

Report on the Summer Student project „Improved background estimation for the $t\bar{t}H$ analysis”

Karl Ehatäht

Supervisors: *Lorenzo Bianchini, Mario Kadastik, Joosep Pata*

August 21, 2014

The Yukawa coupling of Higgs boson to top quark, y_t , is yet still to be determined. One way to measure the constant is to study loop-induced processes in which top-Higgs interaction takes place, however, it may also include the effects of Beyond the Standard Model particles. A more direct probe of y_t can be provided in those processes in which top quark radiates a Higgs boson [1]. The process of interest, $pp \rightarrow t\bar{t}H$ (or simply $t\bar{t}H$), has the largest cross section when Higgs decays into a pair of b quarks [3] (see Fig. 1).

The large multijet background in the final state can be reduced by selecting only the events with one or two charged isolated leptons. However, the signal is still dwarfed by QCD (quantum chromodynamics) background ($t\bar{t} + b\bar{b}$, $t\bar{t} + c\bar{c}$ and $t\bar{t} + jj$) after the cut [4]. Although the former subprocess, $pp \rightarrow t\bar{t}b\bar{b}$, remains still irreducible¹, the rest can be suppressed by requiring the jets to be b -tagged. To this purpose, a jet-by-jet discriminant (CSV, combined secondary vertex) is calculated based on the jet kinematics and information about secondary vertex. It is used as a boolean variable to separate b -jets from non- b jets – if the CSV value exceeds some fixed working point, it is b -tagged. All b -tagged jets are treated as b -jets in further analysis, although non- b jets might slip through the cut, depending on the effectiveness of the algorithm.

Since Monte Carlo (MC) simulated samples are finite due to computing reasons, and the large number of required b -tagged jets in the final state results in a small number of surviving events, one needs a better strategy to rescue more events. Fortunately, the distribution of discriminator is known from MC simulations, it can be used probabilistically to reweigh the events. Instead of throwing away most of the MC events, one could assign a weight to each event based on the CSV values of the jets. The method then leads to the enhancement of MC statistics.

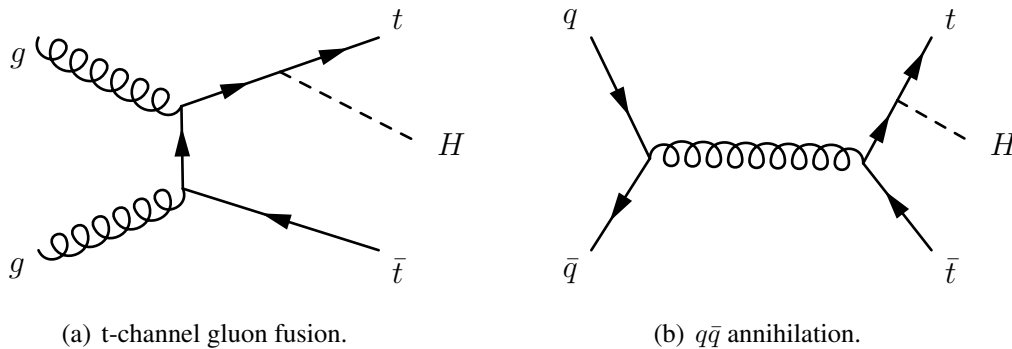


Figure 1: Leading order $t\bar{t}$ production with Higgs radiation.

¹It has the same final state as the signal.

This study concentrates on semi-leptonic (SL) channel² of $t\bar{t}$ + jets MC simulations to test the method. The workflow went as follows:

1. Derivation of histograms of CSV variable (Fig. 2(a)) for specific ranges of jet p_t (transverse momentum of the jet), $|\eta|$ (pseudorapidity) and three flavours (quarks and antiquarks are treated in the same way). The ranges are given in the table below:

variable	ranges
p_t (GeV)	[20, 30), [30, 40), [40, 60), [60, 100), [100, 160), [160, ∞)
$ \eta $	[0, 0.8), [0.8, 1.6), [1.6, 2.5)
flavour	$b(b), c(\bar{c}), g$

