Bayesian PID in AliRoot

I. Belikov

- **Introduction.**
  - The Bayesian philosophy.
  - PID with a single and several detectors.
- **Implementation in AliRoot.**
  - In reconstruction.
    - Mismatching and heavy particle species
  - In ESD/AOD.
- **A few results.**
- **Questions for additional thinking.**
  - Efficiency/contamination.
  - High-momentum limits.

PID task force meeting, CERN, 02 July 2009

What is the Bayesian approach ?

Example: a supervisor choosing a summer student

- What is the probability of choosing a girl or a boy ?
- The answer depends on:
  - Probability with which this supervisor chooses a girl $p(g)$, or a boy $p(b)$
  - The number of application submitted by girls Ng, and by boys Nb (the priors)
- The final probability is given by Bayes' formula:

$$P_g = \frac{p(g)Ng}{p(g)Ng + p(b)Nb} = \frac{p(g)}{p(g) + p(b)[Nb/Ng]}$$

$$P_b = \frac{p(b)Nb}{p(g)Ng + p(b)Nb} = \frac{p(b)}{p(g)/[Nb/Ng] + p(b)}$$

Note: The result depends only on Nb/Ng,
which can be evaluated on a subset of the applications

What is so special about the Bayesian approach ?

- It puts together two quite different sources of information
  - The supervisor's preferences are property of the supervisor. They, probably, do not change with time, and so can be precalculated.
  - The numbers of applications by girls/boys is an example of the conditions that are completely external to the supervisor. These may change year-by-year, and so, fundamentally, cannot be calculated once and forever.

Bayesian approach in PID

Probability to be a particle of ***i***-type (***i* = *e*, *μ*, *π*, *K*, *p*, … **),
if the PID signal in the detector is ***s***:

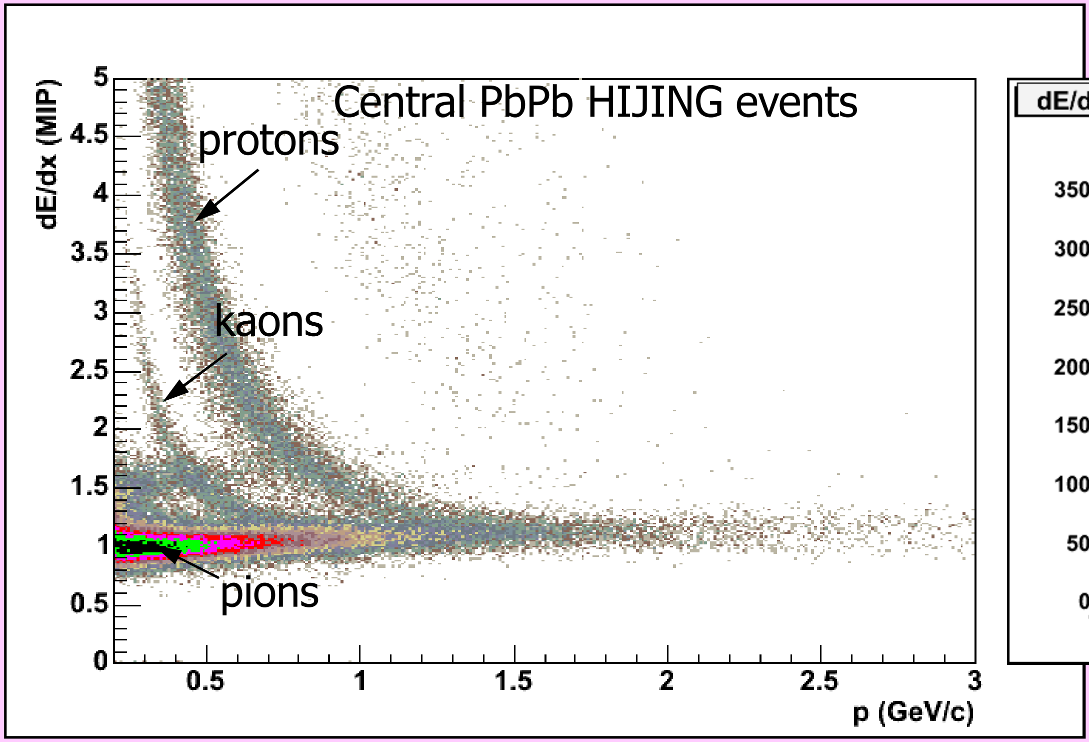
$$\boxed{w(i\,|\,s) = \frac{C_i r(s|i)}{\sum\limits_{k=e,\mu,\pi,\dots} C_k r(s\,|\,k)}}$$

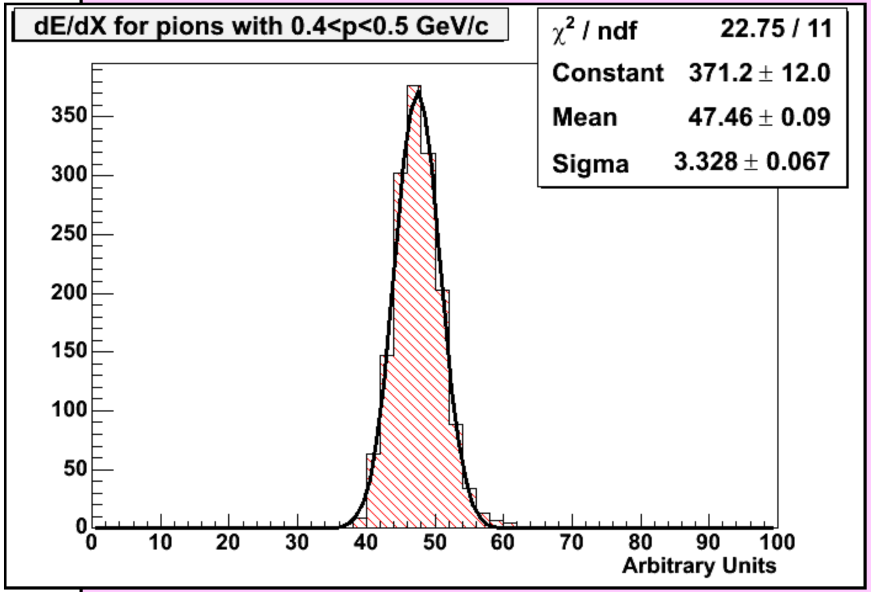
- $\boldsymbol{C_i}$ - *a priori* probabilities to be a particle of the ***i***-type.
  "Particle concentrations", that depend on the event and track selection.
- $\boldsymbol{r(s\,|\,i)}$ – conditional probability density functions to get the signal ***s***, if a particle of ***i***-type hits the detector.
  "Detector response functions", depend on properties of the detector.

