DESY Summer Student
Programme 2011

Estimation of b-tagging performance using $t\bar{t}$ events at CMS

Author: Stefania Vitillo

Supervisor: Wolfgang Lohmann
Tutor: Igor Marfin

Abstract

The capability to detect b-quarks is a necessary prerequisite to study the Higgs decay channel $H \to b\bar{b}$. One method to calibrate b-tagging is to use a sample of pair produced top quarks. Top quarks are abundantly produced in pairs from pp collisions and can be used to isolate jet samples with a highly enriched b-jet content.

1 Introduction

1.1 Large Hadron Collider

The Large Hadron Collider (LHC) is a circular particle accelerator, hosted at the European Organisation for Nuclear Research (CERN), Geneva (Switzerland). Two proton beams circulate in opposite directions around the ring and cross at several points, which are home to large particle detectors. Fig. 1 shows a diagram of the LHC layout.

Figure 1: The layout of the LHC collider with the positions of the experiments ALICE, ATLAS, CMS and LHCb.

The two general-purpose detectors, CMS and ATLAS, sit on opposites sides of the ring, while the two smaller specialty detectors, LHCb and ALICE, sit at the interaction points to either side of ATLAS. Each beam consists of a series of proton bunches, with a maximum of 2835 "buckets" in the beam to be filled by bunches. The bunch structure allows a maximum bunch crossing rate of 40 MHz. The LHC began colliding proton beams in March 2010, quickly reaching its 2010 center of mass operating energy of 7 TeV (3.5 TeV per proton beam). The LHC maximum design energy is 14 TeV (7 TeV per beam). An other important parameter for an hadron collider like the LHC, is the luminosity $\mathcal{L}$ [1]. At the moment $\mathcal{L} \approx 2.4 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$, approacchly the design value.

1.2 Compact Muon Solenoid

A comprehensive description of the Compact Muon Solenoid (CMS) detector is given in [2]. The CMS detector is shown in Fig. 2. The hole structure is arranged concentrically around the interaction point (IP). The main feature of the CMS detector is the 3.8 T magnetic field, created by a superconducting solenoid, which allows an excellent momentum resolution for reconstructed trackss. The fully silicon-based *inner tracking system* comprises a 3-layer silicon-pixel and a 10-layer silicon-strip detector, which allow for an excellent spatial resolution close to the IP, as well as high momentum resolution. In particular, this silicon tracking device allows to perform b-jet identification, subject of this study .
The *electromagnetic calorimeter* is composed of lead-tungstate crystals and provides energy measurements for electrons and photons. A sampling *hadron calorimeter* composed of

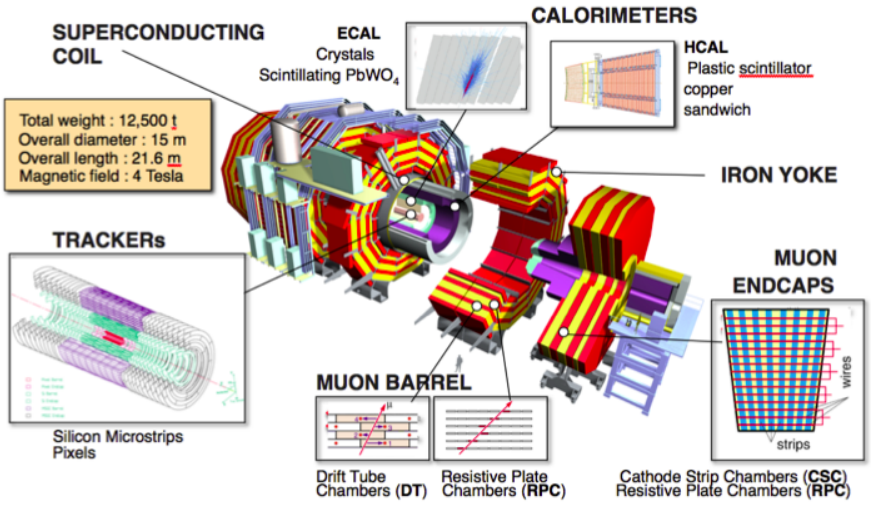


Figure 2: CMS detector.

brass absorber and plastic scintillators completes the calorimetric system, which is hosted inside the solenoid. Furthermore the iron magnet yoke is instrumented with drift chambers to measure muons.

The CMS coordinate system uses the pseudorapidity, based on the polar angle θ, which is defined as $\eta = -\ln\left(\tan(\frac{\theta}{2})\right)$.

2 Studying b-quarks

In this section we will introduce some definitions and tools needed for the study.

2.1 Jets

Jets are an observable manifestation of the quark/gluon structure of matter. When the protons collide, partons from the hard interaction fly apart and hadronize. Hadrons produce bundles, so called jets. The purpose of jet finding algorithms is to reconstruct these and combine them in a way that the kinematic properties of the found jets correspond to the kinematic properties of the partons which have been produced in the hard scattering process. Following variables are used to characterize a jet:

1. energy-momentum 4-vector:

$$p^J = (E^J, \vec{p}^J) = \sum_i (E^i, p_x^i, p_y^i, p_z^i)$$

   where the index i runs over all particles assigned to the jet.

2. transverse momentum:

$$p_T^J = \sqrt{(p_x^J)^2 + (p_y^J)^2}$$

3. rapidity:

$$y_T^J = \frac{1}{2} \ln \frac{E^J + p_z^J}{E^J - p_z^J}$$

4. azimuthal angle:

$$\phi^J = \arctan \frac{p_y^J}{p_x^J}$$

where J denotes a certain jet. There are different clustering algorithms available at CMS for jet reconstruction [4]:

- **Cone Algorithms:** this algorithms starts to look for particles within a cone of radius R in the (η, ϕ) space. All particles i which fulfil

$$\sqrt{\Delta\eta^2 + \Delta\phi^2} \leq \Delta R$$

  are clustered together.