2. Sampling each histogram, i.e. random numbers were drawn from CSV distribution as many times as there were entries in it (see Fig. 2(c)). In order to quantitatively check how compatible the resulting histograms and sampled distributions are, one needs to do statistical tests between them. Kolmogorov-Smirnov (K-S) and χ^2 test are just suitable for this purpose. In a nutshell, K-S finds the maximum distance between the cumulative distribution of the sample and the probability distribution function (PDF), and then converts it to p -value [5]. The closer p -value is to 1.0, the more likely is the sample drawn from reference distribution (which is the null hypothesis of the test). The latter test on the other hand calculates the χ^2 test statistic,

$$\chi^2 = \frac{1}{MN} \sum_{i=1}^r \frac{(Mn_i - Nm_i)^2}{n_i + m_i},$$

where N and M is the total number of events in the histograms (respectively), n_i and m_i the number of events in i -th bin, and r is the number of bins (see [6] and references therein). If the null hypothesis (i.e. two distributions are identical) is true, then the statistic will be drawn from χ^2 -distribution. With the number of degrees of freedom, $r - 1$, and the χ^2 test statistic one can look up the corresponding p -value, the meaning of which is the same as for K-S test. Since all 54 (normalized) histograms passed the tests, they were treated as PDFs from this stage on.

3. Plotting the b-tagging efficiency³ and mistag rate⁴ of c - and light jets for each p_t and η range just to validate the setup. An example is given in Fig. 2(f). The plot reveals that at medium CSV cutting point (working point) CSVM = 0.679 the b-tagging efficiency is around 65%, and the mistag rate of c - and light jets is about 16% and 3%, respectively [7].
4. Tossing random CSV values (according to PDFs) until the generated value of the jet passed the working point (i.e. is greater than or equal to CSVM). The results were the same as if the PDFs were cut off at CSVM (Fig. 2(d)). The number of iterations a jet needs to pass the working was also recorded. Fig. 2(e) shows that the number of iterations falls off exponentially. This is to be expected, because say if the probability that a given jet passes the working point is p_i , the probability of passing after N -th iteration is $P_{N,i} = (1 - p_i)^{N-1} p_i \Rightarrow \log P_{N,i} = (N - 1) \log(1 - p_i) + \log p_i$. Since the probability of the jet passing the working point, p_i , is greater for b -jets than for c - or light jets, it follows that more iterations are needed in the latter case to pass the working point.

²Semi-leptonic (or single-leptonic) channel in the current context means that only one of the W -bosons resulting from $t \rightarrow W^+ b$ ($\bar{t} \rightarrow W^- \bar{b}$), decays leptonically ($W^- \rightarrow \ell \bar{\nu}_\ell$, $W^+ \rightarrow \bar{\ell} \nu_\ell$) and the other hadronically ($W \rightarrow q \bar{q}'$). Thus the final state of SL $t\bar{t}H$ has one tight lepton and 6 jets of which 4 are coming from b -quarks.

³B-tagging efficiency is defined as the number of b -jets that passed the CSV cut over the number of all b -jets.

⁴Mistag rate is the number of non- b jets that passed the CSV cut over the number of all non- b jets (of the same flavour).

5. Only the events which contain exactly N_{jet} jets with $p_t \geq 20$ GeV and $|\eta| < 2.5$ were selected. The jets were sorted by their p_t in descending order. While looping over them $N_{\text{iter}} = 10^4$ times, a CSV value was generated for each jet. If the number of jets that passed CSVM cut coincides with the required number of b-tagged jets, N_{tag} , the number of passes, N_{passes} , was incremented. Probability that a given N_{jet} -jet event contains N_{tag} b-tagged jets is simply $P_M(N_{\text{tag}}, N_{\text{jet}}) = \frac{N_{\text{passes}}}{N_{\text{iter}}}$.
6. The same events (that passed jet multiplicity and kinematical cuts) were once more sampled and their CSV values were recorded. This step is later needed to validate the method.
7. Derivation of cumulative histograms from PDFs in order to easily read the probability that a jet passes the CSVM cut (Fig. 2(b)). While looping over the same jets as in 5. a combined probability was assigned to each event. For example, if $N_{\text{jet}} = 3$, $N_{\text{tag}} = 2$ and individual probabilities that a jet passes the working point are p_1 , p_2 and p_3 , the overall probability that an event passes the cut is

$$P_A(N_{\text{tag}} = 3, N_{\text{jet}} = 2) = p_1 \cdot p_2 \cdot (1 - p_3) + p_2 \cdot p_3 \cdot (1 - p_1) + p_1 \cdot p_3 \cdot (1 - p_2).$$

The code works for any N_{jet} and N_{tag} .

