Data-driven background estimation in CMS

Matti KORTELAINEN[*] **on behalf of the CMS Collaboration**
Helsinki Institute of Physics, Finland
E-mail: matti.kortelainen@helsinki.fi

The data-driven background measurement methods employed in the search for a light charged Higgs boson in the CMS experiment are reviewed. In the τ_h+jets final state analysis these backgrounds include the QCD multijet events and the EWK+$t\bar{t}$ events with a τ lepton in the p_T, $|\eta|$ acceptance. In the $e\tau_h$ and $\mu\tau_h$ final state analyses the background from events with misidentified τ was measured from data.

[*]Speaker.

1. Introduction

The observation of a new boson in the searches for the standard model (SM) Higgs boson by the ATLAS [1] and CMS [2] experiments made the question of the possibility of additional Higgs bosons more interesting. A popular extension on the SM Higgs sector is an addition of a second complex doublet leading to a prediction of the existence of a charged Higgs boson. Such a structure exists, for example, in the minimal supersymmetric extension of the standard model (MSSM) [3, 4, 5, 6, 7, 8, 9, 10]. For values of $\tan\beta > 5$, the charged Higgs boson preferentially decays to a τ lepton and a neutrino, where $\tan\beta$ is defined as the ratio of the vacuum expectation values of the two Higgs field doublets.

CMS [11] has searched for a light ($m_{H^+} < m_{top}$) charged Higgs boson in the top quark decays using $2.3\,\mathrm{fb}^{-1}$ of $\sqrt{s} = 7$ TeV proton-proton collision data recorded in 2011 [12]. The analysis considered the τ_h+jets final state [13] and the $e\tau_h$, $\mu\tau_h$, and $e\mu$ final states [14], where the hadronically decaying τ is denoted as τ_h. An important part of the analysis was the measurements of the backgrounds from data, which are described in this note.

In the τ_h+jets final state analysis the background was divided into QCD multijet events, and events from various electroweak processes, like W+jets, Z+jets, diboson (WW, WZ, ZZ) as well as SM $t\bar{t}$ and tW production. The latter background was further divided into two parts: the first one labelled "EWK+$t\bar{t}$ τ" consists of events where at least one τ lepton in the final state is present with $p_T^\tau > 40\,\mathrm{GeV}/c$ and $|\eta^\tau| < 2.1$, and the second one labelled "EWK+$t\bar{t}$ no-τ" consisting of events with no τ leptons in the final state or with no τ leptons satisfying the above-mentioned criteria. Of these backgrounds the QCD multijet and "EWK+$t\bar{t}$ τ" backgrounds were measured from data and are described further in Sections 2 and 3, respectively. The "EWK+$t\bar{t}$ no-τ" background was estimated using simulation.

In the $e\tau_h$ and $\mu\tau_h$ final state analyses the background arises from two sources: misidentified τ_h, and genuine τ_h. The misidentified τ_h background was measured from data, and is discussed further in Section 4, while the genuine τ_h background was estimated using simulation. The backgrounds of the $e\mu$ final state were estimated using simulation.

2. QCD multijet background measurement for the τ_h+jets final state analysis

In the analysis of the τ_h+jets final state the transverse mass of the τ-ν_τ system, reconstructed from the τ_h and missing transverse energy E_T^{miss}, was employed in the statistical analysis. The normalization and the shape of the transverse mass distribution from the QCD multijet background were measured separately. The measurements are described in the following subsections, a more detailed discussion can be found in Ref. [15].

