

The Simplified Likelihood Framework

Andy Buckley^a, Matthew Citron^b, Sylvain Fichet^c, Sabine Kraml^d,
Wolfgang Waltenberger^e, Nicholas Wardle^f

^a *School of Physics & Astronomy, University of Glasgow, Glasgow, Scotland, UK*

^b *University of California, Santa Barbara, California, USA*

^c *ICTP-SAIFR & IFT-UNESP, R. Dr. Bento Teobaldo Ferraz 271, São Paulo, Brazil*

^d *Laboratoire de Physique Subatomique et de Cosmologie, Université Grenoble-Alpes,
CNRS/IN2P3, Grenoble, France*

^f *Imperial College London, South Kensington, London, UK*

Abstract

We present the Simplified Likelihood framework, a systematic approximation scheme for experimental likelihoods such as those originating from LHC experiments. We present an efficient method to compute the parameters of the simplified likelihood from Monte Carlo simulations. The approach is validated using a realistic LHC-like analysis, and the limits of the approximation are explored. The simplified likelihood data can be conveniently released in the HepData error source format and automatically built from it, making this framework a convenient tool to transmit realistic experimental likelihoods to the community.

1 Introduction

Scientific observations of the real world are by nature imperfect in the sense that they always contain some amount of uncertainty unrelated to data, the *systematic* uncertainty. Identifying, measuring and modeling all the sources of systematic uncertainty is an important part of running a scientific experiment. A thorough treatment of such uncertainties is especially important in exploratory fields like Particle Physics and Cosmology. In these fields of research, today's experiments can be of large scale and can contain a huge number of these uncertainties. For instance in the case of the Large Hadron Collider (LHC) experiments, the experimental likelihood functions used in standard model measurements and searches for new physics can contain several thousands of systematic uncertainties.

Although sources of systematic uncertainty can be numerous and of very different nature, a general feature they share is that they are almost always independent of each other. This property of independence between the systematic uncertainties has profound consequences, and, as we will discuss soon, is the reason why the approach presented in this work is so efficient. Namely, independence of the uncertainties can be used in order to drastically simplify the experimental likelihood function, to the price of an often-negligible error that will be discussed at length in this paper.

The *Simplified Likelihood* framework we present in this paper is a well-defined approximation scheme for experimental likelihoods. It can be used to simplify subsequent experimental analyses, to allow a uniform statistical treatment of published search-analysis data, and to ease the transmission of results between an experiment and the scientific community. We build on the proposals for approximating likelihoods recently suggested in Refs. [1, 2], in which promising preliminary results have been shown.

In the context of the LHC, communicating the full experimental likelihoods via the RooFit/Rootstats software framework [3, 4] has been suggested in Refs. [5, 6]. The presentation method we propose in this paper is complementary in that it is technically straightforward to carry out, without relying on any particular software package. Additionally, the proposal of presenting LHC results decoupled from systematic uncertainties has been pursued in Ref. [7] in the context of theoretical errors on Higgs cross-sections. For Higgs cross-sections and decays, the combined covariance of the Higgs theoretical uncertainties consistent with the simplified likelihood framework presented here has been determined in Ref. [8].

In this paper we unify and extend the initial proposals of Refs. [1, 2], and thoroughly test the accuracy of the approximations using simulated LHC searches for new phenomena. Compared to Refs. [1, 2], an important progress accomplished is that we have been able to rigorously include asymmetries in the combined uncertainties, which is useful in order to avoid inconsistencies such as a negative event yield. Technically this is done by taking into account the next-to-leading term in the limit given by an appropriate version of the Central Limit Theorem (CLT).

The paper is organized as follows. A summary of the main results is given in Section 2. Section 3 contains the formal material, including an interesting result about the next-to-leading term of the CLT and the derivation of the simplified likelihood formula. Practical

considerations regarding the SL flexibility and the release of the SL via HepData are given in Section 4. Finally a validation of the SL framework in a realistic pseudo-search at the LHC is presented in Section 5.

2 From the Experimental Likelihood to the Simplified Likelihood

This section introduces the formalism, presents the main theoretical results and an efficient Monte-Carlo based calculation method. We will focus on the typical experimental likelihood used in searches for new phenomena at particle physics experiments. However we stress that the simplified likelihood approach can be easily generalized to other physics contexts. The data collected in particle physics usually originate from random (quantum) processes, and have thus have an intrinsic *statistical* uncertainty—which vanishes in the limit of large data sets. Our interest rather lies in the *systematic* uncertainties, which are independent of the amount of data.

A likelihood function L is related to the probability to observe the data given a model \mathcal{M} , specified by some parameters,

$$L(\text{parameters}) = \text{Pr}(\text{data}|\mathcal{M}, \text{parameters}). \quad (2.1)$$

We denote the observed quantity as \hat{n} and the expected quantity by n , where n depends on the model parameters. For example, in the case of a particle physics experiment, these quantities can be the observed and expected number of events that satisfy some selection criteria. The full set of parameters includes parameters of interest, here collectively denoted by α , and *elementary* nuisance parameters $\delta = (\delta_1, \dots, \delta_j, \dots, \delta_N)^T$, which model the systematic uncertainties. In the simplified likelihood framework, we derive a set of *combined* nuisance parameters θ . For P independent measurements, there will be P combined nuisance parameters, $\theta = (\theta_1, \dots, \theta_I, \dots, \theta_P)^T$.

The key result at the basis of the simplified likelihood framework is the approximation

$$\begin{aligned} L(\alpha, \delta)\pi(\delta) &= \prod_{I=1}^P \text{Pr}\left(\hat{n}_I \mid n_I(\alpha, \delta)\right)\pi(\delta) \\ &\approx \prod_{I=1}^P \text{Pr}\left(\hat{n}_I \mid a_I(\alpha) + b_I(\alpha)\theta_I + c_I(\alpha)\theta_I^2\right) \cdot \frac{e^{-\frac{1}{2}\theta^T \rho^{-1}(\alpha)\theta}}{\sqrt{(2\pi)^P}} \equiv L_S(\alpha, \theta), \end{aligned} \quad (2.2)$$

$$(2.3)$$

where the first line is the exact “experimental likelihood” and the second line is the simplified likelihood. Here $\pi(\delta)$ is the joint probability density distribution for the elementary nuisance parameters. In our assumptions these are independent from each other, hence the prior factorises as $\pi(\delta) = \prod_{i=1}^N \pi_i(\delta_i)$. The simplified likelihood formalism is relevant for $N > P$, which is the most common case.¹

¹If $P < N$, there are more observed quantities than nuisance parameters. In such case the use of combined uncertainties, although not formally wrong, is inappropriate. The formula Eq. (2.3) still applies but the covariance matrix will be singular.

The a_I , b_I and c_I , and the $P \times P$ correlation matrix $\boldsymbol{\rho} = \rho_{IJ}$ define the simplified likelihood and are in general functions of the parameters of interest. However in concrete cases, this dependence will often be negligible. This is in particular the case in particle physics searches for new physics when the expected event number decomposes into signal (n_s) plus background (n_b) contributions. The parameters of interest that model the new physics enter in n_s while n_b is independent from them. Whenever the expected signal is small with respect to the background, the dominant uncertainties in searches for new physics are those related to the background. In the simplified likelihood, neglecting the systematic uncertainties affecting the signal implies in turn that the parameters of the simplified likelihood are independent of $\boldsymbol{\alpha}$. Hence the simplified likelihood Eq. (2.3) takes the form²

$$L_S(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{I=1}^P \Pr\left(\hat{n}_I \mid n_{s,I}(\boldsymbol{\alpha}) + a_I + b_I \theta_I + c_I \theta_I^2\right) \cdot \frac{e^{-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\rho}^{-1} \boldsymbol{\theta}}}{\sqrt{(2\pi)^P}}, \quad (2.4)$$

which is the expression we use in the rest of this paper. This expression is valid for data with any statistics of observation, however in particle physics, since often the data are observed event counts n_I^{obs} , the data will typically follow Poisson statistics such that

$$\Pr(\hat{n}_I | n_I) \equiv \text{Pois}(n_I^{\text{obs}} | n_I) = \frac{(n_I)^{n_I^{\text{obs}}} e^{-n_I}}{n_I^{\text{obs}}!}. \quad (2.5)$$

The parameters of the simplified likelihood (a_I, b_I, c_I, ρ_{IJ}) have analytical expressions as a function of the variance and the skew of each elementary nuisance parameter (see Sec. 3.2). However, often the elementary uncertainties and the event yields are already coded in a Monte Carlo generator. In such case, an elegant method to obtain the simplified likelihood parameters is the following. From the estimators of the event yields \hat{n}_I , one can evaluate the three first moments of the \hat{n}_I distribution and deduce the parameters of the simplified likelihood directly from these moments. What is needed is the mean $m_{1,I}$, the covariance matrix $m_{2,IJ}$ and the diagonal component of the third moment $m_{3,I} \equiv m_{3,III}$.

