An *Imperfect* machine to search for New Physics: systematic uncertainties in a machine-learning based signal extraction

Gaia Grosso
Dipartimento di Fisica e Astronomia, Universitá di Padova, Italy
CERN, Experimental Physics Department, Geneva, Switzerland
gaia.grosso@cern.ch

Andrea Wulzer
Dipartimento di Fisica e Astronomia, Universitá di Padova, Italy
Institut de Théorie des Phénomenes Physiques, EPFL, Lausanne, Switzerland
andrea.wulzer@cern.ch

Maurizio Pierini
CERN, Experimental Physics Department, Geneva, Switzerland
maurizio.pierini@cern.ch

Marco Zanetti
Dipartimento di Fisica e Astronomia, Universitá di Padova, Italy
marco.zanetti@cern.ch

Raffaele Tito d'Agnolo
Institut de Physique Théorique, Université Paris Saclay, CEA, France
raffaele-tito.dagnolo@ipht.fr

Abstract

We show how to deal with uncertainties on the Standard Model predictions in an agnos- tic new physics search strategy exploiting artificial neural networks. Our approach builds directly on the Maximum Likelihood ratio treatment of uncertainties as nuisance parameters for hypothesis testing that is routinely employed in high-energy physics. After presenting the conceptual foundations of our method, we first illustrate all aspects of its implementation and extensively study its performances on a toy one-dimensional problem. We then show how to implement it in a multivariate setup by studying the impact of two typical sources of experimental uncertainties in two-body final states at the LHC.

1 Introduction

Experimental results in the last several decades consolidated our knowledge of fundamental physics as described by "standard" theoretical models such as the Standard Model (SM) of particle physics or the ΛCDM model of cosmology. On the other hand we lack understanding of the microscopic origin of several ingredients of these models, such as the Dark Matter and Dark Energy densities in ΛCDM , the Electroweak scale and the Yukawa couplings structure in the SM. These considerations, as well as the theoretical incompleteness of our current theory of gravity, guarantee the existence of

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

new fundamental laws waiting to be discovered, but do not sharply outline a path towards their actual experimental discovery. The development of signal-model-independent strategies to search for new physics emerges in this context as a priority of fundamental physics.

In the present work we consider the model-independent method proposed and developed in Ref.s [1, 2] for multivariate data analysis at particle colliders such as the Large Hadron Collider (LHC). The method aims at testing some data sample $\mathcal{D} = \{x_1, \ldots, x_{\mathcal{N}_\mathcal{D}}\}$ for the presence of significant departures with respect to a Reference Model "R". The physical knowledge of the Reference Model (the SM) can be used to produce a synthetic set of Reference data $\mathcal{R} = \{x_1, \ldots, x_{\mathrm{N}_\mathcal{R}}\}$, which plays conceptually the same role as the background dataset in many traditional searches. In general it could be either obtained by a first-principle Monte Carlo (MC) simulation based on the fundamental physical laws of the Reference Model, or with data-driven methods. In both cases, $\mathcal{R}$ results from a knowledge of the Reference Model that is unavoidably *imperfect*. Therefore it provides only an approximate representation of the data distribution in the "R' (or background) hypothesis. In the case of MC samples, for instance, the uncertainties emerge from all the ingredients of the simulations such as the value of the Reference Model input parameters, of the parton distribution functions and of the detector calibration parameters, as well as from the finite accuracy of the underlying theoretical calculations. The impact of all these uncertainties must be assessed and included if needed in any LHC analysis. We define here a strategy to deal with them in a framework for signal-model-independent new physics searches. We ignore instead uncertainties of statistical origin associated with the finite size of the Reference sample. Namely we assume that the Reference sample size $\mathrm{N}_\mathcal{R}$ is a factor from ten to hundred times bigger than the expected number of data events $\mathrm{N(R)}$, such that the data statistical fluctuations are expected (and can be checked) to dominate over the one of the Reference sample.

