# Introductory School to Terascale Physics
# Tutorial on Top Quarks

Prof. Dr. Wolfgang Wagner, Dr. Dominic Hirschbühl, Simon Köhlmann

## Reconstructing Top Quark Events in CDF Data

We are interested in $t\bar{t}$-events decaying via:

| | | |
|---|---|---|
| $t \to bW^+ \to b\ell^+\nu_\ell$ | with $\ell = e, \mu$ | semileptonic |
| $\bar{t} \to \bar{b}W^- \to \bar{b}q\bar{q}$ | with $q = u, d, s, c$ | hadronic |

The charge conjugate states are always implied.

### Setting up

We will analyze the data contained in the directory
`/intro-school/TopTutorial/Samples`.
The root files we use, are:

- `data.root`
  containing measured preselected lepton + jets events (CDF data)
- `tt175.root`
  containing $t\bar{t}$ Monte Carlo (MC) events ($m_t = 175$ GeV/$c^2$)
- `Wbb.root`
  containing background Monte Carlo events
- `tt165.root`
  containing $t\bar{t}$ Monte Carlo events ($m_t = 165$ GeV/$c^2$)

All Monte Carlo samples were passed through a full CDF detector simulation.

- Go to the directory `/intro-school/TopTutorial/` and unpack the file `ttbar.tgz` with the command:
  `tar xzf ttbar.tgz`.

  The Directory `TTbarTutorial` is produced, which contains the files for producing the histograms (`runTTbar.cc`), plotting the histograms (`plotHistos.C`), the Makefile and a shared library (`libTopObjects.so`). In addition the directory contains one sub directories: `results`.
  Go to the directory `TTbarTutorial` with the command:
  `cd TTbarTutorial`

- Before you we can start you have to set up ROOT. This can be done with `source setup_root.sh`. Check if ROOT is set up properly by executing `root`.

- In the main directory you will find the file `runTTbar.cc`, where you can implement your analysis code to select $t\bar{t}$ candidate events. This file has to be compiled and linked with the command `make`. It produces the executable `runTTbar.exe`. Simply execute the command:
  `make runTTbar`

## Exercises

1. **Number of high energetic jets**

Plot the number of high energetic jets of the preselected data sample ($N_{jets} \geq 1$) and of the $t\bar{t}$ Monte-Carlo. How do you interpret this plot? What cut on the variable $N_{jets}$ would you apply to look for $t\bar{t}$-events in measured data?

*Instructions:*

- Run the executable and specify the file you want to analyze. Please omit the file extension `.root` of the file:
`./runTTbar.exe data`
`./runTTbar.exe tt175`
The histograms will be automatically written into the results directory. The histogram for njets is already filled in `runTTbar.exe`.
- Open the resulting file with the ROOT browser:
Start root and load the file automatically:
`root results/dataHistos.root`
Start the browser with:
`new TBrowser`
Click on ROOT Files and click on the file. A folder will open, which contains all histograms. By clicking on them the histograms will show up.
- To use `tt175.root` change line 15 in `plotHistos.C`
`./results/tt165Histos.root` to `./results/tt175Histos.root`
- Using the file `plotHistos.C` you can read in the histogram files and draw them. This file is only a root script. You can execute it with:
Start root: `root`
Load the file: `.L plotHistos.C`
Execute it: `.x plotHistos.C(1)`
You can also use: `root.exe plotHistos.C\(1\)`
The plot is stored as ps-file in the directory `results`.

