Physics Letters B 840 (2023) 137836

Contents lists available at ScienceDirect

Physics Letters B

journal homepage: www.elsevier.com/locate/physletb

Null hypothesis test for anomaly detection

Jernej F. Kamenik [a,b], Manuel Szewc [a,*]

[a] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[b] Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 11 October 2022
Received in revised form 3 March 2023
Accepted 6 March 2023
Available online 8 March 2023
Editor: M. Pierini

ABSTRACT

We extend the use of Classification Without Labels for anomaly detection with a hypothesis test designed to exclude the background-only hypothesis. By testing for statistical independence of the two discriminating dataset regions, we are able to exclude the background-only hypothesis without relying on fixed anomaly score cuts or extrapolations of background estimates between regions. The method relies on the assumption of conditional independence of anomaly score features and dataset regions, which can be ensured using existing decorrelation techniques. As a benchmark example, we consider the LHC Olympics dataset where we show that mutual information represents a suitable test for statistical independence and our method exhibits excellent and robust performance at different signal fractions even in presence of realistic feature correlations.

1. Introduction

The combination of increased experimental sensitivity and no clear leading theoretical guide for how physics beyond the standard model would manifest in current and future particle physics experiments has resulted in increased development of anomaly detection techniques for collider applications, see Ref. [1] for a living review with a continuously updated list of references. These techniques, which make use of state of the art unsupervised and/or weakly supervised algorithms, have the advantage of being sensitive to a large variety of signals at the expense of losing statistical power in comparison to dedicated searches. However, appropriately quantifying said sensitivity is still an open problem [2], with differing proposals, see e.g. Ref. [3]. An especially pressing question is how to evaluate the null hypothesis exclusion sensitivity of an anomaly detection method. The current strategy is to perform cuts using the anomalous score and extrapolate a background model from a control region. This can be problematic for several reasons. First, the use of the anomalous score itself to select events is not guaranteed to yield a robust method that disentangles the underlying processes, see e.g. Ref. [4] for a recent discussion of how ambiguities in the data representation can lead to different notions of anomalous events which vary in their discriminating power. Second, even if the anomaly score is an appropriate event selection tool, the use of cuts, which in an unsupervised search cannot be optimized on a targeted signal model, necessarily introduces a loss in sensitivity by discarding possible signal events. Finally, the use of a control region potentially introduces additional biases when assuming the absence of signal in the control region and/or employing interpolation methods such as the fit to a monotonic mass spectrum in a Bump Hunt.

In this work we aim to address some of the shortcomings outlined above. In particular, we propose a null hypothesis statistical test for anomaly detection which does not rely on fixed anomaly score cuts nor requires background model extrapolations from control regions. We apply it to a specific anomaly detection technique, Classification Without Labels (CWoLa) introduced as a quark/gluon tagger in Ref. [5] and as an anomaly detection technique in Refs. [6,7], and its extension introduced in Ref. [8] incorporating simulation assisted decorrelation of features. We show that by testing for independence between the set of features used in the anomaly score, and those used to define signal and control regions, we can obtain a p-value which avoids false signal-detection and is robust in presence of slight correlations between the two sets of features.

The work is structured as follows. In Section 2 we review CWoLa and introduce the proposed statistical test. In Section 3 we apply our method to a LHC Olympics benchmark to demonstrate its power and limitations. We conclude in Section 4 where we also discuss possible future extensions and improvements. All the necessary code to reproduce our results is available at `GitHub` [9].

* Corresponding author.
E-mail addresses: jernej.kamenik@cern.ch (J.F. Kamenik), manuel.szewc@ijs.si (M. Szewc).

https://doi.org/10.1016/j.physletb.2023.137836

2. Method

Introduced in Ref. [5], CWoLa is a weakly-supervised technique for anomaly detection which aims to learn a monotonic function of the Likelihood Ratio between Signal S and Background B processes for a set of features of interest $\vec{x}$, $\mathcal{L}_{S/B}(\vec{x}) = p(\vec{x}|S)/p(\vec{x}|B)$, with the help of an additional feature y uncorrelated with $\vec{x}$. The latter variable, often but not necessarily the invariant mass of the event, can be used to define two regions of interest: the signal region M_1 and the control (or side-band) region M_2, where the signal-to-background ratio is assumed to be higher in M_1 than in M_2. A weakly-supervised algorithm, CWoLa trains a classifier to distinguish between M_1 and M_2. The obtained output function $s(\vec{x})$ can then be mapped to $\mathcal{L}_{M_1/M_2}(\vec{x})$ through the likelihood ratio trick. The orthogonality of y and $\vec{x}$ guarantees that $\mathcal{L}_{M_1/M_2}(\vec{x})$ is a monotonous function of $\mathcal{L}_{S/B}(\vec{x})$ and thus possesses in principle optimal statistical power.

Usual applications of CWoLa use the learned optimal classifier $s(\vec{x})$ to select events of interest and assign a certain significance to the difference in selected events in M_1 and M_2. The difference in the resulting selection efficiencies $\epsilon_{M_{1,2}}$ is a smoking-gun for the presence of signal in M_1 (and also M_2). However, this is only true in the limit of infinite statistics. In a realistic setting where the dataset is finite, quantifying the degree to which the difference in efficiencies relates to the presence of signal is non-trivial. One common strategy is to assume that there is no signal in M_2 and assess the agreement between the selected events in M_1 and a background extrapolation from M_2.