2 Method

Our treatment follows closely the canonical HEP profile likelihood approach [3]. Each source of imperfection in the knowledge of the Reference Model is associated with a nuisance parameter ν whose (true) value is unknown but statistically constrained by some measurements performed on an auxiliary dataset $\mathcal{A}$ and included in the likelihood as a ν-dependent term $\mathcal{L}(\nu|\mathcal{A})$. The Reference Model prediction for the distribution of the features variable x depends on the set of nuisance parameters ν and is therefore interpreted as a composite (parameter-dependent) statistical hypothesis R_ν, to be identified with the null hypothesis H_0 of the statistical test. The alternative hypothesis H_1 is defined as a local (in the features space) rescaling of the Reference distribution by the exponential of a neural network function $f(x;\mathbf{w})$

$$n(x|\mathrm{H}_{\mathbf{w},\nu}) = e^{f(x;\mathbf{w})} n(x|\mathrm{R}_\nu)\,.$$

The exponential parametrization guarantees $n(x|\mathrm{H}_{\mathbf{w},\nu})$ to be positive and allow to use $f(x;\mathbf{w})$ directly as an approximant of the log ratio of the distributions. Clearly, the H_1 hypothesis is also composite. We denote it as $\mathrm{H}_{\mathbf{w},\nu}$, where $\mathbf{w}$ represents the trainable parameters of the neural network. The neural network architecture and hyper-parameters are problem-dependent. The general criteria for their optimization are discussed in Ref [2] in greater detail. The method exploits the neural network flexibility to approximate the optimal test statistic t based on the Maximum Likelihood log-ratio test statistic between the R_ν and $\mathrm{H}_{\mathbf{w},\nu}$ hypotheses:

$$t(\mathcal{D},\mathcal{A}) = 2\,\log \frac{\max\limits_{\mathbf{w},\nu}\left[\mathcal{L}(\mathrm{H}_{\mathbf{w},\nu}|\mathcal{D},\mathcal{A})\right]}{\max\limits_{\nu}\left[\mathcal{L}(\mathrm{R}_\nu|\mathcal{D},\mathcal{A})\right]}\,.$$

In order to proceed, we consider the special point in the nuisance parameter space that corresponds to their central-value determination as obtained from the auxiliary data alone and we set it as the origin. We further postulate that new physics is absent in the auxiliary data. Namely, that the distribution of the auxiliary data in the $\mathrm{H}_{\mathbf{w},\nu}$ hypothesis is the same one as in hypothesis R_ν

$$\mathcal{L}(\mathrm{H}_{\mathbf{w},\nu}|\mathcal{A}) = \mathcal{L}(\mathrm{R}_\nu|\mathcal{A}) = \mathcal{L}(\nu|\mathcal{A})\,.$$

Therefore the total likelihood of $\mathrm{H}_{\mathbf{w},\nu}$ is

$$\mathcal{L}(\mathrm{H}_{\mathbf{w},\nu}|\mathcal{D},\mathcal{A}) = \mathcal{L}(\mathrm{H}_{\mathbf{w},\nu}|\mathcal{D}) \cdot \mathcal{L}(\nu|\mathcal{A})\,,$$

where $\mathcal{L}(\mathrm{H}_{\mathbf{w},\nu}|\mathcal{D})$ is the extended likelihood

$$\mathcal{L}(\mathrm{H}_{\mathbf{w},\nu}|\mathcal{D}) = \frac{e^{-\mathrm{N}(\mathrm{H}_{\mathbf{w},\nu})}}{\mathcal{N}_{\mathcal{D}}!} \prod_{x_i \in \mathcal{D}} n(x_i|\mathrm{H}_{\mathbf{w},\nu}) \,.$$

