BDT Study in 1tau1l(2016)

Search for Four Top in Tau Final States

Anshul Kapoor[1] Huiling Hua[1] Hongbo Liao[1]

[1]IHEP

November 15, 2022

Outline

1. Introduction
2. Results
3. Validation of BDT Variables
4. Summary

Section 1

Introduction

MVA optimization method

- Correlation removal method
  - To achieve best performance while keeping the number of input variables smallest
- Process
  - Start training from 50 most powerful(highest seperation power) variables
  - Remove one variable from the 50 variables list, the removal based on which pair has the largest correction, remove the less powerful variable from the pair. This forms a 49 variables list
  - Train BDT using the 49 variable list. Remove one variable using the same principle as above and we get 48 variable list
  - Repeat the above 2 steps until only one variable left
  - For each training, do the application and then feed the output BDT histgram to combine. Get the expected significance and expected limit
  - Plotting the number of training variables as a function of expected significance and limit

Various ways of determining the quality of training

- AUC
- Scanning the maximum significance
- Expected significance from combine
- Expected limit from combine

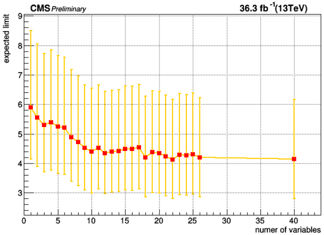
Training setup

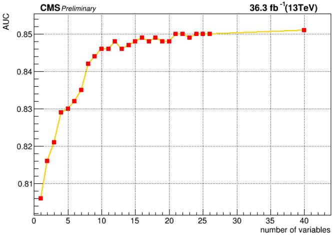
- BDT adaptive boost
- NormMode: None, NumEvents, EqualNumEvents???
  - Using EqualNumEvents
  - To adjust to only care for shape of signal and background. To investigate more
- Treating of negative gen weight events
  - Decision trees can correctly incorporate events with negative weights
  - In cases where a method does not properly treat events with negative weights, it is advisable to ignore such events for the training - but to include them in the performance evaluation to not bias the results
- 60% training, 40% testing
- MC correction: PUweight * EVENT_prefireWeight * EVENT_genWeight

Section 2

Results

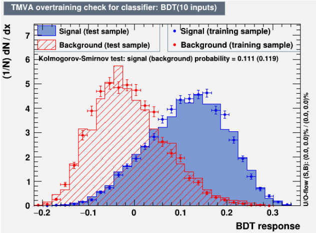
Expected limit Vs number of input variables

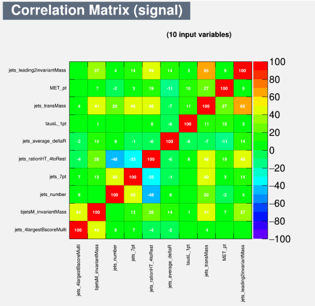




- Reach the plateau at 10 input variables.
- Expected significance 0.6
- We might need to think of some more powerful variables for training as the slope is a bit flat
- The variable list corresponds to the number of variables in the x axis in backup. Correlation matrix for these variables in back up too
- Interpretation of uncertainty of expected limit
  - Expected limit is actually medium limit
  - Expected limit means compare data model with $\mu S + B$ model, test until $f(model, \mu)$ the medium of $f(model, data)$ is 95%
  - Uncertainty: Vary the medium of $f(model, data)$ by up and down 34% and do the test to see for 95% CL the corresponding μ

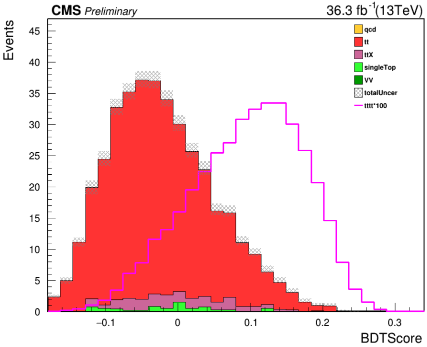
Training performance for 10 input variables





- Why for BDT the score range is [-0.2, 0.4] rather than [-1, 1]?
- Plots for other number of input variables in back up
- ROC to add