In the case of N contributing detectors: $r(s_1,\dots,s_N\,|\,i) \sim \prod\limits_{d=1}^{N} r(s_d\,|\,i)$

Obtaining the conditional PDFs

Example: "TPC response function"





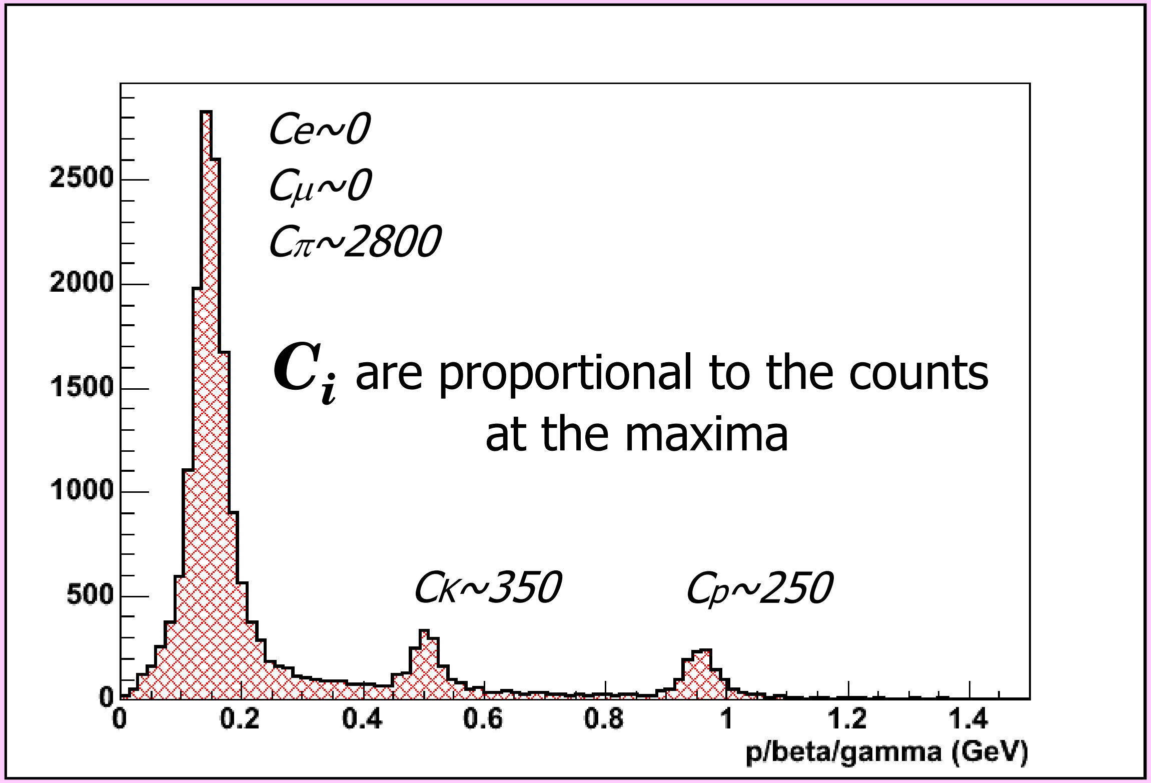
For each momentum p the function $r(s\,|\,i)$ is a Gaussian with

- centroid $<dE/dx>$ given by the Bethe-Bloch formula and
- sigma $\sigma = 0.08<dE/dx>$

This is a property of the detector (TPC). Can be prepared in advance !

Obtaining the *a priori* probabilities

("particle concentrations")



β by TOF, p by TPC

1. Sometimes, we may know the priors (V0s, cascades)
2. Sometimes, we can get the priors by iterating over the data
3. Anytime, we can use the raw PID signals
   - Simple histograming
   - Complicated fits…

The "particle concentrations" depend on the event and track selection.
They cannot be prepared once and for all kinds of analysis !

So, what are the advantages ?

- The explicit factorization of what can be done in reconstruction ("supervisor's preferences"), and what has to be done in physics analysis ("number of applications"). Already at the level of a single detector.
- Computational convenience.
  - when combining the PID information over the contributing detectors.
  - Less parameters to fit during the analysis ("amplitudes" only, because the "centroids and sigmas" are already precalculated during the reconstruction.
- Gain in the disk space at the level of AOD.
  - Because the priors can be estimated on a subset of tracks, the raw PID signals can be stored for this subset only.

The three parts of the PID procedure

- Calibration part, belongs to the calibration software.
  Obtaining the single detector response functions.
  Done by the detector experts.
- "Constant part", belongs to the reconstruction software.
  Calculating (for each track) the values of detector response functions, combining them and writing the result to the ESD.
  Done automatically, in massive reconstruction runs on the Grid.
- "Variable part", belongs to the analysis software.
  Estimating (for a subset of tracks selected for a particular analysis) the concentrations of particles of each type, calculating the final PID weights by means of Bayes' formula using these particle concentrations and the combined response stored in the ESD.
  Done by the physicists involved in this particular analysis.

Implementation in AliRoot

PID classes in reconstruction

- Every detector provides an Ali<*DET*>pidESD class.
  - This class provides the detector response functions, and a few other things.
  - This class has a function:

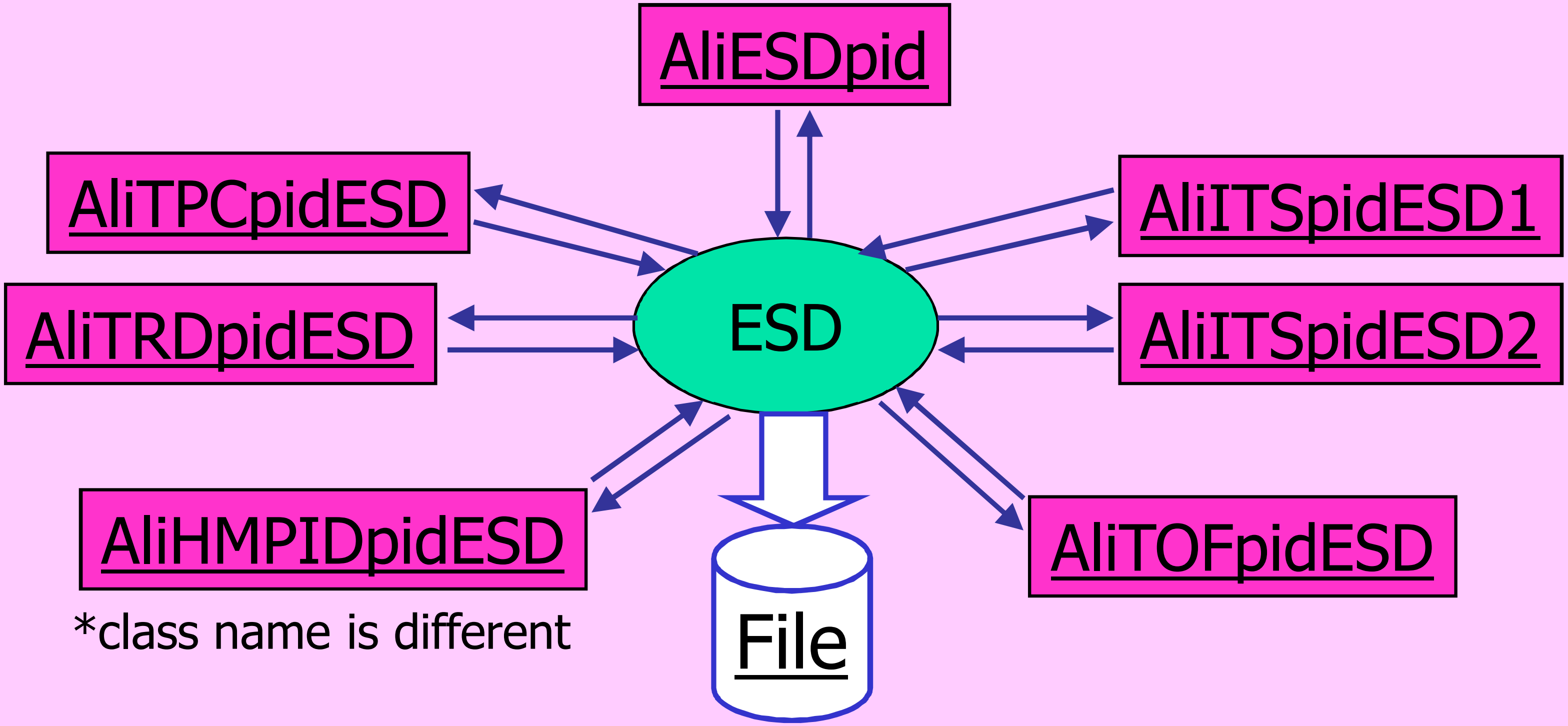
    Int_t Ali<*DET*>pidESD::MakePID(AliESDEvent *event);

    calculating the values of the response functions and puts them to ESD.
- There is an AliESDpid class in STEER/.
  - It also has a function:

    static Int_t AliESDpid::MakePID(AliESDEvent *event);

    resposible for the combining of the PID information from the detectors.
- There is also a helper AliPID class in STEER/.
  - Keeps the particle masses, symbolic constants etc.
  - Implements the Bayes' formula.

