# Effects of Linear Mass Decorrelation on $W$ Boson Tagging with Multivariate Classifiers

Gregory Beauregard

May 8, 2016

**Abstract**

High energy physics multivariate machine learning classifiers that distinguish $W$ boson jets from QCD interaction jets have input variables that are often highly correlated with invariant jet mass; this results in a multivariate classifier output highly correlated with invariant jet mass as well. This thesis investigates the effect of an algorithm to remove all linear correlation with QCD jet mass from input variables of the multivariate classifier. The results of this comparison show removing linear correlation with mass from the input variables is not sufficient to remove the correlation between classifier output and invariant jet mass.

## 1. Introduction

Proton-proton collisions at the Large Hadron Collider (LHC) can produce particles with large $p_{\mathrm{T}}$, or transverse momentum. Fully hadronic decay products of high $p_{\mathrm{T}}$ particles are collimated and may be reconstructed as a single hadronic jet. When investigating physics beyond the Standard Model, many new predicted particles are expected to produce these high $p_{\mathrm{T}}$ jets [1].

Numerous physics analyses seek to classify hadronic jets into the particles they decayed from using variables measured by the LHC. Jets arising from QCD interactions will tend to have an invariant mass near zero but may have larger masses due to the uncertainty principle. On the other hand, jets coming from $W$ bosons will tend to have an invariant mass near 80 GeV, the $W$ boson mass. Unfortunately, kinematic jet variables used in classification are often highly correlated with invariant mass resulting in classifiers that provide redundancy with the information already provided by the invariant mass.

It would be useful to train a classifier that is uncorrelated with invariant mass to remove the redundancy associated with it. In the study presented here a variable transformation that removes linear correlation with QCD background jet mass is investigated for its ability to accomplish this task.

## 2. Background

### 2.1. $W$ Tagging

Hadronically decaying $W$ bosons are one possible source of high $p_{\mathrm{T}}$ jets that show up in many different physics analyses. The process of determining whether or not a given jet comes from a hadronically decaying $W$ boson is referred to as $W$ tagging. The investigation in this paper is based on multivariate classifiers for $W$ tagging with data from a recent ATLAS paper [1].

### 2.2. Multivariate Classifiers

Multivariate analysis, or MVA, is the analysis of multiple statistical variables simultaneously. The way multivariate analysis is often employed in high energy physics is through the use of MVA classifiers, or MVAs for short. Multivariate classifiers are algorithms that can be trained on a data set split into different classes (e.g. signal and background). Once an MVA is trained it takes as input a new variable set and outputs a classification.

In the study presented here, MVAs are used for $W$ tagging. The two MVA classes are $W$ jets (signal) and QCD jets (background). In this case the output of the MVA will be a number on a bounded interval with one end representing signal-like jets and the other end

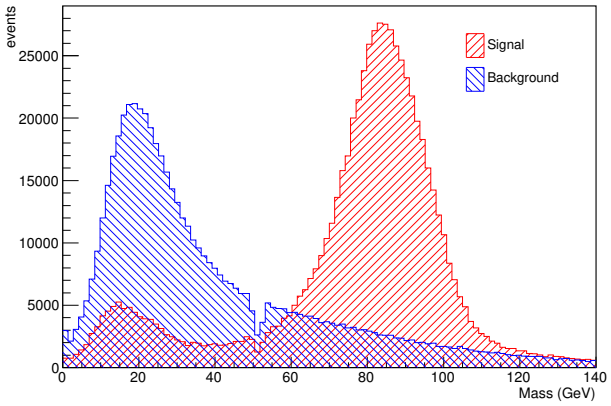



Figure 1: Mass distribution of the signal and background.

representing background-like jets. This study uses a Boosted Decision Tree [2] (BDT) classifier as provided by the TMVA [3] package included in the ROOT [4] data analysis framework.

# 3. Methods

## 3.1. Data Preparation

In order to set up this study, simulated Monte Carlo signal and background jets were retrieved from and prepared similarly to how they were in a recent ATLAS study [1] on $W$ boson tagging. After obtaining the data from this study, a cut was performed on the $p_{\mathrm{T}}$ variable from 200 GeV to 400 GeV. This had the effect of cutting down the number of signal jets from 1 691 398 to 316 925. To facilitate trained classifiers generalizing to a wide range of $p_{\mathrm{T}}$ values, event weights were calculated for the signal jets so that when the weights are applied the signal $W$ jet $p_{\mathrm{T}}$ distribution matches the background QCD jet $p_{\mathrm{T}}$ distribution. The resulting set of jets formed the basis for this study. The mass distribution for this full data set is shown in Figure 1.

