How good is the Standard Model? Machine learning multivariate Goodness of Fit tests

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

Marco Letizia
MaLGa Center - DIBRIS
Università di Genova
INFN, Sez. di Genova
Genova, Italy
marco.letizia@edu.unige.it

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

Andrea Wulzer
Dipartimento di Fisica e Astronomia
Università di Padova
Padova, Italy
andrea.wulzer@cern.ch

Abstract

We formulate the problem of detecting collective anomalies in collider experiments as a Goodness of Fit test of a reference hypothesis (the Standard Model) to the observed data. Several well established Goodness of Fit methods are available for one-dimensional problems but their multivariate generalization is still object of study. We exploit machine learning to build a set of multivariate tests, starting from the outcome of a machine learned binary classifier trained to distinguish the experimental data from the reference expectations, as prescribed in Ref.s [1, 2, 3, 4]. We compare typical one-dimensional test statistics computed on the output of the classifier with less common test statistics built out of standard classification metrics. In the considered setup, the likelihood-ratio test shows a broader model-independent sensitivity to the landscape of the signal benchmarks analysed. A novel test we define, based on event counting with an optimised classifier threshold, is found to perform slightly better than the likelihood-ratio test for resonant signal, but is exposed to strong failures for non-resonant ones.

1 Introduction

As of today, the Standard Model embodies the best description of the current understanding of the world of particle physics. Nevertheless, it does not depict the whole landscape of the physics phenomena governing our Universe; it does not provide an explanation for cosmological observations, like dark matter and dark energy, and it does not include a quantum description for gravity. The high energy physics community is therefore searching for hints of phenomena going beyond the Standard Model, trying to locate where the latter would eventually fail in describing the data. The typical approach to new physics searches is to perform a hypothesis test of the data, taking the Standard Model to be the null hypothesis and a specific theoretical extension of the Standard Model as alternative. This types of test are well understood and routinely applied at collider experiments. However they can only test a specific failure of the Standard Model at the time and therefore only partially answer to the problem of the goodness of the Standard Model fit to the data. A higher degree

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

of inclusiveness can be reached by relaxing the assumptions on the theory behind the alternative model, up to the point of directly comparing the Standard Model expected data distribution with the one experimentally observed. By releasing the dependency on the alternative hypothesis the test becomes sensitive to any possible failure of the Standard Model at once and it moves to the realm of goodness of fit (GoF) tests.
GoF tests can be solved either as a one-sample or a two-sample problem. In the first case, a set of measurements $\{x_i\}_{i=1}^{N}$ is compared to a known target probability distribution $p(x)$. In the second case, the target probability distribution is not analytically known but approximated by the empirical distribution of a second sample $\{z_i\}_{i=1}^{M}$, to be compared to the original one. The GoF problem has no optimal solution, meaning that the power of a test against different discrepancies depends on the model and the strategy adopted to perform the fit. Nevertheless, GoF tests are useful tools in data analysis; recently, for instance, they have been combined with machine learning methods to tackle model-independent searches of statistical anomalies in high energy physics [1, 2, 4, 3, 5, 6]. Machine learning algorithms can indeed be used to push the boundaries of GoF for multivariate problems [7, 8, 9, 10]. GoF can be performed training a binary classifier to compress the discriminant features of a multi-dimensional problem into one dimension (the output of the classifier) over which the standard univariate GoF tests can be computed [7]. But machine learning can be exploited even further and new GoF tests can be defined on top of the classifier quality metrics, such as the AUC and the binary accuracy [4, 9]. More recently [1, 2, 3], it has been proposed to approach the GoF problem as a likelihood-ratio test, where the set of alternative distributions is obtained by extending the Reference prediction with flexible parametric or non-parametric models such as neural networks or kernel methods. This novel approach is a specific incarnation of the classifier-based method, which in particular foresees employing as test statistic the likelihood ratio as evaluated by the trained classifier output. In this work we explore various approaches to perform multivariate GoF tests as a solution for model-independent new physics searches at collider experiments. We start from a binary classifier trained to maximize the lilkelihood-ratio between the true hypothesis of some experimental data (class 1) and the expected normal behavior of the data depicted by a reference sample (class 0). We observe that applying a classifier not only allows to reduce the dimensionality of multivariate problems to a univariate one but it also allows to construct more sensitive tests. In particular, we will show that the likelihood-ratio test is sensitive to the broad landscape of signal benchmarks analysed in this work. Furthermore, we design a novel test, based on event counting with an optimised classifier threshold, that is found to perform slightly better than the likelihood-ratio test for resonant signal, but is exposed to strong failures for non-resonant ones.

