SUPPLEMENTARY MATERIAL FOR MODEL-INDEPENDENT DETECTION OF NEW PHYSICS SIGNALS USING INTERPRETABLE SEMI-SUPERVISED CLASSIFIER TESTS

BY PURVASHA CHAKRAVARTI[1], MIKAEL KUUSELA[2,*] JING LEI[3,‡] AND LARRY WASSERMAN[2,†]

[1]*Department of Mathematics, Imperial College London, p.chakravarti@imperial.ac.uk*

[2]*Department of Statistics & Data Science and NSF AI Planning Institute for Physics of the Future, Carnegie Mellon University, *mkuusela@andrew.cmu.edu; †larry@stat.cmu.edu*

[3]*Department of Statistics & Data Science, Carnegie Mellon University, ‡jinglei@stat.cmu.edu*

1. Exploratory Analysis of the Higgs Boson Data. In this section we explore the Higgs boson data that is used for the experiments in Section 5 of the main paper (Chakravarti et al., 2021b). The Higgs boson machine learning challenge data set is available on the CERN Open Data Portal at `http://opendata.cern.ch/record/328` (ATLAS collaboration, 2014). The data set consists of simulated data provided by the ATLAS experiment at CERN's Large Hadron Collider to optimize the search for the Higgs boson.

We analyze the data set provided by the challenge that has 818,238 observations, where each observation is a simulated proton-proton collision event in the detector. The data set contains $d = 35$ features whose individual details can be found on CERN's Open Data Portal or in Appendix B of Adam-Bourdarios et al. (2015). The data contain information on the properties of the *jets*, which are clustered showers of hadrons, and other objects produced during the collision.

As mentioned in the main paper (Chakravarti et al., 2021b), the data contains primitive "raw" features (names starting with `PRI`) that are measured by the detector, and derived features (names starting with `DER`) that are functions of the primitive features. We use and analyze just the primitive variables ($d = 16$), since the derived features are just functions of the primitive features. We additionally only use the events that produce at least two jets in the collisions (i.e., `PRI_jet_num`$= 2$) in order to avoid structurally absent missing values as mentioned in the main paper (Chakravarti et al., 2021b). This results in 165,027 events; 80,806 background events and 84,221 signal events. Descriptions of the primitive variables used are provided in Table 1.

Among the primitive features, five of them provide the azimuth angle ϕ of the objects generated in the collsion (variables ending with `_phi`). These features are rotation invariant in the sense that the event doesn't change if all of them are rotated together by some angle. The first row of Figure 1 demonstrates the uniform distribution of the ϕ variables. So, the ϕ variables themselves do not contain any information, but the difference between the angles is what contains the information. Hence to interpret these variables more easily using the active subspace methods, we remove the invariance of the azimuth angles by rotating all the ϕ's and setting the azimuth angle of the leading jet at 0 (`PRI_leading_phi`$= 0$). So the new ϕ variables give the difference between the azimuth angles of the objects and the leading jet. The bottom row of Figure 1 demonstrates the importance of the change in the distribution of the angles after the rotation. The symmetry of the distributions is expected as a difference of π radians is the same as a difference of $-\pi$ radians.

TABLE 1

Descriptions of the variables used in the analysis of the Higgs boson machine learning challenge data set (Adam-Bourdarios et al., 2015). Since we only use the primitive variables, we drop the pre-fix `PRI` *from the variable names and further shorten some of the variable names intuitively for convenience.*

Variable	Description
`tau_pt`	The transverse momentum of the hadronic tau.
`tau_eta`	The pseudorapidity η of the hadronic tau.
`tau_phi`	The azimuth angle ϕ of the hadronic tau.
`lep_pt`	The transverse momentum of the lepton (electron or muon).
`lep_eta`	The pseudorapidity η of the lepton.
`lep_phi`	The azimuth angle ϕ of the lepton.
`met`	The missing transverse energy.
`met_phi`	The azimuth angle ϕ of the missing transverse energy.
`met_sumet`	The total transverse energy in the detector.
`lead_pt`	The transverse momentum of the leading jet, i.e., the jet with the largest transverse momentum (undefined if `PRI_jet_num` $= 0$).
`lead_eta`	The pseudorapidity η of the leading jet (undefined if `PRI_jet_num` $= 0$).
`lead_phi`	The azimuth angle ϕ of the leading jet (undefined if `PRI_jet_num` $= 0$).
`sublead_pt`	The transverse momentum of the subleading jet, i.e., the jet with the second largest transverse momentum (undefined if `PRI_jet_num` ≤ 1).
`sublead_eta`	The pseudorapidity η of the subleading jet (undefined if `PRI_jet_num` ≤ 1).
`sublead_phi`	The azimuth angle ϕ of the subleading jet (undefined if `PRI_jet_num` ≤ 1).
`all_pt`	The scalar sum of the transverse momentum of all the jets in the event.
Weight	The event weight.
Label	The event label (string) (s for signal, b for background).