Three data sets were constructed from the prepared jet set for comparison and classifier training. The first, comprising of the full data set, was deemed the *no M cut* set. Performing a mass cut on the no M cut set by removing jets outside a 61 GeV to 92 GeV invariant mass range forms the second data set and was deemed the *M cut* set. Finally, the third data set was formed by applying a transformation that removes linear correlation with background QCD jet invariant mass and was deemed the *mass ortho* set.

In order to remove linear correlation with mass from a variable, a variable transformation needs to be applied. Let an arbitrary variable be represented by the vector $\mathbf{x}$ whose $i$th element corresponds to the $i$th jet. Let the corresponding vector of jet invariant mass be given by $\mathbf{m}$. The process of obtaining a new vector $\mathbf{u}$ that has zero covariance with mass is analogous to Gram–Schmidt orthonormalization. Therefore, it is appropriate to call the linearly decorrelated variables mass orthogonalized. The formula for this transformation is given by

$$\mathbf{u} = \mathbf{x} - \frac{(\mathbf{m} - \langle m \rangle) \cdot (\mathbf{x} - \langle x \rangle)}{(\mathbf{m} - \langle m \rangle) \cdot (\mathbf{m} - \langle m \rangle)} \mathbf{m}. \tag{1}$$

This formula may be rewritten using the definition of covariance and variance to decorrelate a particular $i$th jet. The resulting formula is

$$u_i = x_i - \frac{\mathrm{Cov}(x, m)}{\mathrm{Var}(m)} m_i. \tag{2}$$

As the specific value of the variables is no longer relevant after applying a transformation, it makes sense to go ahead and center the variables at 0 when applying the transformation. The modified formula to accomplish this is given by

$$u_i = x_i - \langle x \rangle - \frac{\mathrm{Cov}(x, m)}{\mathrm{Var}(m)} \left( m_i - \langle m \rangle \right). \tag{3}$$

After calculating $\mathrm{Cov}(x, m)$ and and $\mathrm{Var}(m)$ for the background QCD jets alone, Equation (3) was applied to each of the jet variables in the no M cut set to linearly decorrelate them with background QCD jet invariant mass and create the third mass ortho data set. To help illustrate this process, scatter plots of $C_2^{(\beta)}$, an energy correlation ratio, versus jet mass are contained in Appendix A.

## 3.2. Classifier Training

To narrow down the classifiers used for training, a 20 variable Boosted Decision Tree classifier, or BDT, was trained on the no M cut set using default TMVA package settings as they were found to be reasonable. From these, the TMVA estimated sensitivity for the variables after classifier training was used to select the top 11 variables for use in this study. These variables are detailed in Appendix B.

Using the default TMVA package settings for a BDT, 3-variable classifiers for each of the data sets were trained for every possible 3-variable combination of the 11 chosen variables. From these, the three best classifiers for each training data set were determined by calculating and ranking their sensitivity $z$ on the data set type they were trained on.

If the number of signal jets in a histogram bin is given by $S$ and the background by $B$, the sensitivity of a given classifier is calculated with

$$z = \sum_{i=1}^{n} \frac{S_i}{\sqrt{B_i}} \tag{4}$$

where the summation is over the $n$ bins in a histogram of the classifier output.

# 4. Results

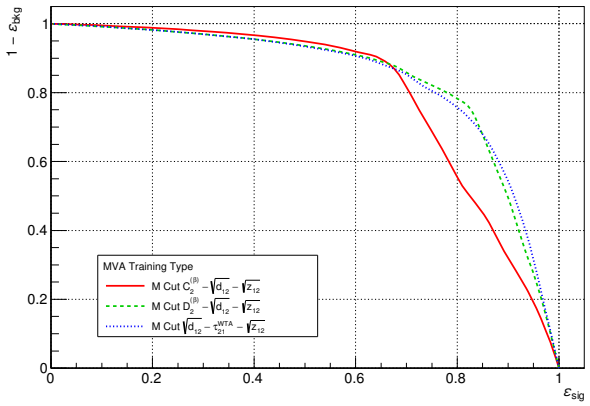
## 4.1. Classifier Performance

The developed classifiers are presented here in the form of receiver operating characteristic (ROC) curves. These were calculated by plotting the signal efficiency against one minus the background efficiency for a progressive cut on the MVA. This is useful since it results in better performing MVAs curving farther from the origin. The ROC curves for the MVAs evaluated on the full data set are shown in Figure 2.