2.1 Number of events

An overview of the method to measure the number of QCD multijet events is shown in Fig. 1. The events were first required to pass the τ-plus-E_T^{miss} trigger, primary vertex, τ_h candidate, rejection of events with isolated electrons or muons, and hadronic jet selections (selections 1–5 in Fig. 1). The details of these requirements can be found in Refs. [12, 13].
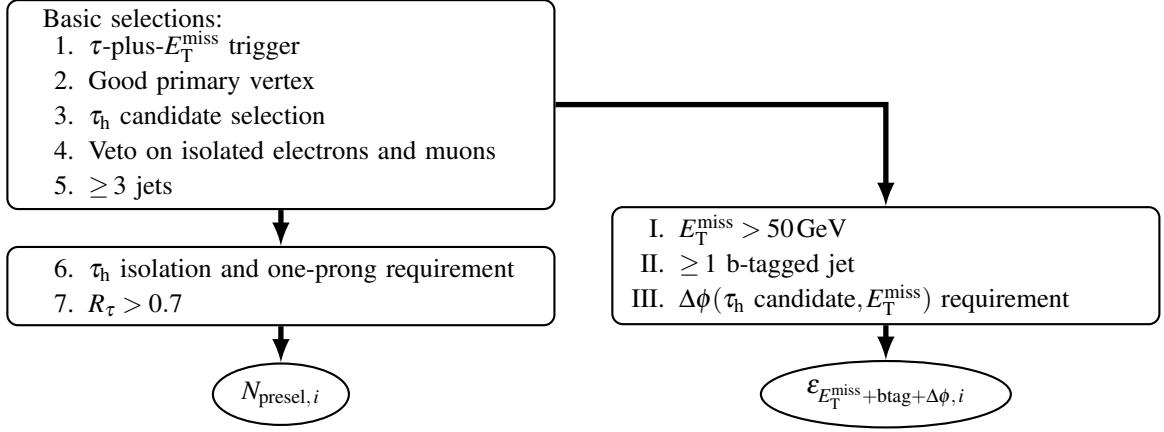
Figure 1: Schematic overview of the measurement of the QCD multijet event yield. The index i denotes the τ_h candidate p_T bin.

Using these selected events, the efficiency of the E_T^{miss}, b-tagging, and $\Delta\phi(\tau_h$ candidate, $E_T^{miss})$ selections (I-III in Fig. 1) was measured in τ_h candidate p_T bins i as

$$\varepsilon_{E_T^{miss}+btag+\Delta\phi,i} = \frac{N_{\text{basic selections}+I+II+III,i}^{\text{data}} - N_{\text{basic selections}+I+II+III,i}^{\text{EWK sim}}}{N_{\text{basic selections},i}^{\text{data}} - N_{\text{basic selections},i}^{\text{EWK sim}}} \ . \tag{2.1}$$

Also the number of events after additional τ_h identification (6–7 in Fig. 1) was recorded ($N_{\text{presel},i}^{\text{data}}$). A small correlation between τ_h candidate p_T and E_T^{miss} was reduced to a negligible level by performing the measurement in τ_h candidate p_T bins. A small amount of EWK+$t\bar{t}$ events, estimated using simulation, was subtracted in both. The estimate for the number of QCD multijet events was then obtained from

$$N^{QCD} = \sum_{i}^{\tau_h \text{ candidate } p_T \text{ bins}} \left(N_{\text{presel},i}^{\text{data}} - N_{\text{presel},i}^{\text{EWK sim}}\right) \times \varepsilon_{E_T^{miss}+btag+\Delta\phi,i} \ . \tag{2.2}$$

The uncertainty on N^{QCD} was evaluated by applying error propagation to Equation (2.2). The considered sources of uncertainties were statistical uncertainties on the number of data and simulated events, and a total 20 % systematic uncertainty on the number of simulated events. These uncertainties translate to 6.5 % and 3.8 % on the final result, respectively.

The final result for the QCD multijet event yield was

$$N^{QCD} = 26 \pm 2\,(\text{stat.}) \pm 1\,(\text{syst.}) \ . \tag{2.3}$$

2.2 Transverse mass distribution shape

The method for measuring the shape of the transverse mass distribution was similar to the event yield measurement discussed in the previous subsection. An overview of the method is shown in Fig. 2. In contrast to the event yield measurement, the efficiency of the τ_h identification was measured in τ_h candidate p_T bins using events which passed the basic selections (selections 1–5 in Fig. 2). The efficiency for τ_h candidate p_T bin i was defined as

