ALICE Statistical Wish-list

Iouri Belikov for the ALICE collaboration
CERN, Geneva, Switzerland

Abstract

A few statistical problems faced by the event reconstruction in ALICE experiment at CERN are discussed in this paper. We outline several ad-hoc extensions of traditional Kalman-filter track finding which seem to increase the quality of tracks reconstructed in high multiplicity events anticipated for Pb–Pb collisions at LHC. These extensions, however, need a stricter formulation and justification from the theoretical side. The particle identification in ALICE is done by combining the information from different detecting systems using a Bayesian method. Having many clear advantages, this approach introduces into the analysis additional complications which are also discussed here.

1 Introduction

A Large Ion Collider Experiment (ALICE) [1] at CERN is a general-purpose heavy-ion experiment designed to study the physics of strongly interacting matter and the Quark-Gluon Plasma in nucleus-nucleus collisions at the LHC. In addition to heavy systems, the ALICE Collaboration will study collisions of lower-mass ions, in order to vary the energy density, and protons (both pp and pA), which primarily provide reference data for the nucleus–nucleus collisions. The pp data will also allow for a number of genuine pp physics studies.

The detector consists of a central part (see Fig. 1), which, event-by-event, measures hadrons, electrons and photons, and of a forward spectrometer to measure muons. The central part, which covers polar angles from $45°$ to $135°$ over the full azimuth, is embedded in the large L3 solenoidal magnet. It consists of an Inner Tracking System (ITS) of high-resolution silicon detectors; a cylindrical Time-Projection Chamber (TPC); three particle identification arrays, a Time-Of-Flight (TOF) detector, a Transition-Radiation Detector (TRD) and a single-arm ring imaging Cherenkov detector (HMPID) and a single-arm electromagnetic calorimeter (PHOS). The forward muon spectrometer (covering polar angles $180° - \theta = 2° - 9°$) consists of a complex arrangement of absorbers, a large dipole magnet, and fourteen planes of tracking and triggering chambers. Several smaller detectors for global event characterization and triggering are located at forward angles.

The detector is optimized for charged-particle density $\mathrm{d}N_{\mathrm{ch}}/\mathrm{d}y = 4000$ and its performance is checked in detailed simulations up to $\mathrm{d}N_{\mathrm{ch}}/\mathrm{d}y = 8000$. The track reconstruction efficiency in the acceptance of the TPC is about 80% down to transverse momentum of $p_{\mathrm{t}} \sim 0.2$ GeV/c and about 90% for tracks with $p_{\mathrm{t}} > 1$ GeV/c. It is limited only by the particle decays and small dead zones between the TPC sectors. Typical momentum resolution obtained with the magnetic field of 0.5 T is $\sim 1\%$ at $p_{\mathrm{t}} \sim 1$ GeV/c and $\sim 4\%$ at $p_{\mathrm{t}} \sim 100$ GeV/c. The secondary vertices can be reconstructed with the precision better then 100 μm.

The detector has excellent particle identification (PID) capabilities. From $p \sim 0.1$ GeV/c to a few GeV/c the charged particles are identified by combining the PID information provided by ITS, TPC, TRD, TOF and HMPID. Statistically, the charged particles can be identified up to a few tens GeV/c using the relativistic rise of $\mathrm{d}E/\mathrm{d}x$ in the TPC. Electrons above 1 GeV/c are identified by the TRD, and muons are registered by the muon spectrometer.

To achieve the benchmarks described above, the ALICE reconstruction has to cope with a few statistical problems. We will discuss some of them in this paper (the details can be found in Chapter 5 of the ALICE Physics Performance Report [2]).