2. **Study of Event Shape Variables**

In this exercise we study event shape variable that could be used to discriminate background ($W + b\bar{b}$) and $t\bar{t}$ signal Monte-Carlo.

a) As the first step select events with at least three jets ($N_{\text{jets}} \geq 3$).
*Instructions:*
Add a selection cut, to select $N_{\text{jets}} \geq 3$. You have to modify `runTTbar.cc` accordingly and recompile. The code for the selection is already there, it only has to be commented out. You have to execute `runTTbar.exe` for the $t\bar{t}$ Monte Carlo (option tt175) and for the $W + b\bar{b}$ background (option Wbb). In order to check, if the selection works, plot again the njet distribution with: `plotHistos.C`, with option 1.For example by

```
root.exe plotHistos.C\(1\)
```

b) Now look at the variables

- $H_T$, the total transverse energy in the event,
- the missing transverse energy,
- the transverse momentum and
- the pseudorapidity of the reconstructed charged lepton.

Plot the distributions for background and signal in one histogram. Which variable has the largest discrimination power? The histograms for the four quantities are already declared and filled in `runTTbar.cc`.

Then draw the histograms of the event shape variables using `plotHistos.C`, with option 2.

3. **Extraction of the $t\bar{t}$ Signal ($N_{\text{jets}} \geq 4$)**

   Use the variable $H_T$ as a discriminating variable in a template fit to determine the ttbar and the background fraction in the observed data. What $t\bar{t}$-signal fraction do you get? How can you improve that?

   *Instructions:*
   Change the selection to $N_{\text{jets}} \geq 4$ and recompile. Execute `runTTbar.exe` for the data, signal (option tt175) and background Monte Carlo (option Wbb).
   Check the selection of $N_{jets} \geq 4$ events by looking at the $N_{jets}$ plots: `plotHistos.C(1)`

   *Explanation of fitting code:*
   The code for the fit can be found in `HTfit.h`. The expectation for each bin $i$ consist of two parts, namely the $t\bar{t}$ signal contribution
   `nData*par[0]*ValSig[i]`
   and the background contribution `nData*(1-par[0])*ValBkg[i]`. Here `nData` is the total number of entries of the data, `par[0]` is a free parameter and gives the $t\bar{t}$-signal fraction. `ValSig[i]` and `ValBkg[i]`, respectively, are the to unity normalized signal and background distributions. These arrays and also the corresponding arrays for the data are filled from the histograms you produced before. In order to get that theory prediction, which describes the data in the best way, the negative log-likelihood function $f$ has to be minimized:

$$
\begin{aligned}
f \quad = \quad & \sum_{i}^{\texttt{nbins}} 2 \cdot (\texttt{expectation-ValData[i]} \cdot \ln(\texttt{expectation}) + \\
& \texttt{ValData[i]} \cdot \ln(\texttt{ValData[i]}) - \texttt{ValData[i]})
\end{aligned}
$$

   In the code, $f$ is labeled with `fMin`. `ValData[i]` is the number of data events in bin $i$ and `nbins` is the number of bins used for the fit. The fit itself is then performed by the ROOT class `TMinuit`.

   *Instructions:*
   Execute `plotHistos.C` with option 3. The result of the fit will be automatically visualized.

4. **Improvement of the Signal to Background Ratio ($S/B$)**

   Extract the $t\bar{t}$-signal for the three additional cases:

   (a) $N_{\text{jets}} \geq 4$ and no $b$-tag (was done in section 3)
   (b) $N_{\text{jets}} \geq 3$ and at least one jet which has a $b$-tag

(c) $N_{\text{jets}} \geq 4$ and at least one jet which has a $b$-tag

How does the $S/\sqrt{B}$ ratio change? For which scenario is the significance $S/\sqrt{B}$ largest (useful for the discovery and cross section measurement)? Which scenario is the best for the determination of the top mass (clean sample needed)?

| cut scenario | $N_{\text{data}}$ | sig. fraction | $S/B$ | $S/\sqrt{B}$ |
|---|---|---|---|---|
| (a) | | | | |
| (b) | | | | |
| (c) | | | | |

*Instructions:*
Change the event selection for each test case and execute `runTTbar.exe` for the data, signal and background Monte Carlo, respectively.
The information of the number of $b$-tagged jets is stored in `ev->nbtags`. Execute `plotHistos.C` with option 3 for each scenario and fill the results in the table above.
You get a calculator with: `xcalc &`
To get $N_{\text{data}}$ you can edit `plotHistos.C`. Write at the end of the (exercise 3) loop

```
double nData = hData -> GetEntries();
cout << nData << endl;
```

