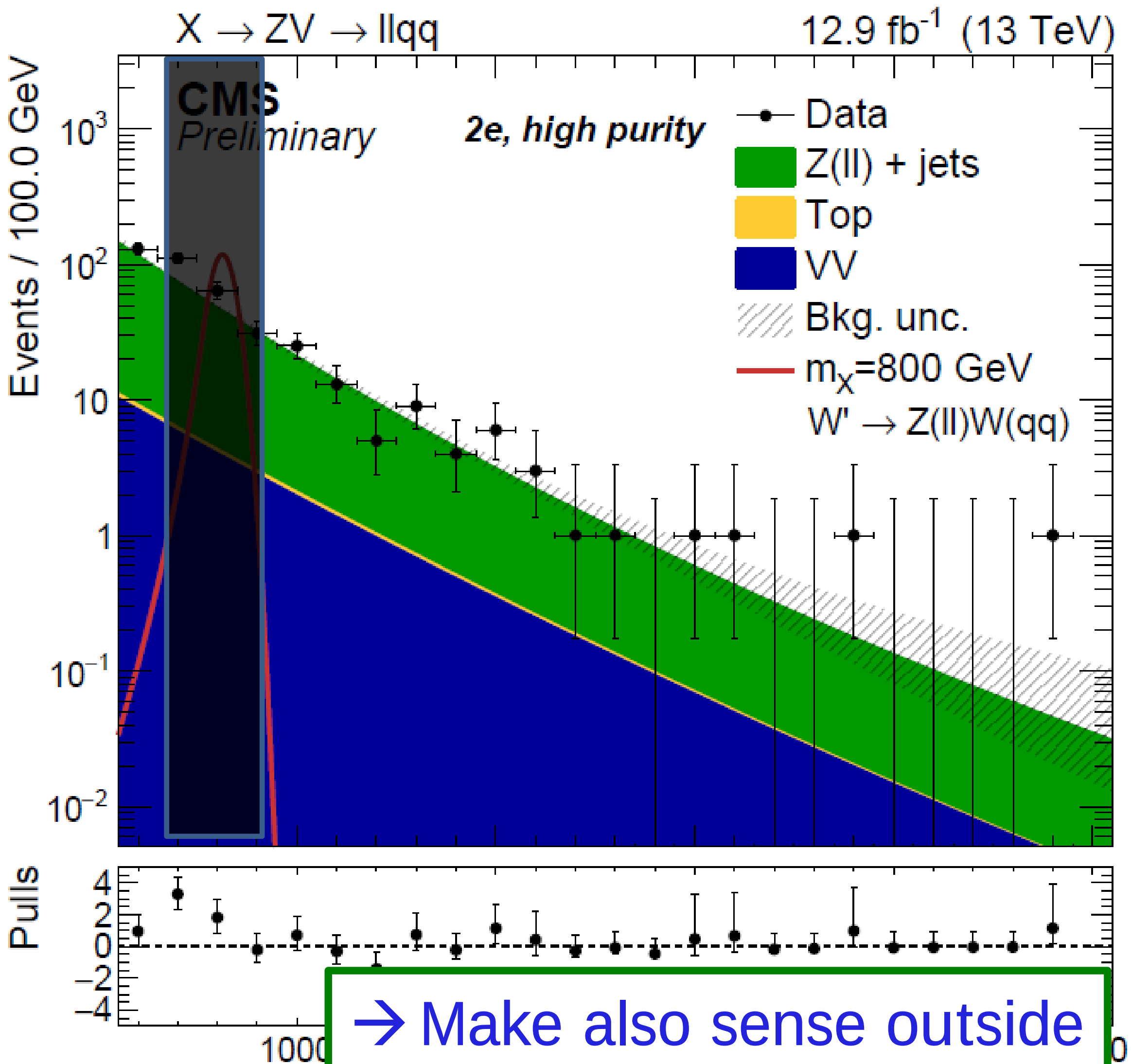
H0: $\mu=0$, H1: $\mu>0$

Neyman-Pearson:
L(H1)/L(H0)
→ max. power test
→ Use for discovery

→ Proper modelling of background over whole range is essential → do GOF-tests for H0!

CMS-PAS-B2G-16-022

Searches with Likelihood ratio



$$L(\mu) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)}$$

H0: μ=0, H1: μ>0

Neyman-Pearson:
L(H1)/L(H0)
→ max. power test
→ Use for discovery

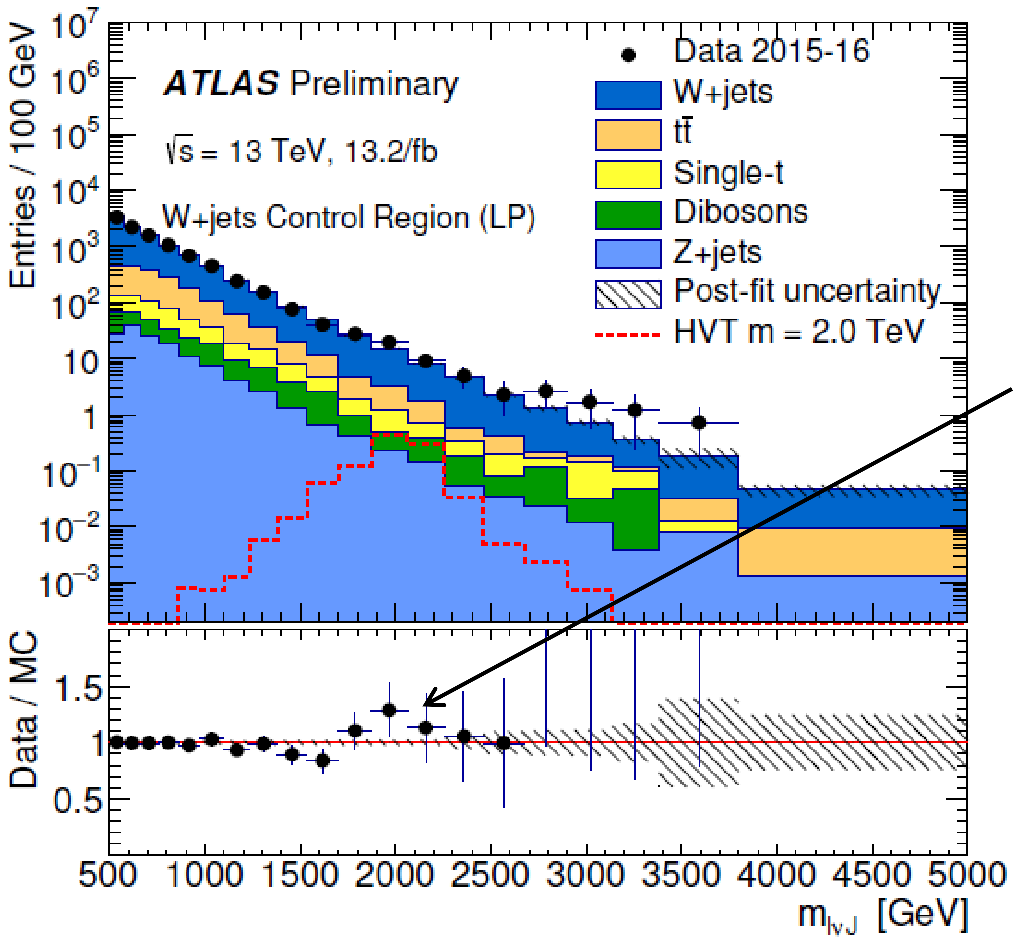
→ Proper modelling of background over whole range is essential → do GOF-tests for H0!

→ Make also sense outside blinded signal region!

CMS-PAS-B2G-16-022

Goodness-Of-Fit Tests in ATLAS and CMS searches

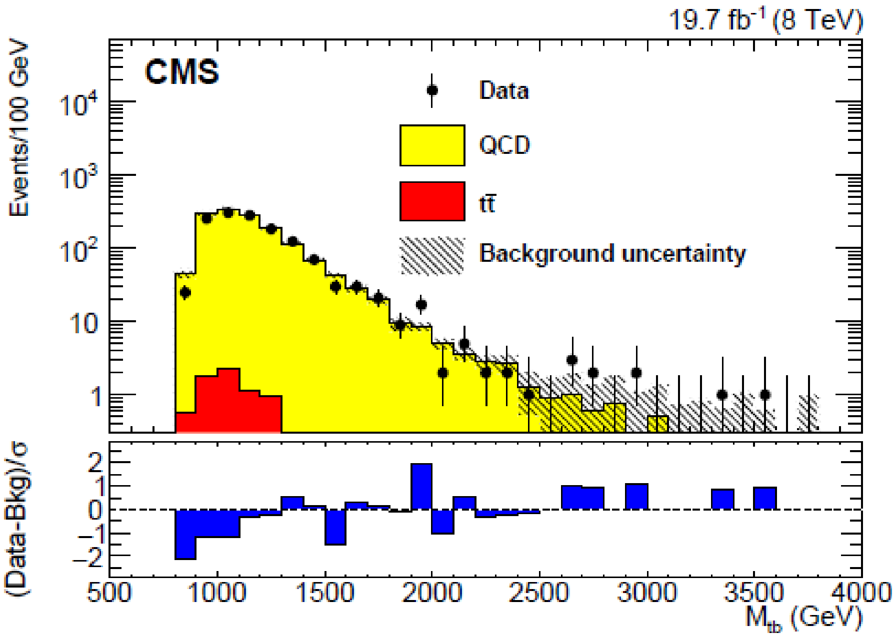
ATLAS-CONF-2016-062 Diboson resonance

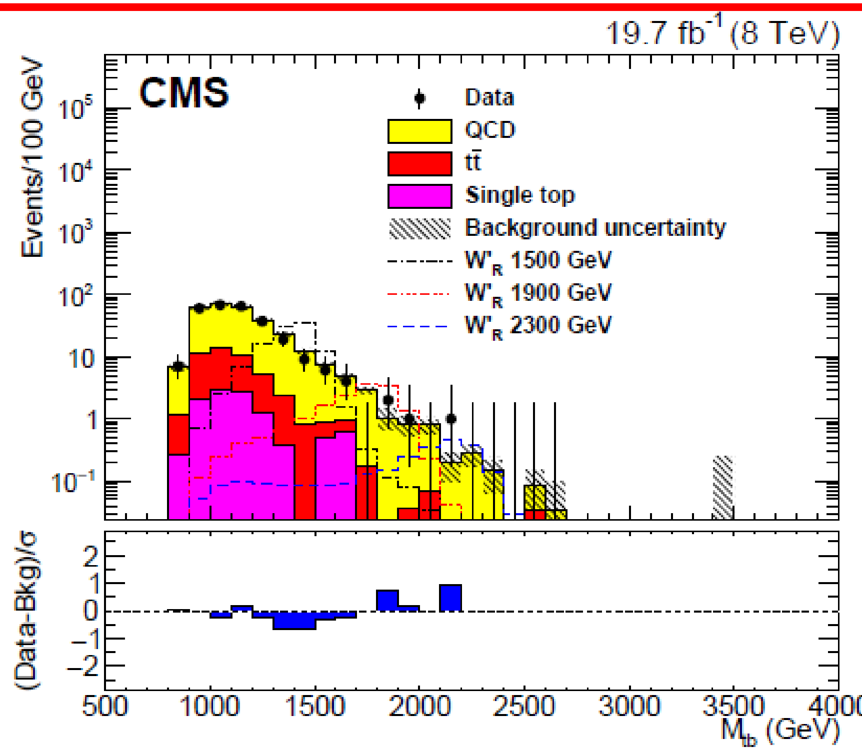


"Good agreement (optical inspection of ratio) between the data and the background prediction"

Goodness-Of-Fit Tests in ATLAS and CMS searches

arXiv:1509.06051v2 W' → tb



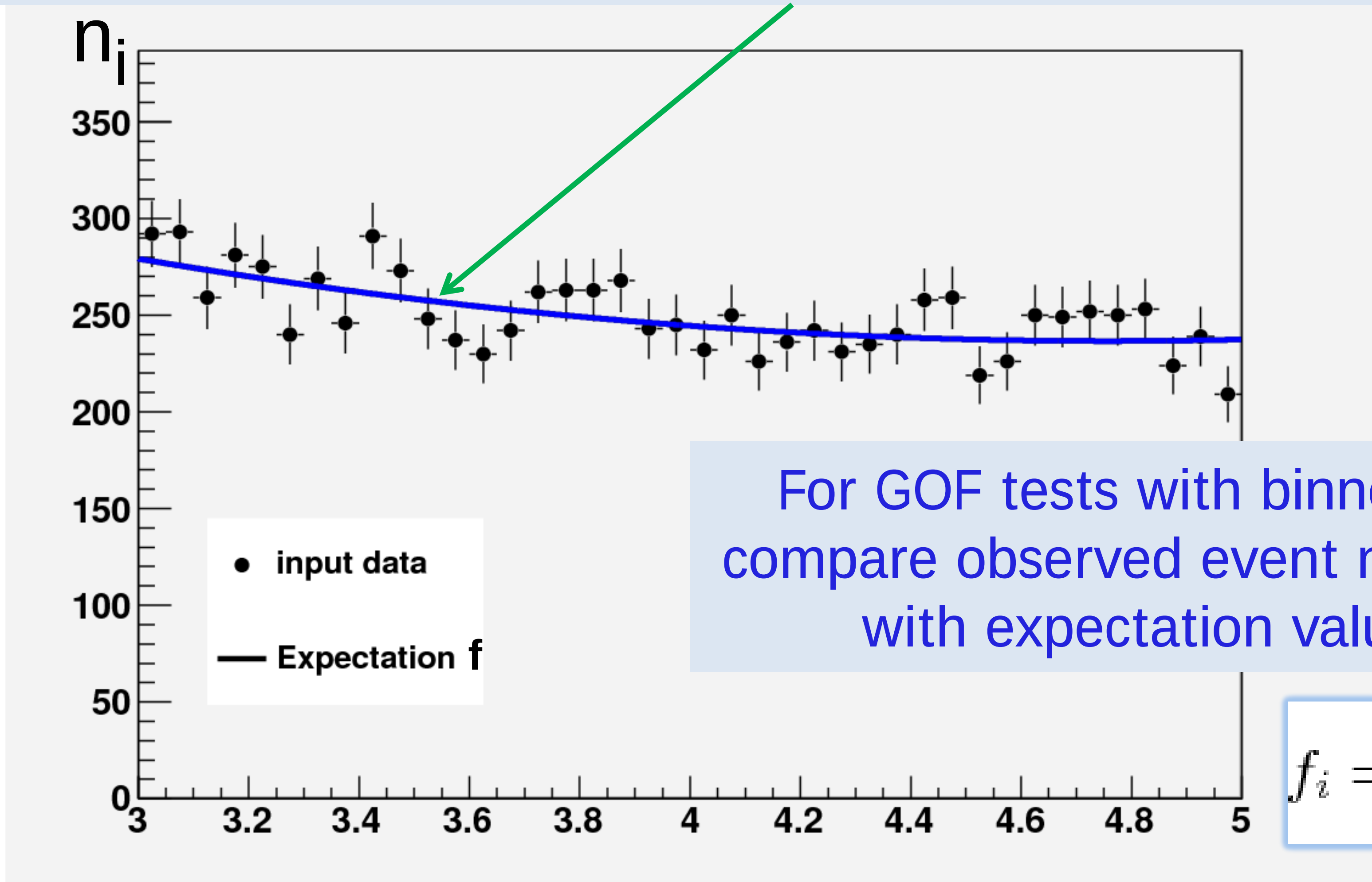


"This test (optical pull inspection) shows good agreement between data and SM"

→ Optical inspection of bin-wise pulls **is crucial** but should do also global tests as discussed in the following slides!

Goodness-Of-Fit Tests – basics

Basic question: how well does H_0 describe the data?



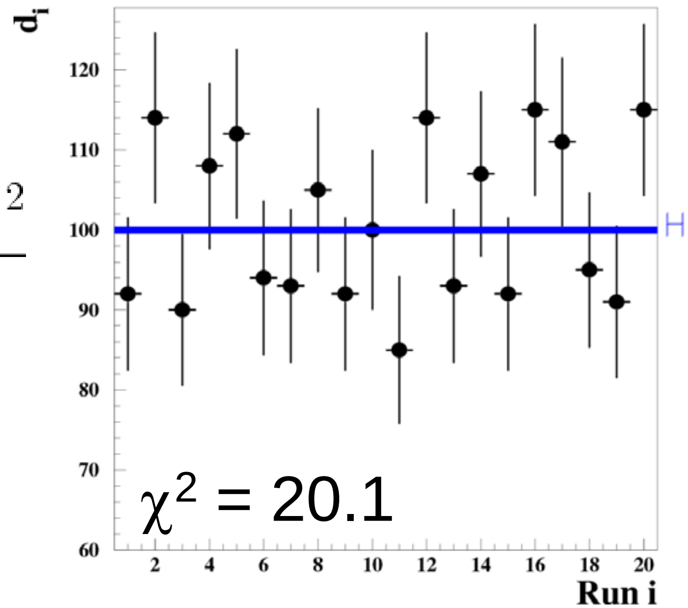
For GOF tests with binned data: compare observed event numbers n_i with expectation values f_i

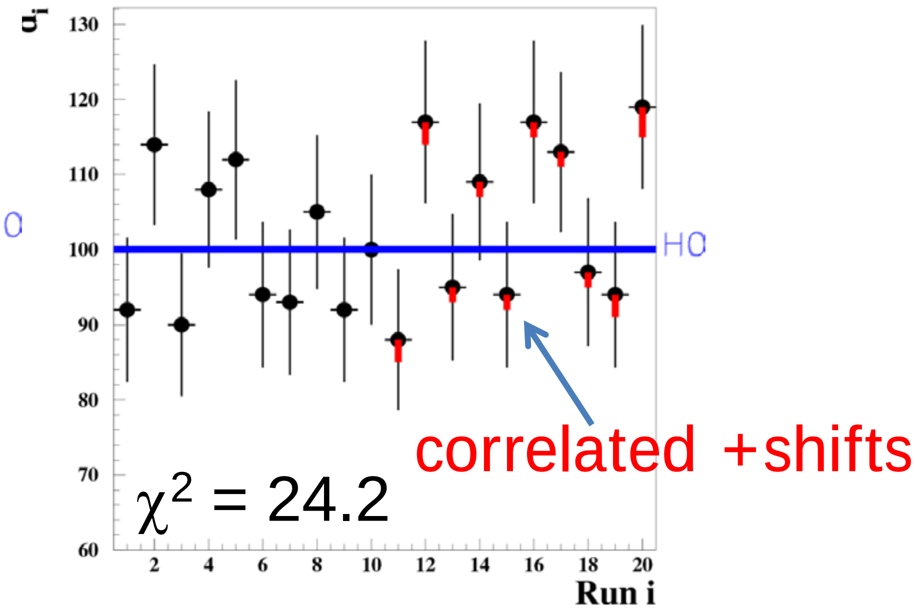
$$f_i = \int_{bin\ i} f\, dm$$

→ Since no H_1 specified → many different GOF tests possible

Goodness-Of-Fit Test – χ^2 tests

$$\chi^2 = \sum_i \frac{(f_i - n_i)^2}{\sigma_i^2}$$





→ χ^2 throws away all sign and order info → not very sensitive to correlated shifts in a certain region.

