RanBox_SS: A Semi-supervised Learning Method for the Search of Rare Processes in LHC Data

Hevjin Yarar[1,2], Alexander Demin[3], Tommaso Dorigo[2], Luca Quagliarella[1], and Andrey Ustyuzhanin[3]

[1]University of Padova, Italy
[2]INFN - Sezione di Padova, Italy
[3]Higher School of Economics, Russia

January 31, 2022

Abstract

We discuss the customization of a unsupervised learning algorithm for anomaly detection, RanBox, for the semi-supervised task of searching for a known signal in a large background for which a precise model is not available. The algorithm is meant to be used for the search of the rare $B_s \to \tau\tau$ decay in proton-proton collision data produced by the Large Hadron Collider at CERN and collected by the CMS experiment. The RanBox method consists in a systematic scan of subspaces of a standardized copula of the feature space, where the extremum of a suitable figure of merit is sought by gradient descent. In this document we first characterize the performance of the proposed algorithm and show its effectiveness using simulated data, and then show its adaptation to a semi-supervised version, RanBox_SS, which may be useful in the search of a B_s signal that we are performing on proton-proton collisions data collected at the Large Hadron Collider by the CMS detector.

1 Introduction

1.1 Searches for new phenomena in collider physics

The word *anomaly* has its roots in the ancient Greek word $\alpha\nu\grave{\omega}\mu\alpha\lambda o s$; in common parlance, anomalous means "different, peculiar, or not easily classified" [1]. Consistently with the common usage of the word, in Statistics an anomalous datum is one which does not conform to the others in a set, because its observable features θ single it out as unlikely to have been sampled from the probability density function[1] $p_b(\theta)$ which the rest of the data conform to.

In the context of searches for new phenomena in high-energy particle physics (HEP), the identification of anomalies is of great interest. At the Large Hadron Collider (LHC), which produces the most energetic subnuclear reactions ever achieved in a laboratory, the CMS [2] and ATLAS [3] experiments compare the observable features of the final state of proton-proton collisions to extremely precise predictions yielded by the Standard Model (SM) of particle physics [4], in search for signals of new physics that the SM does not account for. Each collision can be typically summarized, through a complex reconstruction of tens of millions of digitally recorded signals, into few dozens of high-level features. The comparison of the distribution of those features with the ones expected from SM processes yields sensitivity to new physics phenomena. Given the large dimensionality of the problem, the typical approach followed by CMS and ATLAS in their searches is supervised classification: Monte Carlo simulations of both SM processes and hypothetical new physics phenomena inform the training of a classifier, whose output enables inference on the existence of new physics.

The above *modus operandi* rests on two pillars: reliance on an extremely precise model of $p_b(\theta)$, from which SM processes are sampled, and availability of theoretical models predicting the possible distribution of the density function $p_s(\theta)$ of signal events. It should be clear that the conditions on which those two pillars are based are difficult to satisfy in practice.

The SM, while superbly tested through decades of experimental probing [5], is subjected to uncertainties arising from imperfect knowledge of its underlying parameters, as well as from the purely empirical description of some of the fundamental ingredients playing a role in the particle collisions –*e.g.*, the parton distribution functions of the colliding protons, which determine the relative probability of different processes and their energy release. The resulting systematic uncertainties are especially impactful in those poorly-investigated regions of phase space which, thanks to the LHC's unprecedented reach, are probed by experimental data for the first time, and which are consequently the most likely hiding place of new processes. In addition to systematic uncertainties, new physics models suffer from the limited range of possibilities that they explore. New physics may manifest itself in ways that theoreticians have not yet ventured to speculate; if the resulting high-level features of signal events are not striking enough, or if they do not result in conspicuous modifications of some of the marginals of $p_b(\theta)$, they may be overlooked.

For the above reasons we planned for a systematic, unsupervised exploration of the feature space of LHC collider data. The algorithm we developed, called `RanBox` [6], was originally designed to exploit and fit to the characteristics of collider data: in particular, the wide range of values taken by $p_b(\theta)$, which calls for a vigorous standardization procedure at preprocessing stage; its local smoothness in the feature space; and the typically limited phase space where a signal density $p_s(\theta)$ may contribute in an observable way to recorded data.

In parallel to the plan of a unsupervised scanning of collected data in search for anomalous signals of new physics, we found that `RanBox` can also be excellently repurposed to handle the case when the

[1] The subscript b denotes it as the distribution of the "background", which is the name commonly associated to the non-anomalous component of unknown data.

signal that is being sought can be well specified and simulated with Monte Carlo programs, while the background – due to its complex composition in terms of contributing processes, when one selects data with triggers that do not characterize them sufficiently well – is hard to model. In this case, the algorithm can be taught to search for a region of phase space that, while rich in the expected signal, is comparatively less populated by backgrounds. A local knowledge of the background density may come from investigation of contiguous regions to the signal region of the feature space –which we call "sidebands" in the following. A figure of merit which minimizes the expected upper limit on the cross section for the rare process of interest is the appropriate one to use for this specific application, since we do not expect that the current data collected by the CMS experiment will be sufficient to observe the rare process, due to its expected very small branching fraction (of the order of a fraction of a millionth).

The plan of this document is as follows. In the remainder of this Section we briefly formalize the problem we wish to solve, for the unsupervised version of `RanBox`. In Sec. 2 we describe the `RanBox` algorithm in its original form. In Sec. 3 we demonstrate the performance of the algorithm on a synthetic dataset of simplified characteristics. In Sec. 4 we proceed to exemplify the results of the application of `RanBox` on the HEPMASS dataset, a simulation of proton-proton collisions produced for machine learning studies by the University of California Irvine. In Sec. 5 we discuss the peculiarities of the semi-supervised version of the algorithm and a first look at its performance on the HEPMASS dataset. We offer some concluding remarks in Sec. 6.

1.2 Problem Statement

We consider a set of N data examples $x \in \mathcal{S} \subseteq \mathbf{R}^{\mathcal{D}}$ sampled from a unknown multivariate density function $p(x)$. In general, $p(x)$ can be written as the sum of a background component $p_b(x)$ and a possible signal contamination $p_s(x)$,

$$p(x) = (1 - f_s)p_b(x) + f_s p_s(x) \tag{1}$$

where f_s is the signal fraction. An anomaly detection problem may be defined as one of finding a localized region of the feature space $\mathcal{S}$ that contains a density of data examples significantly higher than that of its surroundings, as defined by some suitable metric. This problem may be cast as a semi-supervised or a unsupervised one, depending on whether the (by definition) non-anomalous density of the background component is assumed known or not, or even if we instead assume known the signal component, as is of interest for the $B_s \to \tau\tau$ search. In the case when the signal is unknown, a central issue is how to retain sensitivity to a wide variety of anomalous contaminations, which may produce distortions of the density in a subset of the $\mathcal{D}$ features; in the case when it is the background which is unknown, the method to extrapolate its density to a region of interest becomes the focus of attention.

1.3 The idea of `RanBox`

We describe here the unsupervised version of the problem, which offers the benefit of avoidance of any model-related uncertainties, and we focus on new physics signals that characteristically produce localized, compact variations in the overall density of the feature space.

We wish to construct an algorithm that searches the feature space $\mathcal{S}$ by considering a "box", *i.e.*, a multidimensional interval constructed in a subspace of $\mathcal{S}$. The random nature of the box lays not only in the endpoints x_i^{min}, x_i^{max} of its intervals in each marginal, $x_i \in [x_i^{min}, x_i^{max}]$, but also in the involved subspace $\mathcal{S}' \subseteq \mathbf{R}^{\mathcal{D}'}$ of $\mathcal{S}$ described by a subset of the x_i. Alternatively, one may think of the box as having restricted intervals in only a subset $\mathscr{D} - \mathscr{D}'$ of the dimensions of $\mathscr{S}$. If we consider for the time being the case $f_s = 0$ and N data points in $\mathscr{S}$ sampled from a multi-dimensional uniform density $p_b(x) = \mathcal{U}(x)$, such a box will contain a predictable fraction of the total data: given the box volume V_{box} and the total volume V of the feature space $\mathscr{S}$, the expectation value of the number of events in the box is $N_{exp} = NV_{box}/V$. Conversely, if $f_s > 0$, the observed number of events captured within the box boundaries N_{obs} may yield an estimate of the density of the total sampling distribution in the corresponding region of $\mathscr{S}$, contributed by both $p_b(x)$ and $p_s(x)$:

$$\hat{p}(x) = \frac{N_{obs}}{VN_{exp}} = \frac{N_{obs}}{NV_{box}}. \tag{2}$$

The above estimate may be used to construct a test statistic sensitive to an anomalous local overdensity of the data; *e.g.*, one may simply define the test statistic to equal the estimated excess of events in the box, $N_{obs} - N_{exp}$, or a significance measure of its non-null value. The maximization of such a test statistic will be appropriate for searches of anomalies that preferentially populate well-confined regions of the feature space, such as those of interest in collider searches for new physics, but also relevant to other branches of science as, *e.g.*, astrophysical observations, or industrial applications such as process control, fraud detection, or spam filtering. Conversely, we expect little sensitivity to multi-modal signals, and (by construction) no sensitivity to broad deformations of a nearly-uniform background distribution $p_b(x)$. The locality of the signal to be detected, however, is the only assumption we allow ourselves to take in the construction of our anomaly detection procedure. The assumption of uniformity on which the estimate in Eq. 2 is based can be loosened if we work in the copula space, as discussed more in detail in Sec. 2.

2 Algorithm Description

2.1 Starting considerations

The multitude of subnuclear particles resulting from proton-proton collisions recorded by LHC experiments, which we take as our target application in the construction of the algorithm, yield tens of millions of electronic signals in the detectors. This large body of information is summarized by a process called “event reconstruction” through the extraction of several tens of high-level features that describe the measurement of energy and direction of all observed particles (*e.g.*, energetic electrons or muons) or sets of particles (hadronic jets) [2]. Even if we focus on specific interesting subsets of the available data, any energy-related feature of the observed particles will show a highly dis-uniform distribution, with a peak at low values and long tails extending to higher energy (see, *e.g.*, Fig. 1). The variation in density between those peaks and tails may amount to orders of magnitude, and is due to the corresponding large variation in the probability that the collision is originated by quarks or gluons carrying a low or a high fraction of their parent’s total momentum.

Because of the above, it seems natural to proceed by first pre-processing the data with an integral transform of all the features, such that each marginal becomes uniform by construction. The algorithm will then work in the copula space, examining the data structure with a metric unaffected, at least to first order, by the original strong density variations in the feature space.

[2] In HEP it is thus customary to call *events* the observed data examples, and we will stick to that convention in this work.

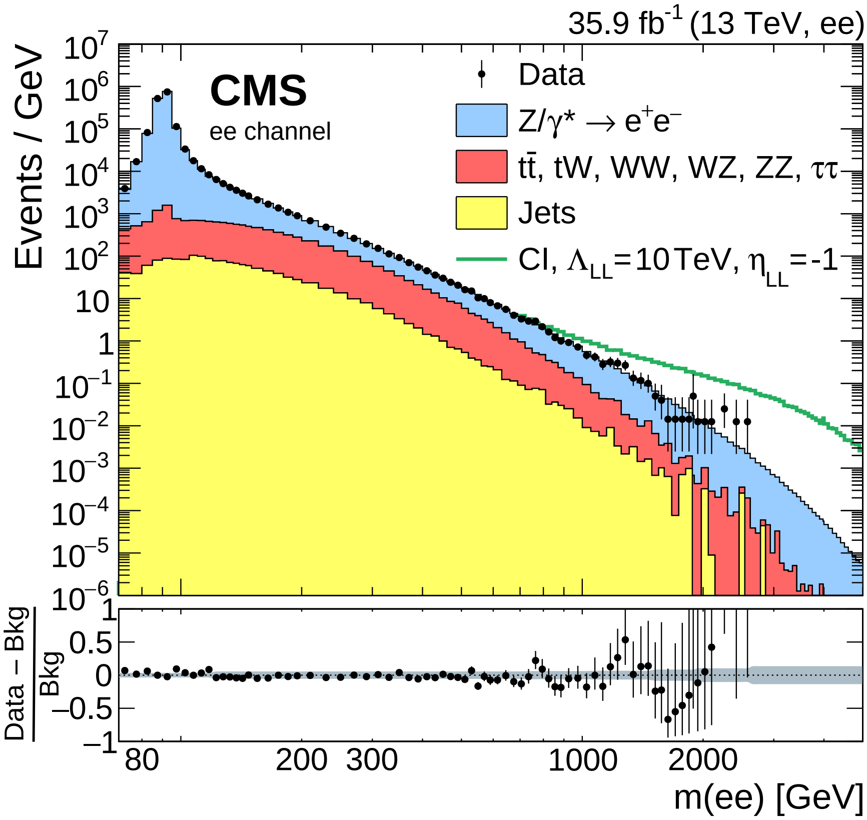


Figure 1: Distribution of the invariant mass of candidate electron-positron pairs observed by the CMS experiment in 36 fb^{-1} of Run 2 LHC collisions [7]. The data show a variation in density by several orders of magnitude as a function of mass. The cited reference reports on searches for a new physics contribution involving contact interactions, which could contribute to the distribution at its high-end tail (green curve).

2.2 Data preprocessing

The probability integral transform of a function $f(x)$ is defined by setting

$$F(x) = \int_{-\infty}^{x} f(t)dt, \tag{3}$$

which is such that $y = F(x)$ is uniform in $[0, 1]$:

$$\begin{aligned} F_y(y) &= P(Y \leq y) \\ &= P(F_X(X) \leq y) \\ &= P(X \leq F_x^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) = y. \end{aligned} \tag{4}$$

Once each of the variables of the feature space x_i is transformed as above into the corresponding one in the set y_i, information once contained in the interdependence of the x_i is retained in the copula, which is the joint distribution function of variables with uniform marginals (Sklar's theorem) [8]. The advantage of the transformation is evident: a search for overdensities in the space spanned by y_i will not be spoiled by uneven marginals, and will correctly concentrate on the regions of space which are dense because of interdependence of the features. An additional bonus of working with the y_i variable basis is that the feature space is now a unit hypercube, with volume $V = 1$.