Using the definition $n_I = a_I + b_I \theta_I + c_I \theta_I^2$, we have the relations

$$m_{1,I} = \mathbf{E}[\hat{n}_I] = a_I + c_I \quad (2.6)$$

$$m_{2,IJ} = \mathbf{E}[(\hat{n}_I - \mathbf{E}[\hat{n}_I])(\hat{n}_J - \mathbf{E}[\hat{n}_J])] = b_I b_J \rho_{IJ} + 2c_I c_J \rho_{IJ}^2 \quad (2.7)$$

$$m_{3,I} = \mathbf{E}[(\hat{n}_I - \mathbf{E}[\hat{n}_I])^3] = 6b_I^2 c_I + 8c_I^3, \quad (2.8)$$

where \mathbf{E} denotes the expectation value. Inverting these relations — while taking care to pick the relevant solutions to quadratic and cubic equations — gives directly the parameters

²We have substituted $a_I(\alpha) \rightarrow a_I + n_{s,I}(\alpha)$, $b_I(\alpha) \rightarrow b_I$ and $c_I(\alpha) \rightarrow c_I$.

of the simplified likelihood. We find

$$c_I = -\text{sign}(m_{3,I}) \sqrt{2m_{2,II}} \cos\left(\frac{4\pi}{3} + \frac{1}{3} \arctan\left(\sqrt{8\frac{m_{2,II}^3}{m_{3,I}^2} - 1}\right)\right) \quad (2.9)$$

$$b_I = \sqrt{m_{2,II} - 2c_I^2} \quad (2.10)$$

$$a_I = m_{1,I} - c_I \quad (2.11)$$

$$\rho_{IJ} = \frac{1}{4c_I c_J} \left(\sqrt{(b_I b_J)^2 + 8c_I c_J m_{2,IJ}} - b_I b_J \right). \quad (2.12)$$

These formulae apply if the condition $8m_{2,II}^3 \geq m_{3,I}^2$ is satisfied. This limit is approached when the asymmetry becomes large. Near this limit, the approximation typically tends to become inaccurate because higher order terms $O(\theta_I^3)$ would need to be included in the expressions of the simplified likelihood Eq. (2.3). In practice, however, this requires a high skewness of the nuisance parameters, and the Simplified Likelihood framework up to quadratic order is sufficient for most applications.

This method will be used in the examples shown in the rest of the paper. This means that if one is provided with the moments m_1 and m_3 for each bin and the covariance matrix $m_{2,IJ}$, the simplified likelihood parameters are completely defined. Moreover, in the case where the nuisance parameters are affecting only the background rate Eq. (B.3), this computation has to be realized once and the resulting likelihood can be used for any kind of signal by appropriate substitution of $n_s(\alpha)$.

3 The simplified likelihood from the central limit theorem

This section contains the derivation of the simplified likelihood formula Eq. (2.3). The reader interested only in the practical aspects of the simplified likelihood framework can safely skip it. In Section 3.1 we lay down a result about the next-to-leading term of the central-limit theorem. Then in Section 3.2 we demonstrate Eq. (2.3) and give the analytical expressions of the simplified likelihood parameters as a function of the elementary uncertainties.

3.1 Asymmetries and CLT at next-to-leading order

The CLT is often used in its asymptotic limit where the distribution becomes exactly normal. In the context of the simplified likelihood framework, however, it is mandatory to keep the next-to-leading term in the CLT's large N expansion. This next-to-leading term encodes skewness: this is the main information about the asymmetry of the distribution. This asymmetry is a relevant feature for the analyses hence it is in principle safer to keep this information. But keeping the asymmetry is truly critical for a slightly different reason. A normal distribution has a support on \mathbf{R} , while quantities like event yields are defined on \mathbf{R}^+ . Hence having a nuisance parameter with a normal distribution can give inconsistencies, for instance a negative yield. This is a problem both conceptually and concretely when running the analyses. This issue occurs because the Gaussian limit of

the CLT loses information about the support. Using this limit is simply a too rough approximation. Instead, to have an asymmetric support such as \mathbf{R}^+ , the distribution must be asymmetric, therefore the skew must be taken into account.

The deformed Gaussian obtained when keeping the skew into account does not seem to have in general an analytical PDF. However, by using further the large N expansion, we have been able to develop a trick to express the CLT at next-to-leading order in a very simple way. We realize that a random variable Z with characteristic function

$$\varphi_Z(t) = \exp \left(-\frac{\sigma^2 t^2}{2} - i \frac{\gamma t^3}{6\sqrt{N}} + O \left(\frac{t^4}{N} \right) \right) \quad (3.1)$$

can, up to higher order terms in the large N expansion, be equivalently be expressed in terms of an exactly Gaussian variable θ in the form

$$Z = \theta + \frac{\gamma}{3\sqrt{N}} \theta^2, \quad \text{with} \quad \theta \sim \mathcal{N}(0, \sigma^2). \quad (3.2)$$

We will refer to this type of expression as “normal expansion”. Details about the derivation are given in App. A.

Equation (3.2) readily gives the most basic CLT at next-to-leading order when assuming $Z = N^{-1/2} \sum_{j=1}^N \delta_j$, where the δ_j are independent identically distributed centred nuisance parameters of variance σ^2 and third moment γ . The trick applies similarly to the Lyapunov CLT, *i.e.* when the δ_j are not identical, in which case one has defined $\sigma^2 = N^{-1} \sum_{j=1}^N \sigma_j^2$, $\gamma = N^{-1} \sum_{j=1}^N \gamma_j$,

Finally, our trick applies similarly to the multidimensional case where various linear combinations of the δ_j give rise to various Z_I . The Z_I have covariance matrix Σ_{IJ} and a skewness tensor $\gamma_{IJK} = \mathbb{E}[Z_I Z_J Z_K]$. For our purposes, we neglect the non-diagonal elements of γ , keeping only the diagonal elements, noted $\gamma_{III} \equiv \gamma_I$. These diagonal elements encode the leading information about asymmetry, while the non-diagonal ones contain subleading information about asymmetry and correlations. With this approximation, we obtain the multidimensional CLT at next-to-leading order,

$$Z_I \rightarrow \theta_I + \frac{\gamma_I}{3\sqrt{N}} \theta_I^2, \quad N \rightarrow \infty \quad \text{with} \quad \theta_I \sim \mathcal{N}(0, \Sigma). \quad (3.3)$$

This result will be used in the following. Again, for $\gamma_I \rightarrow 0$, one recovers the standard multivariate CLT.

3.2 Calculation of the simplified likelihood

Here we prove Eq. (2.3). The dependence on the parameters of interest α is left implicit in this section. We will first perform a step of propagation of the uncertainties, then a step of combination. This is a generalization of the approach of [1]. Here we take into account the skew, hence there is no need to use an exponential parameterization like in [1].

In this section the elementary nuisance parameters δ_i are independent, centered, have unit variance, and have skew γ_i , *i.e.*

$$\mathbb{E}[\delta_i] = 0, \quad \mathbb{E}[\delta_i^2] = 1, \quad \mathbb{E}[\delta_i^3] = \gamma_i. \quad (3.4)$$

It is convenient to use a vector notation for the set of these elementary nuisance parameters, $(\delta_i) \equiv \boldsymbol{\delta}$.

As a first step, we want to propagate the systematic uncertainties at the level of the event numbers. For an event number n depending on a quantity Q subject to uncertainty, we have

$$n[Q] \equiv n[Q_0(1 + \Delta_Q \delta)]. \quad (3.5)$$

The propagation amounts to performing a Taylor expansion with respect to Δ_Q . This expansion should be truncated appropriately to retain the leading effects of the systematic uncertainties in the likelihood. It was shown in [1] that the expansion should be truncated above second order.

For multiple sources of uncertainty, we have a vector $\boldsymbol{\delta}$ and the relative uncertainties propagated to n are written as

$$n \equiv n^0 \left(1 + \Delta_1^T \cdot \boldsymbol{\delta} + \boldsymbol{\delta}^T \cdot \Delta_2 \cdot \boldsymbol{\delta} + O\left(\frac{n^{(3)}}{n^0} \Delta_Q^3\right) \right) \quad (3.6)$$

with

$$\Delta_1 = \frac{1}{n^0} \left(\frac{\partial n}{\partial \delta_1} \Delta_{Q,1}, \dots, \frac{\partial n}{\partial \delta_p} \Delta_{Q,p} \right)_{\delta=0}^T, \quad \Delta_2 = \frac{1}{2n^0} \left(\frac{\partial^2 n}{\partial \delta_i \partial \delta_j} \Delta_{Q,i} \Delta_{Q,j} \right)_{\delta=0}. \quad (3.7)$$

The $n^{(3)}$ denotes schematically the third derivatives of n .

The second step is to combine the elementary nuisance parameters. We introduce combined nuisance parameters θ_I which are chosen to be centered and with unit variance without loss of generality, and whose correlation matrix is denoted ρ_{IJ} , *i.e.*

$$\mathbf{E}[\theta_I] = 0, \quad \mathbf{E}[\theta_I^2] = 1, \quad \mathbf{E}[\theta_I \theta_J] = \rho_{IJ}. \quad (3.8)$$

Moreover we define the expected event number in terms of the combined nuisance parameters as

$$n_I = n_I^0 (1 + \Delta_{1,I} \cdot \boldsymbol{\delta} + \boldsymbol{\delta} \cdot \Delta_{2,I} \cdot \boldsymbol{\delta}) \equiv a_I + b_I \theta_I + c_I \theta_I^2. \quad (3.9)$$

The a_I, b_I, c_I parameters together with the correlation matrix ρ_{IJ} fully describe the combined effect of the elementary uncertainties.

