Higgs mass reconstruction in H→ττ using Boosted Regression Trees

Natascha Hedrich
Thompson Rivers University

Supervisor: Dr. Dugan O'Neil (SFU), Dr. Andres Tanasijczuk (SFU)

CERN Summer Student Lectures
August 14 2013

Outline:

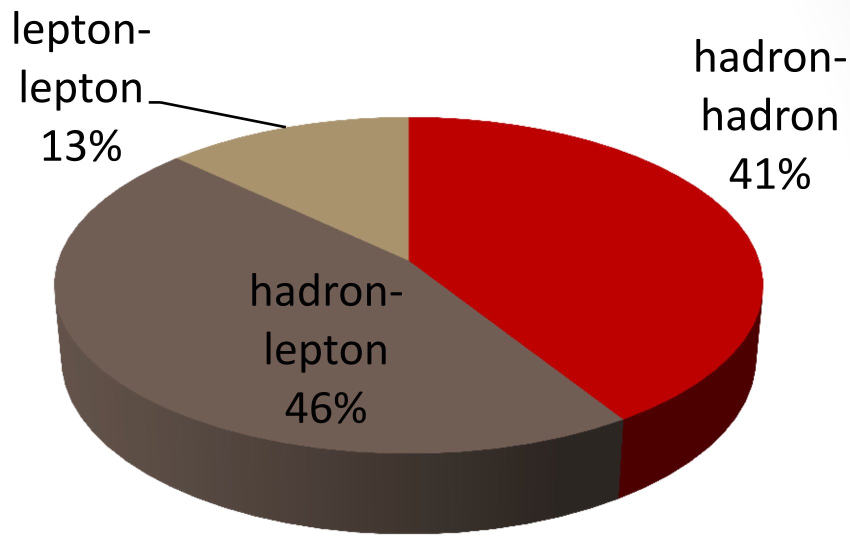
- Boosted Regression Trees (BRT's)
- TMVA
- Results
- Conclusions

Introduction: Why BRT's?

- Higgs → ττ → hadron, e, μ + ν

Lots of neutrinos!!

Difficult to reconstruct the Higgs mass



- Currently, we use the Missing Mass Calculator (MMC)
  - It's slow to evaluate
  - must be retuned for each new data set → a lot of work
- Boosted Regression Trees (BRT's)
  - are fast to evaluate
  - re-training is trivial- can we use them to reconstruct the Higgs mass?

BRT's (a quick introduction)

- **What:** Binary tree structure designed to approximate a target (Higgs mass)

