Eur. Phys. J. C (2019) 79:289
https://doi.org/10.1140/epjc/s10052-019-6787-3

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Theoretical Physics

Guiding new physics searches with unsupervised learning

Andrea De Simone[1,2,a], **Thomas Jacques**[1,2,b]

[1] SISSA, Via Bonomea 265, 34136 Trieste, Italy
[2] INFN Sezione di Trieste, Via Bonomea 265, 34136 Trieste, Italy

Received: 12 December 2018 / Accepted: 14 March 2019

Abstract We propose a new scientific application of unsupervised learning techniques to boost our ability to search for new phenomena in data, by detecting discrepancies between two datasets. These could be, for example, a simulated standard-model background, and an observed dataset containing a potential hidden signal of New Physics. We build a statistical test upon a test statistic which measures deviations between two samples, using a Nearest Neighbors approach to estimate the local ratio of the density of points. The test is model-independent and non-parametric, requiring no knowledge of the shape of the underlying distributions, and it does not bin the data, thus retaining full information from the multidimensional feature space. As a proof-of-concept, we apply our method to synthetic Gaussian data, and to a simulated dark matter signal at the Large Hadron Collider. Even in the case where the background can not be simulated accurately enough to claim discovery, the technique is a powerful tool to identify regions of interest for further study.

Contents

1 Introduction

The problem of comparing two independent data samples and looking for deviations is ubiquitous in statistical analyses. It is of particular interest in physics, when addressing the problem of searching for new phenomena in data, to compare observations with expectations to find discrepancies. In general, one would like to assess (in a statistically sound way) whether the observed experimental data are compatible with the expectations, or there are signals of the presence of new phenomena.

In high-energy physics, although the Standard Model (SM) of particle physics has proved to be extremely successful in predicting a huge variety of elementary particle processes with spectacular accuracy, it is widely accepted that it needs to be extended to account for unexplained phenomena, such as the dark matter of the Universe, the neutrino masses, and more. The search for New Physics (NP) beyond the SM is the primary goal of the Large Hadron Collider (LHC). The majority of NP searches at the LHC are performed to discover or constrain specific models, i.e. specific particle physics extensions of the SM. Relatively less effort has been devoted to design and carry out strategies for model-independent searches for NP [1–9]. At the current stage of no evidence for NP in the LHC data, it is of paramount importance to increase the chances of observing the presence of NP in the data. It may even be already there, but it may have been missed by model-specific searches.

Recently, there has been growing interest in applying machine learning (ML) techniques to high-energy physics

[a] e-mail: andrea.desimone@sissa.it

[b] e-mail: thomas.jacques@sissa.it

Published online: 29 March 2019

problems, especially using supervised learning (see e.g. Refs. [10–27] and in particular the recent work of Ref. [8] with which we share some ideas, although with a very different implementation). On the other hand, applications of unsupervised learning have been relatively unexplored [10,28,29]. In unsupervised learning the data are not labeled, so the presence and the characteristics of new phenomena in the data are not known a priori. One disadvantage of unsupervised learning is that one cannot easily associate a performance metric to the algorithm. Nevertheless, unsupervised methods such as anomaly (or outlier) detection techniques, or clustering algorithms, provide powerful tools to inspect the global and local structures of high-dimensional datasets and discover 'never-seen-before' processes.

In this paper, we propose a new scientific application of unsupervised learning techniques to boost our ability to search for new phenomena in data, by measuring the degree of compatibility between two data samples (e.g. observations and predictions). In particular, we build a statistical test upon a test statistic which measures deviations between two datasets, relying on a Nearest-Neighbors technique to estimate the ratio of the local densities of points in the samples.

Generally speaking, there are three main difficulties one may face when trying to carry out a search for the presence of new processes in data: (1) a model for the physics describing the new process needs to be assumed, which limits the generality of the method; (2) it is impossible or computationally very expensive to evaluate directly the likelihood function, e.g. due to the complexity of the experimental apparatus; (3) a subset of relevant features needed to be extracted from the data, otherwise the histogram methods may fail due to the sparsity of points in high-dimensional bins.

A typical search for NP at LHC suffers from all such limitations: a model of NP (which will produce a signal, in the high-energy physics language) is assumed, the likelihood evaluation is highly impractical, and a few physically motivated variables (observables or functions of observables) are selected to maximize the presence of the signal with respect to the scenario without NP (the so-called background).

Our approach overcomes all of these problems at once, by having the following properties:

1. it is *model-independent*: it aims at assessing whether or not the observed data contain traces of new phenomena (e.g. due to NP), regardless of the specific physical model which may have generated them;
2. it is *non-parametric*: it does not make any assumptions about the probability distributions from which the data are drawn, so it is likelihood-free;
3. it is *un-binned*: it partitions the feature space of data without using fixed rectangular bins; so it allows one to retain and exploit the information from the full high-dimensional feature space, when single or few variables cannot.

The method we propose in this paper is particularly useful when dealing with situations where the distribution of data in feature space is almost indistinguishable from the distribution of the reference (background) model.

Although our main focus will be on high-energy particle physics searches at the LHC, our method can be successfully applied in many other situations where one needs to detect incompatibilities between data samples.

The remainder of the paper is organized as follows. In Sect. 2 we describe the details of the construction of our method and its properties. In Sect. 3 we apply it to case studies with simulated data, both for synthetic Gaussian samples and for a more physics-motivated example related to LHC searches. We outline some directions for further improvements and extensions of our approach, in Sect. 4. Finally, we conclude in Sect. 5.

2 Statistical test of dataset compatibility

In general terms, we approach the problem of measuring the compatibility between datasets sampled from unknown probability densities, by first estimating the probability densities and then applying a notion of functional distance between them. The first task is worked out by performing density ratio estimation using Nearest Neighbors, while the distance between probability densities is chosen to be the Kullback–Leibler divergence [30]. We now describe our statistical test in more detail.

2.1 Definition of the problem

Let us start by defining the problem more formally. Let $\{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^{N_T}$ and $\{\boldsymbol{x}'_i | \boldsymbol{x}'_i \in \mathbb{R}^D\}_{i=1}^{N_B}$ be two independent and identically distributed D-dimensional samples drawn independently from the probability density functions (PDFs) p_T and p_B, respectively:

$$\mathcal{T} \equiv \{\boldsymbol{x}_i\}_{i=1}^{N_T} \overset{\text{iid}}{\sim} p_T, \tag{2.1}$$

$$\mathcal{B} \equiv \{\boldsymbol{x}'_i\}_{i=1}^{N_B} \overset{\text{iid}}{\sim} p_B. \tag{2.2}$$

We will refer to $\mathcal{B}$ as a 'benchmark' (or 'control' or 'reference') sample and to $\mathcal{T}$ as a 'trial' (or 'test') sample. The $\mathcal{T}$, $\mathcal{B}$ samples consist of N_T, N_B points, respectively. The $\mathbb{R}^D$ space where the sample points $\boldsymbol{x}_i$, $\boldsymbol{x}'_i$ live will be referred to as 'feature' space.

The primary goal is to check whether the two samples are drawn from the same PDF, i.e. whether $p_B = p_T$. In other words, we aim at assessing whether (and to what significance level) the two samples are compatible with each other. More formally, we want to perform a statistical test of the null hypothesis $\{H_0 : p_T = p_B\}$ versus the alternative hypothesis $\{H_1 : p_T \neq p_B\}$.

This problem is well-known in the statistics literature as a *two-sample* (or *homogeneity*) test, and many ways to handle it have been proposed. We want to construct a statistical hypothesis test of dataset compatibility satisfying the properties 1–3 outlined in the introduction.

First, the $\mathcal{B}, \mathcal{T}$ samples are going to be analyzed without any particular assumptions about the underlying model that generated them (property 1); our hypothesis test does not try to infer or estimate the parameters of the parent distributions, but it simply outputs to what degree the two samples can be considered compatible.

Second, if one is only interested in a location test, such as determining whether the two samples have the same mean or variance, then a t test is often adopted. However, we assume no knowledge about the original PDFs, and we want to check the equality or difference of the two PDFs as a whole; therefore, we will follow a non-parametric (distribution-free) approach (property 2).

Third, we want to retain the full multi-dimensional information of the data samples, but high-dimensional histograms may result in sparse bins of poor statistical use. The popular Kolmogorov–Smirnov method only works for one-dimensional data, and extensions to multi-dimensional data are usually based on binning (for an alternative method that instead reduces the dimensionality of the data to one, see Ref. [16]).

Alternative non-parametric tests like the Cramér–von Mises–Anderson test or the Mann–Whitney test require the possibility of ranking the data points in an ordinal way, which may be ill-defined or ambiguous in high-dimensions. Thus, we will employ a different partition of feature space not based on fixed rectangular bins (property 3), which allows us to perform a non-parametric two-sample test in high dimensions.

So, in order to construct our hypothesis test satisfying properties 1–3, we need to build a new test statistic and construct its distribution, as described in the next sections.

2.2 Test statistic

Since we are interested in measuring the deviation between the two samples, it is convenient to define the ratio of probability densities to observe the points in the two samples, in the case $p_B \neq p_T$ (numerator) relative to the case $p_B = p_T$ (denominator)

$$\lambda \equiv \frac{\prod_{\boldsymbol{x}'_j \in \mathcal{B}} p_B(\boldsymbol{x}'_j) \prod_{\boldsymbol{x}_j \in \mathcal{T}} p_T(\boldsymbol{x}_j)}{\prod_{\boldsymbol{x}'_j \in \mathcal{B}} p_B(\boldsymbol{x}'_j) \prod_{\boldsymbol{x}_j \in \mathcal{T}} p_B(\boldsymbol{x}_j)} = \prod_{\boldsymbol{x}_j \in \mathcal{T}} \frac{p_T(\boldsymbol{x}_j)}{p_B(\boldsymbol{x}_j)}. \tag{2.3}$$

The above quantity may also be thought of as a likelihood ratio. However, as we are carrying out a non-parametric test, we prefer not to use this term to avoid confusion.