- **Clustering Algorithms:** the k_T algorithm, the Cambrige/Aachen and the Anti-k_T algorithm. Each of this algorithms combines the input four vector af two particles, i and j, pair-wise according to the distance of the two particles d_{ij} and their individual distance to the beam d_{iB}. For all input particles the smallest distance among d_{ij} and d_{iB} is determined. If $d_{ij} < d_{iB}$ the two particles are recombined, adding their four momenta, otherwise particle i is removed from the list of particles and considered for another jet. The procedure is repeated until all particle are clustered in jets.

2.2 B-tagging

2.2.1 Theoretical reasons and practical difficulties

The presence of b quarks in many interesting physics channels makes it desirable to have efficient labelling or *tagging* of reconstructed b jets. The potential Higgs discovery channel $H \to b\bar{b}$ is a good example for an analysis which strongly depends on an excellent b-tagging performance as it can be seen in Fig. 3. In the Standard Model, for a Higgs boson mass below 130 GeV, the Higgs decays predominantly in b-quarks. In models beyond the Standard Model, this final state may be even produced with larger cross setion, such that a discovery is possible.

Higgs production with associated b-quarks is expected in supersimmetric models. The final state consists of 4 b-jets (Fig. 4) which need to be identified to be able to reconstruct the Higgs boson, where, at the same time, an efficient rejection of light jets in the background channels is required.
Due to the relatively long lifetime of hadrons with b quarks as their constituents, **jets stemming from a b quark have a different topology** than jets from light quarks. Due to a lifetime of 1.5 ps, the B hadron can travel several millimeters before it decays creating a signature of a jet with a reconstructable secondary vertex (as illustrated in Fig. 5). To be sensitive to these topological differences between b jets and light jets, a high-precision vertex detector like the CMS silicon tracker, which reconstructs tracks and their impact parameters with very high resolution, is necessary. These tracks are along with the calorimeter based jets the basic ingredients for b-tagging. They are used to reconstruct the primary vertex, to calculate the impact parameters with respect to the primary vertex and to fit secondary vertices. These are the key quantities for efficient b-tagging.

2.3 Simulation and Event Reconstruction

An event is the result of a single proton-proton collision inside the CMS detector or a simulated collision event. **Monte Carlo generators** are used to simulate events. The

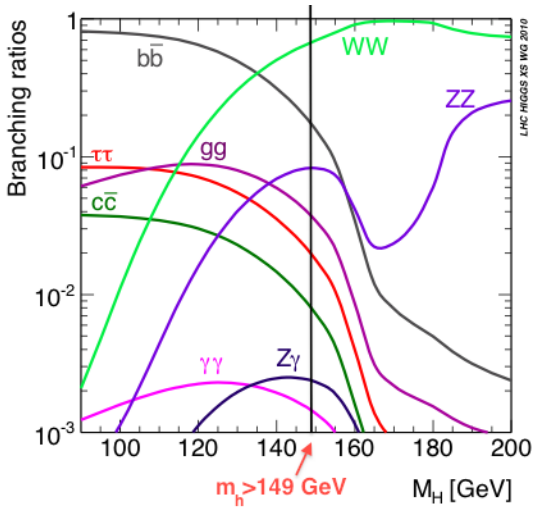


Figure 3: Branching fractions of the Higgs boson in several decay channels as a function of the Higgs boson mass in the electroweak Standard Model. The current lower limit for the Higgs boson in the SM is 114 GeV.

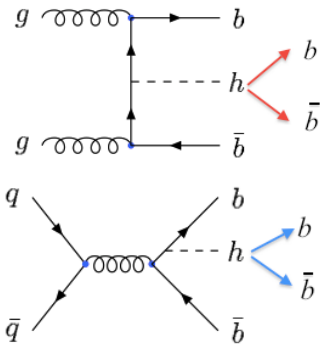


Figure 4: Diagrams for the production of a Higgs boson in pp interactions with associated b-quarks

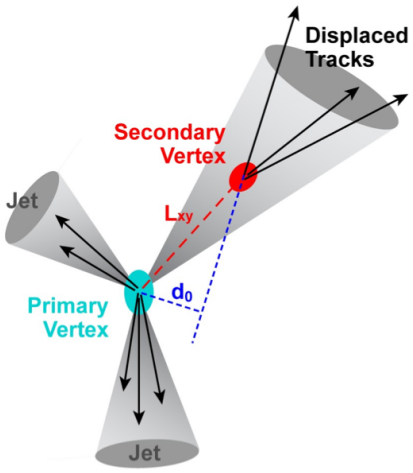


Figure 5: The impact parameter d_0 of a particle originating from a B hadron decay.

calculation of the production process originating from an inelastic proton-proton scattering and the subsequent creation of stable quarks and leptons, is called the *generation* step. In the second step, quark hadronise and form hadrons of different lifetimes that may decay inside the detector. The response of the CMS detector is simulated in the *simulation* step. Leptons and hadrons are then transported through the detector and energy depositions inside are converted into a data format, which is similar to the real detector output, and referred to as the *digitisation* step. The data format is committed to the *reconstruction* step. The so obtained depositions in all subdetectors are used to fully reconstruct the event, i.e. tracks of all charged particles and depositions of charged and neutral particles in the calorimeters. The aim of the **event reconstruction** is to determine the detector response, reconstruction efficiencies, as well as the resolution of reconstructed objects (Fig. 6).

Different Monte Carlo (MC) generators are used in order to simulate signal and back-

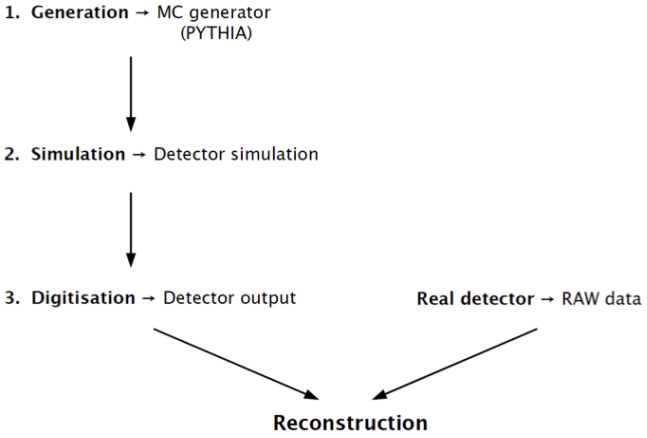


Figure 6: The generation and reconstruction of a Monte Carlo simulated event.

ground processes for physics analysis. For the generation of simulated event samples, which are used for this study, the event generator PYTHIA6 is used [3].