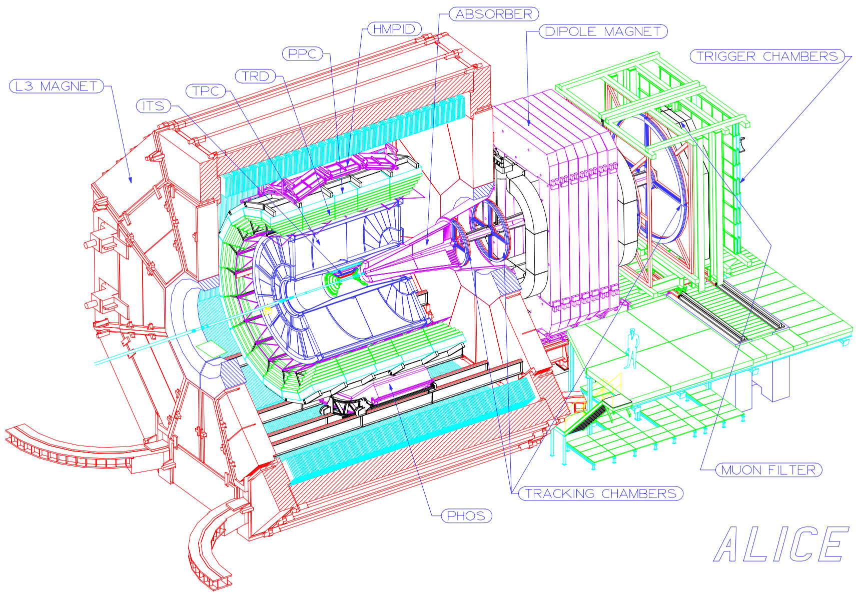

Fig. 1: Schematic layout of the ALICE detector.

2 Statistical problems with track finding in ITS

The track reconstruction in ALICE starts in the TPC, and then the tracks have to be prolonged in the ITS. This is difficult, because the distance between the inner wall of the TPC and the outer layer of the ITS is rather large and the track density inside the ITS is so high that there are always many ITS clusters found within the prolongation 'window' defined, mainly, by the multiple scattering in the material. The same often happens between the ITS layers as well (see Fig. 2). All this leads to a non-negligible probability of assigning to tracks many wrong clusters, if we use just the criterion of minimal χ^2 at each layer. Therefore we have to find the ways to improve the classical Kalman filter track-finding procedure [3].

For each event, we do two reconstruction passes over the set of clusters in the ITS: first, with a 'primary vertex constraint' (see below) and then without the constraint. In the both cases, we try to assign to a track, one by one, all the hits within the predicted window that have a χ^2 below a given limit, and not only the one with minimal χ^2. This way, for each track from TPC, we build a whole 'tree' of all possible prolongations in ITS. To speed up building the tree, the branches are sorted after each layer according to χ^2 and only a restricted number of acceptable branches are propagated further. Finally, we choose the most probable track candidate (i.e. the path along the tree) taking into account the quality of the whole path (sum of χ^2s at the layers, total number of assigned clusters and a few other criteria).

Because most of tracks are expected to be primary, the first reconstruction pass is done applying an ad-hoc 'primary vertex constraint' (see Fig. 3). Since the primary vertex in ALICE can be reconstructed in advance sufficiently well, the idea is to use this additional information during the track finding. When going over the clusters within the 'window', we take into account not only the positions of clusters and the track intersection point with the layer, but also the direction towards the primary vertex. Technically, this is done by extending the vector of measurement m

$$m^T = \{y, z\} \rightarrow \{y, z, \sin(\phi), \tan(\lambda)\},$$

where $\{y, z\}$ are the coordinates of the cluster position and the angles $\{\phi, \lambda\}$ define the direction to the primary vertex and are calculated using the current value of the track curvature. The elements of the covariance matrix of the extended measurement vector that correspond to the two angles are evaluated considering the material which this track would cross on its way to the primary vertex. The subsequent evaluation of the χ^2 and update of the track parameters become thus 4-dimensional problem.

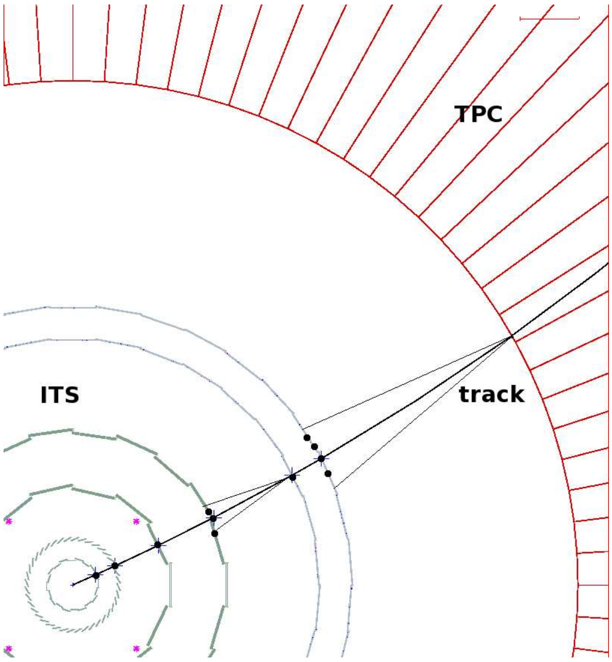

Fig. 2: The problem of track finding in ITS in high multiplicity events: Several clusters are found within the prolongation 'window' from one layer to another.