→ apply further GOF tests to check all data/model facets!

Note: p-values for χ^2: TMath::Prob(χ^2_{obs},ndf)

GOF-tests: exemplary analysis

Hypothetical pp data@100 TeV

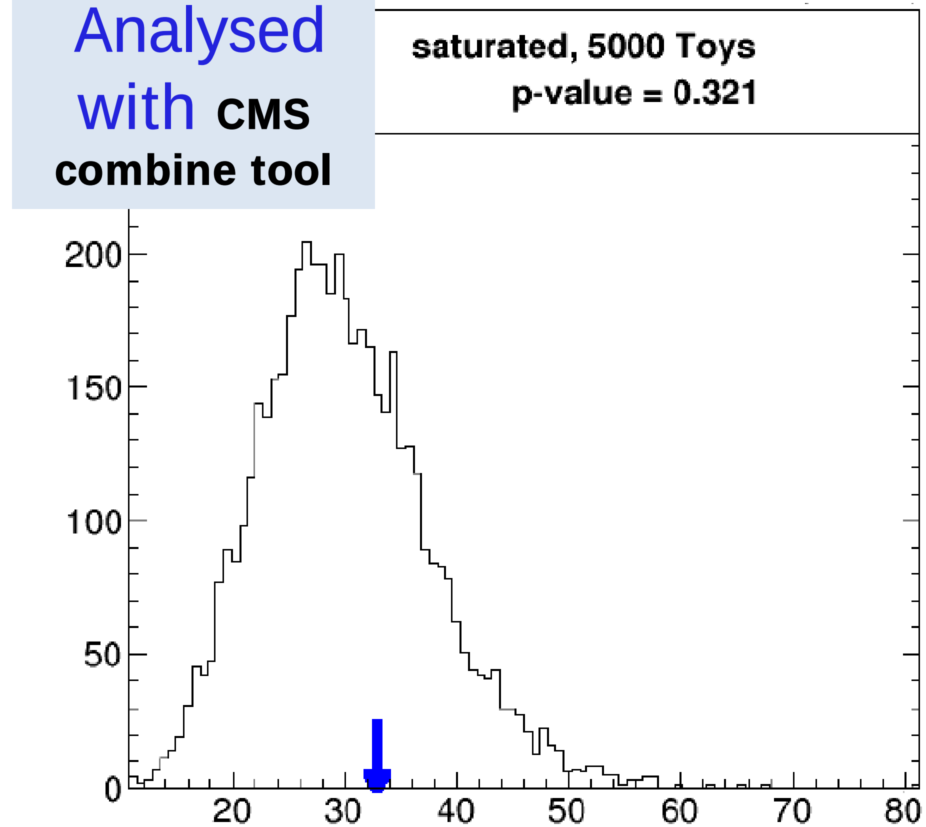
Likelihood ratio improved χ^2 test

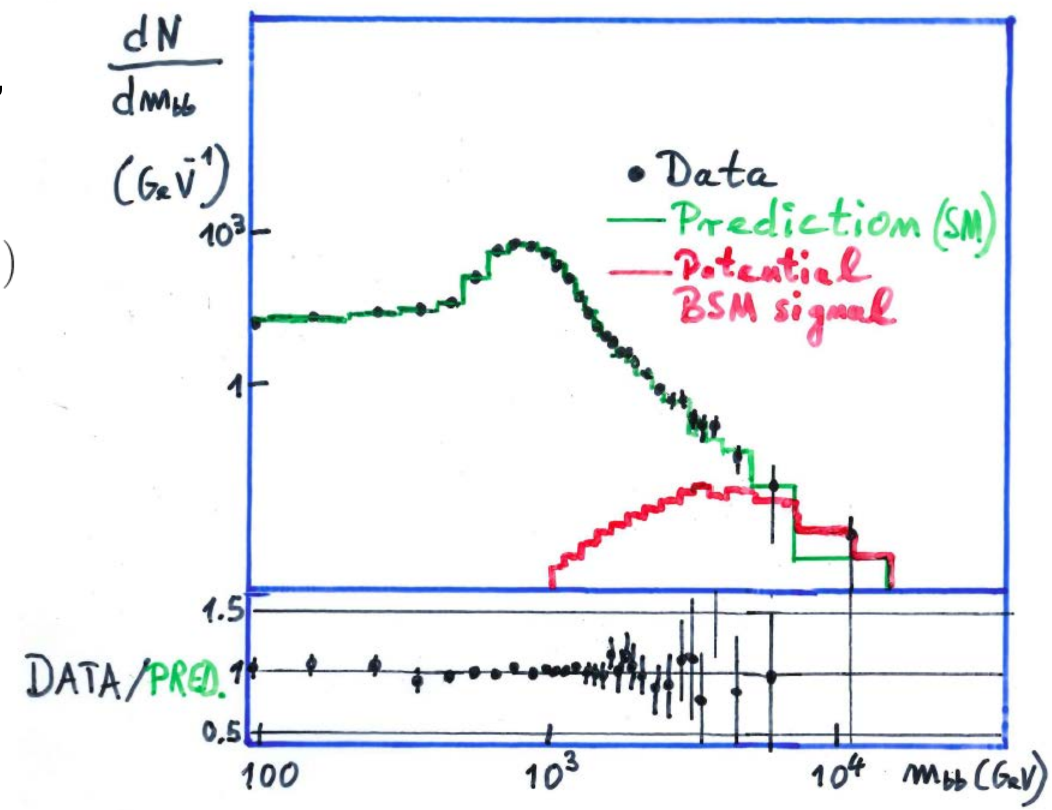
S. Baker & R.D. Cousins, NIM 221 (1984) 437

$$\tilde{\chi}^2 = -2\ln\left(\frac{L(\boldsymbol{n}|\boldsymbol{f})}{L(\boldsymbol{n}|\boldsymbol{n})}\right) = 2\sum_i f_i - n_i + n_i \ln(n_i/f_i)$$

Saturated model

Analysed with **CMS combine tool**





→ Test result ok

GOF-tests: exemplary analysis

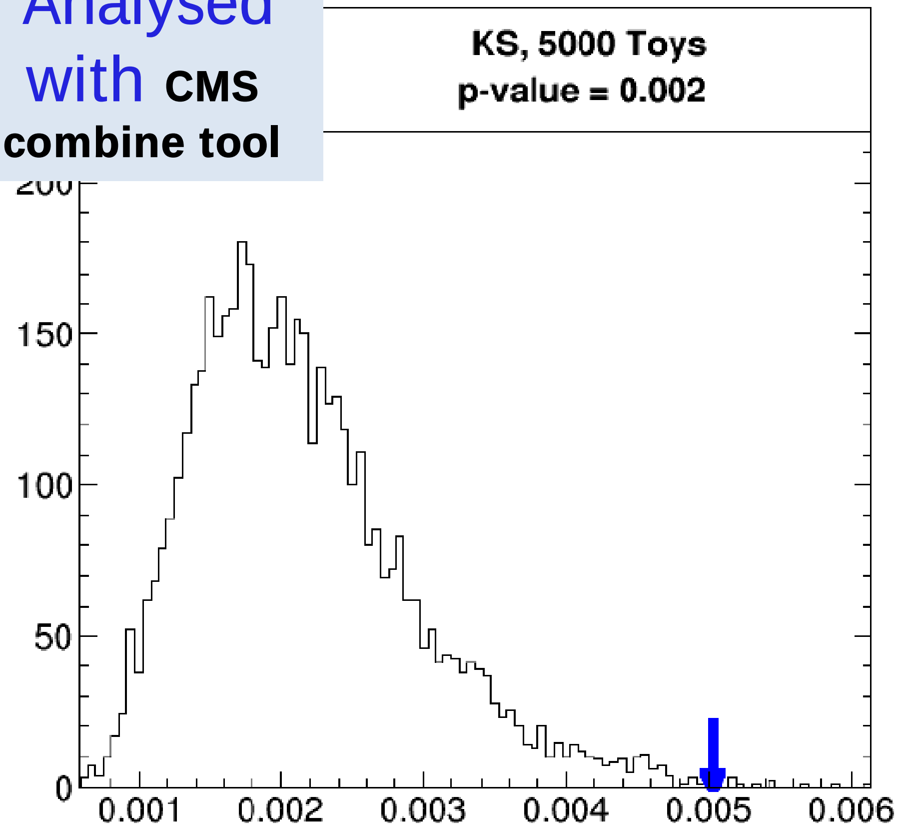
Kolgomorov-Smirnov test

F_c: Cumulative distribution function

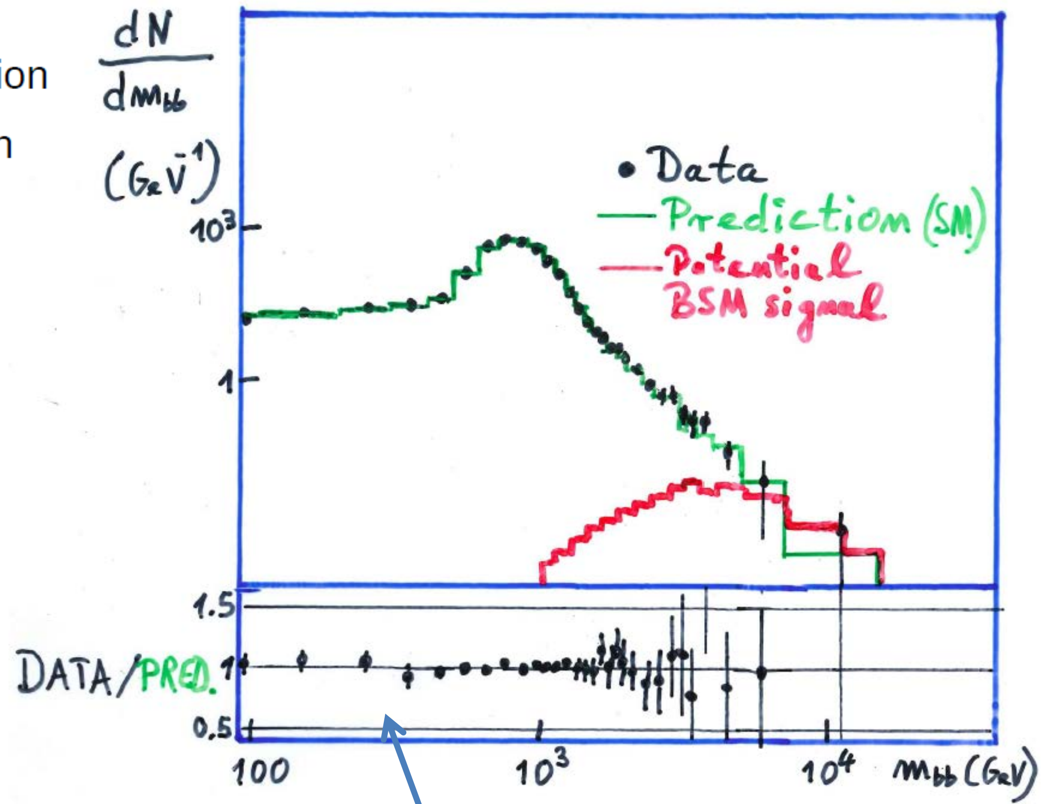
F_e: Empirical distribution function

$$q_{GoF,KS} = \sup \left| F_c(x) - F_e(x) \right|$$

Analysed with CMS combine tool



Hypothetical pp data@100 TeV



correlated shift?

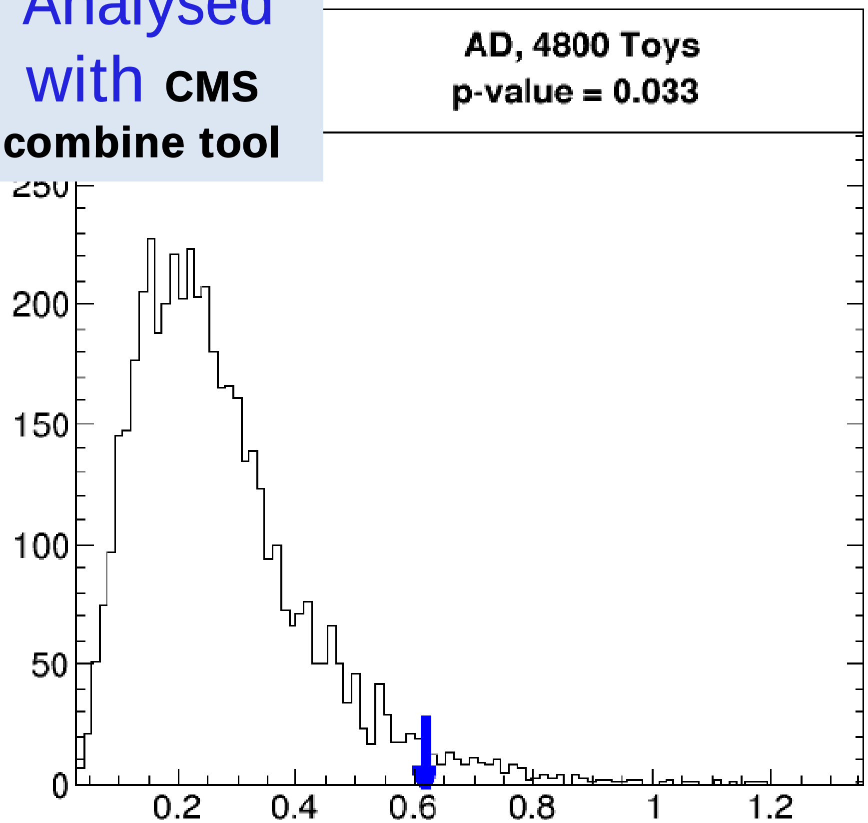
→ Test result bad!

GOF-tests: exemplary analysis

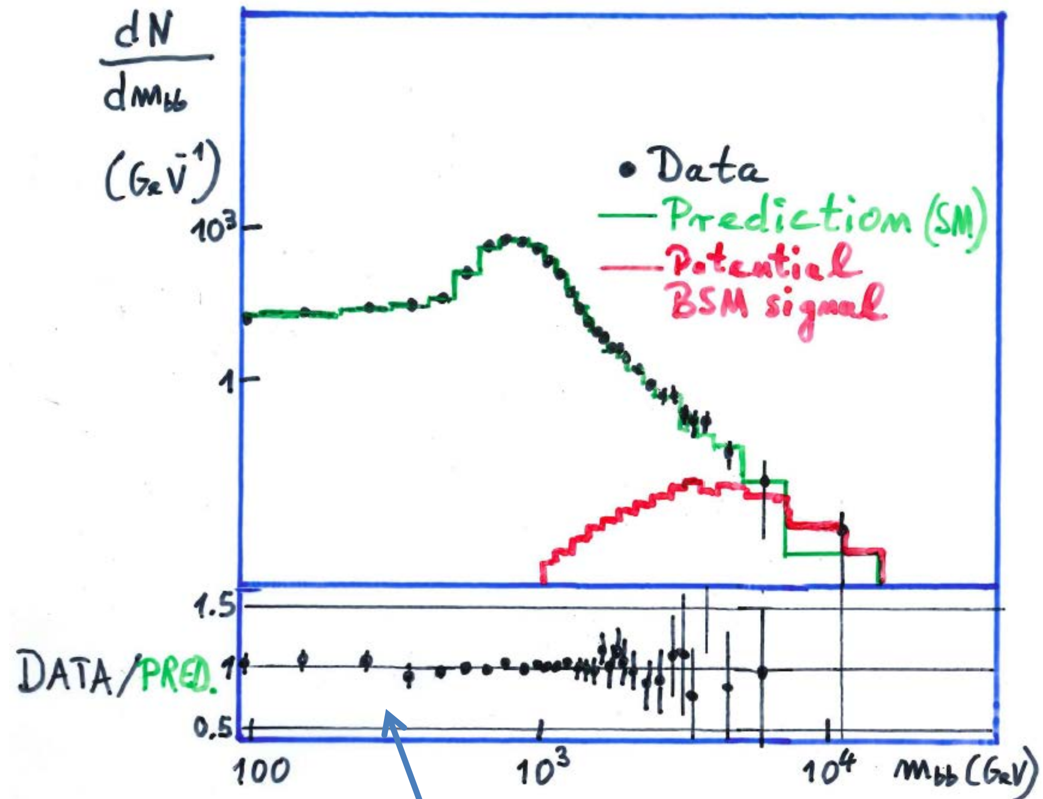
Anderson-Darling test

$$q_{GoF,AD} = n \cdot \int dF_e(x) \frac{\left(F_c(x) - F_e(x)\right)^2}{F_e(x) \cdot \left(1 - F_e(x)\right)}$$

Analysed with **CMS combine tool**



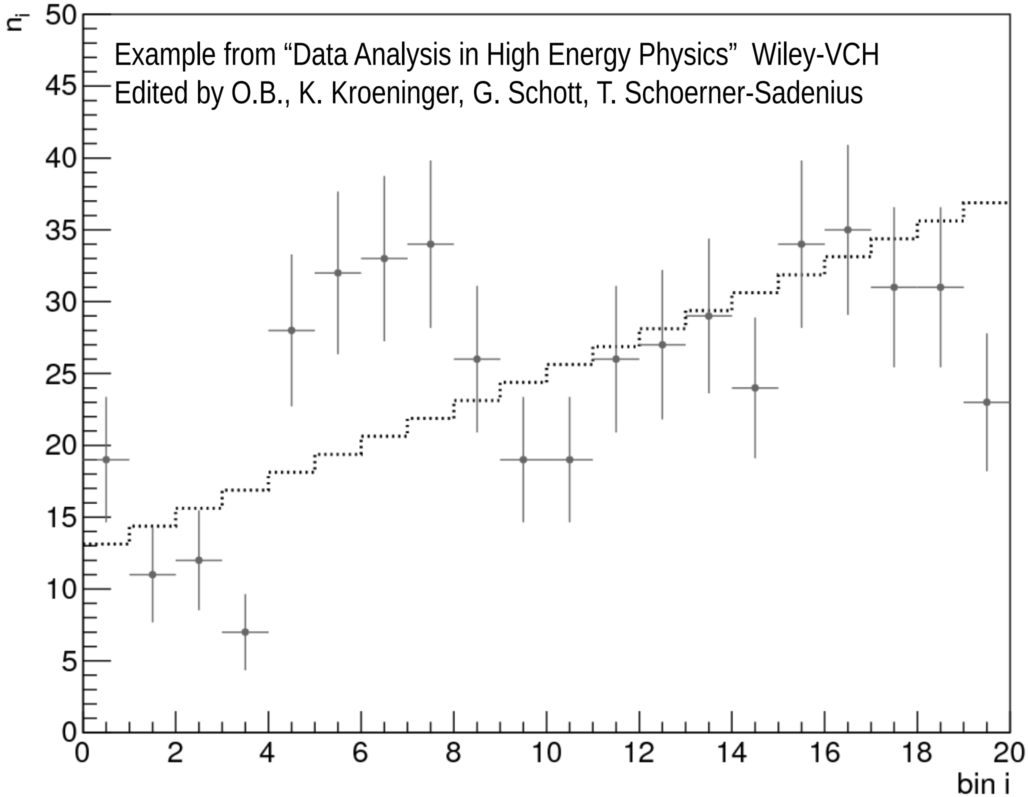
Hypothetical pp data@100 TeV



correlated shift?

