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*
[e]
[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. This framework can be used to simplify data analyses and to transmit realistic experimental likelihoods to the community. 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 toy search. ...

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. [**?** **?**], in which promising preliminary results have been shown.

In the context of the LHC, communicating the full experimental likelihoods via the `RooFit/Roostats` software framework [**?** **?**] has been suggested in Refs. [**?** **?**]. 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. [**?**] 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. [**?**].

In this paper we unify and extend the initial proposals of Refs. [**?** **?**], and thoroughly test the accuracy of the approximations using simulated LHC searches for new phenomena. Compared to Refs. [**?** **?**], 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. As 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.

Section 4 contains details about LHC likelihoods. A first validation of the simplified likelihood framework is done in Section 4.1.

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,

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

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 $\boldsymbol{\alpha}$, and *elementary* nuisance parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_j \ldots, \delta_N)^{\mathrm{T}}$, which model the systematic uncertainties. In the simplified likelihood 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 simplified likelihood framework is the approximation

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

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

where the first line is the exact experimental likelihood and the second line is the simplified likelihood. 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 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 form [1]

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

which is the expression we use in the rest of this paper. 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,

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

The parameters of the simplified likelihood $(a_I, b_I, c_I, \rho_{IJ})$ have analytical expressions as a function of the variance and the skew of each elementary nuisance parameter (see 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

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

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

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

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

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

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

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

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

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

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 $n_s(\boldsymbol{\alpha})$.

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 truely 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

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

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 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 $\gamma_{IJK} = \mathrm{E}[Z_I Z_J Z_K]$. 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,

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

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

Here we 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 generalization of the approach of [**?**]. Here we take into account the skew, hence there is no need to use an exponential parameterization like in [**?**].

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 [**?**] 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

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

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.*

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

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. [**?**].[2]

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

[2]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 solve the system shown in Eqs. (2.6)-(2.8), (2.9)-(2.12).

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 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 [**?**]. We can for instance use Lyapunov's condition on the third moment, and the theorem reads: If

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

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 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

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

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 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.

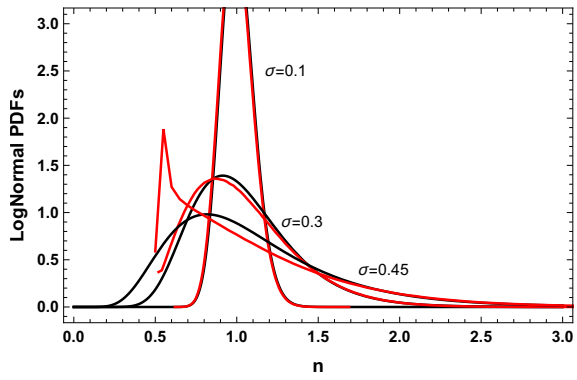

Figure 1. The normal approximation for a Poisson and a log-normal distribution. Black curves show the true distributions, red curve show the approximate one.

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 realistic toy search 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 Simplified likelihoods for LHC searches for new physics

This section contains LHC-specific comments and details related to the simplified likelihood framework.

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 toy search for new physics, inspired by those performed at the LHC, and derive the simplified likelihood parameters for it.

4.1 A toy search for new physics

In order to illustrate the construction of the simplified likelihood, a toy 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 the separation needed to identify a potential signal. 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.

4.2 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)[3] number of background events, due to a particular process, in a given bin (I) in Eqn 2 is denoted by

$$n_{b,I}(\boldsymbol{\delta}) \equiv f_I(\boldsymbol{\delta})N(\boldsymbol{\delta}), \tag{4.1}$$

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 is the total number of expected events for that process in

[3]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 $\underline{\delta} = 0$.

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

$$f_I(\boldsymbol{\delta}) = f_I^0 \cdot \frac{1}{F(\boldsymbol{\delta})} \prod_j p_{Ij}(\delta_j), \tag{4.2}$$

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 toy 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,

$$p_{Ij}(\delta_j) = \begin{cases} \frac{1}{2}\delta_j(\delta_j - 1)\kappa_{Ij}^- - (\delta_j - 1)(\delta_j + 1) + \frac{1}{2}\delta_j(\delta_j + 1)\kappa_{Ij}^+ & \text{for } |\delta_j| < 1 \\ \left[\frac{1}{2}(3\kappa_{Ij}^+ + \kappa_{Ij}^-) - 2\right]\delta_j - \frac{1}{2}(\kappa_{Ij}^+ + \kappa_{Ij}^-) + 2 & \text{for } \delta_j > 1 \\ \left[2 - \frac{1}{2}(3\kappa_{Ij}^- + \kappa_{Ij}^+)\right]\delta_j - \frac{1}{2}(\kappa_{Ij}^+ + \kappa_{Ij}^-) + 2 & \text{for } \delta_j < -1 \end{cases} \tag{4.3}$$

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 one, respectively. The choice of using a quadratic interpolation and linear extrapolation is to avoid assuming to large a variation beyond the known values[4].