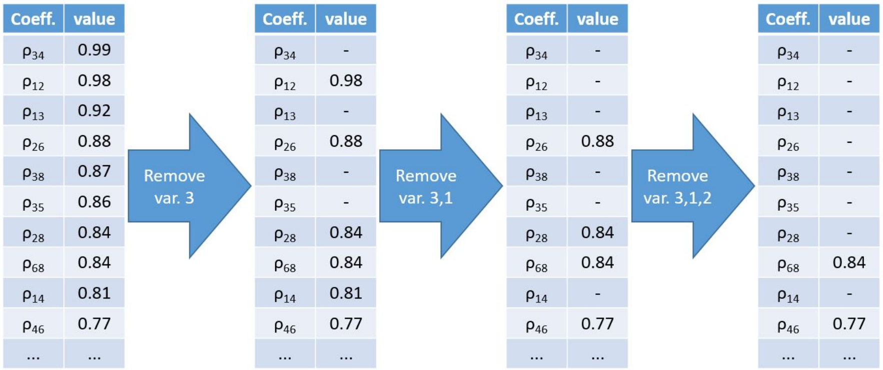
2.3 Dimensionality reduction

The dreaded "curse of dimensionality" [9] affects any search in high-dimensional spaces populated by sparse data. In the typical applications considered in this work, the total data size N lays in the few thousands to few hundreds of thousands range; consequently, an investigation of subspaces $\mathcal{S}'$ of the feature space $\mathcal{S}$ quickly becomes meaningless as their dimensionality grows larger than about $\mathcal{D}' = 12 - 15$, when Poisson fluctuations prevent any reasonable multi-dimensional density estimate.

An additional optional preprocessing step, which may prove useful to reduce the dimensionality in cases when $\mathcal{D}$ is larger than a few tens, is the application of Principal Component Analysis (PCA) to the feature space. PCA essentially consists in fitting a hyper-ellipsoid to the data, and remapping the feature space in a space spanned by the principal axes of the ellipsoid. One may then use the principal components, which are those on which the data exhibit the largest variance, and ignore the last few in the ordered list of components, which are likely to contain the least information. PCA can be useful for `RanBox` in cases when the search for subspaces of limited dimensionality $\mathcal{D}'$ of the feature space proves impractical because of the large binomial coefficient $\binom{\mathcal{D}}{\mathcal{D}'}$, which makes the exploration of a meaningful fraction of the possible $\mathcal{D}'$-dimensional subspaces too CPU-intensive. However, in our investigations we have found that PCA is generally liable to reduce the power of the search for overdense regions of the feature space when the data are composed of a large background component and a small signal contamination to which we wish to be sensitive. The typical reason of this effect is connected with the fact that a variable which exhibits little variance on the majority of the data, and is thus discarded by PCA, may still be very distinctive for a small signal.

A viable alternative to reduce the dimensionality of the problem, which may facilitate the identification of small signals, is to exploit the correlation matrix of the features, by removing features which add little information. This is an attractive option when searching for small anomalous components in a background-rich dataset: by identifying and removing variables that are highly correlated with others on the majority component of the data, we reduce the possibility that such correlations affect negatively the chance of the algorithm to identify localized overdensities genuinely due to a clustering of multiple distinguishing features of a minority component. As a telling example, if in a $\mathcal{D} = 30$-dimensional feature space one of the variables were identically repeated 10 times, and `RanBox` performed a search in $\mathcal{D}' = 10$-dimensional subspaces, the algorithm would be very likely to end up focusing on the same narrow interval (any one would do) of each of those features: *e.g.*, a 10-dimensional box of width 0.1 in each of the correlated features would have a volume of 10^{-10}; if there were $N = 10,000$ events in the space, such a box would be predicted to contain $N_{exp} = 10^{-6}$ events, while it would in fact contain exactly 1000 events!

Our correlated variable removal (CVR) procedure, which performs the identification of variables to be discarded, works as follows. We first compute the correlation coefficients ρ_{ij} among all pairs of variables ij, and order them in a list by decreasing absolute value $|\rho_{ij}|$. Then we choose the number of variables to be removed N_{void}; in order to identify these we consider that if the k-th variable is removed, all correlation coefficients that include k as one of the two indices will become irrelevant. We thus find the combination of N_{void} variables which, when removed, minimizes the value of the highest surviving correlation coefficient. A graphical example of the technique is shown in Fig. 2.



Coeff.	value
ρ_{34}	0.99
ρ_{12}	0.98
ρ_{13}	0.92
ρ_{26}	0.88
ρ_{38}	0.87
ρ_{35}	0.86
ρ_{28}	0.84
ρ_{68}	0.84
ρ_{14}	0.81
ρ_{46}	0.77
…	…

Coeff.	value
ρ_{34}	-
ρ_{12}	0.98
ρ_{13}	-
ρ_{26}	0.88
ρ_{38}	-
ρ_{35}	-
ρ_{28}	0.84
ρ_{68}	0.84
ρ_{14}	0.81
ρ_{46}	0.77
…	…

Coeff.	value
ρ_{34}	-
ρ_{12}	-
ρ_{13}	-
ρ_{26}	0.88
ρ_{38}	-
ρ_{35}	-
ρ_{28}	0.84
ρ_{68}	0.84
ρ_{14}	-
ρ_{46}	0.77
…	…

Coeff.	value
ρ_{34}	-
ρ_{12}	-
ρ_{13}	-
ρ_{26}	-
ρ_{38}	-
ρ_{35}	-
ρ_{28}	-
ρ_{68}	0.84
ρ_{14}	-
ρ_{46}	0.77
…	…

Figure 2: Graphical description of the CVR procedure available in the preprocessing stage of **RanBox**. *The ordered list of absolute values of correlation coefficients among the variables defining the $\mathcal{D}-$dimensional feature space is scanned by searching for all possible combinations of N_{void} variables which, once removed, minimize the largest surviving correlation coefficient. In the figure, for $N_{void} = 3$ the removal of variables 3, 1, 2 (shown in succession for clarity) reduces the highest surviving correlation most effectively.*

2.4 Choices of a test statistic for the unsupervised learning task

We consider two estimates of the expected number of events contained in a multi-dimensional region of the unit hypercube resulting from the standardization procedure, both corresponding to a binomial ratio. The first one is simply

$$N_{exp,V} = NV_{box}. \tag{5}$$

As the total copula space volume is $V = 1$, the above estimate is only driven by the extension of the box volume V_{box}. The expectation results from assuming that the data distribute in the feature space with a constant density, and is useful in cases when $p_b(x)$ contains little structure in its copula, as departures from that assumption can then easily be associated with anomalous contaminations. This measure is the default one for the studies of algorithmic performance presented in Sec. 3, which are performed on synthetic datasets where the assumption above is identically true in the limit $f_s = 0$. A second estimate, affected by higher statistical uncertainty than the former but conversely much less affected by a non-uniform density $p_b(x)$ in the copula space, may be obtained by defining a sidebands (SB) region that surrounds the search box (see Fig. 3). In this case, no reliance is made on overall constancy of the density for non-anomalous events, and the estimate leverages the density of data in the immediate neighborhood of the search box. If $[x^i_{min}, x^i_{max}], i = 1....\mathcal{D}'$ are the boundaries of the search box, the SB region is defined by the following relations:

$$\delta_i = 0.5(x^i_{max} - x^i_{min})(2^{1/\mathcal{D}'} - 1), \tag{6}$$

$$x^i_{min,SB} = \max(0, x^i_{min} - \delta_i), \tag{7}$$

$$x^i_{max,SB} = \min(1, x^i_{max} + \delta_i), \tag{8}$$

with $x^i \notin [x^i_{min}, x^i_{max}]$ for at least one i, *i.e.* the SB volume does not include the search box volume. The SB then has a volume at most as big as the search box volume; it is in general smaller than that, as some of the intervals cannot extend on each side of the search box by the required length δ_i, due to the hard boundaries at 0 and 1 (see again Fig. 3). If one observes a number of events N_{out} in the

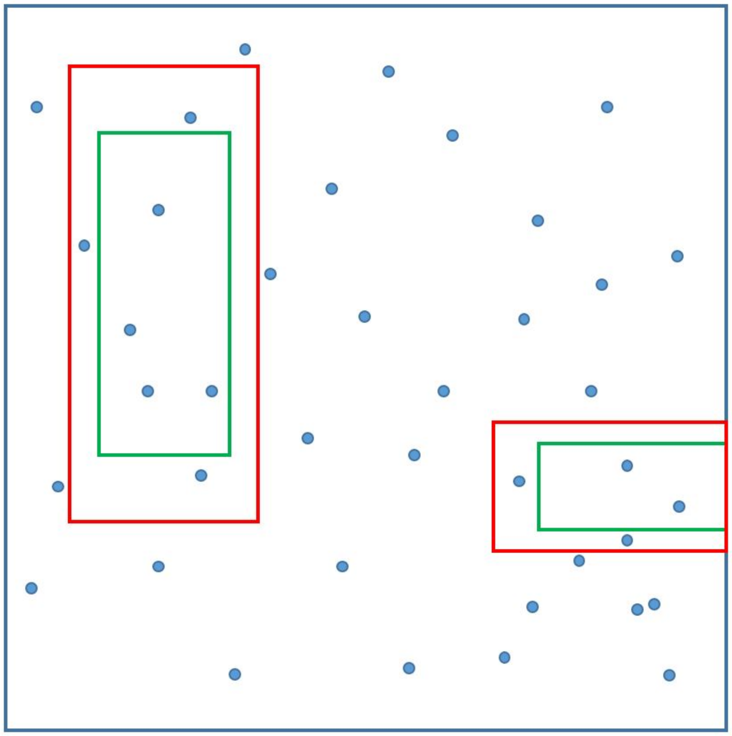


Figure 3: Representation of the sideband method for box density estimation. Two possible search boxes in a 2-dimensional space are shown in green; the relative sidebands, constructed according to the recipe of Eq. 8, are the regions between the red and the green rectangles. The sideband region on the lower right can only extend horizontally to the left, and the area it defines is thus smaller than that of the related search box.

sideband region, the expectation value of the number of events in the search box in the assumption of uniformity may be written as

$$N_{exp,\tau} = \tau N_{out}, \tag{9}$$

where

$$\tau = \frac{V_{box}}{V_{SB}} \tag{10}$$

is defined by the volumes of sideband region V_{SB} and search box V_{box}. A slight modification of the recipe for the expectation value above, which we have found to be effective, is operated when the number of observed sideband events N_{out} is zero. In that case, which is frequent for large dimensionality searches and small statistics of the data sample, it is useful to reset N_{exp} to the full-volume prediction, Eq. 5. We stick to this recipe in our applications of the sideband method in the studies described in this work.

To formulate the problem in its generality through the above definition of the extrapolation variable τ, we observe that the full-volume estimate in Eq. 5 corresponds to setting

$$\tau = \frac{V_{box}}{1 - V_{box}} \tag{11}$$

and $N_{out} = N - N_{in}$. In either case a likelihood-ratio-based test statistic may now be defined as follows:

$$Z_{PL} = \sqrt{2}\Bigg\{ N_{in} \ln \left[(1+\tau) \left(\frac{N_{in}}{N_{in}+N_{out}} \right) \right] + N_{out} \ln \left[\frac{1+\tau}{\tau} \left(\frac{N_{out}}{N_{in}+N_{out}} \right) \right] \Bigg\}^{0.5} \tag{12}$$

The above defined function has been shown [10] to be a good approximation of the Z-score corresponding to the binomial probability of observing an excess of events $N_{in} - N_{exp,\tau}$ in the box. It is to be noted, however, that Z_{PL} cannot be considered a genuine signal significance, because in real applications "non-anomalous" data contain structure in the copula due to interdependence of their features; as a result, the Z_{PL} test statistic for the null hypothesis has fatter tails at positive values than a Normal distribution. In addition, as discussed *infra* in more detail, `RanBox` effectively operates multiple testing on the dataset, hence Z_{PL} cannot be used as a significance measure in the absence of a Bonferroni or similar correction [11]. Despite the above caveats, the fact that Z_{PL} is a principled proxy to the significance of an excess in a Binomial counting experiment makes it a sound choice for a test statistic when the focus is the search for significant, anomalous signals.

We have observed that the Z_{PL} test statistic is especially useful when anomalies are sought which may interest wide volumes of the feature space, with N_{exp} correspondingly being not very small — typically in the range of several tens to a hundred of events. Conversely, when the expectation N_{exp} in the overdense region amounts to only a few events or less, an attractive alternative is to use the function R_{reg} defined as

$$R_{reg} = \frac{N_{in}}{N_{exp} + N_{reg}}, \tag{13}$$

with, *e.g.*, the regularization term set to $N_{reg} = 1$. The maximization [3] of R_{reg} may identify more effectively small anomalies well confined in the search volume, in cases when the copula space of non-anomalous events has a rich structure, capable of producing high values of Z_{PL} in regions of large volume and thus diverting the algorithm's attention from small, well-confined anomalies.

2.5 Box seeding

The search for the most overdense multi-dimensional interval in a feature space populated by sparse data points is complicated by the presence of a large number of local extrema; hence, a careful initialization of the box location and dimensions may significantly improve the performance of the algorithm. Although we tried several recipes for this task, here we only describe three of them, which we found the most suitable for our applications.

The baseline method, "Algorithm 0", consists in a fixed initialization of the box to a multi-dimensional interval of total volume V_{box}, set to equal a given fraction of the unit volume of the full feature space hypercube. The box, which lives in a $\mathscr{D}'$-dimensional subspace of the copula, is constructed by defining intervals x^i_{min}, x^i_{max} (with $i = 1, ..., \mathscr{D}'$) as follows:

$$\Delta = \frac{1 - V_{box}^{1/\mathscr{D}'}}{2} \tag{14}$$

$$x^i_{min} = \Delta \tag{15}$$

$$x^i_{max} = 1 - \Delta \tag{16}$$

An optimization of the initial value of V_{box} is of course impossible in a unsupervised search, where neither non-anomalous or anomalous data have a specified density. However, our tests suggest that

[3] In this work we stick to the setting $N_{reg} = 1$, and consequently address the test statistic as R_1.

setting $V_{box} = 0.1$ is a reasonable choice when, as is the case in several of our considered applications, $\mathscr{D}'$ lays in the 6-10 dimensions range. *E.g.,* with $\mathscr{D}' = 6$ one obtains starting intervals equal to $[0.16, 0.84]$, and with $\mathscr{D}' = 10$ intervals equal to $[0.10, 0.90]$. Note that this corresponds to a relatively large box, in terms of its extension along each marginal. When combined with a search algorithm that considers initial expansions or shrinkages in each of the box dimensions by amounts sufficient to extend all the way to the unit hypercube boundaries, the above initialization ensures that no overdensity laying close to the boundary of a coordinate will be overlooked by the search algorithm taking a step in the wrong direction at the start of the search.

