CEPC NOTE

May 31, 2019

Separation of 2 fermions from 4 fermions with event-shape variables at CEPC

Zhu Yongfeng[a]

[a]*Institute of High Energy Physics*
[a]*University of Chinese Academy of Sciences*

Abstract

Event shape variables are used to describe event structure in high energy physics. Different event shape variables have different definitions, and can represent different aspect of an event. Using the $e^+e^- \to WW \to 4$ light-quarks (four-fermion) and $e^+e^- \to 2$ light-quarks (two-fermion) simulation samples at CEPC with center-of-mass energy of 240 GeV, this note demonstrates that several event shape variables can separate two fermions samples from four fermions samples with $efficiency \times purity$ higher than 90%. After combining these event shape variables with Adaptive Boosting (AdaBoost, named BDT by TMVA), the $efficiency \times purity$ can reach 0.9397.

E-mail address: zhuyf@ihep.ac.cn

Contents

1 Introduction

Event-shape variables characterize geometrical properties of final states observed in high energy collisions, it can also describe the energy and momentum of final state particles. This note introduces several event shape variables to describe the event structure of $e^+e^- \rightarrow WW \rightarrow 4$ light-quarks and $e^+e^- \rightarrow 2$ light-quarks processes at MCTruth level and Reconstruction level at CEPC. Furthermore, this note uses these variables to separate two fermions samples from four fermions samples with a high $efficiency \times purity$. Section 2 introduces the sample used in this note. Section 3 introduces the event shape variables used in this note, and shows the distribution of each variable in two processes mentioned above, followed by the distribution of efficiency, purity and $efficiency \times purity$ of each variable when separate two fermions from four fermions. Section 4 combines the event-shape variables mentioned in 3 with BDT in TMVA to separate two types of events. Section 5 gives a brief conclusion about this note.

2 Sample

The $e^+e^- \rightarrow 2$ light-quarks, called two-fermion process, acts as the signal and the $e^+e^- \rightarrow WW \rightarrow 4$ light-quarks, called four-fermion process, acts as the background. For the process of two fermions, there would be an energetic initial state radiation, which takes away around 120 GeV of energy. The MCTruth energy distribution of two quarks in the $e^+e^- \rightarrow 2$ light-quarks process is shown as Figure 1.

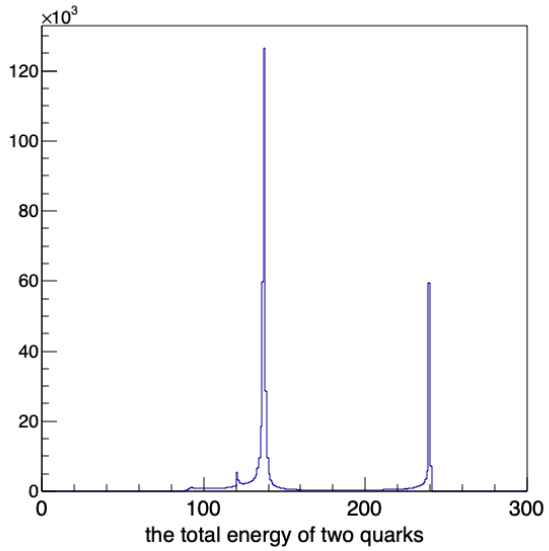

Figure 1: The energy distribution of two quarks in the process $e^+e^- \rightarrow$ 2light-quarks at MCTruth level.

For the two-fermion process, this note only considers the samples with the total energy of two quarks larger than 230 GeV at MCTruth level. When the CEPC operates at 240 GeV, the two-fermion process has 9.09 times the statistics of four-fermion process. The sample used in this note had been normalized according to luminosity at CEPC.

3 Event shapes

Event shapes are required to be infrared safe by definition; soft particle emission or collinear parton splitting do not change the value of an event shape for a particular final state. This section introduces

some event shape variables used in electron-positron collision.

3.1 Thrust

- The canonical event shape is thrust. To evaluate the thrust of an event, one first determines the thrust axis n_T, which is the direction of maximum momentum flow. The thrust is then defined as the fraction of particle momentum flowing along the thrust axis.
- The thrust is defined by the expression

$$T = max_{n_T}\left(\frac{1}{\sum_{j=1}^{N_{particles}} |P_j|} \sum_{i=1}^{N_{particles}} |P_i \cdot n_T|\right)$$

  where P_i is the 3-momentum of particle i and n_T is a unit vector, chosen in the direction along which the value of the thrust is maximal. For a two particle final state the value of the thrust is trivially T = 1, since then the thrust vector is aligned with both back-to-back momenta. For final states with more than two particles the thrust takes values in the range $0.5 \leq T \leq 1$. A value T = 1 coincides with the limit of two jets of collinear particles, while T = 0.5 occurs in the limit of a spherically symmetric distribution.
- For every event, we can get the unit vector and corresponding value of T.
- The distribution of T for two processes are shown as following.

