Strategy for Measurement of $t\bar{t}$ Production Cross-Section with the ATLAS Detector

Matt LeBlanc
Supervisor: Anadi Canepa

Acadia - TRIUMF - IPP - CERN

Summer 2010

1 Introduction

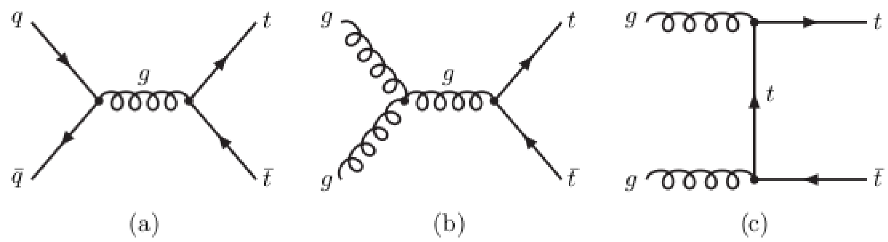
1.1 Top Physics with the ATLAS Detector

The discovery of the top quark by the CDF and $D\emptyset$ collaborations at Fermilab in 1995 confirmed its existence, predicted twenty years earlier by Kobayashi and Maskawa [1], consequently winning them the 2008 Nobel Prize. The confirmation of the Top's existence suffered such a long delay in part due to its unique properties amongst the particles of the Standard Model: its mass is by far the greatest of the Standard Model particles, at 172.0 ± 0.2 GeV/c^2 [2], placing the possibility of top production well outside of the reach of any accelerator in use before the Tevatron upgrade in 1994.

As a consequence of its large mass, the top quark is extremely short-lived: it exists for a mean lifetime of only 5×10^{-23}s. This time scale is $\mathcal{O}(10)$ smaller [3] than typical hadronization processes, giving physicists a unique opportunity to study the properties of a single quark decaying free from the effects of confinement.

Top quarks are produced almost entirely by either $q\bar{q}$ annihilation (fig. 1a) or gluon-fusion interactions (fig. 1b). At LHC energies, the gluon-fusion channel accounts for nearly 85% of all observed events, and $q\bar{q}$ annihilation the other 15%[1]. Once produced, a top will decay into a W-boson and b-quark 99.9% of the time. The W-boson will in turn decay into either a lepton-neutrino pair or a quark-antiquark pair, offering a set of three possible final decay states: dileptonic (a branching ratio of 10%), semi-leptonic (a branching ratio of 44%, pictured in fig. 2), and fully hadronic (a branching ratio of 46%).

[1] A third leading-order channel, due to double-gluon splitting (fig. 1c), offers a negligible contribution.



t

Figure 1: Leading-order Feynman diagrams illustrating the three Top-Antitop pair production via the strong interaction, specifically through (a) quark-antiquark annihilation, (b) gluon fusion, and (c) double-gluon splitting [3].

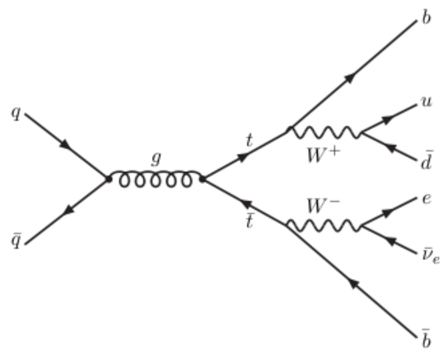


Figure 2: A $t\bar{t}$ decay in the semi-leptonic channel [3].

CERN's Large Hadron Collider (LHC) is expected to produce nearly 80 million top pair-production events annually at design luminosity and centre-of-mass energy[2], meaning that one of the most dominant backgrounds in both the hunt for the Higgs Boson and also when searching for physics beyond the Standard Model will be $t\bar{t}$ events. The large statistics will also help shed light on certain aspects of the Standard Model which are still not well-understood, such as the Electroweak Symmetry Breaking. Its large mass also provides a window through which the Standard Model may be extended, through new couplings directly to the top and other exotic interactions. It is therefore critical to these new theories that the top pair production events observed by the ATLAS detector are characterized precisely and accurately, in order to minimize the uncertainty in these new measurements.

