QCD for Higgs invisible analysis

Considerations for proper normalisation, and application to signal-selection BDT

A.-M. Magnan
Imperial College London

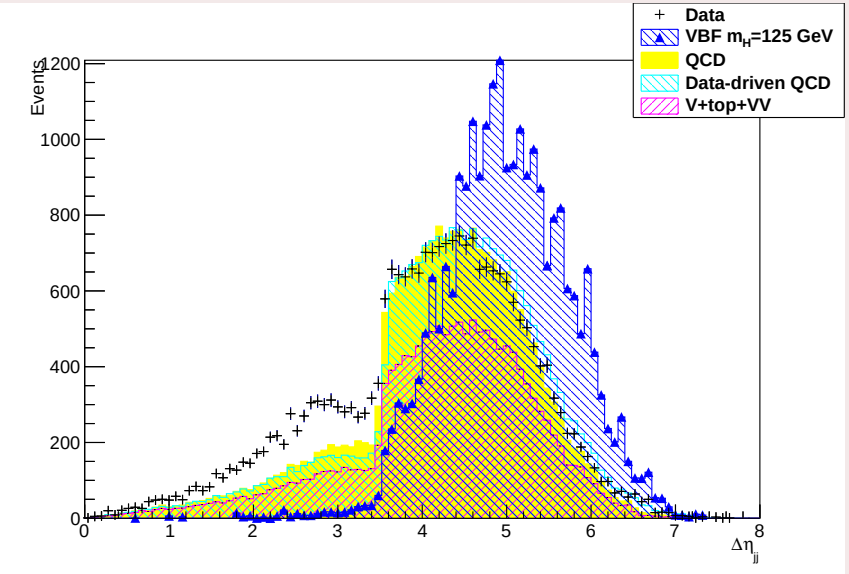
06/05/2014, IC $H \rightarrow \tau\tau$,inv analysis meeting

Steps to finding a proper preselection...

- Preselection: dijet pair with jets pT>50 GeV, eta < 4.7.
- Plotting standard variables $\Rightarrow$ clear problem in $\Delta\eta_{jj}$ variable $\Rightarrow$ add $\Delta\eta_{jj} > 3.6$ to the preselection stage.
- Given available QCD-VBF samples: select METsignificance>3 to enhance real MET contribution $\Rightarrow$ OK: about 90% efficient on the signal.
- Looking at $\Delta\phi_{jj}$: it is clear that normalising QCD to region $\Delta\phi_{jj} > 2.6$ will end-up in disagreement with QCD-VBF samples !!
- Looking at minimum $\Delta\phi$(jets,MET): clear that signal is > 1.5, whereas QCD has two populations. Select min$\Delta\phi$(jets,MET)> 1.5 $\Rightarrow$ in addition to MET significance: about 75% efficient on the signal.
- Sample composition in the end, in events: Data = 17271, EWKbkg = 11816.1, QCD=data-ewk= 5454.9, signal = 912.6 (out of 1211.35).
- Normalisation factor: 1.44.
- Clearly: need to find a less unbiased way to normalise the QCD !!
- Other try: data-driven QCD sample by reverting cut on MET significance (slice 2-2.6).

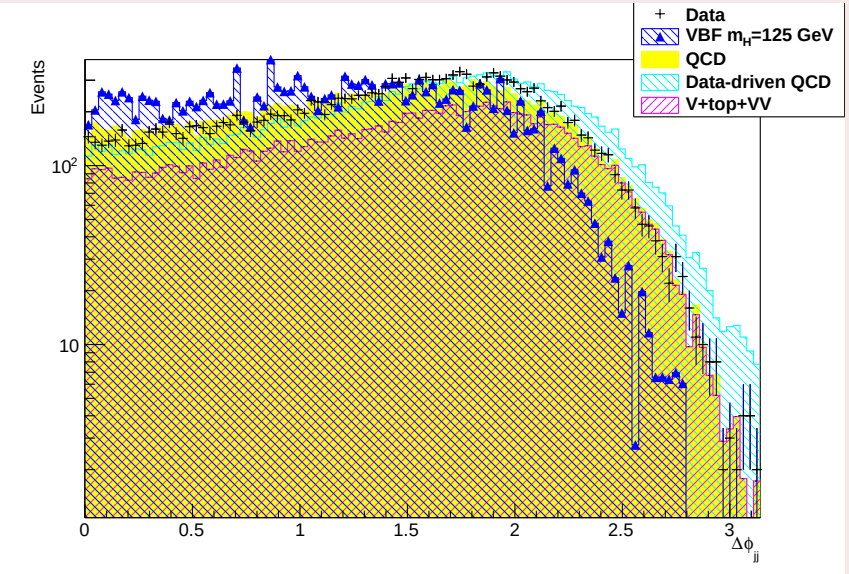
Applying all cuts except variable plotted