→ Test result bad!

Goodness of Fit - Run test



$$(+ - - - + + + + + - - - - - - + + - - -)$$

Runs

Idea: count runs = regions with same sign of deviation

r = #runs

N_+ = #bins data>model

N_- = #bins data<model

r should follow Binomial statist.

$$E[r] = 1 + \frac{2N_+N_-}{N_+ + N_-}$$

$$V[r] = \frac{2N_+N_-(2N_+N_- - N_+ - N_-)}{(N_+ + N_-)^2(N_+ + N_- - 1)}$$

$$Z = \frac{r - E[r]}{\sqrt{V[r]}}$$

Approximate Significance

Here: r=6, E(r)=10.6±2.1

→ p-value = 0.0285

→ Easy to do test!

1 Summary GOF tests of modelling

Perform GOF tests through various analysis stages:

✓ **Control plots (!)**

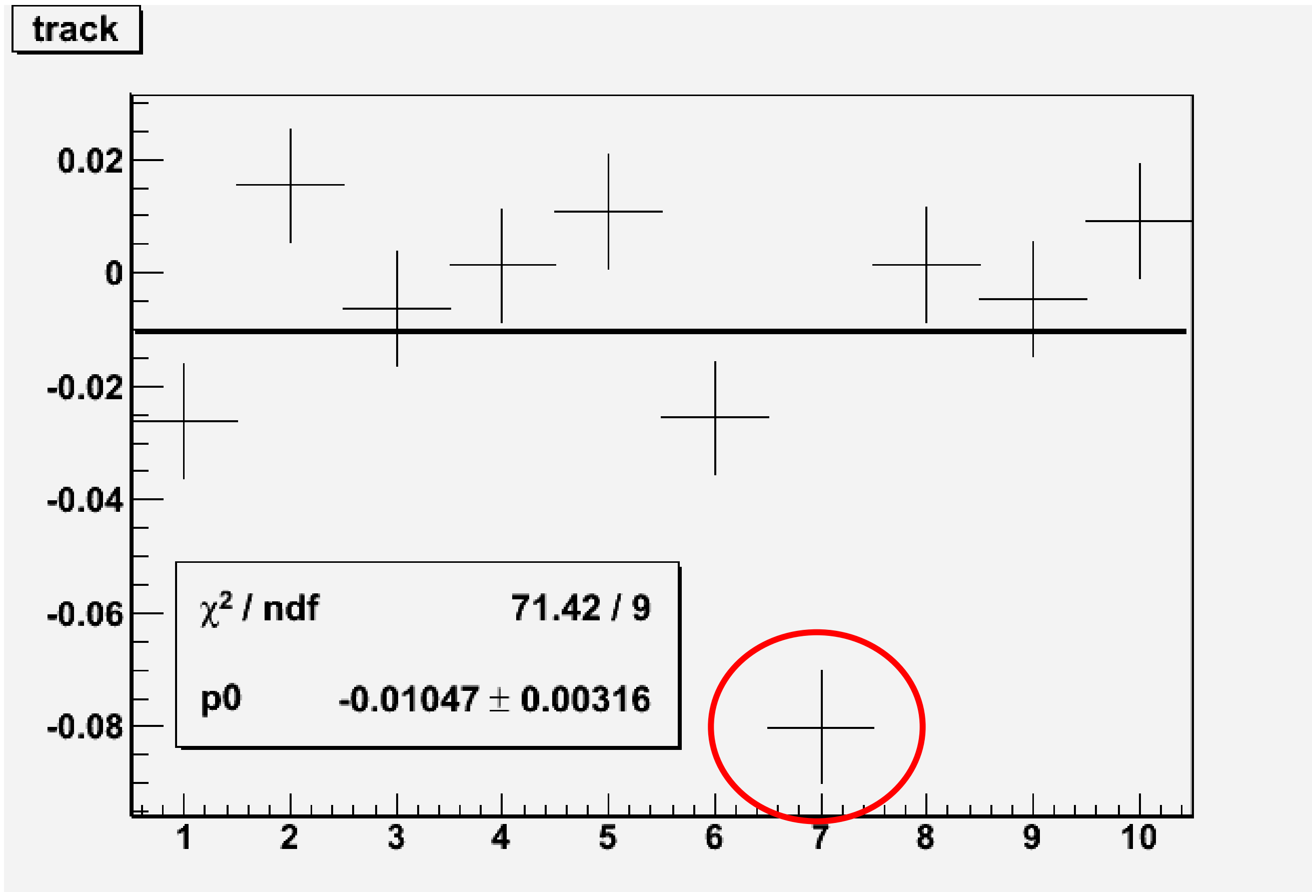
✓ **Signal Extraction (!!)**

✓ **Comparisons to theory (!!)**

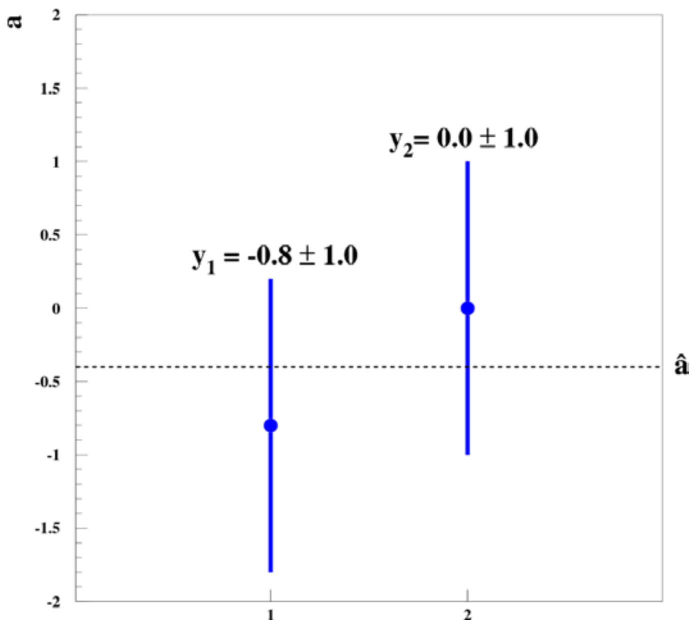
→ essential for understanding/control of analysis results and theory!

→ Apply ≥ two different tests, e.g $\tilde{\chi}^2$ and K.S.

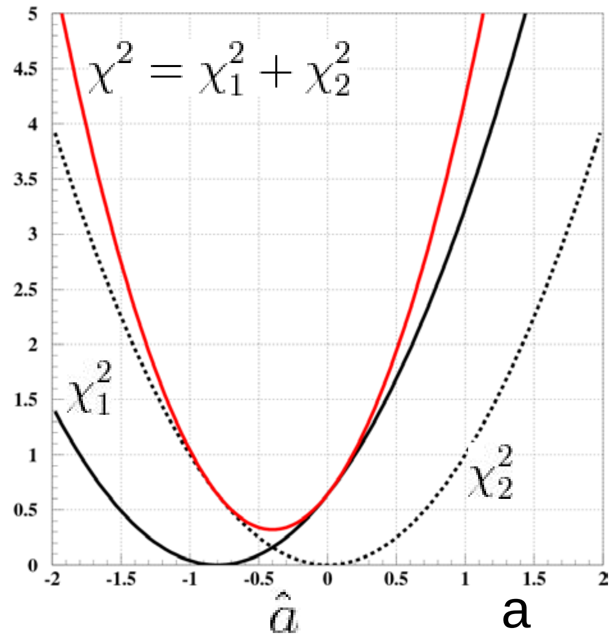
2 Use χ^2 tests for *outlier rejection*



Role of χ^2: Combination of two measurements



$$\chi_i^2 = \frac{(a - y_i)^2}{\sigma^2} \qquad L \sim e^{-\chi_1^2/2}\, e^{-\chi_1^2/2} = e^{-\chi^2/2}$$



$$\chi^2(a) = \chi^2(\hat{a}) + H(a - \hat{a})^2$$

$$\hat{a} = (y_1 + y_2)/2; \quad H = 2/\sigma^2 \qquad \chi^2(\hat{a}) = \frac{(y_1 - y_2)^2}{2\sigma^2}$$

← Squared pull

$$L \sim e^{-\frac{(y_1 - y_2)^2}{2\sigma^2}} \cdot e^{-H(a-\hat{a})^2/2}$$

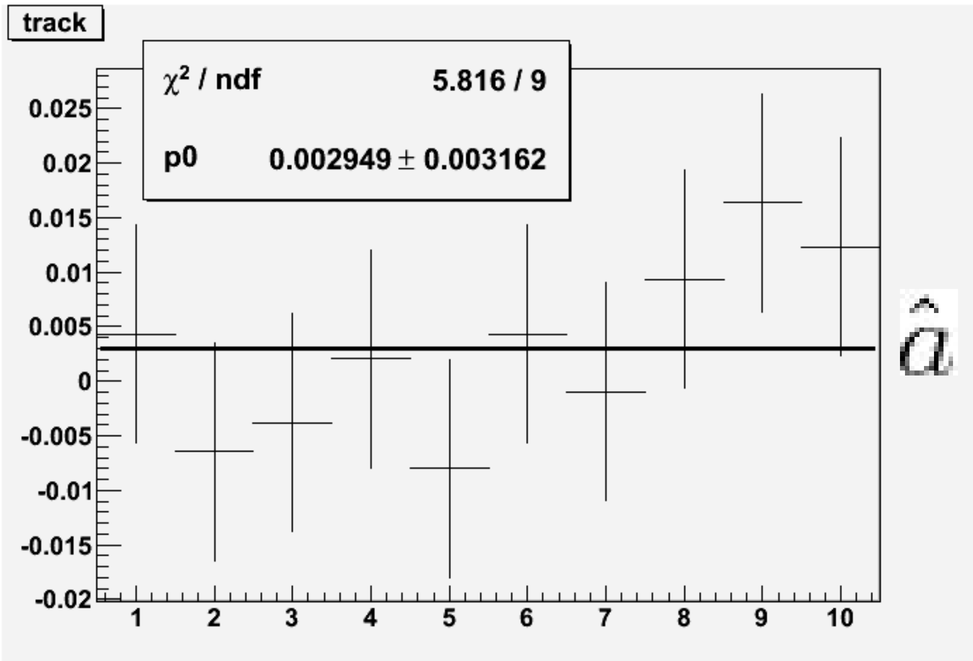
GOF-test

Info on a

→ GOF test and parameter Info decoupled!

Combination of n measurements

Example: track fit of horizontally flying particle in n detector layers



Weighted average position

Repeated experiments: $\chi^2_{min} = \chi^2(\hat{a})$ follows f(χ^2,n-1) distribution

$$f(\chi^2, n) = \frac{1}{\Gamma(n/2)2^{n/2}} \cdot (\chi^2)^{n/2-1} \cdot e^{-\chi^2/2}$$

$$\text{with} \quad \Gamma(n/2) = \int_0^\infty dt\, e^{-t} t^{n/2-1}$$

$$\text{TMath::Prob}(\chi^2, n) = \int_{\chi^2}^{\infty} f(\chi'^2, n)\, d\chi'^2$$

uniformly distr. in [0,1], why?

Track fits to 10 hits – **Interactive work with ROOT, Run 1000 Fits**

1) No noise: uncertainty of â, means of χ^2 and prob(χ^2 ,9) distr.

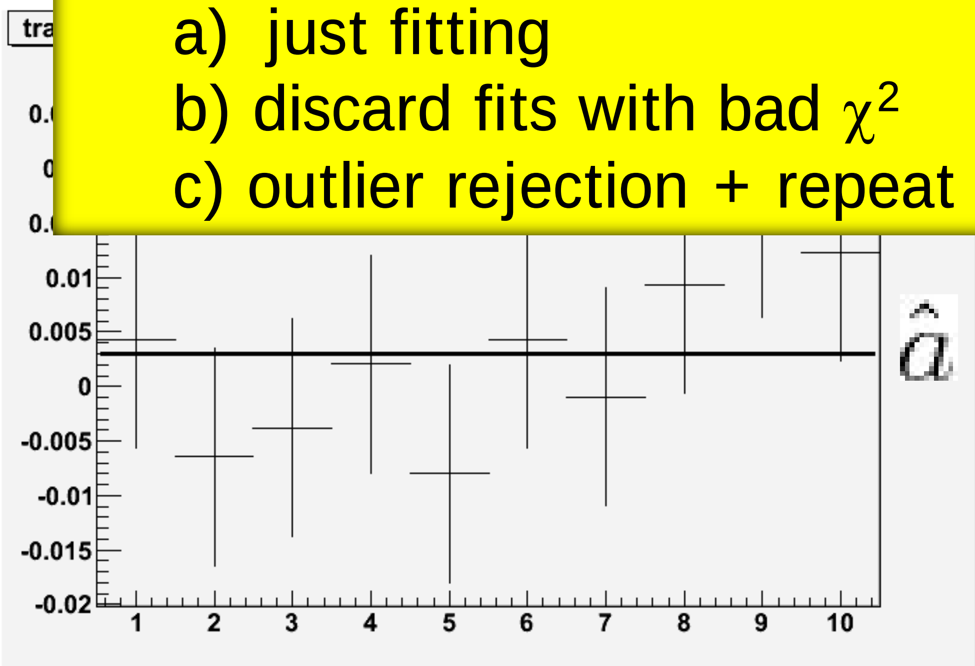
2) Repeat with noise hits

a) just fitting

b) discard fits with bad χ^2

c) outlier rejection + repeat track fit

Ex



$\hat{a}$ ← Weighted average position

Repeated experiments: $\chi^2_{min} = \chi^2(\hat{a})$ follows f(χ^2,n-1) distribution

$$f(\chi^2, n) = \frac{1}{\Gamma(n/2)2^{n/2}} \cdot (\chi^2)^{n/2-1} \cdot e^{-\chi^2/2}$$

$$\text{with} \quad \Gamma(n/2) = \int_0^\infty dt\, e^{-t} t^{n/2-1}$$

$$\text{TMath::Prob}(\chi^2, n) = \int_{\chi^2}^{\infty} f(\chi'^2, n)\, d\chi'^2$$

uniformly distr. in [0,1], why?

Track fits to 10 hits – **Interactive work with ROOT, Run 1000 Fits**

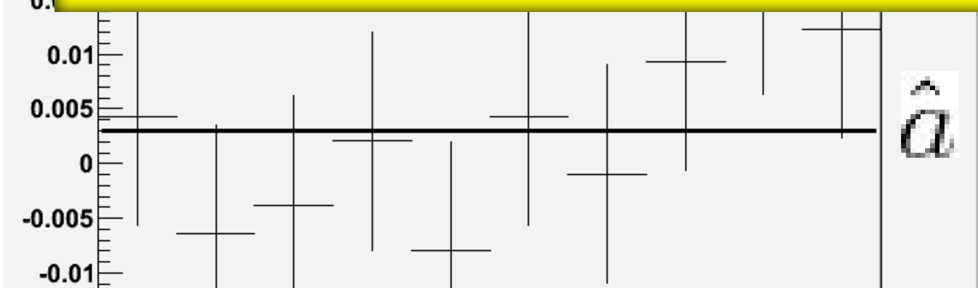
1) No noise: uncertainty of â, means of χ^2 and prob(χ^2 ,9) distr.