2.3.1 Generation of $Z \to b\bar{b}$ events

Using PYTHIA6, as an exercise, $Z \to b\bar{b}$ events were generated, simulated and reconstructed. The Z boson was reconstructed trough the invariant mass of two quarks (whose p_T andη distributions are plotted in Fig.7) in the final state (as shown in Fig.8):

$$M_Z^2 = \left(\sum_{b-quarks} E \right)^2 - \| \sum_{b-quarks} p \|^2 \tag{1}$$

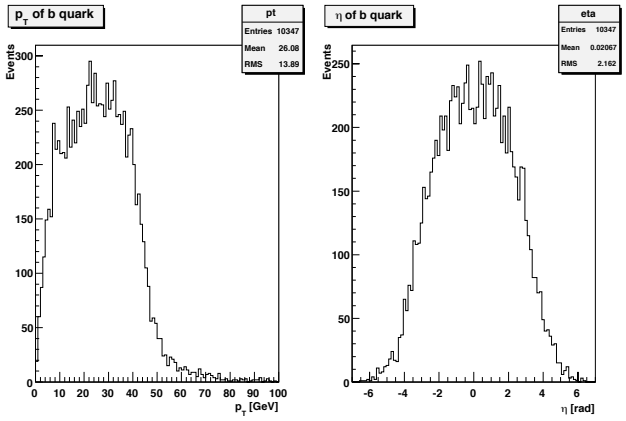


Figure 7: Distributions of the transverse momentum and the pseudorapidity for b quarks originated from Z -boson decays.

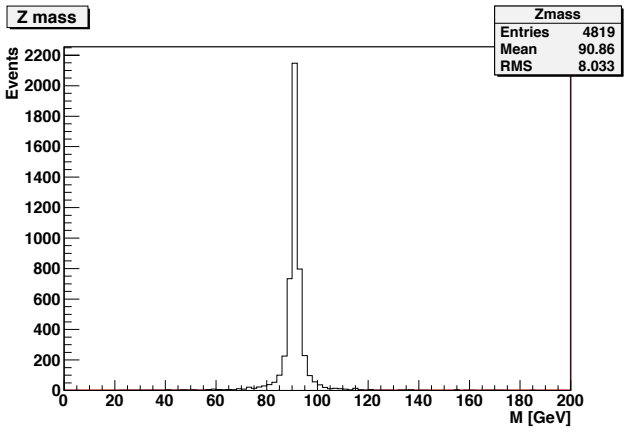


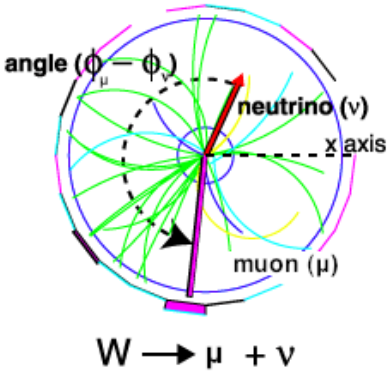
Figure 8: The invariant mass of the two b-quarks originated from the Z-boson (m_Z = 91.18 GeV[5]).

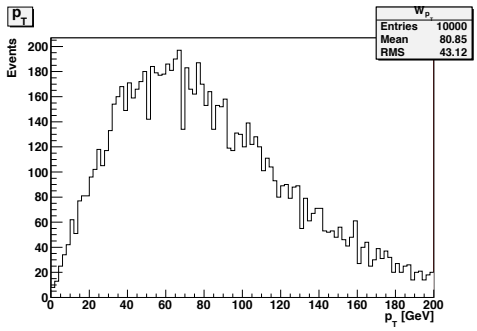
2.3.2 Generation of $pp \to t\bar{t}$

From the generation of pairs of top quark that are known to decay with BR[1] $\approx 100\%$ to $W^{\pm}b$, we selected among all the possible W decay modes, the decay $W \to \mu\nu_\mu$ and represented p_T and η distributions of the selected W bosons (shown in Fig.9)

Then from the decay products, as exercise, we reconstructed the W transverse mass (as shown in Fig. 10) using the formula:

$$M_T^W = \sqrt{2E_T^{\mu}\ E_T{}^{\nu}(1 - \cos\Delta\phi_{\mu\nu})}.$$





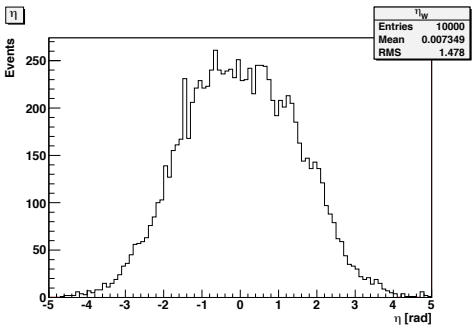


Figure 9: Distributions of the transverse momentum and the pseudorapidity for W bosons coming from the decay $t \to Wb$.

[1]The branching fraction (BR) for a decay is the fraction of particles which decay by an individual decay mode with respect to the total number of particles which decay.

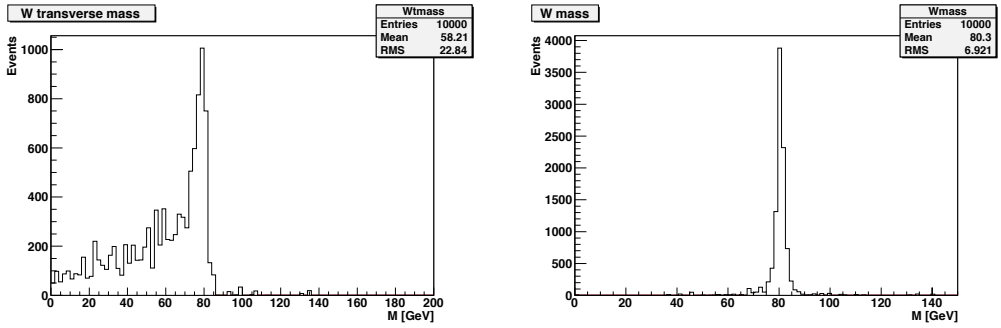


Figure 10: Reconstructed W transverse mass (on the left) and invariant mass (right, from formula 1). ($m_W = 80.39$ GeV [5])