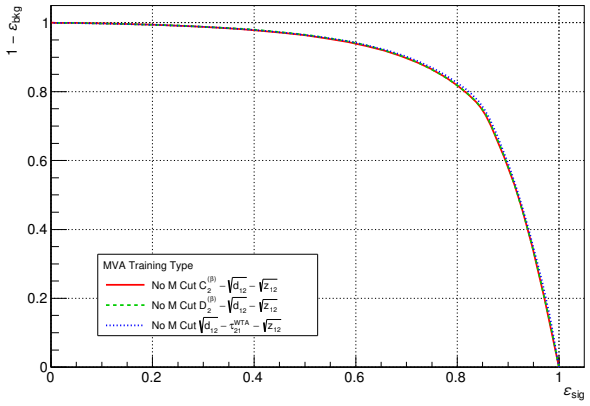
Figure 2 shows ROC curves evaluated on the full data set. In Figure 2a, the M cut MVAs perform worse than the other classifiers at signal efficiencies above 0.7. This is expected since Figure 2a was trained with a 61 GeV to 92 GeV mass cut but is being evaluated on the full data set. Notably, however, two of the best classifiers in this case were able to generalize their performance.

Figures 2b and 2c were trained on the full data set and appear to show similar classification performance in spite of Figure 2c being trained with the mass ortho set. For a refined view, the ROC curves after a 61 GeV to 92 GeV mass cut are plotted in Figure 3.

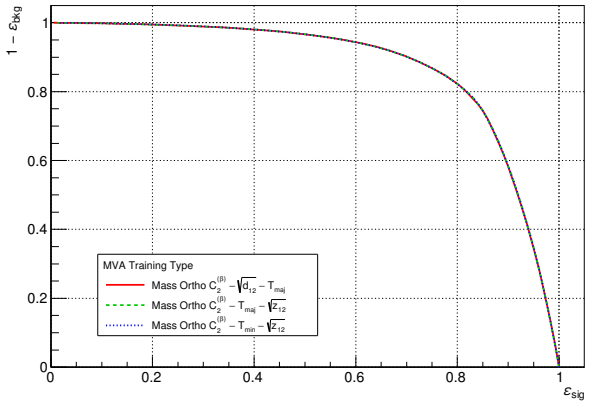
Figure 3 shows ROC curves evaluated after a $61 < M < 92$ mass cut on the full data set. In Figure 3a, the M cut trained classifier is shown to perform very well when evaluated on events with the same mass cut applied. However, the classifiers trained on the full data set, Figures 3b and 3c, appear to have competitive performance. Notably there is




(a) M cut trained MVA.

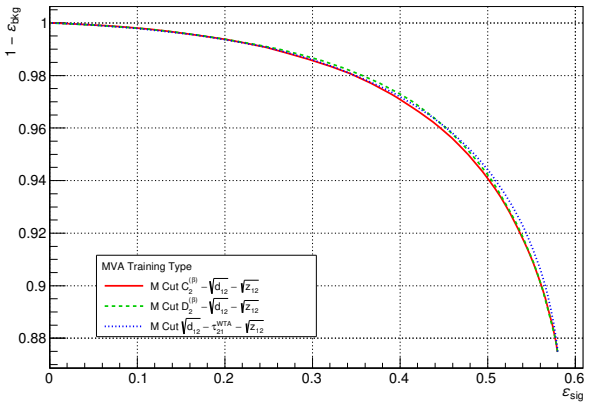



(b) No M cut trained MVA.

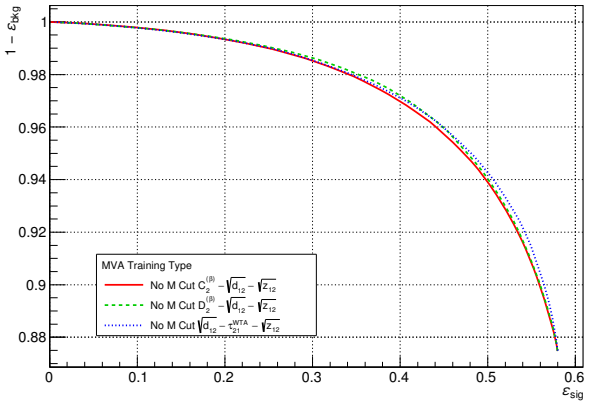



(c) Mass ortho trained MVA.