8. All events of interest passed the probability normalization check,

$$\sum_{N_{\text{tag}}=0}^{N_{\text{jet}}} P_i(N_{\text{tag}}, N_{\text{jet}}) = 1, \quad i = A, M.$$

Also, the probabilities for each event obtained in 5. and 7. were really close (up to 3% of difference), although the possible values in 7. are limited.

9. Looping over all events and summing the probabilities found in 5. and 7. indicates that the integrals were nearly exact (up to 3% of difference):

$$\sum_{\text{events}} P_M(N_{\text{tag}}, N_{\text{jet}}) \approx \sum_{\text{events}} P_A(N_{\text{tag}}, N_{\text{jet}}).$$

By looking at the randomly generated CSV values in 6., there was up to 2% of discrepancy between the number of events in which the number of b-tagged jets equals to N_{tag} , and the sum of probabilities $\sum P_i$ ($i = A, M$)⁵.

10. Jet p_t , $|\eta|$ and measured CSV were plotted with both of the following methods (Fig. 3):

- (a) loop over all N_{jet} jets, read generated CSV value and check if the number of b-tags in an event equals to N_{tag} ; if so, fill the histogram (the so-called hard cut method);
- (b) loop over all N_{jet} jets and fill the histogram with weight $P_i(N_{\text{tag}}, N_{\text{jet}})$ ($i = A, M$) (reweighting method).

Since the two histograms passed Kolmogorov-Smirnov and χ^2 tests, it is reasonable to only use analytical probabilities as weights because method 5. is computationally intensive.

In conclusion, the proposed reweighting method works with CSV distributions of specific p_t , $|\eta|$ and flavor ranges – full MC statistics is preserved. All events that didn't pass the hard cuts were saved (see Table 1 for $N_{\text{jet}} = 6$ case)⁶. Also, the analytical method, when applicable, is faster and as

⁵The discrepancy was found for $N_{\text{jet}} = 6$ and $N_{\text{tag}} = 4$. Of course it depends on the number of events that pass the cuts and the number of required b-tags – the lower the number of survived events, the larger the difference.

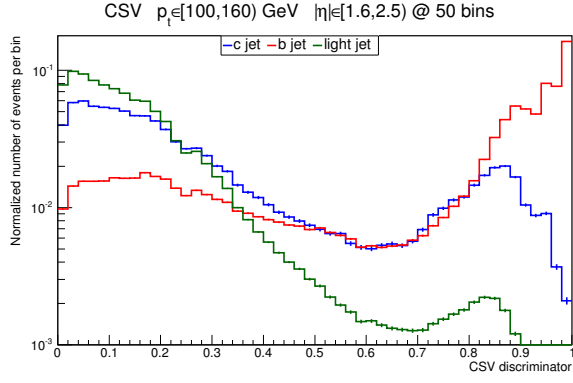
⁶It is true as long as not too many jets are required to be b-tagged. For $N_{\text{tag}} = 5$ and $N_{\text{tag}} = 6$ case few events pass the hard cuts, thus the distribution of some kinematical variable is statistically incompatible with the full MC version.

precise as MC sampling. The next step would be a proof that the method works with „measured” CSV values (i.e. the ones obtained in full MC simulation), and test the effect of correlations.

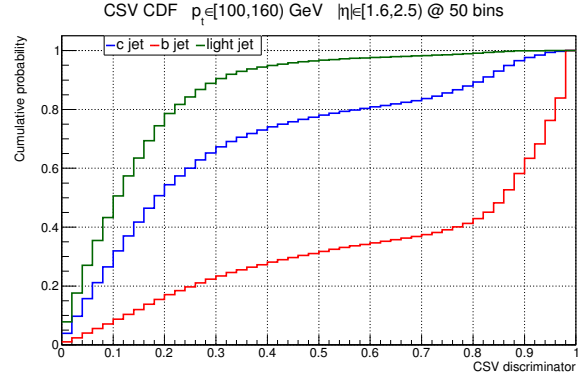
N_{tag}	Number of events (counting b-tags)	Sum of analytical weight	Percentage of all events (counting b-tags)
0	319392	319516.28	13.81%
1	992953	993980.81	42.95%
2	850589	851460.75	36.79%
3	138667	137588.78	5.99%
4	9566	9330.78	0.413%
5	351	335.00	0.015%
6	4	4.288	$\sim 0\%$
total	2311522	2312216.69	100%

