BAYESIAN APPROACH FOR COMBINED PARTICLE IDENTIFICATION IN ALICE EXPERIMENT AT LHC

I. Belikov, P. Hristov, M. Ivanov, T. Kuhr, K. Šafařík, CERN, Geneva, Switzerland

Abstract

Implementation of a Particle Identification (PID) procedure in a Bayesian way is discussed in this report. The algorithm is capable of combining PID signals of quite different nature. All the necessary conditional probability density functions and *a priory* probabilities can be obtained from the analyzed data. The approach has been applied for identifying particles in the ALICE experiment at LHC. Its efficiency and contamination have been estimated using the ALICE offline simulation/reconstruction framework.

INTRODUCTION

Particle identification over a large momentum range and for many particle species is often one of the main design requirements of high energy physics experiments. The ALICE experiment at LHC [1] is able to identify particles with momenta from 0.1 GeV/c and, in some cases, up-to 10 GeV/c. This can be achieved by combining several detecting systems that are efficient in some narrower and complementary momentum sub-ranges. The situation is complicated by the amount of data to be processed (about 10^7 events with about 10^4 tracks in each). Thus, the particle identification procedure should satisfy the following requirements:

1. It should be as much as possible automatic.
2. It should be able to combine PID signals of different nature (*e.g.* dE/dx and time-of-flight measurements).
3. When several detectors contribute to the PID, the procedure must profit from this situation by providing an improved PID.
4. When only some detectors identify a particle, the signals from the other detectors must not affect the combined PID.
5. It should take into account the fact that, due to different event and track selection, the PID depends on the kind of analysis.

In this report we will demonstrate that combining PID signals in a Bayesian way satisfies all these requirements.

BAYESIAN PID WITH A SINGLE DETECTOR

Let $r(s|i)$ be a conditional probability density function to observe in some detector a PID signal s if a particle of $i-$type ($i = e, \mu, \pi, K, p, ...$) is detected. The probability to be a particle of $i-$type if the signal s is observed, $w(i|s)$, depends not only on $r(s|i)$, but also on how often this type of particles is registered in the considered experiment (*a priory* probability C_i to find this kind of particles in the detector). The corresponding relation is given by the Bayes' formula:

$$w(i|s) = \frac{r(s|i)C_i}{\sum_{k=e,\mu,\pi,...} r(s|k)C_k} \quad (1)$$

Under some reasonable conditions, C_i and $r(s|i)$ are not correlated so that one can rely on the following approximation:

- The functions $r(s|i)$ reflect only properties of the detector ("detector response functions") and do not depend on other external conditions like event and track selections.
- On contrary, the quantities C_i ("relative concentrations" of particles of i-type) do not depend on the detector properties, but do reflect the external conditions, selections *etc*.

The PID procedure is done in the following way. First, the detector response function is obtained. Second, a set of values $r(s|i)$ is assigned to each track. Third, the relative concentrations C_i of particle species are estimated for a subset of events and tracks selected in a specific physics analysis. Finally, an array of probabilities $w(i|s)$ is calculated (see Eq. 1) for each track within the selected subset.

The probabilities $w(i|s)$ are often called PID weights.

Obtaining the conditional probability density functions

The conditional probability density functions $r(s|i)$ (detector response functions) can be always parameterized with sufficient precision using available experimental data.

Let's consider, for example, the ALICE Time Projection Chamber (TPC) [2]. Currently, the ALICE reconstruction software uses the following parametrization. For each track reconstructed in the TPC $r(s|i)$ (s is the assigned dE/dx measurement) is Gaussian with the centroid $< dE/dx >$ given by the Bethe-Bloch formula and the width calculated as $\sigma = \kappa < dE/dx >$, where the coefficient κ is approximately constant over all the momentum region and for all the particle species. In case of simulated central HIJING [4] PbPb $\sqrt{s_{NN}} = 5.5$ TeV events, κ is about 0.07 (see Fig. 1).