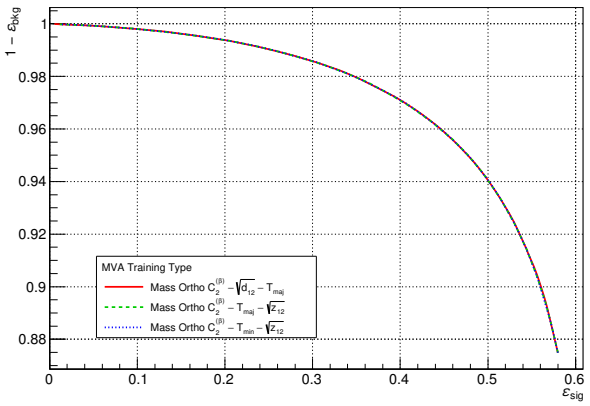
Figure 2: ROC curves of the full data set on the three best 3-variable MVAs trained with a $61 < M < 92$ mass cut, no mass cut, and mass orthogonalization.




(a) M cut trained MVA.




(b) No M cut trained MVA.




(c) Mass ortho trained MVA.

Figure 3: ROC curves after $61 < M < 92$ mass cut on three best 3-variable MVAs trained with a $61 < M < 92$ mass cut, no mass cut, and mass orthogonalization.

still a lack of any apparent performance degradation in the mass ortho trained MVA.

### 4.2. Mass Distributions

To investigate the behavior of the mass orthogonalization classifier in detail, the mass distributions after $-0.2$, $-0.1$, 0.0, 0.1, and 0.2 MVA cuts are plotted in Figure 4 for the best MVA of each training data set.

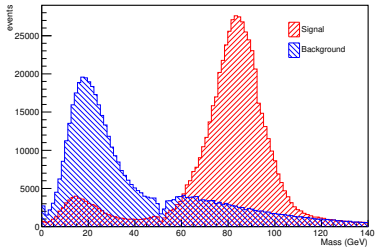
Figures 4b to 4n along the center column show the mass distributions of the no M cut trained classifier as progressively more aggressive MVA cuts are applied. Since this classifier was trained on the full data set, it was expected to do well classifying the background mass region and provides a baseline by which the other classifiers can be compared.

Figures 4a to 4m of the M cut trained classifier illustrate the expected behavior of a classifier having trouble distinguishing background mass; for mild MVA cuts, the classifier cannot significantly shape the background mass distribution as much as it could in the no M cut case. This makes sense as the M cut classifier was trained without events outside a 61 GeV to 92 GeV mass window.

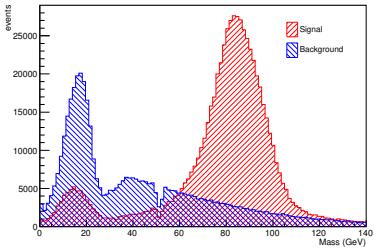
Figures 4c to 4o of the mass ortho trained classifier show the background mass distribution behaving almost identically to the classifier trained without an M cut. This is contrary to the expectation that removing linear correlation with background mass would result in difficulty classifying it. Further, these plots corroborate the ROC curve finding that the mass ortho classifier has no apparent performance degradation when compared to the no M cut classifier.

## 5. Conclusion

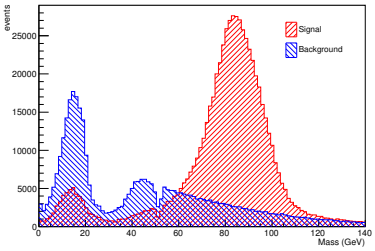
Despite removing linear correlation with mass, the mass ortho trained classifiers showed no degradation in classification performance or lack of ability to cut out background mass events. This is certainly unexpected, but likely has a simple explanation. By picking the best three classifiers on the mass ortho data set the chosen variables were likely to have high nonlinear correlation with mass. It is proposed the trained BDTs were able to effectively make use of the nonlinear correlation present in these variables to distinguish background mass. Hence, removing linear correlation with background mass is not sufficient to



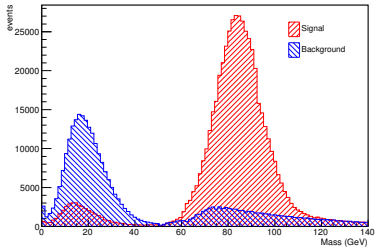

(a) M cut, −0.2.



