Prompt/non-prompt lepton discrimination study with BDT in the ttH multilepton final states at ATLAS

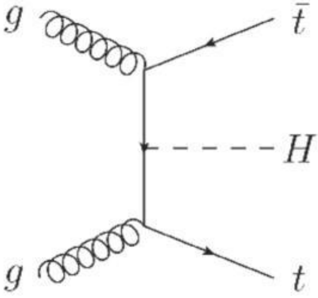
Nazim Huseynov [JINR]   Yuriy Ilchenko [U of Texas]

Physics Computing Russian Institutes meeting

September 22, 2015

Physics Motivation

- ttH measures directly the top quark Yukawa coupling
- Fundamental constant of nature – proportional to top quark mass
- $H \to bb$ is a dominant decay mode - irreducible $t\bar{t} + bb$ background
- $H \to WW$ is a second largest branching ratio
- $\sigma_{ttH} = 130 fb^{-1} at\ 8 TeV - expect \approx 2700\ ttH$ events with $21 fb^{-1}$
- $\sigma_{ttH} = 509 fb^{-1} at\ 13 TeV$



Predicted theory events
$\int L = 21 fb^{-1}\ at\ \sqrt{s} = 8 TeV$

	N_{events}
$ttH \to 4W + 2b$	588
$ttH \to 4WW\tau\tau + 2b$	172
$ttH \to WWZZ + 2b$	72

Multilepton channels

5 analysis channels ($H \to WW, \tau\tau, ZZ$)

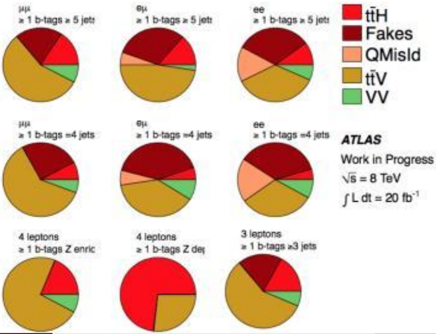
- $ttH \to$ *same sign* 2-*lepton*
- $ttH \to$3-*lepton*
- $ttH \to$4-*lepton*
- $ttH \to$2-*lepton* $+ 1\tau$
- $ttH \to$1-*lepton* $+ 2\tau$

Signal and background composition

Dominant backgrounds

Channel	S/B	Signal	Background	top (%)	ttW (%)	ttZ (%)	VV (%)	Z+jets (%)
2lee5j	0.16 ± 0.04	0.73 ± 0.03	4.56 ± 1.17	28.8	32.7	24.4	10.6	3.5
2lem5j	0.20 ± 0.03	2.13 ± 0.05	10.62 ± 1.54	24.0	47.3	22.6	3.5	2.7
2lmm5j	0.19 ±0.03	1.41 ± 0.04	7.57 ± 1.31	23.3	51.8	14.4	9.0	1.6
2lee4j	0.04 ± 0.01	0.44 ± 0.02	10.16 ± 2.43	49.1	20.4	9.5	7.6	13.5
2lem4j	0.06 ± 0.01	1.16 ± 0.03	18.51 ± 2.54	44.2	33.9	11.4	10.4	0.0
2lmm4j	0.07± 0.01	0.74 ± 0.03	10.26 ± 1.82	36.1	46.4	10.1	5.2	2.2
3l	0.24± 0.03	2.34 ± 0.04	9.63 ± 1.33	12.6	28.2	46.6	9.2	3.3
4lZenr.	0.22 ± 0.02	0.19 ± 0.01	0.83 ± 0.07	0.5	0.8	88.0	9.1	0.0
4lZdep.	4.17 ± 2.42	0.03 ± 0.003	0.01 ± 0.004	16.7	33.3	50.0	0.0	0.0

$t\bar{t}$ production with non−prompt leptons is a major background for the few ttH channels, e.g. 2l, 3l



Motivation of BDT implementation

- Non-prompt leptons from hadron decays is a significant background for a number of channels
- $t\bar{t}$ is one of major processes contributing non-prompt leptons in the signal region
- Idea is to use TMVA - Toolkit for Multivariate Analysis – boosted decision tree (BDT) – to separate prompt from non-prompt leptons
- Original roadmap is described in the proposal note – http://www.yuraic.web.cern.ch/yuraic/notes/note_tth_bdt_proposal.pdf

Samples and Strategy

Samples

- Full simulation sample.

  group.physhiggs.ttHlep.117050.NTUP_COMMON.e1728_s1581_s1586_r3658_r3549_p1575_vFullTruth4_Special/

- Altfast simulation sample.

  group.physhiggs.ttHlep.117050.NTUP_COMMON.e1727_a188_a171_r3549_p1575_vFullTruth4_SpecialV2/

