PROCEEDINGS OF SCIENCE

Probing the heavy flavour content of ttbar events in proton-proton collisions at CMS

Antonio Tropiano[*†]
Università degli Studi di Firenze e INFN
E-mail: antonio.tropiano@cern.ch

In the framework of the standard model (SM), the top quark is expected to decay to a W-boson and a b-quark 99.8% of the times due to the Cabibbo-Kobayashi-Maskawa (CKM) matrix element V_{tb} being close to unity. The current experimental limits from the Tevatron on V_{tb} from top-quark pairs and single-top production are consistent with the SM predictions. The higher energy of proton-proton collisions and larger top quark production cross section at the Large Hadron Collider (LHC) may provide an improved reach in the measurement of V_{tb}. We present analysis strategies dedicated to measure ratios of branching ratios of the top quark using ttbar events collected with the CMS detector, in which either one or both W-bosons from the top-quark decays lead to a lepton and a neutrino. These di-leptonic and semi-leptonic final states provide high cross section with small background. The sensitivity of the measurement is evaluated after particle identification and detector reconstruction. Data-driven techniques to control the backgrounds are discussed and the expected simulation results are presented for a center-of-mass energy of 10 TeV. We also discuss how the method can be used to measure directly from data the efficiency of the algorithms used to discriminate jets coming from the hadronization of b quarks from the lighter quarks and gluons (b-tagging).

The Xth Nicola Cabibbo International Conference on Heavy Quarks and Leptons,
October 11-15, 2010
Frascati (Rome) Italy

[*]Speaker.
[†]On behalf of the CMS collaboration.

PoS(HQL 2010)041

1. Introduction

Top quarks decay mostly to Wb, while the final states Wd and Ws are suppressed by the small values of the CKM matrix elements V_{td} and V_{ts}. Besides single top studies, V_{tb} can be obtained also through top pairs production, by measuring $R = B(\frac{t\to Wb}{t\to Wq})$, with $q = d,s,b$, and assuming that exactly 3 generations of quarks exist, as the Standard Model (SM) predicts; indeed, by imposing the unitarity of the 3×3 CKM matrix, such ratio is $R = |V_{tb}|^2/(|V_{td}|^2 + |V_{ts}|^2 + |V_{tb}|^2) = |V_{tb}|^2$.

Without any assumption on the number of generations of quarks, an R measurement is still useful to put constraints on V_{tb} and it can give a clue on the existence of a fourth generation; indeed in such scenario, R is appreciably less than the SM value [1]. With the CMS experiment [2], two feasibility studies of the measurement of R have been carried on, one using selected semileptonic $t\bar{t}$ events [3], the other using selected dileptonic $t\bar{t}$ events [4]. Both the analyses use data-driven methods in order to estimate the irreducible background contribution and consider the number of b-tagged jets as the physical observable.

2. Event selection in the dileptonic channel

The event selection in this channel is tuned to identify leptonic final states with two prompt, isolated leptons with high transverse momenta in the CMS detector.

Data samples are triggered by requiring a non-isolated single muon ($p_T > 9$ GeV/c) or a single electron ($E_T > 15$ GeV). Lepton candidates are reconstructed with $p_T > 20$ GeV/c in the fiducial region $|\eta| < 2.4$ of the detector. The track assigned to each lepton candidate is required to have an impact parameter compatible with prompt production: $|d_0| < 400\mu$m.

Identification and isolation requirements are imposed on the lepton candidates. Relative tracker isolation (I_{trk}) is defined as the fraction of momentum carried by the track assigned to the lepton candidate with respect to the total momentum of tracks in a cone $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} < 0.3$ built around the lepton track. A similar definition that uses the energy of the calorimeter cluster assigned to the lepton is applied. $I_{trk} > 0.9$ is required for each lepton candidate and $I_{cal} > 0.9$ (> 0.8) is required for the muon (electron) candidates.

Jets are reconstructed using the Seedless Infrared Safe cone algorithm [5] and are required to have at least one assigned track so that the b-tagging algorithms can be applied. The energy of the jets is corrected for the η dependence and absolute E_T using Monte Carlo based corrections for generator level jets. Taggable jets are selected with $E_T > 30$ GeV/c and $|\eta| < 2.4$. Jet candidates are further required to be separated from the selected leptons by ΔR(jet, lepton) > 0.3 and to have an electromagnetic fraction EMF<0.98. The total missing transverse energy is corrected for the energy deposited by muons and is required to be above 30 GeV.

