Missing values treatment in event classification

Aleksandr Petukhov, Evgeny Soldatov, Konstantin Savelev
National Research Nuclear University MEPhI

ICNFP-2021, Crete
27.08.2021

What are the missing values?

In machine learning:

- data corruption
- failure to record data

Such data are usually **not considered** in a high energy physics analysis.

However, machine learning algorithms are coming up with ways to treat these values **without assigning** them anything.

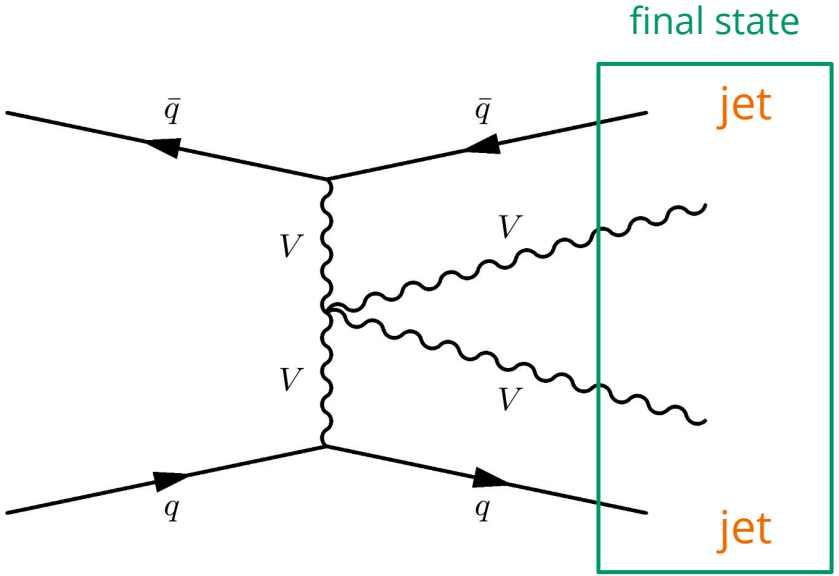
Is there an application in high energy physics?

What are the missing values?

In high energy physics:

- parameters of particles not present in all of the considered events

Example: vector boson scattering (VBS) processes with VVjj final states in proton collisions.

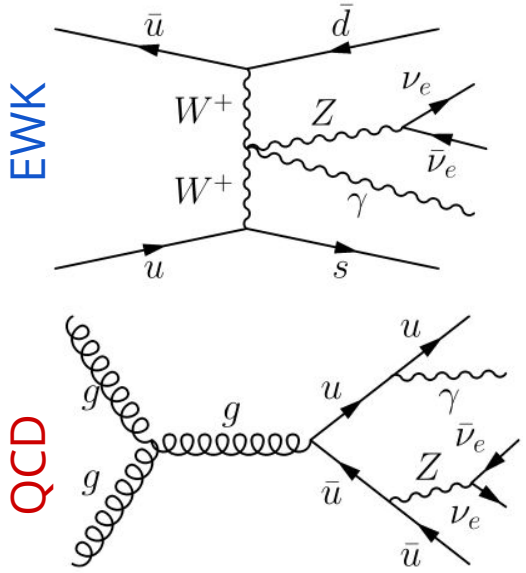


- ≥ 2 hadron jet final state
- additional hadron jet variables could be used for discrimination

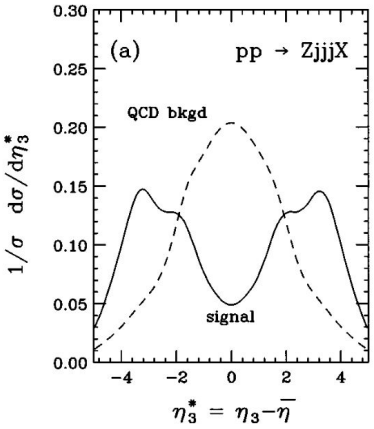
Can we use machine learning approaches to the missing values for this case?

