Reliably estimating the statistical significance of a new signal by exploiting GPUs

CHARM 2020 @ Mexico City (31.05-04.06 2021)

10th International Workshop on Charm Physics

VIRTUAL TALK

Alexis Pompili & Adriano Di Florio

UNIVERSITÀ degli STUDI di BARI & I.N.F.N. Sezione di Bari

INFN Bari

Outline

In particle physics we often have to deal with "signals" that highlight a discrepancy with what the current theoretical models predict. These signals can be already known or completely new.

In any case when a signal is observed, we need to assess the statistical significance, local or global.

Part 1 :

Part 2 :

Outline

- In particle physics we often have to deal with "signals" that highlight a discrepancy with what the current theoretical models predict. These signals can be already known or completely new.
  In any case when a signal is observed, we need to assess the statistical significance, local or global.

- **Part 1 : Local statistical significance of a *known* physics signal with GPUs**
  - `GooFit` **framework capabilities** thanks to **GPUs acceleration**
  - **Test use case: estimation of the local statistical significance of a known/to-be-confirmed signal**
  - **Exploring the applicability limits of Wilks' Theorem & the asymptotic behaviour of a likelihood ratio test statistics (Asymptotic Formula by Cowan *et al.*)**

- **Part 2 :**

Outline

- In particle physics we often have to deal with "signals" that highlight a discrepancy with what the current theoretical models predict. These signals can be already known or completely new.
  In any case when a signal is observed, we need to assess the statistical significance, local or global.

- **Part 1 : Local statistical significance of a *known* physics signal with GPUs**
  - `GooFit` framework capabilities thanks to GPUs acceleration
  - Test use case: estimation of the local statistical significance of a known/to-be-confirmed signal
  - Exploring the applicability limits of Wilks' Theorem & the asymptotic behaviour of a likelihood ratio test statistics (Asymptotic Formula by Cowan *et al.*)

- **Part 2 : Global statistical significance of a *new* physics signal with GPUs**
  - Role of Look-Elsewhere-Effect in a new signal search
  - Test use case: estimation of the global statistical significance of a new signal
  - Exploring the approximation of the Gross-Vitells method ("Trial Factors")

Scope of the work

The effort presented in this talk has somehow changed scope in time

- Part 1 was developed in 2015-2016
  - In 2014 we started using the *nVidia Tesla* GPUs newly acquired in the ReCas-Bari Data Center. **At that time the idea of using GPUs for HEP data analysis was rather pioneering** and started collaborating with **`GooFit`** ("**`Roofit`** for GPUs") developers (M. Sokoloff team in LHCb-Cincinnati). Our interest developed in the framework of our involvement in hadron spectroscopy searches in CMS.
  - Presentations at conferences: ACAT2016, ICHEP2016, Stat. Session of XIIQCHS(2016)

Scope of the work

The effort presented in this talk has somehow changed scope in time

- Part 1 was developed in 2015-2016
  - In 2014 we started using the *nVidia Tesla* GPUs newly acquired in the ReCas-Bari Data Center. **At that time the idea of using GPUs for HEP data analysis was rather pioneering** and started collaborating with **`GooFit`** ("**`Roofit`** for GPUs") developers (M. Sokoloff team in LHCb-Cincinnati). Our interest developed in the framework of our involvement in hadron spectroscopy searches in CMS.
  - Presentations at conferences: ACAT2016, ICHEP2016, Stat. Session of XIIQCHS(2016)
- Part 2 was developed in 2017-2018
  - Presentations at conferences: ACAT2017, Stat. Session of XIIIQCHS(2018)
-

Scope of the work

The effort presented in this talk has somehow changed scope in time

- Part 1 was developed in 2015-2016
  - In 2014 we started using the *nVidia Tesla* GPUs newly acquired in the ReCas-Bari Data Center. **At that time the idea of using GPUs for HEP data analysis was rather pioneering** and started collaborating with **`GooFit`** ("**`Roofit`** for GPUs") developers (M. Sokoloff team in LHCb-Cincinnati). Our interest developed in the framework of our involvement in hadron spectroscopy searches in CMS.
  - Presentations at conferences: ACAT2016, ICHEP2016, Stat. Session of XIIQCHS(2016)
- Part 2 was developed in 2017-2018
  - Presentations at conferences: ACAT2017, Stat. Session of XIIIQCHS(2018)
- Nowadays we use often **`GooFit`** for our more complicated Unbinned Maxiumum Likelihood fits; multi-process & multi-thread approaches are being introduced also in **`ROOT`**/**`Roofit`**/**`PyROOT`**.

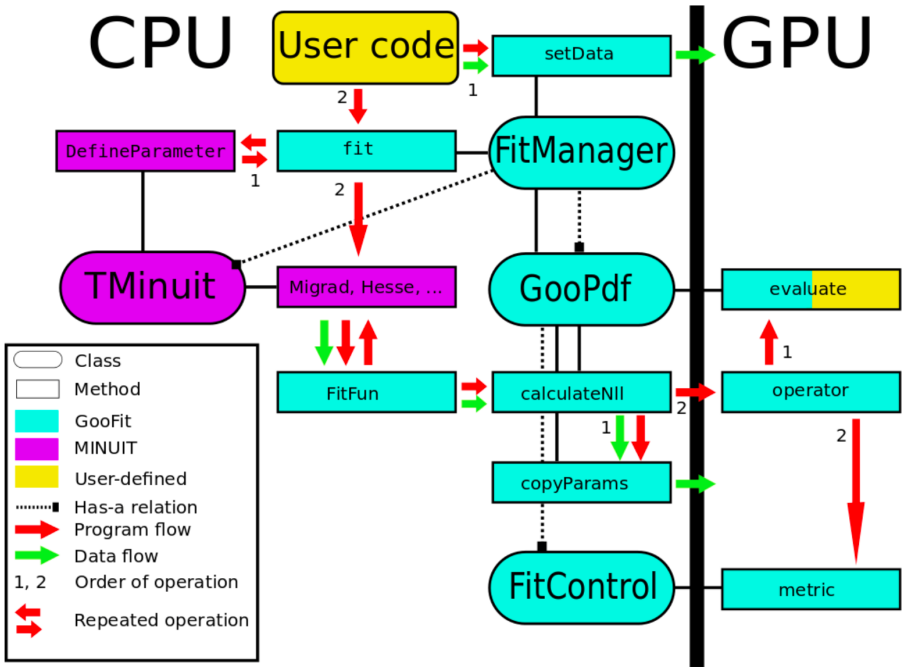
  The capabilities of GPU acceleration are being massively used in HEP (data analysis, Python-based ML/DL algorithms for reconstruction & identification, …)
  - This work is currently used in **Ph.D. lectures** about *Statistics in Data Analysis*
  - **It is still valid as** (unique?) **reference in exploring (and confirming) the validity of asymptotic results/methods now commonly used in HEP, by using huge amount of MC toys run on GPUs.**

Part 1 : Local statistical significance

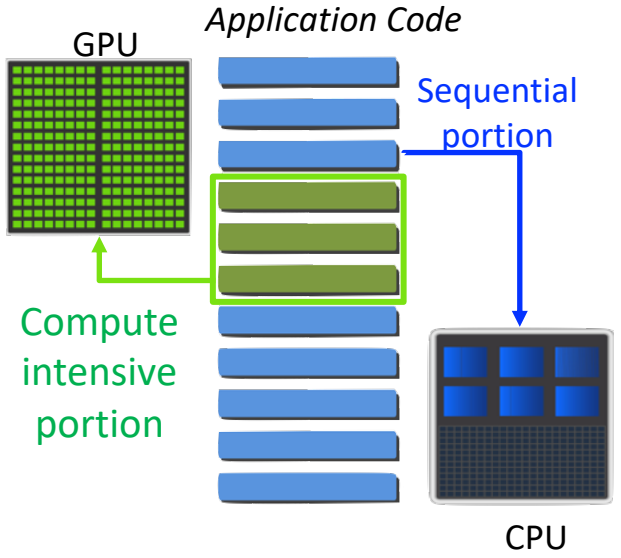
GPU computing in HEP: the GooFit framework

Heterogeneous GPU-accelerated computing is the use of a Graphics Processing Unit to accelerate scientific applications

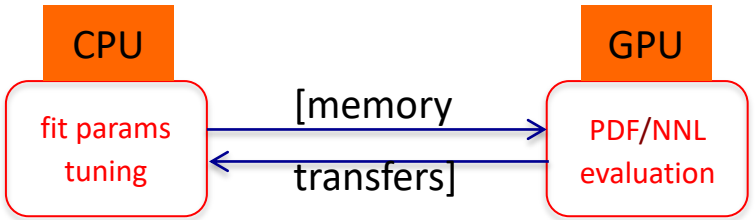
We explored the capabilities of GPU computing in the context of the 'end-user HEP analyses' by using *GooFit*.



From the user's perspective? Applications simply run significantly faster! How much faster ? It depends - of course - on the application… We tested it firstly with the estimation of the local significance of a known signal.



Goofit is a data analysis tool for HEP, that interfaces ROOT/RooFit to CUDA parallel computing platform on *nVidia* GPU. It also supports OpenMP.



Since v2.0 **Goofit** is completely integrated in python through **PyBindings** and it can run within jupyter notebooks that makes its use even easier.

A preliminary example of GPU capabilities

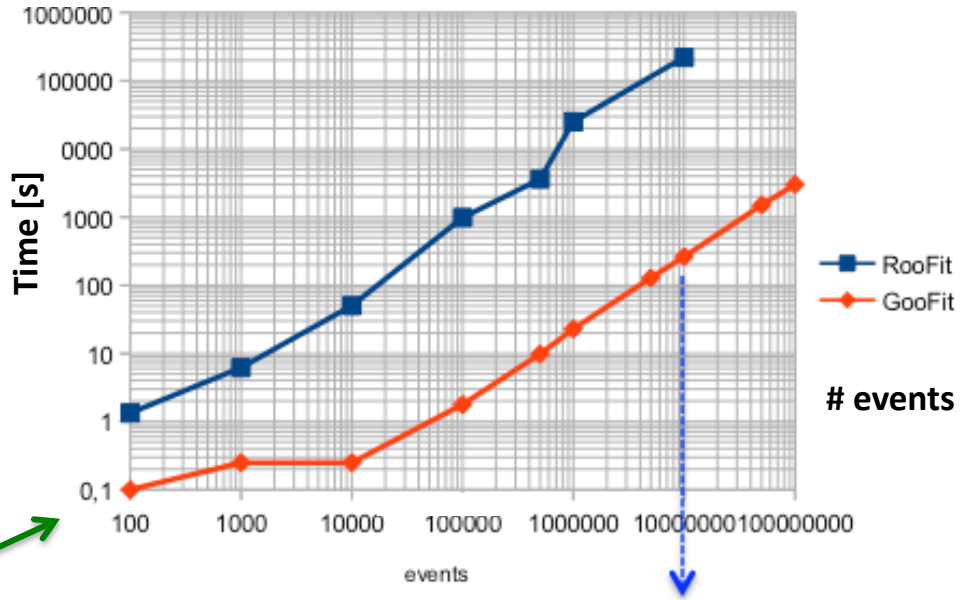
Parameter estimation is a crucial part of many physics analyses.

PDF evaluation on large datasets is usually the bottleneck in the MINUIT algorithm.

***GooFit* acts as an interface between the MINUIT minimization algorithm and a parallel processor which allows a Probability Density Function to be evaluated in parallel.**

A preliminary test was done with an Unbinned ML fit either by using a single CPU and by using an additional GPU (an nVIDIA Tesla C2070 hosted @ Bari T2).

Events according to a Voigtian model (convolution is CPU-intensive) are generated & fitted. The time needed (the negligible generation time is not included) **is studied as a function of the #events:**



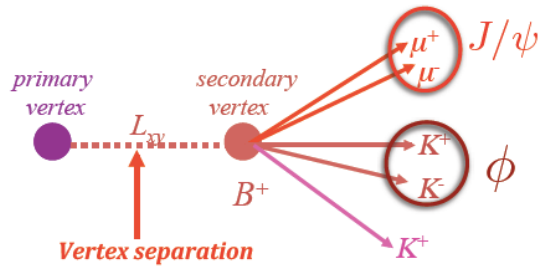
For 10M events: *RooFit* needs 61h+23m & *GooFit* takes 4m+39s : speed-up ~ 750

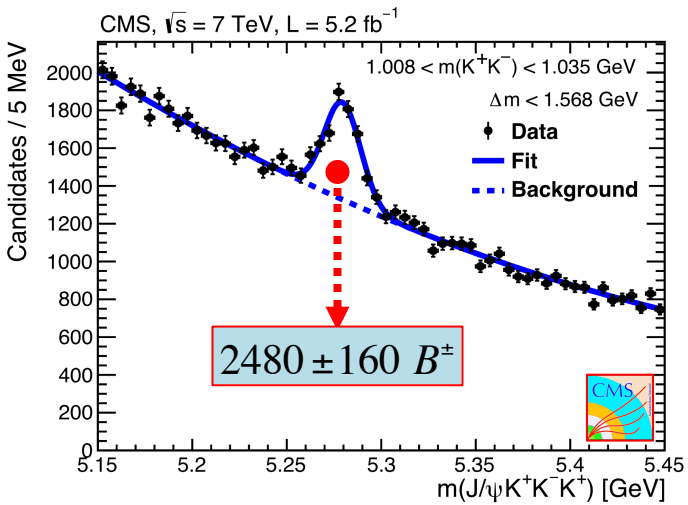
For 1M fitted events with *RooFit* … you need to wait overnight,
for 10M fitted events with *GooFit* … you need to take an espresso!

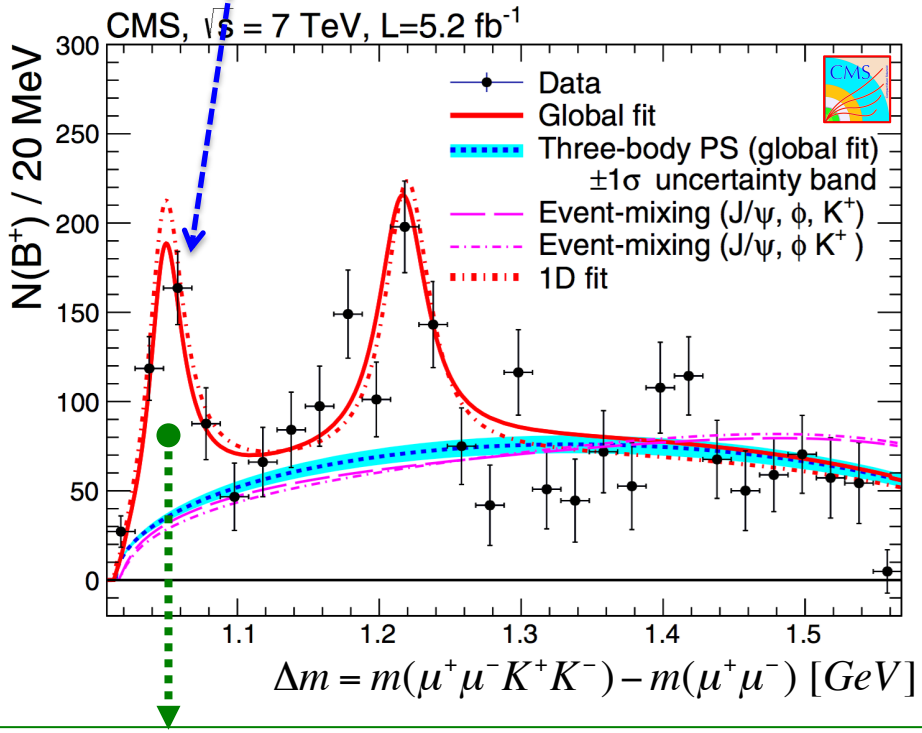
As expected, for a Binned ML fit, the speed-up ranges from few units to few dozens (with #bins).

Test application : the Physics case

To test the computing capabilities of GPUs with respect to CPU cores: a high-statistics toy Monte Carlo technique has been implemented both in *ROOT/RooFit* and *GooFit* frameworks with the aim to estimate the (local) statistical significance of the structure observed by CMS close to the kinematical boundary of the $J/\psi\phi$ invariant mass in the 3-body decay $B^+ \to J/\psi\phi K^+$ [PLB 734 (2014) 261]







Structure parameters [compatible with Y(4140) by CDF]:

- $m = 4148.0 \pm 2.4(stat.) \pm 6.3(syst.)\ MeV$
- $\Gamma = 28^{+15}_{-11}(stat.) \pm 19(syst.)\ MeV$

Test application : the pseudo-experiments method

MC pseudo-experiments are used to estimate the probability (*p-value*) that background fluctuations would - alone - give rise to a signal as much significant as that seen in the data.

Toy MC fit cycle (for each generated fluctuation):

- Generation of fluctuated background binned distribution (3-body phase-space model)
  [total #entries fixed by that in the data (ignoring Poisson fluctuations) ⇨ fits with not-extended ML]
- Null Hypothesis binned ML fit performed with the phase-space model only
-

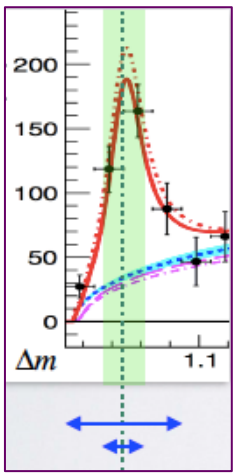
Test application : the pseudo-experiments method

MC pseudo-experiments are used to estimate the probability (*p-value*) that background fluctuations would - alone - give rise to a signal as much significant as that seen in the data.

