WWW.OEAW.AC.AT

ÖAW AUSTRIAN ACADEMY OF SCIENCES

ALICE

Machine Learning Methods in the Analysis of Low-mass Dielectrons in ALICE

Sebastian Lehner and Aaron Capon**
on behalf of the ALICE collaboration
**Stefan Meyer Institute, Vienna, Austria*

One of the most promising probes to study heavy-ion collisions are dileptons ($\mu^+\mu^-$ and e^+e^-) since they reach the detector without significant final-state interactions. Therefore, the low-mass dielectron spectrum is of great interest for the study of the quark-gluon plasma (QGP).

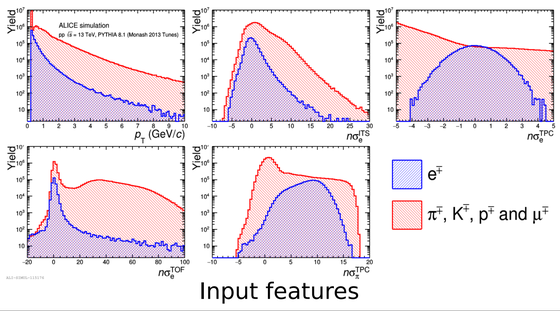
In order to precisely measure the low-mass dielectron spectrum a high purity sample of e^+e^- pairs is required. While traditional cut-based methods can provide high purity samples, they suffer from low efficiency. A multivariate analysis (MVA) for particle identification could in future be used to mitigate this drawback.

The main background in the analysis of dielectrons are combinatoric e^+e^- pairs. This background contribution can be suppressed by rejecting e^+ and e^- tracks that originate from photon conversions. Numerous features allow discrimination of background from signal dielectrons which motivates a multivariate approach in the classification of e^+e^- pairs.

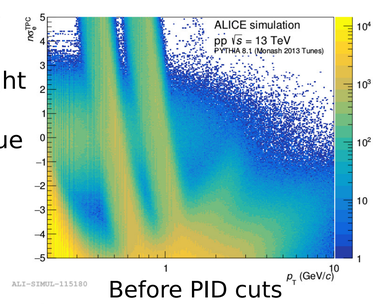
Cut-Based Particle Identification

Electrons are identified by using their specific energy loss within the Inner Tracking System (ITS) and Time Projection Chamber (TPC), as well as their time-of-flight (TOF), which are represented as the number of standard deviations, nσ, away from the expected value for a given species and p_T.

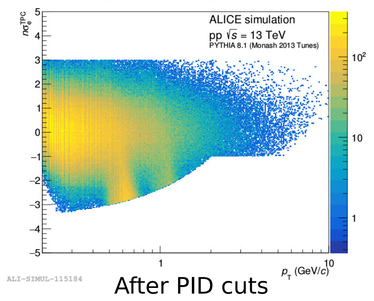
The standard approach then sequentially cuts on these values in order to remove the background.



Input features



Before PID cuts



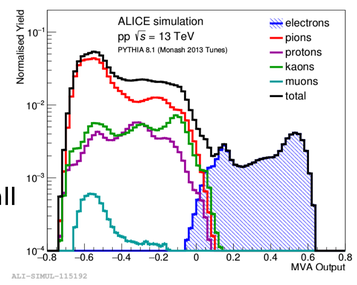
After PID cuts

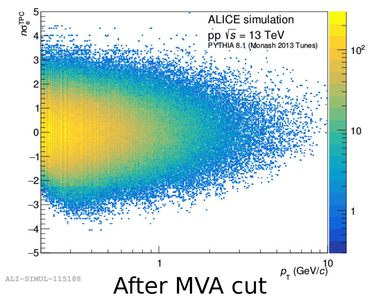
MVA Particle Identification

A boosted decision tree was trained with the same features used for the standard cuts.

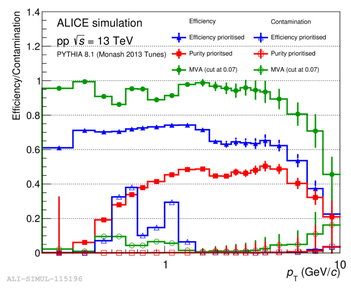
Cutting on the MVA output exploits the interplay of all detector features simultaneously.

