Physics Letters B 778 (2018) 35–42

Contents lists available at ScienceDirect

Physics Letters B

www.elsevier.com/locate/physletb

On new physics searches with multidimensional differential shapes

Felipe Ferreira [a], Sylvain Fichet [b,*], Veronica Sanz [c]

[a] Departamento de Física, Universidade Federal da Paraíba, Caixa Postal 5008, 58051-970, João Pessoa, Paraíba, Brazil
[b] ICTP-SAIFR & IFT-UNESP, R. Dr. Bento Teobaldo Ferraz 271, São Paulo, Brazil
[c] Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK

ARTICLE INFO

Article history:
Received 20 July 2017
Received in revised form 4 January 2018
Accepted 6 January 2018
Available online 9 January 2018
Editor: G.F. Giudice

Keywords:
LHC
Statistical analysis
Bayesian statistics
Effective theory

ABSTRACT

In the context of upcoming new physics searches at the LHC, we investigate the impact of multidimensional differential rates in typical LHC analyses. We discuss the properties of shape information, and argue that multidimensional rates bring limited information in the scope of a discovery, but can have a large impact on model discrimination. We also point out subtleties about systematic uncertainties cancellations and the Cauchy–Schwarz bound on interference terms.

1. Introduction

In modern High Energy Physics, the use of large datasets has become commonplace. In two areas in particular, Particle Physics and Cosmology, the forefront of discoveries and characterisation of new phenomena relies on extraction of information from complex datasets produced by experiments like Planck [1] and the LHC [2]. In both fields, a precise theoretical paradigm is used to interpret the data (ΛCDM and SM, respectively) and the search for new phenomena depends then on identifying subtle deviations within the data, often relying on machine learning techniques. For example, the discovery rare SM processes, like mono-top [3] and Higgs decays to tau-leptons [4], has been achieved using this methodology.

On the theoretical side, these multivariate techniques obscure the physical understanding of which variables drive the analysis, making the re-interpretation of results very difficult and in general hindering the public use of the data. Yet more detailed information, in particular differential rates, is required to advance the programme of searching for a new paradigm beyond the standard one. For example, the use of differential information on Higgs production [5] has proven key to pushing the limits of understanding the impact of possible new phenomena in the Higgs boson properties.

In this paper we investigate the advantages and limitations of multidimensional shape information in searching for new physics and present two case studies, the new physics search in the context of the SM Effective Field Theory (SMEFT) and the characterisation of the quantum numbers of a new resonance. These case studies, together with the material collected in the Appendix, give examples of how differential distributions can be exploited by theorists. Currently, experiments provide mostly one-dimensional distributions, and only rarely two-dimensional information – a notable exception to this trend is provided by the ATLAS analysis [6] of $h \to \gamma\gamma$, which made public the 2D differential distributions in p_T of the Higgs and number of jets. This work is also meant as an incentive for experiments to provide systematically differential distributions in an exploitable form.

2. Statistical basics

In this section we set the notation of the statistical analysis. We denote phase space by $\mathcal{D}$, and consider a binning of $\mathcal{D}$ in d dimensions. The bins are set along a dimension $i \in (1 \dots d)$ and labelled by r_i, with the coordinates $(r_1, \dots, r_d)$ of a bin denoted r, and the associated piece of phase space $\mathcal{D}_r$. The observed event number in the bin r is denoted $\hat{n}_r$, and the expected event number for a given value of the underlying parameter θ is denoted $n_r(\theta)$. Total number of observed events is $\hat{n} = \sum_r \hat{n}_r$ and the expected total number events is $n = \sum_r n_r$.

For further convenience one also introduces the d-dimensional density of expected events $f_X(x)$, where $X = (X_i)$ denotes the set of binned variables. $f_X(x)$ is simply the differential event rate

* Corresponding author.
E-mail addresses: ff83@sussex.ac.uk (F. Ferreira), sylvain@ift.unesp.br (S. Fichet), v.sanz@sussex.ac.uk (V. Sanz).

https://doi.org/10.1016/j.physletb.2018.01.008

normalised by the total event rate. The expected event number in a bin r is then given by

$$n_r = n \int_{\mathcal{D}_r} f_X(x)dx\,. \tag{1}$$

2.1. Likelihood

The likelihood function L is defined as the conditional probability of obtaining the observed data given a hypothesis, taken as a function of this hypothesis. For a hypothesis H with a set of parameters θ,

$$L(\theta) \equiv \Pr(\text{data}|H,\theta)\,. \tag{2}$$

The likelihood function can be defined up to an overall constant factor.

The events counted in each of the bins are statistically independent, hence the likelihood factorises as

$$L = \prod_r L_r\,. \tag{3}$$

The event number in every bin follows a Poisson statistics, so that the likelihood function in the bin r is given by

$$L_r(\theta) = n_r(\theta)^{\hat{n}_r} e^{-n_r(\theta)}\,. \tag{4}$$

For a given integrated luminosity $\mathcal{L}$, $n_r(\theta)$ is given by the event rate on the bin, $n_r(\theta) = \mathcal{L}\sigma_r(\theta)$.