Now, since the true PDFs $p_{B,T}$ are not known, we follow the approach of finding estimators $\hat{p}_{B,T}$ for the PDFs and evaluate the ratio λ on them

$$\hat{\lambda} = \prod_{\boldsymbol{x}_j \in \mathcal{T}} \frac{\hat{p}_T(\boldsymbol{x}_j)}{\hat{p}_B(\boldsymbol{x}_j)}. \tag{2.4}$$

We then define our *test statistic* TS over the trial sample as

$$\mathrm{TS}(\mathcal{B}, \mathcal{T}) \equiv \log \hat{\lambda}^{1/|\mathcal{T}|} = \frac{1}{N_T} \sum_{j=1}^{N_T} \log \frac{\hat{p}_T(\boldsymbol{x}_j)}{\hat{p}_B(\boldsymbol{x}_j)}, \tag{2.5}$$

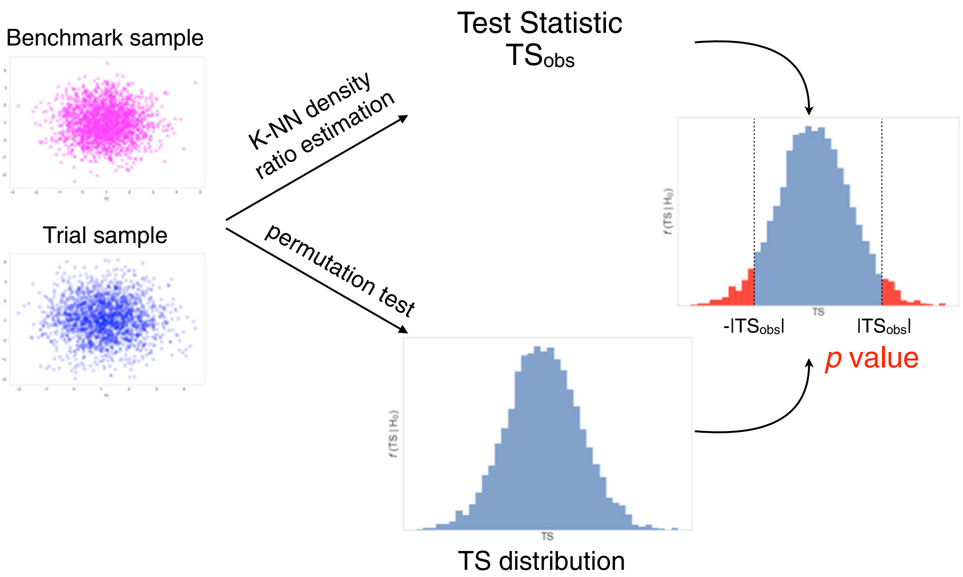
where $|\mathcal{T}| = N_T$ is the size of the trial sample. This test statistic will take values close to zero when H_0 is true, and far from zero (positively or negatively) when H_0 is false.

The test statistic defined in Eq. (2.5) is also equal to the estimated Kullback–Leibler (KL) divergence $\hat{D}_{\mathrm{KL}}(\hat{p}_T || \hat{p}_B)$ between the estimated PDFs of trial and benchmark samples, with the expectation value replaced by the empirical average [see Appendix A and in particular Eq. (5.2)]. The KL divergence plays a central role in information theory and can be interpreted as the relative entropy of a probability distribution with respect to another one. Our choice is also motivated by the fact that the log function in Eq. (2.5) makes the test statistic linearly sensitive to small differences between the distributions. Of course, other choices for the test statistic are possible, based on an estimated divergence between distributions other than the KL divergence, e.g. the Pearson squared-error divergence. The exploration of other possibilities is beyond the scope of this paper and is left for future work.

Ultimately, we want to conclude whether or not the null hypothesis can be rejected, with a specified significance level α (e.g. $\alpha = 0.05$), therefore we need to associate a p value to the null hypothesis, to be compared with α. To this end, we first need to estimate the PDFs $\hat{p}_{B,T}$ from the samples, then compute the test statistics $\mathrm{TS}_{\mathrm{obs}}$ observed on the two given samples. Next, in order to evaluate the probability associated with the observed value $\mathrm{TS}_{\mathrm{obs}}$ of the test statistic, we need to reconstruct its probability distribution $f(\mathrm{TS}|H_0)$ under the null hypothesis H_0, and finally compute a *two-sided* p value of the null hypothesis.

The distribution of the test statistic is expected to be symmetric around its mean (or median), which in general may not be exactly zero as a finite-sample effect. Therefore, the two-sided p value is simply double the one-sided p value.

Fig. 1 Schematic view of the proposed method to compute the *p* value of the null hypothesis that the two samples are drawn from the same probability density



A schematic summary of the method proposed in this paper is shown in Fig. 1.

In the remainder of this section we will describe this procedure in detail.

2.3 Probability density ratio estimator

We now turn to describing our approach to estimating the ratio of probability densities $\hat{p}_B/\hat{p}_T$ needed for the test statistic. There exist many possible ways to obtain density ratio estimators, e.g. using kernels [31] (see Ref. [32] for a comprehensive review). We choose to adopt a nearest-neighbors (NN) approach [33–39].

Let us fix an integer $K > 0$. For each point $\boldsymbol{x}_j \in \mathcal{T}$, one computes the Euclidean distance[1] $r_{j,T}$ to the Kth nearest neighbor of $\boldsymbol{x}_j$ in $\mathcal{T}\backslash\{\boldsymbol{x}_j\}$, and the Euclidean distance $r_{j,B}$ to the Kth nearest neighbor of $\boldsymbol{x}_j$ in $\mathcal{B}$. Since the probability density is proportional to the density of points, the probability estimates are simply given by the number of points (K, by construction) within a sphere of radius $r_{j,B}$ or $r_{j,T}$, divided by the volume of the sphere and the total number of available points. Therefore, the local nearest-neighbor estimates of the PDFs read

$$\hat{p}_B(\boldsymbol{x}_j) = \frac{K}{N_B}\frac{1}{\omega_D r_{j,B}^D}, \tag{2.6}$$

$$\hat{p}_T(\boldsymbol{x}_j) = \frac{K}{N_T - 1}\frac{1}{\omega_D r_{j,T}^D}, \tag{2.7}$$

(for any $\boldsymbol{x}_j \in \mathcal{T}$) where $\omega_D = \pi^{D/2}/\Gamma(D/2+1)$ is the volume of the unit sphere in $\mathbb{R}^D$. So, the test statistic defined in Eq. (2.5) is simply given by

$$\mathrm{TS}(\mathcal{B}, \mathcal{T}) = \frac{D}{N_T}\sum_{j=1}^{N_T}\log\frac{r_{j,B}}{r_{j,T}} + \log\frac{N_B}{N_T - 1}. \tag{2.8}$$

The value of the test statistic on the benchmark and trial samples will also be referred to as the ‘observed’ test statistic $\mathrm{TS}_{\mathrm{obs}}$. The NN density ratio estimator described above has been proved to be consistent and asymptotically unbiased [35,36,38], i.e. the test statistic TS (2.8) built from the estimated probability densities converges almost surely to the KL divergence between the true probability densities in the large sample limit $N_B, N_T \to \infty$.

Two advantages of the NN density ratio estimator are that it easily handles high-dimensional data, and its calculation is relatively fast, especially if k–d trees are employed to find the nearest neighbors. As a disadvantage, for finite sample sizes, the estimator (2.8) retains a small bias, although several methods has been proposed to reduce it (see e.g. Refs. [35,40]). Such a residual bias is only related to the asymptotic convergence properties of the test statistic to the estimated KL divergence $\hat{D}_{\mathrm{KL}}(\hat{p}_T||\hat{p}_B)$, and does not affect the outcome and the power of our test in any way.

The use of NN is also convenient as it allows the partition of the feature space not into rectangular bins, but into hyperspheres of varying radii, making sure they are all populated by data points.

The test statistic TS in Eq. (2.8), being an estimator of the KL divergence between the two underlying (unknown) PDFs, provides a measure of dataset compatibility. In the construction of TS we have chosen a particular K as the number of nearest neighbors. Of course, there is not an a priori optimal value of K to choose. In the following analyses we will use a particular choice of K, and we will comment on the possibility of extending the algorithm with adaptive K in Sect. 4.1.

[1] Other distance metrics may be used, e.g. a L^p-norm. We do not explore other possibilities here.

Now that we have a test statistic which correctly encodes the degree of compatibility between two data samples, and its asymptotic properties are ensured by theorems, we need to associate a probability with the value of the TS calculated on the given samples, as described in the next section.

2.4 Distribution of the test statistic and p value

In order to perform a hypothesis test, we need to know the distribution of the test statistic $f(\mathrm{TS}|H_0)$ under the null hypothesis H_0, to be used to compute the p value. Classical statistical tests have well-known distributions of the test statistics, e.g. normal, χ^2 or Student t. In our case, the distribution of TS is not theoretically known, for finite sample sizes. Therefore, it needs to be estimated from the data samples themselves. We employ the resampling method known as the permutation test [41,42] to construct the distribution $f(\mathrm{TS}|H_0)$ of the TS under the null hypothesis. It is a non-parametric (distribution-free) method based on the idea of sampling different relabellings of the data, under the assumption they are coming from the same parent PDF (null hypothesis).

In more detail, the permutation test is performed by first constructing a pool sample by merging the two samples: $\mathcal{U} = \mathcal{B} \cup \mathcal{T}$, then randomly shuffle (sampling without replacement) the elements of $\mathcal{U}$ and assign the first N_B elements to $\tilde{\mathcal{B}}$, and the remaining N_T elements to $\tilde{\mathcal{T}}$. Next, one computes the value of the test statistic on $\tilde{\mathcal{T}}$. If one repeats this procedure for every possible permutation (relabelling) of the sample points, one collects a large set of test statistic values under the null hypothesis which provides an accurate estimation of its distribution (exact permutation test). However, it is often impractical to work out all possible permutations, so one typically resorts to perform a smaller number N_{perm} of permutations, which is known as an approximate (or Monte-Carlo) permutation test. The TS distribution is then reconstructed from the N_{perm} values of the test statistic obtained by the procedure outlined above.