At MCTruth level, the distribution of T for two kinds of samples are shown as Figure 2, the two top graphs are at MCTruth level and two bottom graphs are at reconstruction level. The max value of $efficiency \times purity$ can reach 0.9282 and 0.9270 for MCTruth level and reconstruction level, respectively.

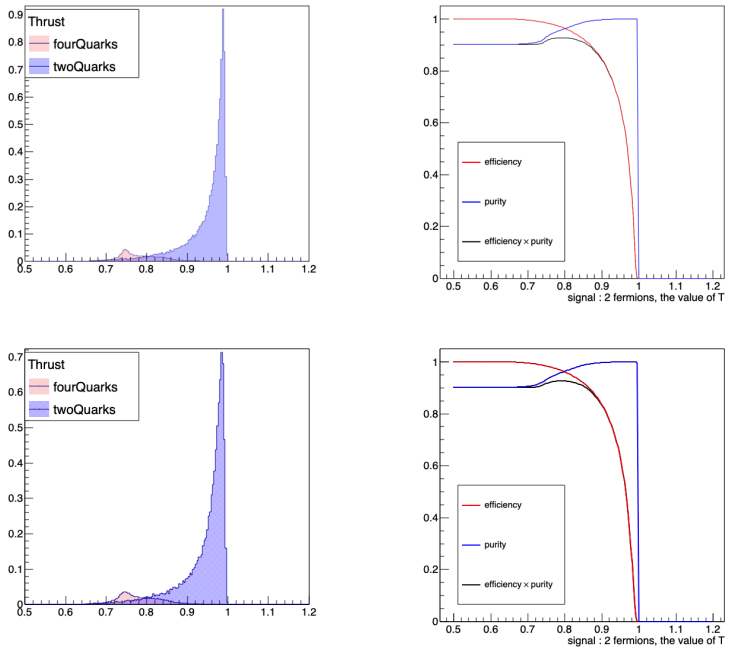

Figure 2: The distribution of T for two kinds of samples, the two top graphs are at MCTruth level, while two bottom graphs at reconstruction level. The max value of $efficiency \times purity$ can reach 0.9282 and 0.9270 for MCTruth level and reconstruction level, respectively.

3.2 Heavy-Jet mass

- We can introduce a plane that is perpendicular to the thrust axis and use it to divide space into two hemispheres.
- Each hemisphere H_i can be assigned an invariant mass, shown as the following function,

$$M_1^2/s = \frac{1}{E_{vis}^2}(\sum_{i=1,P_i\cdot n_T>0}^{N_{particles}} P_i)^2$$

$$M_2^2/s = \frac{1}{E_{vis}^2}(\sum_{i=1,P_i\cdot n_T<0}^{N_{particles}} P_i)^2$$

  where E_{vis} is the total energy of final state particles and P_i is the 4-momentum of particle i.
- Heavy jet mass :

$$M_h^2/s = max(M_1^2/s, M_2^2/s)$$

- Note that in the case of two massless final state partons, the value of the parameter is trivial $M_h^2/s = 0$.

The distribution of heavy jet mass, M_h^2/s is shown as Figure 3. The max value of $efficiency \times purity$ can reach 0.9009 and 0.9010 for MCTruth level and reconstruction level, respectively.

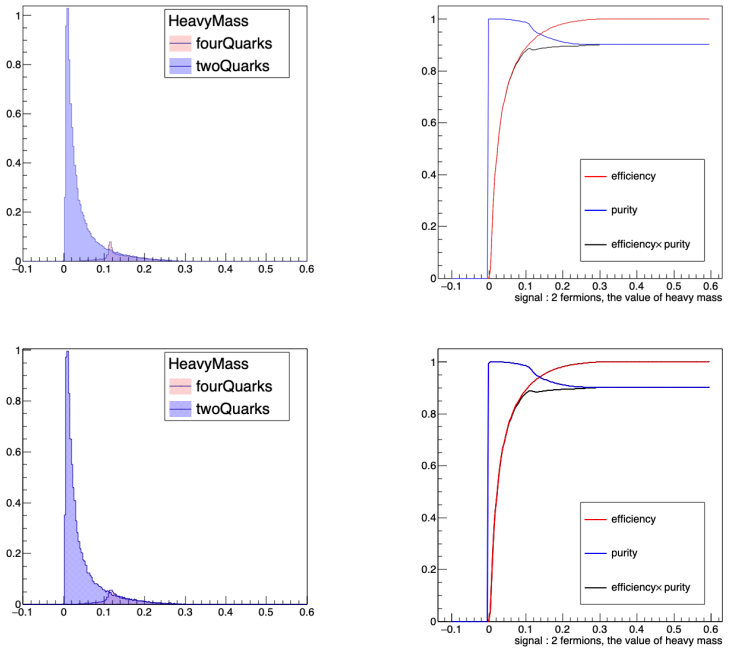

Figure 3: The distribution of heavy jet mass, M_h^2/s, for two kinds of samples, the two top graphs are at MCTruth level, while two bottom graphs at reconstruction level. The max value of $efficiency \times purity$ can reach 0.9009 and 0.9010 for MCTruth level and reconstruction level, respectively.