Our method constitutes an alternative to assess how the learned output $s(\vec{x})$ encodes differences between M_1 and M_2 caused by the presence of a signal. To introduce it, we focus on the density estimation framing of CWoLa, which clearly defines a background-only or null hypothesis. At its heart, CWoLa is a mixture model where $\vec{x}$ and y are assumed to be conditionally independent given the process label $z = \{S, B\}$. After defining M_1 and M_2 using y, the trained classifier output is a function $s(\vec{x})$ that inherits the conditional independence with respect to y. The statistical model can be explicitly written as

$$p(s(\vec{x}), y|\pi) = (1-\pi)\, p(s(\vec{x})|B)p(y|B) + \pi\; p(s(\vec{x})|S)p(y|S)\,, \tag{1}$$

where π is the signal probability. The background-only hypothesis is explicitly written as $p(s(\vec{x}), y|\pi = 0)$ and corresponds to the case where the observed data shows independence between $s(\vec{x})$ and y. This is the key observation for our strategy. For a given measured dataset of pairs $\{s(\vec{x}_i), y_i\}$, one can assess whether they are statistically independent. If statistical independence is ruled out, the background-only hypothesis is ruled out, provided conditional independence holds. Conversely, if statistical independence cannot be ruled out, one has a clear statement about the incapability of CWoLa to discern whether any difference between M_1 and M_2 originates from the presence of a signal or is due to statistical fluctuations in the data.

Several tests of statistical independence exist for both discrete and continuous distributions, including mutual information [10], Hoeffding's D independence test [11] and distance correlation [12]. For simplicity, in the present work we focus on the use of the estimated mutual information (MI) I of the measured probability distribution. MI encodes the exact property we want to test as it measures the difference between the joint distribution and the marginals:

$$I(s, y) = D_{\rm KL}(p(s, y)||p(s)p(y)) \tag{2}$$

$$= \int ds\, dy\; p(s, y) \log \frac{p(s, y)}{p(s)p(y)}\,, \tag{3}$$

where $D_{\rm KL}(p, q)$ is the Kullback-Leibler divergence between two probability distributions, capturing how much information is lost when approximating the distribution p with the distribution q. The MI thus captures how well one can approximate the joint distribution by the product of its marginals and it is trivial to show that it vanishes for independent variables. Conditional Independence can then be expressed as a vanishing MI conditioned on a given process

$$I(s, y|z) = \int ds\, dy\; p(s, y|z) \log \frac{p(s, y|z)}{p(s|z)p(y|z)} = 0\,. \tag{4}$$

On the other hand, for the full dataset the possible mixture between the two processes encoded in $\pi \in [0, 1]$ results in

$$I(s, y) \geq 0\,, \tag{5}$$

with the equality achieved when there is only one process or the two processes have the same probability distributions.

A very nice feature of the MI is that it has well behaved asymptotic properties in the limit of small MI and large sample size [13]. Thus, we can estimate it from the measured sample of N events and obtain the p-value of said estimator $\hat{I}(s, y)$ under the null hypothesis $I(s, y) = 0$. Assuming a two dimensional binning of (s, y) with d_s and d_y the number of chosen bins per variable, the estimator $\hat{I}$ is a random variable that under the null hypothesis $I = 0$ follows a Gamma distribution with shape parameter $\frac{(d_s-1)(d_y-1)}{2}$ and scale parameter N.

To estimate $\hat{I}$ we need to estimate $\hat{p}(s, y)$, with $\hat{p}(s)$ and $\hat{p}(y)$ obtained by marginalizing. We estimate $\hat{p}(s, y)$ through two-dimensional histogram event counts with the aforementioned d_s and d_y chosen bins. Because we are dealing with continuous variables, the use of binning introduces additional hyperparameters. In this work we bin s and y in such a way that each bin has a relative statistical uncertainty equal or lower than 1%. Other criteria for statistical independence that deal explicitly with continuous variables such as Hoeffding's D independence test or distance correlation could be used to avoid the introduction of binning at the expense of increased computational cost. We choose MI as it is straightforward to implement with a general signal-blind binning criteria and it suffices to establish the relevance of the strategy detailed in this work.

We emphasize that the role of CWoLa is to provide a one-dimensional observable $s(\vec{x})$ which can then be combined with y to test for statistical independence. Once $s(\vec{x})$ is obtained, the rest of the test relies only on data without the need to introduce additional cuts or labels. If testing for statistical dependence between $\vec{x}$ and y directly was feasible, then one would not need to introduce any learnable function. However, this is often not the case. One in general needs a high-dimensional $\vec{x}$ to ensure discriminative power between possible signals and the background, which is in turn converted by our method into statistical power to exclude statistical independence. On the other hand working directly with a high-dimensional set of features renders any statistical test problematic either due to the test being designed for two variables, as is the case for Hoeffding's D independence test and distance correlation, or due to the necessary density estimation suffering from the course of dimensionality as is the case for the mutual information test presented in this work.

The method relies on the assumption of conditional independence between $\vec{x}$ and y. In a realistic application this is not ensured, specially when considering highly-discriminative variables between the background and potential signals. The presence of correlation between $\vec{x}$ and y will result in non-null MI for each

process separately. Thus, the p-value obtained from the MI estimation will be merely testing for conditional independence, not the presence of a single process. In other words, the null hypothesis ceases to be equal to the background-only hypothesis. This challenge is already present in current implementations of CWoLa, with correlations resulting in loss of classification power.