Benchmark processes

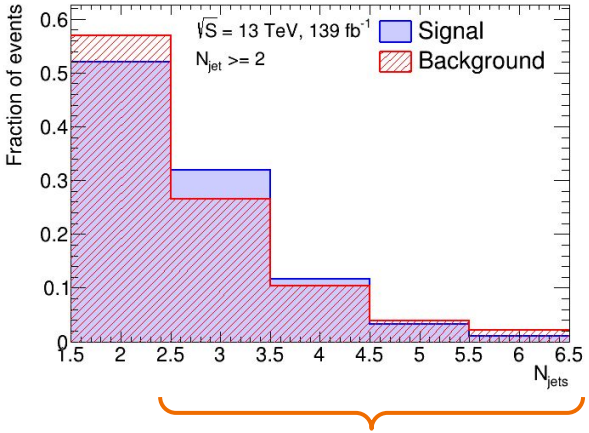
Signal: electroweak (EWK) associated Z(νν̄)γjj production in pp-collisions
Background: QCD associated Z(νν̄)γjj production



- defined by 2 jets in the final state
- 3 jet variables could be used for disctimination



[source]



~50% events have
3 or more jets

Machine learning algorithms

Two main Boosted Decision Tree algorithms have an automated way to treat missing variables

- XGBoost
- **LightGBM**

Preliminary test show no difference in the results, so the **LightGBM** was adopted as a faster one

Performance metric: statistical significance

$$\frac{S}{\sqrt{S+B}}$$

S - number of signal events
B - number of background events

Studied approaches

Base. ≥2 jets. No 3rd jet variable used. Reference.

Using 3rd jet information

Clustering. Split the samples into two categories, with two classifiers:

1) =2 jets — using no 3rd jet variables
2) ≥3 jets — using 3 jet variables

Combine the result.

Imputation. ≥2 jets. Use 3rd jet variables but set the distinct values for events with 2 jets.

Automated (from LightGBM)**.** ≥2 jets. No special treatment for 3rd jet variables in events with 2 jets

Dataset used for the study

Process: $pp \rightarrow Z\gamma jj$, $Z \rightarrow \nu\bar{\nu}$
Z-boson is observed with missing transverse energy (E_T^{miss})

Data:

- MadGraph + Pythia8 + Delphes (with ATLAS card)
- $\sqrt{S}$ = 13 TeV
- Normalized for L = 139 fb^{-1}

} 2015-2018 ATLAS datataking

Final state objects:

- ≥ 2 hadron jets
- high energy photon
- E_T^{miss} with large magnitude

See backup for details on the selection

Studied variables

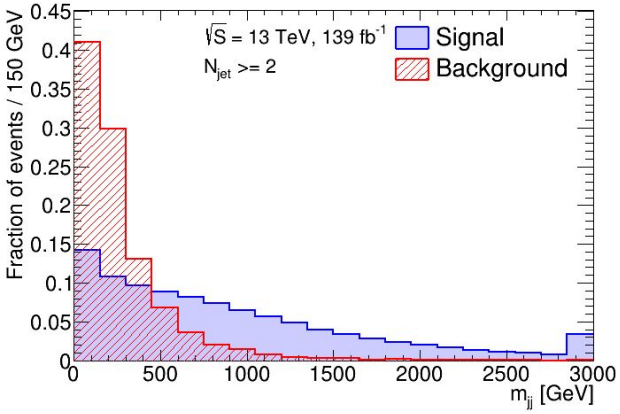
34 variables with photon, E_T^{miss} and 2 leading jet parameters