3.3 C and D parameter

- The linearized sphericity is defined as

$$L^{ab} = \frac{1}{\sum_{j=1}^{N_{particles}} |P_j|^2} \sum_{i=1}^{N_{particles}} \frac{P_i^a P_i^b}{|P_i|}$$

  where P_i is the 3-momentum of particle i and P_i^a denotes the component a of the 3-momentum of the particle i.

- C parameter : $C = 3(\lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3)$, where λ is the eigenvalue of L^{ab}.

- D parameter : $D = 27\ \lambda_1\ \lambda_2\ \lambda_3$

- In the case of two final state partons with momenta along the z axis, the tensor has only one nonzero value and thus only one nonzero eigenvalue. Thus in that case, C = 0.

The distribution of C parameter is shown in Figure 4. The max value of $efficiency \times purity$ can reach 0.9288 and 0.9278 for MCTruth level and reconstruction level, respectively.

The distribution of D parameter is shown as Figure 5. The max value of $efficiency \times purity$ is 0.9294 and 0.9279 for MCTruth level and reconstruction level, respectively.

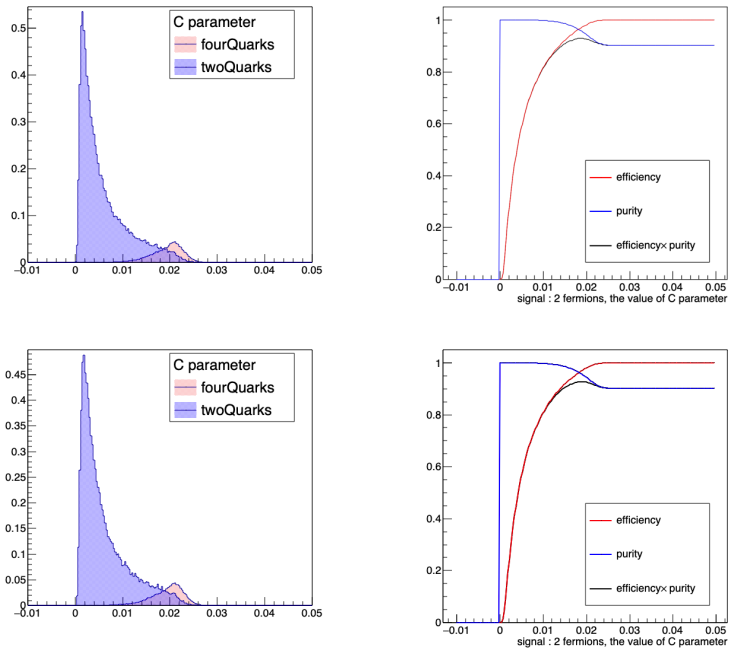

Figure 4: The distribution of C parameter for two kinds of samples, the two top graphs are at MCTruth level, while two bottom graphs are at reconstruction level. The max value of $efficiency \times purity$ can reach 0.9288 and 0.9278 for MCTruth level and reconstruction level, respectively.

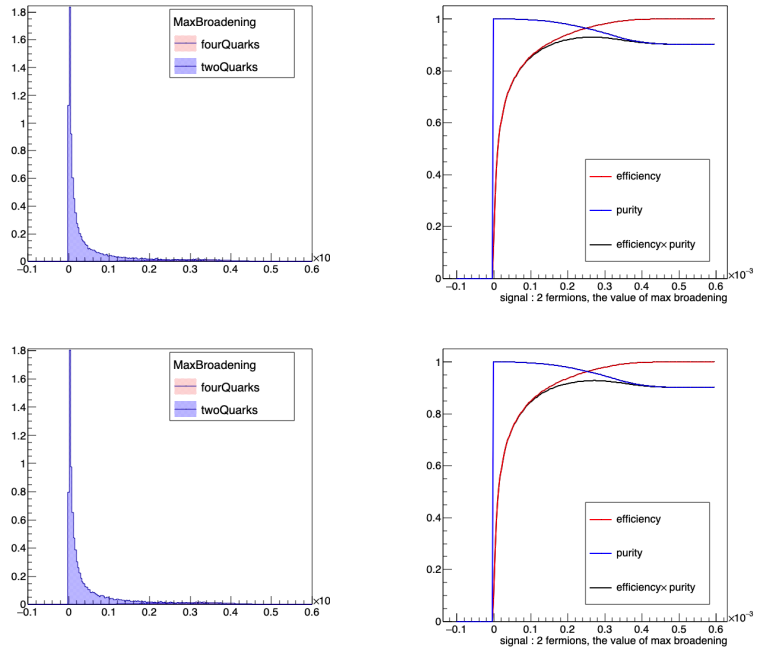

Figure 5: The distribution of D parameter for two kinds of samples, the two top graphs are at MCTruth level, while two bottom graphs are at reconstruction level. The max value of $efficiency \times purity$ can reach 0.9294 and 0.9279 for MCTruth level and reconstruction level, respectively.