The distribution of the test statistic under a permutation test is asymptotically normal with zero mean in the large sample limit $N_B, N_T \to \infty$ [42], as a consequence of the Central Limit Theorem. Furthermore, when the number N_{perm} is large, the distribution of the p value estimator approximately follows a normal distribution with mean p and variance $p(1-p)/N_{\mathrm{perm}}$ [41,43]. For example, if we want to know the p value in the neighborhood of the significance level α to better than $\alpha/3$, we need $N_{\mathrm{perm}} > 9(1-\alpha)/\alpha$, which is of the order of 1000 for $\alpha = 0.01$.

Once the distribution of the test statistic is reconstructed, it is possible to define the critical region for rejecting the null hypothesis at a given significance α, defined by large enough values of $\mathrm{TS}_{\mathrm{obs}}$ such that the corresponding p value is smaller than α.

As anticipated in Sect. 2.2, for finite samples the test statistic distribution is still approximately symmetric around the mean, but the latter may deviate from zero. In order to account for this general case, and give some intuitive meaning to the size of the test statistic, it is convenient to standardize (or 'studentize') the TS to have zero mean and unit variance. Let $\hat{\mu}, \hat{\sigma}$ be the mean and the variance of test statistic under the distribution $f(\mathrm{TS}|H_0)$. We then transform the test statistic as

$$\mathrm{TS} \to \mathrm{TS}' \equiv \frac{\mathrm{TS} - \hat{\mu}}{\hat{\sigma}}, \tag{2.9}$$

which is distributed according to

$$f'(\mathrm{TS}'|H_0) = \hat{\sigma} f(\hat{\mu} + \hat{\sigma}\mathrm{TS}'|H_0), \tag{2.10}$$

with zero mean and unit variance. With this redefinition, the two-sided p value can be easily computed as

$$p = 2 \int_{|\mathrm{TS}'_{\mathrm{obs}}|}^{+\infty} f'(\mathrm{TS}'|H_0) d\mathrm{TS}'. \tag{2.11}$$

2.5 Summary of the algorithm

The pseudo-code of the algorithm for the statistical test presented in this paper is summarized in Table 1. We implemented it in Python and an open-source package is available on GitHub.[2]

2.6 Extending the test to include uncertainties

So far we have assumed that both $\mathcal{B}$ and $\mathcal{T}$ samples are precisely known. However, in several situations of physical interest this may not be the case, as the features may be known only with some uncertainty, e.g. when the sample points come from physical measurements. There can be several factors affecting the precision with which each sample point is known, for instance systematic uncertainties (e.g. the smearing effects of the detector response) and the limited accuracy of the background (Monte-Carlo simulation), which may be particularly poor in some regions of the feature space.

Of course, once such uncertainties are properly taken into account, we expect a degradation of the results of the statistical test described in the previous sections, leading to weaker conclusions about the rejection of the null hypothesis.

Here we describe a simple and straightforward extension of the method described in this section, to account for uncertainties in the positions of the sample points. We consider the test statistic itself as a random variable, which is a sum of the test statistic TS defined in Sect. 2.2, and computed on the original $\mathcal{B}$, $\mathcal{T}$ samples, and an uncertainty fluctuation (noise) U, originating when each point of $\mathcal{B}$ (or $\mathcal{T}$ or both) is shifted

[2] https://github.com/de-simone/NN2ST.

Table 1 Pseudo-code for the two-sample test algorithm, using nearest neighbors density ratio estimation

```
Algorithm 1 Nearest-Neighbors Two-Sample Test
Require: Benchmark sample: $\mathcal{B} = \{\boldsymbol{x}'_i | \boldsymbol{x}'_i \in \mathbb{R}^D\}_{i=1}^{N_B}$, Trial sample: $\mathcal{T} = \{\boldsymbol{x}_j | \boldsymbol{x}_j \in \mathbb{R}^D\}_{j=1}^{N_T}$
Input: $K, N_{\text{perm}} \in \mathbb{N} \setminus \{0\}$.
Output: $p$-value of the null hypothesis.

for $j = 1$ to $N_T$ do
    $r_{j,B} \leftarrow$ distance of $K$th-NN in $\mathcal{B}$ from $\boldsymbol{x}_j \in \mathcal{T}$
    $r_{j,T} \leftarrow$ distance of $K$th-NN in $\mathcal{T}$ from $\boldsymbol{x}_j \in \mathcal{T}$
end for
$\text{TS}_{\text{obs}} \leftarrow \frac{D}{N_T} \sum_{j=1}^{N_T} \log \frac{r_{j,B}}{r_{j,T}} + \log \frac{N_B}{N_T - 1}$ {observed value of test statistic}

for $n = 1$ to $N_{\text{perm}}$ do {permutation test}
    $\mathcal{U}_n \leftarrow$ randomly reshuffle $\mathcal{B} \cup \mathcal{T}$
    $\tilde{\mathcal{B}} \leftarrow$ first $N_B$ elements of $\mathcal{U}_n$
    $\tilde{\mathcal{T}} \leftarrow$ remaining $N_T$ elements of $\mathcal{U}_n$
    for $j = 1$ to $N_T$ do
        $\tilde{r}_{j,B} \leftarrow$ distance of $K$th-NN in $\tilde{\mathcal{B}}$ from $\tilde{\boldsymbol{x}}_j \in \tilde{\mathcal{T}}$
        $\tilde{r}_{j,T} \leftarrow$ distance of $K$th-NN in $\tilde{\mathcal{T}}$ from $\tilde{\boldsymbol{x}}_j \in \tilde{\mathcal{T}}$
    end for
    $\text{TS}_n \leftarrow \frac{D}{N_T} \sum_{j=1}^{N_T} \log \frac{\tilde{r}_{j,B}}{\tilde{r}_{j,T}} + \log \frac{N_B}{N_T - 1}$ {test statistic on permutation $n$}
end for

$f(\text{TS}|H_0) \leftarrow \{\text{TS}_n\}$ {probability distribution of TS under $H_0$}
$\hat{\mu}, \hat{\sigma}^2 \leftarrow$ mean and variance of TS under $f$
$\text{TS}' \leftarrow (\text{TS} - \hat{\mu})/\hat{\sigma}$
$f'(\text{TS}'|H_0) \leftarrow \hat{\sigma} f(\hat{\mu} + \hat{\sigma}\text{TS}'|H_0)$ {probability distribution of TS$'$ under $H_0$}
$p \leftarrow 2 \int_{|\text{TS}'_{\text{obs}}|}^{+\infty} f'(\text{TS}'|H_0) d\text{TS}'$
```

by a random vector: $\text{TS}_u = \text{TS} + U$. The trial and benchmark samples with uncertainties are then given by

$$\mathcal{T}_u = \{\boldsymbol{x}_i + \Delta\mathbf{x}_i\}_{i=1}^{N_T}, \tag{2.12}$$

$$\mathcal{B}_u = \{\boldsymbol{x}'_i + \Delta\mathbf{x}'_i\}_{i=1}^{N_B}, \tag{2.13}$$

which represent a point-wise random shift, where the error samples $\Delta\mathbf{x}_i, \Delta\mathbf{x}'_i \in \mathbf{R}^D$ are independent random variables drawn from the same distribution, according to the expected (or presumed) distribution of uncertainties in the features, e.g. zero-mean multivariate Gaussians.

Next, one can compute the test statistic on the 'shifted' samples as

$$\text{TS}_u \equiv \text{TS}(\mathcal{B}_u, \mathcal{T}_u) = \text{TS}(\mathcal{B}, \mathcal{T}) + U. \tag{2.14}$$

Since the TS computed on the original $\mathcal{B}, \mathcal{T}$ samples is given by the observed value TS_{obs}, the value of U for any random samplings of the error samples is simply $U = \text{TS}(\mathcal{B}_u, \mathcal{T}_u) - \text{TS}_{\text{obs}}$. By repeating the calculation of U many (N_{iter}) times, each time adding a random noise to $\mathcal{B}$ (or $\mathcal{T}$ or both) we can reconstruct its probability distribution $f(U)$, which is asymptotically normal with zero mean in the large-sample limit $N_B, N_T \to \infty$.

The resulting distribution of the test statistic TS_u, being the sum of two i.i.d. random variables, is then given by the convolution of the distribution $f(\text{TS}|H_0)$, computed via permutation test on $\mathcal{B}, \mathcal{T}$, and the distribution $f(U)$ with mean set to zero. This is motivated by the desire to eliminate the bias in the mean of the distribution of U coming from finite-sample effects. As a result of this procedure, the distribution $f(\text{TS}_u|H_0)$ will have the same mean as $f(\text{TS}|H_0)$ but a larger variance.

The p value of the test is computed from TS_{obs} with the same steps as described in Sect. 2.4, but with the distribution of the test statistic with uncertainties given by $f(\text{TS}_u|H_0)$, rather than $f(\text{TS}|H_0)$. Since $f(\text{TS}_u)$ has larger variance than $f(\text{TS})$, the p value will turn out to be larger, therefore the equivalent significance Z will be smaller. This conclusion agrees with the expectation that the inclusion of uncertainties leads to a degradation of the power of the test.

