PROCEEDINGS OF SCIENCE

QCD backgrounds in charged Higgs boson searches

Alexandros Attikis[*†]
University of Cyprus
E-mail: Alexandros.Attikis@cern.ch

The search for a light charged Higgs boson at the LHC will greatly depend on how well the QCD background can be controlled, especially for the fully hadronic final state where it is expected that QCD is overwhelmingly the dominant source of background. As the QCD production cross-section at the LHC is large and more importantly poorly known, specific methods to suppress and consequently predict the remaining QCD multi-jet background are required. Several methods of suppressing this QCD background and data-driven methods to estimate the remaining in the signal region are discussed for both the lepton+jets and fully hadronic final state signal channels.

Third International Workshop on Prospects for Charged Higgs Discovery at Colliders - CHARGED2010, September 27-30, 2010
Uppsala Sweden

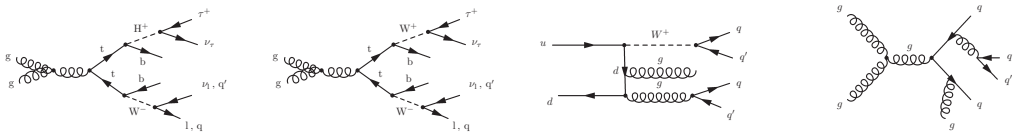
[*]Speaker.
[†]on behalf of the CMS Collaboration.

PoS(Charged 2010)012

1. Introduction

In the Minimal Supersymmetric Standard Model (MSSM) [1 – 3] there are five Higgs scalar mass eigenstates, consisting of a light and a heavy CP-even neutral scalars h^0 and H^0, one CP-odd neutral scalar A^0, and a charged scalar H^+ with its charge conjugate analogue H^-. The Large Hadron Collider (LHC) is expected to provide access to the majority of the MSSM parameter space, and the CMS [4] experiment alone has already recorded proton-proton collision data at 7 TeV which amount to 43.17 pb^{-1} of integrated luminosity. The dominant production mechanism of light charged Higgs bosons ($m_{H^\pm} < m_t$) at the LHC is through gluon fusion in $t\bar{t}$ events, with the $t \to bH^\pm$ decay. Production of heavy charged Higgs bosons ($m_{H^\pm} < m_t$) proceeds mainly through the associated production processes $gb \to tH^\pm$ and $gg \to tbH^\pm$. The branching ratio of top decay to charged Higgs boson depends on both $m_{H^\pm}$ and $\tan\beta$. For $m_{H^\pm} < m_t$ the charged Higgs boson decays almost exclusively through the $H^\pm \to \tau^\pm \nu_\tau$ decay mode with $BR(H^\pm \to \tau^\pm \nu_\tau) \approx 1$, while for $m_{H^\pm} > m_t$ the aforementioned mode is the sub-dominant one, the dominant being the $H^\pm \to tb$ decay mode instead. For both charged Higgs boson mass regions however, the most promising discovery channel is the $H^\pm \to \tau^\pm \nu_\tau$ decay mode [5]. Henceforth, only channels involving this decay mode with the tau lepton decaying hadronically will be considered.

For $m_{H^\pm} < m_t$, there are two different final states for $t\bar{t} \to bW^\mp \bar{b} H^\pm$ events, depending on the $W^\mp$ decay to leptons or jets, as shown in Fig. 1(a). The mode in which the $W^\mp$ decays leptonically has a final state with $(l\nu_l)(\tau\nu_\tau)b\bar{b}$ and is referred to as lepton+jets, whereas the mode in which the $W^\mp$ decays hadronically is referred to as the fully hadronic final state with $\left(q\bar{q}'\right)(\tau\nu_\tau)b\bar{b}$. The $t\bar{t}$ and $W+3/4$ jet processes, shown in Fig. 1(b) and 1(c) respectively, can contribute to the background through genuine τ's originating from $W^\pm \to \tau^\pm \nu_\tau$ decays. These events can be suppressed with a high E_T cut and by exploiting the opposite state of the τ polarisation resulting from the scalar charged Higgs boson and the vector W boson [7]. For the lepton+jets final state, the main source of background comes from processes containing electrons/muons, τ leptons and jets in the final state. The dominant background for this final state is $t\bar{t}$ events with one or more



(a) (b) (c) (d)

Figure 1: Feynman diagrams for (a) the production of a charged Higgs boson via gluon fusion which is the dominant production process at the LHC for $m_{H^\pm} < m_t$, (b) $t\bar{t}$ background events to $H^\pm$ searches, (c) W+jets background events to $H^\pm$ searches, (d) QCD multi-jet background events to $H^\pm$ searches.