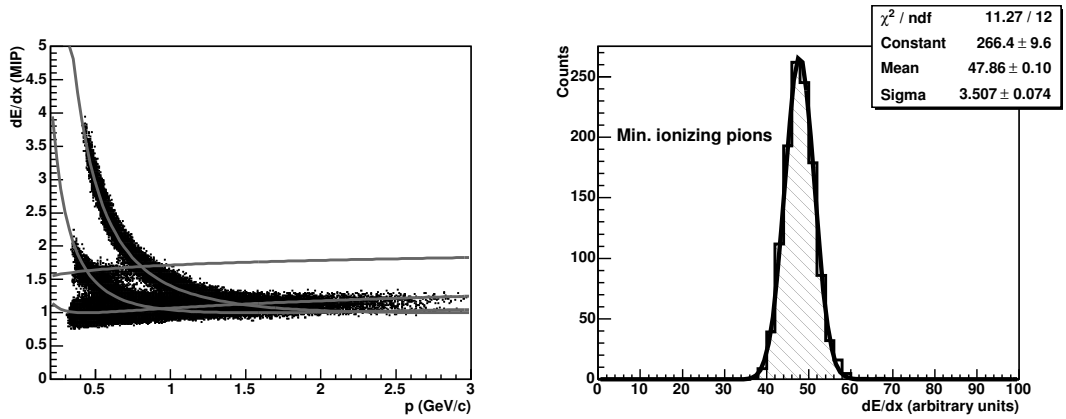

Figure 1: dE/dx response of the ALICE TPC (left) and its profile for minimum ionizing pions (right) for simulated central HIJING PbPb $\sqrt{s_{NN}} = 5.5$ TeV events.

Obtaining the a priory probabilities

In the simplest approach, the *a priori* probabilities C_i (relative concentrations of particles of i-type) to observe a particle of i-type can be assumed to be equal.

However, in many cases one can do better. For example in ALICE, when doing PID in the TPC for the tracks that are registered both in the TPC and in the Time-Of-Flight detector (TOF) [3], these probabilities can be estimated using the measured time-of-flight. One simply fills a histogram of the following quantity:

$$m = \frac{p}{\beta\gamma} = p\sqrt{\frac{c^2t^2}{l^2} - 1}, \qquad (2)$$

where p and l are the reconstructed track momentum and length and t is the measured time-of-flight. Such a histogram peaks near the values m that correspond to the masses of particles.

Under ALICE conditions, because the width of the peaks is mainly defined by the time resolution and is almost the same for all the particle types (see Fig. 2), the counts at the maxima of the histogram are proportional to the C_i. The absolute normalization of C_i is not important (see Eq. 1). Therefore, one can use straightaway $0 < C_e < 10$, $0 < C_\mu < 100$, $C_\pi \sim 2800$, $C_K \sim 350$ and $C_p \sim 250$ for the event and track selection shown in the Fig. 2.

Forcing some of the C_i to be exactly zeros excludes the corresponding particle type from the PID analysis and such particles will be redistributed over other particle classes (see Eq. 1). This can be useful for the kinds of analysis when, for the particles of a certain type, one is not concerned by the contamination but, at the same time, the efficiency of PID is of particular importance.

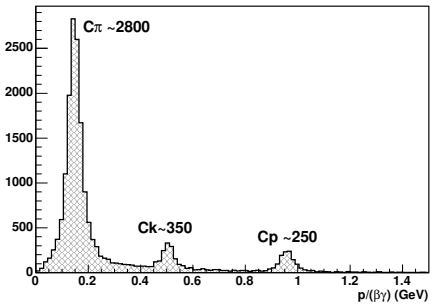

Figure 2: *A priory* probabilities C_i estimated using time-of-flight measurements (see the text).

PID COMBINED OVER SEVERAL DETECTORS

This method can be easily applied for combining PID measurements from several detectors. Considering the whole system of N contributing detectors as a single "super-detector" one can write the combined PID weights $W(i|\bar{s})$ in the form similar to that given by Eq. 1 :

$$W(i|\bar{s}) = \frac{R(\bar{s}|i)C_i}{\sum_{k=e,\mu,\pi,...} R(\bar{s}|k)C_k}, \qquad (3)$$

where $\bar{s} = s_1, s_2, ..., s_N$ is a vector of PID signals registered in the first, second and other contributing detectors, C_i are the *a priory* probabilities to be a particle of the i-type (the same as in Eq. 1) and $R(\bar{s}|i)$ are the combined response functions of the whole system of detectors.