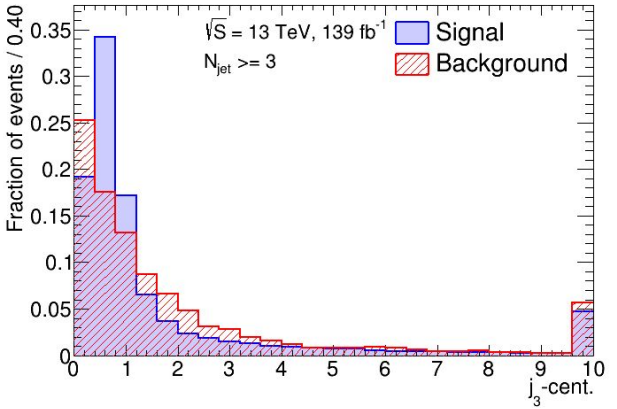
- each object parameters
- m[jj], jet pair invariant mass
- photon centrality
- ΔY[jj]

11 variables with 3rd jet parameters

- m[jjj], invariant mass of 3 jets
- 3rd jet centrality

$$X\text{-centrality} = \left| \frac{y(X) - \frac{y(j_1)+y(j_2)}{2}}{y(j_1)-y(j_2)} \right|$$





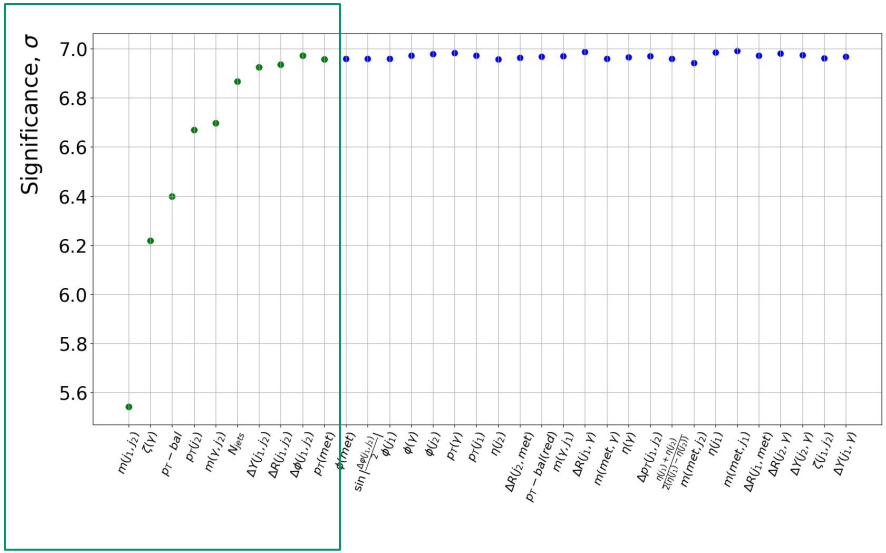
all variables are listed in the backup

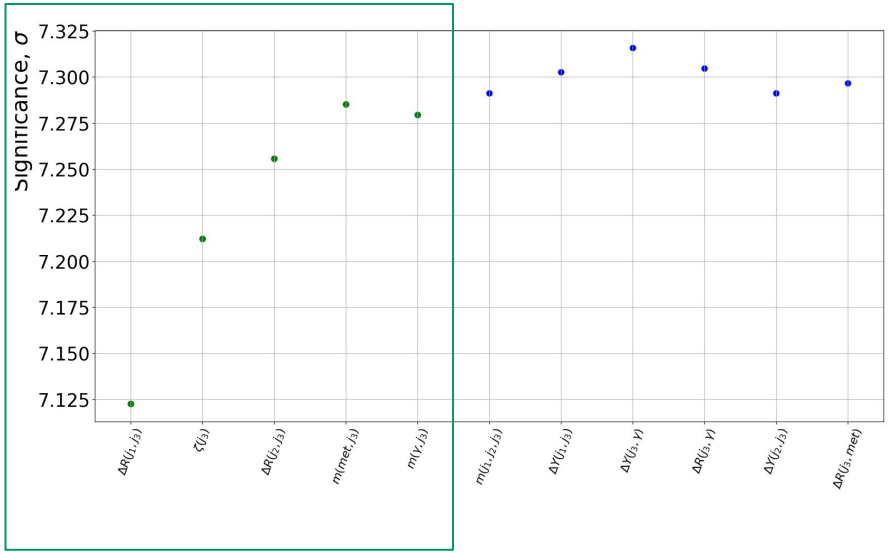
Classifier optimization

Default hyperparameters of the **Base** approach used to select the nominal variable set.