The central-value Reference hypothesis R_0 predicts a distribution for the variable $n(x|\mathrm{R}_0)$, that can be regarded as the best pre-fit guess for the actual SM distribution of x. The likelihood of R_0 is thus conveniently used to "normalize" the numerator and denominator likelihoods in our test statistic formulation and thus express the latter as the difference of two terms

$$t(\mathcal{D},\mathcal{A}) = \tau(\mathcal{D},\mathcal{A}) - \Delta(\mathcal{D},\mathcal{A}) \,,$$

where τ involves the maximization over the neural network parameters $\mathbf{w}$ and over ν

$$\tau(\mathcal{D},\mathcal{A}) = 2 \max_{\mathbf{w},\nu} \log \left[\frac{\mathcal{L}(\mathrm{H}_{\mathbf{w},\nu}|\mathcal{D})}{\mathcal{L}(\mathrm{R}_0|\mathcal{D})} \cdot \frac{\mathcal{L}(\nu|\mathcal{A})}{\mathcal{L}(0|\mathcal{A})} \right] ,$$

while the "correction" term Δ does not contain the neural network and involves exclusively the Reference hypothesis

$$\Delta(\mathcal{D},\mathcal{A}) = 2 \max_{\nu} \log \left[\frac{\mathcal{L}(\mathrm{R}_\nu|\mathcal{D})}{\mathcal{L}(\mathrm{R}_0|\mathcal{D})} \cdot \frac{\mathcal{L}(\nu|\mathcal{A})}{\mathcal{L}(0|\mathcal{A})} \right] .$$

The τ term can be seen as a measure of the total discrepancy between the distribution of the data and the one expected in the R_0 hypothesis; the Δ term is instead a measure of the only discrepancies that can be shaped as systematic effects. Both τ and Δ are positive-definite and since they contribute with opposite sign, the test statistic t will emerge from a cancellation between these two terms. t is then a measure of the discrepancies that are relevant in the search for New Physics. The cancellation is more and more severe the more the data happen to favour a value of ν that is far from the central value.

Learning the effect of nuisance parameters The correction term Δ is of interest for any statistical analysis to be performed on the dataset $\mathcal{D}$, as it provides a first gross indication of the data compatibility with the Reference hypothesis. In particular a sizeable departure of the best-fit nuisance parameters from the central values should be monitored the same way a large "pull" of a nuisance parameter is monitored in a ML fit.

To evaluate Δ, we employ an un-binned $\mathcal{L}(\mathrm{R}_\nu|\mathcal{D})$ likelihood, obtained by reconstructing the ratio between the $n(x|\mathrm{R}_\nu)$ and $n(x|\mathrm{R}_0)$ distributions locally in the features space. This is achieved by a rather straightforward adaptation of "likelihood-free inference" techniques developed in the literature. In particular our implementation follows closely the "Quadratic Classifier" approach of Ref [4] to which we refer the reader for a more in-depth exposition. The basic idea, is to employ a polynomial approximation for the dependence of the distribution on the nuisance. The coefficient functions of the polynomial will be approximated by suitably trained neural networks. For instance in the case of a single nuisance parameter ν, we would write

$$r(x;\nu) \equiv \frac{n(x|\mathrm{R}_\nu)}{n(x|\mathrm{R}_0)} = \exp\left[\nu\, \delta_1(x) + \frac{1}{2}\nu^2\, \delta_2(x) + \ldots \right] ,$$

with the Taylor series expansion in the exponent truncated at some finite order. The exponential parametrization here adopted guarantees that the ratio is positive and thus a well defined argument for the logarithmic function which is applied to it when the test statistic is computed.

Maximum likelihood from minimal loss The τ term involves the $\mathrm{H}_{\mathbf{w},\nu}$ hypothesis, which foresees possible non-SM effects (i.e., departures from the Reference Model) in the distribution of x that can be parametrized by the neural network $f(x;\mathbf{w})$. The calculation of τ involves the maximization over the neural network parameters $\mathbf{w}$ and the nuisance parameters ν, that will be performed by running a learning algorithm considering both $\mathbf{w}$ and ν as trainable parameters. The algorithm will exploit the knowledge of the δ coefficient functions that is provided by neural networks as previously explained. Since these networks describe our knowledge of the systematic effects on the input variables, they don't have to be fitted to the data; they are thus pre-trained and treated as constant during the evaluation of τ.