One of the most crucial parameters is certainly the cross-section of the $t\bar{t}$ interaction: a measurement of the probability that an event will occur in

[2]Which has given the collider its nickname, the 'Top Factory.'

any given pp collision, calculated using equation 1:

$$\sigma_{t\bar{t}} = \frac{N_{Observed} - N_{background}}{\epsilon \cdot \mathcal{L}} \quad (1)$$

where σ is the cross-section; $N_{Observed}$ and $N_{Background}$ are the number of signal and background events detected; ϵ is the efficiency and $\mathcal{L}$ the integrated luminosity of the data sample being examined. A rough idea of the associated uncertainty in this measurement is described by error propagation in equation 2, and offers some guidance to consider when attempting to make a more precise cross-section measurement:

$$\frac{\Delta\sigma}{\sigma} = \sqrt{\frac{\Delta N_{obs}^2 + \Delta N_{bg}^2}{(N_{obs} - N_{bg})^2} + \left(\frac{\Delta\mathcal{L}}{\mathcal{L}}\right)^2 + \left(\frac{\Delta\epsilon}{\epsilon}\right)^2} \quad (2)$$

In standard 'cut-and-count' analysis procedures carried out by the ATLAS Top Working Group, a minimum threshold cut will be defined on certain kinematic variables and events which do not meet this threshold will be left out of the analysis. This tends to result in large acceptance losses, which is not in itself an issue when performing an analysis with normal statistics. However; when attempting an analysis with early data (≈1pb), the need for maximizing the statistics given by the accessible data becomes much greater. In the remainder of this report, an alternative method of measuring the $t\bar{t}$ interaction cross-section is examined, which proposes to maximize the number of good signal events selected from early LHC data by using looser event selection criteria than those allowed by standard 'cut-and-count' methods, in combination with a statistical binned-likelihood curve fitting technique to use the shapes of the spectra of kinematic signal- and background-variables themselves to predict the true number of signal and background events observed in a set of data. Although it accounts for only 10% of total $t\bar{t}$ events, preliminary studies have focussed on the di-lepton channel due to the relatively low QCD background when compared to the other decay modes, allowing a more precise initial measurement to be performed. The signature of this channel will therefore be: a pair of high-p_t leptons, jets and large missing transverse energy. For the sake of brevity, the main focus of results presented will be those from the $\mu\mu$ decay mode.

1.2 Dominant Backgrounds

In the di-lepton channel, several background processes mimic the properties of top events in any given sample of data. These interactions and their relative contributions in a data sample of 200pb^{-1} are summarized in table 2.

Table 1: Prediction of the number of events passing standard cleaning cuts in a data sample of 200pb^{-1} [4].

Event Type	ee Channel	$\mu\mu$ Channel
$t\bar{t}$ Di-Lepton	209^{+6}_{-6}	327^{+7}_{-7}
Other $t\bar{t}$	$0^{+0.1}_{-0}$	$0^{+0.1}_{-0}$
Single Top	7^{+1}_{-1}	11^{+2}_{-2}
Drell-Yan	17^{+4}_{-2}	54^{+5}_{-3}
diboson	2^{+1}_{-1}	5^{+1}_{-1}

The dominant background in this case, Drell-Yan production (illustrated in fig. 3), accounts for over half of the background events. In this type of interaction, an incoming quark-antiquark pair annihilates to form a virtual Z and photon, which then decays into a pair of leptons. Either incoming quark may emit gluon radiation before annihilating, forming jets in the calorimetry systems. The major difference between Drell-Yan and $t\bar{t}$ signals is the nature of the missing transverse energy[3] ($\not{E}_t$), which is not physically generated in Drell-Yan interactions. Despite this, it is still possible for the event reconstruction software to associate some 'fake' $\not{E}_t$ with the events due to mis-measurement in the calorimetry systems.

