Studies of Energy Regression to b-jets in SUSY Higgs Searches at the LHC.

Sergio Calvente, Universidad Autonoma de Madrid, Spain.

Supervisors: Gregor Hellwig, Rainer Mankel and Roberval Walsh.

DESY Summer School

September 4, 2013

Abstract

Multivariate regression analysis have been applied to correct the b-jets energy in Higgs searches in the context of the modified P4 scenario of the next-to-minimal supersymmetric extension of the standard model (NMSSM). The main result is an improvement of 11% in the mass resolution of the lightest Higgs (H_1) mass peak. For the other peaks these improvements are lower and the position of the mass peaks are over-corrected, this effect needs further investigation.

Contents

1 Introduction.

In this report I will explain my project during the DESY Summer School 2013. I will start with an introduction to the theory and motivation of the project. Then I will state the procedure to do the task. I will present and discuss the obtained results and I will finish with a summary and outlook.

A Higgs boson has been discovered at the LHC but its nature still is unclear. Several models that extends the SM predicts additional Higgs bosons and in some scenarios a Higgs boson could even be lighter than the Z boson and the decay is largely into b quarks. For instance, in the NMSSM (next-to-minimal supersymetric extension of the standard model) such light Higgs boson decaying into b quarks could be observed at the LHC in SUSY cascades.

The b-jet energy needs specific corrections because the b hadrons decay largely semileptonically, therefore, emitting a neutrino that is not detected and carries away part of the jet energy. In the context of these corrections the multivariate analysis regression is applied.

2 NMSSM Higgs Searches.

2.1 Motivation for NMSSM.

Supersymmetric models (SUSY) allow to solve some problems of the SM, such as the Hierarchy problem [1–4], why the Higgs mass is so much lower than M_{GUT}[1], as one would expect when large loop corrections are take into account up to high energy scales. Unless there is a miraculous cancellation between the individual corrections. This is solved by adding a symmetry between fermions and bosons (each SM particle gains a superpartner). Then the superparticle contributes with the opposite sign to the loop corrections and they cancel out to avoid this fine-tuning. SUSY also offers dark matter candidates; as the LSP (lightest supersymmetric particle) in models with R-parity conserved, this particle is stable, heavy (cold dark matter) and weakly interacting.

The MSSM, minimal supersymmetric extension of the SM, is the simplest SUSY model that solves these problems. Nevertheless, this model still suffers from some problems, such as the μ problem [5]. This μ coupling factor has mass dimension, and its value is expected to be zero or at the order of GUT scale. However, the electroweak symmetry breaking (EWSB) mechanism leads to a μ in the electroweak scale ($\sim M_W, M_Z$), in order to get a phenomenologically acceptable theory.

[1] Grand Unification Theories (GUT), M_{GUT} is the energy scale at which the three fundamental forces are expected to converge ($\sim 10^{16}$ GeV).

This problem can be solved using the NMSSM, next-to-minimal supersymmetric extension of the SM [6, 7]. This model solves the μ problem by introducing a new singlet superfield. The singlet enters in the Lagrangian with a dimensionless λ term, which dynamically generates an effective coupling μ_{eff}. This new singlet superfield produces additional physical particles with respect to the MSSM, one extra neutralino (mostly singlino), and two extra Higgs bosons (one CP-even and one CP-odd).

2.2 Higgs Sector.

In the NMSSM there are seven physical Higgs bosons. Some scenarios predict Higgs bosons that can be lighter than the Z bosons but evaded detection at LEP due to the suppressed interaction with the gauge bosons.

In this project one of these light Higgs scenarios, the modified P4 benchmark scenario, will be used. In this scenario the seven Higgses are three CP-even: H_1 (lighter than the Z boson), H_2 (SM-like $\sim$ 125 GeV) and H_3 (much heavier); two CP-odd: A_1 (a bit heavier than H_2) and A_2; and two charged Higgses: $H^{\pm}$ (mass $\sim M_{H_3}, M_{A_2}$). In this report only the lightest neutral Higgses (H_1, H_2 and A_1) will be analysed.

2.3 Light Higgs Searches.

The lightest Higgs (H_1) has strongly suppressed interactions with the gauge bosons, due to its large singlet component. Therefore it is very hard to detect and produce in the standard channels at the LHC. For that reason it is necessary to look at SUSY cascade decays [9], as the one shown in Figure 1.