This approach greatly improves the efficiency in all bins while also further supressing contamination from kaons and protons ($p_T \sim 0.4$ - 1 GeV/c).

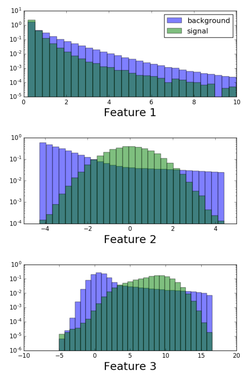




After MVA cut

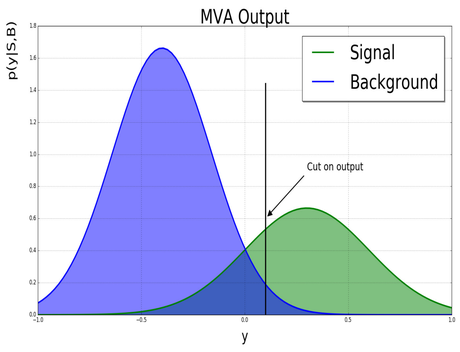


Multivariate Classification



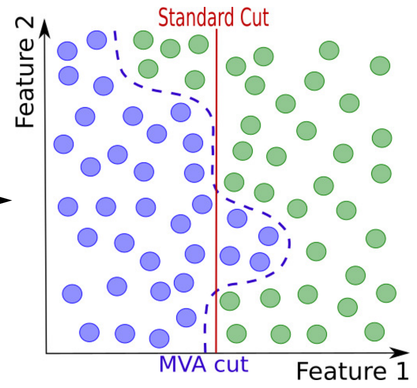
Aim: Find a mapping from a D-dim. input feature space to a 1-dim. output (y), such that signal and background have the least overlap.

The mapping is found in a training phase which utilises Monte Carlo (MC) data.



A single cut is then applied on y (MVA cut).

The cut, which corresponds to y = const, defines the decision boundary in the feature space.

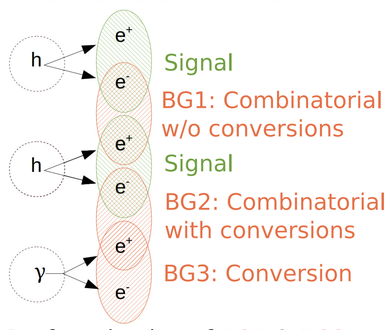


Signal & Background Dielectrons

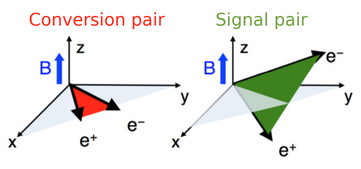
All unlike sign (US) tracks in the same event are paired.
Four classes of US pairs: 1 Signal & 3 Background (BG)

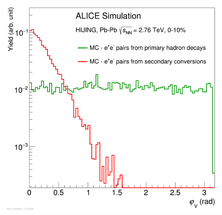
Background Reduction
Statistical significance = $\frac{S}{\sqrt{S+B}}$
Aim: Reduction of the dominant BG classes, i.e. combinatorial pairs with a conversion track (BG2) and conversions (BG3) while preserving the Signal

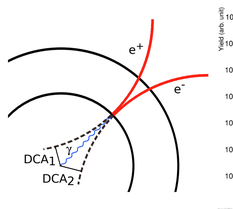
Features
20 features are used for BG rejection, e.g.:
φ_v for rejection of BG3: angle between normal vector of plane spanned by e^+e^- tracks and B field

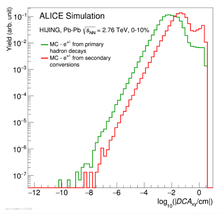


DCA_{xy} for rejection of BG1 & BG2: distance of closest approach of track to the primary vertex in the xy-plane









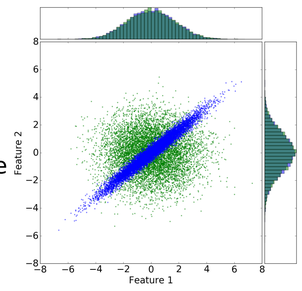
Multivariate MC-Data Adaption

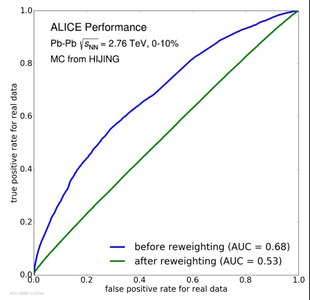
Decision boundary for S/B classification is determined on MC. Therefore, MC must accurately represent data.