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

$$\frac{n_{b,I}(\boldsymbol{\delta})}{n_{b,I}^0} = \prod_j (1 + \epsilon_{Ij})^{\delta_j}. \tag{4.4}$$

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,

$$N(\boldsymbol{\delta}) = N^0 \cdot \prod_j (1 + K_j)^{\delta_j}, \tag{4.5}$$

[4]The 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.

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 toy 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 4.2 in addition to one of these normalization terms. In our toy 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 4.2, 4.4 and 4.5 yields the full parameterization,

$$n_{b,I}(\boldsymbol{\delta}) = N^0 \cdot \prod_j (1 + K_j)^{\delta_j} \cdot f_I^0 \cdot \frac{1}{F(\boldsymbol{\delta})} \prod_j p_{Ij}(\delta_j) \cdot \prod_j (1 + \epsilon_{Ij}\delta_j). \tag{4.6}$$

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.

4.3 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 4.6. 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.

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 infinitessimal value: the toy model linear approximation has a negative fraction of $\sim$11.6%, while the quadratic form has a negative fraction of 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 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,

$$n_{s,I}(\boldsymbol{\alpha}) = \mu n_{s,I}, \tag{4.7}$$

where the yields $n_{s,I}$ here refer explicitly to the expected contributions from signal for a specified hypothesis. In order to remove the dependance 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 it maximum for a given set of n^{obs_I}. FiXme: Note, here I am using L to mean the whole thing but in Eqn 1 we refer to L as only part of the likelihood, i.e we need to multiply by pi(d). I wonder if it makes sense to define L_{simp} as the production of these things and make it a function of the thetas ? FiXme!

$$L^{\max}(\mu) = \max_{\theta_\text{I}} \{\text{L}(\mu, \boldsymbol{\theta})\}\,. \tag{4.8}$$

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

$$t_\mu = -2 \ln \frac{L^{\max}(\mu)}{L^{\max}}, \tag{4.9}$$

where $L^{\max}$ denotes the maximum value of $L_{\max}(\mu)$ for any value of μ[5]. Similarly, likelihood ratios are also used for quantifying some excess in the case of the discovery of new physics [**?**].

Figure 6 shows a comparison of the value of t_μ as a function of μ for the toy 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 [**?**] 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 $\mu_{\text{up}} < 1$.

FiXme: WOLFGANG: Plots showing profile & marginalisation upper-limit extractions between true, symm/linear, and asymm/quadratic forms FiXme!

5 Construction and distribution of Simplified Likelihood data

FiXme: Give an introduction to the experimental/stats issues in extraction of stable and consistent 2nd and 3rd order correlation moments. Note the necessity of cross-checking and sanity-checking moment estimates: covariances have previously been published which, due to numerical rounding issues, are singular and hence unusable. FiXme!

Figure 7 shows the RMS of the simplified likelihood coefficients for the three bins, $i = 4, 50$ and 86 relative to the values determined from 100,000 pseudo-datasets.

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 [**?**].

Unfortunately, at present there is no standard semantic representation of second order (i.e. covariance) correlation data, let alone the third order "skew" information. At present a review of the correlation information in HepData and on the experiments' analysis websites reveals a mixture of second-order data presentation styles:

- 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 a "primary" dataset of values/first moments it is impossible for computer codes to unambiguously construct the relevant likelihood. In the case of a normalised correlation representation cf. ρ, the primary dataset must also provide the diagonal variances.

- A breakdown by error-source, e.g. a series of labelled $\pm$ terms for each value in the primary dataset. From this, with some conventions (e.g. a "stat" label to be a purely diagonal contribution, a "lumi" label to be 100% correlated across all bins, and all other labelled uncertainties treated as orthogonal) the correlation or covariance matrices can be constructed.

- 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 forms are an impediment to automated use.