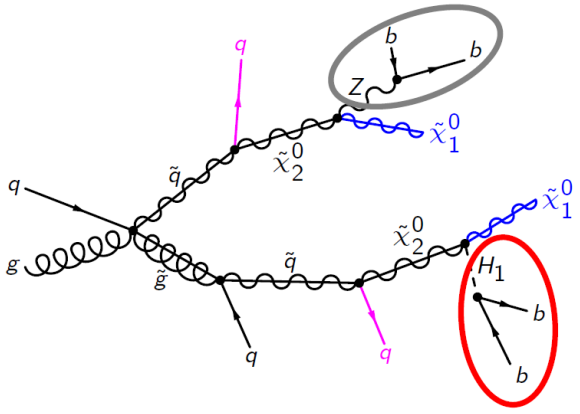


Figure 1: *SUSY cascade decays.*

The gluons and quarks interact in the LHC producing squark and gluinos (their supersymmetric partners). These supersymmetric particles produce a cascade decay in which at some point heavier neutralinos (χ_i^0), or charginos ($\chi_i^\pm$), decay in a lighter one and produce one of the lightest Higgs (or Z boson), see equations (1) and (2) below. All these cascades end in the lightest neutralino that is the LSP and therefore escapes from the detector, so there is high missing transverse energy in these events.

$$\chi_i^0 \to \chi_j^0 H_k \qquad \chi_i^0 \to \chi_j^0 A_k \tag{1}$$

$$\chi_i^\pm \to \chi_j^\pm H_k \qquad \chi_i^\pm \to \chi_j^\pm A_k \tag{2}$$

The lightest Higgs (H_1) is produced in these cascades and decays mainly ($\sim$90%) in a pair of b-jets. As it has lower mass, it carries a higher momentum and therefore the b-jets are more boosted and closer to each other than the b-jets from other bosons in the cascade. These b-jets have a harder momentum distribution than the b-jets from $t\bar{t}$ (main background), which allows to separate the signal from the background.

The strategy to search for this lightest Higgs is to reconstruct the invariant mass of the two closest b-jets of each event and look for resonances in the spectrum. Some cuts are applied to reduce the background. The success in this search depends on b-jet identification efficiency and the di-b-jet mass resolution.

3 Multivariate-analysis Regression.

The measured energy and momentum of the jets need to be corrected to account for detector effects. The b-jets need further corrections than the light jets. The reason is that a third of the B hadrons decay leptonically and semileptonically, and due to the lepton number conservation many neutrinos are produced in these jets. Neutrinos leave the detector without interacting, so measurements of energy and momentum are underestimated. We can only try to account for this missing transverse energy.

This further correction can be performed by using regression multivariate-analysis (MVA) techniques. This section explains how this works and which methods can be used.

3.1 Regression.

A regression analysis estimates the form of a function, which predicts the value of a response variable (target), in terms of the values of given known variables. A multivariate regression technique is a "supervised learning" algorithm which makes use of training events, whose output is known, to determinate an approximation of the functional behaviour of the target variable [10].

For correction of the b-jets only reconstructed variables can be used as input, because they are the only available variables in the real data. Using Monte Carlo simulations a MVA function can be trained to estimate the true value of a variable, such as transverse momentum, from the values of some reconstructed variables. In order to get a good estimation these variables must be as strongly correlated as possible with the target variable. If the ratio between generated and reconstructed transverse momentum is used as target, the output of the regression is already the correction factor for the b-jets.

3.2 Methods.

To perform the multivariate regression analysis, the software package Tool kit for Multivariate Analysis (TMVA) [10] in the ROOT framework has been used. In particular two different techniques have been tried. They will be explained below.

3.2.1 Boosted Decision Trees (BDT).

A decision tree is a binary tree structured regressor where repeated yes/no decisions are taken on one variable at a time until a stop criterion is fulfilled. The node splitting is performed on the variable that gives the maximum decrease in the average square error when a constant value is assumed for the target as output of the node.

The boosting of a decision tree extends the concept from one to many trees to form a forest. The trees are produced with the same training sample, and are combined in a single regressor as a weighted average of all the regressor trees. The weights are calculated in relation with a loss function that accounts for the deviation of the output from the true value. This is performed with a Gradient Boost, that minimize this loss function. The boosting is combined with bagging, using of a fraction of the sample (randomly selected) in each tree. The boosting and bagging lead to a greater stability with respect to statistical fluctuations in the training sample and enhance the performance of a single tree.

3.2.2 Neural Network.

An Artificial Neural Network (ANN) is a simulated set of interconnected neurons, where each neuron produces a response at a given input signal. The TMVA used method is the Multi-Layer Perceptron (MLP), where neurons are organised in layers and only interact with the previous and the following layers. The first layer is for input variables, the last layer has one neuron per target (this allows for more than one target), and there are intermediate hidden layers. Each neuronal interaction has an associated weight which is used to scale the input signals to a neuron.

