Data driven background determination for SUSY searches with ATLAS

Aleksej Koutsman on behalf of the ATLAS collaboration
Nikhef National Institute for Subatomic Physics, Science Park 105, P.O. Box 41882, NL - 1009 DB Amsterdam, Netherlands

Abstract. Supersymmetry (SUSY) is an attractive extension of the Standard Model possibly solving many standing issues in particle physics and cosmology. The general purpose ATLAS detector at the Large Hadron Collider (LHC) is an experiment capable of discovering or excluding TeV SUSY. However discovery can only be claimed when the Standard Model backgrounds are understood and are under control. The expectations at the LHC are that Monte Carlo simulation predictions may not be sufficient to achieve this and the backgrounds will have to be determined from data itself. In this note we will highlight some data driven methods developed to estimate backgrounds and detect a possible SUSY excess.

1. Introduction
In order to prevent rapid proton decay a new quantum number called R-parity is conserved in SUSY models, with two clear consequences for SUSY searches with ATLAS. The first one is the lightest supersymmetric particle is absolutely stable, giving a signature of missing energy in the detector. Secondly we expect SUSY events to exhibit relatively large multiplicity of high $p_T$ jets, as SUSY particles with strong interaction couplings have highest production rates at the LHC and these will go through long decay chains. Furthermore it was decided to study SUSY with different search strategies based on exclusive requirement of zero, one or more leptons.

The aim of the described studies was to develop data driven techniques to estimate the major Standard Model (SM) backgrounds such as $W/Z$ bosons with associated jets, top quark pair production and jet production from QCD processes. SUSY should show up as an excess over the predicted SM events in the so-called “signal” region, where new physics may be present. The prediction is done by extrapolating from a “control” region, which should be as close as possible to the signal region, give an unbiased estimate, have sufficient statistics and small theoretical uncertainties. The first requirement has the effect of contaminating the control region with SUSY events, which results in overestimating the SM background in the signal region, but this contamination can be taken into account.

We assumed the LHC running with $\sqrt{s} = 14$ TeV/$c^2$ and a collected integrated luminosity of 1 fb$^{-1}$. Due to space limitations the methods described in this note are a subset of all the methods developed and the reader is referred to [1] for a complete description.

2. One-lepton search mode
The one-lepton search mode requires an isolated lepton. This will strongly suppress the QCD background that dominates over other processes by orders of magnitude and has high theoretical
and instrumental uncertainties. A lepton in our studies is either an electron or a muon with 
\( p_T \) of at least 20 GeV/c. To avoid overlap with the di-lepton mode, we veto events with a 
second identified lepton with a \( p_T \) of more than 10 GeV/c. Assuming R-parity conservation we 
demand at least four jets with \(|\eta| < 2.5\) and \( p_T > 50\) GeV/c out of which one jet must have 
\( p_T > 100\) GeV/c. The missing transverse energy \( E_{T}^{\text{miss}} \) should be larger than 100 GeV/c 
and above \( 0.2M_{\text{eff}} \), where \( M_{\text{eff}} \) is the effective mass. Our last selection criterion is that transverse 
sphericity \( S_T \) is larger than 0.2. All definitions of variables can be found in [1].

2.1. Combined fit method

After the selection the only non-negligible backgrounds left are W bosons with associated 
jets (W+jets) and top quark pairs. The latter we divide into two categories: semileptonic 
\((t\bar{t} \rightarrow b\bar{b}\ell\nu\bar{q})\) and dileptonic \((t\bar{t} \rightarrow b\bar{b}\ell\nu\ell)\) top quark pairs, as these have different shapes 
and yields. Dileptonic top quark pair events end up in our one lepton sample because either 
the second lepton was mis-/not reconstructed or the second lepton was a tau lepton decaying 
hadronically, respectively constituting one third and two-thirds of all dileptonic events.

We fit the SM backgrounds in three observables: \( E_{T}^{\text{miss}}, M_T\) and \( m_{t_{\text{top}}} \). \( M_T\) is the invariant 
mass of \( E_{T}^{\text{miss}} \) vector and the lepton \( p_T \), while \( m_{t_{\text{top}}} \) is the invariant mass of the three jets with 
largest vector-summed \( p_T \).

Taking physics features into account we construct probability density functions (p.d.f.’s) that 
model each of the three main backgrounds in the three observables. For example the semileptonic \( t\bar{t} \) shows a clear peak in \( m_{t_{\text{top}}} \) if we find the correct three quarks from top quark decay.