The first two of these forms may be readily extended for automated correlation handling. In the first case, the HepData table headers may be augmented to express relationships between tables, i.e. identifying the second (and third) moment data tables associated with a primary dataset in a standardised fashion. And in the second case, the SL covariance matrices may be represented in error-source form, with each bin I reporting N different error sources of value $\sqrt{m_{2,IJ}}$ as well as specially identified skew and statistical error components.

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

Both forms require some keyword standardisation: in the first case to express the semantic types of datasets and the relationships between them, and in the second case to identify the statistical and skew error sources which should be treated distinctly from the SL systematic uncertainties. The final scheme should be capable of equally applying to search-analysis signal regions as discussed here, and to bins of differential cross-section observables where correlations may also be available between the bins of different distributions/datasets. For such observables, the Poisson–Gaussian likelihood form presented in this paper may not be the appropriate one to use – e.g. a pure multivariate Gaussian cf. χ^2 testing is more appropriate for normalised differential observables – and so a further record annotation will be required to identify the appropriate likelihood family into which to cast the provided correlation data.

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

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

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. 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, and an `SLParams` class which computes these and applies them in combination with observed and expected signal yields to calculate profile and marginal likelihoods, log likelihood-ratios, and limit-setting test statistics. 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:

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

$$\Rightarrow \qquad \frac{\mathrm{d}\ln L}{\mathrm{d}\mu} = \sum_I^P \left(\frac{n_I^{\text{obs}}}{\mu n_{s,I} + n_{b,I}(\theta_I)} - 1\right) \cdot \mu n_{s,I} \tag{B.2}$$

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

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

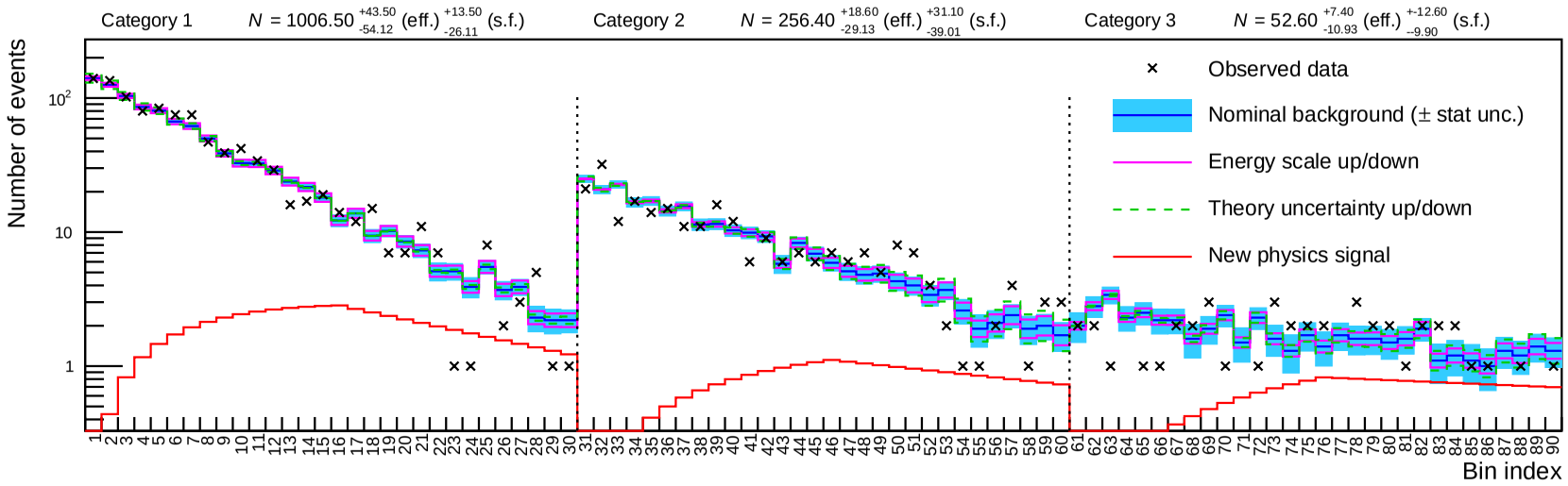

Figure 2. Toy search for new physics. The search is performed across three event categories, each divided into 30 bins to make a total of 90 search regions. The nominal expected contribution in each bin from the background and from the new physics signal is shown by the blue and red lines, respectively.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