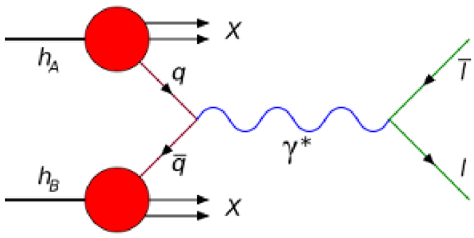
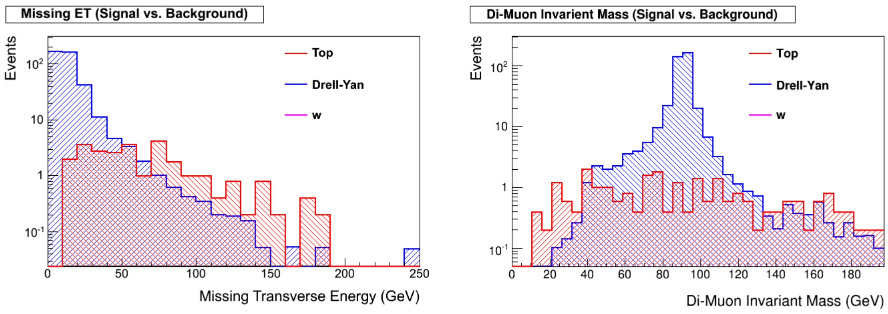


Figure 3: Leading-order Drell-Yan production Feynman diagram [5].

It is instructive to consider the kinematics of this type of event, in order to develop some concept of how its signature may vary from that of a $t\bar{t}$ decay. Two of the best discriminating variables which have been examined are the missing transverse energy ($\not{E}_t$) and the invariant mass of the leading two leptons. For the $\not{E}_t$ (fig. 5a), the background curve peaks near zero and drops sharply as expected (due to the high probability of there being

[3]Because LHC protons collide head-on, the total energy in the plane transverse to the beam direction before and after collisions is zero. Because of this, the missing transverse energy is defined as the negative of the vector sum of all energy in this plane: the vector quantity which, when added to the event's sum-transverse-energy, will yield zero and maintain energy conservation.

a small amount of $\not{E}_t$ being associated with an event during reconstruction, versus the relatively low probability of there being a high amount). For the signal, the curve peaks far from zero due to the neutrinos produced by the W decay. For the invariant mass (fig. 5b), there is a clear peak in the background curve due to the on-shell Z decay which is not present in the uniform signal curve. Both of these variables offer promising grounds with which to distinguish signal from background using statistical means.



(a) Missing Transverse Energy

(b) Leading Lepton Invariant Mass

Figure 4: A signal-vs.-background comparison of the most promising discriminating variables found thus-far. Note the clear peak around 91.2 GeV in fig.5b, the mass of the produced Z.

2 Methodology

2.1 Event Selection

To identify $t\bar{t}$ events which occur in the ATLAS detector, the top analysis group has outlined a standard set of event selection cuts based on the kinematics of $t\bar{t}$ interactions in the dilepton channels:

- To select the leptonic decay of the W's; require 2 isolated leptons (ee, $\mu\mu$ or ee) with $P_t > 20$ GeV.
- Veto events when the combined invariant mass of leading leptons falls near the Z peak, in order to suppress dominant Drell-Yan backgrounds ($|Mass(ee) - 91| < 5$ or $|Mass(\mu\mu) - 91| < 10$).
- To select the b-jets produced by the t decay; require at least 2 jets with $P_t > 20$ GeV.

- Require large missing transverse energy ($\not{E}_t > 40$ GeV ee or $\not{E}_t > 30$ GeV $\mu\mu$) or large H_t[4] ($H_t > 150$ GeV $e\mu$).

In order to validate the performance of the analysis code, results using the above selection criteria were compared to those results of other institutions running over a common Monte Carlo $t\bar{t}$ sample, and yielded results falling within the range of those accepted by other validated analysis groups.

Table 2: Summary of $t\bar{t}$ analysis code validation, comparing results to the validation results from TRIUMF and Simon Fraser University.

