PROCEEDINGS OF SCIENCE

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

Alexis Pompili^{a,b,*} **and Adriano Di Florio**^{a,c}

^{a}*Istituto Nazionale di Fisica Nucleare - Sezione di Bari, via E.Orabona 4, 70126 Bari, Italy*
^{b}*Dipartimento Interateneo di Fisica, Università degli Studi di Bari, via G.Amendola 175, 70126 Bari, Italy*
^{c}*Dipartimento Interateneo di Fisica, Politecnico di Bari, via E.Orabona 4, 70126 Bari, Italy*

E-mail: alexis.pompili@ba.infn.it, adriano.diflorio@ba.infn.it

GPUs represent one of the most sophisticated and versatile parallel computing architectures that have recently been introduced in the High Energy Physics (HEP) field. `GooFit` is an open source tool interfacing ROOT/RooFit to the CUDA platform that allows to manipulate probability density functions and perform fitting tasks. The striking performances and computing capabilities of GPUs, in comparison to traditional CPU cores, have been exploited in the application of a high-statistics pseudo-experiment method implemented in `GooFit`, with the purpose of estimating the local or global statistical significance of a physics signal, already known or new respectively. When dealing with an unexpected new signal, a global significance must be estimated to take into account the Look-Elsewhere-Effect and this is accomplished coupling a clustering-based scanning technique to the pseudo-experiments method, also without introducing any relevant systematic uncertainty.

By means of these tools it has been possible to investigate the approximation characterizing modern, and currently widely used, statistical methods. In particular two studies have been carried out: 1) the asymptotic behaviour of a likelihood ratio test statistics (Cowan-Cranmer-Gross-Vitells) has been investigated while estimating the local statistical significance of a known signal, 2) the approximation of the Gross-Vitells method (*trial factors*) has been explored while estimating the global statistical significance of a new signal.

These studies have been collected and presented here coherently with a didactic approach. Indeed this work is currently used in lectures about Statistics for Data Analysis. However the presented results can be a useful reference for the confirmation - by means of GPUs - of the validity of few asymptotic formulas/methods now commonly used in HEP.

**** 10th International Workshop on Charm Physics (CHARM2020), ****
**** 31 May - 4 June, 2021 ****
**** Mexico City, Mexico - Online ****

[*] Speaker

PoS(CHARM2020)002

1. Introduction

The HEP researchers often have to deal with "signals" that highlight a discrepancy with the current theoretical models predictions. These signals can be either *already known* or *completely new*. To claim their confirmation or first observation respectively, their statistical significance, correspondingly *local* or *global*, must be assessed. Modern statistical methods commonly used to estimate statistical significances have been introduced at the beginning of the LHC era, when scientific computing with GPUs was only starting to be explored and not yet introduced in HEP as a resource. Nowadays the capabilities of GPU acceleration are being enough commonly used in HEP (in data analyses as well as in algorithms for event reconstruction and particle identification).

The word *GPU-accelerated computing* refers to an enhancement of application performances that can be obtained by offloading compute-intensive portions of the code to the GPU, while the remaining parts still runs on the CPUs. The computing capabilities are enhanced once a sequence of elementary arithmetic operations are performed in parallel on a huge amount of data. In the HEP context `GooFit` [1] is an under development open source data analysis tool, used in applications for parameter estimation, that interfaces the commonly used ROOT/RooFit to the CUDA [2] parallel computing platform on nVidia's GPUs (it also supports OpenMP). The Probability Density Function (PDF) evaluation on large datasets is typically the bottleneck in the MINUIT algorithm. `GooFit` acts as an interface between the MINUIT minimization algorithm and a parallel processor which allows a PDF to be evaluated in parallel. Fit parameters are estimated at each negative-log-likelihood (NLL) minimization step on the *host side* (CPU) while the PDF/NLL is evaluated on the *device side* (GPU). Applications using, even recursively, a series of several fits with complicated PDFs, can evidently take advantage of the GPU acceleration by using `GooFit`. Once implemented within `GooFit`, Monte Carlo pseudo-experiments represent a very good example of an application with these characteristics, as discussed in the next sections. Description and details about `GooFit` can be found elsewhere and especially in Refs. [1].

The studies, that have been here collected and coherently discussed, were carried out along the period 2015-2018 and already presented in a few conferences [3] but in a disaggregated way.

2. Pseudo-experiments method for local statistical significance estimation

Many searches for new physical phenomena look for a peak in a distribution that typically is a reconstructed invariant mass and the peaking structure may represent a resonance/particle. The location (mass) of a peak (particle) is known in some cases such as 1) in searches for rare decays of a known particle, or 2) when an experiment is looking to confirm a new particle/claimed by another experiment, and/or 3) when one or more theoretical models predicts it. The local statistical significance associated to a peak (at m_0) can be estimated in terms of a local p-value, expressed as

$$p(m_0) = \int_{q_{obs}(m_0)}^{\infty} f(q|m_0, \mu = 0) dq \tag{1}$$

Monte Carlo pseudo-experiments (MC toys) 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. To test the computing capabilities of GPUs with respect to CPU cores, a high-statistics MC

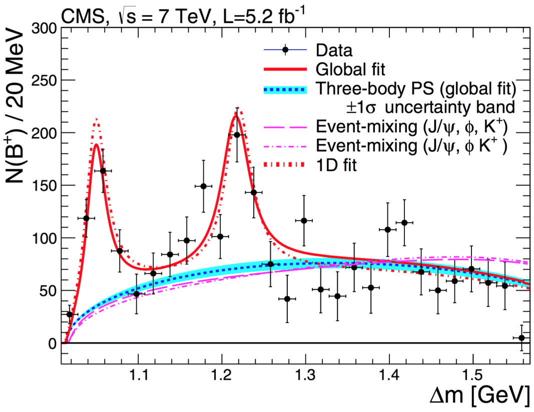

Figure 1: Fits to the background-subtracted $J/\psi\phi$ invariant mass in the $B^+ \to J/\psi\phi K^+$ decay, as presented in Ref.[4]. The left peaking structure, close to the kinematic threshold and laying on a residual phase-space background, was already observed by the CDF experiment. Its significance has been re-estimated - in this study - with a MC toys technique implemented in the `GooFit` framework.

toys technique has been implemented both in ROOT/RooFit and `GooFit` frameworks [3] with the aim to estimate a *p-value* and specifically the local statistical significance of the structure observed by the CMS experiment close to the kinematic threshold of the $J/\psi\phi$ invariant mass in the $B^+ \to J/\psi\phi K^+$ decay, presented in Ref. [4] and reported in Fig.1 as well.