The very preliminary results has shown that the performance of this method is worse and efforts were concentrated in the studies on the BDT method.

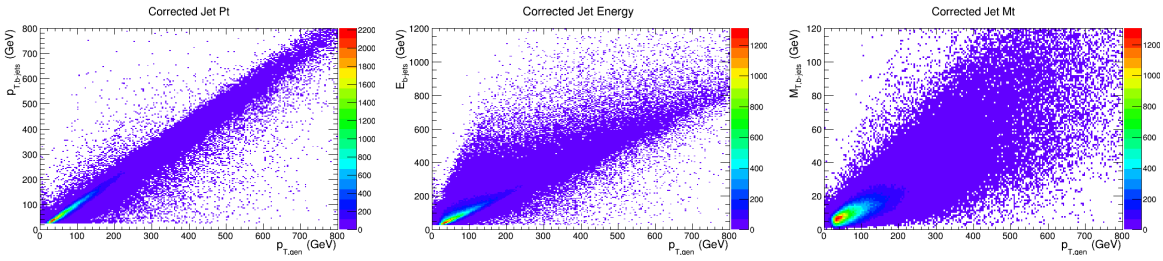
4 Correlations.

The variables used for training the regression must be correlated with the target, as it was stated before. The variables were chosen following the line of the CDF regression [11]. There are 15 variables in three groups that come from the calorimeter, the tracking and the vertexing. The performance with these 15 variables was better than only with 7 of them before. Further studies concerning the optimal and minimal set of variables need to be performed.

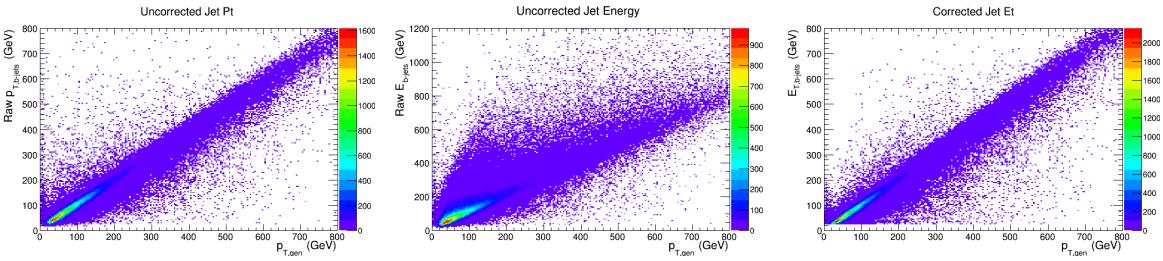
The variables from the calorimeter: corrected transverse momentum (p_T), corrected energy, transverse energy (E_T), transverse mass (M_T), Jet Energy Correction (JEC) uncertainty, number of constituents of the jet, and the uncorrected transverse momentum and energy of the jets. The variables related with tracking are the maximum transverse momentum of a track, and the sum of the transverse momenta of all the tracks. The variables related with the leptons, which could be more correlated with the missing energy carried away by neutrinos, could not be added to the analysis due to time constraints.

The variables related with vertexing are: the mass, transverse momentum and number of tracks, of the secondary vertex (SV), and three dimensional (3D) distance, and error of the distance, between the SV and primary vertex. These variables can be useful for the regression because they are related with the b-tagging and could be correlated with the neutrinos. Since the missing transverse energy is mostly carried away by the lightest neutralino, as it was pointed out in the section 2.3, this variable cannot be used for the training.

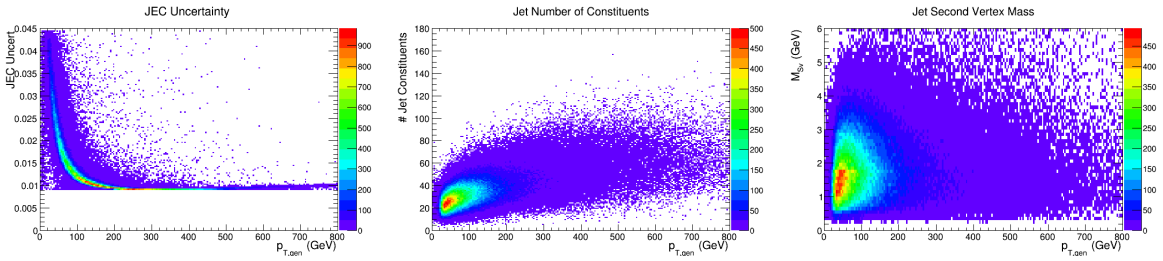
The correlations between these variables and the target can be seen in the Figures 2 and **??** below. There are examples of linear correlation, in Figure 2(a), functional correlation in Figure 2(g), or greater correlation at low p_T, as in Figure 2(j).