2) Repeat with noise hits:

   a) just fitting

   b) discard fits with bad χ^2

   c) outlier rejection + repeat track fit



← Weighted average position

ROOT 6 Macro available at
www.desy.de/~obehnke/stat/school_18feb/p0toyf.C

Instructions available at
www.desy.de/~obehnke/stat/school_18feb/compueb_p0toyf.pdf

$$f(\chi^2, n) = \frac{1}{\Gamma(n/2)2^{n/2}} \cdot (\chi^2)^{n/2-1} \cdot e^{-\chi^2/2}$$

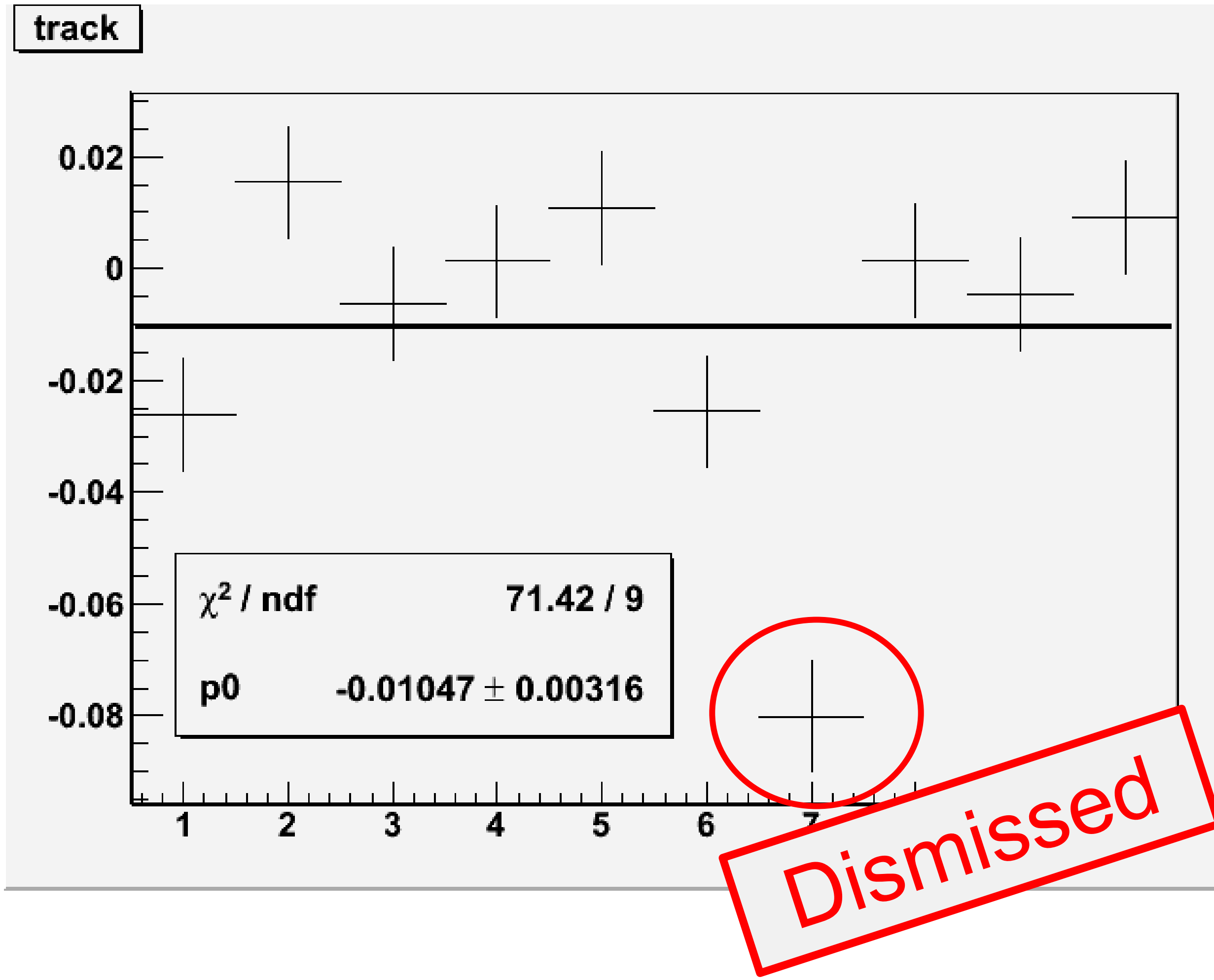
$$\text{with} \quad \Gamma(n/2) = \int_0^\infty dt\, e^{-t} t^{n/2-1}$$

$$\text{TMath::Prob}(\chi^2, n) = \int_{\chi^2}^{\infty} f(\chi'^2, n)\, d\chi'^2$$

uniformly distr. in [0,1], why?

2

χ^2 tests for outlier rejection - Summary

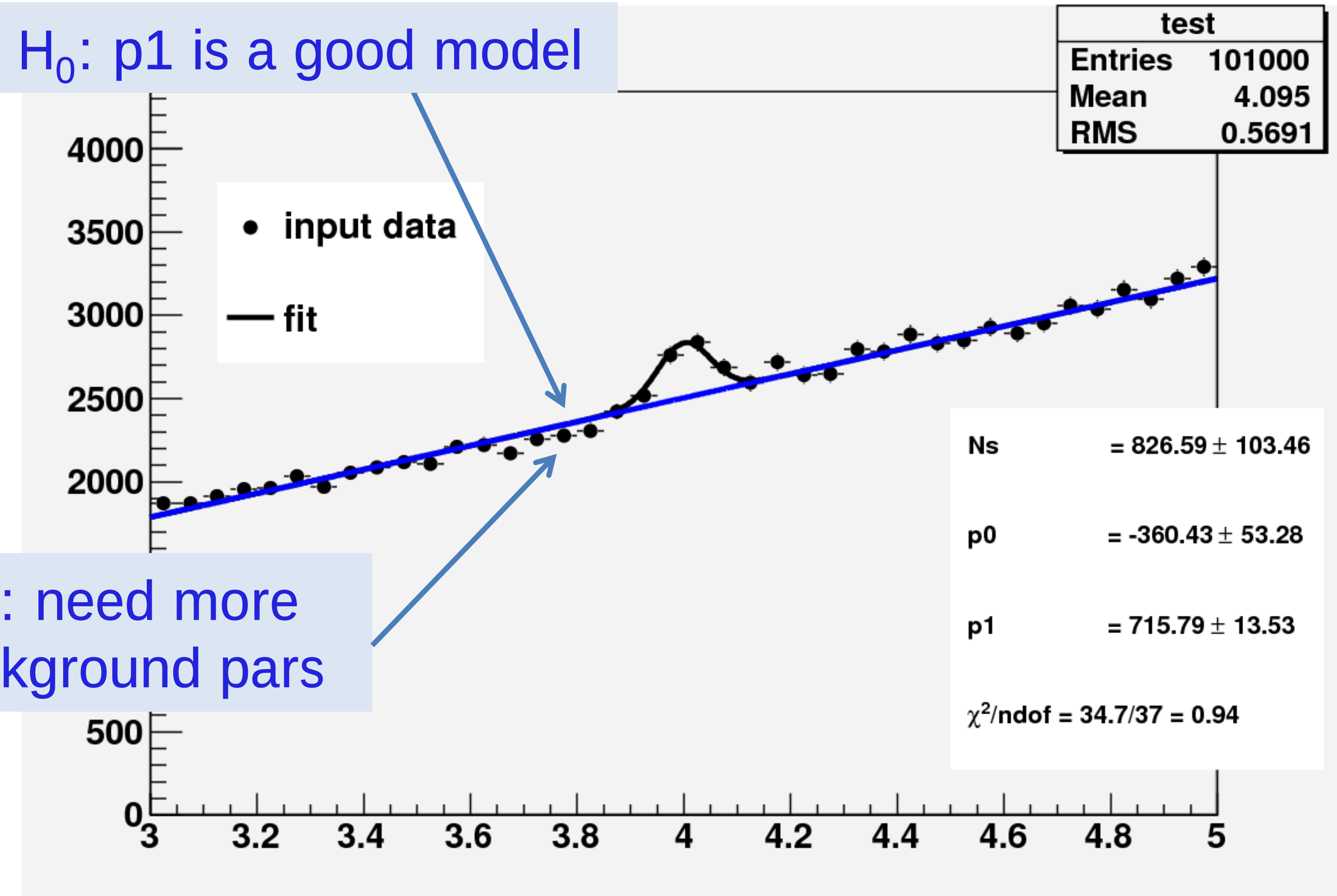


→ Can use χ^2 tests as powerful tool for Pattern recognition tasks

Note: Rejecting hits with χ^2 >5 is hard cut, tune (e.g. try 10)

3 GOF+Likelihood-ratio tests for *optimal background parametrisation*

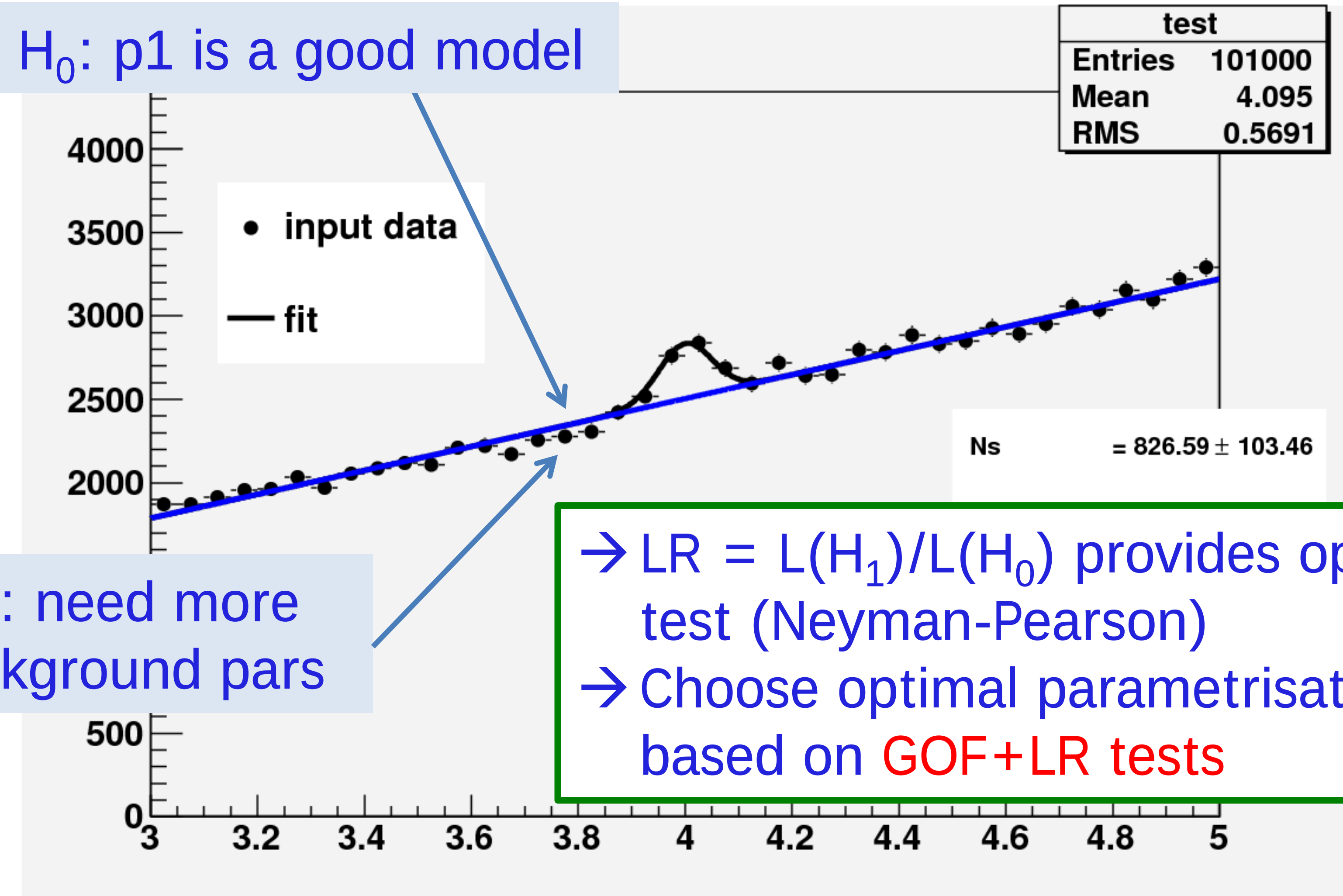
H_0: p1 is a good model



H_1: need more background pars

3 GOF+Likelihood-ratio tests for *optimal background parametrisation*

H_0: p1 is a good model

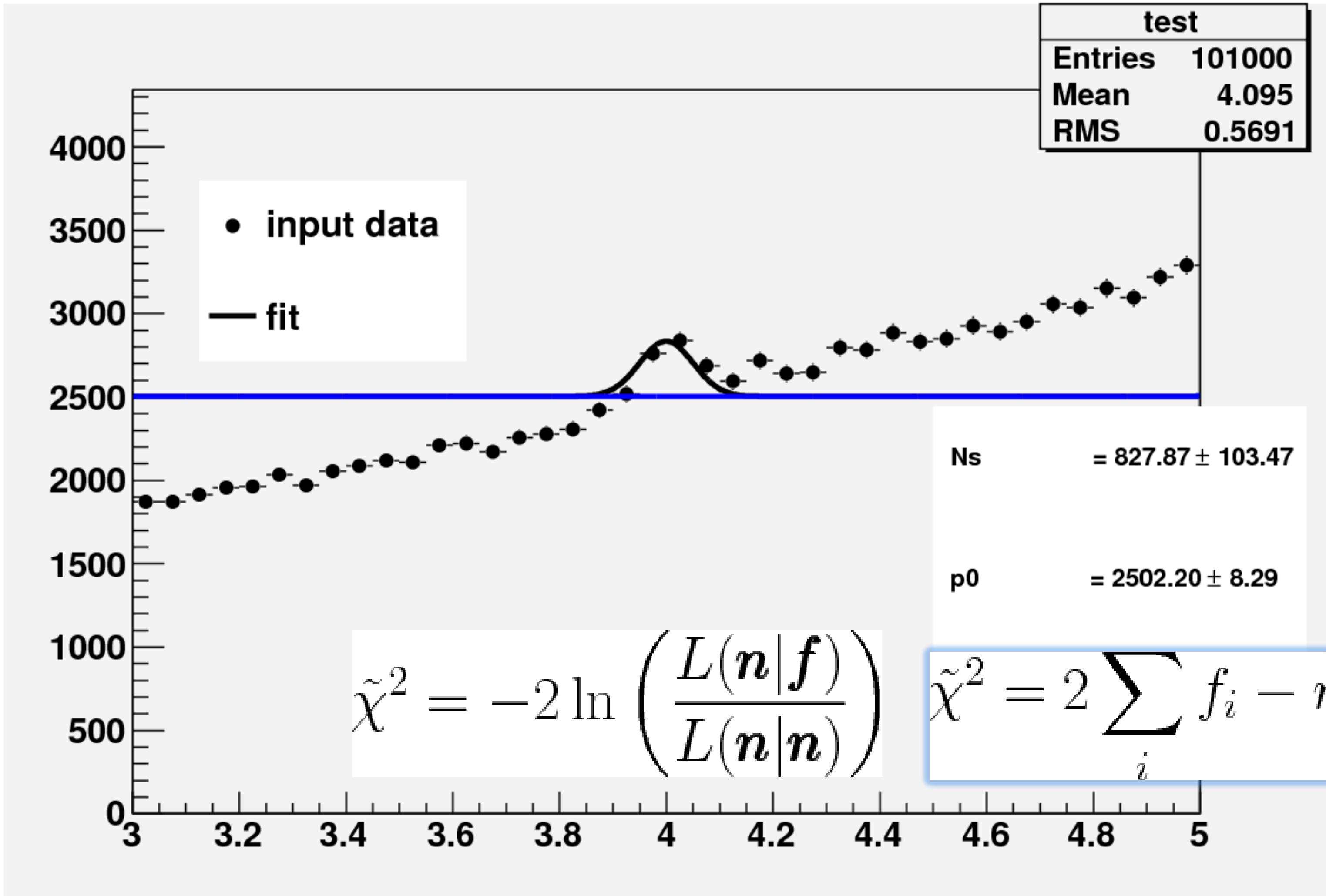


H_1: need more background pars

→ LR = $L(H_1)/L(H_0)$ provides optimal test (Neyman-Pearson)

→ Choose optimal parametrisation based on GOF+LR tests

#of background fit pars – How many are needed?



$$\tilde{\chi}^2 = -2\ln\left(\frac{L(\boldsymbol{n}|\boldsymbol{f})}{L(\boldsymbol{n}|\boldsymbol{n})}\right)$$

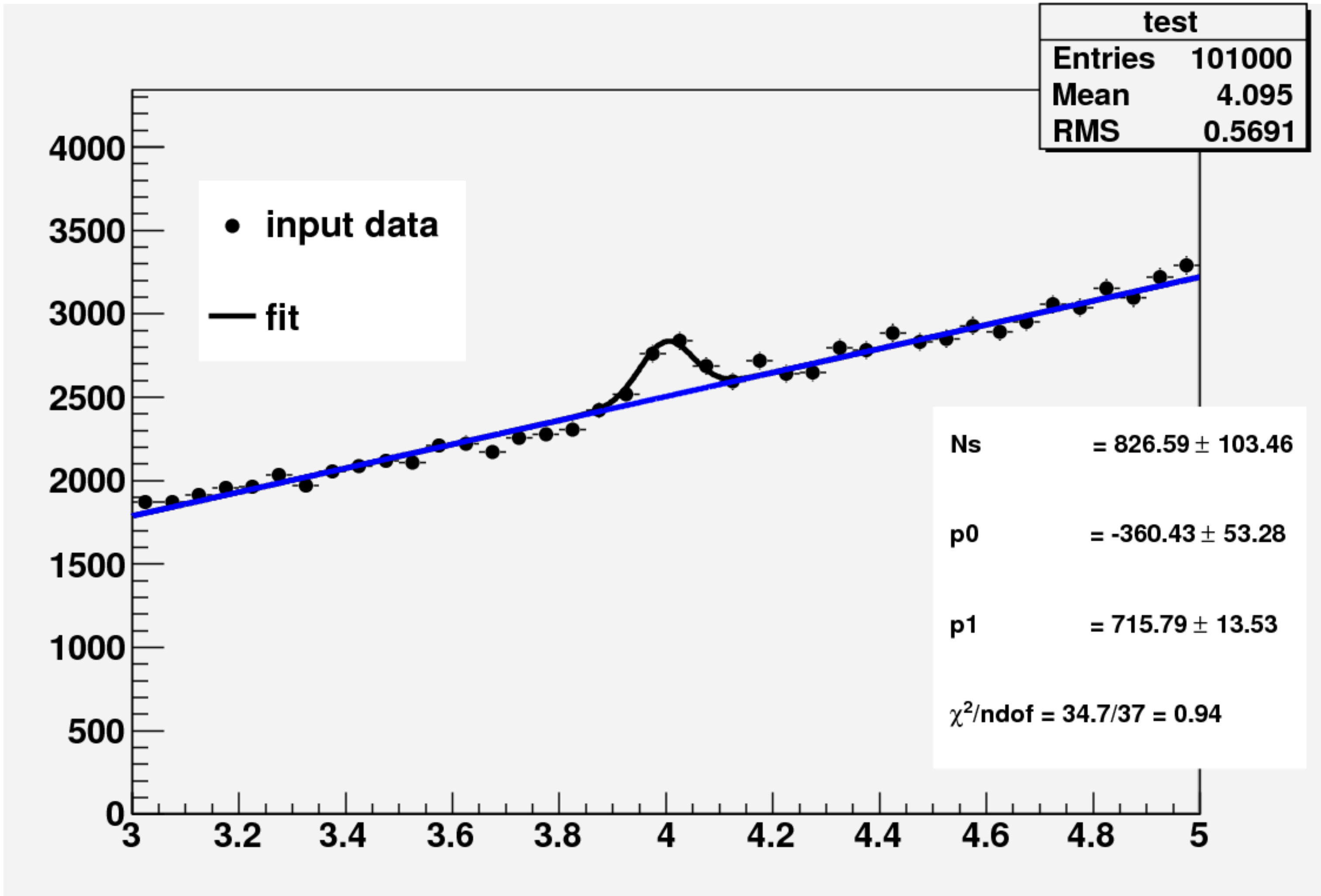
$$\tilde{\chi}^2 = 2\sum_i f_i - n_i + n_i \ln(n_i/f_i)$$

Fit function:
gauss+p0

$\tilde{\chi}^2$ = 2880

→ Very poor fit: TMath::Prob(2880,38) = 0.

