KIT-Fakultät für Physik
Institut für Experimentelle Teilchenphysik
Prof. Dr. Ulrich Husemann
Dr. Matthias Schröder
Daniela Schäfer, Michael Waßmer

Karlsruher Institut für Technologie

Übungen zu Teilchenphysik I

Wintersemester 2017/18

Übungsblatt Nr. 11 Bearbeitung bis 25. Januar 2018

Search for New Physics at CMS

Data-Driven Background Estimation

Part II

This is the second part of a two parts exercise. It requires the results of the first part as input.

1 Introduction

As we have seen in the previous section, an important SM background arises from the production of W bosons in association with jets ($\mathrm{W}(l\nu) + \mathrm{jets}$) and from the production of top-antitop quark pairs ($\mathrm{t\bar{t}} + \mathrm{jets}$). We will now study closer in what cases these events can pass the baseline selection and thus contribute to the signal region.

Since top quarks decay almost exclusively to W bosons and bottom quarks, both processes lead to the presence of W bosons and jets in the event. The W bosons decay either into a quark-antiquark pair (*hadronically*) or into a lepton-neutrino pair (*leptonically*). In the first case, there is only very little $\not{H}_\mathrm{T}$ in the event, produced only from jet mismeasurements. Hence, the events are efficiently rejected by the $\not{H}_\mathrm{T}$ and $\Delta\phi$ selection criteria (steps 4 and 5 of the baseline selection). In the latter case of leptonically decaying W bosons, the events are to first approximation also rejected because events with isolated leptons are rejected in the baseline selection (step 1).

There are two important cases, however, in which the lepton veto fails, and hence, $\mathrm{W}(l\nu)+\text{jets}$ and $\mathrm{t\bar{t}}+\text{jets}$ events enter the search region:

- *Lost lepton*: The lepton from the W decay is not reconstructed due to either
  - the limited geometric and kinematic **acceptance** of the detector (there is no detector installed at large $|\eta|$, only leptons with $p_\mathrm{T} > 10\,\mathrm{GeV}$ are considered); or
  - the inefficiency of the **reconstruction** algorithm; or
  - the lepton is **not isolated** because it geometrically overlaps with a jet.
- *Hadronic τ*: The W boson decays into a τ lepton that decays to hadrons that form a jet.

In both cases, there will not be a well-reconstructed and isolated lepton present in the event such that the lepton veto is passed. At the same time, there can be sufficient $\not{H}_\mathrm{T}$ to pass the $\not{H}_\mathrm{T} > 200\,\mathrm{GeV}$ criterion caused by the neutrinos from the W (and subsequent τ) decays.

- **Question: 1.1.1** What does lepton isolation mean? Why do we only consider isolated leptons when applying the lepton veto?

In this analysis, the $\mathrm{W}(l\nu)+\text{jets}$ and $\mathrm{t\bar{t}}+\text{jets}$ backgrounds are determined via two methods which address separately the lost-lepton and the hadronic-τ case, cf. Fig. 1. Here, we will discuss the 'lost-lepton (prediction) method' because it is a good examples of data-driven background prediction methods on which many important analyses at the LHC rely, in particular searches for new physics, and the addressed background processes are important for this SUSY search. The method is described in full detail in [1].

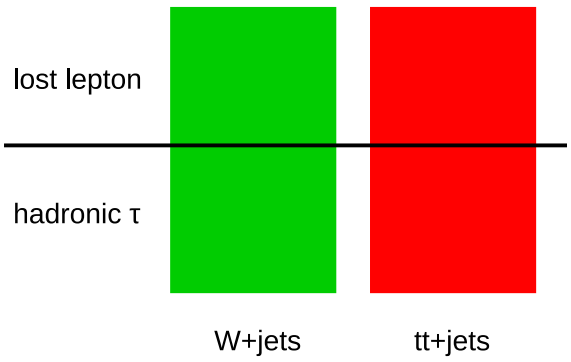

Figure 1: Illustration of the lost-lepton and hadronic-τ contributions to the $\mathrm{W}(l\nu)+\text{jets}$ and $\mathrm{t\bar{t}}+\text{jets}$ backgrounds.

Remember: Both the lost-lepton and the hadronic-τ method are used to predict the background expected from $\mathrm{W}(l\nu)+\mathrm{jets}$ and $\mathrm{t\bar{t}}+\mathrm{jets}$ events, but each method predicts different parts of the backgrounds that occur due to different effects. The total number of $\mathrm{W}(l\nu)+\mathrm{jets}$ and $\mathrm{t\bar{t}}+\mathrm{jets}$ background events is obtained by adding the predictions from the lost-lepton and the hadronic-τ methods.