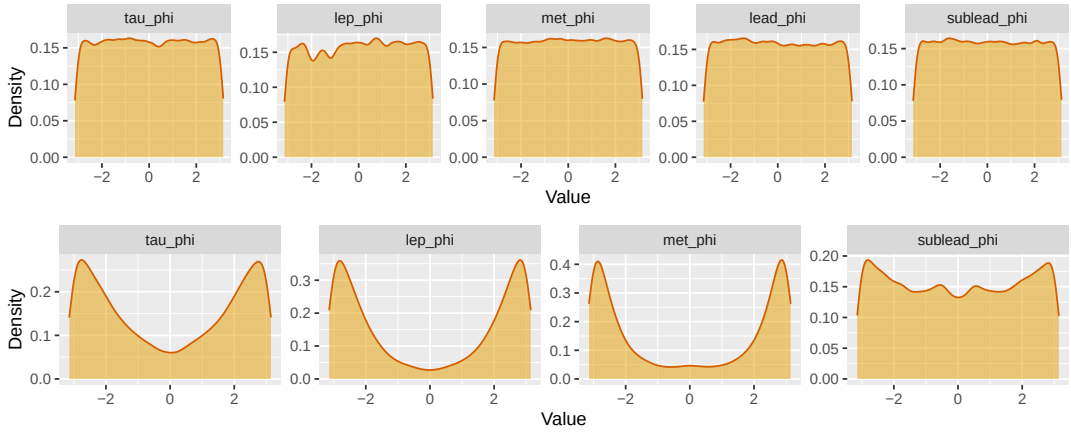


Fig 1: Top row gives the azimuth angles (ϕ) before rotation. Bottom row gives the azimuth angles after rotation such that the angle of the leading jet is set to 0 and the angles of the other objects give the difference of the azimuth angle between the object and the leading jet.

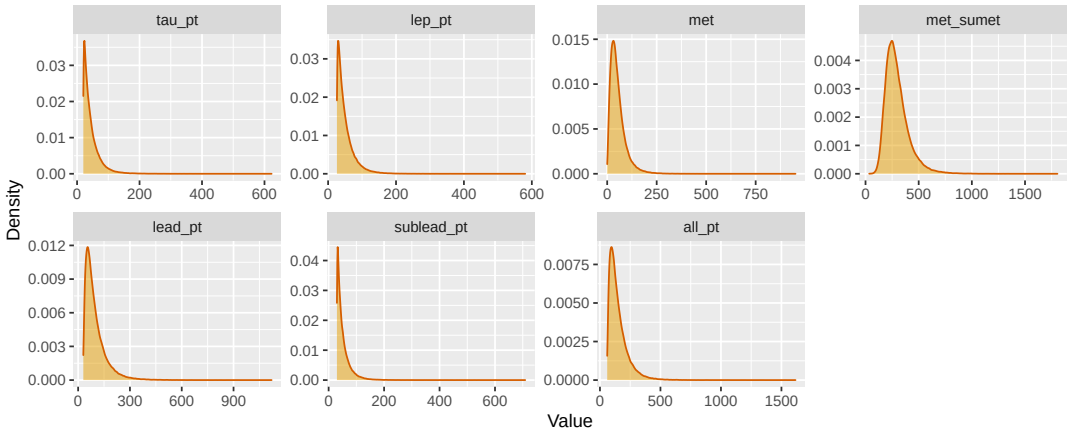


Fig 2: Distributions of the transverse momentums of the particles produced, the missing transverse energy and the total transverse energy in the detector. These are the variables for which we consider a log transformation due to the skewness in their distribution.

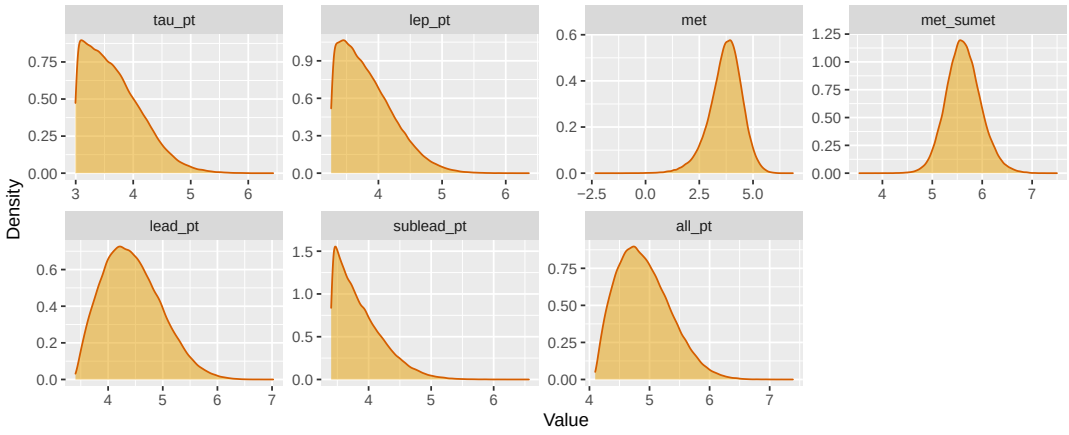


Fig 3: Distributions after a log transformation of the transverse momentums of the particles produced in the collision, the missing transverse energy and the total transverse energy in the detector.