#of background fit pars



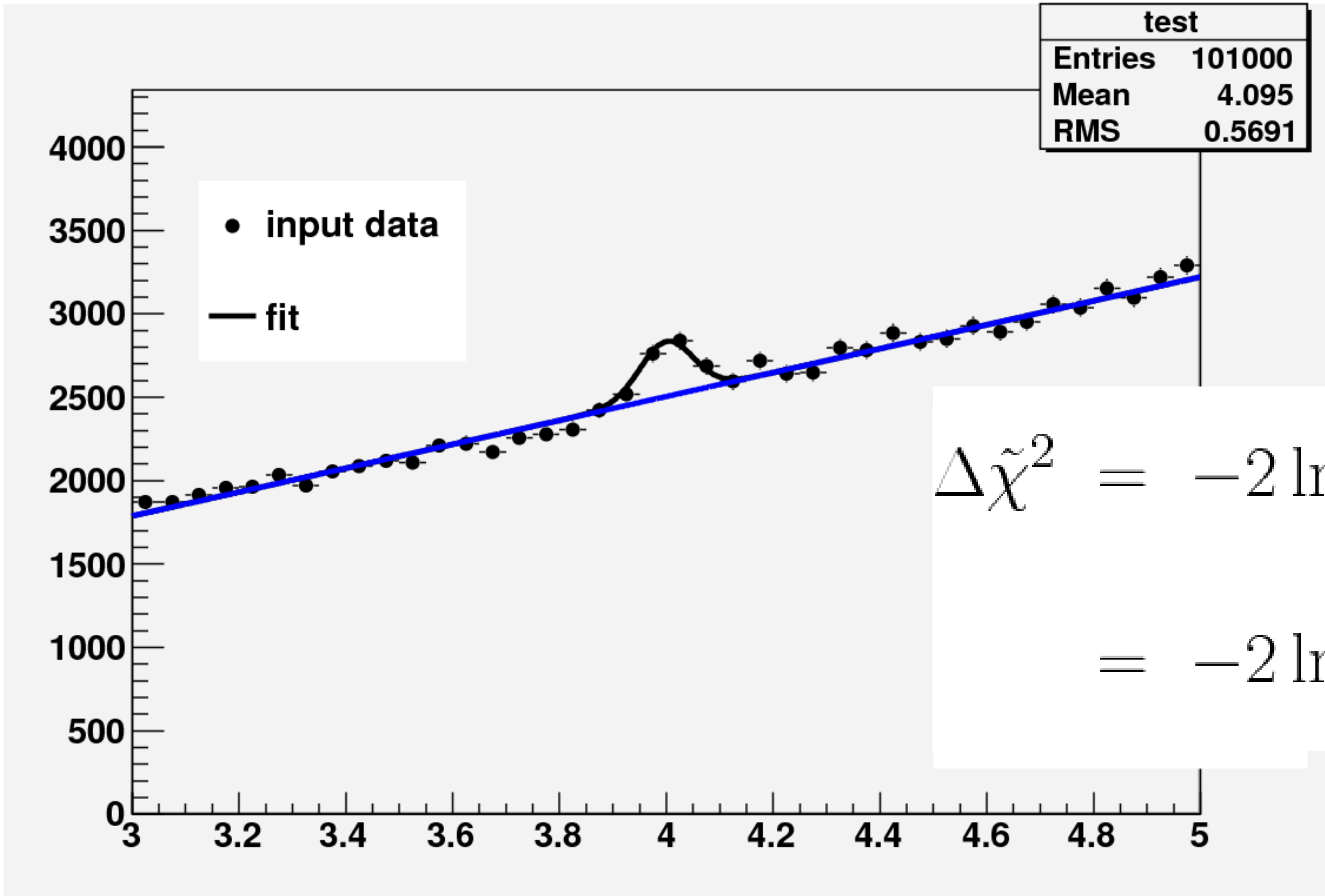
Fit function: gauss+p1

$\tilde{\chi}^2 = 34.7$

→ Reasonable $\tilde{\chi}^2$ TMath::Prob(34.7,37) = 0.58

Should we stop here?

#of background fit pars



Fit function: gauss+p1

$\tilde{\chi}^2$ = 34.7

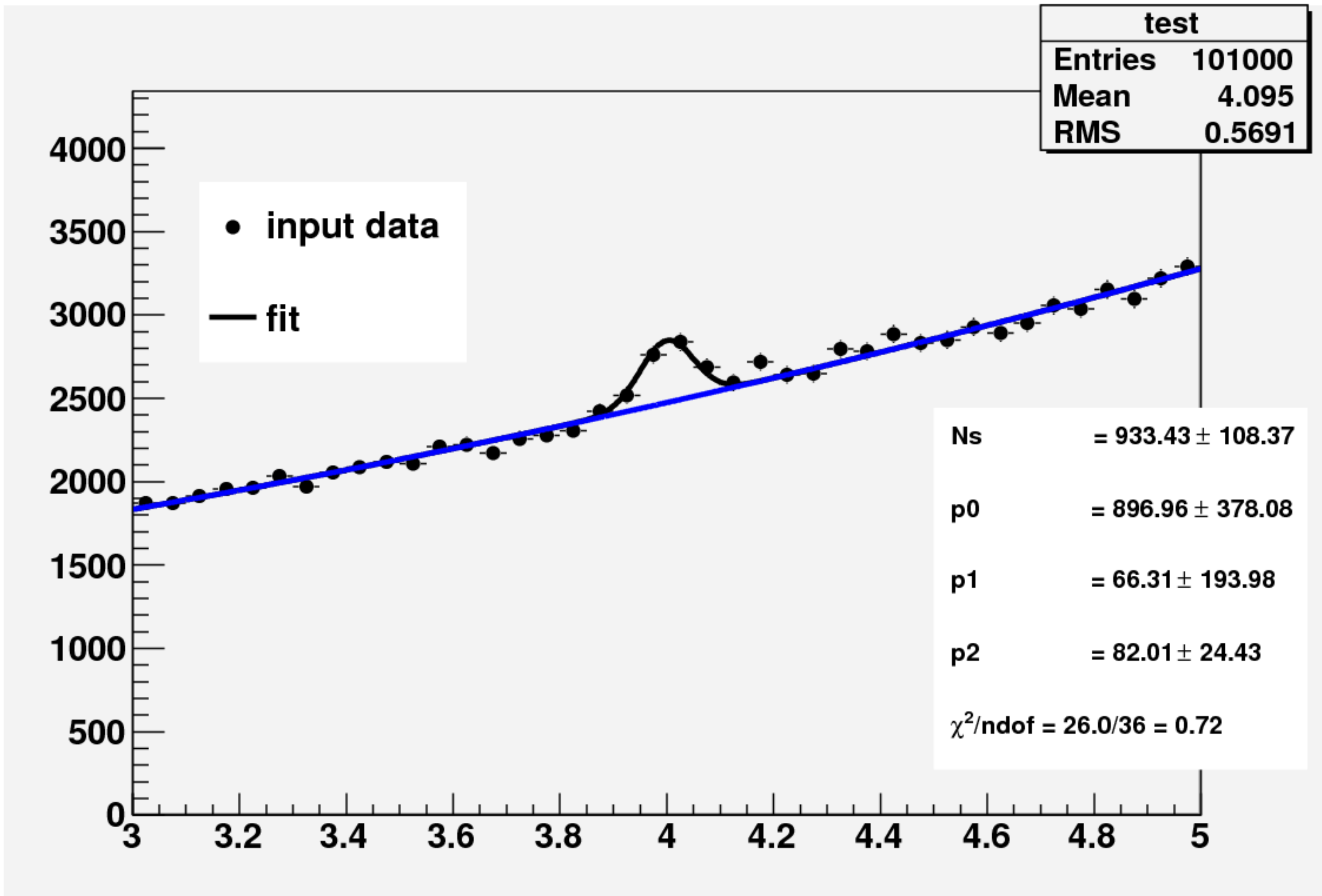
$$\Delta\tilde{\chi}^2 = -2\ln\left(\frac{L(H_1)}{L(H_0)}\right) = -2\ln\left(\frac{L(g+p_1)}{L(g+p_0)}\right)$$

= -2845.3

→ Reasonable $\tilde{\chi}^2$ TMath::Prob(34.7,37) = 0.58

Should we stop here?

#of background fit pars



Fit function: gauss+p2

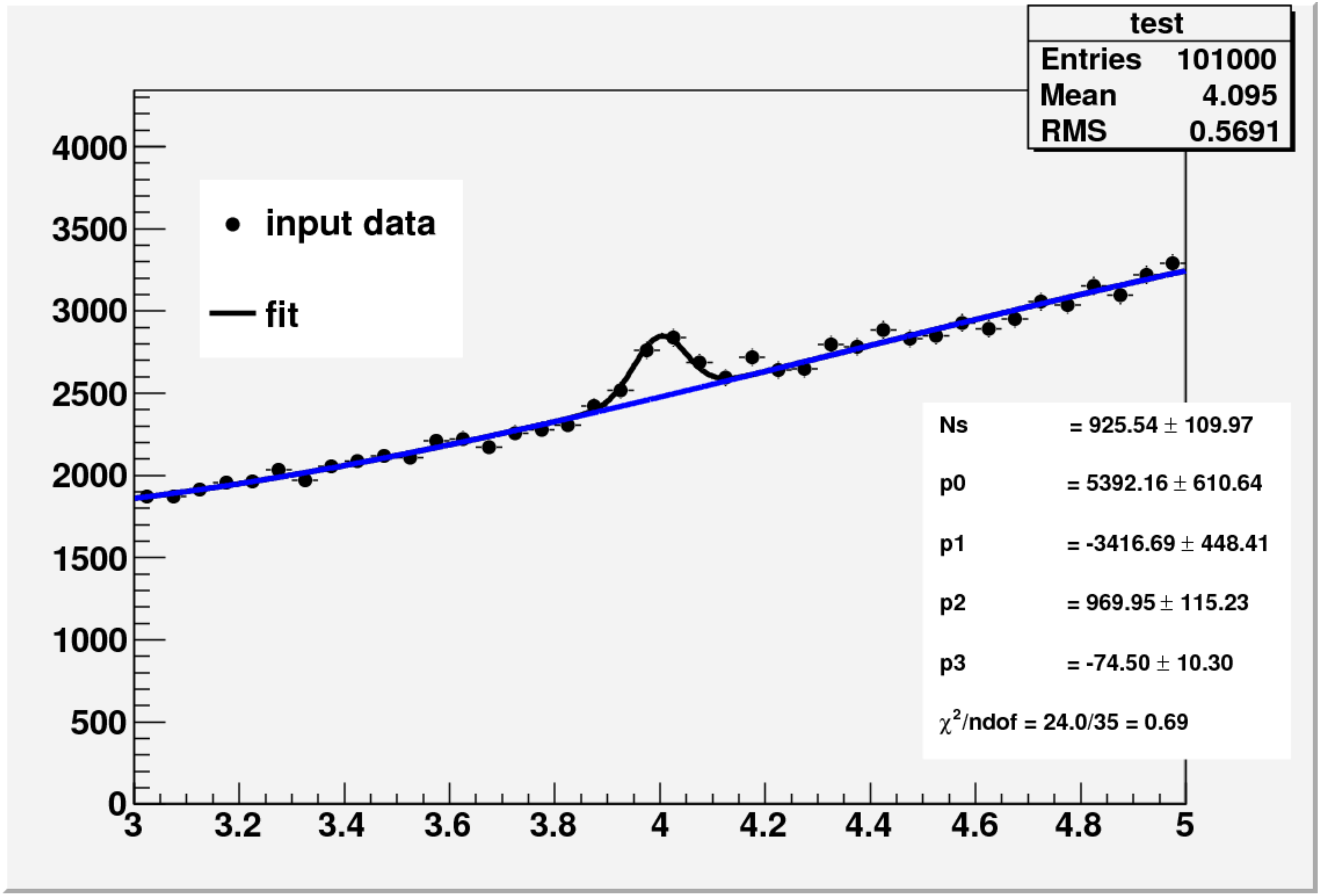
$\tilde{\chi}^2 = 26.0$

$\Delta\tilde{\chi}^2 = -8.7$

→ Reasonable $\tilde{\chi}^2$ TMath::Prob(26.0,36) = 0.89

Should we stop here?

#of background fit pars



Fit function: gauss+p3

$$\tilde{\chi}^2 = 24.0$$

$$\Delta\tilde{\chi}^2 = -2.0$$

→ Reasonable $\tilde{\chi}^2$ TMath::Prob(24.0,35) = 0.92

Lets stop here ☺

#of background fit pars

See also: www.pd.infn.it/~dorigo/rolkelrvsftest.pdf

g+p0 $\tilde{\chi}^2$ = 2880

g+p1 $\tilde{\chi}^2$ = 34.7 $\Delta\tilde{\chi}^2$ = -2845.3

g+p2 $\tilde{\chi}^2$ = 26.0 $\Delta\tilde{\chi}^2$ = -8.7

g+p3 $\tilde{\chi}^2$ = 24.0 $\Delta\tilde{\chi}^2$ = -2.0

When to stop adding further parameters?

H_0 Hypo: Additional parameter not needed (= zero)

If H_0 correct then according to Wilks' theorem: $-\Delta\tilde{\chi}^2$ should follow χ^2 function with ndf=1 (in asymptotic regime of large n)

TMath::Prob(8.7,1) = 0.003 → g+p2 favoured over g+p1
Tmath::Prob(2.0,1) = 0.15 → g+p3 not favoured over g+p2

Gaussian z-scores: √8.7~3 and √2=1.4

Wilks' theorem

Samuel S. Wilks
(1906-1964)

- H_0: Additional parameters (as predicted by H_1) not needed (= zero)
- If H_0 correct then according to Wilks' theorem:
  $-\Delta\tilde{\chi}^2$ = -2ln[L(H_1)/L(H_0)] should follow for n→∞
  χ^2 function with ndf = #added parameters
  (e.g. ndf = 3 for p2 → p5)

Wilks' theorem only applies for nested hypotheses:
✓ H_0: 1st order polynomial → H_1: 2nd order polynomial
✘ H_0: 1st order polynomial → H_1: a·exp(bx+cx²)

3 Optimal background parametrisation - Summary

Stop adding parameters k → k+1 when

- $\mathrm{TMath::Prob}(\tilde{\chi}^2_k, ndf) > 5\%$
- $\mathrm{TMath::Prob}(\tilde{\chi}^2_{k+1} - \tilde{\chi}^2_k, 1) > 5\%$

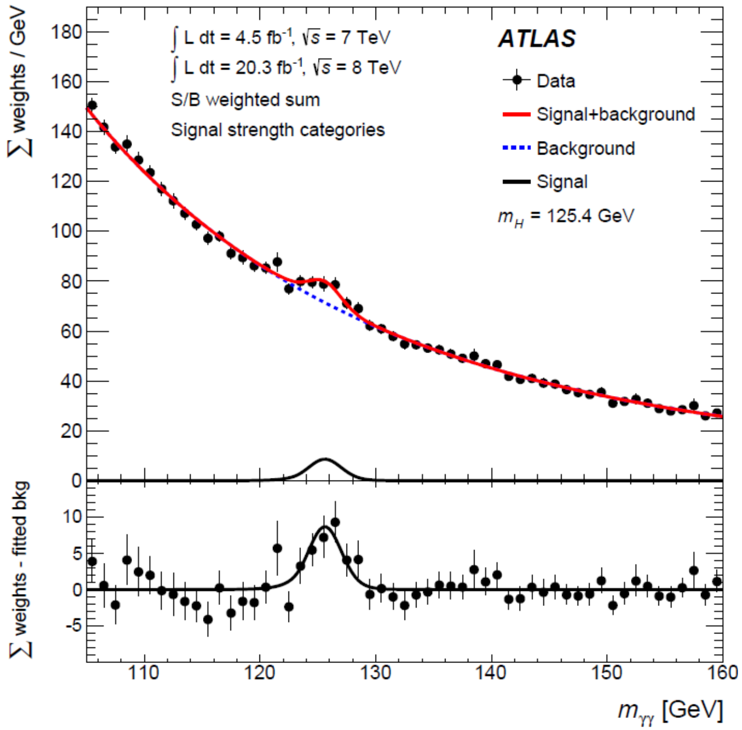
Equivalent $\tilde{\chi}^2_{k+1}$ vs $\tilde{\chi}^2_k$ test:

- Fisher F-test

What about background shape systematics?

→ Discuss next

4 Background shape systematics: intro+spurious signal



† arXiv:1408.7084
Phys. Rev. D90, 112015 (2014)

Conventional shape systematics:

- repeat fits with different functions (e.g. polynomials, exponential)
- changes on signal strength $\mu \rightarrow \Delta\,\mu_{sys}^{bgr}$

Spurious signal idea†: absorb systematics in fit function $f = \mu \cdot signal + \mu' \cdot signal + \text{bgr}$ with μ' = extra fit par. for spurious signal

constraint on μ'

$$\tilde{\chi}^2 = 2\Big[\sum_i f_i - n_i - n_i \ln\frac{f_i}{n_i}\Big] + \left[\frac{\mu'}{\Delta\mu_{sys}^{bgr}}\right]^2$$

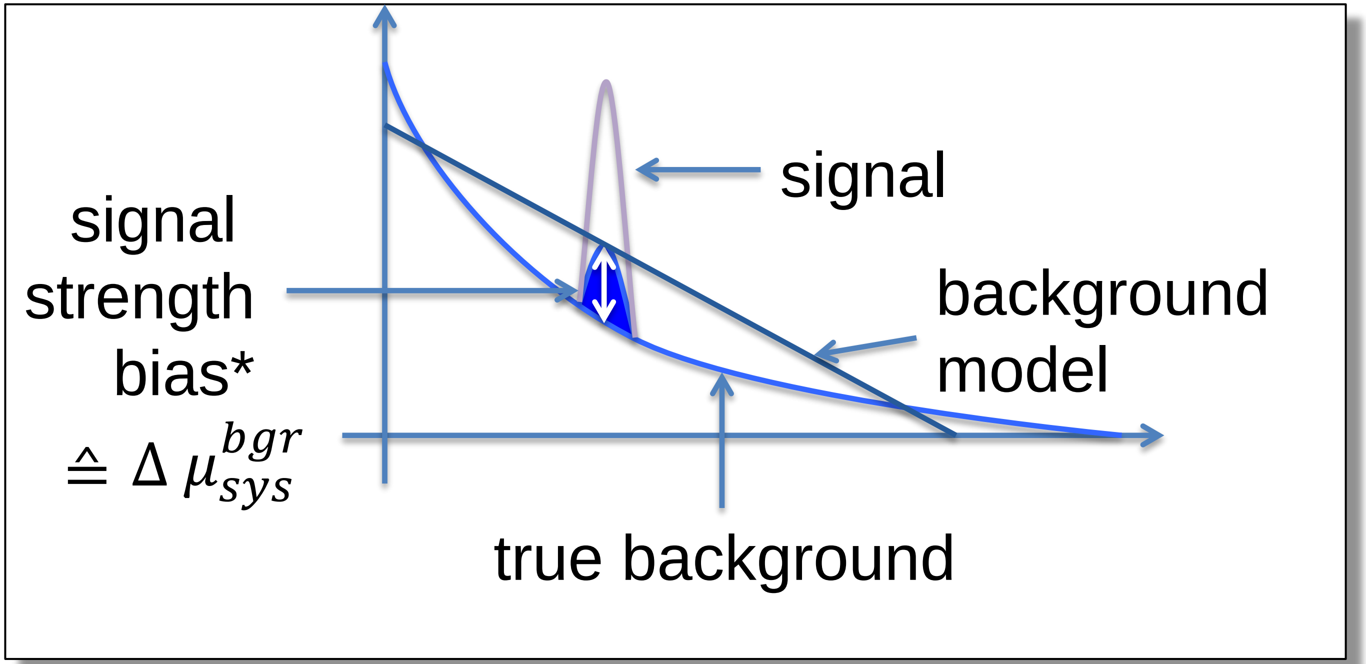
→ effective way of treating systematics as statistical uncertainty
→ Perhaps looks a bit ‘ugly’?

