13TeV VBF $H \rightarrow \gamma\gamma$ analysis

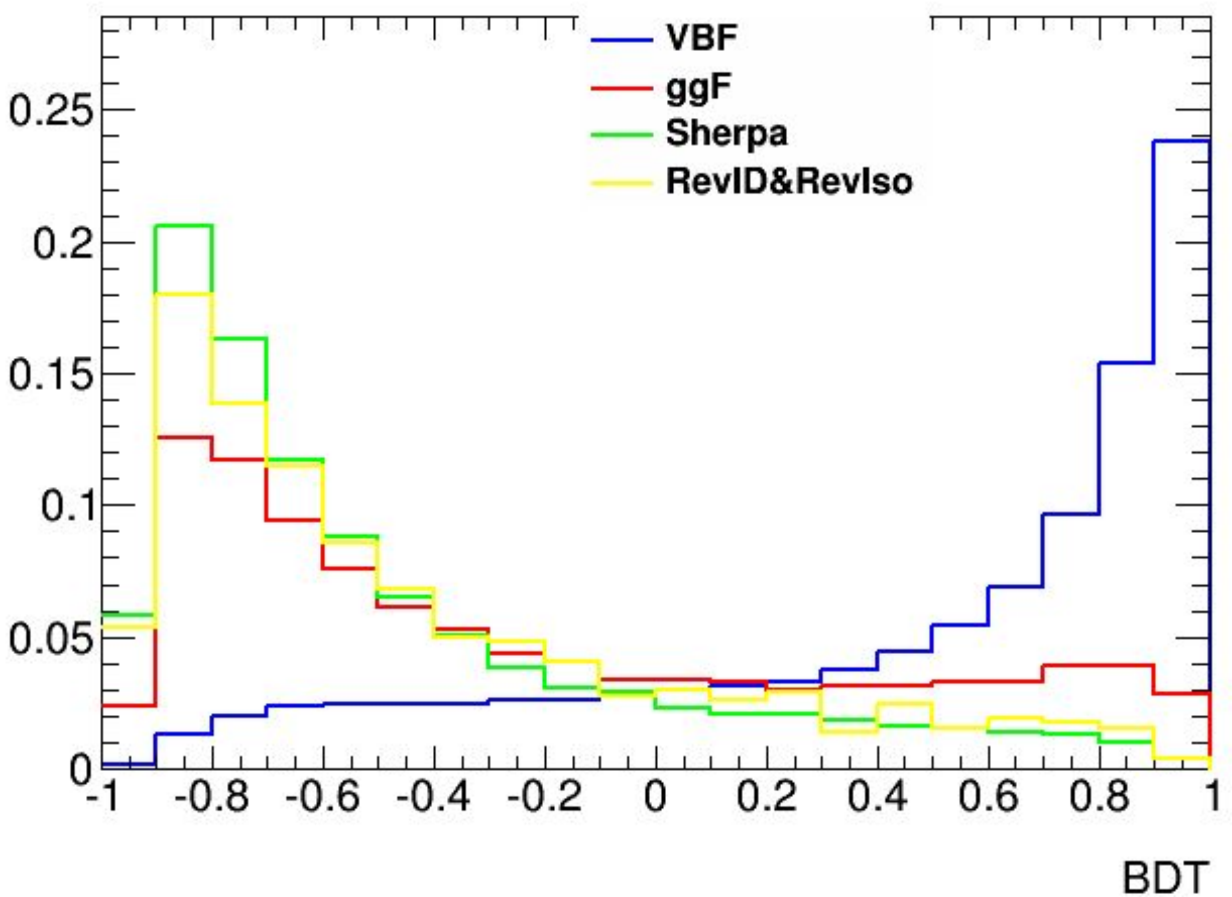
Yu Zhang
09-21

MVA

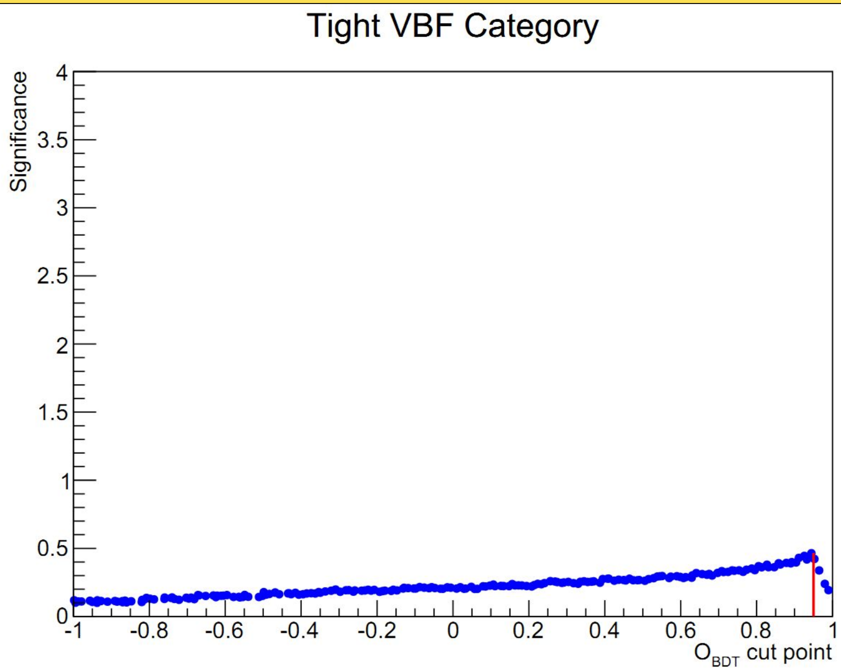
- input variables
  - m_{jj},$\Delta\eta_{jj}$,$\Delta\Phi_{yy,jj}$,p_{Tt},$\Delta R^{min}_{y,j}$,η^{*}
- preselection
  - Njet>=2,$\Delta\eta_{jj}$>2,η^{*}<5
- configuration
  - default one same as run1
- training sample
  - SignalTree:VBF