$$\varepsilon_{\tau_h \text{ ID},i} = \frac{N_{\text{basic selections}+6+7,i}^{\text{data}} - N_{\text{basic selections}+6+7,i}^{\text{EWK sim}}}{N_{\text{basic selections},i}^{\text{data}} - N_{\text{basic selections},i}^{\text{EWK sim}}} \ . \tag{2.4}$$
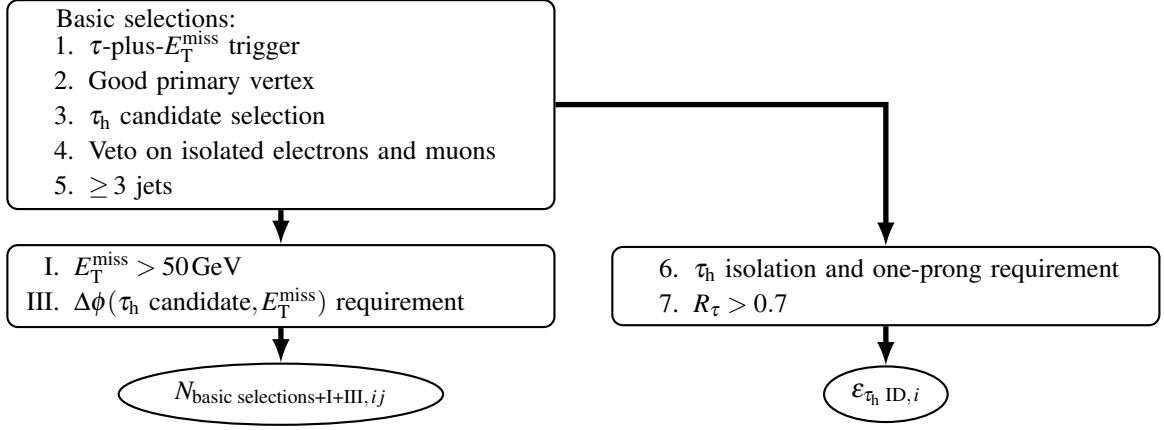
Figure 2: Schematic overview of the measurement of the shape of the transverse mass distribution in QCD multijet background. The index i denotes the τ_h candidate p_T bin, and the index j refers to a bin in the transverse mass distribution.

The correlation between tagging of b jets and the shape of the transverse mass distribution was found to be negligible, and therefore the b-tagging was left out from the selections. The transverse mass was reconstructed from the τ_h candidate and E_T^{miss} in bins of the τ_h candidate p_T using events which passed the basic selections, as well as E_T^{miss} and $\Delta\phi(\tau_h$ candidate$, E_T^{miss})$ requirements. The number of events in the transverse mass histogram bin j for the τ_h candidate p_T bin i was then given by

$$N(m_T)_{ij} = \left(N^{\text{data}}_{\text{basic selections}+\text{I}+\text{III},ij} - N^{\text{EWK sim}}_{\text{basic selections}+\text{I}+\text{III},ij}\right) \times \varepsilon_{\tau_h \text{ ID},i} . \qquad (2.5)$$

Summing over the τ_h candidate p_T bins gives the number of events in the transverse mass histogram bin j

$$N(m_T)_j = \sum_i^{\tau_h \text{ candidate } p_T \text{ bins}} \left(N^{\text{data}}_{\text{basic selections}+\text{I}+\text{III},ij} - N^{\text{EWK sim}}_{\text{basic selections}+\text{I}+\text{III},ij}\right) \times \varepsilon_{\tau_h \text{ ID},i} . \qquad (2.6)$$

The final step was to normalize the area of the histogram to the QCD multijet event yield (N^{QCD}). The uncertainties were, again, evaluated by applying the error propagation to Equation (2.6).

The resulting transverse mass distribution, normalized to unit area, is shown in Fig. 3 without and with the b-tagging requirement for three $\Delta\phi(\tau_h, E_T^{miss})$ requirement values. The good agreement between the shapes of the distributions without and with the b-tagging requirement justifies the use of events without the b-tagging for the transverse mass distribution.

The bump in the distributions in the 100–$150\,\text{GeV}/c^2$ region was found to be caused by multijet events where the hadronic jet misidentified as a τ_h and the $\vec{E}_T^{miss}$ are in back-to-back configuration in the transverse plane. The amount of such events can be suppressed by imposing a requirement on $\Delta\phi(\tau_h, E_T^{miss})$. The effect of this requirement can be clearly seen in Fig. 3. The relative height of the bump was reduced already with the $\Delta\phi < 160°$ requirement, and it was almost completely removed with the tighter $\Delta\phi < 130°$ selection. The $\Delta\phi < 160°$ option was chosen in the analysis.
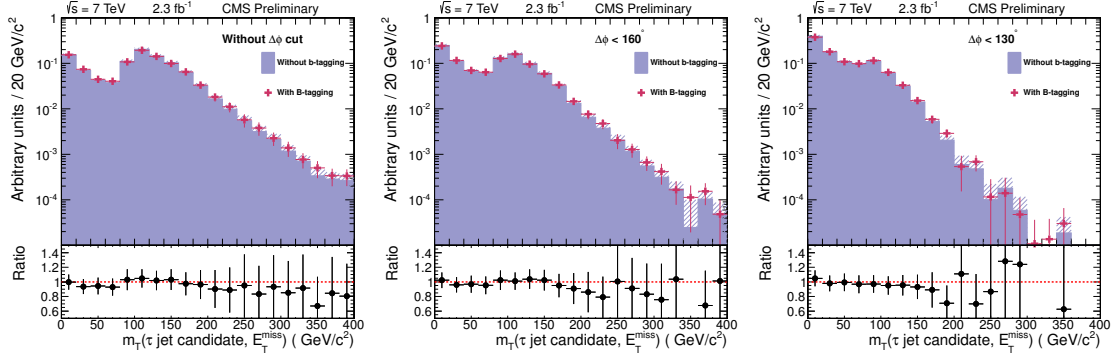
Figure 3: The transverse mass distributions of the QCD multijet background without the $\Delta\phi(\tau_h \text{ candidate}, E_T^{\text{miss}})$ requirement (left), with $\Delta\phi < 160°$ (middle; used in the analysis), and with $\Delta\phi < 130°$ (right). The histograms were normalised to unit area in order to compare their shapes. The crosses were obtained with b-tagging, and the filled histograms without it. The statistical uncertainties are shown as the error bars (with b-tagging, ratio) and as the shaded area (without b-tagging) [15].