Default hyperparameters the **Automated** approach used to further select the additional set of 3rd jet variables.

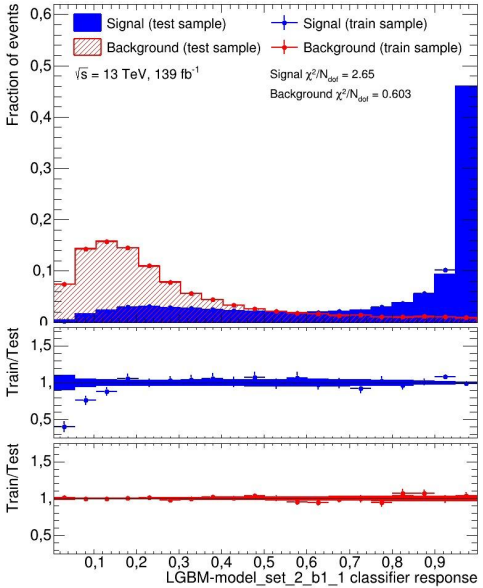
Selected variable sets are used to optimize the hyperparameters of all of the approaches.



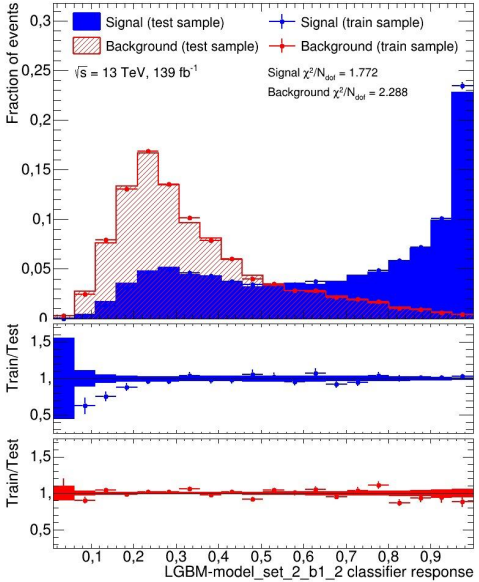


Problem with clusterization

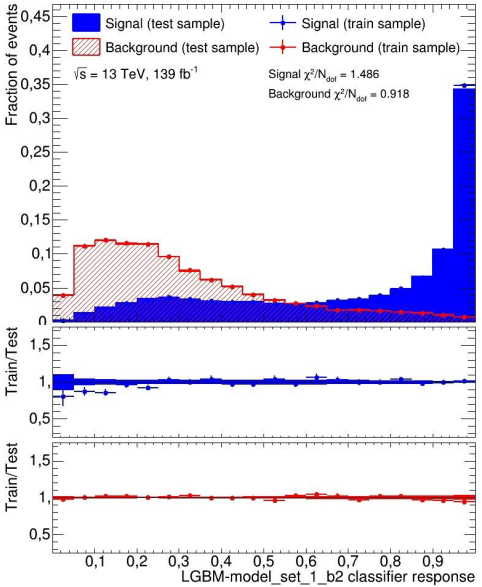
=2 jets, **Base** hyperparameters



≥3 jets, **Base** hyperparameters



≥3 jets, tuned hyperparameters



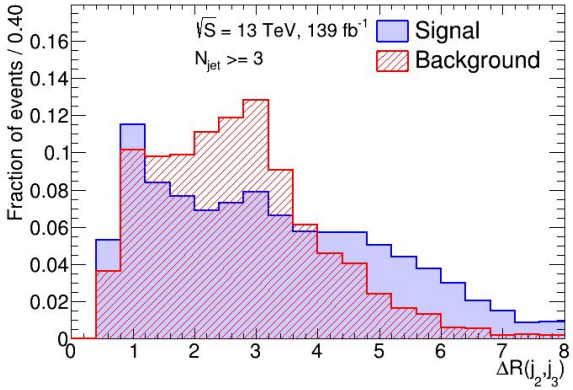
The model for the ≥3 jets region is prone to overtraining → requires additional optimization → more time and computer resources required