Toy MC fit cycle (for each generated fluctuation):

- Generation of fluctuated background binned distribution (3-body phase-space model)
  [total #entries fixed by that in the data (ignoring Poisson fluctuations) ⇨ fits with not-extended ML]
- Null Hypothesis binned ML fit performed with the phase-space model only
- Alternative Hypothesis binned ML fit performed with the phase-space model + Voigtian PDF
  [the latter is truncated to correctly account for the kinematical threshold; the Gaussian resolution function has width fixed @ 2MeV]. Signal yield constrained > 0.

  Note: for each bin, the PDF value is estimated by ROOT integration over the bin
  [*time-consuming* but needed : steep signal w.r.t. bin size]



- Fit performed 8 times within the region of interest (from CDF: no LEE) trying different starting values (2 masses & 4 widths).

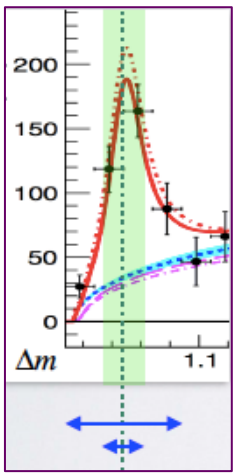
Test application : the pseudo-experiments method

MC pseudo-experiments are used to estimate the probability (*p-value*) that background fluctuations would - alone - give rise to a signal as much significant as that seen in the data.

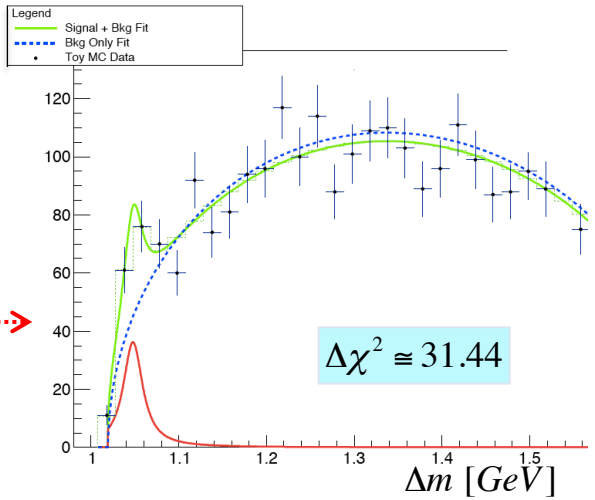
Toy MC fit cycle (for each generated fluctuation):

- Generation of fluctuated background binned distribution (3-body phase-space model)
  [total #entries fixed by that in the data (ignoring Poisson fluctuations) ⇨ fits with not-extended ML]
- Null Hypothesis binned ML fit performed with the phase-space model only
- Alternative Hypothesis binned ML fit performed with the phase-space model + Voigtian PDF
  [the latter is truncated to correctly account for the kinematical threshold; the Gaussian resolution function has width fixed @ 2MeV]. Signal yield constrained > 0.

  Note: for each bin, the PDF value is estimated by ROOT integration over the bin
  [*time-consuming* but needed : steep signal w.r.t. bin size]

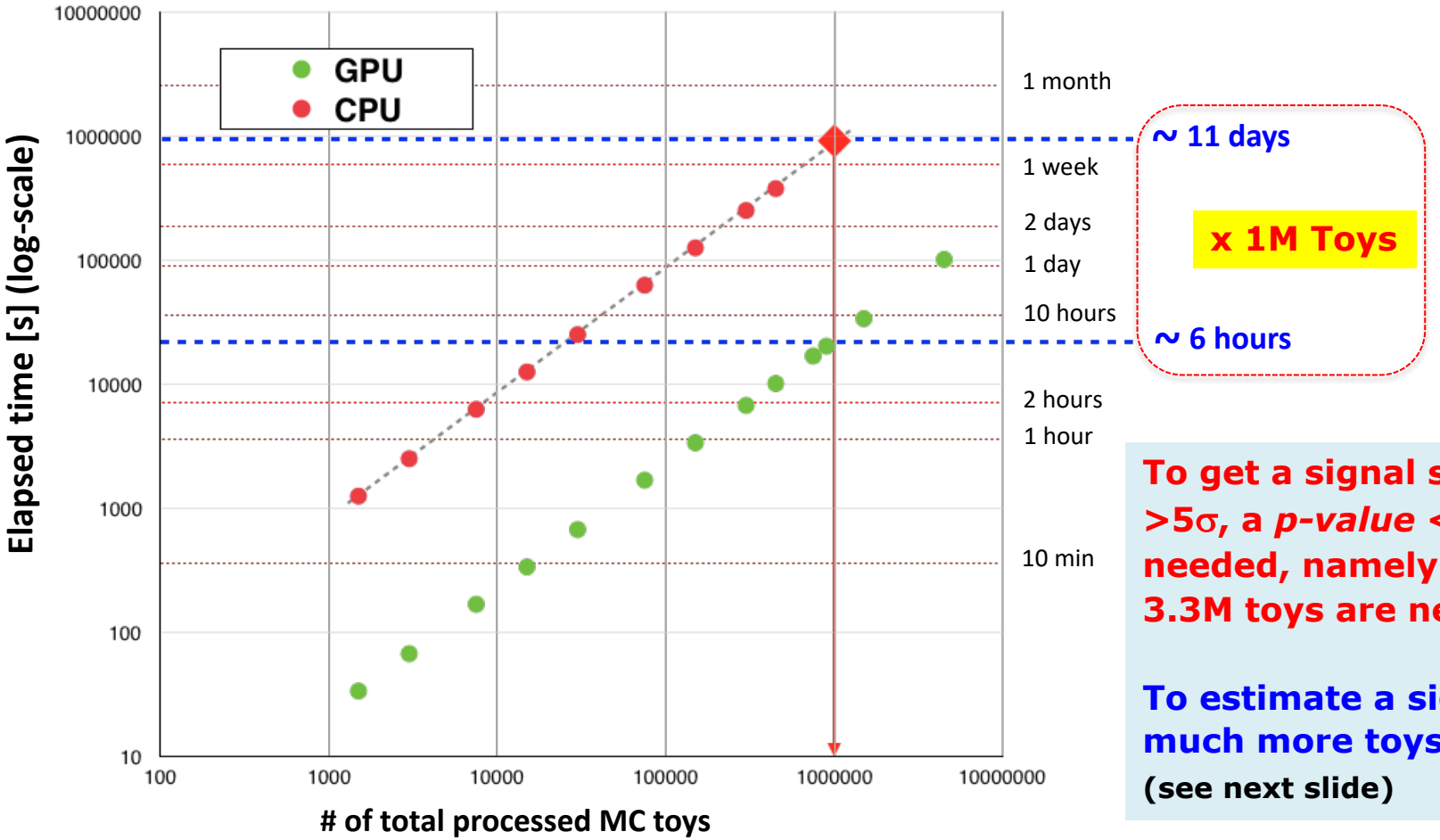


- Fit performed 8 times within the region of interest (from CDF: no LEE) trying different starting values (2 masses & 4 widths).
- For each fit calculate a $\Delta\chi^2$ w.r.t. the Null Hypothesis fit; the best $\Delta\chi^2$ fit among the 8 alternative fits is chosen !
- A $\Delta\chi^2$ (our test statistic) distribution is obtained over the sample of MC toys.



Roofit/Proof-Lite vs GooFit performances

- A performances' comparison can be done from the point of view of the end-user/analyst and the time needed to deliver the pseudo-experiments' task.
  Let us assume he has at his own disposal the full computational power used in these studies:
  2 servers equipped with 3 GPUs (2 TK20 & 1 TK40) and 72 CPU cores (36 physical cores + HyperThr).

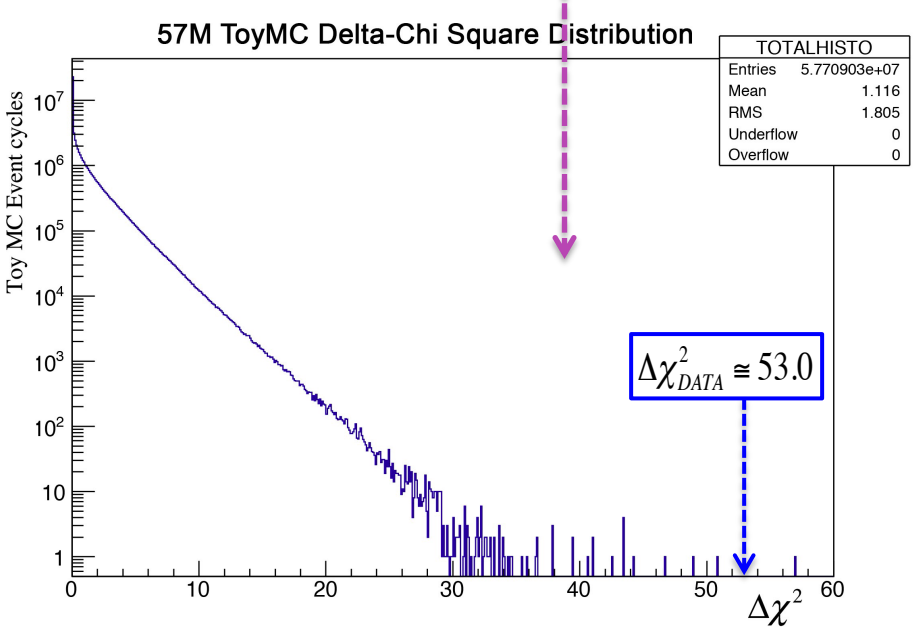


To get a signal significance >5σ, a *p-value* < 3×10^{-7} is needed, namely at least 3.3M toys are needed.

To estimate a signal signif. much more toys are needed
(see next slide)

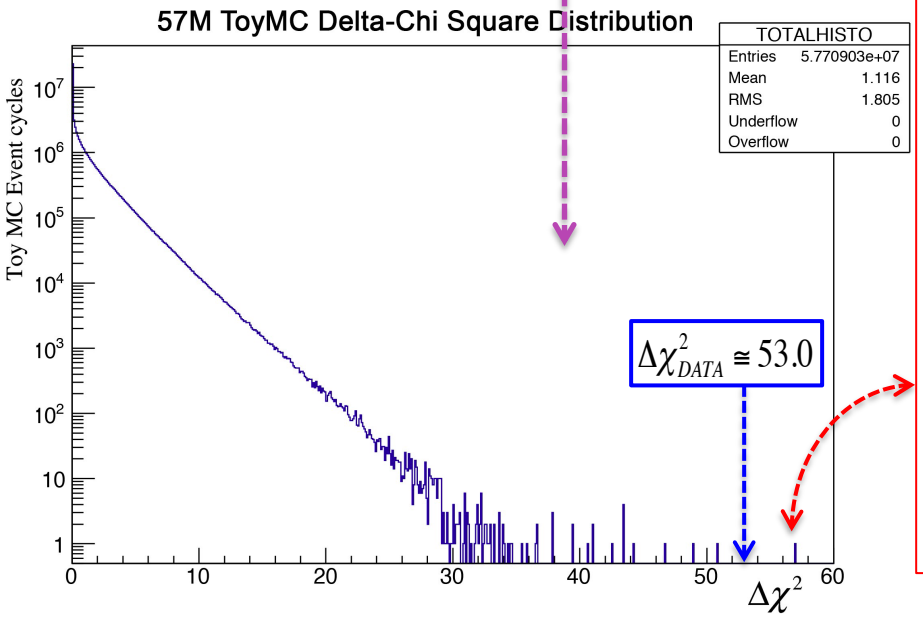
P-value & local statistical significance estimation

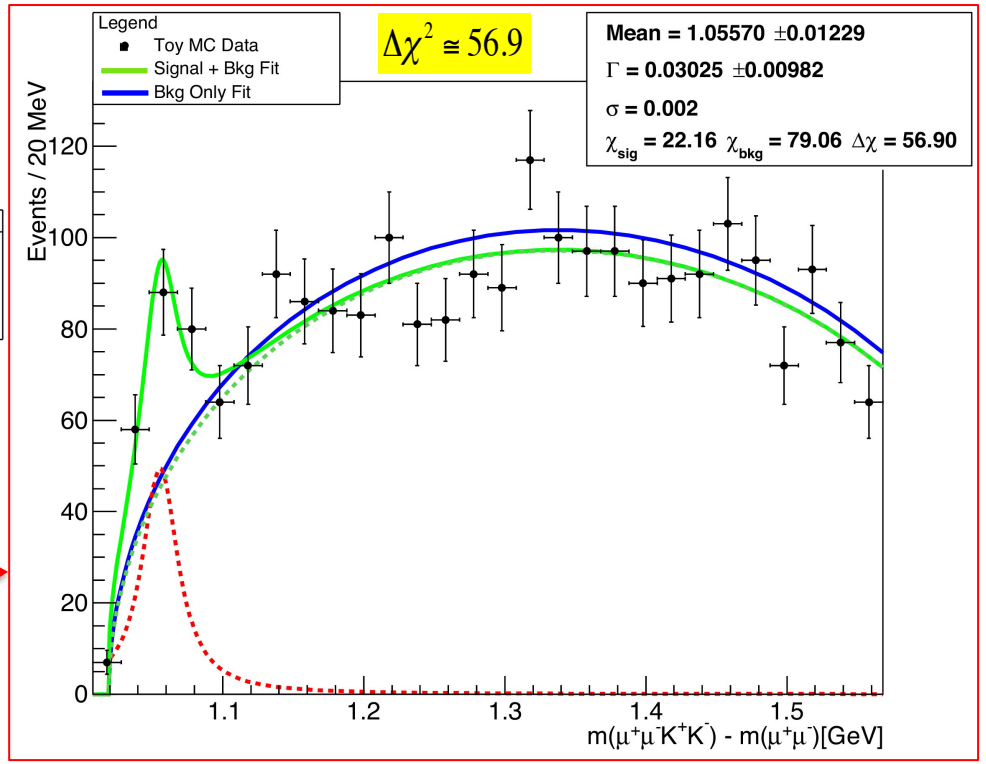
The **final obtained $\Delta\chi^2$ distribution**
(MC toys production was stopped once a fluctuation with $\Delta\chi^2 > \Delta\chi^2_{DATA}$ was found)



P-value & local statistical significance estimation

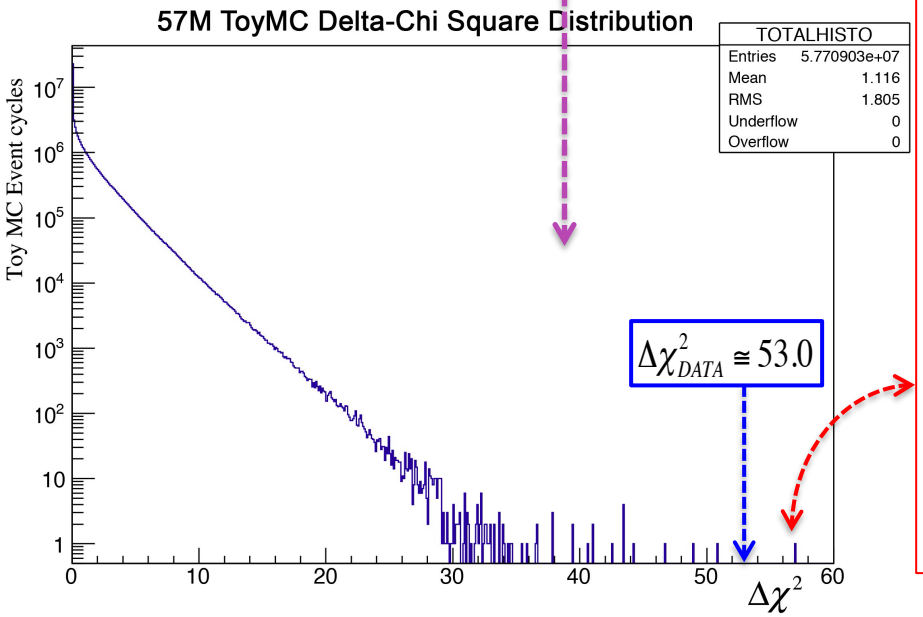
The **final obtained $\Delta\chi^2$ distribution**
(MC toys production was stopped once a fluctuation with $\Delta\chi^2 > \Delta\chi^2_{DATA}$ was found)

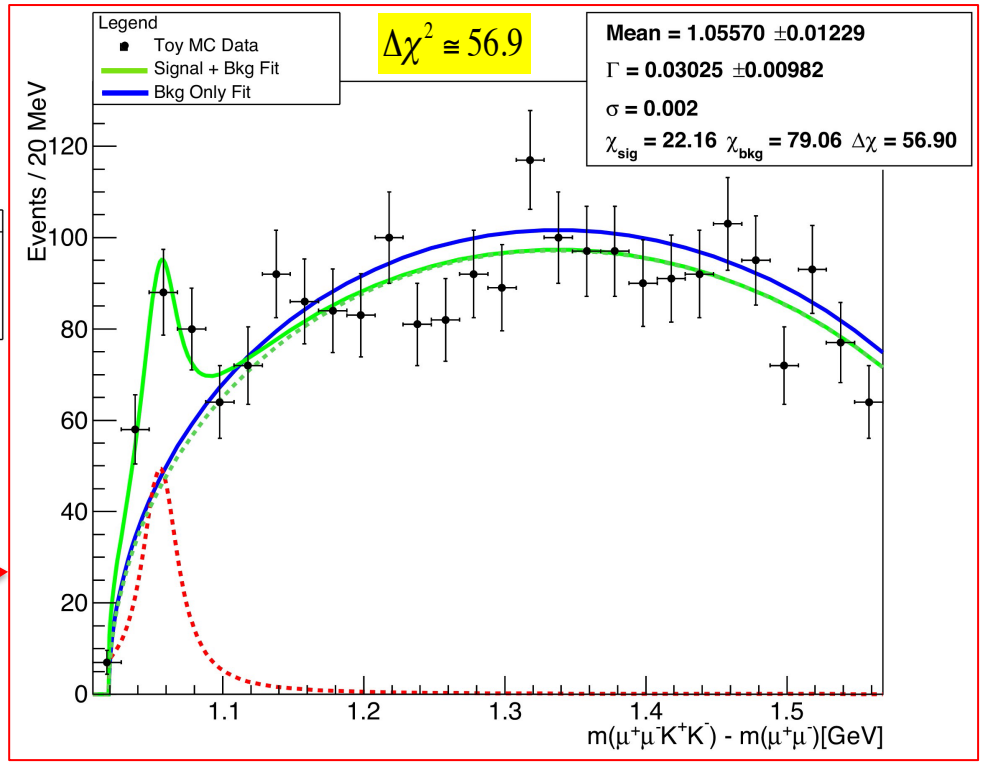




P-value & local statistical significance estimation

The **final obtained $\Delta\chi^2$ distribution**
(MC toys production was stopped once a fluctuation with $\Delta\chi^2 > \Delta\chi^2_{DATA}$ was found)





The **p-value estimation is straightforward:**

$$p-value\ :\ P = \int_{\Delta\chi^2_{DATA}}^{+\infty} \Delta\chi^2 \approx \frac{1}{57.7\cdot 10^6} \cong 1.73\cdot 10^{-8}$$

Equivalent (gaussian) statistical significance:

$$Z\sigma = \Phi^{-1}(1-P)\sigma \cong 5.52\sigma$$

Inverse function of the cumulative distribution of the standard gaussian

Compatible with the lower limit of 5σ for the statistical significance quoted in the CMS paper **PLB 734 (2014) 261** on the basis of 50.5 millions of MC toys (by *RooFit)*.