$\Delta\eta_{jj}$



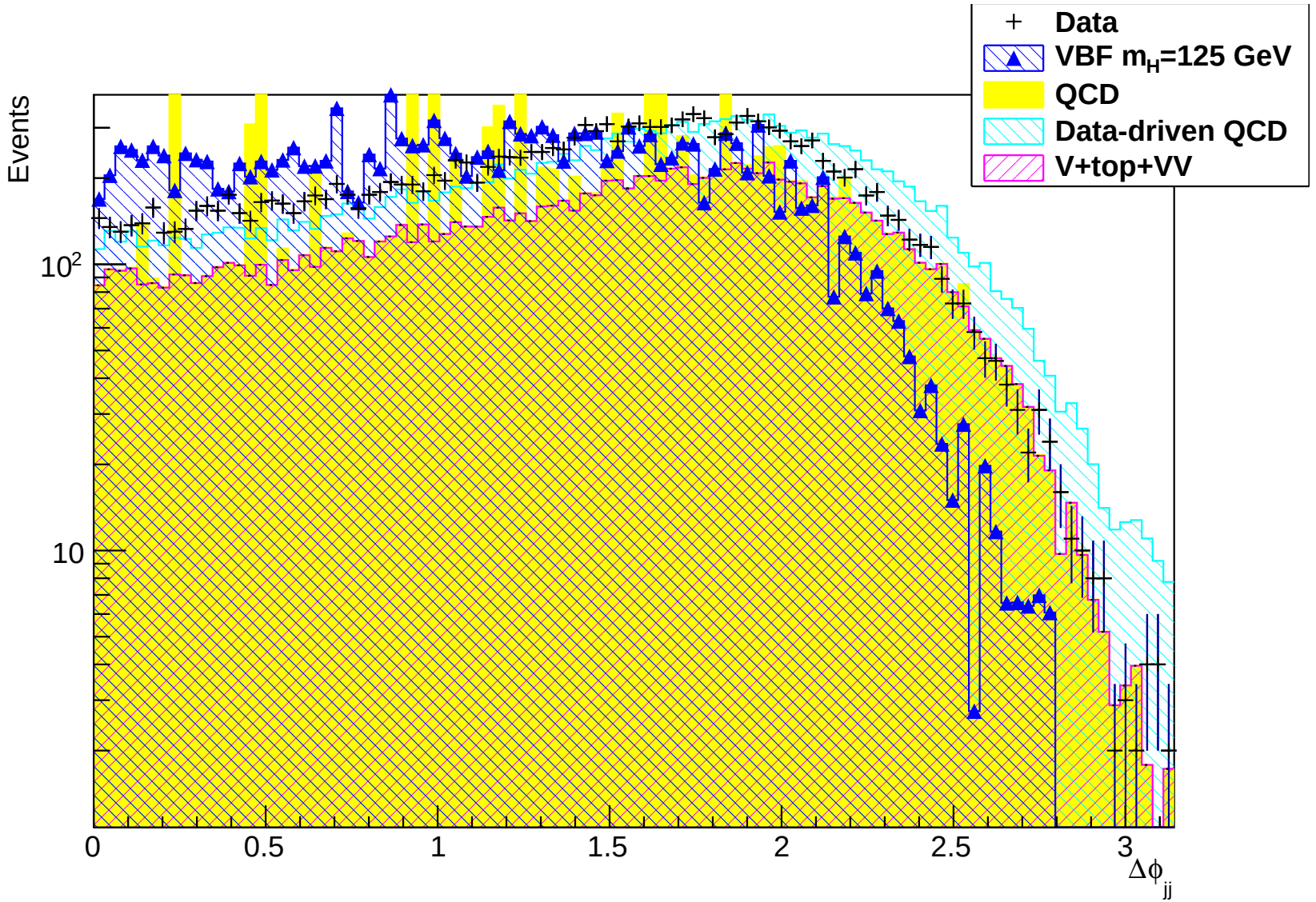
- QCD MC is normalised to region about 3.6.
- step below 3.6: trigger effect, but much less MC than data.
- signal already cut by trigger: marginal effect.
- Data-driven QCD looks fine too here.