By making use of the parametrization of $n(x|\mathrm{H}_{\mathbf{w},\nu})$ in terms of the neural network and by combining it with the definition of $r(x;\nu)$ previously introduced we can rewrite τ in the form

$$\tau(\mathcal{D},\mathcal{A}) = 2 \max_{\mathbf{w},\nu} \left\{ \sum_{x_i \in \mathcal{D}} \left[f(x_i;\mathbf{w}) + \log(r(x_i;\nu)) \right] - \mathrm{N}(\mathrm{H}_{\mathbf{w},\nu}) + \mathrm{N}(\mathrm{R}_0) + \log \left[\frac{\mathcal{L}(\nu|\mathcal{A})}{\mathcal{L}(0|\mathcal{A})} \right] \right\} .$$

The first, third and fourth terms in the curly brackets are easily accessible. The first one depends on the neural network $f(x_i;\mathbf{w})$, as well as on the coefficient functions δ through $r(x_i;\nu)$. The second term, which is the total number of events in the $\mathrm{H}_{\mathbf{w},\nu}$ hypothesis, is not easily available and requires us to employ the Reference data set $\mathcal{R} = \{x_1, \ldots, x_{\mathrm{N}_\mathcal{R}}\}$, to approximate it as

$$\mathrm{N}(\mathrm{H}_{\mathbf{w},\nu}) \simeq \sum_{\mathrm{e} \in \mathcal{R}} w_\mathrm{e} \exp\left[f(x_\mathrm{e};\mathbf{w}) + \log(r(x_\mathrm{e};\nu)) \right] .$$

Here each event is weighted in such a way that the normalization sums up to $\mathrm{N}(\mathrm{R}_0)$. Since the accuracy of the approximation improves with the size of the Reference sample, we require $\mathrm{N}_\mathcal{R} \gg \mathrm{N}(\mathrm{R}_0) \sim \mathcal{N}_\mathcal{D}$; in this way the statistical variability of τ is expectedly dominated by the statistical fluctuation of the data sample $\mathcal{D}$. We can then express τ as

$$\tau(\mathcal{D},\mathcal{A}) = -2 \min_{\mathbf{w},\nu} \left\{ L \left[f(\cdot;\mathbf{w}),\, \nu;\, \widehat{\delta}(\cdot) \right] \right\} ,$$

where L has the form of a loss function for a supervised training between the $\mathcal{D}$ and $\mathcal{R}$ samples

$$L \left[f(\cdot;\mathbf{w}),\, \nu;\, \widehat{\delta}(\cdot) \right] = - \sum_{x_i \in \mathcal{D}} \left[f(x_i;\mathbf{w}) + \log(r(x_i;\nu)) \right] + \sum_{\mathrm{e} \in \mathcal{R}} w_\mathrm{e} \left[e^{f(x_\mathrm{e};\mathbf{w}) + \log(r(x_\mathrm{e};\nu))} - 1 \right] - \log \left[\frac{\mathcal{L}(\nu|\mathcal{A})}{\mathcal{L}(0|\mathcal{A})} \right] .$$

The training procedure is relatively straightforward to implement in standard deep learning packages, provided the loss depends on ν through analytically differentiable functions. This is the case for $r(x;\nu)$, and typically also for the auxiliary likelihood ratio.

Our strategy to evaluate τ is an extension of the one developed in Ref.s [1,2]. In the absence of nuisance parameters, the loss L reduces to the one of Ref.s[1,2], plus the auxiliary log likelihood ratio that carries all the dependence on ν and can be minimized independently. The latter term however cancels in the test statistic t when subtracting the correction term Δ and the results of Ref.s[1,2] are fully recovered.

