Prepared for submission to JHEP

The Simplified Likelihood Framework

Andy Buckley,[a] **Matthew Citron,**[b] **Sylvain Fichet,**[c,d] **Sabine Kraml,**[e] **Wolfgang Waltenberger,**[f,g] **Nicholas Wardle**[h]

[a] *School of Physics & Astronomy, University of Glasgow, Glasgow, Scotland, UK*
[b] *University of California, Santa Barbara, Santa Barbara, California, USA*
[c] *Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena, CA 91125, California, USA*
[d] *ICTP-SAIFR & IFT-UNESP, R. Dr. Bento Teobaldo Ferraz 271, São Paulo, Brazil*
[e] *Laboratoire de Physique Subatomique et de Cosmologie, Université Grenoble-Alpes, CNRS/IN2P3, 53 Avenue des Martyrs, F-38026 Grenoble, France*
[f] *Institut für Hochenergiephysik, Österreichische Akademie der Wissenschaften, Nikolsdorfer Gasse 18, 1050 Wien, Austria*
[g] *University of Vienna, Faculty of Physics, Boltzmanngasse 5, 1090 Wien, Austria*
[h] *Imperial College London, South Kensington, London, UK*

E-mail: andy.buckley@ed.ac.uk, matthew.citron@cern.ch, sylvain.fichet@gmail.com, sabine.kraml@lpsc.in2p3.fr, walten@hephy.oeaw.ac.at, nckw@cern.ch

Abstract: We discuss the simplified likelihood framework as a systematic approximation scheme for experimental likelihoods such as those originating from LHC experiments. We develop the simplified likelihood from the Central Limit Theorem keeping the next-to-leading term in the large N expansion to correctly account for asymmetries. Moreover, 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. Finally, we discuss how 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.

Contents

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 modelling 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. In the case of the Large Hadron Collider (LHC) experiments, for instance, the experimental likelihood functions used in Standard Model measurements and searches for new physics can contain several hundreds of systematic uncertainties.

Although sources of systematic uncertainty can be numerous and of very different nature, a general feature they share is that their most elementary components tend to be independent from each other. This property of independence between the elementary systematic uncertainties has profound consequences, and, as discussed below, is the reason

why the approach presented in this work is so effective. Namely, independence of the uncertainties can be used to drastically simplify the experimental likelihood function, for the price of an often-negligible error that will be discussed at length in this paper.

The *simplified likelihood* (SL) framework we present in this paper is a well-defined approximation scheme for experimental likelihoods. It can be used to ease subsequent numerical treatment like the computation of confidence limits, 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 experimental likelihoods, in their full form or in convenient approximations, was advocated in Refs. [3, 4]. One possibility is to communicate the full experimental likelihoods via the `RooFit/Roostats` software framework [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 SL 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 refinement is that we provide a way 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 organised as follows. Section 2 introduces the formalism and key points of our approach. The formal material, including an in-depth discussion of the next-to-leading term of the CLT and the derivation of the SL formula, is presented in Section 3. 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. Section 6 contains our summary and conclusions. Two appendices give some more useful details: Appendix A contains a 1D example of how the skew appears in the asymptotic distribution, and Appendix B presents a reference implementation of the SL written in Python.

2 From the experimental likelihood to the simplified likelihood

This section introduces the formalism and an efficient Monte-Carlo based calculation method. Some preliminary remarks are in order. From the conceptual point of view, the SL framework relies only on the convergence of the CLT. In practice however, the representation of the SL will depend on broad, structural features of the dataset under consideration. The case considered in this paper is a set of P independent observables in the presence of N independent sources of uncertainties, with $N \geq P$. For a dataset with different structure, the SL would take a different form, but this is not a fundamental limitation of the approach *per se*. Moreover, in the scope of a given problem, *e.g.* the search for new physics in our case, additional approximations may simplify the formalism. Again, this should not be understood as a fundamental limitation, as such approximations could be removed in a different application. A summary of the validity conditions for the SL treated in this paper will be given in Section 6.

In the following, we will focus on the typical experimental likelihood used in searches for new phenomena at particle physics experiments. However, as argued above, the SL approach can be easily generalised to other physics contexts. The data collected in particle physics usually originate from random (quantum) processes, and have thus 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 Pr to observe the data given a model $\mathcal{M}$, specified by some parameters,

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

We denote the observed quantity as n^{obs} 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 $\boldsymbol{\alpha}$, and *elementary* nuisance parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_j \ldots, \delta_N)^{\mathrm{T}}$, which model the systematic uncertainties. In the SL framework, we derive a set of *combined* nuisance parameters $\boldsymbol{\theta}$. For P independent measurements, there will be P combined nuisance parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I, \ldots, \theta_P)^{\mathrm{T}}$.

The key result at the basis of the SL framework is the approximation

$$L(\boldsymbol{\alpha}, \boldsymbol{\delta})\pi(\boldsymbol{\delta}) = \prod_{I=1}^{P} \Pr\Big(n_I^{\text{obs}} \,\Big|\, n_I(\boldsymbol{\alpha}, \boldsymbol{\delta})\Big)\pi(\boldsymbol{\delta}) \tag{2.2}$$

$$\approx \prod_{I=1}^{P} \Pr\Big(n_I^{\text{obs}} \,\Big|\, a_I(\boldsymbol{\alpha}) + b_I(\boldsymbol{\alpha})\theta_I + c_I(\boldsymbol{\alpha})\theta_I^2\Big) \cdot \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\rho}^{-1}(\boldsymbol{\alpha})\boldsymbol{\theta}}}{\sqrt{(2\pi)^P}} \equiv L_{\mathrm{S}}(\boldsymbol{\alpha}, \boldsymbol{\theta})\,, \tag{2.3}$$

where the first line is the exact "experimental likelihood" and the second line is the SL. Here $\pi(\boldsymbol{\delta})$ 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

$\pi(\boldsymbol{\delta}) = \prod_{i=1}^{N} \pi_i(\delta_i)$. The SL formalism shown here is relevant for $N \geq P$, which is also the most common case.[1] The derivation is shown in Sec. 3.