- **How:**

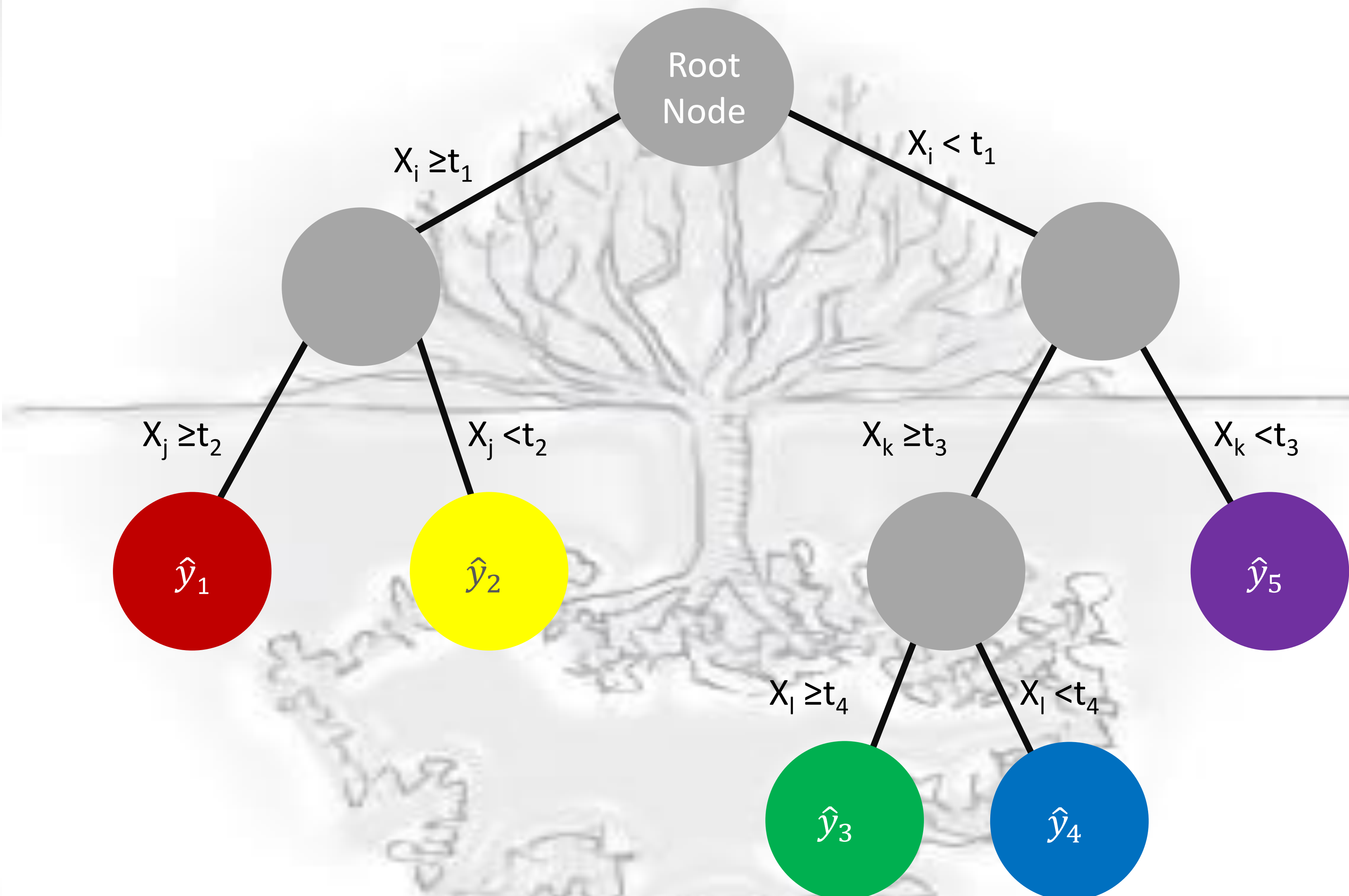
Take a set of input variables $\{X_1, X_2, \ldots , X_3\}$
Eg. MET, leadJetPt, etc…
And a target Y

At each node, apply a cut t_i to variable X_i to minimize the AVERAGE SQUARED ERROR

$$\frac{1}{N}\sum^{N}(y-\hat{y})^2$$

Output is = $\{\hat{y}_1, \ldots \hat{y}_n\}$
→ Estimates of the target
→ a mean over all training events in the node

- **Boosting?**
  - → reweight the misclassified events more heavily and repeat
    - Improves robustness against statistical fluctuations- final output is an average of the "forest" of trees

