Project Proposal

Robert Brockman, Justin DeVito, Ricky LeVan

March 17, 2013

Statement of Problem

With the discovery of the Higgs-like particle at the Large Hadron Collider (LHC), all of the important predictions of the Standard Model of particle physics have been tested. The focus of particle physics has now shifted to the search for new particles associated with extensions of the Standard Model. One such search involves the search for a particular particle associated with an extension to the Standard Model known as supersymmetry (SUSY). In particular, physicists at the Compact Muon Solenoid (CMS) detector group at the LHC are interested in verifying a SUSY model that predicts specific p-p collision events which generate the super-symmetric partner to the top quark known as the *stop* squark. Verification of these models requires an efficient classifier capable of separating signal *stop* squark events from background events. More efficient classifiers allow discoveries to be made with fewer events, less expensive beam time, and thus lower cost.

Objectives

Our initial objectives are as follows:

- Use Neural Networks to build a classifier for distinguishing two types of collision events. One class of event indicates the presence of supersymmetric particles (specifically, a *stop squark*), and the other class of event represents a troublesome background.
- Our goal is to make a classifier that is better and more efficient at separating the signal from the background than models currently being developed and used. We will compare our results against two other methods of classification:
  - Theoretical particle physicists have come up with a means to separate signal and background using a series of cuts on a set of eight derived parameters (discussed below). [1]

  - Onkur Sen, a physics senior working with Dr. Paul Padley at Bonner Lab here at Rice, is attempting to improve upon this with a boosted decision tree algorithm, using the same derived parameters.

- The efficiency of the classifiers will be compared by measuring the discovery significance per unit beam luminosity.

If these objectives can be achieved quickly and easily, we can proceed to accomplish larger goals:

- We would like to create a production version of the neural network algorithm. This would include parallelizing the code so that it can run efficiently on Bonner Lab and RCSG computing clusters at Rice. We want to create a working tool that CMS can use to evaluate real physics data.

- Another further goal is to build an unsupervised clustering method to detect the occurrence of new types of events in simulated collisions.

Technical Approach to Initial Goals

Our classifier will be trained on simulated data generated by PYTHIA, a particle physics simulator which can be tuned to represent the appropriate SUSY model. If the SUSY theory under consideration is correct, a classifier which works on the simulated data will detect the *stop* squark on the real data from the CMS detector.

Physics of the Signal Events

The event type we are trying to detect is shown below:

$$
\begin{aligned}
p + p &\rightarrow \tilde{t} + \bar{\tilde{t}} \\
\tilde{t} &\rightarrow t + \widetilde{\chi}_1^0 \\
\bar{\tilde{t}} &\rightarrow \bar{t} + \widetilde{\chi}_1^0 \\
t &\rightarrow b + j + j \\
\bar{t} &\rightarrow \bar{b} + j + j
\end{aligned}
$$

Here components of two protons colliding roughly head on down the accelerator beam axis combine to form two *stop squarks*, represented by $\tilde{t}$ and $\bar{\tilde{t}}$. (The bar on top represents an antiparticle.) The stop squarks decay into top quarks and neutralinos. Neutralinos escape the

experiment undetected, carrying missing energy with them. In the signal event, the top quarks then decay into three other quarks, one of which is a b quark and the other two are lighter quarks from the 1st two particle generations. The quarks form recognizable jets of particles in the detector, and the jets from the b quarks can be differentiated from the others. Thus the detector will see this event as 6 jets, 2 of which are from b quarks, and there will be considerable missing transverse energy caused by the escape of the neutralinos. (The total transverse momentum is the vector sum of all of the momenta perpendicular to the beam axis of the detected particles . If it is not zero, momentum conservation demands that there be undetected particles which have escaped with the missing momentum.)

The relevant raw numbers from this event are therefore the energy-momentum four-vectors for the four normal jets (j) and two b jets (b), for a total of 24 scalars per event.