If the single detector PID measurements s_j are uncorrelated (which is approximately true in the case of the ALICE experiment), the combined response function is product of single response functions $r(s_j|i)$ (the ones in Eq. 1) :

$$R(\bar{s}|i) = \prod_{j=1}^{N} r(s_j|i). \qquad (4)$$

One obtains the following expression for the PID weights combined over the whole system of detectors:

$$W(i|s_1, s_2, ..., s_N) = \frac{C_i \prod_{j=1}^{N} r(s_j|i)}{\sum_{k=e,\mu,\pi,...} C_k \prod_{j=1}^{N} r(s_j|k)} \quad (5)$$

In the program code, the combined response functions $R(\bar{s}|i)$ do not necessarily have to be treated as analytical. They can be "procedures" (C++ functions, for example). Also, some additional effects like probabilities to obtain a mis-measurement (mis-matching) in one or several contributing detectors can be accounted for.

The formula Eq. 5 has the following useful features:

- If for a certain particle momentum one (or several) of the detectors is not able to identify the particle type (*i.e.* $r(s|i)$ are equal for all $i = e, \mu, ...$), the contribution of such a detector cancels out from the formula.
- When several detectors are capable of separating the particle types, their contributions are accumulated with proper weights, thus providing an improved combined particle identification.
- Since the single detector response functions $r(s|i)$ can be obtained in advance at the calibration step and the combined response can be approximated by Eq. 4, a part of PID (calculation of the $R(\bar{s}|i)$) can be done track-by-track "once and forever" by the reconstruction software and the results can be stored in the Event Summary Data. The final PID decision, being dependent via the *a priory* probabilities C_i on the event and track selections, is then postponed until the physics analysis of the data.

RESULTS

Let's define the efficiency of the PID as N_{corr}/N_{true} and the contamination as $N_{incorr}/(N_{incorr} + N_{corr})$, where N_{corr} is the number of correctly identified, N_{incorr} number of mis-identified particles and N_{true} is the true number of particles of a certain type in the PID procedure. These efficiencies and contaminations were estimated using the ALICE simulation/reconstruction framework (ALIROOT [5]) for central HIJING PbPb $\sqrt{s_{NN}} = 5.5$ TeV events.

The results of identifying charged kaons using the ALICE Inner Tracking System (ITS) [6], TPC and the TOF as stand-alone detectors (see Eq. 1) and the result for the combined PID (see Eq. 5) are shown in Fig. 3. Only tracks reconstructed simultaneously in all the detectors were selected for the analysis, and the set of *a priory* probabilities was C_e=0, C_μ=0, C_π=0.70, C_K=0.15 and C_p=0.15.

One can see from this picture that

- the efficiency and the contamination of the combined PID are significantly weaker functions of the momentum than in the case of a single detector particle identification;