Additionally, we take logarithmic transformations of the variables that give the transverse momentum of the particles produced (variables ending with `_pt`), the missing transverse energy (`PRI_met`) and the total transverse energy in the detector (`PRI_met_sumet`). This ensures that our analyses are not affectced by the skewness demonstrated by these variables in Figure 2. Taking a log transformation of these variables in Figure 3 fixes the problem upto some extent.

Our goal is to detect the presence of the Higgs boson signal in the experimental data, using this data set. The difficulty of this problem is demonstrated by Figure 4 which shows that the distributions of the signal and the background data are not very different. Particularly, when we are searching for signal that is just around 15% of the experimental data or even less, these minute differences are difficult to detect. Note that the experimental data is a mixture of the background and the signal data, so with just 15% of the experimental data being signal, the distribution of the experimental data appears to be almost indistinguishable from

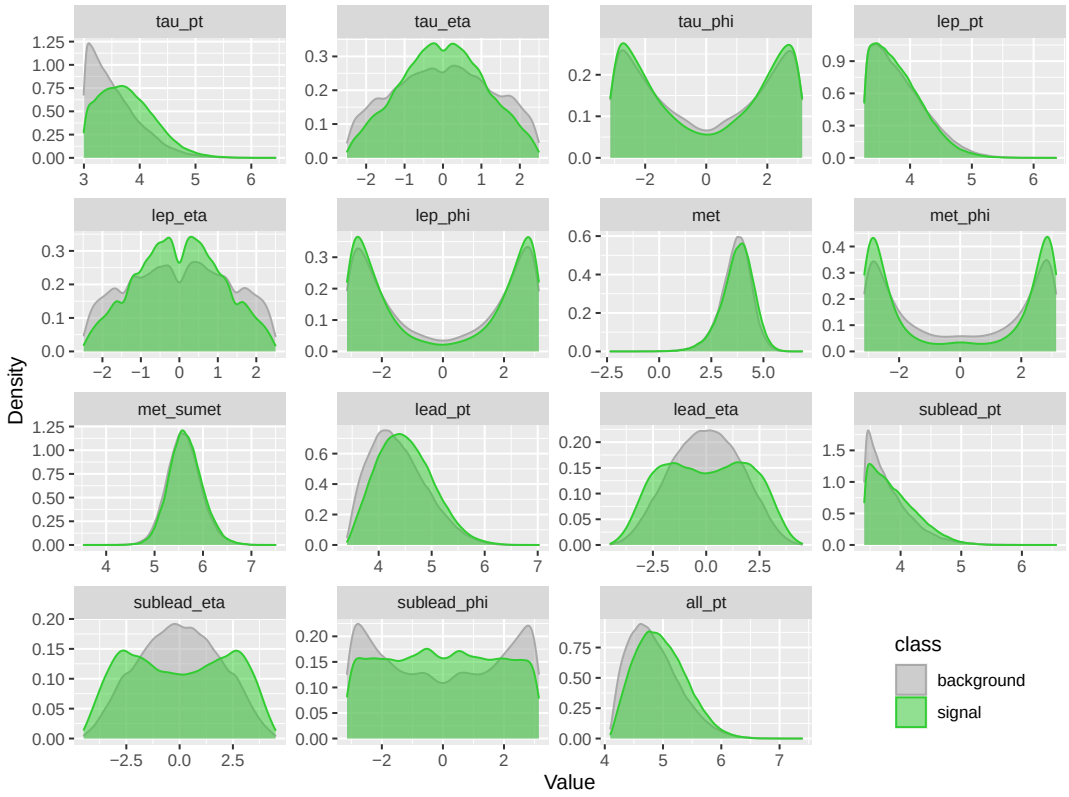


Fig 4: Histograms of the 15 variables for signal (green) data as well as background (grey) data.

the distribution of just the background data. But as seen in Section 5.2 of the main paper (Chakravarti et al., 2021b), in the case where $\lambda = 0.15$, almost all the semi-supervised tests conclude that the experimental data's distribution is different from the background data's distribution, hence detecting a signal in the experimental data. In the next section, we explore the output of the random forest classifier that detected the signal successfully in the experimental data in one of the 50 simulations explored in Section 5.2 when $\lambda = 0.15$, i.e., 15% of the experimental data is from the signal sample.