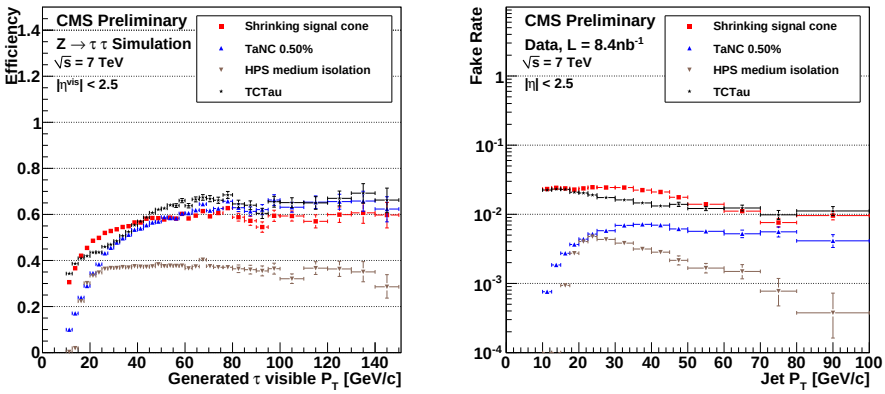
tau leptons in the final state. The QCD multi-jet background events, which mainly produce non-isolated leptons in semi-leptonic decays are less important for this channel, due to the presence of isolated leptons which can be used to efficiently reject this QCD multi-jet background. Contrary to the leptons+jets final state however, the background for the fully hadronic final state is dominated by QCD multi-jet events, while $t\bar{t}$ remains a significant background. The QCD multi-jet events can fall to the signal area due to the mis-identification of a hadronic jet as a tau-jet and due to

uncertainties in the Missing Transverse Energy (MET) measurement, which can result primarily from jet resolution effects. Such background events can be suppressed with a tau-identification algorithm based on an efficient tracker isolation.

However, besides the fake tau-jet[1] rates and the fake MET, QCD multi-jet events have a large and poorly known cross-section and are thus a very dangerous background. Therefore, QCD must not only be adequately suppressed, but also its contribution to the signal region must be accurately estimated. To achieve this, data-driven methods must be developed and their effectiveness demonstrated by use of collision data. These proceedings concentrate on data-driven methods for estimating the number of expected QCD multi-jet background events in the light charged Higgs boson searches.

2. Tau-jet fake-rate with early data

Proton-proton collision events collected with the CMS experiment at LHC at $\sqrt{s}$= 7 TeV in 2010 have been used to commission the algorithms for reconstruction and identification of tau lepton hadronic decays [8]. The efficiency with which genuine tau lepton hadronic decays are identified was estimated from MC with a sample of simulated $Z \to \tau^+\tau^-$ events, as the number of tau leptons present in the collision data analysed was insufficient for such measurement (Fig. 2(a)). It was possible however to determine the probabilities with which generic QCD jets pass the selection criteria of several tau identification algorithms (fake-rates), with collision data (Fig. 2(b)). The measured fake-rates were compared to MC predictions and were found to be systematically under-estimated by the latter, with preliminary studies suggesting that the modeling of hadronisation processes in the MC could be the reason for these discrepancies.



(a) (b)

Figure 2: (a) Efficiencies of the tau algorithms to identify genuine tau lepton hadronic decays as function of p_T at generator level and using simulated $Z \to \tau^+\tau^-$ events. (b) Measured probabilities of quark/gluon jets to pass the tau identification of the algorithms as a function of jet p_T[8].

[1] A tau-jet is a hadronically decaying tau lepton.