One possible strategy introduced in Ref. [8] is to ensure that $s(\vec{x})$ is agnostic to the correlation between $\vec{x}$ and y through the addition of a simulated background dataset during the training stage. In this approach, named Simulation Assisted Classification Without Labels (SA-CWoLa), the loss function is modified with an additional term that incorporates the simulation dataset. Following Ref. [8], we define the loss function as

$$\mathcal{L}_{\text{SA-CWoLa}}[s] = -\left(\sum_{\vec{x}_n\in M_1^{\text{data}}} \log s(\vec{x}_n) + \sum_{\vec{x}_n\in M_2^{\text{data}}} \log\left(1-s(\vec{x}_n)\right)\right) - \lambda\left(\sum_{\vec{x}_n\in M_1^{\text{sim.}}} \log\left(1-s(\vec{x}_n)\right) + \sum_{\vec{x}_n\in M_2^{\text{sim.}}} \log s(\vec{x}_n)\right), \tag{6}$$

that inverts the labelling in the simulation so as to penalize learning background differences between M_1 and M_2, with λ the hyper-parameter that controls the relative importance of said penalization.[1] Note that the specific choice of the loss function is not relevant as long as it ensures proper decorrelation. Similarly, the simulated dataset does not need to be perfect, it only needs to encode accurately enough the correlation between $\vec{x}$ and y for the background process. Learning to ignore the correlations in the background guarantees that excluding the null hypothesis $I(s,y)=0$ corresponds to excluding the background-only hypothesis $p(s,y)=p(s,y|\pi=0)$ and not merely excluding conditional independence $p(s,y|B)=p(s|B)p(y|B)$. The main drawback of introducing decorrelation is that the learned function $s(\vec{x})$ ceases to be optimal and looses classification power. Thus, one should tune λ with a given criteria that balances learning to decorrelate between $\vec{x}$ and y for B and learning to distinguish between B and S through discriminating between M_1 and M_2.

In this proof-of-principle, we are satisfied with presenting results for fixed λ that is large enough to ensure decorrelation in the simulation sample and at the same time small enough so that $s(\vec{x})$ is sensitive to the presence of signal in the measured sample and improves over the naive significance estimation $S/\sqrt{B}$. To verify that decorrelation is enforced, we follow Ref. [8] and compute the Area-Under-Curve (AUC) for the $s(\vec{x})$ classifier for the M_1 and M_2 samples in the simulation dataset. If the AUC is approximately 0.5, we have a classifier that is not better than a random classifier and thus it has learned to ignore any possible correlations between $\vec{x}$ and y. We emphasize that the sole purpose of the simulation dataset is to ensure decorrelation, and we never compare data to simulation to obtain a significance after training. This makes the test more robust against background mismodelling than other unsupervised methods for anomaly detection which avoid the use of anomaly cuts at the expense of performing data-to-simulation hypothesis tests such as Refs. [14–19]. This is only possible because we assume that the simulations are precise enough to capture qualitative correlations between features in data. The simulator precision needed for decorrelation to be effective is considerably less than what is needed for a full multivariate comparison with measurements which often suffers both from systematic biases in the simulations and from the computational cost of achieving a given statistical precision.

[1] We always reweigh the events from both data and simulation during training in such a way that each of the four subset of events $\{M_1^{\text{data}}, M_2^{\text{data}}, M_1^{\text{sim.}}, M_2^{\text{sim.}}\}$ has the same total weight.

3. Application: LHC Olympics

In order to demonstrate its power, we apply our method to the LHC Olympics R&D labelled dataset [20]. The dataset is comprised of dijet events from two different sources: SM quantum chromodynamics (QCD) processes (background B), and the production of a hypothetical new resonance W' with mass $m_{W'}=3.5$ TeV, decaying to two intermediate particles X and Y with masses $m_X=500$ GeV and $m_Y=100$ GeV, which in turn both decay promptly to pairs of quarks producing two large-radius jets with a two-prong substructure (signal S). Our variable of interest y is the reconstructed dijet invariant mass m_{jj} of the two hardest (in p_T) jets in the event. Mimicking the selection criteria of Ref. [8], the selected events have a reconstructed dijet mass $m_{jj}\in[3.1,3.9]$ TeV. To perform CWoLa [5], we define two orthogonal regions $M_1\equiv\{m_{jj}\in[3.3,3.7]\text{ TeV}\}$ and $M_2\equiv\{m_{jj}\in[3.1,3.3]\text{ TeV}\cup[3.7,3.9]\text{ TeV}\}$. In the following, we refer to S and B as the total number of Signal and Background events in $M_1\cup M_2$.

For anomaly score input features $\vec{x}$, we choose a set of variables based on the invariant masses and the first N-subjettiness ratios [21,22] of the two selected jets. Ordering the jets by mass, with $j=1$ being the heavier jet, our variables are

$$\vec{x}=\{m_1-m_2, m_2, \tau_{21,1}, \tau_{21,2}\}. \tag{7}$$

The correlation between $\vec{x}$ and m_{jj} is mostly concentrated in the correlations between $\{m_1,m_2\}$ and m_{jj}. To illustrate how important they are, we define

$$\Delta_c^{ab}(m_{jj}^{\text{bin}}) = \frac{\mathbb{E}[m_c|m_{jj}\in m_{jj}^{\text{bin}}, a]-\mathbb{E}[m_c|b]}{\mathbb{E}[m_c|b]}, \tag{8}$$

where $a,b\in\{B,S\}$ and $c\in\{1,2\}$. $\Delta_c^{ab}(m_{jj}^{\text{bin}})$ represents the relative difference between the average of m_c in a given m_{jj} bin m_{jj}^{bin} for process a and the average of the same observable over the whole m_{jj} range for process b. If $a=b$, this observable conveys the presence of a correlation between m_c and m_{jj} for a given process. For $a\neq b$, this observable conveys the difference in the m_c probability distributions for the two processes and its dependence on m_{jj}. By comparing Δ_c^{BB} with $(S/B)\Delta_c^{SB}$ we can check whether the correlations between m_c and m_{jj} can obscure the differences between signal and background that CWoLa aims to learn. The prefactor S/B is introduced to account for the fact that the data contains less signal than background and originates from comparing the full data distribution to the background-only distribution and separating the signal and the background contributions:

$$\Delta_c^{S+B,B} = \frac{B}{S+B}\Delta_c^{BB} + \frac{S}{S+B}\Delta_c^{SB} \approx \Delta_c^{BB} + (S/B)\Delta_c^{SB}.$$

