CERN Summer Student Programme Final Report: Optimization of Signal Region for Dark Matter Search at the ATLAS Detector

Yip Long Sang Kenny
CERN Summer Student Programme
The Chinese University of Hong Kong
Supervised by Priscilla Pani
ATLAS, Stockholm University

August 14, 2015

Abstract

This report focused on the optimization of signal region for the search of dark matter produced in proton-proton collision with final states of a single electron or muon, a minimum of four jets, one or two b-jets, and missing transverse momentum at least 100 GeV. A brute-force approach was proposed to scan for the optimal signal region in rectangularly discretized parameter space. Analysis of the leniency of signal regions motivated event-shortlisting and loop-breaking features that allowed efficient optimization of the signal region. With the refined algorithm for the brute-force search, the computation time slimmed from an estimation of three months to one hour, in a test run of a million Monte-Carlo simulated events over densely discretized parameter space of four million signal regions. Further studies could focus on manipulating random numbers, and the interplay between the maximal figure of merit and the lower bound imposed on the background.

1 Introduction

Astronomical and cosmological observations urged the search for dark matter at the ATLAS detector. We would only consider events where dark matter is produced in associated with a pair of top quarks. We consider final states of a single electron or muon, a minimum of four jets, one or two b-jets, and missing transverse momentum at least 100 GeV. Owing to the enormous Standard Model background events producing similar final states as the dark matter signal, it was necessary to optimize the signal region for data analysis.

1.1 Dark matter candidates beyond the Standard Model

Astronomical and cosmological observations of disk galaxies provided compelling evidence for the existence of dark matter [1]. Standard model particles fails to explain the origin of dark matter. All the particles in the Standard model interact electromagnetically, except neutral Higgs, Z-bosons, and neutrinos. However, Higgs and Z-bosons decay readily into electromagnetically active products, and neutrinos failed to account for the relic density of the current universe [2]. Standard Model could not well-explain the origin of dark matter. Therefore, it is natural to seek for dark matter candidate beyond the Standard Model.

1.2 Search for dark matter at the LHC

In this report, we would focus on events with final state consisting of a single electron or muon, a minimum of four jets, one or two b-jets, and missing transverse momentum at least 100 GeV. The Feynman diagram of the signal processes is schematically shown in Figure 1.

At the LHC, various Standard Model backgrounds of similar final states would interfere with the signal. The dominant backgrounds are pair-production of top quarks (ttbar), production of W bosons in association with jets (Wjets), and top-vector boson associated production (ttV) [3]. To confirm or reject the existence of dark matter, it is necessary to perform event selection with appropriate cuts, in order to obtain a better statistical conclusion with the collision data. Throughout this report, the region in the parameter space specified by the cuts is also called the signal region.

1.3 Objective and flow of the report

The objective of this project is to propose and test an efficient and trustworthy algorithm for the brute force optimization of signal region over a discretized parameter space. In Section 2, the optimization problem would be further explained, with reference to the signal region in related literature and the interplaying variables of the parameter space. In Section 3, an algorithm for the optimization of event selection would be described, focusing on the features to speed up the computation. We would also provide a theoretical explanation of the speeding-up features in terms of leniency of selection cuts based on a natural partial ordering among the signal regions. In Section 4, we would present results of the test runs of the algorithm using ntuples from DC14 Monte-Carlo samples. In Section 5, the possibility of alternative algorithms of the optimization of event selection would be explored. We would also present a few problems with the DC14 Monte-Carlo samples, and the interplay between to key parameters in the study.

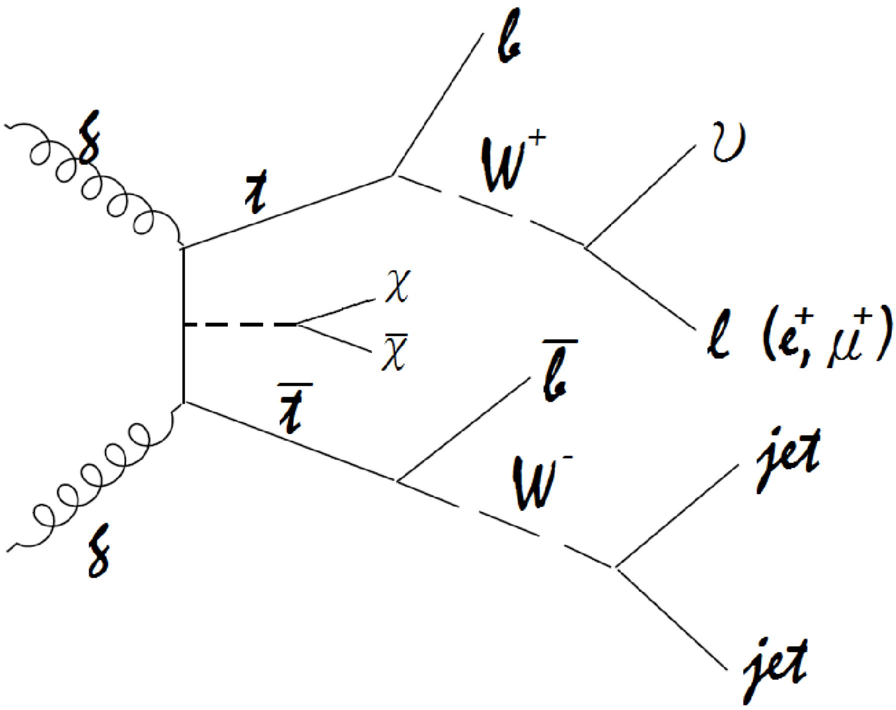


Figure 1: In the study, we would focus on events with dark matter produced in association with top-antitop pair. We would focus on final states consisting of a single electron or muon, a minimum of four jets, one or two b-jets, and missing transverse momentum at least 100 GeV. The large missing transverse momentum is carried away by the neutrinos and dark matter, undetected by the ATLAS detector.