In order to identify the flavor of the jets, specific algorithms are used. For this study, the Track Counting (TC) and Jet Probability (JP) algorithms are used to tag the b-jets [6].

Table 1 shows the expected event yield for an integrated luminosity of 250 pb^{-1}, for signal events. After the opposite-charge requirement, a signal to background ratio of approximately 10/1 is expected. The largest background contribution comes from single top (3.3%) and W/Z+jets (3.0%). A detailed study was performed by relaxing some of the selection cuts, which indicates that background due to QCD sources is small.

Selection	Total	tt dileptons
Triggered	$(426 \pm 1) \cdot 10^6$	6251 ± 25
$\geq$ 2 leptons (>20 GeV/c)	$(204.7 \pm 0.5) \cdot 10^3$	2595 ± 16
1 e and 1 μ	2531 ± 32	1344 ± 12
$\geq$ 2 jets (>30 GeV)	1041 ± 12	914 ± 10
$E_T \geq 30$ GeV	884 ± 10	789 ± 9
Opp. sign leptons	867 ± 10	787 ± 9

Table 1: Expected event yield for an integrated luminosity of L=250 pb^{-1} using a MADGRAPH sample. Only statistical uncertainties from the Monte Carlo samples are shown.

3. Probing the heavy flavor content

The heavy flavor content of the selected events can be probed from the *b*-tagging multiplicity distribution. In the selected events, jets are *b*-tagged if the discriminator is larger than a given threshold.

Despite small contributions from other background processes there is a non negligible probability that at least one jet from a $t\bar{t}$ decay is either missed because it was not reconstructed or because it did not pass the jet selection criteria, and another jet is chosen instead, such as, for example, jets from initial (ISR) or final (FSR) state radiation. This will be referred to as "jet misassignment" and an estimate of the jet misassignment level has to be made from data. The estimate is done in terms of probability weights α_i , where i = 0, 1, 2 is the number of jets from top decays correctly reconstructed and selected.

3.1 Determining the heavy flavor content from data

The expected b-tagging multiplicity can be modelled as:

$$P_k = R^2 P_k(bb) + 2R(1-R)P_k(bq) + (1-R)^2 P_k(qq) \tag{3.1}$$

where P_k is the probability to observe k *b*-tags written as a combination of R (i.e. the ratio of branching fractions of the top quark to *b* quarks) and the contributions from events in which the $t\bar{t}$ system decays to 2, 1 or 0 *b*-jets, indicated as $P_k(bb)$, $P_k(bq)$ and $P_k(qq)$ (*b*=heavy flavor, *q*=light flavor), respectively. These contributions are a function of the *b* tagging efficiency ε_b , the mis-tagging probability ε_q and α_i.

Using this model, R (or ε_b) can be fit by a likelihood function in different ways:

- Fit R or ε_b: checks the consistency of the measurement;
- choose one exclusive jet multiplicity bin: checks model consistency, and allows one to individually choose the bins which may be affected differently by systematic uncertainties;
- use all selected events inclusively;
- estimate α_2 from data, and leave α_0 as a free parameter ($\alpha_1 = 1 - \alpha_2 - \alpha_0$): fits simultaneously R (or ε_b) and the background contribution to the dilepton channel.

3.2 Jet misassignment estimate

The selected events are a combination of three different categories: events with no jet selected from the top decays (background-dominated), events with only one jet correctly assigned to the top decay (combination of signal and background), events with two jets correctly assigned to the top decays (signal-dominated).

The contributions of these three classes of events are defined by the weights α_i. The weights α_i can be parametrized in terms of a binomial combination of α, the probability of correctly assigning individual jets. The value of α can be estimated using the kinematic properties of the events directly from data. A correlation can be sought in the lepton-jet pairs originating from the same top quark decay [7] and it is possible to show that no pair with $M_{l,b} > M_{l,b}^{max} = \sqrt{m_t^2 - m_W^2} = 156$ GeV/c^2 should be observed (spectrum endpoint). It can be shown that the combinations with $M_{lepton,jet} \geq 190$ GeV/c^2 are dominated by misassignments.

Two methods are proposed to emulate the invariant mass distribution of the misassigned jets: "swapping" the jet in the assigned lepton-jet pair, with a jet from a different event, or "randomly rotating" the momentum vector of the selected leptons.

