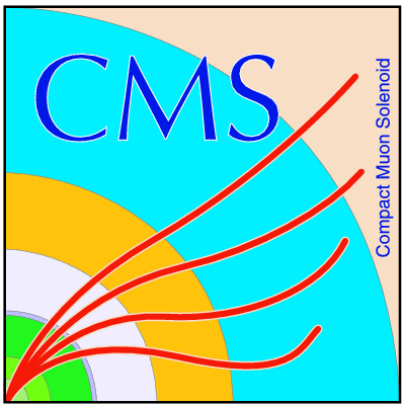
Determining Optimal Control Regions in the MSSM Higgs Search at CMS

DESY Summer Student Programme, 2013

Ewen Lawson Gillies IX
University of Edinburgh, UK

Igor Marfin, Wolfgang Lohmann



31st of August 2013

Abstract

The CMS detector at the LHC is currently analyzing data in a effort to find signatures of the supersymmetric neutral Higgs-like bosons. As a part of this analysis, a control region with a high background to signal ratio must be determined. ROOT's Toolkit for Multivariate Analysis was used to find the optimal cuts across nine input kinematic variables. This analysis returned a viable control region, with a background to signal ratio of 7.6, while retaining a background selection efficiency of 60%.

Contents

1 Introduction

Over the past few decades, numerical theory has become an integral part of scientific approaches. This is especially true when it comes to solving partial differential equations, many of which have no known analytical solution. These types of equations are often useful in describing physical processes, especially in astrophysics. This paper aims to outline some of the fundamentals of differential numerical theory by exploring several different types of numerical differentiation and integration. These methods will be compared and applied to a model differential equation.

1.1 Background

1.2 Theory

The CMS detector is one of two general purpose detectors at the LHC. It is equipped with several layers of detectors, each of which is designed to measure different properties of the final state particles.

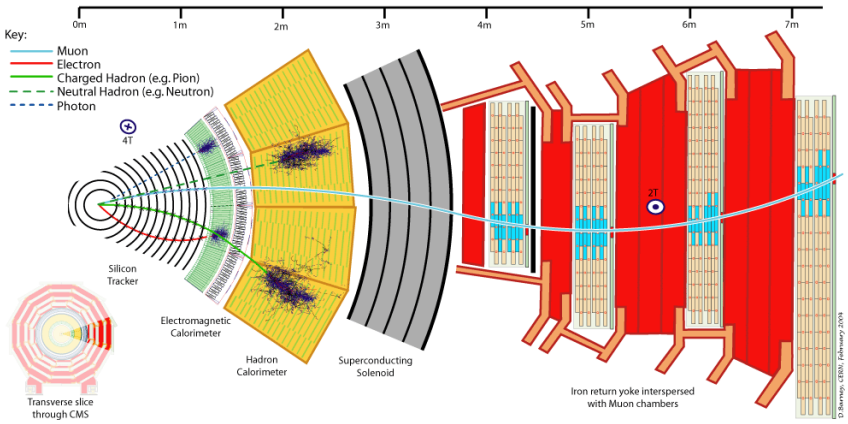


Figure 1: The path of particles as they pass through all the CMS detectors [2]

The first layer is the silicon tracker detector. It reconstructs the path, or *track*, of post collision muons, electrons and charged hadrons as they pass through. The particles then reach the electromagnetic calorimeter. As the name suggests it measures the energy of electromagnetic particles, namely electrons and photons. The leftover particles, unaffected by the electromagnetic calorimeter, now pass through the hadron calorimeter. This measures the energy of hadrons, which are particles made of quarks and gluons. The muon detector sits on the outermost layer of CMS. [3, 4].

The combination of all these layers allows the CMS collaboration to accurately measure final state particles from a collision. From this data, physicists are able to reconstruct the interactions of the full event, including signatures of currently undiscovered particles.

1.3 Jets

The main physical objects that this paper deals with are *jets*. Jets result from the fundamental nature of the strong force. Since the strong force coupling increases with distance, unbound quarks are never observed in nature. Instead, they exist in hadronic bound states. In proton-proton collisions at high energies, the partons involved in the hard interaction may fly apart to form several hadrons. This process is referred to as *hadronization*.

The resulting particles are unstable. This causes them to decay into lower energy hadrons. The particles in these bunches all travel in approximately the same direction, and hence are referred to as jets. Jets are reconstructed from final state particles using *clustering algorithms*. Furthermore, jets can be described by their own set of kinematic properties [5].