3. EWK+$t\bar{t}$ τ background measurement for the τ_h+jets final state analysis

The "EWK+$t\bar{t}$ τ" background consists of $t\bar{t} \to W^+bW^-\bar{b}$, W+jets, Z/γ^* +jets, single top quark, WW, WZ, and ZZ events with a τ lepton within the $p_T > 40\,\text{GeV}/c$, $|\eta| < 2.1$ acceptance. This background was measured by using $\mu + \geq 3$ jets events and by applying the *tau embedding method*, similarly to Refs. [16, 17, 18]. The measurement is discussed in more detail in Ref. [19]. The method relies on the list of particles reconstructed by the particle-flow algorithm [20].

An overview of the method is shown in Fig. 4. First, a control sample of $\mu + \geq 3$ jets was selected. The muons were required to be isolated using a method similar to the τ_h isolation, but looser. Events with additional muons or electrons were rejected. The muon p_T distribution after the control sample selection is shown in Fig. 5, showing that the data agrees reasonably well with the simulation.

A decay of a τ lepton was simulated and reconstructed up to the particle-flow particles with the momentum of the τ lepton equal to the measured momentum of the μ. The μ was removed from the list of particle-flow particles of the event, and the particles from the simulated τ decay were embedded to the list. Thus the produced *hybrid* event contains a τ decay from simulation, and all other physics objects from data. The jets, τ_h, and E_T^{miss} were reconstructed from the particles of the hybrid event. The remaining event selections were applied on the hybrid event. These selections include the τ_h identification, E_T^{miss} requirement, b-jet tagging, $\Delta\phi(\tau_h, E_T^{\text{miss}})$ requirement, and the reconstruction of the transverse mass of the τ-ν system.

The following effects were taken into account in the normalization of the embedded sample

1. muon trigger and offline identification efficiency ($\varepsilon_{\mu\,\text{sel}}$), measured using $Z \to \mu\mu$ events and the tag-and-probe method [21],

2. fraction of $W \to \tau\nu_\tau \to \mu\nu_\mu\nu_\tau$ events of all selected $W \to \mu\nu$ events ($f_{W\to\tau\to\mu}$), estimated using simulation,

