Maximising AMS by classifying Higgs Boson Data

Kieran Morris, Cecina Babich Morrow and Daniella Montgomery

2023-11-07

This RMarkdown document is a companion piece to a Latex document submitted detailing our work on this project. For discussion including references to other works and justification for the techniques and methods used, please refer to that document. This will be an indepth exploration of the statistical results and `R` code only.

The source code for this Rmarkdown and all annotated `R` code can be found in the following GitHub repository:

```
https://github.com/babichmorrowc/sm1_large_hadron_collider
```

Additionally this project comes with a package which contains all necessary functions which can be downloaded via the following command in the `R` console:

```
devtools::install_github("babichmorrowc/higgiesmalls")
```

Preliminary Observations

Our data consists of events detected at the Large Hadron Collider which were classified as either Background (b) or Signal (s) - Signal being a detection of a higgs boson.

Types of Data

We have a few types of data to consider:

- Variables `KaggleSet` and `KaggleWeight` can be ignored as they denote which data points were in the provided Kaggle challenge and their relative weights, which is irrelevant to us.
- The discrete classification variable `Label` which takes values in $\{b, s\}$
- The continuous variable `Weight`, which will be used to compute the AMS, and will not be used in classification.
- The discrete variable `PRI_jet_num` denotes the amount of jets from each event and takes values in $\{0, 1, 2, 3\}$.
- We have multiple variables which are the angle of detection for some observation.
- All other variables are either direct or indirect measurements and are continuous.

Missing Data

What is missing and why?

Below are the variables which contain undefined values. By convention it was provided to use with values `-999` which is out of range for every observation.

Show 11 entries Search:

	skim_variable	n_missing	complete_rate
1	DER_mass_MMC	124602	0.8477191232868676
2	DER_deltaeta_jet_jet	580253	0.2908505838154669
3	DER_mass_jet_jet	580253	0.2908505838154669
4	DER_prodeta_jet_jet	580253	0.2908505838154669
5	DER_lep_eta_centrality	580253	0.2908505838154669
6	PRI_jet_leading_pt	327371	0.5999073619167039
7	PRI_jet_leading_eta	327371	0.5999073619167039
8	PRI_jet_leading_phi	327371	0.5999073619167039
9	PRI_jet_subleading_pt	580253	0.2908505838154669
10	PRI_jet_subleading_eta	580253	0.2908505838154669
11	PRI_jet_subleading_phi	580253	0.2908505838154669

Showing 1 to 11 of 11 entries Previous 1 Next

Notice that every column besides `DER_mass_MMC` is a jet variable, and in those, we only have two values for `completion_rate`. In fact these correspond to different values of `PRI_jet_num`:

- 0 jets: All jet variables were missing data.
- 1 jet: only `PRI_leading_pt`, `PRI_leading_eta` and `PRI_leading_phi` have data.
- 2 or more jets: All jet variables have data.

The above result can be found in the handbook for the variables provided with the Kaggle challenge. Unfortunately `DER_mass_MMC` does not have such an explanation and may be a result of some event during measurement. However it is still assigned the same `-999` value as the other missing data.

Impact of the Missing Data

Despite understanding the cause of (most of) our missing data, we still don't know the impact of it, the following section is dedicated to understanding the correlation between the missing data from each variable and its classification.

Below we compute the percentage of `NA` data in Background and Signal respectively. If this missing data is distributed uniformly across Background and Signal then the percentages should be very close.