The summary of the algorithm to compute the distribution $f(U)$ can be found in Table 2. Once $f(U)$ is computed, it needs to be convolved with $f(\text{TS}|H_0)$, which was previ-

Table 2 Pseudo-code for the algorithm to find the distribution $f(U)$ of the test statistic noise U

```
Algorithm 2 Distribution of the test statistic noise
Require: Benchmark sample: $\mathcal{B} = \{\boldsymbol{x}'_i | \boldsymbol{x}'_i \in \mathbb{R}^D\}_{i=1}^{N_\mathcal{B}}$, Trial sample: $\mathcal{T} = \{\boldsymbol{x}_j | \boldsymbol{x}_j \in \mathbb{R}^D\}_{j=1}^{N_\mathcal{T}}$
Input: $K, N_{\text{iter}} \in \mathbb{N} \setminus \{0\}$
Input: $F_\mathcal{B}(\mathbf{x}), F_\mathcal{T}(\mathbf{x})$: distributions of feature uncertainties for $\mathcal{B}, \mathcal{T}$ samples
Output: $f(U)$: distribution of the test statistic noise $U$

TS_obs ← TS($\mathcal{B}, \mathcal{T}$) {observed value of test statistic}
for $j = 1$ to $N_{\text{iter}}$ do
    $\mathcal{E}_\mathcal{T} = \{\Delta\mathbf{x}_i\}_{i=1}^{N_\mathcal{T}}$ randomly drawn from $F_\mathcal{T}(\mathbf{x})$
    $\mathcal{E}_\mathcal{B} = \{\Delta\mathbf{x}'_i\}_{i=1}^{N_\mathcal{B}}$ randomly drawn from $F_\mathcal{B}(\mathbf{x})$
    $\mathcal{T}_u \leftarrow \mathcal{T} + \mathcal{E}_\mathcal{T}$ {point-wise sum}
    $\mathcal{B}_u \leftarrow \mathcal{B} + \mathcal{E}_\mathcal{B}$ {point-wise sum}
    $\text{TS}_u \leftarrow \text{TS}(\mathcal{B}_u, \mathcal{T}_u)$
    $U_j \leftarrow \text{TS}_u - \text{TS}_{\text{obs}}$
end for
$f(U) \leftarrow \{U_j\}$ {distribution of $U$}
```

ously found via permutation test, as described in Sect. 2.4, to provide the distribution of the test statistic with uncertainties needed to compute the p value.

3 Applications to simulated data

3.1 Case study: Gaussian samples

As a first case study of our method let us suppose we know the original distributions from which the benchmark and trial samples are randomly drawn. For instance, let us consider the multivariate Gaussian distributions of dimension D defined by mean vectors $\boldsymbol{\mu}_{\mathcal{B},\mathcal{T}}$ and covariance matrices $\Sigma_{\mathcal{B},\mathcal{T}}$:

$$p_\mathcal{B} = \mathcal{N}(\boldsymbol{\mu}_\mathcal{B}, \Sigma_\mathcal{B}), \quad p_\mathcal{T} = \mathcal{N}(\boldsymbol{\mu}_\mathcal{T}, \Sigma_\mathcal{T}). \tag{3.1}$$

In this case, the KL divergence can be computed analytically [see Eq. (5.4)]. In the large sample limit, we recover that the test statistic converges to the true KL divergence between the PDFs (see Fig. 2 and Appendix A). Of course, the comparison is possible because we knew the parent PDFs $p_\mathcal{B}$, $p_\mathcal{T}$.

For our numerical experiments we fix the benchmark $\mathcal{B}$ sample by the parameters $\boldsymbol{\mu}_\mathcal{B} = 1_D$, $\Sigma_\mathcal{B} = \boldsymbol{I}_D$, and we construct four different trial samples $\mathcal{T}_{G0}$, $\mathcal{T}_{G1}$, $\mathcal{T}_{G2}$, $\mathcal{T}_{G3}$ drawn by Gaussian distributions whose parameters are defined in Table 3. Each sample consists of 20,000 points randomly drawn from the Gaussian distributions defined above. Notice that the first trial sample $\mathcal{T}_{G0}$ is drawn from the same distribution as the benchmark sample.

As is customary, we associate an equivalent Gaussian significance Z to a given (two-sided) p value as: $Z \equiv \Phi^{-1}(1 - p/2)$, where Φ is the cumulative distribution of a standard (zero-mean, unit-variance) one-dimensional Gaus-

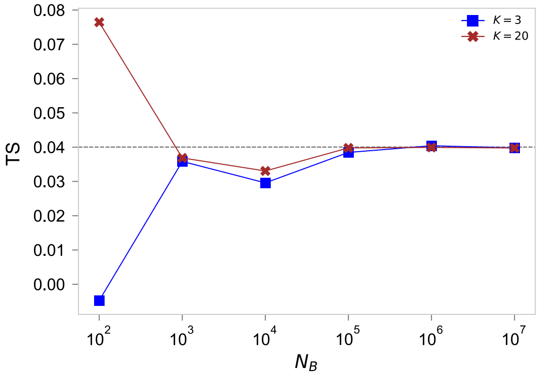


Fig. 2 Convergence of the test statistic to the exact KL divergence (dashed horizontal line) between two 2-dimensional Gaussian distributions, in the large-sample limit. The $\mathcal{B}$, $\mathcal{T}$ samples have the same size $N_\mathcal{B} = N_\mathcal{T}$, and they are sampled from 2-dimensional Gaussian distributions with $\boldsymbol{\mu}_\mathcal{B} = 1.0_2$, $\boldsymbol{\mu}_\mathcal{T} = 1.2_2$, $\Sigma_\mathcal{B} = \Sigma_\mathcal{T} = \boldsymbol{I}_2$. Two different choices for the number of nearest neighbors are shown: $K = 3$ (blue squares) and $K = 20$ (red crosses)

Table 3 Definition of the Gaussian datasets used for the numerical experiments. Each sample consists of $N_\mathcal{B} = N_\mathcal{T} = 20{,}000$ points randomly drawn from D-dimensional Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Dataset	$\boldsymbol{\mu}$	Σ
$\mathcal{B}$	1_D	$\boldsymbol{I}_D$
$\mathcal{T}_{G0}$	1_D	$\boldsymbol{I}_D$
$\mathcal{T}_{G1}$	1.12_D	$\boldsymbol{I}_D$
$\mathcal{T}_{G2}$	1_D	$\left(\begin{array}{cc\|c} 0.95 & 0.1 & \multirow{2}{*}{$\mathbf{0}$} \\ 0.1 & 0.8 & \\ \hline \multicolumn{2}{c\|}{\mathbf{0}} & \boldsymbol{I}_{D-2} \end{array}\right)$
$\mathcal{T}_{G3}$	1.15_D	$\boldsymbol{I}_D$

Table 4 Summary of the results comparing $\mathcal{B}$ with 4 trial samples, for different dimensionality D. The samples are defined in Table 3. We set $K = 5$ and $N_{\text{perm}} = 1000$

Trial Dataset	$D = 2$		$D = 5$		$D = 10$	
	p value	Z	p value	Z	p value	Z
$\mathcal{T}_{G0}$	8.2×10^{-1}	$0.2\,\sigma$	6.9×10^{-1}	$0.4\,\sigma$	6.9×10^{-1}	$0.4\,\sigma$
$\mathcal{T}_{G1}$	2.8×10^{-2}	$2.2\,\sigma$	1.5×10^{-7}	$5.2\,\sigma$	3.6×10^{-13}	$7.3\,\sigma$
$\mathcal{T}_{G2}$	4.0×10^{-4}	$3.5\,\sigma$	8.8×10^{-8}	$5.3\,\sigma$	9.4×10^{-9}	$5.7\,\sigma$
$\mathcal{T}_{G3}$	1.2×10^{-6}	$4.9\,\sigma$	1.4×10^{-19}	$9.1\,\sigma$	1.9×10^{-30}	$11.5\,\sigma$

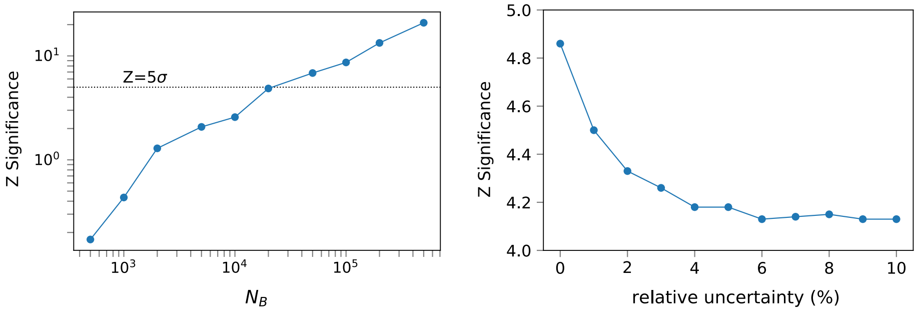


Fig. 3 We compare $\mathcal{B} = \mathcal{T}_{G0}$ and $\mathcal{T} = \mathcal{T}_{G3}$, as defined in Table 3, with $D = 2$, using $K = 5$ and $N_{\text{perm}} = 1000$. The $\mathcal{B}$,$\mathcal{T}$ samples have the same size $N_B = N_T$. Left panel: the Z significance of the test for different sample sizes. Right panel: the Z significance for different relative uncertainties added to $\mathcal{B}$ only, with fixed $N_B = N_T = 20{,}000$. The U distribution has been computed with $N_{\text{iter}} = 1000$ random samplings from the distribution of feature uncertainties

sian distribution. In Table 4 we show the p values and the corresponding Z significance of the statistical tests for different dimensions D. The results are interpreted as follows. For $D = 2$, the first two trial samples $\mathcal{T}_{G0}, \mathcal{T}_{G1}$ are not distinguished from the benchmark $\mathcal{B}$ at more than 99% CL ($p > 0.01$), while $\mathcal{T}_{G2}, \mathcal{T}_{G3}$ are distinguished ($p \leq 0.01$, or equivalently $Z \geq 2.6\sigma$). Therefore, one would reject the null hypothesis at more than 99% CL and conclude that the PDFs from which $\mathcal{T}_{G2}, \mathcal{T}_{G3}$ are drawn are different from the benchmark PDF p_B. It is remarkable that our statistical test is able to reject the null hypothesis with a large significance of 4.9σ for two random samples $\mathcal{B}, \mathcal{T}_{G3}$ drawn from 2-dimensional distributions which only differ by a shift of the mean by 15% along each dimension. For higher dimensionality of the data, the discriminating power of the test increases, and the null hypothesis is rejected at more than 5σ significance for all trial samples $\mathcal{T}_{G1}, \mathcal{T}_{G2}, \mathcal{T}_{G3}$. The running time to compute the p value on a standard laptop for two 2-dimensional samples of 20,000 points each, and for 1000 permutations, was about 2 minutes. The running time scales linearly with the number of permutations.