3. Data-driven methods for estimating QCD multi-jet background

Generic QCD backgrounds are significant in SUSY and $t\bar{t}$ studies due to jets faking electrons (e) in the e+jets final state and muons (μ) in hadronic jets from heavy flavours in the μ+jets final state. For this reason, several data-driven methods have been developed in CMS, to estimate the QCD background contribution in the signal regions, including the fake-rate application methods [9], extrapolation methods using relative isolation [10] and the kinematical variable α_T[11], and fits to discriminating variables using template fit methods [9]. The fake-rate, the extrapolation and template methods have been tested for the selection of top-like events in the di-lepton and lepton+jets channels [10]. Given the similarities in the final state topology of this channel and the lepton+jets final state in light charged Higgs boson searches, methods studied for the former can also be used for the latter. However, this is not the case for the fully hadronic final state, as the difference in final state topologies and background composition suggests that if any such methods can be used in QCD background estimation measurements, it would require variations in the technique. All the aforementioned methods for measuring QCD backgrounds are discussed in detail in the following sub-sections and their relevance to searches for light charged Higgs bosons is established.

3.1 Fake-rate application method

Tau identification fake-rates have been used in $t\bar{t}$ studies to estimate the W+$\geq$3 jets background in the lepton+jets final state, by use of the fake-rate application method [9]. In order to estimate this background from data, the probability for a jet to pass the tau-identification was evaluated from all jets in a jet-dominated sample and parameterized as a function of jet p_T. More specifically, the fake-rate was calculated from the ratio of the p_T distribution of the jets satisfying the tau-tagging algorithm divided by the inclusive jet p_T distribution. By applying this fake probability to the p_T spectrum of all jets in the events passing the event selection in a multi-jet sample, an event weight can be derived and by summing over all such event weights, an estimate on the number of the expected background or fake events can be obtained. The results obtained from multi-jet samples were compared to results obtained with photon+jet samples and their average was found to be approximately 20% away from MC expectations. The method was found to be strongly dependent on the jet-selection and the sample composition, while separate results obtained by grouping jets into different categories were found to have up to 25% discrepancies, an indication of the biases introduced.

For the fully hadronic final state, the fake-rate application method can also be used to determine QCD backgrounds in the signal region, provided that an appropriate High Level Trigger (HLT) and event selection are chosen. It is anticipated that after the signal HLT[2], the resulting data sample is still QCD dominated, with the level of contamination from $t\bar{t}$ and W+jets assumed to be negligible. By selecting only offline tau-jet candidates that match the HLT tau object and applying the fake-rate to them, a weight can be derived for each event. Then, by defining a QCD-efficiency factor which describes the fraction of QCD events surviving all selection cuts apart from tau-identification, the number of estimated QCD events in the signal region can be derived by applying this factor to the sum of event weights. This QCD-efficiency factor can be determined from

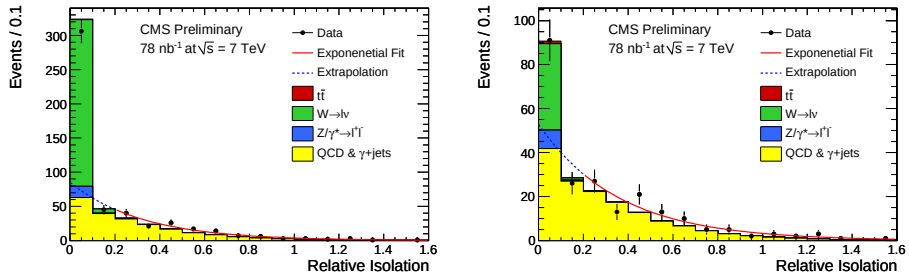
[2]Currently set to HLT_IsoTau20_Trk15 + HLT_MET20, seeded by the level-1 single tau trigger (L1SingleTau20).

data, by applying the signal trigger and event selection but with the tau-identification factorised out. The systematic uncertainties related to this method are dominated by the purity of the QCD sample and the statistics of the QCD events available for the measurement. The possible bias introduced by the assumption that the contamination from $t\bar{t}$ and W+jets is negligible can be checked with MC simulations. Another possibility is to select a QCD enriched sample with a single jet trigger (e.g. HLT_Jet30U) and then apply an anti-isolation cut on the tau candidates, thereby increasing the purity of the resulting QCD enriched sample. After applying an event selection that emulates the signal final state topology, the efficiency with which QCD events pass the selection can be determined, while the Electro-Weak contamination can again be checked with MC. This efficiency factor could then by applied to the signal trigger sample to estimate the spill of QCD multi-jet events in the signal region after the full signal selection.