The second method, "Algorithm 1", is instead based on clustering the data based on a specialized Nearest-Neighbour (NN) search. First, the nearest neighbour j is found for every event i in the data, by using as a distance the following function:

$$d_{ij} = \prod_{k=1}^{\mathscr{D}'/2} |x^i_{o_k(ij)} - x^j_{o_k(ij)}| \tag{17}$$

where $o_k(ij)$ are the $\mathscr{D}'/2$ indices identifying the spatial coordinates for which the intervals $|x^i - x^j|$ are the smallest. In other words, the map d_{ij} determines the minimum volume of a $\mathscr{D}'/2$-dimensional box that includes events i and j. Once d_{ij} is defined for all i and j, one may compute for every event i the number of neighbouring events $j = j_1 ... j_{N_{cl}}$ that have i as their closest event according to that metric. The event i_{maxNN} with the maximum number $N_{cl,maxNN}$ of such neighbours now allows to identify all $N_{2^{nd}order}$ events which have any of the $N_{cl,maxNN}$ events as their own nearest neighbours. The box can finally be initialized as the smallest $\mathscr{D}'$-dimensional interval that includes all the $N_{2^{nd}order}$ neighbours. A graphical description of the algorithm is provided in Fig. 4.

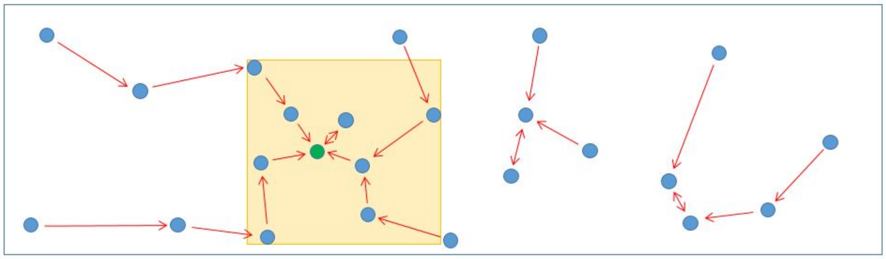


Figure 4: Graphical description of the clustering algorithm used for box initialization with Algorithm 1. Blue points indicate the position of events in the two shown variables of the feature space. Arrows pointing from an event to another indicate the location of the closest neighbour of the event originating the arrow (according to a metric described in the text). The green point is the closest to four others, and it provides the seed of the box: the collection of all events which point to those four events define the boundaries of the box.

A third initialization method, "Algorithm 2", uses instead a kernel estimation of the density for the identification of starting box boundaries. The density is evaluated at the position of each of the N events as the sum of N $\mathscr{D}$-dimensional Gaussian distributions centered at the location of every event in the sample, and with equal diagonal covariance matrices $C = k^2 I_D$, with I_D the $\mathscr{D}$-dimensional identity matrix and k a tunable parameter which must be chosen according to the total dataset size and the dimensionality of the $\mathscr{D}'$ subspaces scanned by `RanBox`; its default value, used in the applications described in this work, is $k = 0.2$. Once the point of highest density x_{HD} is identified, the box is initialized as the multi-dimensional interval whose extension in each coordinate x is

$$[\max(x_{HD} - \delta_2, 0.), \min(x_{HD} + \delta_2, 1.)], \tag{18}$$

with the provision that if x_{HD} is at less than δ_2 distance from the boundary at 0 (1), the interval defaults to $[0., 2\delta_2]$ or $[1. - 2\delta_2, 1.]$, respectively. The default value of δ_2 is 0.2; *e.g.* this corresponds, for a 10-dimensional subspace search, to initial boxes of volume equal to or smaller than 0.0001: the expected number of events within a 10,000-event sample contained in a random box of that volume is 1.0, which is a suitable starting point for the background expectation in the test statistic maximization. Given that the initialization provided by Algorithm 2 offers a good candidate for an overdense region, the focusing on a small initial region of feature space has been observed to be effective in the tested applications of our interest: those are in fact cases when a small, overdense region exists in the first place.

2.6 Maximization of the test statistic

A search for the multi-dimensional interval providing the highest value of the chosen test statistic (either Z_{PL} or R_{reg} as defined in Sec. 2.4 above) in a $\mathcal{D}'$-dimensional subspace of the feature space can be performed as follows.

Step 1: The initialization of the box is performed with the algorithm of choice. A set of step parameters are set to the starting value $\lambda_i = 0.5\,(i = 1...\mathcal{D}')$. A loop counter N_{GD} is set to zero.

Step 2: Seven possible modifications are considered for each of the $\mathcal{D}'$ intervals defining the box:

$(x^i_{min})'$	$(x^i_{max})'$
$\max(x^i_{min} - \lambda_i, 0)$	x^i_{max}
$\min(x^i_{min} + \lambda_i, x^i_{max} - \epsilon)$	x^i_{max}
x^i_{min}	$\max(x^i_{max} - \lambda_i, x^i_{min} + \epsilon)$
x^i_{min}	$\min(x^i_{max} + \lambda_i, 1)$
$\max(x^i_{min} - \lambda_i, 0)$	$\max(x^i_{max} - \lambda_i, \epsilon)$
$\min(x^i_{min} + \lambda_i, 1 - \epsilon)$	$\min(x^i_{max} + \lambda_i, 1)$
$r^i_{min} = \min(r_1, r_2)$	$r^i_{max} = \max(r_1, r_2)$

where ϵ is a parameter determining the coarseness of the algorithmic scan in the feature space, fixed in applications described in this work to $\epsilon = 0.01$. In the last line the values r_1, r_2 determining a "random jump" in the i-th interval are random numbers sampled from a uniform distribution in $[0, 1]$. The values of $(x^i_{min})'$ and $(x^i_{max})'$ defined above are rounded off to two decimal places in all cases. For each of these $7\mathcal{D}'$ variations, an associated SB region is defined by the recipe described *supra*; this determines the numbers N_{in} and N_{out} and consequently the $7\mathcal{D}'$ values of the test statistic of choice.

Step 3: If the highest among the $7\mathcal{D}'$ values of the test statistic corresponding to the tentative box modifications is higher than the current maximum value, the box is modified to the corresponding new multi-dimensional interval, and all λ_i values for the coordinates not affected by the change are reduced as follows:

$$\lambda_i \to \max\left(f\lambda_i, \epsilon\right) \tag{19}$$

where the factor f is set to 0.9. In addition, if the box modification is chosen based on one of the $\mathcal{D}'$ random intervals $[r^i_{min}, r^i_{max}]$, a counter j_i is incremented by one; once a j_i reaches a maximum value (10 by default), no more random jumps are allowed for the intervals in variable i. This recipe allows to control the convergence of the algorithm as well as the trade-off between its CPU consumption and its freedom in exploring new box configurations in the considered feature space dimensions.

If, instead, the current value of the test statistic is higher than all of the $7\mathcal{D}'$ new values, no modifications to the box boundaries are applied, and λ_i values are reduced as in Eq. 19.

Step 4: The loop counter N_{GD} is incremented by one. If N_{GD} reaches a limiting value (set to 100 by default) the algorithm stops; the algorithm also stops if all values λ_i have reached the value ϵ. Otherwise, steps 2, 3, and 4 above are repeated.

Despite its simplicity, the procedure described above typically converges in 30 to 50 iterations for $\mathcal{D}' = 6 - 10$, which are typical values for the considered applications of fixed-subspace searches.

3 Performance studies with synthetic data

3.1 Event generation

A synthetic dataset sampled from a multi-dimensional Uniform distribution $p_b(x) = \mathcal{U}(x)$, with $x \in [0,1]^{\mathcal{D}}$, may be generated by repeated calls to the TRandom3→Uniform() routine [4] of the ROOT package [12], which we employ in our `c++` implementations of `RanBox`. Such a dataset may be considered the ideal background for an anomaly search: by lacking any internal structure in the copula, it constitutes a best-case scenario for performance evaluations of the algorithm in a controlled setting. The unknown signal may instead be generated by drawing samples from a multi-dimensional Gaussian distribution in a subset $x_g, g = 1, ..., N_g$ of the features, $x_g \in \mathbf{R}^{N_g}$, and the remaining ones $x_u, u = N_g + 1, ..., \mathcal{D}$ from a uniform density. While Gaussians have support on the real axis, the generation ensures that the drawn features are also contained in the $[0, 1]$ interval, as detailed below.

We define the following default set of parameters:

- (for background) $x_i = \mathcal{U}(0, 1)$;
- (for signal) $x_u = \mathcal{U}(0, 1)$;
- sigma $\sigma_{gg} = \mathcal{U}(0.01, 0.1)$;
- mean $\mu_g = \mathcal{U}(3\sigma_{gg}, 1 - 3\sigma_{gg})$;
- $r_{gh} = \mathcal{U}(-1, 1)$ (with $g, h \in \{1, ..., N_g\}, g \neq h$).

A random choice of σ_{gg} and r_{gh} values as defined above will not in general generate a positive-definite covariance matrix C with variances σ^2_{gg} and $\sigma^2_{gh} = r_{gh}\sigma_{gg}\sigma_{hh}$; hence the procedure of generating C is repeated until a Cholesky-Banachiewicz (CB) decomposition $LL^T = C$ into a lower-triangular matrix L [13] is found, which guarantees the positive-definite nature of C. Once successful, the CB decomposition allows to easily draw samples from the multi-dimensional Gaussian distribution by posing, for every g,

- (for signal) $x_g = \mu_g + \sum_{h=1..N_g} L_{gh} n_h$

[4] The random generation is based on the Mersenne primes, and has a periodicity of about 10^{6000}.

with n_h sampled from a Normal distribution. During event generation, if a coordinate sampled from the multivariate Gaussian exceeds the range $[0, 1]$, it is simply resampled. This truncation has the effect that Gaussians with μ_g values close to the boundaries have an up to twice higher local density than Gaussians closer the center of the $[0, 1]$ interval. For this reason, in most tests we limit μ_g values to the range stated above, except when we explicitly study the performance at the edges (see *infra*). Although the background is already generated with flat marginals, after the inclusion of signal we of course re-standardize the dataset by using Eq. 4.

When performing power tests of the algorithm, we avoid the random effect of varying σ parameters, and use reference samples with a more narrowly defined signal component, by fixing all Gaussian sigmas to $\sigma_{gg} = 0.05$. In this case correlation coefficients r_{gh} are chosen at random within the discrete set $\{-max(r_{gh}), 0., max(r_{gh})\}$ by posing $max(r_{gh}) = 0.2$, and we allow means μ_g to vary at random in their default range, $[0.15, 0.85]$. The different signals that correspond to varied means and correlations have equal chance of being identified by the algorithm. For example, Fig. 5 shows average Z_{PL} values from runs of the algorithm with the following choice of parameters:

- $N_b = 4950$ background events
- $N_s = 50$ signal events
- $\mathscr{D} = 20$ active dimensions of feature space
- $N_G = 6$ Gaussian features in signal component
- $\mathscr{D}' = 6$ dimensions for box definition
- $N_{trials} = 1$ subspace sampled per dataset, of features coincident with the N_G in which signal component has a Gaussian distribution.
- $N_{rep} = 50$ datasets generated and searched
- Algorithm 0 (random box initialization) and 2 (kernel density) used
- No dimensionality reduction (PCA or correlated features removal) performed

By only considering, through the above choices, the subspace which yields the highest probability of locating a signal-rich box, we reduce the effect of randomness and allow for a more precise study of the impact of the tested parameters. In Fig. 5 the values of the test statistic Z_{PL} appear stable as a function of the sampled ranges $max(r_{gh})$ and $\Delta\mu_g = \mu_g^{max} - \mu_g^{min}$, indicating that the search algorithm is capable of locating overdensities regardless of their position in the space [5], and that the correlation between Gaussian-distributed variables does not affect the chance of identifying overdense multi-dimensional intervals. Similar results are obtained by initializing the box dimension with Algorithm 1 (kNN-seeded clustering), and/or by using R_1 as a test statistic.

[5] The observed slightly lower performance of searches initialized by Algorithm 0 for $\Delta\mu_g$ values close to 1 is an effect of the higher chance of central signals to be initially contained in randomly-initialized boxes. Instead, Algorithm 2 allows to exploit the slightly higher maximum density reached by signals with one or more features close to the boundaries of the space, due to the already mentioned truncation we operate outside the $[0, 1]$ range.

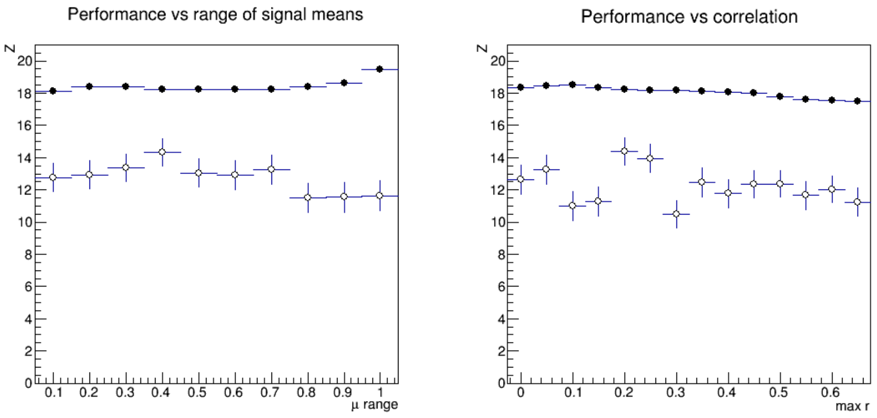


Figure 5: Mean values of the test statistic Z_{PL} as a function of the characteristics ($\Delta\mu_g$ for $r_{gh} = 0.2$ (left), and r_{gh} for $\Delta\mu_g = 0.7$ (right)) of the signal component, from 50 repetitions of searches in synthetic datasets each composed of 50 signal events and 4950 background events. Black points correspond to searches initialized with Algorithm 2, empty points correspond to searches initialized with Algorithm 0. For reference, the critical region (for $\alpha = 0.05$) corresponds to $Z_{PL} = 7.1(7.2)$ for Algorithm 0 (2, respectively). See the text for other detail.

3.2 Power tests of the unsupervised `RanBox`

While in a unsupervised search one cannot in general define a hypothesis test, given the absence of hypotheses for the sampling distributions, we are still interested in verifying the ability of `RanBox` to locate overdense regions of the feature space as a function of its free parameters for a set of different benchmark datasets. This will provide a scale of the algorithm sensitivity. Hence we construct a "flat" dataset containing events uniformly distributed in the feature space, and "signal" datasets where a fraction of the events are sampled from a PDF which includes, for some of the features, a multivariate Gaussian component (see *supra*). Once a type-I error rate α is defined, the tail integral of the test statistic distribution $f(TS|H_1)$, output by `RanBox` searches on alternative hypotheses H_1 corresponding to datasets contaminated with events having multivariate Gaussian features, allows to construct a power function $1 - \beta(\alpha)$ as

$$1 - \beta(\alpha) = \int_{x_{cr}(\alpha)}^{\infty} f(x|H_1)dx, \tag{20}$$

where $x_{cr}(\alpha)$ is defined by the relation

$$\alpha = \int_{x_{cr}(\alpha)}^{\infty} f(x|H_0)dx. \tag{21}$$

To check the performance of the algorithm in a controlled setting, we define signal parameters by fixing the Gaussian sigma values in signal events to $\sigma_{gg} = 0.05$, and allow means and correlations to vary in the range $\mu_g \in [0.15, 0.85]$ and $r_{gh} \in$ {-0.2,0.,0.2}, respectively. We consider again samples of 5000 events, and study the power $1 - \beta$ for the three choices $\alpha = 0.05, 0.01, 0.001$, using $\mathcal{D} = 20$ space dimensions. We also set the following algorithm hyperparameters:

- Algorithm = 0
- $N_{trials} = 1000$ subspaces scanned for each dataset

- test statistic used: Z_{PL}
- expectation value of events in the box: $N_{exp,V}$.

