EVENT SELECTION CRITERIA OPTIMIZATION FOR THE DISCOVERY OF NEW PHYSICS WITH THE ATLAS EXPERIMENT

In large experiments, the researchers must analyze millions of events and choose the candidates for discovery. For this purpose, they do not inspect them one by one but use techniques which you will learn in this exercise.

Introduction

First let's talk about the detection of Z bosons through their decay to an opposite charged muon or electron pair. The Z boson has a small probability (branching ratio) to decay through these channels (~3% in each), but the signature it leaves in the detector is easily observable. Moreover, since its mass was measured with high precision by electron-positron accelerators, we can use the decay products to calibrate the detector. An example of Z production diagram via quarks is shown on Figure 1

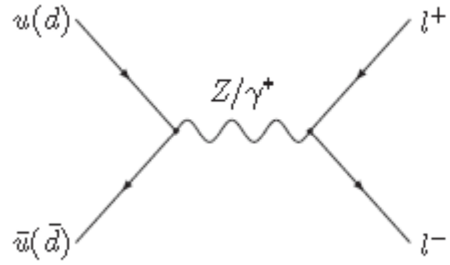


Fig. 1. Example of Z boson production by hadrons

The wiggly line of Figure 1 (Z/γ*) indicates that together with the Zs, virtual photons are produced in a wide range of masses (Drell-Yan mechanism). The lepton pairs from the decay of those virtual photons, with masses in the Z mass region, constitute the so-called irreducible background, because all the distributions of the kinematic variables are similar to the ones of the signal. In addition to this background, there exist other backgrounds which can be reduced (reducible) by appropriate selections (cuts). Such a background is the lepton pairs originating from W + jets, where the W decays in a lepton (plus a neutrino) and one of the heavy quarks or antiquarks (b or b̄ or c or c̄) decays to a lepton (by the so-called semileptonic decay), an anti-neutrino and other particles, so that the final state is two leptons of opposite charge. A similar background can result from the production of two jets or from a pair of $t\bar{t}$ quarks. In the following, we will examine how we can use the different kinematic distributions to reduce the background in the Z boson mass region and at the same time use it as the first step of the optimization of the criteria (cuts) for the "discovery" of the Higgs boson.

The Higgs boson was discovered by the ATLAS and CMS experiments in July 2012 with a mass ~125.5 GeV and published in Physics Letter B in September 2012. This concluded a nearly 50-year search, since Prof. P. Higgs and others introduced the Higgs mechanism to explain why the elementary particles Z and W have non-zero mass.

The Higgs bosons in the LHC are mainly produced by the “fusion of two gluons”, each one originating from each colliding proton, as in Figure 2.

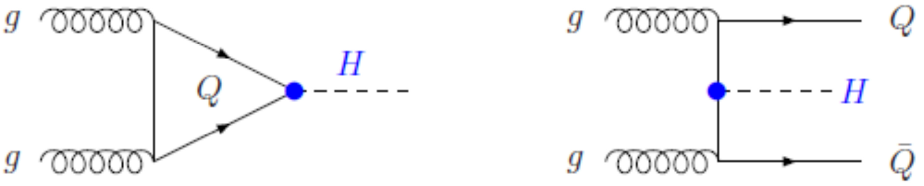


Fig.2 Higgs production in proton-proton accelerators by the fusion of two gluons

One of the easiest -to detect- decays of the Higgs boson is to a pair of Z bosons. Afterwards, each Z decays into two leptons of opposite charge but of the same flavor, so as observable final states we have the following:

- H → 2 e^+ + 2 e^-
- H → 2 μ^+ + 2 μ^-
- H → e^+ + e^- + μ^+ + μ^-

Figure 3 shows the invariant mass of the four leptons from data measurements, showing a peak at ~125 GeV (Higgs mass) along with the distribution of the background (calculated or estimated from data). The background is primarily from the direct production of two Z, i.e. a direct ZZ pair production, which does not originate from Higgs resonance decays (as shown in Figure 4).