The number of sample points ($N_{B,T}$) plays an important role. As an example, we sampled the same datasets $\mathcal{B}, \mathcal{T}_{G0}, \mathcal{T}_{G1}, \mathcal{T}_{G2}, \mathcal{T}_{G3}$ with $N_B = N_T = 2000$ points, i.e. ten times less points than for the cases shown in Table 4. The results for the equivalent significance for $\mathcal{T}_{G0}, \mathcal{T}_{G1}, \mathcal{T}_{G2}, \mathcal{T}_{G3}$ with $D = 2$ are $Z = 1.4\sigma, 1.9\sigma, 1.9\sigma, 2.3\sigma$, respectively. Clearly, the test is not able to reject the null hypothesis at more than 99% CL (the p value is never below 0.01, or equivalently $Z < 2.6\sigma$) in none of the cases. As another illustration of this point, we run the statistical test for $\mathcal{B} = \mathcal{T}_{G0}$ vs $\mathcal{T} = \mathcal{T}_{G3}$ for $D = 2$ and different sample sizes $N_B = N_T$, and show the resulting Z significance in Fig. 3 (left panel). We find that for $N_B \leq 10^4$, the test is not able to reject the null hypothesis at more than 99% CL. Therefore, the power of our statistical test increases for larger sample sizes, as expected since bigger samples lead to more accurate approximations of the original PDFs.

We have also studied the power performance of our statistical test with respect to parametric competitors. We ran 200 tests of two samples drawn from multivariate Gaussian distributions with $D = 1, 2, 5$, with sample sizes $N_B = N_T = 100$, and computed the approximated power as the fraction of runs where the null hypothesis is rejected with significance level 5% ($p < 0.05$). We considered normal location alternatives, with $\mathcal{B} \sim \mathcal{N}(0_D, \boldsymbol{I}_D)$ and $\mathcal{T} \sim \mathcal{N}(\Delta_D, \boldsymbol{I}_D)$, where Δ varies from 0.05 to 1.0. As competitor, we choose the Student's t test (or its generalization Hotelling's T^2-test, for $D > 1$). We find that our test shows a power comparable to its competitor, in some cases lower than that by at most a factor of 3, which is satisfactory given that the T^2-test is parametric and designed to spot location differences.

Next, we run the statistical test by including uncertainties, as described in Sect. 2.6. For the uncertainties, we assume uncorrelated Gaussian noise, so the covariance matrix of the uncertainties is a D-dimensional diagonal matrix $\text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$ where each eigenvalue is proportional to the relative uncertainty ϵ of the component x_i of the sample point $\mathbf{x}$: $\sigma_i = \epsilon x_i$.

In Fig. 3 (right panel) we show how the significance of rejecting the null hypothesis degrades once uncorrelated relative uncertainties are added to the $\mathcal{B}$ sample. For $D = 2$, the initial $4.9\,\sigma$ when comparing $\mathcal{B} = \mathcal{T}_{G0}$ and $\mathcal{T} = \mathcal{T}_{G3}$ without noise goes down to about $4.1\,\sigma$ with 10% relative error.

3.2 Case study: Monojet searches at LHC

A model-independent search at the LHC for physics Beyond the Standard Model (BSM), such as Dark Matter (DM), has been elusive [3–6]. Typically it is necessary to simulate the theoretical signal in a specific model, and compare with data to test whether the model is excluded. The signal-space for DM and BSM physics in general is enormous, and despite thorough efforts, the possibility exists that a signal has been overlooked. The compatibility test described in Sect. 2 is a promising technique to overcome this challenge, as it can search for deviations between the expected simulated Standard Model signal and the true data, without any knowledge of the nature of the new physics.

In a real application of our technique by experimental collaborations, the benchmark dataset $\mathcal{B}$ will be a simulation of the SM background, while the trial dataset $\mathcal{T}$ will consist of real measured data, potentially containing an unknown mix of SM and BSM events. As a proof-of-principle, we test whether our method would be sensitive to a DM signature in the monojet channel. For our study, both $\mathcal{B}$ and $\mathcal{T}$ will consist of simulated SM events (‘background’), however $\mathcal{T}$ will additionally contain injected DM events (‘signal’). The goal is to determine whether the algorithm is sensitive to differences in $\mathcal{B}$ and $\mathcal{T}$ caused by this signal.

Model and simulations

The signal comes from a standard simplified DM model (see e.g. Ref. [44] for a review) with Fermion DM χ and an s-channel vector Z' mediator [45,46]. Our benchmark parameters are $g_\chi = 1$, $g_q = 0.1$, $g_\ell = 0.01$, in order to match the simplified model constraints from the ATLAS summary plots [47]. We use a DM mass of 100 GeV, and mediator masses of (1200, 2000, 3000) GeV, in order to choose points that are not yet excluded but could potentially be in the future [47].

Signal and background events are first simulated using MG5_aMC@NLO v2.6.1 [48] at center-of-mass energy $\sqrt{s} = 13$ TeV, with a minimal cut of $E_T^{\text{miss}} > 90$ GeV, to emulate trigger rather than analysis cuts. We use Pythia 8.230 [49] for hadronization and Delphes 3.4.1 [50] for detector simulation. The so-called ‘monojet’ signal consists of events with missing energy from DM and at least one high-p_T jet. The resulting signal cross-section is $\sigma_{\text{signal}} = (20.4,\ 3.8,\ 0.6)$ pb for $M_{\text{med}} = (1200,\ 2000,\ 3000)$ GeV respectively. For the background samples, we simulate 40,000 events of the leading background, $Z \to \nu\bar{\nu} + nj$ where n is 1 or 2, resulting in a cross section of $\sigma_{\text{background}} = 202.6$ pb.

The Delphes ROOT file is converted to LHCO and a feature vector is extracted with Python for each event, consisting of p_T and η for the two leading jets; the number of jets; missing energy E_T^{miss}; Hadronic energy H_T; and $\Delta\phi$ between the leading jet and the missing energy. Together this gives an 8-dimensional feature vector ($D = 8$), which is scaled to zero-mean unit-variance based on the mean and variance of the background simulations. This feature vector is chosen to capture sufficient information about each event while keep running time of the algorithm reasonable. Other choices of the feature vector could be chosen to capture different aspects of the physical processes, including higher- or lower-level features, such as raw particle 4-vectors. Application of high-performance computing resources would allow the feature vector to be enlarged, potentially strengthening results. A full study of the choice of feature vector is left to future work. Our simulation technique is simple and designed only as a proof of principle; we do not include sub-leading SM backgrounds, nor full detector effects, adopting a generic Delphes profile.

Test statistic distribution under null hypothesis

Following the technique described in Sect. 2, for each of the 3 considered points in signal model parameter space, we first construct an empirical distribution of the test statistic under the null hypothesis, $f(\text{TS}|H_0)$, and we then measure TS_{obs} and compute the p value to determine the compatibility of the datasets. We choose $K = 5$ and $f(\text{TS}|H_0)$ is constructed over $N_{\text{perm}} = 3000$.

The pool sample $\mathcal{B} \cup \mathcal{T}$ consists of the 40,000 background events, along with a number of signal events proportional to the signal cross-section. We define $\mathcal{B}$ and $\mathcal{T}$ as having an equal number of background events, so that $N_{\text{signal}} = 20{,}000 \times \sigma_{\text{signal}}/\sigma_{\text{background}}$, $N_{\mathcal{T}} = 20{,}000 + N_{\text{signal}}$. The resulting distribution of TS under the null hypothesis is shown in Fig. 4. The simulations are relatively fast, taking approximately an hour per 1000 permutations on a standard laptop, although computation time grows as a power-law with the number of events, such that further optimization and high-performance computing resources will be a necessity for application to real LHC data with many thousands of events. The statistics of $f(\text{TS}|H_0)$ converge quickly, as shown in Fig. 5, consistent with the discussion of N_{perm} in Sect. 2.4, and showing that N_{perm} is more than sufficient.

Note that since $\tilde{\mathcal{B}}$, $\tilde{\mathcal{T}}$ are chosen from permutations of $\mathcal{B} \cup \mathcal{T}$, it is not necessary to specify how the 40,000 background

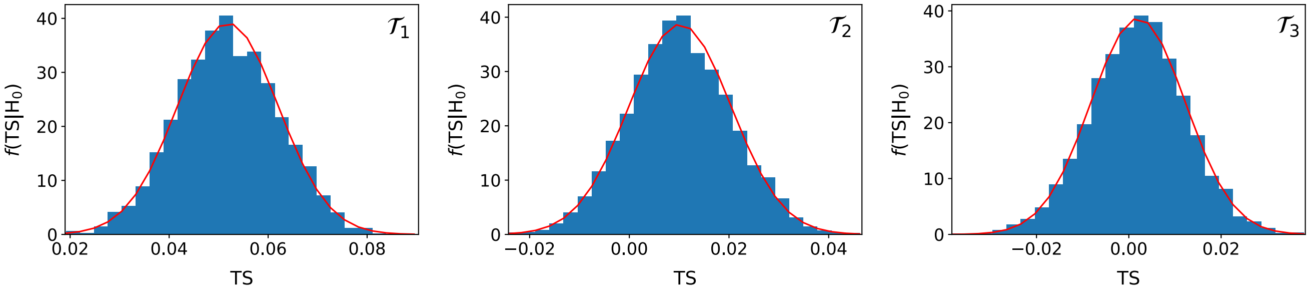


Fig. 4 Distribution of the test statistic under the null hypothesis for our 3 signal points. Overlayed is a Gaussian distribution with the same mean and standard deviation as the data

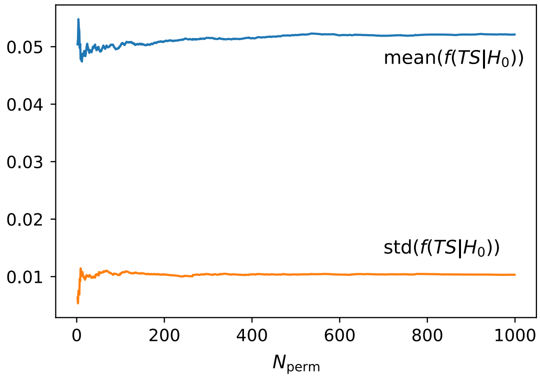


Fig. 5 Effect of N_{perm} on the null-hypothesis test statistic for the mono-jet study with $\mathcal{T}_2$

events are divided between $\mathcal{B}$ and $\mathcal{T}$; It is only necessary to specify N_B and N_T at this point.