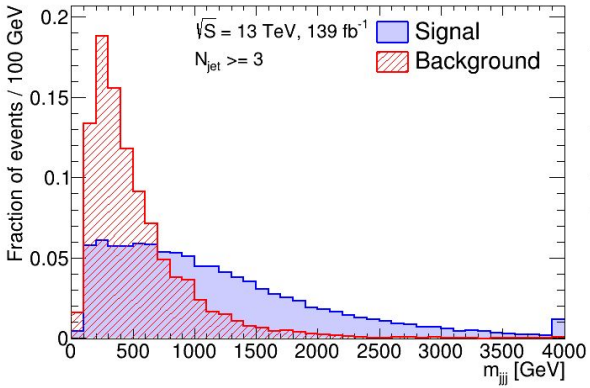
Imputation

Used only variables with positive values.

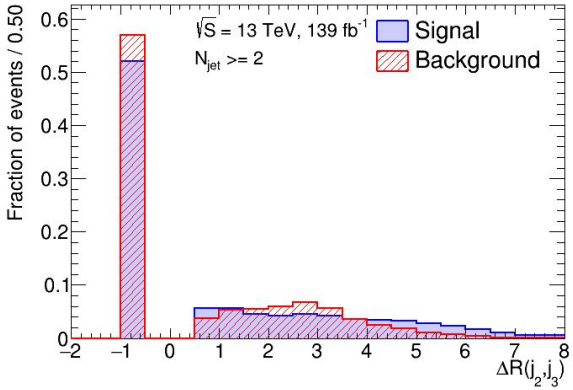
For events with 2 jets the values of the variables were set to -1.

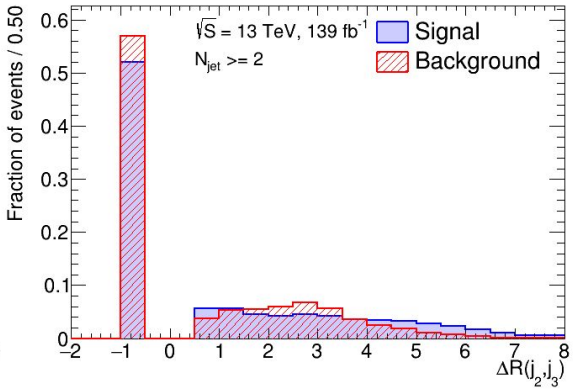
≥3 jets





≥2 jets





Results

	Algorithm	Signal events	Background events	Significance, σ
	Base	81.5 ± 0.5	53.0 ± 1.0	7.03 ± 0.04
Using 3rd jet information	Clustering	97.8 ± 0.6	76.8 ± 1.3	7.40 ± 0.05
	Imputation	92.4 ± 0.5	64.7 ± 1.2	7.37 ± 0.04
	Automated	99.2 ± 0.6	84.6 ± 1.3	7.31 ± 0.04

Use of 3rd jet information allows for a better statistical significance.

Clustering and **imputation** approaches provide the largest increase, but **automated** algorithm together with not very significant loss in performance requires the least computation time or manual variable modification

Conclusion

- Algorithms that use 3rd jet information were used to discriminate VBS signal from it's main background.
- Use of 3rd jet information allowed for a higher statistical significance
- The best performing 3rd jet variables: $\Delta R(j_1, j_2)$, j_3-cent., $\Delta R(j_2, j_3)$, $m(E_T^{miss}, j_3)$, $m(\gamma, j_3)$
- **Clustering** and **imputation** provide the largest increase in significance, however they have some limitations
- **Automated** algorithm from **LightGBM** together with not very significant loss in performance requires the least time and manual modification
- All of the described approaches may be used in further VBS searches in proton collisions. However, **automated** algorithm shows more universality