In a first test we fix the number of features where the signal component exhibits a Gaussian distribution to $N_g = 15$, and vary the number of signal events in the generated samples. The critical region is directly obtained for $\alpha = 0.05$ from the distribution $f(TS|H_0)$ obtained by repeating 500 times the procedure of generation and 1000-subspace-search of datasets including no signal. For the two smaller values of α (0.01, 0.001), we instead rely on the modeling of the distribution of $f(TS|H_0)$ with a Gamma function (see Fig. 6) to determine the corresponding x_{cr} values. For each studied value of the signal component we obtain 50 values of $f(TS|H_1)$, from which we extract the power as the fraction of values in the critical regions corresponding to the three chosen values of α. The results of this test are shown in Fig. 7 (top row). We observe that `RanBox` is fully capable of spotting localized accumulations due to a multivariate Gaussian signal, down to few-per-mille contaminations of the data sample.

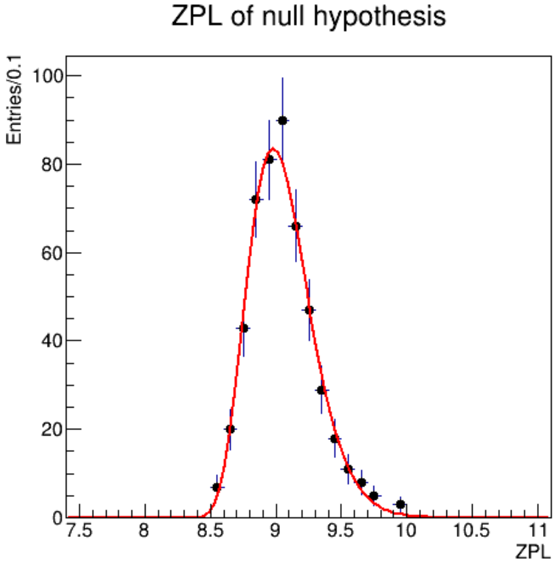


Figure 6: Distribution of the Z_{PL} test statistic for 500 repetitions of `RanBox` tests of the null hypothesis in 5000-event background-only samples; a fit to a Gamma function is overlaid. 1000 subspaces are scanned with Algorithm 0 for the box initialization. See the text for other details.

In a second study we determine, with the same procedure described *supra*, the power of `RanBox` as a function of the number of Gaussian dimensions N_g of the signal component, by fixing the signal fraction to $f_s = 1\%$ (*i.e.*, 50 signal events and 4950 background events). We observe in Fig. 8 (top row) that there is sensitivity to multivariate Gaussian signals that involve even only few (4 and above) of the 20 dimensions of the feature space.

In Fig. 9 and Fig. 10 we provide a visualization of sample results of a `RanBox` run. The first figure shows marginal distributions of the six features where `RanBox` identifies an anomalous signal, in the copula space (where the total dataset has by definition uniform marginals before the selection). The subspace where the best box is found is one where the signal exhibits Gaussian distributions in all the features, and all the events in the box are in fact due to the signal component. The scatterplots of Fig. 10 show two-dimensional distributions of the full data sample and the data selected as the best

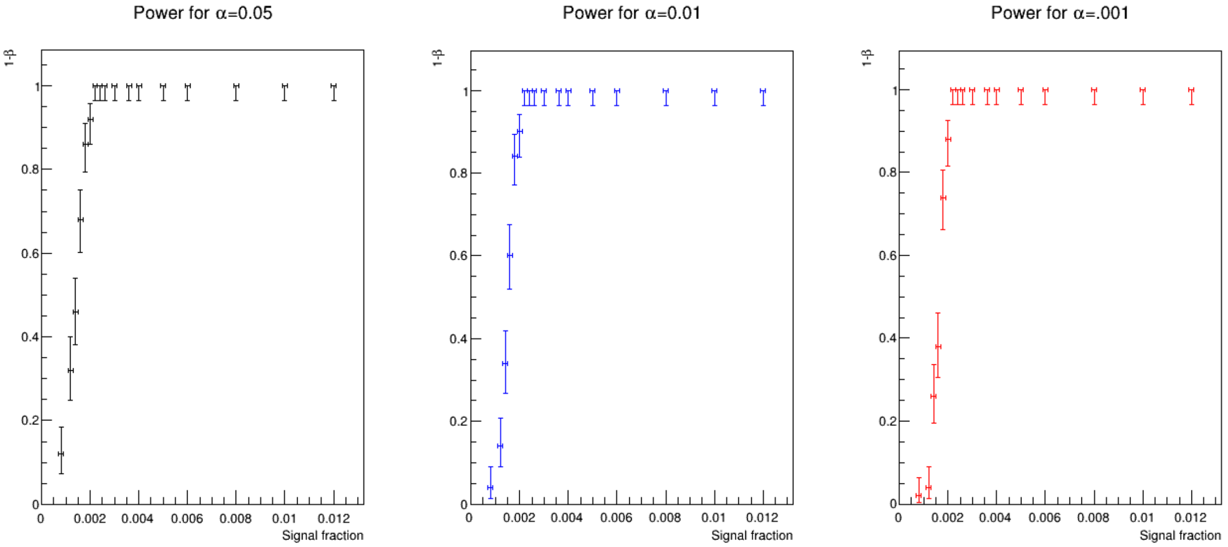


Figure 7: `RanBox` power curves for Z_{PL} as a function of the fraction of signal in 5000-event samples. The black points (left) correspond to $\alpha = 0.05$, the blue points (center) to $\alpha = 0.01$, and the red points (right) to $\alpha = 0.001$; the critical region for the latter two tests are obtained from extrapolated values of Z_{PL} for the null hypothesis. 68.3% intervals are computed with the Clopper-Pearson method for the Binomial ratio. See the text for other details.

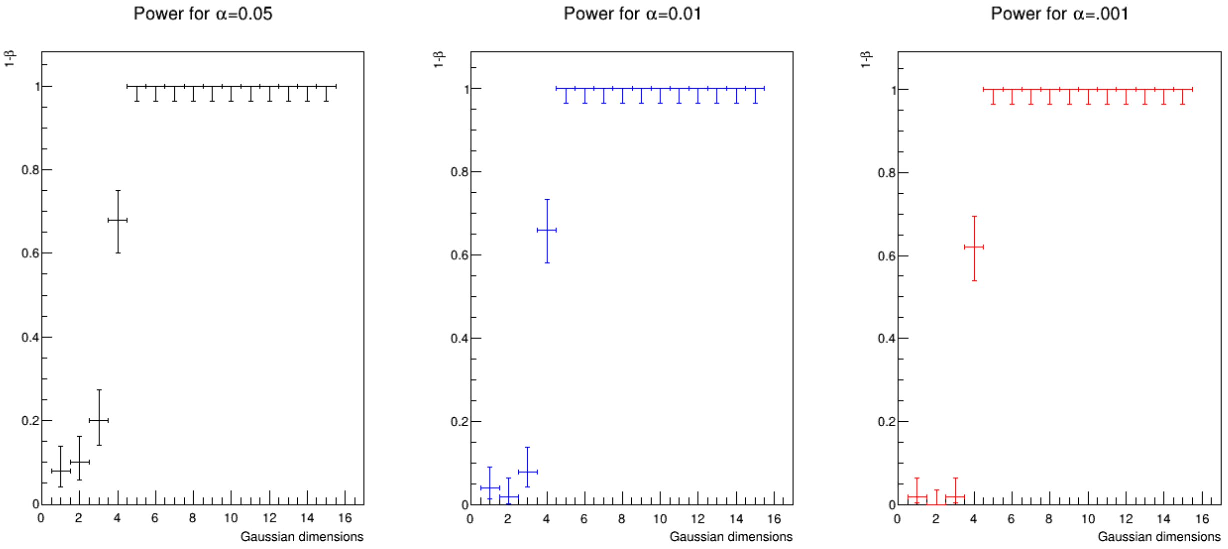


Figure 8: `RanBox` power curves for Z_{PL} as a function of the number of Gaussian features in signal events, in samples containing 50 signal events and 4950 flat-distributed events. The black points correspond to $\alpha = 0.05$, the green points to $\alpha = 0.01$, and the red points to $\alpha = 0.001$; the latter two are obtained from extrapolated values of the critical region. 68.3% intervals are computed with the Clopper-Pearson method for the Binomial ratio. See the text for other details.

box. This further demonstrates the correct working of the algorithm, which can effectively extract the overdense region from an apparently flat distribution. The conclusions we draw are that the

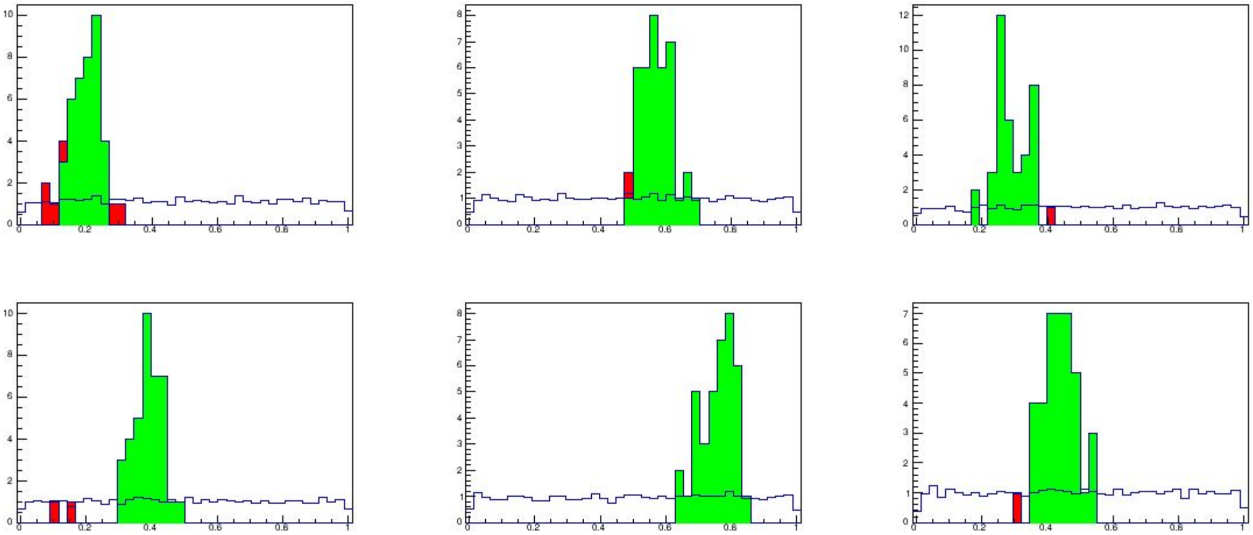


Figure 9: Distribution of the six features defining the subspace where **`RanBox`** *finds the highest-Z_{PL} box in a run on 5000 synthetic events, 4950 of them generated from a $D = 20$-dimensional uniform distribution and the remaining 50 "signal" events generated with 11 features drawn from a multidimensional Gaussian distribution. The blue histograms show the totality of the data; the filled green histograms show the distribution of events contained in the highest-Z_{PL} box; the filled red histograms show the distribution of events that fail to be contained in the box only because of their value on the displayed variable. See the text for other details.*

algorithm performs as expected when run on a synthetic data sample and in controlled conditions.

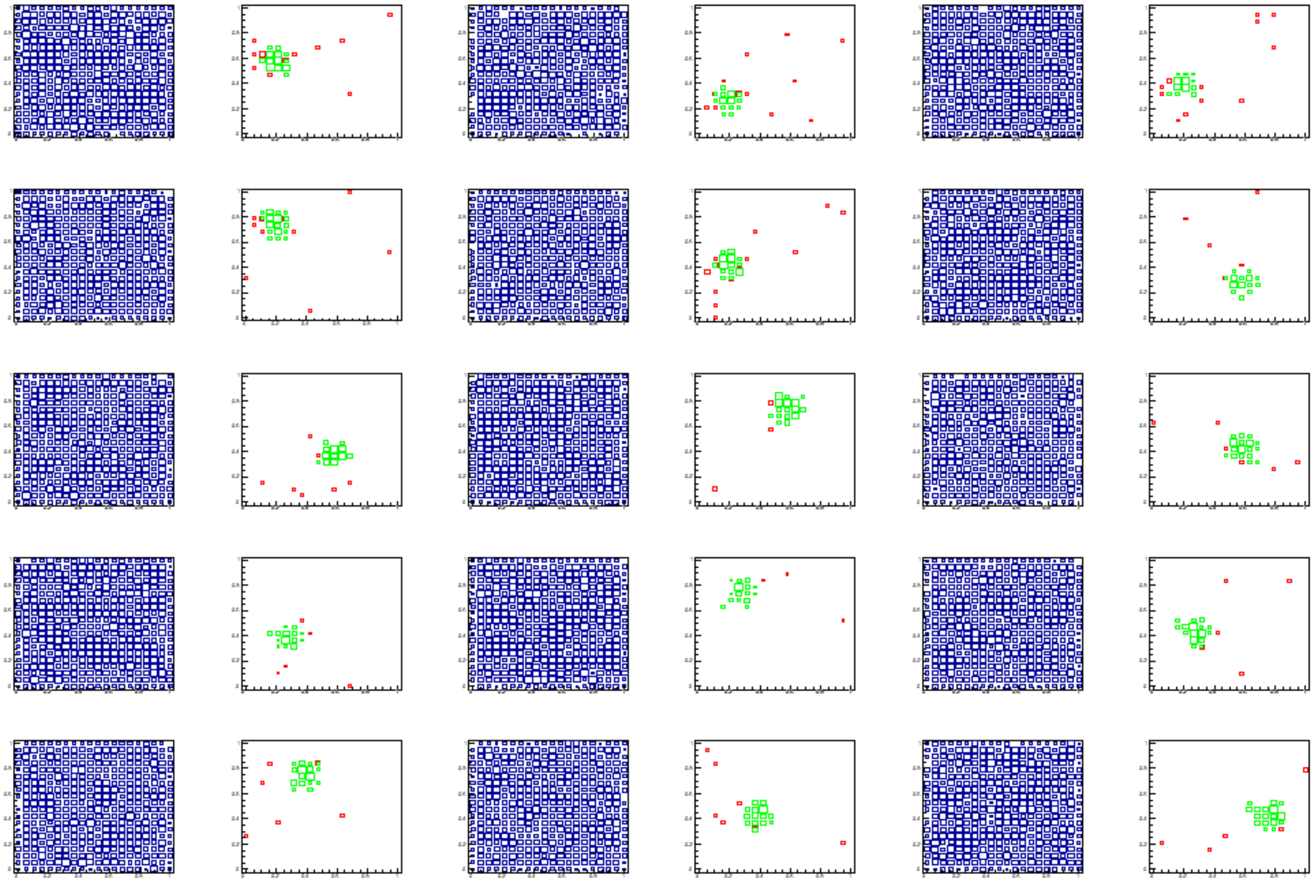


Figure 10: Scatterplots of the six features defining the subspace where **`RanBox`** *finds the highest-Z_{PL} box in a run on 5000 synthetic events, 4950 of them generated from a $D = 20$-dimensional uniform distribution and the remaining 50 "signal" events generated with 11 features drawn from a multidimensional Gaussian distribution. The distribution of the totality of the data is shown in blue on the left of each pair of graphs, while the distribution of selected events (in green) is shown in green on the corresponding right graph; in red are events that fail to be included in the highest-Z_{PL} box only because of their value of the shown features. From top to bottom and left to right each pair of graph describes the spaces (v_1, v_2), (v_1, v_3), (v_1, v_4) (first row), (v_1, v_5), (v_1, v_6), (v_2, v_3) (second row), (v_2, v_4), (v_2, v_5), (v_2, v_6) (third row), (v_3, v_4), (v_3, v_5), (v_3, v_6) (fourth row), and (v_4, v_5), (v_4, v_6), (v_5, v_6) (fifth row). See the text for other details.*