$\Delta\phi_{jj}$



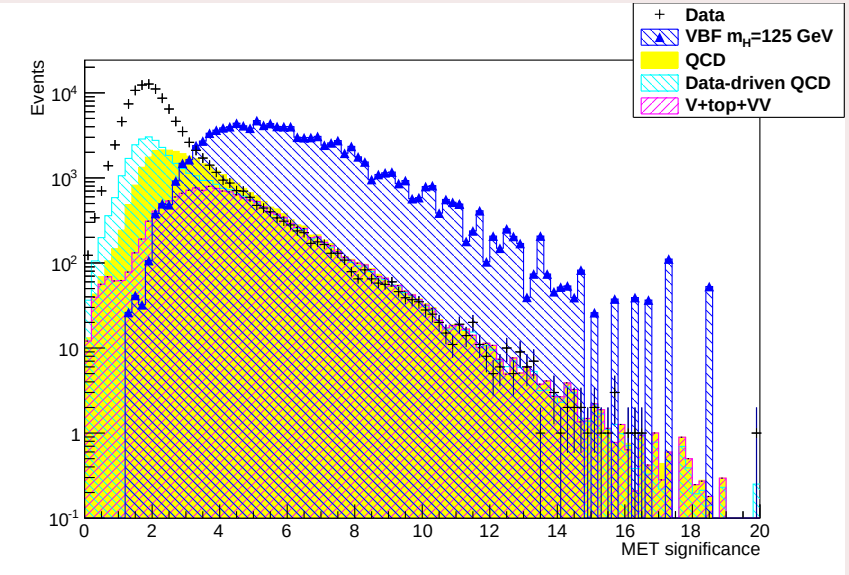
- Excellent agreement: high $\Delta\phi_{jj}$ region completely cut by min$\Delta\phi$(jets,MET)$>$ 1.5 selection.
- Data-driven QCD not so good.
- Still some discrimination with signal for further selection.

Note on inclusive QCD samples



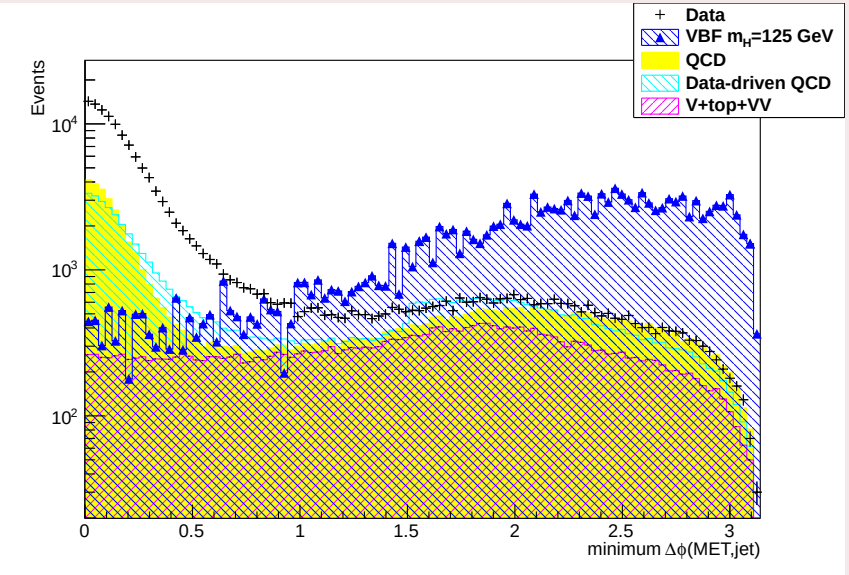
Applying all cuts except variable plotted

METsignificance



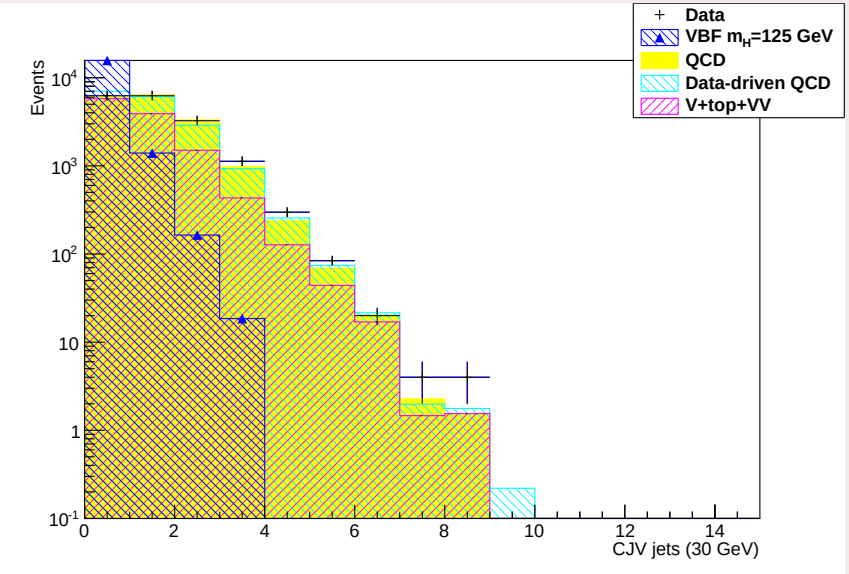
- Data in good agreement with no-QCD for *METsig* > 5.0.
- Data-driven QCD: Data-BKG for $2 <$ *METsig* < 2.6 normalised to Data-BKG for *METsig* > 3.0, factor 0.22.