2. Experimental Data when Signal Strength $\lambda = 0.15$. As described in Section 5 of Chakravarti et al. (2021b), for the semi-supervised model-independent tests that detect the presence of the Higgs boson signal in the experimental data, we consider a training sample of $m_1 =$ 20,403 background events and $n_1 =$20,403 experimental events to train a random forest classifier to differentiate between the background and the experimental events. We then test for the presence of signal using a test sample of $m_2 =$20,000 background events and $n_2 =$20,000 experimental events. Both the training and the test experimental samples contain signal events with probability $\lambda = 15\%$. That is, the number of signal events in the training and the test experimental data is randomly distributed as $\text{Bin}(n_1, \lambda)$ and $\text{Bin}(n_2, \lambda)$ respectively, where $\lambda = 0.15$. Note that the trained classifier successfully differentiates between the training background sample and the training experimental sample, which differ from each other very slightly, as can be seen in Figure 5, since the experimental data is a mixture of background and signal events. As seen in Figure 5, the distributions of the background and the experimental data are almost indistinguishable visually from the histograms.

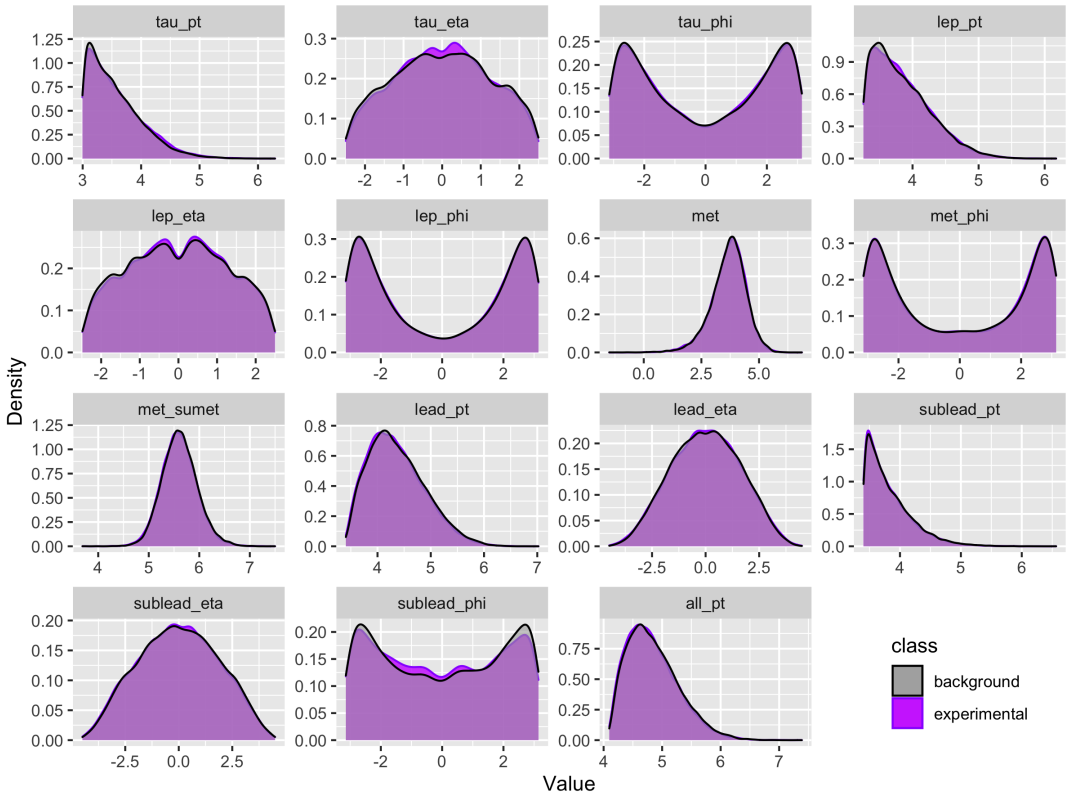
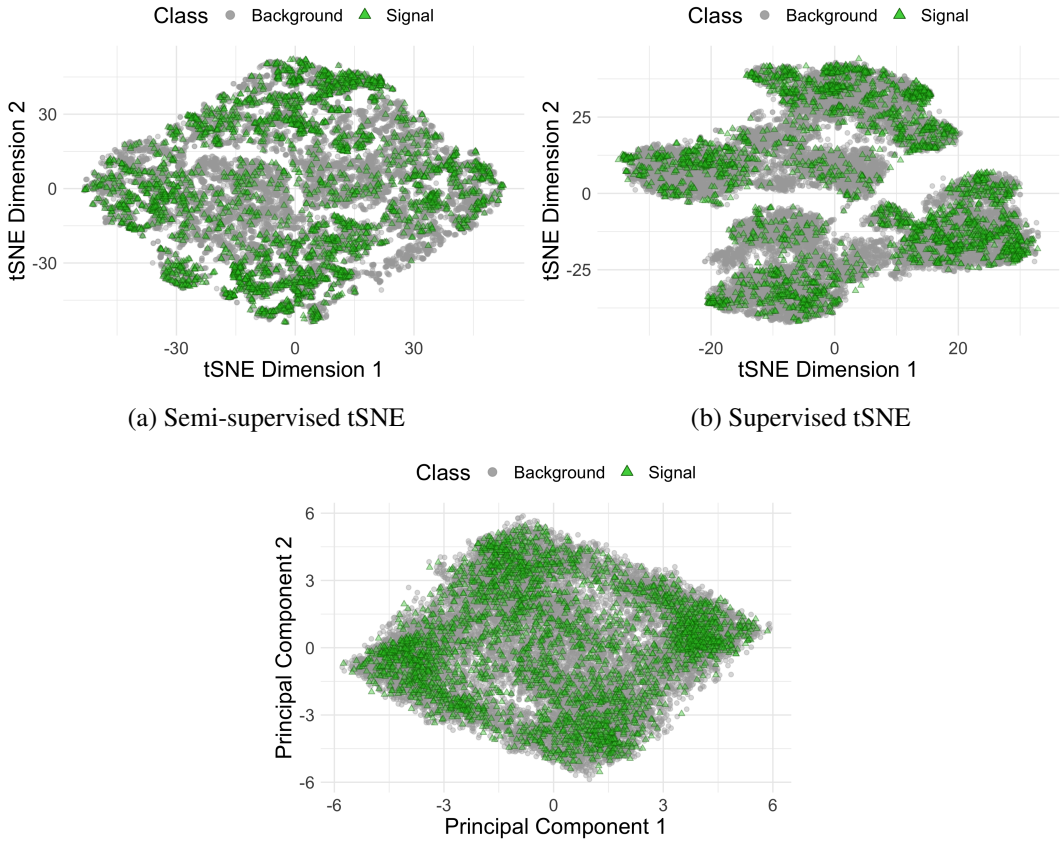