Figure 1 (left) shows the invariant mass distribution reconstructed for all lepton-jet assignments found in each event (signal and background samples combined). The distribution of the swapped and randomly rotated pairs, normalized to fit the high-end part of the distribution, is superimposed. The two background models provide a good estimate of the fraction of misassigned pairs with $M_{lepton,jet} > 190$ GeV/c^2 (Figure 1, right). The excess at $M_{l,j} < 150$ GeV/c^2 is due to $t \to Wb$ background events. The normalization factor applied to the distribution of the swapped (randomly rotated) pairs is related to the misassignment fraction, $1 - \alpha$.

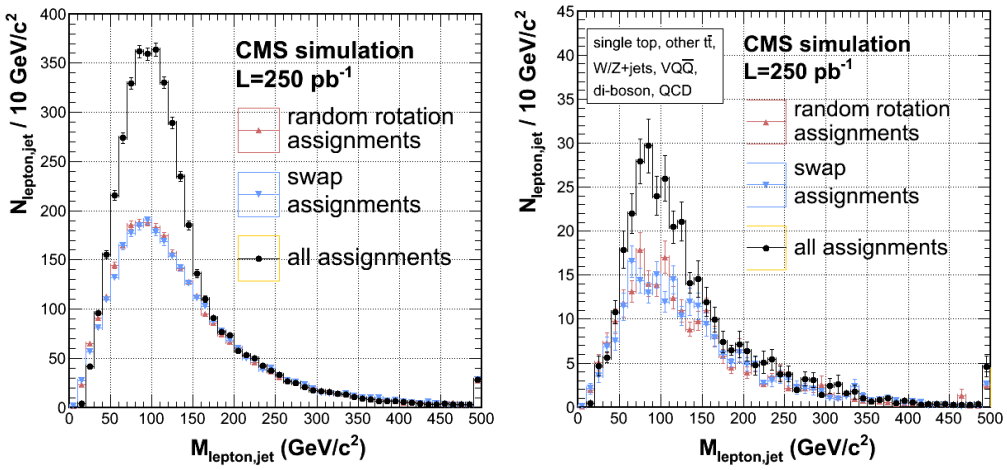

Figure 1: Invariant mass of the lepton-jet pairs for the all lepton-jet assignments found in each event (reconstruction level). (Left) all "data"; (right) background contributions only. MADGRAPH samples are used. The two additional curves (rescaled to fit the tail of the spectrum) correspond to the random rotation and swap models.

3.3 Measurement of R

The measurement of R is now discussed. Here ε_b is taken as an input. In this measurement,

the value of α is measured. Figure 2 shows the results obtained by fitting R (or R and α_0, the background level) using jets tagged with the Jet Probability algorithm.

