QCD Modelling

A. Magnan, P. Dunne, J. Pela

Imperial College London

2014-05-12

Today's presentation

Topics

- QCD VBF samples definition and motivation.
- Definition of a preselection to select real MET QCD.
- Analysis strategy.

QCD VBF sample definition and motivation

This initial idea was to make a QCD sample with real MET and jets with VBF characteristics.

Motivations and limitations

- The ability to actually make QCD samples with enough statistics to compare with 2012 dataset ($\sim 40\ fb^{-1}$).
- Samples with manageable size ($\sim 10\ TB$).
- CPU capability able to be produced with Imperial resourced in under 1 month.
- RECO filters are prohibitively CPU expensive to be usable. Generator filter are the only feasible option.

Generator level filters

Generator MET Filter:

- $MET_{Generator} = 40.0\ GeV$ (here the MET is the vectorial sum of all neutrinos $p_{\perp}$)

Generator Jets Filter:

- Jet selection:
  - $Jet(p_{\perp}) > 20\ GeV$
  - $-5.0 < Jet(\eta) < 5.0$
- Dijet selection:
  - $3.2 < Dijet(\Delta\eta) < 10.0$
  - $700 < Dijet(m_{jj}) < 50000$

Filter efficiencies

	Gen. Ev	Pass MET	Pass Dijet	MET Filter Eff	Dijet Filter Eff	Efficiency
QCD-Pt-50to80-pythia6	1000000	127	3	0,00013	0,024	0,000003
QCD-Pt-80to120-pythia6	1000000	1172	41	0,00117	0,035	0,000041
QCD-Pt-120to170-pythia6	1000000	4276	293	0,00428	0,069	0,000293
QCD-Pt-170to300-pythia6	1000000	9315	1012	0,00932	0,109	0,001012
QCD-Pt-300to470-pythia6	1000000	17956	2598	0,01796	0,145	0,002598
QCD-Pt-470to600-pythia6	1000000	23913	4187	0,02391	0,175	0,004187

Approach to pre-selection definition

We want to determine a pre-selection that selects mostly real met events in such a way that our QCD VBF + real MET sample together with all MC background samples can describe data (where backgrounds dominated).

Baseline selection

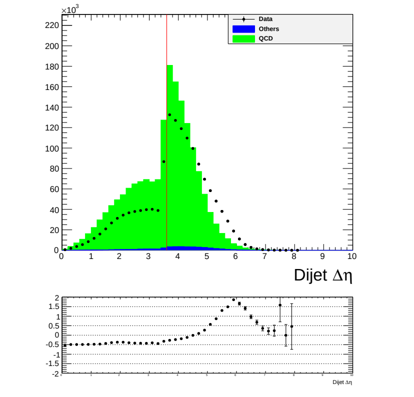
- Trigger bit selection (include weighting for MC to data matching)
- Lepton Veto
- Dijet selection (at least 2 jets with):
  - $p_{\perp} > 50\ GeV$
  - $|\eta| < 4.7$

Now we, add cut by cut methodically removing generator bias on distributions in order to reduced QCD fake MET content (not described by MC) to a manageable/negligible level.

$\Delta\eta$ cut

First we need to cut above the turn on the generator dijet $\Delta\eta$ cut in order to remove that bias.

Dijet $\Delta\eta$

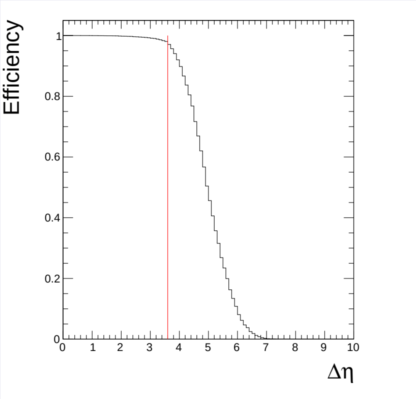


Methodology

- We can see a shape rise around $\Delta\eta \sim 3.5$ this is most likely due to the GenJet filter of $\Delta\eta > 3.2$
- We will cut at $\Delta\eta > 3.6$.
- We assume that we do not understand distribution on left of the cut so we normalise on the right side:
- QCD normalised on $\Delta\eta > 3.6$ with $factor = \frac{n_{Data} - n_{\text{other bkg}}}{n_{QCD}}$
- Distributions still does not agree in shapes.

$\Delta\eta$ cut - signal efficiency

Signal Eff. vs $\Delta\eta$

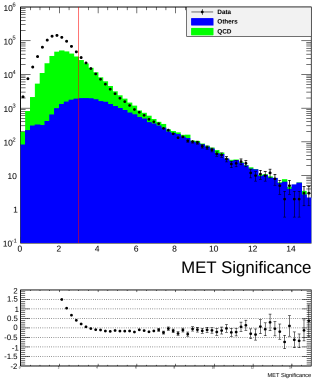


Efficiency

- We can now calculate how much is our signal efficiency for $m_{Higgs} = 125$ GeV and $BR(Inv) = 100\%$
- Signal Efficiency $\Delta\eta$ (3.6) $= \sim 0.97$
- Our signal is almost untouched by this cut.