	Validation Results	TRIUMF	SFU
Before Cuts	**9990**	9990	9990
2 Good Leptons	**105** ee**, 152** $\mu\mu$	106 ee, 157 $\mu\mu$	105 ee, 156 $\mu\mu$
Z-Mass Veto	**100** ee**, 134** $\mu\mu$	100 ee, 141 $\mu\mu$	99 ee, 140 $\mu\mu$
$\not{E}_t$ Cut	**70** ee**, 110** $\mu\mu$	70 ee, 116 $\mu\mu$	69 ee, 115 $\mu\mu$
≥ 2 Good Jets	**56** ee**, 95** $\mu\mu$	56 ee, 99 $\mu\mu$	55 ee, 98 $\mu\mu$

2.2 Jet Cleaning Cuts

After preliminary event-selection has occurred, it is crucial to ensure that all remaining events have jets[5] which are well-behaved. Due to the complex nature of jet objects and the nature of their particular reconstruction process, there are several considerations which must be made [6] [7]:

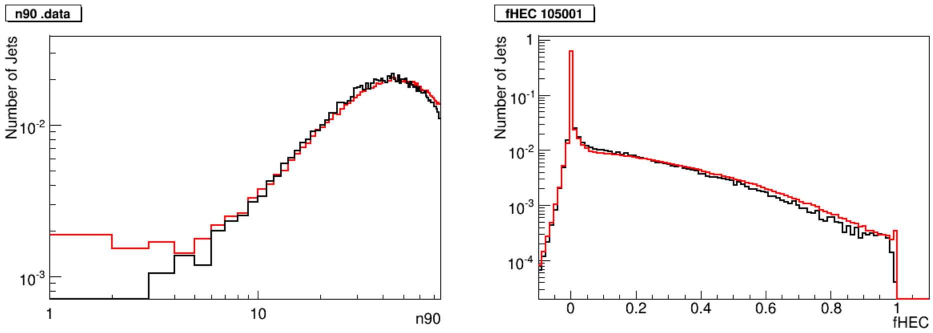
- Hardware failures in certain HEC cells occasionally cause noise bursts, which manifest themselves as large-energy events in extremely localized areas (on the order of 10-15 GeV in one or two cells), regardless of whether or not collisions are occurring in ATLAS. To remove these events, it is required that no jet may deposit over 90% of its energy in fewer than 5 unique calorimeter cells ($n_{90} < 5$), while at the same time the fraction of energy deposited in the calorimeter by the jet (f_{HEC}) is greater than 80% of the total event energy.

- In order to eliminate bad-quality jets in the electromagnetic calorimeter, it is required that no jet with greater than 95% of its energy in the EM calorimeter ($f_{EM} > 95$) possess a fitting quality factor[6] of more than 0.8 ($f_Q > 0.8$).

[4] H_t is defined as the sum energy of all selected objects (jets, leptons) in the transverse plane.

[5] All jets considered are reconstructed using the K_T jet reconstruction algorithm, using a distance parameter of R=0.4.

[6] The fitting quality factor is a measure of the difference between the readout pulse from the electromagnetic calorimeter, and the reference curve used in reconstruction software.



(a) n_{90}

(b) f_{HEC}

Figure 5: Histograms of n_{90} and f_{HEC} before jet cleaning cuts are applied, with overlaid Monte Carlo prediction **[BLACK]** and early run data **[RED]** (area-normalized in order to allow comparison).

- In order to remove events with jets resulting from collisions of beam particles which arrive either before or after the bunch ("beam gas" events), as well as events with jets from photons due to cosmic ray muon decay, an event timing cut requires that the time-window in which a jet occurs fall within 50ns of the bunch crossing associated with the event.

2.3 Likelihood Fitting

When attempting to perform an analysis with initial LHC data, it becomes clear that another approach may be needed in order to perform precision measurements of parameters such as $\sigma_{t\bar{t}}$. As shown in eqn. 2, the cross-section uncertainty is inversely proportional to the number of signal events accepted, revealing one possible area in which the uncertainty may be reduced by alternative analysis methods. One method to increase the acceptance is to loosen the event selection criteria, preserving our signal events but also allowing more background events to pass selection. In order to distinguish between the signal and background events after preliminary event selection has taken place, a binned likelihood fit was constructed to complete the analysis.