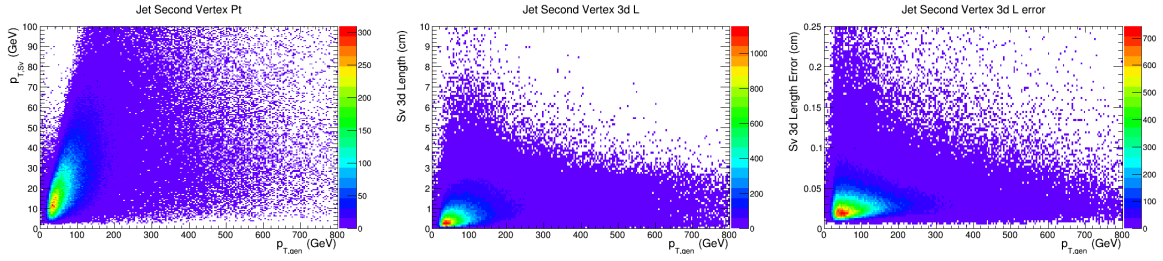
(a) Corrected Jet p_T (b) Corrected Jet Energy (c) Jet M_T



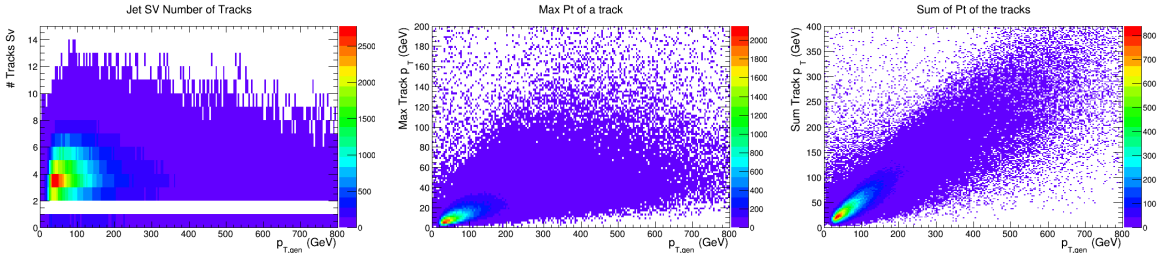
(d) Uncorrected Jet p_T (e) Uncorrected Jet Energy (f) Jet E_T



(g) JEC Uncertainty (h) Jet Number of Constituents (i) SV Mass



(j) SV p_T (k) SV 3D Length (l) SV 3D Length error



(m) SV Number Tracks (n) Maximum p_T Track (o) Sum p_T Tracks

Figure 2: *Graphs showing the correlation between the input variables of the training and the generated p_T.*

5 Regression Results.

In this section the details of the analysis and the results will be exposed, explained and discussed.

5.1 Training and Application.

For training the regression, only the b-jets, with a loose b-tagging, coming from H_1, H_2 and A_1 are used. If all the b-jets are used the regression does not improve the resolution neither the mass scale. In a first attempt only the b-jets from H_1 were used, but this leads to a good improvement of the resolution only in this mass peak, and shifts to lower masses all the other mass peaks. The use of b-jets from the 3 Higgs reduces this effect, but not completely what is under further investigation. The b-jets from the Z boson, available in the sample, are not used in order to use Z boson b-jets data to validate the method later.

There are two possible scenarios, one with $m_{H_1} < \frac{1}{2} m_{H_2}$ where H_2 decays predominantly to two H_1, so the H_2 is difficult to see and the H_1 is enhanced, but with a bit more difficult topology. This scenario has been analysed, but as it is more complicated will not be showed and discussed in this report.

In the other scenario, the H_2 cannot decay in H_1, it is kinematically forbidden. There are 5 different H_1 mass (and therefore different A_1 mass) simulation samples available: 65, 70, 75, 80 and 85 GeV. The first attempt was training only with 65 GeV as benchmark, but later all the masses samples where used for training to avoid the introduction of a bias in the mass of the H_1 that will be searched in data.

The cuts applied to the training sample were looser than the analysis cuts, again to avoid to introduce bias in the method. The available sample was divided in two parts, one part was used for training and test, and the other part for application of the created regression function. In the application, a benchmark point of H_1 mass 65 GeV was chosen. All the analysis cuts were used, and the two closest b-jets (tight b-tagging), using the angular separation, were selected to calculate the invariant mass of these pairs. The results of this application are shown below.

5.2 p_T Regression.