  - BkgTree:Sherpa $\gamma\gamma$+jets,RevID,RevIso
  - Ratio: 1:1
- optimization
  - scan BDT cut value
  - get threshold value for VBF tight and loose category
  - plot signal and bkgshape in each category

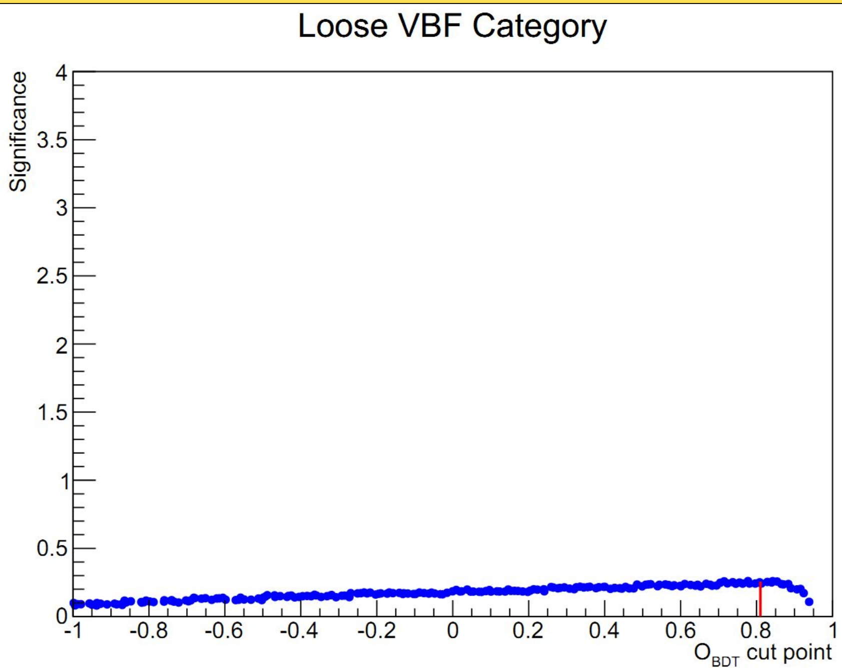
BDT value

- obtain weight xml file after training
- calculate BDT value with xml file
- plot BDT value of signal and bkg samples