Wilks' theorem & the need of MC toys - I

- The Wilks[*] theorem is often used to estimate the p-value associated to a new/unexpected signal :

  Given two hypotheses:
  - Null hypotheses H_0 with ν_0 d.o.f.
  - Alternative hypotheses H_1 with ν_1 d.o.f.

  … any test statistic t, defined as a likelihood ratio $-2\ln\lambda = -2\ln\left(\frac{L_{H_0}}{L_{H_1}}\right)$

  [or similarly (in the asymptotic limit) as a $\Delta\chi^2 = \chi^2_{H_0} - \chi^2_{H_1}$],

  approaches a χ^2 distribution with $\nu = \nu_1 - \nu_0$ d.o.f., provided that these regularity conditions hold :

  - H_0 and H_1 are nested (H_1 "includes" H_0)
  - while $H_1 \to H_0$ the H_1 parameters are well behaving (defined and not approaching some limit)
  - asymptotic limit (of a large data sample)

- Once this theorem holds, the p-value associated to the signal is given by : $P = \int_{t_{obs}}^{\infty} \chi^2_{\nu_1-\nu_0}(t)\,dt$

  The use of pseudo-experiments to estimate the p-value is not needed (but still suggested)

- When null hypothesis is background-only and the alternative is background+signal, often the above regularity conditions are not all satisfied, and MC toys are mandatory !

[*] S.S.Wilks, *Ann.Math.Stat.* **9** (1938) 60-62

Wilks' theorem & the need of MC toys - II

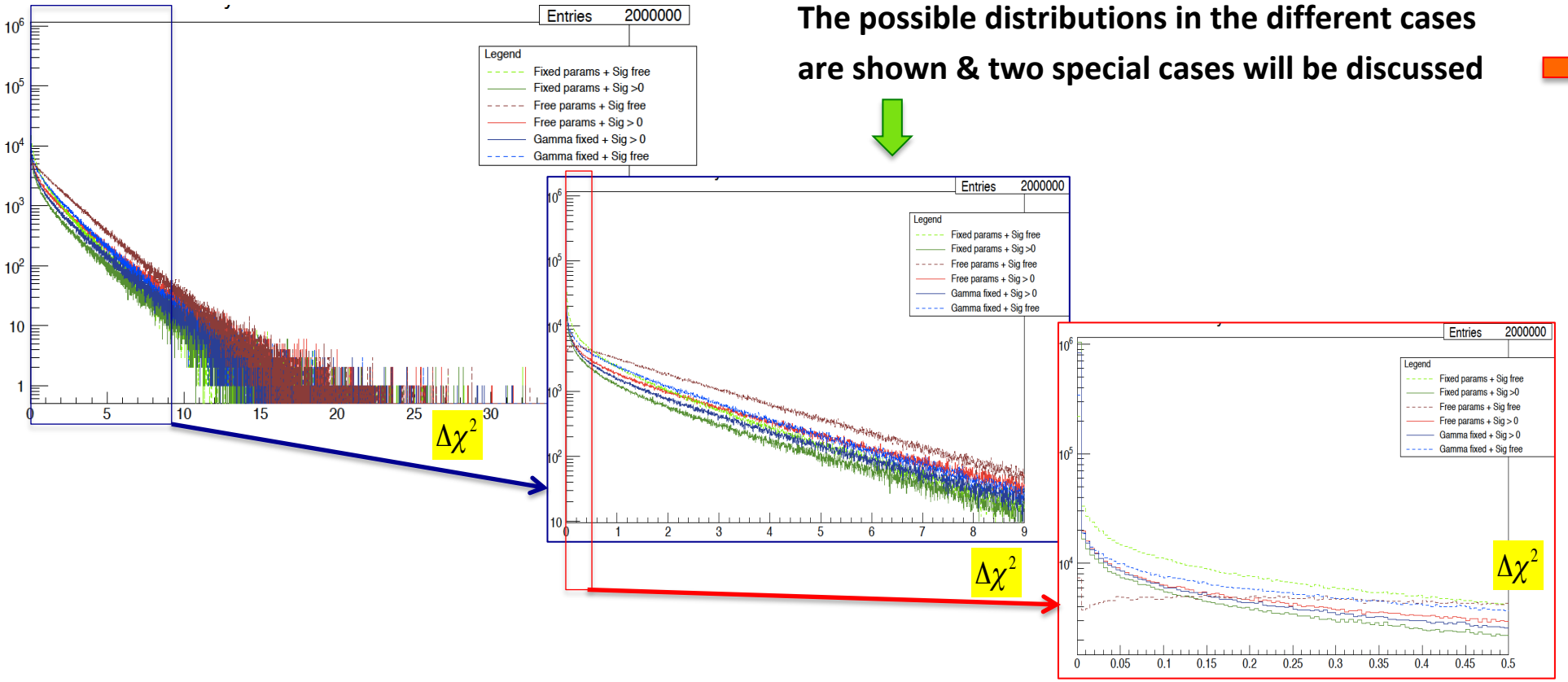
- **Indeed this is the case we are dealing with, here!**

  The signal parameters in the model of H_1 hypothesis are mass (m), width (Γ) and yield ($\mu \geq 0$).

  When $H_1 \to H_0$ the problem is that : 1) m and Γ are not well defined, 2) μ tend to the null limit.

  This explains why we have used pseudo-experiments.

- **The distributions of test statistic are in general nonpredictable and can be extracted from MC toys!**



The possible distributions in the different cases are shown & two special cases will be discussed

Special case in which Wilks' theorem holds

Consider the test statistic $t_\mu = -2\ln\lambda(\mu)$ [μ : *strength parameter*] as the basis of the statistical test.
This could be a test of $\mu = 0$ for purposes of establishing the existence of a signal process, or
… of $\mu \neq 0$ for purposes of obtaining a confidence interval.

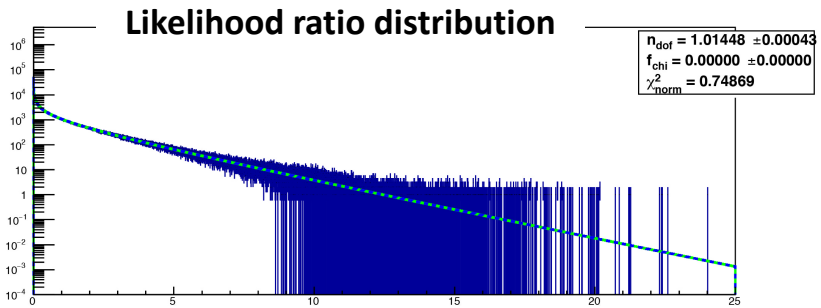
In this case following Cowan *et al.* [*] the PDF of the test statistic approaches a chi-square distribution for 1 d.o.f. :
[in agreement with Wilks theorem !]

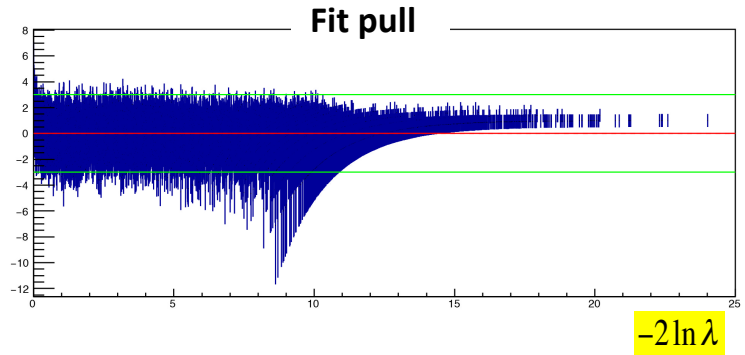
$$f(t_\mu|\mu) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{t_\mu}}e^{-t_\mu/2}$$

Let us fix the m & Γ parameters,
(to the CMS estimates from the fit to data)
while leaving μ free in our ML fits
(μ is not properly a signal yield).

By fitting our likelihood ratio distrib. we indeed get :

$$\text{d.o.f.} \approx 1.014 \pm 0.001$$





$$\chi^2_{norm} = 1.009 \quad P(fit) = 0.118$$

[*] Cowan *et al.*, EPJ C71 (2011) 1554

Special case : asymptotic formula by Cowan *et al.* [*] holds

Consider the special case of the test statistic t_μ with the purpose to test $\mu = 0$ in a class of model where we assume $\mu \geq 0$. Rejecting $\mu = 0$ (the null hypothesis) leads to the discovery of a new signal.

In this case following Cowan *et al*. the test statistic is :

$$q_0 = \begin{cases} -2\ln\lambda(0) & with \begin{cases} \hat{\mu} \geq 0 \\ \hat{\mu} < 0 \end{cases} \\ 0 \end{cases}$$

Cowan *et al*. derive analitically that the PDF of q_0 is an equal mixture of a delta function at 0 & a chi-square distribution for 1 d.o.f. :

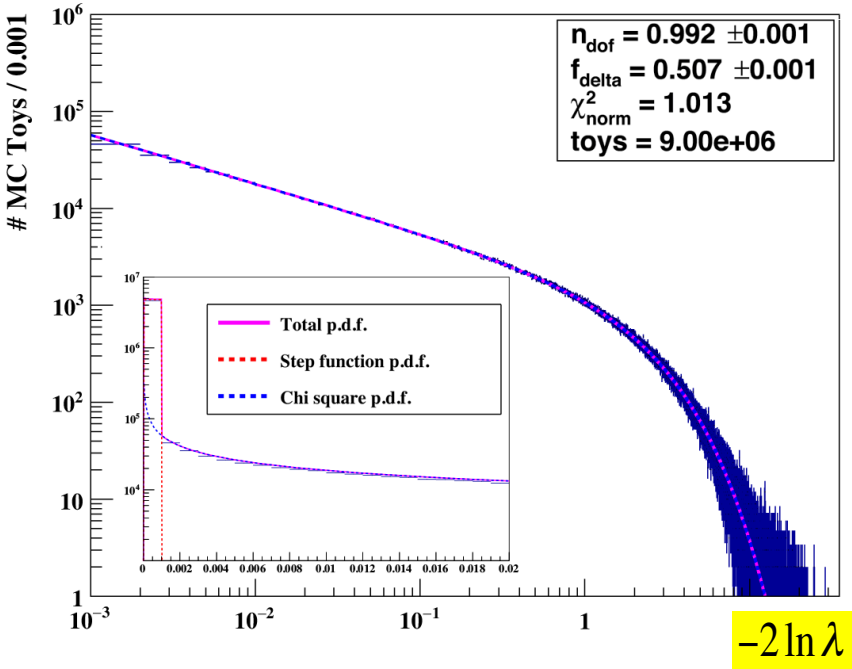
$$g(q_0|\mu = 0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}\right]$$

Let us fix the m & Γ parameters (to the CMS estimates from fit to data) while constraining $\mu \geq 0$ in our ML fits (μ represents a signal yield here).

By fitting our likelihood ratio distrib. we indeed get :

$$\text{d.o.f.} \approx 0.992 \pm 0.001$$

$$\text{weight } C_{\chi^2} \approx 0.507 \pm 0.01$$



[*] Cowan *et al.*, EPJ C71 (2011) 1554

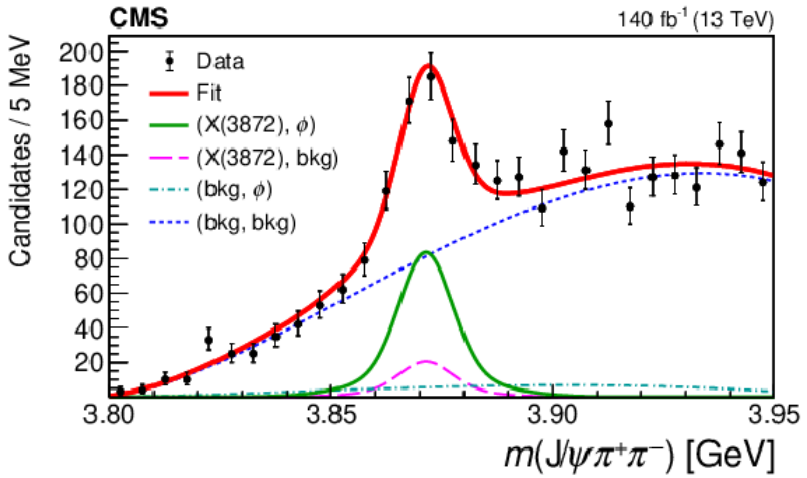
Part 2 : Look-Elsewhere-Effect & Global statistical significance

Many searches for new physical phenomena look for a peak in a distribution, typically a reconstructed invariant mass. The peaking structure may represent a resonance/particle.

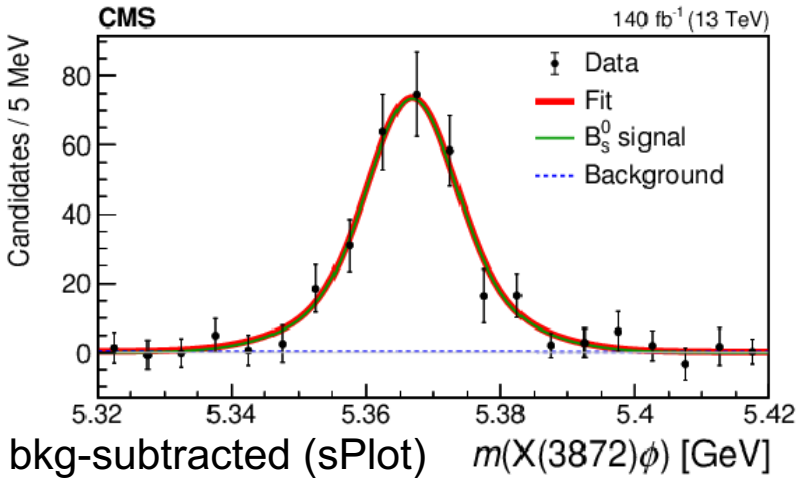
In some cases the location (mass) of a peak (particle) is known …
1) like in searches for rare decays of **known** particles

Recently observed new decay mode involving the $X(3872)$

$X(3872) \to J/\psi\,\pi^+\pi^-$



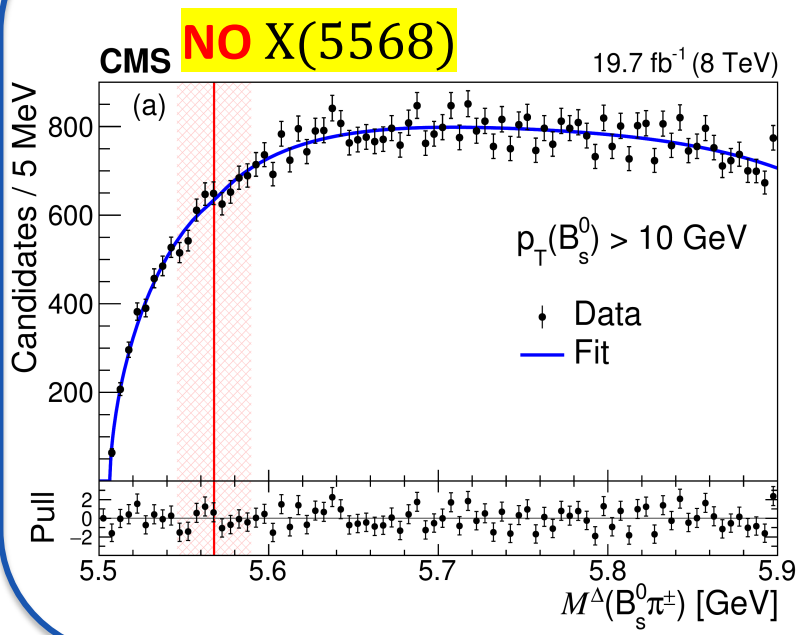
$B_s^0 \to X(3872)\phi$



- Many searches for new physical phenomena look for a peak in a distribution, typically a reconstructed invariant mass. The peaking structure may represent a resonance/particle.

