Optimisation of MVA Approach for $\tilde{\tau}$ Search

Summerstudent: Natalia Lenshina
Lomonosov Moscow State University

Supervisors: Ilya Bobovnikov, Isabell Melzer-Pellmann
Deutsches Elektronen-Synchrotron, Hamburg

September 6, 2017

Abstract

In this work, we present an optimisation workflow of Multivariable Analysis, specifically Boosted Decision Trees, for the search of superpartner of τ lepton at LHC. The previous approach was based on rectangular cuts, but one yielded no enough sensitivity in studied $\tilde{\tau}$ masses region. By using more intelligent methods, that belong to the family of supervised learning algorithms, there is a possibility to get much more sensitivity. Boosted Decision Trees have a similar concept as the cut-based approach: one divides phase space into hypercubes, labeled either as a signal or as a background, but the advantage is that BDT applies cuts automatically. In the results of the project, one increased the sensitivity to the future possible signal by using MVA approach. The optimisation from this paper will be used in further analysis.

Contents

1 Introduction

1.1 Supersymmetry

Supersymmetry is a proposed type of spacetime symmetry that relates two basic classes of elementary particles: bosons, which have an integer-valued spin, and fermions, which have a half-integer spin. Each particle from one group is associated with a particle from the other, known as its superpartner, the spin of which differs by a half-integer.

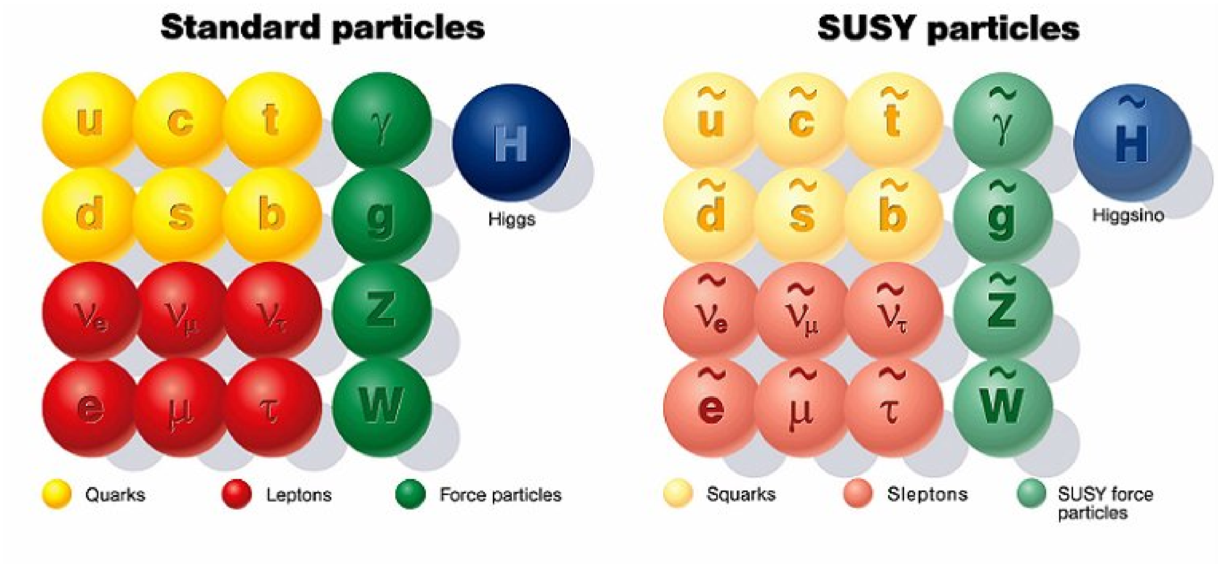

Figure 1. SM particles and their superpartners.

There are several motivation points to introduce such an extension of SM:

1. **The hierarchy problem**
   There is no scientific consensus on why, for example, the weak force is 10^24 times as strong as gravity. More technically, the question is why the Higgs boson is so much lighter than the Planck mass. For such a little mass there is a very fine-tuning cancellation between the quadratic radiative corrections and the bare mass.

2. **Gauge coupling unification**

3. **Dark Matter**
   The existence of dark matter would explain a number of otherwise puzzling astronomical observations. One of the candidates is Lightest Sypersymmetrical Particle (neutralino – a mixture of neutral higgsinos, the bino and the neutral winos).

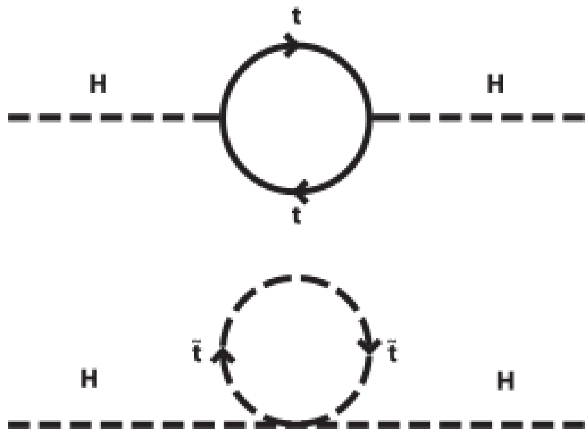

Figure 2. Cancellation of the Higgs boson quadratic mass renormalization between fermionic top quark loop and scalar stop squark tadpole Feynman diagrams in a supersymmetric extension of the Standard Model.