A single toy fit cycle consists in the following sequence of steps:

1) generation of fluctuated background binned distribution according to the 3-body phase-space model (the number of entries are fixed to that in the data thus ignoring Poisson fluctuations);

2) a Binned Maximum Likelihood (BML) fit is performed with the phase-space model (null hypothesis H_0); the number of entries are fixed to that in the data thus ignoring Poisson fluctuations;

3) 8 BML fits are performed by adding to the phase-space a Voigtian model truncated to account for the kinematic threshold (alternative hypothesis H_1); the Gaussian resolution function has fixed width ($2MeV$) and the signal yield is constrained to be positive. For each bin the PDF value is estimated by integration over the bin since the signal is steep with respect to the bin size. The 8 H_1 fits differ by the starting values (2 masses and 4 widths) within the region of interest defined from the available values from the CDF experiment.

4) For each fit a $\Delta\chi^2$ value is calculated with respect to the H_0 fit and the best (higher) value is chosen among the 8 H_1 fits. The final $\Delta\chi^2$ (the test statistic) distribution, $f(\Delta\chi^2)$, is obtained over the whole sample of MC toys and is shown in Fig. 2.

The MC toys production was stopped after 57.7M toys, once a fluctuation with $\Delta\chi^2 > \Delta\chi^2_{obs}$ was found (Fig.3). The p-value is estimated to be:

$$p = \int_{\Delta\chi^2_{obs}}^{\infty} f(\Delta\chi^2) d(\Delta\chi^2) \simeq (57.7 \cdot 10^6)^{-1} \simeq 1.73 \cdot 10^{-8}$$

By the inverse function of the cumulative distribution of the standard Gaussian, this p-value corresponds to the statistical significance $Z\sigma = \Phi^{-1}(1-p)\sigma \simeq 5.52\sigma$, which is compatible with the lower limit of 5σ quoted in Ref. [4], on the basis of 50.5M toys obtained by means of RooFit.

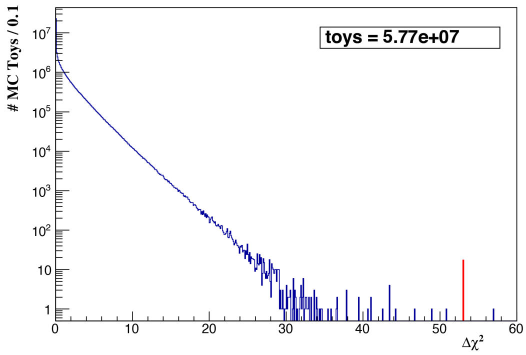

Figure 2: Final $\Delta\chi^2$ distribution with one toy (shown in Fig. 3) exceeding the value observed in the data ($\simeq$ 53.0) marked by a red tick.

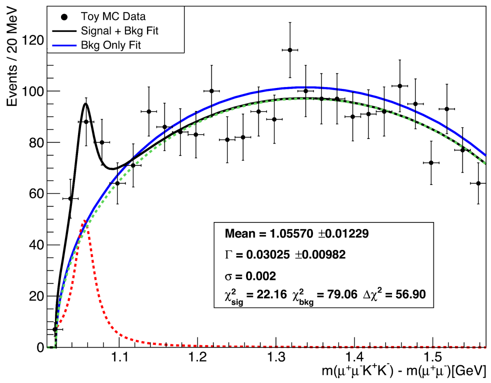

Figure 3: Bkg-only and bkg+signal fits to the MC toy characterized by a $\Delta\chi^2 \simeq 56.9$.

2.1 `GooFit` performances in the execution of the pseudo-experiments task

The used hardware setup consists in two servers, one equipped with two nVidia TeslaK20 and 32 cores (16 + 16 by Hyper-Threading) and the other with one nVidia TeslaK40 and 40 (20 + 20) cores. To efficiently run RooFit MC toys on the 72 CPUs available on the two servers hosting the GPUs, the ROOT/PROOF-Lite tool has been used. On the other hand the *nVidia* Multi Process Service tool allows the execution of - up to 16 - simultaneous processes on the same GPU acting as a scheduler and allowing a balanced full usage of the GPU. The optimized `GooFit` application running on GPUs has provided (see details in Refs. [3]) striking speed-up performances with respect to the RooFit application parallelized on multiple CPUs by means of PROOF-Lite. In particular, from the point of view of the end-user analyst, having at its own disposal all the 72 CPU cores and the three GPUs, it has been measured that 1M of toys can be produced in about 11days with RooFit/PROOF-Lite and in about 6 hours only with `GooFit`/MPS, as shown in Fig.4. For a reference significance $\geq 5\sigma$ a p-value $\leq 2.87 \cdot 10^{-7}$ is needed, namely at least $3.48M$ toys are needed. However, as in the case under study, the significance estimation may require many MC toys more.