3 A multivariate case study: two-body final state at the LHC

We test the previously outlined strategy on a multivariate case study which is inspired by the realistic problem of model-independent new physics searches in two-body final states at the LHC (a detailed description of the simulated datasets can be found in Ref [2]). A two-body final state can be characterized in terms of the five kinematical variables, the transverse momenta and the pseudorapidities of the individual particles and their relative azimuthal angle, that we choose as input features for our method. We separately analyze three different scenarios given either by opposite-signed muons, electrons or taus in the final state. The kinematical distributions are quite different in the three cases, but we do not expect these differences to impact the technical viability of our strategy, which we aim at demonstrating. The total cross-section of the process is also different. We compensate for this effect by selecting the integrated luminosity of the dataset that makes the total number of expected events $\mathrm{N}(\mathrm{R}_0)$ equal to 8500 for all the cases. In this way, the only relevant difference between muons, electrons and taus final states resides in the increasing large systematic uncertainties that affect the corresponding SM predictions.
We consider two Gaussian nuisance parameters ν_N and ν_S describing the uncertainty on the event yield normalization and on the transverse momenta scale as exponential factors. We adopt a simple modeling of the normalization uncertainties by a global factor with standard deviation $\sigma_\mathrm{N} = 2.5\%$. As for the scale factor, we consider three representative scenarios: a 5×10^{-4} and 15×10^{-4} uncertainty errors respectively for $|\eta| < 2.1$ and $|\eta| \geq 2.1$ in the muon-like regime; 3×10^{-3} and 9×10^{-3} for the electron-like; 3×10^{-2} at any η in the tau-like regime. The effects in the two regions of η are fully correlated. A lower threshold to the two leptons invariant mass is applied to discard the Z pole from the dataset and thus keeping the scale effects manageable. This lower threshold is set to 100 GeV for the muon and electron-like regimes and to 120 GeV for the tau-like regime.

For the neural network model involved in the computation of τ, we consider a fully connected feed-forward neural network with sigmoid activation functions for the hidden layers and a single linear output. We define a weight clipping parameter which sets up an upper bound constraint to all the trainable parameters magnitude. In Ref [2] it was showed that the weight clipping parameter can be tuned to make the distribution of the test statistic under null hypothesis compatible with an asymptotic χ^2_{df} distribution whose number of degrees of freedom equals the number of trainable parameters of the chosen neural network. This consideration still applies to the extended version of the algorithm here proposed. According to this prescription, we select a neural network with three layers and five nodes each and a weight clipping value of 2.16 for the muon and electron like regime and 2.38 for the tau-like regime. Moreover, we model the effect of $\nu_{\rm N}$ and $\nu_{\rm S}$ in $r(x;\nu_{\rm N},\nu_{\rm S})$ using an analytical description for the normalization effect and the un-binned likelihood approach mentioned in the previous section for the momentum scale effect. For the latter, a quadratic polynomial approximation is used in all three systematic uncertainties regimes. For the parametric models set up, there are no further prescriptions than those already mentioned in Ref [4]. For all δ coefficients we consider architectures of five layers of ten nodes each and relu activation function. The quality of the approximation is verified by validating the method. This also allows for an optimal choice of the training epochs. The

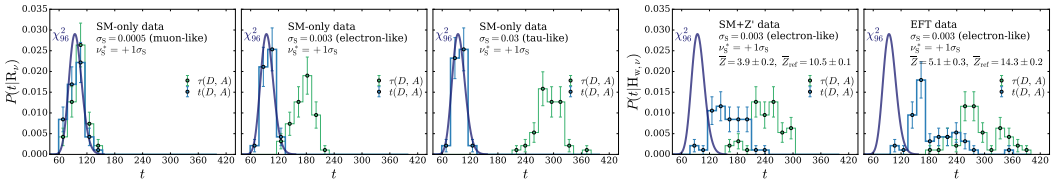

Figure 1: Five examples of the empirical test statistic distribution for 100 experiments, before (τ) and after (t) applying the Δ correction, under different true underlying hypothesis of the data. The blue line denotes the target χ^2 distribution for the test statistic in the reference hypothesis; the number of degrees of freedom, 96, coincides with the number of trainable parameters of the neural network model used for the experiments.