The coefficients a_I, b_I and c_I, and the $P \times P$ correlation matrix $\boldsymbol{\rho} = \rho_{IJ}$ define the SL 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 of 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. Neglecting the systematic uncertainties affecting the signal implies in turn that the parameters of the SL are independent of $\boldsymbol{\alpha}$. Hence the SL Eq. (2.3) takes the form [2]

$$L_{\mathrm{S}}(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{I=1}^{P} \Pr\Big(n_I^{\mathrm{obs}} \,\Big|\, n_{s,I}(\boldsymbol{\alpha}) + a_I + b_I\theta_I + c_I\theta_I^2\Big) \cdot \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\rho}^{-1}\boldsymbol{\theta}}}{\sqrt{(2\pi)^P}}, \tag{2.4}$$

which is the expression we use in the rest of this paper. (Non-negligible signal uncertainties will be commented on in Sec. 4.1.) The expression Eq. (2.4) is valid for data with *any statistics of observation*. Since the data in particle physics are often observed event counts, n_I^{obs}, they will typically follow Poisson statistics such that

$$\Pr(n_I^{\mathrm{obs}}|n_I) \equiv \mathrm{Pois}(n_I^{\mathrm{obs}}|n_I) = \frac{(n_I)^{n_I^{\mathrm{obs}}}\,\mathrm{e}^{-n_I}}{n_I^{\mathrm{obs}}!}\,. \tag{2.5}$$

However, as mentioned, the formalism presented here applies regardless of the dependence on the parameters of interest. For example, the likelihood can very well be multimodal in the parameters of interest; this does not affect the validity of the approach.

The parameters of the SL $(a_I, b_I, c_I, \rho_{IJ})$ have analytical expressions as a function of the variance and the skew of each elementary nuisance parameter (see Section 3.2). However, often the elementary uncertainties and the event yields are already coded in a Monte Carlo (MC) generator. In this case, an elegant method to obtain the SL 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 SL 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 $m_{3,I} \equiv m_{3,III}$.

[1] If $P < N$, there are more observed quantities than nuisance parameters. In such case, using the SL at the level of the event rates, although not formally wrong, is inappropriate. Equation (2.3) still applies but the covariance matrix will be singular. In the case of Higgs theoretical uncertainties for example, a more appropriate combination is done at the level of cross-sections and branching ratios, as realised in [8]. Another example is the one of unbinned likelihoods, for which parametric functions for the signal and background probability densities are typically used to construct the experimental likelihood. The systematic uncertainties are then on the parameters of the signal and background functions. Notice that in such case, the shapes can be directly provided in their analytic form by the experimental collaborations.

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

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

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

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

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

where $\mathbf{E}$ denotes the expectation value. Inverting these relations, while taking care to pick the relevant solutions to quadratic and cubic equations, gives the parameters of the SL. We find

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

$$b_I = \sqrt{m_{2,II} - 2c_I^2}\,, \tag{2.10}$$

$$a_I = m_{1,I} - c_I\,, \tag{2.11}$$

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

These formulae apply if the condition $8m_{2,II}^3 \geq m_{3,I}^2$ is satisfied. Near this limit, the asymmetry becomes large and the approximation inaccurate because higher order terms $O(\theta_I^3)$ would need to be included in Eq. (2.3). In practice, however, this requires a high skewness of the nuisance parameters, and the SL 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 SL parameters are completely defined. Moreover, in the case where the nuisance parameters affect only the background rate Eq. (2.4), this computation has to be realised only once and the resulting likelihood can be used for any kind of signal by appropriate substitution of $n_s(\boldsymbol{\alpha})$.

A reference code implementing the SL and subsequent test statistics is described in Appendix B and publicly available at `https://gitlab.cern.ch/SimplifiedLikelihood/SLtools`.

3 The simplified likelihood from the central limit theorem

This section contains the derivation of the SL formula Eq. (2.3). The reader interested only in the practical aspects of the SL framework can safely skip it. In Section 3.1 we lay down a result about the next-to-leading term of the CLT. In Section 3.2 we then demonstrate Eq. (2.3) and give the analytical expressions of the SL parameters as a function of the elementary uncertainties. The precision of the expansion is discussed in Section 3.3.

3.1 Asymmetries and CLT at next-to-leading order

In the classical proof of the CLT, a Taylor expansion is applied to the characteristic functions of the random variables. Within this Taylor expansion, usually only the leading term

is considered, resulting in an asymptotically Normal behavior for the sum of the random variables. In the context of the SL framework, however, the next-to-leading term in the CLT's large N expansion is also considered. This next-to-leading term encodes skewness, which encodes 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. Another reason to take the asymmetry into account is that event yields are defined on $\mathbf{R}^+$, while the normal distribution takes values on $\mathbf{R}$. Thus, keeping only the leading order distribution can lead to negative yields. Such unphysical results can be interpreted as an indicator that the leading order approximation (namely the normal distribution) is too inaccurate. When taking the next-to-leading term into account, an asymmetric support – such as $\mathbf{R}^+$ – becomes possible, such that the issue of negative yields disappears. Concrete examples of this feature will be shown in Fig. 1. [3]

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

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

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

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

We will refer to this type of expression as "normal expansion". Details about its derivation are given in Appendix 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 method works similarly with the Lyapunov CLT, *i.e.* when the δ_j are not identical and have different moments σ_j^2, γ_j, 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 approach applies similarly to the multidimensional case where various linear combinations of the δ_j give rise to various Z_I. The Z_I have a covariance matrix Σ_{IJ} and a skewness tensor $\gamma_{IJK} = \mathrm{E}[Z_I Z_J Z_K]$. For our purposes, we neglect the non-diagonal elements of γ, keeping only the diagonal elements, denoted $\gamma_{III} \equiv \gamma_I$. 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,

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

[3] It is in principle possible to truncate the Gaussian prior by requiring that the expected background plus signal be positive. However in the presence of signal uncertainties with truncated Gaussian prior, the posterior can become improper (see *e.g.* [9]). This can be understood as a pathology of such approach. In contrast, the alternative we propose does not require truncation.