3.2 Extrapolation methods using Relative Isolation

Muons and electrons originating from W-boson decays are commonly consistent with originating from the interaction point (IP), are typically isolated from the rest of the event activity and have small impact parameters. Given that non-isolated muons originate mostly from QCD generic jets, one can define a variable to quantify the degree of isolation of a lepton. The relative isolation variable for a lepton ($\text{isol}^{\text{l}}_{\text{rel}}$) is defined as the sum of the track momenta and electromagnetic and hadronic calorimeter energy deposits around the lepton track in a cone of ΔR=0.3 (excluding the contribution from the lepton candidate itself) and normalised to the lepton's transverse momentum. With this definition, samples containing genuine W-boson decays like $t\bar{t}$ events should be characterized by small values (signal region), while QCD multi-jet events (background region) are expected to occupy a region around higher values. So, for a given data sample one can define a control region at high values of relative isolation which is expected to be dominated by QCD multi-jet events. Then, by fitting a function to the variable distribution shape in the control region, and then extrapolating its shape into the signal region, an estimate of the number of QCD multi-jet events there can be obtained.

This technique has been deployed in studies involving top-like events with the CMS detector at $\sqrt{s}$ = 7 TeV, for both the μ+jets and e+jets analyses [10] and example fit results for events containing $\geq$0 and $\geq$1 jet in the e+jets analysis are shown in Fig. 3. By attempting various fit intervals and binning schemes, good agreement was achieved between the estimated and predicted (MC) number of events for the e+jets analysis with discrepancies $\leq$10%, whereas for the μ+jets analysis the results were less encouraging. In similar fashion to the lepton relative isolation, the isolation criteria used in tau identification algorithms are known to be strong suppressors of QCD multi-jet events, and are powerful discriminators between signal and QCD multi-jet events. Therefore, a similar method to the one employed in the μ/e+jets analyses could also be applied in the fully hadronic final state of charged Higgs boson searches, by simply replacing the isolated lepton with the tau-jet and by defining a similar isolation variable for the tau-jet. Possible complications can arise from the fact that there are mild isolation requirements for the tau-jet already at the HLT level. Therefore, with the present trigger menu, the extrapolation method with an isolation variable appears to be a less attractive option for the fully hadronic final state, especially when one considers that there are alternative methods available for estimating the QCD multi-jet background (sub-sections 3.3 and 3.4).



(a) (b)

Figure 3: Fit and extrapolation of the relative isolation variable in e+jets events with a high-p_T electron candidate and (a) $\geq$0-jets or (b) $\geq$1-jet in data using an exponential function [10]. Shaded histograms denote expected signal and background processes based on simulation and are normalized to the integrated luminosity of the data (78 nb^{-1}).

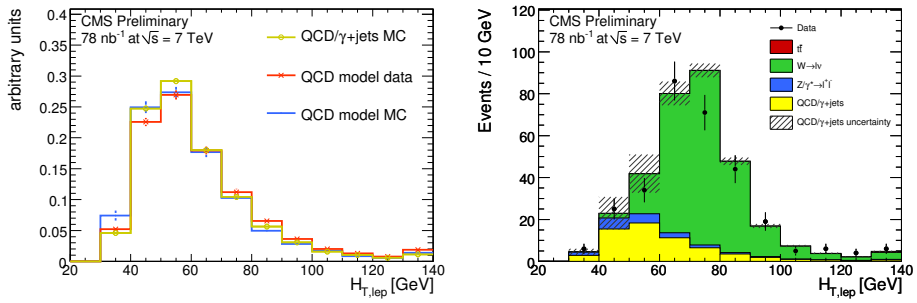
3.3 Fits to discriminating variables using a Template fitting method

Another technique also explored for the lepton+jets analyses in the top-like event studies aims at extracting the QCD contribution to the signal region from fits to event-level kinematic variables with discriminating power between QCD multi-jet and $t\bar{t}$ events [10]. Such variables include the missing transverse energy and the scalar sum of the former and the lepton transverse energy ($H_{T,lep} = E_T^{miss} + E_T^{lep}$). In the following only one of the two variables will be discussed, specifically $H_{T,lep}$, but the method can be applied in the same way to any discriminating variable. In this method, two models were considered in order to obtain template distributions for QCD multi-jet events; the so called background electron template which employs electron candidates that marginally fail the electron selection criteria and the jet electron template which uses positively identified jet objects with large electromagnetic fraction (EMF) that closely resemble electron candidates. According to simulations, both models yielded a QCD purity of 99% and similar QCD shapes. In Fig. 4(a) the normalised sum of the two shapes used in the estimation is shown as determined from MC and data, along with the corresponding true simulated QCD distribution.