In some cases the location (mass) of a peak (particle) is known …
1) like in searches for rare decays of a know particles
2) **when an experiment is looking to confirm a new particle discovered/claimed by another experiment** (we discussed in detail an example in the first part)
3)

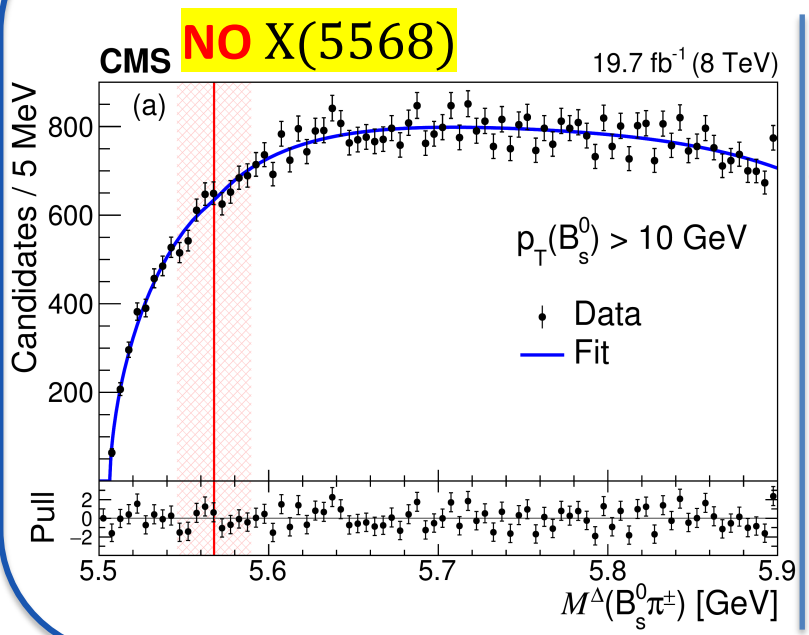
The investigated invariant mass is the same: $B_s^0\pi^\pm$

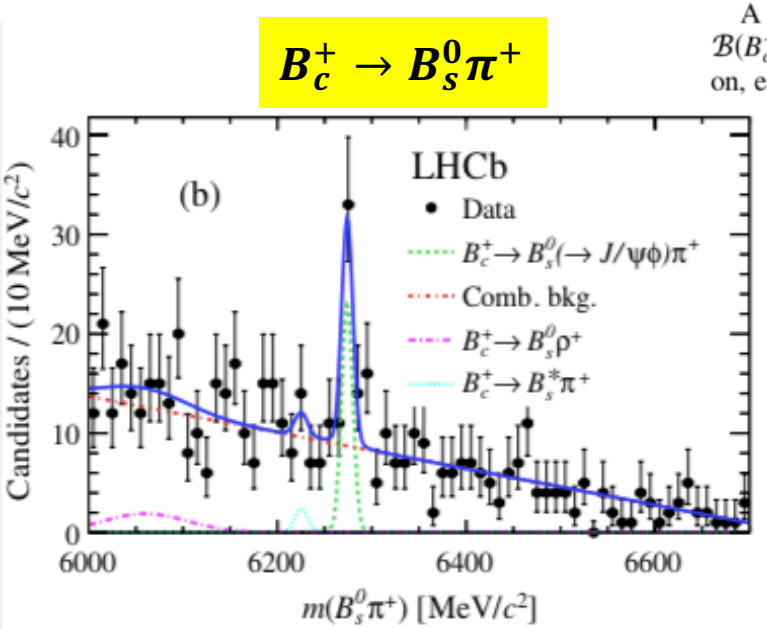


- Many searches for new physical phenomena look for a peak in a distribution, typically a reconstructed invariant mass. The peaking structure may represent a resonance/particle.

In some cases the location (mass) of a peak (particle) is known …
1) like in searches for rare decays of a know particles
2) when an experiment is looking to confirm a new particle discovered/claimed by another experiment (we discussed in detail an example in the first part)
3) **or when one (or more) theoretical model(s) predicts it**

The investigated invariant mass is the same: $B_s^0\pi^\pm$





A wide range of predictions for the branching fraction $\mathcal{B}(B_c^+ \to B_s^0\pi^+)$ exists, between 16.4% and 2.5%, based on, e.g., QCD sum rules [9,10], or quark-potential models

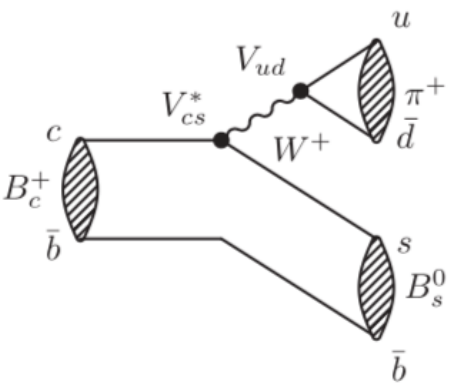
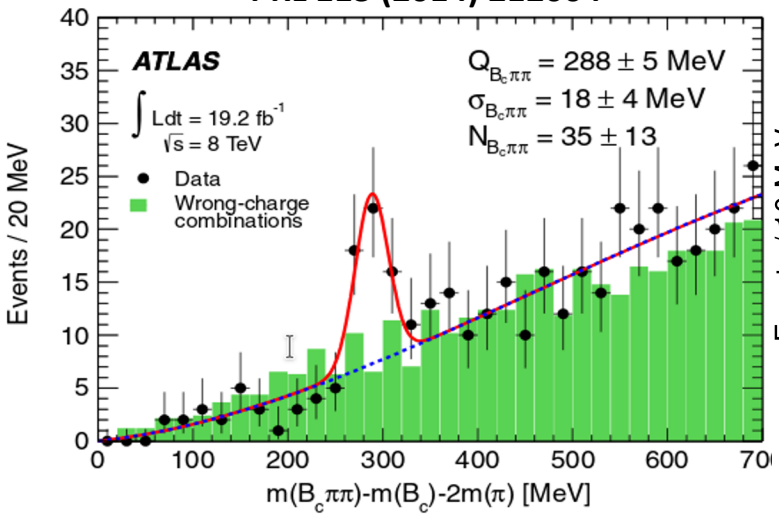


FIG. 1. Leading-order Feynman diagram of the decay $B_c^+ \to B_s^0\pi^+$.

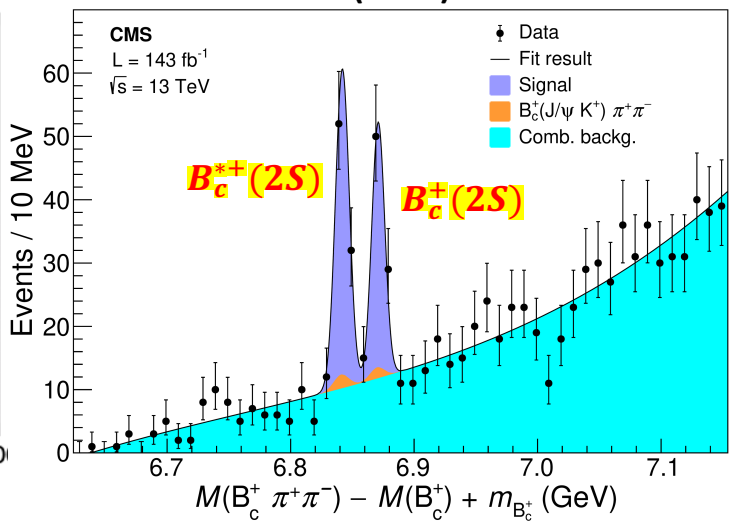
Many searches for new physical phenomena look for a peak in a distribution, typically a reconstructed invariant mass. The peaking structure may represent a resonance/particle.

In some cases the location (mass) of a peak (particle) is known …
1) like in searches for rare decays of a know particles
2) when an experiment is looking to confirm a new particle discovered by another experiment
3) or when one (or more) theoretical model(s) predicts it
4) **or even when (2) and (3) both happen and hold**

PRL 113 (2014) 212004



PRL 122 (2019) 132001



$B_c^{*+}(2S) \rightarrow B_c^{*+}\pi^+\pi^-, B_c^{*+} \rightarrow B_c^+\gamma$

very soft: undetected

Predictions indicate that

$$[m(B_c^{*+}(1S)) - m(B_c^+(1S))] > [m(B_c^{*+}(2S)) - m(B_c^+(2S))]$$

imply $B_c^{*+}(2S)$ peak is assumed …
… to be the lower one

Local significance exceeding 6.5σ for observing 2 peaks rather than 1

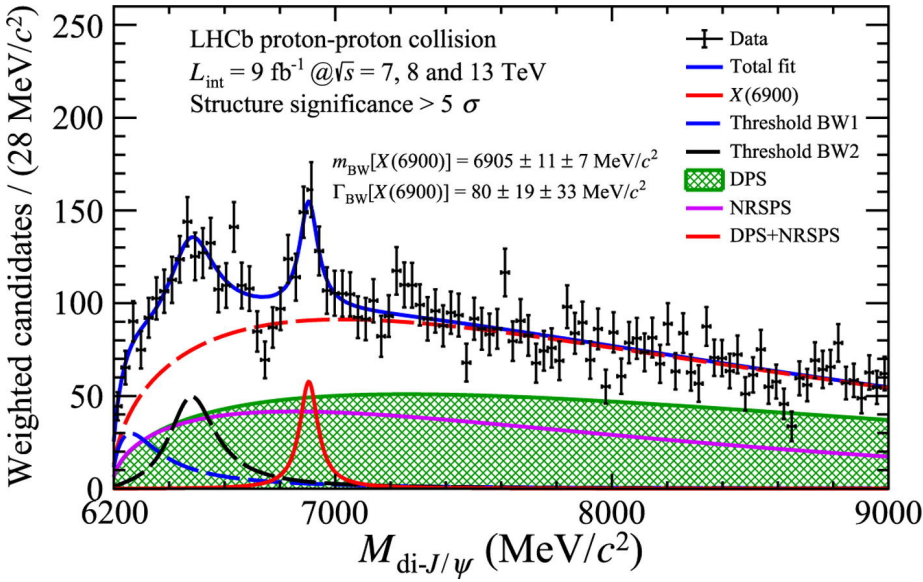
For both: single peak significance >5σ

Look-Elsewhere-Effect

In the case of searches for new particles whose mass is not predicted by theory (like the Higgs boson) or unexpected at all …

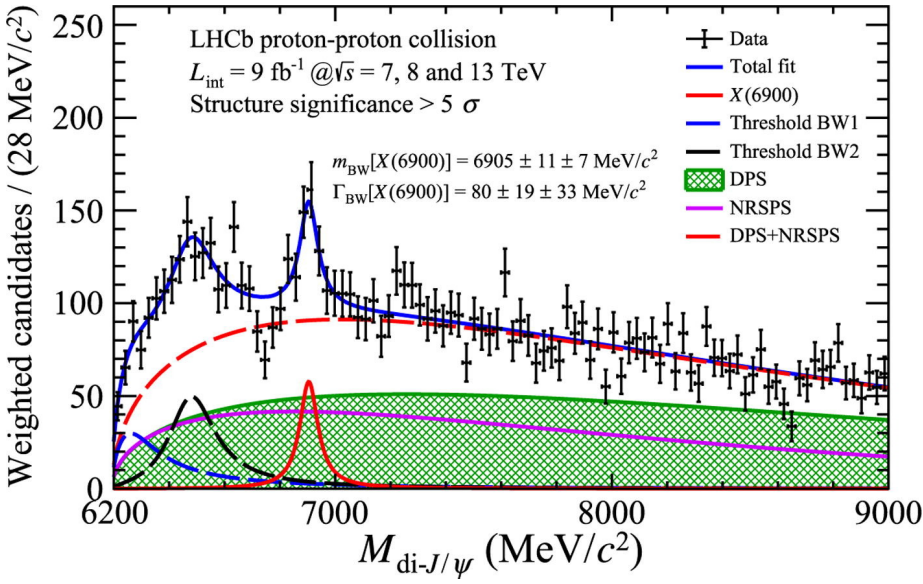
… if an excess in data, compared with the back-ground(s) expectation(s) is found at any mass value - in principle produced either by the presence of a real signal or by a backgrund fluctuation - **it could be interpreted as a possible signal of a new resonance (**in **any** position in the investigated mass window**).**

In this case …



Look-Elsewhere-Effect

In the case of searches for new particles whose mass is not predicted by theory (like the Higgs boson) or unexpected at all …



… if an excess in data, compared with the back-ground(s) expectation(s) is found at any mass value - in principle produced either by the presence of a real signal or by a backgrund fluctuation - **it could be interpreted as a possible signal of a new resonance (**in any position in the investigated mass window**).**

In this case … the mass is not fixed but estimated from data

and … the local significance must be replaced by a global significance based on:

~~Local~~ **p-value** Global $p(m_{\not\phi}) = \int_{q_{obs}(m_{\not\phi})}^{\infty} f(q|m_{\not\phi}, \mu = 0)\, dq$

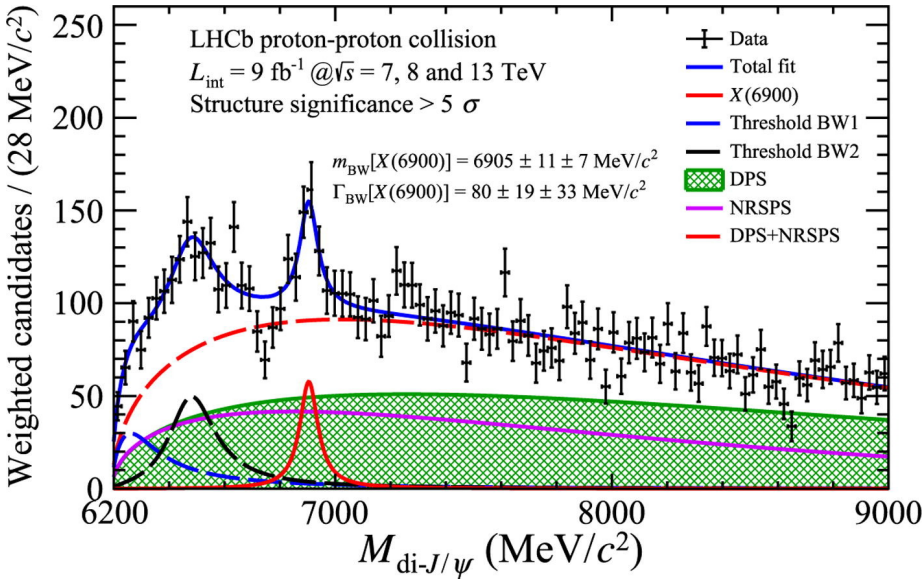
PDF of the adopted test statistic q

… that gives the probability that a background fluctuation at ~~a fixed~~ any mass value, in the range of interest, results in a value of q greater or equal the observed value $q_{obs}(m_{\not\phi})$

Look-Elsewhere-Effect

- **In the case of searches for new particles whose mass is not predicted by theory (like the Higgs boson) or unexpected at all …**

… if an excess in data, compared with the back-ground(s) expectation(s) is found at any mass value - in principle produced either by the presence of a real signal or by a backgrund fluctuation - **it could be interpreted as a possible signal of a new resonance (**in any position in the investigated mass window**).**



In this case … the mass is not fixed but estimated from data

and … the local significance must be replaced by a global significance based on:

~~Local~~ Global ***p*-value** $p(m_{\not\phi}) = \int_{q_{obs}(m_{\not\phi})}^{\infty} f(q|m_{\not\phi}, \mu = 0)\, dq$

PDF of the adopted test statistic q

… that gives the probability that a background fluctuation at ~~a fixed~~ any mass value, in the range of interest, results in a value of q greater or equal the observed value $q_{obs}(m_{\not\phi})$

- **In general: Global *p*-value > Local *p*-value ⇒ Global *significance* < Local *significance***

This effect of reduction of significance is called Look-Elsewhere-Effect (LEE)

Global *p*-value determination - I

- More in general, when an experiment is looking for a signal where one or more parameters of interest ($\vec{\theta}$) are unknown (i.e. both mass and width or other properties of a new particle), the global *p*-value can be computed using, as test statistic, the largest value of the parameter estimator over the entire parameter range:

$$q^{glob} = \sup_{\substack{\theta^i_{min} < \theta^i < \theta^i_{max} \\ i = 1,..,m}} q(\vec{\theta}, \mu = 0) = q(\hat{\vec{\theta}}, \mu = 0)$$

Set of parameters of interest that maximize $q(\vec{\theta}, \mu = 0)$

- The global *p*-value can be determined from the distribution of the test statistic q^{glob} assuming background only, given the observed value q^{glob}_{obs}: $p^{glob} = \int_{q^{glob}_{obs}}^{\infty} f(q^{glob}|\mu = 0)\, dq^{glob}$

- Let's remain in the simplest 1D case of a resonance search (thus $\vec{\theta} = m, \Gamma$) where the peak width is dominated by the experimental resolution if the intrinsic width is relatively small ($\Gamma \ll \Gamma^{exp}_{res} \equiv \Gamma_0$): $\theta = m$ and Γ_0 fixed (taken from simulation).