Fig 5: Histograms of the 15 variables for training data containing the experimental (purple) data as well as the background (grey) data. Note that the experimental data is a mixture of the background and the signal data, with the probability of signal event being $\lambda = 0.15$.

To demonstrate this further, we visually analyze the multivariate dependencies for the two data sets as well. We demonstrate two different approaches. First we use Principal Component Analysis on just the background data to find the two principal components of the background data and then project the test experimental data on those axes. Figure 6c shows that the signal is not very distinguishable from the background. We then use t-distributed stochastic neighbor embedding proposed by Maaten and Hinton (2008) to visulaize the data in two dimensions. First we train the algorithm to distinguish experimental data from the background data. We see in Figure 6a that the method doesn't appear to be able to spatially separate the signal data from the background data. Since this approach fails, we directly train the algorithm to distinguish the signal data from the background data. As shown in Figure 6b, this approach fails to separate the signal data from the background data as well. This emphasizes the difficulty of the problem to detect differences between the background data and the experimental data.

Despite the difficulty of the problem, random forests demonstrate power in detecting the differences between the background and the experimental data, hence detecting the presence of signal in the case of $\lambda = 0.15$. So, it is important to understand, charaterize and interpret how the variables influence the classifier output, in order to understand the random forest, which otherwise would be a black box. This is not an easy task as demonstrated by Figure 7, which shows the random forest classifier output (estimated probability of being an experimental event) marginally as a function of each of the variables in the data.



(a) Semi-supervised tSNE

(b) Supervised tSNE

(c) PCA

Fig 6: Experimental and background test data containing signal events (green) and background events (grey). (a) t-distributed stochastic neighbor embedding (tSNE) trained on experimental versus background training samples. (b) t-distributed stochastic neighbor embedding (tSNE) trained on signal versus background training samples. (c) Principal component analysis (PCA) trained on background training samples.

We notice that the random classifier seems to depend on the transverse momentums of all the particles produced (variables ending with `_pt`), as well as the missing transverse energy (`met`) and the total transverse energy in the detector (`met_sumet`).

To understand better how these variables affect the classifier we use the active subspace method introduced in Section 4 and show the results in Section 5.4 of the main paper (Chakravarti et al., 2021b). As mentioned in the paper, we use a Gaussian kernel for the local linear smoother used in the active subspace method. In order to select the smoothing parameter for the local linear smoother, we use the standard deviation of the variables scaled by a factor h as the bandwidth for the multivariate local linear smoother. We explore a few scaling factors h and calculate the standardized gradients as well as the mean of the standardized gradients for every choice of h. Following the active subspace method (Method 4.1) in the main paper (Chakravarti et al., 2021b), we find the projection of the experimental test data on the mean standardized gradient vector. We finally choose a scaling factor h that visually demonstrates the maximum amount of distinguishability between the signal and the background distributions when the experimental data is projected along the corresponding mean standardized gradient vector.