Spurious signal in practice

arXiv:1408.7084
Phys. Rev. D90, 112015 (2014)

Determine $\Delta\,\mu_{sys}^{bgr}$ from MC background toys

- generate with one function → fit with another function + signal



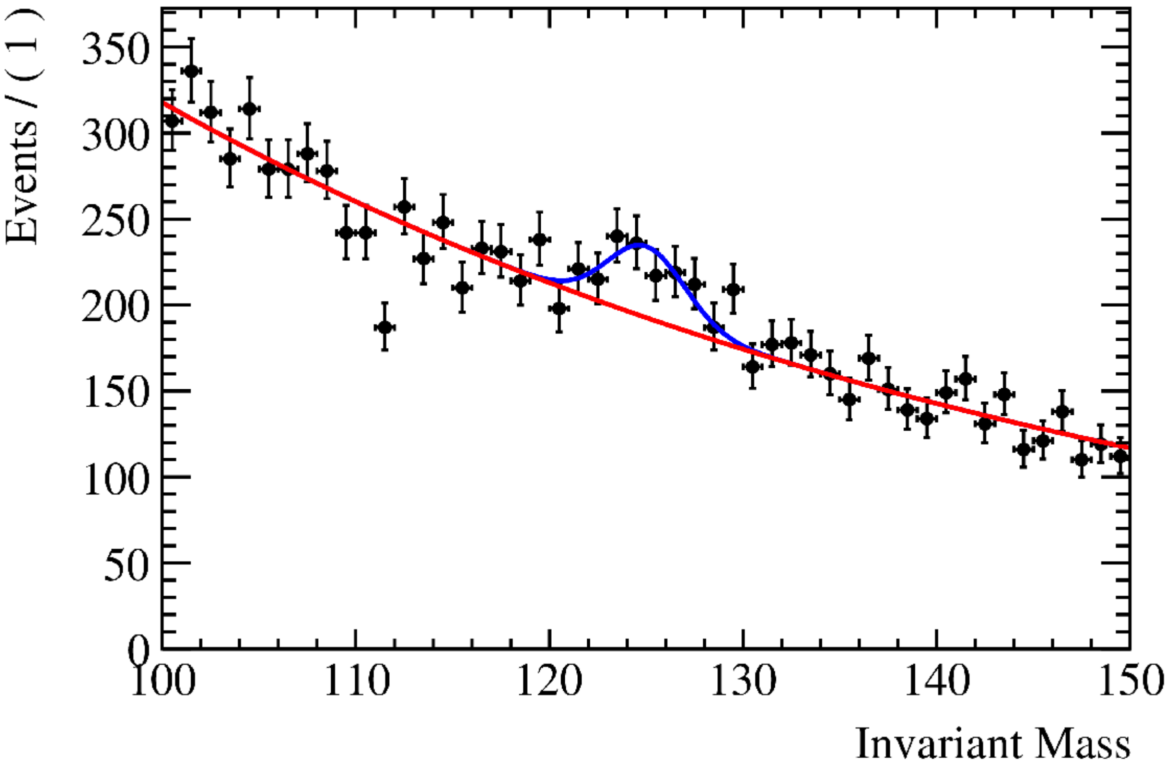
*too little signal measured in this case

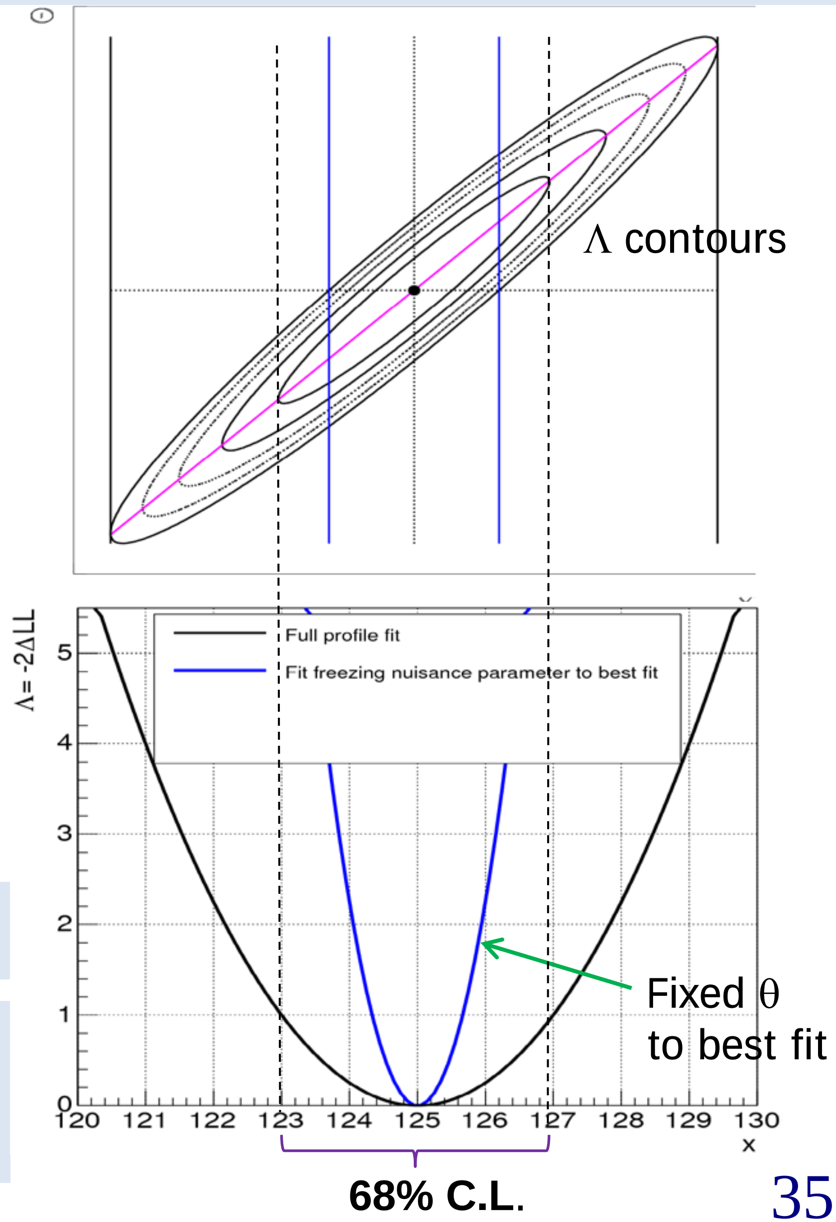
→ Lets look now at another method used in CMS

Discrete profiling method

P. Dauncey, M. Kenzie, N. Wardle and G. Davies
JINST 10 P04015 [arXiv:1408.6865]

Standard Profile likelihood:
Scan Λ =-2Δln(L) vs x; profiling θ



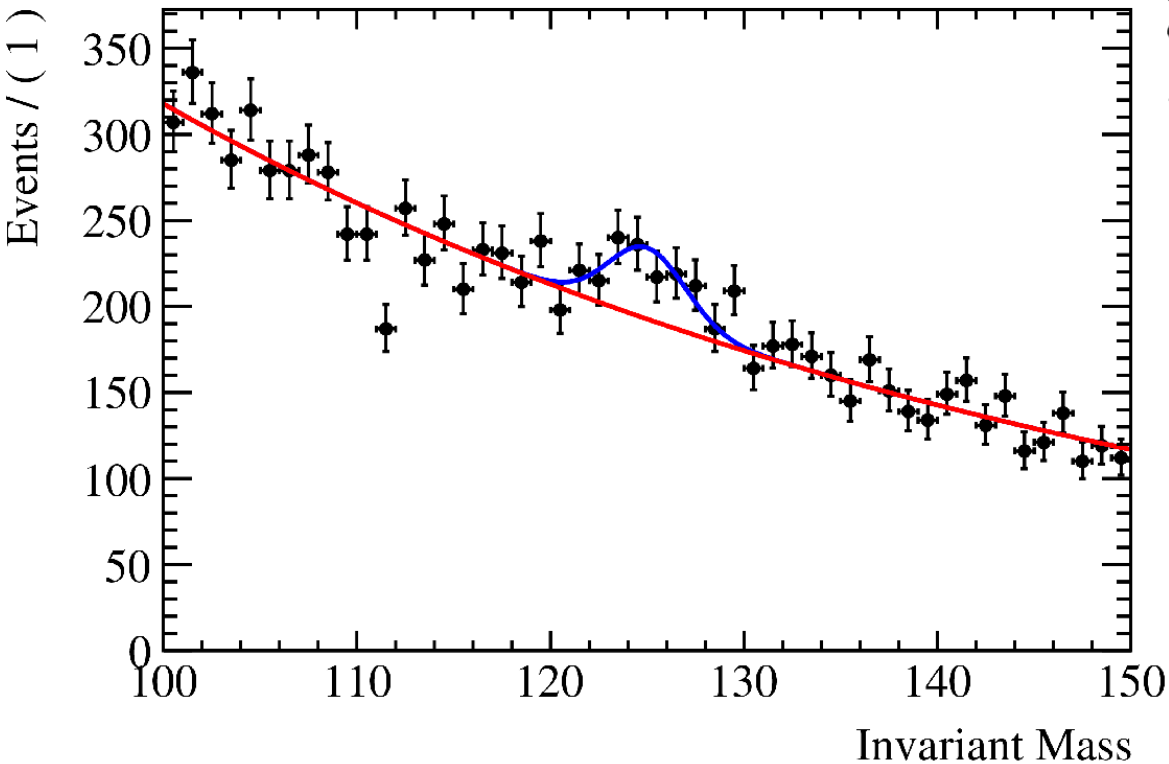


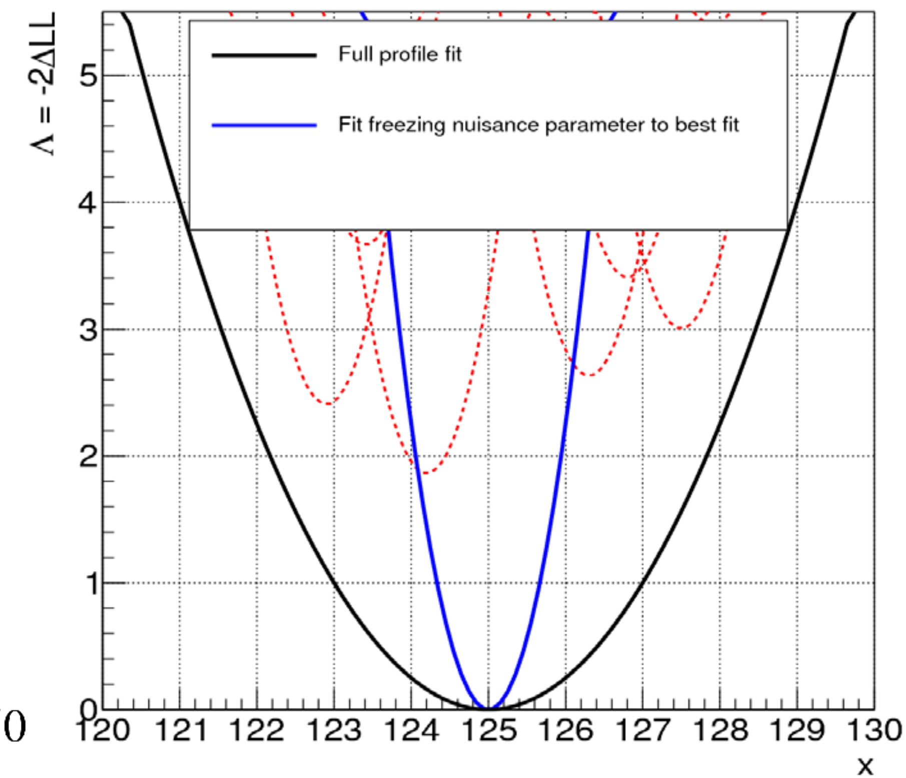
Fit gaussian signal + exponential background

Parameters: x = signal mass; θ = background exponential slope (nuisance)

Play around with nuisance parameter

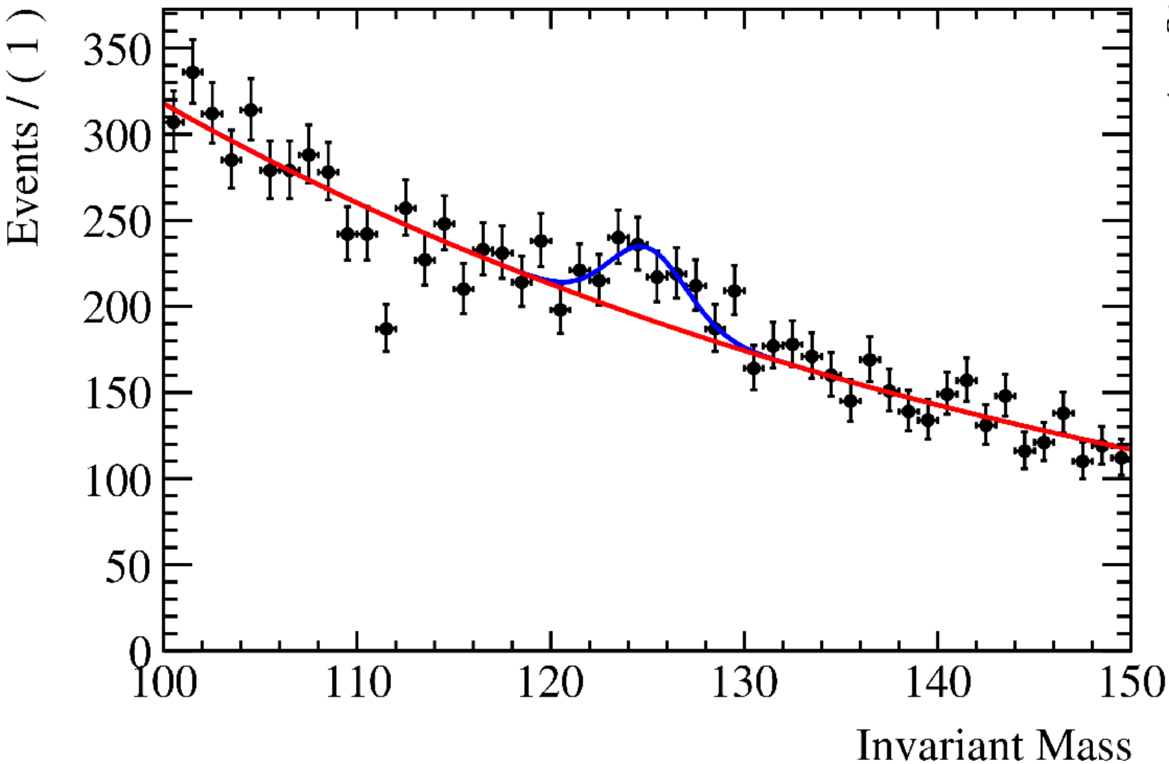
Fix θ to a few random values → red dashed lines