Show 11 entries Search:

% of Missing Background and Signal Data

	Signal	Background	Difference
DER_mass_MMC	3.287	21.4252	18.1382
DER_deltaeta_jet_jet	61.9026	75.5921	13.6895
DER_mass_jet_jet	61.9026	75.5921	13.6895
DER_prodeta_jet_jet	61.9026	75.5921	13.6895
DER_lep_eta_centrality	61.9026	75.5921	13.6895
PRI_jet_leading_pt	29.6369	45.3922	15.7553
PRI_jet_leading_eta	29.6369	45.3922	15.7553
PRI_jet_leading_phi	29.6369	45.3922	15.7553
PRI_jet_subleading_pt	61.9026	75.5921	13.6895
PRI_jet_subleading_eta	61.9026	75.5921	13.6895
PRI_jet_subleading_phi	61.9026	75.5921	13.6895

Showing 1 to 11 of 11 entries Previous 1 Next

Notice that `Signal` consistently has $13 - 18\%$ less missing data than `Background`. Considering the size of our dataset (\$ ^5) this is statistically significant. This means that in both the `DER_mass_MMC` and jet-variables cases, the amount of missing data is indicative of a classification. Meaning that including `-999` can help with classification and will not be removed.

Variable Selection

Co-Dependency of Data

We have the following algebraic codependencies:

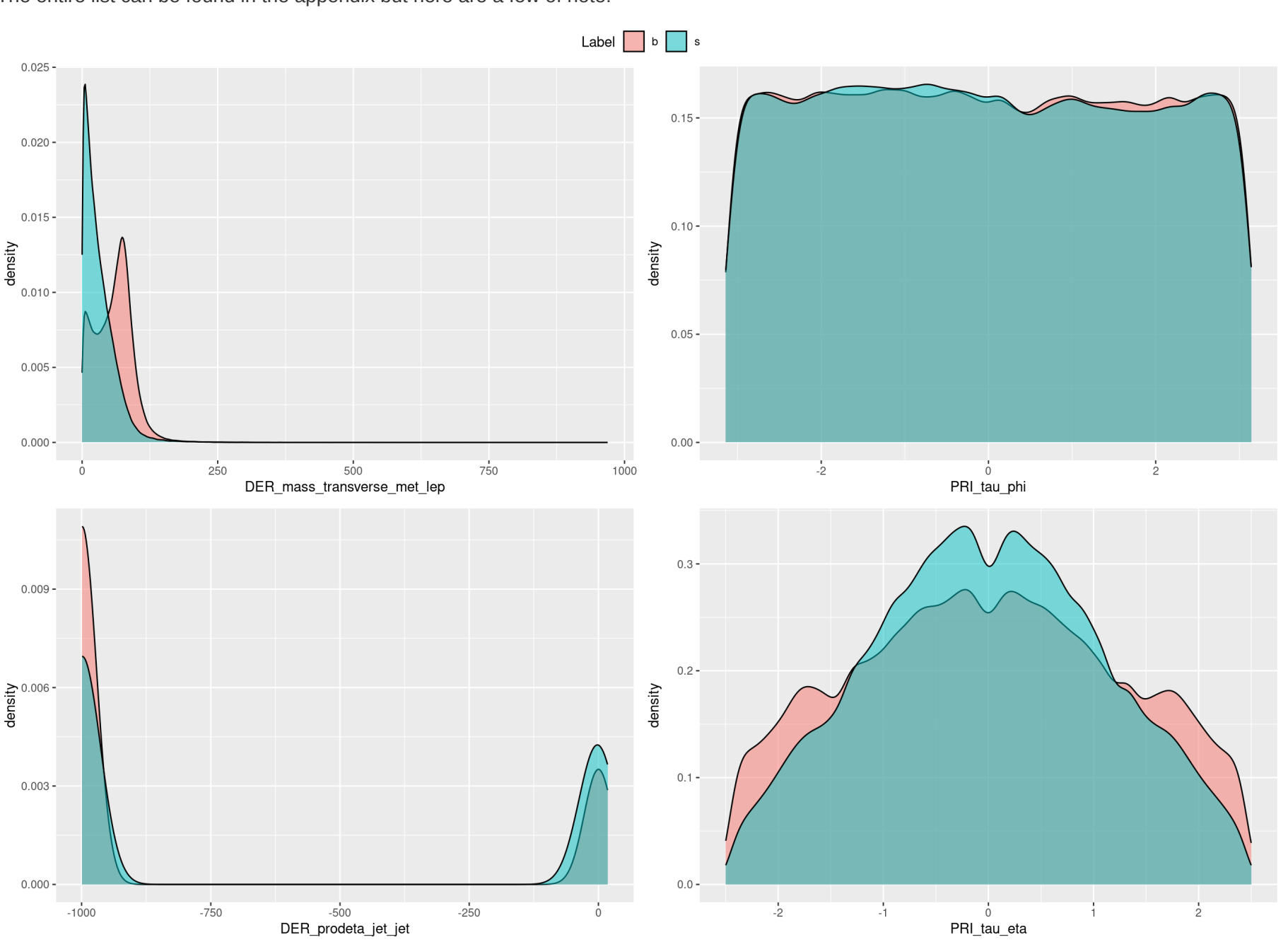
- `PRI_jet_all_pt = PRI_jet_leading_pt + PRI_jet_subleading_pt`
- `PRI_lep_pt = DER_pt_ratio_lep_tau*PRI_tau_pt`