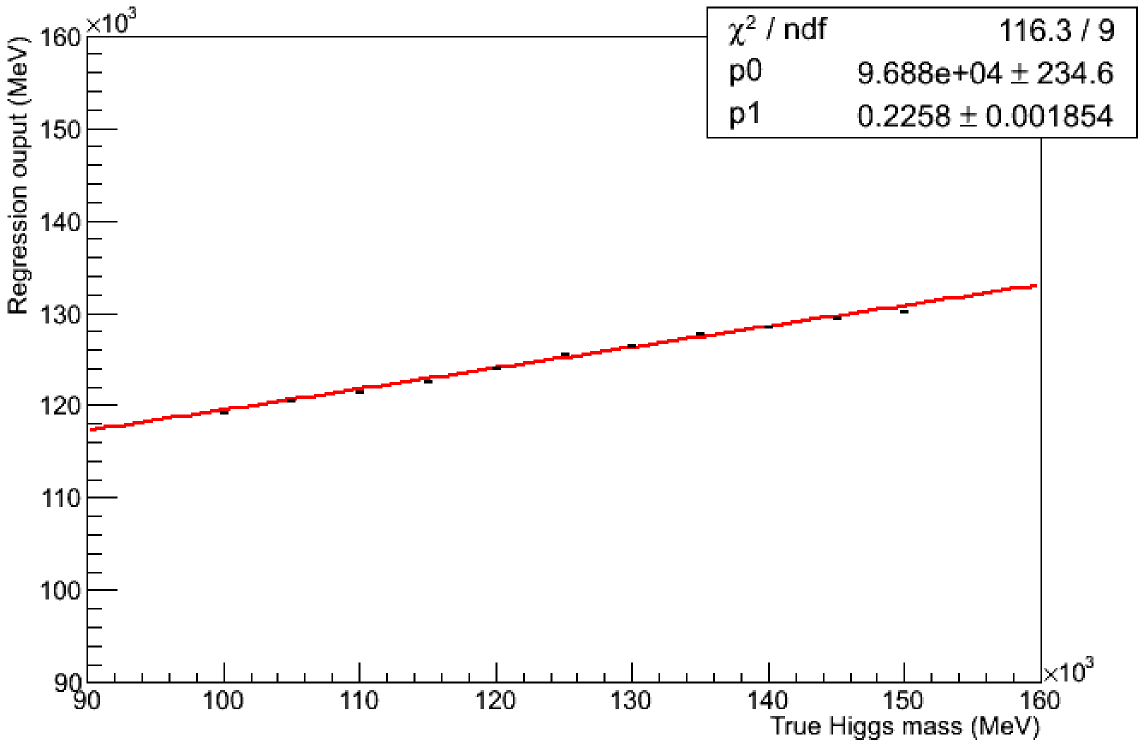


- $\{X_1, X_2, \ldots, X_n\}$ = training variables
- $\{\hat{y}_1, \hat{y}_2, \ldots \hat{y}_n\}$ = outcomes
- $\{t_1, t_2, \ldots, t_n\}$ = cuts

TMVA

- **Factory**: used to train, test and evaluate various MVA methods (in this case, BDT's)
  - **Training & Testing** – train on samples of H→ ττ with masses from 100 GeV to 150 GeV in 5 GeV increments
    - optimize cuts and save in a binary file
    - test for overtraining
  - **Evaluation** – determines regression performance and variable correlations
- **Reader**: used to apply the MVA method to an independent testing sample using the binary file produced during training
  - Read out the regression output for each event and fill into a histogram.
  - Obtain the mean and RMS of the histogram and plot mean against the true mass of the Higgs

Initial Results



Target: true_higgs_mass
(discrete distribution)

Input Variables:

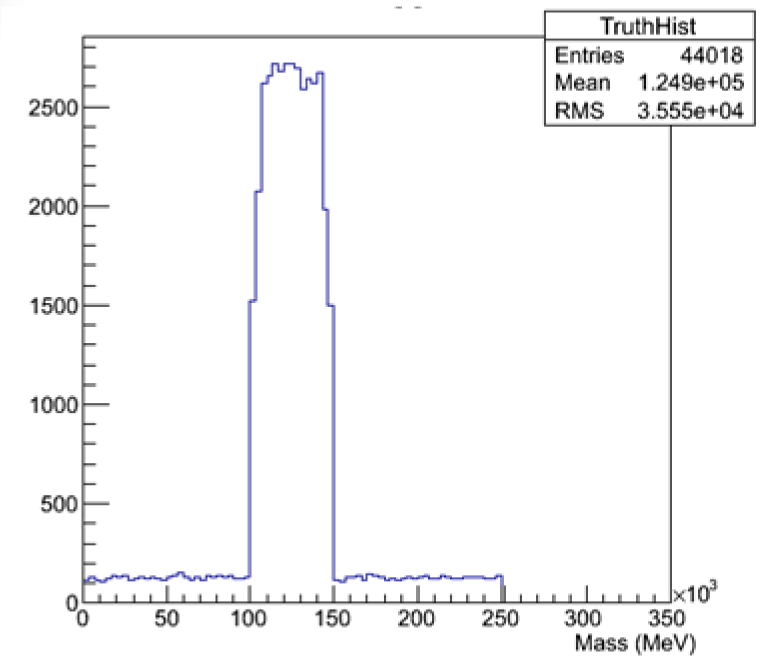
MET
LeadJetPt
mass_vis_tau_lep
sumPt
tau_fourvect.fE
…