1.3.1 Properties

The fundamental property a jet is its four momentum, which is simply a sum of the four momentum of all its constituent particles. From this, the transverse momentum, traverse energy, rapidity, pseudorapidity, azimuthal angle, and polar angle can be defined. For a jet J, consisting of N particles, each denoted by index i, these properties are defined as the following [5]:

- Four Momentum: $p^J = \left(E^J, \mathbf{p}^J\right) = \sum_i^N \left(E^i, p_x^i, p_y^i, p_z^i\right)$
- Tranverse Energy: $E_T^J = \sqrt{\left(m^J\right)^2 + \left(p_T^J\right)^2}$
- Tranverse Momentum: $p_T^J = \sqrt{\left(p_x^J\right)^2 + \left(p_y^J\right)^2}$
- Rapidity: $y_T^J = \ln \sqrt{\frac{E^J + p_z^J}{E^J - p_z^J}}$
- Pseudorapidity: $\eta_T^J = \ln \sqrt{\frac{|\mathbf{p}^J| + p_z^J}{|\mathbf{p}^J| - p_z^J}}$
- Azimuthal Angle: $\phi^J = \arctan \frac{p_y^J}{p_x^J}$
- Polar Angle: $\theta^J = \arccos \frac{p_z^J}{p_T^J}$

1.3.2 Identification

In order to group final state particles into a jet, the "anti-k_t" jet-clustering algorithm is used. This process compares effective distances between entities i and j, and entity i and the beam, B. The goal of this algorithm is to identify all high-p_T, or *hard*, particles as the origin of a jet, grouping the appropriate low-p_T, or *soft*, particles to this jet. This is done by identifying the smallest distance between entities. The distances are defined by the following, where $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, and R is a radius parameter defined in the analysis.

$$d_{ij} = \min\left[(p_T^i)^{-2}, (p_T^j)^{-2}\right] \frac{\Delta_{ij}^2}{R^2} \qquad d_{iB} = (p_T^i)^{-2}$$

The algorithm groups particles i, j into the same jet for the case where d_{ij} is smaller than d_{iB}. Otherwise, it defines a new jet for particle i. Since hard particles have a much higher transverse momentum than soft particles, this minimum function in this algorithm ensures hard particles group with soft particles long before soft particles will group to each other. The R parameter is analogous to a cone parameter around the hard particle, inside which all soft particles will be included in the jet [6].

1.4 Supersymmetry

One of the most developed framework for physics beyond the Standard Model is Supersymmetry (SuSy). It was first formulated in the mid 1970's as a solution to the "fine tuning" problem of the SM [1]. Since much of the SM is built around the observed symmetries in nature, supersymmetry looks to extend this success by positing an as of yet unobserved symmetry. It predicts that every particle has a superpartner, or a corresponding particle with similar properties, yet opposite spin statistics. Hence, for every boson, there is a corresponding particle that obeys fermionic statistics and vice versa.

[1] This problem requires a delicate balancing of divergent terms, due to loop-level contributions of particle interactions. See [7] for more details

1.4.1 SuSy Breaking

It is important to note that no supersymmetric particles have been observed yet. While the symmetry itself suggests that all SM particles have a superpartner of the same mass, the fact that none have been observed suggests that supersymmetry is a broken symmetry [8]. The favoured supersymmetric theories are ones that contain soft SUSY breaking, which means that the mass between the SM particles and the SUSY counterparts is different, but only by an order of magnitude or so at the most. As experiments reach higher and higher energies, the absence of observed superpartners pushes this theory to more and more unnatural territory.

1.4.2 Minimal Supersymmetric Model

The most minimal form is appropriately named the Minimal Super Symmetric Model (MSSM). Even in this economical form, there exist two Higgs doublet superfields, each comprised of a fermionic component, called a Higgsino, and a scalar component, the Higgs boson. These doublet fields are described by $H_u = \left(H_u^+ \; H_u^0\right)$ and $H_d = \left(H_d^0 \; H_d^-\right)$. Due to chirality considerations, supersymmetry demands the existence of at least two such fields, where H_u gives mass to the "up-like" quarks, and H_d gives mass to the "down-like" quarks. Each of these multiplets contain an electrically neutral component, denoted with a 0, and a charged component, denoted with a plus or minus respectively. All of these components can be complex [7].