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

3.2 Calculation of the simplified likelihood

Let us now prove Eq. (2.3). The dependence on the parameters of interest $\boldsymbol{\alpha}$ 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 generalisation of the approach of [1]. Here we take into account the skew, hence there is no need to use an exponential parameterisation 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.*

$$\mathbf{E}[\delta_i] = 0\,, \quad \mathbf{E}[\delta_i^2] = 1\,, \quad \mathbf{E}[\delta_i^3] = \gamma_i\,. \tag{3.4}$$

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

$$n[Q] \equiv n[Q_0(1 + \Delta_Q \delta)]\,. \tag{3.5}$$

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

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

with

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

and the $n^{(3)}$ denoting 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 centred and with unit variance without loss of generality, and whose correlation matrix is denoted ρ_{IJ},*i.e.*

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

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

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

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

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

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

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

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

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

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

$$n_I \approx n_I^0 \left(1 + \Delta_{1,I} \cdot \boldsymbol{\delta}\right) \tag{3.14}$$

from Eq. (3.9). We now make the key observation that this quantity is the 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 [10]. We can for instance use Lyapunov's condition on the third moment, and the theorem reads as follows. If

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

then

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

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, 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).

[4] For 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 use the exact solutions of the system, Eqs. (2.9)–(2.12).

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 situation in which the next-to-leading CLT, Eq. (3.3), tends to fail. It is instructive to check on a simple case how the normal expansion approximates the true distribution, and in which way 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

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

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 Section 2). This is the limit 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 $n > a - b^2/4c$. Below this limit on σ, the lower bound on n is roughly $n \gtrsim 0.5$, therefore the approximation can never produce a negative event yield.

To check numerically how well the approximation performs, the true and approximate PDFs are compared in Figure 1 for various values of σ. Since the approximate PDF never gives $n < 0.5$, it can only be a good approximation if the true PDF is vanishing in this region. This is the case for asymmetries $\sigma \lesssim 0.3$, and as can be seen in the figure the normal approximation indeed works very well. For larger asymmetries, $\sigma = 0.45$ in our example, the true PDF becomes sizeable in the region $n < 0.5$. The approximation still performs reasonably well for larger n, however, near $n \sim 0.5$, the approximate PDF tends to increase and become peaked to account for the area at $n < 0.5$ that it cannot reproduce. This behaviour will also be observed for certain bins in the LHC-like analysis implemented in Sec. 5.

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 SL framework will be tested in a realistic setup in Sec. 5.

4 Practical aspects of the simplified likelihood framework

4.1 Range of application

An important feature of the SL 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 SL to be accurate. There is thus a freedom in

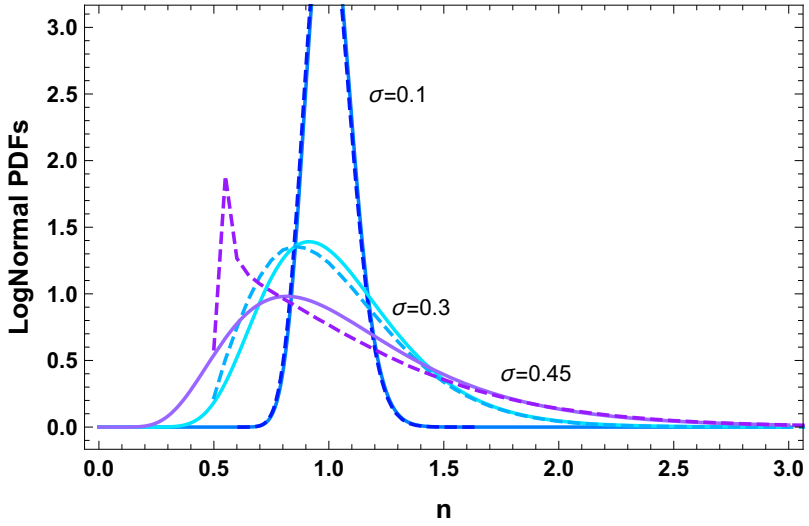

Figure 1. The log normal PDFs and corresponding normal approximations for $\sigma = 0.1$, 0.3 and 0.45 are shown in blue, cyan and purple respectively. Solid curves show the true distributions, dashed curves show the approximate distributions.

partitioning the set of systematic uncertainties, giving rise to variants of the SL that can be either equivalent or slightly different upon marginalising.

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

$$L_{\mathrm{S}}(\boldsymbol{\alpha},\boldsymbol{\theta}) = \prod_{I=1}^{P} \Pr\Big(\hat{n}_I \,\Big|\, a_I(\boldsymbol{\alpha}) + b_I\theta_I + c_I\theta_I^2 + \Delta_I\delta\Big) \cdot \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\rho}^{-1}\boldsymbol{\theta}}}{\sqrt{(2\pi)^P}} \cdot \pi(\delta)\,. \tag{4.1}$$

Similarly, if two subsets of systematic uncertainties $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ tend to separately satisfy the CLT condition, they can be separately combined, giving

$$L_{\mathrm{S}}(\boldsymbol{\alpha},\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = \prod_{I=1}^{P} \Pr\Big(\hat{n}_I \,\Big|\, a_I(\boldsymbol{\alpha}) + b_I\theta_I + c_I\theta_I^2 + \tilde{b}_I\tilde{\theta}_I + \tilde{c}_I\tilde{\theta}_I^2\Big) \cdot \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\rho}^{-1}\boldsymbol{\theta}}}{\sqrt{(2\pi)^P}} \cdot \frac{\mathrm{e}^{-\frac{1}{2}\tilde{\boldsymbol{\theta}}^{\mathbf{T}}\tilde{\boldsymbol{\rho}}^{-1}\tilde{\boldsymbol{\theta}}}}{\sqrt{(2\pi)^P}}\,. \tag{4.2}$$

The SL 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 not only for informative purpose but also for further interpretations. For example the theoretical uncertainties may be improved later on and it is clearly of advantage if their effect can be re-evaluated without having to re-analyse the whole data (which could only be done by collaboration insiders).[5] 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

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

to take into account exactly. In order to profit from the versatility of the SL, an equally versatile format is needed to release the SL data. This will be the topic of next subsection.