Strategy

- Run the ntupler (UT)
  - Save only needed lepton variables for decision tree
  - Matching procedure: mark leptons as prompt or non-prompt based on the truth information in MC
- Employ BDT from TMVA
- Comparison with standard cuts
- Try to optimize standard cuts with MVA

Object selection

- Muon cuts
  - Matching $\Delta R = 0.05$
  - $|eta| < 2.47$, excluding $1.37 < |eta| < 1.52$
  - Muon ID: Tight
- Electron cuts
  - Matching $\Delta R = 0.1$
  - $|eta| < 2.47$, excluding $1.37 < |eta| < 1.52$
  - Electron ID: VeryTightLH

BDT parameters and statistics

Training and Test samples

- Samples
  - Training - 8000 events for signal and background
  - Test - 8000 events for signal and background
- Decision Tree
  - Boosting algorithm, BoostType=Grad
  - Maximum cell tree depth, NNodesMax=3
  - A variable range granularity nCuts=20

BDT input paramters

We use lepton variables only (no event global variables are included)

Muons

Variable	Separation
1: *mu_etcone20/mu_pt*	6.172e–01
2: *mu_ptcone20/mu_pt*	4.955e–01
3: *mu_sigd0PV*	3.870e–01
4: *mu_z0SinTheta*	3.010e–01
5: *mu_pt*	3.357e–01
6: *mu_eta*	7.432e–03

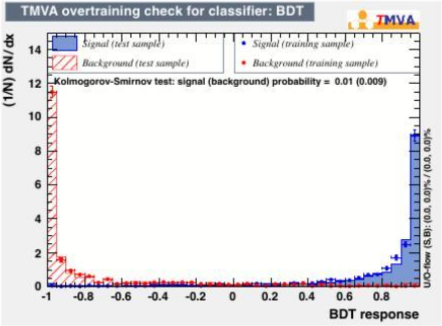
Electrons

Variable	Separation
1: *el_etcone20/el_pt*	4.677e–01
2: *el_ptcone20/el_pt*	4.290e–01
3: *el_sigd0PV*	3.344e–01
4: *mu_z0SinTheta*	3.103e–01
5: *el_pt*	1.692e–01
6: *el_eta*	2.291e–02

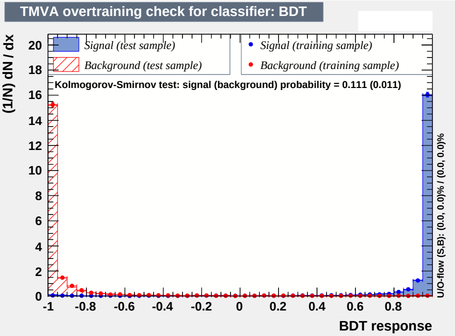
BDT Response

Training and Test samples

Muons



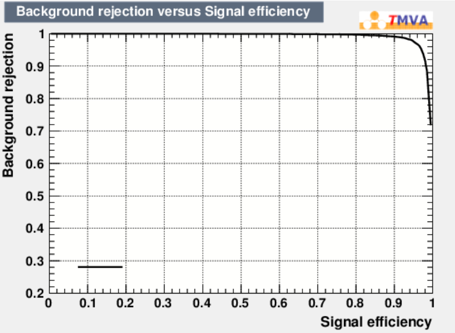
Electrons



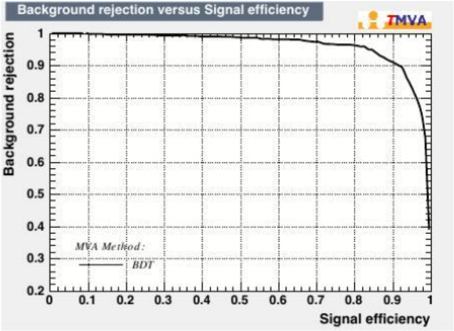
Good separation between prompt(signal) and non-promt (background) leptons

Receiver-operating characteristic (**ROC**) curves allows us to assess the decision tree performance