Upon comparison of the template obtained from simulation with the true simulated distribution, good agreement was observed with a small bias being visible on the shape of the two distributions, which is attributed to the different event selections. By defining the signal region as $H_{T,lep} >$ 60 GeV, an estimate of the QCD contribution there was obtained by fitting a sum of templates (QCD template from data and W/Z+jets template from simulation) to the data in the background region (QCD dominated region) at small values of $H_{T,lep}$, and extrapolating into the signal region. The estimated number of QCD events is then simply the integral of the QCD template (shown in Fig. 4(b)) in the signal region. The results obtained for both the MET and $H_{T,lep}$ variable, were similar with about $\leq$50% discrepancy from MC. Such template fit methods could be applied to obtain estimates of the QCD multi-jet background in the fully hadronic final state of light charged Higgs boson searches, with an appropriate selection of event-level kinematic variable being the missing transverse energy and/or the dimensionless kinematical variable α_T (sub-section 3.4).

3.4 Extrapolation methods using kinematical variable α_T

The last data-driven method considered involves the kinematical variable α_T which has been



(a) (b)

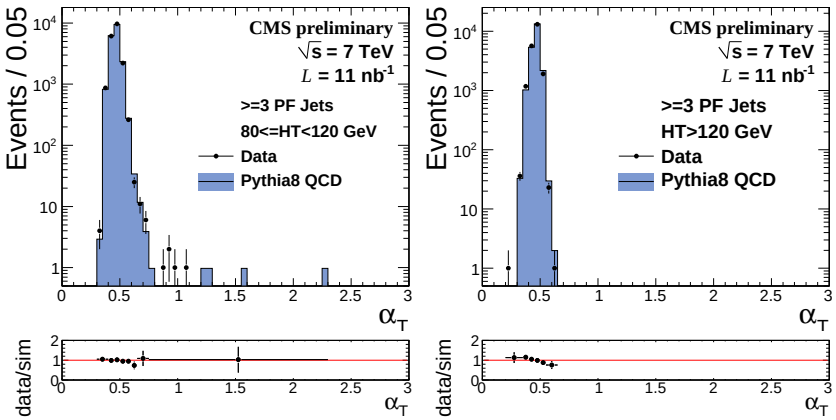
Figure 4: (a) Shape comparison for $H_{T,lep}$ between the template models using MC and data, and the true distribution constructed from simulated events. The QCD template obtained from MC (labeled QCD model MC) is shown as blue solid line with a dot marker, while the QCD template obtained from data (labeled QCD model data) is shown in a red solid line with a cross marker. The predicted QCD shape as determined from simulation (labeled QCD/γ+jets MC) is shown in yellow solid line with an open circle marker.
(b) Distributions of the $H_{T,lep}$ variable for any jet multiplicity in the e+jets mode with data corresponding to an integrated luminosity of 78 nb^{-1}. Predictions from simulation are overlaid, normalized to the data. The last bins include overflows [10].

explored in SUSY searches with the CMS detector [11 – 14]. This dimensionless variable describes the transverse momentum imbalance of an event, and like missing transverse energy is a powerful discriminator against QCD multi-jet background. For multi-jet events, this variable is constructed by using all the jets in an event to create two pseudo-jets and it takes the form:

$$\alpha_T = \frac{1}{2}\frac{H_T - \Delta H_T}{\sqrt{H_T^2 - (MHT)^2}} \tag{3.1}$$

where $H_T = \sum_j^{jets} |\vec{p}_{T_j}|$ is the total transverse energy of the event obtained by summing over all the jet transverse momenta and $\Delta H_T = p_T^{pseudo-jet_1} - p_T^{pseudo-jet_2}$ is the difference in H_T between the pseudo-jets. The quantity MHT= $|\sum_j^{jets} -\vec{p}_{T_j}|$ is the missing H_T of the event and is closely analogous to MET but based only on transverse energy clustered into jets and is thus dependent on jet thresholds. Since there are multiple ways to form two pseudo-jets in a given event, a unique configuration must be chosen and it is the one that provides the maximum balance in transverse momenta; the pseudo-jets configuration that minimises ΔH_T.