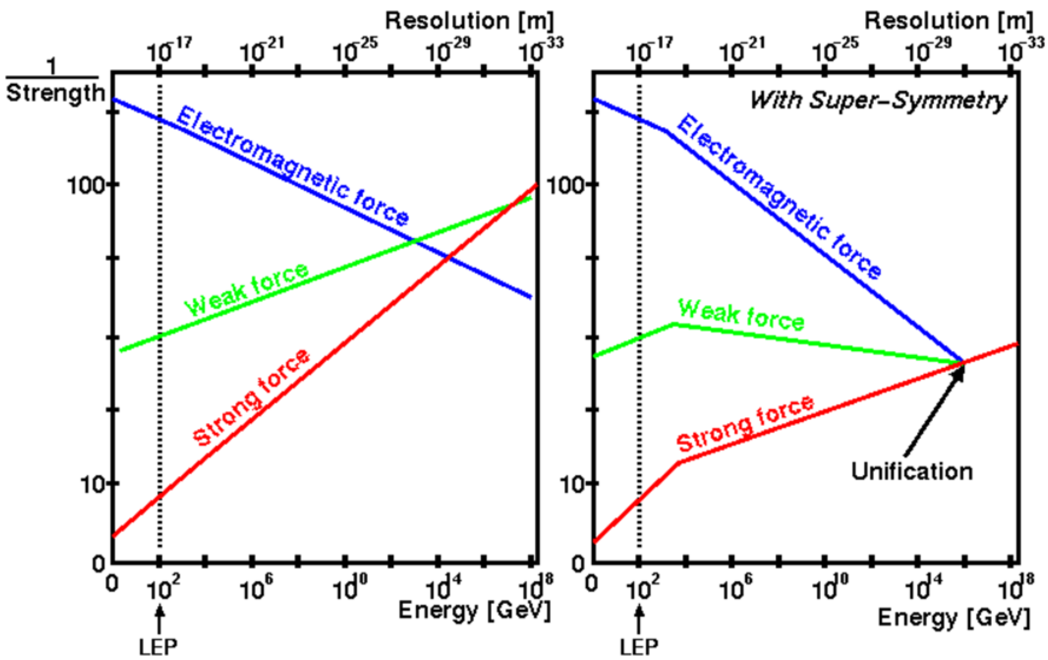

Figure 3. Grand unification of the forces with Minimal Supersymmetric Standard Model extension.

1.2 Analysis Motivation

In some SUSY models $\tilde{\tau}$, superpartner of τ-lepton, is the second lightest SUSY particle after neutralino and so it makes sense to look for $\tilde{\tau}$ LHC. During LEP experiment a region of masses of $\tilde{\chi}$ and $\tilde{\tau}$ was excluded, as one can see on the plot under the blue curve. At LHC people excluded one more $\tilde{\tau}$ mass close to 100 GeV in 2008.

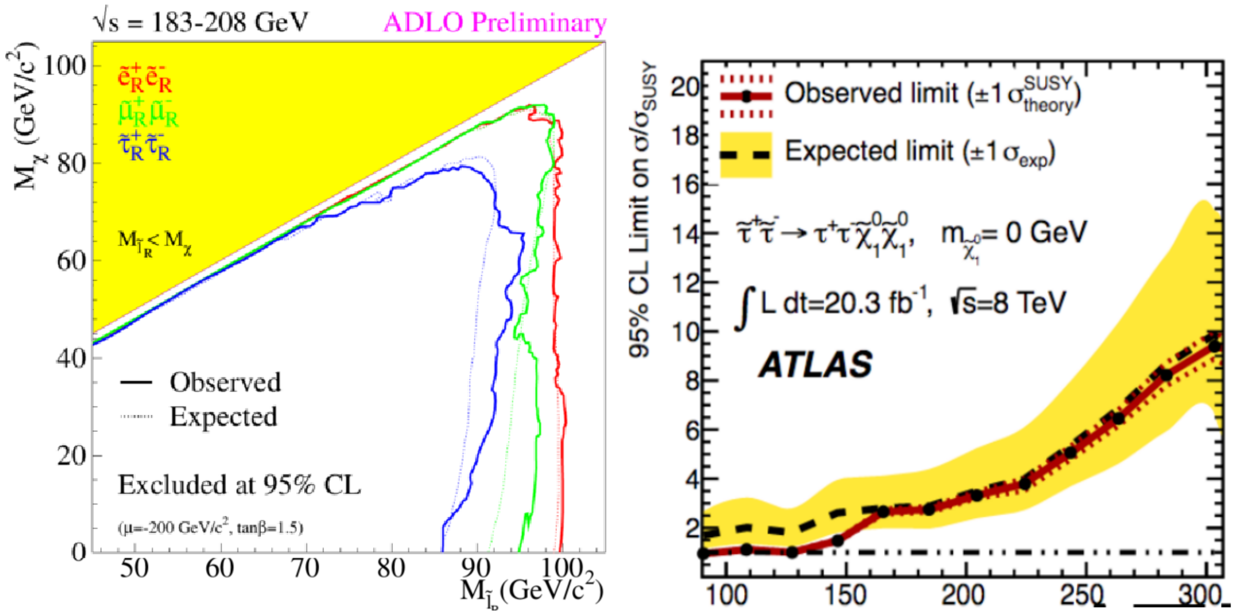

Figure 4. $\tilde{\tau}$ mass limits at LEP and LHC

2 Analysis

2.1 Studied Process

During the summerproject, I was studying direct $\tilde{\tau}$ pair production. In this process each of $\tilde{\tau}$-s decays to $\tilde{\chi}$ and τ-lepton. There are four channels of two τ-leptons (Muon-hadron channel has been studied):

1. Electron-hadron (22.8%)
2. **Muon-hadron (22.8%)**
3. Electron-muon (12.3%)
4. Hadron-hadron(42.1%)

2.2 Cut-Based Approach

Previous approach in analysis was rectangular cut-based one, which means applying physically appropriate cuts on chosen complex and simple variables and dividing the phase space formed by these variables into regions. But before applying cuts for the three selected physically appropriate variables, one need to suppress the background and enhance the ratio of signal over background by using event selection (μ is lepton 1 and τ in lepton 2):

- **Baseline Cuts**
  1. Two leptons ($\mu\tau$) with with $\Delta R < 3.5$ and special properties as given in figure below;