3.4 Jet Broadening

- Values related to the transverse momentum measured with respect to the thrust axis.
- After getting the two hemispheres, we introduce the notion of jet broadening, which is defined as

$$B_1 = \frac{1}{2\sum_{j=1}^{N_{particles}} |P_j|} \sum_{i=1, P_i \cdot n_T > 0}^{N_{particles}} |P_i \times n_T|$$

$$B_2 = \frac{1}{2\sum_{j=1}^{N_{particles}} |P_j|} \sum_{i=1, P_i \cdot n_T < 0}^{N_{particles}} |P_i \times n_T|$$

  where P_i is the 3-momentum of particle i.

- total jet broadening : $B_T = B_1 + B_2$
- wide jet broadening : $B_W = max(B_1, B_2)$
- narrow jet broadening : $B_N = min(B_1, B_2)$
- $B_1 = 0$ and $B_2 = 0$ in the case of two final state partons.

The distribution of total jet broadening, B_T, is shown in Figure 6. The max value of $efficiency \times purity$ can reach 0.9253 and 0.9245 for MCTruth level and reconstruction level, respectively. Figure 7 shows the distribution of wide jet broadening, B_W. The max value of $efficiency \times purity$ can reach 0.9008 and 0.9009 for MCTruth level and reconstruction level, respectively.

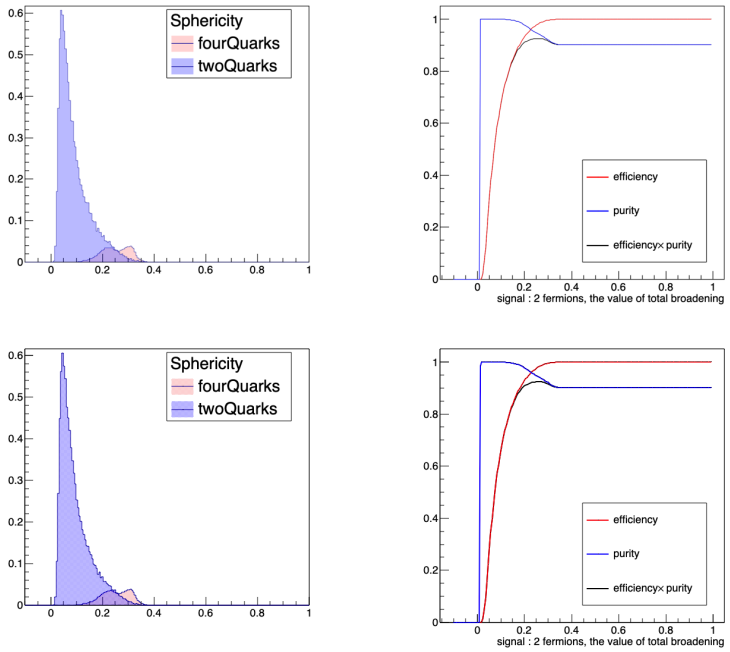

Figure 6: The distribution of total jet broadening, B_T, for two kinds of samples, the two top graphs are at MCTruth level, while two bottom graphs are at reconstruction level. The max value of $efficiency \times purity$ can reach 0.9253 and 0.9245 for MCTruth level and reconstruction level, respectively.

