ALL HADRONIC $t\bar{t}H(b\bar{b})$ TRIGGER STUDY AND ANALYSIS USING THE BDT METHOD
CERN Summer Student 2015 Report

Maria Giovanna Foti
Chapter 1

Introduction

The aim of this report is to describe the project on which I have worked as a CERN Summer Student 2015.
The project consists of two parts:

1. Study of the trigger efficiencies for all hadronic $t\bar{t}H(bb)$ analysis;
2. All hadronic $t\bar{t}H(bb)$ analysis using the BDT (and Fisher) methods.

All hadronic $t\bar{t}H(bb)$

Figure 1.1: A leading-order Feynman diagram for $t\bar{t}H$ production, illustrating the top-quark pair system decay channel considered here, and the $H \rightarrow b\bar{b}$ decay mode to which the analysis refers.

The analysis focuses on the search for a Higgs boson produced in association with a pair of top quarks ($t\bar{t}H$) conducted at the CMS experiment.
All hadronic signal events are characterised by:

- 3 jets coming from the top quark decay;
- 3 jets coming from the antitop quark decay;
- 2 more jets coming from the decay of the Higgs boson to two b-quarks.

Therefore, we look for events with:

- $\geq 8$ jets;
- $\geq 4$ b-jets;
- no leptons (and no missing energy).

The whole study is carried out on Monte Carlo 25ns $t\bar{t}H$, $t\bar{t}$ and $QCD$ samples.
Chapter 2

Study of the trigger efficiencies for all hadronic $ttH(b\bar{b})$

Trigger Paths

The first part of the project has focused on the study of the trigger efficiencies of four HLT trigger paths. In particular, the selected signal triggers are:

- **Trg 0**: HLT\_PFHT450\_SixJet40\_PFBTagCSV Selects events with $H_T > 450$ GeV, $\geq 6$ jets with $p_T > 40$ GeV and one b-jet CSVM ($csv > 0.814$).

- **Trg 2**: HLT\_PFHT400\_SixJet30\_BTagCSV50p5\_2PFBTagCSV Selects events with $H_T > 400$ GeV, $\geq 6$ jets with $p_T > 30$ GeV and two b-jets CSVM ($csv > 0.814$).

Whilst the selected reference triggers are:

- **Trg 1**: HLT\_PFHT450\_SixJet40 Selects events with $H_T > 450$ GeV, $\geq 6$ jets with $p_T > 40$ GeV.

- **Trg 7**: HLT\_PFHT350 Selects events with $H_T > 350$ GeV.

Preselection Requirements

Based on the topology of the signal events, the samples used in the analysis contain events with:

- a number of jets (nJets) $\geq 6$;
6 jets with $p_T > 30\ \text{GeV}$ and $|\eta| < 2.4$;

- $H_T > 400\ \text{GeV}$;

- a number of b-jets ($nBJets \geq 2$), tagged with the CSVM algorithm ($csv > 0.814$).

**Two-dimensional efficiency and fraction of retained signal**

As a first step, the two-dimensional projection of the trigger efficiency versus $(H_T; p_T^{j_5})$ is considered. By analysing this two-dimensional map the offline kinematic cuts can be inferred; moreover, by comparing different maps, information about the effect of different triggers (with and without b-tag) on different samples ($t\bar{t}H$, $t\bar{t}$, $QCD$) can be extracted.

As expected, and as it can be inferred by looking at Figure 2.1, Trg 1 (containing only kinematics) has higher efficiency than Trg 0 (same kinematics + b-tag) in the considered phase space.

Moreover, both, Trg 0 and Trg 1, have higher efficiency on the signal ($t\bar{t}H$), rather than on $t\bar{t}$ and $QCD$ (both sources of background in this analysis). Again, this is expected since these triggers are designed to perform best on the signal events.

Ideally, we would like to deal with a trigger efficiency equal to 1, but this would imply cutting too high in $H_T$ and jet $p_T$. Consequently, not only on the amount of background but also of the signal is reduced significantly.

In order to properly estimate the amount of signal we are able to retain by cutting above a certain threshold, we define the variable $N_{signal}$:

$$N_{signal}(H_{T,min}; p_{T,min}^{j_5}) = \int_{H_{T,min}}^{H_{T,max}} \int_{p_{T,min}^{j_5}}^{p_{T,max}^{j_5}} \varepsilon_{TRG}(H_T; p_T^{j_5}) \cdot f(H_T; p_T^{j_5}) \, dp_T^{j_5} \, dH_T$$

where $f(H_T; p_T^{j_5})$ is the signal distribution and $H_{T,min}$ and $p_{T,min}^{j_5}$ are the aforementioned thresholds.