Training Parameters:

100 Trees
20 layers
min 5 events per node

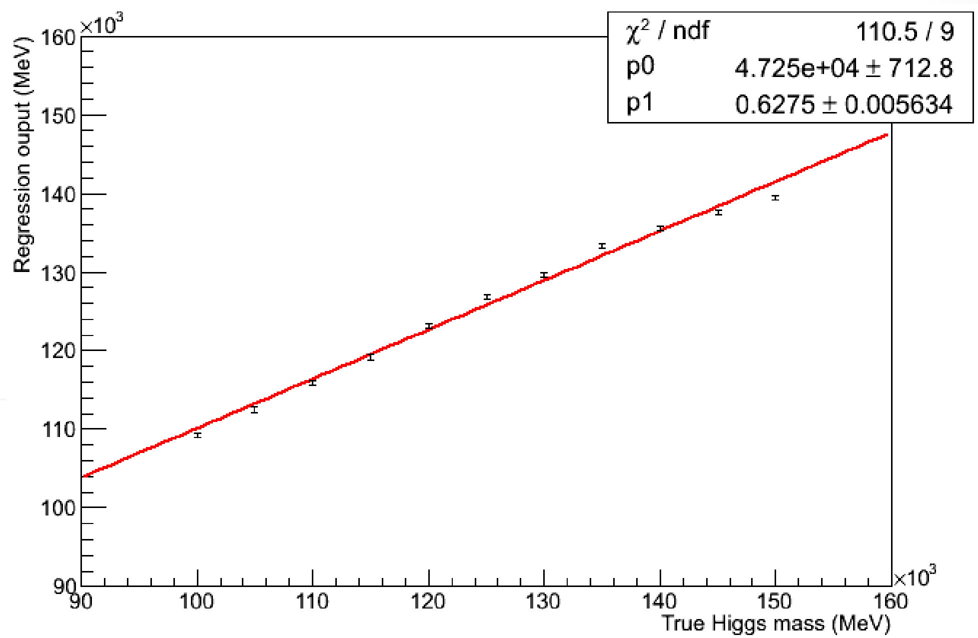
Goal → slope = 1

Improvements:



(a) Target Distribution

Introduce artificial mass target & adjust target distribution → tails are important