3 B-tag calibration

Various b-tagging algorithms have been implemented in CMS. However the efficiencies of these algorithms will have to be calibrated while the detector takes data. One way to calculate this efficiency is using the formula:

$$\varepsilon_b = \frac{x_{tag} - \varepsilon_0(1 - x_b)}{x_b} \tag{2}$$

where:

- $x_b = \frac{b_{jets}}{all_{jets}}$ the fraction of all b-jets among all jets from MC samples;
- $x_{tag} = \frac{tag_{jets}}{all_{jets}}$ is the fraction of b-tagged jets selected from Data samples;
- ε_0 is the mistag rate estimation, that is not precisely determined from Data/MC.

Trying to improve x_b will be our goal, to gain a b-tag efficiency that will be independent from ε_0: in fact if $x_b \to 1 \Rightarrow$ it is possible to suppress from Eq.2 the dependence on the systematic uncertainties coming from the estimation of the mistag rate ε_0.

In this section we will go trough one of the calibration methods.

3.1 Top-quark based method [6]

With a branching fraction $BR(t \to Wb) \approx 100\%$ [5], $t\bar{t}$ events can be used to isolate jet samples with a highly enriched b-jet content. These jet samples can then be used to calibrate b-jet identification algorithms.

We will study the semimuonic $t\bar{t}$ decay with final state $b\bar{b}qq'\mu\nu_\mu$, where q and q' are light quarks (shown in Fig. 11). The experimental signature of the final state is: a muon, four jets and missing transverse momentum (of the neutrino, that can't be detect directly). This will be the signal that we will search in MC and Data samples.

3.1.1 Preselection of b-jet enriched sample from semimuonic $t\bar{t}$ events

A b-jet enriched sample from $t\bar{t}$ events, is obtained by applying the following criteria to the events requesting :

- at least 4 jets with high p_T in the central region of the detector;

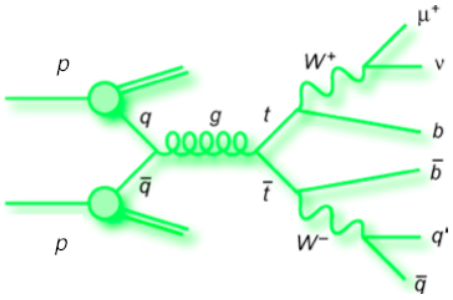


Figure 11: $t\bar{t}$ decay diagram. One W boson decays into a muon and a neutrino and the other couples to a weak quark doublet.

- at least one isolated muon with high p_T in the central region of the detector;
- no isolated high p_T electrons in the central region of the detector.

3.1.2 Preselection on MC and Data samples

On MC samples, chosen to cover all signal and background processes that we are looking for (shown in Fig.11), it is possible to apply a preselection of the events requested for the searched final state with a High Level Trigger.[2]

Each type of event is produced with a specific cross section σ. So, knowing the total number of generated events (N_{tot} in Tab.1), it is possible to calculate the integrated luminosity of each MC sample with the following formula:

$$L = \int \mathcal{L} dt = \frac{N_{tot}^{evt}}{\sigma} \tag{3}$$

The integrated luminosity of each sample will be used when we will compare all signal and background events from the MC samples to the Data[3]. In that case we will use the rescaling factor between them and given by:

$$R = \frac{L_{Data}}{L_{MC}} \tag{4}$$

in which the integrated luminosity from Data recorded by the HLT trigger is 324pb^{-1}.

We applyied the following selection on both Data and MC samples, requiring events with at least one muon with: $|\eta| < 2.1$ and $p_T > 20\,\text{GeV}$, and at least 4 jets with: $|\eta| < 2.4$ and $p_T > 30\,\text{GeV}$.

The selected events are listed in Tab.1. After this preselection the ratio between signal and background to the process that we are looking for is:

$$\frac{S}{B} = 1.72 \pm 0.6. \tag{5}$$

To compare graphically MC samples of both signal and background to Data, we represented in the same histogram a kinematical variable for Data events and the same kinematical variable for the sum of all MC events. We did it for many kinematical variables like:

[2] In this case we used $HLT\_IsoMu17\_v5$, $HLT\_IsoMu17\_v6$, and $HLT\_IsoMu17\_v7$

[3] New summer 2011 data: $/SingleMu/Run2011A-PromptReco-v4/AOD$ for runs in the range [160404, 167151]with L=869pb^{-1}.

Process	$N_{tot}^{(evt)}$	σ(pb)	N_{evt} after preselection[4]	$\varepsilon_{sel} = \frac{N_{evt}^{selected}}{N_{tot}}$
$t\bar{t} \to \mu+$ jets	4522 ± 276	14.0 ± 0.1	280 ± 17	0.011 ± 0.001
$t\bar{t} \to$ Non($\mu+$ jets)	26101 ± 1588	80.6 ± 0.8	1030 ± 60	0.23 ± 0.02
QCDMuEnriched	$(96 \pm 6) \cdot 10^9$	$(297 \pm 3) \cdot 10^6$	$(12.3 \pm 3.5) \cdot 10^4$	$(13 \pm 4) \cdot 10^{-7}$
$W \to \mu\nu_\mu+$ jets	2559276 ± 155683	7899 ± 79	695 ± 46	$(27 \pm 2) \cdot 10^{-5}$
$W \to \tau\nu_\mu+$ jets	2559276 ± 155683	7899 ± 79	51 ± 6	$(20 \pm 3) \cdot 10^{-6}$
$Z \to \mu\mu+$ jets	797747 ± 48533	2462 ± 25	196 ± 15	$(25 \pm 2) \cdot 10^{-5}$
Zbb	4517 ± 275	13.9 ± 0.1	9.9 ± 0.6	$(22 \pm 2) \cdot 10^{-4}$
Single t (s)	758 ± 46	2.34 ± 0.02	2.1 ± 0.2	$(28 \pm 3) \cdot 10^{-4}$
Single t (t)	2412 ± 147	7.44 ± 0.07	51 ± 3	0.021 ± 0.002
Single t (tw)	2302 ± 140	7.1 ± 0.1	42 ± 2	0.18 ± 0.002
Drell-Yan $\to \mu\mu$	421200 ± 25624	1300 ± 13	66 ± 5	$(16 \pm 2) \cdot 10^{-5}$
Drell-Yan $\to \tau\tau$	421200 ± 25624	1300 ± 13	33 ± 3	$(8 \pm 1) \cdot 10^{-5}$
Data	1877395 ± 1370	–	2497 ± 50	
$\sum N_{evt}^{selected}$ from MC			2456 ± 82	