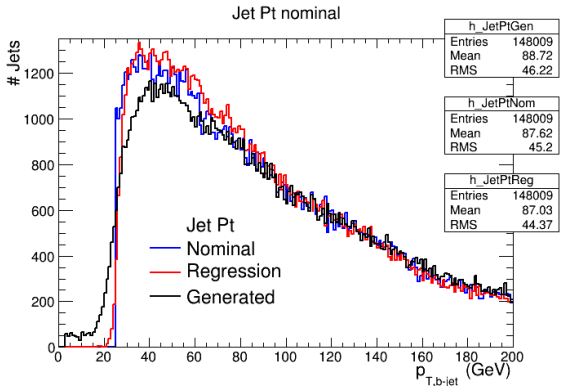


Figure 3: *Generated (black), nominal (blue) and regression (red) p_T distributions.*

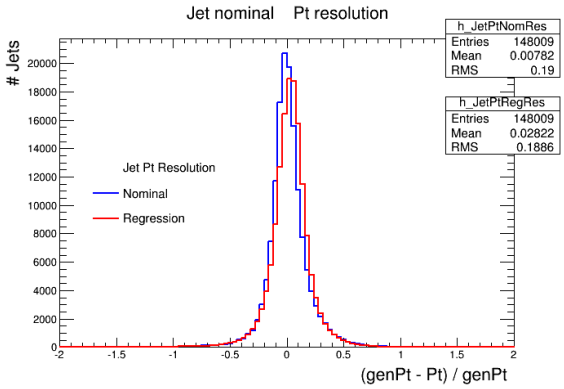


Figure 4: *Nominal (blue) and regression (red) p_T resolutions.*

In Figure 3 p_T distribution are shown for generated, nominal and regression corrected b-jets. It is difficult to appreciate the effect of the regression, because the non-resonant background b-jets are unaffected by the regression and are included in these distributions.

Figure 4 shows the relation of generated minus reconstructed over generated p_T for both nominal and regression b-jets. After regression it shifts to the right, with is consistent with the p_T distribution going to lower values, and also the mass distribution as it will be shown now. The RMS of this quantity is reduced which is consistent with the resolution improvement of the regression.

5.3 Invariant Mass

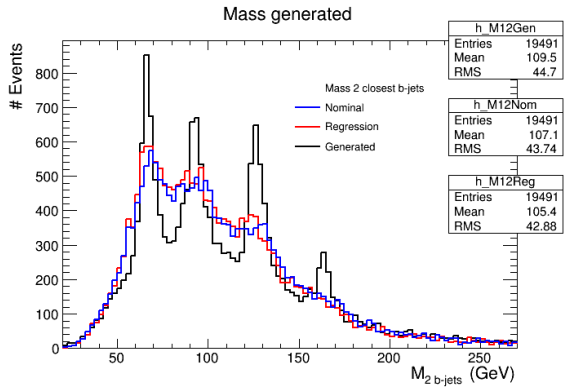


Figure 5: *Generated (black), nominal (blue) and regression (red) invariant mass of pairs of closest b-jets. The generated distribution shows , from left to right, the H_1, Z, H_2 and A_1 resonant peaks.*

In Figure 5 the invariant mass of the pairs of closest b-jets are shown for the generated, nominal and regression b-jets. It is noticeable that the regression distribution is shifted to lower masses, consistently with the p_T shifted to lower values. To get a better view of the effect of the regression in the resonant peaks, the b-jets coming from each Higgs or Z boson will be treated separately.

5.3.1 H_1 Peak.

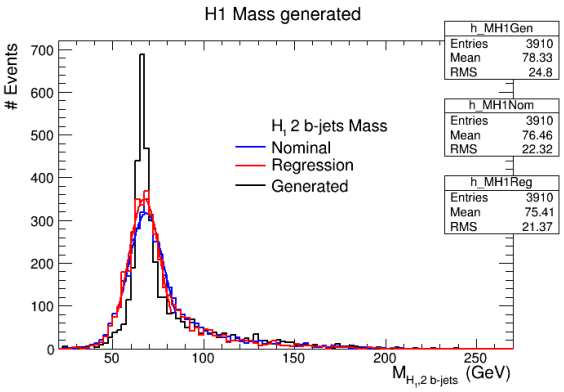


Figure 6: *Generated (black), nominal (blue) and regression (red) distributions for the invariant mass of two b-jets from H_1.*