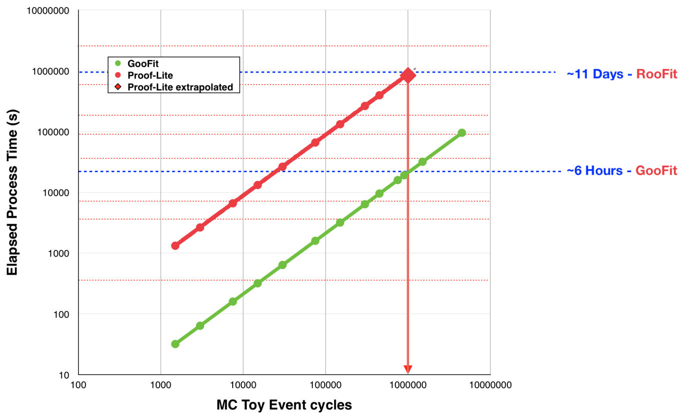

Figure 4: Comparison for the elapsed time employed with two TeslaK20 and one TeslaK40 together as a function of the number of MC toys; `GooFit`/MPS runs 48 concurrent processes while RooFit/PROOF-Lite runs on 72 CPUs. For 1M toys the red diamond point shows the extrapolated time (about 11days) for the RooFit application (red curve). The green curve represents the `GooFit` time.

3. Exploring the applicability of Wilks' theorem and Cowan's asymptotic formula

The Wilks theorem [5] is often used to estimate the p-value associated to a physical signal. Given two hypothesis, the *null* one, H_0, with ν_0 degrees of freedom (dof) and an *alternative* one, H_1, with ν_1 dof, any test statistic t, defined as a likelihood ratio $-2ln\lambda = -2ln(L_{H_0}/L_{H_1})$, 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 $\nu = \nu_1 - \nu_0$ dof, provided that the following regularity conditions hold:

1. H_0 and H_1 are nested (H_1 includes H_0);
2. while $H_1 \longrightarrow H_0$, the H_1 parameters are well behaving, namely well defined and not approaching some limit;
3. asymptotic limit, namely in the enough large data sample regime.

Once this theorem holds, the p-value associated to the signal is

$$p = \int_{t_{obs}}^{\infty} \chi^2_{\nu_1 - \nu_0}(t)dt$$

and the use of pseudo-experiments to estimate the p-value is not needed in principle, even if still suggested. When null hypothesis is background-only and the alternative one is background plus signal, often the above conditions are not all satisfied, and the MC toys are mandatory.

By means of `GooFit`, which massively allows fits over millions of MC toys, it has been enough effortless to explore the (asymptotic) behaviour of a likelihood ratio test statistic in different situations in which the Wilks' theorem may apply or may not apply because its regularity conditions are not satisfied.

As in the previous with the signal parameters in the model of H_1 hypothesis being the mass (m), the width (Γ) and the yield ($\mu \geq 0$), when considering $H_1 \longrightarrow H_0$ not only m and Γ are not well defined but also μ tends to the null limit.

In general the distributions of a test statistic are not predictable and thus need to be extracted from pseudo-experiments. MC toys according to the previously discussed procedure and physics case have been generated for each of the following four cases:

- case (1): m and Γ fixed, μ free;
- case (2): m and Γ fixed, μ free but constrained to be positive;
- case (3): m and Γ free, μ free;
- case (4): m and Γ free, μ free but constrained to be positive.

The $\Delta\chi^2$ distributions for the four cases are reported and superimposed in Fig. 5. Case (4) was the one studied in Section 2 (with much higher statistics). Let us focus on the first two special cases: in case (1) Wilks' theorem must hold whereas in (2) a Cowan's asymptotic formula should apply. Both circumstances have been verified by means of the pseudo-experiments handled with `GooFit`.

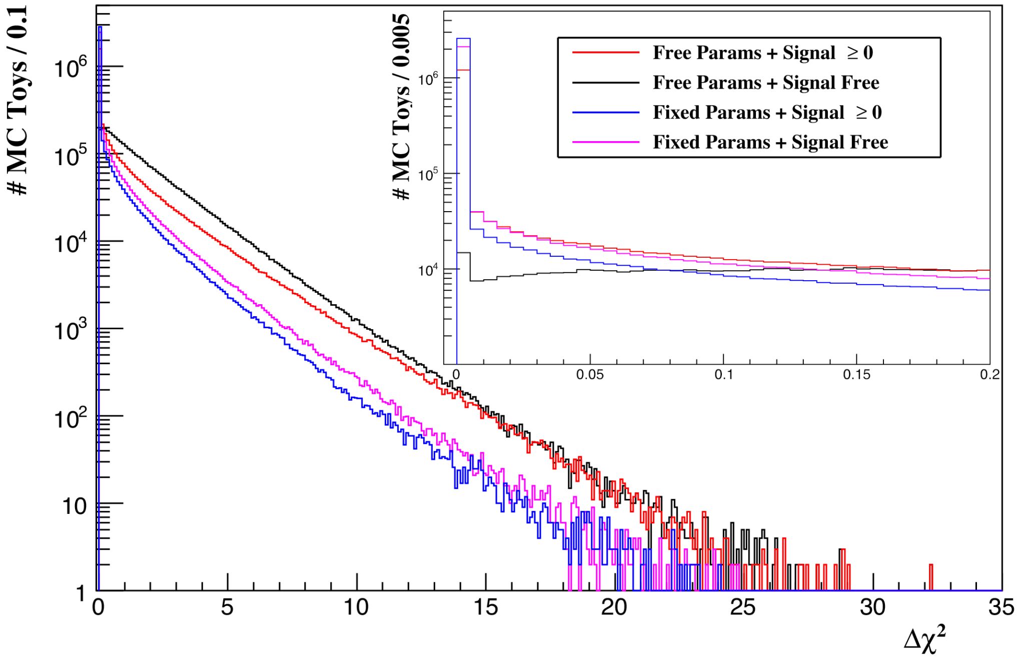

Figure 5: Different test statistic ($\Delta\chi^2$) distributions for the 4 cases discussed in the text, with the same number ($2M$) of MC toys. Case (1) is in fuchsia, case (2) in blue whereas case (4) in red.