This model is the simplest supersymmetric extension and has been referred to as the "standard model" for physics beyond the SM. As illuminated before, the scalar Higgs fields in MSSM consists of two complex doublets, H_u and H_d, yielding a total of eight degrees of freedom. When the electroweak symmetry is broken, the three would-be the Goldstone bosons of the potential are absorbed into the longitudinal degrees of freedom of the Z^0 and $W^\pm$ bosons [5]. The remaining five degrees of freedom manifest as two CP-even neutral scalar bosons, h^0 and H^0, one CP-odd neutral scalar boson A^0, a positively charged scalar boson H^+ and a negatively charged scalar boson H^- [8]. By convention, h^0 is of smaller mass than H^0, which is theoretically limited by $m_{h^0} \lesssim 135$ GeV [2] [8]. This h^0 boson is a superposition of the neutral components of H_d and H_u, hence it couples to both "up" and "down" like quarks. Given the upper limit on its mass, it closely resembles the SM Higgs boson, and can naïvely be considered the MSSM equivalent [8].

1.5 MSSM Higgs Search

The ultimate goal of this study is to search for neutral scalar decays into two bottom quarks at CMS. This involves colliding two protons, which creates a scalar particle, which then decays into two bottom quarks. This decay channel can be visualized in the adjacent Feynman diagram in Figure 2.

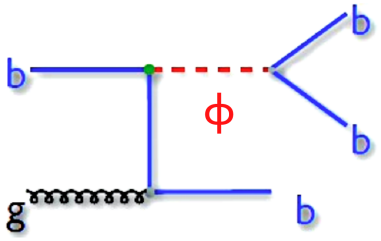


Figure 2: Decay channel

The desired final state is characterized by three b-quark jets that fall within the tracker acceptance and satisfy kinematic parameters. A *signal* event occurs when two of these jets result from the one of the desired neutral scalar bosons. These jets are expected to have the highest two transverse momenta in the interaction, with the third jet following. The *background* for this process is either composed of three genuine b-jets from QCD processes, or two b-jets and one misclassified jet originating from a lighter parton.

MSSM has three candidates for the scalar boson in this process: the h^0, the H^0 and the A^0 bosons. As such, additional mass resonances are expected mass spectrum of the two resulting b-quark jets. This is referred to as the dijet mass spectrum. This study is looking at a medium mass scenario which requires that the mass of the scalar particle(s) is above 180 GeV [9].

[2] This limit assumes that all supersymmetric particles that contribute to the $m_{h^0}^2$ term in loops do not have a mass that exceeds 1 TeV.

In order to verify the results of this study, a control region must be determined where no signal is expected. By probing both control regions and signal regions, this study can confirm that any resulting signal is not a result of event selection bias. The optimal control region is determined using multivariate analysis techniques.

1.6 Multivariate Analysis

In order to determine the optimal control region, cuts must be made across many variables. Making simple cuts across each variable independently results in a low sensitivity since these variables are not independent. Determining the best cut necessitates multivariate analysis techniques which are conveniently provided by ROOT's Toolkit for Multivariate Analysis (TMVA). TMVA provides several classification tools, all of which follow the same general, two step "supervised learning" structure. The goal of these tools is to be able to neatly classify events as background or signal.

The first step makes use of Monte Carlo generated training events, for which the desired classification is already known. From these events, the classifier builds a mapping function that describes a "decision boundary." This boundary represents the optimal cuts across all variables to distinguish background from signal. Another set of statistically independent events are then passed through the classifier to test if this decision boundary holds for any event set [10].

1.7 Input Kinematic Variables

Events are characterized by their measurable final state kinematic properties. The best variables for classification are then determined. Ideally, each variable used has a different distribution for background and signal events and no two variables are correlated. Furthermore, to avoid sample biasing, none of the preselection parameters can be used, as discussed in section 2.3. This study uses nine input kinematic variables, each of which are outlined below.

For the following variables, note that the input jets are ranked by transverse momentum. It is expected that the first two leading jets, i.e. jet 1 and jet 2, result from the decay of the scalar boson. The third jet is expected to result from the third b-quark in the process.

- `dphijet2jet3_boost13`: The difference in the polar angles of the second and third jet in the center of mass frame of the first and third jet system.
- `Et3byEt2`: The transverse energy of the third jet divided by the transverse energy of the second jet. Note that this is always less than 1 by definition.
- `phijet3_boost12`: The polar angle of the third jet in the center of mass frame of the first and second jet system
- `thetajet1_boost13`: The polar angle of the first jet in the center of mass frame of the first and third jet system
- `ptd1`: The area of the leading jet upon detection.
- `ptd2`: The area of the sub-leading jet upon detection.