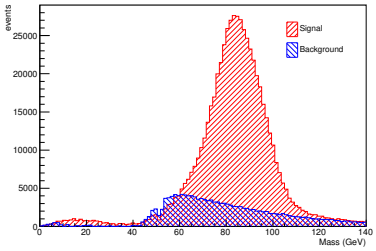

(b) No M cut, −0.2.



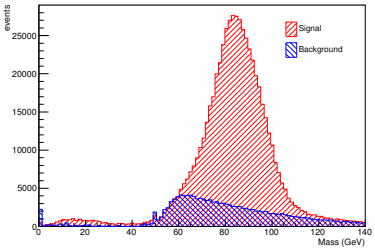

(c) Mass ortho, −0.2.



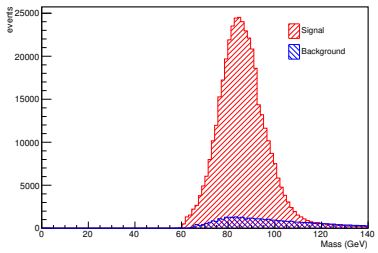

(d) M cut, −0.1.



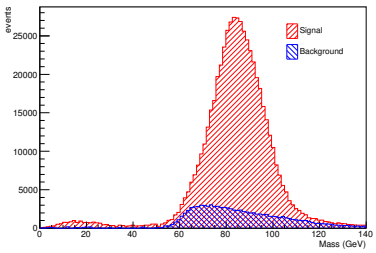

(e) No M cut, −0.1.



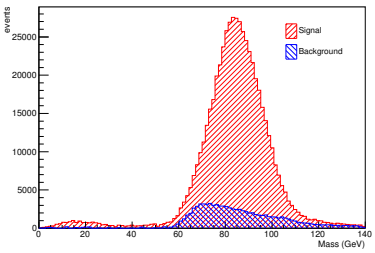

(f) Mass ortho, −0.1.



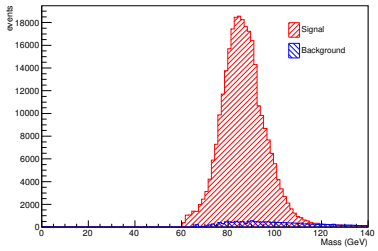

(g) M cut, 0.0.



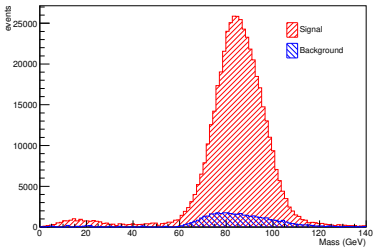

(h) No M cut, 0.0.



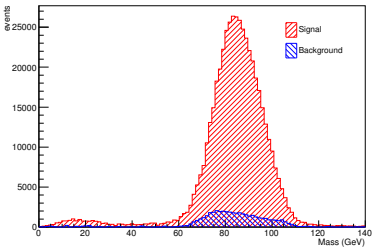

(i) Mass ortho, 0.0.



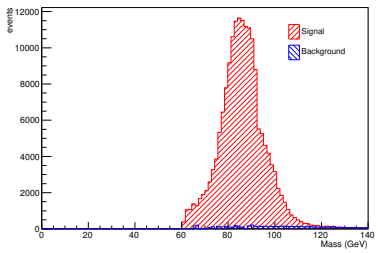

(j) M cut, 0.1.



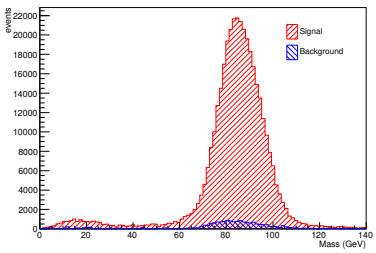

(k) No M cut, 0.1.



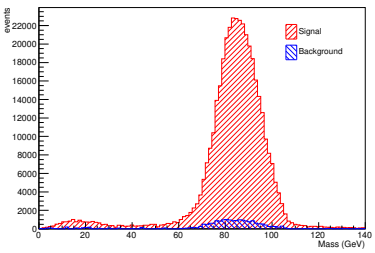

(l) Mass ortho, 0.1.




(m) M cut, 0.2.




(n) No M cut, 0.2.




(o) Mass ortho, 0.2.

Figure 4: Mass histograms after various cuts (−0.2 to 0.2) on best 3-variable MVA trained with a $61 < M < 92$ mass cut, no mass cut, and mass orthogonalization.

produce MVAs unable to make use of background mass information.

## References


[1] ATLAS Collaboration, "Identification of boosted, hadronically decaying $W$ bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8\,\text{TeV}$," The European Physical Journal C **76**, 1–47 (2016).