Property	$\mu\tau_h$ (MuHad)	$e\tau_h$ (EleHad)	τ	μ (EleMu)	e (EleMu)
$p_T >$ (GeV)	25	26	20	(Mu23 && $\mu p_T > 24$ GeV && Ele12 && $e p_T > 13$ GeV) OR (Mu8 && $\mu p_T > 10$ GeV && Ele23 && $e p_T > 24$ GeV)	
$\lvert d_{xy}\rvert <$ (cm)	0.045	0.045	–	0.045	
$\lvert d_z\rvert <$ (cm)	0.2	0.2	0.2	0.2	
$\lvert\eta\rvert <$	2.4	2.1	2.3	2.4	2.5
RelIso $<$	0.15	0.1	–	0.15	0.1
Id	medium	non-trig. MVA	see [*]	medium	non-trig. MVA
Pair	OS with $0.3 < \Delta R < 3.5$			OS with $\Delta R > 0.3$	
[*] τ candidates	`byTightIsolationMVArun2v1DBoldDMwLT` & `decayModeFindingOldDMs` & `againstElectronVLooseMVA6` & `againstMuonTight3`				
Matching to trig.	sel. lepton has to match HLT object within $\Delta R < 0.5$				

Figure 5. Summary of the required $\mu\tau$ properties.

2. No additional leptons with such conditions;
3. $n_{\text{jet}} \leqslant 1$, where jets are $P_T^{\text{jet}} > 20$ GeV and $|\eta^{\text{jet}}| < 2.4$;
4. $n_b = 0$, no b-tagged jets;
5. $M_T > 120$ GeV to suppress W+jets background;
6. $M_{l_1 l_2} < 30$ GeV to suppress Drell-Yan process background;

- **Signal Region Cuts**
  1. $\Delta|\eta|(P^{l_1}, P^{l_2}) < 2$;
  2. $M_{l_1 l_2} > 20$ GeV;
  3. $M_{T_{\text{sum}}} > 30$ GeV;
  4. $\Delta|\eta|(J_0, l_1) < 1.5$ (1-jet category only);
  5. $\Delta R(J_0, \tau) < 4$ (1-jet category only);

- **Stau Cuts**
  1. $\text{Iso}(\tau) > 0.85$;
  2. $\Delta|\eta|(l_1, l_2) < 1.5$;
  3. $\Delta\Phi(l_1, l_2) > 1.5$;
  4. $2 < \Delta R(l_1, l_2) < 3.2$;
  5. $M_{l_1, l_2} > 50$ GeV;
  6. $M_{T_{\text{sum}}} > 50$ GeV.

After applying event selection cuts, one can use three search variables, which have a shape sensitive to differences between signal and background:

1.

$$M_{T2}(m_s, \vec{s}, m_t, \vec{t}, \vec{p}_{\mathrm{T}}^{\text{miss}}; \chi_1, \chi_2) = \min_{\vec{p},\vec{q};\vec{p}+\vec{q}=\vec{p}_{\mathrm{T}}^{\text{miss}}} \left\{\max[M_{\mathrm{T}}(m_s, \vec{s}, \chi_1, \vec{p}), M_{\mathrm{T}}(m_t, \vec{t}, \chi_2, \vec{q})]\right\}, \quad (2.1)$$

where the transverse mass is given by:

$$M_{\mathrm{T}}(m, \vec{v}, \chi, \vec{p}) = \sqrt{m^2 + \chi^2 + 2\sqrt{m^2 + |\vec{v}|^2}\sqrt{\chi^2 + |\vec{p}|^2} - 2\vec{v}\cdot\vec{p}}; \quad (2.2)$$

2.

$$p_{\mathrm{T}}^{\mathrm{miss}}; \quad (2.3)$$

3.

$$D\zeta = \vec{P}^{\mathrm{miss}} \cdot \vec{\zeta} - 0.85(\vec{P}^{l_1} + \vec{P}^{l_2}) \cdot \vec{\zeta}, \quad (2.4)$$

where $\vec{\zeta}$ is bisector between directions of the two leptons.

Further cuts for these three variables are based on their shapes after the event selection. Bins with than 100 background events are taken as search regions. Cuts for 53 SR bins are explained below.

$p_{\mathrm{T}}^{\mathrm{miss}}$	M_{T2}	$D\zeta$	#SR 1-24	Category
<40	<40	<-100	1	0-Jets events
		>-100 & <0	2	
		>0	3	
	>40	>-500	4	
>40 & <80	<40	<-100	5	
		>-100 & <50	6	
		>50	7	
	>40 & <80	<-100	8	
		>-100	9	
	>80	>-500	10	
>80 & <120	<40	<-100	11	
		>-100	12	
	>40 & <80	<-150	13	
		>-150	14	
	>80	>-500	15	
>120 & <250	<40	<-100	16	
		>-100	17	
	>40 & <80	<-150	18	
		>-150 <-100	19	
		>-100	20	
	>80 & <100	>-500	21	
	>100 & <120	>-500	22	
	>120	>-500	23	
>250	>0	>-500	24	

$p_{\mathrm{T}}^{\mathrm{miss}}$	M_{T2}	$D\zeta$	#SR 25-53	Category
<40	<40	<-150	25	1-Jet events
		>-150 & <-100	26	
		>-100 & <0	27	
		>0	28	
	>40	>-500	29	
>40 & <80	<40	<-100	30	
		>-100 & <50	31	
		>50	32	
	>40 & <80	<-100	33	
		>-100	34	
	>80	>-500	35	
>80 & <120	<40	<-100	36	
		>-100	37	
	>40 & <80	<-150	38	
		>-150	39	
	>80 & <120	>-500	40	
	>120	>-500	41	
>120 & <250	<40	<-150	42	
		>-150 & <-100	43	
		>-100	44	
	>40 & <80	<-150	45	
		>-150 <-100	46	
		>-100	47	
	>80 & <100	>-500	48	
	>100 & <120	>-500	49	
	>120	>-500	50	
>250	>80 <100	>-500	51	
	>100 & <120	>-500	52	
	>120	>-500	53	

Figure 6. Signal region cuts.