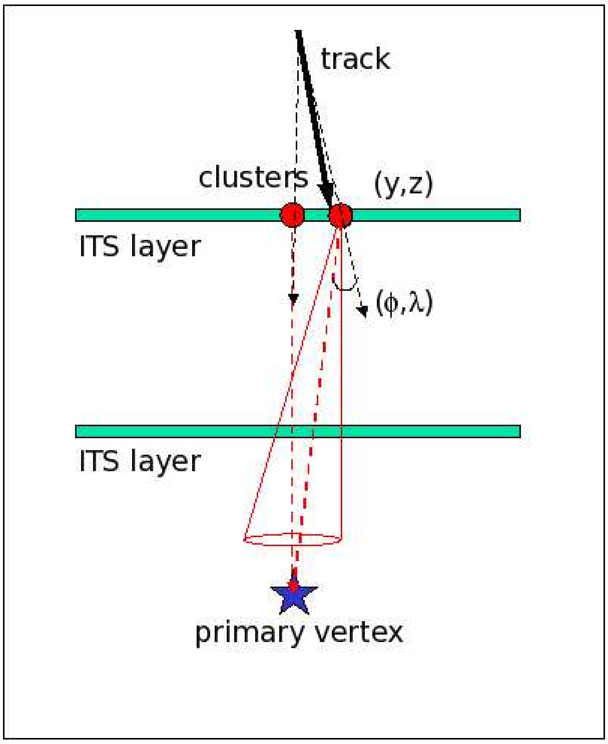

Fig. 3: Taking into account the information about the primary vertex position by applying a 'vertex constraint' (see the text).

Detailed Monte-Carlo studies performed with ALICE offline simulation and reconstruction framework AliRoot [1] show that the outlined ad-hoc 'vertex constraint' significantly reduces the probability of wrong cluster assignment, and so the quality of reconstructed tracks improves. Unfortunately, the procedure is not free of flaws. For example, for each of the tracks, it uses several times (even though with different 'weights') the same information about the primary vertex position. This is done as many times per a track as there are detector layers. Consequently, one of the undesirable features is that the resulting covariance matrix of the track parameters becomes underestimated (which can be overcome by an additional refitting step, however).

In future, we would like to incorporate the vertex constraint into the Kalman-filter track finding in a stricter way. A possible solution can probably be found by introducing the information about the primary vertex position in the form of Bayesian priors.

3 Statistical problems with particle identification

The ALICE experiment is able to identify particles with momenta from 0.1 GeV/c and up-to a few tens GeV/c (statistically, on the relativistic rise of $\mathrm{d}E/\mathrm{d}x$ in TPC). 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 (PID) 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.* $\mathrm{d}E/\mathrm{d}x$ and time-of-flight measurements).
3. When several detectors contribute to the PID, as it is shown in Fig. 4, the procedure must profit from this situation by providing an improved PID.
4. When some of the detectors can not separate the particle species, 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.

The method described here is similar to that in Ref. [4]. Let $r(s|i)$ be a conditional probability density function to observe in some detector a PID signal s if a particle of type i ($i = \mathrm{e}, \mu, \pi, \mathrm{K}, \mathrm{p}, ...$) is detected. The probability to be a particle of type i 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 experiment (*a priori* probabilities C_i to find a particle of i-type in the detector). The corresponding relation is given by Bayes's formula:

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

If C_i and $r(s|i)$ are not strongly correlated we 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 the contrary, the quantities C_i ('relative concentrations' of particles of type i) do not depend on the detector properties, but reflect the external conditions, selections etc.