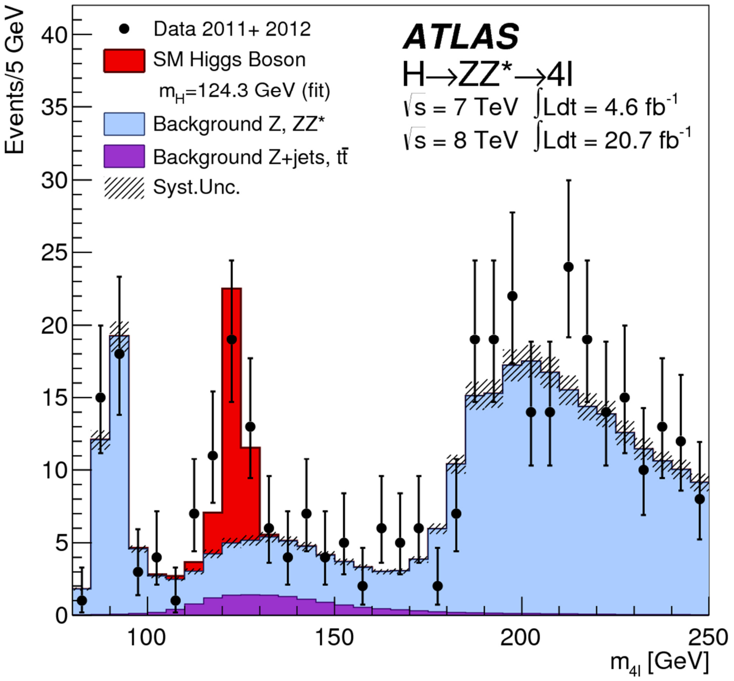
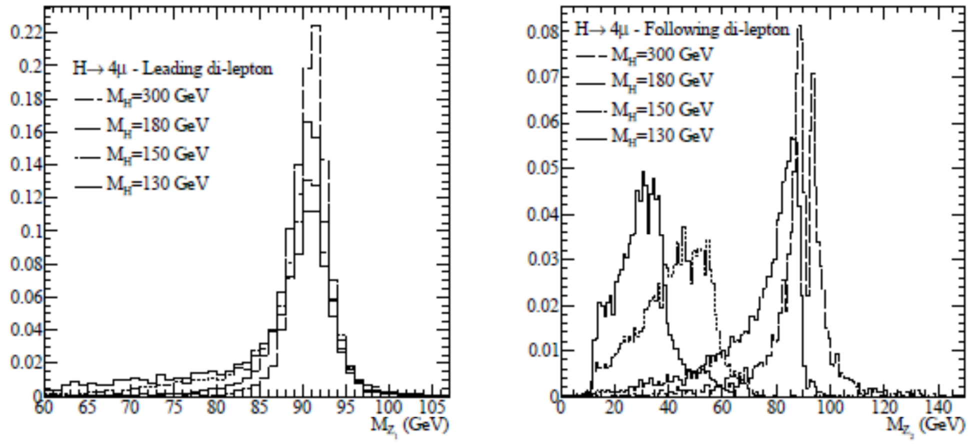


Fig.3 Four lepton invariant mass from data collected by ATLAS in 2011-2012

Because now we know that the Higgs boson has a mass less than the sum of the masses of the two Zs, we say that it decays into a real Z and a "virtual" Z (Z*). So if you plot the distribution of mass of the larger (in mass) opposite charge but of same flavor lepton pair (let's call it m_{12}), the distribution will have a peak close to the mass of the Z (Figure 5a). However, the distribution of masses of the other pair (let's call it m_{34}) has a broad peak at masses of about 30 GeV (see Figure 5b distribution for

m_H~130 GeV), while there will also be a concentration of events into larger masses, i.e. m~180 GeV where both Z's are real, As explained above, we can use those distributions to separate the signal from the background.



(a)

(b)

Fig.5 Mass distributions for two leptons of opposite charge and same flavor from simulated events for different values of the Higgs mass (a) for m_{12} and (b) for m_{34}

The ZZ* background, with the two Z decaying into two leptons each, is called **irreducible** background, because all the distributions of the kinematic variables are similar to the signal. Besides this background, we have other backgrounds which we can reduce (**reducible**) by suitable selection of restrictions (cuts). Such a background is the production of Z + jets, where in addition to the Z decaying into two leptons we also have the production of a heavy quark-antiquark pair ($b\bar{b}$ or $c\bar{c}$) with each of these quarks decaying through the so-called semi-leptonic decay into a lepton (antilepton) plus an anti-neutrino (neutrino) and other particles. The resulting final state is (as in the Higgs four lepton decay case) four leptons of which a pair gives a mass close to the mass of the Z.