For a perfectly balanced multi-jet event with no real MET, the α_T variable is expected to have values close to 0.5, and a recent SUSY study [11] found that QCD events are largely confined in the region $\alpha_T < 0.5$. More specifically, results were reported on the shape of the α_T distributions in QCD samples and their dependence on the total transverse energy of the event. In Fig. 5(a) and 5(b) the α_T distributions for different H_T slices are shown, and below each plot the ratio between data and MC is also shown, demonstrating that there is good agreement. For both H_T slices it was observed that most QCD multi-jet events are confined in the region $\alpha_T < 0.5$, with the lower threshold H_T bin exhibiting a significant tail. This tail is reduced dramatically in the higher H_T slice, as shown in Fig. 5(b), which is understood in terms of jet resolution effects as they become



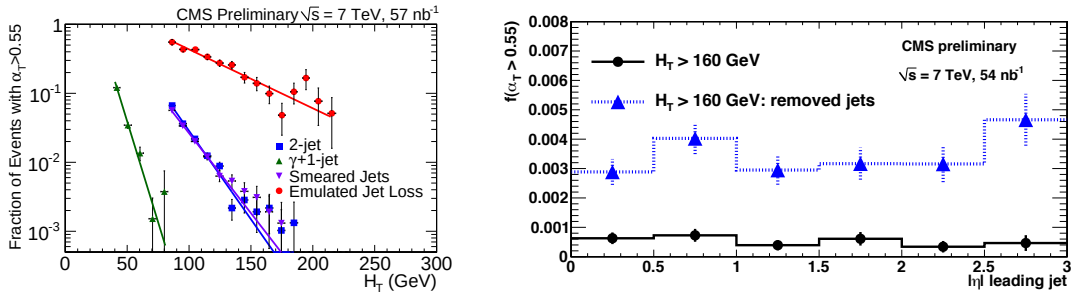
(a) (b)

Figure 5: Plots of the α_T distribution in multi-jet events with particle flow jets for (a) $80 < H_T < 120$ GeV and (b) $H_T > 120$ GeV, demonstrating the reduction of the α_T tail as H_T increases. Below each plot, Data/MC ratio plots are shown [11].

less significant at higher H_T values making it easier for the pseudo-jets to balance. Contrary to QCD events however, SUSY and light charged Higgs boson searches are expected to have α_T distributions which extend well above this value, with the effect increasing at higher Higgs masses due to the more energetic neutrinos from the charged Higgs side.

In order to quantify this H_T dependence of α_T, the fraction of QCD events passing the cut $\alpha_T > 0.55$ was investigated as a function of H_T, with data. It was found that there is an exponential decrease of surviving QCD events as a function of increasing H_T, for both 2-jets and $\geq$3 jets. More interestingly, this behaviour was also found to hold even with mis-measurement biases, like extreme jet losses (simulated by artificially removing a jet from the event) and jet energy smearing, as shown if Fig. 6(a). Therefore, the QCD background estimation planned for SUSY searches exploits the dependence of events with $\alpha_T > 0.55$ on the pseudo-rapidity of the leading jet. The plot shown in Fig. 6(b) shows the fraction of events surviving the $\alpha_T > 0.55$ cut in minimum-bias data, as a function of leading jet pseudo-rapidity (η). In contrast with a typical SUSY signal, the pseudo-rapidity distribution for QCD events is uniform, even in the case of large fake MET which can be simulated by the random removal of a jet from the event. According to MC studies, SUSY events are expected to be more central than QCD events, and so one can define a control region at high $|\eta|$ where the signal is depleted to estimate the QCD shape and extrapolate to the signal region of low values of $|\eta|$, in a similar fashion with the relative isolation and template fit methods.