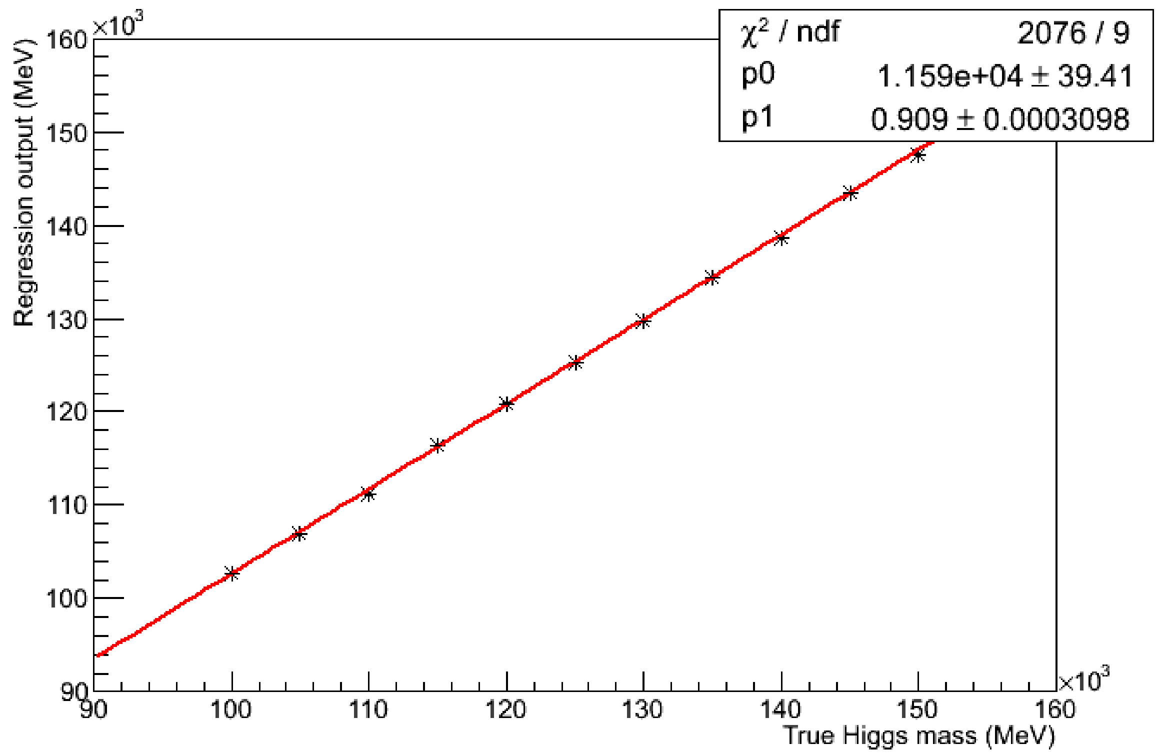


(b) Regression output

- Vary the training parameters → little improvement
- Change the input variables to truth level → large improvement

Results (truth level)

$H \rightarrow \tau\tau$ MC samples generated using PowHeg interfaced with Herwig



Target: true_higgs_mass

Input Variables:

- tau_E
- tau_px
- tau_py
- tau_pz

Training Parameters:

- 100 Trees
- 20 layers
- min 200 events per node

Outstanding Questions:

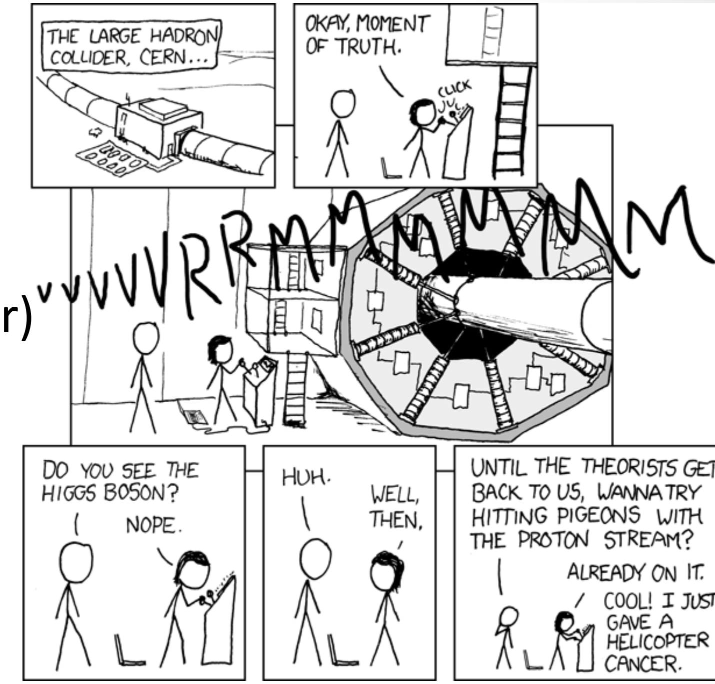
What If: we use the Higgs four vector as the input variables?

we smear the truth variables→ mimicking the reconstructed resolution?

Acknowledgements:

Thank you to:

Dr. Dugan O'Neil (SFU, supervisor)
Dr. Andres Tanasijczuk (SFU, co-supervisor)
SFU HEP group
CERN & Summer Student Team



Contact: hedrich.natascha@gmail.com

Further information about MVA and Boosted Regression Trees:

John Lu, Z. Q. "The elements of statistical learning: data mining, inference, and prediction." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173.3 (2010): 693-694.

Hoecker, Andreas, et al. "Tmva-toolkit for multivariate data analysis." *arXiv preprint physics/0703039* (2007).