2 Machine learning Goodness of Fit tests of the Standard Model

For the classifier training procedure we adopt two different prescriptions, based respectively on neural networks (NN) [1, 2] and kernel methods (KM) [3]. We assume that a set of experimental data $\mathcal{D} = \{x_i\}_{i=1}^{\mathcal{N}_\mathcal{D}}$ has been collected and the aim is testing its agreement with a reference (null) hypothesis. The problem is set up as a log-likelihood-ratio test of the data comparing the null hypothesis (according to which the data follow the reference distribution) and an alternative, which is not specified a priori but machine learnt from the data themselves. As it is typically the case in high energy physics experiments, we assume the total number of collected events to be poissonian distributed around a number of expected events (N(R) for the reference, N(D) for the alternative) so that the correct likelihood of the data is the extended likelihood. Moreover, while the alternative hypothesis is not known and should be retrieved from the limited statistics experimental data, the reference hypothesis (which in the context of high energy physics experiments is the Standard Model) is assumed to be well known and accurately reproducible thanks to a large amount of simulated training events, $\mathcal{R} = \{x_i\}_{i=1}^{\mathcal{N}_\mathcal{R}}$ with $\mathcal{N}_\mathcal{R} \gg \mathcal{N}_\mathcal{D}$.

In what follows we define the set of GoF tests computed on the output of the trained classifier.

LRT: is the extended likelihood ratio test computed according to [1]

$$t_{\mathbf{LRT}}(\mathcal{D}) = -2\left[\frac{\mathrm{N(R)}}{\mathcal{N}_\mathcal{R}}\sum_{x\in\mathcal{R}}(e^{f(x;\widehat{\mathrm{w}})}-1) - \sum_{x\in\mathcal{D}} f(x;\widehat{\mathrm{w}})\right].$$

For both the NN and the KM implementations, $f(x; \widehat{\mathrm{w}})$ is the (linear) functional output of the model with $\widehat{\mathrm{w}}$ the model parameters at the end of training.

Pois: is inspired to the figure of merit used in particle physics to quantify the statistical significance[1] of an excess on a single bin analysis. The test is computed on the right tail of the classifier output with sigmoid activation ($\bar{f}(x; \widehat{\mathrm{w}})$) and maximised over the lower boundary of the bin (thr). It quantifies the poissonian probability of observing $n_{\mathcal{D}}(\bar{f} > \mathrm{thr})$ data events above the threshold given the expected number of events under the Reference hypothesis $n_{\mathcal{R}}(\bar{f} > \mathrm{thr})$

$$t_{\mathbf{Pois}}(\mathcal{D}) = \max_{\mathrm{thr} \in [0,\,1]} \left[P_{\mathrm{pois}}(n_{\mathcal{D}}(\bar{f}(x; \widehat{\mathrm{w}}) > \mathrm{thr});\, \lambda = n_{\mathcal{R}}(\bar{f}(x; \widehat{\mathrm{w}}) > \mathrm{thr})) \right].$$

ACC: is the accuracy of the classifier maximised over the classification threshold (thr)

$$t_{\mathrm{ACC}}(\mathcal{D}) = \max_{\mathrm{thr} \in [0,\,1]} \left[\frac{n_{\mathcal{D}}(\bar{f}(x; \widehat{\mathrm{w}}) > \mathrm{thr}) + n_{\mathcal{R}}(\bar{f}(x; \widehat{\mathrm{w}}) < \mathrm{thr})}{\mathcal{N}_{\mathcal{D}} + \mathrm{N(R)}} \right].$$

AUC ext: is a variation of the Area Under the Curve of the Receiver Operating Characteristic (ROC), accounting for the normalization of the distributions (by analogy with the extended likelihood). In the construction of the ROC curve, rates are replaced by the absolute number of true positives and false positives (normalized to the expected background $\mathrm{N(R)}$). The resulting AUC does not sum to unity and it is sensitive to the number of events.