With this approach one yields no sensitivity on every signal point, the significance of the signal is less than 0.2, which will be not enough to recognize future signals in real data.

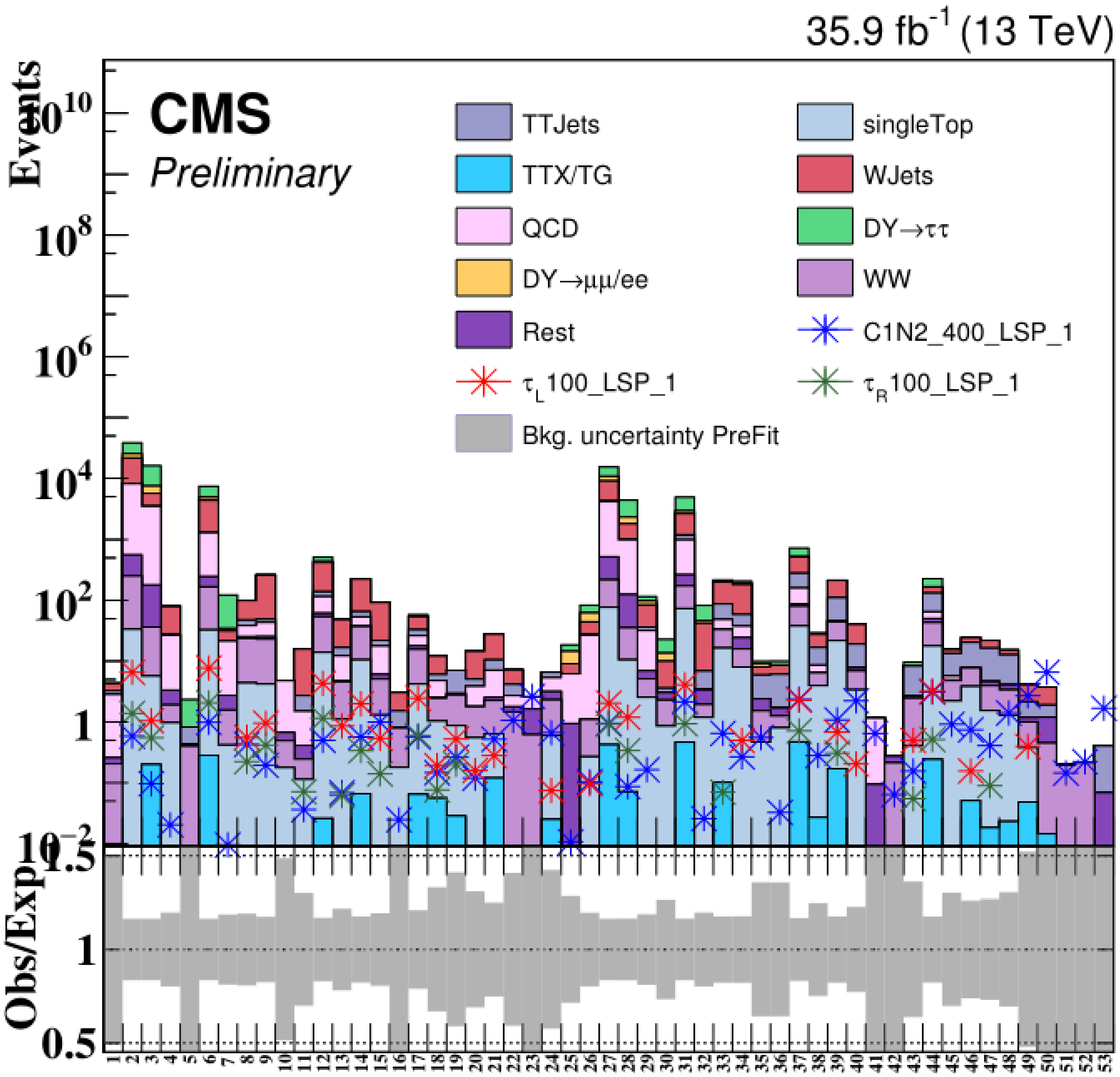

Figure 7. Example of insufficient signal over background ratio with cut-based approach.

2.3 MVA Approach

All multivariate techniques belong to the family of "supervised learning" algorithms. They make use of training events, for which the desired output is known, to determine the mapping function that either describes a decision boundary (classification) or an approximation of the underlying functional behavior defining the target value (regression). The mapping function can contain various degrees of approximations and may be a single global function, or a set of local models. Currently implemented classifiers and regression methods in TMVA:

- Rectangular cut optimisation
- Projective and multidimensional likelihood estimator (incl. regression)
- k-Nearest Neighbor algorithm (incl. regression)
- Fisher and H-Matrix discriminants
- Function discriminant
- Artificial neural networks (3 multilayer perceptron implementations) (incl. regression)
- **Boosted/bagged decision trees** (incl. regression)
- Rule Fitting
- Support Vector Machine (incl. regression)