The likelihood can be formally factored in a Poisson term L_{tot} containing the information about the total rate and a term L_{shape} containing the information about the shape of the differential distribution, so that $L(\theta) = L_{\text{tot}}(\theta)L_{\text{shape}}(\theta)$ with

$$L_{\text{tot}}(\theta) = n(\theta)^{\hat{n}} e^{-n(\theta)} \tag{5}$$

$$L_{\text{shape}}(\theta) = \prod_r \left(\frac{n_r(\theta)}{n(\theta)}\right)^{\hat{n}}\,. \tag{6}$$

We stress this feature remains valid beyond Poisson statistics, when systematic uncertainties are also included in the likelihood. Indeed, it is always possible to split the nuisance parameters into as subset affecting only $L_{\text{tot}}(\theta)$ and a subset affecting only $L_{\text{shape}}(\theta)$.

Finally, in the limit where bin size is small enough so that every bin contains only zero or one event, $L_{\text{shape}}(\theta)$ tends to the classical "unbinned" likelihood expressed in terms of the continuous probability density function of the events along the previously-binned variable X,

$$L^{\text{unbinned}}_{\text{shape}}(\theta) = \prod_{I=1}^{\hat{n}} f_X^\theta(x_I)\,, \tag{7}$$

where the x_I are the values of X associated to each of the observed events. The f_X^θ has been defined in Eq. (1).

2.2. Credible regions and hypothesis testing

We adopt the framework of Bayesian statistics.[1] The model parameters are given an a-priori probability density $\pi(\theta)$, called "prior", that can encode both subjective and objective information. The "posterior" density is defined as $p(\theta) \propto L(\theta)\pi(\theta)$, it provides the preferred regions of θ ones data are taken into account. The shape of the prior becomes irrelevant once enough data are accumulated, *i.e.* when the posterior is data-dominated.[2]

A so-called $1-\alpha$ credible region of highest density is defined by the domain $\Omega^{1-\alpha} = \{\theta \,|\, p(\theta) > p_{1-\alpha}\}$, where $p_{1-\alpha}$ is determined by the fraction of integrated posterior

$$\frac{\int_{\Omega^{1-\alpha}} d\theta\, p(\theta)}{\int_\Omega d\theta\, p(\theta)} = 1-\alpha\,, \tag{8}$$

Ω being the whole parameter space. We will use the credible regions associated with $1-\alpha = \{68.27\%, 95.45\%, 99.73\%\}$.[3]

Comparison between two hypotheses H_0 and H_1 is done by means of the Bayes factor

$$B_{01} = \frac{\int_{\Omega_1} L(\theta_1)\pi_1(\theta_1)}{\int_{\Omega_0} L(\theta_0)\pi_0(\theta_0)}\,, \tag{9}$$

where the $\pi_{0,1}$ are the priors for hypotheses $H_{0,1}$ respectively. The Bayes factor is interpreted using the Jeffreys' scale [7], which associates weak, moderate and strong evidence in favour of H_0 to the threshold values $\log B_{01} \sim 1, 2.5, 5$ (*i.e.* $B_{01} \sim 3, 12, 150$).

The Bayes factor framework can be used in the context of new physics searches. In order to assess that the data favour a hypothesis where a parameter θ is different from a given value θ_0 one has to compare the H_1 hypothesis to $H_0 \equiv H_1|\theta = \theta_0$ (see also [7], [8]). In the context of effective operators, H_1 can for instance be the SM deformed by higher dimensional operators (the SMEFT), while H_0 is the SM. Defining $B_0 \equiv 1/B_{01}$, we have

$$B_0 = \frac{\int_\Omega L(\theta)\pi(\theta)}{L(\theta_0)}\,, \tag{10}$$

that we refer to as the *discovery Bayes factor*. The test assesses that $\theta \neq \theta_0$ for $B_0 > 1$, using the thresholds given above.

2.3. Asimov (projected) data

In order to evaluate the sensitivity of a future analysis, measurement, or experiment, one can rely on imaginary, speculative data. That is, instead of introducing actual observed data in the likelihood Eq. (2), one can instead introduce speculative data coming for instance from a simulation of the experiment. We refer to these as *projected data*.

An important subtlety, well discussed in [9], is that an assumption has to be made on the statistical fluctuations present in the projected data. Along this paper, we will simply consider the case where no statistical fluctuations are present in the projected data. A dataset satisfying this condition is sometimes referred to as an "Asimov" dataset [9].

The event numbers in the projected dataset assuming no statistical fluctuations and the presence of an operator with coefficient c' are then simply given by $\mathscr{L}\sigma_r(c')$. In practice, these rates have to be estimated by MonteCarlo simulations, just like the expected ones.

[1] The Bayesian framework is consistent with the "likelihood principle", which states that all experimental information is encoded in the likelihood function. This is not the case of, for example, frequentist p-values.

[2] The prior has however to satisfy basic physical conditions, such as keeping an event number positive in order to avoid singularities in the posterior. As a general rule, the posterior should be data-dominated in the limit of many data, otherwise the inference process cannot happen.

[3] Note that confidence regions are not uniquely defined, but the method of highest density is the most commonly used and arguably the most natural.