PID classes in reconstruction



*class name is different

- Since 2005 all the charged-PID detectors contribute to the combining.
- Everything works also starting from already existing ESDs.
  - There is a possibility to re-calculate the PID combining over a subset of the detectors.

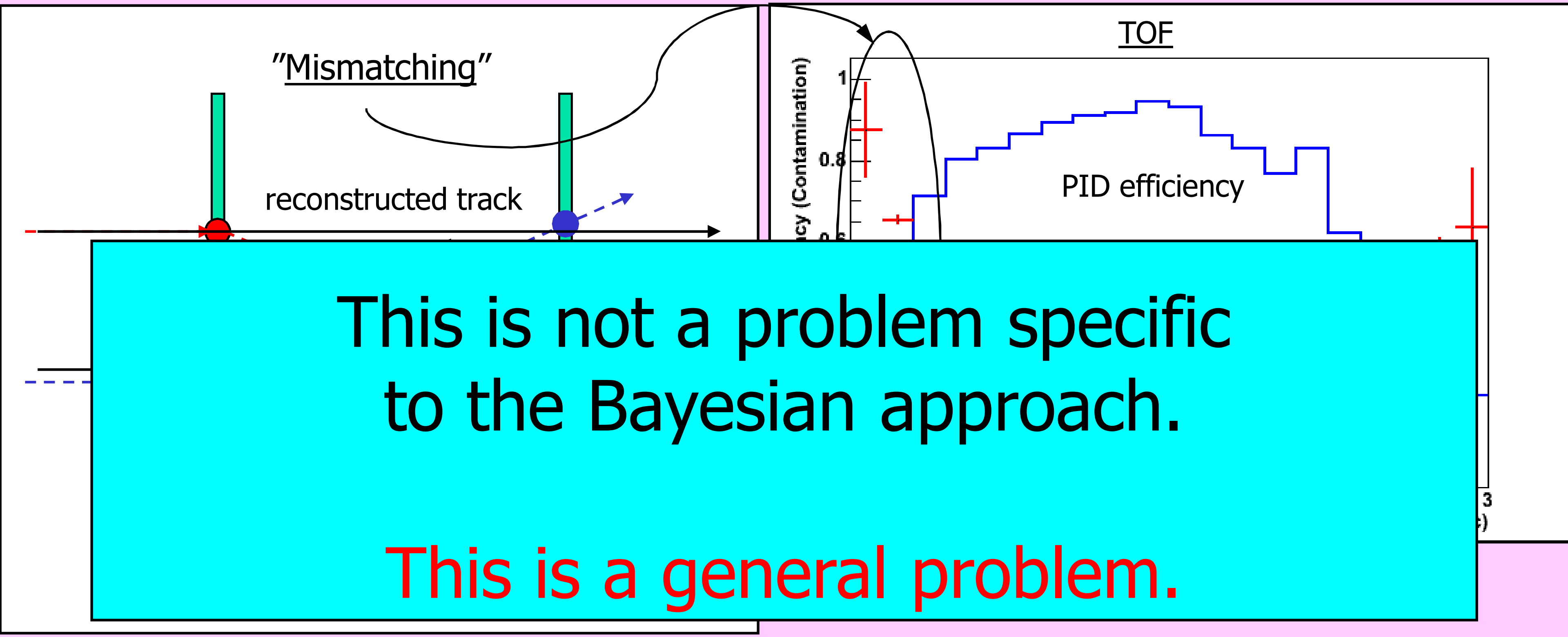
Implementation in ESD

- Every ESD track has
  - The raw PID signal reconstructed in all the contributing detectors.
    - Accessed by different getters.
  - An array of AliPID::kSpecies elements containing the $r(s|i)$ values for all the contributing detectors.
    - Accessed by Get<DET>pid(Double_t *) methods.
  - An array of AliPID::kSpecies elements containing the $r(s_1,\dots,s_N|i)$ values combined over a (sub)set of contributing detectors.
    - Accessed by GetESDpid(Double_t *) method.
  - A bit mask containing the bits corresponding to the contributing detectors
    - Checked by IsOn(AliESDtrack::k<DET>pid) methods.
- The AliTPCpidESD and AliTOFpidESD classes for n-sigma cut operations.
  - Because the Bayesian and n-sigma cut approaches share the same detector response functions.

Implementation in AOD

- Every AOD track has
  - An array of AliPID::kSpecies elements containing the $r(s_1,\ldots,s_N|i)$ values combined over a (sub)set of contributing detectors.
    - Accessed by GetPID(Double_t *) method (and a few others).
  - A bit mask containing the bits corresponding to the contributing detectors
    - Checked by IsOn(AliESDtrack::k<DET>pid) method.
- Every AOD track above a certain pt and a fraction of AOD tracks below this pt (parameter !).
  - A special AliAODPid object containing the raw PID signals from the contributing detectors.
    - Accessed by GetDetPid() method.