We show in Fig. 1 the resulting distributions for m_1 and m_2, with the largest S/B considered in this work, $S/B=0.01$. We observe how the correlation between features for the background process, as evidenced by the monotonic increase of Δ_c^{BB} from negative to positive values towards larger m_{jj}, is sizable and crucially more pronounced compared to the S/B weighted difference between S and B as traced by Δ_c^{SB}. In other words, the correlation between m_c and m_{jj} in the background can easily mask the presence of a small signal. For lower S/B, the correlations become even more dominant and can lead to a strongly biased anomaly score. In addition, in our approach the correlations will also induce statistical dependence between $s(\vec{x})$ and $M_{1,2}$ even in absence of

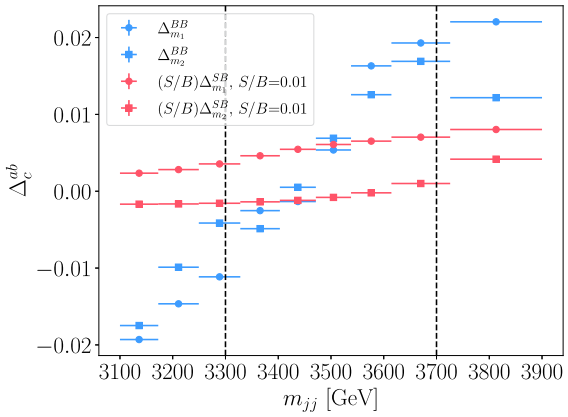

Fig. 1. Distribution of Δ^{ab}_c defined in Eq. (8) comparing correlations between $\{m_1, m_2\}$ and m_{jj} to the differences between S and B for the largest S/B considered. Vertical error bars signify statistical uncertainties, while horizontal bars indicate m_{jj} bin width. Vertical dashed lines denote boundaries between signal (M_1) and side-band (M_2) regions. See text for details.

a signal and thus jeopardize the validity of the null-hypothesis test.

In order to address this crucial issue, we follow Ref. [8] and incorporate a set of simulated events into the training stage. As simulation, we consider the background provided in the labelled version of the Black Box 1 (BB1) dataset. We use BB1 as simulation to take advantage of the larger signal sample provided in the R&D dataset. Using the loss function defined in Eq. (6), the classifier is then trained to distinguish M_1 and M_2 in data but not in simulation, obtaining a $s(\vec{x})$ that is agnostic to correlations between $\vec{x}$ and y for QCD. As a classifier, we use a very similar set-up as in Ref. [8]: we train a Neural Network composed of three hidden layers with 64 nodes each and `ReLU` activation function with a sigmoid function applied to the output to ensure $s(\vec{x}) \in [0, 1]$ for 20 epochs using the `ADAM` [23] optimizer. The Neural Network is implemented in `PyTorch` [24]. We have trained our classifier using k-fold cross-validation with $k = 10$ to avoid overfitting by ensuring that every $s(\vec{x}_n)$ is obtained by combining the data point $\vec{x}_n$ with a classifier which has not seen $\vec{x}_n$ during training. We also perform several random weight initializations to ensure better convergence. At the end of training, we evaluate the AUC score between the M_1 and M_2 simulated samples to ensure that decorrelation is achieved.

We show in Fig. 2 the learned $s(\vec{x})$ for different S/B with $B = 250$k and $\lambda = \{0.0, 1.0\}$. In each plot, the binning is chosen in such a way that each bin has a relative statistical uncertainty lower than or equal to 1%. This choice ensures good performance of the density estimation needed for the hypothesis test. We can appreciate how as S/B increases, $s(\vec{x})$ goes from being mostly centered around $s = 0.5$ to yielding higher s values, indicating improved learning of signal features. However, the binning choice obscures somewhat how much the background and signal are separated (within the highest s bin), as the whole signal is grouped together with the necessary background events to obtain a 1% statistical uncertainty. In absence of correlation mitigation (for $\lambda = 0$), anomaly score bias causes clearly unbalanced classification of events in M_1 and M_2 even in absence of any signal. The introduction of $\lambda = 1$ causes the events to be even more centered around 0.5, specially for low S/B. However, more importantly, $\lambda \geq 0$ forces the training to ignore possible correlations between $\vec{x}$ and y for (simulated) background and thus approaches a random classifier for $S/B \to 0$.

As expected, the impact of λ is even more pronounced when testing for statistical independence. For each dataset of $\{s, m_{jj}\}$ values, we estimate mutual information and obtain the p-value associated with the null hypothesis as detailed in Section 2. We show in Fig. 3 the resulting estimated $\hat{I}_{\text{data}}$ and their corresponding p-values. We also ran a series of pseudo-experiments to verify that the asymptotic limit is appropriate. We only present results for $\lambda = 1$ because for $\lambda = 0$ we are able exclude $I(s, y) = 0$ for all S/B with very high confidence (p-value $< 10^{-14}$). This, as detailed in Section 2, is because we are excluding conditional independence in the background process due to correlations between $\vec{x}$ and y. This is specially important for $S/B = 0$, where the effect of correlations can mislead CWoLa to falsely exclude the background-only hypothesis.