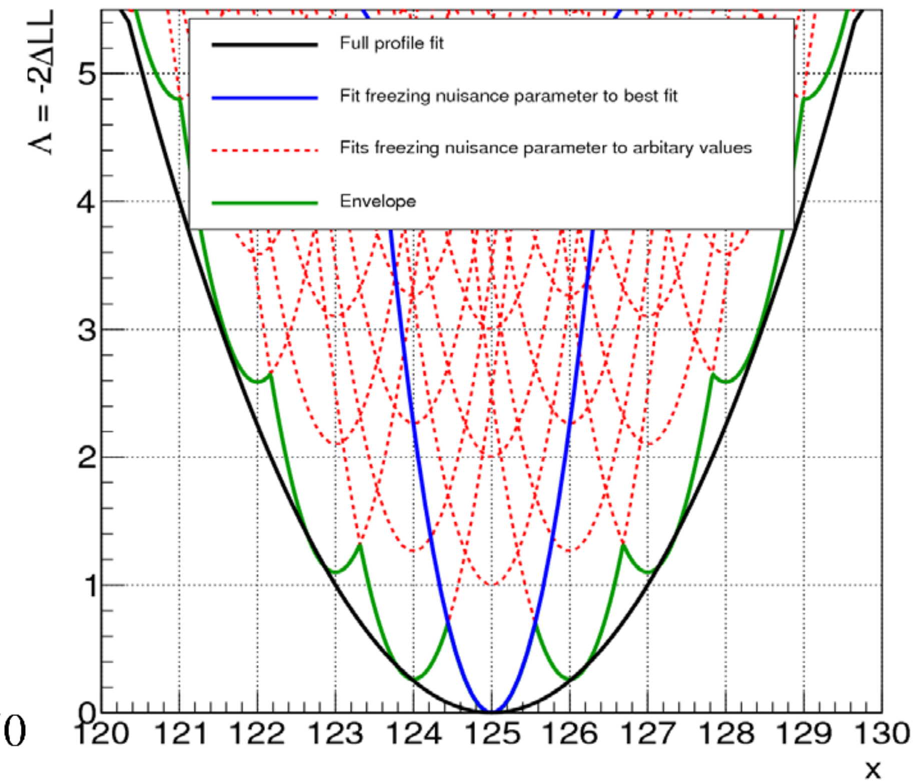




Play around with nuisance parameter

Fix θ to **many** random values
→ more red dashed lines

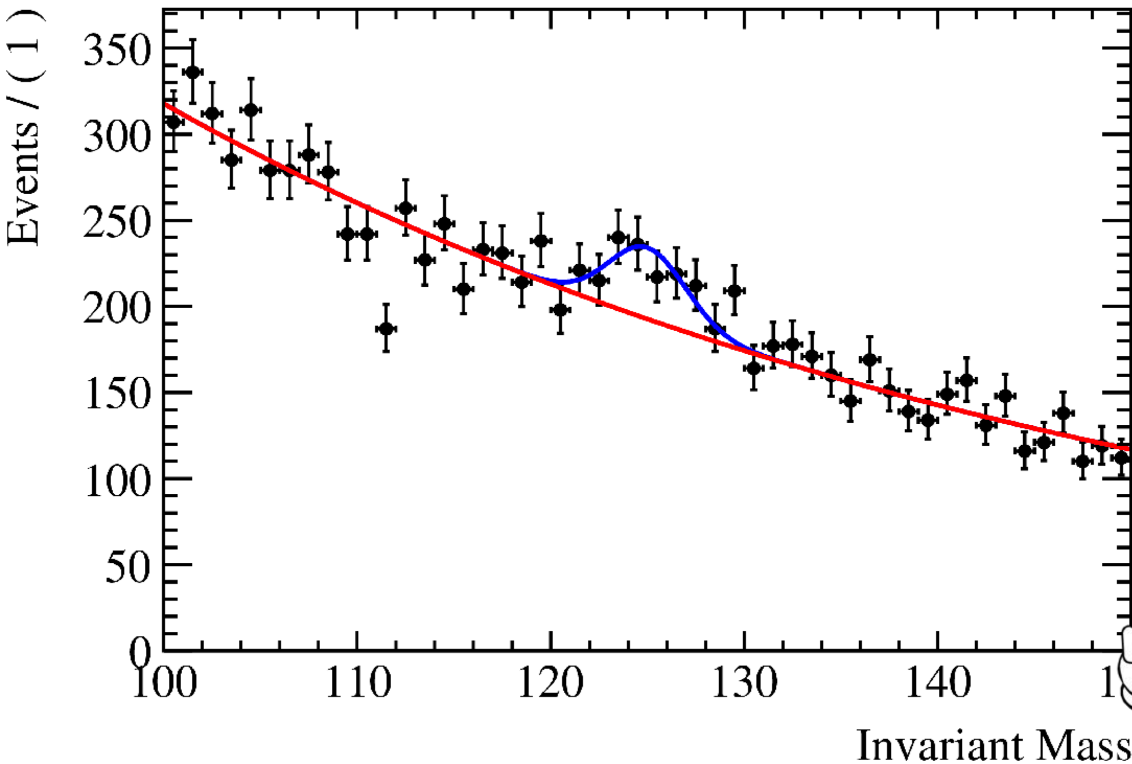


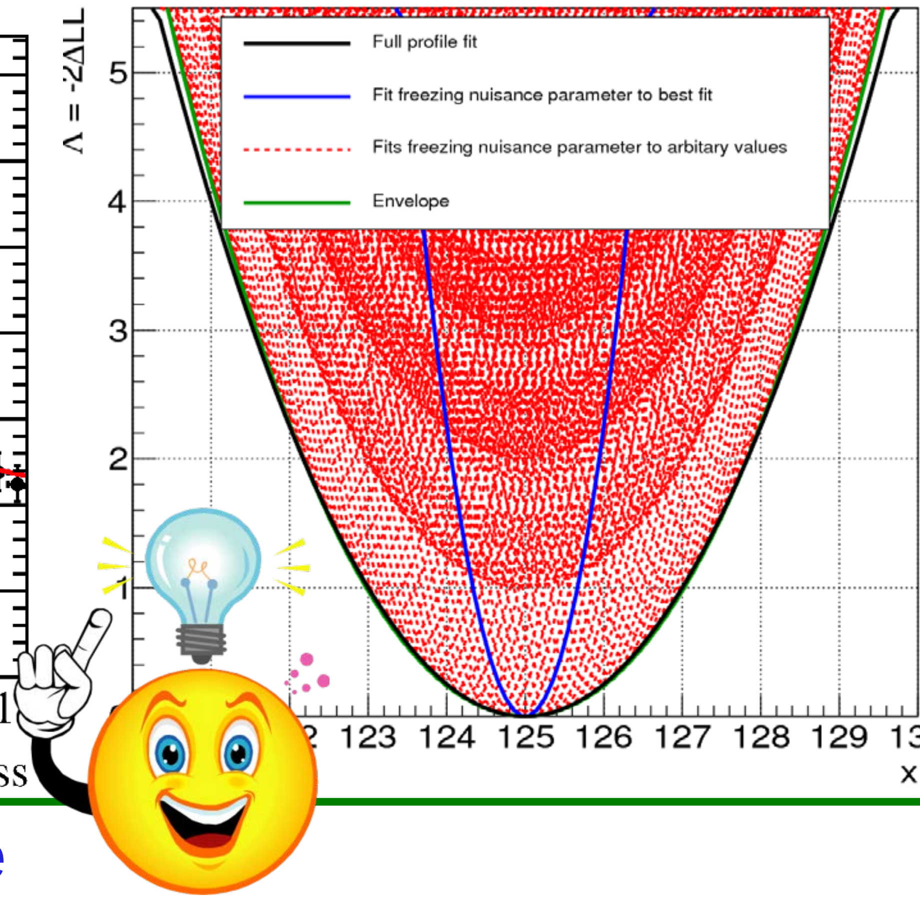


Draw minimum envelope (green line)
→ begin to recover original curve

Play around with nuisance parameter

Fix θ to **huge number** of random values → more red dashed lines





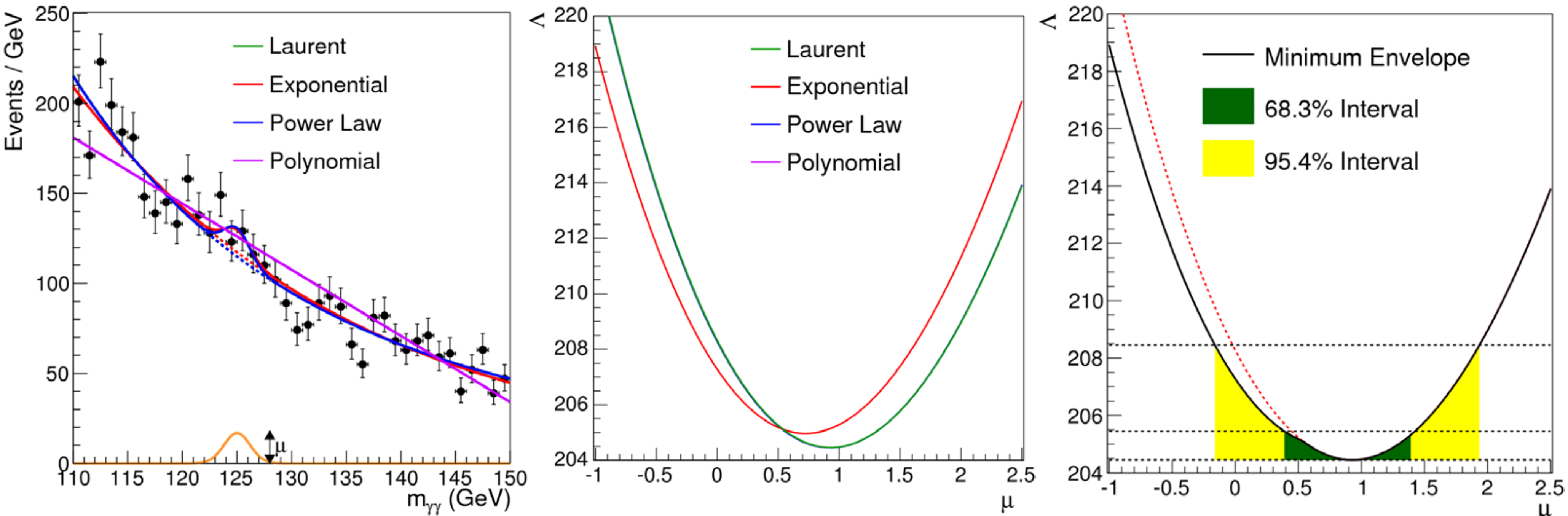
Minimum envelope = original curve
→ One can mix discrete nuisance parameters with continuous ones

A more realistic example

Fit μ ·signal-model + background

$$\Lambda = 2[\sum_i f_i - n_i - n_i \ln \frac{f_i}{n_i}]$$ (Baker-Cousins $\tilde{\chi}^2$)

Test background functions with same #fit parameters



→ Minimum envelope provides:

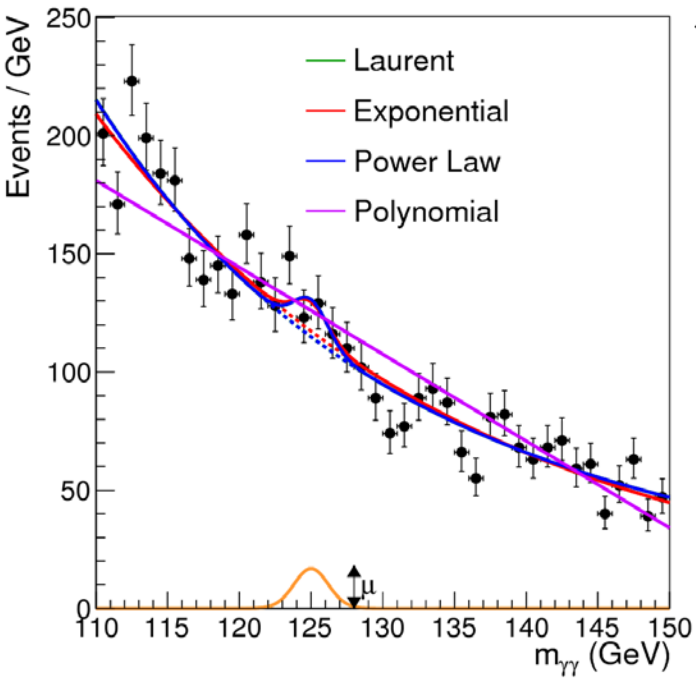
- best fit value $\hat{\mu}$
- Confidence interval ($\Delta\Lambda \leq 1$)
- Systematic from background model choice

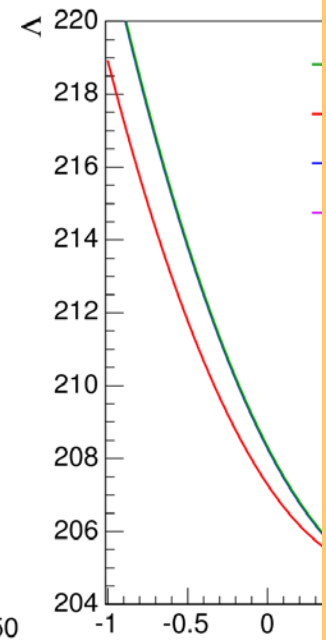
A more realistic example

Fit μ ·signal-model + background

$$\Lambda = 2[\sum_i f_i - n_i - n_i \ln \frac{f_i}{n_i}]$$ (Baker-Cousins $\tilde{\chi}^2$)

Test background functions with sar



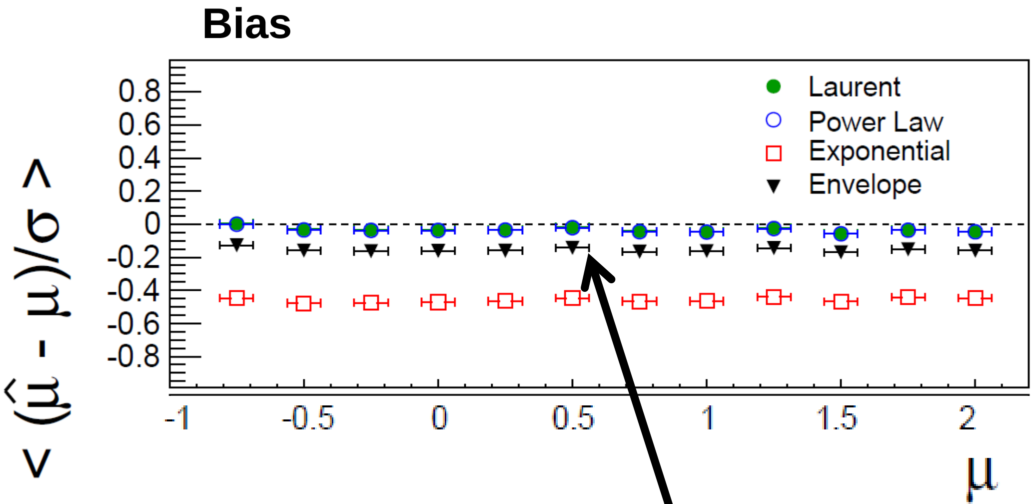


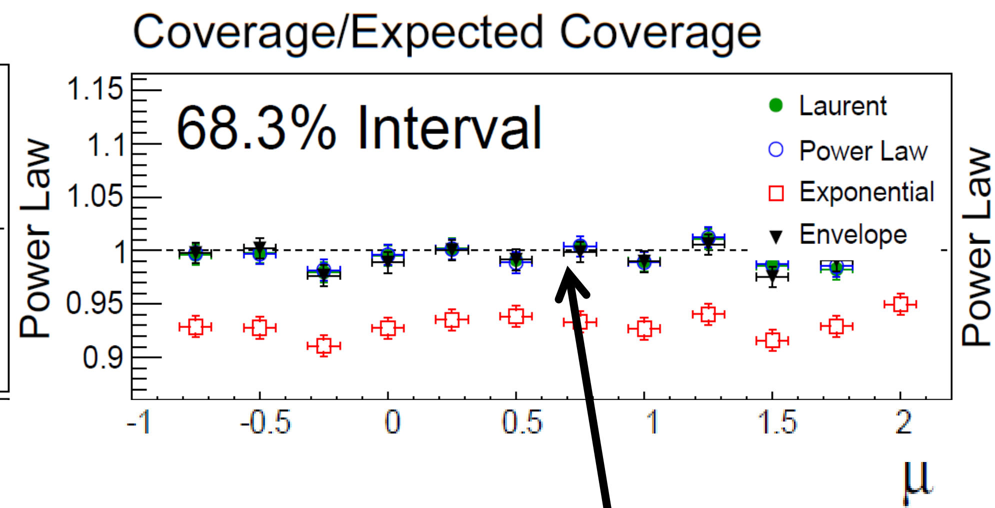
→ Minimum envelope provides:

- best fit value $\hat{\mu}$
- Confidence interval ($\Delta\Lambda \leq 1$)
- Systematic from background model choice

Bias and Coverage

Generate toy MC from various background hypotheses and study bias and coverage† of fitted $\hat{\mu}$ as function of generated true μ

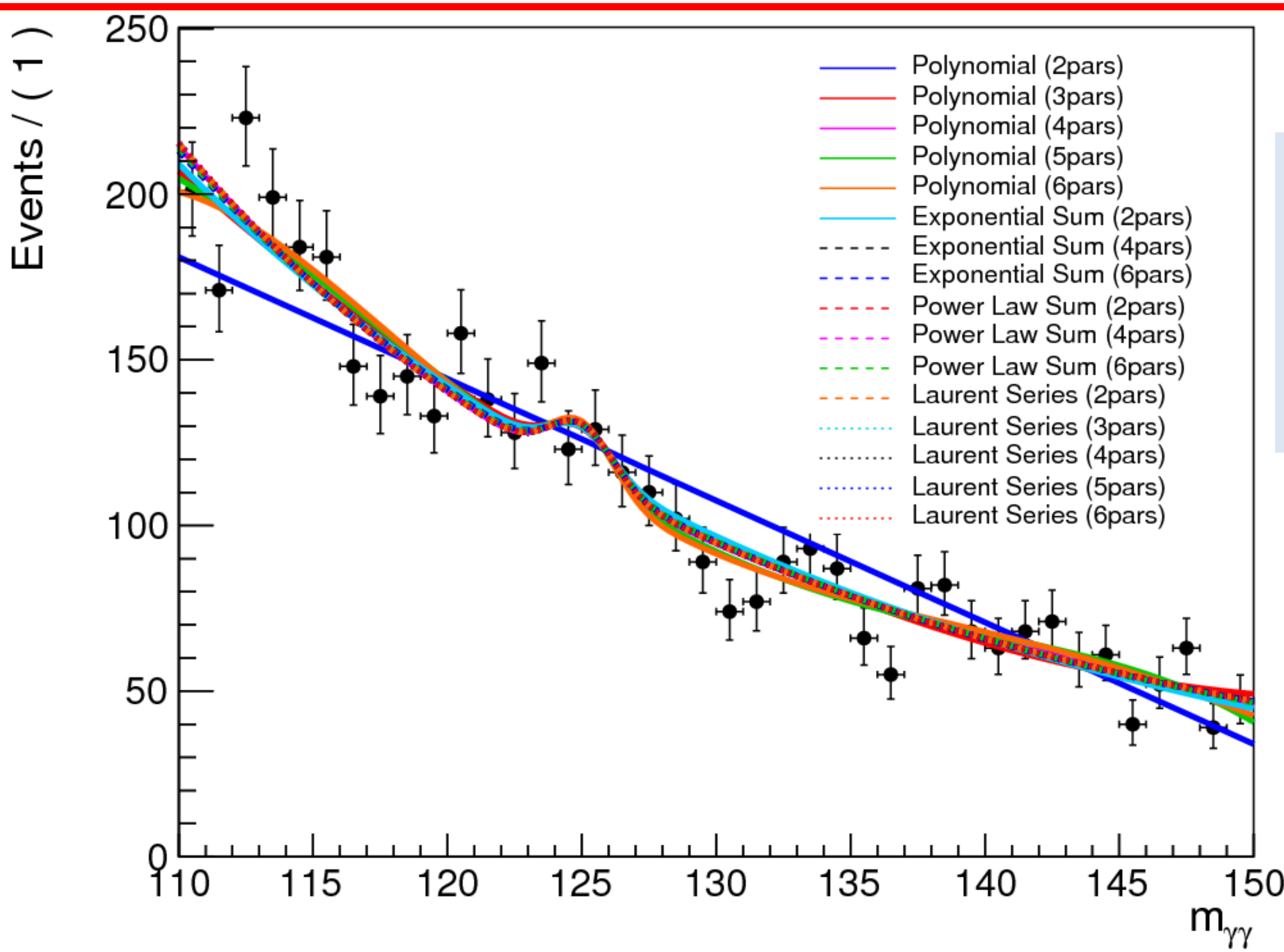




→ Minimum envelope provides small bias and good coverage

†*Coverage: correct coverage means that in 68.3% of repeated experiments the true parameter value is contained within the estimated ±1 sigma region.*