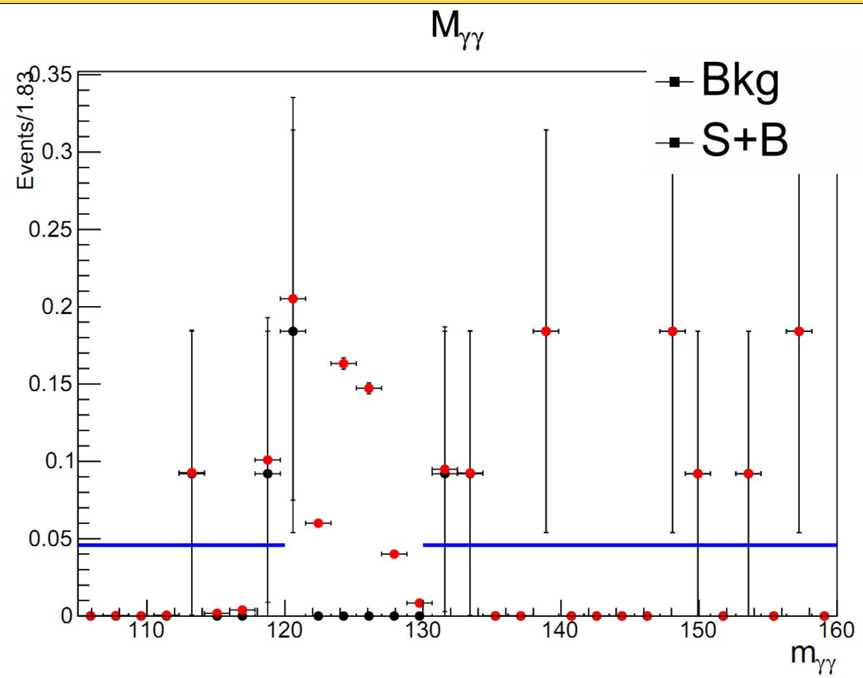
category optimization

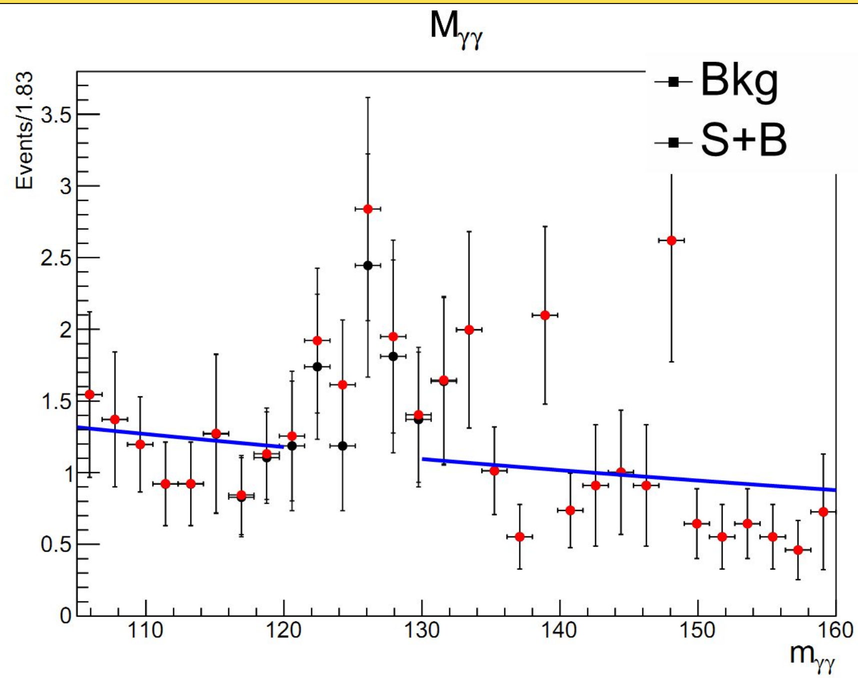




- significance VS BDT cut value
- tight [0.95,1],0.565σ
- loose [0.81,0.95],0.326σ
- combined 0.652σ

fit sideband





- left is VBF tight ,right is VBF loose
- fit sideband with exponential
- get the bkg events number
- calculate bkg number in [122.5,127.5]GeV
- red point is ggF+VBF+bkg,black point is bkg
- in loose category ,bkg has a peak near 125GeV
- BDT is correlated to $m_{\gamma\gamma}$?

event number after BDT cut

signal region : $m_{\gamma\gamma}$ [122.5,127.5]GeV

	VBF tight Category	VBF loose category	all VBF category
ggF	0.069	0.359	0.428
VBF	0.296	0.629	0.925
bkg	0.122	3.152	3.274
significance	0.565	0.326	0.652

$$\sigma_{VBF} = \sqrt{2 \times ((N_{VBF} + N_{ggF} + N_{Background}) \times \ln(1 + \frac{N_{VBF}}{N_{ggF} + N_{Background}}) - N_{VBF})}$$

- when s+b is not large enough ,this formular may not work
- toy throwing is ongoing

review 8TeV result

8TeV
20.3/fb

Events	Cut-based loose	Cut-based tight	MVA loose	MVA tight
VBF signal	2.6	3.5	3.96	3.22
ggF	1.35	1.05	2.59	0.72
Fitted background	34.8	10.6	40.1	6.7
VBF purity	65.6	76.6	60.0	81.6
VBF significance	0.43	0.98	0.60	1.11
Combined significance	1.07		1.26	