Table 1: Selection of events containing at least one muon and at least four tight jets, on both MC and Data samples. We will exclude *QCDMuEnriched* events because of the small preselection efficiency. Selected events from Data and MC are comparable. [Work in Progress]

- $\sum_i E_{Ti}$ [GeV]: sum of all energy deposits in the transverse plane (that should be $\neq 0$ if a neutrino is expected, and this is the case);
- E_T [GeV] of the first jet;
- $\frac{\min M(j_m, j_n)}{\sum_k M(j_k)}$: fraction between the lower massive jet in respect of the total jet mass sum;
- MET [GeV]: the missing transverse energy (of the neutrino), given by the the difference between the total transverse energy deposit in the calorimeter and $\sum_i E_{Ti}$;
- $\Delta(\Phi_{MET}, \Phi_\mu)$: difference in polar angle Φ between the neutrino and the muon.
- $|\eta_\mu|$: absolute value of the muon pseudorapidity;
- $\Delta\eta(j_2, j_3)$: difference in pseudorapidity between two jets, labelled as 2 and 3.

These kinematical distributions and the ratio between Data and the sum of all MC samples are shown in Fig. 12 and 13. There is not a perfect agreement in the Data/MC ratio for control plots of jet energy based kinematical variables (Fig.12) due to pileup [9] in events and different resolution for Data events despite to MC events.

3.2 MVA performance

Using TMVA[8] (Toolkit for Multivariate Data Analysis with ROOT), we tried to improve the ratio $\frac{S}{B}$ (shown in Eq.5) in the preselected events, selecting events enriched by signal. To do this we used the MVA likelihood ratio method (that can be found in more detail in section 8.2 of [8]), that is here briefly described.

[4] after HLT trigger and offline preselection

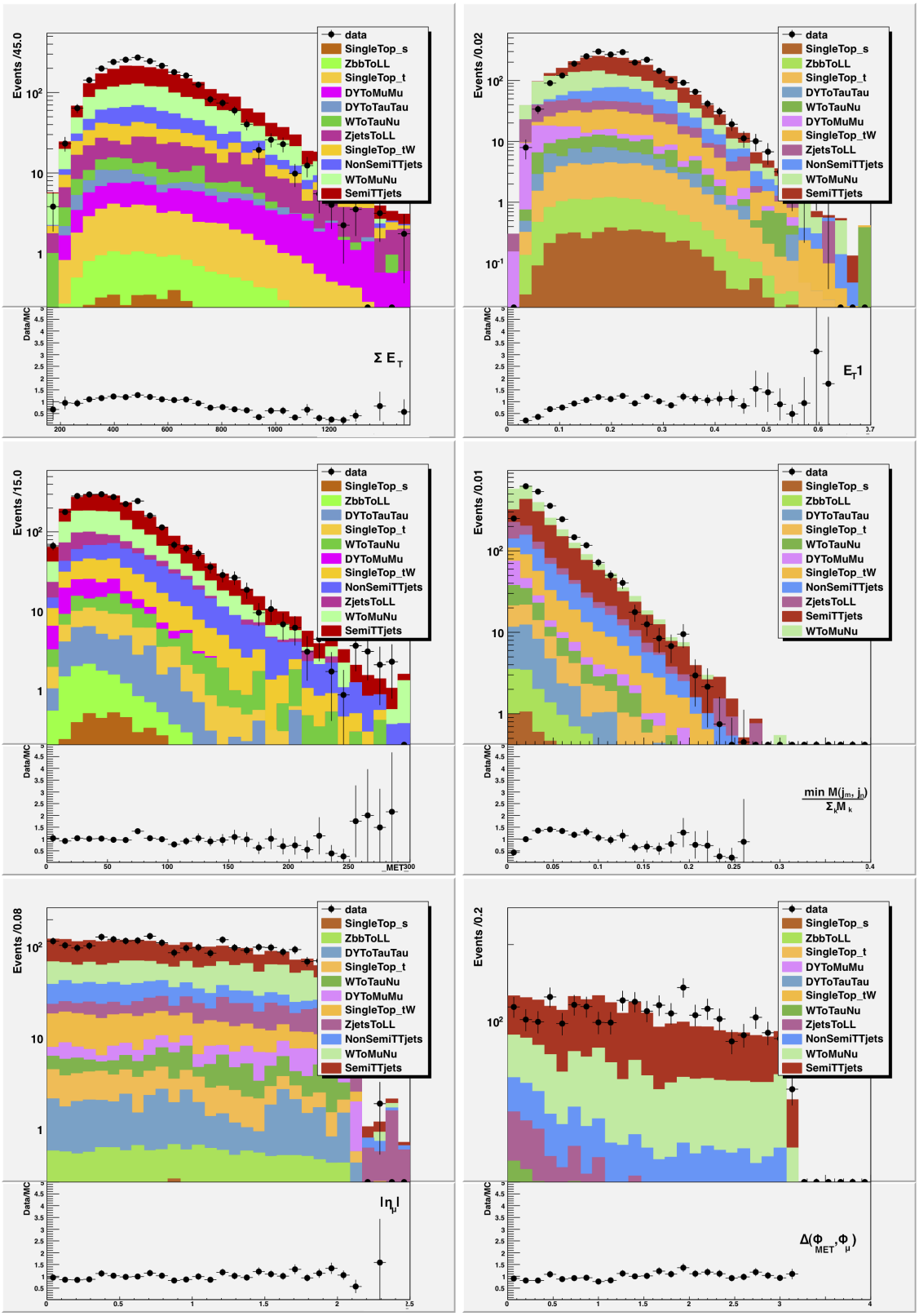


Figure 12: Control plots: kinematical distributions comparison between Data and MC samples after preselection of the events.