2 Problem

The signal region, SR_mix, in Reference [3] embarked the whole story of signal region optimization. To quantify the goodness of a signal region, we introduced the figure of merit, which would be maximized in the optimization. To prevent overly tight selection cuts, we imposed a constraint on the minimum number of background events.

2.1 SR_mix

The SR_mix would serve as a blueprint of the signal region optimization [3]. Here, we would summarize the key features and variables of the SR_mix. For further details, you may refer to SR_mix Chapter 8 of Reference [3], and SR4 in the recent ATLAS publication for dark matter search [4].

As an example, we would illustrate the cut $E_T^{miss} > 270$ GeV. Neutrinos and dark matter, if present, passes through the ALTAS detector without any interactions, signified by a large missing transverse momentum. In the normalized plot of E_T^{miss} in Figure 2, the line showed the cut in SR_mix on E_T^{miss} at 270 GeV. The idea was to discard most of the background events while retaining the signals with an appropriate set of cuts.

Another example is the cut $m_T > 130$ GeV. Recall that the final state contains a single electron or muon, a neutrino, and possibly dark matter if it is a signal event. m_T is the transverse mass of the sum of the four-momentum of the lepton and the missing transverse momentum.

Table 1: List of cut conditions of SR_mix from Reference [3].

Variable	Cuts in SR_mix
Jet multiplicity	≥ 4
First leading jet, j_1 p_T	> 80 GeV
Second leading jet, j_2 p_T	> 70 GeV
Third leading jet, j_3 p_T	> 50 GeV
Fourth leading jet, j_4 p_T	> 25 GeV
Number of b-jets	≥ 1
Leading b-jet, bj_1 p_T	> 60 GeV
$E_T^{miss}/\sqrt{H_T}$	> 9 GeV$^{1/2}$
E_T^{miss}	> 270 GeV
m_T	> 130 GeV
am_{T2}	> 190 GeV
m_{jjj}	< 360 GeV
topness	> 2
$\Delta R(j_1, l)$	< 2.75
$\Delta R(bj_1, l)$	< 3
$\lvert\phi(j_1, p_T^{miss})\rvert$	> 0.6
$\lvert\phi(j_2, p_T^{miss})\rvert$	> 0.6
$\lvert\phi(l, p_T^{miss})\rvert$	> 0.6

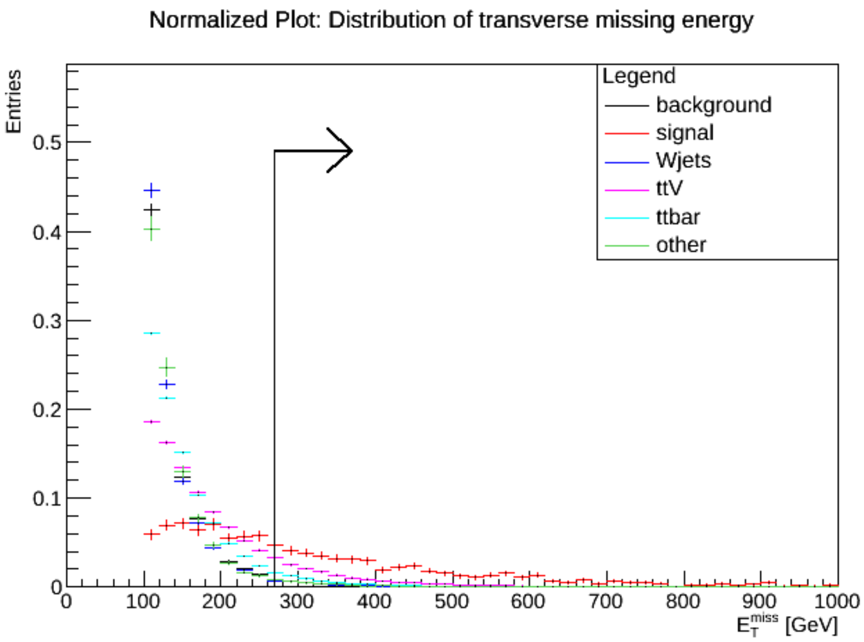


Figure 2: SR_mix cut on E_T^{miss} on high-E_T^{miss} regime to discard background events