Muons

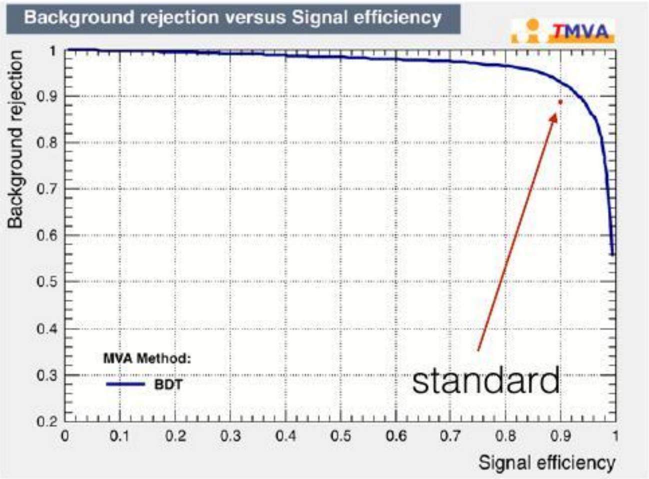


Electrons



Comparison BDT vs cuts for electrons

Electrons

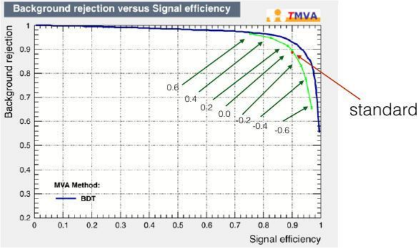


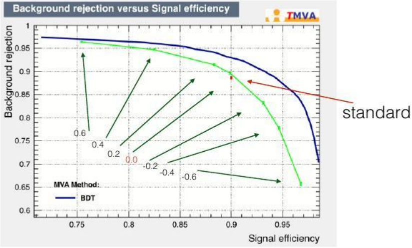
BDT is better by $30 - 40\%$ in number of non-prompt electrons passed

Comparison BDT vs cuts (zoomed) for electrons

TMVA BDT

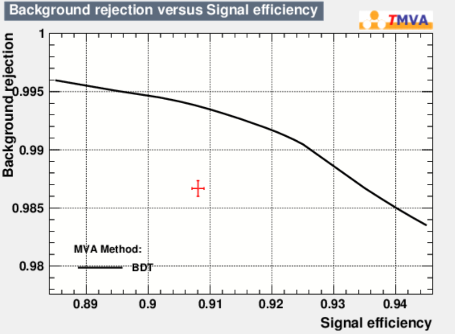
BDT score cuts : 0.6, 0.4, 0.2, 0.0, -0.2, -0.4, -0.6

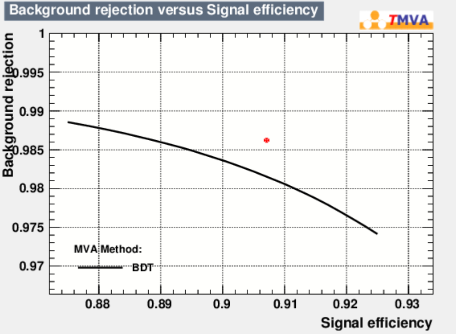




Comparison BDT vs cuts (zoomed) for muons

Muon ROC curve: low-stat (left), large-stat(right)



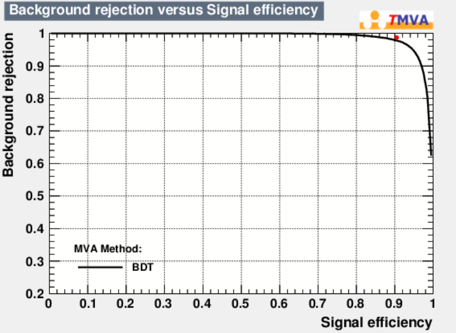


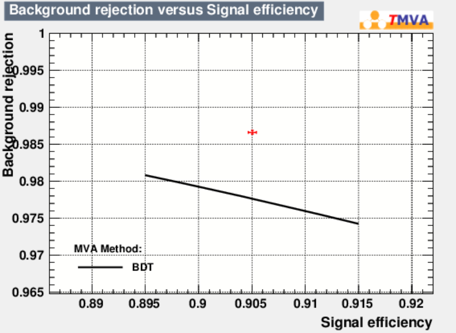
Decided to take into account pile-up by proper sampling
Produced two samples 10% and 33% of the size of fullsim sample

Comparison BDT vs cuts (zoomed) for muons

TMVA BDT

ROC curve comparison: 10% fullsim samples





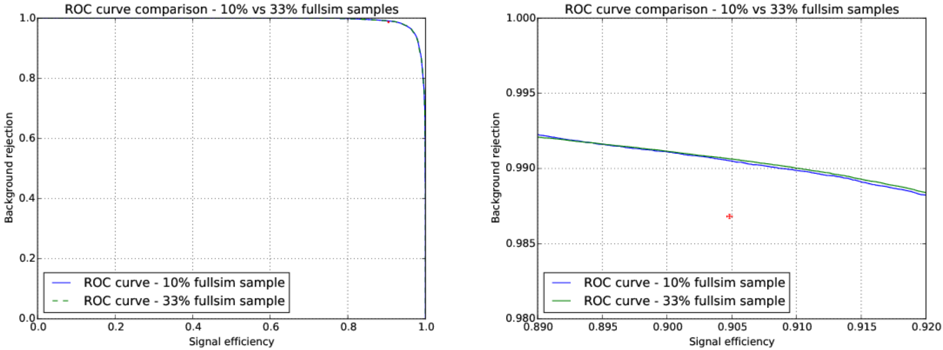
TMVA is unable to construct BDT on 33% sample (errors)