Templates for combine for 10 input variables



- For other variable list the template distribution in backup

Datacard to combine 10 variables

```
imax *
jmax *
kmax *
--------------------------------------------------------------------------------
shapes * *  /publicfs/cms/user/huahuil/tauOfTTTT_NanoAOD/TMVAoutput/2016/v3extra1tau1lCut_v41addVertexSelection/1tau1l_v0/AppResults_30bins/TMVApp_1tau1l_10var_f
bin          SR_1tau1l
observation  -1
--------------------------------------------------------------------------------
bin       SR_1tau1l         SR_1tau1l         SR_1tau1l         SR_1tau1l         SR_1tau1l
process   tttt              tt                ttX               VV                singleTop
process   0                 1                 2                 3                 4
rate      -1                -1                -1                -1                -1
--------------------------------------------------------------------------------
SR_1tau1l   autoMCStats  10
```

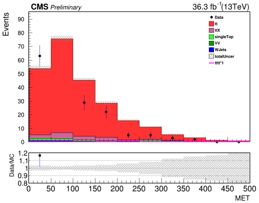
- Combine commands
  - text2workspace.py datacard.txt workspace.root
  - combine -M AsymptoticLimits workspace.root –run blind –name name
  - combine -M Significance workspace.root -t -1 –expectSignal=1 –name name
  - combine not working in CMSSW_12_2_4
- When feed histgrams to combine, should we sum bg samples or not?
  - To me, it doesn't matter. We group for better undertanding for the reading of the datacard

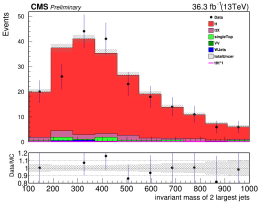
Section 3

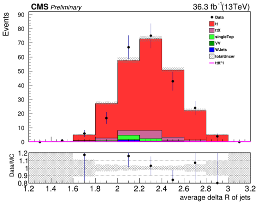
Validation of BDT Variables

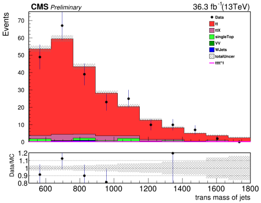
List of final input variables(10)

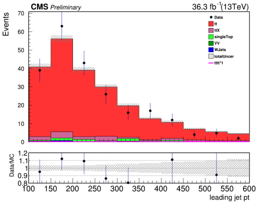
Input variables in tt control region(CR0)

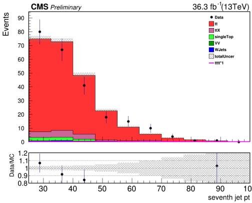






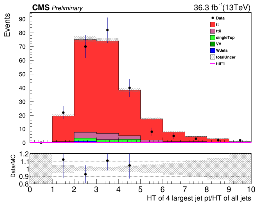


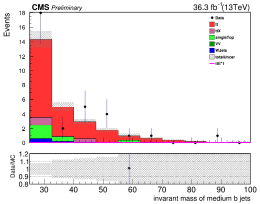


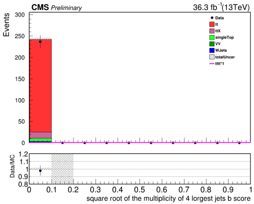


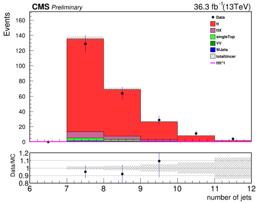
- Only statistic uncertainty included
- MC correction: prefiring reweighting, pileup reweighting
- MET correctionneeded

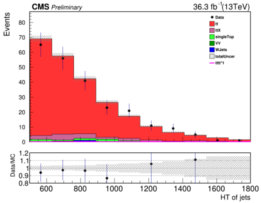
Input variables in tt control region(CR0)

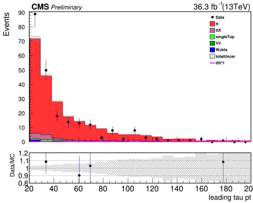






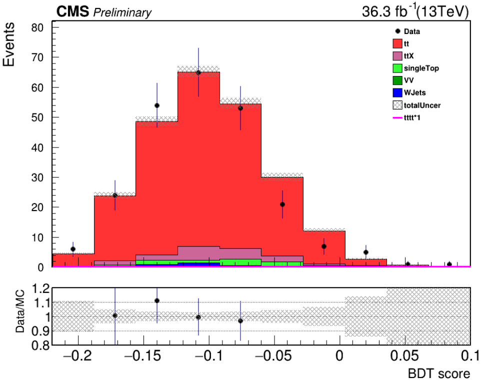






- We need to implement b tag corrections to see if agreement of the b jet related variables improves
- Variables in CR2 and control region definition in back up