To determine them we shall identify the three first moments on each side of Eq. (3.9). We obtain

$$a_I = n_I^0 \left(1 + \text{tr} \Delta_{2,I} - \frac{1}{6} \sum_{i=1}^N \gamma_i (\Delta_{1,I,i})^3 + O(\Delta^4) \right), \quad (3.10)$$

$$b_I = a_I \left(\Delta_{1,I}^T \cdot \Delta_{1,I} + 2 \sum_{i=1}^N \gamma_i \Delta_{1,I,i} \Delta_{2,I,i} + O(\Delta^4) \right)^{1/2}, \quad (3.11)$$

$$\rho_{IJ} = \frac{a_I a_J}{b_I b_J} \left(\Delta_{1,I}^T \cdot \Delta_{1,J} + \sum_{i=1}^N \gamma_i (\Delta_{1,I,i} \Delta_{2,J,i} + \Delta_{1,J,i} \Delta_{2,I,i}) \right) + O(\Delta^4), \quad (3.12)$$

$$c_I = \frac{a_I}{6} \sum_{i=1}^N \gamma_i (\Delta_{1,i})^3 + O(\Delta^4) \quad (3.13)$$

where the $O(\Delta^4)$ denotes higher order terms like $\text{tr}(\Delta_{2,I}^T \cdot \Delta_{2,I})$, $(\text{tr} \Delta_{2,I})^2$, $\Delta_{1,I}^T \cdot \Delta_{1,I} \text{tr} \Delta_{2,I}$ which are neglected. When $\gamma_i \rightarrow 0$ one recovers the expressions obtained in Ref. [1].³

Importantly, the Δ_2 term contributes at leading order only in the mean value a_I and always gives subleading contributions to higher moments. Hence, for considerations on higher moments – which define the shape of the combined distribution, we can safely take the approximation

$$n_I \approx n_I^0 (1 + \Delta_{1,I} \cdot \delta) \quad (3.14)$$

from Eq. (3.9). We now make the key observation that this quantity is a sum of a large number of independent random variables. These are exactly the conditions for a central limit theorem to apply. As all the elementary uncertainties have in principle different shape and magnitudes we apply Lyapunov's CLT [9]. We can for instance use Lyapunov's condition on the third moment, and the theorem reads: If

$$\frac{\mathbf{E}[(n_I - \mathbf{E}[n_I])^3]}{\mathbf{E}[(n_I - \mathbf{E}[n_I])^2]^{3/2}} \sim \frac{6c_I}{b_I} \rightarrow 0 \quad \text{for } N \rightarrow \infty \quad (3.15)$$

then

$$\theta_I \sim \mathcal{N}(0, \rho) \quad \text{for } N \rightarrow \infty. \quad (3.16)$$

Furthermore we can see that the expression of n_I in terms of the combined nuisance parameters, $n_I = a_I + b_I \theta_I + c_I \theta_I^2$ (first defined in Eq. (3.9)), takes the form of a normal expansion as defined in subsection 3.1 (see Eq. (3.3)). This means that the $c_I \theta_I^2$ term corresponds precisely to the leading deformation described by the next-to-leading term of the CLT. This deformation encodes the skewness induced by the asymmetric elementary uncertainties. We have therefore obtained a description of the main collective effects of asymmetric elementary uncertainties, which is dictated by the CLT. The resulting simplified likelihood is given in Eq. (2.3).

3.3 Precision of the normal expansion

The accuracy of the normal expansion $n = a + b\theta + c\theta^2$ with $\theta \sim \mathcal{N}(0, 1)$ — and thus of the simplified likelihood — is expected to drop when only a few elementary uncertainties are present and these depart substantially from the Gaussian shape. This is the combination of conditions for which the next-to-leading CLT Eq. (3.3) tends to fail. It is instructive to check on a simple distribution how the normal expansion approximates the true distribution, and in which way the discrepancies tend to appear.

We consider the realistic case of a log-normal distribution with parameters μ, σ . We fix $\mu = 0$ without loss of generality. The three first centered moments are

$$m_1 = e^{\frac{\sigma^2}{2}}, \quad m_2 = e^{2\sigma^2} - e^{\sigma^2}, \quad m_3 = e^{\frac{9\sigma^2}{2}} - 3e^{\frac{5\sigma^2}{2}} + 2e^{\frac{3\sigma^2}{2}} \quad (3.17)$$

³For simplicity we show here the expressions assuming $c_I \ll b_I$, as it is sufficient in the scope of the proof. For sizeable c_I one should instead solve the system shown in Eqs. (2.6)-(2.8), (2.9)-(2.12).

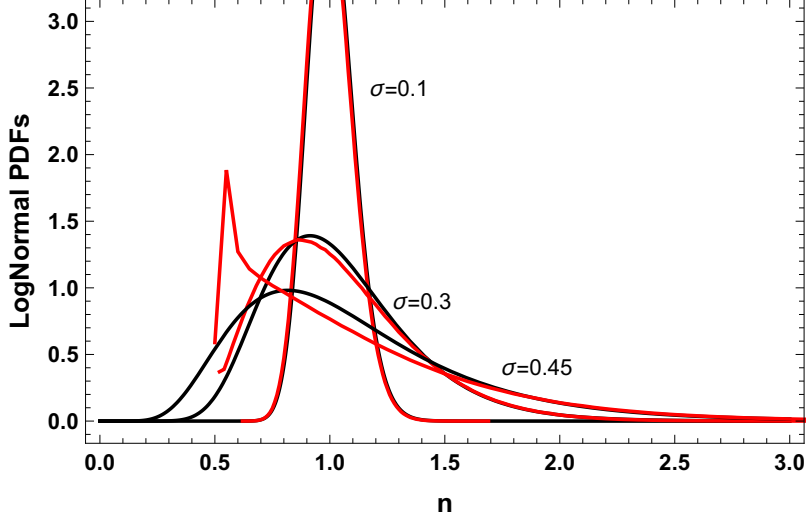


Figure 1. The normal approximation for a Poisson and a log-normal distribution. Black curves show the true distributions, red curve show the approximate one.

and a, b, c are obtained using Eqs. (2.9)-(2.12).

For $\sigma \sim 0.69$, the bound $8m_2^3 \approx m_3^2$ is reached (see Sec. 2). This is the bound where the distribution is so asymmetric that the variance comes entirely from the θ^2 term. Beyond this bound the normal expansion cannot be used at all as Eqs. (2.9)-(2.12) have no solutions. The distribution has $c > 0$ thus n has a lower bound given by $n > a - b^2/4c$. It turns out that apart near the above mentioned limit on σ , the lower bound on n is roughly $n \gtrsim 0.5$, therefore the approximation can never produce a negative event yield.

Let us finally check how well the approximation performs. The true and approximate densities are shown in Fig. 1. When the true density is vanishing in the region $n < 0.5$, the approximation is excellent. For larger asymmetry, $\sigma \sim 0.35 - 0.4$, the true density starts to be sizeable in the region $n < 0.5$. The approximate density is still reasonably good away from $n \sim 0.5$. However near this boundary, the approximate density tends to increase and possibly peak, somehow to account for the area at $n < 0.5$ that it cannot reproduce. This behaviour can be seen in Fig. 1, and will also be observed for certain bins in the LHC-like analysis implemented in next section.

Overall, through this example, we can see that the normal approximation tends to become inaccurate for a skewness of $\sim 100 - 150\%$. This is a moderate value, however one should keep in mind that these considerations apply to the combined uncertainties, for which small skewness is typical. The accuracy of the simplified likelihood framework will be tested in next section in a realistic setup.

4 Practical aspects of the simplified likelihood framework

4.1 Range of application

An important feature of the simplified likelihood is that it is flexible in the sense that the combination of the systematic uncertainties does not have to be applied to the whole set. The only requirement to combine a subset of the uncertainties is that it should have a convergent enough CLT behaviour in order for the simplified likelihood to be accurate. There is thus a freedom in partitioning the set of systematic uncertainties, giving rise to variants of the simplified likelihood that can be either equivalent or slightly different upon marginalizing.