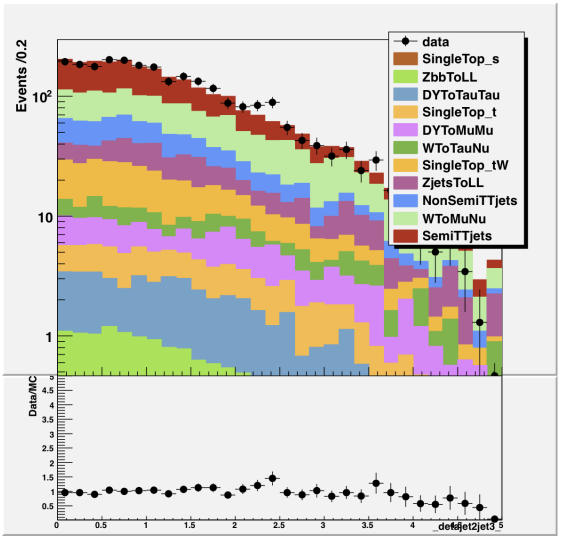


Figure 13: Kinematical distributions comparison between Data and MC samples after preselection of the events. [Work in Progress]

3.2.1 Likelihood Ratio Method

For a given event, the likelihood for being of signal type is obtained by multiplying the signal probability densities of all input variables, which are assumed to be independent, and normalising this by the sum of the signal and background likelihoods. The likelihood ratio $y_{\mathcal{L}}(i)$ for event i is defined by:

$$y_{\mathcal{L}}(i) = \frac{\mathcal{L}_S(i)}{\mathcal{L}_S(i) + \mathcal{L}_B(i)} \tag{6}$$

where

$$\mathcal{L}_{S(B)}(i) = \prod_{k}^{N_{var}} p_{S(B),k}(x_k(i))$$

and where $p_{S(B),k}$ is the normalized to 1 signal (background) probability density function (PDF) for the k_{th} input variable x_k [8] (in our case the input variables are the different kinematical variables that we choosed).
It can be shown that in absence of model inaccuracies (such as correlations between input variables), the ratio in Eq.6 provides optimal signal from background separation for the given set of input variables.
We decided to apply this method to our selected MC events, building the PDF's of the kinematical variables that we studied. So, first we checked the correlation matrixes between the input variables and choosed only that kinematical input variables that were almost uncorrelated (see the correlation matrixes for signal and background events in (Fig. 14)).

The likelihood distribution for the selected MC events, that contain both signal and background events to the $t\bar{t}$ semimuonic decay channel, is shown in Fig.16. Looking at this figure (Fig.16) we decided that the optimal value of likelihood response cut was around $y_{\mathcal{L}} \approx 0.81$, because a cut of all the events that have $y_{\mathcal{L}} \leq 0.81$, would reduce not only signal events, but also a great amount of background events. And in this case that cut in background events is larger than that for signal. The result would be an increased $\frac{S}{B}$ ratio.

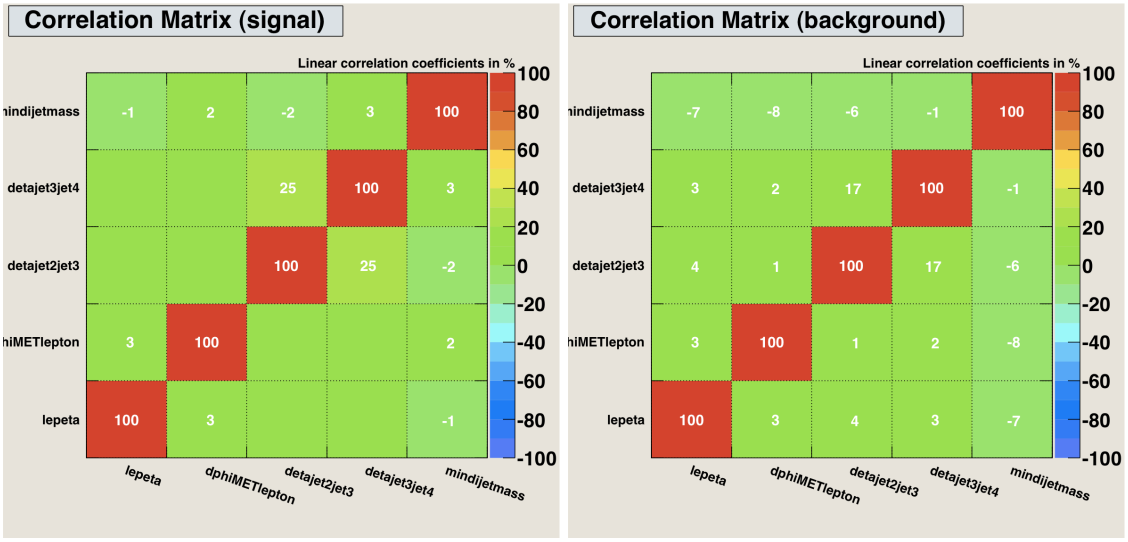


Figure 14: Correlation matrix for the chosen input variables for signal and background. [Work in Progress]