2 Preparation: Setting up the Code

The code and input files for this exercise are provided in

`/home/staff/mschrode/TP1_WS17/susy/susy_ex_part2.tar.gz`

They require the code and input files of the first part of the exercise (Übungsblatt 10) and have to be integrated into that setup: Copy the archive `susy_ex_part2.tar.gz` to your working area `susy_ex` of the previous exercise and extract it there. This will add the file `HadTau_Data_Prediction.root` as well as the directories `LostLepton` and `Results`. Afterwards, your setup should look like (output of `ls -1 susy_ex`):

```
data
General
HadTau_Data_Prediction.root
LostLepton
Result
Utils
```

We will reuse the results of the previous exercise in Section 4, so please do not delete any of the old results.

3 Lost-Lepton Method

This method aims at determining the number of $\mathrm{W}(l\nu)+\mathrm{jets}$ and $\mathrm{t\bar{t}}+\mathrm{jets}$ events in which the lepton escapes identification, and hence, does not trigger the lepton veto. This happens because the lepton is either out-of-acceptance, not well-reconstructed, or not isolated, as illustrated in Fig. 2.

The procedure starts from events with exactly one well-reconstructed, isolated muon. For each event, the search variables $\mathrm{N_{jets}}$, H_T, and $\not{H}_\mathrm{T}$ are computed and the baseline selection steps *except for the lepton veto* are applied. The sample of events surviving this selection forms our *control sample*: these are events that almost entirely originate in $\mathrm{W}(l\nu)+\mathrm{jets}$ and $\mathrm{t\bar{t}}+\mathrm{jets}$ production. Then, the expected number of background events, i.e. the number of events surviving the full baseline selection including the lepton veto, is estimated from the control sample by essentially weighting

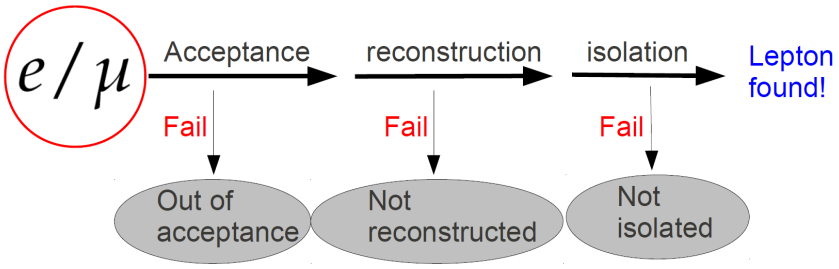

Figure 2: Illustration of the lost-lepton background: if the lepton is either out-of-acceptance, not well-reconstructed, or not isolated, it will not be identified. Hence, the event passes the lepton veto and contributes to the background.

each event by the probability w that the lepton survives the lepton veto. In this way, we obtain the full kinematic properties of the background events directly from data and do not need to worry about its simulation!

The crucial ingredient to the method is the lost-lepton probability w. This probability $w(\alpha, \epsilon_{\text{reco}}, \epsilon_{\text{iso}})$ depends on the probability α that the lepton is within the geometric and kinematic acceptance as well as on the lepton reconstruction and isolation efficiencies ϵ_{reco} and ϵ_{iso}, respectively.

As an example, consider for the time being only the acceptance and ignore the reconstruction and isolation. In this case, an event contributes to the lost-lepton background if the lepton is out of acceptance. The probability that the lepton is out of acceptance is $(1-\alpha)$, and thus, if the true number of $\text{W}(l\nu)+\text{jets}$ and $\text{t}\bar{\text{t}}$ events is N_{true}, we expect

$$N = (1-\alpha) \cdot N_{\text{true}}$$

lost-lepton background events. However, N_{true} is not equal to the number N_{CS} of events in the control sample, because the control sample itself was selected requiring one lepton to be measured in the event, i.e. requiring one lepton to be within acceptance. Hence, it is

$$N_{\text{CS}} = \alpha \cdot N_{\text{true}} \,.$$

Thus, we can predict the number of lost-lepton background events N from the measured control sample via

$$N = \frac{(1-\alpha)}{\alpha} \cdot N_{\text{CS}} \,,$$

where the lost-lepton probability is given by

$$w(\alpha) = \frac{(1-\alpha)}{\alpha} \,. \tag{1}$$

In order to generalise this approach to take into account also the reconstruction and isolation, it is important to precisely define α, ϵ_{reco}, and ϵ_{iso}. We will use the following definition:

- α: the probability that a lepton is within the geometric and kinematic acceptance of the detector;
- ϵ_{reco}: the probability that a lepton that is within acceptance is also reconstructed; and
- ϵ_{iso}: the probability that a lepton that is in acceptance and reconstructed is also isolated.