3.1 Special case (1) and Wilks' theorem: m and Γ fixed, μ free

Let us consider a likelihood ratio test statistic $t_\mu = -2ln\lambda(\mu)$, where μ is the `strength parameter`, as the basis of the statistical test. This can be a test of $\mu = 0$ with the purpose of establishing the existence of a signal process, thus μ is free to be either positive or negative (and it does not properly represent a signal yield). Following [6], the PDF of the test statistic

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

asymptotically approaches a $\chi^2_{\nu=1}$ distribution, in agreement with the Wilks' theorem and with the difference of degrees of freedom being one.

A fit to the test statistic distribution with a χ^2_ν model has been performed, where the likelihood ratio distribution has been obtained by the fit procedure already discussed in Section 2, provided that the values of mass and width parameters have been set to the CMS (or CDF), while leaving μ free. The best estimate obtained for the number of dof is $\hat{\nu} \simeq 1.014 \pm 0.001$, thus enough close to the theoretical prediction; the goodness of fit is checked using a chi-square test that returns a 11.8% probability [3].

3.2 Special case (2) and Cowan's asymptotic formula: m and Γ fixed, μ free but ≥ 0

Let us now consider the special case of the test statistic t_μ with the purpose to test $\mu = 0$ in a class of models where $\mu \geq 0$ is assumed; rejecting the null hypothesis ($\mu = 0$) leads to the discovery of a signal. In this case, following Ref. [6], the test statistic is $q_0 = -2ln\lambda(0)$ if the estimated signal strength $\hat{\mu} \geq 0$ while is null otherwise, with $\lambda(0)$ being the profile likelihood ratio for $\mu = 0$. Cowan, Cranmer, Gross and Vitells [6] derive analytically that an asymptotic approximation for the PDF of the statistic q_0 under assumption of the background-only ($\mu = 0$) hypothesis is an equal mixture of a delta function at zero and a chi-square distribution for one dof:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi q_0}}e^{-q_0/2}.$$

A fit to the test statistic distribution with a model consisting in a linear combination of a χ^2_ν function and a narrow step function at zero has been performed (Fig. 6), where the likelihood ratio distribution has been obtained by the fit procedure already discussed in Section 2 in the case of the values of mass and width parameters are set to the CMS estimates previously obtained, while leaving μ free. The best estimates obtained for the number of dof and the coefficient/weight in front of the step function are $\hat{\nu} \simeq 0.992 \pm 0.001$ and $\hat{c} \simeq 0.507 \pm 0.001$ respectively, namely close to the approximate theoretical prediction. A chi-square test returns a 3.5% probability for this fit [3].

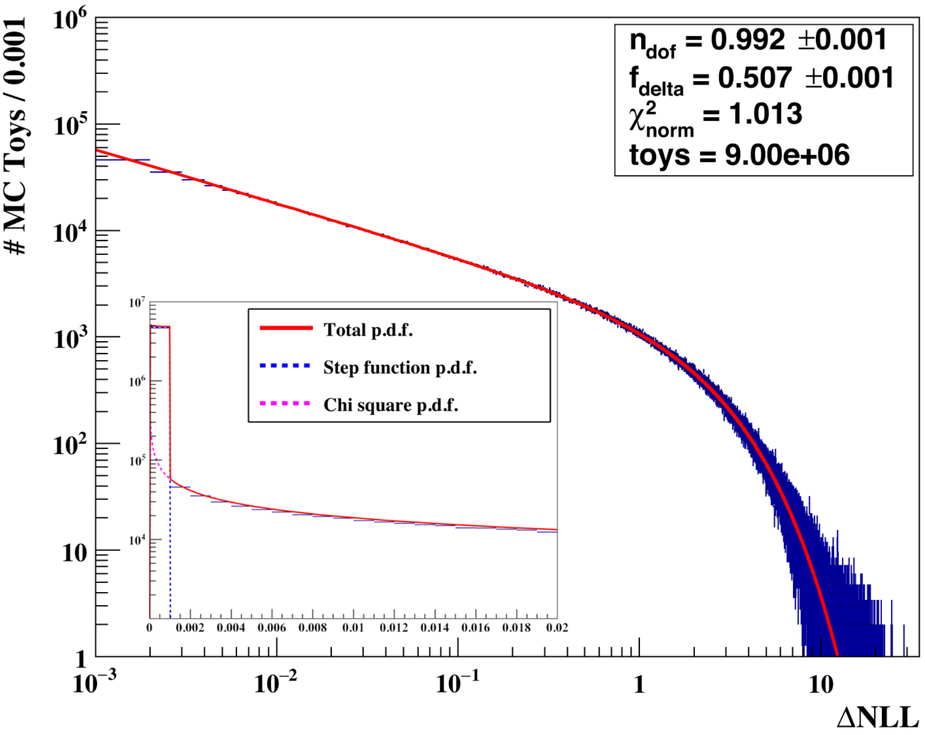

Figure 6: Fit to the ΔNLL distribution for case (2). The fit model has a linear combination of two components: a very narrow step function at zero and a χ^2_ν.

4. Global statistical significance and the Look-Elsewhere-Effect

Unlike Sections 2 and 3, let us now consider the case of searches of new particles whose mass is not predicted by the theory or are not expected at all. If an excess in the data, compared with the background(s) expectation(s), is found at any mass value - in principle produced either by the presence of a real signal or by a background fluctuation - it could be interpreted as a possible signal of a new resonance (in any position of the investigated window of a mass spectrum). In this case the mass is not fixed but estimated from data and the local significance, given by Eq. 1, must be replaced by a global significance of the associated peak. The corresponding global p-value is

$$p(m) = \int_{q_{obs}(m)}^{\infty} f(q|m, \mu = 0)dq \tag{2}$$

where the function $f(q|m, \mu = 0)$ is the PDF of the adopted test statistic q. This p-value gives the probability that a background fluctuation at any mass value, in the mass range of interest, results in a value of q greater or equal the observed value $q_{obs}(m)$. The global p-value is greater than the local one and thus the global significance is lower than the local one. This effect of reduction of significance is called *Look Elsewhere Effect* (LEE) [7][8].