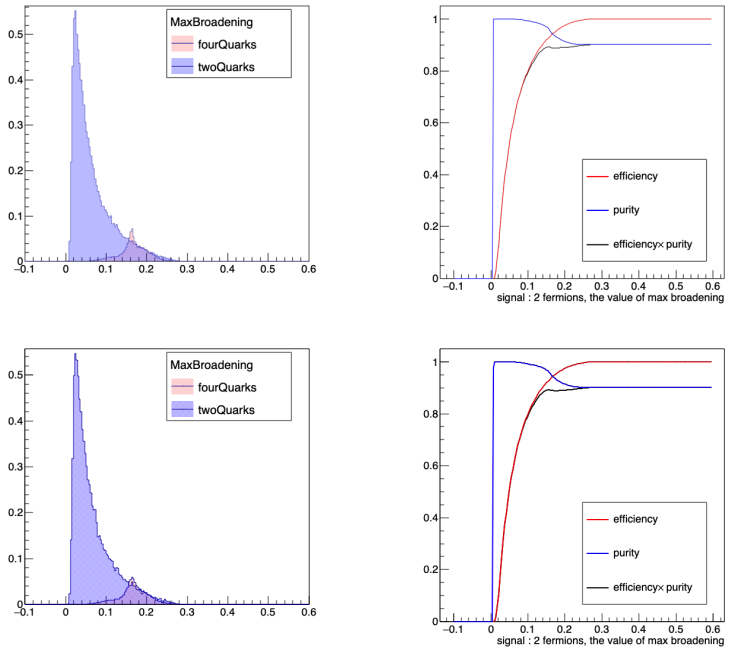

Figure 7: The distribution of wide jet broadening, B_W, for two kinds of samples, the two top graphs are at MCTruth level, while two bottom graphs are at reconstruction level. The max value of $efficiency \times purity$ can reach 0.9008 and 0.9009 for MCTruth level and reconstruction level, respectively.

4 Combining with BDT

The six event shape variables mentioned above demonstrate good performance in separating two fermions from four fermions. What the separation performance would be after combining these six variables? Next, this note will analyse the correlation of these six variables and get the max value of efficiency×purity after combining them with BDT.

4.1 Adaptive Boosting, also called BDT by TMVA

Boosting is a typical method in machine learning, it refers to any ensemble method that can combine several weak learners into a strong learner. The general idea of most boosting methods is to train predictors sequentially, each trying to correct its predecessor. One way for a new predictor to correct its predecessor is to pay a bit more attention to the training instances that the predecessor under-fitted. This results in new predictors focusing more and more on the hard cases. This is the technique used by AdaBoost.

For example, to build an AdaBoost classifier, a first base classifier (Decision Tree used in this note) is trained and used to make predictions on the training set. The relative weight of misclassified training instances is then increased. A second classifier is trained using the updated weights and again it makes predictions on the training set, weights are updated, and so on. The whole process is repeated and the algorithm stops when the desired number of predictors is reached, or when a perfect predictor is found. To make predictions, AdaBoost simply computes the predictions of all the predictors and weights them using the predictor weights, which is computed depending on their overall accuracy on the weighted training set.

4.2 Correlation between event shape variables

Figure 8 shows the correlation matrix between event shape variables for signal and background at MCTruth level. It shows that there are strong correlations between these variables. Figure 9 shows the correlation matrix at reconstruction level. Figure 10 shows the correlations between six variables for signal events, while Figure 11 shows that for background events. They are both at MCTruth level.

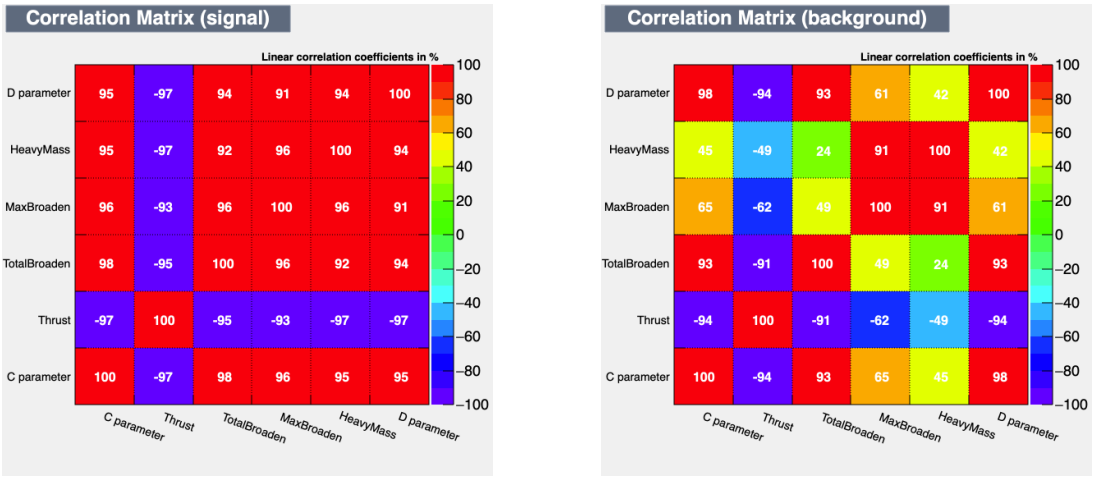

Figure 8: The correlations between these six event shape variables at MCTruth level. The left graph is about the signal, while the right graph is about the background.