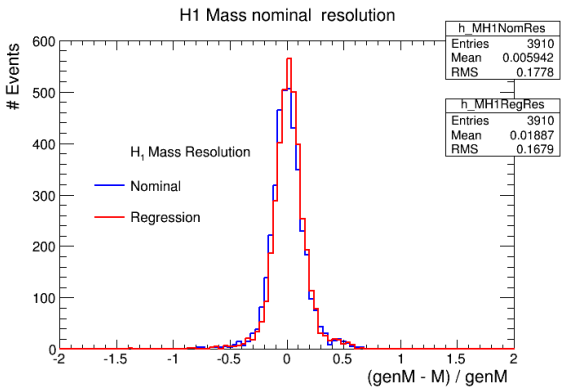


Figure 7: *Nominal (blue) and regression (red) resolutions of the invariant mass of two b-jets from H_1.*

In Figure 6 we can see the invariant mass of the pairs of b-jets coming just from H_1. A Gaussian fit has been performed to get the peak and width values, shown in Table 1, to avoid the effect of the long tails. From the fit values displayed in Table 1, it is appreciable that an improvement of 11% in the mass resolution is obtained after regression. In addition, the mass scale gets closer to the generated value.

In Figure 7 the generated minus reconstructed over generated mass is shown before and after regression. The regression shifts this to the right, consistently with the mass shifted to lower values.

Table 1: *Fit parameters of the H_1 mass peak.*

Parameter	Nominal	Regression	Generated
Peak (GeV)	68.6	67.5	66.9
Width (GeV)	9.9	8.8	4.3

5.3.2 Z Peak.

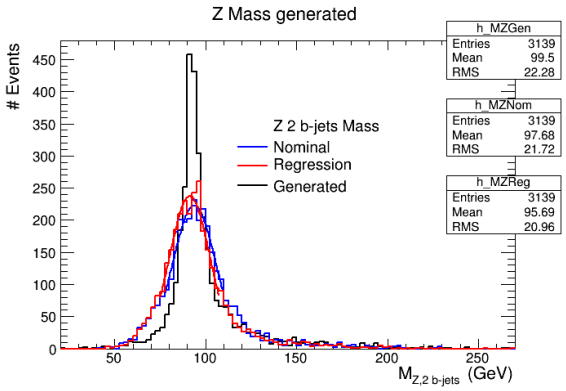


Figure 8: *Generated (black), nominal (blue) and regression (red) distributions for the invariant mass of two b-jets from Z.*

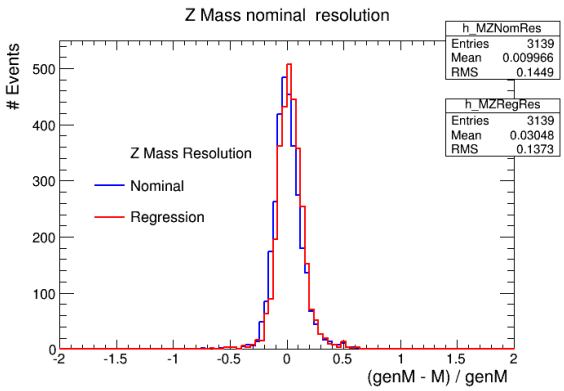


Figure 9: *Nominal (blue) and regression (red) resolutions of the invariant mass of two b-jets from Z.*

In Figure 8 we can see the invariant mass of the pairs of b-jets coming just from Z. A Gaussian fit has been performed to get the peak and width values, shown in Table 2, to avoid the effect of the long tails. From the fit values displayed in Table 2, the improvement obtained after regression in the mass resolution now is smaller ($\sim 6\%$). The mass scale gets lower than the generated one, the mass is over-corrected. This effect is under investigation.

In Figure 9 the generated minus reconstructed over generated mass is shown before and after regression. The regression shifts this to the right, consistently with the mass shifted to lower values.

Table 2: *Fit parameters of the Z mass peak.*

Parameter	Nominal	Regression	Generated
Peak (GeV)	93.9	91.4	93.1
Width (GeV)	11.7	11.0	6.0

5.3.3 H_2 Peak.

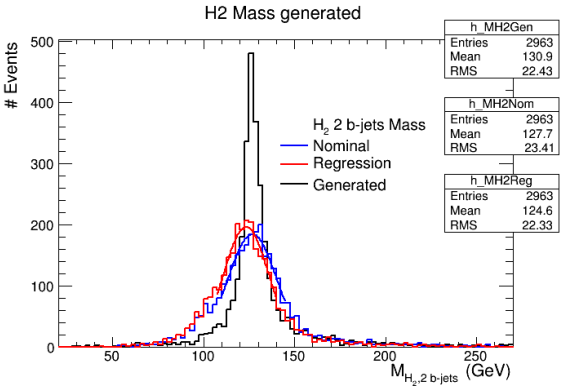


Figure 10: *Generated (black), nominal (blue) and regression (red) distributions for the invariant mass of two b-jets from H_2.*