Problem: How to check equivalence of multidimensional distributions? Comparing projections is not sufficienct!
→ Train multivariate classifier (e.g. GBDT) to separate MC and data. Classifier identifies regions in 20-dim. feature space that over- and under-populated in MC w.r.t. to data.

Problem: How to correct for found discrepancies?
→ Population differences in feature space are compensated by reweighting. Weighting factors are chosen such that corresponding multidimensional bins in MC and data have the same effective content n_i.
Weighting factor for bin i in MC, w_i, according to: $w_i = \frac{n_i^{data}}{n_i^{MC}}$.

MC is matched multivariately to data using all information on MC and data available. After reweighting MC and data are hardly distinguishable. False and true positive rates for classifying a pair as real data are almost the same.
(A. Rogozhnikov, arXiv:1608.05806 [physics.data-an])

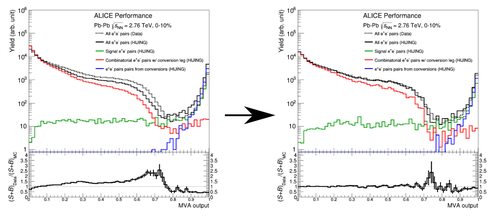




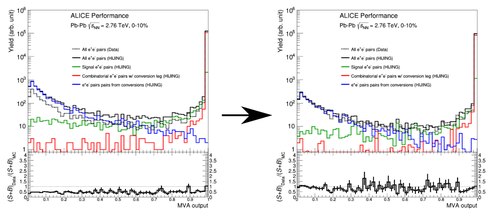
Classification with Neural Networks

Two neural networks are trained to identify background pairs:

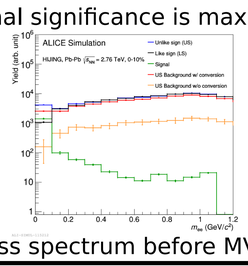
Combinatorial pairs with conversion (BG2)



Conversion pairs (BG3)

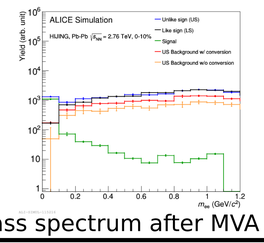


Weights from MC-data adaption are taken into account in training. After reweighting, the deciding quantity for classification, the MVA output, is in agreement between MC and data for both classifiers. Cuts on the MVA outputs of both classifiers are placed such that the gain in signal significance is maximised.



Mass spectrum before MVA cut

MVA cut →



Mass spectrum after MVA cut

Summary and Outlook

Particle Identification

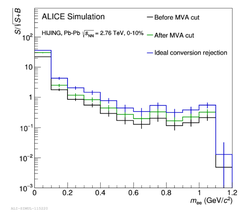
	Efficiency (%)	Purity (%)
Cut Method Efficiency Prioritised	70	91
Cut Method Purity Prioritised	13	99
Multivariate Method	95	96

Momentum integrated (0.2 < p_T < 10 GeV/c)

Dielectron Classification

	m_{ee} < 0.1 GeV/c^2	m_{ee} > 0.1 GeV/c^2
S Efficiency	0.77	0.73
B Efficiency	0.28	0.25

Significance gain



To complete these feasability studies a detailed investigation of the systematic uncertainties of the presented classification techniques will be carried out.

After finishing the evaluation of these methods, they will be employed in the analyses of low-mass dielectrons for pp, p-Pb, Pb-Pb Run 2 data.

∫dkΠ Doktoratskolleg Particles and Interactions

FTE NATIONALSTIFTUNG FORSCHUNG | TECHNOLOGIE | ENTWICKLUNG

SMI

SMI - STEFAN MEYER INSTITUTE FOR SUBATOMIC PHYSICS