BDT discriminant in CR0



- Agreement could improve after apply MET and b tag correction and re train

Section 4

Summary

Summary

- Summary
  - BDT training achived good results for 1tau1l
  - Seems we need to add b tag and MET correction in improve MC modeling of b tag and MET variables
- Next step
  - Add b tag and MET correction
  - Hypeparameter optimization for BDT
  - Repeat for 2017 and 2018

Section 5

Back up

V5 selection

- V5 added good vertex selection compared to v4
- Event yield for v5

1tau1l control region definition

	N_τ	N_l	N_{jets}	N_b
SR	1	1	>=7	>=2
CR0	1	1	>=7	1
CR1	1	1	>=7	0
CR2	1	1	6	>=2
CR3	1	1	6	<2

Table 1: 1tau1l

Total variable list and their meaning

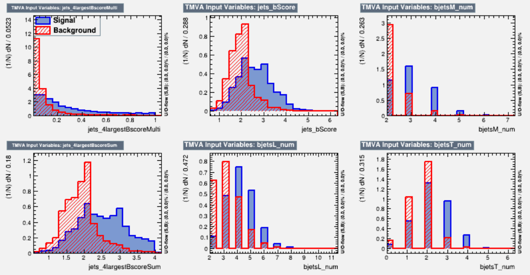
Hyperparameter optimization

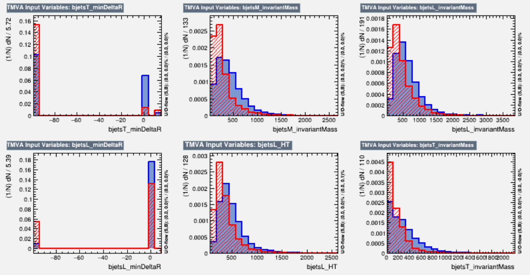
- Hypeparameters for BDT

Questions for my self

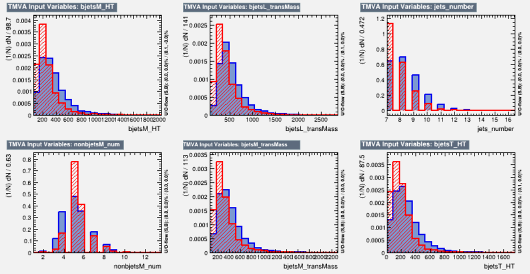
- How the BDT works much better compared to traditional cut based methods? How to understand this?

