
Charged Particle Identification

Shohei Nishida (KEK),
Alessandro Gaz (SLAC)

5th PBF Workshop @ KEK
Nov. 21, 2011

Charged Particle Identification

- No update since June.
- We basically consider that the present version is close to final, though we might need more polishing.
 - ✓ SN might add a few plots that show Belle PID performance.
 - ✓ Any comments are welcome.

5.2 PID

Editors:
Alessandro Gaz (BABAR)
Shohei Nishida (Belle)

5.2.1 Introduction

In this section the implementation and the performance of particle identification (PID) at Belle and BABAR will be presented.

After a brief introduction, the algorithms and statistical tools used by the two experiments will be presented. Some examples on the typical performance of the PID factors will be given, along with some discussion on the estimation of systematic uncertainties.

5.2.1.1 Definitions

The performance of a PID selector dedicated to the identification of charged particles of type α is characterized in terms of efficiency and mis-identification probabilities.

The first is computed as the fraction of successfully identified α tracks among all the α tracks reconstructed by the detector, while the second are the probabilities that particles of type β, γ, \dots are incorrectly identified as α .

In many cases the quantities defined above depend on the momentum, on the polar and azimuthal angles of the tracks, so the performance of PID selectors is studied and presented in bins of (p, θ, ϕ) .

5.2.1.2 Detector and physics sources of PID

BABAR uses as input for the PID selectors the information from all its subdetectors. Independent measurements of the dE/dx of a charged track are provided by the SVT and the DCH. The number of Cherenkov photons and the measurement of the aperture angle are provided by the DIRC, while the EMC is responsible for the measurement of the deposited energy and of the quantities (as the lateral and the Zemlin moments) related to the shape of the electromagnetic and hadronic showers associated to a charged track. Finally most information relevant to the identification of muons is provided by the IFR.

Belle also use similar input informations. dE/dx from CDC, time of flight information from TOF and the number of Cherenkov photons at the ACC are used mainly for hadron identifications. Informations from ECL and KLM are also necessary for electron and muon identification, respectively.

5.2.2 PID algorithms and multivariate methods

In the early phases of the experiment, PID selectors were mostly based on cuts applied to the most relevant variables for every particle type (e.g. E/p for electrons, the distance travelled in the return yoke for muons, the Cherenkov angle for the K/π separation, ...). A better performance is obtained with the use of likelihood based selectors, in which the information for the different subdetectors is used to compute the likelihood L_α for the candidate particle of belonging to type α :

$$L_\alpha = \prod_i P_i(x_i; \alpha), \quad (21)$$

where $P_i(x_i; \alpha)$ is the probability density function for particles α in the input variable x_i . The observables used for the construction of the likelihood are assumed to be uncorrelated. The individual likelihoods for the particle types among which we want to discriminate are then combined in likelihood ratios, and the selection is performed applying cuts of different tightness on the likelihood ratios.

Cut and likelihood based selectors are very robust and do not need re-tuning to compensate for the aging of the detector and the changes introduced by the reprocessing of the data. However, significant improvement can be achieved by considering a larger set of variables, even some with very mild discrimination power, in the implementation of PID selectors. Given the large number of input variables (up to 36) and the significant correlations among those, the use of more sophisticated statistical tools, such as Neural Networks (NN), *Begged Decision Trees* (BDT), and *Error Correcting Output Code* (ECOC) algorithms becomes mandatory.

Due to their higher sensitivity, the selectors based on multivariate methods need to be trained on high purity data control samples (see Section 5.2.3) after every major change in the reconstruction algorithms. Particularly important for BABAR, which was affected by large variations in the performance of the IFR, is the possibility to consider the data taking period as one of the input variables, in order to compensate for the loss of efficiency of specific regions of the detector.

In the following sections the algorithms which give the ultimate performance at Belle and BABAR will be described.

5.2.2.1 Belle algorithms

The PID at Belle is based on the likelihood method. Belle have different PID packages for hadron- and lepton-identification. The basic PID package called `atc_pid` combines the information of dE/dx , time of flight and response of ACC. This packages can provide PID information for any charged particles, but is mainly used for the separation among π, K and p . For electron identification, Belle has another package (`eid`) that combines the information from ECL and `atc_pid` to provide separation between electron and hadron. For muon, `mu2d` package calculates the package only from the information from KLM.

5.2.2.2 BABAR algorithms

In BABAR, the ultimate performance in the selection of muons is achieved with an algorithm based on *Begged Decision Trees* (Narsky (2005a)). The algorithm takes as input 30 variables: besides the variables related to the length and the shape of the IFR cluster associated to the candidate track and the measurement of the energy deposited in the EMC, also the variables related to the shape of the cluster in the calorimeter, the number of Cherenkov photons and the aperture angle of the Cherenkov cone, and the number of DCH hits and the dE/dx measured in the DCH are used.