min$\Delta\phi$(jets,MET)



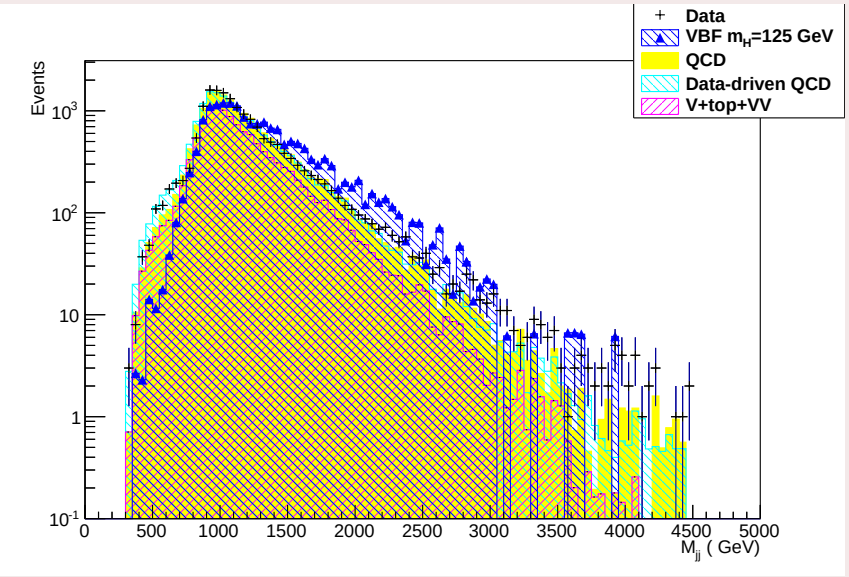
- Good agreement above 1.5, also with data-driven QCD.
- Below: data has a plateau when QCD-VBF decreases $\Rightarrow$ can expect that decreasing the cut would increase data-MC disagreement....

CJV jets $p_T > 30$ GeV



- Very good description of the number of additional jets.

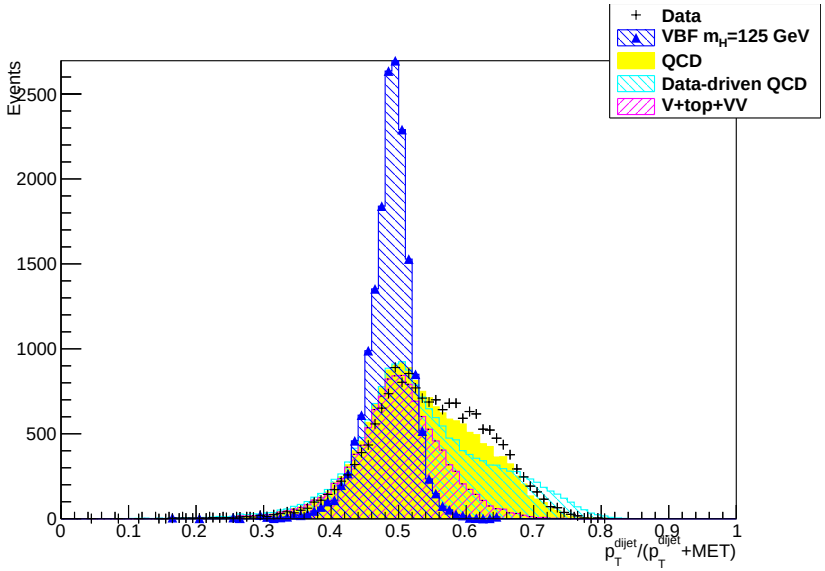
M_{jj}



- Suspiciously good description of Mjj too !
- If we trusted other backgrounds: things would be fine, but.....

Considerations for the absolute normalisation

- Ideas for better normalisation:
  - Work out proper NLO/LO k-factors, using fastNLO QCD dijet calculations from Klaus Rabbertz. Have instructions available, have not tried yet...
  - Use variable like dijet pT fraction which seems to have region with QCD-only ... Bit clumsy !
  - Use a template fit: need to trust other backgrounds first, and be in signal-free region: maybe using CJV jet cut ?