SFU

THOMPSON RIVERS UNIVERSITY

Additional Slides

Input Variables/Parameters

MET
dphi_met_lep
dr_tau_lep
leadJetPt
mass_transverse_met_lep
mass_transverse_met_tau
mass_vis_tau_lep
pt_ratio_tau_lep
pt_vector_sum_all
sumPt
tau_fourvect.fE
lep_fourvect.fE

factory.BookMethod():

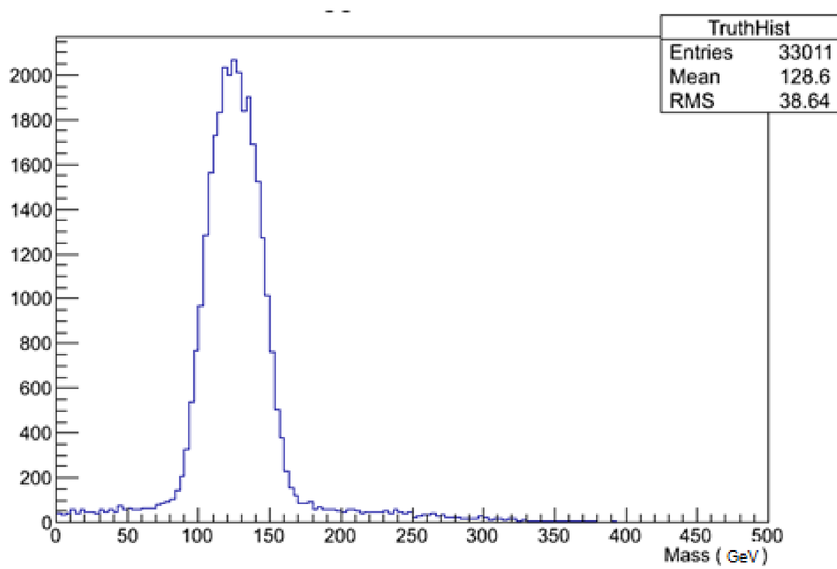
- NTrees=100
- nEventsMin=5
- MaxDepth=20
- BoostType=AdaBoost
- AdaBoostBeta=0.2
- SeparationType=RegressionVariance
- nCuts=20

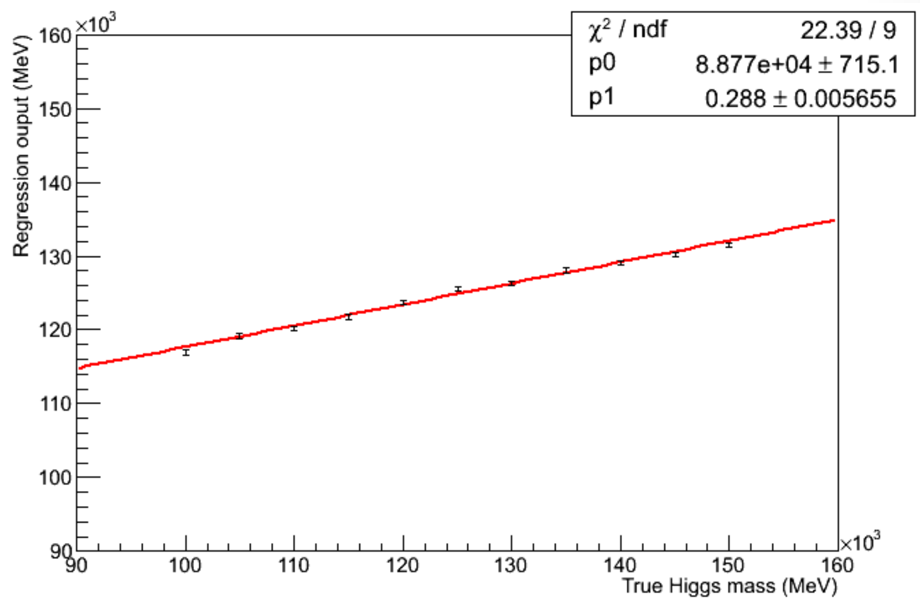
Variables were chosen based on a standard signal/background analysis

N.Hedrich (TRU)
August 14, 2013

What happens if we use Gaussian distributions?

→For each mass m, use a gaussian distribution centered at m with a standard deviation of 50 GeV for 100 GeV<m< 150GeV and with a standard deviation of 100 GeV for 100 GeV and 150 GeV.





Variation of Training Parameters

Question: Can we still improve the performance by other means?