KS out: is the conventional Kolmogorov-Smirnov test computed on the classifier output $\bar{f}(x; \widehat{\mathrm{w}})$.

$\mathbf{N(S)/\sqrt{N(R)}}$**:** is the significance due to normalization only. $\mathrm{N(S)}$ is the difference between the total number of events in the dataset $\mathcal{D}$ and the expectation under the Reference hypothesis $\mathrm{N(R)}$.

For the one-dimensional experiment considered in the next section, we additionally compute two standard one-dimensional GoF tests: the Kolmogorov-Smirnov test and the χ^2 test. For the latter, we assume the reference hypothesis to be analytically known so that the expected number of events in each bin is error free. We divide the distribution in equally populated bins and in order to study how the power of the test depends on the choice of binning, we consider n_i =5, 10, 100 events per bin.

Following Refs. [1, 2, 3], every time a new set of data $\mathcal{D}$ is analysed, the models are retrained from scratch and every test statistic is computed in-sample. The empirical distribution of each test is obtained running around 400 experiments for the null hypothesis and 100 for the alternative.
The kernel-based models have been trained sequentially on a single machine with a NVIDIA Quadro RTX 6000 with 24 GB of VRAM while the neural network models have been trained in parallel on a CPU cluster. This choice of training resources is dictated by the average time needed to train a single classifier in the two implementations, which is $\mathcal{O}(\mathrm{hours})$ for the NN and $\mathcal{O}(\mathrm{seconds})$ for the KM.

The results are reported in terms of power curves: for different confidence levels α we scatter the fraction of true positives $P(Z > Z_\alpha)$ as a function of the score Z_α.

3 Results

One-dimensional toy model. As a first example, we consider a univariate problem with a reference hypothesis characterized by an exponentially falling distribution with unit mean. The experimental luminosity is such that we expect 2000 background events on average. We consider four signal benchmarks mimicking qualitatively different new physics effects:

- H_1: a peak in the tail of the exponential distribution, modelled as a gaussian distribution with mean 6.4 and standard deviation 0.16.
- H_2: a quadratically growing excess in the tail ($p(x) \propto x^2 e^{-x}$). For this signal we consider the case of shape only effect, keeping the normalization the same as in the reference hypothesis, and the case of both shape and normalization effects.

[1] We adopt the standard definition $Z = \Phi^{-1}(1-p)$, where p is the p-value and Φ^{-1} is the quantile of the Gaussian distribution.

H_3: a peak in the bulk of the exponential distribution, modelled as a gaussian distribution with mean 1.6 and standard deviation 0.16.

H_4: a defect in the tail of the distribution obtained simply applying an upper boundary to the data at $x_4 = 5.07$.

For the NN model we use a 1-4-1 architecture with sigmoid activation function in the hidden layer and a linear output. We apply a weight clipping of 9 and we run each training on 300k epochs using ADAM optimizer. The model is implemented in TENSORFLOW [11]. The KM approach is a kernel logistic regression model with a L2 regularizarion ($\lambda = 10^{-10}$), a gaussian kernel ($\sigma = 2.3$) and column subsampling ($m = 5000$ Nyström centers). The model is built using the Falkon library [12]. More detailed descriptions of model selection and training schemes can be found in Ref.s [3, 13]. As a reference training sample we generate 200k events (100 times larger than the expected background).

Figure 1 shows the performances of conventional one-dimensional test statistics computed on the input variable x (top row) and those of test statistics that have been computed on the classifier output using the NN and KM models (bottom row). The **LRT** (NN) is reported on every plot as a reference. We observe that the test statistics based on the **LRT** and the one based on the right tail cut-and-count (**Pois**) outperform on average all the other tests. The performances of the $\boldsymbol{\chi^2}$ test highly depends on the binning choice, while the **KS** is more powerful in case of non localized shape effects. **LRT** is the only test sensitive to H_4.