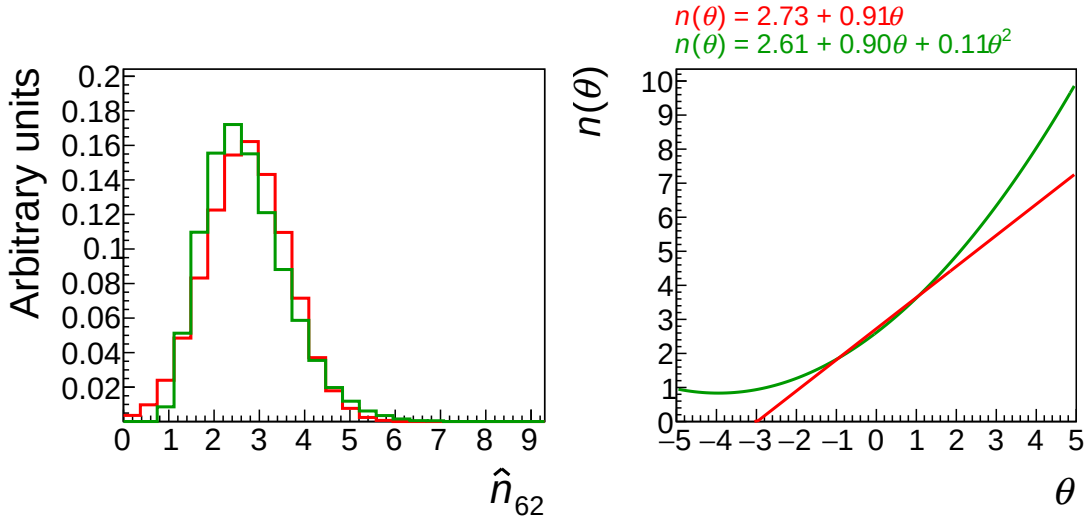

Figure 3. Distributions of $\hat{n}_I$ for $I = 62$ for the simplified likelihood. The functions $n_I(\theta_I)$ assuming the simplified likelihood form (green line), and when neglecting the third moment (red line), are shown in the right panel while the distributions of $\hat{n}_I$ obtained for these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