For instance if a single systematic uncertainty is left apart from the combination, the simplified likelihood takes the form

$$L_S(\boldsymbol{\alpha}, \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \prod_{I=1}^P \Pr\left(\hat{n}_I \mid a_I(\boldsymbol{\alpha}) + b_I\theta_I + c_I\theta_I^2 + \Delta_I\delta + \right) \cdot \frac{e^{-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\rho}^{-1} \boldsymbol{\theta}}}{\sqrt{(2\pi)^P}} \cdot \pi(\delta) \quad . \quad (4.1)$$

Similarly, if two subsets of systematic uncertainties tend to separately satisfy the CLT conditions, they can be separately combined, giving the simplified likelihood

$$L_S(\boldsymbol{\alpha}, \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \prod_{I=1}^P \Pr\left(\hat{n}_I \mid a_I(\boldsymbol{\alpha}) + b_I\theta_I + c_I\theta_I^2 + \tilde{b}_I\tilde{\theta}_I + \tilde{c}_I\tilde{\theta}_I^2\right) \cdot \frac{e^{-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\rho}^{-1} \boldsymbol{\theta}}}{\sqrt{(2\pi)^P}} \cdot \frac{e^{-\frac{1}{2}\tilde{\boldsymbol{\theta}}^T \tilde{\boldsymbol{\rho}}^{-1} \tilde{\boldsymbol{\theta}}}}{\sqrt{(2\pi)^P}} \quad . \quad (4.2)$$

The simplified likelihood naturally accommodates any such partitions. It is actually commonplace in LHC analyses to present systematic uncertainties combined in subsets, for example “theoretical”, “experimental”, “luminosity”, “MC” uncertainties. This is useful for mere informative purpose but also for further reinterpretations. For example the theoretical uncertainties can be improved later on, hence members of the community should be able to change them with no loss of consistency.⁴ Another reason to single out a nuisance parameter from the combination (as shown in Eq. (4.1)) is if it has a large non-Gaussian PDF that one prefers to take into account exactly. In order to take profit from the versatility of the simplified likelihood, an equally versatile format is needed to release the simplified likelihood data. This will be the topic of next subsection.

Finally, if systematic uncertainties on the *signal* are taken into account, the b, c, ρ parameters become dependent on the parameter of interest $\boldsymbol{\alpha}$. While there is no conceptual difference with the case of background-only systematics, there are important practical differences. Numerical evaluations become heavier since the parameters of the simplified likelihood – especially $\rho_{IJ}(\boldsymbol{\alpha})$ which requires a matrix inversion – have to be evaluated for each value of $\boldsymbol{\alpha}$. Presentation of the simplified likelihood data may also become more evolved.

⁴Such combination of theoretical uncertainties has been done in [8] for the Higgs production and decay rates and can be implemented in a Higgs simplified likelihood.

4.2 Construction and presentation

There are in principle two ways of releasing the data needed to build the simplified likelihood. One way could be to release the whole set of elementary systematic uncertainties. The other is to release the three first moments of the PDF of the combined systematic uncertainties. While the former is in principle doable, we will focus only on the latter. Indeed, the elementary uncertainties are usually already coded by the experimentalists in Monte Carlo generators, hence it is straightforward to evaluate these moments and one should take profit of it. ⁵

We thus focus on the release of the simplified likelihood data via the $m_{1,I}$, $m_{2,IJ}$, $m_{3,I}$ moments of the PDF of the combined systematic uncertainties, already defined in Eqs. (2.6)-(2.8), where $m_{3,I}$ is the diagonal part of the third-rank tensor $m_{3,IJK}$. Evaluating these moments via MC toys is straightforward for the experimental analysis. However their presentation needs to be discussed in great details, taking into account the available tools and the current practices. This is the purpose of this subsection.

Key to the usefulness of any likelihood data for analysis reinterpretation is the availability of that data in a standard format. For global fits, where tens or hundreds of analyses may be used simultaneously, it is crucial that this format be unambiguously parseable by algorithms without human assistance. A standard location is also necessary, for which the obvious choice is the longstanding HEP data repository, HepData [10].

Let us fix some terminology. It is convenient to refer to the data in terms of the order of the moment from which they originate. We will use the term “ n -th order data” to refer to information coming from a moment of order n ; here, n will go only up to 3. Second-order data includes the covariance matrix, correlation matrix, and/or diagonal uncertainties: these can be given either in a relative or absolute parametrization. There is the same kind of freedom for third-order data but this does not need to be discussed here. In addition to the moments of the combined systematic uncertainties, this terminology will also apply to the observed central values and statistical uncertainties usually presented by the experiments.

Let us review the current formats of presentation of likelihood data. There is no problem of format with the presentation of first-order data and no third-order data are usually given. Regarding second-order data there is unfortunately no standard representation currently established. A review of the second-order data in HepData and on the experiments’ analysis websites reveals a mixture of presentation styles:

- *Table format:* 2D histograms of either covariance or correlation matrices. This has the difficulties that the convention used is not made clear (other than by inspection of the matrix diagonal), and without a structural association with first order dataset it is impossible for computer codes to unambiguously construct the relevant likelihood. In the case of the presentation of a correlation (as opposed to covariance) matrix, the diagonal variances must be provided with the first-order dataset.

⁵ Using the elementary uncertainties maybe more convenient when one wishes to include the systematic uncertainties on the signal, *i.e.* α -dependent b, c, ρ . Since these systematics are not crucial for new physics searches we do not take them into account here.

- *Error source format:* A vector of labeled \pm terms associated to each element of the first-order dataset. The correlations between the error sources is indicated via the labels, (e.g. a "stat" label to be a purely diagonal contribution, a "lumi" label to be 100% correlated across all bins, and all other labeled uncertainties treated as orthogonal). The correlation or covariance matrices can be constructed using [TODO?]. This format presents the second-order data in the form of "effective" elementary uncertainties.
- *Auxiliary files in arbitrary format:* the *ad hoc* nature of these makes them impossible to be handled by unsupervised algorithms. This includes 2D histograms in ROOT data files, since variations in path structure and the ambiguity between covariance or correlation matrices are an impediment to automated use. This presentation style will be disregarded below.

The table and error source formats may be readily extended for automated data handling and are thus appropriate to release simplified likelihood data.

In the case of the table format, in addition to the observed central values and statistical uncertainties usually released, extra HepData tables can encode the $m_{1,I}$, $m_{2,IJ}$, $m_{3,I}$ moments describing the combined nuisance parameters. However the HepData table headers will have to be augmented in a standardised fashion to express the relationships between tables, i.e. unambiguously identifying the moment data tables associated with a first-order dataset. While the format is conceptually straightforward, introducing the semantic description of the tables is at present highly impractical. We hence recommend the error source format for which identifying the associations between datasets is trivial.

In the error source format, the $m_{1,I}$, $m_{2,IJ}$, $m_{3,I}$ moments are *all* encoded in the form of labeled vectors. The $m_{2,IJ}$ matrix is reconstructed via a sum of the form

$$m_{2,IJ} = \sum a_{I,i} a_{J,i} \quad (4.3)$$

where the $a_{I,i}$ are the released error sources. The vector of third order data can be indicated via a special label. There is not limit in the number of labels associated to an element hence this format is very flexible. For instance the $a_{I,i}$ error sources corresponding to the decomposed covariance can just get bland names such as "sys,NP1", but can also be extended with *e.g.* a "th" prefix to allow separation of experimental and theory systematics (since the theory can in principle be improved on in future reinterpretations).

This format requires some keyword standardization. The final scheme should be capable of equally applying to any kind of experimental data and systematic uncertainties. In particular it should be valid for event counts, differential cross-sections with bins correlated by the systematic uncertainties, correlations between the bins of different distributions/datasets, and so on.

Summarizing, our recommendation is to release the moments of the combined uncertainty distributions via the HepData error source format, which has built-in semantics of arbitrary complexity and can thus make the most of the simplified likelihood framework.

5 Simplified likelihood in a realistic LHC-like analysis

In this section we introduce a realistic pseudo-analysis which is representative of a search for new physics at the LHC. This analysis will be used to validate the simplified likelihood method and to test its accuracy in realistic conditions. It is also used to validate the simplified likelihood reference code provided in [...]. Finally this pseudo-analysis provides a concrete example of simplified likelihood data release via the HepData error source format. The results of this pseudo-analysis can be reproduced using [codes...].

As already mentioned in Sec. 2, the dominant systematic uncertainties relevant in searches for new physics are those related to the background processes. Indeed, any mis-estimation of the background could result in an erroneous conclusion regarding the presence (or absence) of a signal. There are a number of different ways in which an experimentalist may assess the effect of a given systematic uncertainty, but generally, these effects are parameterized using knowledge of the variations observed in a given process when some underlying parameter of the simulation model, theory, detector resolution etc. Estimates of the contribution from background processes are obtained either from simulation or through some data-driven method. Often such methods are accompanied by associated uncertainties which reduce the ability to make precise statements about the expectation from the background. These systematic uncertainties are important for new physics searches. In the following section, we describe a pseudo-search for new physics, inspired by those performed at the LHC, and derive the simplified likelihood parameters for it.

5.1 A LHC-like pseudo-search for new physics

In order to illustrate the construction of the simplified likelihood, a model has been constructed which is representative of a search for new physics at the LHC. Typically in these searches the observed events are binned into histograms in which the ratio of signal to background contribution varies with the bin number. A search performed in this way is typically referred to as a ‘shape’ analysis as the difference in the distribution (or shape) of the signal events, compared to that of the background, provides crucial information to identify a potential signal.