validation of the strategy consists in running experiments using SM-like synthetic datasets generated with systematic effects $\nu^*_{\rm S,N}=\pm\sigma_{\rm S,\,N}$. We observe that any nuisance-induced source of discrepancy detected by the τ term is correctly canceled by the corresponding Δ term, resulting in a final test statistic t which is compatible with the asymptotic χ^2_{df} (three panels on the left side of Figure 1).

Furthermore, we apply the strategy on two BSM benchmark scenarios. In the first one a new vector boson with the same couplings to SM fermions as the SM Z boson and mass of 300 GeV is injected on top of the SM background (Z' scenario); in the second one we insert a dimension-6 4-fermion contact interaction $\frac{c_W}{\Lambda}J^a_{L\mu}J^\mu_{La}$ (EFT scenario). We run experiments on data following the BSM models and with generated systematic effects $\nu^*_{\rm S,N}=\pm\sigma_{\rm S,N}$. A residual discrepancy surviving the cancelation hints at the presence of New Physics (two panels on the right side of Figure 1). The median Z-score of the method can be compared to that of a typical model-dependent analysis ($\overline{Z}_{\rm ref}$); we observe similar gains in both the resonant and non-resonant scenarios.

4 Conclusion and outlook

In this work we present a new model independent strategy to compute a p-value for the discovery of unspecified New Physics models in presence of systematic uncertainties on the reference hypothesis. In the case of resonant signals, similar results can be achieved with alternative techniques, like the BumpHunter [5], provided that the variable manifesting the resonance is known. On the other hand, the proposed approach is able to detect resonances even when the variable of interest is not considered as an input of the strategy, learning non trivial combinations out of low level features. Moreover, it is sensitive to various New Physics signatures at the same time and with similar performances; general approaches that can handle multivariate problems and are sensitive to non-resonant signatures are still missing in the landscape of the model independent analysis strategies. In this respect our solution is a novel and promising tool for discovery new phenomena at the LHC experiments.

Acknowledgments

M.P. and G.G. are supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n^o 772369). A.W. acknowledges support from the Swiss National Science Foundation under contract 200021-178999.

References

[1] D'Agnolo RT, Wulzer A. *Learning new physics from a machine.* Physical Review D. 2019 Jan 8;99(1):015014.

[2] D'Agnolo RT, Grosso G, Pierini M, Wulzer A, Zanetti M. *Learning multivariate new physics.* The European Physical Journal C. 2021 Jan;81(1):1-21.

[3] ATLAS Collaboration, CMS collaboration, *Procedure for the LHC Higgs boson search combination in Summer 2011* ATL-PHYS-PUB-2011-011, CERN-CMS-NOTE-2011-005, 2011

[4] Chen S, Glioti A, Panico G, Wulzer A. *Parametrized classifiers for optimal EFT sensitivity.* Journal of High Energy Physics. 2021 May;2021(5):1-39.

[5] Choudalakis, G. (2011).*On hypothesis testing, trials factor, hypertests and the BumpHunter.* arXiv preprint arXiv:1101.0390.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] the assumptions on which the method relies are specified throughout Section 1 and Section 2
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] A reference is provided where the details of the datasets are given; the main instructions needed to reproduce the experimental results can be found in Section 3.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] the distributions reported in Figure 1 represent the statistical behaviour of the method on experiments run with different random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] considerations on this matter will be available in an extended version of this work
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]