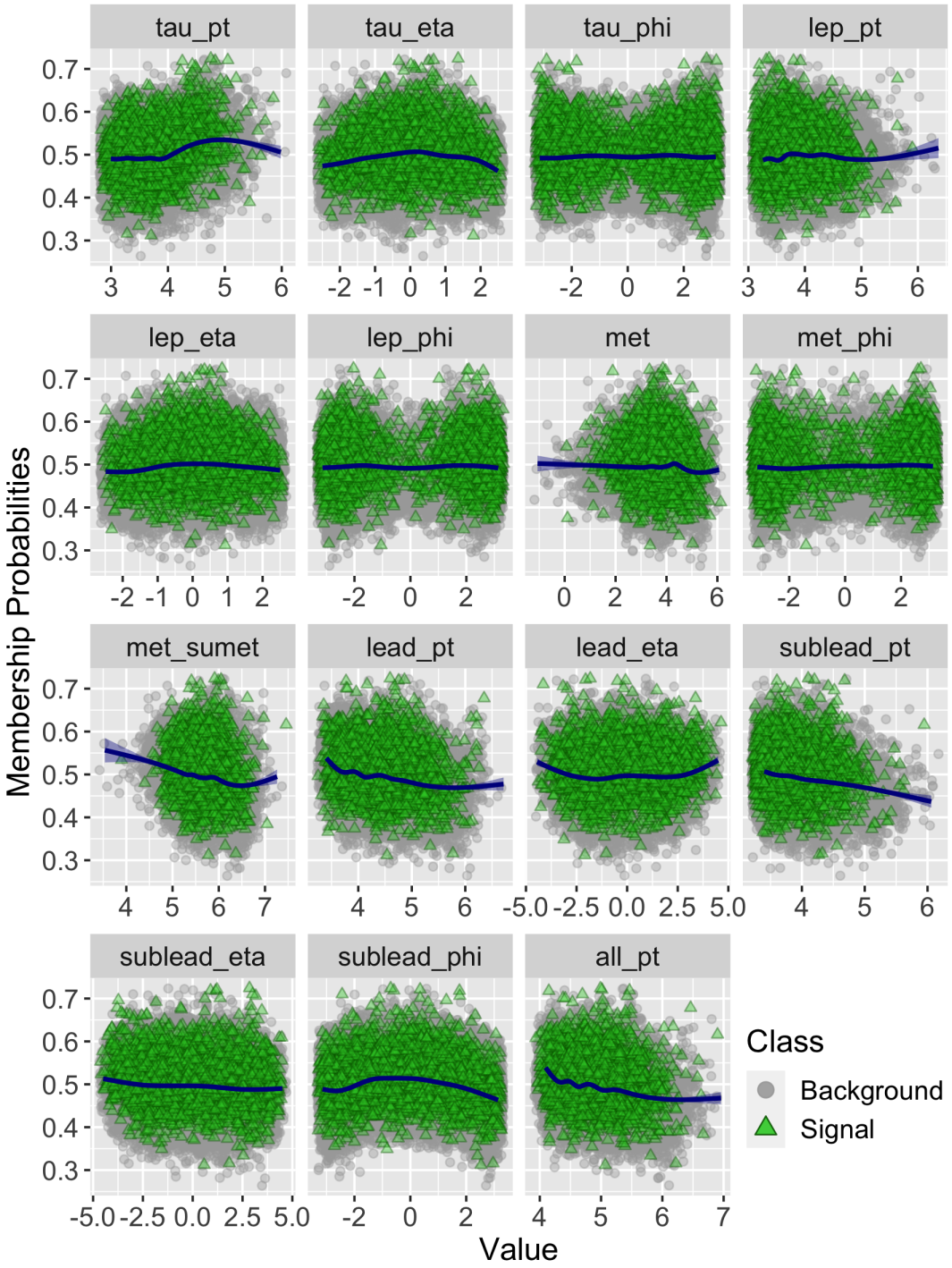


Fig 7: Experimental membership probabilities (the random forest output) versus all the variables for the test data sets. Signal events in green and background events in grey.