Our pseudo-search requires to make assumptions for an “observed” dataset, for the corresponding background, and for the new physics signal. These ingredients are summarized in Figure 2. Figure 2 shows the distribution of events, in each of the three categories along with the expected contribution from the background, along with its uncertainties, and from some new physics signal. The ‘nominal’ background follows a typical exponential distribution where fluctuations are present, representing a scenario in which limited Monte Carlo simulation (or limited data in some control sample) was used to derive the expected background contribution. The uncertainties due to this, indicated by the blue band, are uncorrelated between the different bins. Additionally, there are two uncertainties which modify the ‘shape’ of backgrounds, in a correlated way. The effects of these uncertainties are indicated by alternate distributions representing ‘up’ and ‘down’ variations of the systematic uncertainty. Finally, there are two uncertainties which effect only the overall expected rate of the backgrounds. These are indicated in each category as uncertainties

on the normalisation N of the background. These uncertainties are correlated between the three categories and represent two typical experimental uncertainties; a veto efficiency uncertainty (eff.) and the uncertainty from some data-simulation scale-factor (s.f.) which has been applied to the simulation.

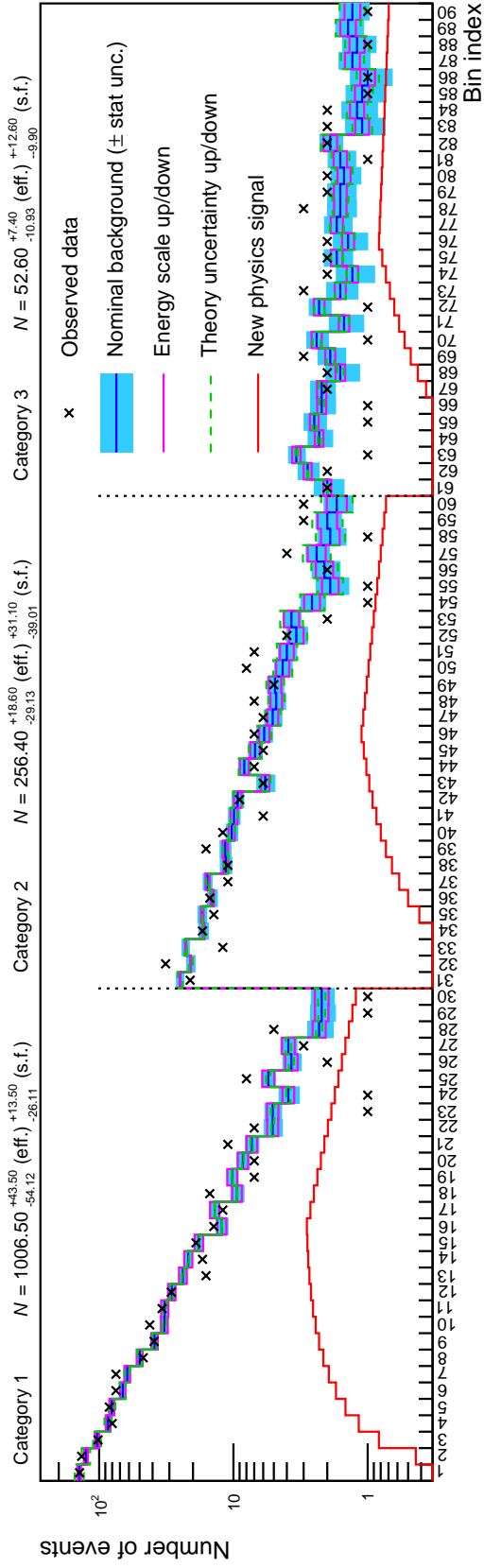


Figure 2. LHC-like search for new physics. The search is performed across three event categories, each divided into 30 bins to make a total of 90 search regions. The nominal expected contribution in each bin from the background and from the new physics signal is shown by the blue and red lines, respectively. The solid and dashed lines show the $\pm 1\sigma$ correlated uncertainty in each bin expected due to an experimental and theoretical uncertainty while the blue shaded band shows the uncorrelated uncertainty in each bin due to limited Monte Carlo simulation. The observed number of events in data in each bin is indicated by the black points.

5.2 Parameterization of backgrounds

It is typical in experimental searches of this type to classify systematic uncertainties into three broad categories, namely; those which affect only the normalization of a given process, those which effect both the ‘shape’ or ‘distribution’ of events of that process in addition to its normalization, and those which affect only a small number of bins or single bin in the distribution and are largely uncorrelated with the other bins (eg uncertainties due to limited Monte Carlo simulation).

The expected (or nominal)⁶ number of background events, due to a particular process, in a given bin (I) in Eqn 2 is denoted by

$$n_{b,I}(\boldsymbol{\delta}) \equiv f_I(\boldsymbol{\delta})N(\boldsymbol{\delta}), \quad (5.1)$$

where the process index (k) is suppressed here as we only have a single background process. The functions $N(\boldsymbol{\delta})$ and $f_I(\boldsymbol{\delta})$ are the total number of expected events for that process in a particular category and the fraction of those events expected in bin I , respectively, for a specified value of $\boldsymbol{\delta}$. Often, these functions are not known exactly and some interpolation is performed between known values of n_I at certain values of $\boldsymbol{\delta}$. For each uncertainty, j , which affect the fractions, f_I , a number of different interpolation schemes exist. One common method however is to interpolate between three distribution templates representing three values of δ_j . Typically, these are for $\delta_j = 0$, the nominal value, and $\delta_j = \pm 1$ representing the plus and minus 1σ variations due to that uncertainty.

The interpolation is given by

$$f_I(\boldsymbol{\delta}) = f_I^0 \cdot \frac{1}{F(\boldsymbol{\delta})} \prod_j p_{Ij}(\delta_j), \quad (5.2)$$

where $f_I^0 = f_I(\boldsymbol{\delta} = 0)$ and $F(\boldsymbol{\delta}) = \sum_I f_I(\boldsymbol{\delta})$ ensures that the fractions sum to 1. In our pseudo-search, as there are three event categories, there are three of these summations, each of which runs over the 30 bins of that category. The polynomial $p_{Ij}(\delta_j)$ is chosen to be quadratic between values of $-1 \leq \delta_j \leq 1$ and linear outside that range such that,

$$p_{Ij}(\delta_j) = \begin{cases} \frac{1}{2}\delta_j(\delta_j - 1)\kappa_{Ij}^- - (\delta_j - 1)(\delta_j + 1) + \frac{1}{2}\delta_j(\delta_j + 1)\kappa_{Ij}^+ & \text{for } |\delta_j| < 1 \\ \left[\frac{1}{2}(3\kappa_{Ij}^+ + \kappa_{Ij}^-) - 2 \right] \delta_j - \frac{1}{2}(\kappa_{Ij}^+ + \kappa_{Ij}^-) + 2 & \text{for } \delta_j > 1 \\ \left[2 - \frac{1}{2}(3\kappa_{Ij}^- + \kappa_{Ij}^+) \right] \delta_j - \frac{1}{2}(\kappa_{Ij}^+ + \kappa_{Ij}^-) + 2 & \text{for } \delta_j < -1 \end{cases} \quad (5.3)$$

The values of κ_{Ij}^- and κ_{Ij}^+ are understood to be determined using the ratios of the template for a -1σ variation to the nominal one and the $+1\sigma$ variation to the nominal

⁶It should be noted that the expectation value for $n_{b,I}$ is *not* necessarily the same as the mean value. For this reason, we typically refer to this as the ‘nominal’ value since it is the value attained when the elementary nuisance parameters are equal to their expectation values $\underline{\delta} = 0$.

one, respectively. The choice of using a quadratic interpolation and linear extrapolation is to avoid assuming too large a variation beyond the known values⁷.

For uncertainties which directly modify the expected number of events n_i of distribution, an exponent interpolation is used as the parameterization. This is advantageous since the number of events, in any given bin, for this process is always greater than 0 for any value of δ_j . For a relative uncertainty ϵ_{Ij} , the fraction varies as

$$\frac{n_{b,I}(\boldsymbol{\delta})}{n_{b,I}^0} = \prod_j (1 + \epsilon_{Ij})^{\delta_j}. \quad (5.4)$$

This is most common in the scenario where a limited number of Monte Carlo simulation events are used to determine the value of $n_{b,I}^0$ and hence some uncertainty is associated. As these uncertainties will be uncorrelated between bins of the distributions, most of the terms ϵ_{Ij} will be 0.

Systematic uncertainties which only affect the overall normalization, are also interpolated using exponent functions,

$$N(\boldsymbol{\delta}) = N^0 \cdot \prod_j (1 + K_j)^{\delta_j}, \quad (5.5)$$

where $N^0 = N(\boldsymbol{\delta} = 0)$ and j runs over the elementary nuisance parameters. A simple extension to this arises if the uncertainty is ‘asymmetric’, as in our pseudo-search; the value of K_j is set to K_j^+ for $\delta_j \geq 0$ and to K_j^- for $\delta_j < 0$. Furthermore, any uncertainty which affects both the shape and the normalization can be incorporated by including terms such as those in Eqn 5.2 in addition to one of these normalization terms. In our pseudo-search, there will be a separate $N(\boldsymbol{\delta})$ term for each category which provides the total expected background rate summing over the 30 bins of that category.