We can either remove `PRI_jet_all_pt` or remove both `PRI_jet_leading_pt` and `PRI_jet_subleading_pt` to reduce the dimension by 1 or 2. This same idea can be applied to to the other (multiplicative) dependence. Taking us down to possibly 26 dimensions.

Another 'dependence' is the fact that `PRI_jet_num` characterises the missing data in jet-variables. Meaning we could remove all of these in favour of keeping `PRI_jet_num` however this is very extreme and may produce detrimental effects.

Density Plots

To get an understanding of the behavior of variable given their classification, we condition each variable on whether $y = b$ or $y = s$ respectively. The entire list can be found in the appendix but here are a few of note:



Some key takeaways from the density plots are:

- We see that some these angle variables: `PRI_tau_phi`, `PRI_lep_phi`, `PRI_met_phi`, `PRI_jet_leading_phi` and `PRI_jet_subleading_phi` are approximately uniformly distributed and that there is very little variation between the classifications.
- We have many distributions with distinct translations in their peaks: `DER_mass_MMC`, `DER_mass_transverse_met_lep`, `DER_mass_vis`, `PRI_tau_pt` and `DER_pt_ratio_lep_tau`.
- Some distributions had higher variance in the background case (see `DER_deltar_tau_lep` or `PRI_tau_eta`), indicating they feel the difference, but may not be as useful for classification.
- We can see that many jet-related variables are dominated by their `-999` values, additionally we see that the proportion of `-999` values is higher in the background case - as expected.

