Integration of ROOT Notebooks as a Web-based ATLAS Analysis tool for public data releases and outreach

August 19, 2016

Student: Tea Banda
University of Donja Gorica

Supervisors: Dr. Arturo Rodolfo Sanchez Pineda
University of Naples "Federico II" and INFN

Dr. Francesco Conventi
University of Naples "Parthenope" and INFN
ABSTRACT

The project consists in the initial development of ROOT notebooks for a Z boson analysis in C++ programming language that will allow students and researches to perform fast and very useful data analysis, using ATLAS public data and Monte-Carlo simulations. Several tools are considered: ROOT Data Analysis Framework, Jupyter Notebook Technology and CERN-ROOT computing service so-called SWAN.
CONTENTS

1 Introduction 4

2 Integration of notebooks 4
   2.1 Z Analysis ................................. 4
   2.2 Development of the Notebook .................. 5
       2.2.1 Defining the chains, adding data to chains and defining variables 5
       2.2.2 Analysis (filtering data) .................. 6
       2.2.3 Filling the histograms .................... 7
       2.2.4 Scale the histograms ...................... 7
   2.3 Complete Analysis Plots ........................ 7

3 Conclusions 8
1 INTRODUCTION

The ATLAS collaboration [1] is releasing an official dataset, open to the public for educational use. The dataset consists of real data with an integrated luminosity \( \approx 1 \text{ fb}^{-1} \) and a centre-of-mass energy of \( \sqrt{s} = 8 \text{ TeV} \) paired with matching simulated data.

The dataset is intended to provide the means for doing hands-on particle physics exercises in the context of higher education, e.g., laboratory courses or introductory exercises for undergraduate students as well as provide data for teaching materials that may be used in lectures and public talks. Furthermore, it may be used by people with a data analysis but not necessarily a physics background as a test dataset for studying and developing analysis techniques as was done in the Kaggle Higgs Boson Machine Learning Challenge[2].

Together with datasets, ATLAS Collaboration is releasing a set of analysis examples written in Python for educational purposes.

The main idea of the project I was working on is the representation of one of the physics analysis that look for the reconstruction of the Z boson [3], but in different programming language and platform. The project is consisted of the development of that analysis in C++ programming language using ROOT [4] as a framework and Jupyter [5] notebook technology. It contains detailed steps of the whole process, from starting points to producing final plots. Each step is explained in the following sections.

2 INTEGRATION OF NOTEBOOKS

The integration of the ROOT data analysis framework with the Jupyter Notebook technology presents the potential of enhancement and expansion of educational and training programs.

One representation of this new approach of interactive data analysis is the physics analysis developed during this project: the so-called Z analysis, that uses as inputs the latest ATLAS public release of data and Monte-Carlo(MC) [6] and that under C++ programming language allows to the user to review and understand how this kind of research is done. This analysis is part of a series of seven analysis that were released [3] together with the datasets. The overall aim is to get all the seven analysis into ROOT notebook in C++ and Python. My aim is to create one of firsts.

Through the report the development of the notebook is shown using example plots relative to the leading lepton variables.

2.1 Z ANALYSIS

Many analyses selecting leptons suffer from Z + jets as a contributing background due to its large production cross section. It is therefore vital to check the correct modelling of this process by the MC simulated data. It is important to measure well known Standard Model particles, to confirm that we understand properly the detector and software. We are then ready to search for
2.2 Development of the Notebook

The process of developing the notebook is consisted of six crucial steps:

- Defining the ROOT TChain (Figure 1) objects, those are necessary to get and merge the input samples for the single analysis. Once the chains are defined and the input obtained, we proceed to read and define the variables coming from the input that we will use later on. See Section 2.2.1 for some details about adding data to chains and defining variables that will be used.

- Defining set of histograms that will be filled per each variable of each object under consideration in the Z analysis. The number of histograms is proportional to the different input samples as well. This means that we are dealing with several 100's of histograms.

- The actual Analysis: here is where the requests and the conditions are applied to each event into the datasets. Section 2.2.1 shows how this process is done in the actual notebook.

- Filling the histograms: Once the conditions of the analysis have been fulfilled by the current event, the histograms corresponding to that particular sample are filled with the corresponding variables. For variables that have units of energy, the GeV is used. See section 2.2.3.

- Scaling the histograms: Because we know the simulated MC samples and their production rate, we can calculate the number of events that should be contained in the total real data in used, e.g. 1fb⁻¹. This procedure rescale de histograms of the MC samples to agree with the Standard model predictions and the total amount of data to be compared. See Section 2.2.4 for details how this procedure is performed.

- Create the analysis plots: This procedure pass for the merging of common MC smaples histograms into groups that will be used later to create staked plots to be compared with the real data histograms.

These are the major steps of developing. Each of them has a subset of actions needed to be done in order to get the precise results which will be explained in the next sections.

2.2.1 Defining the Chains, Adding Data to Chains and Defining Variables

First steps in developing the notebook are setting the data that will be used in analysis. The way to access the data is by making the chain in ROOT and adding data to it. One of advantages of this technology is that data can be used online, doesn’t need to be downloaded. As shown in the Figure 1. we create the MC chain and added simulated samples for Z analysis in it:
Now that we have access to datasets, the next step needed for setting the environment is to declare and initialise the variables we will need in order to fill the histograms. We set branches for each variable and read them from the files.