Observed test statistic

To test whether the null hypothesis would be excluded in the event of an (otherwise unobserved) DM signal hiding in the data, we calculate TS_{obs} using $\mathcal{B}$ containing only background, and $\mathcal{T}$ containing background plus a number of signal events proportional to the relative cross section. In a practical application of this technique by the experimental collaborations, $\mathcal{B}$ would instead correspond to background simulations, while $\mathcal{T}$ would be the real-world observation; therefore only one measurement of TS_{obs} would be performed.

However, in our case the distribution of TS under the null hypothesis is insensitive to the way the 40,000 background events are divided between $\mathcal{B}$ and $\mathcal{T}$. Therefore we can simulate multiple real-world measurements of TS_{obs} by dividing the 40,000 background events between $\mathcal{B}$ and $\mathcal{T}$ in different permutations (always keeping 20,000 background events in each sample). This allows us to be more robust: since TS_{obs} is itself a random variable, multiple measurements of TS_{obs} allows us to avoid the claim of a small p value, when in reality the algorithm may not be sensitive to a small signal.

The calculation of TS_{obs} is performed for 100 random divisions. The p value and significance Z of each TS_{obs} are calculated with respect to the empirical distribution $f(\text{TS}|H_0)$ where possible. In many cases, TS_{obs} is so extreme that it falls outside the measured range of $f(\text{TS}|H_0)$, in which case p and Z are determined from a Gaussian distribution with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$. This is equivalent to assuming that $f(\text{TS}|H_0)$ is well-approximated by a Gaussian, which is true to a good approximation, as seen in Fig. 4. To be conservative, the technique is only considered sensitive to the signal if all simulated observations of TS exclude the null hypothesis, i.e. we show the minimum Z significance (and maximum p value). These results are shown in Table 5, where we see that the background-only hypothesis is strongly excluded for $\mathcal{T}_1$ and $\mathcal{T}_2$, even though these points are not yet excluded by traditional LHC searches. Bear in mind that this is a proof-of-concept, and real-world results are unlikely to be as clean, as discussed in Sect. 3.3.

Inclusion of uncertainties

To test the sensitivity of this technique to uncertainties and errors in the background simulation, we use the method outlined in Sect. 2.6 to estimate the drop in significance when uncertainties are taken into account. Uncorrelated Gaussian noise with $\epsilon = 10\%$ (as defined in Sect. 3.1) is added to $\mathcal{B}$, allowing the construction of $f(\text{TS}_u|H_0)$ using $N_{\text{iter}} = 1000$. Note that while the primary result without uncertainties is agnostic as to how the overall background sample is divided between $\mathcal{B}$ and $\mathcal{T}$, this is not the case when applying uncertainties. We construct $f(\text{TS}_u|H_0)$ by repeatedly applying different noise to the same $\mathcal{B}$, and so $\mathcal{B}$ and $\mathcal{T}$ must be defined from the outset, leaving just one measurement of TS_{obs}, for a random draw of $\mathcal{B}$ and the background component of $\mathcal{T}$ from the overall pool of background simulations. For $(\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3)$, we find that without noise $Z = (40, 13, 2.7)$. Note that as expected, these are larger than the minimum values over 100 observations reported in Table 5. With $\epsilon = 10\%$, we find that this reduces to $Z = (26, 12, 2.5)$ for the 3 samples, respectively. This is in line with expectations: while this is a powerful technique, limited knowledge of the expected background will degrade the results. With this in mind, we reiterate that

Table 5 Summary of monojet results comparing $\mathcal{B}$ (background only) with $\mathcal{T}$ (background plus DM signal). The cross section corresponding to the trial sample is simply given by $\sigma_{\mathcal{T}} = \sigma_{\text{background}} + \sigma_{\text{signal}}$. The p value and Z statistic show the compatibility between $\mathcal{B}$ and $\mathcal{T}$; Large Z indicates that $\mathcal{T}$ is not consistent with the background-only hypothesis. Note that these results will be weakened by application of uncertainties (see text for details)

Sample	M_{med}	$\sigma_{\mathcal{T}}$ [pb]	σ_{signal} [pb]	max (p value)	min (Z)
$\mathcal{T}_1$	1.2 TeV	223.0	20.4	$< 10^{-50}$	$> 15\,\sigma$
$\mathcal{T}_2$	2 TeV	206.4	3.8	5.7×10^{-25}	$10\,\sigma$
$\mathcal{T}_3$	3 TeV	203.2	0.6	0.90	$0.13\,\sigma$

results based on simulations alone should be taken with a grain of salt. They show the strengths of the statistical test we are proposing and prove it is worthwhile to investigate it further, but they will be weakened in a real-world situation.

As an application to experimental data, our technique could be applied by seeding the simulated background $\mathcal{B}$ with noise associated with uncertainties in the Monte-Carlo background estimation, or seeding the measured data sample $\mathcal{T}$ with noise associated with systematic uncertainties.

Discussion

To study the threshold to which this technique is sensitive, we can construct $\mathcal{T}$ by adding an arbitrary number of signal events to the background, without reference to the relative signal cross-section. The result is shown in Fig. 6 (left panel), using the signal dataset with $M_{\text{med}} = 2$ TeV. For each value of N_{sig}, the distribution $f(\text{TS}|H_0)$ is constructed over 1000 permutations, and the Z significance is determined through taking the minimum value of Z over 100 measurements of TS_{obs} for different background permutations. There is a clear threshold, below which the significance is negligible and constant, and above which the significance grows as a power-law. The number of signal events in $\mathcal{T}_2$ crosses this threshold while $\mathcal{T}_3$ does not, explaining the rapid drop in the significance.

The strength of the technique is also sensitive to the number of samples. Figure 6 (right panel) demonstrates this, again using the signal dataset with $M_{\text{med}} = 2$ TeV, $N_{\text{perm}} = 1000$, and taking the minimum Z over 100 measurements of $\mathcal{T}_{\text{obs}}$. It shows an approximately power-law growth in the significance, consistent with the same growth in the significance with number of signal events. Clearly, the more data the better.

3.3 Future application to real data

In a practical application of this technique by experimental collaborations, $\mathcal{B}$ would correspond to simulations of the SM background, while $\mathcal{T}$ would be the real-world observation, consisting of an unknown mix of signal and background events. Both $\mathcal{B}$ and $\mathcal{T}$ could be constructed under the same set of minimal cuts, imposed based on trigger requirements rather than as a guide to finding new physics. While the technique itself is model-independent, there is freedom to apply physical knowledge in the choice of minimal cuts to keep the background simulation and data load manageable, and in the choice of feature vector, which can either be low-level (raw 4-vectors of reconstructed objects, or even pixel hits) or high-level (missing energy, hadronic energy etc.).

Even though we have only applied our method to a generic monojet signal, the strength of the algorithm is that it is sensitive to unspecified signals, and is limited only by the accuracy of the background simulation. We emphasize that our case study in Sect. 3.2 is a proof of concept with a generic signal and a naïve estimation of the background.

Accurately estimating SM backgrounds at the LHC is a significant challenge in the field and must be considered carefully in any future application of this technique. Currently used techniques of matching simulations to data in control regions still allow the use of our method, although this introduces some model-dependent assumptions. Alternatively, one may apply our statistical test in the context of data-driven background calculation, as a validation tool to measure the compatibility of Monte-Carlo simulations with data in control regions.

For instance, it is common practice to tune the nuisance parameters in order to make the Monte-Carlo simulation of the background match the data in control regions. When one deals with more than one control region, this procedure results in a collection of patches of the feature space, in each of which the background simulation is fit to the data. The statistical test we propose in this paper can be used to determine to what extent (significance) the background simulation is representative of the data at the global level, in all control regions. And in case of discrepancies, it can pinpoint the regions of feature space where the mismatch between data and simulations is the largest.

As we have shown by implementing sample uncertainties in our statistical test, the test alone may not be sufficient to claim discovery in cases where background simulations are not sufficiently accurate, but this does not weaken the value of the method. It remains valuable as a tool to identify regions

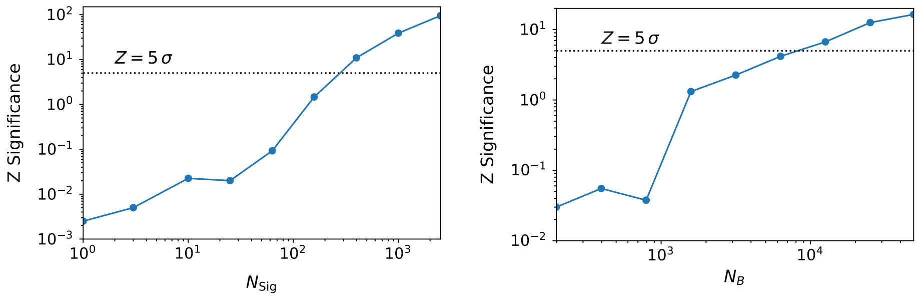


Fig. 6 The effect of N_{sig} (left) and N_B (right) on the ability of the algorithm to distinguish $\mathcal{B}$ and $\mathcal{T}$. For the left figure, $N_B = 20{,}000$ background events and $N_T = N_B + N_{\mathrm{sig}}$. (Based on the actual simulated signal and background cross-sections, the true value is $N_{\mathrm{sig}} = 375$.) In the right figure, $N_T = N_B + N_{\mathrm{sig}}$, where N_{sig} varies in proportion to N_B and the relative signal/background cross-sections. In both cases, we use the trial sample $\mathcal{T}_2$ corresponding to the signal with $M_{\mathrm{med}} = 2$ TeV

of excess in a model-independent way, allowing follow-up hand-crafted analyses of potential signal regions.

4 Directions for extensions

In this section we summarize two main directions to extend and improve the method proposed in this paper. We limit ourselves to just outlining some ideas, leaving a more complete analysis of each of these issues to future work.