In general, when an experiment is searching for a new signal where one or more parameters of interest ($\vec{\theta}$) are unknown (i.e. both mass and width or other properties of the new state), the global p-value can be determined from the distribution of the test statistic q^{glob} assuming background only hypothesis, given the observed value q^{glob}_{obs}, according to the expression (a generalization of Eq. 2)

$$p^{glob} = \int_{q^{glob}_{obs}}^{\infty} f(q^{glob}|\mu = 0)dq^{glob} \tag{3}$$

where the test statistic $q^{glob} = q(\hat{\vec{\theta}}, \mu = 0)$ is the one corresponding to the the largest values obtainable for the parameters' estimators over the entire parameter range, having denoted with $\hat{\vec{\theta}}$ the set of parameters of interest that maximizes $q(\vec{\theta}, \mu = 0)$ [8]. For the purpose of simplification of the notation let us remain in the simplest one-dimensional case of a resonance search ($\vec{\theta} = m, \Gamma$) 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 is settled (taken from simulation).

Even in this 1D case (only mass as free parameter) and even if the test statistics q is derived by 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 entering the numerator and denominator of the likelihood ratio would not be nested [8]. However how can q^{glob} be evaluated? There are two viable approaches: 1) compute it by means of the method of pseudo-experiments, thus requiring a large amount of MC toys and a huge demand of CPU time and the aid from GPUs is crucial, or 2) estimate it in an approximate way (still taking into account the LEE) by the method of *Trial Factors* [7], namely relaying on the asymptotic behaviour of likelihood-ratio estimators. In the next two sections the two methods will be addressed separately and eventually the full compatibility between them will be discussed in Section 7.

5. Pseudo-experiments with a clustering-based scanning approach to address LEE

Taking the LEE into account implies to consider, within the same background-only fluctuation and everywhere in the relevant mass spectrum, any random peaking behaviour with respect to the expected shape associated to the background model. For this purpose a scanning technique based on a clustering approach has been developed, as described below.

Beforehand a pseudo-data invariant mass distribution of 15K candidates in a generic region of interest, namely $[1, 18]$GeV, has been generated according to a fictitious 7^{th} order Polynomial background model on the top of which any desired amount of a *significant* signal, mimicked by a Voigtian model, was artificially added close to 8GeV (as in Fig.7). At this mass value a 60MeV mass resolution is considered.

The fits to the pseudo-data distribution of Fig.7 are performed accordingly: the background-only model (the *Null Hypothesis* H_0) is a 7^{th} order Polynomial function whereas the background+signal model (the *Alternative Hypothesis* H_1) is obtained by adding a Voigtian function. The resolution values in the latter are reasonably increased as a function of the increasing invariant mass while satisfying the 60MeV constraint at 8GeV [3]. By performing the H_0 and H_1 fits, the local statistical significance of this peaking structure is $Z\sigma = 5.5\sigma$ with Z approximately estimated by means of the formula

$$Z \simeq \sqrt{-2[ln(L_{H_1}) - ln(L_{H_0})]} \tag{4}$$

where L_{H_0} (L_{H_1}) is the likelihood evaluated for the H_0 (H_1) hypothesis [9].

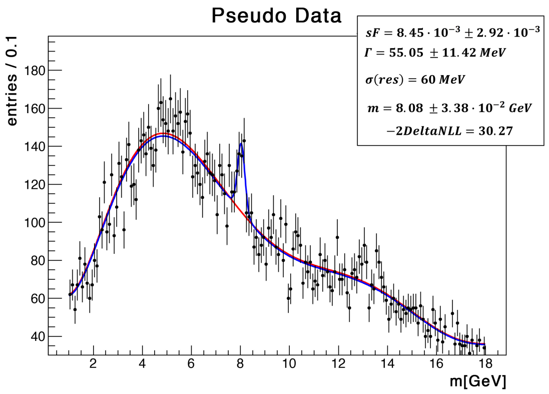

Figure 7: Simulated invariant mass distribution (pseudo-data). $H_0(H_1)$ fit is in red (blue); in the top right box the best values for the estimated parameters of the H_1 model are given.

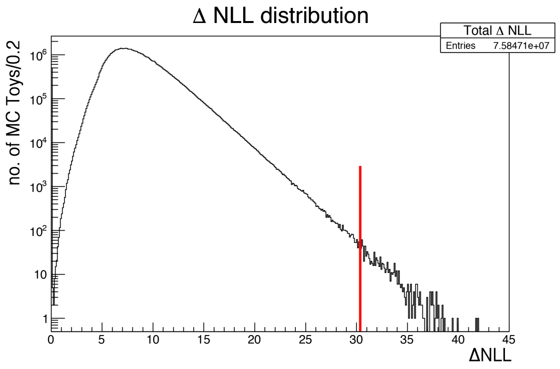

Figure 8: ΔNLL distribution for about 76M toys for the baseline configuration of clustering technique. The red line indicates the $\Delta NLL_{data} \simeq 30.27$ value for the pseudo-data distribution in Fig. 7.

The MC toys method is configured as follows. As first step of each toy iteration, a distribution based on the background-only model is generated over the whole mass spectrum and the H_0 fit is performed. As a second step the clustering technique acts on each generated pseudo-experiment as follows:

1. search for a main *seed* bin, namely for a bin whose content fluctuates more than $x\sigma$ strictly above the value of the background function in the center of that bin (σ is the statistical error associated to the considered bin).

