Ideal τ tagging with TMVA multivariate data-analysis toolkit

A. Heikkinen[1], P. Kaitaniemi, V. Karimäki, R. Kinnunen, M. J. Kortelainen, T. Lampén, S. Lehti, T. Lindén, and L. Wendland

Helsinki Institute of Physics, P.O. Box 64, FIN-00014 University of Helsinki, Finland.

March 25, 2009

[1] aatos.heikkinen@cern.ch

Outline

- **Motivation**: our experience on using ROOT package TMVA for multivariate data analysis [2]
- **Use case:** τ tagging in the framework of heavy charged MSSM Higgs boson searches at the LHC
- **Goal:** to investigate how in the ideal case (no detector effects) τ tagging could be performed and signal τs separated from the background

- **Data:** a generator level Pythia/Tauola data for 14 TeV p-p collisions
  - Signal: hadronic τ decay
  - Background: hadronic jets from QCD multi-jets (p_T = 120-170 GeV/c)

- **Method/Results:** The τ tagging efficiency and background rejection are studied with various MVA classifiers
  - Moderate customisation was required to map our problem into the TMVA framework

[2] T. Lampen *et. al.*, Testing TMVA software in b-tagging for the search of MSSM Higgs bosons at the LHC, CHEP07, Journal of Physics: Conference Series 119 (2008) 032028

Data

The data for QCD multi-jet background was generated using Pythia8.108 for LHC 14 TeV p-p collisions.

The signal was generated with Pythia6.4.19, and required to consist only of genuine τ jets from the (heavy charged) MSSM $H^{\pm} \rightarrow \tau^{\pm}\nu_{\tau}$ decay with τ polarization simulated with Tauola 2.6[3].
($m_{H_{\pm}} = 217 GeV/c^2$, $\tan\beta = 30$, max. m_H SUSY scenario [4].)

Our analysis is based on the following variables:

Variable	Name
Jet E_T	`jetEt`
Jet η	`jeteta`
Track isolation (ΔR=0.50)	`isolMaxPt50`
Isolation of electromagnetic energy (ΔR=0.10-0.50)	`ecalIsolEt10_50`
Neutral hadron rejection	`hcalRatio`
(i.e. track p matching to hadronic energy deposition)	
R_{τ} = p(leading track) / E(jet)	`rtau`

[3] D.P. Roy, Phys. Lett. B 459 607-614
[4] arXiv:hep-ph/9912223

Production and Preselection Cuts

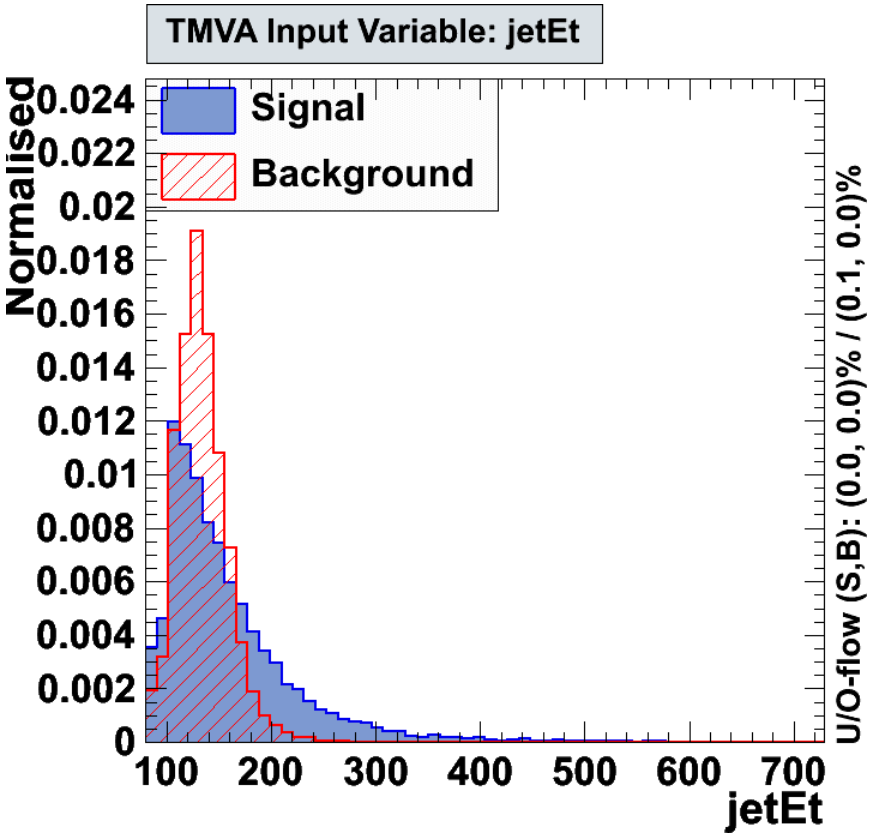
At Monte Carlo level the following preselection cuts were used[5]:

- `jetEt` $>$ `100`
- `abs(jeteta)` $<$ `2.2` (Fig. 1)
- Jet leading track $p_T > 20$ GeV, matching cone 0.1
- Isolation:
  - track $p_T > 0.5$
  - abs(track η) < 2.5
  - abs(d_{IP}-z) < 0.2
  - around the leading track:
    - signal cone $\Delta R < 0.04$
    - isolation cone 0.04-0.5
  - signal tracks 0, 1 or 2
  - isolation tracks 0 or 1

Further, before analysis we choose:

- 1-prong signal final states (`signalTracks==1`)
- MC matched τ jets coming from $H^{\pm}$ (`tauDecayType==4`)

[5] τ jet tagging: The European Physical Journal C - Particles and Fields, Volume 46, Supplement 1, July 2006; $H^+ \rightarrow \tau\nu$: J. Phys. G: Nucl. Part. Phys. 34 (2007) 2307-2455



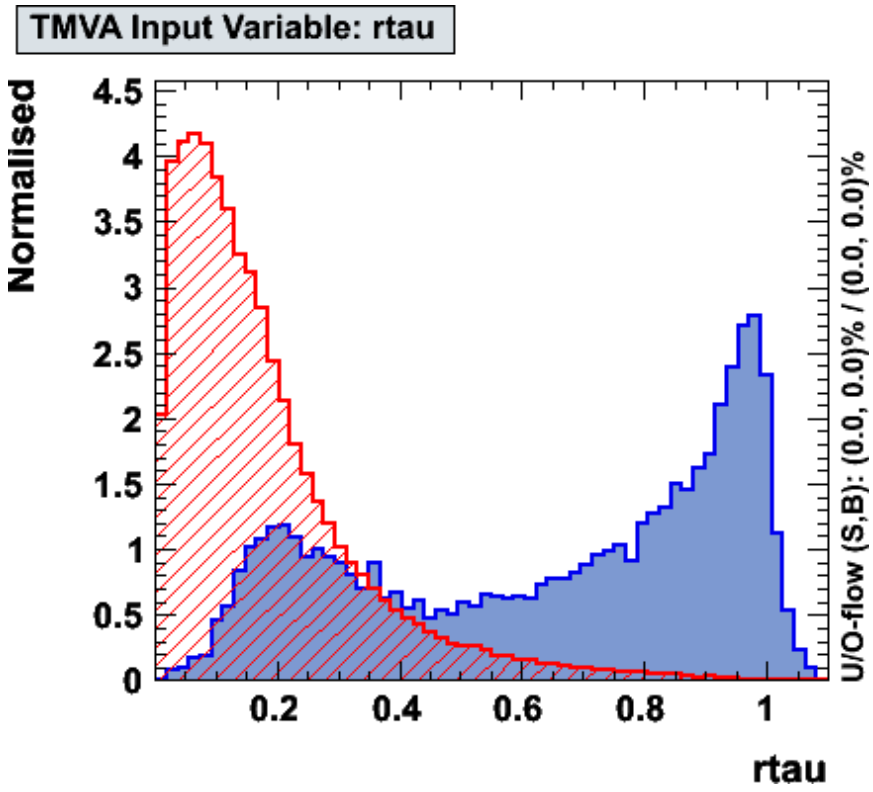


Fig. 1: Example of data used in τ tagging: distribution of jet E_T (`jetEt`) and R_τ (`rtau`).