Global *p*-value determination - II

- **Even in this 1D case (**only mass as free parameter**) and even if the test statistic *q* is derived, as usual, from a likelihood ratio, Wilks' theorem cannot be applied because the value of the mass is undefined for $\mu = 0$: in case of background only *q* would no longer depend on *m* and the two hypotheses assumed at the numerator and denominator of the likelihood ratio would not be *nested*.**

- **Then… how to evaluate q^{glob} ? There are again two approaches:**

Global *p*-value determination - II

- **Even in this 1D case (**only mass as free parameter**) and even if the test statistic *q* is derived, as usual, from a likelihood ratio, Wilks' theorem cannot be applied because the value of the mass is undefined for $\mu = 0$: in case of background only *q* would no longer depend on *m* and the two hypotheses assumed at the numerator and denominator of the likelihood ratio would not be *nested*.**

- **Then… how to evaluate q^{glob} ? There are again two approaches:**

  - **Compute it with the method of pseudo-experiments (MC toys)**

    As we know from the 1st part: large significance values, corresponding to very low *p*-values, require a considerably large amount of toys and a huge demand for CPU time.

    Again... **GPUs will get us to the rescue !**

  -

Global *p*-value determination - II

- **Even in this 1D case (**only mass as free parameter**) and even if the test statistic *q* is derived, as usual, from a likelihood ratio, Wilks' theorem cannot be applied because the value of the mass is undefined for $\mu = 0$: in case of background only *q* would no longer depend on *m* and the two hypotheses assumed at the numerator and denominator of the likelihood ratio would not be *nested*.**

- **Then… how to evaluate q^{glob} ? There are again two approaches:**

  - **Compute it with the method of pseudo-experiments (MC toys)**

    As we know from the 1st part: large significance values, corresponding to very low *p*-values, require a considerably large amount of toys and a huge demand for CPU time.
    Again... **GPUs will get us to the rescue !**

  - **Estimate it in an *approximate way* (**still taking into account the LEE**) relying on the asymptotic behaviour of likelihood-ratio estimators : method of *Trial Factors***

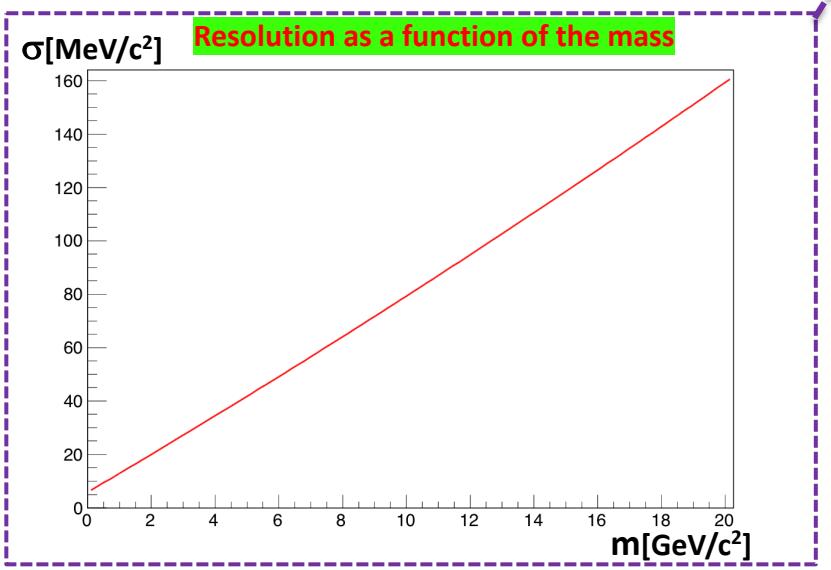
Pseudo-experiments & LEE - starting point

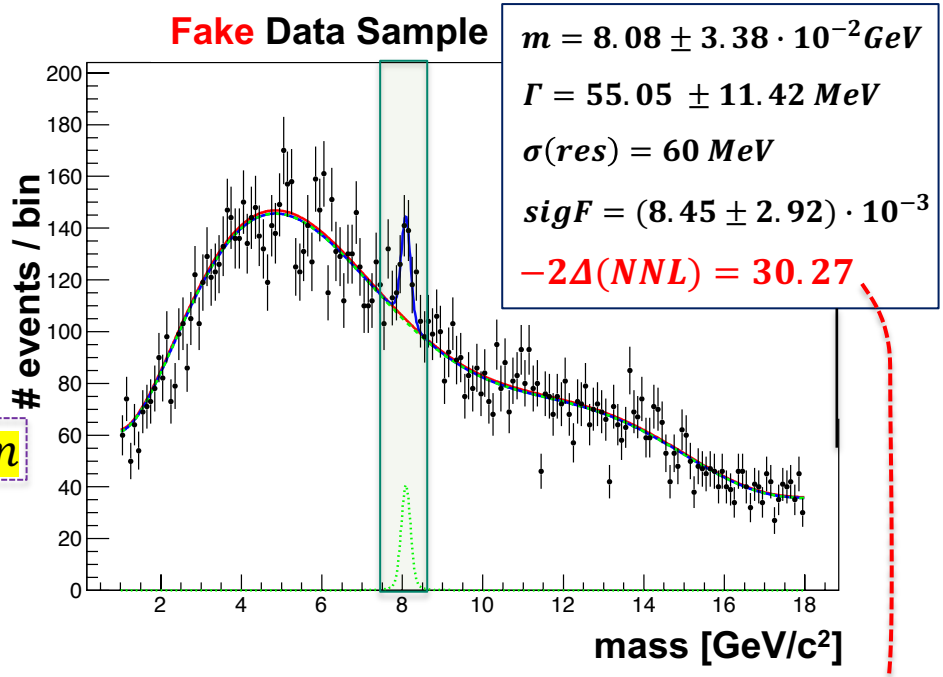
To take into account the LEE and calculate the global *p*-value we need to extend the MC toys method (earlier discussed) by introducing a (clustering-based) scanning technique.

A **pseudo-data inv. mass distribution** of **15K** candidates in a generic region of interest (1-18GeV) **was generated ad hoc**:

In the generation model & subsequent fit :

- **BKG-model:** 7th-order *Polynomial*
- **SIGNAL-model:** $BW \otimes Gaussian\ resolution\ function$
  [artificially added at ~8GeV]





From the approximation:

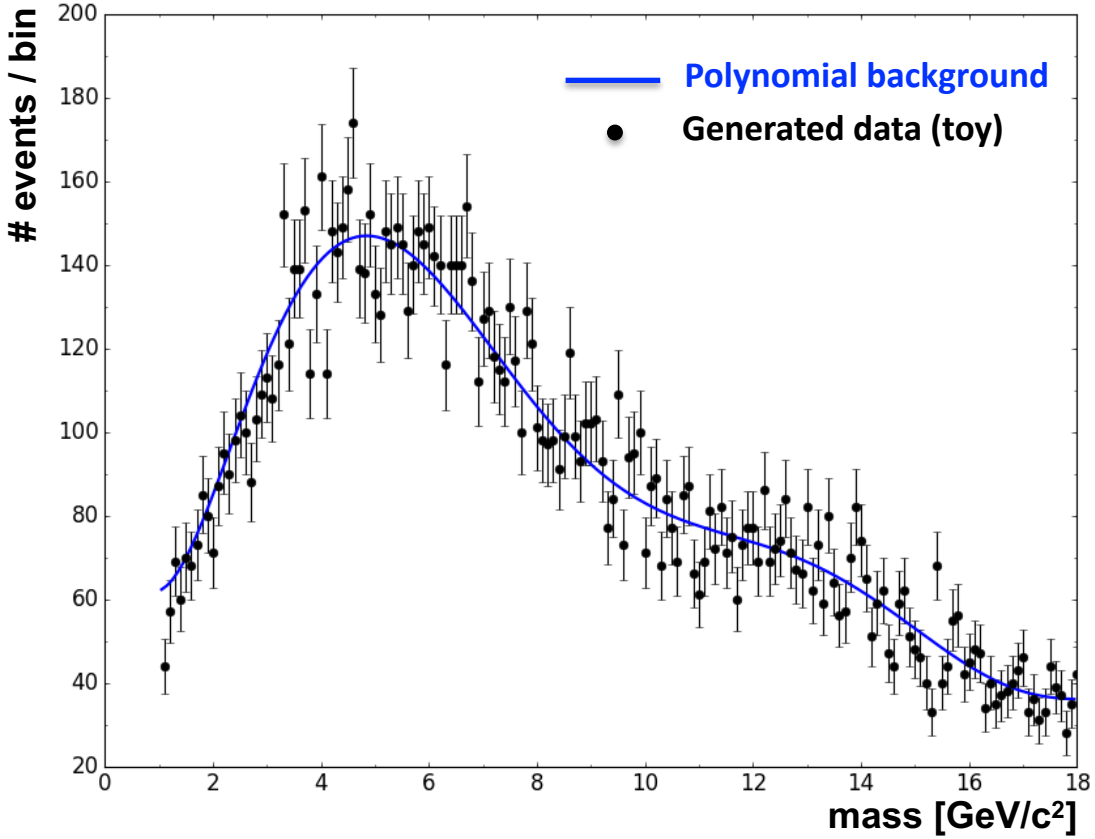
$$Z \cong \sqrt{-2[\ln(\mathcal{L}_{H1}) - \ln(\mathcal{L}_{H0})]} \equiv \sqrt{-2\Delta(NNL)} \cong 5.5$$

Pseudo-experiments & LEE - scanning technique [steps 1-2]

The **scanning technique** has been configured on the basis of a **clustering approach** and has been designed in advance with the aim to satisfy two concurrent requirements:

(A) Do not miss any relevant fluctuation

(B) Do not select too many small fluctuations



The procedure:

1) For **each MC Toy iteration** a distribution based on the **background PDF** model is generated.

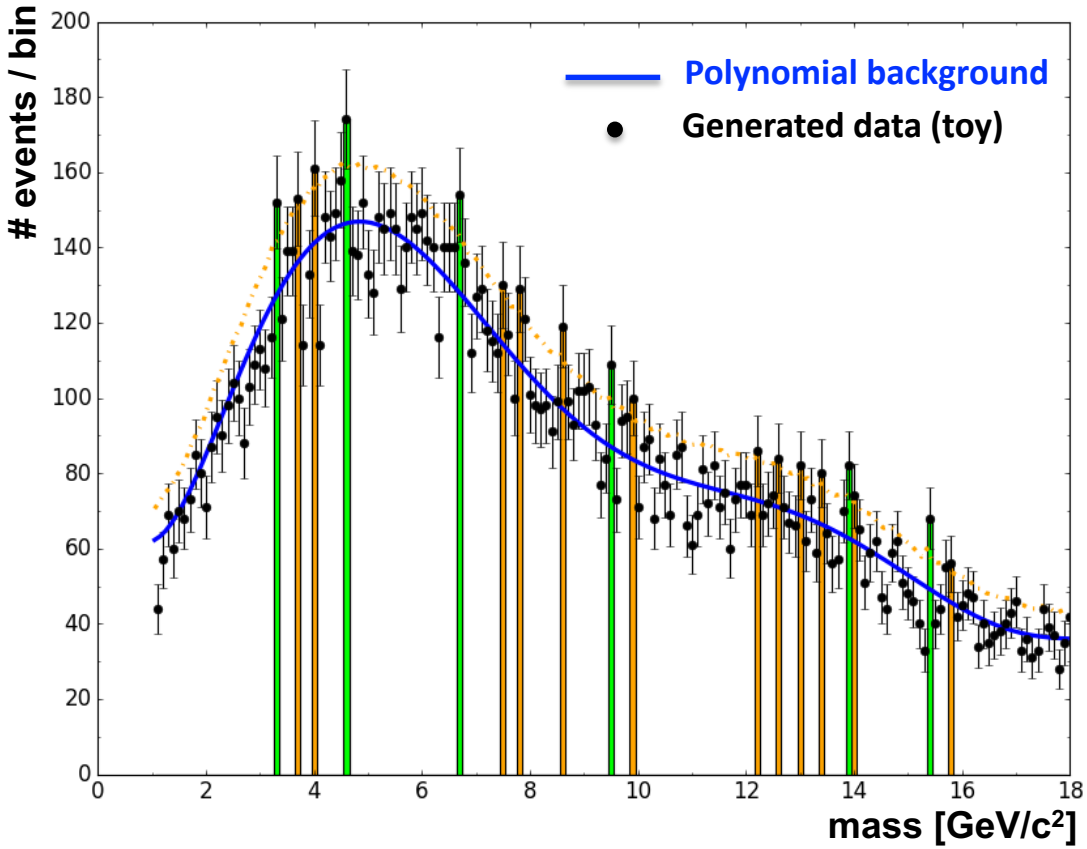
2) The ***H_0 Null Hypothesis*** fit is performed with the background function only.

Pseudo-experiments & LEE - scanning technique [steps 1-4]

The **scanning technique** has been configured on the basis of a **clustering approach** and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do not miss any relevant fluctuation

B) Do not select too many small fluctuations



The procedure:

1) For **each MC Toy iteration** a distribution based on the **background PDF** model is generated.

2) The ***H_0 Null Hypothesis*** fit is performed with the background function only.

3) **A first scan is performed to search for a main seed defined as a bin the content of which fluctuates more than $x\sigma$ strictly above the value of the background function.**

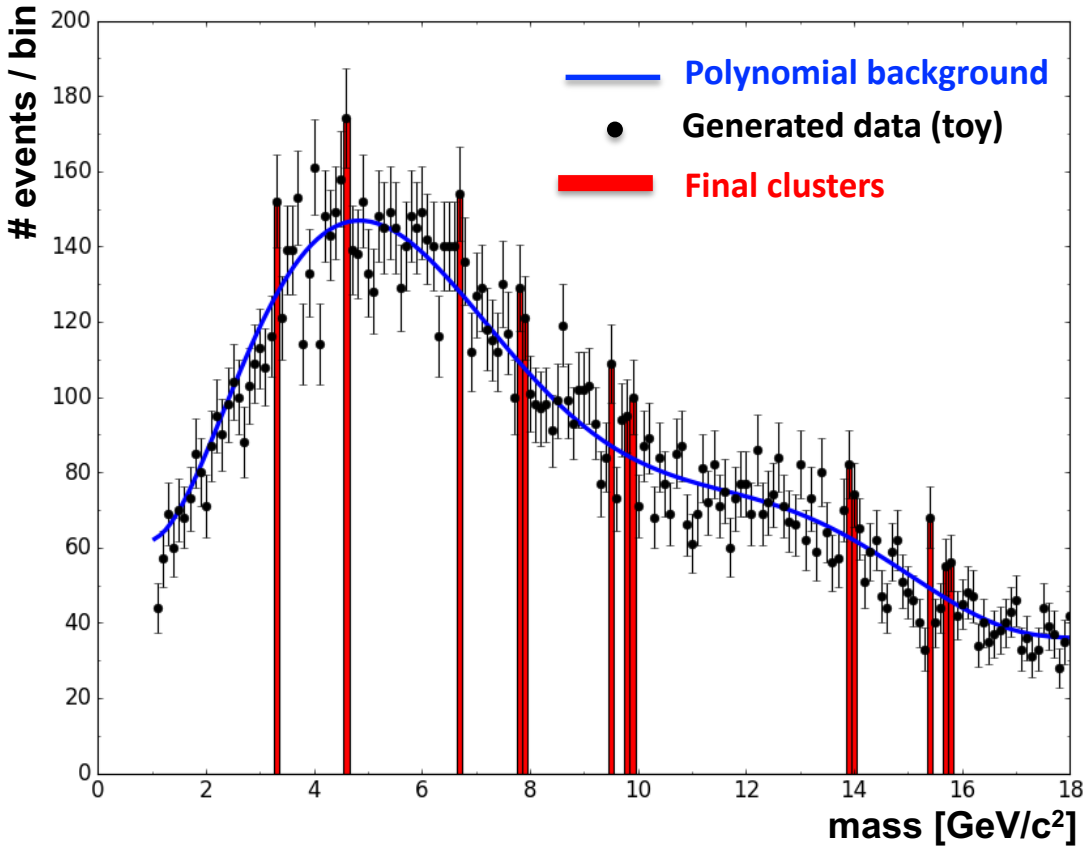
4) **A second scan is performed to search for a light seed defined as a bin the content of which fluctuates more than $y\sigma$ ($y<x$) strictly above the value of the background function.**

Pseudo-experiments & LEE - scanning technique [steps 5-7]

The **scanning technique** has been configured on the basis of a **clustering approach** and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do not miss any relevant fluctuation

B) Do not select too many small fluctuations



The procedure:

5) A final scan is performed to search for a side seed defined as a bin the content of which fluctuates more than $z\sigma$ ($z<y<x$) strictly above the value of the background function.