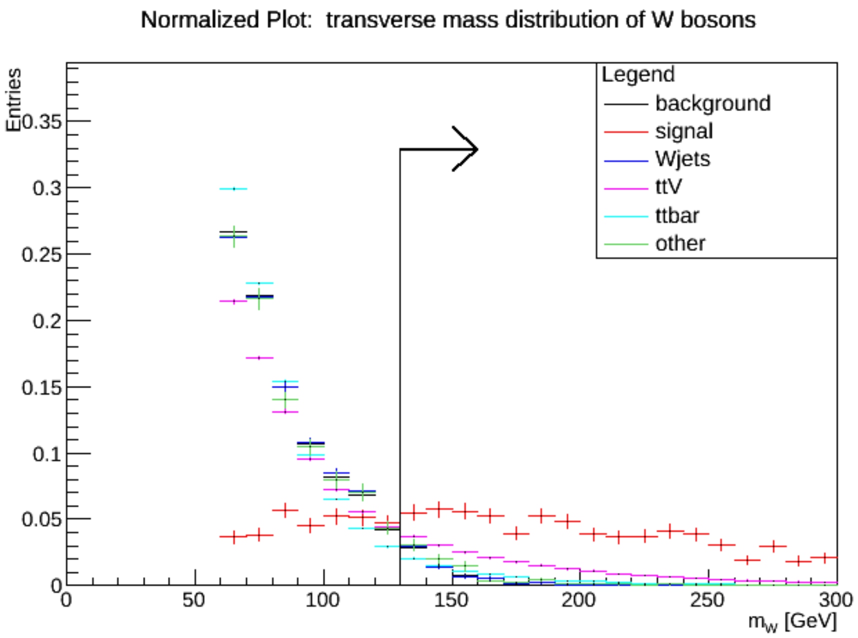


Figure 3: SR_mix cut on m_T on high-m_T regime to discard background events

In contrast to the background, dark matter contributes additionally to the missing transverse momentum, causing a long-tail in the m_T distribution. As could be seen in Figure 3, a great deal of background events falls out of the selection cut.

The third example demonstrates the selection cut $j_1 p_T > 80$ GeV. As could be seen in Figure 4, the signal carries a pronounced long tail in the transverse momentum distribution, compared to most background types. Top-vector boson associated production (ttV) is a reluctant class of background, having a similar tail as the dark matter signal. One could intuitively notice that selection on the jets' p_T would not be as effective as E_T^{miss} and m_T.

2.2 Figure of merits to be maximized

To claim a discovery, a null hypothesis (H_0) and a research hypothesis (H_1) are considered. The null hypothesis and the research hypothesis state that the observations are in perfect agreement with the Standard Model, and dark matter signal respectively.

H_0: The observations are Standard Model background only.
H_1: The observations are Standard Model background plus dark matter signal.

To evaluate a signal region, we could calculate the figure of merit S for each signal region of interest, defined by Equation (1).

$$S = \frac{N_{sig}}{\sqrt{N_{bkg} + (sys \times N_{bkg})^2}} \tag{1}$$

N_{sig} is the number of signal events in the signal region;
N_{bkg} is the number of background events in the signal region; and
sys is an estimation of the systematic uncertainty, taken to be $sys = 0.2$ throughout this project.

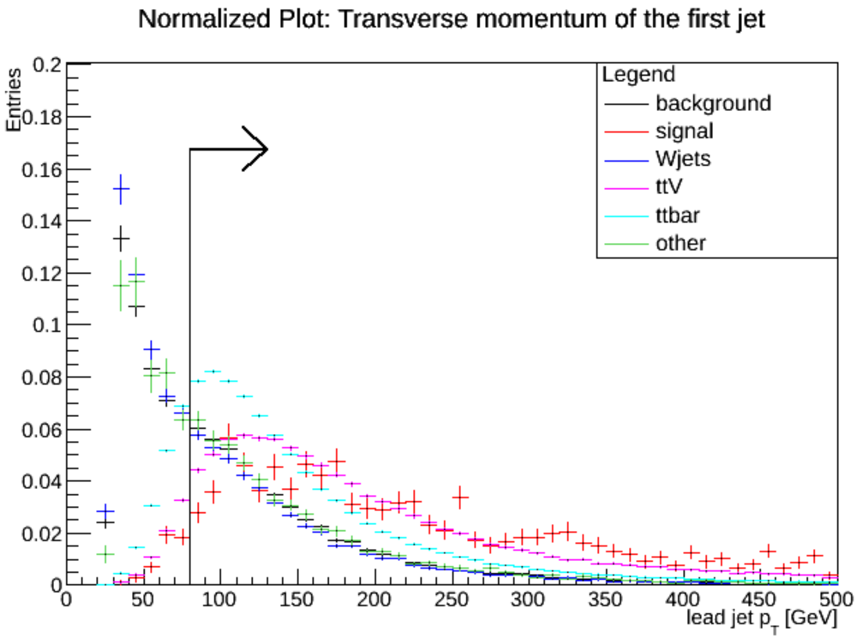


Figure 4: SR_mix cut on j_1 p_T on high-p_T regime to discard background events

The figure of merit defined above tells how unlikely the signal detected is caused by statistical fluctuations of the background, in other words, the degree of which the data rejects the null hypothesis. Poisson distribution of the number of background events gives an error of $\sqrt{N_{bkg}}$, and the systematic uncertainty is estimated by $sys \times N_{bkg}$. Combining the two types of errors yield the overall uncertainty of the number of background events in the signal region, as in the denominator of Equation (1). The deviation from the expected background divided by the overall uncertainty yields an estimate how unlikely the null hypothesis is true.

We set to maximize the figure of merit among the set of signal regions of interest, subject to a constraint of the minimum number of background events in the signal region. In actual analysis, it is necessary to further analyze the confidence level (CL) to claim a discovery, and the systematic uncertainty of the background. Readers could refer to other references for the discussion of the CL_s techniques and systematic uncertainties [3] [5].

2.3 Constraint on the background

We impose an additional constraint (2) to avoid selection cuts that are too tight. Excessively tight selection cuts are characterized by the small number of background events included. However, such a tight signal region would be heavily influenced by the systematic uncertainties, and the approximation in Equation 1 becomes invalid. They are uninteresting in the physical analysis.

$$N_{bkg} \geq N_{bkg}^{thres} \tag{2}$$

N_{bkg}^{thres} is a lower bound of the background we impose to eliminate overly tight selection cuts.

3 Method

In this Section, we would outline a brute-force optimization scheme to determine the rectangular signal regions with the largest figure of merit S, subject to the constraint (2). For the sake of computation, the parameter space would be discretized into vertices, with each vertex associated with a rectangular signal region. Directly looping over all the vertices would give rise to a computational complexity scaling up exponentially with the product of the number of points chosen on each parameter axis. Features to speed up the optimization, without harming the generality, would be discussed. It involves preselecting events passing through the most lenient signal region of interest, and arranging the signal regions in the order of leniency to break uninteresting loops.

