Data-driven background measurements in CMS

Matti Kortelainen\textsuperscript{*} on behalf of the CMS Collaboration

Helsinki Institute of Physics, Finland

E-mail: matti.kortelainen@cern.ch

The data-driven background measurement methods employed in the searches for a charged Higgs boson in the CMS experiment are reviewed. In the $\tau_h$+jets final state analysis these backgrounds include the multijet events, and the EWK+t$\bar{t}$ events where the selected $\tau_h$ originates from a hadronically decaying $\tau$ lepton. In the $\mu$+$\tau$ final state analysis the backgrounds from events with misidentified $\tau_h$, and from Drell-Yan events with one genuine $\tau_h$ and the other $\tau$ decaying to a muon are measured from data.

\textsuperscript{*}Speaker.
1. Introduction

The CMS and ATLAS collaborations have observed a boson \(m_H = 125 \text{GeV}\) and properties consistent with the predicted standard model (SM) Higgs boson [3, 4, 5, 6]. The observation makes models with an extended Higgs sector more interesting, although the measured properties also constrain the models. 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.

CMS [7] has conducted the following searches for a light \((m_{H^+} < m_{t\bar{t}} - m_b)\) charged Higgs boson in the top quark decays (\(\tau_h\) denotes the hadronic decay of the \(\tau\) lepton):

- \(\tau_h+\text{jets}\) final state for the \(H^+ \rightarrow \tau\nu\tau\) decay channel [8, 9], and
- \(\mu+\text{jets}\) final state for the \(H^+ \rightarrow c\bar{s}\) decay channel [10, 11],

and the following searches for a heavy \((m_{H^+} > m_{t\bar{t}} - m_b)\) charged Higgs boson

- \(\tau_h+\text{jets}\) final state for the \(H^+ \rightarrow \tau\nu\tau\) decay channel [8, 9], and
- \(\mu+\tau_h\) and ee/e\(\mu/\mu\mu\) final states for the \(H^+ \rightarrow \tau\nu\tau\) and \(H^+ \rightarrow t\bar{b}\) decay channels [12, 13]

using the proton-proton collision data at \(\sqrt{s} = 8\ \text{TeV}\) recorded in 2012 corresponding to an integrated luminosity of 19.7 \(\text{fb}^{-1}\). Important parts of the analyses of \(\tau_h+\text{jets}\) and \(\mu+\tau_h\) final states are the data-driven background measurements, which are described in this note.

In the \(\tau_h+\text{jets}\) final state analysis the background is divided into multijet events, and events from various electroweak processes, like W+jets, Z+jets, diboson (WW, WZ, ZZ) as well as from the SM \(t\bar{t}\) and tW production. The latter background is further divided into two parts: the first one labelled “EWK+\(t\bar{t}\) with \(\tau_h\)” consists of events where the selected \(\tau_h\) originates from a hadronically decaying \(\tau\) lepton, and the second one labelled “EWK+\(t\bar{t}\) no \(\tau_h\)” consists of events where an electron, a muon, or a jet is misidentified as the \(\tau_h\). Of these backgrounds the multijet and “EWK+\(t\bar{t}\) with \(\tau_h\)” backgrounds are measured from data as described in Sections 2.1 and 2.2, respectively. The “EWK+\(t\bar{t}\) no \(\tau_h\)” background is estimated using simulation.

In the \(\mu+\tau_h\) final state analysis the background arises from two sources: misidentified \(\tau_h\), and genuine \(\tau_h\). The misidentified \(\tau_h\) background is measured from data, as is also the \(Z \rightarrow \tau\tau \rightarrow \tau_h\mu\) component of the genuine \(\tau_h\) background. The measurements are described in Sections 3.1 and 3.2, respectively. The remaining genuine \(\tau_h\) background is estimated using simulation.

The backgrounds in the remaining analyses, i.e. of \(\mu+\text{jets}\) and ee/e\(\mu/\mu\mu\) final states, are estimated using simulation. Common themes in the background measurements are the fake-rate technique for jets misidentified as the \(\tau_h\), and the embedding technique for events with genuine \(\tau_h\).