Combining Eqns 5.2, 5.4 and 5.5 yields the full parameterization,

$$n_{b,I}(\boldsymbol{\delta}) = N^0 \cdot \prod_j (1 + K_j)^{\delta_j} \cdot f_I^0 \cdot \frac{1}{F(\boldsymbol{\delta})} \prod_j p_{Ij}(\delta_j) \cdot \prod_j (1 + \epsilon_{Ij} \delta_j). \quad (5.6)$$

As already mentioned, a typical search for new physics will have contributions from multiple background processes, each with their own associated systematic uncertainties. Only by summing over all of these backgrounds (i.e $n_{b,I} = \sum_p n_{b,p,I}$ for different background processes p) is the likelihood fully specified.

5.3 Validation of the simplified likelihood

We constructed 100,000 pseudo-datasets by taking random values $\hat{\boldsymbol{\delta}}$, generated according to $\pi(\boldsymbol{\delta})$, and evaluating $n_{b,I}(\hat{\boldsymbol{\delta}})$ for each dataset according to the Eqn 5.6. Figure 3 shows the distribution of \hat{n}_i , for an example bin, $i = 62$, from the simplified likelihood. The values of m_1 , m_2 and m_3 are calculated using the pseudo-datasets and subsequently used to calculate the coefficients for the simplified likelihood. For comparison, the distribution obtained if the third moment is neglected is also shown.

⁷The validity of this interpolation scheme can (and frequently is) tested by comparing the interpolation to templates for additional, known values of f_I for δ_j values other than 0, -1 and 1 .

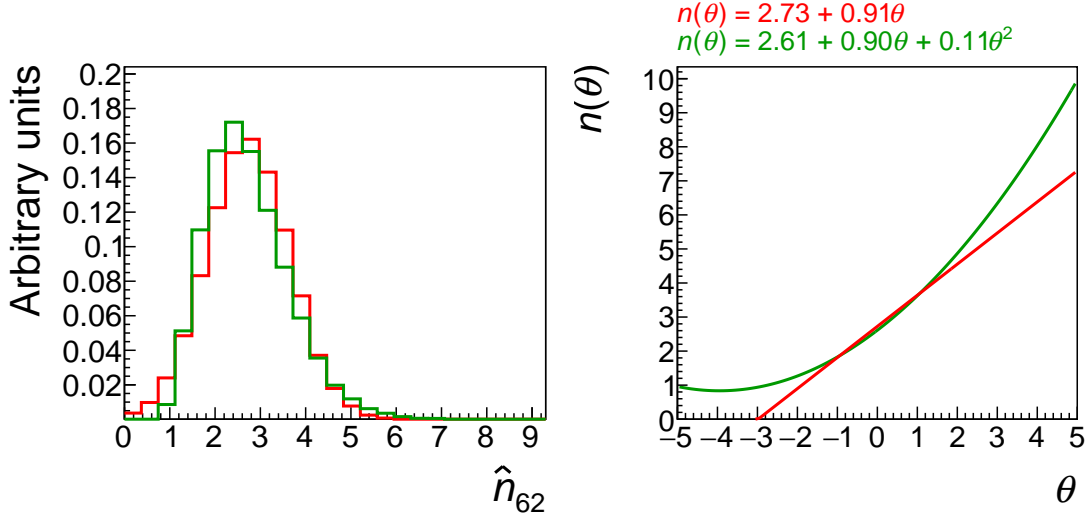


Figure 3. Distributions of \hat{n}_I for $I = 62$ for the simplified likelihood. The functions $n_I(\theta_I)$ assuming the simplified likelihood form (green line), and when neglecting the third moment (red line), are shown in the right panel while the distributions of \hat{n}_I obtained for these two cases letting $\hat{\theta}_I \sim \mathcal{N}(0,1)$ are shown in the left panel.

In Figure 4 2D projections of the background distributions are shown between four pairs of signal-region bins: bin pair (4,7) shows a projection for high-statistics bins where both the linear and quadratic forms of the SL agree closely with the true distribution (that obtained in the pseudo-datasets); the true distribution in (4,62) starts to display deviations from the multivariate normal approximation which are well captured by the quadratic approximation. This is expected when the skew, defined as $m_{3,I}/(m_{2,II})^{\frac{3}{2}}$, is small. However, in the bottom pair of plots with bins 4 and 62 joint with the low-statistics bin 86, the proximity of the mean rate to zero induces a highly asymmetric Poisson distribution which neither approximation can model well. In these last two plots, it can be seen that the quadratic-order SL peaks at too low a value, near a sudden cutoff also seen in Figure 3, while the linear form peaks at too high a value. Systematic relaxation of the quadratic-form cutoff to more closely model the true pdf would require evaluation of higher-order coefficients (and/or off-diagonal skew terms) and hence higher moments of the experimental distributions.

An advantage of the quadratic form cutoff is that a strictly positive approximate distribution can be guaranteed, while the linear form can have a significant negative yield fraction as seen in the figures for bin 86. Sampling from the linear SL form, e.g. for likelihood marginalisation, requires that the background rates be positive since they are propagated through the Poisson distribution. The quadratic SL provides a controlled solution to this issue, as opposed to *ad hoc* methods like use of a log-normal distribution or setting negative-rate samples to zero or an infinitesimal value: the linear approximation has a negative fraction of $\sim 11.6\%$, while the quadratic form has a negative fraction of

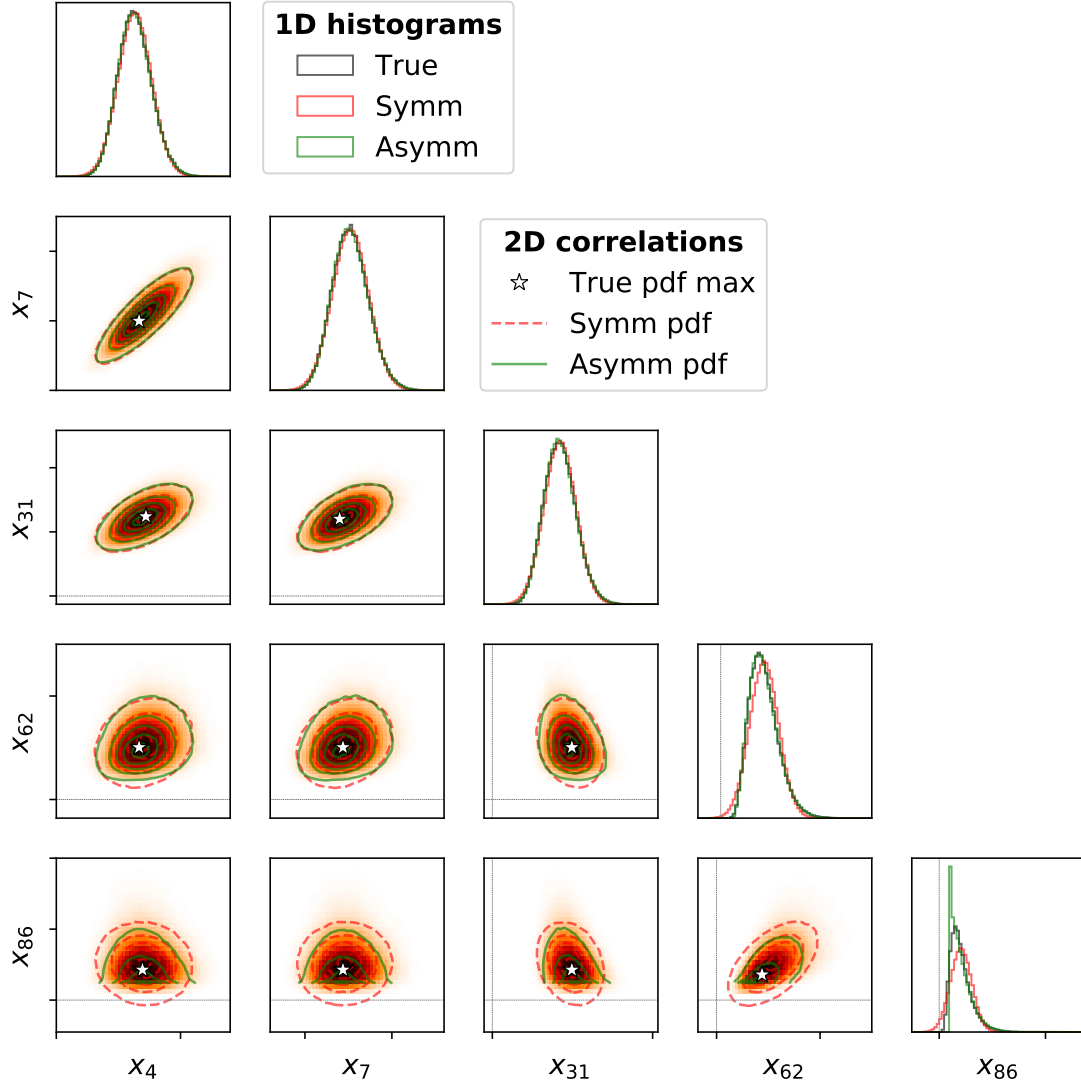


Figure 4. 2D distributions of $\hat{n}_{b,I}$ against $\hat{n}_{b,J}$ for the LHC-like experimental pseudo-search (black points) as described in the text. The background heat map is generated from 100,000 samples from the true model, the dashed red contours from the linear SL form, and the solid green contours from the quadratic SL form. The diagonal panels show the 1D distribution in each of the bins for the toys (black histograms), and the symmetric (red histograms) and asymmetric (green histograms) simplified likelihood approximations. In the pair of high-statistics bins in the top-left plot, clear agreement is seen between the linear and quadratic SL forms; in the top-right, deviations start to appear, and in the low-statistics bin $J = 86$ of the bottom plot the asymmetry is seen to become very significant, and the linear SL form has a significant probability density fraction in the negative-yield region.

exactly zero.