We observe that for $\lambda = 1$ the proposed test has the required behavior: for $S/B = 0$ the test yields results consistent with statistical independence, while an increase of S/B leads to an increasingly strong exclusion of the null hypothesis. The use of SA-CWoLa thus ensures that we can identify the null hypothesis with the background-only hypothesis. For $S/B > 0$, we also compute the discovery significance $Z = \Phi^{-1}(1 - p)$, where Φ is the unit Gaussian cumulative distribution function, and compare it to the naive counting significance $Z_0 = S/\sqrt{B}$. We observe how our method presents an increased discovery significance even compared to the case of perfect (up to statistical fluctuation) knowledge of background yields.

Overall, Fig. 3 shows how $\hat{I}$ can be used to infer whether the data presents a deviation from the null hypothesis, defined as the case where a single process (or an m_{jj} independent mixture of processes) is present for which s and m_{jj} are independent. This is the analogous to the p-value obtained using the Bump Hunt in a "traditional" implementation of CWoLa. However, contrary to existing approaches, here there is no selection cut to be optimized, and no extrapolation of the background into the signal region is required.

In principle the method could learn the likelihood-ratio between signal and background, which is the optimal test-statistic. However, the presence of the decorrelation term in Eq. (6) reduces the optimality of s and consequently $\hat{I}$. This can be seen by the decrease in Z/Z_0 as S/B increases. When S/B is large enough, $s(\vec{x})$ will ignore the small correlations in each individual process even for $\lambda = 0.0$. A non-null λ will thus only worsen the performance of the algorithm. However, as we are interested in low S/B cases where anomaly detection is useful, a more conservative approach which is robust to correlations even when there is no signal present is warranted.

In the previous paragraphs, we have shown how the proposed method yields an appropriate test statistic which is different from existing approaches. In Table 1 we provide a significance comparison of the proposed method to two traditional implementations of CWoLa which we denote as "Anomaly cuts" and "Bump Hunt". The former, implemented e.g. in Ref. [25], assumes that no signal is present in the side-band region M_2 and estimates the total number of background events in the signal region M_1 through the use of cuts on the anomaly score. For a fixed efficiency in the side-band region ϵ_2, the estimated background event yield in signal region is $\epsilon_2 N_1$. If the measured efficiency in the signal region ϵ_1 is larger than ϵ_2, then there is an excess of events with a significance of

$$Z = \begin{cases} \frac{(\epsilon_1 - \epsilon_2)N_1}{\sqrt{\epsilon_2(N_1+N_2)}}, & \text{if } \epsilon_1 \geq \epsilon_2\,, \\ 0, & \text{if } \epsilon_1 < \epsilon_2\,. \end{cases}$$

Similarly, the Bump Hunt method also assumes no signal is present in M_2 but estimates the background event yield in M_1