Binned likelihood fitting [8] operates by comparing a theoretical model to a particular set of data $\{x_1, x_2, ..., x_N\}$, and attempts to maximize the Poisson probability that the data in each bin corresponds to the model by varying a free parameter - in this case, this number of events (or, the integral of the histogram). The likelihood is defined by the product of each bin's individual probabilities:

$$L(x_1, x_2, ..., x_N; a) = \prod_i P(x_i; a) \tag{3}$$

Signal and background templates for each background channel will be constructed from Monte Carlo models, in order to extract the number of signal and background events in the data being analyzed. By adding together signal and background curves of each kinematic variable, it is possible to create probability distributions which possess the theoretical shape of real data distributions. From these models, as well as knowledge of how many events are theoretically expected in a data sample of a certain size, it is possible to generate toy data on which to attempt a likelihood fit [7]. By performing many pseudoexperiments where this toy data is fit with the statistical model, it is possible to create a likelihood distribution which may be maximized.

As the likelihood fitting algorithm is developed, it will be simultaneously validated: since the number of true (expected) signal and background events for a given data sample is input, it will be easy to determine whether or not the fit is behaving as it is expected - crucial information to understand during these Monte Carlo trials, since this knowledge of the 'truth' will be hidden when running over LHC data.

The focus of the analysis now shifts to an effort to determine the most discriminating kinematic variables: those variables whose signal and background spectra are distinct enough for the fit to be able to distinguish the shape of each curve from their sum. Each variable's relative effectiveness will be evaluated in terms of three 'figures of merit':

- **Mean number of events predicted**, should be equal to true number of events, N_{true}.
- **Statistical Uncertainty**, should ideally be close to $\sqrt{N_{true}}$.
- **Pull**, the difference between the number of events predicted by the fit and the true number. Should be a gaussian distribution centred at 0 and of width 1.

Via the implementation of Likelihood fitting, it is possible to loosen the event selection cuts in order to allow more events to pass the selection criteria. **By reducing the $\not{E}_t$ cut to a minimum of 10 GeV and removing the Z-mass veto outright**, our un-normalized acceptance in the signal is increased by 19% in the $\mu\mu$ channel (from 95 to 113 events). The background acceptance also benefits from the loosened event selection, and is increased by nearly a factor of 14.

[7]Much of the statistical analysis discussed here was performed using RooFit, a ROOT toolkit designed to aid in the modelling of physical distributions.

3 Preliminary Results

Preliminary examinations of kinematic variables in the di-electron and di-muon channels have yielded some promising early results. For both channels, kinematic variables have been found which allow the fit to converge and extract the true number of signal and background events from the data model. For the results presented below, the fitting procedure was run over a monte carlo sample normalized to 10pb^{-1}, containing 9 signal and 75 background events. Using the $\mu\mu$ invariant mass (fig. 6) and the $\mu\mu$ missing transverse energy (fig. 7), it is clear that the fitting procedure is capable of matching the 'most likely' value of signal and background events to the truth, with statistical errors peaking near $\sqrt{N_{signal}}$ and $\sqrt{N_{background}}$ for each case and pull widths generally no higher than 10-15% more than ideal.

It is important to note that not all kinematic variables which have been tested resulted in such success. Most variables either resulted in a divergent fit, or had insufficiently distinguishable signal and background spectra to allow the fitting procedure to accurately function (as shown in fig. 8, when the fitting procedure was applied to the $\Delta\phi$ between leading leptons). Empty bins in the signal and background histograms also pose a problem for the fit, causing it to diverge when applied to variables such as the invariant lepton mass when the Z-veto is being applied.