TMVA - Toolkit for Parallel Multivariate Data Analysis

ROOT-integrated TMVA is a framework the training, testing and performance evaluation of multivariate classification techniques.

TMVA works in **transparent factory mode to guarantee an unbiased performance comparison between the classifiers**, such as:

- Rectangular cut optimisation
- Projective likelihood estimation (PDE approach)
- Multidimensional k-nearest neighbour classifier
- Linear discriminant analysis (H-Matrix and Fisher discriminants)

In this presentation, we are focusing to following classifiers:

- Artificial neural networks (ANN)
- Boosted/Bagged decision trees (BDT)
- Support Vector Machine (SVM)

Customisation of TMVA for τ tagging 1/2

Customisation of TMVA included:

- Configuration files
- Evaluation of **event efficiencies** with `TMVA::Reader`, taking into account the MC-level preselection efficiencies and TMVA preselection efficiencies
- Timing profiling with `TStopwatch`
- Using signal (event) efficiency at background (event) **efficiency level** 10^{-5} and 10^{-6}
- Also we tabulating data related to signal and background **jet efficiency** at 10^{-5}

(See example listing on next slide)

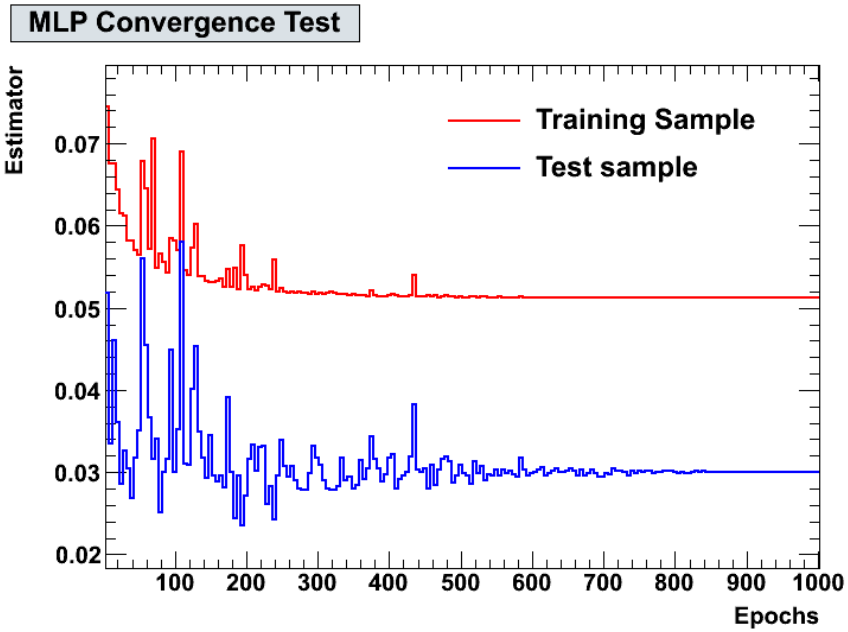
Customisation of TMVA for τ tagging 2/2

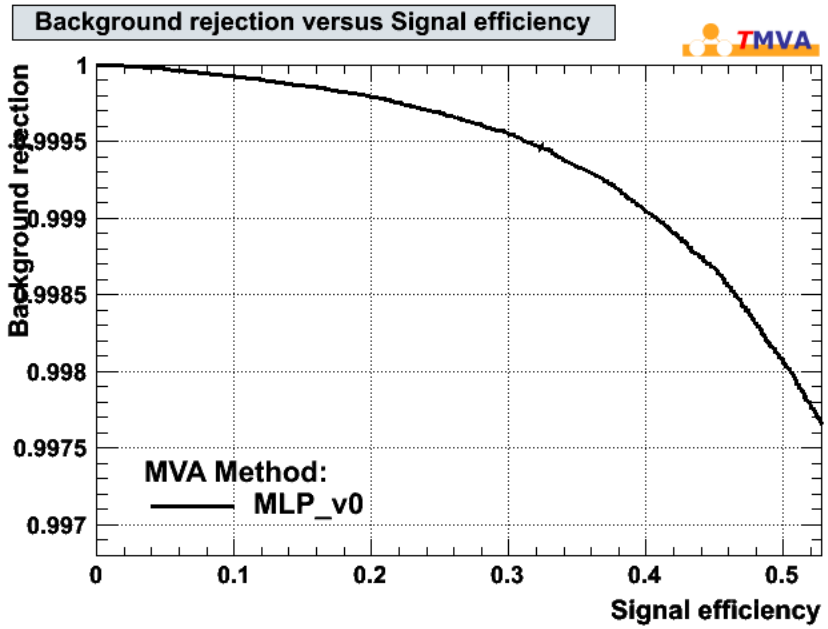
```
MyEvaluate : Testing classifiers in order to obtain signal/background efficiencies for EVENTS
MyEvaluate : (not jets, which are the input to TMVA)
MyEvaluate :
MyEvaluate : ------------------------------------------------------------------------
MyEvaluate :              |              Events                |               Jets
MyEvaluate :              |   Signal   eff       Bkg    eff |   Signal  eff       Bkg  eff
MyEvaluate : ------------------------------------------------------------------------
MyEvaluate : Generated    :   200000         89799840        |      N/A            N/A
MyEvaluate : Event presel :    47866 0.239    1470627 0.016 |   206079        4345182