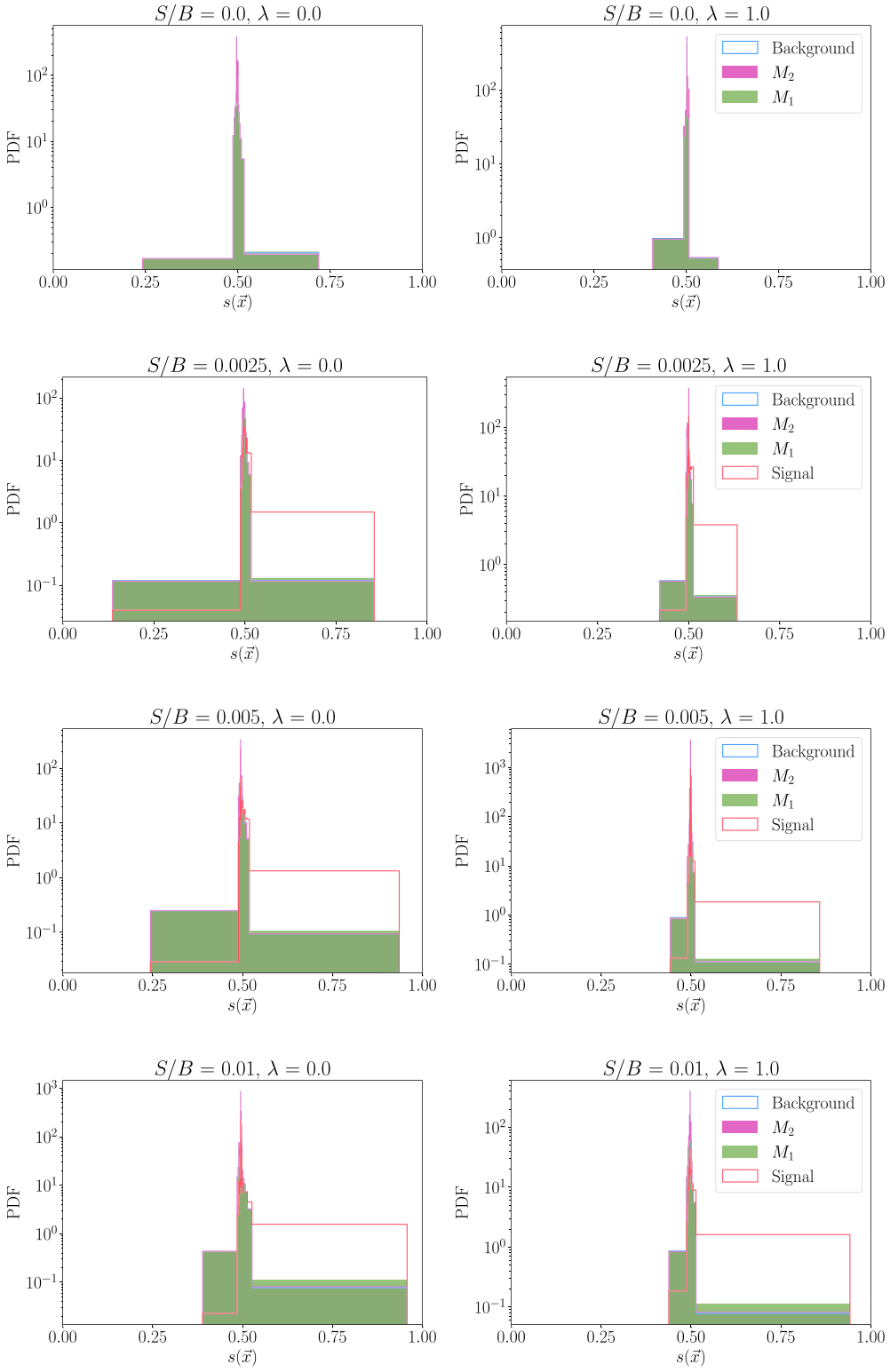

Fig. 2. Anomaly score $s(\vec{x})$ probability density function (PDF) after training for different event labellings. Each row corresponds to a given S/B and each column to a given λ. Non-uniform binning ensures that when considering the full dataset each bin has a relative statistical uncertainty equal or lower than 1%, resulting in between 15 and 25 bins per plot. The data is shown both labelled according to the (observable) values of m_{jj} (defining M_1 and M_2) as well as according to the (unobservable) truth labels (background and signal).

differently. In this approach, the background m_{jj} distribution is explicitly modelled and a Profile Likelihood Ratio fit of the number of signal events in M_1 is performed. We follow Ref. [8] and model the background distribution for $m_{jj} \in [3.1, 3.9]$ TeV as

$$\frac{d\sigma}{dm_{jj}} = \frac{p_0 \left(1 - \frac{m_{jj}}{\sqrt{s}}\right)^{p_1}}{\left(\frac{m_{jj}}{\sqrt{s}}\right)^{p_2 + p_3 \log \frac{m_{jj}}{\sqrt{s}}}},$$

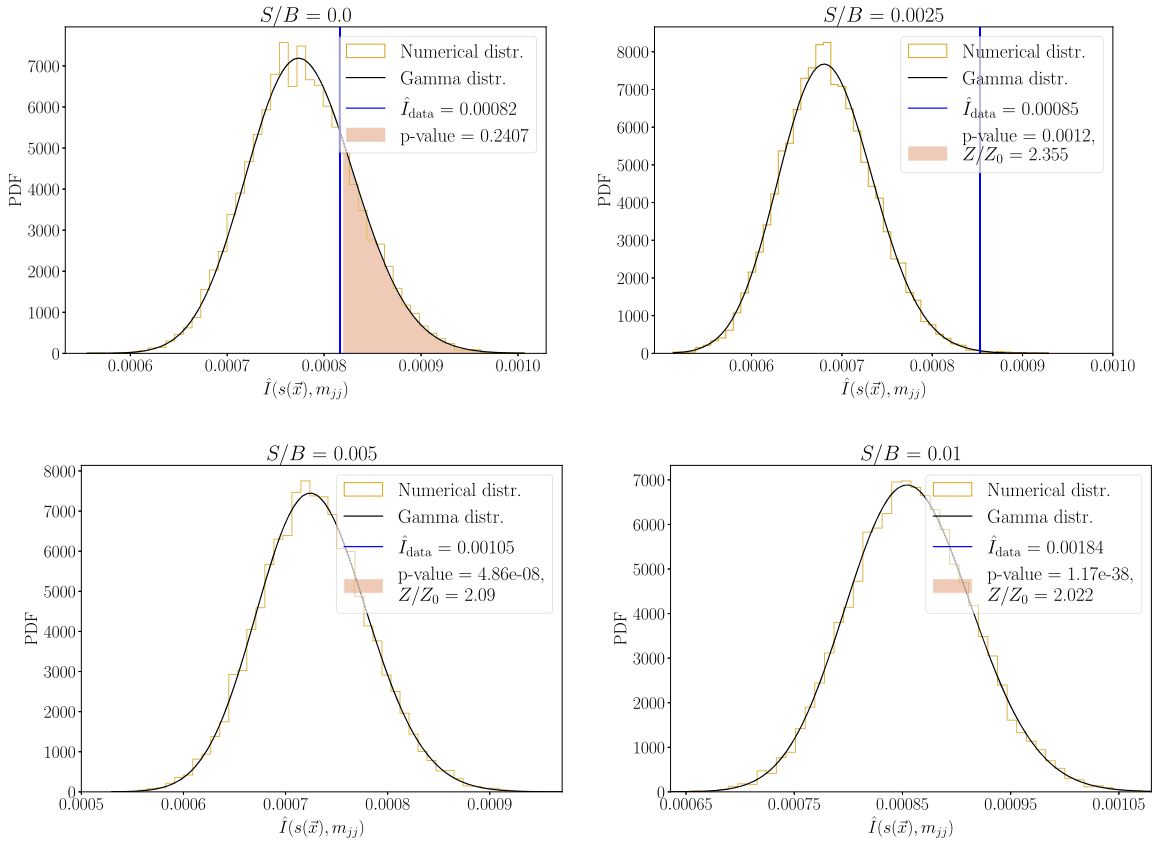

Fig. 3. PDFs of estimated mutual information, its numerical distribution under the null hypothesis estimated through resampling, and its asymptotic distribution under the null hypothesis, for different considered benchmark datasets. Each plot corresponds to a different choice of S/B with $\lambda = 1$. We combine the estimated $\hat{I}$ with their asymptotic distribution to obtain the resulting p-values. When $S/B > 0$, we also compute the discovery significance $Z = \Phi^{-1}(1-p)$, where Φ is the unit Gaussian cumulative distribution function, and compare it to $Z_0 = S/\sqrt{B}$.

