CDF 8453
Version 1.0
August 11, 2006

A Method to Disentangle Distributions of Kinematical Variables of Different Particle Species

M. Casarsa[1], P. Catastini[2], M.A. Ciocci, G. Punzi, L. Ristori, A.M. Zanetti

Abstract

This note describes a statistical method to disentangle the distributions of kinematical variables of different particle species with their correct relative normalization using PID (Particle IDentification) information. The proposed procedure avoids possible biases caused by different particle species having different unknown distributions of kinematical variables correlated with PID response (e.g. momentum and *arclength*) and is shown to work very well on a toy Montecarlo sample.

1 Particle Fraction Estimation

Consider a mixture of known particle types, for example *Pions*, *Kaons*, *Protons* and *Electrons* . We indicate by f_π, f_K, f_P and f_e respectively the unknown fractions of each particle type contained in the sample. Using Particle Identification *PID* information we want to estimate f_π, f_K, f_P and f_e and possibly obtain distributions of relevant kinematical variables for different particle types.

We will consider two methods for particle identification, both available to CDF: one is based on the measurement of the energy loss of charged particles in the gas of the COT (dE/dx), the other is based on the measurement of the Time−of−Flight (*TOF*) of the particle from the production vertex to the scintillator bars just outside the COT.

A tricky feature of *PID* devices based on the above principles is that the separation power between different particles is strongly momentum dependent. This is shown in Fig. 1 where the dE/dx mean response of different particles is plotted as a function of momentum. Assuming that the resolution of the measurement is constant, the separation power dramatically changes in a short momentum range. A similar situation occurs for Time−of−Flight though the momentum dependence is different.

As a consequence of this, the *pdf*'s that enter in the expression of the likelihood and describe the response of the *PID* devices for different particle types are strongly momentum dependent and we need to be very careful in order to avoid the *Punzi*

[1]casarsa@fnal.gov
[2]catasti@fnal.gov

effect [1]. A multicomponent Maximum Likelihood fit may be affected by strong biases whenever the dependence on event observables of the *pdf*'s used in the likelihood is not correctly taken into account. Therefore care must be taken to properly define the likelihood function as discussed below.

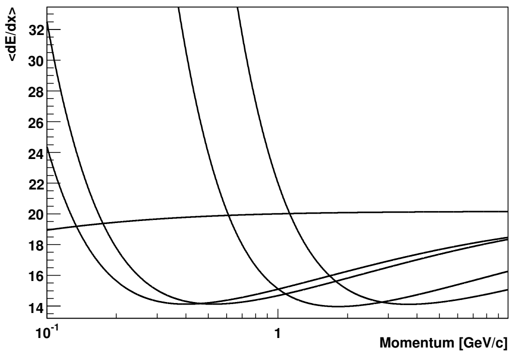

Figure 1: Mean dE/dx response of the CDF COT for different particle types as a function of momentum.

2 Avoiding the *Punzi Effect*

Consider, for example, the *PID* information provided by dE/dx. Our observables are the measured dE/dx response (that we will indicate as *pid*) and the momentum of the particle (*mom*). We then indicate as *hyp* the particular particle hypothesis.

Unfortunately, we cannot simply write the Likelihood function as:

$$L(f_j) = \prod_i \Big(\sum_{j=\pi,K,P} f_j P(pid_i | mom_i, hyp_j) \Big) \tag{1}$$

using expression (1) may give strongly biased results if our additional variable, the momentum, has different distributions for different particle types. As discussed in [1] and shown in [2], whenever the templates used in a multi-component fit depend on additional observables, it is necessary to use the complete Likelihood expression, explicitly including the probability distributions of all observables. In our particular case this implies that we need to include in the Likelihood the momentum distributions of each particle type.

We than write the correct Likelihood function as:

$$\begin{aligned} L(f_j) &= \prod_i \Big(\sum_{j=\pi,K,P} f_j P(pid_i, mom_i | hyp_j) \Big) \qquad (2) \\ &= \prod_i \Big(\sum_{j=\pi,K,P} f_j P(pid_i | mom_i, hyp_j) \\ &\qquad \times P(mom_i | hyp_j) \Big) \end{aligned}$$

With the condition:

$$\sum_{j=\pi,K,P} f_j = 1 \tag{3}$$

3 What if the distributions of additional observables are unknown?

Writing the complete Likelihood function is relatively straightforward in principle. On the other hand, in practice, we often have poor information about the distributions of the additional observables ($P(mom|hyp_j)$ in our example). Sometimes they are completely unknown. This is the case, for example, of the particles produced during the hadronization of B mesons: the momentum distribution of each particle type is unknown and the correct Likelihood function as defined in (2) cannot be constructed.