MyEvaluate : TMVA  presel :    35615 0.744    1087375 0.739 |    35615 0.17   2230315 0.51
MyEvaluate : Total presel :          0.178             0.012 |
MyEvaluate : ------------------------------------------------------------------------
MyEvaluate :
MyEvaluate : Thus 1e-5 overall bkg event efficiency corresponds to 8.3e-04 bkg event
MyEvaluate : efficiency by TMVA.
MyEvaluate :
MyEvaluate : Signal and background event efficiencies have been scaled with the preselection
MyEvaluate : event efficiencies
MyEvaluate : Evaluation results ranked by best signal efficiency at 1e-5
MyEvaluate : ------------------------------------------------------------------------
MyEvaluate : MVA              Signal event efficiency at bkg event efficiency (error):
MyEvaluate : Methods:         @B=1e-6      @B=1e-5      @B=1e-4      @B=1e-3      @B=0.01
MyEvaluate : ------------------------------------------------------------------------
MyEvaluate : SVM_Gauss_2_5   : 0.0212(007)  0.0556(012)  0.1096(016)  0.1670(019)  0.1774(020)
MyEvaluate : ------------------------------------------------------------------------
```

Results - Neural Networks 1/2

- **ANN:** Of the three Multilayer Perceptrons (MLP) implementations supported by TMVA we used `MLP` (`TMVA::Types::kMLP`).
- **Variables:** `jetEt`, `jeteta`, `isolMaxPt50`, `ecalIsolEt10_50`, `hcalRatio`, and `rtau`
  - Transformation: `log(jetEt)` and Principal Component Analysis (`VarTransform=PCA`)
- **Training:** Data for 6-15-1 MLP configuration with neurons of sigmoid type we trained 1k cycles
  - 5k signal jets and 40k background jets
- **Testing:** Rest of the data was used for testing (35k signal and 2230k background jets)
  - Systematic uncertainty was estimated by repeating the full analysis with independent background data

Results - Neural Networks 2/2





Convergence of ANN over 1000 training cycles and background rejection vs. signal efficiency.