Table 1: Percentage of N_{tag} b-tagged events for $N_{\text{jet}} = 6$. CSV values are generated according to 6., and then the number of b-tags was later counted for each event (second column). For comparison, all weights are added as in 9. (third column). Total number of events here means the number events with $N_{\text{jet}} = 6$ jets which passed the kinematical cuts. In case of $N_{\text{jet}} = 6$ and $N_{\text{jet}} = 4$ only 0.413% of all events would have passed the hard cuts.

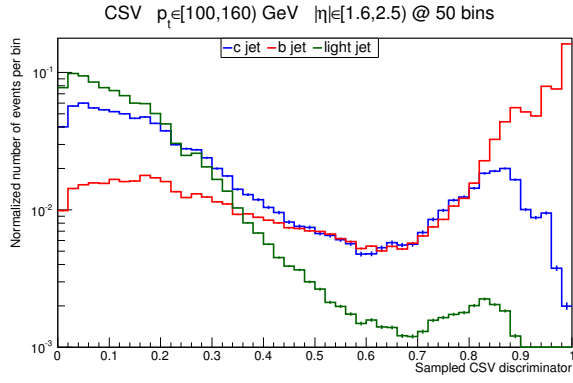
Appendix



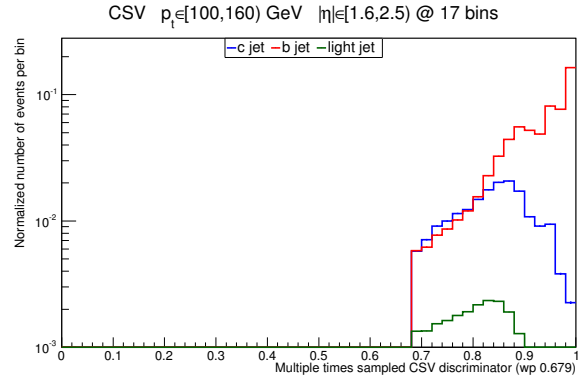
(a) CSV PDF derived from MC simulations.



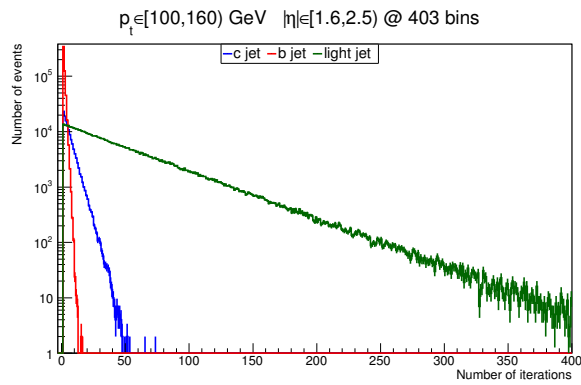
(b) Cumulative distribution function (CDF) of the same PDF.



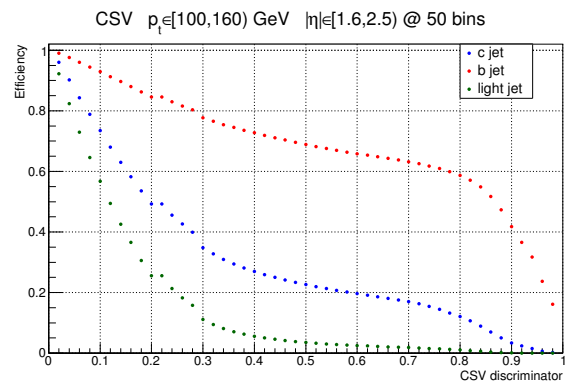
(c) Sampled CSV distribution.



(d) Sampled CSV distribution until the jet passes the working point. All events are included.