The problem of track mismatching

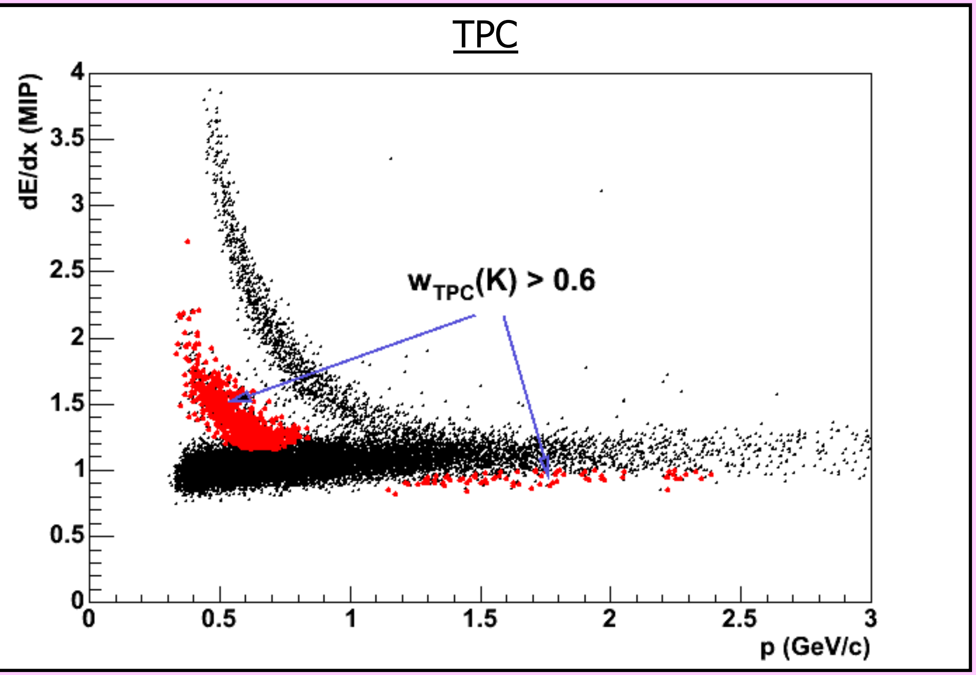


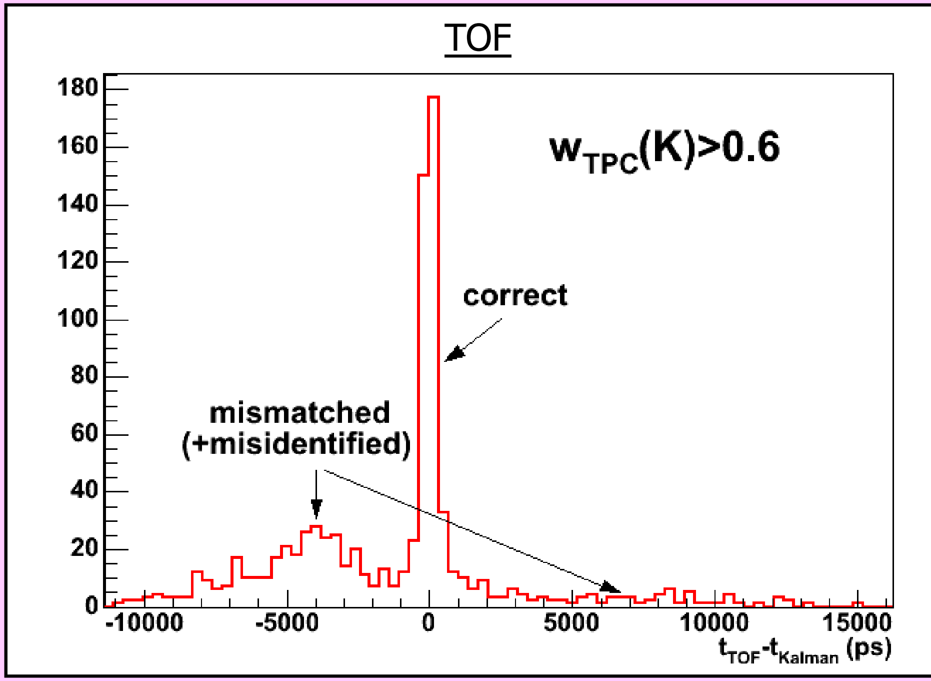
This is not a problem specific to the Bayesian approach.

This is a general problem.

The track mismatching biases the combining (any kind of !) the PID information, because the main assumption that all the detectors register the same particle, is not satisfied…

Ad-hoc treatment for the mismatching in TOF





Observing in one of the detectors the distribution of signals for a clean sample of particles pre-selected in other detectors, we can get the range of signals, where the probability of mismatching is "high" → Do not include into combinining

A question: Can it be somehow generalized ? Made "smooth" ? Optimized ?

P. Hristov's idea: $w = (1-p_{12})w_1 + p_{12}w_{12}$ (p_{12} - prob. of a correct matching)

Implementation of the mismatching and the heavy particle species

- If the **probability of mismatching** is higher than the probability of the correct matching, make the *r(s|i)* be equal (nice feature of Bayes' formula !).
  - Implemented in TOF as a ~1/(pβ) parameterization (one free parameter).

- If the PID signal is "**heavier than proton's**", do not calculate the *r(s|i)* and do not set the PID bit in the track status word
  (then, such a track is simply ignored by the PID procedure)
  - Implemented in ITS, TPC and TOF (one free parameter)

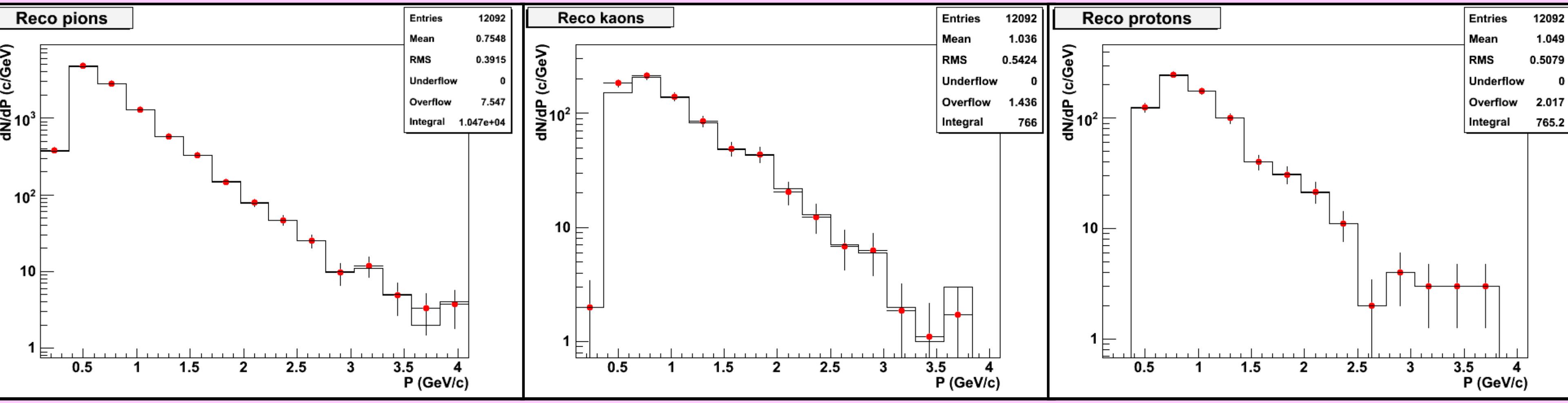
A complementary approach: n-sigma cuts

- Guarantees a definite and constant over the momentum efficiency. Does not deal with the priors
  - Does not tell anything about the contamination
  - Does not maximize the significance
- Everything that concerns the response functions is the same as for the Bayesian
- An important piece of software can (and must) be shared by the two approaches.