In [2] it was shown that a possible solution is to use a series expansion of the unknown distributions. The momentum term of the Likelihood function (2) is then written as:

$$P(mom_i, type_j) = \sum_m a_{mj} U_m(mom_i) \tag{4}$$

and

$$L(f_j) \rightarrow L(f_j, a_{mj}) \tag{5}$$

Where m is the order and U_m are the base vectors used for the series expansion. The a_{mj} terms are treated as free parameters to be determined by the fit.

Note that the maximization of (2) combined with (4) is not only an unbiased estimation of the particle content of the sample, but it provides also an approximation of the momentum distributions of each particle type in the form of a Series Expansion of order m of the true distributions.

A different possible strategy, used for example in [3], is based on the idea that if the fit is performed in small momentum intervals, the bias due to the use of the likelihood function (1) is small and it goes to zero as the momentum interval width decreases. This approach requires an independent maximum likelihood fit in *each* momentum bin.

In this note we take this idea to the extreme and use very small momentum bins and in each one of them we perform a separate fit to estimate particle fractions. In this way any bias caused by momentum dependence of the *pdf*'s is negligible and we show that we can extract the unknown momentum distributions of each particle type in a completely data driven mode without any assumption on the corresponding functional forms.

4 The Likelihood Function

Once we have restricted ourselves to particles contained in a sufficiently small momentum bin, the likelihood function we need to maximize in order to extract the fractions of each particle type takes the form:

$$Log(L) = \mathcal{L} = \sum_{i=0}^{N} Log(\sum_{j=1}^{M} N_j pdf_j(R_i)) \tag{6}$$

where N is the total number of particles in a particular momentum bin, M is the number of different particle species, N_j is the number of particles of type j and R_i is the PID response.

The likelihood must be maximized with respect to the free parameters N_j with the constraint:

$$N = \sum_{j=1}^{M} N_j \tag{7}$$

The discriminating power comes from the fact that the PID response has different distributions for different particle types (pdf_j). The *Punzi effect* is avoided because, once we restrict ourselves to sufficiently small momentum bins, the pdf_j do not depend on any other quantity but the particle type.

The constrained maximum likelihood fits to obtain particle fractions in each momentum bin are performed using the iterative procedure called Channel Likelihood (CL) described in Appendix A.

5 A Toy Monte Carlo study

We test our method on toy Montecarlo samples. Each sample consists of 3500 particles. Different P_T spectra used to generate each particle type are taken from Pythia and the corresponding fractions are fixed at:

$$f_\pi = 0.74, \quad f_K = 0.17, \quad f_P = 0.07, \quad f_e = 0.02 \tag{8}$$

We then divide the sample into 50 momentum bins from 0.45 GeV/c up to 6.0 GeV/c. A maximum likelihood fit is made in each momentum bin to estimate the fractions of each particle species. Empty bins are ignored. We repeat the fit on 100 toy samples and plot the distributions of the residuals of the estimates.

Fig.2 and Fig.3 show the plots of the residuals of the estimators of the total fractions of Pion, Kaon, Proton and Electron, summed over all momentum bins. No significant bias is observed (a small shift can be noticed for the electron estimator, but it is less than one σ). Fig.4 shows the plots of the residuals of the estimators as a function of the bin number (i.e. momentum). In each plot, each point represents the mean value,