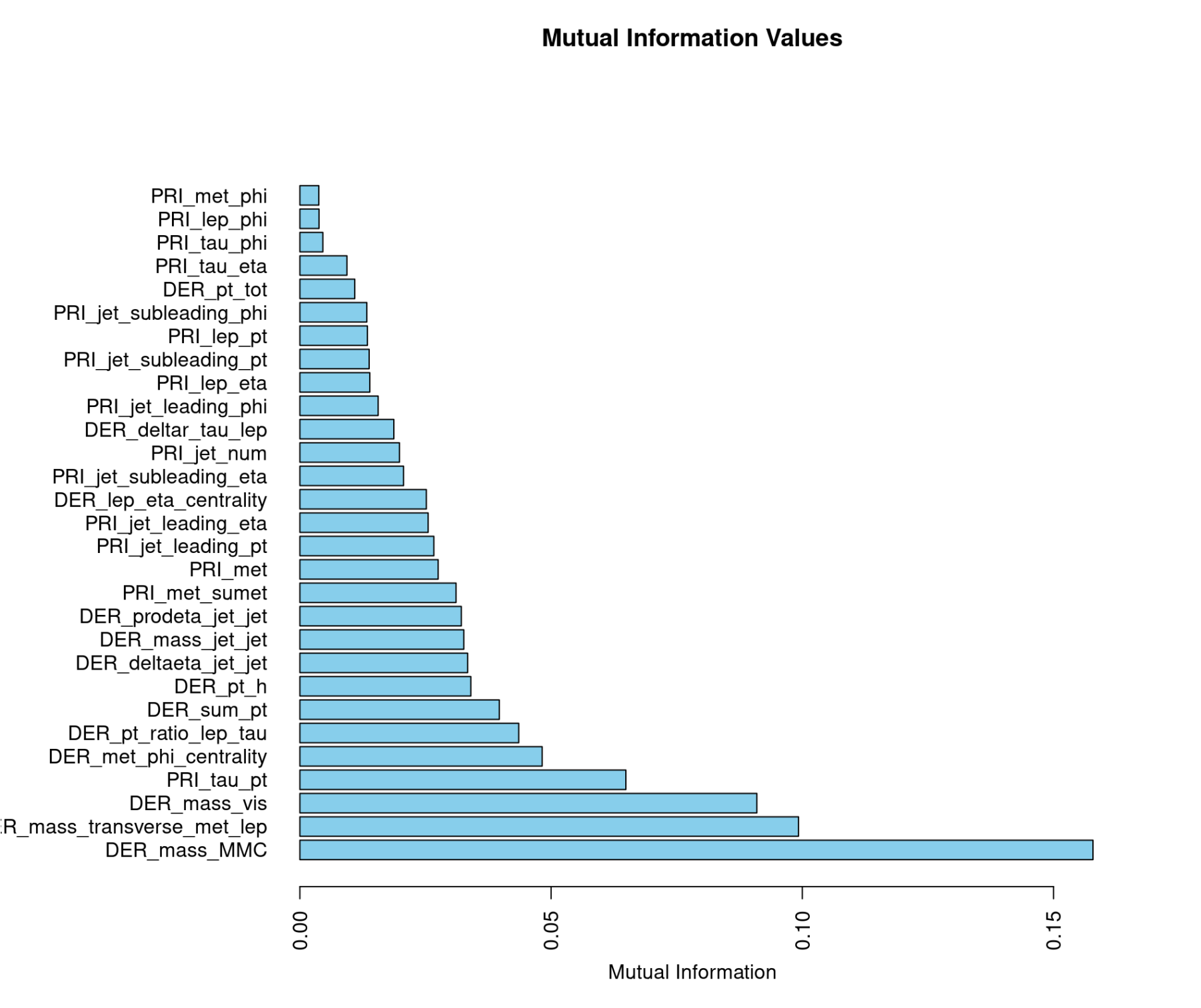
Variables with clear translational peaks make great contenders for classification, while jet-variables and ones with higher variance may need some sort of feature transform. We hypothesis that the uniform angle variables only provide noise to the problem and have no bearing on classification at all.

Before we decide on our preferences of variables, we employ an information theoretic technique to measure the impact of each variable: The Mutual Information.

Performing Mutual Information Tests

For each variable V we compute its mutual information with `Label`, variables with high mutual information are better indicators for classification. We hope this will validate our observations from plotting the data.

Our results are below, see the appendix for the exact values.



Thankfully these results justify many of our direct observations from plotting the densities:

- The variables with clear translational difference in their peaks (`DER_mass_MMC`, `DER_mass_transverse_met_lep`, `DER_mass_vis`, `PRI_tau_pt` and `DER_pt_ratio_lep_tau`) had very high mutual information. Additionally, `DER_mass_MMC`: the only jet-variable with `-999` values, had the highest mutual information than any other variable, by a significant margin.
- Variables which were around the middle in terms of mutual Information were a mixture of higher variance and jet-variables.
- The mostly uniform variables ranked lowest in terms of mutual information.