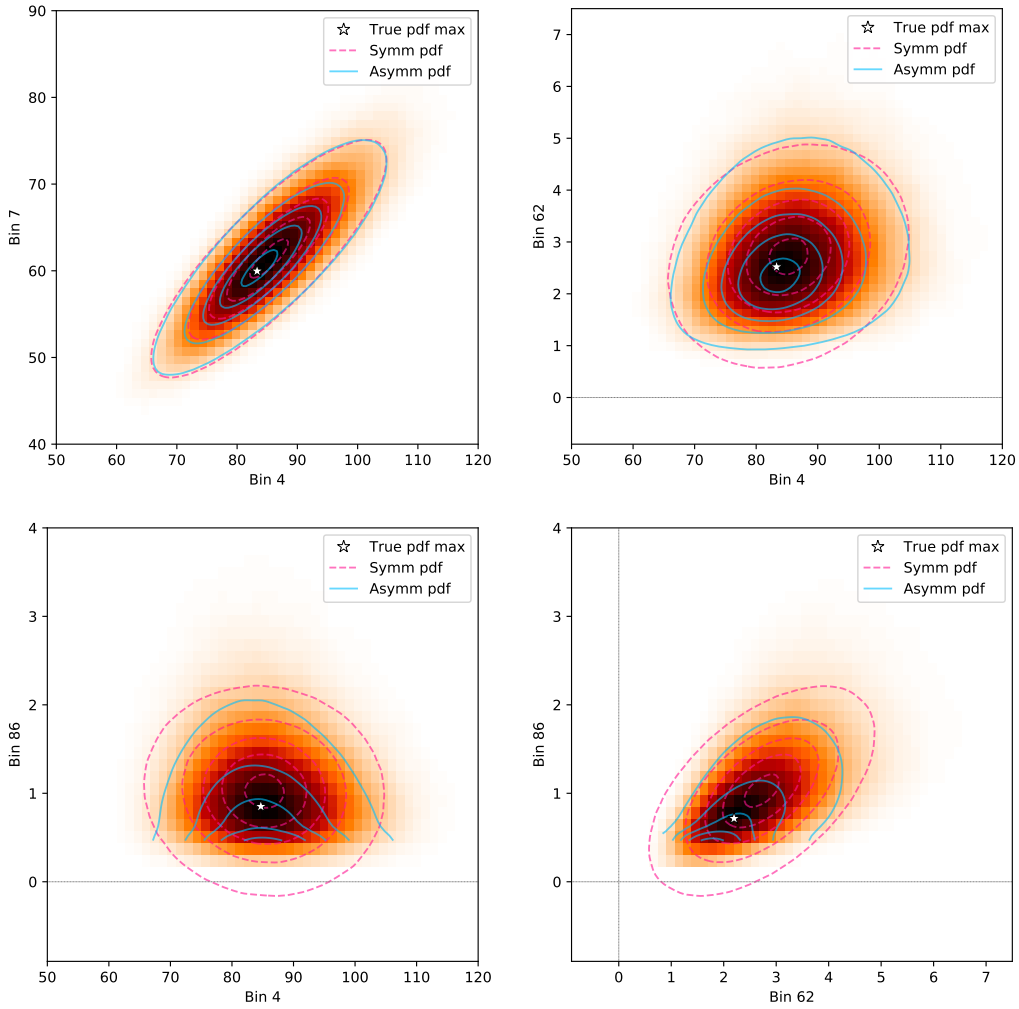

Figure 4. 2D distributions of $\hat{n}_{b,I}$ against $\hat{n}_J$ for $I = 4, J = 7$ (top left), $I = 4, J = 62$ (top right), $I = 4, J = 86$ (bottom left), and $I = 62, J = 86$ (bottom right) in pseudo-datasets generated from the experimental toy search (black points) as described in the text. The background heat map is generated from 100,000 samples from the true toy model, the dashed pink countours from the linear SL form, and the solid light-blue contours from the quadratic SL form. 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.