→ vary the training parameters

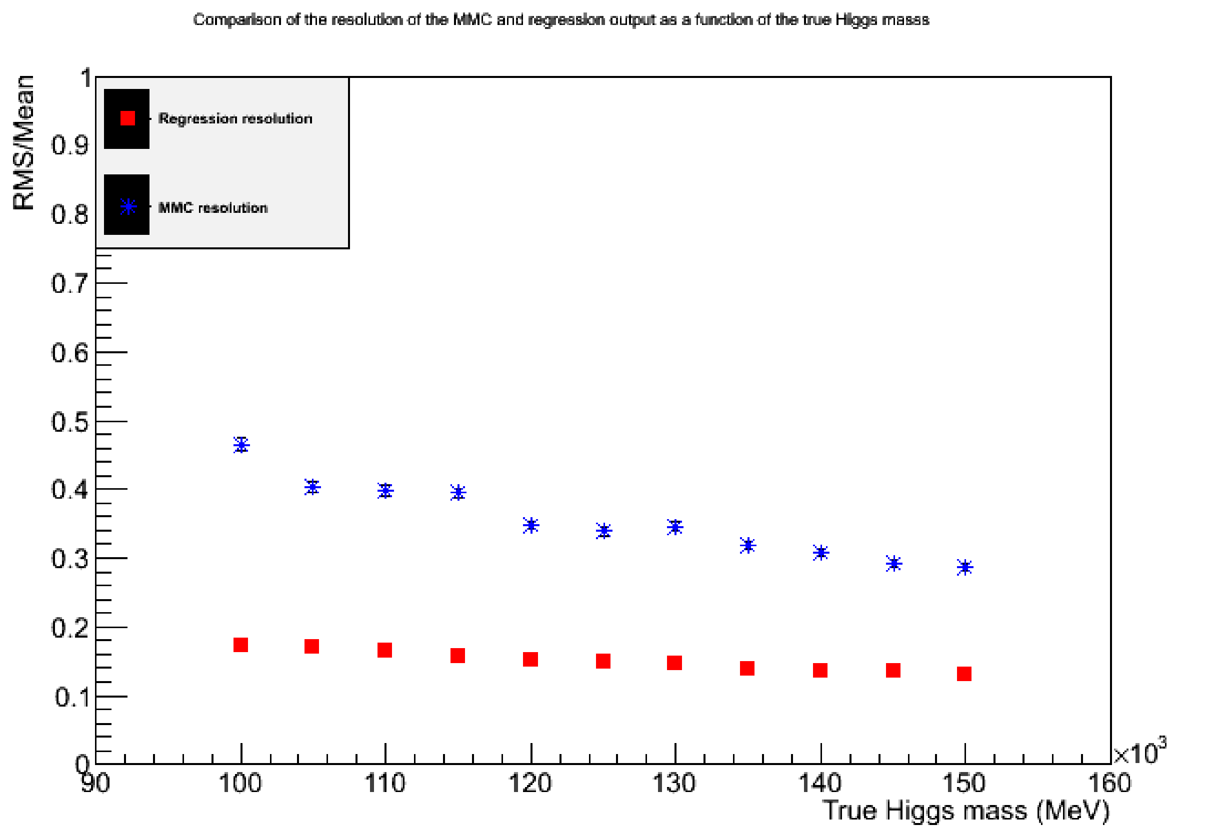
1. Use more trees
   NTrees = 1000 --> slope = 0.6527
2. Increase the minimum number of events per node
   nEventsMin = 50 -->slope = 0.71
3. Make the trees deeper
   MaxDepth = 200 --> slope = 0.5943
4. Increase the boosting parameter
   AdaBoostBeta = 0.5 --> slope = 0.6295
5. Decrease the boosting parameter
   AdaBoostBeta = 0.05 --> slope = 0.5917

Conclusion: there is little improvement from changing the training parameters except when increasing nEventsMin

Note: increasing Ntrees → increase in training time
With a low nEventsMin, there is some evidence of overtraining

Resolution

- Comparing the resolution between the MMC and regression, they are very comparable. Regression seems a bit better, but in reality is due to compressed range.



SFU

THOMPSON RIVERS UNIVERSITY