A few results

Identified particle spectra

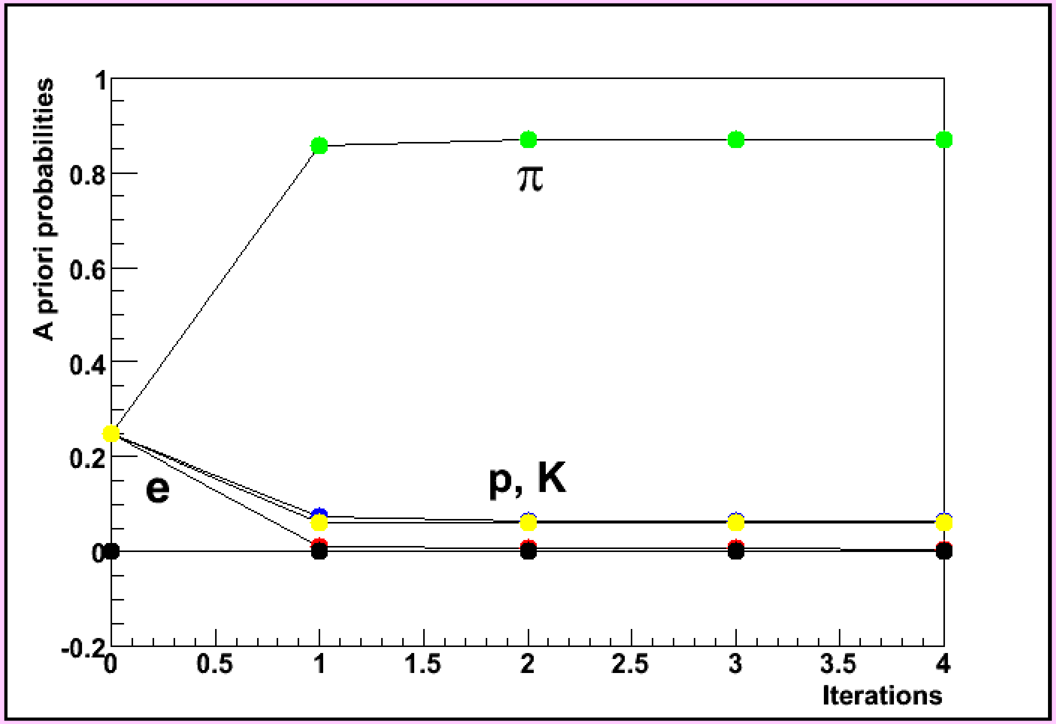
(IB’s talk at PWG1 during the Physics Week in Prague)



- In this momentum region - no particular problem

Identified particle spectra: the priors

(PWG1 in Prague)

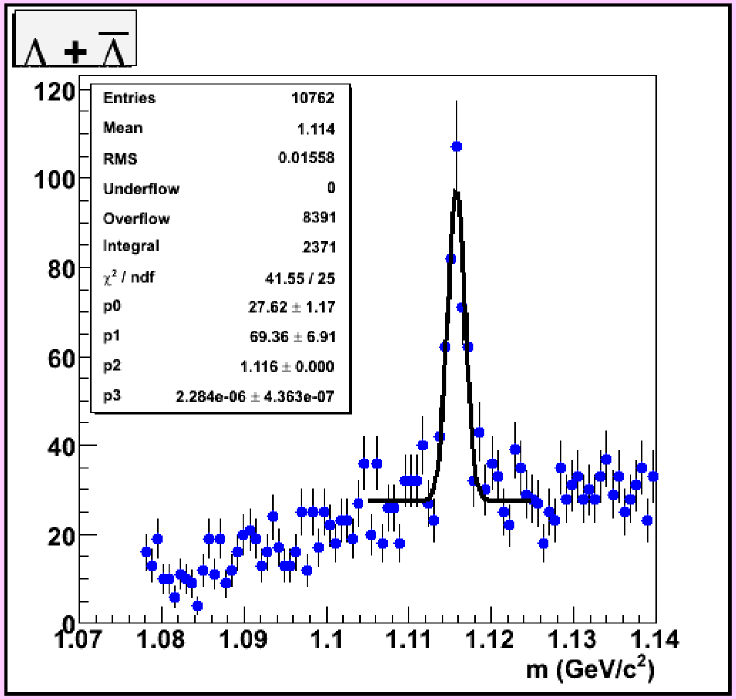


- Initial approximation for the priors: (¼ , 0 , ¼ , ¼ , ¼)
- The first iteration gives a good estimate of the priors
- This estimation is stable with respect to the subsequent iterations

Λ reconstruction

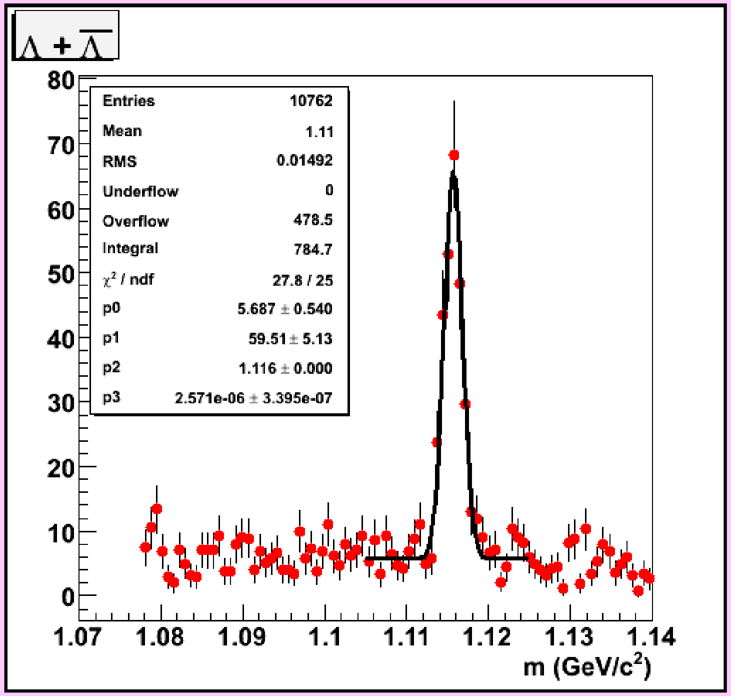
(PWG1 in Prague)

No PID



Entries	10762
Mean	1.114
RMS	0.01558
Underflow	0
Overflow	8391
Integral	2371
χ^2 / ndf	41.55 / 25
p0	27.62 ± 1.17
p1	69.36 ± 6.91
p2	1.116 ± 0.000
p3	2.284e-06 ± 4.363e-07