3. τ trigger efficiency ($\varepsilon_{\tau\,\text{trg}}$) in bins of τ_h p_T, measured using $Z \to \tau\tau \to \tau_\mu\tau_h$ events and the tag-and-probe method, and
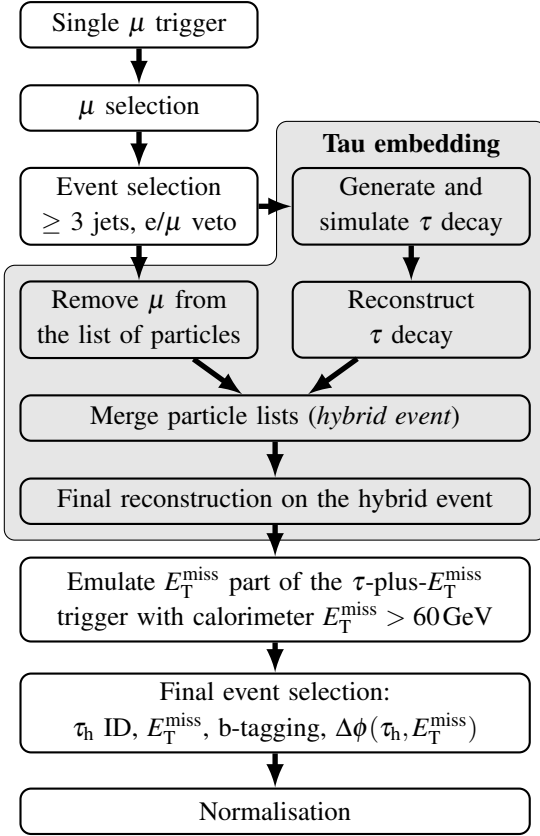
Figure 4: Schematic overview of the tau embedding method for the EWK+t$\bar{\text{t}}$ genuine τ background measurement.
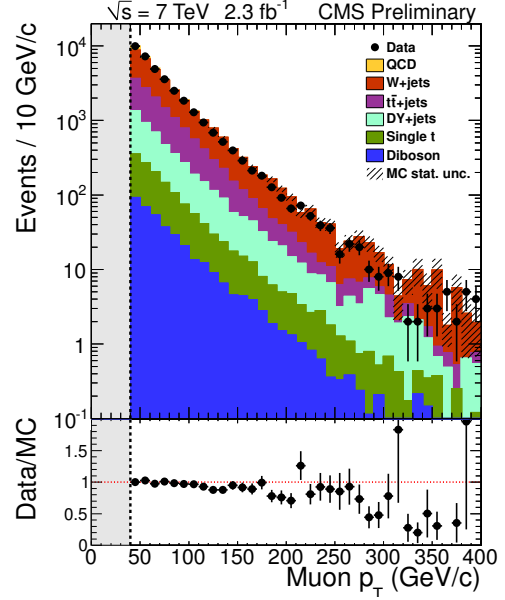
Figure 5: Distribution of the selected muon p_T in the control sample events for the data and the simulated samples.

4. E_T^{miss} trigger, incorporated by calculating calorimeter E_T^{miss} of the hybrid event as an approximation of the trigger-level E_T^{miss}, and accepting only events with calorimeter $E_T^{\text{miss}} > 60\,\text{GeV}$.

By denoting the number of hybrid events passing the final event selection and the calorimeter E_T^{miss} requirement in the τ_h p_T bin i with $N_i^{\text{EWK+t}\bar{\text{t}}\tau}$, the estimate for the number of the EWK and t$\bar{\text{t}}$ genuine τ events becomes

$$N^{\text{EWK+t}\bar{\text{t}}\,\tau} = (1 - f_{\text{W}\to\tau\to\mu}) \times \frac{\sum_i^{\tau_h\,p_T\,\text{bins}} N_i^{\text{EWK+t}\bar{\text{t}}\,\tau}\varepsilon_{\tau\,\text{trg}}}{\varepsilon_{\mu\,\text{sel}}}. \tag{3.1}$$