4.1 Adaptive choice of the number of nearest neighbors

The procedure for the density ratio estimator described in Sect. 2.3 relies on choosing the number K of NN. As mentioned earlier, it is also possible to make the algorithm completely unsupervised by letting it choose the optimal value of K.

One approach is to proceed by model selection as in Refs. [31,39,51]. We define the loss function as a mean-squared error between the true (unknown) density ratio $r(\boldsymbol{x}) = p_T(\boldsymbol{x})/p_B(\boldsymbol{x})$ and the estimated density ratio $\hat{r}(\boldsymbol{x}) = \hat{p}_T(\boldsymbol{x})/\hat{p}_B(\boldsymbol{x})$ over the benchmark PDF $p_B(\boldsymbol{x})$,

$$L(r,\hat{r}) = \frac{1}{2}\int \left[\hat{r}(\boldsymbol{x}') - r(\boldsymbol{x}')\right]^2 p_B(\boldsymbol{x}')d\boldsymbol{x}' \tag{4.1}$$

$$= \frac{1}{2}\int \hat{r}(\boldsymbol{x}')^2 p_B(\boldsymbol{x}')d\boldsymbol{x}' - \int \hat{r}(\boldsymbol{x})p_T(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2}\int r(\boldsymbol{x}')^2 p_B(\boldsymbol{x}')d\boldsymbol{x}', \tag{4.2}$$

where the last term is constant and can be dropped, thus making the loss function independent of the unknown ratio $r(\boldsymbol{x})$. The estimated loss function is obtained by replacing the expectations over the unknown PDF p_B with the empirical averages

$$\hat{L}(r,\hat{r}) = \frac{1}{2N_B}\sum_{\boldsymbol{x}'\in\mathcal{B}} \hat{r}(\boldsymbol{x}')^2 - \frac{1}{N_T}\sum_{\boldsymbol{x}\in\mathcal{T}} \hat{r}(\boldsymbol{x}). \tag{4.3}$$

So, one can perform model selection by minimizing the estimated loss function (4.3) with respect to the parameter K and choosing this value of K as the optimal one. However, this procedure may be computationally intensive as it requires running the full algorithm several times (one for each different value of K).

Another approach is to implement the Point-Adaptive k-NN density estimator (PAk) [52–54], which is an algorithm to automatically find a compromise between large variance of the k-NN estimator (for small K), and large bias (for large K) due to variations of the density of points.

4.2 Identifying the discrepant regions

Suppose that after running the statistical test described in this paper one finds a p value leading to a rejection of the null hypothesis, or at least for evidence of incompatibility between the original PDFs. This means that the absolute value of the test statistic on the actual samples $|\mathrm{TS}_{\mathrm{obs}}|$ is large enough to deviate from zero significantly (to simplify the discussion, we assume in this subsection that $\mathrm{TS}_{\mathrm{obs}} > 0$ and the distribution of TS has zero mean and unit variance: $\hat{\mu} = 0, \hat{\sigma} = 1$). Then, our algorithm has a straightforward by-product: it allows to characterize the regions in feature space which contribute the most to a large $\mathrm{TS}_{\mathrm{obs}}$.

From the expression of the test statistic in Eq. (2.8) we see that we may associate a density field $(\boldsymbol{x}_j)$ to each point $\boldsymbol{x}_j \in \mathcal{T}$ as

$$u(\boldsymbol{x}_j) \equiv \log\frac{r_{j,B}}{r_{j,T}}, \tag{4.4}$$

such that the test statistic is simply given by the expectation value (arithmetic average) of $u(\boldsymbol{x}_j)$ over the whole trial sample $\mathcal{T}$

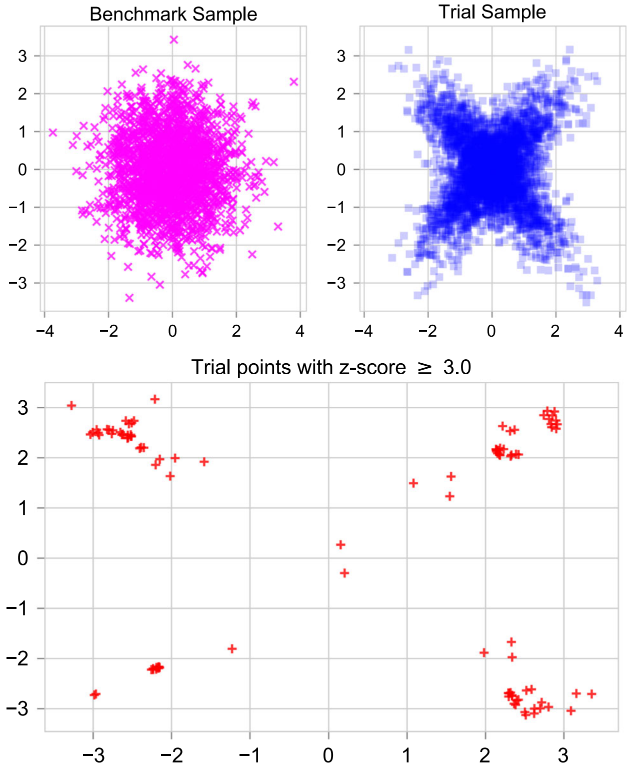


Fig. 7 Upper panel: benchmark (magenta crosses, left) and trial (blue squares, right) samples. Lower panel: points of trial sample with $z > 3.0$; this condition isolates the regions where most of the discrepancy between samples occurs

$$\mathrm{TS}_{\mathrm{obs}} = D \cdot \mathrm{E}_{\mathcal{T}}[u(\boldsymbol{x}_j)] + \log \frac{N_B}{N_{\mathcal{T}} - 1}. \tag{4.5}$$

It is then convenient to define a z-score field over the trial sample, by standard normalization of $u(\boldsymbol{x}_j)$ as

$$z(\boldsymbol{x}_j) \equiv \frac{u(\boldsymbol{x}_j) - \mathrm{E}_{\mathcal{T}}[u(\boldsymbol{x}_j)]}{\sqrt{\mathrm{Var}_{\mathcal{T}}[u(\boldsymbol{x}_j)]}}. \tag{4.6}$$

One can then use this score field to identify those points in $\mathcal{T}$ which are significantly larger than $\mathrm{TS}_{\mathrm{obs}}$, and they can be interpreted as the regions (or clusters) where the two samples manifest larger discrepancies.

This way, the z-score field provides a guidance for characterizing the regions in feature space where the discrepancy is more relevant, similar in spirit to regions of large signal-to-background ratio. For instance, the points $\boldsymbol{x}_j$ with $z(\boldsymbol{x}_j)$ larger than a given threshold, e.g. $z(\boldsymbol{x}_j) > 3$, are the points where one expects most of the "anomaly" to occur. An example of this is shown in Fig. 7, where a circular $\mathcal{B}$ sample is compared with a cross-like $\mathcal{T}$ sample. As expected, the z-field has higher density in correspondence of the corners of the cross.

Such regions of highest incompatibility between trial and benchmark samples may even be clustered using standard clustering algorithms, thus extending the method studied in this paper with another unsupervised learning technique.

Once they have been characterized and isolated, these high-discrepancy regions in feature space can provide a guidance for further investigation, in order to identify what causes the deviations. For example, they can be used to place data selection cuts.

5 Conclusions

Many searches for new phenomena in physics (such as searches for New Physics at the LHC) rely on testing specific models and parameters. Given the unknown nature of the physical phenomenon we are searching for, it is becoming increasingly important to find model-independent methods that are sensitive to an unknown signal hiding in the data.

The presence of a new phenomenon in data manifests itself as deviations from the expected distribution of data points in absence of the phenomenon. So, we propose a general statistical test for assessing the degree of compatibility between two datasets. Our method is model-independent and non-parametric, requiring no information about the parameters or signal spectrum of the new physics being tested; it is also un-binned, taking advantage of the full multi-dimensional feature space.

The test statistic we employ to measure the 'distance' between two datasets is built upon a nearest-neighbors estimation of their relative local densities. This is compared with the distribution of the test statistic under the null hypothesis. Observations of the test statistic at extreme tails of its distribution indicate that the two datasets come from different underlying probability densities.

Alongside an indication of the presence of anomalous events, our method can be applied to characterize the regions of discrepancy, providing a guidance for further analyses even in the case where one of the two samples (e.g. the background) is not known with enough accuracy to claim discovery.

The statistical test proposed in this paper has a wide range of scientific and engineering applications, e.g. to decide whether two datasets can be analyzed jointly, to find outliers in data, to detect changes of the underlying distributions over time, to detect anomalous events in time-series data, etc.

In particular, its relevance for particle physics searches at LHC is clear. In this case the observed data can be compared with simulations of the Standard Model in order to detect the presence of New Physics events in the data. Our method is highly sensitive even to a small number of these events, showing the strong potential of this technique.

Acknowledgements We would like to thank A. Davoli and A. Morandini for collaboration at the early stages of this work, and D. Barducci, R. Brasselet, R.T. D' Agnolo, A. Farbin, F. Gieseke, E. Merelli, E. Merényi, A. Laio, L. Lista and A. Wulzer for insightful discussions.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors' comment: There are no external data associated with the manuscript.]

Funded by SCOAP^3.

Appendix: Kullback–Leibler divergence

The Kullback–Leibler (KL) divergence (or distance) is one of the most fundamental measures in information theory. The KL divergence of two continuous probability density functions (PDF) P, Q is defined as

$$D_{\mathrm{KL}}(P||Q) \equiv \int P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})} d\boldsymbol{x}, \tag{5.1}$$

and it is a special case of f-divergences.