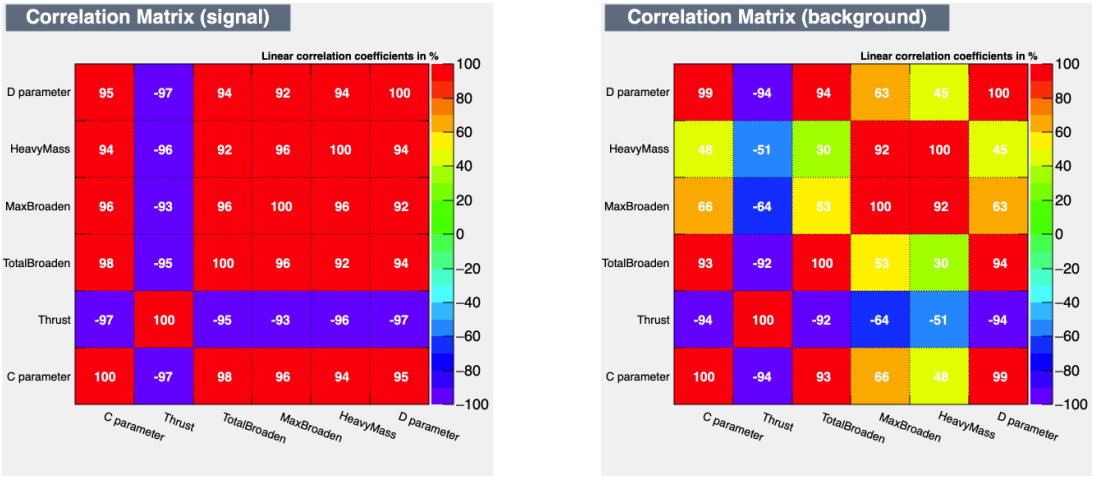

Figure 9: The correlations between these six event shape variables at reconstruction level. The left graph is about the signal, while the right graph is about the background.

Figure 10 and Figure 11 show the correlations between each event shape variable at MCTruth level, they are highly correlated.

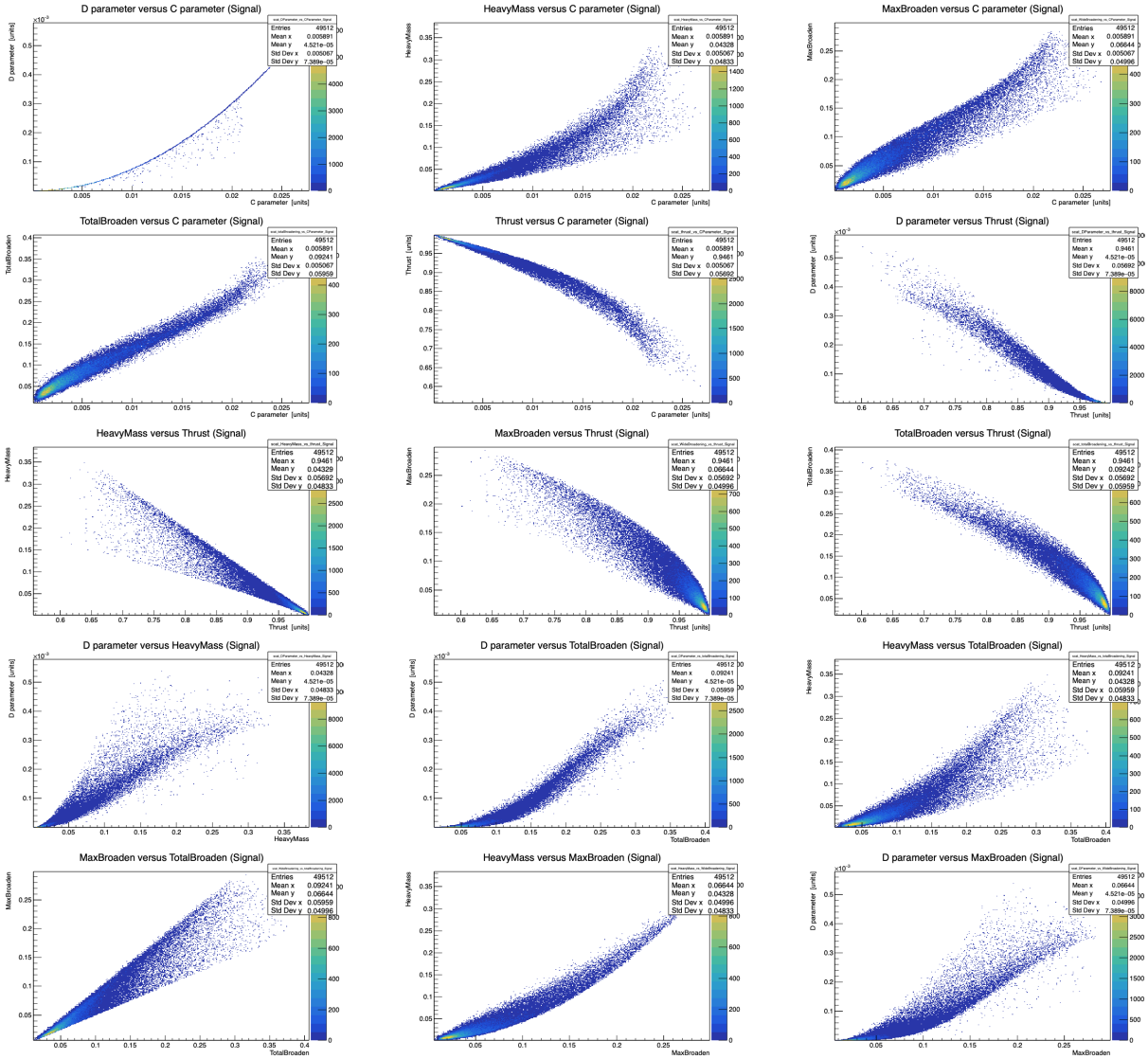

Figure 10: The correlations between six event shape variables for signal samples at MCTruth level.