6) The final step consists of cleaning up the seeds

- all the main (x) seeds are retained;
- the light (y) seeds are kept only if at least one of the side bins is a seed (of any kind);
- the side (z) seeds are kept only if at least one of the side bins is a main or light seed.

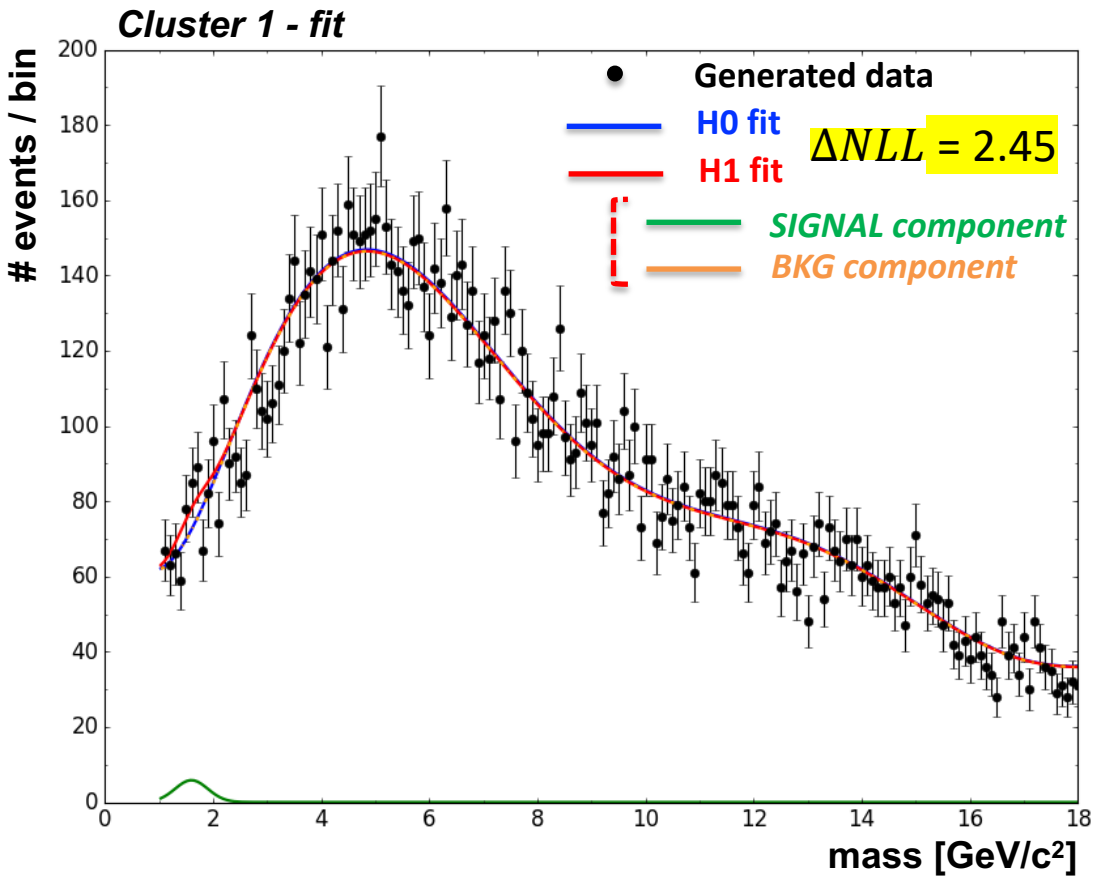
7) The clusters are thus formed.

Pseudo-experiments & LEE - scanning technique [step 8/fit 1]

The **scanning technique** has been configured on the basis of a **clustering approach** and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do not miss any relevant fluctuation

B) Do not select too many small fluctuations



The procedure:

8) For **each cluster** the **Alternative Hypothesis H1** fits are performed with the ***polynomial H0*** **+** a ***Convolution*** of a BW (signal) and a Gaussian (resolution) for the peak.

A set of fits is performed **changing the** range & the starting values of the **parameters** (***m***, ***Γ***, ***σ***):

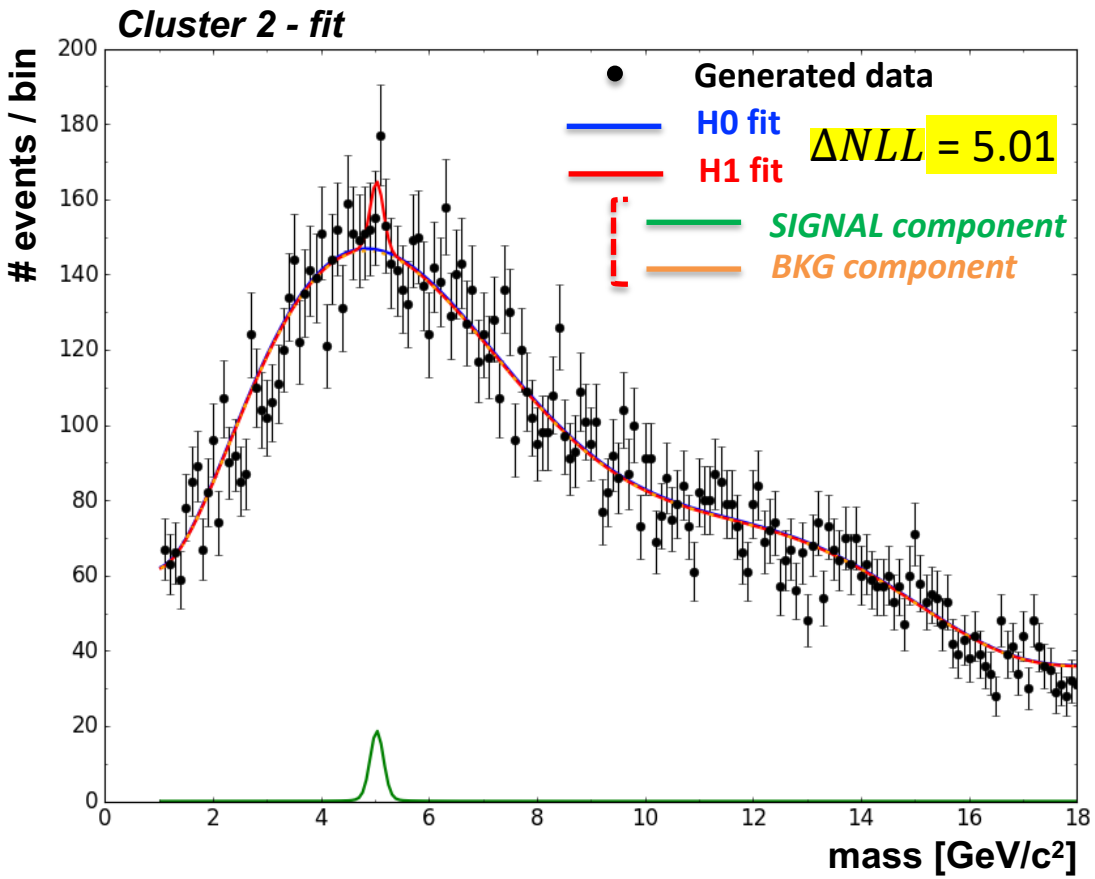
- mass values (***m***) are changed scanning the whole cluster;
- width value (***Γ***) is changed from 1MeV to the whole cluster width [anyway always limited to 300MeV] ;
- resolution value (***σ***) is varied according to the **function of the resonance** mass (shown earlier)

Pseudo-experiments & LEE - scanning technique [step 8/fit 2]

The **scanning technique** has been configured on the basis of a **clustering approach** and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do not miss any relevant fluctuation

B) Do not select too many small fluctuations



The procedure:

8) For **each cluster** the **Alternative Hypothesis H1** fits are performed with the ***polynomial H0*** **+** a ***Convolution*** of a BW (signal) and a Gaussian (resolution) for the peak.

A set of fits is performed **changing the** range & the starting values of the **parameters** (***m***, ***Γ***, ***σ***):

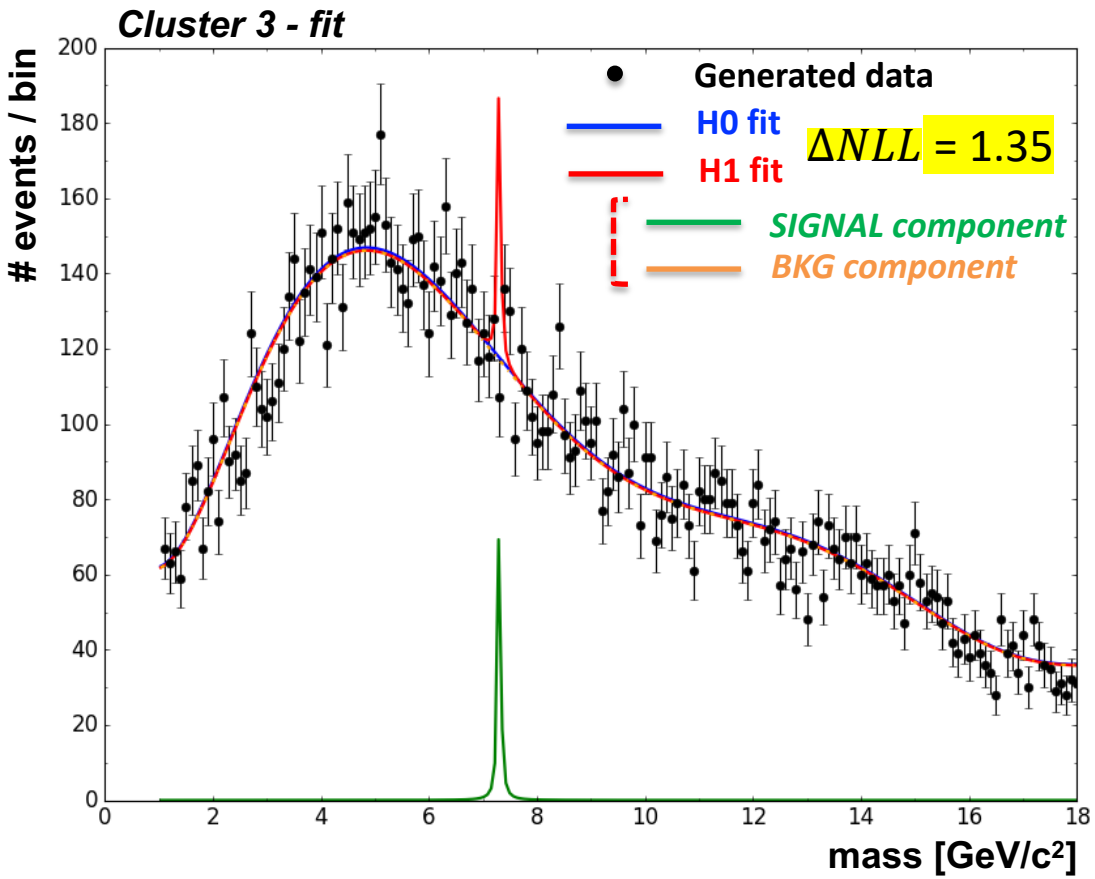
- mass values (***m***) are changed scanning the whole cluster;
- width value (***Γ***) is changed from 1MeV to the whole cluster width [anyway always limited to 300MeV] ;
- resolution value (***σ***) is varied according to the **function of the resonance** mass (shown earlier)

Pseudo-experiments & LEE - scanning technique [step 8/fit 3]

The **scanning technique** has been configured on the basis of a **clustering approach** and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do not miss any relevant fluctuation

B) Do not select too many small fluctuations



The procedure:

8) For **each cluster** the **Alternative Hypothesis H1** fits are performed with the ***polynomial H0*** **+** a ***Convolution*** of a BW (signal) and a Gaussian (resolution) for the peak.

A set of fits is performed **changing the** range & the starting values of the **parameters** (***m***, ***Γ***, ***σ***):

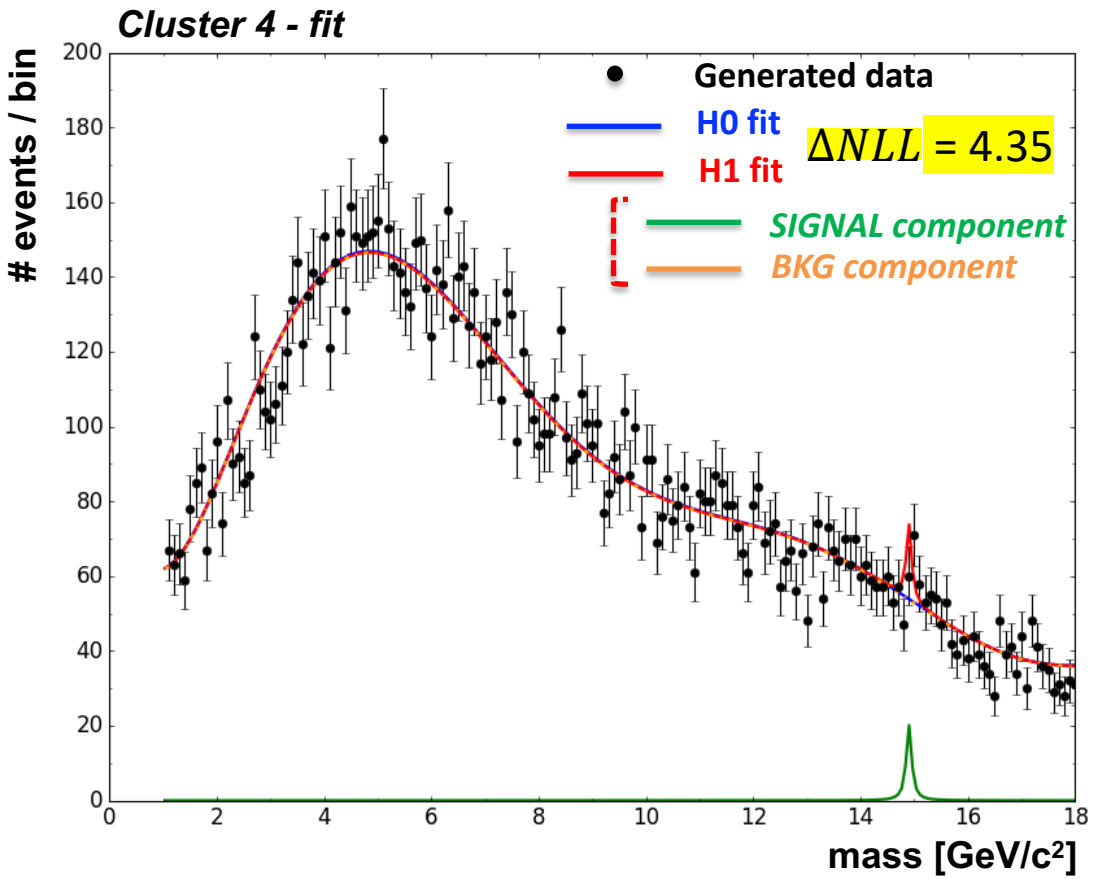
- mass values (***m***) are changed scanning the whole cluster;
- width value (***Γ***) is changed from 1MeV to the whole cluster width [anyway always limited to 300MeV] ;
- resolution value (***σ***) is varied according to the **function of the resonance** mass (shown earlier)

Pseudo-experiments & LEE - scanning technique [step 8/fit 4]

The **scanning technique** has been configured on the basis of a **clustering approach** and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do not miss any relevant fluctuation

B) Do not select too many small fluctuations



The procedure:

8) For **each cluster** the **Alternative Hypothesis H1** fits are performed with the ***polynomial H0*** **+** a ***Convolution*** of a BW (signal) and a Gaussian (resolution) for the peak.