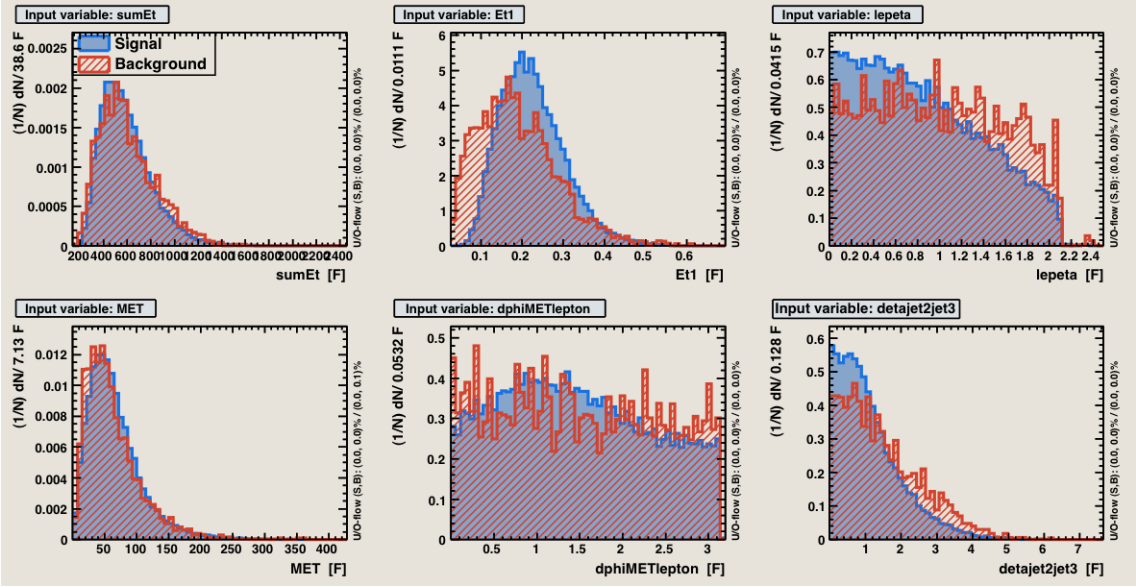


Figure 15: Input distributions used to create PDF's of kinematical variables for signal and background. In order to improve efficiency of the method (likelihood ratio method) we have to find other kinematical variables. [Work in Progress]

The effect of applying MVA selection to the events is shown in Tab.2. In this way we improved the signal to background ratio:

$$\frac{S}{B} \approx 2.38$$

that compared to the result after the offline preselection (shown in Eq.5), provides a 138% gain in purity of slection. We made control plots for events passing MVA likelihood discrimination, to compare kinematical distributions for Data and MC (see Fig.17).

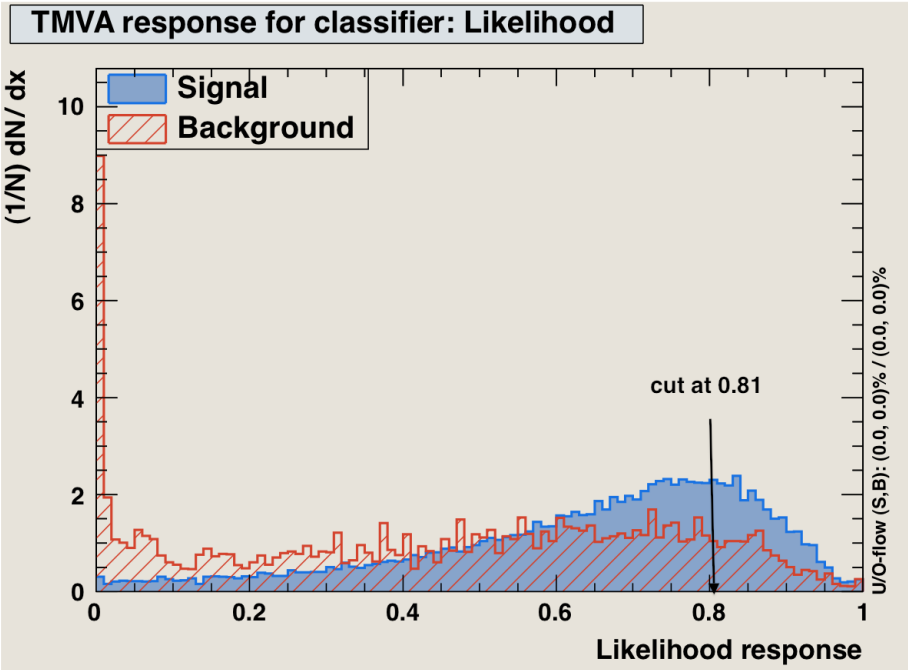


Figure 16: Likelihood distribution ($y_{\mathcal{L}}$) for signal and background and choice of the optimal cut value. [Work in Progress]