For a broad range of SUSY parameters it was observed that the shapes of low \( E_{T}^{\text{miss}} \) and low 
\( M_T \) distributions have little dependence with the chosen model point. Thus we can construct a 
model-independent Ansatz shape to describe the SUSY contamination at low energy.

The combined model used in the fit is the addition of p.d.f.’s of each background sample and 
the SUSY Ansatz shape with a yield parameter for each separate component. The first step in 
the procedure is to fit the \( E_{T}^{\text{miss}}, M_T\) and \( m_{t_{\text{top}}} \) distributions with shapes obtained from Monte Carlo simulation. To make it more data-driven is to release in the fit as many of the shape 
parameters as possible. Figure [1] shows the result of the fit with floating yields and floating 
shape parameters to 1 fb\(^{-1}\) of data.

The final step is to extrapolate the yields of the SM background components to the 
signal region. Table 2.1 shows the extrapolated yields to the signal region (SIG), defined

![Figure 1](image-url)

**Figure 1.** Distribution of \( E_{T}^{\text{miss}} \) (left), \( m_T \) (center) and \( m_{t_{\text{top}}} \) (right) of a 1 fb\(^{-1}\) mix of \( t\bar{t} \) and W+jets SM events and SUSY SU3 events overlaid with projections of the combined model 
fit to the mix of events with floating yield and shape parameters. For each projection the contributions of the semileptonic \( t\bar{t} \) contribution (dark blue), dileptonic \( t\bar{t} \) contribution (light blue), W+jets contribution (red) and Ansatz SUSY contribution (black) are shown.
Table 1. Yields from the combined fit with either fixed or floating shapes extrapolated to the full parameter space, the truth yields in full parameter space, the extrapolated yields into the signal region and the truth yields in the signal region.

| Component     | Extrap. Yield in FULL | True | Extrap. Yield in SIG | True |
|---------------|-----------------------|------|----------------------|------|
|               | Shape Fixed           | Shape Floating |                       | Shape Fixed | Shape Floating | SIG |
| W + jets      | 205 ± 45              | 227 ± 68 | 173                   | 0.5 ± 0.4 | -1.2 ± 2.7 | 2   |
| 1-lepton $t\bar{t}$ | 476 ± 35              | 485 ± 59 | 502                   | 0.4 ± 0.2 | -1.1 ± 3.9 | 0   |
| 2-lepton $t\bar{t}$ | 62 ± 38               | 17 ± 54  | 70                    | 4.5 ± 2.9 | 4.7 ± 7.9  | 5   |
| SUSY SU3      | 273 ± 33              | 287 ± 38 | 271                   | 92.7 ± 2.8 | 95.6 ± 4.0 | 91  |

by $E_T^{\text{miss}} > 200$ GeV/c and $M_T > 150$ GeV/c$^2$, while propagating all correlated parameter uncertainties with fixed and floating shapes. For comparison the same table shows extrapolated yields to the full parameter space (FULL) from the control region. The yields that we find are in good agreement with the truth values of the fitted event mix within statistical uncertainties.

2.2. HT2 method
If we add an extra cut of $M_T > 100$ GeV/c$^2$ to our event selection the only significant background we are left with is dileptonic $t\bar{t}$. The HT2 method estimates this background by using two near independent variables HT2 and $E_T^{\text{miss}}$ significance defined as:

$$\text{HT2} \equiv \sum_{i=2}^{4} p_{\text{jet}}^{i} + p_{\text{lepton}}^{i}, \quad E_T^{\text{miss}} \text{significance} \equiv E_T^{\text{miss}} /[0.49 \cdot \sqrt{\sum E_T}] \quad (1)$$

The leading jet $p_T$ was excluded from HT2 in order to reduce the correlation with $E_T^{\text{miss}}$. The correlation between the highest-$p_T$ jet and $E_T^{\text{miss}}$ is a consequence of kinematics, as the rest of the event recoils from this jet. As the $E_T^{\text{miss}}$ resolution depends on $\sum E_T$ that is clearly related to HT2, $E_T^{\text{miss}}$ significance was used instead of $E_T^{\text{miss}}$ to further remove any correlation.