However, in the case of the reducible background, the various distributions differ (slightly) from those of the Higgs and the purpose of this exercise will be **to select the optimal constraints (cuts) to reject/limit the background without losing too much signal**, i.e. -as said in the language of analysis- **to optimize the signal-to-background separation**.

Execution Instructions

1. Now that you have visually inspected a few dozen events you are ready to “run” many thousands of events using HYPATIA in “batch mode” i.e. “Batch Process Events” to analyze large samples of events, like the real researchers do. HYPATIA can be found at: **hypatia.iasa.gr/app-en**

2. From the dropdown menu (Figure 6), select “Batch Process Events”

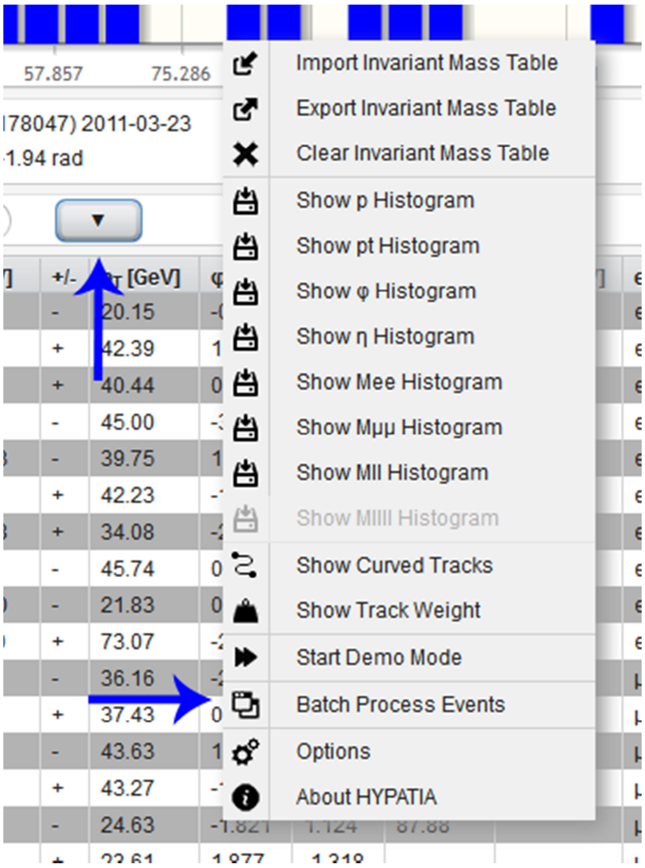
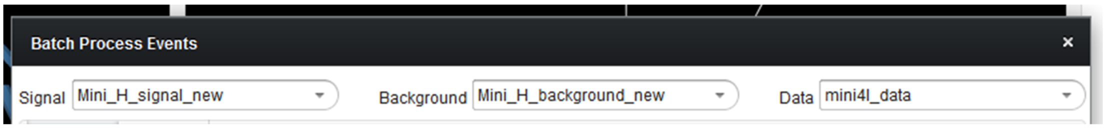


Fig.6 Dropdown options menu

The main window will appear where you will work to analyze your signal, i.e. to optimize your selection criteria (cuts) to get the best signal-background separation (as explained in the introduction).

3. At the top of the window select the samples to be analyzed (Figure 7)



Signal Mini_H_signal_new Background Mini_H_background_new Data mini4l_data

Fig.7 Window for event sample selection for analysis

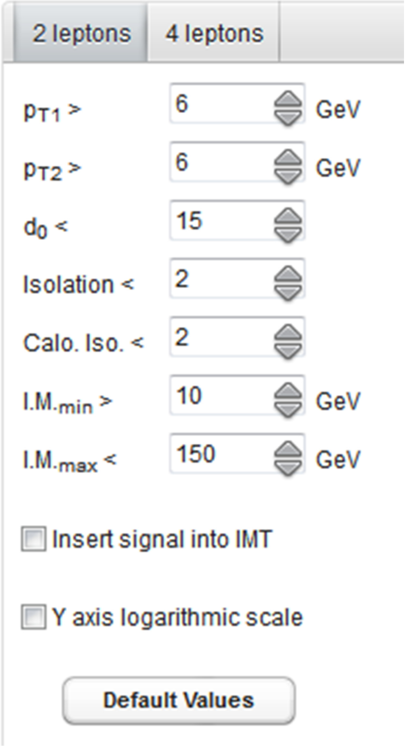
Choose:
A) For the study of the Z bosons decays to two leptons:
as **Signal** = **mini_ll_signal**
as **Background** = **mini_ll_bkg** (summed backgrounds from various sources)
and **Data** = **mini_ll_data** (actual data from the experiment corresponding to 1/10 fb^{-1})