5. **Reconstructed Top Mass Distribution**

To measure the top-quark mass $m_t$ one needs to find an observable that is sensitive to $m_t$. The easiest way to calculate a distribution, which is highly correlated with $m_t$, is to calculate the invariant mass of the three jets with the highest transverse momentum, the so-called tri-jet mass $m_{3j}$. For signal events this combination represents the hadronically decaying top quark. Why is the mean of this distribution higher than the actual top quark mass of the used Monte Carlo sample of 175 GeV/$c^2$?

*Instructions:*
For the determination of the mass the cut scenario (c) is used ($N_{\text{jets}} \geq 4$) and at least one jet, which has a $b$-tag.
Calculate the invariant mass of the three jets with the highest $p_T$ by adding the `TLorentzVector`s of the first 3 jets of the vector `jets`. They are already ordered py $p_T$. The invariant mass of the resulting `TLorentzVector` can be calculated with the member function `.M()`. After compiling the code you have to execute `runTTbar.exe`

with the parameter tt175. Then, you can plot the histograms using the ROOT script `plotHistos.C` with the parameter 4.

6. **Determination of the True Top Quark Mass**

Determine the true top mass in the data from the reconstructed tri-jet mass. To get a "clean" distribution the background has to be subtracted before. Since the reconstructed tri-jet mass is not exact the true top quark mass, one has to find a calibration. We assume that a linear dependence between the mean of $m_{3j}$ and the true top quark mass is a good approximation.

$$m_{t,true} = \alpha \cdot m_{3j} + \beta$$

Determine the mean of the reconstructed tri-jet mass for two different true top quark masses (165 GeV/$c^2$ and 175 GeV/$c^2$). Afterward calculate the two coefficients for the linear dependence. What is the result for $m_{t,true}$ in the data?

*Instructions:*
Execute `runTTbar.exe` with `tt175, tt165, Wbb, data`. The mean value of the tri-jet mass distribution is displayed in the plot, if you execute `plotHistos.C` with option 4.
To obtain, the correct value in data, at first you have to subtract the correctly scaled background from the data histogram. Use the signal fraction obtained in exercise 1.5. Add code for this in `plotHistos.C` for the already prepared option 5. Look for the variable `sigFrac`. After subtraction obtain the mean value and calculate the measured top quark mass. This is done automatically if you provide the mean values for the both $t\bar{t}$ samples. Look for the variables `m165` and `m175` and give them the obtained values.

Check the public CDF webpages
`http://www-cdf.fnal.gov/physics/new/top/public_mass.html`
for official top quark mass measurement in the the lepton + jets channel.

## 7. Appendix: Data Structure of a Root Tree

| Variable Name | Type | Definition |
|---|---|---|
| `run_number` | int | Number of run |
| `event_number` | int | Number of event |
| **Charged lepton variables ($e$ or $\mu$) - reconstructed** | | |
| `primLep.v4` | TLorentzVector | Four-momentum of the charged lepton |
| `primLep.id` | int | Electron $\rightarrow$ 1; muon $\rightarrow$ 2 |
| `primLep.charge` | double | Charge of lepton |
| **Missing energy - reconstructed** | | |
| `met` | TVector2 | Missing transverse energy vector (Corresponds to px, py of neutrino) |
| **High energetic jets - reconstructed** | | |
| `njets` | int | Number of high energetic jets in event |
| `nbtags` | int | Number of tagged jets in event |
| `jets` | std::vector<TopJet> | Vector containing all jets in the event |
| `jets[i].btag` | int | b-tag information for each jet (array of `njets` entries) |
| `jets[i].v4` | TLorentzVector | Four-momentum of each jet (array of `njets` entries) |

Table 1: Description of the used preselected trees.