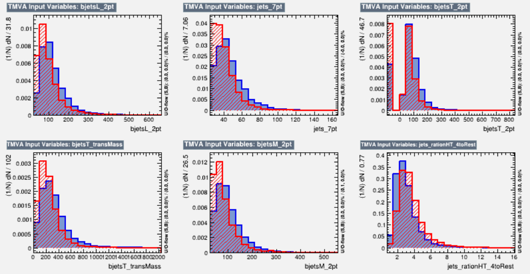
BDT input variable shape



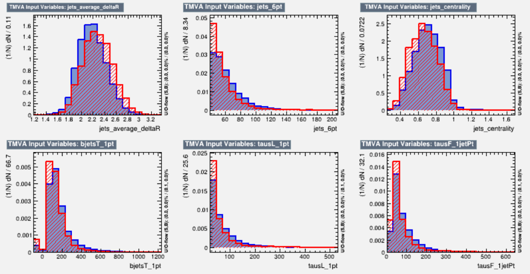


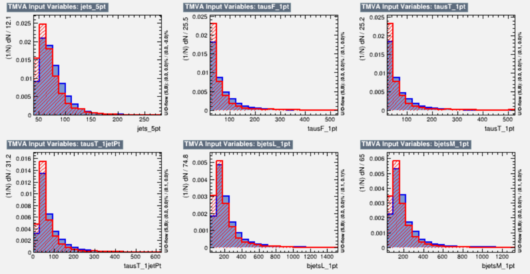
BDT input variable shape



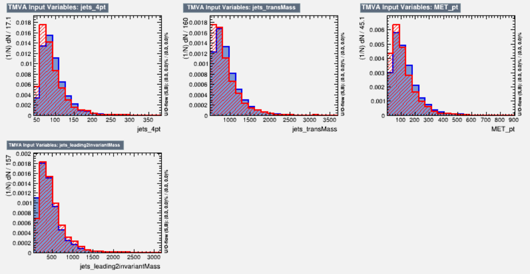


BDT input variable shape

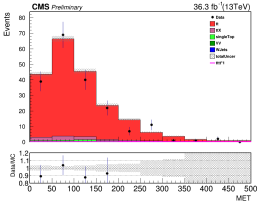


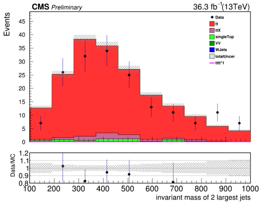


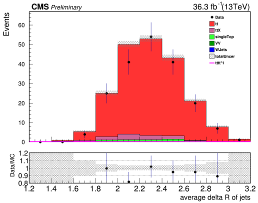
BDT input variable shape

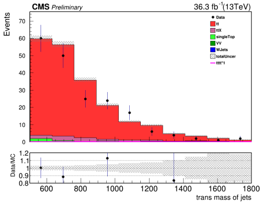


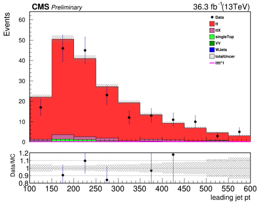
Input variables in tt control region(CR2)





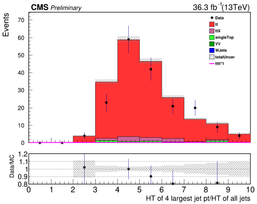


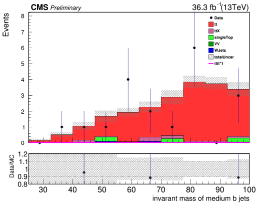


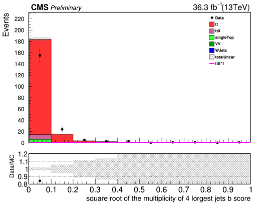


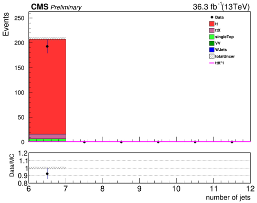
- Only statistic uncertainty included
- MC correction: prefiring reweighting, pileup reweighting
- MET correctionneeded

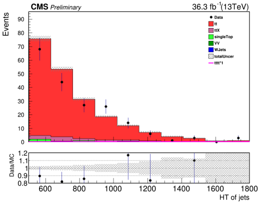
Input variables in tt control region(CR2)

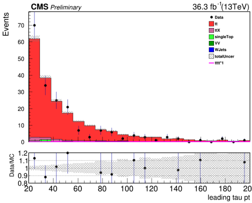






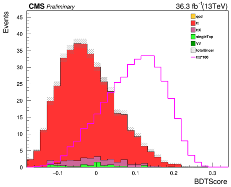


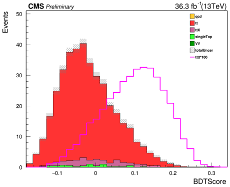


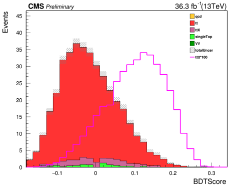


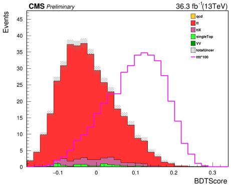
- We need to implement b tag corrections to see if agreement of the b jet related variables improves
- Variables in CR2 and control region definition in back up

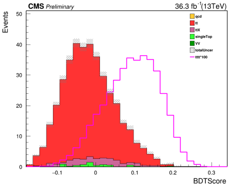
Templates for combine

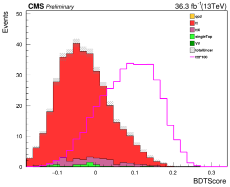






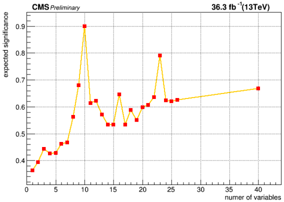






- 10, 11, 12
- 14, 23, 24

Expected limit and significance Vs number of input variables



- Why this fluctuation of expected significance? Intrensic quality of expected significance or something wrong with input templtaes or something wrong with training?
- Why compared to expected significance, expected limit seems more steady?
- How to take into account the uncertainty assioated with expected significance and limit?
- What it expected limit so sensitive to?