13TeV
4/fb

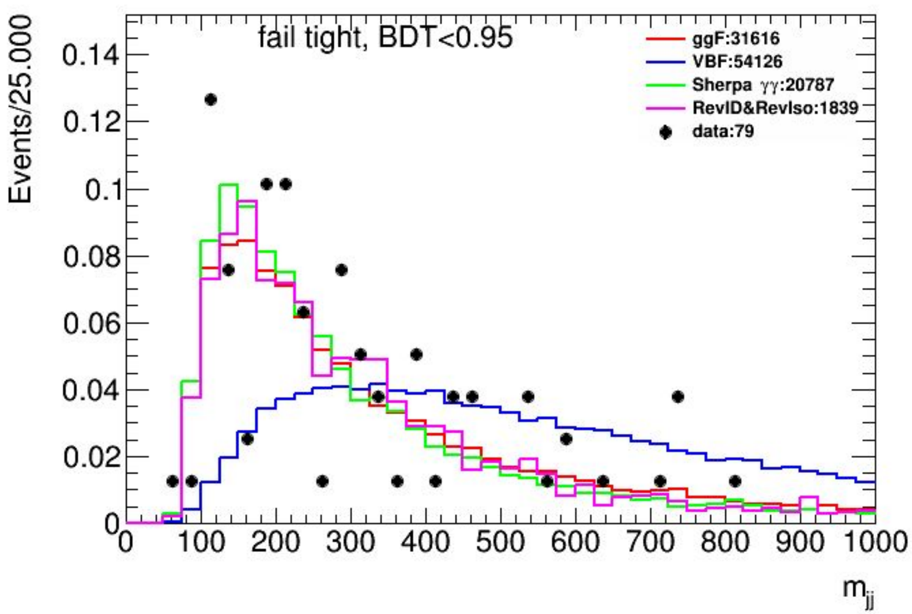
	MVA loose	**MVA tight**
VBF signal	0.629	0.296
ggF	0.359	0.069
fitted background	33.035	1.343
VBF purity	0.647	0.814
VBF significance	0.326	0.565
Combined significance	0.652	