Using PID



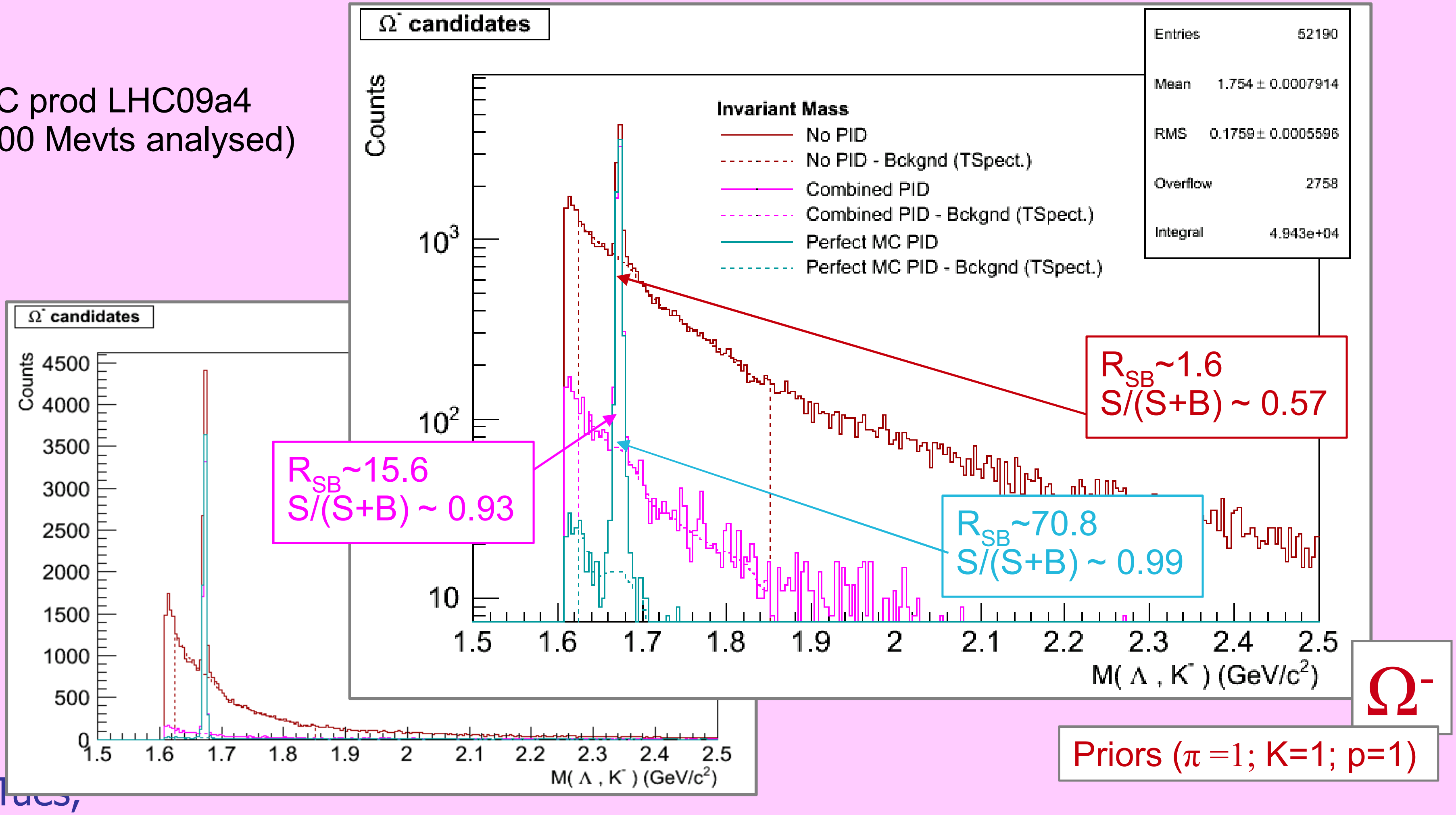
Entries	10762
Mean	1.11
RMS	0.01492
Underflow	0
Overflow	478.5
Integral	784.7
χ^2 / ndf	27.8 / 25
p0	5.687 ± 0.540
p1	59.51 ± 5.13
p2	1.116 ± 0.000
p3	2.571e-06 ± 3.395e-07

- PID gains a factor 3 in S/B, “without much loss of signal”.

Example of Ω^-

(A. Maire, PWG2 on 30.06.2009)

MC prod LHC09a4
(100 Mevts analysed)



R_{SB}~1.6
S/(S+B) ~ 0.57

R_{SB}~15.6
S/(S+B) ~ 0.93

R_{SB}~70.8
S/(S+B) ~ 0.99

Ω^-

Priors (π =1; K=1; p=1)

More examples:

- Particle spectra on the relativistic rise (A. Mastroserio)
- ϕ-meson reconstruction (A. Pulvirenti).
- ...

Conclusions

- The framework for doing the Bayesian PID (both single-detector and combined) has been available in AliRoot since 2004 (talk at CHEP04 in Interlaken).
- Within the limits of application it work reasonably well.
- There is still quite some room for improvement, mainly at the level of detector response functions.

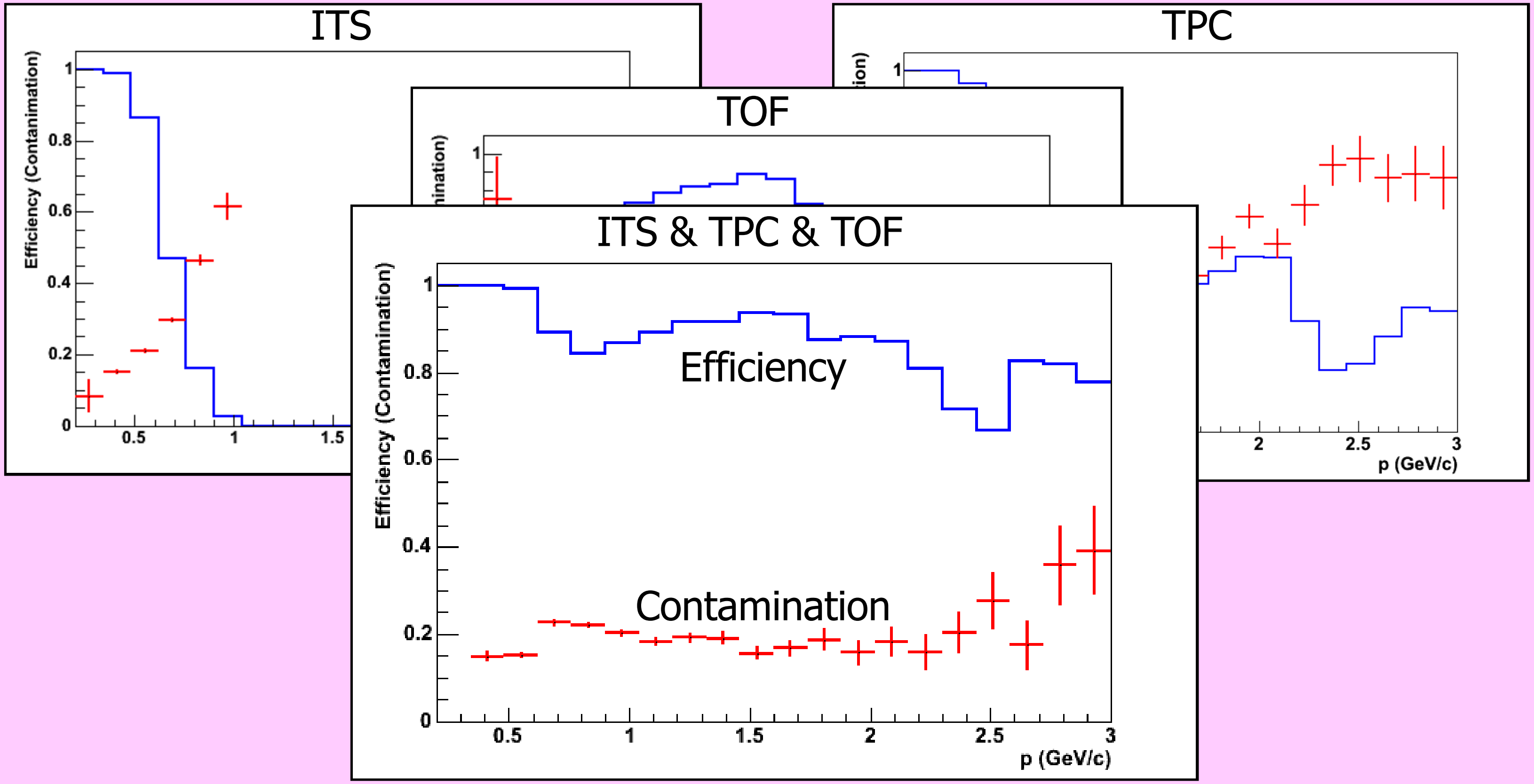
- And, there are questions for additional thinking…

Questions for additional thinking…

- Corrections for the efficiency/contamination
- High-momentum limits for PID

Example: Kaon identification with ITS,TPC and TOF

(central PbPb HIJING events)



Efficiency of the combined PID is higher (or equal) and the contamination is lower (or equal) than the ones given by any of the detectors stand-alone.