2.4 Boosted Decision Trees

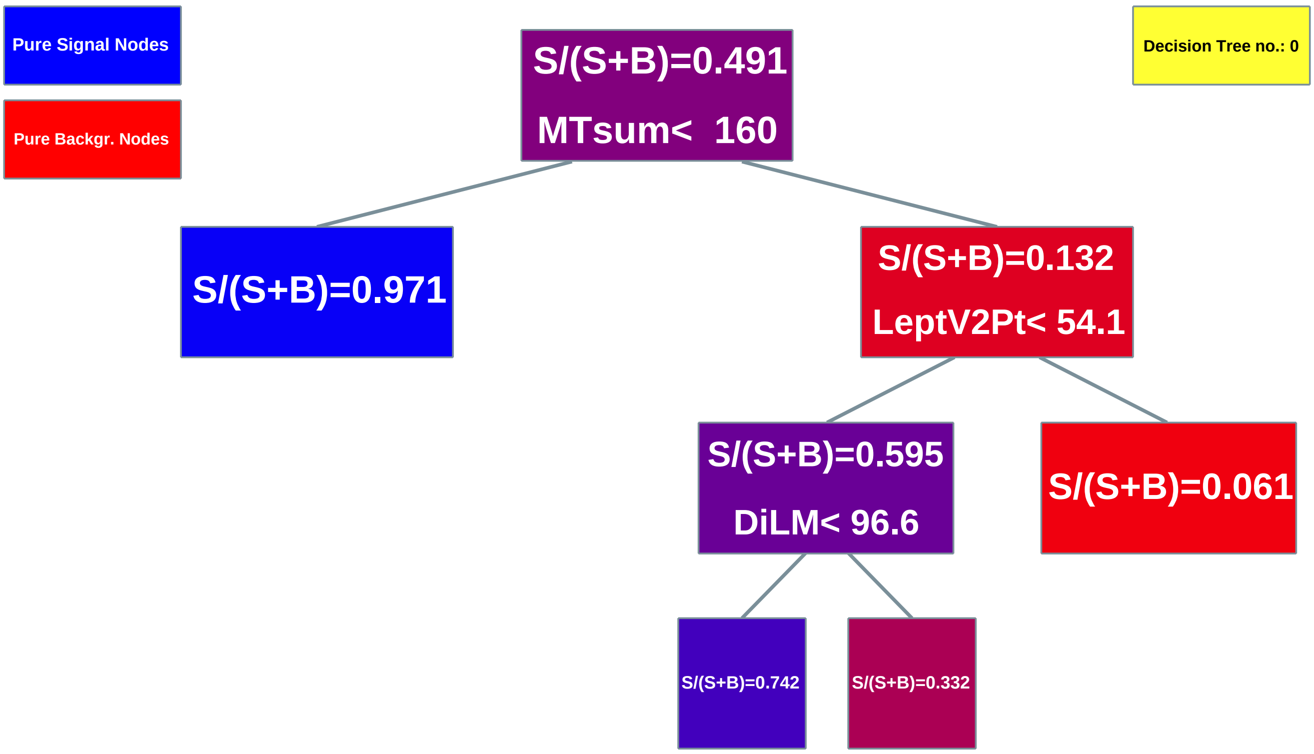

Figure 8. Example of insufficient signal over background ratio with cut-based approach.

A decision tree is a binary classifier, which starts from the root node. On each node, BDT chooses a variable for the list and a cut on this variable that divides

Figure 9. Forest of trees

phase space into 2 regions in the best way. After BDT classifier repeats the procedure more times till one of stop criteria is fulfilled. The result of one tree – several regions in phase space labeled according to the significance of the node (purity of the node is represented by color on the example above).

The problem is that using one tree for analysis one clashes instability with respect to statistical fluctuations in the training sample from which the tree structure is derived. This problem is overcome by constructing a forest of decision trees and classifying an event on a majority vote of the classifications done by each tree in the forest. All trees in the forest are derived from the same training sample by boosting procedure. After building one tree BDT adjusts weights to events, that have been classified correctly, and grow another tree from the reweighted sample. Boosting increases the statistical stability of the classifier and is able to drastically improve the separation performance compared to a single decision tree.

3 BDT Workflow

3.1 Variables

The input of BDT analysis is a list of variables of test samples. The most important property of chosen variables is different shapes of the background and known signal. In this case, it will be more efficient to apply cuts on them.

From about 40 variables I have ended up only 13 of them. After training BDT for the first time in the log file one yields BDT ranking list, which based on an information how many times each of variables has been chosen to apply cut.

In the optimisation of the list, I've noticed that after adding unimportant ones result does not change.

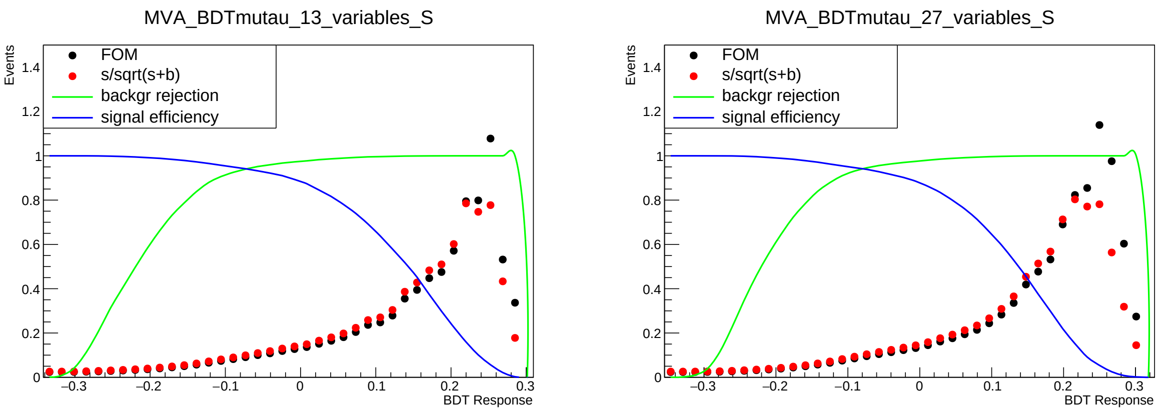

Figure 10. No increasing sensitivity after adding more than 13 variables, e.g. 27.