TMVA shows bad performance

Comparison BDT vs cuts (zoomed) for muons

scikit-learn BDT

Applied scikit-learn - widely used open-source MVA library, an industry standard
Produces stable ROC-curve



Summary

- TMVA produces instable results
- scikit-learn BDT gives a better discrimination vs standard cuts for muons
- standard cut optimization with BDT does not seem to be possible but requires further investigation
- at a given signal efficiency BDT rejection is better than the standard cuts
- estimated improvement in background rejection is about 25% (scikit-learn)

Plans

- understand why TMVA misbehaves (ask within community and/or get in touch with authors possibly?)
- repeat the study for electons
- use ttH signal for prompts and ttbar for non-prompt
- employ either scikit-learn or fixed TMVA to implement MVA object selection cuts
- produce analysis cutflow and compare with the standard cutflow

THANKS FOR YOU ATTENTION!

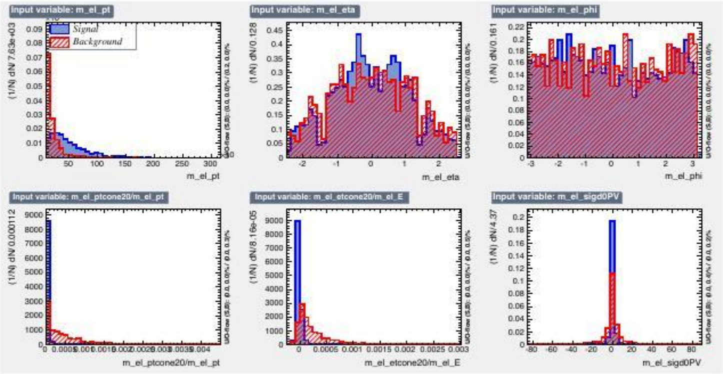
Backup

Electron variables distribution

Training sample

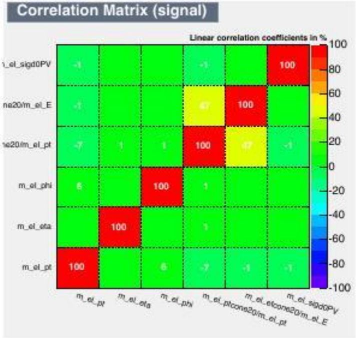
- Signal: blue
- Background: red

** Bottom left and center plots scale by 1000 x-axis*

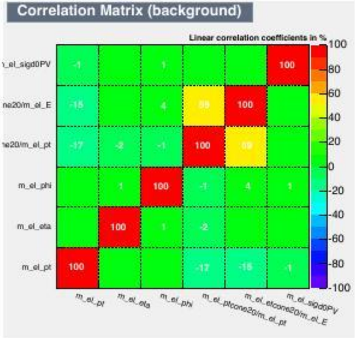


13

Electron variables correlations



Signal sample

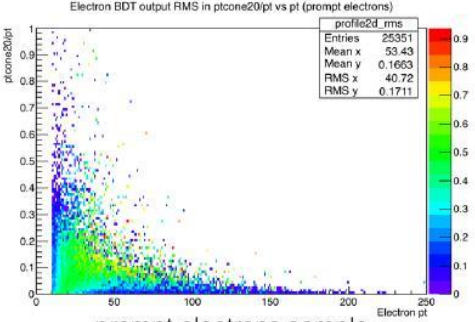


Background sample

14

Electron BDT score RMS

- BDT score RMS in pt vs ptcone20/pt coordinates
  - Electrons with low pt and low ptcone20/pt have large BDT score RMS (green area); it means information from other variables is used

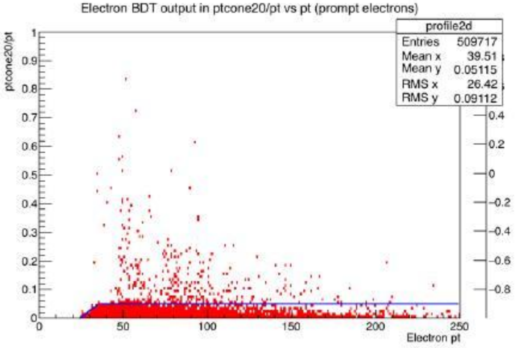


prompt electrons sample

6

Isolation cut optimization

- Standard: **ptcone20/pt=0.05**
  - Events with BDT score = 0.6 selected (subset of the left plot from slide #5)
  - Optimized isolation cut is selected as follows,



pt: [24, 35] GeV

cut = 0.05/11*pt - 0.11

pt: [35, inf] GeV

cut = 0.05

prompt electrons sample 7