Using these definitions and considering the three cases in which the lepton passes the veto, answer the following

- **Question 3.1** How does the weight $w(\alpha, \epsilon_{\text{reco}}, \epsilon_{\text{iso}})$ depend on α, ϵ_{reco}, and ϵ_{iso}?

3.1 Acceptance and efficiency determination

We will now learn how α, ϵ_{reco}, and ϵ_{iso} can be determined from MC. Here, we will do this only for muons using W + jets events. This includes both direct decays of the W boson to a muon and decays of the W boson to a τ lepton which then decays to a muon. In the actual analysis, the numbers are determined also for electrons, of course, but it works analogous to the muon case and one does not learn anything new from it.

Before starting the implementation, think about the following questions.

- **Question 3.1.1** How would you technically determine α, ϵ_{reco}, and ϵ_{iso}?
- **Question 3.1.2** Would you apply any event selection to the simulated W+jets events?
- **Question 3.1.3** Is it sufficient to consider the W + jets sample?

Please go the `LostLepton` directory in your `susy_ex` working area. Have a look at the script `lostLepton1.C`. It is an example of how to determine the muon acceptance α and the muon-reconstruction efficiency ϵ_{reco} from the W + jets sample (`id=11`). Investigate how this is implemented.

- What event variables are being used?
- How are the acceptance and efficiency computed and stored?

Now, execute

```
root -l -b -q lostLepton1.C+
```

By default, this script runs over the W + jets sample. The produced acceptance and efficiency maps are stored in `LostLepton_MuonEfficienciesFromWJetMC.root` in `susy_ex/data` (as ROOT `TH2` objects). Investigate the result; you can conveniently execute the plotting script

```
root -l plotMuonEfficiencies.C+
```

for this.

As you can see (and have already seen when investigating the `lostLepton1.C` script), the efficiencies are parametrized as functions of N_{jets} and $\not{H}_T$ and, in case of ϵ_{reco} and ϵ_{iso}, also H_T. In general, the efficiencies depend on the lepton kinematic quantities and also the event topology, e.g. the ϵ_{iso} is small if the lepton is close to a jet. However, it turns out that parametrizing the efficiencies as function of lepton p_T and the distance to the closest jet leads to difficulties in determining the efficiencies, essentially because in the problematic regions at low p_T and close to the jet, there are too few events in the MC to properly determine the efficiencies. Instead, better performance has been obtained with a parametrization in the search variables N_{jets}, H_T, and $\not{H}_T$.

So far, we have only determined the acceptance α and the reconstruction efficiency ϵ_{reco}. Extend the `lostLepton1.C` script to also determine the isolation efficiency ϵ_{iso}. (The required histograms are already present, you just need to fill them.) Insert your code after the statement

```
//////// Implement isolation efficiency calculation here
```

following the instructions given there.

3.2 Validation of the method

We will now validate the lost-lepton method. We will do this by verifying that we can correctly predict the number of lost-lepton background events in the W + jets MC sample, where we know the correct answer. We use the simulated events because here we have the generator-truth information at hand, and thus, can compare the prediction to the truth. Thus, this is a consistency check of the method (often also referred to as *closure test*).

At first, we will validate the method for the simplified example discussed above with Eq. (1) considering only the muon acceptance. Have a look at `lostLepton2.C`. This is an example of how the prediction can be implemented. Note for example that the efficiency maps are automatically loaded and accessed via a `LeptonAcceptance` object and note the implementation of the lost-lepton weight Eq. (1) in line 94.

The script also shows how to obtain the true number of lepton-out-of-acceptance events against which we want to validate our prediction. Note that the latter and the control sample are a partition of all events.

Now, execute the script

```
root -l -b -q lostLepton2.C+
```

which again by default runs over the W+jet sample. Its output, the predicted and the true H_{T}, $\not{H}_{\mathrm{T}}$, and $\mathrm{N}_{\mathrm{jets}}$ distributions for out-of-acceptance-muon events, are written to the file `LostLepton_ClosureMuonAcceptance.root`. They can be conveniently plotted and compared using the script `plotClosureMuons.C`,

```
root -l plotClosureMuons.C+\(\"LostLepton_ClosureMuonAcceptance.root\"\)
```

- **Question 3.2.1** Are the predicted yields compatible with the truth?
- **Question 3.2.2** Which uncertainties do the printed error bars represent? Are these sufficient?

Now, we will perform and validate the full lost-lepton method. Hence, we will perform the prediction as we would do it on real data and compare to the expectation from the MC truth. Have a look at `lostLepton3.C`. Here, the full data-driven prediction from the muon control sample for the number of lost lepton events is implemented. Note how the prediction does not rely on any MC-truth based quantities! The essential part is the computation of the lost-lepton weight

```
const double llw = lostLeptonProb(muAcc,muEffReco,muEffIso)
                 * controlSampleCorr(muAcc,muEffReco,muEffIso);
```

which is the generalisation of Eq. (1) taking into account also the reconstruction and isolation efficiencies. However, so far the two functions `lostLeptonProb` and `controlSampleCorrection` are dummies. Implement them, following your solution to Question 3.1!