3.1 Rectangular discretization of the parameter space

Each rectangular signal region is uniquely associated with a point in the parameter space. Recall the selection cuts of SR_mix outlined in Section 2. Each parameter axis is associated with a cut value. For example, the missing transverse momentum E_T^{miss} is associated with a cut value of 270 GeV. Hence, the SR_mix could be described by a point in the high-dimensional parameter space. This association assumes a direction of each inequality of the selection cuts, taken to be the same as SR_mix. For example, the selection cut $E_T^{miss} > 270$ GeV carries greater-than ($>$) inequality. It would be ambiguous to say the parameter E_T^{miss} is associated with the cut value of 270 GeV, without mentioning the direction and the strictness. We would assume the direction and the strictness of all the selection cuts in the SR_mix.

The high-dimensional parameter space would be discretized rectangularly into vertices. Each parameter axis would be partitioned by a few selection values, with the separation determined by the characteristic scale of the corresponding parameter axis. Consequently, the parameter space as a whole is discretized rectangularly. The number of rectangular signal regions to be considered is given by the product of the number of selection values on each parameter axis.

3.2 Partial ordering of signal regions in terms of leniency

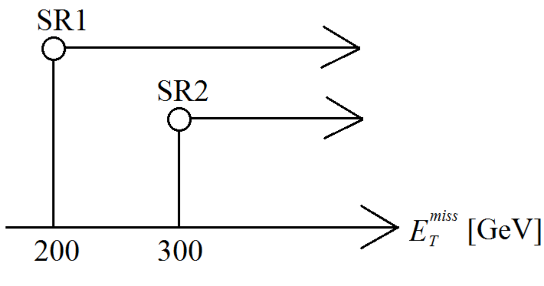


Figure 5: SR1 is more lenient than SR2.

The set of all interested signal regions inherits a partial ordering in terms of the leniency of the selection cuts. The understanding of the ordering structure could allow us to speed up the optimization through event preselection and loop breaking. The signal region (SR1) is said to be more lenient than the signal region (SR2), if for any event in SR2, it is also in SR1. In other words, SR2 is a subset of SR1. It is more casual to say that SR2 is tighter than SR1, or the selection cuts of SR2 are tighter than SR1. For example, suppose SR1 carries the cut $E_T^{miss} > 200$ GeV, SR2 carries the cut $E_T^{miss} > 300$ GeV, and assuming the same cut values on other parameter axes. Any event passing $E_T^{miss} > 300$ GeV must pass $E_T^{miss} > 200$ GeV. Hence, SR1 is more lenient than SR2.

The ordering is partial, but not total, as not all signal regions are comparable in terms of the leniency. For example, suppose SR1 carries the cuts $E_T^{miss} > 200$ GeV $m_T > 150$ GeV, and SR2 carries the cut $E_T^{miss} > 300$ GeV $m_T > 100$ GeV. An event E1 with $(E_T^{miss}, m_T) = (250, 200)$ GeV would lie in SR1, and not in SR2. Similarly, an event E2 with $(E_T^{miss}, m_T) = (350, 130)$ GeV would lie in SR2, and not in SR1. Hence, one could not decide whether SR1 or SR2 is more lenient. In such case, the leniency of the two signal regions are said to be incomparable.

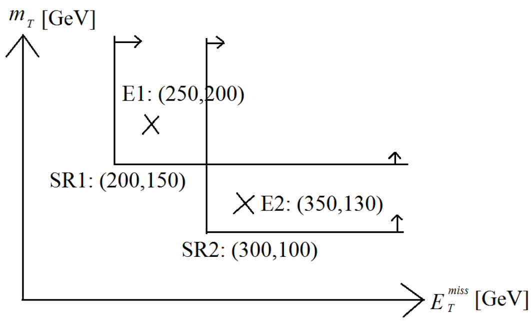


Figure 6: SR1 is incomparable with SR2.

3.3 Algorithm for the brute-force optimization

The idea behind the brute-force optimization scheme is straightforward. The figure of merit of each signal region is determined by counting the number of signal events and background events within. Then, the signal region with the maximum figure of merit, subject to the constraint of the number of background events, is selected. It would have taken days, if not months, to loop over all the signal regions and all the events. A few features to speed up the computation are introduced to shrink the computational time. The procedures of the optimization would be outlined, followed by an explanatory note.

1. The parameter space is discretized into rectangular vertices, with the cut spacing proportional to the typical scale of parameter.
2. The most lenient signal region is chosen from the set of all interested signal regions.
3. The signal and background events from the Monte-Carlo simulations were preselected by the most lenient signal region.
4. Among the list of preselected events, each event is looped over all the signal regions of interest. If the event fails a particular selection cut, with some skills of ordering the signal region by leniency, one could break the loop and skip other tighter cuts. The event weight, normalization factor, and the luminosity have to be taken into account when counting the number of signal and background events.
5. Among the signal regions satisfying the constraint on the background, the signal region of the maximum the figure of merit is selected.
6. Repeat with finer parameter space discretization, if a better signal region is desired.

There is no loss of generality to speed up the programme by shortlisting events by the most lenient signal region, and breaking the loop to skip tighter cuts. By discretizing the parameter space rectangularly, the most lenient signal region is guaranteed to exist uniquely. If an event does not lie in the most lenient signal region of interest, it would fall out of all other tighter signal regions, and therefore could be omitted with impunity. As discussed before, some signal regions could be ordered in terms of leniency. If an event fails a particular selection cut, it fails other tighter cuts. For this event, we are sure that it lies out of a big truck of tighter signal regions prior to checking tighter the selection cuts one by one.