Typically in searches for new physics, limits on models for new physics are determined using ratios of the likelihood at different values of the parameters of interest. In the simplest

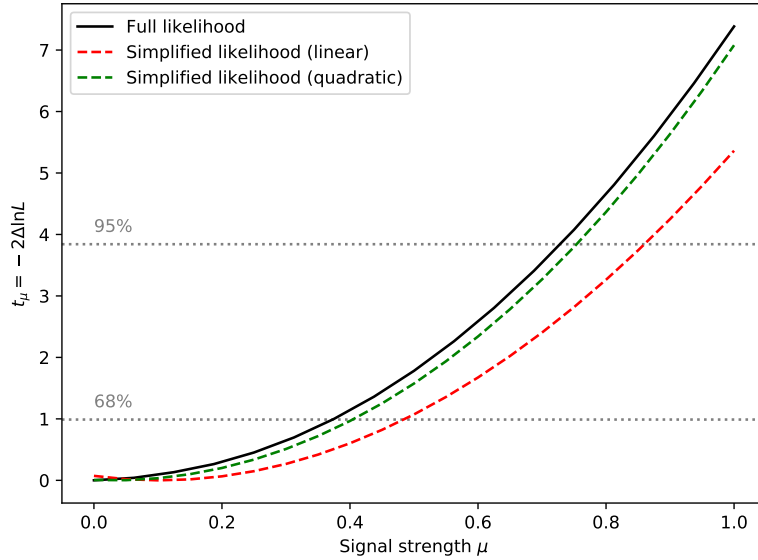


Figure 5. Value of t_μ as a function of μ for the pseudo-search assuming the experimental likelihood (black solid line) and simplified likelihood retaining (green dashed line) or not (red dashed line) the contribution from the quadratic term. The horizontal lines drawn at $t_\mu = 1$ and 3.86 represent the values for which the 68% and 95% CL exclusions can be determined, assuming certain asymptotic properties of the distribution of t_μ .

case, a single parameter of interest is defined as μ , often referred to as the signal strength, which multiplies the expected contribution, under some specific signal hypothesis, of the signal across all regions of the search, giving,

$$n_{s,I}(\boldsymbol{\alpha}) = \mu n_{s,I}, \quad (5.7)$$

where the yields $n_{s,I}$ here refer explicitly to the expected contributions from signal for a specified hypothesis. In order to remove the dependence of the likelihood on the nuisance parameters, $\boldsymbol{\theta}$, two approaches are commonly adopted. The first, termed ‘profiling’, involves replacing the nuisance parameters with the values at which the likelihood attains its maximum for a given set of n_I^{obs} .

$$L_S^{\text{max}}(\mu) = \max_{\boldsymbol{\theta}_I} \{L_S(\mu, \boldsymbol{\theta})\}. \quad (5.8)$$

The test-statistic t_μ is then defined using the ratio,

$$t_\mu = -2 \ln \frac{L_S^{\text{max}}(\mu)}{L_S^{\text{max}}}, \quad (5.9)$$

where L_S^{max} denotes the maximum value of $L_S^{\text{max}}(\mu)$ for any value of μ ⁸. Similarly, such likelihood ratios are also used for quantifying some excess in the case of the discovery of

⁸The precise definition of the test-statistic used as searches at the LHC and the procedures used to determining limits are slightly different to that presented here and are detailed in Ref. [11].

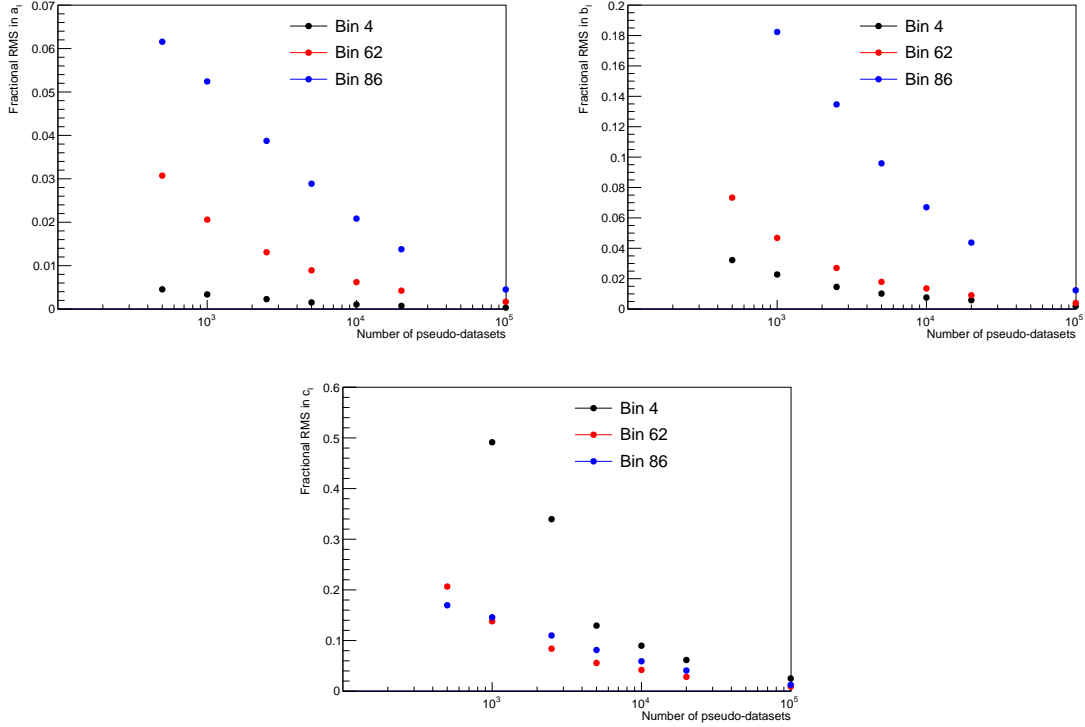


Figure 6. RMS of the simplified likelihood coefficients relative to the mean coefficient value determined from 100,000 pseudo-datasets for a_I (left), b_I (center), and c_I (right). The distributions are shown for $I = 4$ (black line), $I = 62$ (red line) and $I = 86$ (blue line).

new physics [11]. The test-statistic can also be constructed for the experimental likelihood $L(\mu, \delta)\pi(\delta)$, where the same substitution as in Eqn. 5.7 is applied, by profiling the elementary nuisance parameters δ . A direct comparison of the test-statistic for the full and simplified likelihoods is therefore possible.

Figure 5 shows a comparison of the value of t_μ as a function of μ for the pseudo-search between the full (experimental) likelihood and the simplified likelihood. In addition, the result obtained using only the linear terms of the simplified is shown. As expected, the agreement between the full and simplified likelihood is greatly improved when including the quadratic term. A horizontal line is drawn at the value of $t_\mu = 3.86$. The agreement in this region is particularly relevant due to the fact that asymptotic approximations for the distributions of t_μ [12] allow one to determine the 95% confidence level (CL) upper limit on the signal strength, μ_{up} . The signal hypothesis is ‘excluded’ at 95% CL if $\mu_{\text{up}} < 1$.

We performed a study of the rate of convergence of the simplified likelihood coefficients by calculating them using different numbers of pseudo-datasets. The coefficients are calculated for three bins, $i = 4, 50$ and 86 , using independent sets of a fixed number of pseudo-datasets. The root mean square (RMS) of the resulting distribution of each coefficient provides an estimate for how much variation can be expected in the calculation of the coefficients given a limited pseudo-data sample size. The RMS values are normalised to the RMS of the distributions resulting from a sample size of 100,000 pseudo-datasets to

give a relative RMS. Figure 6 shows the relative RMS of the distribution of the coefficients calculated using increasing numbers of pseudo-datasets.

The coefficients a and b can be calculated with relatively high precision using only 1000 pseudo-datasets in each case. The determination of the c coefficient for bin 4 however is slower to converge, requiring 5000–10,000 pseudo-datasets to calculate accurately. However, since the value of m_3 for this bin is relatively small compared to m_2 , the coefficient c is less relevant in this case and so a poor accuracy will have little effect on the accuracy of the simplified likelihood.

6 Conclusion

The transmission of highly complex LHC likelihoods from the experimental collaborations to the public has been a long standing issue. The simplified likelihood (SL) framework, besides having a sound theoretical basis, should satisfy both sides.

Although the SL is accurate, it is still a mere approximation of the actual experimental likelihood, hence experimentalists do not have to release their true likelihood. Meanwhile, for the public, having a good approximation of the true likelihood is arguably much enough for phenomenology purposes. Moreover the SL is very simple to transmit hence it is neither a big effort for the experimentalists to release it nor a big challenge for the user to construct it. Besides, with some standardization effort, part of this transmission process can be automatized.