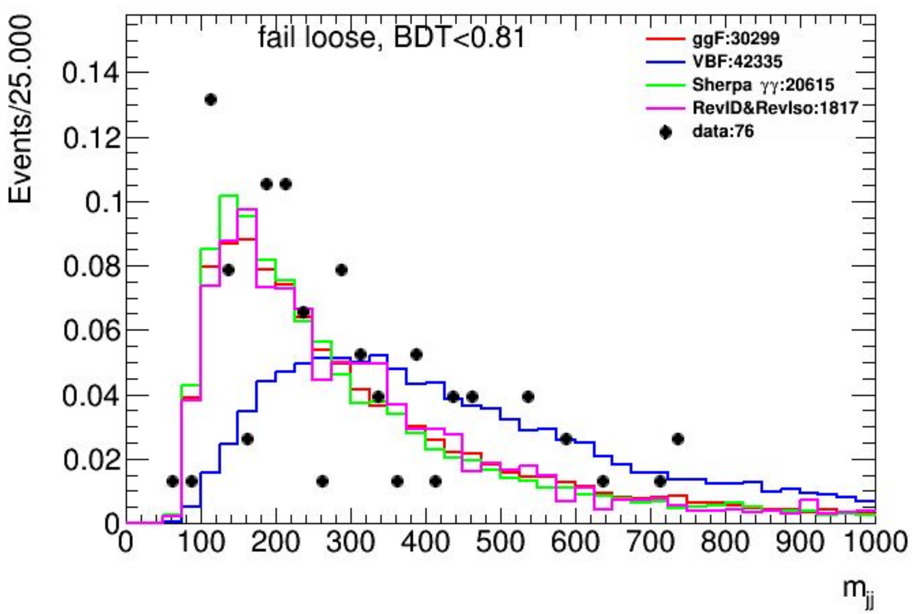
13TeV
10/fb

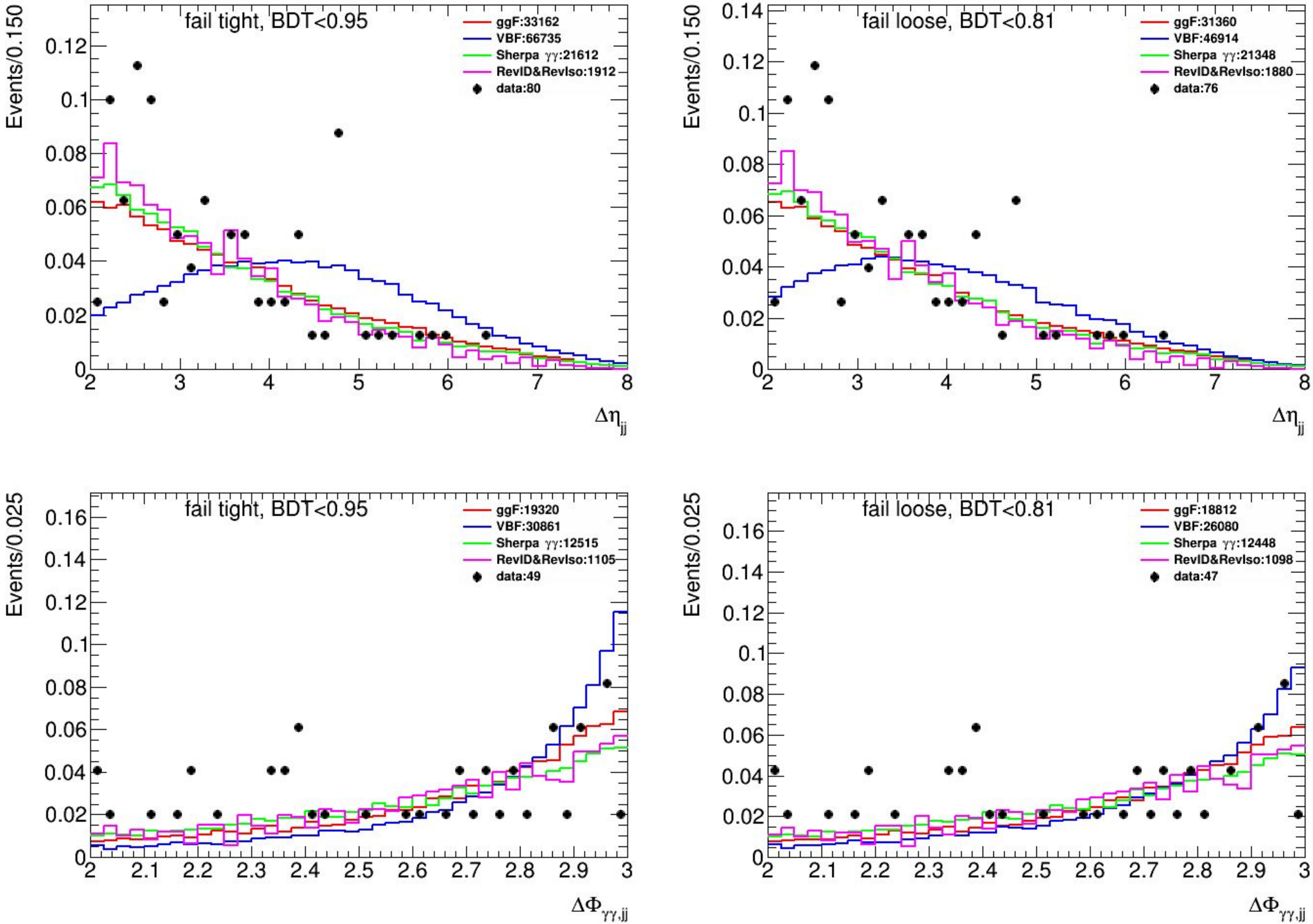
VBF signal	1.572	0.740
ggF	0.898	0.173
fitted background	82.588	3.380
VBF purity	0.647	0.814
VBF significance	0.516	0.900
Combined significance	1.030	