Backup slides

Preselection

- $p_T^{\text{jet}} > 20$ GeV
- ≥ 2 hadron jets
- no leptons
- $E_T^{\text{miss}} > 120$ GeV
- $p_T^{\gamma} > 150$ GeV
- $\frac{\sum_{\Delta R < 0.4} p_T}{p_T^{\gamma}} < 0.05$ — photon isolation

Nominal variable set. 1/2

1. $p_T(j_1)$;
2. $\varphi(j_1)$;
3. $\eta(j_1)$;
4. $p_T(j_2)$;
5. $\varphi(j_2)$;
6. $\eta(j_2)$;
7. p_T^{γ};
8. $\varphi(\gamma)$;
9. $\eta(\gamma)$;
10. E_T^{miss};
11. $\varphi(\vec{p}_T^{\,\text{miss}})$;
12. m_{jj}, jet pair invariant mass;
13. γ-centrality$= \left| \frac{y(\gamma) - \frac{y(j_1)+y(j_2)}{2}}{y(j_1)-y(j_2)} \right|$;
14. p_T-balance$= \frac{|\vec{p}_T^{\,\text{miss}} + \vec{p}_T^{\,\gamma} + \vec{p}_T^{\,j_1} + \vec{p}_T^{\,j_2}|}{E_T^{\text{miss}} + E_T^{\gamma} + p_T^{j_1} + p_T^{j_2}}$;
15. p_T-balance(reduced)$= \frac{|\vec{p}_T^{\,\gamma} + \vec{p}_T^{\,j_1} + \vec{p}_T^{\,j_2}|}{E_T^{\gamma} + p_T^{j_1} + p_T^{j_2}}$;
16. N_{jets}, number of hadron jets;
17. $\Delta Y(j_1, j_2)$;
18. $\Delta Y(j_1, \gamma)$;
19. $\Delta Y(j_2, \gamma)$;
20. $\Delta R(j_1, j_2)$, where $\Delta R = \sqrt{(\Delta\varphi)^2 + (\Delta\eta)^2}$;
21. $\Delta R(j_1, \gamma)$;
22. $\Delta R(j_2, \gamma)$;
23. $\Delta R(j_1, \vec{p}_T^{\,\text{miss}})$;
24. $\Delta R(j_2, \vec{p}_T^{\,\text{miss}})$;

Highlighted variables were used to create the classifiers

Nominal variable set. 2/2

25. $m(\vec{p}_T^{\,\text{miss}}, \gamma)$;
26. $m(\vec{p}_T^{\,\text{miss}}, j_1)$;
27. $m(\vec{p}_T^{\,\text{miss}}, j_2)$;
28. $m(\gamma, j_1)$;
29. $m(\gamma, j_2)$;
30. $\Delta\varphi(j_1, j_2)$;
31. $\sin\left(\left|\frac{\Delta\varphi(j_1, j_2)}{2}\right|\right)$;
32. $|p_T(j_1) - p_T(j_2)|$;
33. $\text{jet-centrality} = \frac{p_T^{j_1} - p_T^{j_2}}{E^{j_1} + E^{j_2}}$;
34. $\left|\frac{y(j_1)+y(j_2)}{2(y(j_1)-y(j_2))}\right|$