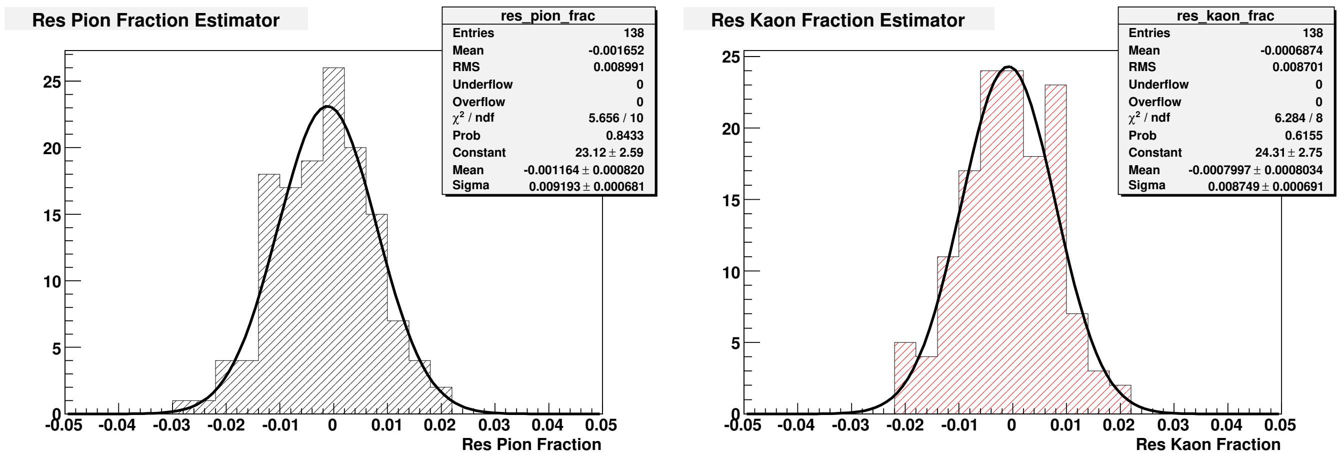

Figure 2: $Fraction_{meas}$ - $Fraction_{true}$ distributions (integrated on all momentum bins) for 100 toy samples: left Pions, right Kaons.

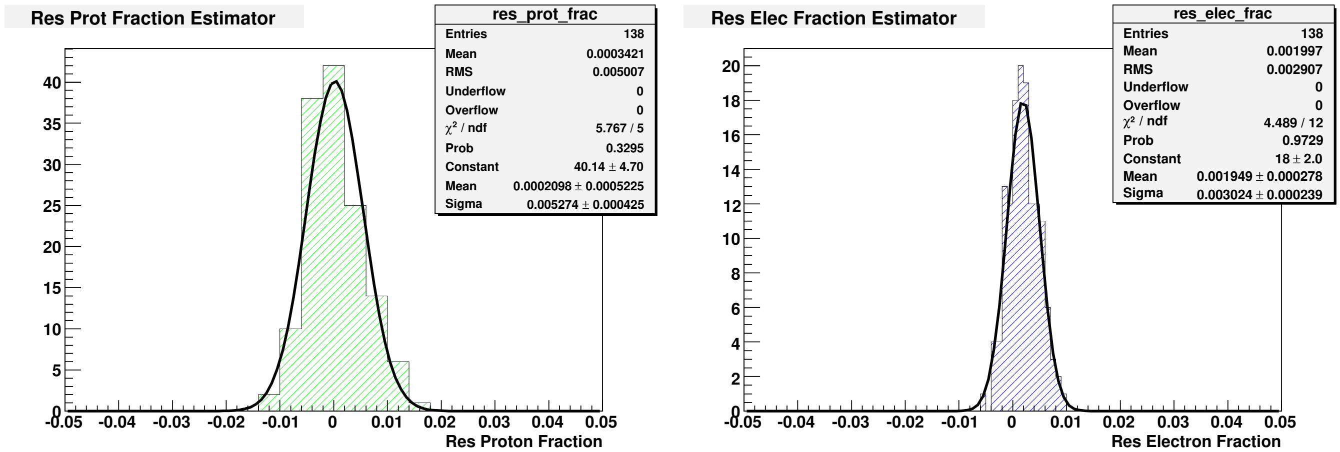

Figure 3: $Fraction_{meas}$ - $Fraction_{true}$ distributions (integrated on all momentum bins) for 100 toy samples: left Protons, right Electrons.

while the error bar represents the width of the distribution of the residual. We observe no significant bias in all cases.

Examples of momentum distributions obtained from the fit are shown in Fig.5. A very good agreement between the true distributions (filled histograms) and the fit (dots) is observed for all particle types. The fits provide a very good estimation of the unknown distributions.