To further speed up the optimization, one may also consider the following practices.

- The loop over events should enclose the loop over the set of all interested signal regions, as accessing TTree Object with ROOT takes a longer time.
- It is advisable to switch off branches irrelevant to the optimization, if there are any.
- The variables of the event should be stored in local variables when interest signal regions were looped over. Multiple access of variables from the TEvent object within the loop would slow down the computation.
- Fourth, one could divide the job into several components, and process them in parallel. It is possible because the analysis of an event is independent from the others.

4 Results

The brute-force algorithm described in Section 3 was tested with the DC14 ntuples generated by Monte-Carlo simulations. We would present the reduction of computation time as a result of the preselection and loop-breaking features. The optimization was performed in two steps - a coarse discretization, followed by a fine discretization. The best cuts obtained from the optimization would be compared to SR_mix. Readers should focus on the algorithm and the sense of the output, instead of the particular signal region generated, because there were a few problems with the DC14 ntuples. The issues with the DC14 ntuples would be briefly presented in Section 5.

The following parameters were set to be: $N_{bkg}^{thres} = 5$, sys $= 0.2$. The luminosity was set to be 5 fb^{-1}.

4.1 Coarse discretization and fine discretization

Table 2 shows the selection values on each parameter axis for the coarse discretization. The cut values were chosen manually by inspecting the normalized distributions of each parameter, so as to cut away the peaks of the background and retaining the long-tails of the signal distributions. There were a total of 4,644,864 signal regions considered. We neglected the variable m_{jjj} for some technical reasons. The variable $|\phi(l, p_T^{miss})|$ was also dropped, for its similar distributions in the signal and the background.

Based on the results of the coarse run, the optimization procedures were repeated with a finer parameter space discretization around the optimized signal region obtained in the coarse run. A total of 121,500,000 signal regions were considered in the second run. The cut values were tabulated in Table 3.

Table 2: Parameter space discretization for the coarse run

Variable	Cuts values in coarse run	
Jet multiplicity	≥ 4	
First leading jet, j_1 p_T	$> 80; 120$	GeV
Second leading jet, j_2 p_T	$> 60; 100; 140$	GeV
Third leading jet, j_3 p_T	$> 50; 90$	GeV
Fourth leading jet, j_4 p_T	$> 20; 60$	GeV
Number of b-jets	$\geq 1, 2$	
Leading b-jet, bj_1 p_T	$> 50; 90; 130$	GeV
$E_T^{miss}/\sqrt{H_T}$	$> 8; 12; 16; 20$	$\mathrm{GeV}^{1/2}$
E_T^{miss}	$> 200; 250; 300; 350; 400; 450; 500$	GeV
m_T	$> 120; 150; 180$	GeV
am_{T2}	$> 150; 180; 210; 240; 270; 300$	GeV
topness	$> -5; -2; 1; 4$	
$\Delta R(j_1, l)$	$< 4.0; 3.0; 2.0; 1.0$	
$\Delta R(bj_1, l)$	$< 4.0; 3.0; 2.0; 1.0$	
$\lvert\phi(j_1, p_T^{miss})\rvert$	> 0.6	
$\lvert\phi(j_2, p_T^{miss})\rvert$	> 0.6	

4.2 Shortlisting events by the most lenient signal region of interest

In the optimization algorithm, events passing though the most lenient selection cuts were shortlisted for further analysis. In this rectangular discretization, the most lenient signal region could be obtained by concatenating the leftmost cut values of each variable in Table 2 and Table 3. Table 4 showed the proportion of events passed through the most lenient signal region, 1.0 % in the coarse run, and 0.2% in the fine run. The preselection by the most lenient signal region resulted in a dramatic reduction of the computational time. Without the event-shortlisting and loop-breaking features, the coarse run was estimated to take three months of computation time to analyze 1,126,703 over 4,644,864 signal regions. With these improvements, the script could finish the task in an hour. Also, the preselection applied was selective. In the coarse run, only 1.0 % of the background events passed the preselection, while over 31.6 % signal events were retained. The situation is similar in the fine run. To truly reflect the reduction in computational complexity, the numbers in Table 4 were the number of entries in the ntuples, before applying the effects of the event weight and the luminosity.

4.3 Comparison of SR_mix with the optimization output

The optimized signal regions were compared to SR_mix. The number of signal and background events with the weights and the luminosity applied were shown in Table 5. The cuts were tabulated in Table 6. The signal region obtained in the coarse optimization and fine optimization were denoted by SR_c and SR_f respectively.

The optimized signal regions showed improvements in the figure of merits from 16 in SR_mix to 40 in SR_f. When SR_f was compared to SR_mix, it was noted that the number of signal events were halved, while the number of background was suppressed by a factor of one-ninth. In particular, the dominant ttV and Wjets backgrounds dropped by an order of magnitude.