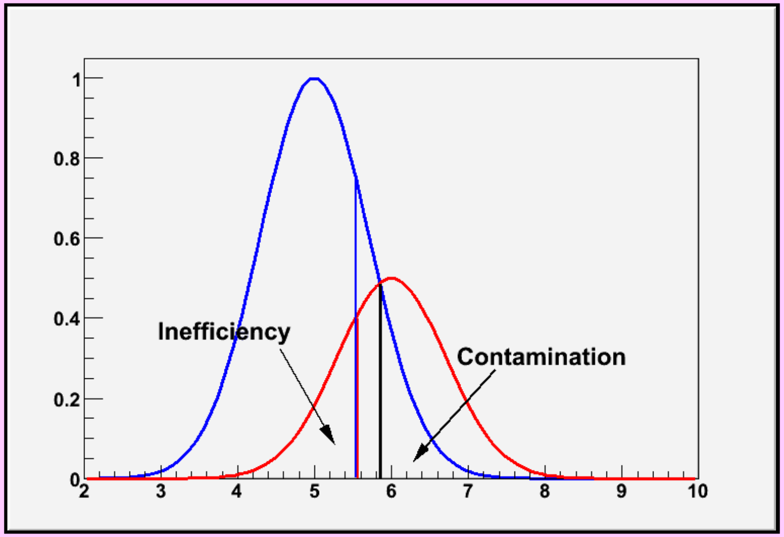
Corrections for the efficiency/contamination

- The problem of the Bayesian PID is the complexity of the efficiency curve as the function of momentum.
  - In the case of the n-sigma approach, this is the contamination that becomes complicated…
- Can we get this curve with the real data ?
- How does the uncertainty of the efficiency/contamination corrections compare with other uncertainties/sysematics ?

(this last is the key question for any PID procedure)

How to correct (1)…

- From the point of view of the corrections, the statistical treatment of the PID weights is quite “self-correcting”:



- Complications in the case of track pairs/triplets (A. Kisiel)…

How to correct (2)…

- If we are forced to decide track-by-track (max. weight, for example), we could try to do the embedding.

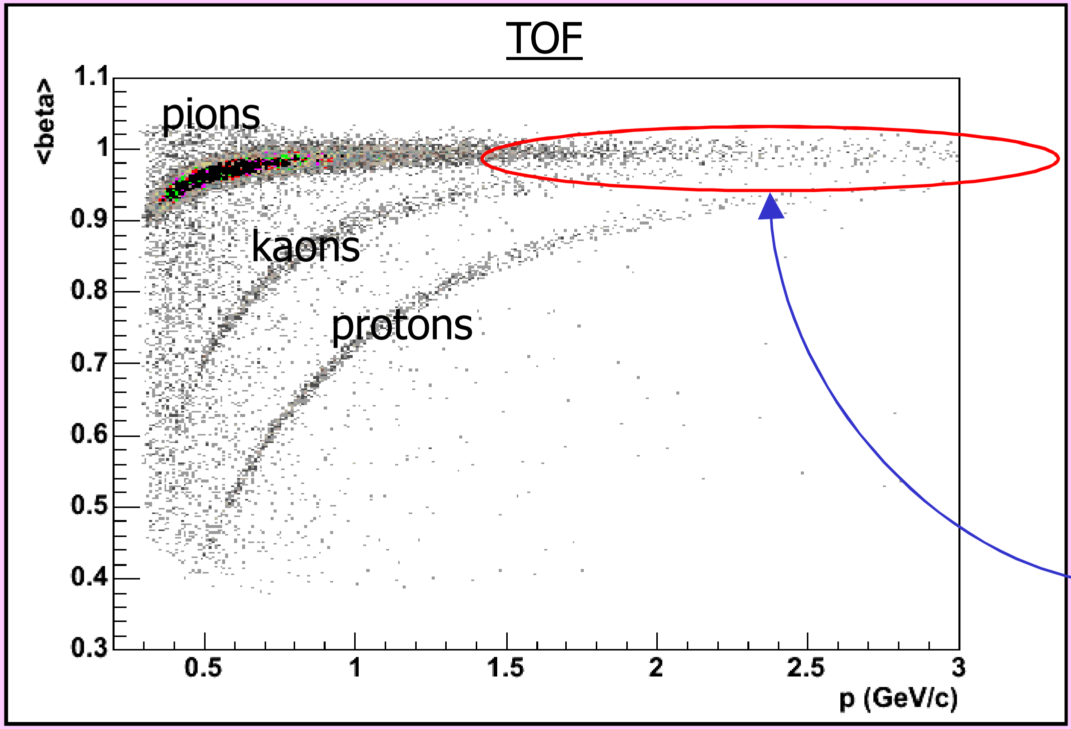
  A bit heavy… ☹

- It seems we can do it in a much more elegant way ! (discussions with B. Hippolyte and A. Maire)

  Just summing up the weights for rejected tracks S_r and the weights for accepted tracks S_a. Then, the efficiency is simply: $S_a/(S_a+S_r)$, which can be binned in pt, eta etc…

  Similar for the contamination !

The high-momentum limits for PID



The Bayesian calculations nicely glue together the momentum sub-ranges, but, as the momentum goes up, the "separation power" vanishes, and…

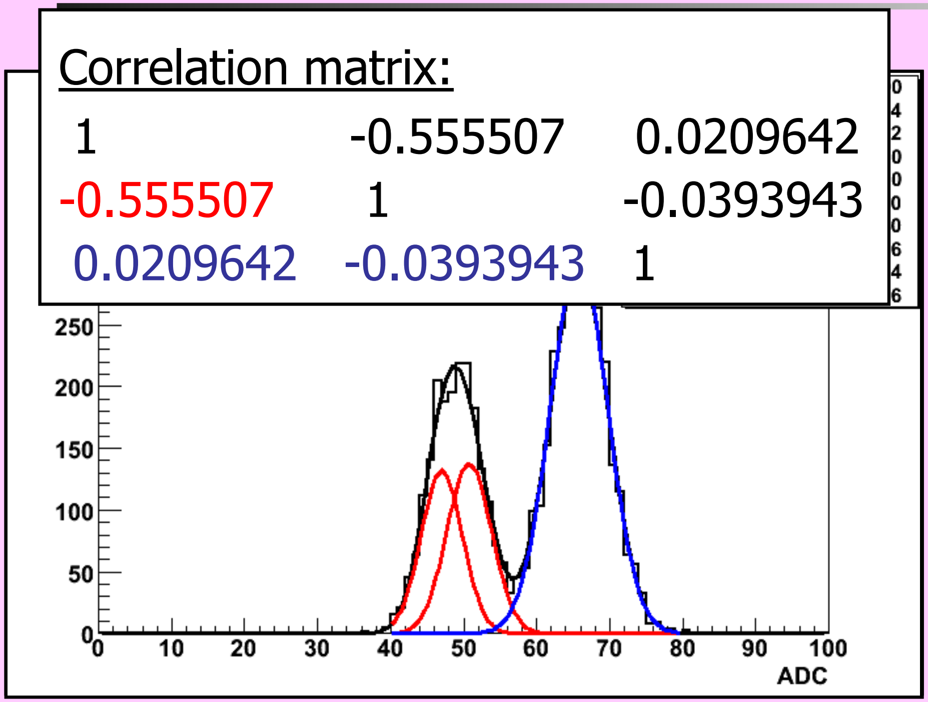
We are left with the bare priors ☹

Questions:

- The influence of the priors on the final result: Can it be somehow quantified ?
- **In any approach:** at what momentum should we stop doing the PID ?