After examining the two-dimensional efficiency plots and the $N_{signal}$ ones (Figures 2.1 and 2.2), initial kinematic cuts can be set. In particular, we have opted for:

$$H_T > 500\text{GeV} \quad \text{and} \quad p_T^{j_5} > 50\text{GeV}$$

with an efficiency of the signal trigger Trg 0 $\geq 0.73$ and a fraction of retained signal $> 0.440$ on the $QCD$ sample and trigger efficiency $\geq 0.83$ and $N_{signal} > 0.592$ on the $t\bar{t}H$ sample.
Figure 2.1: 2D trigger efficiency plots versus \((H_T; p_T^{j5})\) in the case of the \(t\bar{t}H\), \(t\bar{t}\) and \(QCD\) samples.
One dimensional projections

After setting the initial kinematic cuts, we draw the one-dimensional projections of the trigger efficiencies (for both Trg 0 and Trg 1), versus the variables on which the trigger operates: $H_T$, $p_T^{5}$ and $btagmax$, the maximum value the b-tag assumes in a given event.

In the one-dimensional projections (in Figures 2.3, 2.4 and 2.5) it can be seen even better than for the two-dimensional case that Trg 1 (working only on kinematics) has higher efficiency than Trg 0 (working on the same kinematics, with the addition of one b-tag) in our phase space and that both these triggers have higher efficiency on signal ($t\bar{t}H$), rather than on $t\bar{t}$ and $QCD$.

Moreover, as it should be, the efficiency of Trg 0 has a dependency on the $btagmax$, whilst Trg 1 is not affected by it.

The final signal trigger

In order to improve the analysis, Trg 2 is included. The reason for this is that Trg 2 performs better at low energies, whilst Trg 0 shows a better performance at higher energies. As a final result, the logic OR between the
(a) 1D projection of the trigger efficiency versus $H_T$

(b) 1D projection of the trigger efficiency versus $p_T^{j5}$

(c) 1D projection of the trigger efficiency versus $btagmax$

Figure 2.3: $t\bar{t}H$ sample, 1D trigger efficiency plots versus $H_T$, $p_T^{j5}$ and $btagmax$ for Trg 0 (with b-tag, in black) and Trg 1 (without b-tag, in red).
Figure 2.4: $t\bar{t}$ sample, 1D trigger efficiency plots versus $H_T$, $p_T^{j5}$ and $btagmax$ for Trg 0 (with b-tag, in black) and Trg 1 (without b-tag, in red).
Figure 2.5: QCD sample, 1D trigger efficiency plots versus $H_T$, $p_T^{j5}$ and $btagmax$ for Trg 0 (with b-tag, in black) and Trg 1 (without b-tag, in red).
two is expected to have higher efficiency than both the separate cases.

Therefore, new offline kinematic cuts need to be set and as a first step the two-dimensional projection of the trigger (Trg 0 || Trg 2) efficiency versus \((H_T; p_T^{j5})\) is plotted.

As expected (see Figure [2.6]), the logic OR between Trg 0 and Trg 2 has higher efficiency than the separate cases, and, as in the case of the two triggers already analysed, it has also higher efficiency on the signal \(t\bar{t}H\), rather than on \(t\bar{t}\) and \(QCD\).

In order to set the new kinematic cuts, the sensitivity \(S/\sqrt{QCD}\) is considered (see Figure [2.7]).

Therefore, given the plots in Figures [2.6] and [2.7] the kinematic cuts for the analysis, are:

\[
H_T > 450\text{GeV} \quad P_T^{j5} > 40\text{GeV}
\]

so that the trigger efficiency (of Trg 0 || Trg 2) is greater than 0.65 on QCD and greater than 0.86 on the signal \(t\bar{t}H\), without rejecting areas of the phase space in which there is the highest sensitivity.

One-dimensional projections and first hypothesis for a reference trigger

After setting the new kinematic cuts for the signal trigger, the one-dimensional projections of the efficiencies are drawn. With these, a first evaluation of the reference trigger to be used on data for the all hadronic \(t\bar{t}H\) analysis is performed. The chosen candidate reference trigger is:

\[
HLT_{\text{-PFHT350}}.
\]

In the one-dimensional projections in Figure [2.8] we can see that our signal trigger (Trg 0 || Trg 2) depends on \(H_T, p_T^{j5}, \text{btagmax}\) as expected, and also on \(nJ\text{ets}\), since it actually selects at least six jets. Instead, no dependency is shown neither on \(\eta^{j5}\), nor on \(n\text{Vtx}\), the number of vertices in the event. The latter means that the performance of the considered trigger should not be affected significantly by the pileup.