Principle Component Analysis

Conclusions on Variable Selection

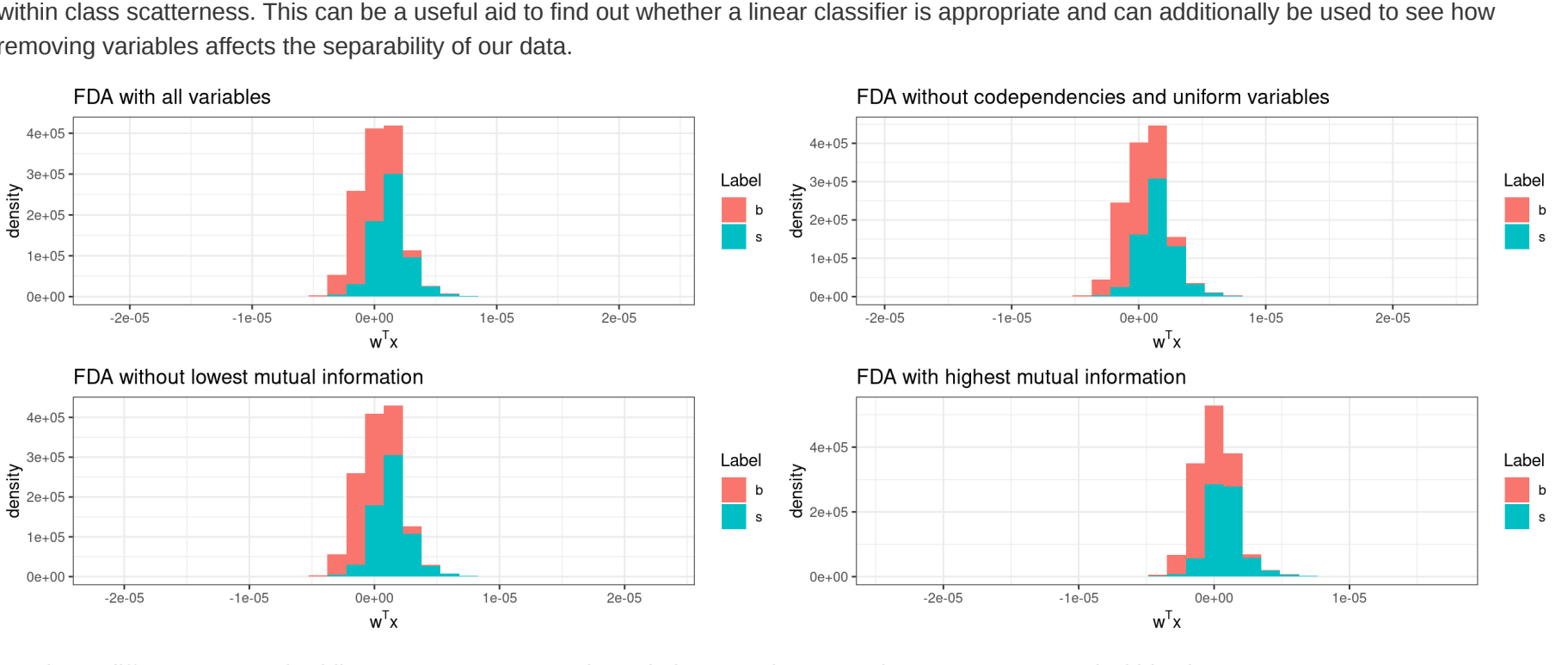
After considering our multiple ways of prioritising these variables, we conclude that we will remove our variables heirarchically in the following order:

1. Include all variables
2. Remove all 4 algebraic co-dependencies (`PRI_jet_leading_pt`, `PRI_jet_subleading_pt`, `DER_pt_ratio_lep_tau`, `PRI_tau_pt`).
3. Remove the above and any uniform variables (`PRI_lep_phi`, `PRI_met_phi`, `PRI_tau_phi`).
4. Only use the top ten highest ranking mutual information variables (see above).