Physics of the Background Events

Unfortunately, there is a troublesome background event that looks very similar to the signal event, but contains no interesting new physics:

$$
\begin{aligned}
p + p &\rightarrow t + \bar{t} + j + j \\
t &\rightarrow b + j + j \\
\bar{t} &\rightarrow \bar{b} + W^- \\
W^- &\rightarrow e^- + \nu_e
\end{aligned}
$$

Here the initial collision produces two top quarks directly as well as two low-mass quarks which form jets. One of the top quarks decays into a b quark and two low mass quarks as in the signal event. The other top quark decays into a b quark, a lepton (electron or muon) and a neutrino which carries off missing energy. The big problem is that lepton detection is not perfect, and when the lepton is missed, this event also appears as 6 jets, 2 of which are from b quarks, and with missing transverse energy, just like the signal event. Our task is to study the four-vectors of the 6 jets in each of the event types to see if we can distinguish them.

Derived Scalars - Physics Perspective

Dutta, et. al. have derived 8 scalar features to help in identifying the background event. [1]. The basic strategy comes from the observation that two of the jets from low-mass quarks (hereafter j) emerge directly from the initial collision in the background event, whereas in the signal event the j's are paired and each pair is associated with a b quark. Thus, the patterns in the angles between jets and the energy of the jet groupings should be different.

Python scripts provided by Onkur Sen compute the scalar features for each event and output them in a format which we can feed into a classifier. The scripts group the quark jets from each event into two groups of 3 jets each, referred to as System A and System B. Each group contains one b quark and 2 normal jets. The scripts then compute the following:

- M3 : The invariant mass of all three jets in a group. This scalar does not change with reference frame, and should correspond to the rest mass of the top quark that spawned the jets. This is computed for both groups, so we have M3A and M3B.
- M2: The invariant mass of the two non-b jets in a group. This should correspond to the difference between the top and b quark rest masses in the signal case. In the background case, two non-b jets emerged from the initial collision and thus there should be less correlation. This is computed for both groups, so we have M2A and M2B.
- B_ANGLES: The azimuthal angle for the b quark in System B.
- J1_ANGLES: The azimuthal angle for the 1st non-b quark in System B.
- J2_ANGLES: The azimuthal angle for the 2nd non-b quark in System B.
- MISSING_E: The missing transverse energy.

Derived Scalars - Machine Learning Perspective

The data we initially plan to classify with an ANN thus consists of vectors of 8 real-valued scalars, 5 of which are energies in GeV (range 0-1000, mostly around 200) and 3 of which are angles in radians (range 0-π). Each of the vectors is labelled as signal or background, so we have only 2 classes. Right now we have 2302 signal events and 2847 background events generated by Onkur's scripts, but we can get more. Thus we have enough data to split them 50-50 into a training and test set.

The figures in the section below show the histograms that Onkur has created of some of the above features. Angles are in radians, whereas the units for M3, M2, and MISSING_E are GeV. The scaling for signal and background in these graphs isn't great – we will need to make our own, better graphs soon.

Use of ANNs

We will run the data through a supervised learning paradigm to teach a network to respond appropriately to the presence of the absence of the *stop* particle. In this we will use our standard, n-layer backpropagation algorithm, beginning with 2 layers and adding more if necessary. We

will have as many input PE's as prominent features (initially, the 8 derived features), and since we are working with a simple classifier, we will simply have 2 output PE's.

Our code still needs to be extended to allow for time-decreasing learning rates. Turning our algorithm into production code usable by CMS will require parallelizing the batch learning functionality.

Additional Work

Contingent upon the time taken to complete the steps mentioned thus far, we may be able to further analyze this data using unsupervised learning methods. We may be able to find a way to classify these different events by looking at the way in which the 2-D competitive SOM became organized. This may help to characterize the separation between the two different types of events. We can also compare these methods to the derived parameters to see if the detected separating parameters coincide with the derived parameters. Alternatively, we could use the number of significant dimensions from a GHA based approach to indicate how complex the events are.