MET Significance cut

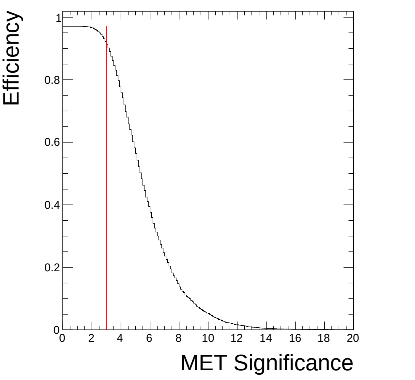
MET Significance cut



- We know that QCD Fake MET events will typically have low MET Significance
- We can slide or cut and normalisation window from high to low values and stop when QCD sample does not describe data well anymore due to the raising percentage of QCD fake MET events.
- A good value is $MET_{Significance} = 3.0$ (or 3 sigma significance)
- QCD normalised on $MET_{Significance} > 3.6$ with $factor = \frac{n_{Data} - n_{\text{other bkg}}}{n_{QCD}}$
- Above the cut distributions match reasonably.

MET Significance cut - signal efficiency

Signal Eff. vs MET Significance



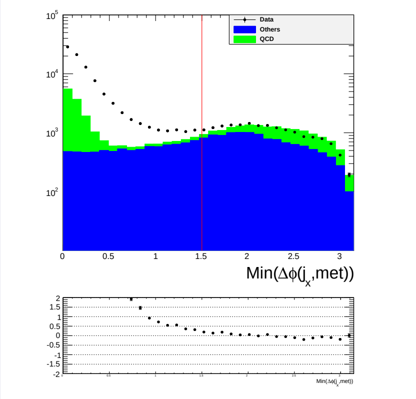
Taking into account the previous cut we can calculate:

- Again for signal of $m_{Higgs} = 125$ GeV and $BR(Inv) = 100\%$
- Signal Efficiency MET Significance (3.0) $= \sim 0.91$
- Our signal still remains largely untouched.

Imperial College London

$Min(\Delta\phi(jet_x, MET))$ cut

$Min(\Delta\phi(jet_x, MET))$

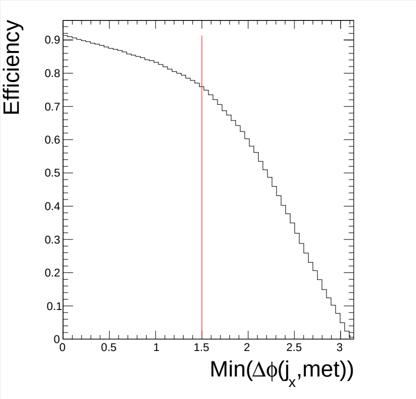


Methodology

- We also know that one of the main reasons for fake MET is the miss measurement of the energy of one of the leading jets.
- This typically results in MET being aligned with with the jets directions, which is reflected by the variable $Min(\Delta\phi(jet_x, MET))$
- Again o repeat the procedure to used for MET significance
- A good value is $Min(\Delta\phi(jet_x, MET)) = 1.5$
- QCD normalised on $Min(\Delta\phi(jet_x, MET)) > 1.5$ with $factor = \frac{n_{Data} - n_{\text{other bkg}}}{n_{QCD}} = 1.50$ (compatible with NLO correction)
- Distribution matches well above cut.

$Min(\Delta\phi(jet_x, MET))$ cut - signal efficiency

Signal Eff. vs MET Significance



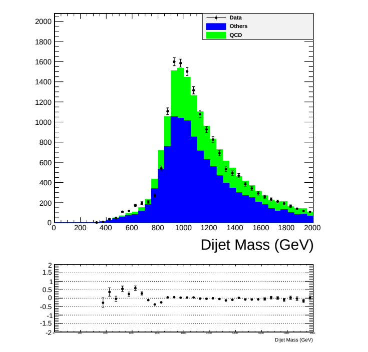
Taking into account the previous cuts we can calculate:

- Again for signal of $m_{Higgs} = 125$ GeV and $BR(Inv) = 100\%$
- Signal Efficiency $\Delta\phi$ (1.5) $= \sim 0.76$
- Our signal is still at a comfortable level.