4 Experiments

The power tests described in Sec. 3 are about as far as one can go to characterize the performance of the unsupervised version of RanBox, since on any real-life dataset the specificities of the data structure and the lack of generalization power of the algorithm will make it pointless to investigate in a systematic way its optimal settings and resulting sensitivity. For this reason, in this Section we free ourselves of the need to assess confidence intervals on all the reported statistics, which would also entail a quite significant computing burden [6], and prefer to offer sets of results of single runs of the algorithm on samples of data taken from a dataset offered by particle physics research.

4.1 Exotic signals in LHC data

The search of new phenomena in LHC proton-proton collisions data is the very application that the unsupervised version of RanBox was designed to address. A signal of new physics may manifest itself as a localized increase in density in some of the features derived from particle interactions in the detector. A model-independent search should consider a complete set of kinematical features describing the observed particles in the final state of the collision events, and perform an unbiased scan of their combined multi-dimensional distribution.

For a test of RanBox on the above use case we rely on the large dataset of simulated proton-proton collisions available in the University of Irvine's repository [14], a dataset known by its nickname "HEPMASS". This dataset was generated explicitly to test multivariate algorithms for classification and search of small signals in large background datasets. The generated signal is that of an exotic resonant particle X, with a mass of 1000 GeV, which decays to a pair of top quarks, $X \to t\bar{t}$, when the top quarks successively produce in their decay a single-lepton final state characterized by a high-energy electron or muon, a neutrino, and four hadronic jets. Background samples describe all Standard Model processes that produce a similar final-state signature. The ATLAS experiment is considered as the detector that performs the reconstruction of the produced particle signals; more detail on the generated dataset and the simulation are available in [15].

The data are characterized by reconstruction-level variables from a fast simulation. An idealized reconstruction of a proton-proton collisions yielding top quark pairs is performed, identifying the observed jets, leptons, and b-jets [7]. From the reconstruction of the event, the low-level kinematic features obtained are particle momenta: the momentum of the leading lepton, the momentum of the four leading jets (in decreasing order of transverse momentum) and related b-tagging information, and magnitude and azimuthal angle of the so-called "missing transverse momentum" vector. The latter is defined as the opposite of the sum of the momentum vectors of all observed particles, calculated in the transverse plane of the particle beams [8].

The high-level features of the set are the values of the invariant masses of the intermediate objects calculated using the low-level kinematic features, in the hypothesis that a correct identification of decay objects and assignment to final state particles has been obtained. These are: $m_{\ell\nu}$ from the decay process $W \to \ell\nu$, m_{jj} from the $W \to qq'$ process, m_{jjj} from the $t \to Wb \to bqq'$ process, $m_{j\ell\nu}$ from the $t \to Wb \to \ell\nu b$ process, and the combined m_{WWbb} mass of the decay products assumed for X. Table 1 lists identity and information of the 27 features.

[6] The tests we report in this work overall cost several thousand hours of single-machine CPU by themselves.

[7] We call "b-jet" a hadronic jet which has been originated from a b-quark. When classified as such by a software algorithm, the jet is said to be "b-tagged".

[8] Missing transverse momentum carries information on the momenta of neutrinos, particles typically produced in weak boson decays that do not leave a traceable signal in the detector but can still be inferred from the imbalance of

Number	Feature	Description
1-3	P_T^ℓ, η_ℓ, ϕ_ℓ	3-momentum of primary lepton
4	P_T^{miss}	Missing transverse momentum
5	$\phi_{P_T^{miss}}$	Missing transverse momentum azimuthal angle
6	N_{jets}	Number of additional jets
7-9	$P_T^{j_1}$, η_{j_1}, ϕ_{j_1}	3-momentum of first jet
10	$P_{tag}^{j_1}$	First jet b-tag information
11-13	$P_T^{j_2}$, η_{j_2}, ϕ_{j_2}	3-momentum of second jet
14	$P_{tag}^{j_2}$	Second jet b-tag information
15-17	$P_T^{j_3}$, η_{j_3}, ϕ_{j_3}	3-momentum of third jet
18	$P_{tag}^{j_3}$	Third jet b-tag information
19-21	$P_T^{j_4}$, η_{j_4}, ϕ_{j_4}	3-momentum of fourth jet
22	$P_{tag}^{j_4}$	Fourth jet b-tag information
23	$m_{\ell\nu}$	Mass of reconstructed lepton-neutrino system
24	m_{jj}	Mass of jets from $W \to qq'$ decay products
25	m_{jjj}	Mass of reconstructed $t \to Wb \to bqq'$ decay system
26	$m_{j\ell\nu}$	Mass of reconstructed $t \to Wb \to l\nu b$ decay system
27	m_{WWbb}	Mass of hypothetical X resonance

Table 1: List of the 27 features of signal and background events in the HEPMASS dataset. The first 22 are low-level features, the last 5 are higher-level ones produced by combining the low-level features into physics-motivated observables. See the text for more detail.

In [16] several ROC curves are presented to compare the performance of parametrized and non-parametrized neural networks on the HEPMASS signal discrimination problem. Those are the result of supervised classification, which duly exploits *a priori* knowledge of the signal density. As can be seen in Fig. 12, the non-mass-parametrized neural network achieves a background efficiency of about 3% for a signal efficiency of 80%, *e.g.*. We will use these approximate values for a qualitative comparison to the performance of RanBox, bearing in mind all the caveats of any comparison of supervised and unsupervised classification methods.

In this section we use a mixture of signal and background events from the HEPMASS dataset to test under what conditions RanBox is capable of evidencing feature space regions with a dominant signal contamination. Since the feature space is rich with interdependencies among the features, the task of a unsupervised algorithm is considerably harder than in the case of the synthetic dataset studied in Sec. 3, as significant overdensities are expected to arise from the structure of background processes alone. Furthermore, in a real-life application of RanBox, the user would be unable to extract the distribution of the test statistic under the null hypothesis, as even slight differences between simulation and real data would distort the output. We consider therefore that in that case RanBox would be used by running it on real data as they come, without any pretense of assessing a significance level of the returned overdense regions or of studying the power of a selection criterion, but rather with the aim of focusing the attention of researchers on the combinations of features that exhibit interesting localized overdensities.

We proceed with exploratory runs of the RanBox algorithm on the HEPMASS dataset as we would perform them on real data. We construct a dataset comprised of 250 signal and 4750 background events: the 5% signal fraction is small enough to make the signal indistinguishable in the marginal

the momenta of observed particles.

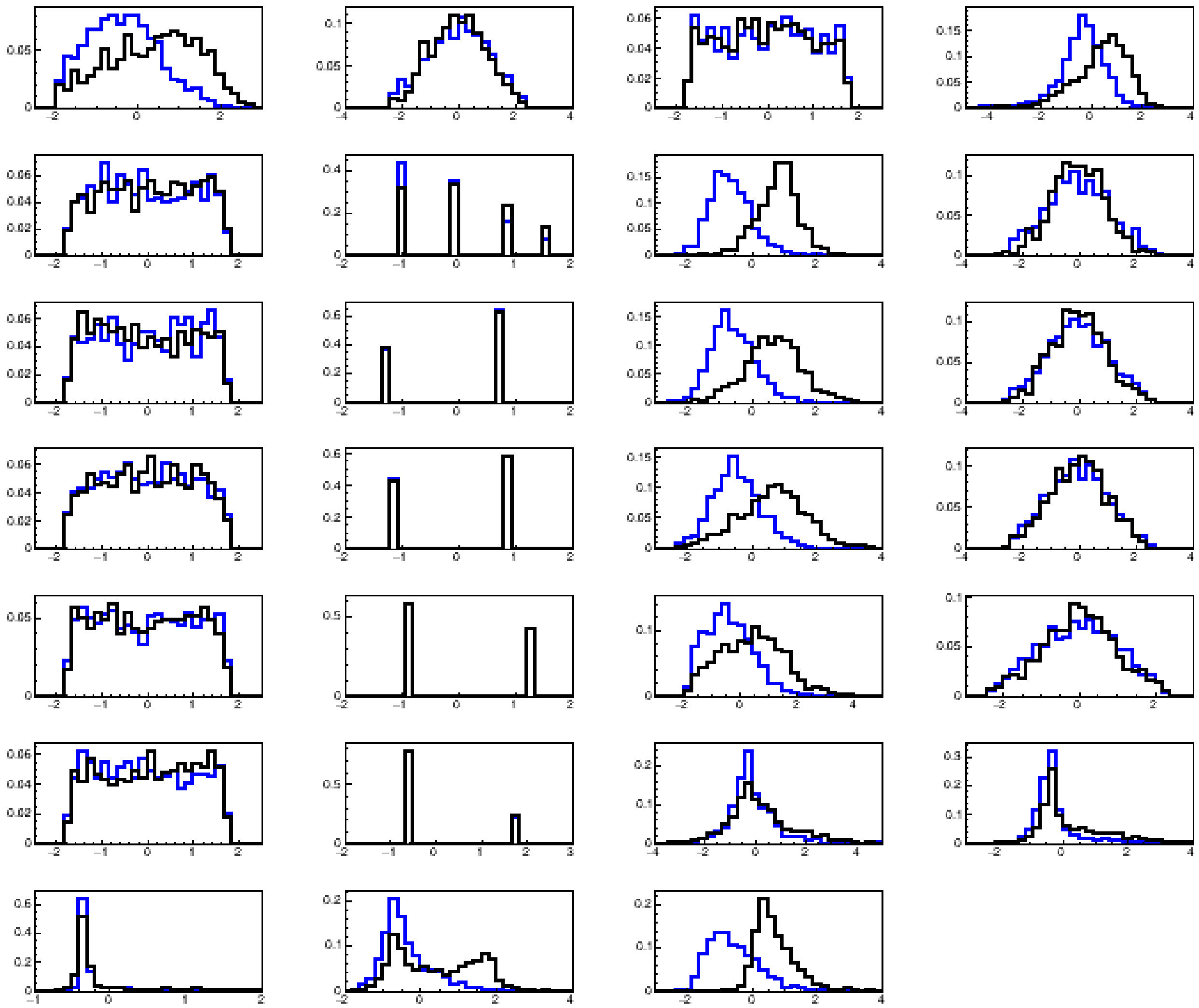


Figure 11: Normalized and standardized distributions of the 27 features of HEPMASS data for signal (black) and background (blue).

Algorithm	Init.	T.S.	Extrap.	$\mathscr{D}'$	D red.	N_{iter}	N_{best}
`RanBox`	A2	R_1	SB	$\mathscr{D}' = 12$	no	10,000	N/A

Table 2: Run parameters of the **`RanBox`** *algorithm for a test on the HEPMASS dataset with a 5% signal contamination. "Init." indicates the method defining the initial dimension of the search box; "Extrap." identifies the way by which a prediction of events in the box is computed; $\mathscr{D}'$ is the dimensionality of the subspaces scanned by* **`RanBox`***; "Dim. red." indicates whether the dimensionality of the feature space was reduced with PCA or by discarding the most correlated variables; N_{iter} is the number of searched subspaces by* **`RanBox`***.*

distributions of even the most discriminating variables, as shown in Fig. 13. We run `RanBox` with the parameters listed in Table 2. They constitute a reasonable choice for a run on HEPMASS. In particular, since we wish to be sensitive to a small signal contamination rather than having the algorithm get distracted by broader-scale background correlations, we initially consider that the R_1

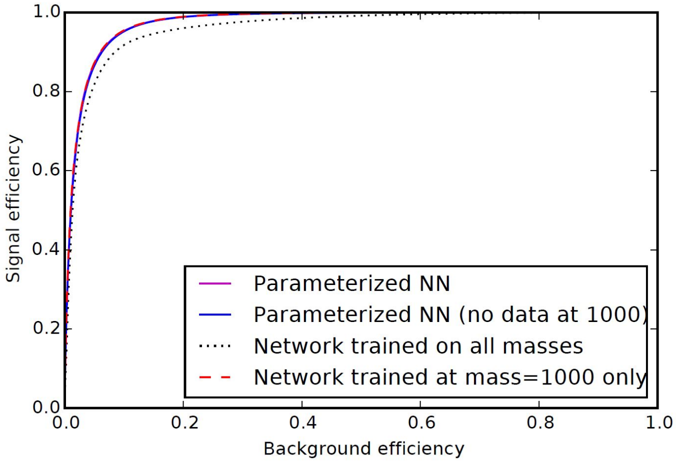


Figure 12: Comparison of signal and background efficiency curves for four classes of neural networks on the HEPMASS dataset. Of relevance here is the dashed red curve, which correspond to a non-parametrized network trained and tested on the sample of reference, with a resonance mass of 1000 GeV. Reprinted with permission from [16].

test statistic might be more sensitive to a signal component. Also, we use the sidebands method to extrapolate the density in the search box, as this better factors out the local disuniformities in the data. The choice of dimensionality of the scanned subspaces is instead driven by preconceptions on the fact that a signal of new physics will most likely exhibit distinctive features only in a subset of the considered kinematical variables [9]; 12 is anyway close to the maximum meaningful choice for that parameter, as is clear if we consider that in a 12-dimensional space a box of sides equal to half the range of each feature will on average contain only $5000 \times 2^{-12} = 1.2$ events out of 5000. Finally, we do not apply any dimensionality reduction to the input data, as we observe that the maximum two-variable correlation coefficient (0.757) in the mixture dataset is not particularly high.
From the results listed in Table 3 we may draw a few interesting conclusions. First of all, the search of 10,000 subspaces performed by `RanBox` returns a good number of signal-rich regions, as five of the ten most significant boxes are dominated by the signal component, and two more are also considerably signal-enriched, by factors above six [10]. Such an output, and in particular the most significant box alone, would certainly allow experimentalists to focus on the small signal now evident in the identified regions, hence we consider this output a success of the anomaly detection task. We also note that the scan of 10,000 12-dimensional subspaces costs nearly 10 hours of running on a single CPU; the scan of all 12-dimensional subspaces of the 27-dimensional feature space is instead not an easily viable option, as this would require $27!/(12!15!) = 1.738 \times 10^7$ iterations, or about two years of CPU on a single machine. Regardless, on the HEPMASS dataset a limited number of combinations of 12 features still allow to evidence a small signal.