If the distributions P, Q are not known, but we are only given two samples $\mathcal{P} = \{\boldsymbol{x}_i\}_{i=1}^{N_P}$ of i.i.d. points drawn from P and $\mathcal{Q} = \{\boldsymbol{x}'_i\}_{i=1}^{N_Q}$ of i.i.d. points drawn from Q, it is possible to estimate the KL divergence using empirical methods. The estimated KL divergence between the estimated PDFs of $\hat{P}$, $\hat{Q}$ is obtained by replacing the PDFs P, Q with their estimates $\hat{P}$, $\hat{Q}$ and replacing the expectation value in Eq. (5.1) with the empirical (sample) average

$$\hat{D}_{\mathrm{KL}}(\hat{P}||\hat{Q}) = \frac{1}{N_P} \sum_{j=1}^{N_P} \log \frac{\hat{P}(\boldsymbol{x}_j)}{\hat{Q}(\boldsymbol{x}_j)}. \tag{5.2}$$

For the special case of Gaussian PDFs, the calculation of the KL divergence is particularly simple. Given two multivariate (D-dimensional) Gaussian PDFs defined by mean vectors $\boldsymbol{\mu}_{1,2}$ and covariance matrices $\Sigma_{1,2}$:

$$P = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_2), \quad Q = \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2), \tag{5.3}$$

the KL divergence in Eq. (5.1) is given by

$$\begin{aligned} D_{\mathrm{KL}}(P||Q) = \frac{1}{2} \Big[& (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \Sigma_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ & + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + \log \frac{\det \Sigma_2}{\det \Sigma_1} - D \Big]. \end{aligned} \tag{5.4}$$

References

1. CDF collaboration, T. Aaltonen et al., Model-independent and quasi-model-independent search for new physics at CDF. Phys. Rev. D **78**, 012002 (2008). arXiv:0712.1311
2. CDF collaboration, T. Aaltonen et al., Global search for new physics with 2.0 fb^{-1} at CDF. Phys. Rev. D **79**, 011101 (2009). arXiv:0809.3781
3. CMS collaboration, MUSIC—An Automated Scan for Deviations between Data and Monte Carlo Simulation. Tech. Rep. CMS-PAS-EXO-08-005, CERN, Geneva (2008)
4. CMS collaboration, Model Unspecific Search for New Physics in pp Collisions at sqrt(s) = 7 TeV. Tech. Rep. CMS-PAS-EXO-10-021, CERN, Geneva (2011)
5. G. Choudalakis, On hypothesis testing, trials factor, hypertests and the BumpHunter, in *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland 17–20 January 2011* (2011). arXiv:1101.0390
6. ATLAS collaboration, A model independent general search for new phenomena with the ATLAS detector at $\sqrt{s} = 13$ TeV. Tech. Rep. ATLAS-CONF-2017-001, CERN, Geneva (2017)
7. P. Asadi, M.R. Buckley, A. DiFranzo, A. Monteux, D. Shih, Digging deeper for new physics in the LHC data. JHEP **11**, 194 (2017). arXiv:1707.05783
8. R.T. D'Agnolo, A. Wulzer, Learning new physics from a machine. arXiv:1806.02350
9. ATLAS collaboration, M. Aaboud et al., A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment. arXiv:1807.07447
10. M. Kuusela, T. Vatanen, E. Malmi, T. Raiko, T. Aaltonen, Y. Nagai, Semi-supervised anomaly detection—towards model-independent searches of new physics. J. Phys. Conf. Ser. **368**, 012032 (2012). arXiv:1112.3329
11. K. Cranmer, J. Pavez, G. Louppe, approximating likelihood ratios with calibrated discriminative classifiers. arXiv:1506.02169
12. P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, D. Whiteson, Parameterized neural networks for high-energy physics. Eur. Phys. J. C **76**, 235 (2016). arXiv:1601.07913
13. J. Hernández-González, I. Inza, J. Lozano, Weak supervision and other non-standard classification problems: a taxonomy. Pattern Recognit. Lett. **69**, 49–55 (2016)
14. S. Caron, J.S. Kim, K. Rolbiecki, R. Ruiz de Austri, B. Stienen, The BSM-AI project: SUSY-AIgeneralizing LHC limits on supersymmetry with machine learning. Eur. Phys. J. C **77**, 257 (2017). arXiv:1605.02797
15. G. Bertone, M.P. Deisenroth, J.S. Kim, S. Liem, R. Ruiz de Austri, M. Welling, Accelerating the BSM interpretation of LHC data with machine learning. arXiv:1611.02704
16. C. Weisser, M. Williams, Machine learning and multivariate goodness of fit. arXiv:1612.07186
17. L.M. Dery, B. Nachman, F. Rubbo, A. Schwartzman, Weakly supervised classification in high energy physics. JHEP **05**, 145 (2017). arXiv:1702.00414
18. G. Louppe, K. Cho, C. Becot, K. Cranmer, QCD-aware recursive neural networks for jet physics. arXiv:1702.00748
19. T. Cohen, M. Freytsis, B. Ostdiek, (Machine) learning to do more with less. JHEP **02**, 034 (2018). arXiv:1706.09451
20. E.M. Metodiev, B. Nachman, J. Thaler, Classification without labels: learning from mixed samples in high energy physics. JHEP **10**, 174 (2017). arXiv:1708.02949
21. S. Chang, T. Cohen, B. Ostdiek, What is the machine learning? Phys. Rev. D **97**, 056009 (2018). arXiv:1709.10106
22. M. Paganini, L. de Oliveira, B. Nachman, CaloGAN: simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. Phys. Rev. D **97**, 014021 (2018). arXiv:1712.10321
23. P.T. Komiske, E.M. Metodiev, B. Nachman, M.D. Schwartz, Learning to classify from impure samples. arXiv:1801.10158
24. K. Fraser, M.D. Schwartz, Jet charge and machine learning. arXiv:1803.08066

25. J. Brehmer, K. Cranmer, G. Louppe, J. Pavez, A guide to constraining effective field theories with machine learning. arXiv:1805.00020
26. J. Brehmer, K. Cranmer, G. Louppe, J. Pavez, Constraining effective field theories with machine learning. arXiv:1805.00013
27. J. Brehmer, G. Louppe, J. Pavez, K. Cranmer, Mining gold from implicit models to improve likelihood-free inference. arXiv:1805.12244
28. A. Andreassen, I. Feige, C. Frye, M.D. Schwartz, JUNIPR: a framework for unsupervised machine learning in particle physics. arXiv:1804.09720
29. J.H. Collins, K. Howe, B. Nachman, CWoLa hunting: extending the bump hunt with machine learning. arXiv:1805.02664
30. S. Kullback, R. Leilber, On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
31. M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, M. Kimura, Least-squares two-sample test. Neural Netw. **24**, 735–751 (2011)
32. M. Sugiyama, T. Suzuki, T. Kanamori, Density ratio estimation in machine learning. (Cambridge University Press, Cambridge, 2012). https://doi.org/10.1017/CBO9781139035613
33. M. Schilling, Multivariate two-sample tests based on nearest neighbors. J. Am. Stat. Assoc. **81**, 799–806 (1986)
34. N. Henze, A multivariate two-sample test based on the number of nearest neighbor type coincidences. Ann. Statist. **16**, 772–783 (1988)
35. Q. Wang, S. Kulkarni, S. Verdù, Divergence estimation of continuous distributions based on data-dependent partitions. IEEE Trans. Inf. Theory **51**, 3064–3074 (2005)
36. Q. Wang, S. Kulkarni, S. Verdù, A nearest-neighbor approach to estimating divergence between continuous random vectors, in *Proceedings—2006 IEEE International Symposium on Information Theory. ISIT* , vol. 2006, pp. 242–246 (2006)
37. T. Dasu, S. Krishnan, S. Venkatasubramanian, K. Yi, An information-theoretic approach to detecting changes in multi-dimensional data streams, in *Proc. Symp. on the Interface of Statistics, Computing Science, and Applications* (2006)
38. F. Pérez-Cruz, Kullback–Leibler divergence estimation of continuous distributions, in *Proceedings of the IEEE International Symposium on Information Theory*, vol. 2008, pp. 1666–1670 (2008)
39. J. Kremer, F. Gieseke, K. Steenstrup Pedersen, C. Igel, Nearest neighbor density ratio estimation for large-scale applications in astronomy. Astron. Comput. **12**, 67–72 (2015)
40. Y.-K. Noh, M. Sugiyama, S. Liu, M.C. du Plessis, F.C. Park, D.D. Lee, Bias reduction and metric learning for nearest-neighbor estimation of Kullback–Leibler divergence. Neural Comput. **30**, 1930–1960 (2014)
41. E. Edgington, *Randomization Tests* (Dekker, New York, 1995)
42. A. van der Vaart, *Asymptotic Statistics* (Cambridge University Press, Cambridge, 1998)
43. B. Efron, R. Tibshirani, *An Introduction to Boostrap* (Chapman & Hall, London, 1993)
44. A.De Simone, T. Jacques, Simplified models vs. effective field theory approaches in dark matter searches. Eur. Phys. J. C **76**, 367 (2016). arXiv:1603.08002
45. G. Busoni et al., Recommendations on presenting LHC searches for missing transverse energy signals using simplified s-channel models of dark matter. arXiv:1603.04156
46. A. Albert et al., Recommendations of the LHC Dark Matter Working Group: Comparing LHC searches for heavy mediators of dark matter production in visible and invisible decay channels. arXiv:1703.05703
47. Atlas summary plot. https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CombinedSummaryPlots/EXOTICS/ATLAS_DarkMatter_Summary_Vector_ModifiedCoupling/history.html
48. J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer et al., The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. JHEP **07**, 079 (2014). arXiv:1405.0301
49. T. Sjstrand, S. Ask, J .R. Christiansen, R. Corke, N. Desai, P. Ilten et al., An introduction to PYTHIA 8.2. Comput. Phys. Commun. **191**, 159–177 (2015). arXiv:1410.3012
50. Delphes. https://cp3.irmp.ucl.ac.be/projects/delphes/
51. M. Sugiyama, K.R. Müller, Model selection under covariate shift, LNCS, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3697 (2005)
52. A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science **344**, 1492–1496 (2014). http://science.sciencemag.org/content/344/6191/1492.full.pdf
53. A. Rodriguez, M. d'Errico, E. Facco, A. Laio, Computing the free energy without collective variables. J. Chem. Theory Comput. **14**, 1206–1215 (2018). https://doi.org/10.1021/acs.jctc.7b00916
54. M. d'Errico, E. Facco, A. Laio, A. Rodriguez, Automatic topography of high-dimensional data sets by non-parametric density peak clustering (2018). arXiv:1802.10549