For the following variables, the thrust and thrust axis must first be defined. The thrust describes the coordination the motion of the final state particles. It is defined as the following quantity, T, where $\mathbf{n}_T$ is the thrust axis. Note that the thrust axis maximizes the thrust by definition and that the index i runs over all particles in the event [11].

$$T = \max_{\mathbf{n}_T} \left(\frac{\sum_i |\mathbf{p}_i \cdot \mathbf{n}_T|}{\sum_i |\mathbf{p}_i|} \right) \qquad T_{\mathrm{Ma}} = \max_{\mathbf{n}_{\mathrm{Ma}} \perp \mathbf{n}_T} \left(\frac{\sum_i |\mathbf{p}_i \cdot \mathbf{n}_{\mathrm{Ma}}|}{\sum_i |\mathbf{p}_i|} \right) \qquad T_{\mathrm{Mi}} = \frac{\sum_i |\mathbf{p}_i \cdot \mathbf{n}_{\mathrm{Mi}}|}{\sum_i |\mathbf{p}_i|}$$

- `major`: Thurst Major, T_{Ma}, which is the defined similarly to the thrust along a different axis, $\mathbf{n}_{\mathrm{Ma}}$. This axis must be perpendicular to $\mathbf{n}_T$ and maximize T_{Ma}.

- `minor` Thurst Minor, T_{Mi}, which is the defined similarly to the thrust along a different axis, $\mathbf{n}_{\mathrm{Mi}}$. This axis is defined as the one perpendicular to both $\mathbf{n}_{\mathrm{Ma}}$ and $\mathbf{n}_T$.

- `Bmin` Narrow jet broadening, B_N, is defined by splitting the event into two hemispheres using a plane through the origin and perpendicular to $\mathbf{n}_T$. The following value is calculated for both hemispheres, where B_N is the lesser of the two values.

$$B_N = \min_{i\in\{1,2\}} \left(\frac{\sum_{k\in H_i} |\mathbf{p}_k \times \mathbf{n}_T|}{2\sum_j |\mathbf{p}_j|} \right)$$

2 Method

This study uses Monte Carlo simulations to build and test multivariate analysis techniques that will eventually be used on real data. This implies that these events must fall within trigger requirements and preselection criteria to be considered. Once a sample is compiled, it is used to build a Boosted Decision Tree (BDT) classifier. This classifier is then tested again on a statistically independent sample.

2.1 Event Selection

Before events are considered for analysis, they must contain at least three jets, each satisfying the following requirements [9].

- $p_T^{J1} > 60$ GeV ; $p_T^{J2} > 53$ GeV ; $p_T^{J3} > 20$ GeV
- $|\eta^J| < 2.2$ for the three leading jets.
- $\Delta R_{12}^2 = (\eta_1 - \eta_2)^2 + (\phi_1 - \phi_2)^2 > 1$

2.2 b-Jet Tagging

Hadrons containing b-quarks have an average lifetime of $\tau_B \approx 1.6$ ps, which corresponds to a flight distance on the order of centimeters when traveling close to the speed of light. This means that any b-hadrons resulting from the initial interaction at the *primary vertex* will decay inside the detector at a *secondary vertex*. Jets originating from b-quarks to be easily tagged by identifying a secondary vertex characteristic of the mean b-hadron lifetime.

In the absence of identifying a secondary vertex directly, shortest distance of approach of particles resulting from a b-hadron are measured from the primary vertex. This is referred to as the impact parameter (IP). Particles originating fro the primary vertex have an IP that is close to the resolution of the detector, whereas those from the secondary vertex may have a much larger IP. This difference, illustrated below in Figure 3, allows for decay products of b-hadrons to be grouped into a b-jet [12].

The secondary vertex and IP methods are combined in Combined Secondary Vertex CSV algorithm. This algorithm outputs a value in the range $0 \leq d \leq 1$, with 1 signifying a "perfect" b-jet. This study opts for a tight CSV parameter of $d > 0.898$. Vertices from other long-lived particles are suppressed by requiring that the secondary vertex is less than 2.5 cm away from the primary vertex. Additionally, no tracks with mass greater than 6.5 GeV or a comparable mass to the K_s^0 meson were considered.