Figure 2. Left: Histogram is observed $E_T^{\text{miss}}$ significance distribution, open circles are true SUSY signal, blue triangles are true SM background, black filled circles is the estimated background. Right: Open circles is true SUSY signal as a function of $E_T^{\text{miss}}$ significance, black is estimated SUSY excess obtained by subtracting the estimated background from the observed $E_T^{\text{miss}}$ significance distribution.
A control sample is defined by HT2 < 300 GeV/c from which the shape of $E_T^{\text{miss}}$ significance is taken. The assumption is that this shape is independent of HT2. The normalization of this shape is obtained by comparing the event counts of the control sample to the signal sample at $8 < E_T^{\text{miss}}$ significance < 14. This low $E_T^{\text{miss}}$ significance region is equivalent to a low $E_T^{\text{miss}}$ region almost unpopulated by SUSY.

The background distribution and the excess of SUSY signal in high $E_T^{\text{miss}}$ significance is shown in Figure 2. The background is somewhat overestimated due to SUSY contamination in the control sample. However, the right histogram of Figure 2 shows that if we would cut higher on $E_T^{\text{miss}}$ significance, the SUSY signal would still be clearly seen over the estimated background.

3. No-lepton search mode
As the name suggests we veto all events with an identified muon or electron of $p_T > 20$ GeV/c. For the rest the no-lepton mode event selection is equivalent to one-lepton mode concerning jets, $E_T^{\text{miss}}$, $S_T$ and $M_{\text{eff}}$. To pass the $E_T^{\text{miss}} > 100$ GeV/c requirement QCD events must contain a poorly reconstructed jet which can be caused by dead material, jet punch-through, pile-up of machine backgrounds and others. $E_T^{\text{miss}}$ will then point to or away from this poorly reconstructed jet. To reject these QCD events the value of the difference in azimuthal angle between the $E_T^{\text{miss}}$ vector and each of the three highest-$p_T$ jets is required to be larger than 0.2.

3.1. Replace method
The $Z \rightarrow \nu\bar{\nu}$ with associated jets process is one of the main backgrounds in the no-lepton search. To estimate its shapes and expected number of events $Z \rightarrow \ell^+\ell^-$ events are selected by requiring: two opposite charged leptons, invariant mass of the two leptons within 10 GeV/c and each of the three highest-$p_T$ jets is required to be larger than 0.2.

Four corrections must be applied to get a correct estimate, as summarized by this formula:

$$N_{Z\rightarrow\nu\bar{\nu}}(E_T^{\text{miss}}) = \frac{N_{Z\rightarrow\ell^+\ell^-}(p_T(\ell^+\ell^-))}{\text{eff}(\eta_{\ell^+}, p_T, \ell^+)} \cdot \text{eff}(\eta_{\ell^-}, p_T, \ell^-) \times c_{\text{Fidu}}(p_T(Z)) \times \frac{\text{Br}(Z \rightarrow \nu\bar{\nu})}{\text{Br}(Z \rightarrow \ell^+\ell^-)}.$$  

where $N_{Z\rightarrow\nu\bar{\nu}}$ is the corrected number of events per bin of $E_T^{\text{miss}}$, $N_{Z\rightarrow\ell^+\ell^-}$ is the raw number of control sample events as a function of $p_T(Z)$, $c_{\text{Fidu}}$ is the kinematic correction due to extra selection criteria, $c_{\text{Fidu}}$ is the fiducial correction since we cannot detect leptons beyond $|\eta| < 2.5$, eff$(\eta_{\ell^+}, p_T, \ell^+)$ and eff$(\eta_{\ell^-}, p_T, \ell^-)$ are corrections for lepton reconstruction efficiency as a function of $\eta$ and $p_T$ and finally we correct for the difference in branching ratios of the Z boson.

This method does not suffer from SUSY contamination as the stringent control sample selection cuts make it practically free of SUSY. The described procedure estimates the $Z \rightarrow \nu\bar{\nu}$ background correctly but its precision is limited by statistics in the control sample.

4. Conclusion
What we have shown in this note are a number of methods to estimate the top quark pair, W+jets, Z+jets and QCD backgrounds from data in regions where we expect to find a SUSY excess. For an integrated luminosity of 1 fb$^{-1}$ many complementary methods have been developed and are shown to reliably assess the SM backgrounds as well as the possible SUSY excess. Ongoing work focussed on the expected early LHC running scenario with a integrated luminosity of 200 pb$^{-1}$ at collision energies close to 10 TeV/c$^2$ shows promising results.

References
[1] The ATLAS Collaboration, *Expected Performance of the ATLAS Experiment, Detector, Trigger and Physics*, CERN-OPEN-2008-020, Geneva, 2008