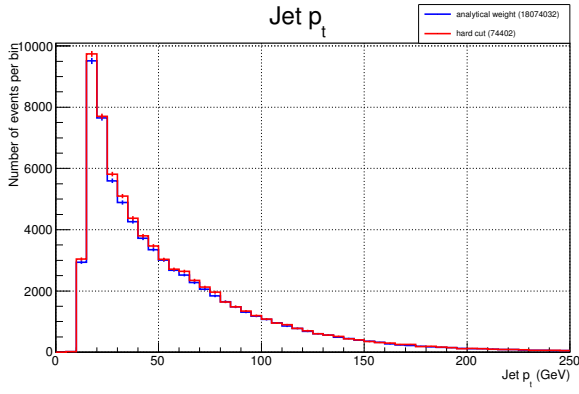


(e) Number of iterations needed to pass the working point.

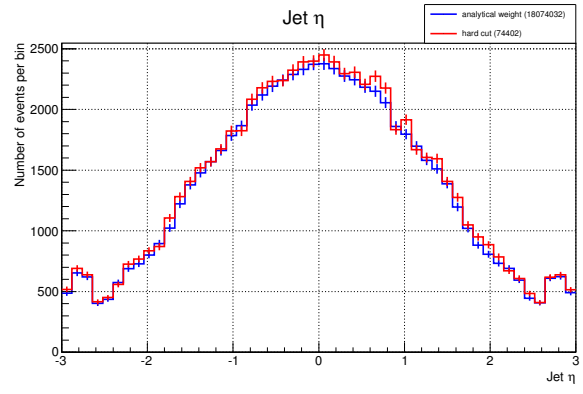


(f) Number of iterations needed to pass the working point.

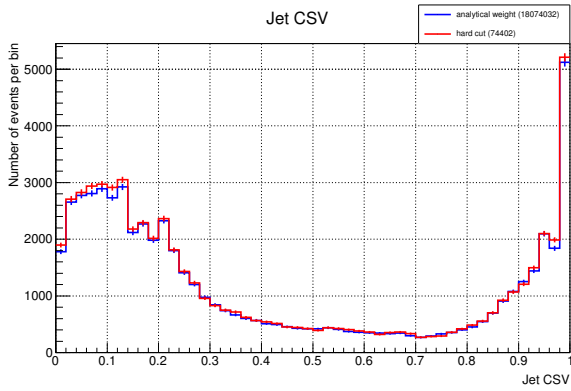
Figure 2: Results obtained from 1.-4.



(a) Jet p_t .



(b) Jet η .



(c) Measured jet CSV.

Figure 3: Hard cuts vs reweighted events ($N_{\text{tag}} = 4$, $N_{\text{jet}} = 6$). All jets of passed events are included to the plot. There are some discrepancies due to low statistics of the hard cut method.

References

- [1] The CMS Collaboration. “Search for $t\bar{t}H$ events in the $H \rightarrow b\bar{b}$ final state using the Matrix Element Method”. In: *CMS Physics Analysis Summary*. CMS PAS HIG-14-010. 2014.
- [2] Marco Farina et al. “Lifting degeneracies in Higgs couplings using single top production in association with a Higgs boson”. In: *JHEP* 1305 (2013), p. 022. DOI: [10.1007/JHEP05\(2013\)022](https://doi.org/10.1007/JHEP05(2013)022). arXiv: [1211.3736](https://arxiv.org/abs/1211.3736) [hep-ph].
- [3] D. Dittmaier et al. *Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables : Report*. pt. 1. CERN, 2011. ISBN: 9789290833581.
- [4] ATLAS Collaboration. “Search for Higgs Boson Production in Association with a Top-Quark Pair and Decaying to $b\bar{b}$ ”. In: *ATLAS Conference Notes*. ATLAS-CONF-2014-011. 2014.
- [5] *Kolmogorov-Smirnov test*. <http://root.cern.ch/root/html/TH1.html#TH1:KolmogorovTest>. Accessed: 2014-08-21.
- [6] χ^2 test. <http://root.cern.ch/root/html/TH1.html#TH1:Chi2Test>. Accessed: 2014-08-21.
- [7] *CSVM definition*. <https://twiki.cern.ch/twiki/bin/viewauth/CMS/TopBTV>. Accessed: 2014-08-21.