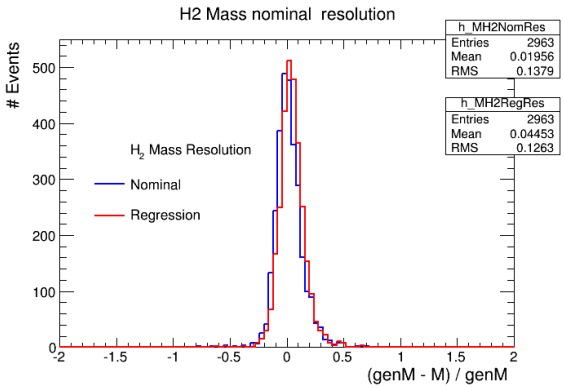


Figure 11: *Nominal (blue) and regression (red) resolutions of the invariant mass of two b-jets from H_2.*

In Figure 10 we can see the invariant mass of the pairs of b-jets coming just from H_2. A Gaussian fit has been performed to get the peak and width values, shown in Table 3, to avoid the effect of the long tails. From the fit values displayed in Table 3, the improvement obtained after regression in the mass resolution now is smaller ($\sim 7\%$). The mass scale gets lower than the generated one, the mass is over-corrected. This effect is greater than for the Z boson.

In Figure 11 the generated minus reconstructed over generated mass is shown before and after regression. The regression shifts this to the right, consistently with the mass shifted to lower values.

Table 3: *Fit parameters of the H_2 mass peak.*

Parameter	Nominal	Regression	Generated
Peak (GeV)	127.0	123.6	127.0
Width (GeV)	13.3	12.4	4.8

5.3.4 A_1 Peak.

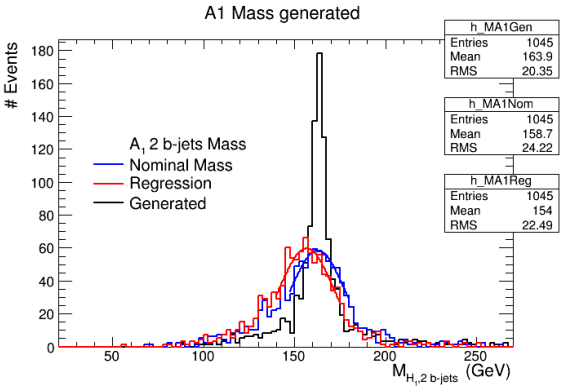


Figure 12: *Generated (black), nominal (blue) and regression (red) distributions for the invariant mass of two b-jets from A_1.*

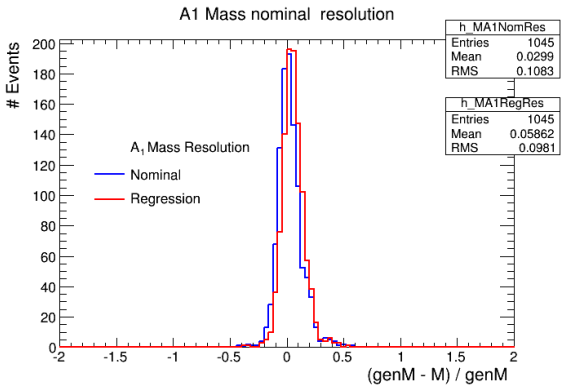


Figure 13: *Nominal (blue) and regression (red) resolutions of the invariant mass of two b-jets from A_1.*

In Figure 12 we can see the invariant mass of the pairs of b-jets coming just from A_1. A Gaussian fit has been performed to get the peak and width values, shown in Table 4, to avoid the effect of the long tails. From the fit values displayed in Table 4, the improvement now is not reliable due to lower statistics in this peak and worse width calculation. The mass scale gets lower than the generated one, the mass is more over-corrected than for the other peaks. This effect is under investigation.

In Figure 13 the generated minus reconstructed over generated mass is shown before and after regression. The regression shifts this to the right, consistently with the mass shifted to lower values.

Table 4: *Fit parameters of the A_1 mass peak.*

Parameter	Nominal	Regression	Generated
Peak (GeV)	162.8	156.8	163.4
Width (GeV)	15	15	5.4

5.3.5 Non-resonant Background.

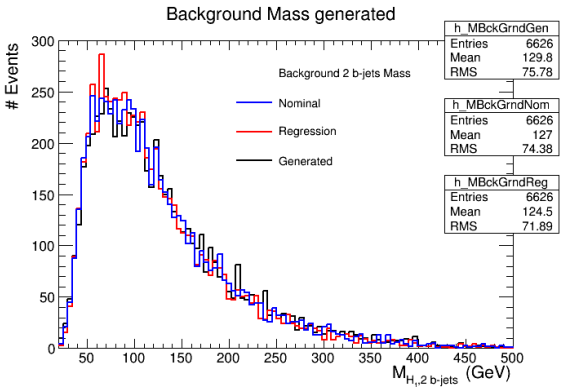


Figure 14: *Generated (black), nominal (blue) and regression (red) distributions for the invariant mass of two b-jets from non-resonant background.*