If we now compute the signal and background efficiency of the regions returned by the algorithm in its exploration of the $f_s = 5\%$ dataset, we notice that the best box identified by `RanBox` contains 46 signal events out of 54, which corresponds to an 85% efficiency; the background efficiency is

[9] Indeed, in the considered search for $X \to t\bar{t}$, apart from the resonant structure of the total invariant mass of the decay products, one expects only minor differences of the signal with respect to the non-resonant $t\bar{t}$ production predicted by the SM.

[10] In the following we take that factor as a threshold to count the number of signal-rich (SR) boxes among the first ten boxes, a number we report as $SR_{1:10}$.

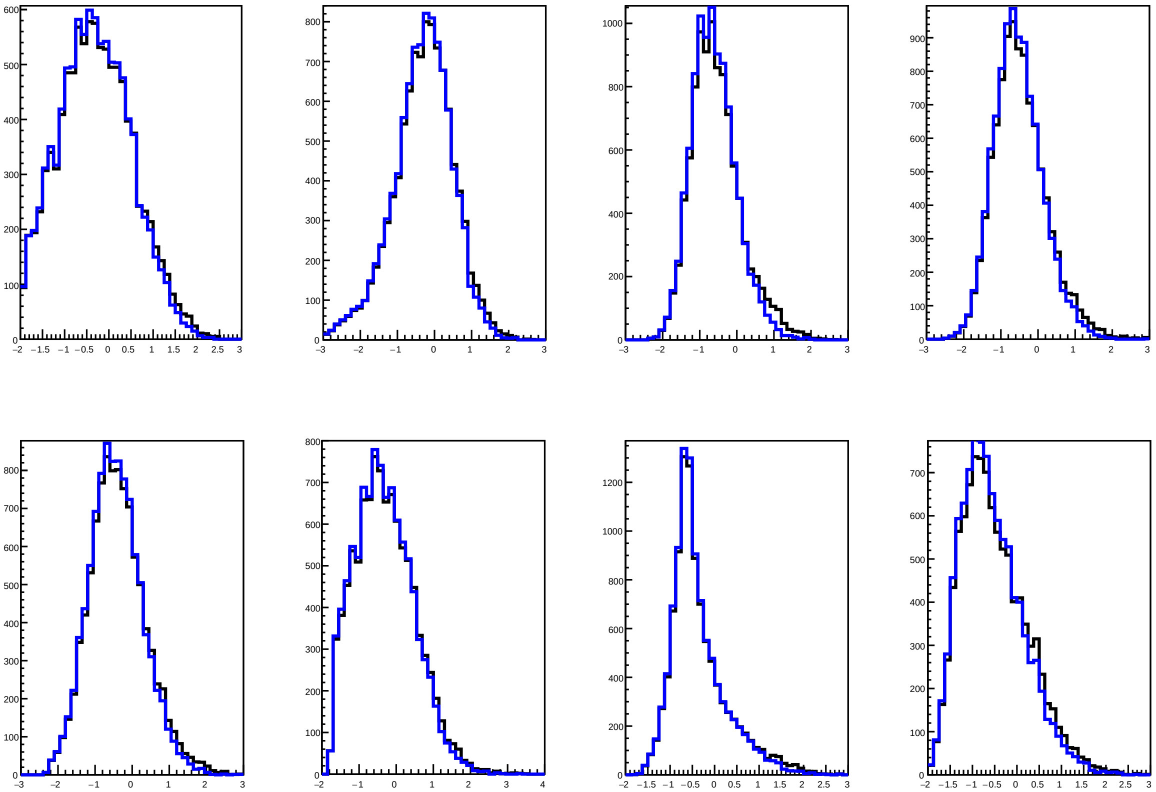


Figure 13: Comparison of the distribution of pure background (blue) and a mixture of 5% signal and background (black) in the most discriminating features in the HEPMASS dataset. Left to right, top to bottom: features 0, 3, 6, 10, 14, 18, 25, and 26.

R_1	N_{in}	N_{exp}	N_s	ϵ_s	Gain	Active features
52.78	54	0.02	46	0.184	17.04	101011100111010000000110001
45.35	48	0.06	38	0.152	15.83	000111100011001011000100011
41.60	46	0.11	33	0.132	14.35	100010010110111011000100001
40.72	46	0.13	18	0.072	7.83	101000110110101000100010011
40.38	44	0.09	41	0.164	18.64	100100100100010001110111001
40.17	47	0.17	0	0.000	0.00	011001000100010111001100011
39.82	44	0.10	0	0.000	0.00	100001010100011001010101011
38.54	44	0.14	0	0.000	0.00	001001101101110001101100000
38.36	44	0.15	30	0.120	13.91	000110101110010000001101011
38.05	43	0.13	14	0.056	6.51	110000100110001110100011001

Table 3: Results of an exploratory `RanBox` *search on the HEPMASS dataset with a 5% signal contamination; data for the 10 most significant boxes are reported. N_s indicates the number of signal events in the search box; ϵ_s is the efficiency of the box selection for the signal component; gain is computed as the increase in the signal fraction of the box over the initial dataset. For other detail see the text.*

instead 8/4950 = 0.16%. These numbers compare quite favourably to those of the neural network results graphically displayed in Fig. 12. We stress again the modest value of this observation, given the improper nature of a comparison of this kind. In particular, the `RanBox` results have unknown generalization properties —they are obtained from a single dataset, on which multiple testing is performed: the performance would be less good on a different testing sample. On the other hand, the search algorithm was only shown a total data sample of 5000 events, a number over two orders of magnitude smaller than the training sample of the neural networks.

Test	N_s/N_b	T.S. max	N_{in}/N_{exp}	N_s	Gain	$SR_{1:10}$	$\overline{\epsilon_s^{1:10}}$
1	250/4750	$Z_{PL} = 28.06$	39/0.00	35	17.95	8	0.097
2	200/4800	$Z_{PL} = 27.46$	1003/133.00	0	0.00	6	0.079
3	150/4850	$Z_{PL} = 24.45$	24/0.00	21	29.17	7	0.140
4	100/4900	$Z_{PL} = 26.95$	45/0.01	0	0.00	1	0.014
5	80/4920	$Z_{PL} = 24.65$	43/0.05	14	20.35	3	0.044
6	70/4930	$Z_{PL} = 23.56$	41/0.01	0	0.00	1	0.010
7	250/4750	$R_1 = 52.78$	54/0.02	46	17.04	7	0.092
8	200/4800	$R_1 = 50.78$	60/0.18	33	13.75	6	0.097
9	150/4850	$R_1 = 43.58$	53/0.21	15	9.44	6	0.071
10	100/4900	$R_1 = 45.29$	49/0.08	0	0.00	3	0.038
11	80/4920	$R_1 = 51.22$	52/0.02	0	0.00	0	0.000
12	70/4930	$R_1 = 43.44$	48/0.10	0	0.00	0	0.000

Table 4: Sample results of `RanBox` *runs on 5000 events from the HEPMASS dataset, with varying signal fraction and the two choices of test statistic. See the text for more details.*

We perform a test using `RanBox`, as detailed in Table 4, using 10,000 trials for the subspace sampling and a subspace dimensionality of $\mathscr{D}' = 12$. This time we start (see test 1) by searching for a 5% signal in a set of 5000 events using the Z_{PL} test statistic. The algorithm returns as the most significant box one which is rich in signal component, and we observe that the three next-best-significance boxes (not reported in Table 4) are similarly enriched in signal events. We gradually reduce the signal fraction in tests 2-6 and observe that results are not uniform: RanBox in some cases identifies as the most significant box one devoid of signal. In general we observe that the number of boxes that are signal-enriched among the first 10 ($SR_{1:10}$) usually decreases as initial signal fraction is reduced; the average signal efficiency also becomes smaller. Yet the algorithm finds significantly signal-enriched boxes among the first 10 even for an initial signal fraction of 1.4% and 1.6%. We also observe that in test 2 the Z_{PL} maximization focuses on a very wide box, an indication of the existence of broad-scale multivariate density variations of the background component of this dataset.

Based on the above observation, in tests 7-12 we turn our attention to the R_1 test statistic, which should give more importance to smaller feature-space regions. This indeed allows RanBox to converge on signal-rich regions when the signal fraction of the data sample is 3% or larger; for smaller signal fractions, however, RanBox becomes unable to evidence the signal component in the reported overdense regions.

The results also allow us to draw some conclusions on the most performing settings of RanBox to be used in the HEPMASS use case. Here, however, we stress one important point: by telling the tale of how these choices may be defined based on sample test results, we are implicitly declaring how the algorithm —but in general, we believe, any unsupervised search— requires an *ad hoc* tuning to perform its task most effectively. This is not to be taken as a demonstration that this kind of search is useless: quite on the contrary, the tool can be a very useful one in examining the properties of multi-dimensional data. It cannot, on the other hand, be employed as a catch-all machine ready to identify an anomaly in an arbitrary dataset: this is nothing else than a by-product of the well-known absence of a universal high-power test statistic, when the alternative hypothesis is not specified.

5 Converting RanBox to a Semi-Supervised Algorithm

When a signal of known kinematic characteristics is sought, but a precise model of the background is not available, the problem lays in middle ground between one of anomaly detection and one a classical supervised learning discrimination. Several methods have been proposed to allow the construction of good discriminators in this situation; for a review see [17].

Here we discuss how RanBox can be adapted to this task, which is of interest to us because of its good match with the needs of the search for the rare $B_s \to \tau\tau$ decay in CMS proton-proton collision data. Indeed, the dataset where we wish to carry out the search is a very complex one, contributed by many different low-p_T processes which Monte Carlo simulations cannot reproduce in high detail. On the contrary, the B_s production and decay mechanisms are well understood and can reliably be simulated.

For the study of the semi-supervised version of RanBox, and for the sake of this report, we still rely on the HEPMASS dataset discussed in the previous section, because a final definition of the most promising kinematic variables useful to discriminate the B_s signal from backgrounds is not ready yet, and until the program is fully tested we prefer to avoid running it on data which will be the basis of a scientific result approved by the CMS collaboration: one should always refrain from performing too many tests on real data, to avoid the chance of unconsciously biasing one's decisions and analysis choices. We thus use here again the HEPMASS data by exploiting our knowledge of the signal portion of the data, to identify a box in multi-dimensional space which is both rich in signal, and poor in background. We act as if the latter information is not available *a priori*, i.e. from a precise model of that process, and only rely on the local background density in sidebands of the search box in the construction of a useful test statistic.

The decision to not exploit the full labels of available data in the same phase space, but use label information for the two classes in different phase space regions, effectively corresponds to a semi-supervised task. It also allows one to remain blind to the real amount of data captured in the search box that the algorithm identifies at the end of the gradient descent procedure that maximizes the test statistic of choice. This enables a data-driven background prediction and avoids biases introduced unconsciously by the analysts, as the procedure is fully automated. However, as we well discuss below, a bias remains in the background estimation, due to the intrinsic correlations between the variables of the feature space, and the similarity of these correlations between the signal and background processes.

5.1 Algorithm description

The semi-supervised version of `RanBox`, which we address in the following as `RanBox_SS`, works as follows.

1. Real data where the search is to be performed is read in. We will call this dataset D1 in the following. For the tests described in this report, D1 is composed of HEPMASS simulated events belonging to the background category, but for specific studies of the performance of the algorithm we include in D1 a fraction of signal simulated events.

2. Simulated signal events are read in a dataset called D2.

3. A pruning of the variable list removes ones that are categorical, as well as ones that have too high correlation with others, using the same procedure as that described in Sec. 2.

4. Dataset D1 is used to define a variable transformation that reduces the feature space into a copula space, as in the original version of `RanBox`.

5. Events from dataset D1 and dataset D2 are both subjected to the same variable transformation. This produces transformed datasets spanning the copula feature space, where the search for the box maximizing a suitable test statistic Z is performed.

6. A test statistic is defined to maximize the search power.

7. A number of dimensions of the subspaces scanned by the algorithm is chosen. This number, $\mathscr{D}'$, should be not larger than 10-12, because of the curse of dimensionality.

8. $\mathscr{D}'$ features are chosen at random from the list of active dimensions of the space.

9. The scan of the corresponding $\mathscr{D}'$-dimensional subspace of the transformed feature space is performed by maximizing the test statistic Z via gradient descent.

10. The previous two steps are repeated a large number of times. At the end of the iteration, the algorithm reports the boundaries of the box that maximizes Z across all subspaces, and related statistics.

A number of details need to be discussed in order to clarify the above procedure. We address them below.

5.1.1 Variable transformation

Because the variable transformation, based on the integral transform, is defined on dataset D1 but is then applied to both datasets D1 and D2, a prescription needs to be given for how to handle variable values which fall outside of the original range of their distribution in dataset D1. For example, it might happen that dataset D1 only contains events for which variable 1 is in the $[-100., 500.]$ range, but dataset D2 has a distribution of variable 1 which extends in the wider range $[-150., 600.]$. This might happen if the signal process produces observable features that are exceedingly rare in backgrounds. Since the range $[-100., 500.]$ is mapped by the integral transform into $[0, 1]$, one needs to define transformed values for variable 1 in the range $[-150., -100.]$ or $[500., 600.]$. This is simply done by assigning transformed value 0 to all values below $-100.$ and transformed value 1 to all values of variable 1 above 500.. This ensures that the copula space remains unaltered when the signal component is considered. However, one must keep note of the fact that such a situation complicates any extrapolation of the data density from sidebands to signal region, when the signal region includes the singular points at 0. or 1: a bias in the sidebands-driven background prediction can be expected in these situations.