We note that one of our algebraic variables is `PRI_tau_pt`, which is in the top ten ranking - so that will be added back in. This allows us to cover both bases: removing significant algebraic codependence, and including variables which have high mutual information. The final choice is certainly a gamble, and is more an exercise is determining when the diminishing effects of dimension reducation take place.

Fisher Discriminant Analysis

FDA creates an embedding that tells us how linearly separated our data can possibly be by maximizing between class scatterness and minimizing within class scatterness. This can be a useful aid to find out whether a linear classifier is appropriate and can additionally be used to see how removing variables affects the separability of our data.



For these different FDA embeddings, we can compare the ratio between between class scatterness and within class scatterness:

Show 10 entries Search:

	Variables included	Scatter ratio
1	All variables	0.3251629060327355
2	Without codependencies and uniform variables	0.3235810938370512
3	Without lowest mutual information	0.3248675982049276
4	Highest mutual information	0.2540494515822303

Showing 1 to 4 of 4 entries Previous 1 Next

We find that removing variables slightly lowers the linear separability of the data, but in general the values are comparable, particularly when comparing the scatter ratio of all variables with that removing the 10 lowest mutual information. Based on the plots, we can conclude that it is not possible to achieve a high degree of linear separability between the signal and background events with any of the variable sets, indicating that a non-linear feature transform is most likely appropriate.

Support Vector Machines

We divided our dataset into a training and testing split, training SVMs on the training data and finding the AMS on the testing data.

By training SVMs with a linear kernel vs. a radial kernel on a 20% training set (using all predictor variables), we found that a radial kernel consistently yielded higher AMS:

Show 10 entries Search:

	Kernel	AMS
1	Linear	2.010106813362876
2	Radial	2.813594259911111

Showing 1 to 2 of 2 entries Previous 1 Next

Based on these results, along with the results of FDA indicating that the dataset is not linearly separable, we moved forward using a radial kernel. We performed 5-fold cross-validation over the 20% training set to tune the two hyperparameters using a grid search: cost and γ (bandwidth). We allowed cost to take on the values 0.5, 1, and 2 and γ to take on the values 0.01, 0.1, and 0.5. Tuning was performed using our function `ams_tune_parallel`, which modifies the `e1071::tune` function to maximise AMS and to conduct cross-validation in parallel, dramatically improving the performance of tuning compared to the original package.

Below are the results of the hyperparameter tuning using four sets of variables: 1) Including all variables; 2) Removing the algebraic co-dependencies and uniformly distributed variables; 3) Removing the 10 variables with the lowest mutual information; 4) Using only the 10 variables with the highest mutual information.

Show 10 entries Search:

	Variables included	Cost	Gamma
5	All variables	1	0.1
6	Without co-dependencies and uniform variables	2	0.1
61	Without lowest mutual information	2	0.1
9	Highest mutual information	2	0.5

Showing 1 to 4 of 4 entries Previous 1 Next

We then used the model using the best hyperparameter values from tuning to find the AMS value on the 80% testing dataset.

Show 10 entries Search:

	Variables included	Testing AMS
1	All variables	2.832435958468929
2	Without co-dependencies and uniform variables	2.856642255556614
3	Without lowest mutual information	2.866093570197652
4	Highest mutual information	2.812699039205425