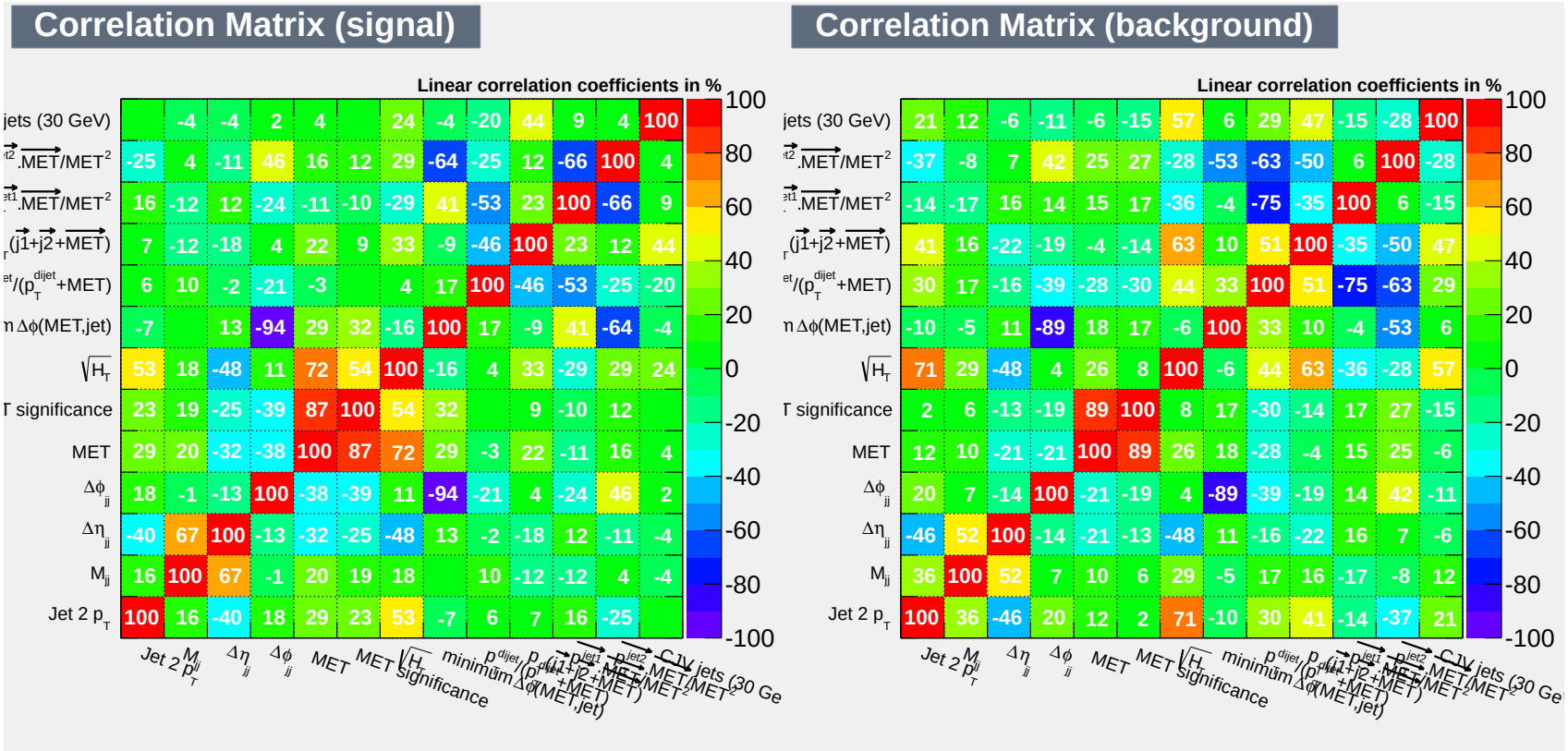


Conclusion on QCD

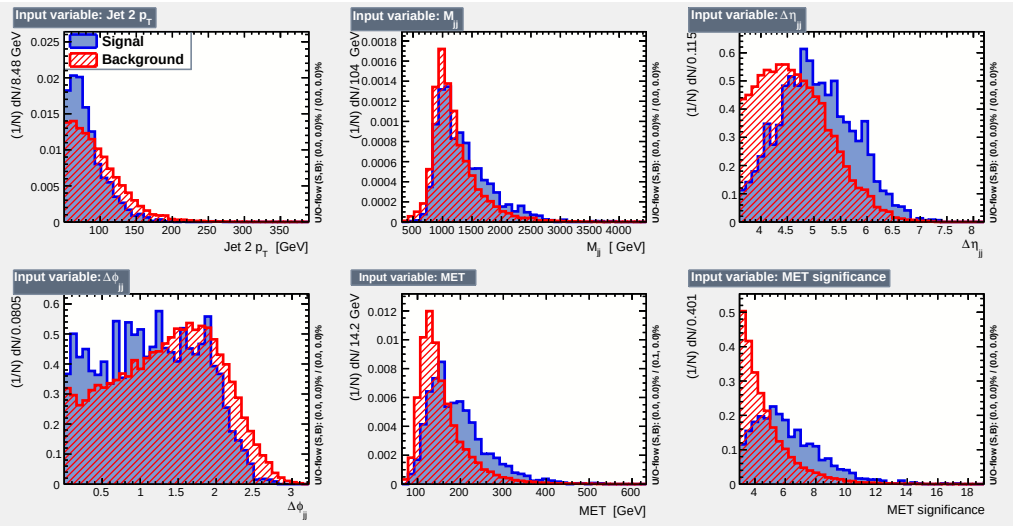
- Really need unbiased way to normalise QCD samples.
- Investigate also data-driven technique based on hypothesis that after this preselection, QCD is coming from pure combinatorics effects: use data-driven sample made with 1-3rd or 2nd-3rd jets for which 1-2nd fails selection to have orthogonal sample ?
- Need to start checking other backgrounds (W,Z) in control regions with leptons, using this preselection:
  - Dijet pair with $p_T > 50$GeV, $|\eta| < 4.7$, $\Delta\eta_{jj} > 3.6$
  - *METsig* > 3.0, min$\Delta\phi$(jets,MET)> 1.5.
- Next slides: started checking potential performance of signal-selection MVA: can we still do better than cut-based analysis with this pre-selection and already 25% loss of signal ?

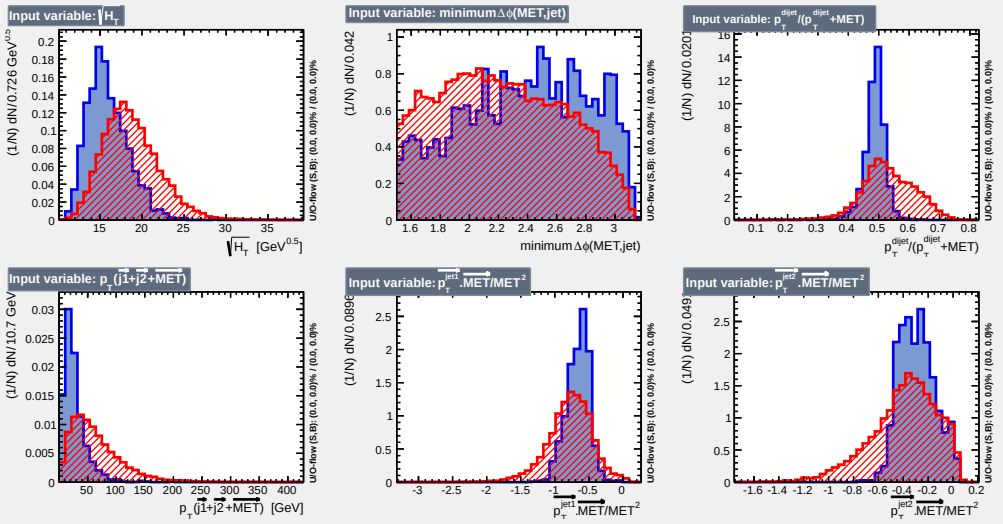
Setting up the signal BDT

- Preselection: background "efficiency" = 10%, signal eff = 75%.
- Note: bkg eff is not real eff (QCD samples are not "complete")
- Select most relevant variables, look at correlations.