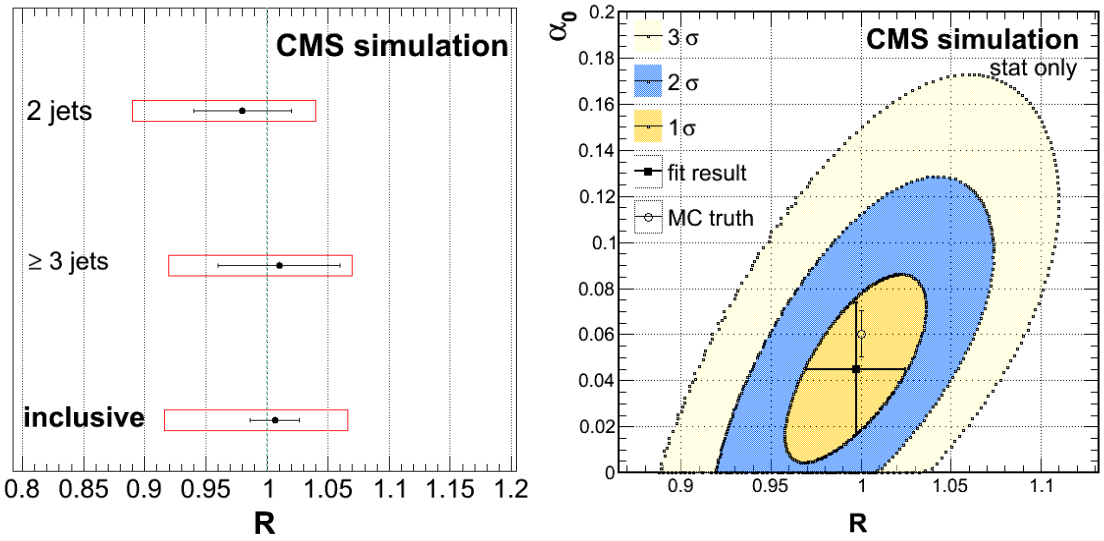

Figure 2: (Left) Fit results for R using the Jet Probability b-tagging algorithm (in the MC R = 1). The fits for the exclusive and inclusive jet multiplicity bins are shown. The inclusive jet multiplicity bin fit uses the value of α derived from data, while MC truth values are used for the other bins. Bars (boxes) represent the statistical (systematic) uncertainty. (Right) Fit to R and α_0 (background level). $\alpha_2$is fixed using a binomial combination α. Only statistical uncertainties are shown.

The total uncertainty (stat+syst) in the measurement of R is 9% for L = 250 pb^{-1} . The systematic uncertainties are dominated by the uncertainty on the b-tagging efficiency. The uncertainty due to different ISR/FSR content in the final sample is expected to be small (< 1%). A closure test is needed in order to evaluate if the method is sensitive to values of R different from 1. In order to make this test, $t\bar{t}$ samples with R= 0, 0.5, 1 are used. All backgrounds are included as before. The b-tag multiplicity distributions obtained this way are sampled according to the expected number of events and fit to determine R. The statistical uncertainty of each fit result is then determined from the width of the distribution of $R_{generated} - R_{measured}$. A good agreement is found between the generated and fitted values of R.

4. Event selection in the semi-leptonic channel

The event selection for this channel is designed to select a final state with four or more jets, a single lepton (electron or muon) and missing transverse energy.

The events have to pass at least one of the two following trigger paths: a single muon trigger, that requires at least a muon reconstructed in the muon system and in the tracker with $p_T > 15$ GeV/c, or a single-electron trigger that requires a loosely isolated electron with $p_T > 18$ GeV/c. The lepton candidate must be reconstructed with pT > 30 GeV/c. An isolation variable is defined as the sum over the tracks with $p_T > 1.5$ GeV/c and $\Delta z < 0.1$ cm in a cone with R=0.2 and an inner cone veto at R=0.02 around the candidate track, plus the sums of the energies of all the calorimeter

towers within a cone of R=0.3 around the lepton candidate. The isolation is required to be less than 0.1.

The jet reconstruction algorithm uses the calorimetric energy deposits and performs an iterative cone procedure with radius ΔR=0.5. Relative and absolute jet energy corrections are applied to account for the dependence of the jet response as a function of η and p_T. The jet candidates are selected by requiring E_T>40 GeV and $|\eta|$<2.4. They have to be separated from the lepton candidate, imposing ΔR(jet,lep)>0.5, and their fraction of electromagnetic energy to the total energy has to be less than 1.

A useful kinematic variable to reduce the background contamination is Centrality. It represents the fraction of the hard scattering going in the transverse plane and it is defined as:

$$Centrality = \frac{\sum E_T}{\sqrt{(\sum E)^2 - (\sum p_z)^2}}$$

where p_z refers to the z component of the jet momentum. All the sums run over the reconstructed jets. It has a good discrimination power especially between the signal and the QCD multi-jet events.

The final step of the event reconstruction is the computation of the invariant masses using the selected reconstructed objects. Among the selected jets, the four with largest E_T are considered as coming from the decays of the two top quarks and of the hadronic W. In order to choose the right combination, a two step association is used. Beforehand the masses and the widths of the hadronic W boson and the tops are obtained from simulation. The distributions of the three invariant masses of the reconstructed objects well matched to the generated particles are used to obtain the parameters m_{WHad} , m_{tHad} , m_{tLep} , $\sigma(m_{WHad})$, $\sigma(m_{tHad})$ and $\sigma(m_{tLep})$. First the hadronic W boson is reconstructed by computing the invariant mass of every pairs of jets among the four. The pair with the nearest invariant mass to the W one, namely ij, is chosen. The following cut is applied:

$$|m_{ij} - m_{WHad}| < \sigma(m_{WHad})$$

The second step is the association of the two remaining jets (k and p) to the partons coming from the direct decay of top quarks. To this end a χ^2 based on the two top quarks masses is defined:

$$\chi^2 = \left(\frac{m_{ijk} - m_{tHad}}{\sigma(m_{tHad})}\right)^2 + \left(\frac{m_{lvp} - m_{tLep}}{\sigma(m_{tLep})}\right)^2$$

where i and j are the 2 jets chosen as coming from the W boson decay. Therefore now the only combinatorial ambiguity lies in the choice of which one of the two remaining jets is associated to which of the two top quarks. The association that minimizes the χ^2 is assumed to be the correct one. We consider the events with a large χ^2 as events which are wrongly reconstructed, so the cut $\chi^2_{min} < 4$ is applied. Table 2 reports the selection efficiencies for the signal and for the most important background processes. The multi-jet QCD expected event number reported in Table 2 is the result of the factorization of two sets of cuts.

The goal of the analysis is to determine the distribution of the number of b-tagged jets among the best four jets in the $t\bar{t}$ semi-leptonic final state and to fit the distribution with the function of Eq. 3.1, to extract the parameter R. The b-tagging algorithm adopted in this analysis takes into account

	ε(HLT)%	ε(lep)%	ε(jets)%	ε(centr)%	$\varepsilon(\Delta M_W)$%	$\varepsilon(\chi^2)$%	N_{events} (1 fb^{-1})
$t\bar{t}$ semil	54	27.5	7.23	5.62	3.79	1.48	2650
$t\bar{t}$ others	22	7.9	$6.3{\cdot}10^{-1}$	$4.9{\cdot}10^{-1}$	$2.4{\cdot}10^{-1}$	$4.7{\cdot}10^{-2}$	109
W + jets	34	16	$9.5{\cdot}10^{-3}$	$6.8{\cdot}10^{-3}$	$3.7{\cdot}10^{-3}$	$6.5{\cdot}10^{-4}$	260
Z + jets	50	19	$1.9{\cdot}10^{-2}$	$1.4{\cdot}10^{-2}$	$7.4{\cdot}10^{-3}$	$1.4{\cdot}10^{-3}$	52
tW	35	18	1.2	$8.4{\cdot}10^{-1}$	$6.4{\cdot}10^{-1}$	$1.8{\cdot}10^{-1}$	52
QCD	2.7	$7.8{\cdot}10^{-3}$	$5.2{\cdot}10^{-6}$	$4.3{\cdot}10^{-6}$	$2.4{\cdot}10^{-6}$	$1.1{\cdot}10^{-6}$	56

Table 2: Expected selected event number after 1fb^{-1} of integrated luminosity and selection efficiency after every selection cut for the signal ($t\bar{t}$ semileptonic) and the main background processes.

the signed significance of the impact parameter of every well reconstructed track in the jet in order to compute a confidence level. The discriminator is defined as the negative logarithm of such a confidence level. All the results reported in the following refer to the working point in which a jet is considered *b*-tagged if it has this discriminator greater than 0.3. For the chosen working point we obtain $\varepsilon_b = (82 \pm 1)\%$ and $\varepsilon_q = (12 \pm 1)\%$ considering all the *b*-jets in the selected sample.

5. Background contribution subtraction strategy

The method developed to evaluate and subtract the background contribution does not use the simulation to obtain the distributions to be subtracted. Semi-leptonic $t\bar{t}$ events, for which the two partons coming from the direct decay of tops are not well matched to any jet among the best four ones, should be considered as background. The χ^2 defined previously, and referred to as χ^2_{normal} in the following, has a peak at low values of χ^2 for correctly reconstructed signal events. Background and incorrectly reconstructed $t\bar{t}$ events lead to low values of χ^2_{normal} only due to random combinatorics. Therefore if the direction of one of the selected jets is artificially changed, the mass χ^2 distribution should remain the same for background events, while we expect the distribution for signal events will appreciably change. We can define a χ^2_{random} just like the χ^2_{normal}, but computed by assigning a random direction to one of the two jets considered as coming directly from the tops. Uniform distributions for ϕ and η have been generated, allowing for ϕ in the range (-π,π) and η in the range (-2.4, 2.4), as for the selected true jets. Then the procedure was repeated leading up to the new combination that gives the minimum χ^2, called χ^2_{random}. Fig. 3 shows the distribution of the χ^2_{min} variable separately for signal and background events. The left distribution, referring to the signal, shows that the difference between χ^2_{normal} (solid) and χ^2_{random} (dashed) distribution in the selected region ($\chi^2 < 4$) is clearly visible. On the contrary the right distribution, which refers to the background sample, shows agreement within the statistical uncertainty, between the Normal (solid) and Random (dashed) distributions. The goal of this approach is to create two distributions of the number of b-tagged jets, to be subtracted bin by bin. The n_{btag} distribution of the events selected after the cut on the χ^2_{normal} will be referred as n_{normal} while the events selected after the cut on the χ^2_{random} will be referred to as n_{random}. If one considers the whole data sample, containing signal and background events, and computes bin-by-bin the difference of the Normal-Random distributions, the resulting n_{btag} distribution will be proportional to the distribution of the signal only. This distribution, after normalization, is to be fitted with equation 3.1.