2. Background measurements in the \(\tau_h+\text{jets}\) final state analysis

2.1 Measurement of the multijet background

The multijet background is measured with a “\(\tau_h\) fake-rate” technique. First, the probability of the \(\tau_h\) candidate to pass the nominal and inverted \(\tau_h\) isolation criteria is measured in bins of
\( \tau_0 \) transverse momentum in an event sample obtained prior applying the \( E_T^{miss} \), b tagging, and \( p_T^{\tau_0} - E_T^{miss} \) back-to-back parts of the event selection. The event sample that passes the nominal \( \tau_0 \) isolation selection contains a non-negligible contribution from EWK+\( t \bar{t} \) events. Therefore the numbers of multijet and EWK+\( t \bar{t} \) events are determined by maximum likelihood fit of the \( E_T^{miss} \) distribution. One such fit of the \( E_T^{miss} \) distribution is performed per \( p_T^{\tau_0} \) bin. The \( E_T^{miss} \) templates are obtained for multijet events from simulation in the nominal region. Figures 1 (a–g) show the multijet template (red dashed line), the EWK+\( t \bar{t} \) template (blue solid line), and the data distribution in the nominal region (black dots) for deriving the number of multijet events in the nominal region in bins of \( p_T^{\tau_0} \). The \( \tau_0 \) isolation criteria in a control region dominated by multijet background, defined by inverting the b-tagging requirement. The comparison is depicted in Fig. 2. The transverse mass distribution for multijet events obtained with the \( \tau_0 \) fake-rate method is found to agree with the \( m_T \) spectrum obtained with nominal \( \tau_0 \) isolation criteria in this control region.

As a cross-check, the transverse mass \((m_T)\) distribution obtained with the \( \tau_0 \) fake-rate technique is compared to the \( m_T \) distribution obtained with the nominal \( \tau_0 \) isolation criteria in a control region dominated by multijet background, defined by inverting the b-tagging requirement. The comparison is depicted in Fig. 2. The transverse mass distribution for multijet events obtained with the \( \tau_0 \) fake-rate method is found to agree with the \( m_T \) spectrum obtained with nominal \( \tau_0 \) isolation criteria in this control region.

Figure 1: (a–g) The multijet template (red dashed line), the EWK+\( t \bar{t} \) template on top of the multijet template (blue solid line), and the data distribution in the nominal region (black dots) for deriving the number of multijet events in the nominal region in bins of \( p_T^{\tau_0} \). (h) The fake-rate probability in \( p_T^{\tau_0} \) bins. The vertical bars represent statistical uncertainty, and the horizontal bars denote the binning.
Figure 2: Comparison of shapes of the transverse mass distributions from the isolated (filled histogram) and anti-isolated (solid line) \( \tau_h \) samples in the inverted b tagging control region. The vertical bars and light grey area represent statistical uncertainty, and the horizontal bars denote the binning.

2.2 Measurement of the “EWK+\( \tau \bar{t} \)” background

The \( m_T \) distribution for the “EWK+\( \tau \bar{t} \)” background is modelled via the so-called \( \tau \) embedding technique. A control sample of \( \mu + \geq 3 \text{ jets} \) is selected by requiring one muon (\( p_T > 41 \text{ GeV}, |\eta| < 2.1 \)) and at least three jets (\( p_T > 30 \text{ GeV}, |\eta| < 2.4 \)). Events with additional muons or electrons are rejected. The selected \( \mu \) is replaced by a simulated \( \tau \) lepton decay, where the \( \tau \) lepton momentum is set to be the same as the momentum of the \( \mu \). The reconstructed \( \tau \) decay products are merged with the original \( \mu + \text{jets} \) event, from which the reconstructed muon is removed. These hybrid events are then subjected to the same event selection as the the \( \tau_h + \text{jets} \) events, i.e. \( \tau_h \) identification, b tagging, \( E_T^{\text{miss}} \) requirement, and the azimuthal angle selections.

The normalization of the “EWK+\( \tau \bar{t} \)” background needs to be corrected for the muon trigger and identification efficiencies, and for the \( \tau + E_T^{\text{miss}} \) trigger efficiency. The efficiency of the \( \tau \) part of the trigger is taken into account by weighting the events, while the \( E_T^{\text{miss}} \) part is included by approximating the \( E_T^{\text{miss}} \) in the trigger with a hybrid calorimetric \( E_T^{\text{miss}} \) constructed from the original event and the simulated \( \tau \) lepton decay. In a fraction of the selected \( \mu + \text{jets} \) events the \( \mu \) originates from a decay of a \( \tau \) lepton, leading to an overestimation of the background by a few percent. The bias is corrected for using a \( p_T^{\mu} \)-dependent event weights derived from simulated \( \tau \bar{t} \) events. A difference is seen in the transverse mass distribution between non-embedded \( \tau_h + \text{jets} \) and embedded \( \mu + \text{jets} \) events in simulated \( \tau \bar{t} \) events, which constitutes about 85\% of the “EWK+\( \tau \bar{t} \)” background. The difference is corrected for with \( m_T \)-dependent event weights derived from simulated \( \tau \bar{t} \) events.
To validate the method, simulated $\mu$+jets events are passed through the embedding procedure and compared to the sum of the expected events from the simulation with the $\tau_h$+jets selection and genuine $\tau_h$. The comparison is shown in Fig. 3, where the $m_T$ distribution from the embedded events is found to be in agreement with the same distribution from the $\tau_h$+jets events with genuine $\tau_h$. In this comparison only the uncertainties that affect each event sample independently of the other are considered.