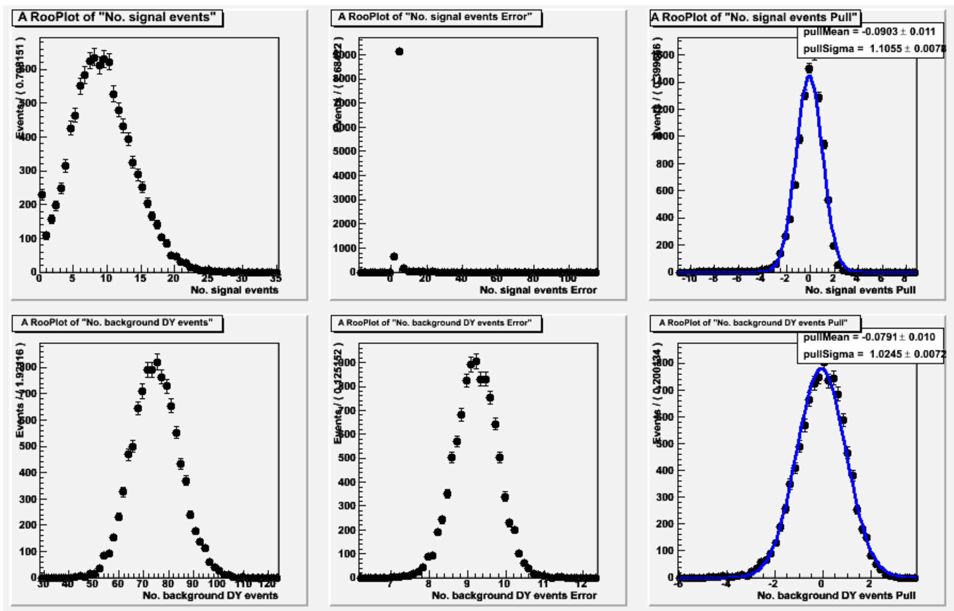


Figure 6: Preliminary results attempting to fit using **invariant muon mass in the di-muon channel** as discriminating variable.

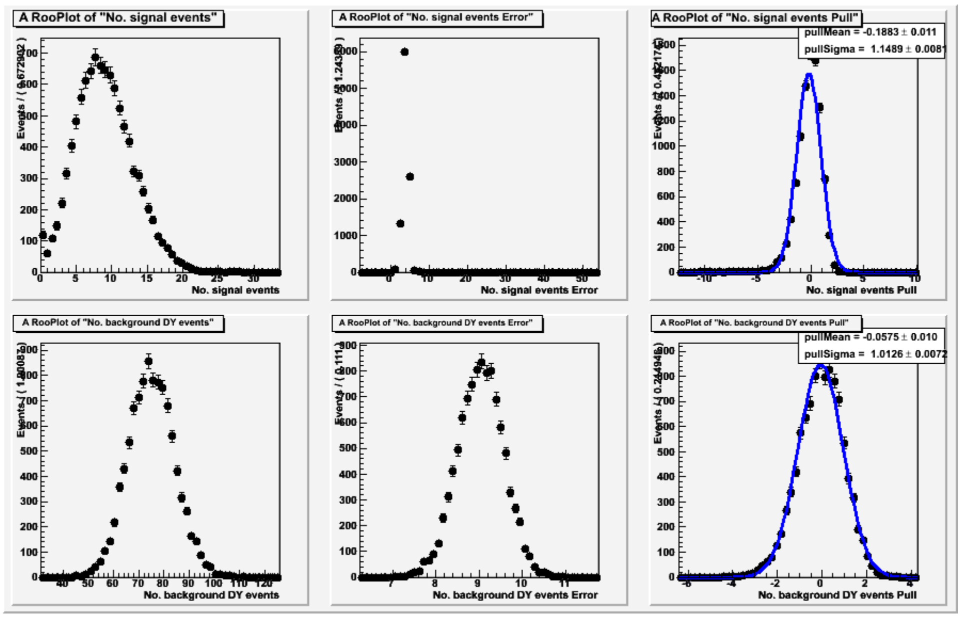


Figure 7: Preliminary results attempting to fit using **missing transverse energy in the di-muon channel** as discriminating variable.