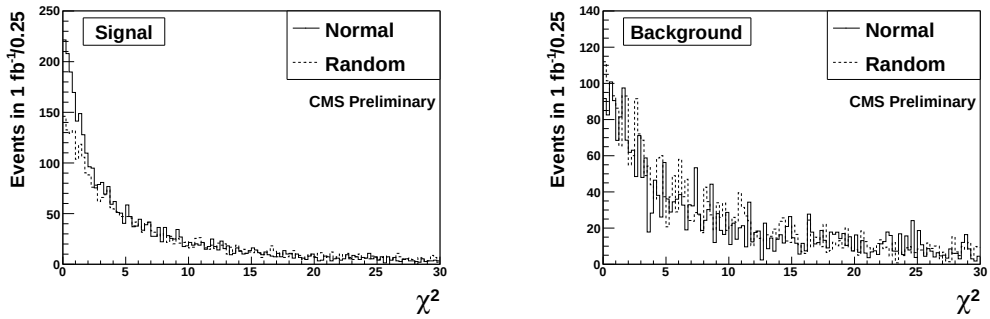

Figure 3: Left: χ^2_{min} distribution of the signal sample (as defined in the text). Right: χ^2_{min} distribution of the complete background sample. Both distributions show the χ^2_{normal} (solid) and the χ^2_{random} (dashed) distributions.

5.1 Results

Different values of R (R_{gen}) were generated in the range [0.9, 1] by properly weighting three samples where the decay of $t\bar{t}$ was forced: $t\bar{t} \to WbWb$, $t\bar{t} \to WbWq$, $t\bar{t} \to WqWq$. Moreover the Normal and Random distribution are correlated, as there is a fraction of events that meets both the requirements $\chi^2_{normal} < 4$ and $\chi^2_{random} < 4$; in the error propagation such correlation has been taken into account.

The measured values of R agree within the statistical uncertainty with R_{gen} in the range R_{gen}=[0.9, 1]. The statistical uncertainty remains steady in all the range and it is σ_R(stat) = 0.11.

6. Conclusions

Two studies of feasibility of the R measurement was presented, one by using selected semi-leptonic $t\bar{t}$ events and the other by using selected di-leptonic $t\bar{t}$ events in the $e\mu$ channel. The expected uncertainties, for the semi-leptonic channel with L = 1 fb^{-1} , are σ_R(stat) = 0.12 and σ_R(sys) = 0.11. For the dileptonic channel, with L = 250 pb^{-1} , the expected uncertainty is σ_R(stat + sys) = 0.09.

Both the studies use data driven methods to subtract the background contribution.

References

[1] J. Alwall et al, *Is Vtb=1 ?*, Eur.Phys.J. C49 **791-801** (2007)

[2] CMS Collaboration, *The CMS experiment at the CERN LHC*, JINST 3 **S08004** (2008)

[3] CMS Collaboration, *Plan for a $B(t \to Wb)/B(t \to Wq)$ measurement in $t\bar{t}$ semi-leptonic decays at $\sqrt{s}$=10 TeV* CMS PAS TOP-09-007 (2009)

[4] CMS Collaboration, *Probing the heavy flavor content of the $t\bar{t}$ dilepton channel in proton proton collisions at $\sqrt{s}$ = 10 TeV* CMS PAS TOP-09-001 (2009)

[5] CMS Collaboration, *Performance of Jet Algorithms in CMS*, CMS PAS JME-07-003 (2007).

[6] CMS Collaboration, *Algorithms for b Jet identification in CMS*, CMS PAS BTV-09-001 (2009).

[7] R. Ellis, W. Stirling, and B. Webber, *QCD and Collider Physics*, Cambridge Monographs on Part. Phys., Nucl. Phys. and Cosmology.