2. Add any *side* bin to the *seed* bin if it holds a content that fluctuates more than $z\sigma$ strictly above the value of the background function in the center of that bin, otherwise the *seed* bin forms a 1-bin cluster.

3. Check also for *light* seeds, namely bins that fluctuate more than $y\sigma$ with $z < y < x$ and with at least a *side* bin fluctuating more than $z\sigma$. In case of positive result a cluster is formed.

In the third step, a series of independent H_1 fits is performed by cycling on the clusters collected in the clustering step. At the end of this step the fit with the best ΔNLL (the test statistic) is chosen. In total a ΔNLL distribution is obtained over all the processed MC toys.

A set of *baseline* clustering parameters $(x, y, z) = (2.25, 1.50, 1.00)$ has been chosen in order to satisfy two concurrent requirements: not missing any possible interesting fluctuation and avoiding selecting too many irrelevant fluctuations. This *baseline* configuration has been run on about 76M pseudo-experiments and the ΔNLL distribution is shown in Fig. 8, with the superimposed red line indicating the ΔNLL_{data} value for the pseudo-data. The global *p-value* is then estimated by:

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

This corresponds to the *global* statistical significance $Z\sigma = \Phi^{-1}(1-p)\sigma \simeq 4.22\sigma$, through the inverse function of the cumulative distribution of the standard Gaussian. As expected by considering the LEE, the *global* significance is relevantly lower than the estimated *local* one.

5.1 Evaluation of the possible systematic uncertainty associated to the clustering

In order to test the behavior of the method and to estimate the possible systematic uncertainty associated to the clustering technique, three sets of configuration parameters, i.e. three values for the (x, y, z) parameters, have been carefully considered. After some tests with different cuts, two further configurations have been chosen besides the baseline clustering cuts: a set of tighter values $(3.00, 1.75, 1.00)$ and a set of looser values $(2.00, 1.25, 1.00)$.

The Tab. 1 reports the details about these three clustering configurations such as the average number of H_1 fits per toy and the fraction of toys with no fit. These three configurations have been run on a same common set of 45M fluctuations and the three corresponding ΔNLL distributions are shown superimposed in Fig. 9. By focusing on the region of interest for the estimation of the statistical significance, namely the tail of the ΔNLL distribution ($\Delta NLL > 20$), it seems that there is no relevant difference among the three configurations. This can be appreciated by inspecting, in

Table 1: Mean number of alternative hypothesis fits per toy ($< fit_{H_1} >$) and fraction of toys with no fit (f_{no-fit}) for the three different clustering configurations described in the text.

Clustering configs.	$< fit_{H_1} >$	f_{no-fit}
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%

Fig. 9 and especially in Fig. 10, the normalized deviations of the type $(x - y)/(x + y)$ of the other two distributions with respect to the baseline distribution. Finally this is also confirmed by examining the estimated *global* significances for the *p-values* corresponding to different values of *local* significances, as reported in Tab. 2. It can be concluded that the systematic uncertainty on the p-values associated to the method is negligible.

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

Table 2: Estimated *global* significances for the 3 clustering configurations with respect to different *local* significance values estimated by Eq. 4.

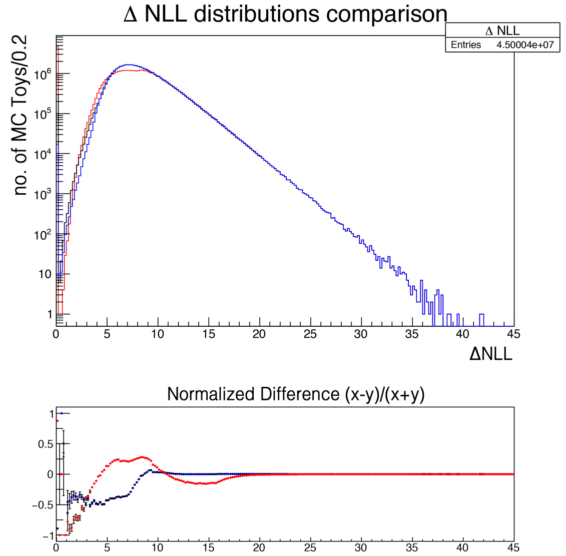

Figure 9: ΔNLL distributions for 45M of common fluctuations for the 3 configurations: baseline (black), tight (red) and loose (blue).

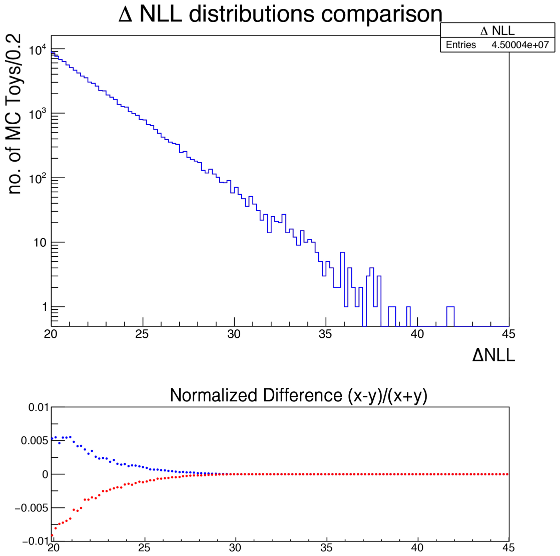

Figure 10: The same 3 ΔNLL distributions of Fig. 9 once zoomed into the range 20-45 to inspect their tail behaviour.

6. Trial factors and Gross-Vitells method for global significance

At the beginning of the LHC experiments, when the acceleration of GPUs embedded in graphics cards was not exploited for scientific computing purposes yet, the authors of Ref. [7] proposed an approximate way to determine the global significance by relying on the asymptotic behaviour of the likelihood ratio estimators [8]. To take into account the LEE, the correct factor that needs to be applied to the local significance in order to obtain the global one is called *trial factor*: $p^{glob} \approx f \cdot p^{loc}$.