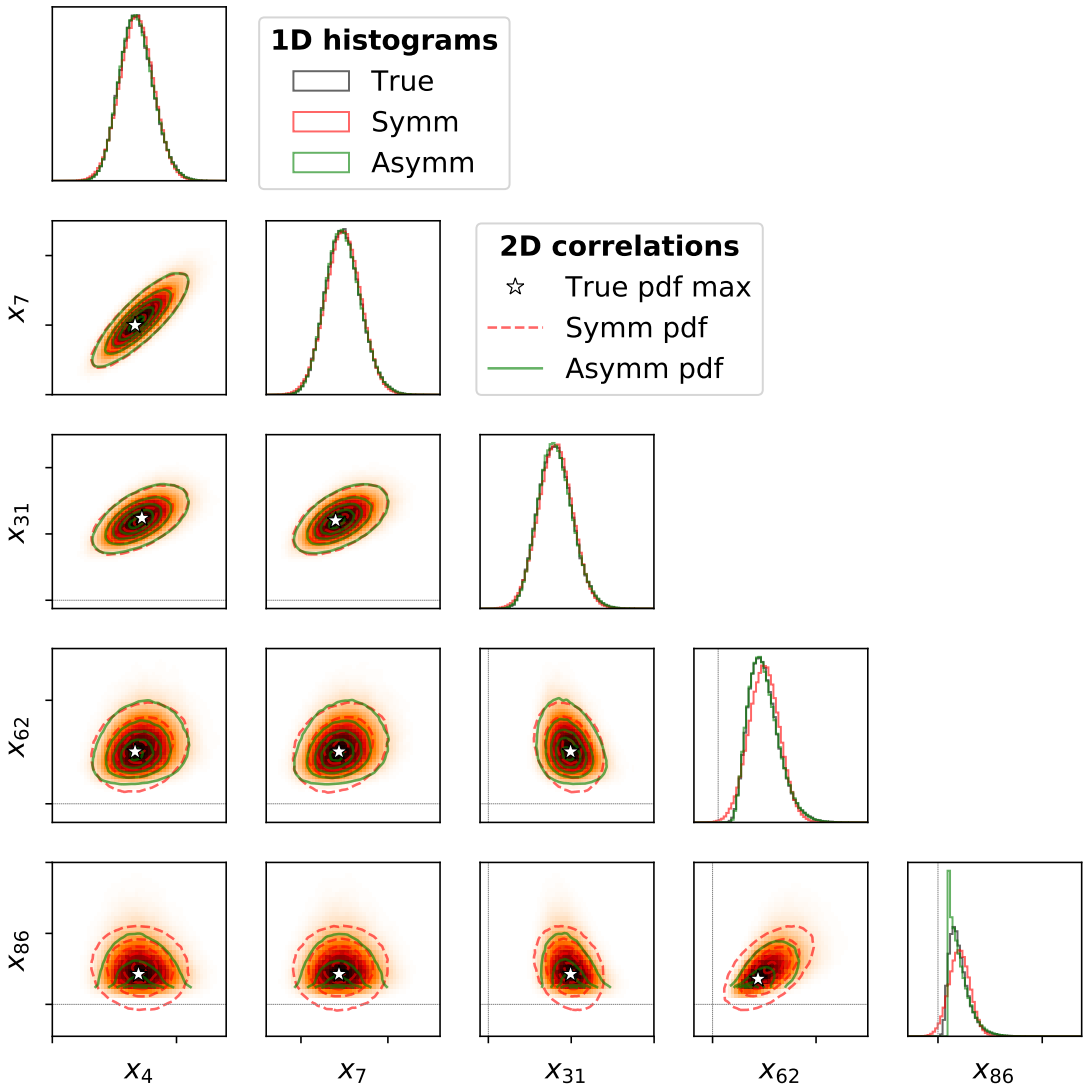

Figure 5. 1D and 2D distributions of $\hat{n}_{b,I}$ against $\hat{n}_J$ for $\{I, J\} = 4, 7, 31, 62, 86$.

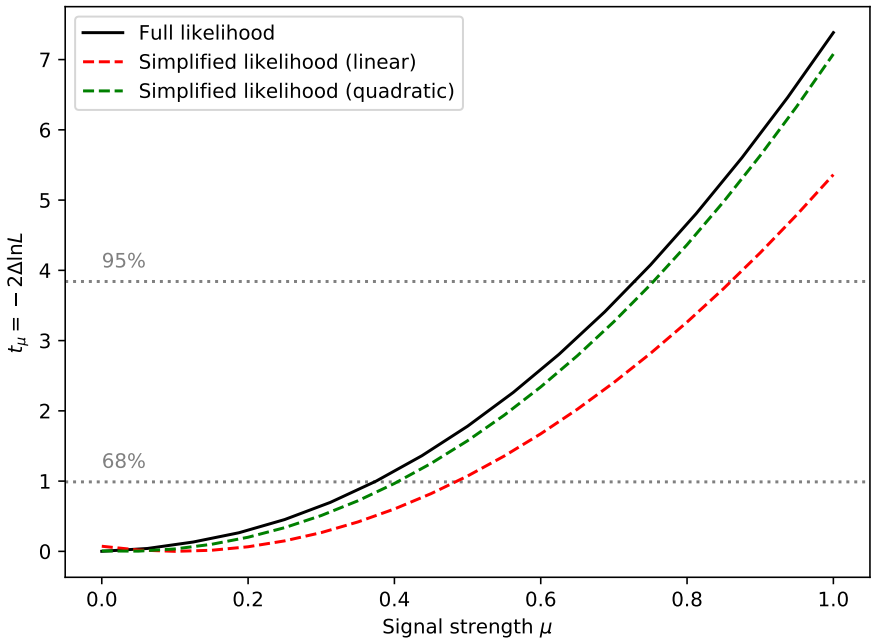

Figure 6. Value of t_μ as a function of μ for the toy search assuming the full likelihood (black solid line) and simplified likelihood retaining (green dashed line) or not (red dashed line) the contribution from the quatratic term. The horizontal lines drawn at $t_\mu = 1$ and 3.86 represent the values for which the 68% and 95% CL exclusions can be determined, assuming certain asymptotic properties of the distribution of t_μ.

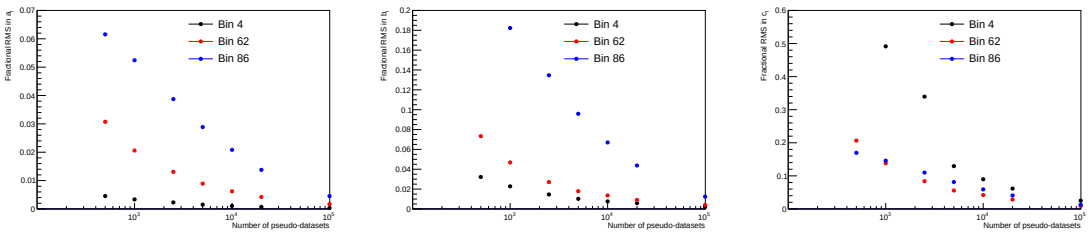

Figure 7. RMS of the simplfied likelihood coefficients relative to the mean coefficient value determined from 100,000 pseudo-datasets for a_I (left), b_I (center), and c_I (right). The distributions are shown for $I = 4$ (black line), $I = 62$ (red line) and $I = 86$ (blue line).