[2] Y. Freund, and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences **55**, 119–139 (1997).

[3] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: toolkit for multivariate data analysis," PoS **ACAT**, 040 (2007).

[4] R. Brun, and F. Rademakers, "ROOT - an object oriented data analysis framework," Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **389**, New Computing Techniques in Physics Research V, 81–86 (1997).

[5] ATLAS Collaboration, "Measurement of the cross-section of high transverse momentum vector bosons reconstructed as single jets and studies of jet substructure in $pp$ collisions at $\sqrt{s} = 7\,\text{TeV}$ with the ATLAS detector," New Journal of Physics **16**, 113013 (2014).


## Acknowledgments


First, I would like to express my gratitude toward my thesis advisor, Dr. Joseph Haley, for guiding the topic of my thesis and supporting me throughout its development. Without him, this thesis would not have been possible.

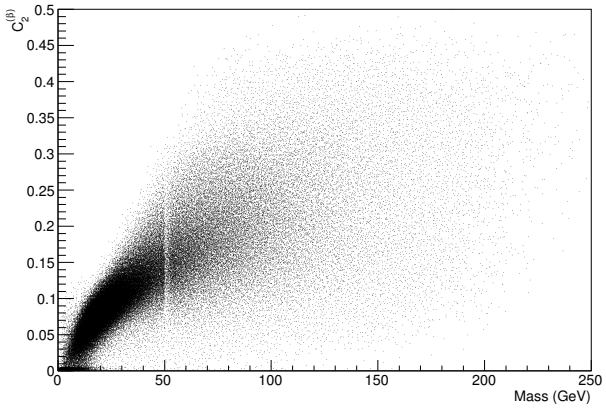
Further, I would like to thank Dr. Flera Rizatdinova for agreeing to be the second reader for this thesis and provide feedback on my work.

I would also like to express gratitude toward all of the physics faculty who have contributed to my understanding of physics since without them this thesis could not have begun.

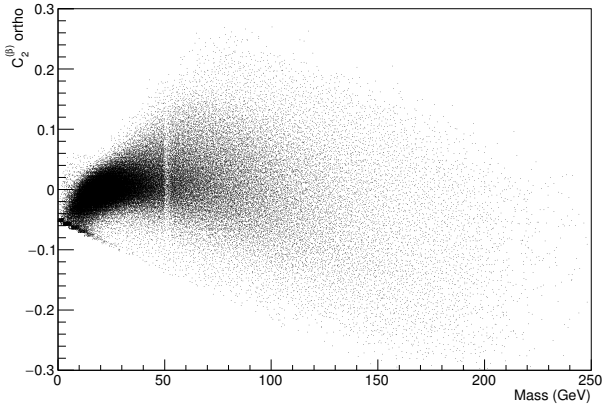
Additionally, I thank my friends who listened to discussions of my work and provided me with invaluable advice and suggestions.

Finally, I would like to thank my parents and family for always supporting me and encouraging me to pursue my interests.


## Appendix A. Variable Scatter Plots




(a) $C_2^{(\beta)}$ before mass orthogonalization.




(b) $C_2^{(\beta)}$ after mass orthogonalization.

Figure 5: Scatter plots of invariant jet mass versus $C_2^{(\beta)}$ before and after mass orthogonalization for background QCD jets.

## Appendix B. MVA Variable List

| Variable | Type | Description |
|---|---|---|
| $\sqrt{d_{12}}$ | Splitting scale | substructure splitting scale |
| $\sqrt{z_{12}}$ | Splitting scale | mass-normalized $\sqrt{d_{12}}$ |
| $\mu_{12}$ | Splitting scale | mass-drop fraction |
| $C_2^{(\beta)}$ | Jet shape | energy correlation ratio |
| $D_2^{(\beta)}$ | Jet shape | energy correlation ratio |
| sphericity | CoM jet shape | derived from sphericity tensor eigenvalues |
| aplanarity | CoM jet shape | derived from sphericity tensor eigenvalues |
| $T_{\mathrm{maj}}$ | CoM jet shape | thrust major |
| $T_{\mathrm{min}}$ | CoM jet shape | thrust minor |
| $R_2^{\mathrm{FW}}$ | CoM jet shape | second to zeroth order Fox–Wolfram moments |
| $\tau_{21}^{\mathrm{wta}}$ | Subjettiness | likelihood composed of $n$ subjets |

Table 1: Variables selected to train 3-variable MVAs [1, 5].