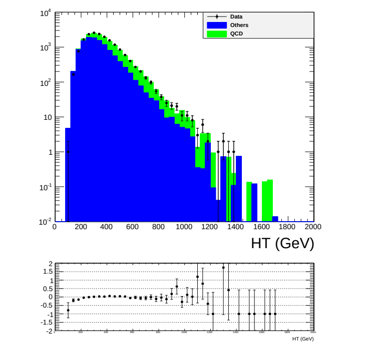
Looking at key variables

At this point we already apparently reduced the fake MET content in data to a level that distributions match to a reasonable level.

Dijet Mass



HT

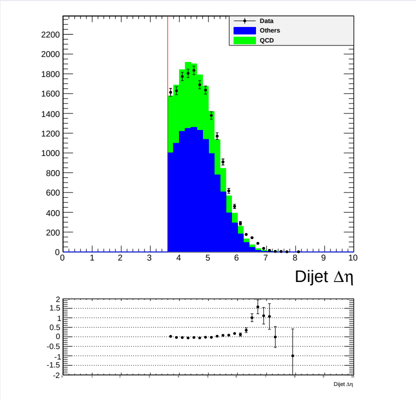


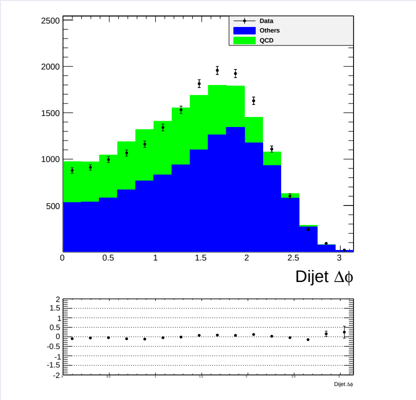
For this point including this slide all distributions are normalised on the total number of expected yield passing all the cuts $factor = \frac{n_{Data} - n_{\text{other bkg}}}{n_{QCD}}$.

This is not correct since we can have signal on this regions, but is a good first approximation.

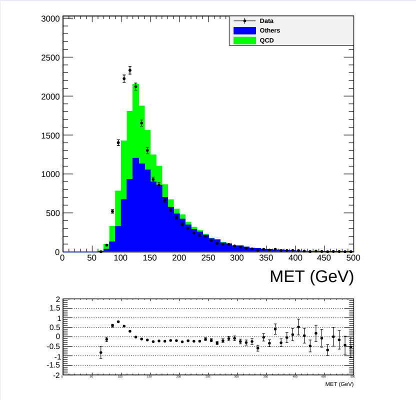
Dijet angles

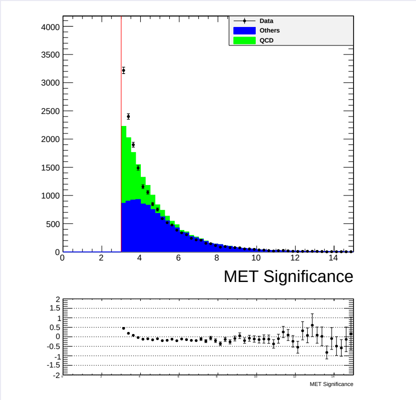
The dijet angles are reasonably described by the MC now and agreement if often of the order of $\sim 10\%$ or better. But there is still room for improvement.



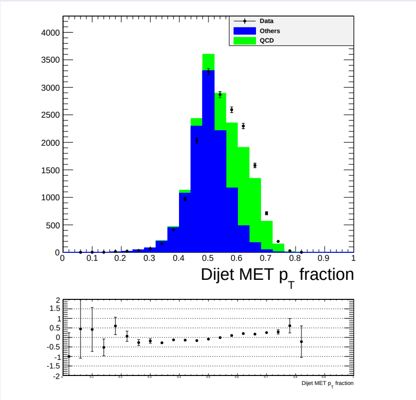


MET variables show some discrepancy but only at low values which are more likely to contain fake MET events. Further studies may allow to improve this.





Effects now described



- Now for the first time we can actually model the excess on data while compared with other backgrounds observed at high values of Dijet-MET $p_\perp$ fraction.
- This excess as expected comes from QCD and this variable is therefore discriminate against that type of backgrounds.

Analysis Strategy

QCD VBF samples

Now that we found a pre-selection that is a reasonable working point where our MC samples can describe data. We can use it as a baseline to the rest of the analysis.

BDT approach

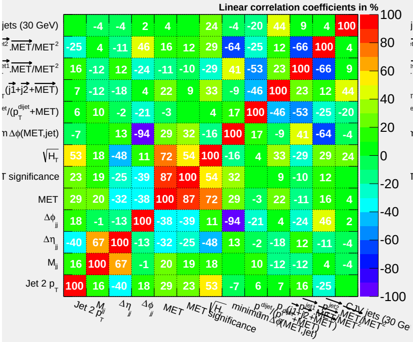
- We can drop the QCD exclusion BDT
- We can make an all background against signal BDT
  - Trained starting from this or a similar pre-selection
  - Use as a basis for training the QCD VBF sample
  - This would be a single BDT approach

Data driven methods still necessary to full understand all steps including pre-selection.