A set of fits is performed **changing the** range & the starting values of the **parameters** (***m***, ***Γ***, ***σ***):

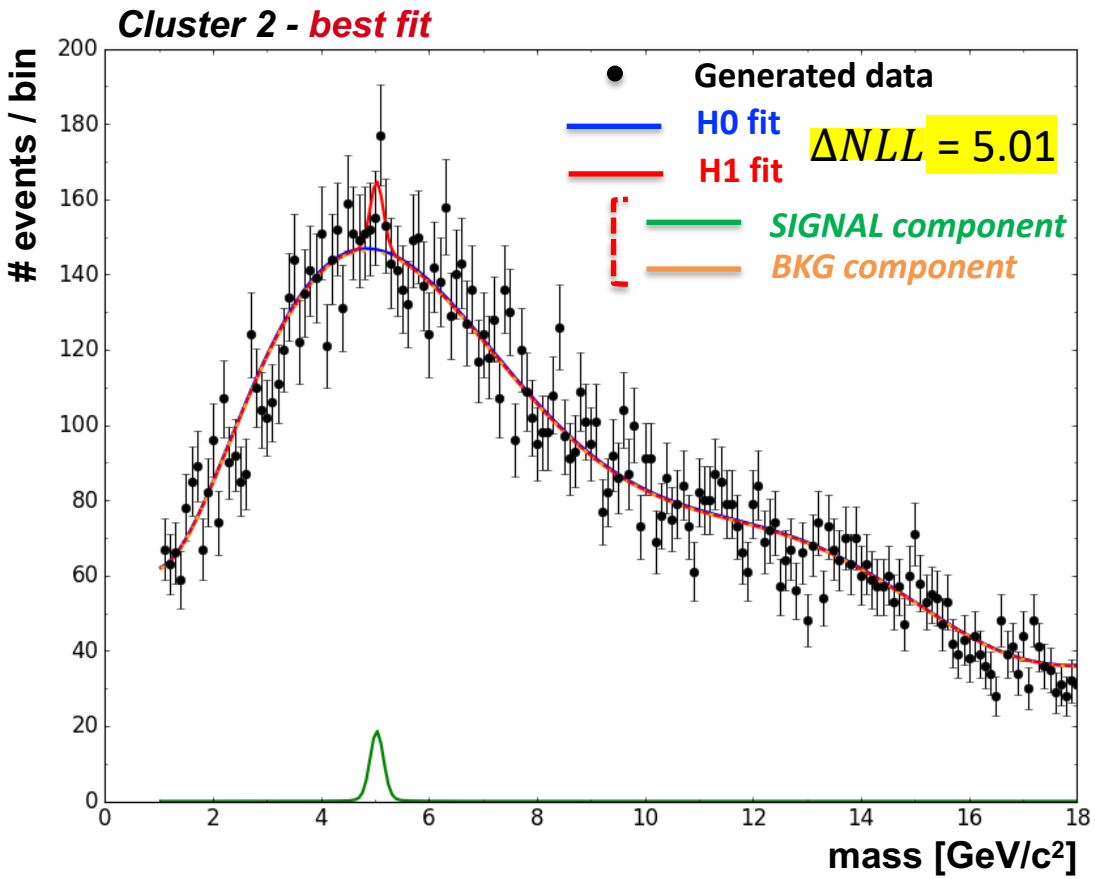
- mass values (***m***) are changed scanning the whole cluster;
- width value (***Γ***) is changed from 1MeV to the whole cluster width [anyway always limited to 300MeV] ;
- resolution value (***σ***) is varied according to the **function of the resonance** mass (shown earlier)

Pseudo-experiments & LEE - scanning technique [step 8/best fit]

The **scanning technique** has been configured on the basis of a **clustering approach** and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do not miss any relevant fluctuation

B) Do not select too many small fluctuations



The procedure:

8) For **each cluster** the **Alternative Hypothesis H1** fits are performed with the ***polynomial H0*** **+** a ***Convolution*** of a BW (signal) and a Gaussian (resolution) for the peak.

A set of fits is performed **changing the** range & the starting values of the **parameters** (***m***, ***Γ***, ***σ***):

- mass values (***m***) are changed scanning the whole cluster;
- width value (***Γ***) is changed from 1MeV to the whole cluster width [anyway always limited to 300MeV] ;
- resolution value (***σ***) is varied according to the **function of the resonance** mass (shown earlier)

Scanning technique : Working Point choice

Once defined the scanning technique, the next step is to tune the parameters of the procedure

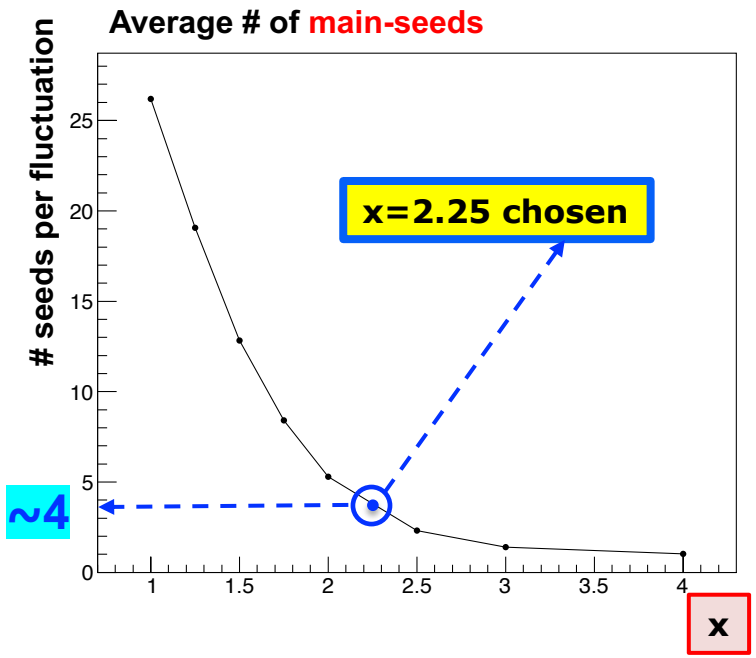
- x (main-seed threshold),

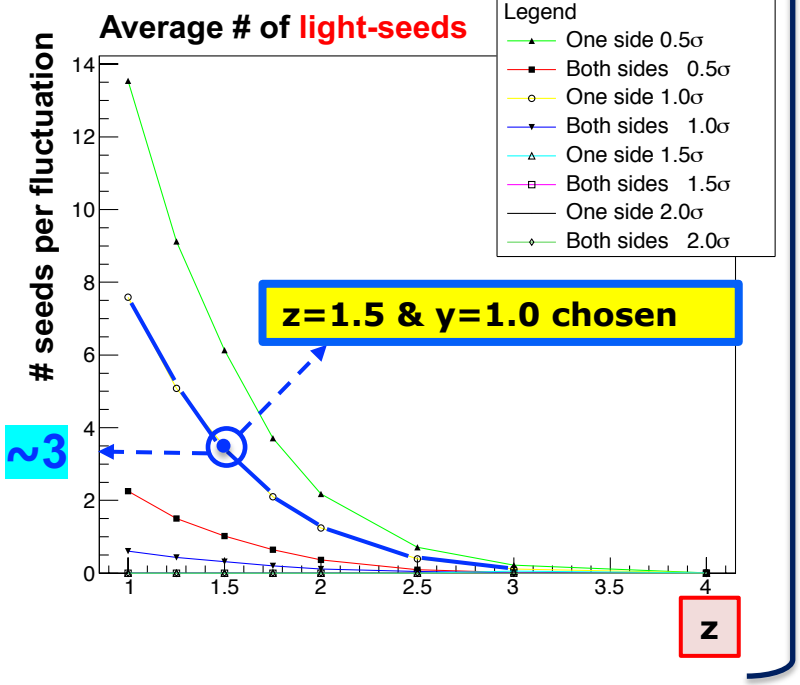
- y (light-seed threshold),

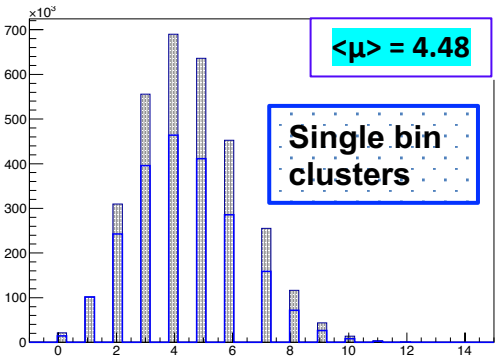
- z (sided-seed threshold) in order to fullfill the requirements (A) & (B).

A set of **1M** toys were produced to estimate the mean value of the distribution...

... of the number of main- and light-seeds per single fluctuation.







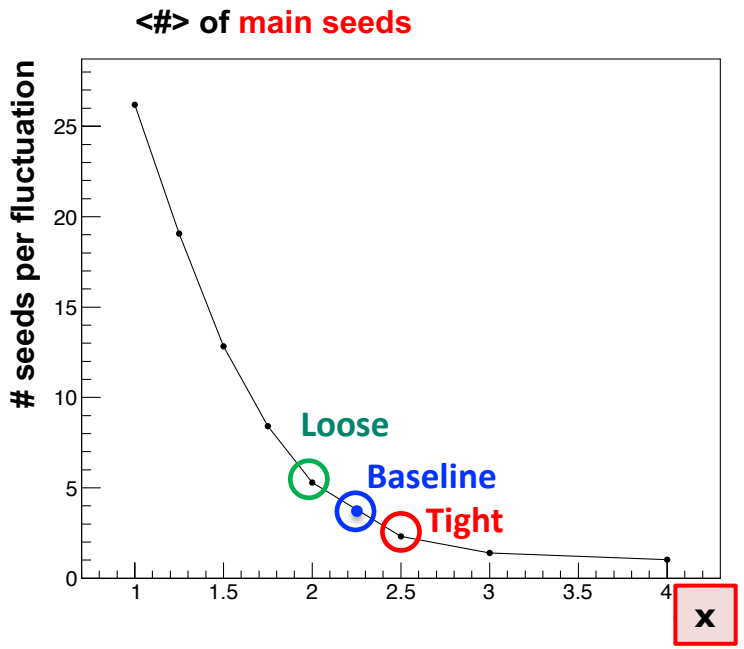
WP(x, y, z)=(2.25, 1.0, 1.5)

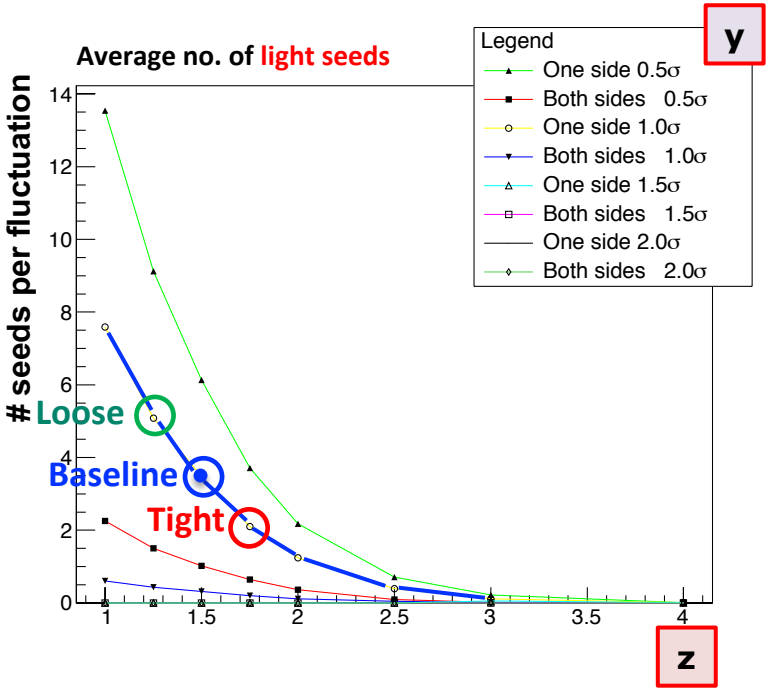
« Baseline»

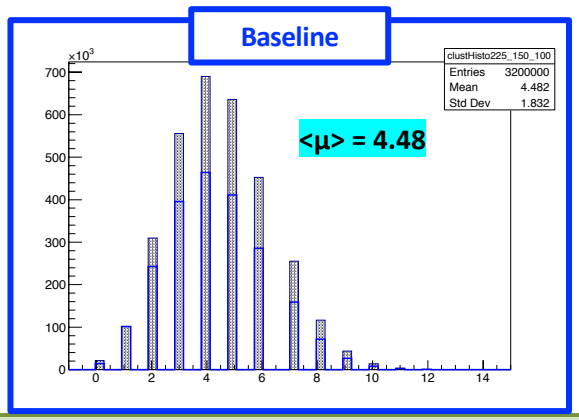
This WP guarantees ~4-5 clusters in average per Toy MC and that at least 1 cluster is found in more than 99% of fluctuations in order to perform at least 1 *H1* fit.

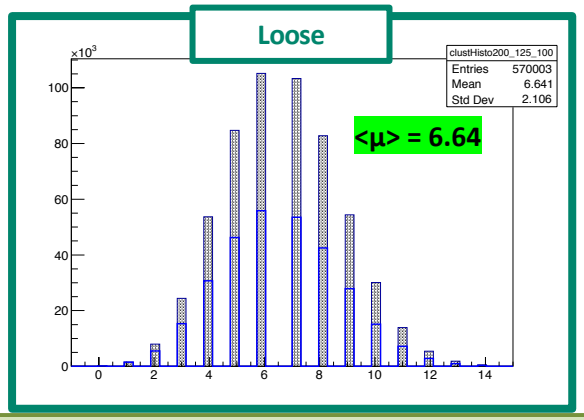
Scanning technique : systematic uncertainties

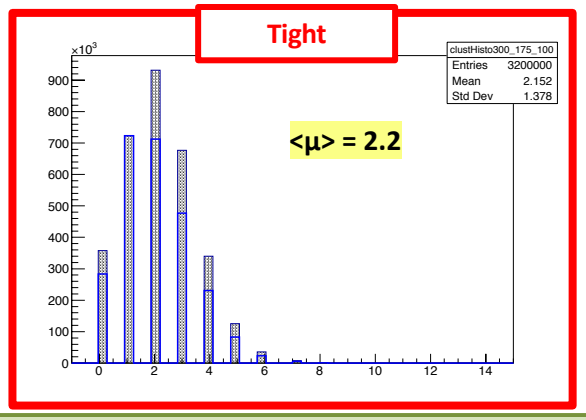
To study the possible **systematic uncertainties** of this clustering method we have ***selected*** also two other combinations of (x,y,z): one **WP *looser*** than the selected one and one **WP *tighter***. In addition, to avoid any possible influence of statistical fluctuations, we have run the MC Toys fitting procedure **three times** for the three different set of cuts on ***the same* set of MC toys fluctuations (**previously independently generated).







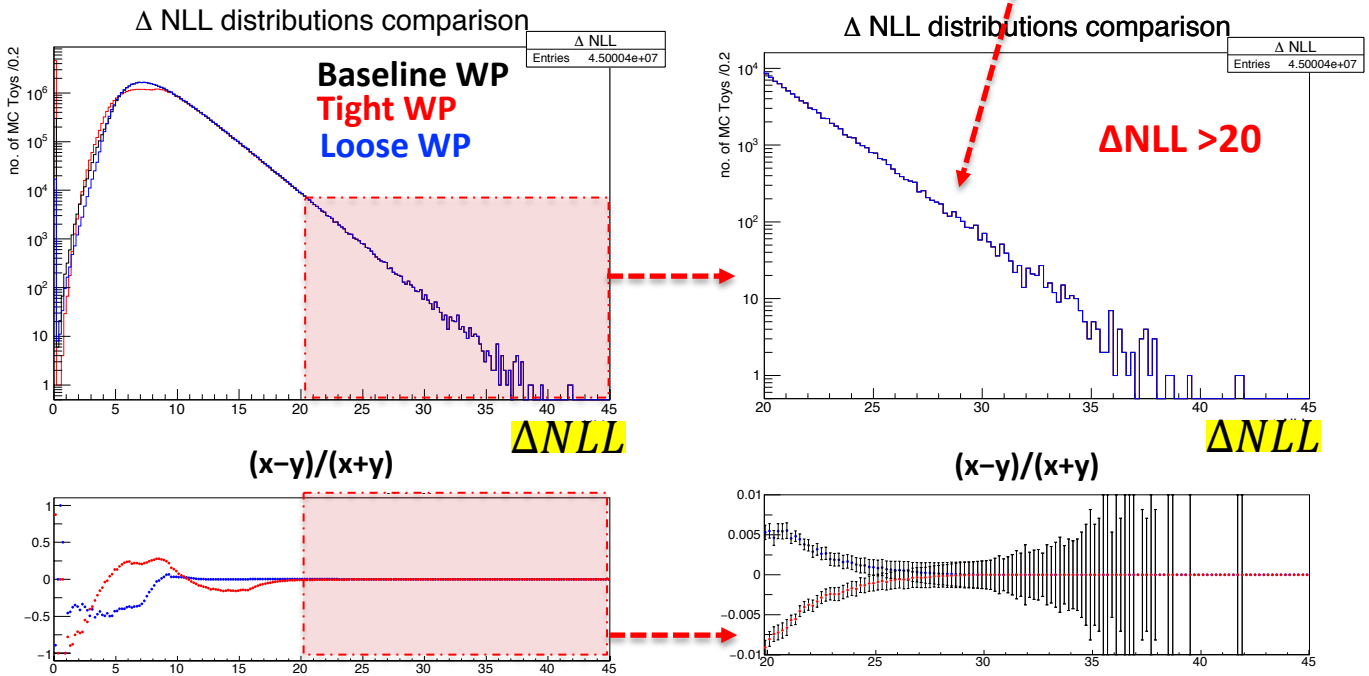




LEE with Scanning technique : results & systematics - I

The resulting distributions from **45M** common MC Toys fluctuations are shown superimposed and compared.