In the case of several detectors, the signal s is replaced by a vector of the PID measurements $\bar{s}$ in the detectors. The response function $r(s|i)$ becomes some 'combined response function' $R(\bar{s}|i)$ of the whole system of the detectors involved (in the simplest case, this is the product of the single-detector PID response functions). The PID procedure is then done in steps:

– First, the detector response functions are obtained (theoretically, or in beam tests). This can be done ‘once and forever’ before the reconstruction starts as a part of detector calibration.
– Second, for each track, a value $R(\bar{s}|i)$ is calculated using the PID signals measured for this track. This is done during the event reconstruction.
– Third, the relative concentrations of particle species C_i are estimated for the subset of events and tracks selected for a specific physics analysis. For obtaining better results, the particle concentrations C_i can be considered as functions of momentum.
– Finally, for each track within the selected subset, the array of probabilities $w(i|\bar{s})$ is calculated using the formula (1). This steps, as well as the previous one, can be done only during the physics analysis of the data.

Doing the particle identification in this way, we naturally satisfy all the requirements mentioned at the beginning of this paragraph. However there are two problems which we are still working on.

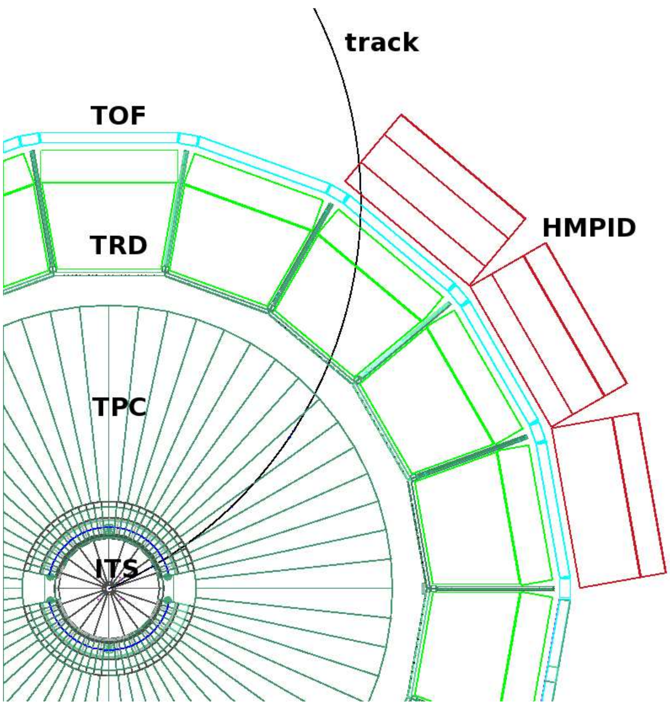

Fig. 4: The particle identification for the shown track is done by combining PID signals from five detectors: ITS, TPC, TRD, TOF and HMPID.

Since the results of such a PID procedure depend explicitly on the choice of *a priori* probabilities C_i (and, in fact, this kind of dependence is unavoidable in any approach), the question of stability of the results with respect to the choice of C_i becomes important. This problem seems to be related to the ‘Punzi effect’ discussed in Ref. [5]. At lower momenta, there is always some momentum region where the single-detector response functions for different particle types of at least one of the detectors do not significantly overlap, and so the stability is guaranteed. The final PID weights $w(i|\bar{s})$ are defined by the detector response functions. The more detectors enter the combined PID procedure, the wider this momentum region becomes and the results are more stable. But, finally, as the momentum goes up, all the detectors lose separation power, and the PID decision is given by the bare priors C_i (which we can not in this case estimate independently). The question is: can we somehow quantify the ‘contribution of priors’ to the final PID weights so that if they become dominant, relative to the ‘contribution of detector responses’, we know when to stop trying to identify particles of higher momenta?

The second problem is of a different nature. The formula (1) fundamentally assumes that all the components of vector $\bar{s}$ are the results of PID measurements done for *the same* particle. In other words, the procedure of assigning clusters to tracks has to be ideal, which is not the case in reality. The consequences are seen, for example, in Fig. 5, where the PID efficiency and contamination in ALICE TOF detector are shown. In spite of the fact that separation of the particle species by the time of flight method improves greatly with decreasing momentum (Fig. 5, upper pad), the actual situation with the PID becomes worse, especially for the particles below 0.5 GeV/c (Fig. 5, lower pad). This is because the low-momentum particles decay, suffer from scattering in material and so have a higher probability of being assigned to the wrong cluster in the TOF detector. This mismatching effect is not taken into account by the formula (1), and so the combined PID result becomes biased at low momenta.