Figure 8 demonstrates the signal and the background distributions of the experimental data when it is projected along the corresponding mean standardized gradient vector for different scaling factors h. We only show the results for a subset of the scaling factors that we considered. Figure 8 shows that scaling the standard deviation by anything larger than 3 appears to

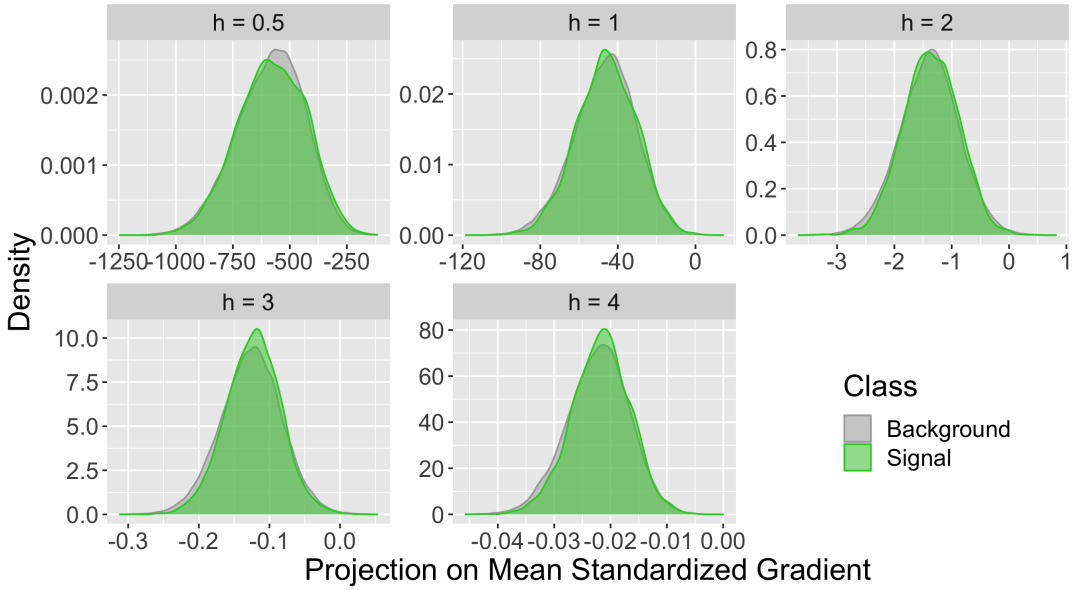
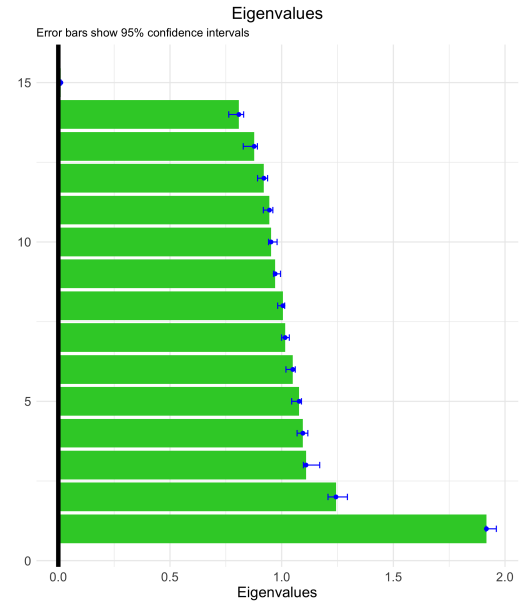


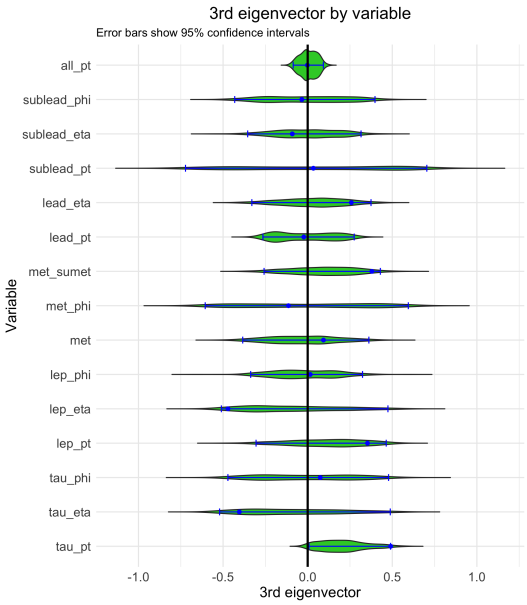
Fig 8: Histograms of the signal (green) and the background (grey) events from the test experimental data projected onto the mean standardized gradient vector when the standard deviation of the variables scaled by a factor (h) is used as the bandwidth for the Gaussian kernel when using the local linear smoother.

give similar results. We also notice that $h = 0.5$ appears to give the most separation between the background and the signal distributions. Note that if two scaling factors give similar results, it is better to pick the smaller scaling factor, which will result in higher bandwidths for the local linear smoother, resulting in a smoother estimate. So we consider $h = 0.5$ and divide the standard deviation in the data by $h = 0.5$ to use that as the bandwidth for the results presented in Section 5.4 of Chakravarti et al. (2021b).

