PanTau

Using TMVA for tau lepton identification

Tim Below[1] Sebastian Fleischmann[2]
Christian Limbach[1] Robindra Prabhu[3]
Peter Wienemann[1]
in collaboration with Mark Hodgkinson[4]

[1]Physics Institute, University of Bonn

[2]University of Wuppertal

[3]University College London

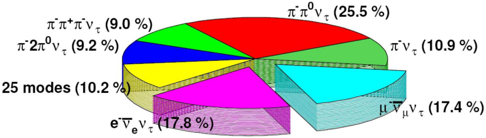
[4]University of Sheffield

21th January 2011
TMVA workshop

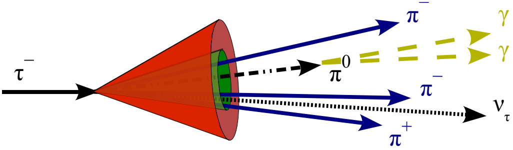
Outline

Hadronic τ decays

- 65% of all decays into hadrons



- decay via resonances (ρ, a_1, K, ...) into $\pi^{\pm}$ and π^0
  - $\pi^0 \rightarrow \gamma\gamma$ ($BR = 99.8\%$) with $\tau_{\pi^0} \approx 8 \cdot 10^{-17}$s
- low charged multiplicity:
  - 1 $\pi^{\pm}$ ("1-prong"): $\pi^{\pm}$, $\rho^{\pm} \rightarrow \pi^{\pm}\pi^0$, $a_1^{\pm} \rightarrow \pi^{\pm}\pi^0\pi^0$, ...
  - 3 $\pi^{\pm}$ ("3-prong"): $a_1^{\pm} \rightarrow \pi^{\pm}\pi^{\pm}\pi^{\mp}$, ...
  - very small fraction (0.1%) of 5-prong decays
- strongly collimated jet



Reconstruction of τ leptons

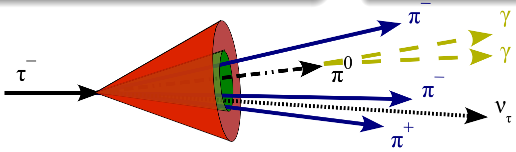
- Only hadronic τ decays are considered in this context

(Inner Detector) Tracking

- Tracks from $\pi^\pm$
  - low multiplicity (1 or 3)
  - collimated
  - isolation from other tracks
- Secondary vertex

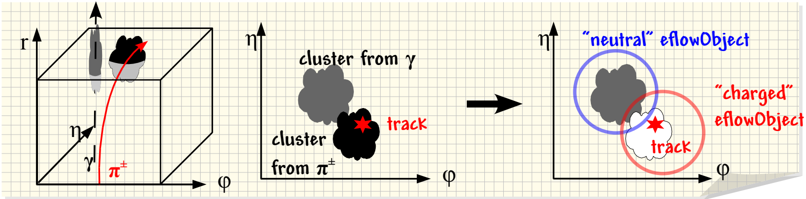
Calorimetry

- Collimated energy deposit
- EM and HAD component
  - strong EM component for 1-prong
  - identify π^0 subclusters
- Isolation



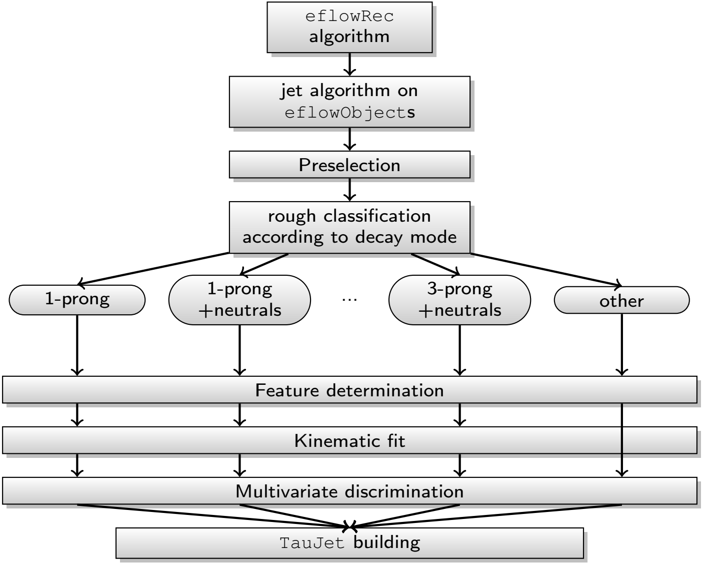
Tau identification with PanTau

- Based on generic energy flow algorithm eflowRec (M. Hodgkinson, D. Tovey, R. Duxfield)
  - Physics motivated approach to combine tracking and calorimetry
  - Use tracking data for charged particles and calorimeter data only for neutrals
  - Therefore: Subtract energy deposition of charged particles from calo clusters



- Separation of detector effects from actual tau identification
- Early categorisation according to (estimated) decay mode