The effect of mismatching can be corrected by excluding from the vector $\bar{s}$ the components that deviate too much from 'reasonable expectation'. This is possible, for example, in the case of the ALICE TOF detector, because we calculate the expected time of flight during the track finding in ITS, TPC and TRD. However, in general case, we may not know what the 'reasonable expectation' is. Also, applying sharp cuts in an otherwise smooth procedure based on formula (1) may cause additional difficulties with finding the best values for the cuts. Thus, a better solution for the problem of dealing with the mismatching is still to be found.

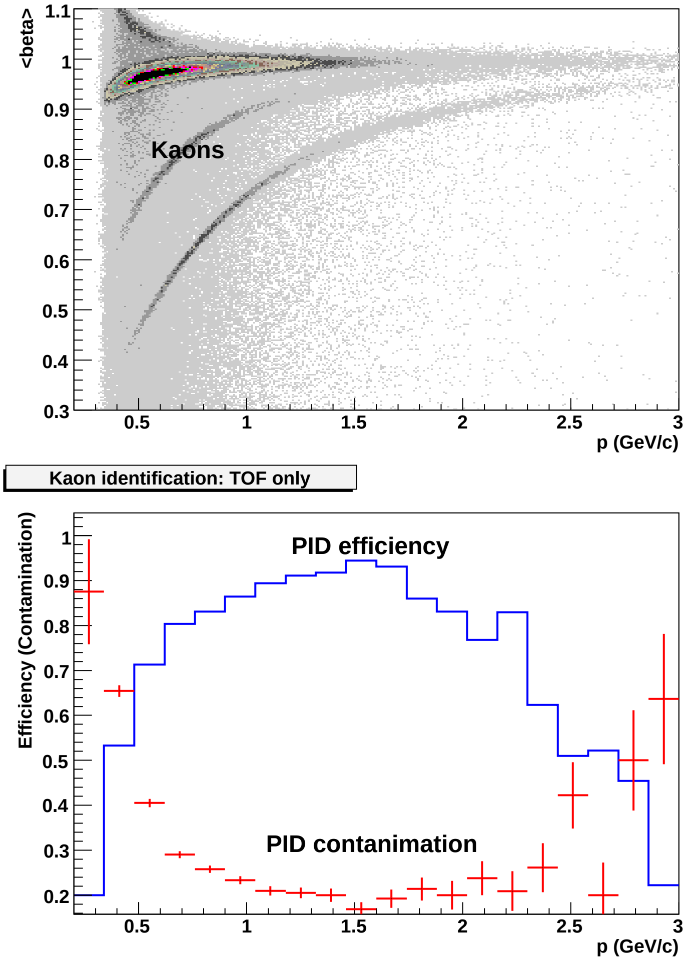

Fig. 5: PID efficiency and contamination for charged kaon identification with ALICE TOF detector. The deterioration of PID at low momenta is the consequence of the mismatching effect (see the text).

4 The wish-list

The statistical problems arising in event reconstruction in ALICE and discussed above represent the ALICE reconstruction statistical wish-list. In short, we would like to find better, theory supported, ways for

1. introducing constraints in the standard Kalman filter (needed for improving the track finding in high-multiplicity events in the ALICE ITS);
2. quantifying the relative importance of prior information and the results of actual measurements when making Bayesian decision (needed to define the highest momentum up-to which the Bayesian particle identification still makes sense);
3. taking into account mismeasurements in Bayesian combination of information (needed for improving the low-momentum particle identification in ALICE).

References

[1] F. Carminati *et al.* [ALICE Collaboration], J. Phys. G **30**, (2004) 1517.

[2] B. Alessandro *et al.* [ALICE Collaboration], J. Phys. G **32** (2006) 1295.

[3] P. Billoir, Nucl. Instrum. Meth. A **225** (1984) 352.

[4] M. Aguilar-Benitez *et al.*, Z. Phys. C **50** (1991) 405.

[5] P. Catastini and G. Punzi, *PHYSTATO5: Statistical Problems in Particle Physics, Astrophysics and Cosmology, Oxford, England, United Kingdom, 12-15 Sep 2005.*