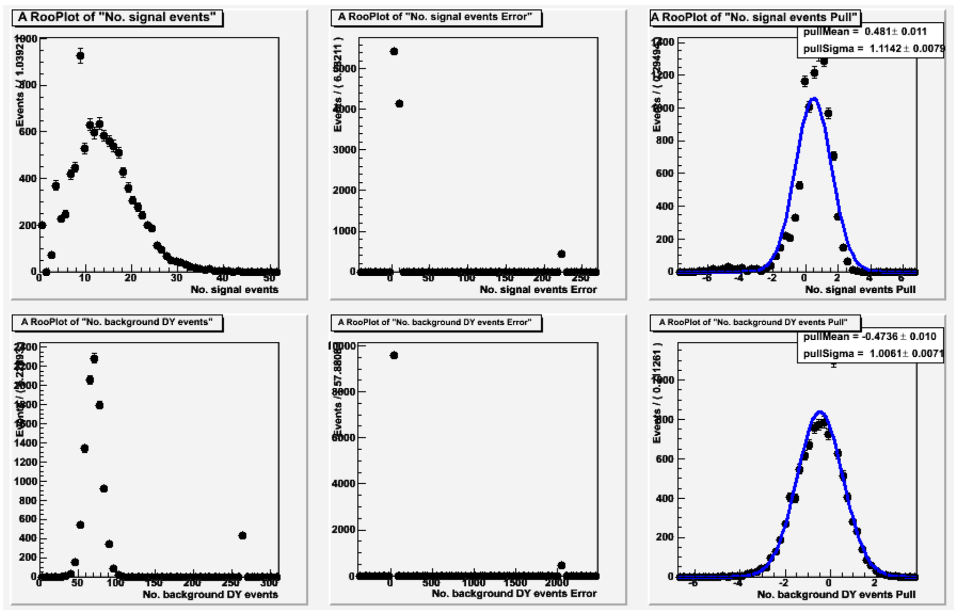


Figure 8: Preliminary results attempting to fit using **the $\Delta\phi$ between leading leptons in the di-muon channel** as discriminating variable. This is an example of a variable for which the fitting algorithm performs sub-optimally.

4 Conclusion

The goal of this project was to obtain an accurate, precise measurement of the $t\bar{t}$ cross-section in the ATLAS detector, despite the low amounts of data expected during the early stages of the LHC's lifespan. The method chosen to cope with this problem and allow event selection cuts to be lowered, based on binned likelihood fitting, seems to be capable of accurately determining the number of signal and background events in generated Monte-Carlo simulations, when applied to certain kinematic variables of the $t\bar{t}$ interactions. While more work remains to be completed in optimizing the fitting procedure and it has yet to be tested on real LHC data samples, the Likelihood fitting procedure remains nonetheless an attractive alternative to standard 'cut and count' analysis methods.

References

[1] Makoto Kobayashi and Toshihide Maskawa. CP-violation in the renormalizable theory of weak interaction. *Progress of Theoretical Physics*, 49(2):652–657, 1973.

[2] C. Amsler *et al.* Particle Data Group. PDGLive particle summary. *Review of Particle Physics*, B667(1), 2008.

[3] Dhiman Chakraborty, Jacobo Konigsberg, and David Rainwater. Review of Top Physics. Mar 2003.

[4] Prospects for measuring top pair production in the dilepton channel with early atlas data at $\sqrt{s}$=10 tev. Technical Report ATL-PHYS-PUB-2009-086, CERN, Geneva, Aug 2009.

[5] Wikipedia. Drell-Yan Process — Wikipedia, the free encyclopedia, 2010. [Online; accessed 17-August-2010].

[6] ATLAS Collaboration. Data-quality requirements and event cleaning for jets and missing transverse energy reconstruction with the ATLAS detector in proton-proton collisions at a center-of-mass energy of $\sqrt{s}$=7 TeV. Jun 2010.

[7] A. Canepa, J.R. Lessard, I. Nugent, H. Okawa, A. Taffard, and B. Toggerson. Missing transverse energy for top physics analyses with early ATLAS data at $\sqrt{s}$=7 TeV. ATLAS collaboration, May 2010.

[8] R. J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series).* WileyBlackwell, reprint edition, 1989.