The training of the selectors is performed on high purity data samples of muons and pions, subdivided in 720 bins of p, θ , and charge. Candidate tracks are randomly discarded in order to have the same number of muons and pions in the same bin. This allows us to use the p, θ , and charge variables in the tree without introducing any bias due to the different (p, θ) spectrum of the source sample. The source sample is then randomly split into a training and a testing sample. Four different levels of tightness are designed for the muon selector (VeryLoose, Loose, Tight, and VeryTight); the cuts on the output of the classifier are designed such that either the muon selection efficiency is kept constant at 90%, 80%, 70%, and 60% respectively across the whole momentum spectrum and for the all main data taking periods, or the pion mis-identification probability is 5%, 3%, 2%, and 1.2%. Two additional selectors, optimized for muons in the momentum range [3.3, 0.7] GeV/c, with a target efficiency of 70% and 60% have been developed. The BDT based muon selectors proved to be significantly more efficient than the selectors based on Neural Networks, as can be seen from Table 5.2.2.2.

For all the other charged particles (electrons, pions, kaons, and protons), a class of selectors based on the *Error Correcting Output Code* algorithms (Dietterich and G. (1995)), are used. The discrimination is based on 36 variables from the four inner subdetectors: SVT, DCH, DIRC, and EMC. Candidate e, π, K , and p are separated by means of several binary classifiers (in our case BDT's) combined through an exhaustive matrix. This ensures the robustness of this type of selectors against potential mis-classifications of some of the binary classifiers. The selectors are trained on high purity data samples (see Section 5.2.3) and the cuts on the outputs of the binary classifiers are tuned in such a way that the selection efficiency matches the one of the analogous likelihood based selectors. Six levels of tightness are provided (SuperLoose, VeryLoose, Loose, Tight, VeryTight, and SuperTight). At the same level of efficiency, the mis-identification rate for the ECOC algorithms is significantly lower than that of the likelihood based selectors (see Table 5.2.2.2).

5.2.3 Measuring PID performance in BABAR

The tuning of the PID selector and the assessment of their performance takes advantage of high purity samples of

Table 2: Average efficiencies and mis-identification rates for different families of muon and pion BABAR selectors. The quoted uncertainty represents the typical statistical uncertainty in each bin of the tables that measure the performance of the supported selectors. No systematic uncertainty has been included.

Muon selector	efficiency (%)	π mis-Id rate (%)
Cut based	95.0 \pm 0.5	1.48 \pm 0.04
Likelihood based	90.5 \pm 0.5	0.97 \pm 0.05
ECOC	59.4 \pm 0.5	0.76 \pm 0.05
Pion selector	efficiency (%)	π mis-Id rate (%)
Cut based	80.2 \pm 0.2	1.89 \pm 0.07
NN	83.0 \pm 0.2	1.47 \pm 0.07
BDT	84.2 \pm 0.2	1.10 \pm 0.07

tracks selected on the real data. A large number of electron and muon tracks is selected from $e^+e^- \rightarrow e^+e^-(\gamma), \mu^+\mu^-(\gamma)$ processes, with minimal cuts on the kinematics of the event and on the quality of both the candidate track and of the other track in the event. For some low-statistics cross-checks also a sample of electrons (muons) from the decays $B \rightarrow J/\psi K^{(*)}, J/\psi \rightarrow e^+e^-(\mu^+\mu^-)$, has been used.

K and π candidates are selected from $D^{*+} \rightarrow D^0\pi^+, D^0 \rightarrow K^-\pi^+$. The K/π assignment is done based on the charge of the soft pion from the D^{*+} decay. The purity of the sample is increased applying quality cuts on the reconstructed tracks and rejecting false D^0 's by cutting on the invariant mass of the reconstructed D^0 candidate and on the probability of the K and π tracks to originate from a common vertex. Additional π samples, especially important for measuring the mistagging of pions as muons at high momentum (where the statistics of the $D_2 \rightarrow K^-\pi^+$ is much lower) are obtained from $K_S^0 \rightarrow \pi^+\pi^-$ decays and from $e^+e^- \rightarrow \tau^+\tau^-$ events where one τ (tag) has one charged particle among its decay products and the other decays to a final state with three charged particles. Finally protons are selected from $\Lambda \rightarrow p\pi^-$ decays, enhancing the purity of the sample by applying cuts on the quality of the candidate tracks and on the probability that the proton and pion tracks are consistent with originating from the same vertex. Some examples of performance of the BABAR selectors are displayed in Table 5.2.2.2 and in Figure 22.