5.1.2 Test statistic

Our focus in this work is to produce a suitable methodology that can be applied to the search for a rare process which we do not expect will be observable in the available data. In fact, current estimates for the branching fraction of the B_s meson decay to tau lepton pairs put this value below 10^{-7}, which corresponds to less than one event in the available LHC luminosity once one accounts for the unavoidably small collection efficiency for identifiable signal events[11].

Because of the above, it makes little sense to maximize a test statistic which is monotonous with the significance of a signal component in the data. Rather, it appears reasonable to define a test statistic which minimizes the upper limit on the signal process under study. This is in fact the reachable scientific objective of the B_s search under way by the CMS collaboration in currently available data.

We therefore consider a counting experiment where a predicted background N_{exp} is compared to an observed event count N_{obs}, the former extracted from a sideband of equal feature space volume to the signal box, possibly after a bias correction (see *infra*). The observed event counts N_{obs} is expected to be sampled from the background prediction, i.e. from a Poisson distribution centered on N_{exp}. If we define a type-I error rate $\alpha = 0.05$, we can compute the expected upper limit at a confidence level $1 - \alpha$ on the number of signal events contributing to the signal box, when N_{exp} is predicted and N_{obs} is observed. An exact calculation involves extracting the tail integral of a Poisson distribution, and is sometimes impractical to perform (when N_{exp} is large). `RanBox_SS` uses a very good approximation given by

$$N^{up} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha) - N_{exp} \tag{22}$$

[11] The data where the search is performed is collected by a trigger which selects events with a muon candidate of transverse momentum above 7-9 GeV (depending on data taking period and running conditions). Once one accounts for the braching ratio of tau decays into muons, the small probability of this giving rise to a muon above the stated momentum threshold, and the additional requirements on the other tau lepton decay, which must include three charged pions, the signal efficiency becomes smaller than 10^{-3}; together with a 10^{-7} $B_s \to \tau\tau$ branching fraction, this means that one reconstructable event corresponds to over 10 billion B_s mesons, which correspond to an integrated luminosity much larger than what is available for the present search.

where F_{χ^2} is the cumulative distribution function of Chi-squared distribution with $2(N_{obs}+1)$ degrees of freedom. We may then convert the upper limit above into an absolute upper limit on the cross section times branching fraction of the studied process by the following formula:

$$\sigma B(B_s \to \tau\tau) = \frac{N^{up}}{L\epsilon_{tot}},\ 95\%\ C.L. \tag{23}$$

Above, the signal efficiency ϵ_{tot} is computed by multiplying a pre-selection efficiency ϵ_{presel} (obtained from the Monte Carlo simulation) that describes the probability that a $B_s \to \tau\tau$ signal event gets included in the analyzed sample, by the fraction of events captured in the signal box ϵ_{box}.

The procedure to define the test statistic which we wish to maximize therefore is to extract from a signal box the number of events N_{sig} in dataset D2 that are contained within it, and to assess the number of expected background events N_{exp} in the same signal box, by counting how many D1 events are collected in a properly defined sideband. Using N_{exp} we obtain, by the formula (22), the 95% C.L. upper limit on the signal, N^{up}; With N_{sig} we compute the signal efficiency ϵ_{box}; and with these inputs we may then define the test statistic to be maximized as

$$Z_{UL} = \frac{\epsilon_{box}}{N^{up}} \tag{24}$$

whose maximization explicitly minimizes the upper limit on the signal cross section times branching fraction.

It is clear what an algorithm tasked with maximizing Z will need to do: find a region of space which contains a large number of signal Monte Carlo events from dataset D2, while having as small as possible predicted contributions from dataset D1. The latter comes from a "non-local" estimate, one derived from a sideband constructed exactly as described for `RanBox` in Sec. 2.

5.1.3 Gradient Descent

The gradient descent procedure used by `RanBox_SS` when studying each subspace of the copula is the same of that of `RanBox`, and it has been already described in Sec. 2. However, due to the fact that the results of `RanBox_SS` are more affected by a biased background prediction, we implement for it a validation technique based on an early stopping criterion. Datasets D1 and D2 are both split evenly into a training and a validation subset, and only the training subset is used for the maximization of the test statistic by the gradient descent procedure. During the procedure, however, the algorithm keeps track of the value of the test statistic on the validation subset of datasets D1 and D2. At the end of the routine, the box which produces the highest value of the test statistic on the validation sample is returned as the best one for the considered subspace.

5.1.4 Box boundaries and sidebands

It should be clear that the initialization of the search box boundaries must be driven by the density that can be assessed from the part of the data from which a local density estimate can be obtained, *i.e.* dataset D2, the signal Monte Carlo. Algorithm 2, described in Sec. 2.5 above, can fulfil that task effectively, and is used for `RanBox_SS`.

The choice we made for `RanBox_SS` of using a sideband of volume equal to the signal box to estimate the background in the signal region is dictated by the need of an estimate affected by low bias. Indeed, bias is a much worse enemy than variance in this particular application, as large correlations between the variables of the feature space have the potential of making any sideband-derived estimate completely unreliable. While non-local, a sideband estimate from events which lie very close to the signal box will suffer a manageable bias even in the presence of large correlations. However, the gradient descent procedure which maximizes the test statistic defined above explicitly tries to shrink the value of N_{exp}, by moving to regions where dataset D1 suffers negative fluctuations. Hence a strong negative bias on that number is anyway expected. We sidestep this problem by constructing a second sideband around the sideband used for the calculation of the test statistic. This second sideband is only used for the final estimate of N_{exp}, and should be unaffected by the gradient descent procedure.

At variance with `RanBox`, in `RanBox_SS` we enforce that sidebands (as well as the second sidebands described below) have a volume exactly equal to that of the signal box. This is a useful property when we need to characterize possible biases in the extrapolation procedure, as the extrapolation factor is always equal to 1.0 and thus is one less parameter to consider in such bias studies. In order to enforce that the sideband has a volume exactly equal to the signal box, we devise an iterative algorithm, described below.

1. The widening factor required to construct a box of volume twice larger than that of the signal box is computed as $f = 2^{1/\mathcal{D}'}$, such that if each side of the signal box were widened by a factor f, the resulting box would have a volume equal to twice the signal box.

2. A loop on the subspace dimensions is performed, and for each dimension the signal box extension B_i and the available region A_i left in the $[0, 1]$ interval once the signal box interval is excluded is stored. So, *e.g.*, if the signal box has an extension of $B_1 = 0.3$, having intervals $[0.2, 0.5]$ in variable 1, the available region on variable 1 is $A_1 = 1. - 0.3 = 0.7$.

3. Available region values are sorted in increasing order.

4. Starting with the smallest available region, the algorithm assigns sideband intervals to each variable i by comparing A_i to $f \times B_i$. If A_i is larger than $f \times B_i$, the interval defining the sideband in dimension i is simply defined by extending the box interval by a factor f (an attempt to construct a symmetric interval is made, and if there is not enough space in one of the sides of the signal box, all the extra space required to extend the interval to $f \times B_i$ is assigned to the sideband on the other side). If, on the other hand, A_i is smaller than $f \times B_i$, the sideband on direction i is defined as $[0, 1]$, and the factor f required to each additional variable to obtain a sideband of volume twice larger than the signal box is recomputed as $f = (2B_i)^{1/(D'-1)}$. A similar rescaling is operated at each successive iteration until the considered A_i grows larger than the current f value.

5. The iteration on every variable continues, until all dimensions of the subspace have been included in the sideband definition. The procedure converges to a sideband of volume equal to twice the signal box (and thus a surrounding region of volume equal to the signal box, once the signal box is vetoed) unless the signal box has a volume larger than 0.5, which is however not allowed in any step of the program (the initialization algorithms, as well as all box extensions in the gradient descent routine, enforce that the signal box has a volume not larger than 0.25 in `RanBox_SS`).

In exact similarity to the algorithm described above, the second sideband is defined as a multi-dimensional interval in the considered $\mathcal{D}'$-dimensional subspace, of total volume exactly equal to three times the signal box, and thus also equal to 1.5 times the sidebands box. Once events in the second sidebands are vetoed if they are contained within the first sideband, the effective volume of the second sideband is equal to that of the signal box, hence the extrapolation factor required to predict the number of events in the signal box from the second sideband is equal to 1.0. A modification of this factor may be required if an estimate of bias is obtained, as discussed below.

5.2 Sample results on the HEPMASS dataset

Below we detail results of running RanBox_SS on the continuous features of the HEPMASS dataset. As it was shown in the previous sections, the signal in that dataset is rather easy to isolate from backgrounds, due to its distinguishing mass-related features: multi-object invariant masses and energy of objects are all high for the signal component. This is quite different from what will be the case of the B_s search, which unfortunately features a signal very difficult to distinguish from backgrounds. We do not expect any variable to have nearly as strong a discrimination power as the most sensitive variables of the HEPMASS dataset. To make the HEPMASS data a better testbed of our algorithm, therefore, we remove from the list of 27 features not only the five categorical variables it contains (variables 6, 10, 14, 18, 22), which would complicate the preprocessing step of the data, but also six of the most discriminating features: variables 4, 7, 11, 15, 26, and 27 (refer to Fig. 11). We are thus left with a set of 16 features, which are in the same ballpark of the dimensionality of the space of discriminating variables we will use for the B_s search, and provide a discrimination problem of similar complexity to the one we are targeting.

To test the working of RanBox_SS, a run maximizing the test statistic Z_{UL} is performed on a dataset D1 composed of 5,000 background events, and a dataset D2 composed of 5000 signal events, by searching in 10,000 different subspaces. The results for the 10 best boxes are shown in Table 5 below.

Box	Z_{UL}	N_{obs}	N_{exp}	Volume	ϵ_{box}	Features
1	10.57	8	12	0.0024	0.044	0 1 2 4 5 6 8 9 10 11 12 14
2	10.39	17	8	0.0040	0.052	0 2 4 5 7 8 9 11 12 13 14 15
3	10.28	12	13	0.0044	0.059	0 1 2 4 5 6 8 9 10 11 12 15
4	10.13	15	8	0.0038	0.055	0 1 2 4 5 7 8 9 11 12 14 15
5	10.01	12	10	0.0034	0.052	0 1 4 5 6 7 9 10 11 12 14 15
6	9.97	18	5	0.0027	0.044	2 3 4 5 6 7 8 10 11 12 13 14
7	9.97	18	5	0.0027	0.044	1 2 4 5 6 8 10 11 12 13 14 15
8	9.97	18	5	0.0027	0.044	2 3 4 5 6 7 8 9 10 12 13 14
9	9.81	15	8	0.0041	0.056	0 2 3 4 7 8 9 10 11 12 13 14
10	9.77	16	9	0.0037	0.043	2 3 6 7 8 9 10 11 12 13 14 15

Table 5: Results of a maximization scan of 5000 subspaces of the HEPMASS feature space, with a D1 dataset composed of 5000 background events, and a D2 dataset made of 5000 signal events. N_{obs} is the number of D1 events in the signal box. The best identified signal boxes are ordered by decreasing value of the Z_{UL} test statistic, whose value is inversely proportional to the estimated $95\% CL$ upper limit on signal cross section achievable by a counting experiment.

The Z_{UL} values listed in Table 5 correspond to upper limits on the signal cross section of $\sigma_{95\%CL} = 1000/(LZ_{UL}\epsilon_{presel})$, where ϵ_{presel} is the fraction of signal events that would be included in the original dataset before the RanBox_SS search, and L is the integrated luminosity corresponding to the analyzed data. For a preselection efficiency of $\epsilon_{presel} = 0.1$, *e.g.*, and an integrated luminosity $L = 1fb^{-1}$, the tabulated highest value of Z_{UL} corresponds to an upper limit of $\sigma_{95\%CL} = 946fb$.

A different test is performed by searching in 5,000 subspaces of a dataset D1 composed by 9500 background and 500 signal events, with a corresponding dataset D2 of 10,000 signal events. Having injected signal in D1, we can check the increase in signal purity of the returned boxes, and observe that the upper limit becomes higher, as the maximum value of Z_{UL} reached is smaller. The results for the five best boxes are shown in Table 6 below. One observes that the maximization of Z_{UL} corresponds to a significant increase in the signal over background fraction of the selected portion of dataset D1, as shown by the second-to-right column.

Box	Z_{UL}	N_{obs}	N_{exp}	$N_{s,in}$	Volume	ϵ_{box}	S/N gain	Features
1	7.13	59	37	22	0.0085862	0.08	7.45783	0 1 2 3 5 6 7 8 10 11 12 14
2	7.1	59	37	22	0.0085004	0.08	7.45783	0 1 2 3 5 6 7 8 10 11 12 15
3	7.01	61	41	22	0.00924	0.08	7.21331	0 1 2 3 4 5 6 7 8 9 13 15
4	6.97	58	32	21	0.0079897	0.08	7.24159	0 1 2 3 4 5 8 9 10 12 13 15
5	6.56	58	31	22	0.00858	0.08	7.58641	0 2 3 4 5 6 8 9 11 12 13 15

Table 6: Results of a maximization scan of 5000 subspaces of the HEPMASS feature space, with a D1 dataset composed of 9500 background events and 500 injected signal events, and a D2 dataset containing 10,000 signal events. N_{obs} is the number of D1 events in the signal box, and $N_{s,in}$ is the number of signal events in dataset D1 captured in the signal box. The best identified signal boxes are ordered by decreasing value of the Z_{UL} test statistic, whose value is inversely proportional to the estimated 95% C.L. upper limit on signal cross section achievable by a counting experiment.

The results of a `RanBox_SS` scan can also be visualized graphically, as shown in Figs. 14 and 15.

5.3 Bias studies

As we discussed above, the estimate of events from dataset D1 in the signal box with events in the second sideband is expected to be negatively biased, due to the intrinsic correlations between kinematic variables defining the feature space. These correlations in many cases affect the signal, which drives the maximization of the numerator of Z_{UL}, and the background in a similar manner – both have to withstand to physical constraints between their kinematical features. We may define the bias as follows:

$$b = 2(N_{obs} - N_{exp})/(N_{obs} + N_{exp}) \tag{25}$$

A principled way to estimate the above bias in N_{exp} is to define a set of alternative signals, simulate their characteristics, obtain a set of alternative datasets $D2_1$, $D2_2$, $D2_3$,... and run `RanBox_SS` on each separately, maximizing Z_{UL} against the same background D1. The resulting mean and variance of the distribution of ratios between observed events in the signal box and expected events in the second sideband are then sound estimators of the bias and its variability, and they can be used to correct the prediction N_{exp} in the case of the original datasets D1 and D2. These alternative datasets might, *e.g.*, be constructed by artificially changing the mass of the tau leptons in the simulation, or the mass of the B_s meson, or even the mass of the ρ meson, as the latter is an almost certain intermediate state in the $\tau \to \rho\pi^- \to \pi^+\pi^-\pi^-$ decay.