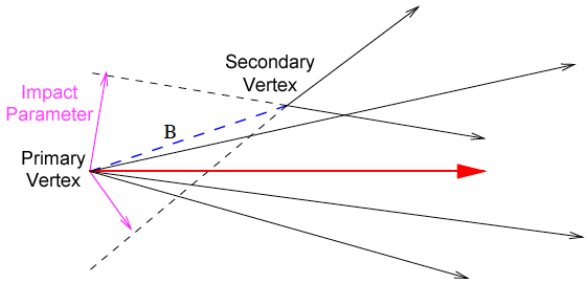


Figure 3: Impact Parameter due to a Secondary Vertex [12]

2.3 Variable Selection

In selecting the ideal kinematic variables, it is important to avoid any variable that introduces a selection bias, i.e. selecting more signal events than background or vice versa. This means that none of the b-tagging criteria are used. By definition, all signal events have b-jets in them, whereas some of the background events have mislabeled jets resulting from lighter quarks. Using b-tagging criteria would then introduce a bias whereby background and signal events would be treated differently.

The result of this study will examine the combined mass of the two leading jets, referred to as dijet mass. In order to avoid selection biasing, this variable and all variables with a correlation of 50% or higher to it were excluded. This is the same as requiring that the final dijet mass distribution is independent of the selection process so that it remains a representative sample of the entire distribution. The selected variables were determined by running TMVA with differing settings many times, until optimal separation was reached.

2.4 Boosted Decision Trees

The classifier used in this study is referred to as a Boosted Decision Tree (BDT). This method utilizes a binary tree structure, as shown in Figure 4. The classifier starts with all events stored in a root node. It determines the variable that will provide the best separation between background and signal, and scans this variable for the optimal cut value. Once the cut is found, the process of selecting the optimal variable and value is repeated on all subsequent daughter nodes until a minimum number of events in each node is reached. The final nodes are referred to as the *leaf nodes*. They are classified as background or signal depending on the majority of events in the node. The testing sample is then run through the same decision tree so that its classification accuracy can be determined.

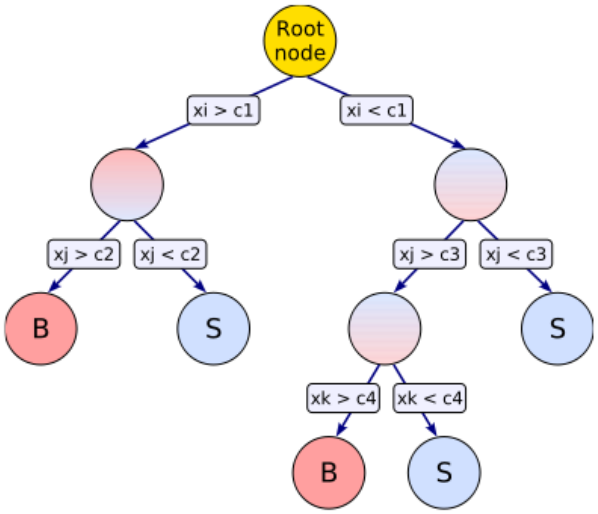


Figure 4: Decision tree

This classification technique has two major sources of error. The first is *overtraining*, which is when the classifier becomes too specific to the training sample, and hence cannot be generalised to a statically independent testing sample. A process called *pruning* is used on the classifier to alleviate this effect. The second is the classifier's instability when dealing with fluctuating data. In order to stabilize the classifier, a process called *boosting* is used, which is analogous to training

many classifiers and averaging their result [10].

2.4.1 Cutting

The goal of MVA is to end up with leaf nodes that are either pure signal or pure background. The purity of a node is defined as the fraction of signal events to total events in the sample, i.e. $p = S/(S + B)$. TMVA then defines a *separation criteria*. This is a function of the purity which helps determine the optimal cut at a given node. This criteria must be maximized for a purely mixed sample, i.e. at $p = 0.5$, and it must fall off to zero for a sample of only one event type. This function must also value pure background as much as pure signal, hence it must be symmetric around its maximum. In this study, the cross entropy was used, as defined below.

$$\text{Cross Entropy}: \quad s(p) = -p \cdot \ln(p) - (1-p) \cdot \ln(1-p)$$

TMVA looks for the variable and cut value that will maximize the difference between the separation index of a parent node, and the sum of the indices of the two daughter nodes. The indices of the daughter nodes are weighted by their relative fraction of events. This study uses the maximum granularity possible when scanning each variable for the best cut value.

The cutting process is repeated until the number of events in a node reaches a minimum value, defined as 69 for this study, or until the separation index is zero. To avoid overtraining, a process called *pruning* is employed once the tree has been grown to its full size [10].

2.4.2 Pruning

The goal of the training phase in TMVA is to build a classifier whose results will be valid for a statistically independent sample. This implies a balance between how well a classifier performs on the training sample and how well it can be generalised to the testing sample. A decision tree that is grown to perfectly separate out all signal events in the training sample will only work well *for that sample.* Such a tree is said to be overtrained. In order to ensure meaningful results from the testing sample, a process called *pruning* is employed.