It is also very important to notice that, above the decided kinematic cuts, the reference trigger is unbiased, especially on the QCD sample (that constitutes the greater part of data).
Figure 2.6: 2D efficiency plots versus \((H_T; p_T^{j5})\) for \((\text{Trg 0} \parallel \text{Trg 2})\), in the case of the \(ttH\), \(t\bar{t}\) and \(QCD\) samples.
Figure 2.7: Signal sensitivity.
Figure 2.8: 1D trigger efficiency plots versus $H_T$, $p_T^{5}$, $btagmax$, $\eta^{5}$, $nJets$ and $nVtx$ for all the samples. The dark points refer to the "absolute" efficiencies, whilst the lighter colours of every series correspond to the efficiencies calculated with respect to the reference trigger.
Chapter 3

All hadronic $t\bar{t}H(bb)$ analysis using the BDT (and Fisher) methods.

In this chapter, two multivariate analysis methods are analysed in terms of their discriminating power between the signal $t\bar{t}H$ and the background given by the $QCD$. Furthermore a first evaluation of the most significantly discriminating variables is carried out.

Preselection Requirements

Based on the topology of the signal events and the previous study of the signal trigger efficiency, the preselection requirements for the multivariate analysis are:

- a number of jets ($nJets$) $\geq 6$;
- a number of b-jets ($nB.Jets$) $\geq 2$, tagged with the CSVM algorithm ($csv > 0.814$);
- no Monte Carlo leptons (therefore $nLeptons = 0$);
- Missing Transverse Energy ($MET$) $< 80$ GeV;
- $H_T > 450$ GeV;
- 6 jets with $p_T > 40$ GeV and $|\eta| < 2.4$;
- the selection by our signal trigger ($Trg$ $0$ $||$ $Trg$ $2$).
Variables involved in the study

The variables involved in the multivariate analysis are:

- the number of jets in each event: \( n_{Jets} \);
- kinematic variables: \( H_T, p_T^0, p_T^1, p_T^2, p_T^3, p_T^4 \) and \( p_T^5 \);
- the angular distance between the closest pair of b-quarks and their invariant mass: \( \Delta R_{bb,\text{min}} \) and \( m_{bb,\text{min}} \);
- event geometry variables: sphericity, aplanarity, \( \text{foxWolfram}[0], \text{foxWolfram}[1], \text{foxWolfram}[2], \text{foxWolfram}[3] \);
- kinematic fit variables: \( m_{Top}, p_T^{\bar{t}}, \Delta R_{bb,\text{top}} \) and \( \chi^2 \). The kinematic fit is performed under two constraints:
  1. Fixed \( m_W \) for both the \( t \)- and \( \bar{t} \)-quarks;
  2. Top quark mass equal to the antitop quark mass in the \( t\bar{t} \) pair.

In order to find the variables with the most significant discriminating power between the signal (\( t\bar{t}H \)) and the background (\( QCD \), in this case), two multivariate methods have been used: the Fisher discriminant and the Boosted Decision Tree (BDT), both with categories.

The different bins in Figures 3.1 and 3.2 represent the areas of the ROC curves obtained by removing the correspondent variable from the training of the MVA method (the black bin refers to the training performed with all the variables).

As far as the Fisher discriminant is concerned, the variables that discriminate the most signal from background are:

- the \( \chi^2 \) of the kinematic fit;
- the top quark mass;
- the aplanarity;
- the \( \text{foxWolfram}[3] \);
- the \( H_T \).

By contrast, according to the BDT method, the variables with the highest discriminating power are:

- the \( \chi^2 \) of the kinematic fit;
Figure 3.1: ROC curves and ROC curves areas obtained with the Fisher discriminant method.

Figure 3.2: ROC curves and ROC curves areas obtained with the BDT discriminant method.
• the top quark mass;
• the number of jets $n_{Jets}$;
• the mass of the closest b pair $m_{bb,\text{min}}$;
• the $\Delta R_{bb,\text{min}}$;
• the aplanarity.

As expected, the areas of the BDT ROC curves are larger than the corresponding Fisher areas, which means that the BDT has a better performance than the Fisher discriminant.
Chapter 4

Conclusions

As far as the trigger study is concerned, the best trigger choice for the study of all hadronic $ttH$ in terms of both trigger efficiency and signal sensitivity is the logic OR between Trg 0 and Trg 2. The optimized kinematic cuts are:

$$H_T > 450 \text{ GeV} \quad \text{and} \quad P_{T_j} > 40 \text{ GeV}$$

Eventually, above these kinematic cuts, the chosen reference trigger is unbiased.

Conversely, when it comes to the multivariate analysis, several BDT trainings have been performed, in order to evaluate the most significant contributions to the classification of signal against background. The variables with the most significant discriminating power are:

- the $\chi^2$ of the kinematic fit;
- the top quark mass;
- the number of jets;
- the mass of the closest $b$ pair.