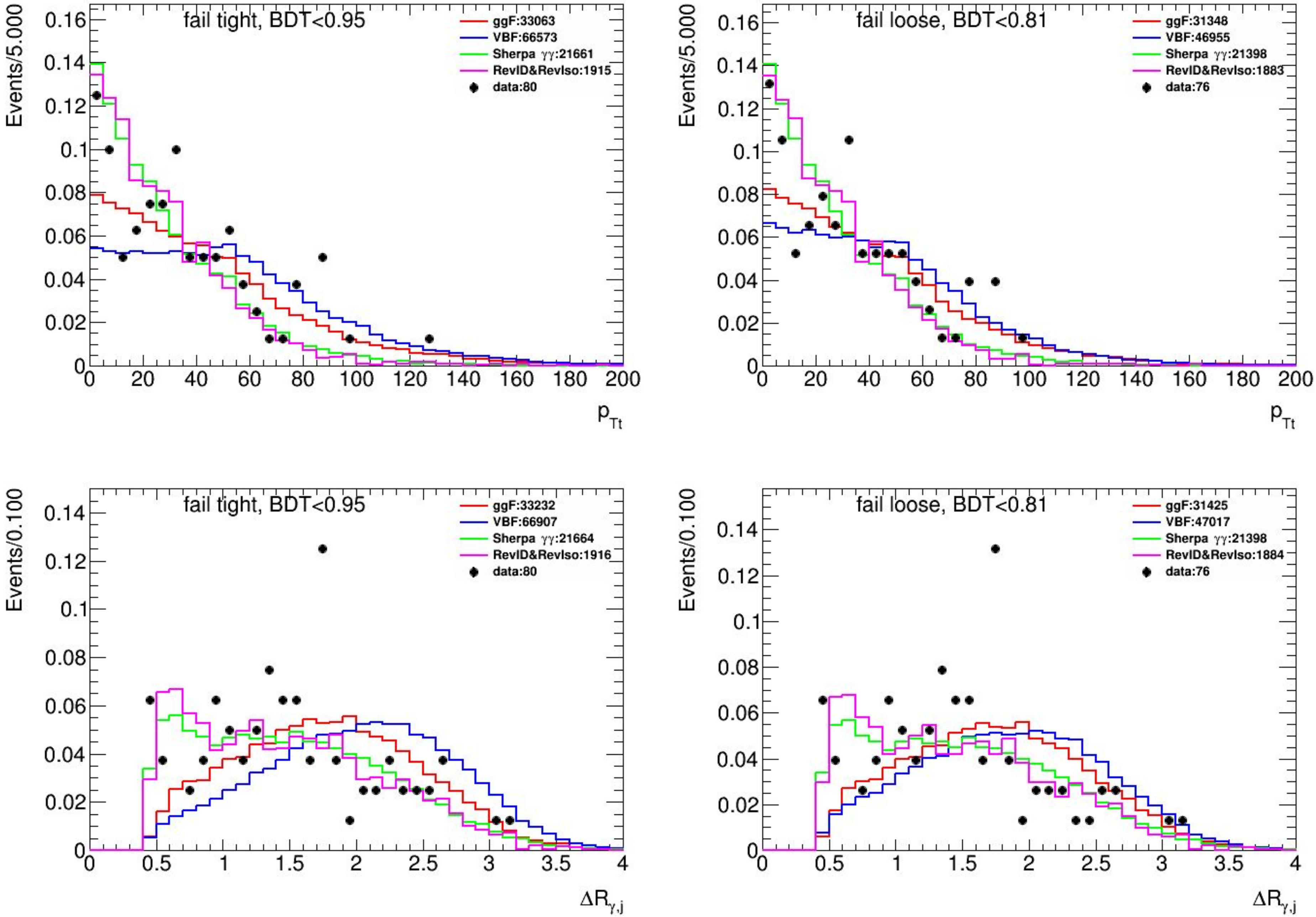
variable modeling validation

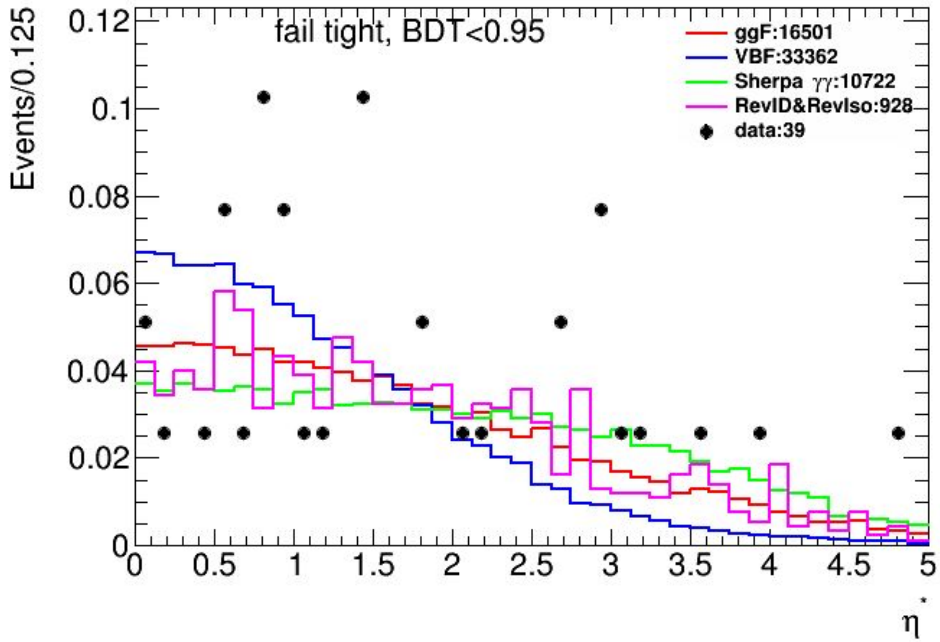
- make sure our background variable modeling (Sherpa + RevIso&RevID) is consistent with real background
- plot variables of samples failing BDT tight/loose selection
- 85 pb^{-1} data
  - 2jets : 179 events
  - VBF preselection : 80 events
  - BDT : no evevts in tight ,4 events in loose