Then, run the script:

```
root -l -b -q lostLepton3.C+
```

It will produce the output `LostLepton_ClosureMuon.root`. Again, prediction and truth can be compared using the script `plotClosureMuons.C`,

```
root -l plotClosureMuons.C+\(\"LostLepton_ClosureMuon.root\"\)
```

Discuss the result.

- **Question 3.2.3** Could the same muon control sample be used to estimate the contributions from the electron channel? What would be the weight factor in this case?

3.3 Background prediction from data

Finally, we will perform the full lost-lepton prediction on data! However, we should not use our efficiency maps because they were determined only for lost muons and only from $W + \text{jets}$ events while the full lost-lepton background uses also $t\bar{t} + \text{jets}$ events and, of course, lost electrons. Therefore, we would have to determine the acceptance and efficiency maps also for electrons and from a realistic mixture of $W + \text{jets}$ and $t\bar{t} + \text{jets}$ events.

We will not do this here, because we would not learn anything new. Instead, the efficiency maps are provided in `susy_ex/data/LLEff-SUS-13-012.root`. Use these maps to perform the lost-lepton prediction on the data sample (`id=1`). You can use the `lostLepton4.C` script, which is an implementation of the lost-lepton method for data:

```
root -l -b -q lostLepton4.C+
```

The result of the prediction can be plotted via

```
root -l plotPrediction.C+
```

which also prints the predicted number of events after the baseline selection.

Let us consider some more important aspects about the lost-lepton method.

- **Question 3.3.1** Recap how we define and predict the lost-lepton background. Are we missing any contributions with our current method?
- **Question 3.3.2** Could potential new-physics events in data bias our prediction? How? What could be done to mitigate such an effect?

Finally, let us think about the systematic uncertainties we would assign to our prediction. In fact, without proper uncertainties the result is just meaningless!

- **Question 3.3.3** What effects do you think impact the prediction and should be considered as a source of systematic uncertainties? How would you determine the size of the uncertainties?
- **Question 3.3.4** In this analysis, the acceptance and efficiencies are determined from a MC simulation. (Essentially, we replace the uncertainty one would get due to using MC simulation to determine the background by the uncertainty on the simulation of the acceptance and efficiencies.) Can you think of a way to verify the numbers with data?

4 Results and Interpretation

Finally, let us compare the data with our data-based background prediction.

The lost-lepton background was one part of the W($l\nu$)+jets and $t\bar{t}$+jets background, the other given by the hadronic-τ background that has not been discussed here. However, we need this part to predict the full W($l\nu$)+jets plus $t\bar{t}$+jets background, cf. Fig. 1; therefore, the results of a simplified hadronic-τ method are provided in `susy_ex/HadTau_Data_Prediction.root`. We can use this together with our lost-lepton result and replace the corresponding simulated predictions in the data-vs-background plot we produced during part I of this exercise.

To do so, please go to the `susy_ex/Result` directory and

- copy the ROOT files with the H_T, $\not{H}_T$, and N_{jets} distributions observed in data and predicted for the background and signal using simulation in this directory:

  ```
  cp ../General/General_*-Yields.root .
  ```

- We do not want to use the W($l\nu$)+jets and $t\bar{t}$+jets predictions from simulation. Therefore, delete them! Instead, copy our data-based lost-lepton prediction as well as the provided hadronic-τ prediction in this directory:

  ```
  cp ../LostLepton/LostLepton_Data_Prediction.root .
  cp ../HadTau_Data_Prediction.root .
  ```

Now you can execute

```
root -l plotDataVsBkg.C+
```

This will compare the H_T, $\not{H}_T$, and N_{jets} distributions spectra observed in data to the expected SM background contributions. As you can see, instead of the individual W($l\nu$) + jets and $t\bar{t}$ + jets processes, the contributions from the lost-lepton and the hadronic-τ background are shown — our results measured from data! For comparison, also the signal expectations are superimposed.

- **Question 4.1** Do you observe any sign of new physics? Is this actually a valid comparison of data and expected backgrounds?
- **Question 4.2** What uncertainties have to be considered for the shown signal expectations? Are these uncertainties also relevant for the backgrounds?

References

[1] A.-R. Dräger, "Prediction of the $t\bar{t}$ and W + Jets Background in a Search for New Physics with Jets and Missing Transverse Energy at CMS". PhD thesis, Universität Hamburg, 2016. DESY-THESIS-2016-005.