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 the data an empirical evaluation, that can be used as a rule of thumb, gives [10]:

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

For the previous considered case: $f \approx 1/3 \cdot (18\text{GeV}/60\text{MeV}) \simeq 100$ which makes sense when considering that going from 5σ ($p \simeq 2.87 \cdot 10^{-7}$) to 4σ ($p \simeq 3.17 \cdot 10^{-5}$) implies a factor 110.

Gross and Vitells proposed a method to estimate an upper limit for the global p-value when the signal hypothesis (H_1) depends on s (nuisance) parameters that are undefined under the null hypothesis (H_0). It is possible to demonstrate [11] [7] that the probability that q^{glob}, that is the profile likelihood ratio test statistic maximized over $\vec{\theta}$, is greater than a given value c is bounded by the following inequality (that can be considered - asymptotically - as an equality):

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

where the first term (related to the local p-value) is a χ^2 distribution with s dof whereas the second term is the average number of *upcrossings*, namely the expected number of times that the local test statistic curve q^{loc} crosses an horizontal line at a given level $q = c$ with a positive derivative. In other words, the second term acts like a correction to the local p-value. For illustration purposes Fig. 11 shows an example of fluctuation and the number of upcrossings for $q = c_0 = 1$.

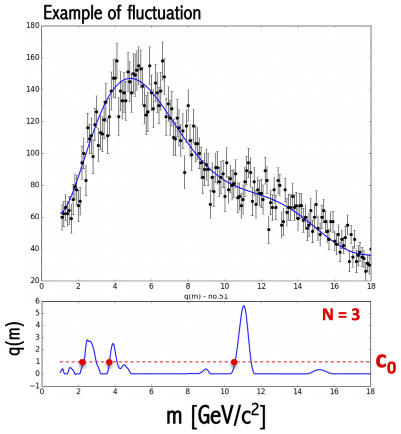

Figure 11: Number of *upcrossings* ($N_{c_0} = 3$), given a threshold $c = c_0 = 1$ (shown by the red line), for a generic example of fluctuation. Upper plot: in blue the H_0 fit. Bottom plot: $q^{loc} = q(m)$ test statistic values.

The $\langle N_c \rangle$ can be typically evaluated as average value over an enough large number of MC toys, but since its value could be very small depending on the level c, in such cases very large MC samples would be required for a precise numerical evaluation [8]. Fortunately, 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 [7]:

$$p^{glob}(c) = P(q(\hat{\hat{\theta}}, \mu = 0) > c) \leq P(\chi_s^2 > c) + \langle N_{c_0} \rangle \cdot \left(\frac{c}{c_0}\right)^{s-1} \cdot e^{-(c-c_0)/2} \tag{6}$$

where it is possible to evaluate $\langle N_{c_0} \rangle$ by generating a not too large number of MC toys and c_0 is chosen in a way that will minimize the resulting uncertainty on the boundary.

A procedure has been setup (within the `GooFit` framework) to estimate $\langle N_{c_0} \rangle$ for the pseudo-data configuration previously used. In this example there are two nuisance parameters, the peak mass m and width Γ, and thus $s = 2$ dof. A binned profile likelihood ratio was used as test statistic, where the number of events (N_i), in each bin i, is assumed to be distributed according to a Poisson distribution with an expected value $E(N_i) = \mu S_i(m, \Gamma) + (1 - \mu) B_i$, where μ is the signal strength parameter (or signal fraction). The chosen test statistics is the ΔNLL. As reference level $c_0 = s - 1 = 1$ has been chosen. The procedure to estimate $\langle N_{c_0} \rangle$ has been set up as follows:

1. 10K MC toys are configured starting from the generation: the mass distribution is based on the random fluctuation of the background-only model H_0.
2. For each toy a H_1-based fit is performed after setting the peak mass value to a certain m value; this fit is repeated 1000 times changing m and Γ values in continuous steps in order to scan the whole mass spectrum.
3. At each mass point m the profile likelihood ratio $q(m)$ is calculated and the distribution $q(m)$ along the mass spectrum is obtained (like in Fig. 11).
4. The number of *upcrossings* of $q(m)$ with respect to the c_0 level is thus easily estimated.

It took about 3 days on a single GPU to carry out this 10K MC-based procedure, namely the time equivalent to 4-5M of MC toys in the clustering approach discussed in Section 5. The result, for $c_0 = 1$, was found to be $\langle N_{c_0} \rangle = 7.3$ with an uncertainty of $\sigma_{N_{c_0}} = 2.4$, and it can be used to evaluate from Eq. 6 the Upper Limit estimated for $p^{glob}(c)$ with this Gross-Vitells method .

7. Comparison between the Gross-Vitells method and the massive pseudo-experiments method with a clustering-based scanning approach

Finally it is possible to compare the Upper Limit (UL) for $p^{glob}(c)$, estimated with the Gross-Vitells method (G-V method) as discussed in Section 6, with the global p-value p^{glob}_{toys} computed by means of the ΔNNL distribution obtained by a massive amount of pseudo-experiments (76M in the baseline configuration) as discussed in Section 5. Specifically, p^{glob}_{toys}, as a function of c, is calculated as a global p-value (as in Eq. 5) now considered as a function of a running chosen value of ΔNLL represented by c:

$$p^{glob}_{toys}(c) = \int_c^\infty f(\Delta NLL) d(\Delta NLL)$$

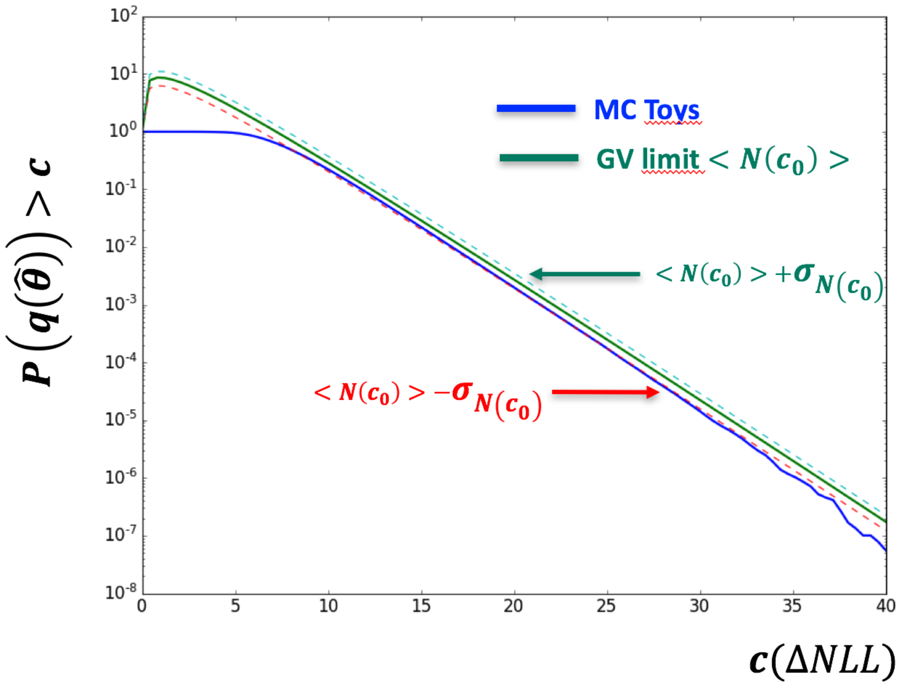

Figure 12: Comparison of two different methods to estimate a global p-value as a function of a threshold c representing a ΔNLL value. The blue curve is obtained by a method exploiting a massive amount of pseudo-experiments (MC toys) and addressing the LEE issue by means of a clustering-based scanning technique. The green curve represents the G-V UL, namely the Upper Limit estimated with the G-V method; dashed green and red curves delimit the band of the statistical uncertainty associated to the UL due to the very limited sample of MC toys used in the G-V approach.

Fig. 12 provides the comparison between the two methods. The estimation of the Upper Limit of the global p-value in the G-V method is compared with the exact function obtained from a huge amount of MC toys integrated with a clustering approach to take into account the LEE. The G-V UL is well compatible with the massive MC toys result. However the G-V UL behaves always more conservatively, overestimating the global p-value and thus underestimating the global significance.

A more quantitative comparison is reported in the Tab. 3 in terms of global statistical significance derived from the previous curves and for a few chosen specific thresholds corresponding to specific local significance values. This table provides the size of the light discrepancy between the two methods, beyond showing again the remarkable size of the LEE shown already in Tab. 2.

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

Table 3: Global statistical significance values corresponding to specific local ones. G-V Upper Limit estimates are compared with the exact values derived by the massive MC toys method (for the baseline clustering configuration).

8. Conclusions

By exploiting the computing acceleration provided by GPUs, through the `GooFit` framework capabilities, it has been possible to explore the applicability limits and investigate the approximation characterizing modern statistical methods by means of the pseudo-experiments technique. The studied methods were introduced in High Energy Physics, at the beginning of the LHC era, before the amplification of the parallel computing paradigm occurred with the advent of GPUs. The asymptotic behaviour of a likelihood ratio test statistics (Cowan-Cranmer-Gross-Vitells) has been studied while estimating the local statistical significance of a known signal. The Look-Elsewhere-Effect has been studied adopting a clustering-based scanning approach to handle a huge amount of MC toys, while estimating the global statistical significance of a new signal. In the latter context the validity of the approximation of the Gross-Vitells method has been investigated as well.

Acknowledgements

We are grateful for valuable support from G.Donvito, manager of the High Performance Cluster hosted by the Bari ReCas [12] Data Center. We thank the organizers for the warm invitation and the Italian Ministery of Foreign Affairs and International Cooperation (MAECI) for the support through the PGR01273/2021, in the framework of the Italy-Mexico bilateral cooperation project MX18MO11 (2018-2020) involving INFN and CINVESTAV.

References

[1] Andreassen R et al., *J. Phys. Conf. Series* **513** (2014) 052003; Schreiner H et al., *J. Phys. Conf. Series* **1085** (2018) 4, 042014.

[2] `https://docs.nvidia.com/`; for these studies: CUDA version 6.5(7.0) for Tesla K20(40).

[3] Pompili A and Di Florio A, *J. Phys. Conf. Series* **762** (2016) 012044; *J. Phys. Conf. Series* **1085** (2018) 4, 042005; *PoS Confinement2018* (2019) **229**.

[4] CMS Collaboration, *Phys. Lett.* **B 734** (2014) 261.

[5] Wilks S S, *Annals Math. Statist.* **9** (1938) 60-62.

[6] Cowan G, Cranmer K, Gross E and Vitells O, *Eur. Phys. J.* **C71** (2011) 1554.

[7] Gross E and Vitells O, *Eur. Phys. J.* **C70** (2010) 525-530.

[8] Lista L, *Statistical Methods for Data Analysis in Particle Physics*, Lectures Notes in Physics **941**, 2^{nd} Edition, Springer, 2017, DOI 10.1007/978-3-319-62840-0.

[9] Narsky I and Porter F C, *Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning*, Wiley, 2013, DOI 10.1002/9783527677320.

[10] Gross E and Vitells O, https://www.birs.ca/workshops/2010/10w5068/files/gross.pdf .

[11] Davies R B, *Biometrika* **74** (1987) 33-43.

[12] ReCas project (funded by the italian MIUR): `http://www.recas-bari.it/index.php/en/`.