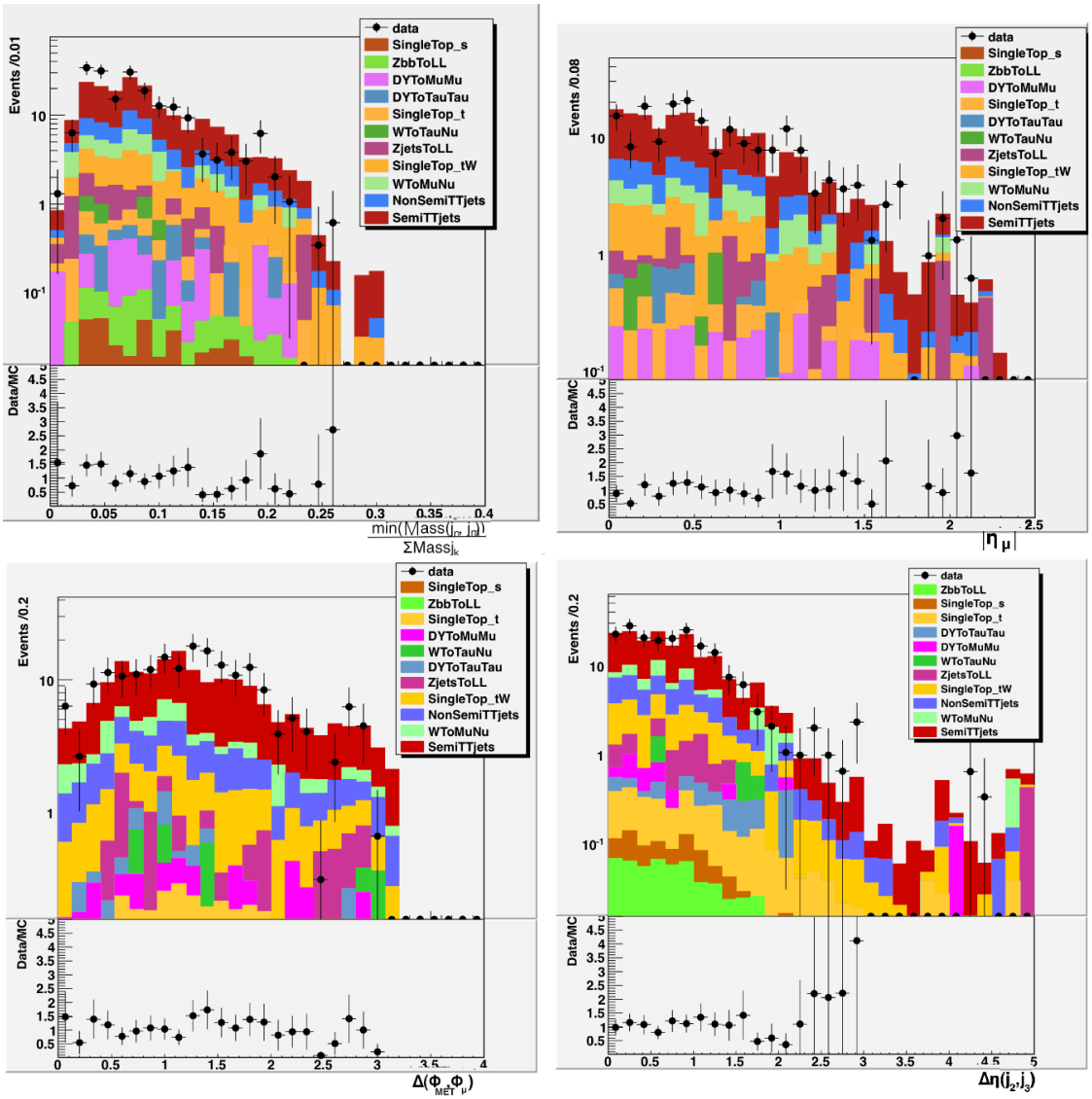


Figure 17: Comparison of kinematic distributions between Data and and MC for various variables, after MVA selection. The agreement in Data/MC ratio is improved for low energy ranges, in comparison to the same distributions before MVA selection (compare it to Fig.12 and 13). [Work in Progress]

Process	N_{tot}^{evt}	N_{sel} after MVA selection
$t\bar{t} \to \mu+$ jets	4522 ± 276	107.77
$t\bar{t} \to$ Non($\mu+$ jets)	26101 ± 1588	26.61
$W \to \mu\nu_\mu+$ jets	2559276 ± 155683	17.11
$W \to \tau\nu_\mu+$ jets	2559276 ± 155683	1.58
$Z \to \mu\mu+$ jets	797747 ± 48533	6.61
Zbb	4517 ± 275	0.59
Single t (s)	758 ± 46	0.41
Single t (t)	2412 ± 147	2.64
Single t (tw)	2302 ± 140	18.87
Drell-Yan $\to \mu\mu$	42100 ± 25624	2.06
Drell-Yan $\to \tau\tau$	421200 ± 25624	1.88
Data	1877395 ± 1370	196
$\sum(N_{evt}^{selected})$ from MC		186.11

Table 2: MVA selection results. [Work in Progress]

4 Conclusions

After a brief introduction to the LHC and to the CMS experiment, we passed through the layout of the variables, tools and theoretical background needed for b-quark studying. After getting practice with the CMS framework software, the first done exercises consisted in Monte Carlo simulations of $pp \to Z \to b\bar{b}$ and $pp \to t\bar{t}$ (selecting the final state of $t\bar{t} \to W^+W^- \to \mu^+\mu-\nu_\mu\bar{\nu}_\mu$), representing some kinematical distributions of these processes. At that point we were able to start with the main goal of this study: find a strategy to estimate b-tag efficiency. We decide to proceed with one of the known methods, the "top quark method". Due to the high production rate of $t\bar{t}$ events at LHC, that are known to decay with BR$\approx$ 100% in b quarks (BR($t \to bW$) $\approx$ 100%), b-enriched jet samples can be selected through the semimuonic $t\bar{t}$ decay channel. From MC samples and Data, we selected events containing at least one tight muon and at least four tight jets and compared them with some control plots containing kinematical distributions of the events. We used the likelihood ratio based method to increase the purity of the selection to have b-enriched samples in the end. The next step for this analysis will be the estimation of b-tagging efficiency.

Acknowledgement

I would like to thank Igor Marfin for teaching me patiently all he could in this few weeks about his work. I wish to thank Wolfgang Lohmann for his advises and corrections to this report. Many thanks to my group colleague and friend Tomek Wojtoń for the constructive discussions about our work. I would like to thank Karl Jansen for giving me the possibility to join this summer student program. Finally, but not less important, I warmly wish to thank Laura Torino and Lucia Ambrogi for spending and enjoying our first german summer together.

References

[1] Particle data group *Accelerator Physics of Colliders* *http://pdg.lbl.gov/2011/reviews/rpp2011-rev-accel-phys-colliders.pdf*

[2] CMS Collaboration, *CMS TriDAS project: Technical Design Report; 1, The Trigger Systems.* Technical Design Report CMS, CERN, 2000.

[3] Torbjorn Sjostrand, Stephen Mrenna, and Peter Skands *PYTHIA 6.4 Physics and Manual* hep-ph/0603175 LU TP 06-13 FERMILAB-PUB-06-052-CD-T March 2006

[4] V. Chetluru, F. Pandolfi, P. Schieferdecker, and M. Zielinski, *et Reconstruction Performance at CMS,* CMS AN-2009/067

[5] Particle data group *Particle Listings http* : *//pdg.lbl.gov*/2011/*listings/contents_listings.html*

[6] S. Lowette, J. D'Hondt, J. Heyninck, P. Vanlaer *Offline Calibration of b-Jet Identification Efficiencies* CMS NOTE 2006/013

[7] O. van der Aa, C.Delaere *The High Level Trigger software for the CMS experiment*

[8] TMVA Toolkit for Multivariate Data Analysis with ROOT *http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf*

[9] H. Sakulin *Methodology for the Simulation of Single-Muon and Di-Muon Trigger Rates at the CMS Experiment in the Presence of Pile-Up* CMS Note 2002/042

.