Fits with background functions of different orders



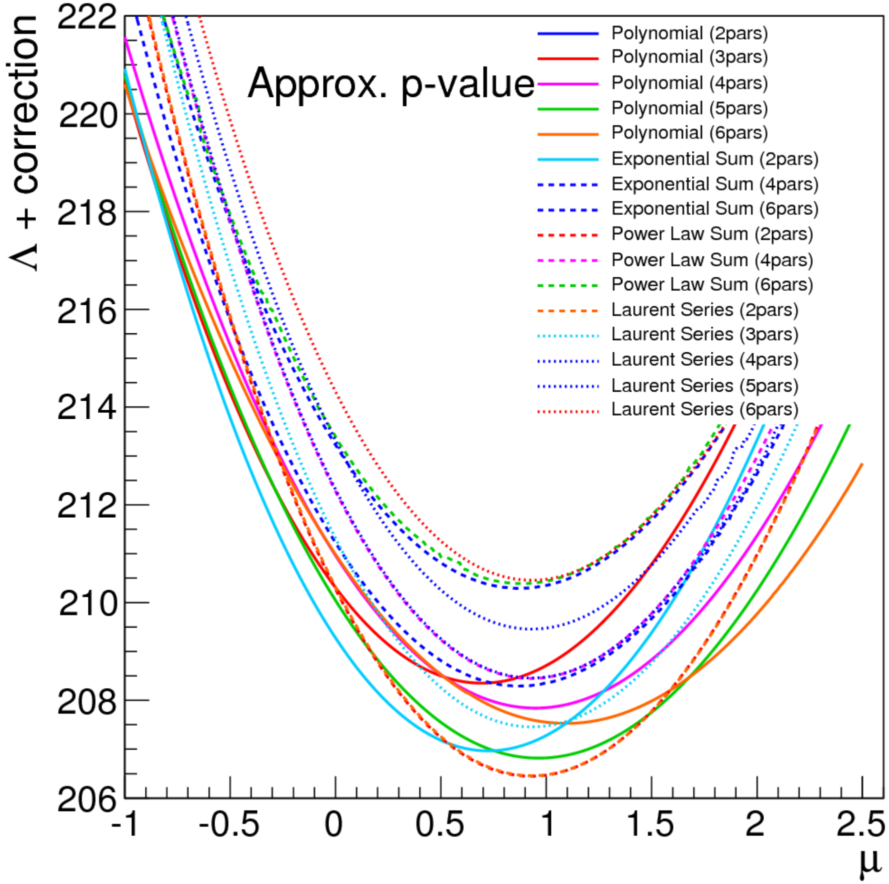
Minimum Λ envelope: functions with large #fit parameters (npars) yield lower Λ

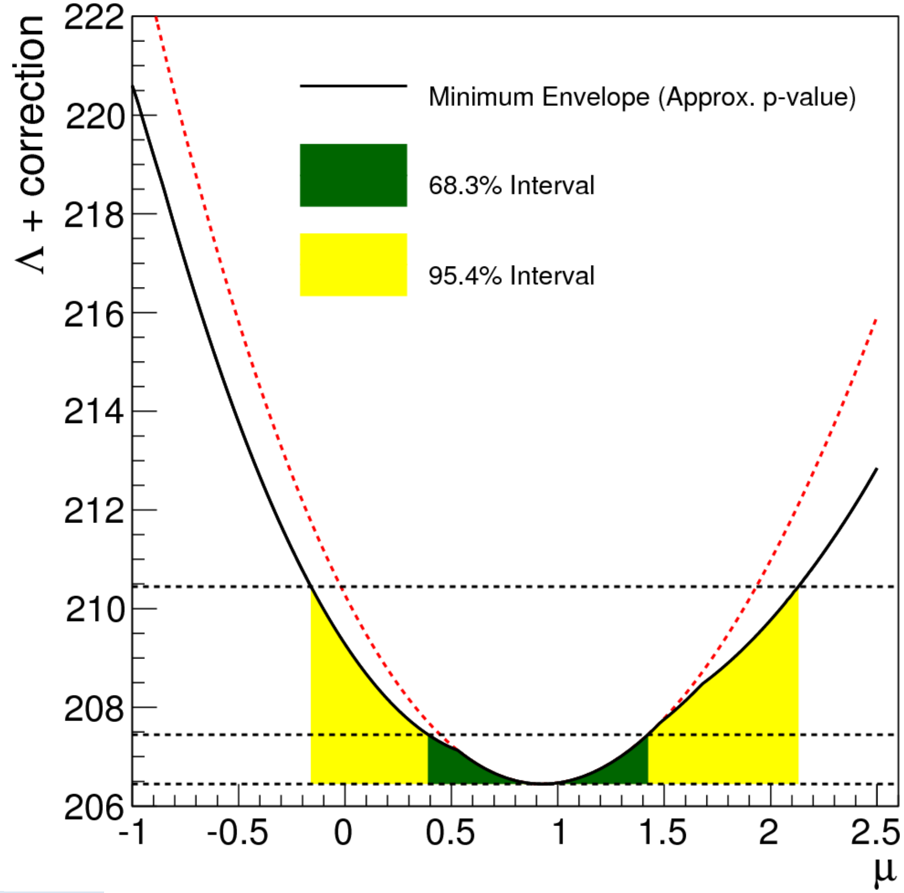
→ need to correct Λ for different npars

Λ = -2ln(L) + c npars

c=1 ≙ “approximate p-value correction”

Λ scans and minimum envelope Λ = -2ln(L) + c npars; **c=1**



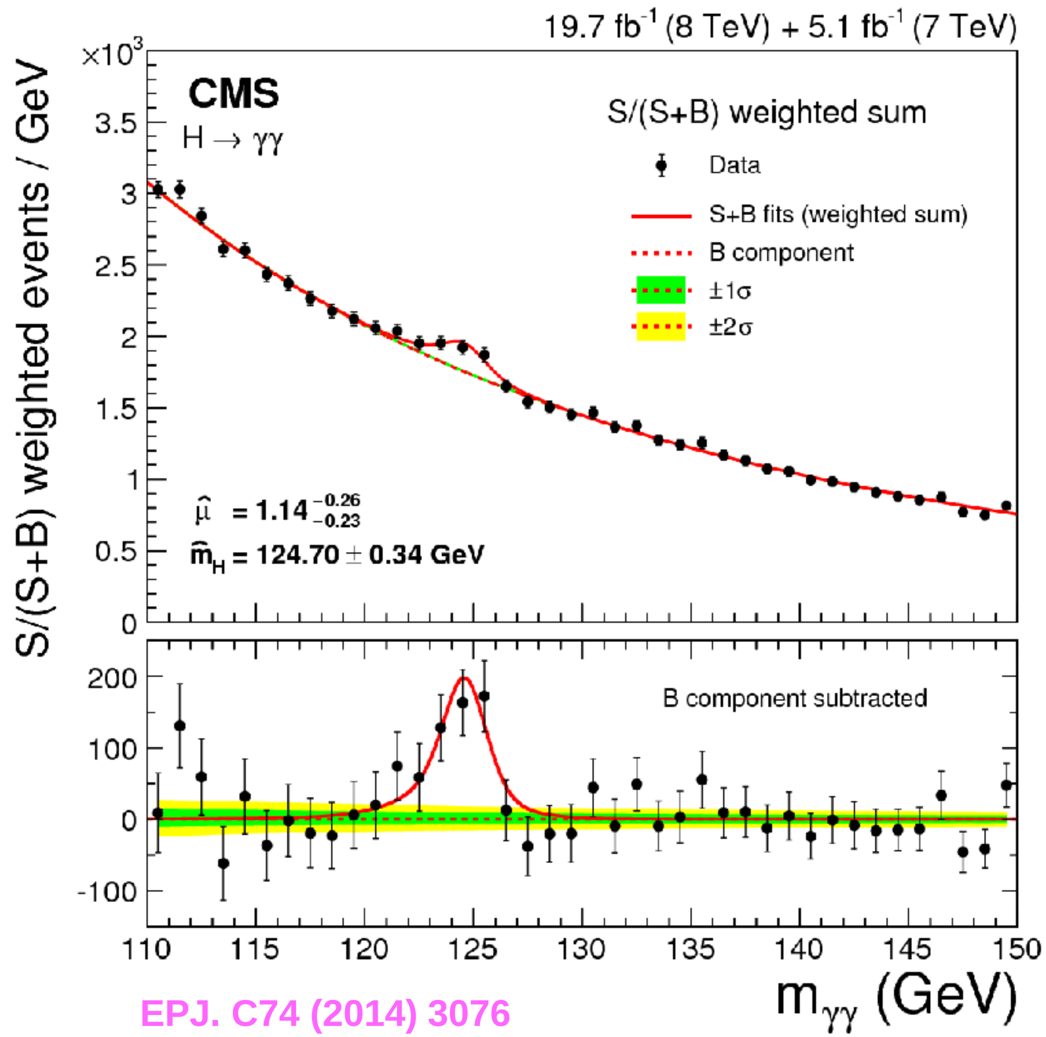


Best fit: 2 parameter power law

Choice of c:
- Large, e.g. 5 →prefer lower order functions →potential biases
- Small, e.g. 0.1 →prefer higher order functions →blow up σ_{stat}

Summary of Discrete Profile Likelihood method

- Method shows good coverage in toy experiments → perform toy experiments for your specific analyses
- Choices (open questions):
  - function models to include?
  - c term
- Method has been used e.g. in CMS H → $\gamma\gamma$ analysis



EPJ. C74 (2014) 3076

Summary

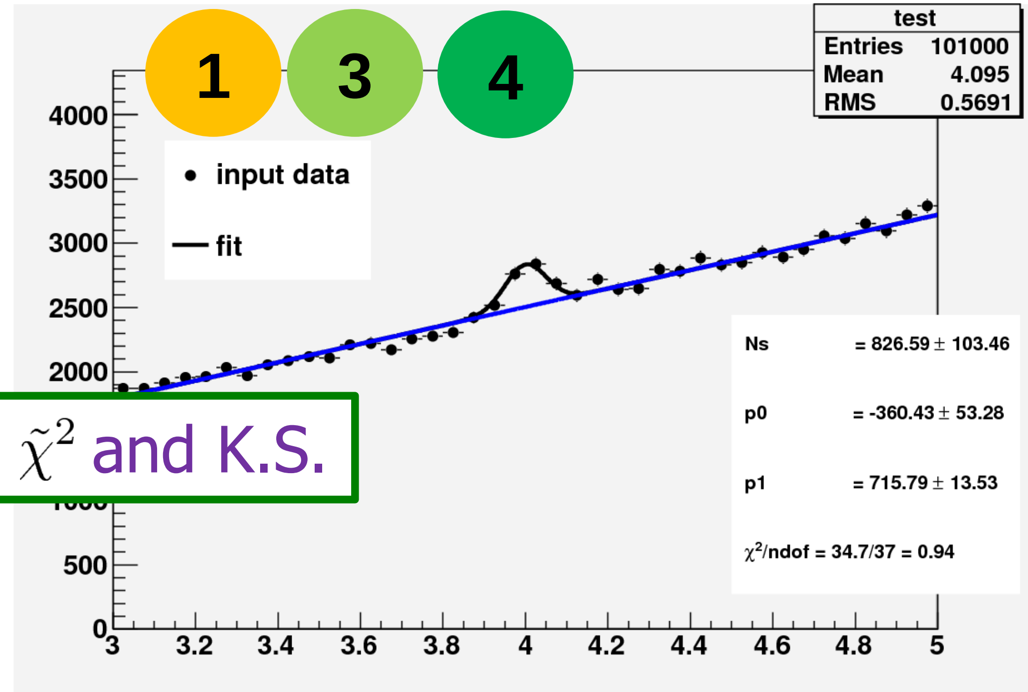
Goodness-of-Fit tests for

1 Checking data modelling

→ Perform ≥ 2 tests, e.g. $\tilde{\chi}^2$ and K.S.

2 Outlier rejection

→ χ^2 tests powerful tool



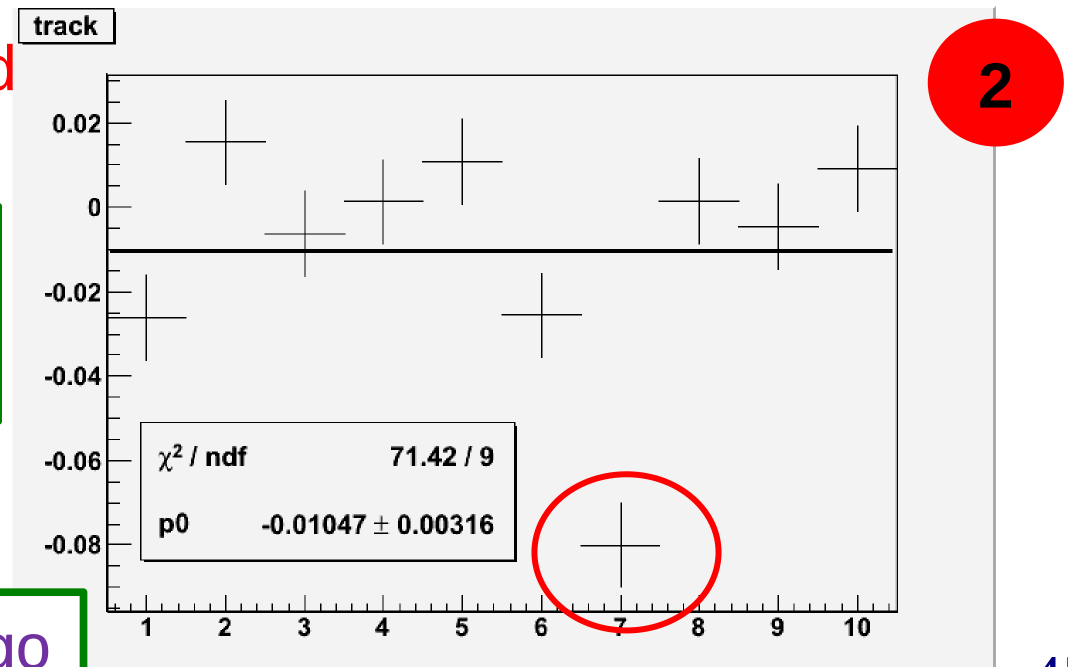
Likelihood ratios for Background

3 Optimal parametrisation

→ Stop k→k+1 when:
$\mathrm{TMath::Prob}(\tilde{\chi}^2_{k+1} - \tilde{\chi}^2_{k}, 1) > 5\%$ &
$\mathrm{TMath::Prob}(\tilde{\chi}^2_{k}, ndf) > 5\%$

4 Shape systematics (discrete profiling)

→ stat+syst error in one go



Final riddle – part I

Meggie has two children and the older one is a girl

→ What is the probability that the other child is also a girl?

Martin Gardner, "the two-child problem," Scientific American, 1959

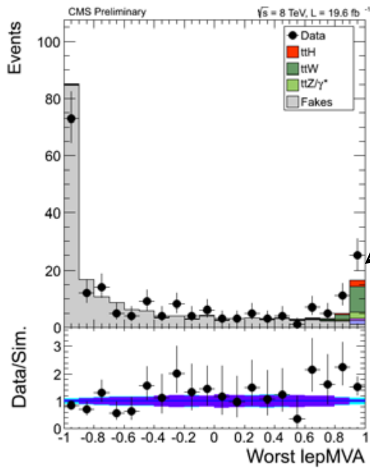
Final riddle – part II

From all families with two children, at least one of whom is a girl, a family is chosen at random → What is the probability that both children are girls?

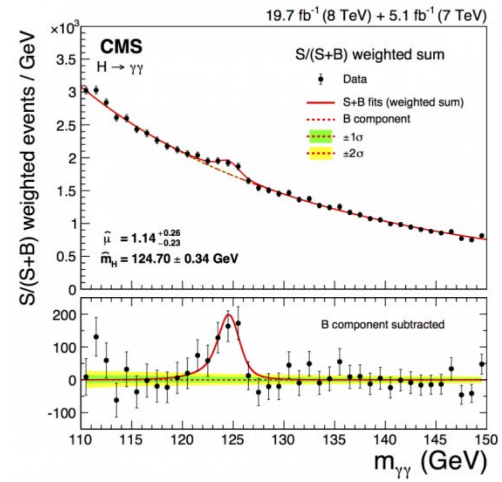
Backup slides

Statistical Data Analysis – typical tasks

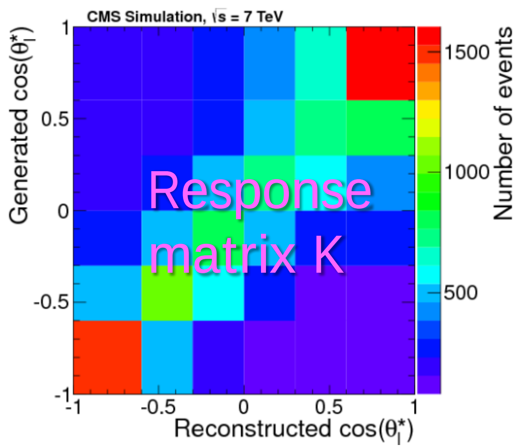
1. Optimal S/B separation



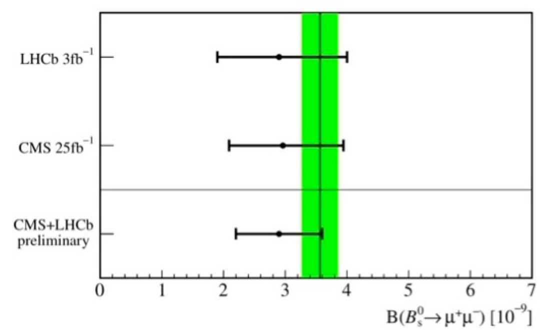
2. Signal searches, fits **of all kind of interesting physics parameters to data** and limits



3. Unfolding differential σ



5. Data combination



4. Systematic uncertainties