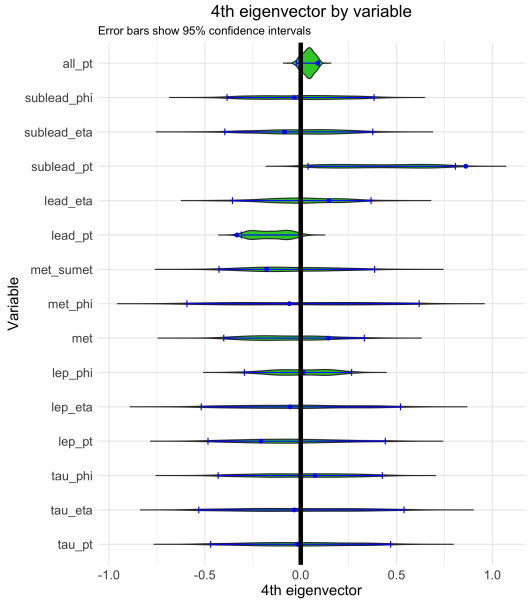
The mean standardized gradient vector as well as the first two active subspace vectors are presented in Section 5.4 of Chakravarti et al. (2021b). Figure 9 presents the third, fourth and the fifth active subspace vectors as well as the eigenvalues. We see that most of the information is contained in the first two eigenvectors, as can be seen from the eigenvalues plot. Similar to the first and the second eigenvector plots, for each of the plots 9b, 9c and 9d, we constrain the variable that has the largest absolute eigenvector value, i.e. $k_j = \arg\max_i |\widehat{M}_{ij}|$, for $j = 3, 4, 5$, to have the same sign in each of the bootstrap iterations as the sign in the original data. That is, for $j = 3, 4, 5$ we constrain $\text{sgn}(\widehat{M}^*_{k_j j}) = \text{sgn}(\widehat{M}_{k_j j})$, where $M_{\cdot j}$ is the j^{th} eigenvector in the data and $M^*_{\cdot j}$ is the j^{th} eigenvector in the bootstrap iterations. We do this to avoid the eigenvalues from being systematically symmetric about zero. We notice from Figure 9 that for most variables their eigenvector values are still symmetric. So, we notice that the third, fourth and the fifth eigenvectors do not really provide any additional information. The fourth eigenvector in Figure 9c shows that the transverse momentum of the sub-leading jet might play a role in detecting the signal. Recollect that this effect was also seen in the first eigenvector in the main paper.



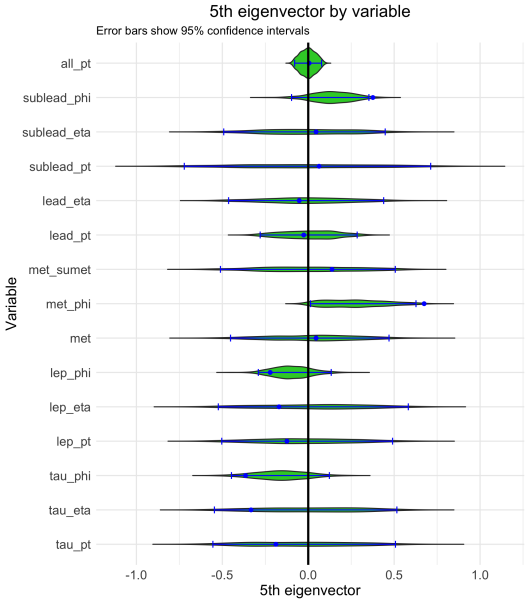
(a) Standardized Gradient Distribution



(b) Third Eigenvector



(c) Fourth Eigenvector



(d) Fifth Eigenvector

Fig 9: The active subspace variables for the classifier trained on data with signal strength $\lambda = 0.15$ computed using a local linear smoother that uses the gaussian kernel with smoothing parameter $h = 0.5$. (a) gives the eigenvalues of the standardized gradients. In (b), (c) and (d), the violin plot and the dashes give the bootstrapped empirical distribution and the bootstrapped uncertainty intervals computed using the empirical quantiles respectively for the third, the fourth and the fifth eigenvectors. The dots represent the eigenvalues, the third, the fourth and the fifth eigenvectors computed on the combined test data.

REFERENCES

ADAM-BOURDARIOS, C., COWAN, G., GERMAIN, C., GUYON, I., KÉGL, B. and ROUSSEAU, D. (2015). The Higgs boson machine learning challenge. In *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning* (G. COWAN, C. GERMAIN, I. GUYON, B. KÉGL and D. ROUSSEAU, eds.). *Proceedings of Machine Learning Research* **42** 19–55. PMLR, Montreal, Canada.

CHAKRAVARTI, P., KUUSELA, M., LEI, J. and WASSERMAN, L. (2021a). Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests.

CHAKRAVARTI, P., KUUSELA, M., WASSERMAN, L. and LEI, J. L. (2021b). Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests.

ATLAS COLLABORATION (2014). Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014. *CERN Open Data Portal*.

MAATEN, L. V. D. and HINTON, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research* **9** 2579–2605.