In this paper we have introduced the formalism of the asymmetric version the SL. The formalism follows directly from the central limit behaviour of the combination of systematic uncertainties: Asymmetry is recognized as the subleading term of the asymptotic distribution dictated by the CLT, which is then recast in a convenient form in the SL formulation. The inclusion of asymmetry was missing in order to obtain a fully reliable SL framework, this question is now settled.

The asymmetric SL can be built either from the elementary systematic uncertainties themselves or from the three first moments of the combination of the systematic uncertainties, which are easily obtained via MC generators. In practice, for the transmission of the SL data from an experiment to the public, our recommendation is to simply release the three first moments of the combined uncertainties, much preferably via the HepData repository in the error source format. The SL framework is flexible in the sense that it can apply to one or more subsets of the systematic uncertainties, and the HepData error source format has adequate flexibility to account for any partitions of the uncertainties the releaser wishes to make.

The construction and use of the SL has been exemplified via a realistic LHC-like pseudo-search for new physics, which is then used to validate the SL method and test its accuracy. This realistic example makes clear that including asymmetry in the SL provides a sensible gain in accuracy, and that it is unlikely that higher moments be ever needed. We find the SL tends to become inaccurate in bins with only a few events. Finally, this pseudo-search is used to provide an example of SL data release. The SL data are available on the HepData repository at [URL...].

Acknowledgements

This work has been initiated at the *LHC Chapter II: The Run for New Physics* workshop held at IIP Natal. AB's work is supported by a Royal Society University Research Fellowship grant. SF's work is supported by the São Paulo Research Foundation (FAPESP) under grants #2011/11973 and #2014/21477-2.

A The CLT at next-to-leading order

Let us show in a 1D example how the skew appears in the asymptotic distribution. Consider N independent centered nuisance parameters δ_j of variance σ^2 and third moment γ . Define

$$Z = \frac{\sum_{j=1}^N \delta_j}{\sqrt{N}}. \quad (\text{A.1})$$

The characteristic function of Z is given by

$$\varphi_Z(t) = \prod_{j=1}^N \varphi_j\left(\frac{t}{\sqrt{N}}\right), \quad (\text{A.2})$$

where $\varphi_j(x) = \mathbf{E}[e^{ix\delta_j}]$. In the large N limit, each individual characteristic function has the expansion

$$\varphi_j\left(\frac{t}{\sqrt{N}}\right) = 1 - \frac{\sigma^2 t^2}{2N} - i \frac{\gamma t^3}{6N^{3/2}} + O\left(\frac{t^4}{N^2}\right). \quad (\text{A.3})$$

It follows that the full characteristic function φ_Z then simplifies to

$$\varphi_Z(t) = \exp\left(-\frac{\sigma^2 t^2}{2} - i \frac{\gamma t^3}{6\sqrt{N}} + O\left(\frac{t^4}{N}\right)\right) \quad (\text{A.4})$$

This characteristic function is simple but has no exact inverse Fourier transform.

To go further, let us observe that the Z random variable could in principle be written in terms of a normally distributed variable $\theta \sim \mathcal{N}(0, \sigma^2)$, with $Z = \phi(\theta)$ where ϕ is a mapping which is in general unknown. At large N however, we know that Z tends to a normal distribution hence ϕ tends to the identity. Thus we can write $Z = \sqrt{N}\phi\left(\frac{\theta}{\sqrt{N}}\right)$ and Taylor expand for large N ,

$$Z = \theta + \frac{c}{2\sqrt{N}}\theta^2 + O\left(\frac{1}{N}\right). \quad (\text{A.5})$$

Let us now compare the characteristic function of this expansion to Eq. (A.4). We find that the characteristic function is given by

$$\varphi_Z(t) = \mathbf{E}\left[e^{it\left(\theta + \frac{c}{2\sqrt{N}}\theta^2 + O\left(\frac{1}{N}\right)\right)}\right] = \exp\left(-\frac{\sigma^2 t^2}{2} - i \frac{ct^3}{2\sqrt{N}} + O\left(\frac{1}{N}\right)\right) \quad (\text{A.6})$$

after using the large N expansion. This function matches Eq. (A.4) for $c = \frac{\gamma}{3}$. Thus we have found the normal expansion provides a way to encode skewness in the large N limit. Namely, we find that the Z variable converges following

$$Z \rightarrow \theta + \frac{\gamma}{3\sqrt{N}}\theta^2, \quad N \rightarrow \infty \quad \text{with} \quad \theta \sim \mathcal{N}(0, \sigma^2). \quad (\text{A.7})$$

When the quadratic term becomes negligible the distribution becomes symmetric, and we recover the usual CLT. As expected, for finite N , we can see that the support of Z is not \mathbf{R} . For example for $\gamma > 0$, we have $Z > -3\sqrt{N}/4\gamma$.

B Reference Code

A reference implementation in Python code is provided in <https://github.com/nucleosynthesis/SL-paper>. This provides functions to calculate the SL a_I , b_I , c_I , and ρ_{IJ} coefficients from provided moments $m_{1,I}$, $m_{2,IJ}$ and $m_{3,I}$; and an `SLParams` class which computes these and higher-level statistics such as profile likelihoods, log likelihood-ratios, and related limit-setting measures computed using observed and expected signal yields. For convergence efficiency, the profile likelihood computation makes use of the gradients of the SL log-likelihood with respect to the signal strength μ and nuisance parameters $\boldsymbol{\theta}$, which we reproduce here to assist independent implementations:

$$\ln(L_S(\mu, \boldsymbol{\theta})\pi(\boldsymbol{\theta})) = \sum_I^P \left[n_I^{\text{obs}} \ln(\mu n_{s,I} + n_I(\theta_I)) - (\mu n_{s,I} + n_{b,I}(\theta_I)) - n_I^{\text{obs}} \right] - \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\rho}^{-1} \boldsymbol{\theta} - \frac{P}{2} \ln 2\pi \quad (\text{B.1})$$

$$\frac{\partial \ln L_S}{\partial \mu} = \sum_I^P \left(\frac{n_I^{\text{obs}}}{\mu n_{s,I} + n_{b,I}(\theta_I)} - 1 \right) \cdot n_{s,I} \quad (\text{B.2})$$

$$\frac{\partial \ln L_S}{\partial \theta_A} = \left(\frac{n_A^{\text{obs}}}{\mu n_{s,A} + n_{b,A}(\theta_A)} - 1 \right) \cdot (b_A + 2c_A \theta_A) - \sum_I^P \rho_{AI}^{-1} \theta_I, \quad (\text{B.3})$$

where $n_{b,I}(\theta_I) = a_I + b_I \theta_I + c_I \theta_I^2$.

The reference code has been written with reverse engineering and comprehensibility of the calculations explicitly in mind. While it computes likelihood statistics on a reasonable timescale, further (but less readable) optimisations can be added for production code.

References

- [1] S. Fichet, *Taming systematic uncertainties at the LHC with the central limit theorem*, *Nucl. Phys. B* **911** (2016) 623 [[1603.03061](#)].
- [2] The CMS Collaboration, *Simplified likelihood for the re-interpretation of public CMS results*, Tech. Rep. CMS-NOTE-2017-001. CERN-CMS-NOTE-2017-001, CERN, Geneva, Jan, 2017.
- [3] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling*, *eConf* **C0303241** (2003) MOLT007 [[physics/0306116](#)].

- [4] L. Moneta, K. Belasco, K. S. Cranmer, S. Kreiss, A. Lazzaro, D. Piparo et al., *The RooStats Project*, *PoS ACAT2010* (2010) 057 [[1009.1003](#)].
- [5] S. Kraml et al., *Searches for New Physics: Les Houches Recommendations for the Presentation of LHC Results*, *Eur. Phys. J. C* **72** (2012) 1976 [[1203.2489](#)].
- [6] F. Boudjema et al., *On the presentation of the LHC Higgs Results*, in *Workshop on Likelihoods for the LHC Searches Geneva, Switzerland, January 21-23, 2013*, 2013, [1307.5865](#), <http://inspirehep.net/record/1244142/files/arXiv:1307.5865.pdf>.
- [7] K. Cranmer, S. Kreiss, D. Lopez-Val and T. Plehn, *Decoupling Theoretical Uncertainties from Measurements of the Higgs Boson*, *Phys. Rev. D* **91** (2015) 054032 [[1401.0080](#)].
- [8] A. Arbey, S. Fichet, F. Mahmoudi and G. Moreau, *The correlation matrix of Higgs rates at the LHC*, *JHEP* **11** (2016) 097 [[1606.0455](#)].
- [9] P. Billingsley, *Probability and Measure*. Wiley, 2012.
- [10] E. Maguire, L. Heinrich and G. Watt, *HEPData: a repository for high energy physics data*, *J. Phys. Conf. Ser.* **898** (2017) 102006 [[1704.05473](#)].
- [11] The ATLAS Collaboration, The CMS Collaboration, The LHC Higgs Combination Group, *Procedure for the LHC Higgs boson search combination in Summer 2011*, .
- [12] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *Eur. Phys. J. C* **71** (2011) 1554 [[1007.1727](#)].