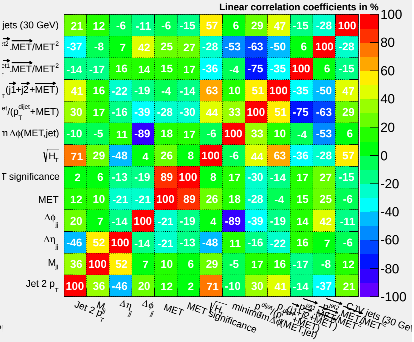
Setting up the signal BDT

- 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.

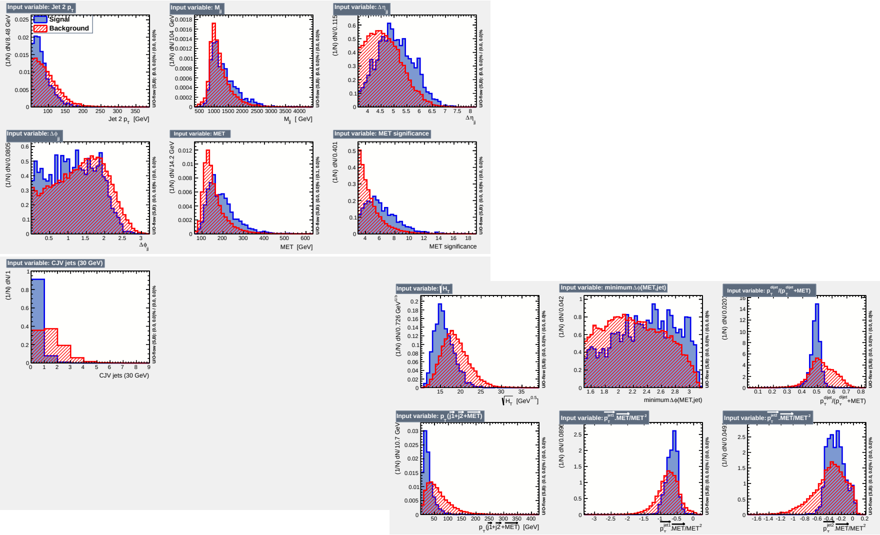
Correlation Matrix (signal)



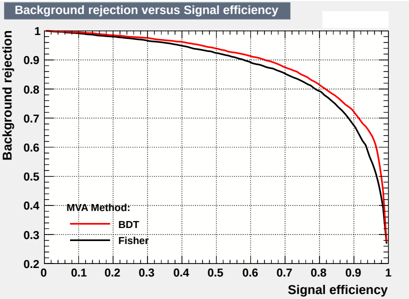
Correlation Matrix (background)

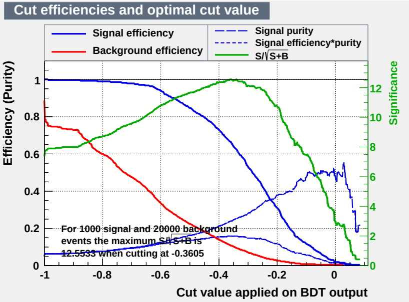


Input variables



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.

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...

Rough limits calculation (P. Dunne)

Assumptions:

- Same background expectation from cut based
- Increasse in signal yield of 20 events (5%)
- Same systematics

The limit on BR(Inv) goes from 65% to 55% so a gain of around 10%.

Summary and next steps

Summary:

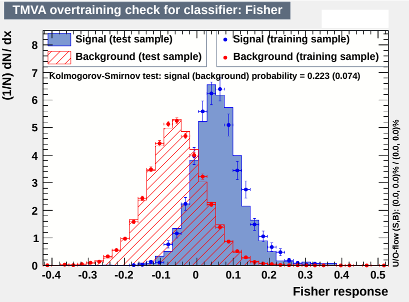
- A good working point for a pre-selection was found
- Reasonable agreement of key variables and data is observed
- Normalisation needs to be revisited to avoid signal contamination
- New possible structure for the analysis presented based on this findings
- With similar working point to current in cut based we can improvement of 10% in the limit

Next Steps:

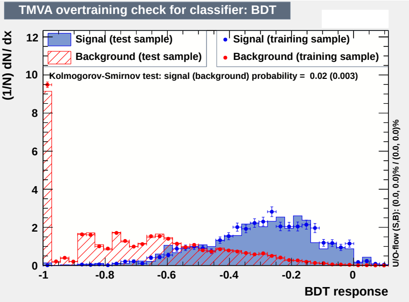
- Optimise pre-selection
- Optimise and study BDT approach
- Calculate yield/limit gains with parked data

Backup Slides

Outputs



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



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