By focusing on the **region of interest** for the estimation of the statistical significance, **i.e. the tail of the ΔNLL distribution (ΔNLL >20)**, it is evident that there is **no relevant difference among the 3 configurations :**



This can furtherly be appreciated by inspecting the normalized deviations **(x−y)/(x+y)** of the other two distributions w.r.t. the baseline distribution

LEE with Scanning technique : results & systematics - II

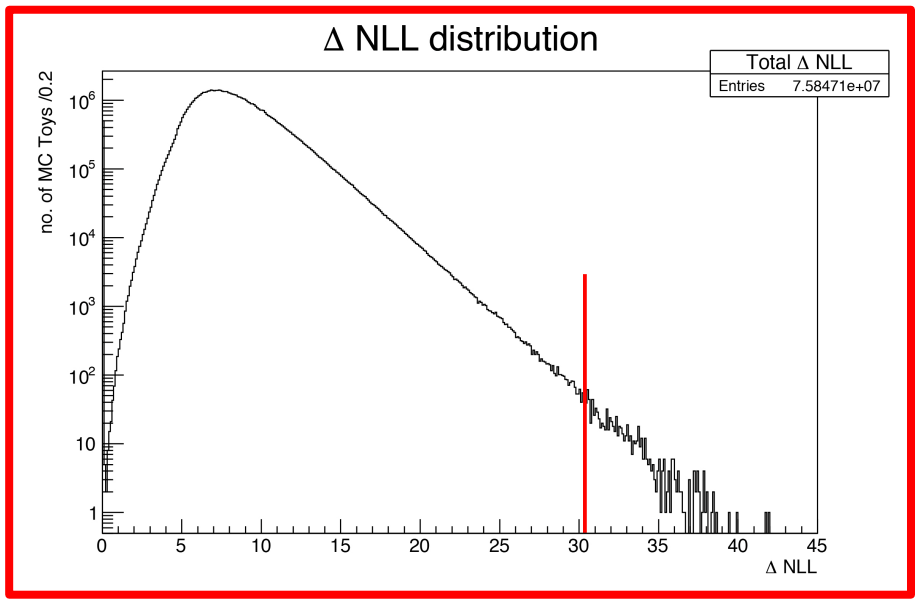
- Also we can examine the **estimated global significances for the *p*-values corresponding to different values of local significances :**

Clustering configs.	$< fit_{H1} >$	f_{nofit}
Tight (3.00, 1.75, 1.00)	2.2	∼10%
Baseline (2.25,1.50, 1.00)	4.5	∼1%
Loose (2.00, 1.25, 1.00)	6.6	0.1%

Local Significance	4.0σ	4.5σ	5.0σ	5.5σ	6.0σ
Tight (3.00, 1.75, 1.00)	2.21	2.91	3.58	4.22	4.87
Baseline (2.25,1.50, 1.00)	2.20	2.91	3.58	4.22	4.87
Loose (2.00, 1.25, 1.00)	2.20	2.91	3.58	4.22	4.87

Conclusion: the systematic uncertainty on the *p*-values associated to the method is negligible

- The **baseline configuration** has been run on about **76M** pseudo experiments and the **ΔNLL** distribution is shown with the superimposed **red line** indicating the ΔNLL data value for our **original pseudo-data:**



The **global p-value** is then estimated by

$$p = \int_{\Delta NLL_{data}}^{\infty} f(\Delta NLL) d(\Delta NLL) \simeq \frac{9.820 \cdot 10^2}{7.584 \cdot 10^7} \simeq 1.295 \cdot 10^{-5}$$

... which corresponds to a **global stat. significance**

$$Z\sigma = \Phi^{-1}(1-p)\sigma \simeq 4.22\sigma$$

Trial Factors

Alternatively, how to avoid millions of MC toys even if run exploting the acceleration of GPUs ?

An approximate way to determine the global significance taking into account the LEE relies on the asymptotic behaviour of likelihood ratio estimators. The correct factor that needs to be applied to the local significance in order to obtain the global one is called *trial factor* :

$$p^{glob} \approx f * p^{loc}$$

Trial factors for the look elsewhere effect in high energy physics

Eilam Gross, Ofer Vitells[a]

Eur. Phys. J. C70 (2010) 525

The trial factor is related to the peak width, which may be dominated by the experimental resolution, if the intrinsic width is relatively small.

When the mass is determined from data (typical for LEE), an empirical evaluation, that can be used as rule of thumb, gives (*) :

(*) https://www.birs.ca/workshops/2010/10w5068/files/gross.pdf

$$f \approx k \frac{search\ mass\ range}{mass\ resolution} \equiv \frac{1}{3} \cdot \frac{\Delta m}{\sigma(m)}$$

For the previous considered case: $f \approx \frac{1}{3} \cdot \frac{18GeV}{60MeV} = \frac{18 \cdot 10^3}{180} \cong \mathbf{100}$

... which makes sense considering that from **5σ** ($p \cong 2.87 \cdot 10^{-7}$) to **4σ** ($p \cong 3.17 \cdot 10^{-5}$) implies a factor 110!

Gross-Vitells method for global statistical significance

G.&V. proposed a method to estimate an upper limit for the global p-value when the signal hypothesis (H_1) depends on *s* parameters that are undefined under the null hypothesis (H_0).

The global test statistic is (the one introduced few slides back**)** $q^{glob} = q\left(\hat{\hat{\vec{\theta}}}, \mu = 0\right)$

... where $\hat{\hat{\vec{\theta}}} \equiv \left(\hat{m}, \hat{\Gamma}\right)$ [or simply $\hat{\theta} = \hat{m}$ if $\Gamma \ll \sigma_{RES}$ (known, for intance from simulation]

Set maximizing $q\left(\vec{\theta}, \mu = 0\right)$

Gross-Vitells method for global statistical significance

G.&V. proposed a method to estimate an upper limit for the global p-value when the signal hypothesis (H_1) depends on *s* parameters that are undefined under the null hypothesis (H_0).

The global test statistic is (the one introduced few slides back) $q^{glob} = q\left(\hat{\hat{\theta}}, \mu = 0\right)$

... where $\hat{\hat{\theta}} \equiv \left(\hat{m}, \hat{\Gamma}\right)$ [or simply $\hat{\theta} = \hat{m}$ if $\Gamma \ll \sigma_{RES}$ (known, for intance from simulation]

Set maximizing $q\left(\vec{\theta}, \mu = 0\right)$

It is possible to demonstrate that the probability that q^{glob} is greater than a given value c is bound by the inequality (that can be considered - asymptotically - as an equality):

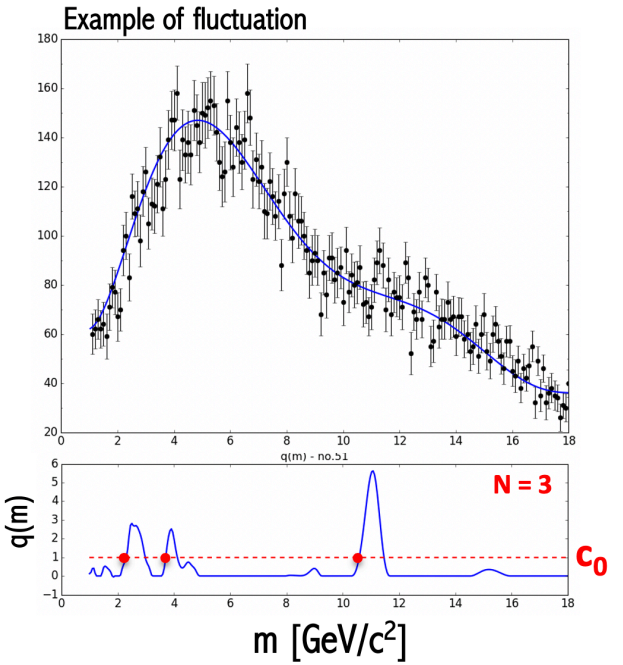
$$p^{glob} = P\left(q\left(\hat{\hat{\theta}}, \mu = 0\right) > c\right) \leq P\left(\chi^2 > c\right) + \langle N_c \rangle$$

Term (related to the local *p*-value) that is a cumulative χ^2 distribution that comes from an asymptotic approximation as a $\chi^2_{v=1}$ (with 1 degree of freedom) of

$$q^{loc} = q\left(\vec{\theta}, \mu = 0\right)$$

Average number of upcrossings i.e. the expected number of times that the local test statistic ***q^loc*** crosses an horizontal line at a given level ***q=c*** with a positive derivative.

It acts like a correction to the Wilks+Cowan local p-value !



Gross-Vitells method : scaling law

- The $<N_c>$ can be typically evaluated using MC toys as average value over a large number of samples. Since its value could be very small (depending on the level of c , and the details of the statistical model), in such cases very large MC samples would be required for a precise numerical evaluation. Luckily a scaling law allows to extrapolate a value $\langle N_{c_0} \rangle$ evaluated at a different level c_0 to the desired level c:

$$P\left(q(\hat{\theta}) > c\right) \leq P\left(\chi^2_s > c\right) + \boxed{\langle N(c_0) \rangle \left(\frac{c}{c_0}\right)^{(s-1)/2} e^{-(c-c_0)/2}} \qquad (*)$$

At this point it is possible to evaluate $\langle N_{c_0} \rangle$ by generating a not too large number of MC toys.

Conclusion: it is possible to move from local to global *p*-value using the asymptotically approximation: $p^{glob} = p^{loc} + \langle N_{c_0} \rangle e^{-(c-c_0)/2}$

Gross-Vitells method : scaling law

- The $<N_c>$ can be typically evaluated using MC toys as average value over a large number of samples. Since its value could be very small (depending on the level of c , and the details of the statistical model), in such cases very large MC samples would be required for a precise numerical evaluation. Luckily a scaling law allows to extrapolate a value $\langle N_{c_0} \rangle$ evaluated at a different level c_0 to the desired level c:

$$P(q(\hat{\theta}) > c) \leq P(\chi^2_s > c) + \boxed{\langle N(c_0) \rangle \left(\frac{c}{c_0}\right)^{(s-1)/2} e^{-(c-c_0)/2}} \qquad (*)$$

At this point it is possible to evaluate $\langle N_{c_0} \rangle$ by generating a not too large number of MC toys.

Conclusion: it is possible to move from local to global *p*-value using the asymptotically approximation: $p^{glob} = p^{loc} + \langle N_{c_0} \rangle e^{-(c-c_0)/2}$

- We set up a procedure [within `GooFit` framework] to estimate $< N(c_0) >$ for **our pseudo-data configuration.** **10k** toys were produced and for each toy a **complete scan** (in **1000** steps) of the mass spectrum is performed.

The procedure took **~3days** on a single GPU, the time equivalent of **~4-5M MC toys (for LEE)** produced.

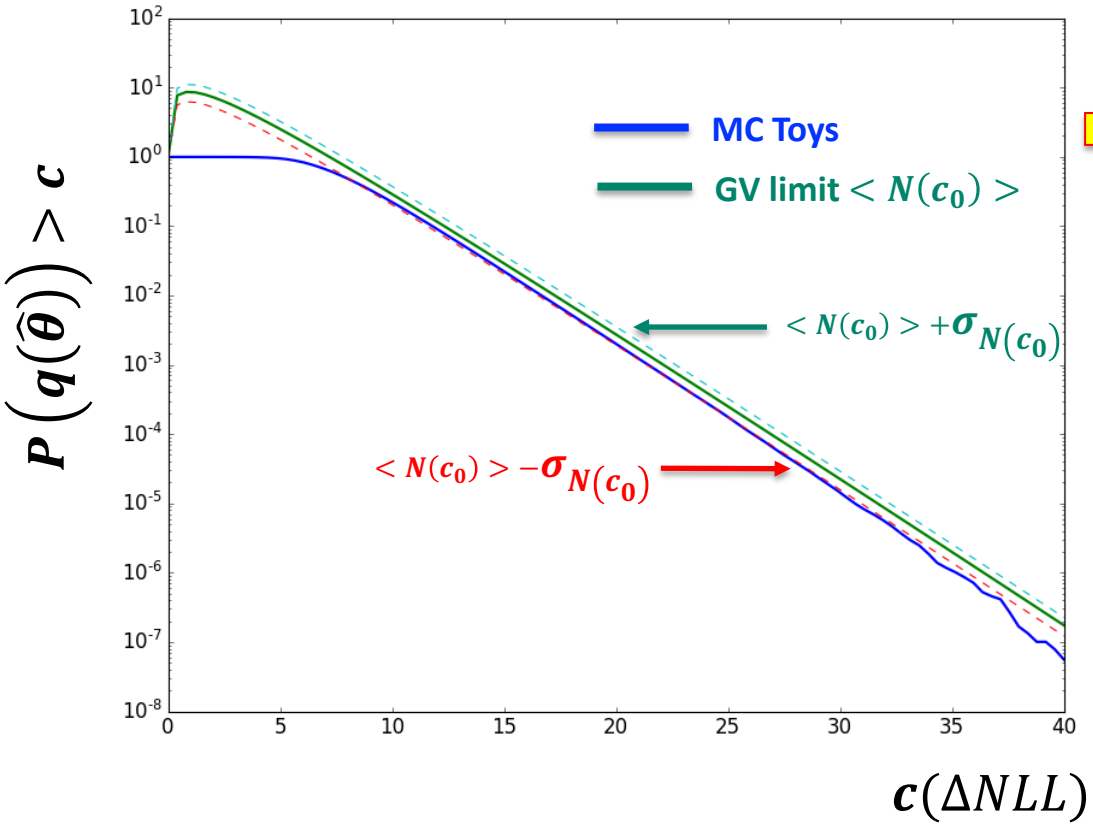
The Upper Limit [G-V result] can be evaluated from (*) with $< N(c_0) > = 7.3$ (for c_0=s-1=1)

The rms of the distribution is given by (useful in next slide): $\sigma_{N(c_0)} = 2.4$

Comparison with G-V method

- Thus we can compare the $\boldsymbol{P\left(q(\hat{\theta})\right)}$ computed from the ΔNLL distribution obtained with MC Toys (in the **baseline** configuration) with the upper limit just **estimated** with the **G-V method**.

In the case of the MC Toys, $\boldsymbol{P\left(q(\hat{\theta})\right)(c)}$ is calculated as the integral $P(q(\hat{\theta}))(c) = \int_c^{\infty} f(\Delta NLL)d(\Delta NLL)$



The G-V Upper Limit result to be conservative w.r.t the MC toys and, for a given ΔNLL value, always underestimate the global statistical significance (see table):

Local Sig.	4.0σ	4.5σ	5.0σ	5.5σ	6.0σ
GV method	2.09	2.82	3.48	4.10	4.71
MC Toys	2.20	2.91	3.58	4.22	4.87

The Upper Limit is perfectly compatible with the results with the MC toys clustering procedure

Summary

- With the advent of GPU acceleration in the field of scientific computation - possible on heterogeneous computing platforms, nowadays available at Science Data Centers - **the pseudo experiment (frequentistic) approach** is **feasible/reliable,** once implemented within the `GooFit` framework, **to estimate the global (local) *p*-value of a signal** within few days **[~1.5M (5M) toys/day** can be produced with a single GPU (nVidia TeslaK40) equipped server **]**

- Thanks to the striking speed-ups allowed by GPUs in dealing with the fitting tasks of MC toys, it was possible to **explore (and confirm) the validity/applicabiity of asymptotic behaviour of likelihood-ratio-based test statistics exploited in statistical methods introduced by Cowan *et al.* & Gross and Vitells, at the beginning of the LHC era (2010-2011) and nowadays commonly used in HEP.**

Bibliography - I

With reference to this work:

- **GPUs for statistical data analysis in HEP: a performance study of GooFit on GPUs vs. RooFit on CPUs**
  CMS Collaboration • Alexis Pompili (Bari Polytechnic and INFN, Bari) et al. (Nov 21, 2016)
  Published in: *J.Phys.Conf.Ser.* 762 (2016) 1, 012044 • Contribution to: ACAT 2016

- **Performance studies of $GooFit$ on GPUs vs $RooFit$ on CPUs while estimating the statistical significance of a new physical signal**
  Adriano Di Florio (Bari Polytechnic and INFN, Bari) (Nov 22, 2017)
  Published in: *J.Phys.Conf.Ser.* 898 (2017) 7, 072036 • Contribution to: CHEP 2016

- **Performance studies of GooFit while estimating the global statistical significance of a new physical signal**
  CMS Collaboration • Alexis Pompili (Bari Polytechnic and INFN, Bari) et al. (Oct 18, 2018)
  Published in: *J.Phys.Conf.Ser.* 1085 (2018) 4, 042005 • Contribution to: ACAT 2017

- **Estimation of global statistical significance of a new signal within the GooFit framework on GPUs**
  Adriano Di Florio (Bari U. and INFN, Bari) (Jul 15, 2019)
  Published in: *PoS* Confinement2018 (2019) 229 • Contribution to: Confinement XIII, 229

Bibliography - II

- With reference to **`GooFit`** :
  - Andreassen R, Meadows B T, de Silva M and Sokoloff M D *J. Phys.: Conf. Series* **513**, 052003 (2014)
  - Schreiner H F et al. *arXiv:1710.08826* (2017) [see also Proceedings of ACAT2017]
  - If you are interested to start **learning** & **working** with *GooFit*, it source code lives in a GitHub repository (https://github.com/GooFit) GitHub and its applications go **way further** than statistical significance estimation. Nowadays is a "common" fitting tool particularly usefull when dealing with (**multidimensional) unbinned likelihood** fit at **high statistics**).
- Useful/inspirational **Statistics textbook** of reference:
  - **Glen Cowan**, ***Statistical Data Analysis***, **Oxford Science Ed., 1998**
  - **Luca Lista**, ***Statistical Methods for Data Analysis in Particle Physics***, **Springer Ed., 2018 (2nd ed.)**

Aknowledgments

We are grateful for valuable support to all the people involved in the maitainance of the High Performance Cluster hosted by the ReCas-Bari Data Center, and especially to its manager Giacinto Donvito.