The statistical precision of the measurement was improved by repeating the embedding procedure ten times using different random number generator seeds. The embedding method turned out to underestimate the background from events with two hadronically decaying τ leptons, mainly from the $Z/\gamma^* \to \tau\tau$ and $WW \to \tau\nu_\tau\tau\nu_\tau$ processes. This was caused by the rejection of events with a second isolated muon in the control sample selection, the rejection of events with a second identified τ_h in the analysis event selection, and the fact that the muon identification efficiency is significantly larger than the τ_h identification efficiency. These residues of background from ditau events were estimated using simulation.
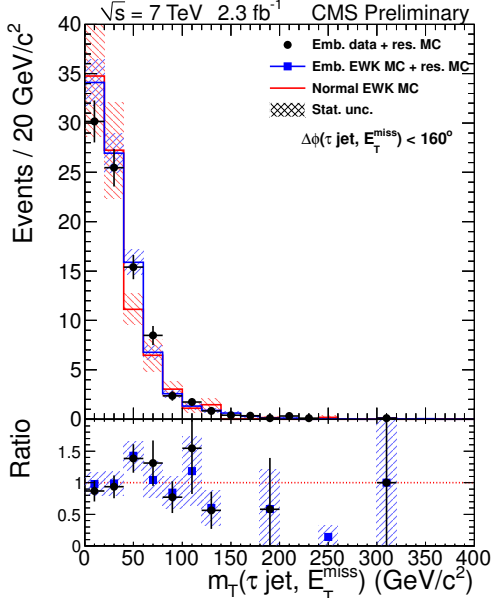
Figure 6: Distribution of the transverse mass reconstructed from the embedded τ_h and the E_T^{miss} ($m_T(\tau_h, E_T^{miss})$) after the final event selection for the embedded data and the embedded simulation, both including the residual $Z/\gamma^* \to \tau\tau$ and $WW \to \tau\nu_\tau\tau\nu_\tau$ background from simulation, and for the normal simulation. In the lower frame, $N^{embedded\ data+res.\ sim.}/N^{normal\ sim.}$ is shown in black dots, and $N^{embedded\ sim.+res.\ sim.}/N^{normal\ sim.}$ in blue squares.
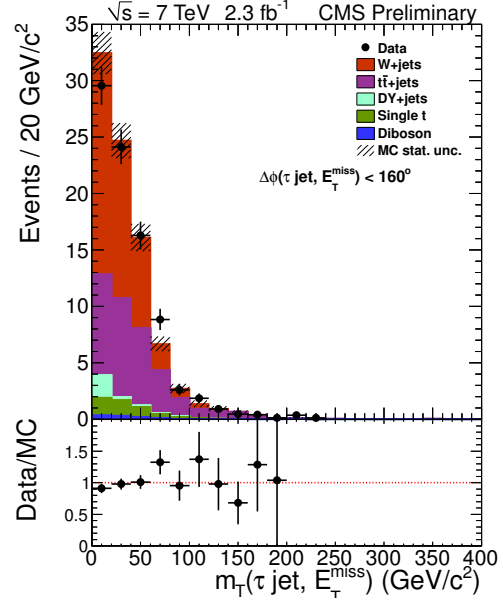
Figure 7: Distribution of the transverse mass $m_T(\tau_h, E_T^{miss})$ after the final event selection. The residual $Z/\gamma^* \to \tau\tau$ and $WW \to \tau\nu_\tau\tau\nu_\tau$ contributions are not included in this figure.

Table 1: Results of the "EWK+$t\bar{t}$ τ" background measurement

Sample	Events
Embedded data	$78 \pm 3\,(\text{stat.}) \pm 11\,(\text{syst.})$
Residual $Z/\gamma^* \to \tau\tau$	$7.0 \pm 2.0\,(\text{stat.}) \pm 2.1\,(\text{syst.})$
Residual $WW \to \tau\nu_\tau\tau\nu_\tau$	$0.35 \pm 0.23\,(\text{stat.}) \pm 0.09\,(\text{syst.})$

The measurement method was extensively validated by comparing embedded simulation, embedded data, and normal simulation [19]. As an example, the transverse mass distribution for these three sets of events are compared in Fig. 6 showing a reasonable agreement.

The transverse mass distribution from the embedded data and embedded simulation is shown in Fig. 7, and the event yields are shown Table 1. The dominant sources of uncertainty were efficiency of the τ-plus-E_T^{miss} trigger (11%), τ_h identification efficiency (6%), τ_h energy scale (6.6%), and the statistical uncertainty of the control sample (3.4%).
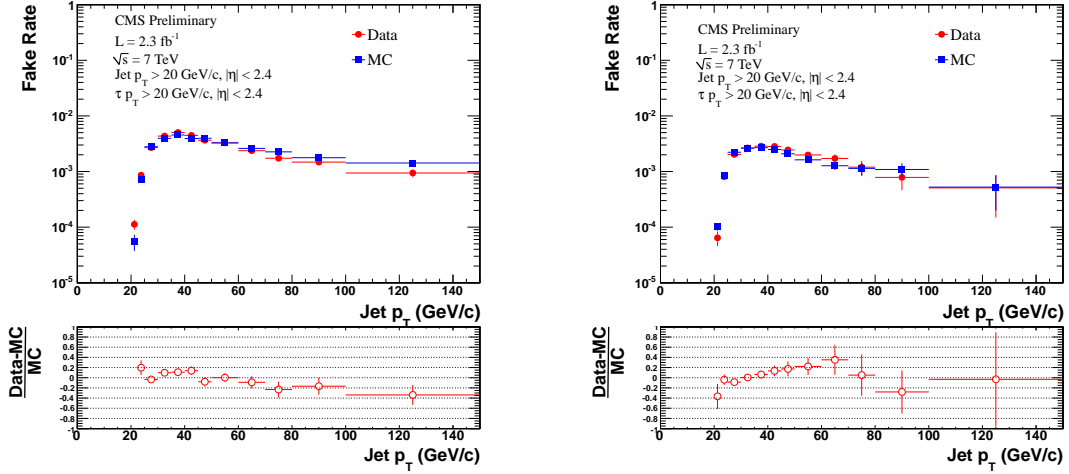
Figure 8: The probability of a jet to be misidentified as a τ_h as a τ_h as a function of the jet p_T evaluated from W+jets type events (left), and from QCD multijet type events (right).