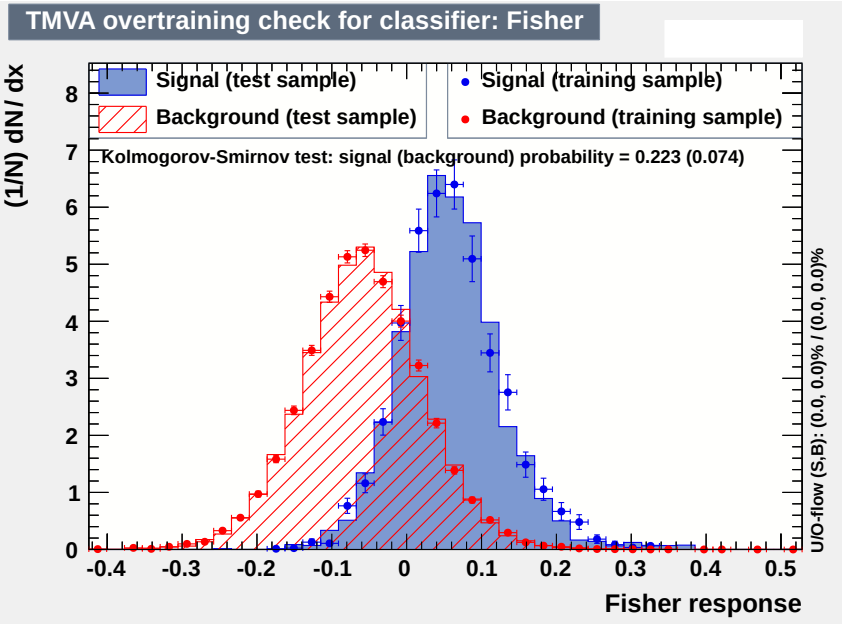


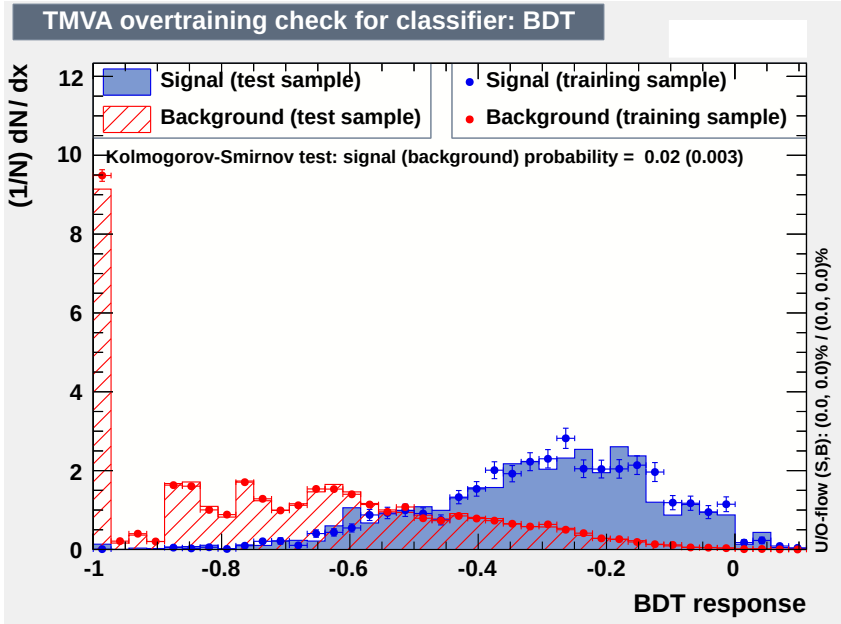
Input variables



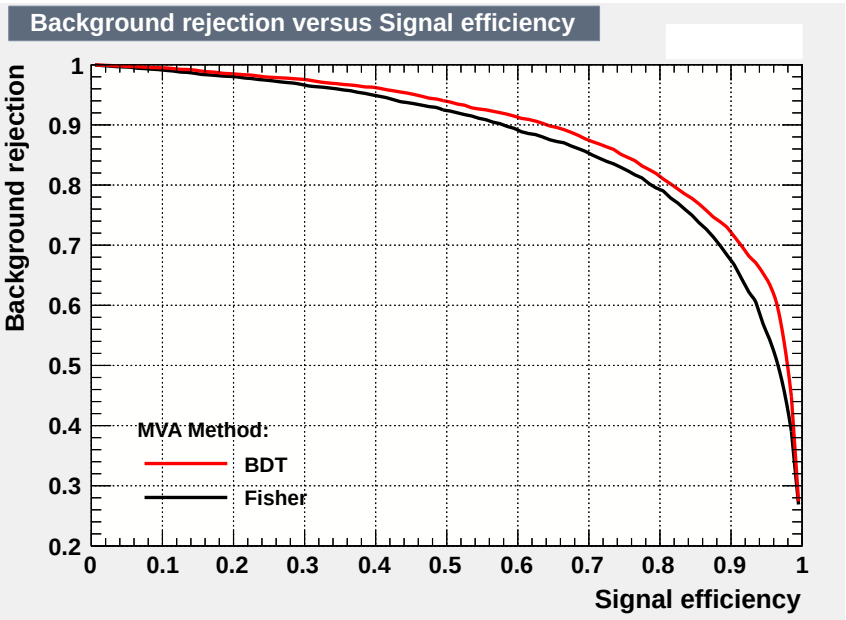


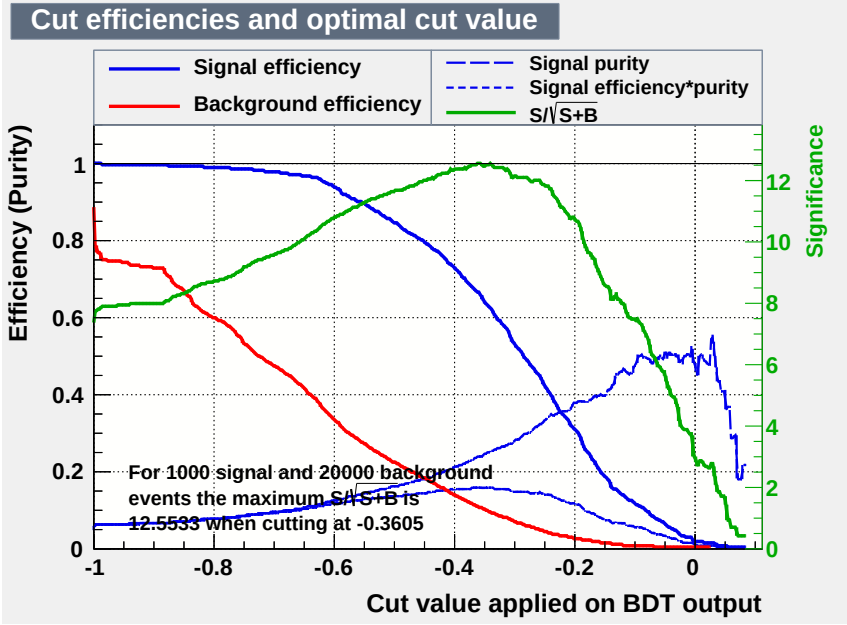
Outputs


TMVA overtraining check for classifier: Fisher
Signal (test sample)
Signal (training sample)
Background (test sample)
Background (training sample)
Kolmogorov-Smirnov test: signal (background) probability = 0.223 (0.074)
(1/N) dN / dx
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%
Fisher response


TMVA overtraining check for classifier: BDT
Signal (test sample)
Signal (training sample)
Background (test sample)
Background (training sample)
Kolmogorov-Smirnov test: signal (background) probability = 0.02 (0.003)
(1/N) dN / dx
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%
BDT response

Efficiency curves

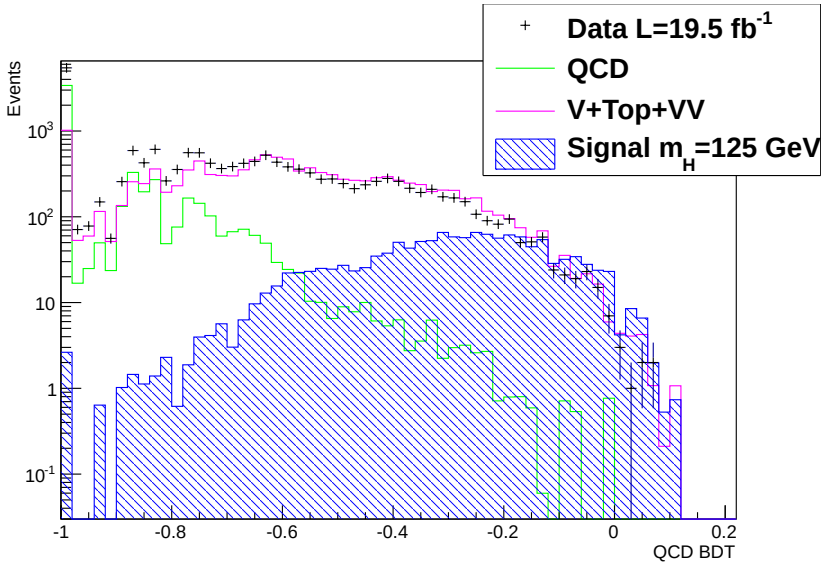




- Best working point: something like 65% signal efficiency for 90% background rejection, $BDT > -0.35$.
- $\Rightarrow$ expect 593 signal events (mH=125 GeV), and 1730 background events.

- Integrals between BDT = -0.36 and 1:

Sample	nsel	ntot	ε
Signal	880.6	1314.1	0.670
Data	1537	17271	0.089
QCD	28.5	5463	0.005
EWK	1828	11816	0.155



- Features in the QCD curve to be understood...
- Everything to be optimised...

Conclusions

- With cut-based analysis: expected 210 $\pm$ 30(stat+syst) signal events, and observed 390 data events.
- With BDT, 210 signal events = 23% signal efficiency compared to preselection applied $\Rightarrow$ expect 0.982 background rejection: 311 events.
- With BDT, 390 background events = 0.977 background rejection $\Rightarrow$ expect 28% signal efficiency...
- So out-of-the-box cutting on the BDT, expect about 20% improvement keeping same working point as cut-based analysis...