The behaviour of the α_T variable for QCD events at high H_T values is encouraging for charged Higgs boson searches, since for the fully hadronic final state it is expected that the event H_T will be significant, due to the presence of multiple hadronic jets in addition to the tau-jet. This means that, with a carefully chosen event selection, the QCD multi-jet events will occupy the region $\alpha_T < 0.5$ and provided that the leading jet $|\eta|$ distribution is flat for events surviving the $\alpha_T > 0.55$ cut, a method to estimate the QCD contamination in the signal region is also possible. However, conclusive statements on this method can only be made after detailed application of the method to



(a) (b)

Figure 6: Fraction of events passing an $\alpha_T > 0.55$ cut, as a function of (a) $H_T \geq 3$ jets and $\gamma + \geq 2$ jets samples. The effects of jet loss and jet smearing are also shown. (b) Fraction of events passing an $\alpha_T > 0.55$ cut as a function of $|\eta|$ of the leading jet, in data and minimum bias MC [11].

light charged Higgs boson searches.

4. Conclusions

The strategy for estimating QCD backgrounds in light charged Higgs boson searches using data-driven methods has been presented in the previous sections. In the lepton+jets final state for which QCD is not the dominant background, detailed studies have already been tested [9, 10] with encouraging albeit mixed results. The fully hadronic final state, for which QCD is overwhelmingly the dominant source of background, presents a more difficult task in devising data-driven methods for estimating it accurately, mostly due to fact that it is more challenging to obtain a pure QCD control sample. Various data-driven methods for estimating these QCD backgrounds have been presented here and are still very much a work in progress. It is anticipated that the proposed methods will be studied in more detail as more data become available, which will not only enable their accurate evaluation but also their refinement. Furthermore, future developments of trigger menus that will unavoidably come due to the expected increase in the instantaneous luminosity of LHC, will open new possibilities of data-driven methods for measuring QCD backgrounds in charged Higgs boson searches.

Acknowledgments

The author of this talk would like to thank the Cyprus Research Promotion Foundation's Framework Programme for Research, Technological Development and Innovation 2008 (DESMI 2008), which is co-funded by the Republic of Cyprus and the European Regional Development Fund.

References

[1] Kenzo Inoue, Akira Kakuto, Hiromasa Komatsu and Seiichiro Takeshita. *Aspects of Grand Unified Models with Softly Broken Supersymmetry*. Progress of Theoretical Physics **68** (3), 927–946 (1982).

[2] Savas Dimopoulos and Howard Georgi. *Softly Broken Supersymmetry and SU(5)*. Nucl. Phys. **B193**, 150 (1981).

[3] N. Sakai. *Naturalness in Supersymmetric Guts*. Zeit. Phys. **C11**, 153 (1981).

[4] R. Adolphi et al. *The CMS experiment at the CERN LHC*. JINST **3**, S08004 (2008).

[5] D. P. Roy. *The Hadronic τ decay signature of a heavy charged Higgs boson at LHC*. Phys. Lett. **B459**, 607–614 (1999).

[6] G. L. Bayatian et al. *CMS technical design report, volume II: Physics performance*. J. Phys. **G34**, 995–1579 (2007).

[7] Sreerup Raychaudhuri and D. P. Roy. *Sharpening up the charged Higgs boson signature using τ polarization at LHC*. Phys. Rev. **D53**, 4902–4908 (1996).

[8] The CMS Collaboration. *Study of tau reconstruction algorithms using pp collisions data collected at sqrt(s) = 7 TeV*. **PFT-10-004** (Jul 2010).

[9] The CMS Collaboration. *Towards the measurement of the ttbar cross section in the e-tau and mu-tau dilepton channels in pp collisions at sqrt(s)=14 TeV*. **TOP-08-004** (Aug 2009).

[10] The CMS Collaboration. *Selection of Top-Like Events in the Dilepton and Lepton-plus-Jets Channels in Early 7 TeV Data*. **TOP-10-004** (Jul 2010).

[11] The CMS Collaboration. *Performance of Methods for Data-Driven Background Estimation in SUSY Searches*. **SUS-10-001** (Jul 2010).

[12] The CMS Collaboration. *SUSY searches with dijet events*. **SUS-08-005** (Oct 2008).

[13] The CMS Collaboration. *Search strategy for exclusive multi-jet events from supersymmetry at CMS*. (Jul 2009).

[14] Lisa Randall and David Tucker-Smith. *Dijet Searches for Supersymmetry at the LHC*. Phys. Rev. Lett. **101**, 221803 (2008).