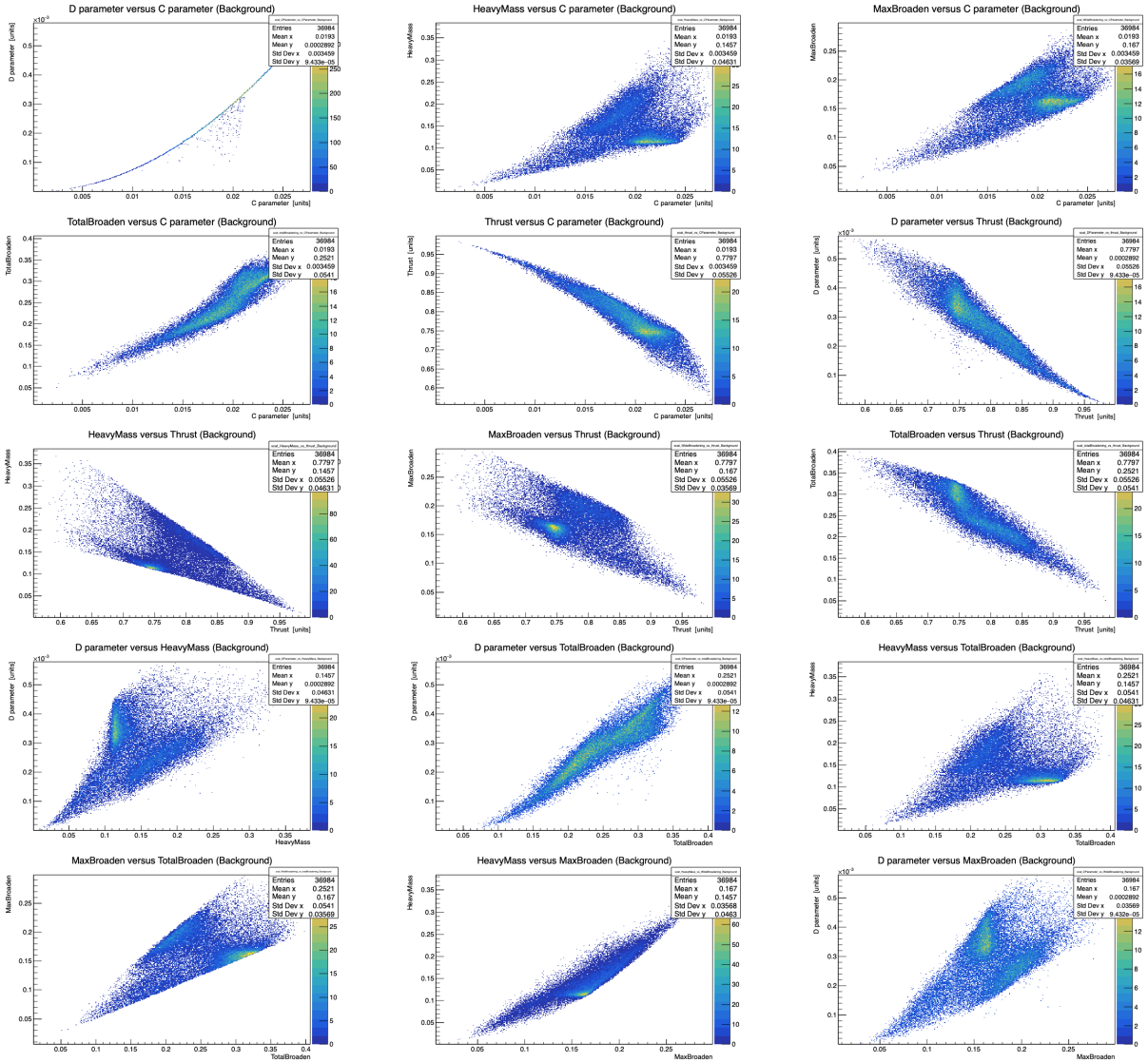

Figure 11: The correlations between six event shape variables for background samples at MCTruth level.

4.3 $Efficiency \times Purity$ with BDT

After combining these six variables with BDT, Figure 12 shows the distribution of the value of BDT for signal and background samples at MCTruth level and reconstruction level, respectively.