B) For the study of the Higgs decays into four leptons:
as **Signal** = **Mini_H_signal_new**
as **Background = Mini_H_backgound_new**. The background consists of the sum of ZZ* irreducible background plus the reducible $Zb\bar{b}$ with $b\bar{b} \rightarrow e^+e^-$ or $\mu^+\mu^-$ with both backgrounds combined according to their respective production probabilities (production cross-sections).
And **Data** = **mini_4l_data** (actual data from the experiment which correspond to 1/25 of the data collected during Run I).

4. Choose what kind of analysis you want to perform from the two available tabs (2 leptons for the Z searches and 4 leptons for the Higgs searches) top left (Figure 8).

Fig.8 Selection of two or four lepton analysis

5. Now you're ready for the main analysis, namely to intelligently try to select the appropriate values of the kinematic variables that are given in the left pane of the window (Figure 9).



p_{T1} > 6 GeV

p_{T2} > 6 GeV

d_0 < 15

Isolation < 2

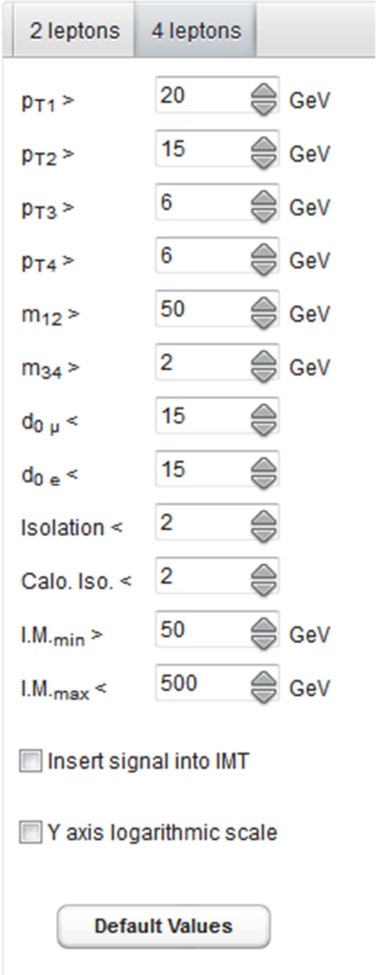
Calo. Iso. < 2

$I.M._{min}$ > 10 GeV

$I.M._{max}$ < 150 GeV

□ Insert signal into IMT

□ Y axis logarithmic scale



p_{T1} > 20 GeV

p_{T2} > 15 GeV

p_{T3} > 6 GeV

p_{T4} > 6 GeV

m_{12} > 50 GeV

m_{34} > 2 GeV

$d_{0\ \mu}$ < 15

$d_{0\ e}$ < 15

Isolation < 2

Calo. Iso. < 2

$I.M._{min}$ > 50 GeV

$I.M._{max}$ < 500 GeV

☐ Insert signal into IMT

☐ Y axis logarithmic scale

Fig.9 Selection of cuts (a) two leptons and (b) four leptons

The options are:

In the left pane of the "Batch Process Events" window we have selected for you the following kinematical distributions

A) For two leptons

- The momenta of the two preselected leptons that make up the invariant mass of the event under study (signal or background) $P_{T1,2}$
- The largest impact parameter ($d_{0,}$ see Appendix A) for one of leptons of the event
- The track isolation parameter (see Appendix B) of the less isolated lepton of the event
- The calorimeter isolation parameter (see Appendix B) of the lepton that has more energy around it.
- Selection of minimum (or maximum) invariant mass of the two leptons

B) Four leptons

- The momenta of the four preselected leptons that make up the invariant mass of the event under study (signal or background) $P_{T1,2,3,4}$
- m_{12} the mass of the lepton pair closest (in value) to the mass of the Z. The lepton pair consists of leptons of opposite charge and same flavor.
- m_{34} the mass of the lepton pair which is the second closest (in value) to the mass of the Z. The lepton pair consists of leptons of opposite charge and same flavor.
- The largest impact parameter (see Appendix A) for one of the muons ($d_{0\mu}$) or one of the electrons (d_{0e}) of the event
- The track isolation parameter (see Appendix B) of the less isolated lepton of the event
- The calorimeter isolation parameter (see Appendix B) of the lepton that has more energy around it.
- Selection of minimum (or maximum) invariant mass of the four leptons

To make the exercise simpler we have preselected some “default cuts”. These are equal to the values you see in Figure 9 when you first open the “Batch Process Events” window.