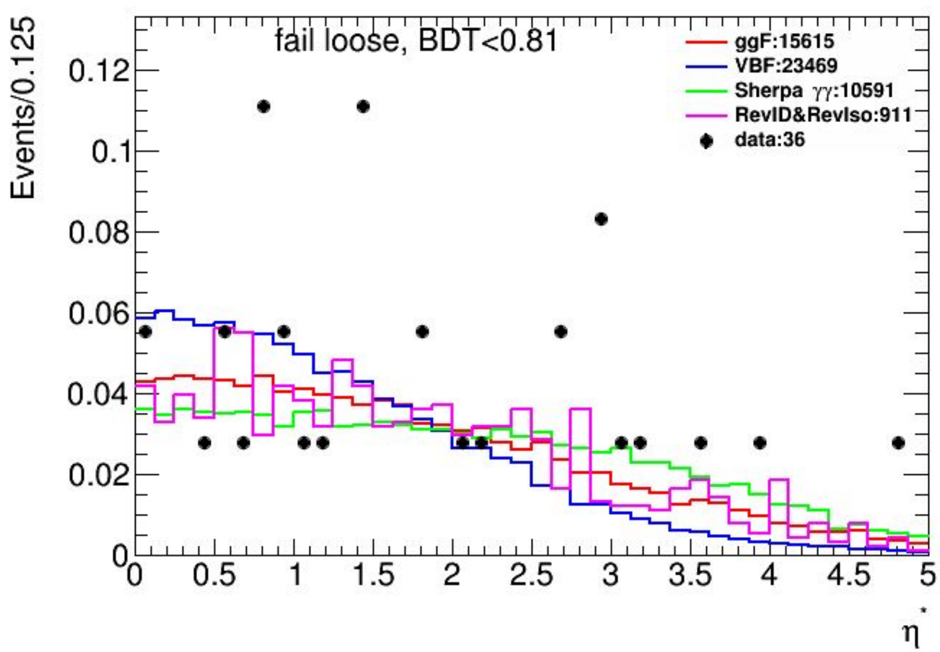





fail tight, BDT<0.95
ggF:33162
VBF:66735
Sherpa γγ:21612
RevID&RevIso:1912
data:80
Events/0.150
Δη_jj
fail loose, BDT<0.81
ggF:31360
VBF:46914
Sherpa γγ:21348
RevID&RevIso:1880
data:76
Events/0.150
Δη_jj
fail tight, BDT<0.95
ggF:19320
VBF:30861
Sherpa γγ:12515
RevID&RevIso:1105
data:49
Events/0.025
ΔΦ_γγ,jj
fail loose, BDT<0.81
ggF:18812
VBF:26080
Sherpa γγ:12448
RevID&RevIso:1098
data:47
Events/0.025
ΔΦ_γγ,jj


fail tight, BDT<0.95
ggF:33063
VBF:66573
Sherpa γγ:21661
RevID&RevIso:1915
data:80
Events/5.000
p_Tt
fail loose, BDT<0.81
ggF:31348
VBF:46955
Sherpa γγ:21398
RevID&RevIso:1883
data:76
Events/5.000
p_Tt
fail tight, BDT<0.95
ggF:33232
VBF:66907
Sherpa γγ:21664
RevID&RevIso:1916
data:80
Events/0.100
ΔR_γ,j
fail loose, BDT<0.81
ggF:31425
VBF:47017
Sherpa γγ:21398
RevID&RevIso:1884
data:76
Events/0.100
ΔR_γ,j


fail tight, BDT<0.95
ggF:16501
VBF:33362
Sherpa γγ:10722
RevID&RevIso:928
data:39
Events/0.125
η*


fail loose, BDT<0.81
ggF:15615
VBF:23469
Sherpa γγ:10591
RevID&RevIso:911
data:36
Events/0.125
η*

other issues

- I find a very stupid mistake......

$$\sigma_{VBF} = \sqrt{2 \times (N_{VBF} + N_{ggF} + N_{Background}) \times ln\left(1 + \frac{N_{VBF}}{N_{ggF} + N_{Background}}\right) - N_{VBF}}$$

- this formula is copied from ATL-COM-PHYS-2013-076
- but it is wrong......the right one should be

$$\sigma_{VBF} = \sqrt{2 \times ((N_{VBF} + N_{ggF} + N_{Background}) \times \ln(1 + \frac{N_{VBF}}{N_{ggF} + N_{Background}}) - N_{VBF})}$$

- which is consistent with Jin's code
- so the cut-base number counting result showed last week should be wrong ,the correct one is in next page

significance ---number counting

- signal region:$m_{\gamma\gamma}$ [121,131]GeV

	tight	weighted	loose	weighted	all	weighted
ggF	3990	1.022	10804	2.767	14794	3.789
VBF	45411	1.937	66780	2.849	112194	4.786
Sherpa	141	12.98	667	61.40	808	74.37
Rev	14	1.208	55	4.746	69	5.954

$$\sigma_{VBF} = \sqrt{2\times((N_{VBF}+N_{ggF}+N_{Background})\times\ln(1+\frac{N_{VBF}}{N_{ggF}+N_{Background}})-N_{VBF})}$$

	tight	loose	all
sigma	0.486648	0.340871	0.5945

compatible with throwing toys!

to do list

- MVA configuration optimization
- 25ns data
  - another ~85pb^{-1}
  - 276262-276954

formula derivation--copy from Cowan

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

Distribution of q_0 in large-sample limit

Assuming approximations valid in the large sample (asymptotic) limit, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case $\mu' = 0$ is a "half chi-square" distribution:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ.

formula derivation

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

Cumulative distribution of q_0, significance

From the pdf, the cumulative distribution of q_0 is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case $\mu' = 0$ is

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right)$$

The p-value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0)$$

Therefore the discovery significance Z is simply

$$\boxed{Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}}$$

formula derivation

$s/\sqrt{b}$ for expected discovery significance

For large $s + b$, $n \to x \sim$ Gaussian(μ,σ) , $\mu = s + b$, $\sigma = \sqrt{(s + b)}$.