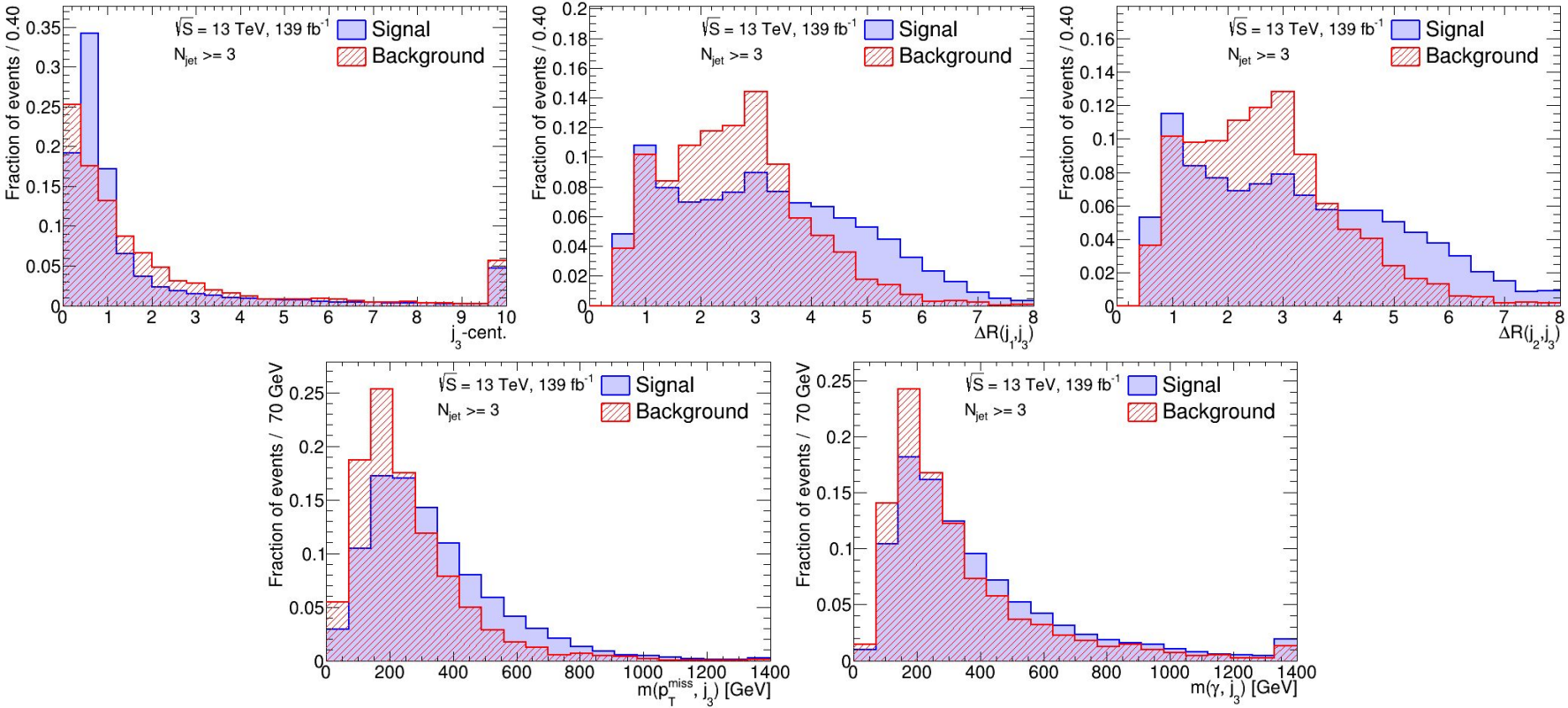
Highlighted variables were used to create the classifiers