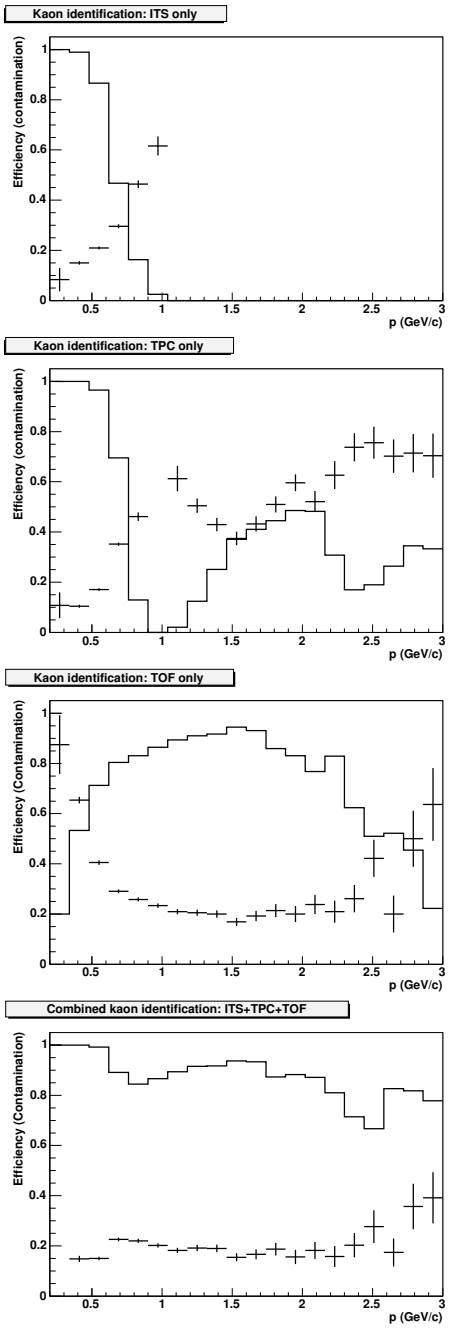

Figure 3: Single detector efficiencies (solid line) and contaminations (points with error bars) of the charged kaon identification with the ITS, TPC and TOF stand-alone and the combined efficiency and contamination using all the detectors working together.

- the efficiency of the combined result is always higher (or equal) than in the case of any of the detectors working stand-alone;
- the combined PID contamination is always lower (or equal) than the contaminations obtained with the single detector PID procedures.

Stability with respect to variations of the a priory probabilities

Since the results of this PID procedure explicitly depend on the choice of the *a priory* probabilities C_i (and, in fact, this kind of dependence is unavoidable in any case), the question of stability of the results with respect to the almost arbitrary choice of C_i becomes important.

Fortunately, in the momentum regions where the single detector response functions for different particle types of at least one of the detectors do not significantly overlap, the stability is guaranteed. The more detectors enter the combined PID procedure, the wider this momentum region becomes and the results are more stable.

Detailed simulations using the ALIROOT framework show that results of the PID combined over all the ALICE central detectors are, within a few per cent, stable with respect to variations of C_i up-to at least 3 GeV/c.

CONCLUSIONS

Particle identification in ALICE experiment at LHC can be done in a Bayesian way. The procedure consists of three parts:

- First, the single detector PID response functions $r(s|i)$ are obtained. This is done by the calibration software.
- Second, for each reconstructed track the combined PID response $R(\bar{s}|i)$ is calculated and effects of possible mis-measurements of the PID signals can be accounted for. The results are written to the Event Summary Data and, later, are used in all kinds of physics analysis of the data. This is a part of the reconstruction software.
- And finally, for each kind of physics analysis, after the corresponding event and track selection is done, the *a priory* probabilities C_i to be a particle of a certain i-type within the selected subset are estimated and the PID weights $W(i|\bar{s})$ are calculated by means of formula Eq. 5. This part of the PID procedure belongs to the analysis software.

The advantages of the described particle identification procedure are

- The fact that, due to different event and track selection, the PID depends on a particular kind of performed physics analysis is naturally taken into account.
- Capability to combine, in a common way, signals from detectors having quite different nature and shape of the PID response functions (silicon, gas, time-of-flight, transition radiation and Cerenkov detectors).
- No interactive multidimensional graphical cuts are involved. The procedure is fully automatic.

ACKNOWLEDGMENTS

We would like to thank Dr. S.Sedykh for the useful discussions and for his help in the preparation of this text.

REFERENCES

[1] ALICE tech. proposal CERN/LHCC 95-71.

[2] ALICE TPC TDR CERN/LHCC 2000-001.

[3] ALICE TOF TDR CERN/LHCC 2002-016.

[4] Comp.Phys.Comm., 83(1994)307.

[5] ALICE PPR v.1 CERN/LHCC 2003-049.

[6] ALICE ITS TDR CERN/LHCC 99-12.