where $\sqrt{s}$ is the center-of-mass energy and p_i are parameters to be fitted from the m_{jj} distribution with the signal region masked. Once the fit is performed, the expected number of background events in M_1 b is estimated from the integral of the background distribution over the signal region. We define the Likelihood function in the signal region as

$$\mathcal{L}(s,\theta) = \mathcal{P}(N_1|s+b+\theta)\mathcal{N}(\theta|0,\sigma)\,,$$

where $\mathcal{P}$ is the Poisson probability mass function of measuring N_1 events, $\mathcal{N}$ is the Normal probability density, θ is a nuisance parameter for background mismodelling and σ is the background yield error propagated from the p_i fit. From this Likelihood we build the usual test statistic

$$q_0 = \begin{cases} -2\,\text{Ln}\,\mathcal{L}(0,\hat{\hat{\theta}})/\mathcal{L}(\hat{s},\hat{\theta}), & \text{if } \hat{s} \geq 0\,, \\ 0, & \text{if } \hat{s} < 0\,, \end{cases}$$

where $\hat{s},\hat{\theta}$ are the maximum likelihood estimates of s and θ and $\hat{\hat{\theta}}$ is the maximum likelihood estimate of θ when keeping s fixed to 0. The resulting significance is $Z = \sqrt{q_0}$. We implement the Bump Hunt by itself and in conjunction with the use of cuts in the anomaly score to enhance the efficiency as would be done in a resonance search.

From Table 1 we observe how in every case the introduction of $\lambda > 0$ reduces the significance. However, it does not imply resilience against spurious signals for all strategies. Both traditional methods are highly dependent on the arbitrary ϵ_2 choice, showing the appearance of spurious significance at $S/B = 0$ for certain values. In general, we observe that our method is better suited for smaller S/B than existing methods. This is mainly because it does not discard potential signal events. The Bump Hunt with no cuts also does this, but its significance is lower for non-null S/B (for the extreme $S/B = 0.01$, its significance is even lower than $S/\sqrt{B}$ due to the presence of the nuisance parameter). From this comparison, we assess that without clear criteria for an optimal anomaly cut (ϵ_2), the mutual information-based method performs better for low to null S/B whereas the cut and count based methods outperform the mutual information test for larger S/B.

Another benefit of our model compared to traditional searches is scalability with sample size. If decorrelation is ensured, the larger the sample size the more powerful the method. This is certainly not true for the Bump Hunt, where the background modelling is an inherent approximation which necessarily introduces bias for large enough sample sizes. Conversely, for smaller datasets our method takes advantage of the full dataset in a better way than through the use of fixed anomaly cuts. The asymptotic approximation for the mutual information CDF, which is vital for the test, has been shown numerically [13] to be valid for $N > 50$ events and true mutual information $I \leq 0.14$, conditions that we expect to always be satisfied in a realistic anomaly search at the LHC.

4. Discussions and outlook

In this work, we have presented a novel strategy to quantify the sensitivity of a specific anomaly detection technique, Simulation-Assisted Classification Without Labels, by testing for statistical independence of the learned $\{s(\vec{x}), y\}$ samples. We have shown that

Table 1
Significances obtained with different strategies for different S/B ratios, see text for details.

Significance	S/B = 0.0	S/B = 0.0025	S/B = 0.005	S/B = 0.01
$S/\sqrt{B}$	0.0	1.29	2.55	6.40
Mutual Info $\lambda = 0.0$	6.40	7.04	7.58	14.1
Mutual Info $\lambda = 1.0$	0.70	3.03	5.33	13.0
Anomaly cuts $\epsilon_2 = 0.1$, $\lambda = 0.0$	3.35	4.78	6.27	11.6
Anomaly cuts $\epsilon_2 = 0.1$, $\lambda = 1.0$	2.48	2.26	4.49	10.0
Anomaly cuts $\epsilon_2 = 0.01$, $\lambda = 0.0$	2.26	4.62	10.1	27.0
Anomaly cuts $\epsilon_2 = 0.01$, $\lambda = 1.0$	0.55	1.66	10.7	27.1
Anomaly cuts $\epsilon_2 = 0.001$, $\lambda = 0.0$	1.39	10.3	17.9	34.2
Anomaly cuts $\epsilon_2 = 0.001$, $\lambda = 1.0$	0.	0.57	13.6	37.0
Bump Hunt	0.95	1.97	2.74	5.30
Bump Hunt $\epsilon_2 = 0.1$, $\lambda = 0.0$	6.41	9.26	10.92	19.5
Bump Hunt $\epsilon_2 = 0.1$, $\lambda = 1.0$	3.81	4.35	6.93	16.0
Bump Hunt $\epsilon_2 = 0.01$, $\lambda = 0.0$	4.77	6.96	14.2	34.7
Bump Hunt $\epsilon_2 = 0.01$, $\lambda = 1.0$	0.97	2.53	14.0	35.0
Bump Hunt $\epsilon_2 = 0.001$, $\lambda = 0.0$	2.98	12.3	20.0	35.8
Bump Hunt $\epsilon_2 = 0.001$, $\lambda = 1.0$	0.29	1.60	15.9	38.7

as long as one can rely on SA-CWoLa to enforce conditional independence of the background processes $p(s, y|B) = p(s|B)p(y|B)$, the null hypothesis of statistical independence is equivalent to the background-only hypothesis. Thus, testing for statistical independence in the observed data corresponds to testing for the background-only hypothesis.