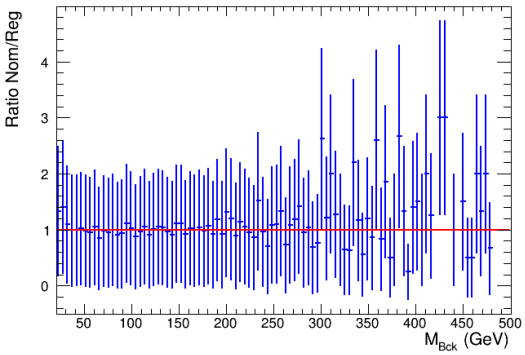


Figure 15: *Ratio between nominal and regression (blue) invariant mass distributions of two b-jets from non-resonant background, with fit to a constant (red).*

In Figure 14 we can see the invariant mass of the pairs of b-jets from the non-resonant background. To see the effect of the regression in these b-jets, Figure 15 shows the ratio between nominal and regression mass distributions. This ratio fits to a constant value of 1.0±0.1, so the regression does not affect the background distribution. The errors bars for this ratio are so large due to the small number of events and therefore large statistical fluctuations.

6 Conclusions.

To conclude, the NMSSM allows light Higgs scenarios which could remain undetected in LEP and Tevatron. The modified P4 scenario has been analysed and a way to detect the lightest Higgs has been proposed. This search depends on the b-jet correction and mass resolution.

The use of MVA regression has been studied to perform this correction. The BDT method performed better than MLP. With the lightest Higgs (H_1) the mass resolution is improved by a 11%, which is a success. For the WH searches in SM at CDF experiment [11], the regression got a 15% of improvement in the resolution, with translated to 20% in the final result, but that is an easier environment and leptonic variables were used. The use of leptonic variables should improve the performance of the regression also in the NMSSM search.

The problem of the shifted peaks to lower masses has been investigated. One possible cause is the dominant number of H_1 b-jets over the sum of H_2 and A_1 b-jets. The H_1 b-jets need to be corrected to lower masses. The reason for this

correction in the H_1 could be that the high cut in the p_T reduces the low p_T events and pushes the reconstructed H_1 mass to higher values. This is consistent with the fact that if this cut is reduced for training, the shifting effect is reduced too.

However, regression performed just with A_1 for the training, led to the same shifting in the masses what discourages this hypothesis. This must be further investigated.

References

[1] Steven Weinberg. Implications of dynamical symmetry breaking. *Phys.Rev.*, D13(4):974–996, Feb 1976.

[2] S. Weinberg. Implications of dynamical symmetry breaking: An addendum. *Phys.Rev.*, D19(4):1277–1280, Feb 1979.

[3] Eldad Gildener. Gauge-symmetry hierarchies. *Phys.Rev.*, D14(6):1667–1672, Sep 1976.

[4] Leonard Susskind. Dynamics of spontaneous symmetry breaking in the weinberg-salam theory. *Phys.Rev.*, D20(10):2619–2625, Nov 1979.

[5] Jihn E. Kim and Hans Peter Nilles. The mu Problem and the Strong CP Problem. *Phys.Lett.*, B138:150, 1984.

[6] Ulrich Ellwanger, Cyril Hugonie, and Ana M. Teixeira. The Next-to-Minimal Supersymmetric Standard Model. *Phys.Rept.*, 496:1–77, 2010.

[7] M. Maniatis. The Next-to-Minimal Supersymmetric extension of the Standard Model reviewed. *Int.J.Mod.Phys.*, A25:3505–3602, 2010.

[8] A. Djouadi, M. Drees, U. Ellwanger, R. Godbole, C. Hugonie, et al. Benchmark scenarios for the NMSSM. *JHEP*, 0807:002, 2008.

[9] Oscar Stal and Georg Weiglein. Light NMSSM Higgs bosons in SUSY cascade decays at the LHC. *JHEP*, 1201:071, 2012.

[10] P. Speckmayer, A. Hocker, J. Stelzer, and H. Voss. The toolkit for multivariate data analysis, TMVA 4. *J.Phys.Conf.Ser.*, 219:032057, 2010.

[11] T. Aaltonen, A. Buzatu, B. Kilminster, Y. Nagai, and W. Yao. Improved b-jet Energy Correction for $H \to b\bar{b}$ Searches at CDF. 2011.