Descriptions of all variables in the final list (in BDT ranking order, μ is lepton 1 and τ in lepton 2 still):

1.

$$M_{T_{sum}} = M_T + M_{T_t} = \\ = \sqrt{2P_T^{l_1}P_T^{\text{miss}}(1-\cos(\Delta\Phi(P^{l_1},P^{\text{miss}}))}+ \\ + \sqrt{2P_T^{l_2}P_T^{\text{miss}}(1-\cos(\Delta\Phi(P^{l_2},P^{\text{miss}}))}; \quad (3.1)$$

2.

$$dR = \sqrt{(\Delta\Phi(P^{l_1},P^{l_2}))^2 + (\Delta\eta(P^{l_1},P^{l_2}))^2}; \quad (3.2)$$

3.

$$P_T^{l_2}; \quad (3.3)$$

4.

$$M_{l_1 l_2} = P_T^{l_1} + P_T^{l_2}; \quad (3.4)$$

5.

$$\text{Mcta} = \sqrt{2P_T^{l_1}P_T^{l_2}(1+\cos(\Delta\Phi(P^{l_1},P^{l_2}))}; \quad (3.5)$$

6.

$$M_{T2}(m_s,\vec{s},m_t,\vec{t},\vec{P}_{\mathrm{T}}^{\text{miss}};\chi_1,\chi_2) = \\ \min_{\vec{p},\vec{q};\vec{p}+\vec{q}=\vec{p}_{\mathrm{T}}^{\text{miss}}} \left\{ \max[M_{\mathrm{T}}(m_s,\vec{s},\chi_1,\vec{p}), M_{\mathrm{T}}(m_t,\vec{t},\chi_2,\vec{q})] \right\}; \quad (3.6)$$

7.
$$P_{\mathrm{T}}^{\mathrm{miss}}; \tag{3.7}$$

8.
$$H_T^{\mathrm{ext}} = P_T^{\mathrm{jet}} + P_T^{l_1}; \tag{3.8}$$

9.
$$D\zeta = \vec{P}^{\mathrm{miss}} \cdot \vec{\zeta} - 0.85(\vec{P}^{l_1} + \vec{P}^{l_2}) \cdot \vec{\zeta}, \tag{3.9}$$

where $\vec{\zeta}$ is bisector between directions of the two leptons;

10.
$$\mathrm{RHT} = \frac{P_T^{l_1}}{P_T^{\mathrm{jet}}}; \tag{3.10}$$

11.
$$\mathrm{HTOsqrMET} = \frac{P_T^{\mathrm{jet}}}{P_T^{\mathrm{miss}}}; \tag{3.11}$$

12.
$$M_T = \sqrt{2P_T^{l_1} P_T^{\mathrm{miss}}(1 - \cos(\Delta\Phi(P^{l_1}, P^{\mathrm{miss}})))}; \tag{3.12}$$

13.
$$P_T^{l_1}. \tag{3.13}$$

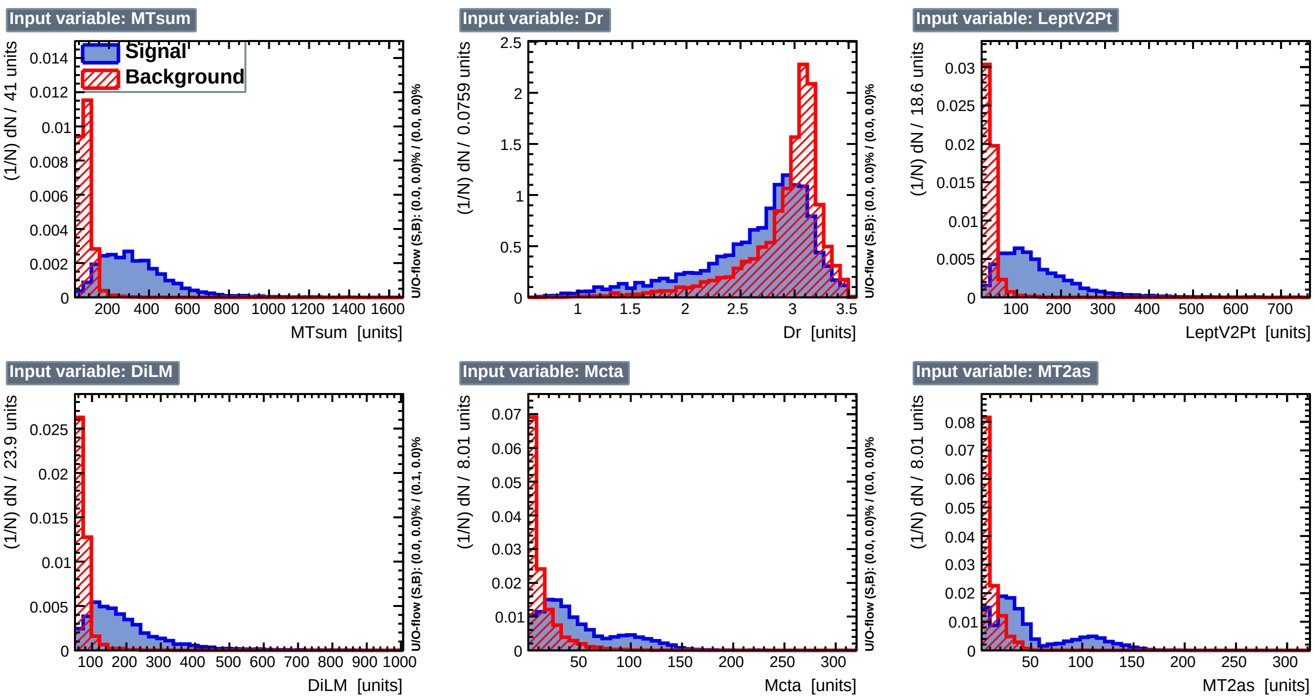

Figure 11. First 6 input variables.

3.2 Pre-Selection Cuts

After optimising the list of variables I applied pre-selection cuts form cut-based approach. All Baseline Cuts were applied, but not all of Signal Region Cuts and Stau Cuts are used, for example, only one cut can make a situation even worse.