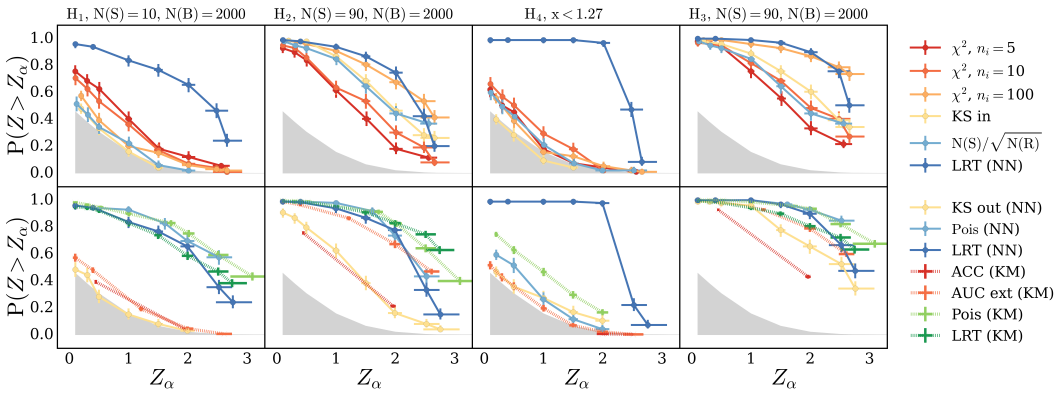

Figure 1: Power curves for different test statistics computed for the four signal benchmarks considered in the 1D experiments. The gray area represents the type I error.

Di-muon final state at the LHC: A more complex setup inspired by the realistic problem of a model-independent new physics search is that of a di-muon final state at the LHC. We perform a multivariate analysis on a set of five variables describing the kinematics of the two-body system: the transverse momenta of the two muons ($p_{T,1}$, $p_{T,2}$), the pseudorapidities (η_1, η_2) and the relative azimuthal angle ($\Delta\phi_{1,2}$). The training data are made available on Zenodo [14] by the authors of Ref. [2], where the reader can find the details of how events are generated. For the sake of this work we neglect the impact of systematic uncertainties that would affect the knowledge of the SM predictions but a strategy to include them is available in Ref. [13].

We consider three signal benchmarks: a new vector boson (Z') with the same couplings to SM fermions as the SM Z boson and mass of 200 and 300 GeV; a non-resonant effect due to a dimension-6 4-fermion contact interaction $\frac{c_W}{\Lambda} J^a_{L\mu} J^{\mu}_{La}$, where J^{μ}_{La} is the $\mathrm{SU}(2)_\mathrm{L}$ SM current. The energy scale Λ is fixed at 1 TeV and the Wilson coefficient c_W determines the coupling strength.

Two different data selections over the two-body invariant mass have been considered: a 100 GeV threshold excluding the Z pole and a lower threshold at 60 GeV which allows to include the Z pole in the dataset making the problem of detecting the signal more challenging. We adjusted the luminosities in order to keep the training statistics comparable for the two configurations: $0.35\,\mathrm{fb}^{-1}$ for the 60 GeV mass threshols and $3.5\,\mathrm{fb}^{-1}$ for the 100 GeV one. We also changed the signal cross sections in order to have comparable ideal scores.

For the NN implementation we consider a model with three layers of 5 neurons each, sigmoid activations for the hidden layers and a linear output. We train with ADAM optimizer and a weight clipping set to 2.15. The KM model has an L2 weight of $\lambda = 10^{-7}$, a gaussian width of $\sigma = 3$ and $m = 5000$ Nyström centers. The reference sample size is five times larger than the expected SM background (see [3] and [13] for more details).

Figure 2 shows the power curves obtained from the various GoF tests computed on the classifier output for the three different signal benchmarks and the two different mass cut choices. Also in this case the **LRT** and **Pois** tests outperform the other test statistics.