Table 3: Parameter space discretization for the fine run

Variable	Cuts values in fine run	
Jet multiplicity	≥ 4	
First leading jet, j_1 p_T	$> 50; 60; 70; 80$	GeV
Second leading jet, j_2 p_T	$> 30; 40; 50; 60$	GeV
Third leading jet, j_3 p_T	$> 20; 30; 40; 50$	GeV
Fourth leading jet, j_4 p_T	$> 0; 10; 20; 30$	GeV
Number of b-jets	$\geq 1, 2$	
Leading b-jet, bj_1 p_T	$> 30; 40; 50; 60$	GeV
$E_T^{miss}/\sqrt{H_T}$	$> 10; 12; 14$	GeV$^{1/2}$
E_T^{miss}	$> 300; 310; 320; 330; 340; 350; 360; 370; 380; 390; 400$	GeV
m_T	$> 180; 210; 240$	GeV
am_{T2}	$> 150; 160; 170; 180; 190; 200$	GeV
topness	$> -1; 0; 1$	
$\Delta R(j_1, l)$	$< 3.0; 2.5; 2.0; 1.5; 1.0$	
$\Delta R(bj_1, l)$	$< 4.0; 3.5; 3.0; 2.5$	
$\lvert\phi(j_1, p_T^{miss})\rvert$	> 0.6	
$\lvert\phi(j_2, p_T^{miss})\rvert$	> 0.6	

Table 4: Event slimming by the most lenient signal region (Event weight and luminosity not applied)

Event type	Original	Coarse run	Fine run
Signal	2174	686 (68.4%)	384(82.3%)
ttV	93367	4140(95.6%)	1018(98.9%)
Wjets	474341	4001(99.2%)	1143(99.8%)
ttbar	370180	1955(99.5%)	244(99.9%)
Other background	188815	1499(98.9%)	386 (99.8%)
Overall	1126703	11595(99.0%)	2791(99.8%)

We could make sense of the relative importance of the variables by comparing the selection cuts in Table 6. It was preferable to loosen the cuts over the jets' transverse momentum, and tighten the cuts on E_T^{miss}, m_T, and $E_T^{miss}/\sqrt{H_T}$. There were several signal regions yielding the same maximum figure of merit in the coarse optimization and fine optimization. Only one was shown in Table 6 for each run.

5 Discussion

In Discussion, we would present the problems with the DC14 ntuples, and the possibilities of refining the algorithm of the brute-force optimization. It was also noted that the best signal region is dependent on the constraint parameter N_{bkg}^{thres}.

5.1 Problems of the DC14 ntuples

The given DC14 ntuples inherited a couple of problems, since they represented an easy simulation of the ATLAS detector. First, the variable m_{jjj} was missing in the CollectionTree. We did not reproduce the variable in situ, as its calculation would involve sophisticated statistical

Table 5: Figure of merits if SR_mix and optimized signal regions

Number of events	No cuts	SR_mix	SR_c	SR_f
Signal	1529.05	184.98	90.03	99.87
ttV	285.44	2.47	0.76	0.85
Wjets	600831.89	14.99	0.97	1.39
ttbar	82650.07	21.41	2.35	1.71
Other background	48085.34	8.01	1.37	1.13
Overall background	731852.74	46.88	5.45	5.08
Merit	0.01	15.93	34.93	40.41

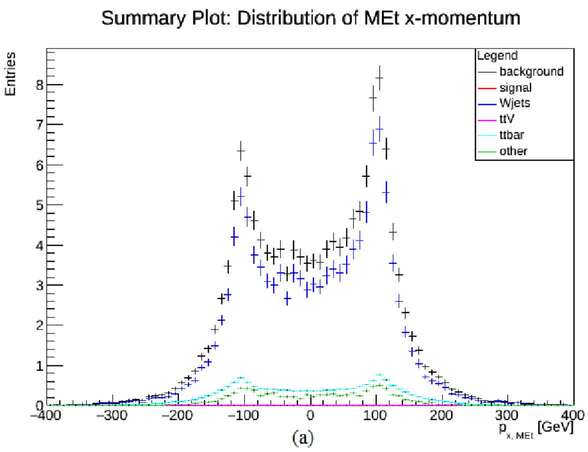
Table 6: Selection cuts of SR_mix and optimized signal regions

Variable	SR_mix	SR_c	SR_f	
Jet multiplicity $\geq$	4	4	4	
First leading jet, j_1 $p_T >$	80	80	80	GeV
Second leading jet, j_2 $p_T >$	70	60	40	GeV
Third leading jet, j_3 $p_T >$	50	50	40	GeV
Fourth leading jet, j_4 $p_T >$	25	20	10	GeV
Number of b-jets $\geq$	1	1	1	
Leading b-jet, bj_1 $p_T >$	60	50	40	GeV
$E_T^{miss}/\sqrt{H_T} >$	9	12	14	GeV$^{1/2}$
$E_T^{miss} >$	270	350	320	GeV
$m_T >$	130	180	210	GeV
$am_{T2} >$	190	180	190	GeV
$m_{jjj} <$	360	NA	NA	GeV
topness $>$	2	1	0	
$\Delta R(j_1, l) <$	2.75	2.0	2.0	
$\Delta R(bj_1, l) <$	3.0	3.0	3.5	
$\lvert\phi(j_1, p_T^{miss})\rvert >$	0.6	0.6	0.6	
$\lvert\phi(j_2, p_T^{miss})\rvert >$	0.6	0.6	0.6	
$\lvert\phi(l, p_T^{miss})\rvert >$	0.6	NA	NA	