A) For two leptons

- The momenta of the two preselected leptons $P_{T1} > 6$ GeV and $P_{T2} > 6$ GeV
- d_0/measurement error < 15
- Track isolation parameter/$P_{Tlepton} < 2$
- Calorimeter isolation parameter/$P_{Tlepton} < 2$
- minimum invariant mass of the two leptons > 10 GeV
- maximum invariant mass of the two leptons < 150 GeV

B) For four leptons

- The momenta of the four leptons $P_{T1,2,3,4}$>20, 15, 6, 6 GeV respectively
- $m_{12} > 50$ GeV
- $m_{34} > 2$ GeV
- $d_{0\mu}$ / measurement error < 15
- d_{0e} / measurement error < 15
- Track isolation parameter/$P_{Tlepton} < 2$
- Calorimeter isolation parameter/$P_{Tlepton} < 2$
- Minimum invariant mass of the four leptons > 50 GeV
- Maximum invariant mass of the four leptons < 500 GeV

To optimize the cut values to be applied to these distributions, you will use a method which is similar to the one used by researchers (one of the methods) that led to the discovery of the Higgs boson. You will study the distribution of the respective variable by selecting it from the top row of the “Batch Process Events” window for either the two lepton analysis (figure 10a) or the four lepton one (Figure 10b).

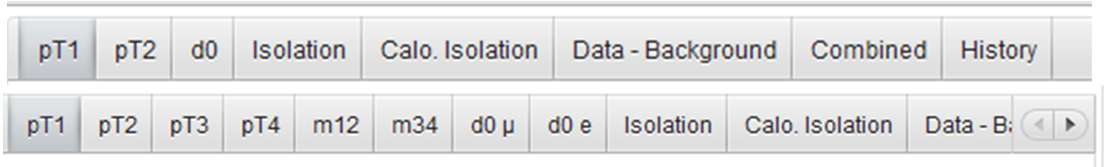
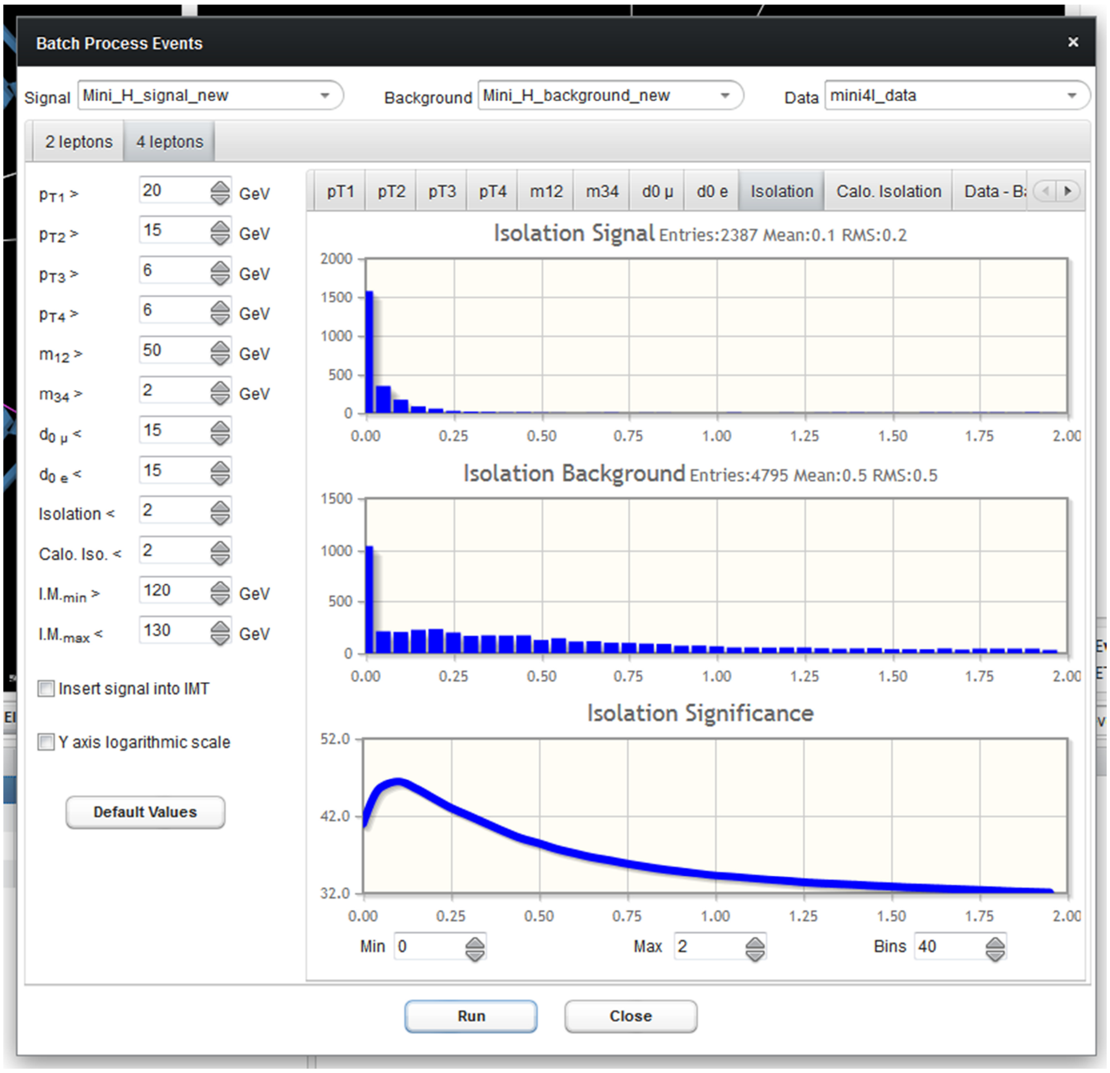