Graphs of Derived Parameters

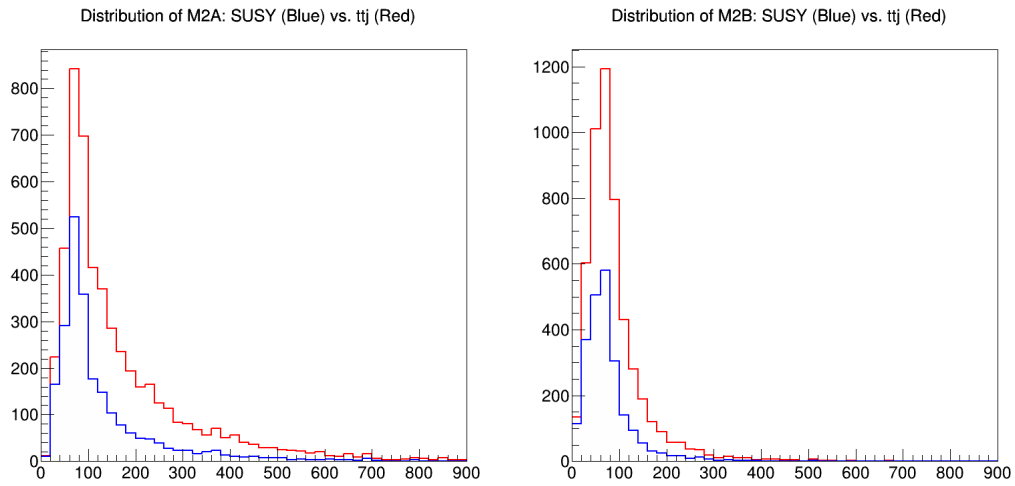

Figure 1: These compare the M2 of the top quark in systems A (LEFT) and B (RIGHT) for signal (Blue) and background (Red). Y-axis is event counts per bin, X-axis is invariant mass in GeV.

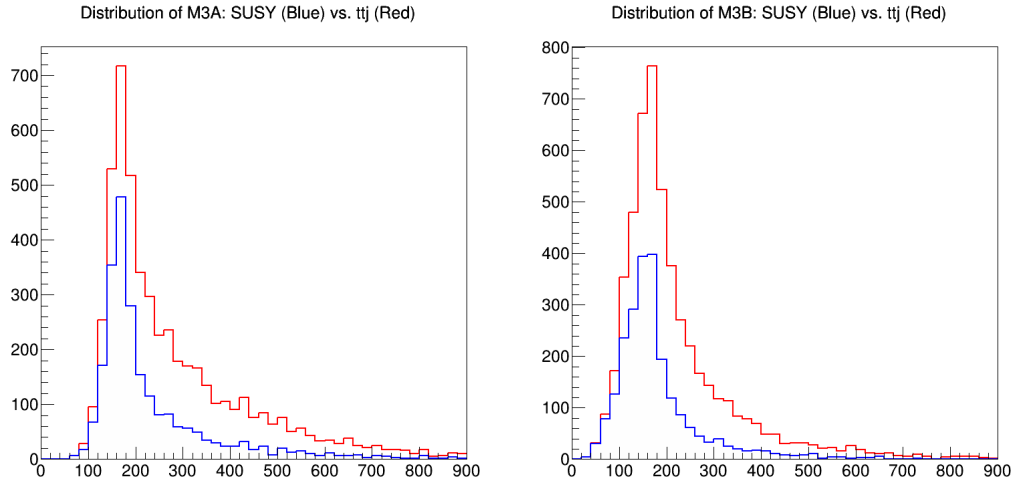

Figure 2: These compare the M3 of the top quark in systems A (LEFT) and B (RIGHT) for signal (Blue) and background (Red). Y-axis is event counts per bin, X-axis is invariant mass in GeV. Observe how M3B for signal is much lower than for background.