Tau identification with PanTau



Multivariate Analysis with TMVA in PanTau

TMVA completely integrated in PanTau

- Various discrimination methods can be used in parallel (with same interface)
- Call TMVA `Reader` inside ATLAS software allows to quickly exchange TMVA training (weight) files. Just by changing python steering files, can be done without recompiling a single line of code
- Currently mainly using projective likelihood ratio
- ID can be re-run outside of ATLAS framework

Multivariate Analysis with TMVA in PanTau

Separate training in categories of estimated decay mode (1-prong, 3-prong, ...) and energy bins

- Splitting helpful, because some variables only available for some categories (e.g. need certain numbers of charged particles) or distributions are very different for different decay modes
- However, very different purity in the categories at the beginning
- Previously used multiple instances of TMVA `Reader` for each class (not possible anymore in some recent TMVA releases due to bug)
- Instead using Category classifier now: Allows to train and evaluate sub-classifiers (projective likelihood, neural network, etc) for disjoint sub-sets of the input data

Combining different reco classes

TMVA output of likelihood classifier:

$$y = \frac{L_S}{L_S + L_B}$$

where

$$L_C = \prod_k p_{C\,k}(x_k), C = S, B \text{ and } \int_{-\infty}^{\infty} p_{C\,k}(x_k) dx_k = 1$$

i.e. signal and background reference distributions are normalised individually. Remark: $\ln\left(\frac{L_S}{L_B}\right) = -\ln\left(\frac{1}{y} - 1\right) \equiv \text{logit}(y)$

Probability for being signal including the a priori probability of the signal class (according to Bayes):

$$p = \frac{aL_S}{aL_S + bL_B} = \frac{aL_S}{aL_S + (1-a)L_B}$$

where a is the a priori probability of the signal class and $b = 1 - a$ the a priori probability of background.

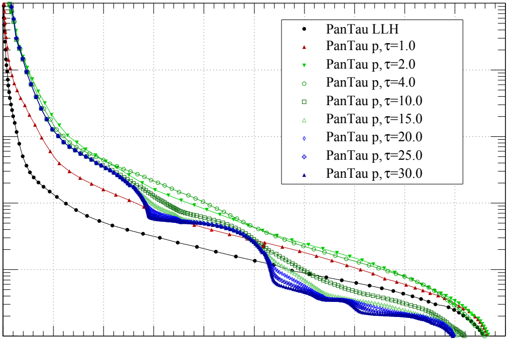
$$\text{logit}(p) = \ln\left(\frac{a}{1-a}\frac{L_S}{L_B}\right) = \ln\left(\frac{a}{1-a}\right) + \ln\left(\frac{L_S}{L_B}\right)$$

Combining different reconstructed categories
Results using the posterior probability

Introduce empirical factor τ to weight the a priori contribution a_i of the reconstructed category i ($i = $ 1-prong, 3-prong, ...)

Reminder: Very different purity in the categories before ID!

$$\mathrm{logit}(a_i) \mapsto \tau \cdot \mathrm{logit}(a_i): \quad \mathrm{logit}(p) = \tau \cdot \mathrm{logit}(a_i) + \ln\left(\frac{L_S}{L_B}\right)$$



Combining different reconstructed categories

Defining additional likelihood ratio

Alternative method to combine classifier output from different reconstructed categories i: Define new probability measure as a function of original likelihood ratio

$$t(y) = \frac{a_i S_i(y)}{a_i S_i(y) + b_i B_i(y)}$$

where $S_i(y)$ and $B_i(y)$ are the probability density functions of signal and background in category i as function of the classifier value y.

- Advantage wrt. previous method
  - Can be defined for any classifier
  - Takes different separation power between signal and background in the different categories into account, not just a priori probabilities
- Preliminary tests with binned PDFs showed further improvements wrt. previous method

Experiences using TMVA in ATLAS TauRec

TMVA also used in the default algorithm for tau ID in ATLAS (TauRec), e.g.

- Boosted Decision Trees
- Cut optimisation

- Good performance, continuous improvement
- TMVA implementation used for BDT training and preliminary performance evaluation
- Weights extracted from XML files into condensed binary file
- Own implementation within ATLAS software of BDT evaluation to reduce memory usage

- Problems:
  - Selected cuts unstable in nearby efficiency bins
  - Cut values not monotonic as function of efficiency, i.e. selected candidates for 30% efficiency may not be subset of candidates for 50% efficiency
- Finally "manual" solution was used for "robust" cut selection with only three variables

Summary

- PanTau is a new (still experimental) algorithm for tau identification in ATLAS based on a generic energy flow approach (eflowRec)
- Tau candidates are categorised according to estimated decay modes at an early stage
- TMVA is used in actual ID step (with category classifier)
- General meta-algorithm to re-combine classifier output of different categories would be useful and can boost overall performance
- TMVA also used in common algorithm for tau ID in ATLAS (TauRec)
  - Very good experiences with BDTs, though using own implementation of evaluation
  - Cut optimisation not suitable for all applications
- Outlook
  - Test new multi-class classifiers

Thanks

- Many thanks to the TMVA developers for their very useful toolkit!