Finally, some considerations are in order regarding *signal uncertainties*. The expected rate n given in Eq. (3.6) splits as $n = s + b$ where s is the signal and b the background. Each elementary nuisance parameter δ_i can in principle affect both s and b. The most general form taken by the expected rate is then

$$\begin{aligned} n = s + b &\equiv s^0 \left(1 + \Delta_{1,s}^T \cdot \boldsymbol{\delta} + \boldsymbol{\delta}^{\mathrm{T}} \cdot \Delta_{2,s} \cdot \boldsymbol{\delta}\right) + b^0 \left(1 + \Delta_{1,b}^T \cdot \boldsymbol{\delta} + \boldsymbol{\delta}^{\mathrm{T}} \cdot \Delta_{2,b} \cdot \boldsymbol{\delta}\right) \\ &= (s^0 + b^0) \left(1 + \Delta_1^T \cdot \boldsymbol{\delta} + \boldsymbol{\delta}^{\mathrm{T}} \cdot \Delta_2 \cdot \boldsymbol{\delta}\right) \end{aligned} \tag{4.3}$$

with

$$\Delta_1 = \frac{s^0 \Delta_1^s + b^0 \Delta_1^b}{s^0 + b^0}, \quad \Delta_2 = \frac{s^0 \Delta_2^s + b^0 \Delta_2^b}{s^0 + b^0}. \tag{4.4}$$

The Δ_1^s, Δ_2^s vectors encode the contributions from the signal, while the Δ_1^b, Δ_2^b vectors encode the contributions to the background. The signal s^0 and possibly Δ_1^s, Δ_2^s depend on the parameters of interest $\boldsymbol{\alpha}$. For discovery or limit-setting, the uncertainties on the signal are a subleading effect. In this paper, as said in Sec. 2, we have neglected signal uncertainties ($\Delta_1^s = \Delta_2^s = 0$). In this approximation, only the background uncertainties remain in Eq. (4.3), and thus the SL does not depend on $\boldsymbol{\alpha}$. A similar discussion of the signal+background case and a toy-model testing the SL in this case has been done in [1].

The inclusion of pure signal uncertainties is fairly straightforward because their contribution factors out from the background ones. In (4.4), this means that the vectors $\Delta_{1,2}$ can be simply organised as the union of the subvectors $\Delta_{1,2} = (\Delta_{1,2}^b, \Delta_{1,2}^s)$. This implies that the pure signal uncertainties do not affect the SL parameters a, b, c, ρ, and can thus be rigorously included directly within the existing SL (for $\Delta_1^s = \Delta_2^s = 0$).

In contrast, for correlated systematic uncertainties affecting both signal and background — for instance in case of measurements as opposed to limit setting, or when interference effects between signal and background are important — the b, c, ρ parameters become dependent on the parameters of interest $\boldsymbol{\alpha}$. This requires to (re-)derive the SL taking into account all the elementary nuisance parameters at once, which is a much heavier task.

Altogether, while there is no conceptual difference regarding the SL formalism with or without signal uncertainties, there are important practical implications. Numerical evaluations become much heavier when the parameters of the SL—especially $\rho_{IJ}(\alpha)$ which requires a matrix inversion—have to be evaluated for each value of $\boldsymbol{\alpha}$. The presentation of the SL data, iscussed in the next subsection, may also become more evolved. Furthermore, and perhaps most importantly, the SL is then valid only for the particular signal assumption it has been derived for.

4.2 Construction and presentation

There are in principle two ways of releasing the data needed to build the simplified likelihood. One way is to release the whole set of elementary systematic uncertainties, the other 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 MC generators, hence it is straightforward to evaluate these moments.[6]

We thus focus on the release of the SL 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 way of presentation needs to be considered in detail, 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 [11].

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 parametrisation. 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. The presentation of first-order data is standardised while currently 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 difficulty that the convention used is not made clear (other than by inspection of the matrix diagonal), and without a structural association with a 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.

- *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

[6] 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.

Eq. (4.5). 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 SL 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

$$m_{2,IJ} = \sum a_{I,i} a_{J,i} \tag{4.5}$$

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 standardisation. 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.

Summarising, 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 SL framework. As a showcase example, we provide the pseudo-data used in the next section as a sandbox HepData record at https://www.hepdata.net/record/sandbox/1535641814.

5 Simplified likelihood in a realistic LHC-like analysis

In this section we introduce a realistic pseudo-analysis that is representative of a search for new physics at the LHC. This analysis will be used to validate the SL method and to

test its accuracy in realistic conditions. It is also used to validate the SL reference code presented in Appendix B. Finally, this pseudo-analysis provides a concrete example of SL data release via the HepData table format (see above). The SL and subsequent results of the pseudo-search can be reproduced using these data.

As already mentioned in Section 2, the dominant systematic uncertainties relevant in searches for new physics are those related to the background processes. Imperfect knowledge of detector effects or approximations used in the underlying theoretical models will lead to uncertainties in the predictions of these processes. 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 parameterised using knowledge of how the estimation of a given process which change under variations of 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 data-driven methods. In the following section, we describe a pseudo-search for new physics, inspired by those performed at the LHC, in which systematic uncertainties are included, and derive the SL parameters for it.

5.1 A LHC-like pseudo-search for new physics

In order to illustrate the construction of the SL, 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 summarised in Figure 2, which shows the distribution of events, in each of three categories along with the expected contribution from the background and the uncertainties thereon, and from some new physics signal. The "nominal" background follows a typical exponential distribution where fluctuations are present, representing a scenario in which limited MC 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.

5.2 Parameterisation 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 normalisation of a given process, those which effect both the "shape" or "distribution" of events of that process in addition to its normalisation, 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 MC simulation).

The expected (or nominal)[7] number of background events, due to a particular process, in a given bin (I) in Eq. (2.3) is denoted by

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

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

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

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,

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

[7] It 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 $\delta = 0$.