As a proof of principle, we have considered mutual information as a test statistic. MI has a known asymptotic distribution under the null hypothesis for binned data and has low computational cost. We have tested our method with LHC Olympics datasets and have shown that the test statistic yields the expected behavior. Most importantly our proposed test statistic provides a clear statement on the presence of signal, i.e. is capable of correctly yielding a no-signal response. This opens the door to testing for new physics in LHC datasets without the need for anomaly score cuts, as well as reducing the need for accurate background modelling.

Possible extensions of the present work could consider other tests for statistical independence such as Hoeffding's D independence test or distance correlation, which can be applied on unbinned $s(\vec{x})$ and y, at the expense of increased computational cost. Similarly, other methods for anomaly score training and decorrelation of features could be explored. Employing the most suitable classification and decorrelation methods can be model and dataset dependent, and has not been the main focus of this work.

Another possibility is to assume that no signal populates the side-bands and identify $p(\hat{s}|M_2) = p(\hat{s}|z = 0)$. This opens the door for an optimal analysis since one can now perform a template fit in the signal region [26]. However, it also potentially introduces additional uncertainties and/or biases due to background modelling that the present method avoids. We leave a more complete study in this direction for future work.

Regarding other physics applications, we emphasize that by dispensing with the need for explicit functional background modelling, our test is especially useful for anomaly detection applications that do not search for predetermined (modelled) signal shapes [2], such as invariant mass resonances as in e.g. Refs. [25, 27]. However, it could also be applied in supervised searches as an additional cross-check to control bias due to modelling at the expense of loss of optimality.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All the necessary code to reproduce our results is available at the Github repository https://github.com/ManuelSzewc/Null_Hypothesis_Test_for_Anomaly_Detection.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (grant No. J1-3013 and research core funding No. P1-0035). MS is grateful to the Mainz Institute for Theoretical Physics (MITP) of the Cluster of Excellence PRISMA$^+$ (Project ID 39083149) and to GGI for their hospitality, support and useful discussions.

References

[1] M. Feickert, B. Nachman, arXiv:2102.02770 [hep-ph], 2021.
[2] P. Shanahan, et al., arXiv:2209.07559 [physics.comp-ph], 2022.
[3] S.E. Park, P. Harris, B. Ostdiek, arXiv:2208.05484 [hep-ph], 2022.
[4] G. Kasieczka, R. Mastandrea, V. Mikuni, B. Nachman, M. Pettee, D. Shih, arXiv:2209.06225 [hep-ph], 2022.
[5] E.M. Metodiev, B. Nachman, J. Thaler, J. High Energy Phys. 10 (2017) 174, arXiv:1708.02949 [hep-ph].
[6] J.H. Collins, K. Howe, B. Nachman, Phys. Rev. Lett. 121 (2018) 241803, arXiv:1805.02664 [hep-ph].
[7] J.H. Collins, K. Howe, B. Nachman, Phys. Rev. D 99 (2019) 014038, arXiv:1902.02634 [hep-ph].
[8] K. Benkendorfer, L.L. Pottier, B. Nachman, Phys. Rev. D 104 (2021) 035003, arXiv:2009.02205 [hep-ph].
[9] Null hypothesis test for anomaly detection, https://github.com/ManuelSzewc/Null_Hypothesis_Test_for_Anomaly_Detection.
[10] C.M. Bishop, Pattern Recognition and Machine Learning, Information Science and Statistics, Springer, New York, NY, 2006, softcover published in 2016.
[11] W. Hoeffding, Ann. Math. Stat. 19 (1948) 546.
[12] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Ann. Stat. 35 (2007) 2769.
[13] B. Goebel, Z. Dawy, J. Hagenauer, J. Mueller, An Approximation to the Distribution of Finite Sample Size Mutual Information Estimates, vol. 2, 2005, pp. 1102–1106.
[14] R.T. D'Agnolo, A. Wulzer, Phys. Rev. D 99 (2019) 015014, arXiv:1806.02350 [hep-ph].

[15] R.T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Eur. Phys. J. C 81 (2021) 89, arXiv:1912.12155 [hep-ph].
[16] R.T. d'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Eur. Phys. J. C 82 (2022) 275, arXiv:2111.13633 [hep-ph].
[17] M. Letizia, G. Losapio, M. Rando, G. Grosso, A. Wulzer, M. Pierini, M. Zanetti, L. Rosasco, Eur. Phys. J. C 82 (2022) 879, arXiv:2204.02317 [hep-ph].
[18] K. Krzyżańska, B. Nachman, arXiv:2203.09601 [hep-ph], 2022.
[19] P. Chakravarti, M. Kuusela, J. Lei, L. Wasserman, arXiv:2102.07679 [stat.AP], 2021.
[20] G. Kasieczka, et al., arXiv:2101.08320 [hep-ph], 2021.
[21] J. Thaler, K. Van Tilburg, J. High Energy Phys. 03 (2011) 015, arXiv:1011.2268 [hep-ph].
[22] J. Thaler, K. Van Tilburg, J. High Energy Phys. 02 (2012) 093, arXiv:1108.2701 [hep-ph].
[23] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014.
[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, pp. 8024–8035.
[25] T. Finke, M. Krämer, M. Lipp, A. Mück, J. High Energy Phys. 08 (2022) 015, arXiv:2204.11889 [hep-ph].
[26] B. Nachman, SciPost Phys. 8 (2020) 090, arXiv:1909.03081 [hep-ph].
[27] E. Alvarez, F. Lamagna, M. Szewc, J. High Energy Phys. 01 (2020) 049, arXiv:1911.09699 [hep-ph].