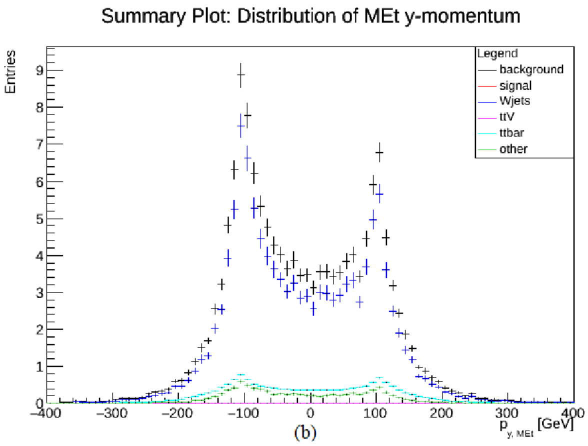
analysis methods. Second, E_T^{miss} showed peculiar distributions. It was expected that the missing transverse momentum would be rotationally symmetrical. However, as shown in Figure 7, the distributions shifted to the positive x-direction and negative y-direction. Also, the angular dependence was a shifted sine curve. These hinted the possibility of incorrect determination of the principle vertex. Third, the statistics of the signal and the background in the high E_T^{miss} and high p_T regime were insufficient, hinted by the degeneracy of the optimized signal regions. In principle, only one best signal region would be obtained, if the number of entries in the Monte-Carlo samples is sufficient. It was a sad news that the updated MC15 samples had not yet been ready by the end of my summer stay. It would be nice to re-run the optimization scripts when the MC15 ntuples are prepared.

5.2 Suggestions for further improvements

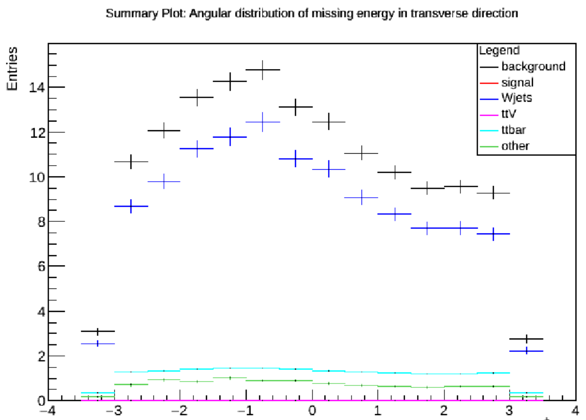
The algorithm could be modified to take the precision of the selection cuts as the input. In the present approach, the parameter space had to be discretized manually, and required a list of cut values of each parameter as the input. However, analysts might prefer specifying the precision



(a)



(b)



(c)

Figure 7: (a) Asymmetrical distribution of the missing transverse momentum in the x-direction; (b) Asymmetrical distribution of the missing transverse momentum in the y-direction; (c) Shifted sinusoidal angular dependence of the missing transverse momentum

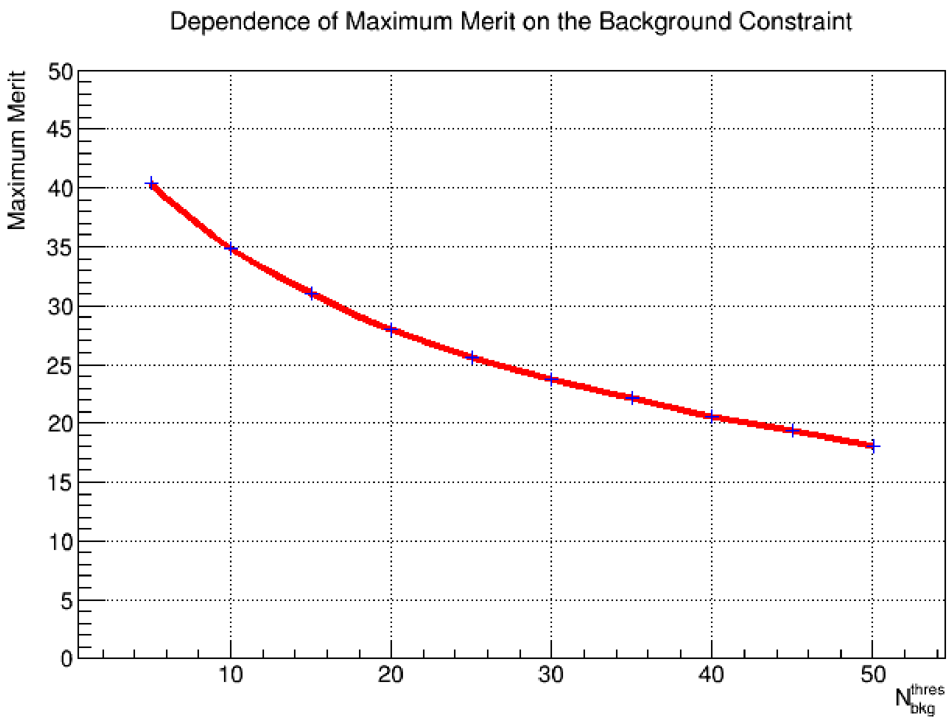


Figure 8: Tradeoff between the maximal figure of merit and N_{bkg}^{thres} using data from the fine discretization

of the cut values instead of the list. In the present algorithm, in order to obtain the best cut value of $E_T^{miss} = 320$ GeV correct to a precision of 10 GeV, we manually narrowed down for search in two steps. In the coarse run, we discretized the E_T^{miss} axis from 200 GeV to 500 GeV in 50-GeV steps. Then, the precision was enhanced to 10 GeV by discretizing the E_T^{miss} axis from 300 GeV to 400 GeV in steps of 10 GeV. It would be nice if we could improve the script to accept the desired precision as the input in a multiple run approach, without sacrificing the speed by overloading the programme by too many signal regions.

It would be interesting to investigate other approaches exploiting random numbers in the optimization. One idea is to start with a vertex close to the best signal in the parameter space. Then, a manageable number of the vertices are obtained by randomly displacing the original vertex in steps proportional to the desirable precision. The vertex yielding the maximum merit is chosen. The process is iterated until a stationary solution is reached. However, an intrinsic problem of this approach is that one might easily stop at the local merit maximum in the parameter space. In contrast, the brute-force approach would filter out the local maxima at the cost of computational complexity. It might be possible to develop advanced algorithm by combining the two methods - random number for speed, and brute force scan for accuracy.