Fig.10 Histogram tabs; Top: two lepton and Bottom: four lepton analysis

When you click on any of these tabs and then press the "Run" button, three histograms will appear in the window (Figure 11). The top one t is the histogram of the sample selected in the left drop-down box of Figure 7 (which is normally the MC "signal"), the middle plot is the corresponding histogram for the sample selected in the middle drop-down box of Figure 7 (which is normally the MC "background") and the third plot is the curve showing the significance (see Appendix C) as a function of the cut value, i.e. how well we distinguish the signal from the background, which will help you to set the value of the cut.


Batch Process Events

Signal: Mini_H_signal_new
Background: Mini_H_background_new
Data: mini4l_data

2 leptons | 4 leptons

p_{T1} > 20 GeV
p_{T2} > 15 GeV
p_{T3} > 6 GeV
p_{T4} > 6 GeV
m_{12} > 50 GeV
m_{34} > 2 GeV
$d_{0\ \mu}$ < 15
$d_{0\ e}$ < 15
Isolation < 2
Calo. Iso. < 2
$I.M._{min}$ > 120 GeV
$I.M._{max}$ < 130 GeV

☐ Insert signal into IMT
☐ Y axis logarithmic scale

Min 0 | Max 2 | Bins 40

Fig.11 Histograms for one of the cuts

The boundaries of the histograms and the number of bins can be modified through the edit boxes at the bottom of the window (Figure 12). To redraw the histograms press the “Run” button again. When you change any of the cuts, you must press the “Run” button to draw the new histograms.

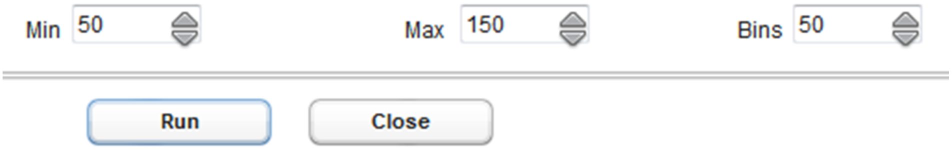
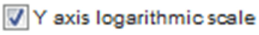

Min 50 Max 150 Bins 50

Fig.12 Selection of the boundaries of the histograms

You can also choose a logarithmic scale on the y-axis of the histogram by ticking the box that can be found under the 'cuts'


☑ Y axis logarithmic scale

Optimization of the selection criteria

Your analysis will be conducted as follows:

- Check that the values of the variables of the selection criteria (cuts) are set at the “default” values and choose one of these variables to optimize. Draw the corresponding histograms of Figure 11 for the variable under study.
- Note that the more different the shape of the distributions (histograms) of the signal and the background are, the more easily you can set the cut (either > or <) to “cut / reject” as much background and less signal as possible.
- Choose the cut value which maximizes the “significance” but without rejecting a large percentage of the signal (this selection is a matter of experience).
- Check how effective your selection was by plotting the “data-background” distribution. There you must have as many signal events and as few background events as possible (compared to what you had before applying the new cut).
- After selecting the value of that particular cut you enter it in the corresponding position of Figure 9 and proceed to examine the next variable. After you have optimized all possible variables (those in Figure 9) you plot the “data-background” distribution and compare it to the original (with the default values of cuts).
- Repeat the process (iteration) a few more times starting from the values of the cuts you have specified in the last step.
- Check the process by examining the “history” tab which displays the cuts and the significance of each step in a tabular way and highlights in green the best choice.

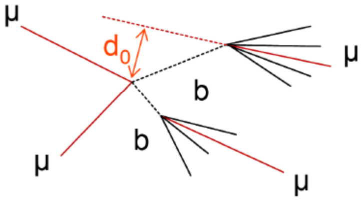
Note: If you want to examine the distributions of the kinematical variables of the **real data** you just choose as sample on the top left tab of Figure 7 the real data (either mini_ll_data or mini_4l_data) and use the tabs of Figure 10 to examine the different available distributions.

Questions

1. If you draw the mass distribution of the real two lepton data (ll_data) starting from very low masses m_{ll}> 2 GeV, what do you notice at these low masses (you cannot do the same with the simulated data (ll_signal) because they were simulated for m_{ll}>50 GeV) ?
2. Examine the mass distributions of two leptons after the best cuts using the «combined» tab. On top are the real data and below the histogram (signal plus background added according to the variable factors shown below the histogram). Compare the two distributions in a mass region close to the Z mass ($70<m_{ll}<110$ GeV) and write out your conclusions.
3. One of the reducible backgrounds for the Z search comes from two leptons produced by a pair of $t\bar{t}$ quarks. Search in the literature to suggest how these leptons can be produced. What properties will they have?
4. Plot the mass distribution of four leptons from the "data-background" histogram for the mass range $80<m_{4l}<170$ GeV and compare it with that of Figure 3. It is similar? Comment.
5. If the number of signal events in the mass range $120<m_{4l}<130$ GeV after the optimal cuts was 100 times smaller than what you have and the number of background events 10 times smaller, what would the significance of the discovery be?
6. The significance as defined corresponds numerically to the number of standard deviations from the normal distribution. What should be the required significance, to exclude at 95% certainty a fake signal, namely a signal which was not detected when the experiment was performed?

Annex

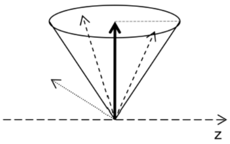
A: Impact parameter (d_0) is the vertical distance of the (secondary) vertex of the leptons from the collision point of the beams (primary vertex) as shown in the Figure below.



7.

If any of the leptons come from a b-quark decay (i.e. from jets or t/b quarks) then due to the finite (not infinitesimal) lifetime of the b-quark, the vertex can be at a (finite) distance away from the primary vertex, meaning it can be "displaced". In the case where all leptons come from Z decays their combined vertex will be very close to (virtually the same as) the primary vertex. This is a cut that we can use to reject the reducible background. Actually we use the d_0 values divided by the error of their measurement. For the Higgs searches, since the measurement of d_0 for electrons and muons is done in different ways, in Figure 9 (b) both values are provided separately.

B: Isolation criterion is a measure of whether each lepton has other particles close to it. The track isolation is defined as follows: use a cone of about 30° and sum the transverse momenta of the tracks (except for the lepton under study) that are within this cone -as in the following shape-



8.

and then divide by the transverse momentum of the lepton under study. The calorimeter isolation is computed in a similar way but summing the energy in the calorimeter close to the track. Although both isolation measurements should yield similar results, since they originate from different detectors, we choose to use both. The leptons from the jets and $Zb\bar{b}$ background are expected to be less isolated than the signal.

C: Significance is the value of

$$significance = \sqrt{2 * \left((S + B) * \ln\left(1 + \frac{S}{B}\right) - S\right)}$$

Where S is the number of signal events summed up to the value where the cut would be applied and B is the corresponding number of background events. The greater the value, the more significant the observation is.