3rd jet information variables

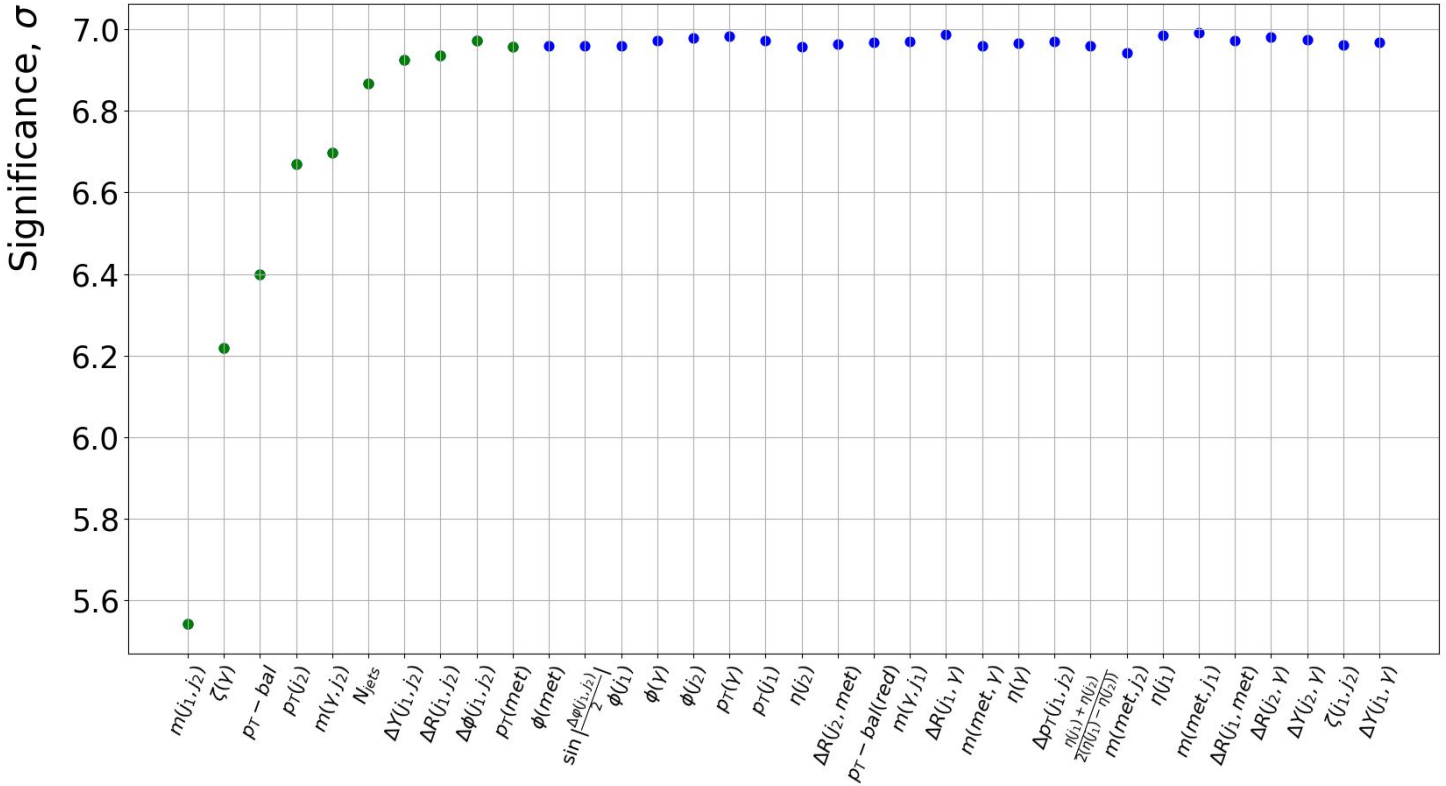
1. m_{jjj};
2. j_3-centrality$= \left| \frac{y(j_3) - \frac{y(j_1)+y(j_2)}{2}}{y(j_1)-y(j_2)} \right|$;
3. $\Delta Y(j_1, j_3)$;
4. $\Delta Y(j_2, j_3)$;
5. $\Delta Y(\gamma, j_3)$;
6. $\Delta R(j_1, j_3)$;
7. $\Delta R(j_2, j_3)$;
8. $\Delta R(\gamma, j_3)$;
9. $\Delta R(\vec{p}_T^{\,\text{miss}}, j_3)$;
10. $m(\vec{p}_T^{\,\text{miss}}, j_3)$;
11. $m(\gamma, j_3)$;

Highlighted variables were used to create the classifiers

Selected 3rd jet variables distributions



Variable selection. Nominal



Variable selection. 3rd jet