5.3 Interplay of N_{bkg}^{thres} and the maximal figure of merit

There was a tradeoff between the maximal figure of merit and N_{bkg}^{thres} as depicted in Figure 8. On one hand, we wish to maximize the figure of merit in order to obtain a signal region that rejects the Standard Model null hypothesis (H_0). On the other hand, the number of background events shall not be too low, as it would be severely influenced by the systematic uncertainties. The rise in systematic uncertainties in the low-background region was not considered in the plot. The interplay between the desirable figure of merit and background signal should deserve more attention in further development of the optimization scheme.

6 Conclusion

Astronomical and cosmological observations stimulated the search for dark matter at the ATLAS detector. This report focused on the events with final states of a single electron or muon, a minimum of four jets, one or two b-jets, and missing transverse momentum at least 100 GeV. Owing to the enormous Standard Model background events producing similar final states as the dark matter signal, it was necessary to design signal regions of high signal-to-background ratio for data analysis.

A signal region could be evaluated by its figure of merit, which describes how unlikely the Standard Model null hypothesis is true. The signal region, SR_mix served as a starting point for further developments. It motivated the problem of signal region optimization based on the variables in SR_mix in a rectangularly discretized parameter space.

A brute-force approach was proposed to scan for the optimal signal region. It could guarantee a correct solution, at the cost of a high computational complexity. Analysis of the leniency of the signal region welcomed the ideas of preselection by the most lenient signal region, and loop-breaking when an event failed a lenient signal region.

The proposed algorithm was tested with ntuples from DC14 Monte-Carlo samples with coarse discretization, followed by fine discretization of the parameter space. The preselection feature by the most lenient signal region sped up the computation by discarding 99.0% uninteresting events in the coarse run, and 99.8% in the fine run. Without the event-shortlisting and loop-breaking features, the coarse run was estimated to take three months to analyze 1,126,703 over 4,644,864 signal regions. With these improvements, the script could finish the task in an hour. The figure of merit of the signal region obtained in the fine optimization(SR_f) was doubled that of SR_mix. One could also compare the relative importance of various parameters by inspecting the tightness of various selection cuts.

Even though the DC14 samples inherited defects of asymmetrical E_x^{miss}, E_y^{miss} distributions and missing variables, they served well as a testing ground of the proposed algorithm. It was suggested to re-run the script when the ntuples of the MC15 Monte-Carlo samples are ready. Further studies in the optimization could focus on incorporating random number methods to enhance the programming efficiency, while cross-checking with the trustworthy brute-force approach. In addition, when designing other optimization scheme, one should take note of the interplay between the maximum figure of merit and N_{bkg}^{thres}.

7 Acknowledgements

The CERN summer student programme was an once-in-a-lifetime experience for my personal and professional development. I have leant a great deal of high-energy physics and programming skills, from the combination of the summer project and the six-week summer lectures series. Moreover, the international working atmosphere at CERN has allowed me to interact with people from diversified cultures. I have learnt to be a team player and a culturally sensitive person. I would like to thank the CERN summer student team for organizing such an amazing summer school.

At the end of the summer programme, I wish to extend my deepest gratitude to my supervisor Priscilla Pani. She has opened the door for my exploration in experimental high-energy physics. Being a beginner in the field, I kept on asking silly questions about the technical aspects of high-energy physics and programming. Despite her tight schedule, Priscilla patiently guided me towards the answers step by step. Moreover, she brought me to the regular meetings of the Stop1L group and the Stockholm third generation group, where I had a glimpse of the daily research work of particle physicists.

The summer stay would not be possible without the support from my home institute and generous donations from various scholarship donors. I would like to thank Prof. M. C. Chu, and the Physics Department of the Chinese University of Hong Kong for offering me the chance of studying and researching aboard, under the Summer Undergraduate Research Exchange (SURE) programme. Also, I would like to mention Prof. Emily S. C. Ching for offering my first research experience and introducing to me scientific computing back in last summer. My summer stay was kindly supported by the donor of Professor Charles K. Kao Research Exchange Scholarship 2014/15; Mr. L.F. Chan, the donor of Professor and Mrs. Yau Wa Chan Scholarship 2014/15; as well as Dr. Alex K. Yasumoto and his wife Mrs. Sacha Yasumoto, the donors of Yasumoto International Exchange Scholarships.

Life is not easy when we live out of town. I would like to express my heartfelt thanks to Wei Wen Wong, Tom Ng and Ceiling Mak for keeping me away from the greasy food at the canteen and the processed food from the supermarket. It is always enjoyable to recall chilling out during weekends. Special thanks go to Anmol Arora, my fellow summer student working under the same supervisor for working together for two months.

I owe my last and biggest thanks to my parents and my girlfriend of five years for their support and understanding.

References

[1] G. Bertone, D. Hooper and J. Silk. *Particle Dark Matter: Evidence, Candidates and Constraints,* pp. 15-19, 2008.

[2] L. Covi. Lecture notes of the 2015 CERN summer student programme lecture series *Astroparticle Physics,* 2015.

[3] P. Pani. *To the bottom of the stop: calibration of bottom-quark jets identification algorithms and search for scalar top-quarks and dark matter with the Run I ATLAS data, 2011,* pp. 111-140, 2014.

[4] The ATLAS Collaboration. *Search for dark matter in events with heavy quarks and missing transverse momentum in pp collisions with the ATLAS detector,* 2015. arXiv:1410.4031v2.

[5] A. L. Read. *Presentation of search results: the CLs technique,* Journal of Physics G: Nuclear and Particle Physics 28 (2002) 26932704.