As the name suggests, pruning is the process of "cutting back" a fully grown decision tree, which means removing leaf nodes that are too specific to the training sample to be applied in general. The tree is first allowed to grow to its full size instead of being interrupted. This is because apparently insignificant cuts can lead to useful cuts down the tree.

TMVA offers two pruning algorithms: *expected error* and *cost complexity.* The method used in this study was expected error,[3] which estimates the statistical error of a parent node and daughter nodes using the binomial error, shown below, where N is the number of nodes. Starting at the leaf level, daughter nodes whose combined error is greater than that of the parent node are recursively deleted. In order to control the strength of the pruning, the daughter nodes can be scaled by a constant factor, `PruneStrength`, denoted C_{pr} below [10].[4]

$$\text{err} = C_{\text{pr}} \cdot \sqrt{\frac{p \cdot (1-p)}{N}}$$

It is worth mentioning that pruning is applied *after* boosting, which is described in the next section.

2.4.3 Boosting

In order to stabilize decision trees, a process called *boosting* is utilized. This can loosely be considered an averaging process which incorporates many different decision trees into the final

[3] for more on cost complexity, see [10]

[4] For the parent node, $C_{\text{pr}} = 1$.

BDT output. This is especially useful for dealing with variables that have both a high separation power along and a high variance in the training sample. In this study, a method called *bagging* was used.[5]

The classifier randomly selects events from the parent sample, allowing the same event to be selected multiple times. It then grows a decision tree based on this resample. This process is repeated many times until a "forrest" of decision trees is grown. Each individual tree will assign each event as signal or background, hence for event $\mathbf{x}$, we define the result of a classifier as $h(\mathbf{x}) = +1$ or -1, respectively. The results are then averaged, as explicitly noted below.

$$y_{\mathrm{Ba}}(\mathbf{x}) = \frac{1}{N_{\mathrm{trees}}} \sum_{i}^{N_{\mathrm{trees}}} h_i(\mathbf{x})$$

Since events can be selected more than once, each resample can be thought of as a representation of the probability density distribution of the parent sample. This parent sample is unaffected throughout the process, which means that each resample will have the same parent distribution, albeit statistically fluctuated. These fluctuations are the key to all boosting methods since they help stabilize the resulting output. This output will fall in $[-1, 1]$ depending on if the event was classified as more signal-like or background-like [10].

2.5 Event Samples

In order to train and test the BDT classifier, four event sample sets were used. They are listed below, along with a brief description.

- `mva_SUSY_combination` : This is the sample contains signal events from a simulation of the medium mass Higgs-like boson scenario. The three masses used were 200 GeV, 300 GeV, and 350 GeV. These masses were all taken to have the same cross section.
- `mva_HTQCD` : This is the sample contains simulated background events.
- `mva_data` : This is data collected by the CMS experiment in 2012 of integrated luminosity of $4\mathrm{fb}^{-1}$.

3 Results

The multivariate analysis returns a viable control region that retains the majority of background events, i.e. high background efficiency, while cutting most of the signal events. The final MVA output was cut along three values, denoted *loose*, *tight*, and *very tight*. After comparing the three, the tight cut was selected as the ideal balance of background efficiency and background to signal ratio.[6]

3.1 Variable Distributions

The kinematic variables used in the MVA show distributions with clear discrimination between signal and background events. Several of the chosen variables proved to be extremely effective, namely the minor thrust and the two jet areas. While background simulations show a high variance, they are in good agreement data points overall. Since the number of background events will heavily outweigh the signal events, this agreement is characteristic of accurate simulations.

The following distributions have been scaled for ease of comparison. The signal has been scaled by a factor of 100, while the background has been scaled by a factor of 2.5.

[5] Bagging is not a "boost" method in the strict sense of the definition, but returns the same result. See [10] for more details

[6] Please note the following is unpublished Data and is still a work in progress