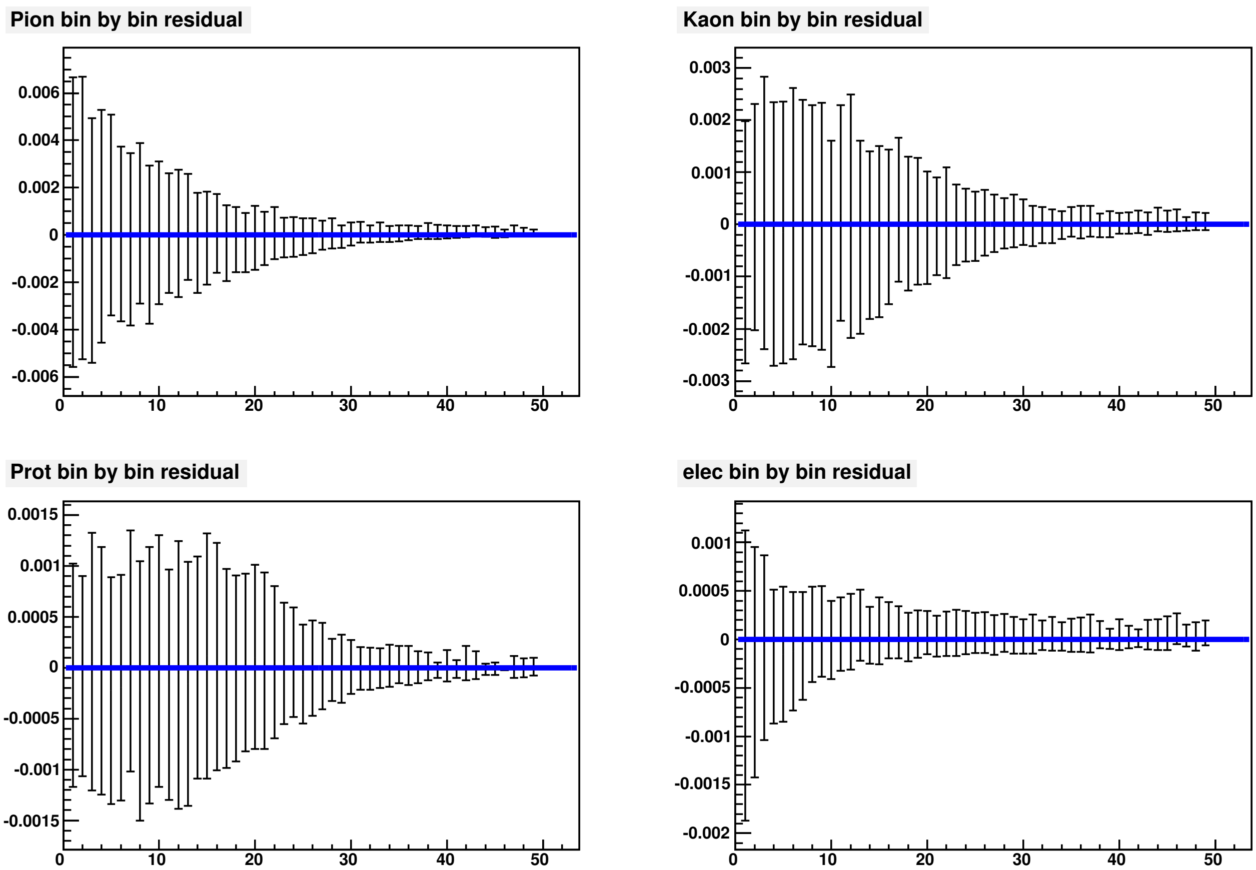

Figure 4: Mean and width of the $Fraction_{meas}$ - $Fraction_{true}$ distributions in each momentum bin.