For observed value x_{obs}, p-value of $s = 0$ is Prob($x > x_{\text{obs}} \mid s = 0$),:

$$p_0 = 1 - \Phi\left(\frac{x_{\text{obs}} - b}{\sqrt{b}}\right)$$

Significance for rejecting $s = 0$ is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\text{obs}} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate s is

$$\text{median}[Z_0|s + b] = \frac{s}{\sqrt{b}}$$

Better approximation for significance

Poisson likelihood for parameter s is

$$L(s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

For now
no nuisance
params.

To test for discovery use profile likelihood ratio:

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{s} \geq 0\,, \\ 0 & \hat{s} < 0\,. \end{cases} \qquad \lambda(s) = \frac{L(s, \hat{\hat{\theta}}(s))}{L(\hat{s}, \hat{\theta})}$$

So the likelihood ratio statistic for testing $s = 0$ is

$$q_0 = -2\ln\frac{L(0)}{L(\hat{s})} = 2\left(n\ln\frac{n}{b} + b - n\right) \quad \text{for } n > b,\ 0 \text{ otherwise}$$

formula derivation

Approximate Poisson significance (continued)

For sufficiently large $s + b$, (use Wilks' theorem),

$$Z = \sqrt{2\left(n \ln \frac{n}{b} + b - n\right)} \quad \text{for } n > b \text{ and } Z = 0 \text{ otherwise.}$$

To find median[$Z|s$], let $n \rightarrow s + b$ (i.e., the Asimov data set):

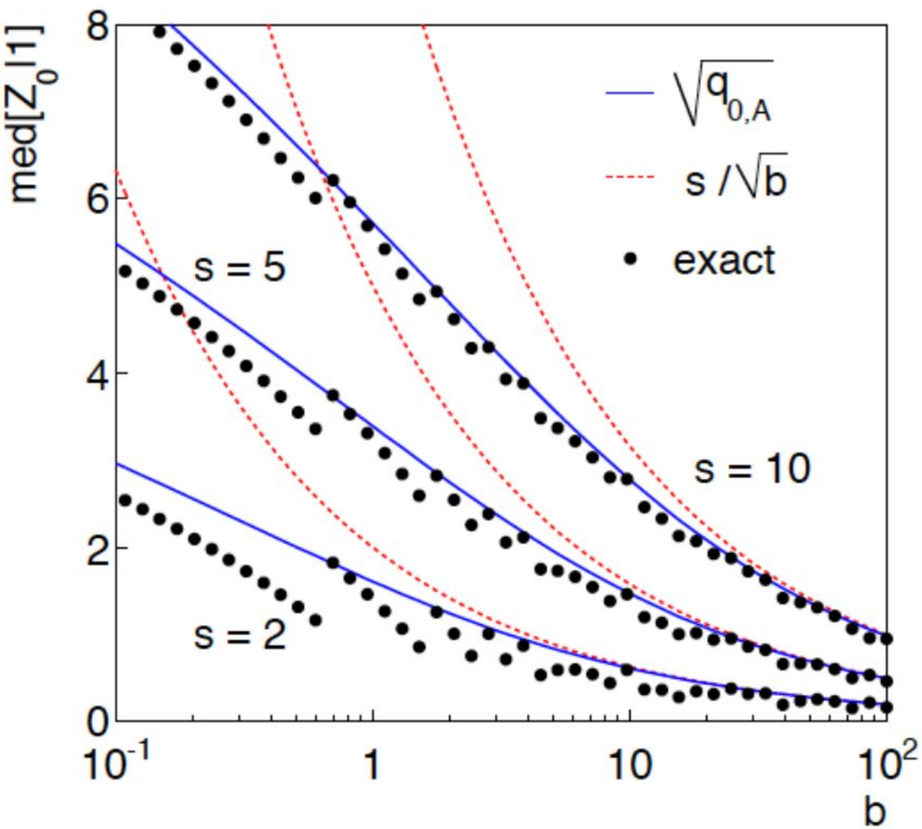
$$Z_{\mathrm{A}} = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)}$$

This reduces to $s/\sqrt{b}$ for s << b.

comparison

$n \sim$ Poisson(s+b), median significance, assuming s, of the hypothesis s = 0

CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727



"Exact" values from MC, jumps due to discrete data.

Asimov $\sqrt{q_{0,\mathrm{A}}}$ good approx. for broad range of s, b.

$s/\sqrt{b}$ only good for $s \ll b$.