Signal efficiency (%)	Background eff.
7.6±0.2 (stat.) ± 0.3 (syst.)	10^{-5}
2.3±0.1 ± 0.3	10^{-6}

Results - Boosted Decision Trees

A decision tree is a binary tree classifier. Repeated left/right (yes/no) decisions are performed on a single variable at a time:

- The phase space is split into regions that are classified as signal or background, depending on the majority of training events that end up in the final leaf nodes

Boosted Decision Trees (BDT) implemented in TMVA represents an extension to a single decision tree to several decision trees derived from the same training sample by reweighting events.

New combined classifier has stabilized response with respect to fluctuations in the training sample.

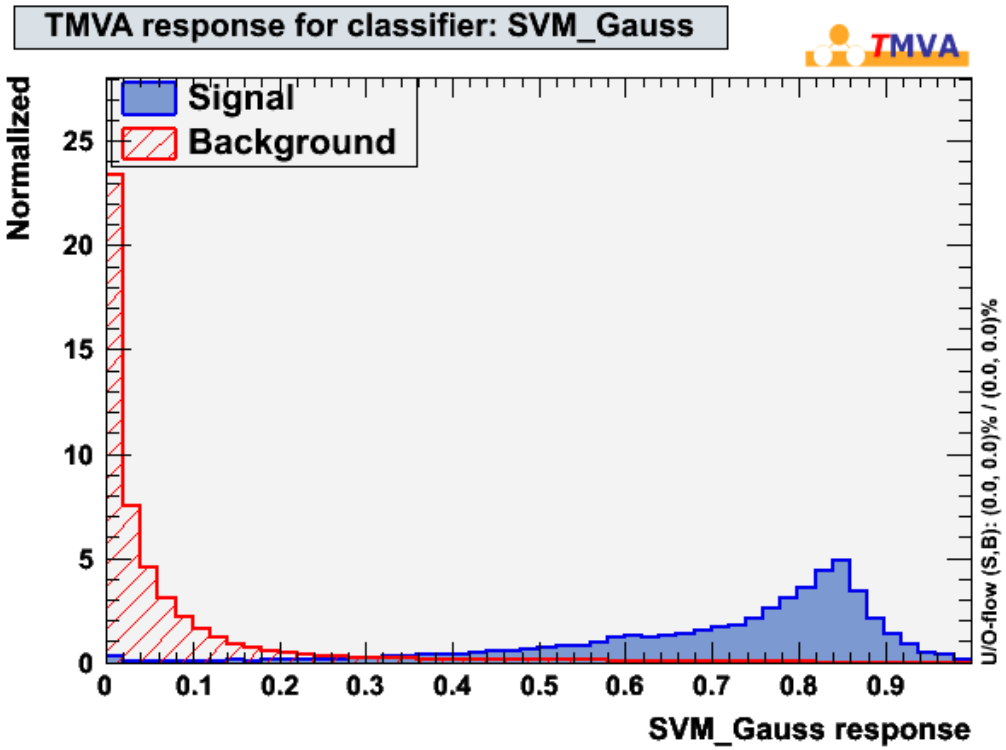
For our τ tagging problem the BDT method was found to give **stable results with only small fluctuations**:

Signal efficiency (%)	Background eff.
7.8±0.1	10^{-5}
3.9±0.1	10^{-6}

Results - Support Vector Machine 1/2

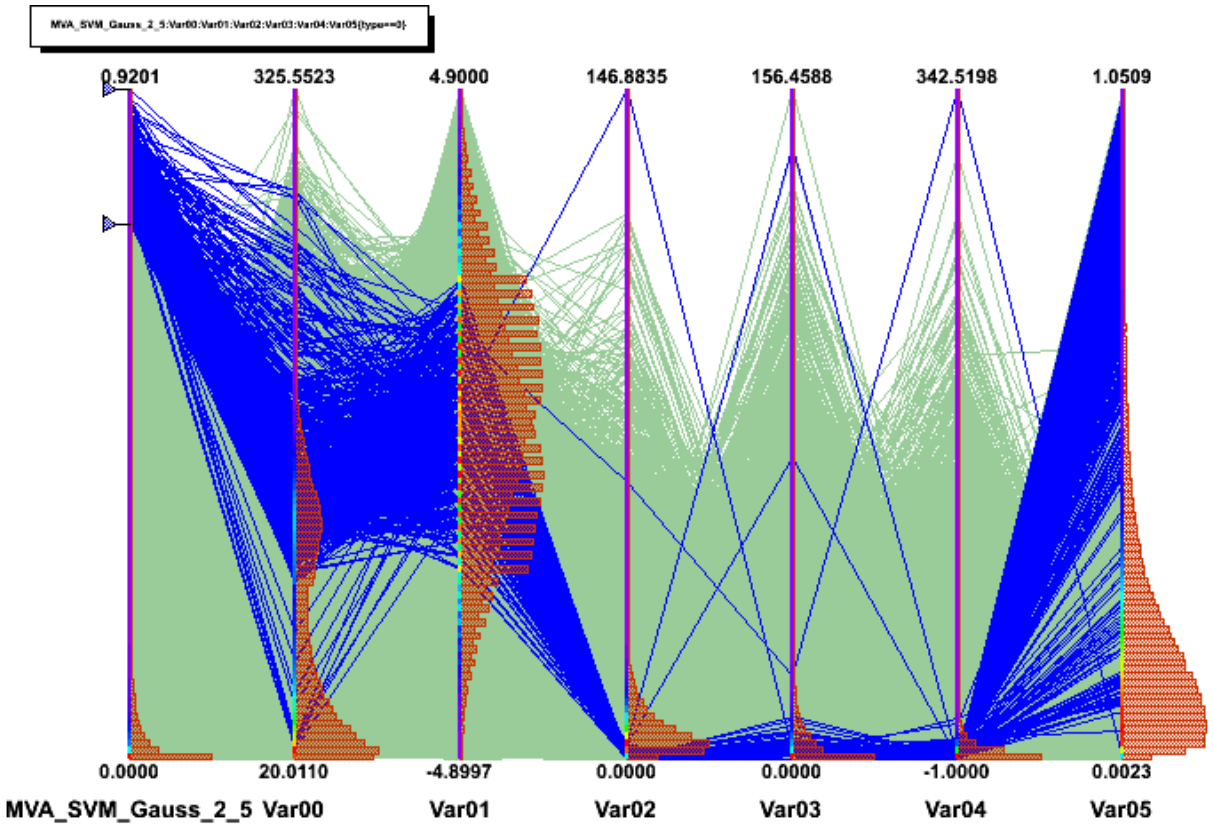
SVM were **trained with 5k signal jets** and **40k background jets**:

- The SVM training time scales as $O(n^2)$, where n is the number of events in the training sample
- In practice training took $\sim$1 h



The output of the SVM classifier with a Gaussian kernel.

Results - Support Vector Machine 2/2



Parallel coordinates -plot indicating poorly classified events.

Signal efficiency (%)	Background eff.
$5.6 \pm 0.1 \pm 0.1$	10^{-5}
$2.1 \pm 0.1 \pm 0.1$	10^{-6}

Conclusions

- We have discussed from the user point of view how TMVA package in ROOT can be used in τ tagging
- We have seen that since CHEP'07 TMVA has matured, and is now fully integrated to ROOT
  - An **interface for adding new classifiers** is provided
- Some customisation was needed to cast our problem into the TMVA framework

- Multivariate data-analysis techniques show a promise in τ tagging:
  - At 10^{-5} background efficiency, TMVA classifiers have 5.6-7.8 % signal efficiency
- Finally, we have defined areas where our approach can be improved (e.g. using additional variables, and yet unused classifiers available in TMVA), and are planning additional research on τ tagging