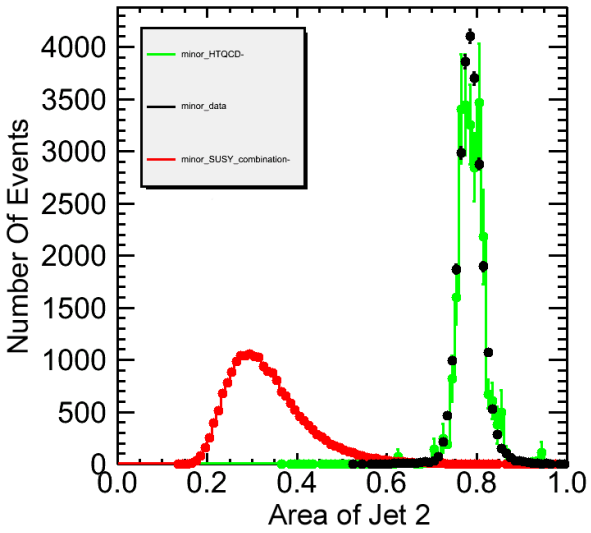


Figure 5: `ptd2` for scaled events

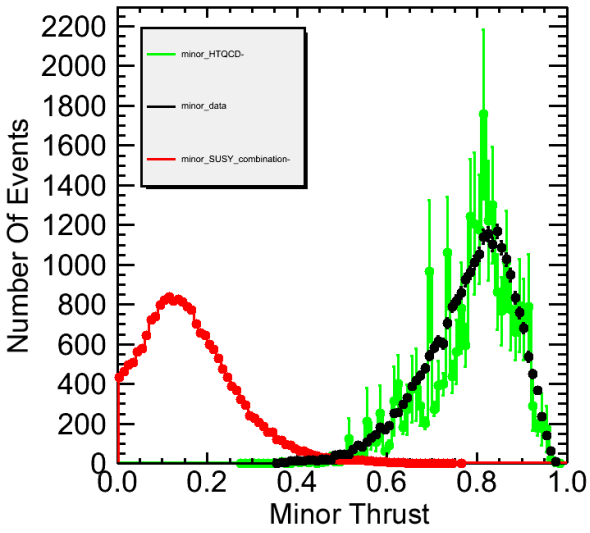


Figure 6: `minor` for scaled events

3.2 Cuts

Once the multivariate analysis was complete, three cuts were made along the MVA output at different tolerances. These tolerances result in varying background, signal, and data selection efficiencies. The tighter cuts resulted in a lower background selection efficiency and a higher background to signal ratio, as expected.

Loose: $y_{\text{Ba}} < 0.45$

Tight: $y_{\text{Ba}} < 0.25$

Very Tight: $y_{\text{Ba}} < 0.00$

Figure 7: Multivariate Analysis Results

Cut	S_{eff}	B_{eff}	D_{eff}	B/S	D/B
L	0.18	0.72	0.65	4.03	0.90
T	0.08	0.60	0.52	7.58	0.87
VT	0.02	0.28	0.24	13.88	0.86

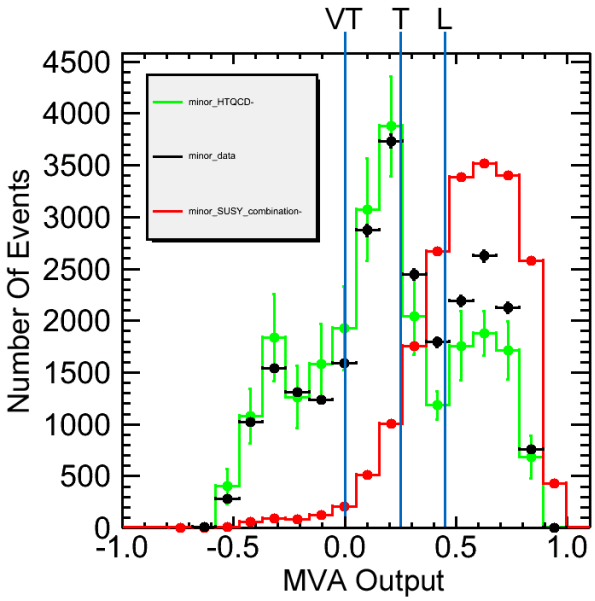


Figure 8: Multivariate Analysis Result

3.3 Signal Supression

After the cuts were made in the MVA output, the resulting dijet mass distributions were compared. All three of the cuts showed significant signal suppression from the uncut set of events, while preserving most of the background set. Shown below is is the tight cut compared to the original sample. This cut was selected for its optimal balance of background efficiency and signal to background ratio, as discussed in section 4.1.