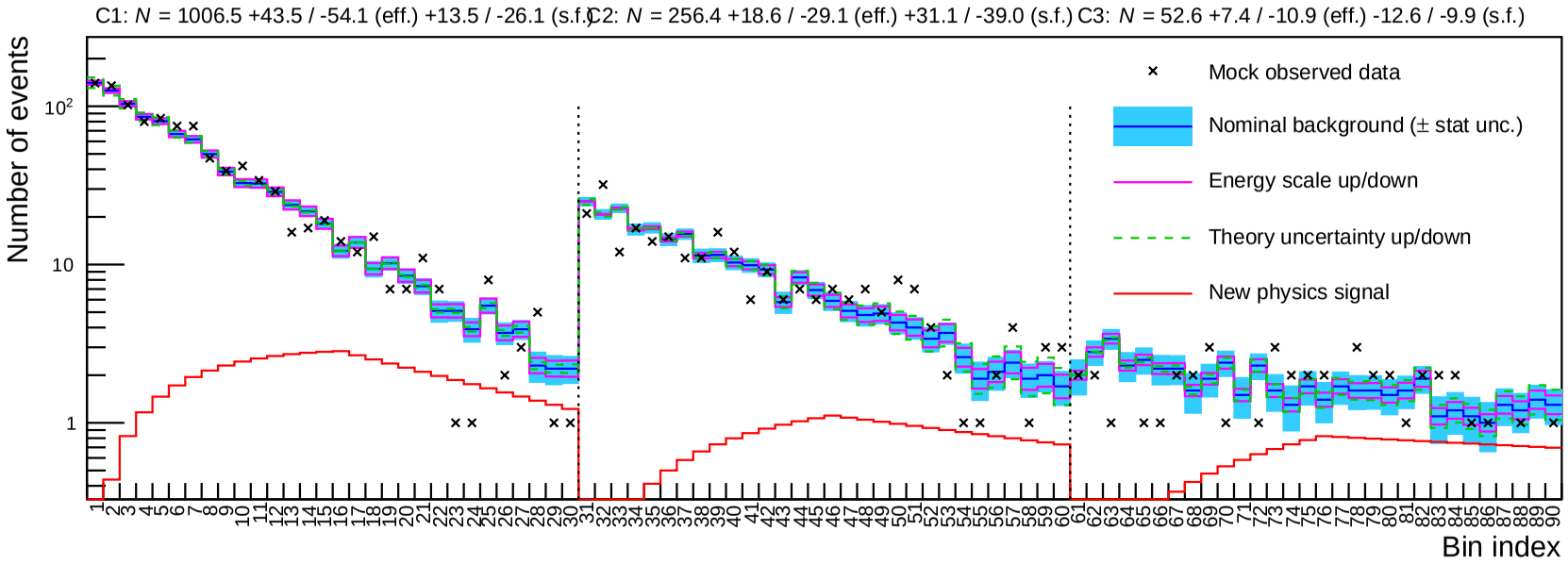

Figure 2. LHC-like search for new physics (mockup). The search is performed across three event categories, each divided into 30 bins to make a total of 90 search regions. The nominal expected contribution in each bin from the background and from the new physics signal is shown by the blue and red lines, respectively. The solid and dashed lines show the $\pm 1\sigma$ correlated variation 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 MC simulation. The "observed" number of events in data in each bin is indicated by the black points.

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 one, respectively[8].

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

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

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

Systematic uncertainties that affect only the overall normalisation are also interpolated using exponent functions,

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

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 normalisation can be incorporated by including terms such as those in Eq. (5.2) in addition to one of these normalisation 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 Eqs. (5.2), (5.4) and (5.5) yields the full parameterisation,

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

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

Here we compare the true and simplified likelihoods arising from the pseudo-search. It is also instructive to consider the simplified likelihood obtained when neglecting the third moments, *i.e.* when setting the coefficients of the quadratic terms c_I to zero in Eq. (2.3). This less accurate version of the SL will be referred to as "symmetric SL", as opposed to the more precise "asymmetric SL" developed in this work.

[8]The accuracy of this interpolation scheme can be (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.

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 Eq. (5.6). Figure 3 shows the distribution of $\hat{n}_i$, for an example bin, $i = 62$, from the SL. 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 SL.

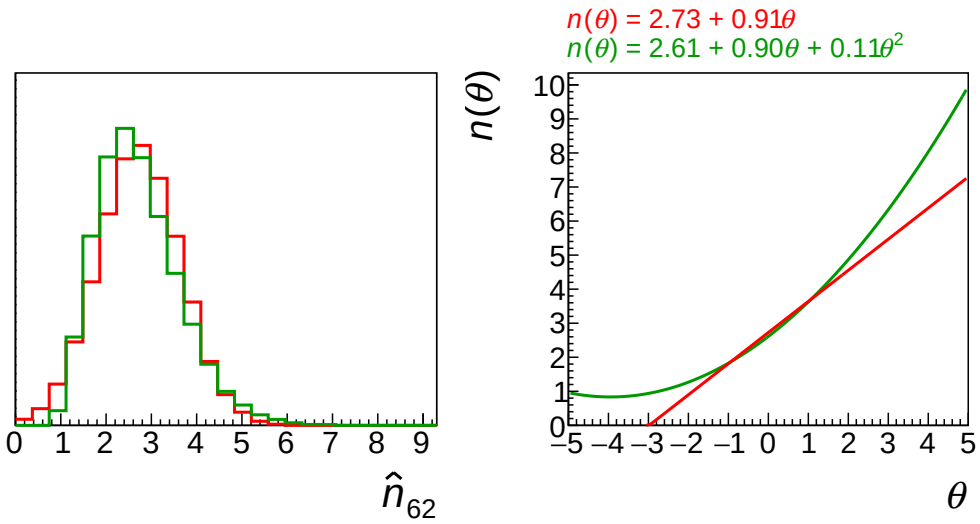

Figure 3. Distributions of $\hat{n}_I$ for $I = 62$ for the SL. The functions $n_I(\theta_I)$ assuming the SL form (green line), and when neglecting the third moment (red line), are shown in the right panel while the distributions of $\hat{n}_I$ obtained for these two cases letting $\hat{\theta}_I \sim \mathcal{N}(0, 1)$ 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 asymmetric and symmetric 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 asymmetric SL. 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 SLs can model well. In these last two plots, it can be seen that the asymmetric SL peaks at too low a value, near a sudden cutoff also seen in Figure 3, while the symmetric SL peaks at too high a value. In this region a better modelling 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 asymmetric SL is that a strictly positive approximate distribution can be guaranteed, while the symmetric SL can have a significant negative yield fraction as seen in the figures for bin 86. Sampling from the symmetric SL, *e.g.* for likelihood marginalisation, requires that the background rates be positive since they are propagated through the Poisson distribution. The asymmetric SL provides a controlled solution to