Since we have been testing `RanBox_SS` on a different sample of data from the one which is our target, we study here a more general technique which is less dependent on the specific kinematic properties of the datasets. The technique consists in searching, for each signal box identified by

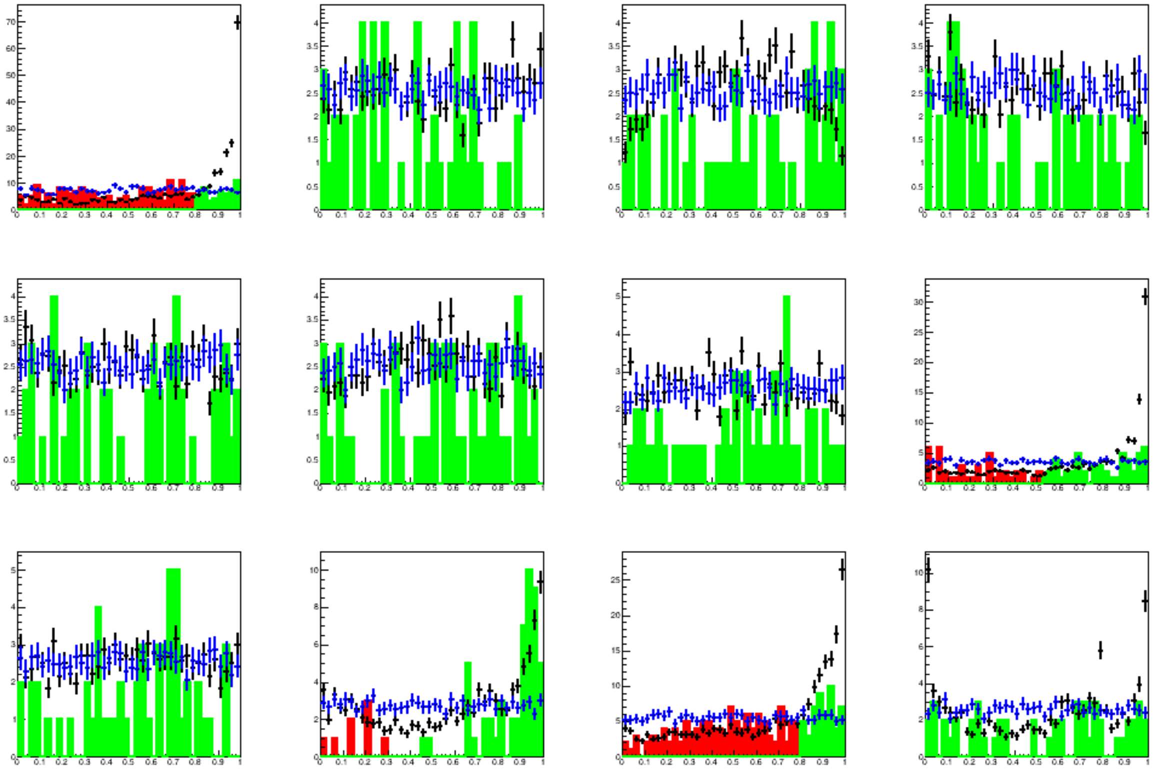


Figure 14: Marginal distributions of the 12 features in the subspace where `RanBox_SS` finds the best box, for a large-statistics run. The green distributions show the selected data, the red distributions show data that are rejected only because of their value of the shown feature (they would otherwise be included in the signal box); the blue distributions show the original unselected dataset D1 (rescaled by an arbitrary factor to fit in the graph), and the black distributions show the unselected dataset D2 (also arbitrarily rescaled). See the text for more detail.

`RanBox_SS` after gradient descent maximization of Z_{UL}, a corresponding alternative region of the considered subspace of the feature space, and correspondingly a sideband and a second sideband to it. The requirements of such an alternative region are the following:

- It must have the same extension in each subspace dimension as the original signal box;
- It must have no overlap with the original signal box;
- It must contain a number of D1 events in the signal box similar to the number of the original signal box.

The alternative box is sought for by random trials, by changing the location of the multi-dimensional interval while keeping its shape unaltered. This is a time-consuming procedure, as it may prove very difficult to fulfil the above criteria, especially if the definition of ”similarity” in the observed event counts from dataset D1 in the signal boxes is too strict. We have observed that, with typical number of events and dimensionality of the subspaces in runs on the HEPMASS dataset, the identification of a box with the above characteristics typically requires less than 2000 trials; in few cases, when `RanBox_SS` has identified by gradient descent a unusually dense and small region of phase space,

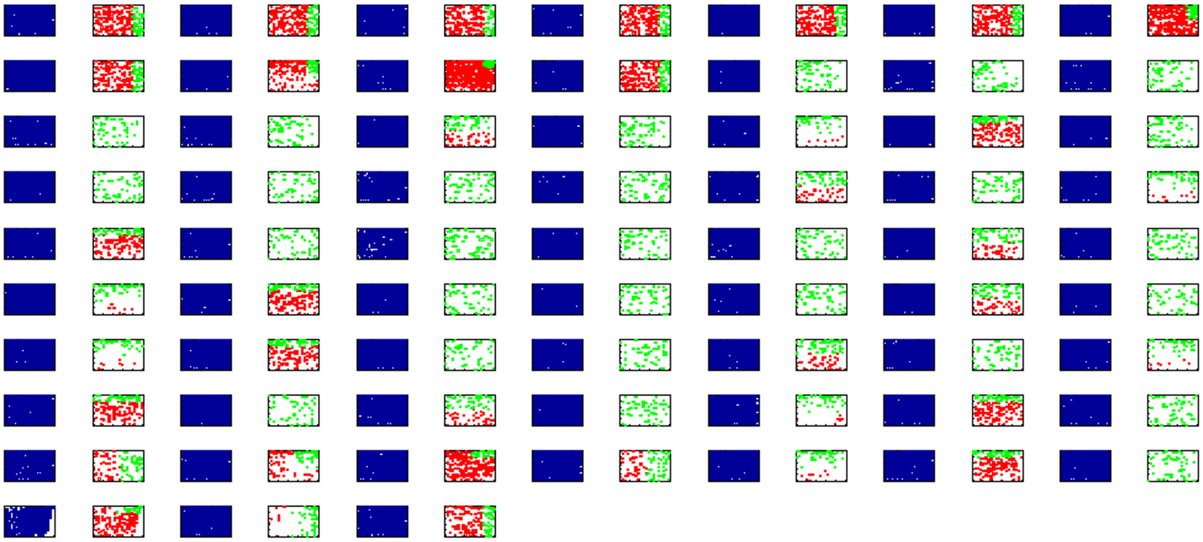


Figure 15: Scatterplots of the 12 features in the copula space for a high-statistics run. Each pair of graphs shows a two-dimensional subspace of the 12-dimensional space where `RanBox_SS` finds the best box. The blue distribution shows data in the D1 dataset before any selection; the corresponding distribution on its right shows the selected data in the best box (in green) and the data that would have been included in the best box if it did not fail the selection on the two displayed features (in red). The 66 pairs of distributions show variables 1 vs 2, 1 vs 3, 1 vs 4, ... 11 vs 12.

which cannot easily be replicated by random sampling, the required iterations diverge. A workable criterion of ”similarity” is to impose that the difference between the observed event counts in the two signal boxes be smaller than 10% of their average value. We have observed that the bias estimates depend very little on the precise value of this criterion.

A run on 10,000 $\mathscr{D}'$-dimensional subspaces of the HEPMASS feature space, using 5000 events in dataset D1 (only composed of background events) and 5000 events in dataset D2, allows to verify the soundness of the above bias estimation procedure. The results are shown in Fig. 16 and reported in Table 7 below.

	Bias	SQM
Original box	0.2685 ± 0.0039	0.392
Alternative box	0.3452 ± 0.0038	0.378

Table 7: Extrapolation bias and its estimate with random boxes. See the text for details.

The estimated and real bias are different, but the difference in their means is not very large. A larger systematic effect on the extrapolation than the one due to the difference in mean biases above is potentially due to the variability of the bias, which is only partly explained by the statistical fluctuation of the observed and expected event counts in the boxes [12]; however, by re-sampling

[12] The typical number of events in the 10,000 signal boxes studied in this run is 37.3 (with a RMS of 36.5); in such conditions, and with the average bias of 0.27 mentioned above, this translates into a typical variability of bias estimates of about 0.21 due to Poisson statistics.

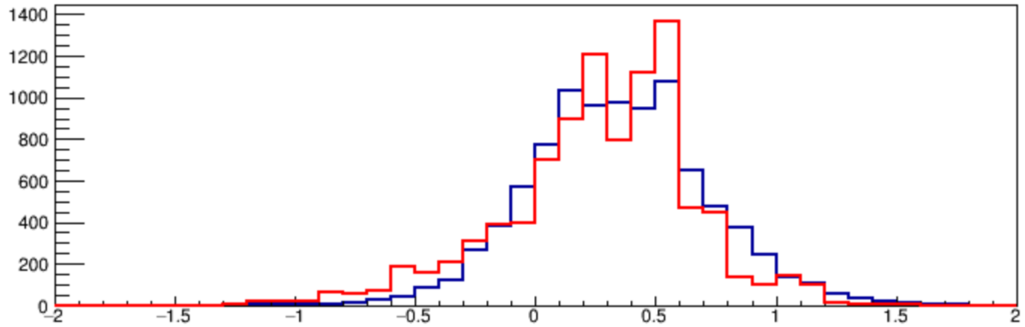


Figure 16: Comparison of the distribution of bias (defined as in Eq. 25) in 10,000 signal boxes returned by the gradient descent search in a HEPMASS data sample of 5000 signal and 5000 background events, with $N_{var} = 12$ (red), with the bias estimated in the alternative boxes by the procedure described in the text (blue).

multiple times random boxes of characteristics similar to the one obtained by gradient descent, it may be possible to reduce this effect.

The above study indicates that the procedure may in fact provide a viable correction to the background estimate provided by the second sideband. Of course, these results are strongly dependent on the characteristics of the studied problem (the inner correlations in the feature space), therefore a separate assessment needs to be carried out in every case. The procedure to handle large biases, which are however not expected in the case of the B_s search due to the less striking characteristics of the signal in that situation, is to run a Z_{UL} maximization to derive a bias estimate, and then to correct the calculation of Z_{UL} including it as a factor in the denominator of that test statistic, so that a second run may converge more precisely to the most advantageous signal region. More studies are needed to finalize this procedure.

6 Conclusion and Outlook

The search for anomalous regions of a complex feature space can be performed proficuously if the space is transformed into a standardized copula, where the marginal density of every feature is uniform. This allows to identify a multi-dimensional interval which captures unusual overdensities, possibly due to anomalous contaminations of the data sample. We have designed an algorithm that performs this search, `RanBox`, and demonstrated that it has considerable power in locating anomalous signals.

In the second part of this document we have shown how to customize `RanBox` to search for a specific, well-defined signal in data that are otherwise hard to model. In this semi-supervised version the algorithm, `RanBox_SS`, is designed to minimize the upper limit on the signal cross section extractable from the identified multi-dimensional interval by a counting experiment that uses as a background prediction the number of data events captured in a suitable sideband in the multi-dimensional space.

Tests of the algorithm show that it is a viable procedure for the search of the B_s meson in LHC collisions data. Future work will allow us to define in an optimal way a feature space where to run `RanBox_SS` and obtain a stringent upper limit on the cross section of that process, which is currently still beyond observability with available LHC data.

References

[1] https://www.merriam-webster.com/dictionary/anomaly.

[2] The CMS Collaboration, *The CMS Experiment at the CERN LHC*, Journ.Inst. 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[3] The ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, Journ. Inst. 3 (2008) S08003, doi:10.1088/1748-0221/3/08/S08003.

[4] S.L. Glashow, *Partial-symmetries of weak interactions*, Nucl. Phys. 22 (1961) 579, doi:10.1016/0029-5582(61)90469-2; S. Weinberg, A Model of Leptons, Phys. Rev. Lett. 19 (1967) 1264, doi:10.1103/PhysRevLett.19.1264; A. Salam, Weak and electromagnetic interactions, in Elementary Particle Physics: relativistic groups and analyticity, N. Svartholm, ed., p.367. Almqvist & Wiskell, 1968. Proceedings of the 8th Nobel symposium.

[5] ALEPH, CDF, D0, DELPHI, L3, OPAL, SLD Collaborations, the LEP Electroweak Working Group, the Tevatron Electroweak Working Group, and the SLD Electroweak and Heavy Flavour Groups, *Precision Electroweak Measurements and Constraints on the Standard Model*, CERN PH-EP-2010-095 (2010).

[6] T. Dorigo, M. Fumanelli, C. Maccani, M. Mojsovska, G.C. Strong, and B. Scarpa, RanBox: Anomaly Detection in the Copula Space, arXiv:2106.05747 [physics.data-an] (2021).

[7] CMS Collaboration, *Search for contact interactions and large extra dimensions in the dilepton mass spectra from proton-proton collisions at $\sqrt{s} = 13\ TeV$*, JHEP 04 (2019) 114, arXiv:1812.10443 [hep-ex], doi:10.1007/JHEP04(2019)114.

[8] A. Sklar, *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst. Statist. Univ. Paris, 8 (1959) 229.

[9] R.E. Bellman, Rand Corporation *Dynamic programming*, Princeton University Press (1957), p. ix. ISBN 978-0-691-07951-6.

[10] T.P. Li and Y.Q. Ma, *Analysis methods for results in gamma-ray astronomy*, Astroph. Journ. 272 (1983) 317, doi:10.1086/161295.

[11] C.E. Bonferroni, *Teoria statistica delle classi e calcolo delle probabilità*, Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

[12] https://root.cern.ch.

[13] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing* (second ed.) (1992). Cambridge University Press, ISBN 0-521-43108-5.

[14] https://archive.ics.uci.edu/ml/index.php.

[15] P. Baldi, P. Sadowski, and D. Whiteson, *Searching for Exotic Particles in High-Energy Physics with Deep Learning*, Nature Comm. 5 (2014) 4308, doi:10.1038/ncomms5308.

[16] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, *Parameterized Machine Learning for High-Energy Physics*, Eur. Phys. Journ. C76,5 (2016) 7, doi:10.1140/epjc/s10052-016-4099-4.

[17] T. Dorigo and P. de Castro Manzano, *Dealing with Nuisance Parameters using Machine Learning in High Energy Physics: a Review*, arXiv:2007.09121 [stat.ML], 2020.