4. Misidentified τ_h background measurement for the eτ_h and $\mu\tau_h$ final state analyses

The background of the $\mu\tau_h$ and eτ_h final states, where a jet is misidentified as the τ_h, was measured with a "fake rate" technique [22]. First a probability for a jet to be misidentified as a τ_h (jet-to-τ_h probability) was measured from data separately for W+jets type events and QCD multijet events. The jets in the former are primarily quark jets, while in the latter the jets are dominantly gluon jets.

The W+jets events were triggered with a single isolated muon trigger ($p_T > 20\,\mathrm{GeV}/c$ and $|\eta| < 2.1$), and the events were required to have at least one jet ($p_T > 20\,\mathrm{GeV}/c$ and $|\eta| < 2.4$) and the transverse mass reconstructed from the muon and E_T^{miss} had to be greater than 50 GeV.

The QCD multijet events were triggered with a single jet trigger ($p_T > 30\,\mathrm{GeV}/c$), and were required to have at least two jets ($p_T > 20\,\mathrm{GeV}/c$, $|\eta| < 2.4$). All jets except the one firing the trigger were used in the misidentification rate calculation in order to avoid bias from the trigger. However, if two or more jets fired the trigger, all jets were used to calculate the misidentification rate.

The jet-to-τ_h probability was parameterized as a function of the jet p_T, η, and radius $R = \sqrt{\sigma_{\eta\eta}^2 + \sigma_{\phi\phi}^2}$ using the k-Nearest Neighbour (kNN) regression [23]. The misidentification rates are shown in Fig. 8 for W+jets events and QCD multijet events. The misidentified τ_h background was then estimated as follows. A set of $\ell + \geq 3$ jets events was selected by requiring 1 isolated electron or muon, E_T^{miss}, at least three jets, and at least one b-tagged jet. The selection thresholds were the same as in the signal event selection [12, 14]. These events are dominated by W+jets and $t\bar{t} \to \ell +$ jets events. The events were weighted by applying the jet-to-τ_h probability to each jet in the events. The selected events contain a small contribution from genuine τ events. The amount of these events was estimated using simulation and subtracted. The true quark and gluon jet composition in these $\ell + \geq 3$ jets lies between the QCD multijet and W+jets events. As a conservative approach, an average of misidentification rate estimates from QCD multijet and W+jets events was taken.

Table 2: Results of misidentified τ_h background measurement for the $e\tau_h$ (top) and $\mu\tau_h$ (bottom) final state. The simulation expectation was obtained by using the generator information, while the simulation estimate was obtained by applying the data-driven method to simulated samples. The residual of genuine τ events was estimated using simulation, and is already subtracted from the data estimate value. The quoted uncertainties are statistical.

Sample ($e\tau_h$)	Simulation expectation	Estimated from simulation	Estimated from data	Residual from simulation
QCD multijet		54.9	64.1	19.6
W+jets	57.9 ± 5.1	78.9	86.7	27.4
Average		66.9 ± 12.0	75.4 ± 11.3	23.5 ± 3.9

Sample ($\mu\tau_h$)	Simulation expectation	Estimated from simulation	Estimated from data	Residual from simulation
QCD multijet		105.1	113.0	34.4
W+jets	120.1 ± 8.1	147.3	144.5	44.3
Average		126.2 ± 21.1	128.8 ± 15.8	39.4 ± 4.9

For the final result, the event weights were multiplied with the efficiency of the opposite-sign requirement. This efficiency was estimated using simulation, and cross-checked using data. The measurement method was validated by applying the data-driven method to simulation, and comparing the result with the expectation obtained from simulation using generator information.

The results of the validation and the background measurement are shown in Table 2 for both $e\tau_h$ and $\mu\tau_h$ final states. The simulation expectation and simulation estimate with the data-driven method agree within statistical uncertainties, validating the method. Also the estimate from data agrees with the estimate from simulation within the uncertainties.

The dominant uncertainties were on the difference in the misidentification rates for quark and gluon jets (12 %), and on the efficiency of the opposite-sign requirement (10 %). Employing the jet radius in the misidentification rate parameterization significantly decreased the uncertainty on the quark and gluon jet misidentification rate difference from about 25 %.