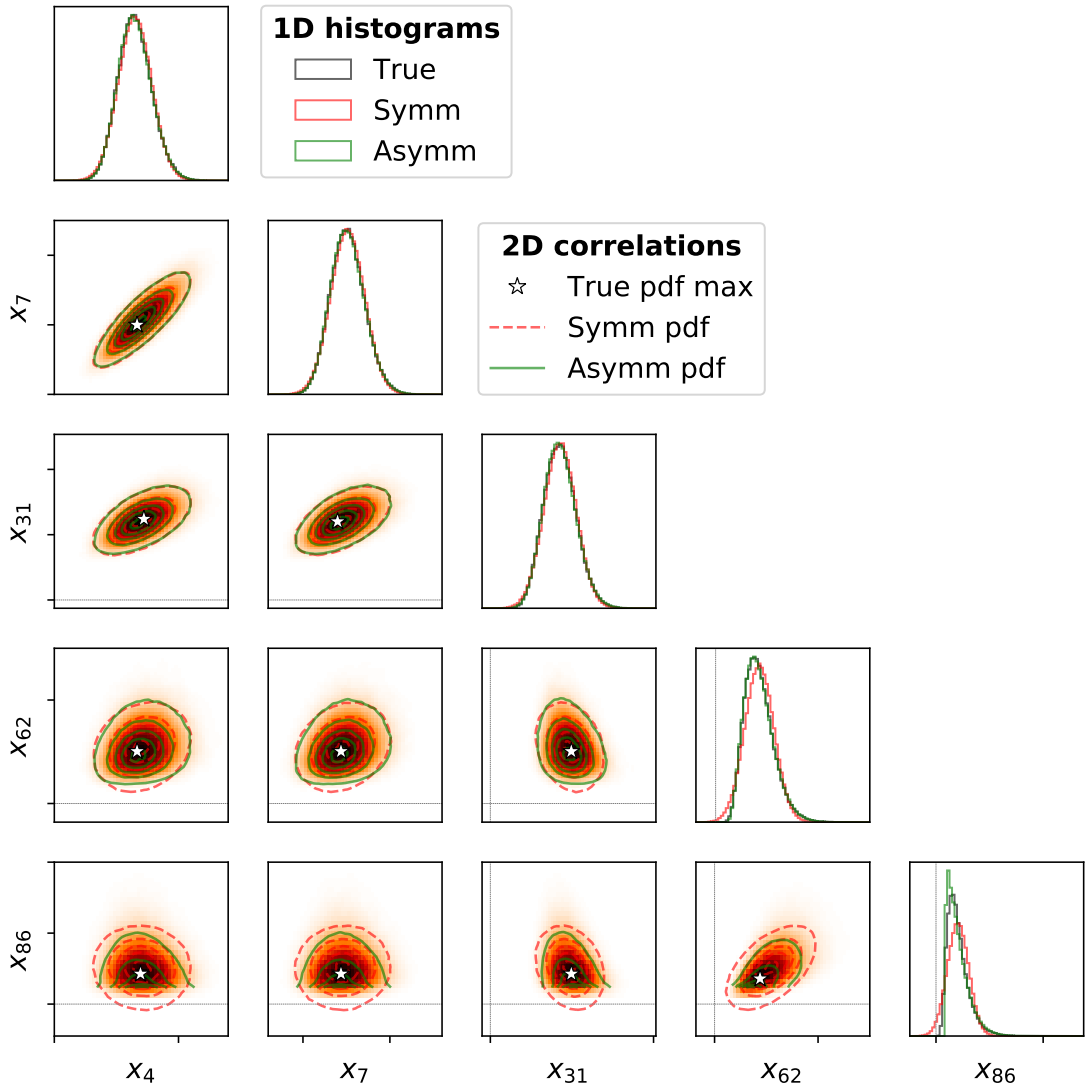

Figure 4. 2D distributions of $\hat{n}_{b,I}$ against $\hat{n}_{b,J}$ for the LHC-like experimental pseudo-search as described in the text. The background heat map is generated from 100,000 samples from the true model, the dashed red contours from the symmetric SL, and the solid green contours from the asymmetric SL. 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) SLs. In the pair of high-statistics bins in the top-left plot, clear agreement is seen between the symmetric and asymmetric SLs; 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 symmetric SL form has a significant probability density fraction in the negative-yield region.

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 symmetric SL has a negative fraction of $\sim$11.6%, while the asymmetric SL has a negative fraction of exactly zero.

Typically in searches for new physics, limits on models for new physics are determined

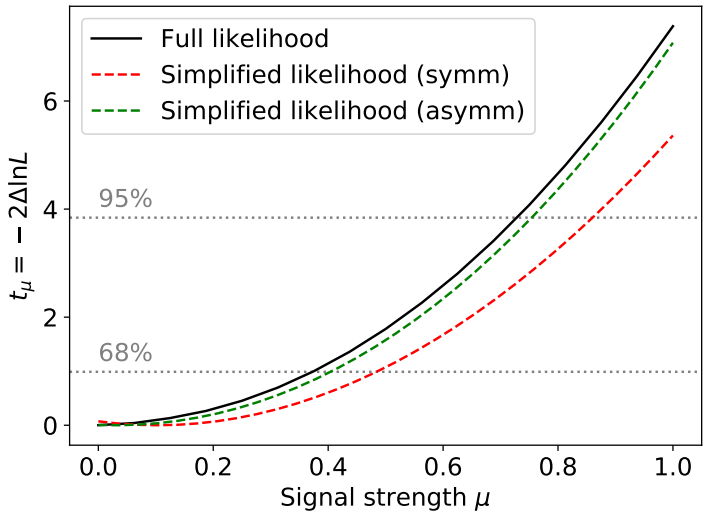

Figure 5. Value of t_μ as a function of μ for the pseudo-search assuming the experimental likelihood (black solid line) and simplified likelihood retaining (green dashed line) or not (red dashed line) the contribution from the quadratic term. The horizontal lines drawn at $t_\mu = 1$ and 3.86 represent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

using ratios of the likelihood at different values of the parameters of interest. In the simplest 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,

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

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, $\boldsymbol{\theta}$, the nuisance parameter values are set to those at which the likelihood attains its maximum for a given set of n_I^{obs}. This is commonly referred to as "profiling" over the nuisance parameters[9].