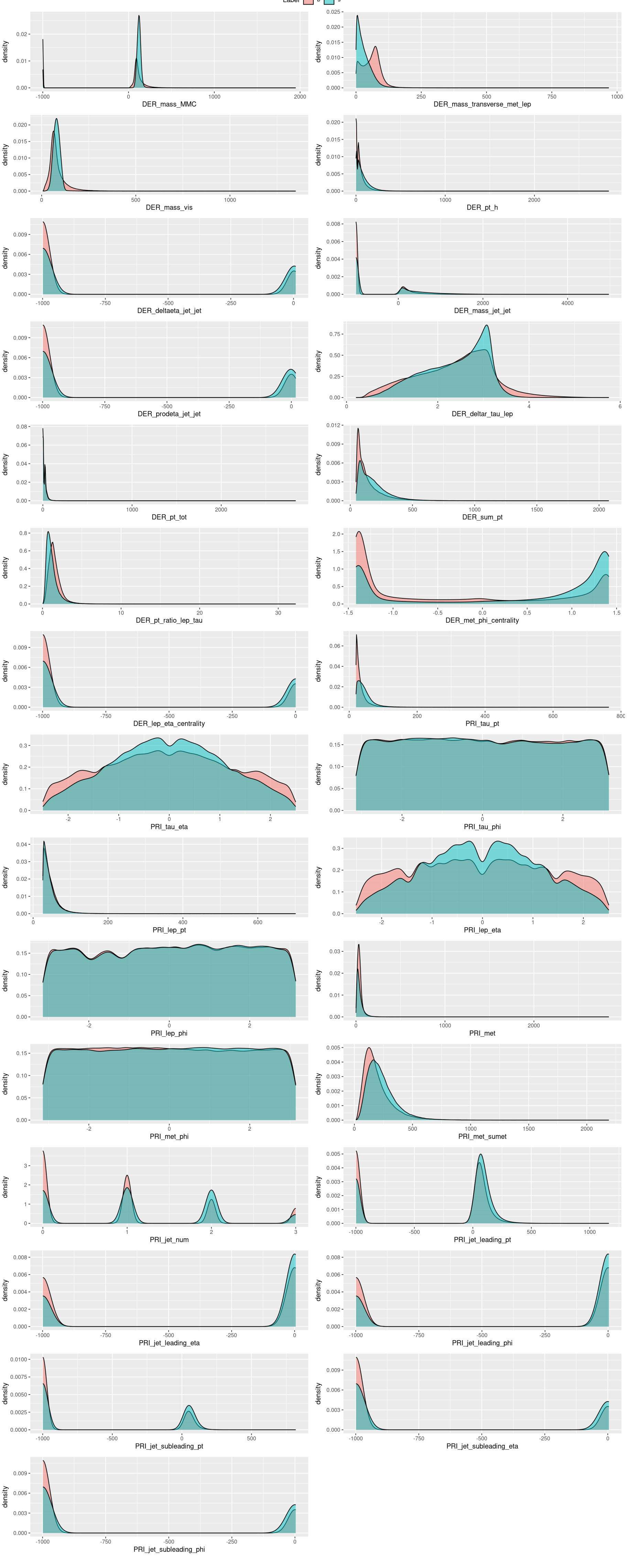
Showing 1 to 4 of 4 entries Previous 1 Next

We see in the above data that removing the ten lowest mutual information produced the highest AMS when using an RBF kernel, followed by removing algebraic codependencies and the uniform variables. Using only the top ten highest mutual information actually yielded the lowest AMS, highlighting the trade-off between prioritizing high mutual information versus removing too many variables.

Appendix

Density plots of each variable

The following are the density plots for our variables, conditioned on whether they are a detection of a higgs boson or not. Note that in the jet-variables and `DER_mass_MMC` we have peaks at `-999` which skew many of the variable plots.



Mutual Information Results Table

```
dt
```

Show 10 entries Search:

	Variable	Mutual Information
1	DER_mass_MMC	0.157847348
2	DER_mass_transverse_met_lep	0.099242021
3	DER_mass_vis	0.0909557
14	PRI_tau_pt	0.064883381
12	DER_met_phi_centrality	0.048218121
11	DER_pt_ratio_lep_tau	0.04355779
10	DER_sum_pt	0.039696267
4	DER_pt_h	0.03402485
5	DER_deltaeta_jet_jet	0.033387741
6	DER_mass_jet_jet	0.032628663

Showing 1 to 10 of 29 entries Previous 1 2 3 Next