5. Summary

The data-driven background measurement methods employed in the search for a light charged Higgs bosons in the CMS experiment were reviewed. In the analysis of the τ_h+jets final state, the main backgrounds were measured from data. The QCD multijet background was measured by factorizing the τ_h isolation and E_T^{miss} selections, and the "EWK+t$\bar{t}$ τ" background was measured using $\mu + \geq 3$ jets events and replacing the μ with a simulated τ decay (tau embedding). In the $e\tau_h$ and $\mu\tau_h$ final state analyses the background from jets misidentified as τ_h was estimated by measuring the misidentification rate, and weighting $\ell +$jets events by the misidentification rate.

References

[1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B*, **716** (1) 1–29, 2012.

[2] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys. Lett. B*, **716** (1) 30–61, 2012.

[3] P. Fayet, *Supergauge invariant extension of the Higgs mechanism and a model for the electron and its neutrino*, *Nucl. Phys. B* **90** 104–102, 1975.

[4] P. Fayet, *Supersymmetry and weak, electromagnetic and strong interactions*, *Phys. Lett. B* **64** (2) 159–162, 1976.

[5] P. Fayet, *Spontaneously broken supersymmetric theories of weak, electromagnetic and strong interactions*, *Phys. Lett. B* **69** (4) 489–494, 1977.

[6] S. Dimopoulos and G. Georgi, *Softly broken supersymmetry and SU(5)*, *Nucl. Phys. B*, **193** (1) 150–162, 1981.

[7] N. Sakai, *Naturalness om supersymmetric GUTs*, *Z. Phys. C* **11** (153), 1981.

[8] K. Inoue et al., *Low Energy Parameters and Particle Masses in a Supersymmetric Grand Unified Model*, *Prog. Theor. Phys.* **67** (6) 1889-1898, 1982.

[9] K. Inoue et al. *Aspects of Grand Unified Models with Softly Broken Supersymmetry*, *Prog. Theor. Phys.*, **68** (3), 927–946, 1982.

[10] K. Inoue et al. *Renormalization of Supersymmetry Breaking Parameters Revisited*, *Prog. Theor. Phys.*, **71** (2) 413–416, 1984.

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

[12] CMS Collaboration, *Search for a light charged Higgs boson in top quark decays in pp collisions at $\sqrt{s}$ = 7 TeV*, *JHEP* **07** (2012) 143.

[13] A. Attikis, *Search for $H^+ \to \tau^+ \nu$ with fully hadronic final state in CMS*, these proceedings.

[14] P. Vischia, *Search for $H^+ \to \tau^+ \nu$ with $\ell^+ \tau(\to had)$ and $\ell\ell$ final states in CMS*, these proceedings.

[15] A. Attikis, *Search for light charged Higgs boson with the $H^+ \to \tau^+ \nu_\tau$ decay in the fully hadronic final state*, PhD thesis, *University of Cyprus*, 2012.

[16] M. Flechl, *Looking for the Charged Higgs Boson: Simulation Studies for the ATLAS Experiment*, PhD thesis, *Uppsala University*, 2009.

[17] ATLAS Collaboration, *Search for charged Higgs bosons decaying view $H^+ \to \tau\nu$ in $t\bar{t}$ events using pp collisions data at $\sqrt{s}$ = 7 TeV with the ATLAS detector*, *JHEP* **06** (2012) 39.

[18] Y. Rozen, *Data-driven background estimation in ATLAS*, these proceedings.

[19] M.J. Kortelainen, *Search for a Light Charged Higgs Boson in the CMS Experiment in pp Collisions at $\sqrt{s}$ = 7 TeV*, PhD thesis, HIP-2012-06, *University of Helsinki*, 2012.

[20] CMS Collaboration, *Particle-flow event reconstruction in CMS and performance for jets, taus, and E_T^{miss}*, *CMS Physics Analysis Summary* **CMS-PAS-PFT-09-001**, 2009.

[21] CMS Collaboration, *Measurements of inclusive W and Z cross sections in pp collisions at $\sqrt{s}$ = 7 TeV*, *JHEP* **01** (2011) 080.

[22] CMS Collaboration, *Performance of the τ-lepton reconstruction and identification in CMS*, 2012 *JINST* **7** P01001.

[23] A. Hoecker et al., *TMVA Users Guide*, arXiv:physics/0703039, 2009.

PoS(CHARGED 2012)012