$$L_{\text{S}}^{\max}(\mu) = \max_{\theta_{\text{I}}} \{\text{L}_{\text{S}}(\mu, \boldsymbol{\theta})\}\,. \tag{5.8}$$

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

$$t_\mu = -2\ln \frac{L_{\text{S}}^{\max}(\mu)}{L_{\text{S}}^{\max}}, \tag{5.9}$$

[9]Other procedures, such as marginalisation, can also be used to remove the dependence on the nuisance parameters. For reviews on how likelihoods, such as the simplified likelihood presented here, are used in searches for new physics, see Refs. [12, 13]

where $L_S^{\max}$ denotes the maximum value of $L_S^{\max}(\mu)$ for any value of μ.[10] Similarly, such likelihood ratios are also used for quantifying some excess in the case of the discovery of new physics [14]. The test-statistic can also be constructed for the experimental likelihood $L(\mu, \boldsymbol{\delta})\pi(\boldsymbol{\delta})$, where the same substitution as in Eq. (5.7) is applied, by profiling the elementary nuisance parameters $\boldsymbol{\delta}$. A direct comparison of the test-statistic for the full and simplified likelihoods, as a function of μ, 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 asymmetric SL. In addition, the result obtained using only the symmetric SL 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 [15] 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 $\mu_{\text{up}} < 1$.

When determining the SL coefficients, we have relied on pseudo-datasets, as we expect this will often be the case for anyone providing SL inputs for real analyses. The accuracy of the SL coefficients will necessarily depend on the number of pseudo-datasets used to calculate them. To investigate this, we have performed a study of the rate of convergence of the SL coefficients by calculating them using several different numbers of pseudo-datasets, the largest being 100,000 pseudo-datasets. The coefficients for the three bins calculated using 100,000 pseudo-datasets are; $a = 84.9,\ b = 8.27,\ c = 0.32$ for bin 4, $a = 2.61,\ b = 0.90,\ c = 0.11$ for bin 62, and $a = 0.90,\ b = 0.47,\ c = 0.13$ for bin 86. The calculation of the coefficients is repeated using many independent sets of a fixed number of pseudo-datasets, resulting in a distribution of calculations for each coefficient.

The root mean square (RMS) of the resulting distributions provides an estimate for how much variation can be expected in the calculation of the SL 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. The relative RMS of the distribution of the coefficients calculated using increasing numbers of pseudo-datasets is shown Figure 6.

The coefficients a and b can be calculated with relatively high precision using only 1000 pseudo-datasets in each case. This is true whether the value of b is large compared to a, as in the case of bin 86, or not, as in the case of bin 4. 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 c for this bin is relatively small compared to b, the coefficient c is less relevant so that a poor accuracy will have little effect on the accuracy of the SL. In bin 86, the value of c is relatively large, compared to b, meaning it will significantly contribute to the SL. In this case, the convergence is quite fast, with only 2,500 pseudo-datasets required to achieve a 10% accuracy in the value of c. We find the property that bins with large c values, compared to b values, require fewer pseudo-datasets

[10] The precise definition of the test-statistic used as searches at the LHC and the procedures used to determine limits are slightly different to that presented here and are detailed in Ref. [14].

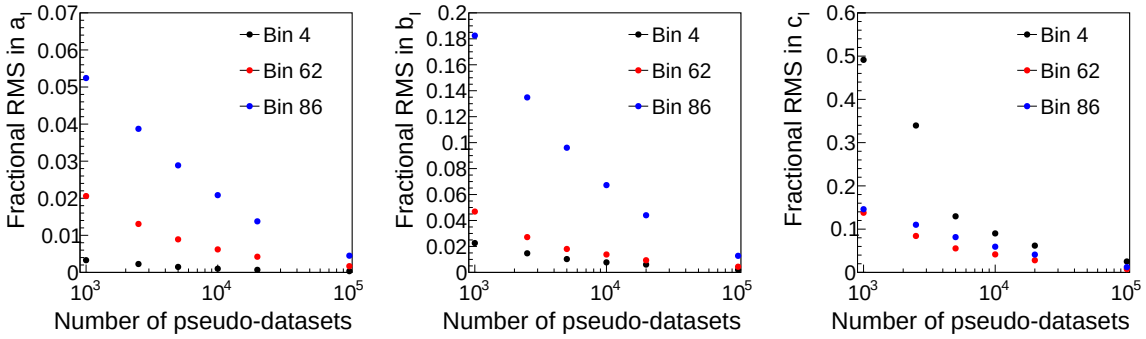

Figure 6. RMS of the SL coefficients relative to the mean coefficient value determined from 100,000 pseudo-datasets for a_I (left), b_I (middle), and c_I (right). The distributions are shown for $I = 4$ (black points), $I = 62$ (red points) and $I = 86$ (blue points).

to achieve a good accuracy than bins for which the c value is less relevant generally holds in this study.

6 Summary and conclusions

The transmission of highly complex LHC likelihoods from the experimental collaborations to the scientific community has been a long standing issue. In this paper, we proposed a simplified likelihood framework which can account for non-Gaussianities as a convenient way of presentation with a sound theoretical basis.

Although the SL is accurate, it is still an approximation of the full experimental likelihood, hence the collaborations do not have to release their full model. Meanwhile, for the public, having a good approximation of the true likelihood is sufficient for most phenomenology purposes. Moreover, the SL is very simple to transmit, requiring neither a substantial effort for the experimentalists to release it nor for the user to construct it. Additionally, with some standardisation effort, part of this transmission process can be automated.

In this paper we introduced the formalism for the asymmetric version of the SL. This formalism follows directly from the central limit behaviour of the combination of systematic uncertainties: asymmetry is recognised 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 completes the SL and provides a fully reliable framework.

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. Using a realistic LHC-like pseudo-search for new physics, we demonstrated that including asymmetry in the SL provides an important gain in accuracy, and that it is unlikely that higher moments will be needed.

The SL formalism discussed in this paper focusses on datasets with more systematic uncertainties than observables (*i.e.* $N \geq P$), and a few extra simplifying approximations have been made. The conditions of its validity are summarised as follows:

- *Convergence of the central limit theorem*: There should be enough independent sources of uncertainties for the combined distribution to tend towards a Gaussian. This is the fundamental condition underlying the SL approach. The leading, asymmetric corrections to the Gaussian can be treated as described in this work.