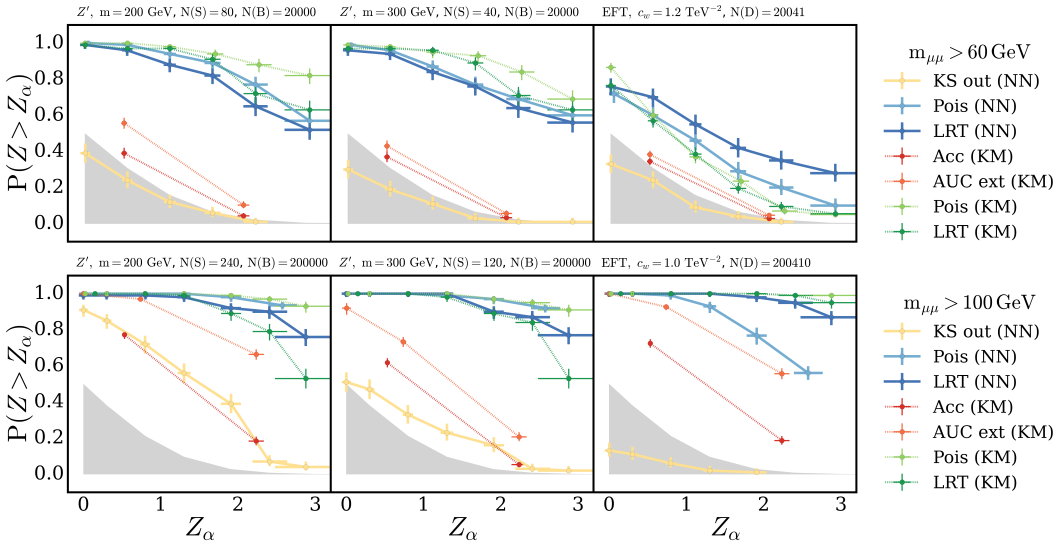

Figure 2: Power curves for different test statistics computed for the three signal benchmarks considered for the dimuon problem. The gray area represents the type I error.

4 Conclusions

The model-independent likelihood ratio test inspired by Ref. [1] can be interpreted as a GoF test and, under the training conditions prescribed in [1, 2, 3], is shown to posses an overall broad sensitivity to the types of signals considered in this work. The **Pois** test, also introduced in this work, obtains comparable or better performances for resonant signals but shows poor sensitivity to non-resonant ones, in particular in the case of a depletion region in the data. A more detailed analysis suggests that the performance gap between **LRT** and **Pois** for certain types of new physics effects resides in the way the signal component is fitted and isolated from the background. Further investigation will be needed to enhance the LRT-based algorithms to their maximum potential.

Acknowledgments and Disclosure of Funding

M.L. acknowledges the financial support of the European Research Council (grant SLING 819789). 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 no 772369). A.W. acknowledges support from the PRIN grant 2017FMJFMW.

References

[1] Raffaele Tito D'Agnolo and Andrea Wulzer. Learning New Physics from a Machine. *Phys. Rev. D*, 99(1):015014, 2019.

[2] Raffaele Tito D'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning multivariate new physics. *Eur. Phys. J. C*, 81(1):89, 2021.

[3] Marco Letizia, Gianvito Losapio, Marco Rando, Gaia Grosso, Andrea Wulzer, Maurizio Pierini, Marco Zanetti, and Lorenzo Rosasco. Learning new physics efficiently with nonparametric methods. *Eur. Phys. J. C*, 82(10):879, 2022.

[4] Purvasha Chakravarti, Mikael Kuusela, Jing Lei, and Larry Wasserman. Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests. 2 2021.

[5] Mike Williams. How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics. *JINST*, 5:P09004, 2010.

[6] Constantin Weisser and Mike Williams. Machine learning and multivariate goodness of fit. 12 2016.

[7] Jerome H. Friedman. On multivariate goodness of fit and two sample testing. *eConf*, C030908:THPD002, 2003.

[8] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR, 2020.

[9] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

[10] Gerda Claeskens and Nils Lid Hjort. Goodness of fit via non-parametric likelihood ratios. *Scandinavian Journal of Statistics*, 31(4):487–513, 2004.

[11] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[12] Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: handling billions of points efficiently. *Advances in Neural Information Processing Systems*, 33:14410–14422, 2020.

[13] Raffaele Tito d'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning new physics from an imperfect machine. *Eur. Phys. J. C*, 82(3):275, 2022.

[14] Gaia Grosso, Raffaele Tito D'Agnolo, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Nplm: Learning multivariate new physics, January 2021.

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]
   (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? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
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] The dataset used for the experiments is available on Zenodo [14]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Details of the implementations have been added as references.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(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)? [Yes]

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? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(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]