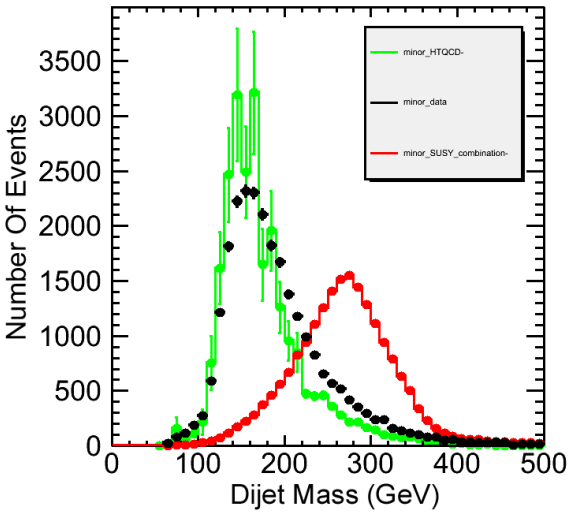


Figure 9: Distribution before cut

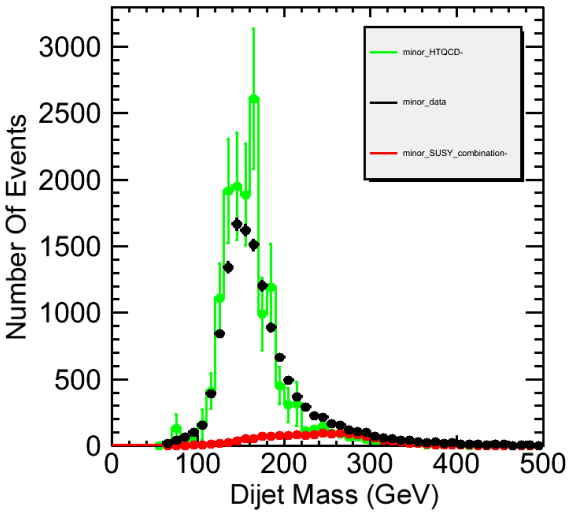


Figure 10: Distribution at "tight" cut

4 Conclusion

Using Monte Carlo generated events along with multivariate analysis techniques, a viable control region for the MSSM Higgs search at the CMS collaboration has been determined. This analysis was done across nine input kinematic variables, each of which displayed moderate to strong separation power. Once the classifier was trained and tested, three cuts were made on the MVA output at different tolerances. These cuts were compared on the basis of background efficiency and background to signal ratio. Ultimately, a middle value of $y_{\mathrm{Ba}} < 0.25$ was chosen because of its ideal balance of these two parameters.

4.1 Ideal Cut Value

In selecting the optimal cut value, two main consideration were made. In order for the final region to be an ideal control region, the background to signal ratio must be maximized, which suggests making a very tight cut. Making such a cut would necessitate that all remaining events be classified similarly by the MVA. This means they would have similar kinematic properties. While none of these properties are strongly correlated to the dijet mass, each has a residual correlation. These residual correlations are magnified when all the remaining events are kinematically similar, resulting in a biased sample. Larger event samples tend to smear out these correlations, reducing their effect. This effect suggests making a looser cut. For this purpose, the MVA parameter was cut at $y_{\mathrm{Ba}} < 0.25$.

This resulted in a background selection efficiency of 60% and a background to signal ratio of 7.6. The signal events were cut to less than 10% of the original sample. The dijet mass distribution for this cut shows similarly shaped background and data samples before and after the cut. This implies that there was a low selection bias, as the form of the distribution was unchanged. The data to background ratio for this cut is 0.87. Since the vast majority of data will be background events, it is expected that this value be close to one.

4.2 Further Studies

With the control region determined, this study can continue on to search for MSSM Higgs signatures in the dijet mass distribution of two bottom quarks. The analysis code for this ongoing work is CMS-AN-13-229.

References

[1] CMS Collaboration. Combined results of searches for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV. *Phys.Lett.*, B710:26–48, 2012.

[2] David Barney. Cms presentations for public. Website, October 2011. `http://cms.web.cern.ch/org/cms-presentations-public`.

[3] Lucas Taylor. Detector overview. Website, November 2011. `http://cms.web.cern.ch/news/detector-overview`.

[4] V Karimäki. *The CMS tracker system project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.

[5] J. Beringer. Review of particle physics. *Phys. Rev. D*, 86:010001, Jul 2012.

[6] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The Anti-k(t) jet clustering algorithm. *JHEP*, 0804:063, 2008.

[7] Ian Aitchison. *Supersymmetry in Particle Physics: An Elementary Introduction*. Cambridge University Press, first edition, 2007.

[8] Stephen P. Martin. A Supersymmetry primer. 1997.

[9] Joerg Behr. Search for higgs production in association with b quarks with cms in pp collisions.

[10] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[11] Hasko Stenzel. Definition of variables, 31/8/2013.

[12] b-jet identification in the cms experiment. Technical Report CMS-PAS-BTV-11-004, CERN, Geneva, 2012.