2.2.2 Analysis (Filtering Data)

In order to fill the analysis histograms we first need to apply the proper analysis selection criteria and calculate some weights (in the case of the MC samples) to proceed with the correct normalisation of the MC histograms. Histograms are being filled using FOR loop that goes through every entry. But, before the filling, each entry has to go through several selection criterias which are defined in the loop.

This is a Z boson analysis where the Z boson decays into a lepton pair. The standard object selection criteria are applied. The event selection criteria are:

- Single electron or muon trigger has fired;
- Event in real data passes the GRL;
- Event has a good vertex (N tracks > 4);
- Exactly two good leptons with pT > 25 GeV;
- Leptons have opposite charge;
- Leptons have same flavour;
- $|m_{ll} - m_Z| < 20$ GeV.

The first three event requirements or cuts in the list above are already applied into our input samples, so we defined and applied the next four requirements for filtering the datasets. As seen in the Figure 2.
First cut filters the data in a way that each event is requested to have at least two good leptons with \( p_T > 5 \) GeV. This is done using the condition: \( \text{lep}_n>1 \) because the leptons saved in this input samples already are filtered to have at least that value in \( p_T \).

Second cut asks that the two leading leptons in each entry have opposite charge. We get this by declaring the variable \( OS \) which will be the product of charges of the first and second lepton. Then we ask if that product is negative and only the entries with negative value in \( OS \) made the cut.

The third condition requires that dilepton invariant mass be close to the Z-boson mass, so that the difference between the two is less than 20 GeV.

In order to get the precise representation of data we need a weight factor for each histogram. We get the weight factor as a product of Scale Factor and Event Weight. These two variables are also read from the MC chain and declared with the rest of the variables in the very first steps of the notebook creation.
2.2.3 FILLING THE HISTOGRAMS

After the data is filtered and the cuts are made, the histograms can be filled, using the ROOT Fill function to feed those histograms (See Figure 3) with the values of the data that made all the analysis requirements described before:

```cpp
if(Montecarlo_runNumber == 147770) {
    float Monte Carlo lep pt_inGeV = MonteCarlo lep pt[0]/1000.;
    h lep pt Zee->Fill(MonteCarlo lep pt inGeV, weight);
    //ETA
    h lep eta Zee->Fill(MonteCarlo lep eta[0], weight);
    //PHI
    h lep phi Zee->Fill(MonteCarlo lep phi[0], weight);
    //E
    float Monte Carlo lep E_inGeV = MonteCarlo lep E[0]/1000.;
    h lep E Zee->Fill(MonteCarlo lep E inGeV, weight);
    //CHARGE
    h lep charge Zee->Fill(MonteCarlo lep charge[0], weight);
    //ETC20
    h lep etc20 Zee->Fill(MonteCarlo lep etccone20[0], weight);
    //PTC30
    h lep ptc30 Zee->Fill(MonteCarlo lep ptcone30[0], weight);
    //ZB
    h lep Z0 Zee->Fill(MonteCarlo lep Z0[0], weight);
    //DO
    h lep DO Zee->Fill(MonteCarlo lep DO[0], weight);
}
```

Figure 3. Filling the histogram

2.2.4 SCALE THE HISTOGRAMS

To proper scale the histograms we need to calculate the so-called scale factors that are specific for each of the simulated data file, (Figure 4)

```cpp
//SC Z
scaleFactorZee=(Luminosity*1241.2972)/(283795455568148.0*0.151816307);
scaleFactorZmmumu=(Luminosity*1241.2972)/(225316022111048.0*0.124706551);
scaleFactorZtautau=(Luminosity*1240.8988)/(31588548303680.9*0.921562294);
```

Figure 4. Calculate the Scale Factor

And then apply the scale factor to each histogram (Figure 5):
2.3 COMPLETE ANALYSIS PLOTS

Once all the analysis steps and individual histograms are set, we can produce final analysis composite plots. The final histograms corresponding to each of the physical variables to analyze are a composition or stack of several merged histograms that have been divided with respect to the Standard Model physical process (e.g. Z, W, WW-WZ-ZZ, top-antitop production, etc.) Finally this total MC histogram is compared with the same histogram-variable for real data to compare the agreement and possible inconsistencies between the theory and the experimental records.

The Figure 6 shows the composite plots of the physical variable: $p_T$ of the leading lepton. The plot on the left plot shows the plot produced in the analysis ROOT notebook while in the right is the plot coming from ATLAS Open Data public note [3].

![Figure 6. Histogram comparison](image-url)
3 CONCLUSIONS

The presented analysis is a part of much bigger project which have potential to expand and make a big difference in further education and training systems. Alone, it gives users a unique opportunity to explore and get a first hand experience in physics analysis of the latest data release of ATLAS experiment. Whole analysis is contained in one notebook, and it pushed the limits of SWAN efficiency. It causes slow reading of the notebook because of the huge amount of data that is being processed. The notebook currently contains around six hundred histograms, for each variable, which are further merged and stacked as into final plots that represent the final results of the analysis. The final step of the analysis which have to be made is adding real data. Since the project is still in development, and ATLAS team is working on it. At the final stage, the website will contain seven full analysis, available online, with all data accessible.

REFERENCES

[1] ATLAS Collaboration
https://atlas.cern

https://www.kaggle.com/c/higgs-boson

[3] Z Boson Analysis
http://cds.cern.ch/record/2203

[4] ROOT a Data Analysis Framework
https://root.cern.ch/

http://jupyter.org/

[6] ATLAS Open Data Website
http://opendata.atlas.cern