As said above, these high purity samples are utilized in the training of the more advanced PID algorithms and in establishing the performance of all the selectors. Depending on the available statistics, the control samples are divided into several bins with different (p, θ) . In the case of the muon selectors at BABAR, the samples are also subdivided in 6 bins of ϕ , to better characterize the degradation of the RPC chambers and the staged upgrade of the barrel section with LST detectors. Each of the selectors supported by the experiments is applied to every bin of the control samples and the efficiency for both the data (ϵ_{data}) and the simulation (ϵ_{MC}) is computed. The tables of efficiencies thus built are then used to correct the simulation so that its PID performance matches that of the

data. One of the most widely used algorithms to apply this correction in BABAR is the so-called *PID-tweaking*. In the case where $\epsilon_{data} = \epsilon_{MC}$, no correction is applied, whereas if $\epsilon_{data} < \epsilon_{MC}$ a MC track that passes the selector is randomly discarded with probability:

$$\frac{\epsilon_{data}}{\epsilon_{MC}} \quad (22)$$

In the case $\epsilon_{data} > \epsilon_{MC}$, a MC track that does not pass the selection is accepted with probability:

$$(\epsilon_{data} - \epsilon_{MC}) \frac{1}{\epsilon_{MC}} \quad (23)$$

At the end of the BABAR experiment, the size of the typical correction applied by the PID-tweaking algorithm was about the percent level.

5.2.4 Belle PID Performance and Systematic Measurements

In Belle, the PID performance of the hadron identification (`attc_pid`) is estimated using the decay $D^{*+} \rightarrow D^0 \pi^+$ followed by $D^0 \rightarrow K^+ \pi^-$, similar to BABAR. On the other hand, two-photon process $e^+e^- \rightarrow e^+e^-\ell^+\ell^-$ ($\ell = e, \mu$) is used to obtain high-statistic electron and muon samples for the study of the lepton identification. Since the above are low-multiplicity events, inclusive J/ψ events ($J/\psi \rightarrow \ell^+\ell^-$) are also used as an control sample, mainly for the estimation of the possible performance differences between low-multiplicity events and hadronic events.

In the study of the hadron identification, the control sample is divided into 384 bins, i.e. 32 momentum (p) bins and 12 polar angle (θ) bins. The division for p is in every 100 (200) MeV/ c step below (above) 3 GeV/ c , while the division for θ is based on the counter types of the ACC. In the study of the lepton identification, the control sample is divided into 70 bins (10 momentum bins in every 500 MeV/ c step and 7 polar angle bins).

For each bin, the efficiency and mis-identification rate for K and π is estimated both for the data and the MC for different PID selections. The relevant value for general analysis is the ratio of the efficiency or mis-identification rate between the data and the MC: $R_l = \epsilon_l^{data}/\epsilon_l^{MC}$ and its error δR_l , where l is the bin index. Therefore, R and δR are provided as a table for general use in Belle. The efficiency (mis-identification rate) ratio and its error, which is quoted as the systematic error from PID, for some physics analysis can be calculate by

$$R = \frac{1}{N} \sum_i n_i R_i \quad (24)$$

and

$$\delta R = \frac{1}{N} \left(\sqrt{\sum_i (n_i \delta R_i^{stat})^2} + \sum_i n_i \delta R_i^{syst} \right) + \delta R_{control} \quad (25)$$

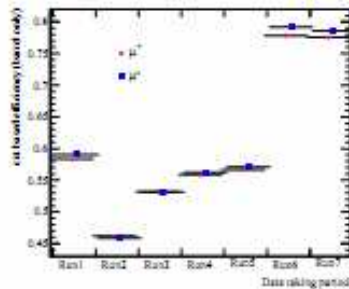


Fig. 23. Muon selection efficiencies for a typical BABAR cut-based muon selector as a function of the data taking period. The efficiencies are computed only for the barrel region. It is very evident the loss in performance due to the degradation of the RPC detector during the early phases of the data taking, and the full recovery with the installation of the LSTs, completed after the end of Run5.

where R_i is the efficiency ratio in a bin i , and n_i is the number of tracks in the bin i (analysis dependent), $N = \sum_i n_i$, R_i^{stat} , R_i^{syst} are respectively statistical and systematic error in bin i obtained from the control sample study, and $\delta R_{control}$ is the systematic error independent to (p, θ) . In this way, the correction factor and the systematic error can be automatically calculated for most of the analyses, unless the analysis needs requires care for PID. The tables for R and δR are also provided separately for positive and negative particles, and the systematic errors for the asymmetry can also be calculated.