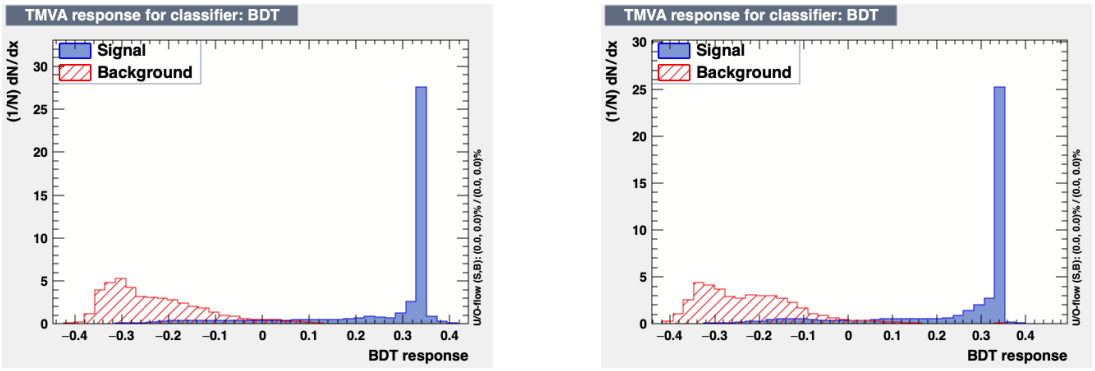

Figure 12: The distribution of the value of BDT for signal and background samples, the left graph is at MCTruth level and right graph is at reconstruction level.

The importance of each variable also can been calculated, the following table shows the importance of each variable.

event-shape variable	T	Heavy Jet Mass	Max Jet Broadening	Total Jet Broadening	C parameter	D parameter
importance factor	0.195	0.045	0.245	0.185	0.245	0.085

After normalizing signal and background according to luminosity at CEPC, Figure 13 shows the variation of efficiency, purity and $efficiency \times purity$ when BDT takes different values. The max value of $efficiency \times purity$ can reach 0.9397 at MCTruth level and 0.9367 at reconstruction level.

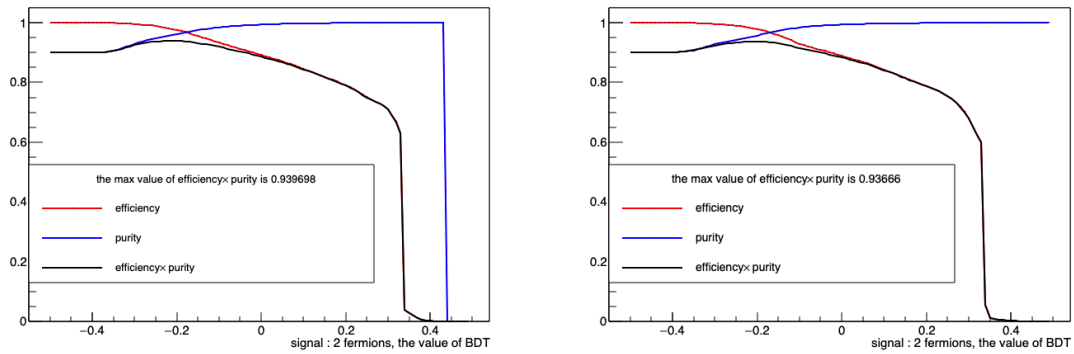

Figure 13: The variation of efficiency, purity and $efficiency \times purity$ when BDT takes different values. The left graph is at MCTruth level, max $efficiency \times purity$ can reach 0.9397. The right graph is at reconstruction level, max $efficiency \times purity$ can reach 0.9367.

5 Conclusion

In section 3, two-fermion samples acted as signal, we can also take four-fermion samples as signal, the max value of efficiency×purity in different conditions is shown in the following table.

max efficiency×purity		T	Heavy Jet Mass	Max Jet Broadening	Total Jet Broadening	C parameter	D parameter
MCTruth	signal : 4 fermions	0.4949	0.4205	0.3896	0.4558	0.4872	0.4891
	signal : 2 fermions	0.9282	0.9009	0.9008	0.9253	0.9288	0.9294
Reco	signal : 4 fermions	0.4874	0.4140	0.3869	0.4489	0.4780	0.4790
	signal : 2 fermions	0.9270	0.9010	0.9009	0.9245	0.9278	0.9279

Since the luminosity of process $e^+e^- \to 2$ quarks is 9.09 times larger than that of the process $e^+e^- \to WW \to 4$ quarks at CEPC, the max value of efficiency×purity is larger when we take two fermions as signal. All event shape variables mentioned above can separate two fermions from four fermions with efficiency×purity higher than 0.9.

After combining these six event shape variables with BDT, the max value of $efficiency \times purity$ can reach 0.9397 at MCTruth level and 0.9367 at reconstruction level.