![Figure 3: The $m_T$ distribution for the “EWK+$t\bar{t}$ with $\tau_h$” background for validating the embedding procedure. The simulated $\mu$+jets events are passed through the embedding procedure (solid line) and compared to the simulated $\tau_h$+jets with a genuine $\tau_h$ (filled histogram). Only the systematic uncertainties that affect only one of them are shown here. The hatched area, light grey area, and vertical black bars depict the statistical uncertainty, while the dark grey area and the brackets at the end of the vertical black lines represent the total uncertainty. The horizontal bars denote the binning.](image)

3. Background measurements in the $\mu+$\tau_h final state analysis

3.1 Measurement of the misidentified $\tau_h$ background

The background from events where a jet is misidentified as the $\tau_h$ is measured with a fake rate technique. A probability for a jet to be misidentified as a $\tau_h$ (jet-to-$\tau_h$ probability) is measured from data separately for W+jets events and 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 are triggered with a single isolated muon trigger ($p_T > 20$GeV and $|\eta| < 2.1$), and the events are required to have at least one jet ($p_T > 20$GeV and $|\eta| < 2.4$) and the transverse mass reconstructed from the muon and $E_T^{\text{miss}}$ is required to be greater than 50GeV. The
Figure 4: The probability of a jet to be misidentified as a $\tau_h$ as a function of the jet (a) $p_T$ and (b) $\eta$ evaluated from W+jets type events. The estimated fake rate is shown as red dots, and the fake rate from simulation truth is shown as blue squares. The horizontal bars denote the binning.

multijet events are triggered with a single jet trigger ($p_T > 30\text{GeV}$), and are required to have at least two jets ($p_T > 20\text{GeV}$, $|\eta| < 2.4$). All jets except the one firing the trigger are used in the misidentification rate calculation in order to avoid bias from the trigger. However, if two or more jets fire the trigger, all jets are used to calculate the misidentification rate. The jet-to-$\tau_h$ probability was parameterized as a function of the jet $p_T$, $\eta$, and radius $R = \sqrt{\sigma_{\eta\eta}^2 + \sigma_{\phi\phi}^2}$ using the k-Nearest Neighbour (kNN) regression [14]. The misidentification rates are shown in Fig. 4 for W+jets events and multijet events.

The misidentified $\tau_h$ background is then estimated as follows. A set of $\mu + \geq 3$ jets events is selected by requiring 1 isolated muon, $E_T^{\text{miss}}$, at least three jets, and at least one b-tagged jet. The selection thresholds are the same as in the signal event selection [12, 13]. This sample is dominated by W+jets and $t\bar{t} \rightarrow \mu \nu + q\bar{q}'$ events. The events are weighted by applying the jet-to-$\tau_h$ probability to each jet in the events. The selected events contain a small contribution from genuine $\tau_h$ events. The amount of these events is estimated using simulation and subtracted. The raw estimates are weighted to take into account the quark and gluon jet composition in these $\mu + \geq 3$ jets events.

3.2 Measurement of the $Z \rightarrow \tau\tau \rightarrow \tau_h\mu$ background

The background from events with $Z \rightarrow \tau\tau \rightarrow \tau_h\mu$ decays, where the hadronically decaying $\tau$ is identified as the $\tau_h$, and the $\tau$ decaying to $\mu$ is identified as the $\mu$, is measured with tau embedding technique. The shape of the b-tagged jet multiplicity distribution is taken from $Z \rightarrow \mu\mu$ events in data where both muons have been replaced with reconstructed particles from simulated $\tau$ lepton decays. The normalization of the distribution is taken from simulation of $Z \rightarrow \tau\tau$ events.
4. Summary

The data-driven background measurement methods employed in the searches for a charged Higgs boson in the CMS experiment were reviewed. The fake-rate technique is used for measuring the multijet background in the $\tau_h+$jets final state analysis, and for measuring the background from jets misidentified as the $\tau_h$ in the $\mu+\tau_h$ final state analysis. The tau embedding technique is employed for the measurement of the “EWK+$t\bar{t}$ with $\tau_h$” background in the $\tau_h+$jets final state analysis, and for measuring the background from $Z\rightarrow\tau\tau\rightarrow\tau_h\mu$ events in the $\mu+\tau_h$ final state analysis.

References


[8] CMS Collaboration, Search for charged Higgs bosons with the $H^+ \rightarrow \tau^+ \nu_{\tau}$ decay channel in the fully hadronic final state at $\sqrt{s} = 8$ TeV, CMS Physics Analysis Summary CMS-HIG-14-020 (2014).

http://cds.cern.ch/record/1950346


http://cds.cern.ch/record/1728343


http://cds.cern.ch/record/1755203