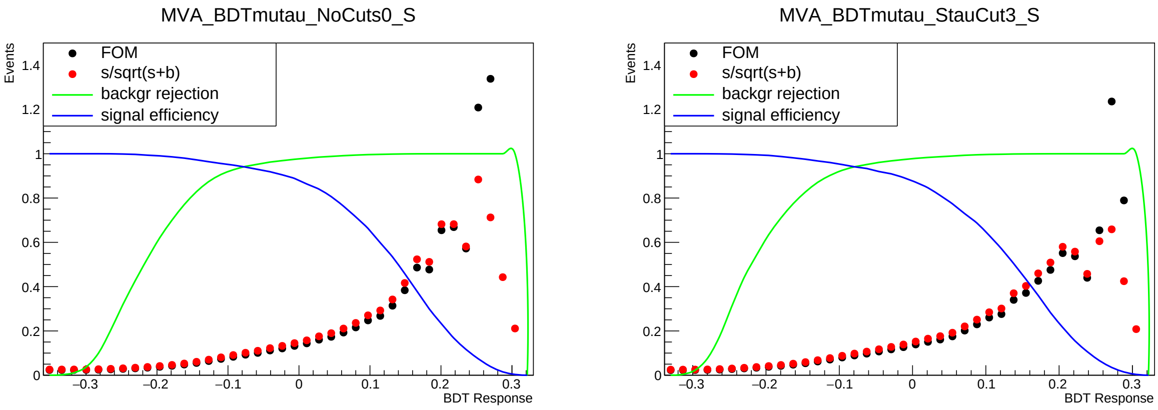

Figure 12. Cut which decreases the significance.

Optimised pre-selection cuts, that will be used in further steps are Baseline Cuts and Iso$(\tau) > 0.85$. We do need not only an increased peak of the significance but more stable to statistical fluctuations, represented by red dots and black ones. Red ones are the significance and black ones are a figure of merit, which formula includes statistical uncertainties and approximately equal to significance.

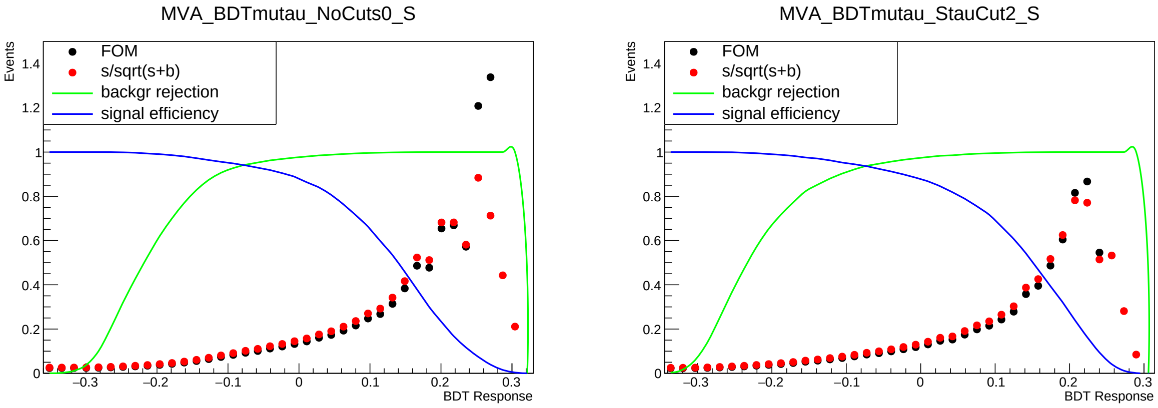

Figure 13. Cut on τ isolation which increases and stabilises the significance.

3.3 BDT Options

The most interesting, complicated and very promising part of optimisation – BDT options. On this step, I generated and compared about 2000 plots to study the influence of different BDT options values and their interference with each other. The most important options and several values (italic for default, red for optimised):

- **NTrees**
  Number of trees in the forest
  600, *800*, 1000, 1200, 1400, 1600

- **MaxDepth**
  Max depth of the decision tree allowed
  3, 5, 7, 9

- **MinNodeSize**
  Minimum percentage of training events required in a leaf node
  2%, 3%,*5%*

- **nCuts**
  Number of grid points in variable range used in finding optimal cut in node splitting
  20, 25, 30, 35, 40, 50

- **BoostType**
  Boosting type for the trees in the forest
  AdaBoost, RealAdaBoost

- **NegWeightTreatment**
  How to treat events with negative weights in the BDT training (particular the boosting)
  Pray, PairNegWeightsGlobal, *InverseBoostNegWeights*

One of the main options is BoostType. The most popular boosting algorithm is the so-called AdaBoost (adaptive boost). Starting with the original event weights when training the first decision tree, the subsequent tree is trained using a modified event sample where the weights of previously misclassified events are multiplied by a common boost weight α:

$$\alpha = \frac{1 - \text{err}}{\text{err}}. \tag{3.14}$$

We define the result of an individual classifier as $h(x)$, with (x being the tuple of input variables) encoded for signal and background as $h(x) = +1$ and –1 (in RealAdaBoost this value is real $[0, 1]$), respectively. The boosted event classification $y_{\text{Boost}}(x)$ is then given by (err – misclassification rate):

$$y_{\text{Boost}}(x) = \frac{1}{N_{\text{collection}}} \sum_{i}^{N_{\text{collection}}} \ln(\alpha_i) h_i(x), \tag{3.15}$$

where the sum is over all classifiers in the collection. Small (large) values indicate a background-like (signal-like) event. The equation above represents the standard boosting algorithm.

Examples of BDT options packs that decrease and increase maximum significance comparing with pre-selected cut result are below.