- *Sufficiently symmetric combined uncertainties*: Alltough we have consistently included skewness in our formalism, it cannot be arbitrarily large, as discussed in Section 3.3. In particular the formulas used to derive the SL coefficients are valid only when the second ($m_{2,II}$) and third ($m_{3,I}$) moments satisfy $8m_{2,II}^3 \geq m_{3,I}^2$.

- *Negligible signal uncertainties*: In order to be re-usable for different signal hypotheses, e.g., for limit setting on different models, the SL must not depend on the parameters of interest. This is ensured if signal uncertainties are negligible to good approximation. While the inclusion of pure signal uncertainties is straightforward, systematic uncertainties that are correlated between signal and background must be included in the derivation of the SL coefficients. Without this, the simplification to Eq. (2.4) is no longer valid, as discussed in Section 4.1.

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, 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.

If adopted by the experimental and theory communities, and provided the above validity conditions are respected, the SL framework has the potential to considerably improve both the documentation and the re-interpretation of the LHC results.

Acknowledgements

This work has been initiated at the *LHC Chapter II: The Run for New Physics* workshop held at IIP Natal, Brazil, 6–17 Nov. 2017. We thank the IIP Natal for hosting the workshop and creating a most inspiring working atmosphere.

AB is supported by a Royal Society University Research Fellowship grant. MC is supported by the US Department of Energy under award number DESC0011702. SF is supported by the São Paulo Research Foundation (FAPESP) under grants #2011/11973, #2014/21477-2 and #2018/11721-4. SK is supported by the IN2P3 project "Théorie LHC-iTools" and the CNRS-FAPESP collaboration grant PRC275431. NW is funded through a Science and Techologies Facility Council (STFC) Fellowship grant #ST/N003985/1.

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

$$Z = \frac{\sum_{j=1}^{N} \delta_j}{\sqrt{N}} \,. \tag{A.1}$$

The characteristic function of Z is given by

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

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

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

It follows that the full characteristic function φ_Z then simplifies to

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

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,

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

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

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

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

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

When the quadratic term becomes negligible the distribution becomes symmetric, and we recover the usual CLT. We can see that for finite N (as opposed to $N \to \infty$) 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, `simplike.py`, is provided in

https://gitlab.cern.ch/SimplifiedLikelihood/SLtools.

It includes 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:

$$\begin{aligned}\ln\big(L_{\mathrm{S}}(\mu,\boldsymbol{\theta})\pi(\boldsymbol{\theta})\big) = &\sum_I^P \Big[n_I^{\mathrm{obs}} \ln\big(\mu n_{s,I} + n_{b,I}(\boldsymbol{\theta})\big) - \big(\mu n_{s,I} + n_{b,I}(\boldsymbol{\theta})\big) - n_I^{\mathrm{obs}}!\Big] \\ &- \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\rho}^{-1}\boldsymbol{\theta} - \frac{P}{2}\ln 2\pi\end{aligned} \tag{B.1}$$

$$\frac{\partial \ln L_{\mathrm{S}}}{\partial \mu} = \sum_I^P \left(\frac{n_I^{\mathrm{obs}}}{\mu n_{s,I} + n_{b,I}(\boldsymbol{\theta})} - 1\right) \cdot n_{s,I} \tag{B.2}$$

$$\frac{\partial \ln L_{\mathrm{S}}}{\partial \theta_A} = \left(\frac{n_A^{\mathrm{obs}}}{\mu n_{s,A} + n_{b,A}(\boldsymbol{\theta})} - 1\right) \cdot \big(b_A + 2c_A\theta_A\big) - \sum_I^P \rho_{AI}^{-1}\,\theta_I \ , \tag{B.3}$$

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

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.

A demo of the construction of the simplified likelihood, and profiling as a function of a signal strength parameter, is given in `simplikedemo.py`. Finally, the SL pseudo-data are available on the HepData repository at https://www.hepdata.net/record/sandbox/1535641814.

References

[1] S. Fichet, *Taming systematic uncertainties at the LHC with the central limit theorem*, *Nucl. Phys.* **B911** (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, Geneva, Jan, 2017.

[3] S. Kraml et al., *Searches for New Physics: Les Houches Recommendations for the Presentation of LHC Results*, *Eur. Phys. J.* **C72** (2012) 1976 [1203.2489].

[4] F. Boudjema et al., *On the presentation of the LHC Higgs Results*, 1307.5865.

[5] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling*, *eConf* **C0303241** (2003) MOLT007 [physics/0306116].

[6] L. Moneta, K. Belasco, K. S. Cranmer, S. Kreiss, A. Lazzaro, D. Piparo et al., *The RooStats Project*, *PoS* **ACAT2010** (2010) 057 [1009.1003].

[7] K. Cranmer, S. Kreiss, D. Lopez-Val and T. Plehn, *Decoupling Theoretical Uncertainties from Measurements of the Higgs Boson*, *Phys. Rev.* **D91** (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] CDF collaboration, L. Demortier, *Objective Bayesian Upper Limits for Poisson Processes*, Tech. Rep. CDF/MEMO/STATISTICS/PUBLIC/5928, 2005.

[10] P. Billingsley, *Probability and Measure*. Wiley, 2012.

[11] E. Maguire, L. Heinrich and G. Watt, *HEPData: a repository for high energy physics data*, *J. Phys. Conf. Ser.* **898** (2017) 102006 [1704.05473].

[12] G. Cowan, C. Patrignani et al., *Probability,* Ch. 38 in: *Review of particle physics*, *Chin. Phys. C* **40** (2016) 100001.

[13] L. Lyons and N. Wardle, *Statistical issues in searches for new phenomena in high energy physics*, *Journal of Physics G: Nuclear and Particle Physics* **45** (2018) 033001.

[14] ATLAS, CMS, LHC Higgs Combination Group collaboration, *Procedure for the LHC Higgs boson search combination in summer 2011*, Tech. Rep. ATL-PHYS-PUB-2011-011, CMS-NOTE-2011-005, CERN, Geneva, 2011.

[15] 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].