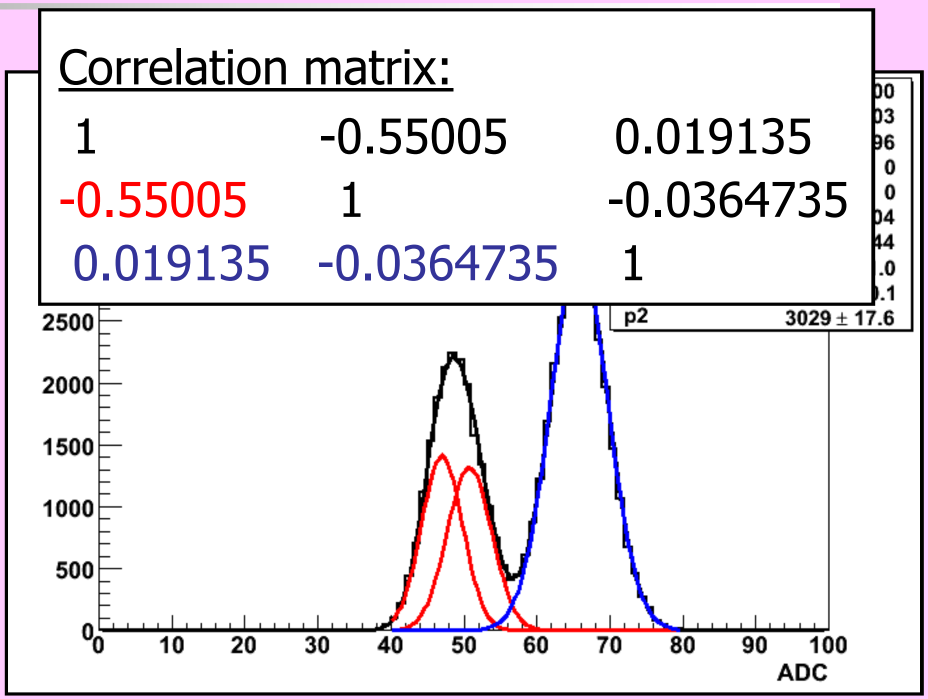
The quantity to try…

(PWG1 in Prague)



Correlation matrix:

1	-0.555507	0.0209642
-0.555507	1	-0.0393943
0.0209642	-0.0393943	1



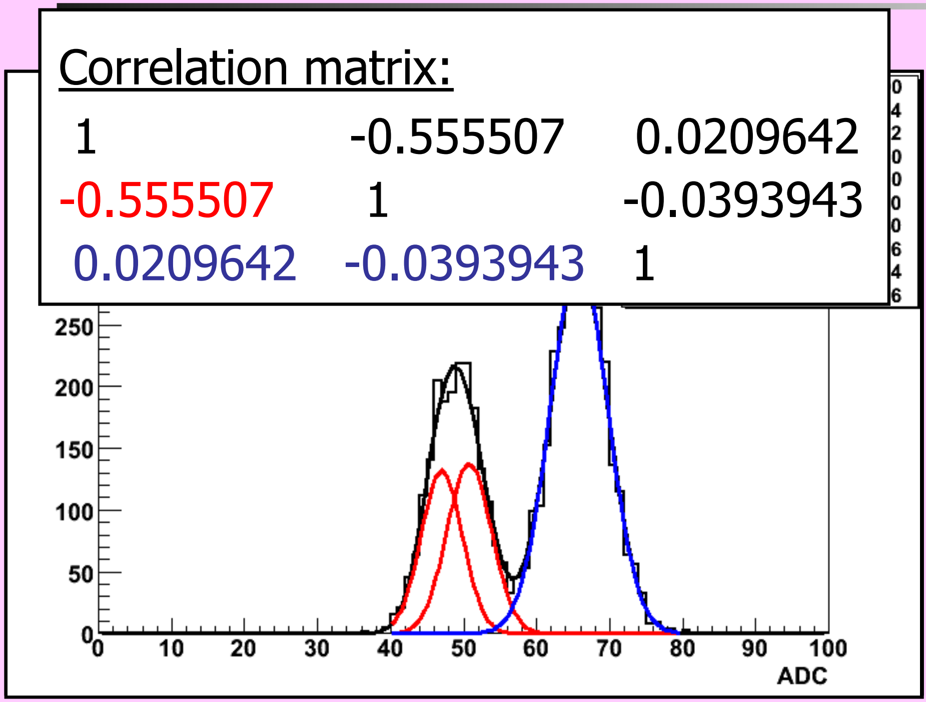
Correlation matrix:

1	-0.55005	0.019135
-0.55005	1	-0.0364735
0.019135	-0.0364735	1

- Fit to a sum of three Gaussians
  - Centroids are fixed by the "Bethe-Bloch formula"
  - Sigmas are fixed at 6% of the centroids
  - The three normalization factors are the parameters of the fit

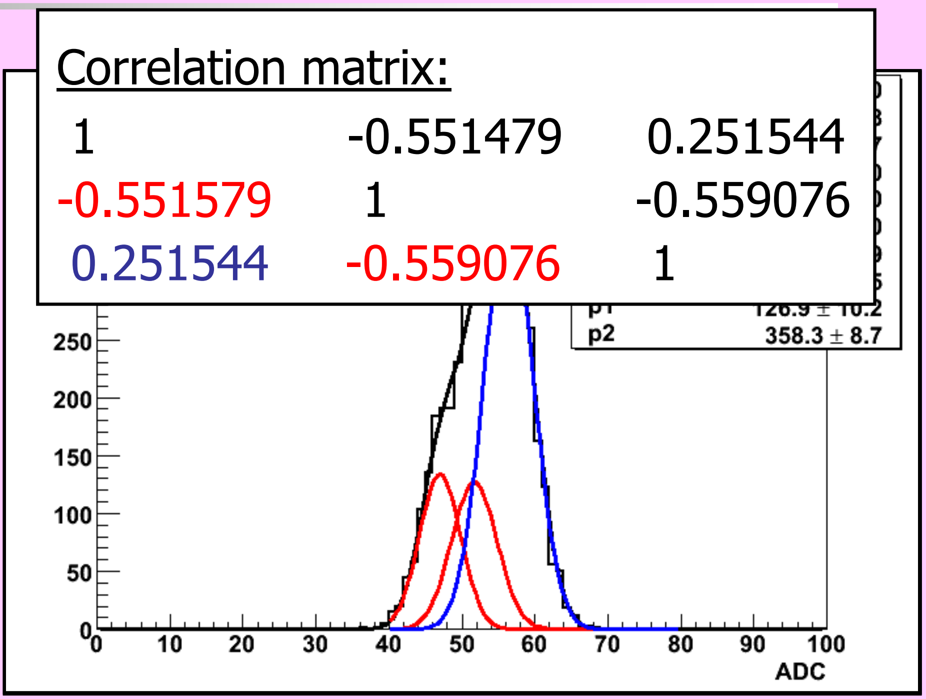
The quantity to try…

(PWG1 in Prague)



Correlation matrix:

1	-0.555507	0.0209642
-0.555507	1	-0.0393943
0.0209642	-0.0393943	1



Correlation matrix:

1	-0.551479	0.251544
-0.551579	1	-0.559076
0.251544	-0.559076	1

- Fit to a sum of three Gaussians
  - Centroids are fixed by the "Bethe-Bloch formula"
  - Sigmas are fixed at 6% of the centroids
  - The three normalization factors are the parameters of the fit