5.2.4.1 History of PID performance in BABAR

For the BABAR experiment, the most important issue affecting the stability of PID performance was the degradation of the efficiency of the RPC chambers. This is very well visible from figure 23, which plots the efficiency of one of the cut-based muon selectors as a function of the data-taking period. This loss of performance was also one of the main motivations to develop muon selectors significantly relying on variables not measured by the IFR.

5.2.5 Systematic effects

As already said above, both experiments rely on high-purity data samples to assess the performance of PID selectors and correct the simulation so it matches the data as much as possible. Several ways exist to estimate the systematic uncertainty related to PID. In general it is not possible to establish a recommended way to proceed for all the analyses, since in general the performance of each

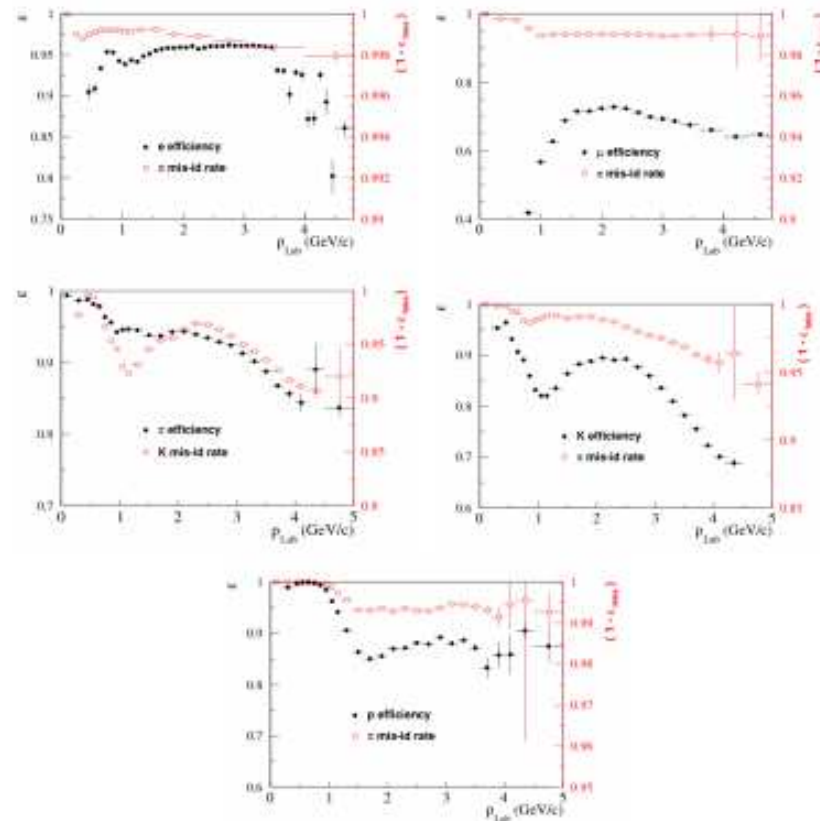


Fig. 22. Performance of some typical BABAR PID selectors for electrons (top left plot), muons (top right), pions (middle left), kaons (middle right), and protons (bottom) as a function of the momentum of the candidate charged track. The full black dots represent the efficiency, which can be read on the left axis, of the particular selectors, while the empty red squares show the complement of the pion (kaon for the pion selector) mis-identification probability (right axis).

selector can be sensitive to the charged and neutral multiplicity of the events studied. For example, the performance of electrons and muons selectors is studied in low multiplicity events, so some care must be taken when applying these selectors to B -decays, where the multiplicity of the final states is sizably higher.

In BABAR, many of the analyses estimate the systematic uncertainty on the PID performance by taking the difference of the signal reconstruction efficiency in the simulation obtained by applying or not applying the correc-

tion (usually the PID-tweaking) based on the tables. For some analyses where the contribution of the PID to the total systematic uncertainty is very relevant, or there is a sizable dependence on the multiplicities and the topologies of the events, alternative strategies have been applied, and where possible the performance of the chosen selector(s) has been checked in control samples with very similar multiplicities and topologies of the channel under study.

Reorganization of Chapter

Current organisation: -----

4. Vertexing
5. Multivariate discriminants
 - 5.1 Analysis optimization
 - 5.2 PID
 - 5.3 Flavor tagging
 - 5.4 Background discrimination
6. B-meson reconstruction
7. Mixing and time-dependent analysis

Proposed new organisation: -----

4. Multivariate methods and analysis optimization
5. Charged particle identification
6. Vertexing
7. B-meson reconstruction
8. B-flavor tagging
9. Background suppression for B-decays
10. Mixing and time-dependent analysis