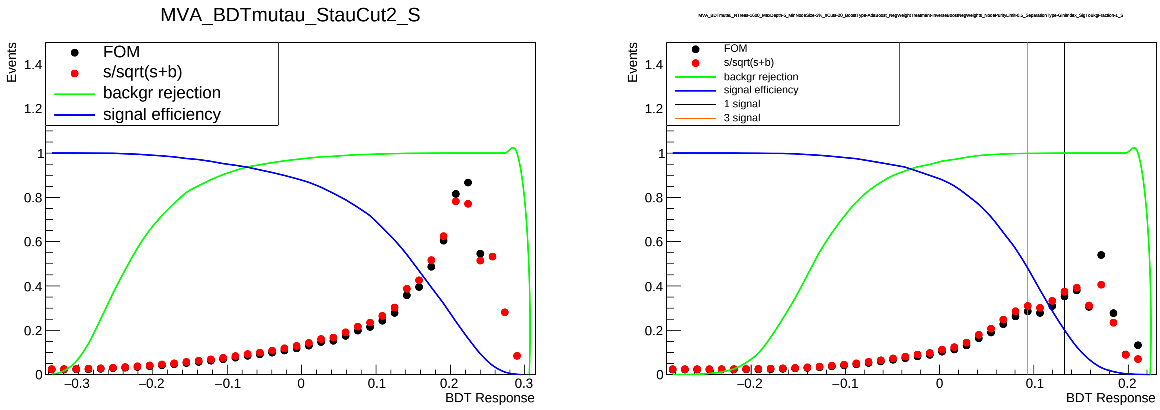

Figure 14. BDT options which decrease the significance.

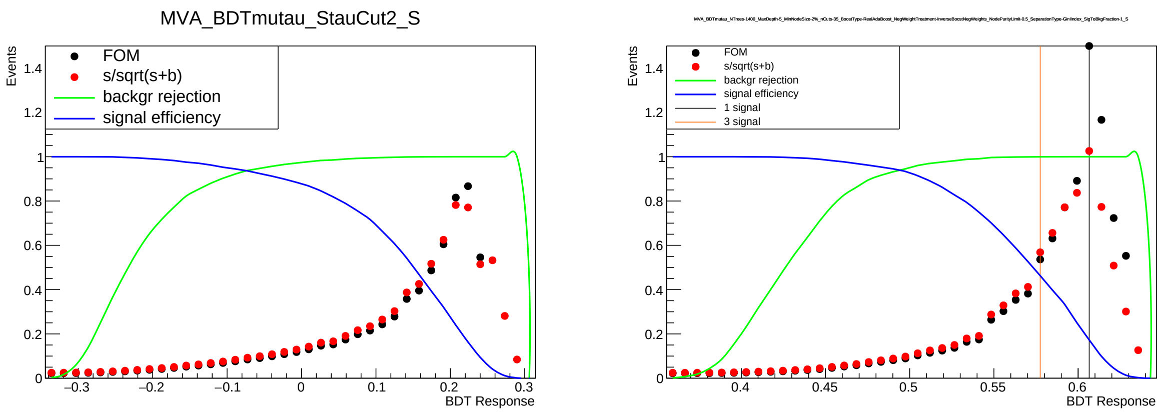

Figure 15. Optimised BDT options which increase the significance.

4 Results

During the summer I learned a lot about instruments that are used particle physics analysis such as ROOT, TMVA, Shell environment, that I've never seen before and a workflow in a SUSY DESY group. I founded DESY is the best place for particle physics work and hope to work inside these walls soon. Great acknowledgments to Ilya an Isabell from SUSY group for helping me all the summer, to Olaf for organising not only work for students but incredible parties too, to all my new friends from all over the world, to staff in the canteen, hostels, secretariat, etc., which were invisible for us, but still very important.

As a result of the project, I would like to accentuate these points of the optimisation:

- List of variables, cuts and BDT options have been **optimized**.
- Significance has been **increased**. Here in the table are results for 4 generated stau points.

	Cuts	BDT
stau100	0.22	**1.05**
stau150	0.4	0.9
stau200	0.41	0.56
stau300	0.21	0.3

- **This research will be used in further analysis for $\tilde{\tau}$ search**.

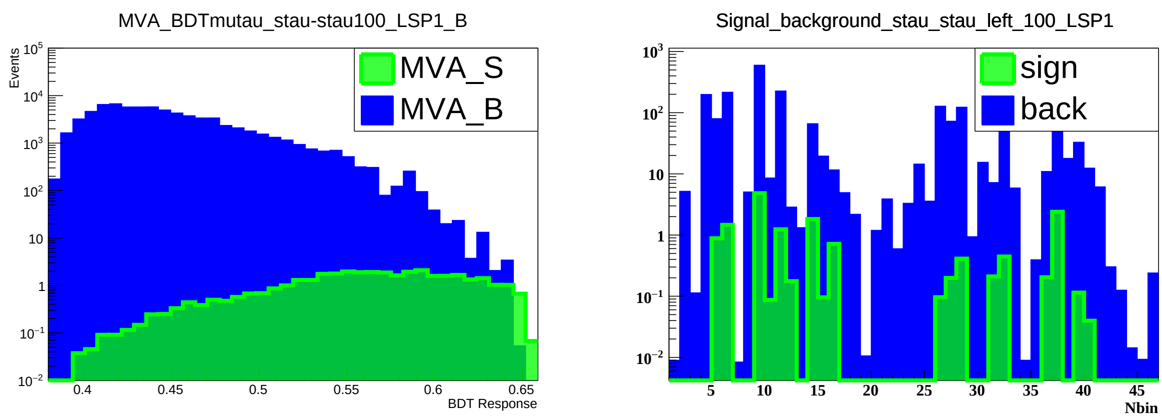

Figure 16. Comparison of the optimised BDT approach and previous cut-based one.

References

[1] Stephen P. Martin. *A Supersymmetry Primer.*

[2] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, A. Christov, D. Dannheim, K. Danielowski, S. Henrot-Versille, M. Jachowski, K. Kraszewski, A. Krasznahorkay Jr., M. Kruk, Y. Mahalalel, R. Ospanov, X. Prudent, A. Robert, D. Schouten, F. Tegenfeldt, A. Voigt, K. Voss, M. Wolter, A. Zemla. *TMVA - Toolkit for Multivariate Data Analysis.*

[3] Ilya Bobovnikov, Alexis Kalogeropoulos, Isabell Melzer-Pellmann, Alexei Raspereza. *Search for supersymmetry with tau leptons.*