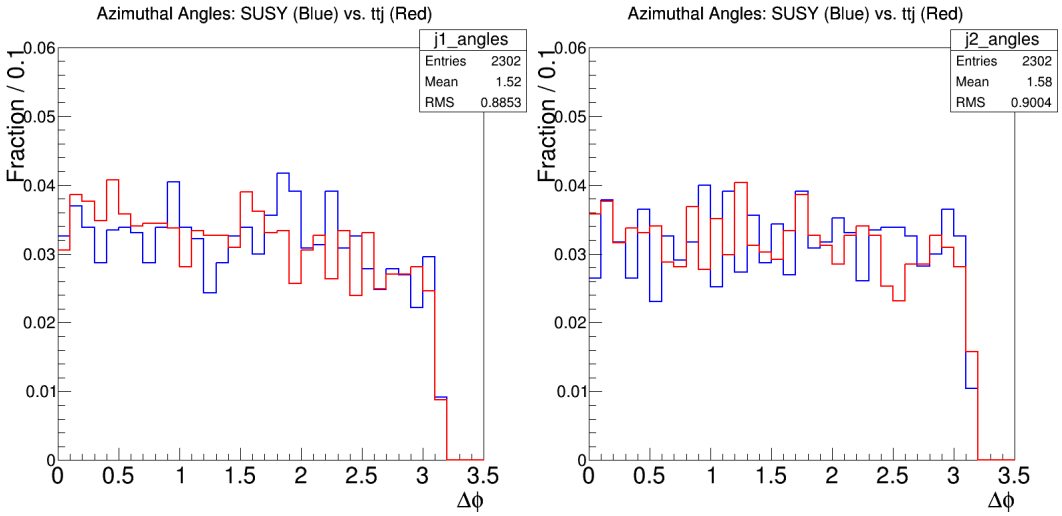

Figure 3: These figures compare the azimuthal angles (in radians) of Jet 1 (LEFT) and Jet 2 (RIGHT) of System B for the signal events (Blue) and background events (Red).

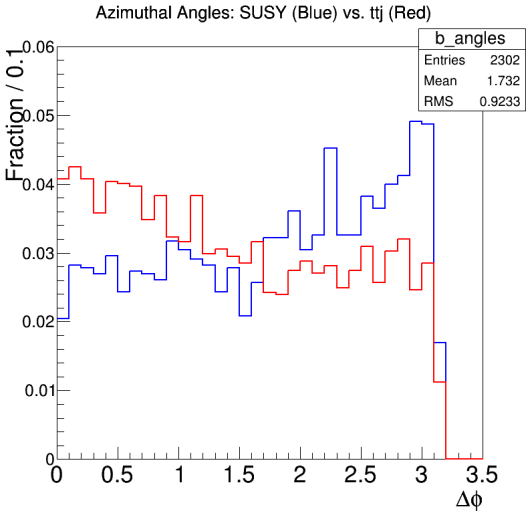

Figure 4: These figures compare the azimuthal angles (in radians) of the b jet of System B for the signal events (Blue) and background events (Red). This parameter looks particularly useful.

Next Steps

- Onkur does not seem to have provided the graph for missing energy, so we should get this or make one ourselves.
- The branching ratio for the signal and background events is not as simple as initially thought, since it depends on the probability that the detector fails to detect a lepton in the background case. We need to know more about the detector itself to get this number – Dr. Padley will know this.
- The data from Onkur comes in two forms: raw PYTHIA events which can be transformed into input vectors by a modified version of Onkur's scripts, and sets of eight files, one for each derived parameter. This latter version of the data is more compact and thus has more data points, but is awkward to work with. The raw data files are huge, and the slice of it we have right now doesn't have nearly enough background data points. As a result, initial results from feeding the raw files through Onkur's scripts and into WEKA weren't very good. Robert will go grab more data from Onkur on Monday.

References

[1] B. Dutta, T. Kamon, N. Kolev, K. Sinha, and K. Wang, "Searching for top squarks at the lhc in fully hadronic final state," *High Energy Physics - Phenomenology* (2012) , arXiv:1207.1873v1 [hep-th].