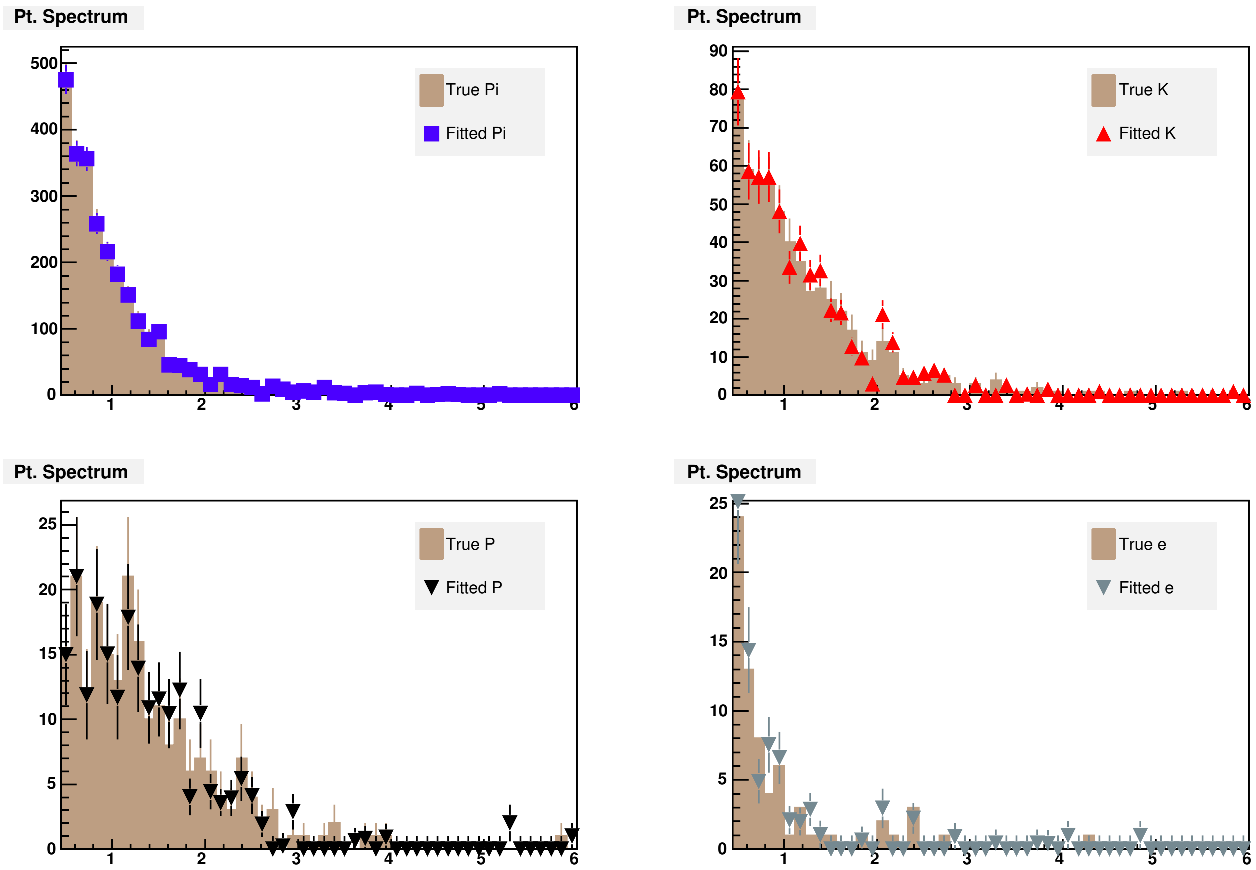

Figure 5: Overlay of the toy momentum distributions (filled histograms) and the ones resulting form the CL fits (dots).

6 The 2-Dimensional case

The method can be easily generalized, at least in principle, to a multidimensional case.

If, for example, our PID information is obtained by the measurement of the Time of Flight, the expression of the expected TOF (TOF_{exp}) is actually a function of two observables:

$$TOF_{exp}(mom, L) = \frac{L}{c}\sqrt{1 + (m_j/mom)^2} \tag{9}$$

where L is the length travelled by the particle (also called $arclength$), c is the speed of light, m_j is the mass of the particle type j and mom is again the momentum of the particle.

The distributions of both momentum and arclength can be different for different particle species and are unknown in general. Therefore both observables could be a source of bias in the estimation of particle fractions.

The consequence is that, since now we have two observables to deal with, we need to build a $2-$dimensional grid of bins in mom and L and estimate particle fractions independently in each bin.

In Fig.6 and Fig.7 the true 2-dimensional distributions of mom and L are shown superimposed to the result of the fit. The agreement is clearly very good also in this case.

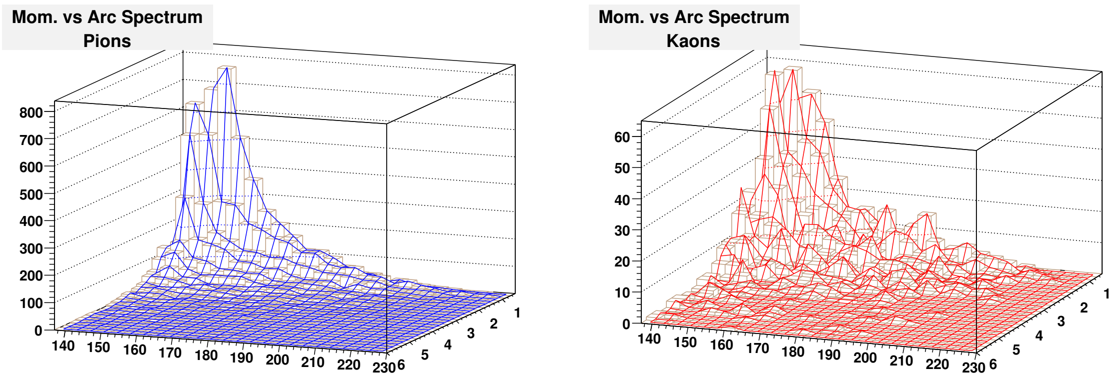

Figure 6: Momentum versus Arclength distributions from Pythia MC (histograms) and the fitter (surface): left Pions, right Kaons.

7 Extracting distributions of other quantities

Besides obtaining momentum and arclength distributions, we are typically interested in obtaining plots of many additional kinematical variables for the different particle species. A method to achieve this is a technique known as $sPlot$ and is based on estimating particle fractions as described above and their covariance matrix [9]. The

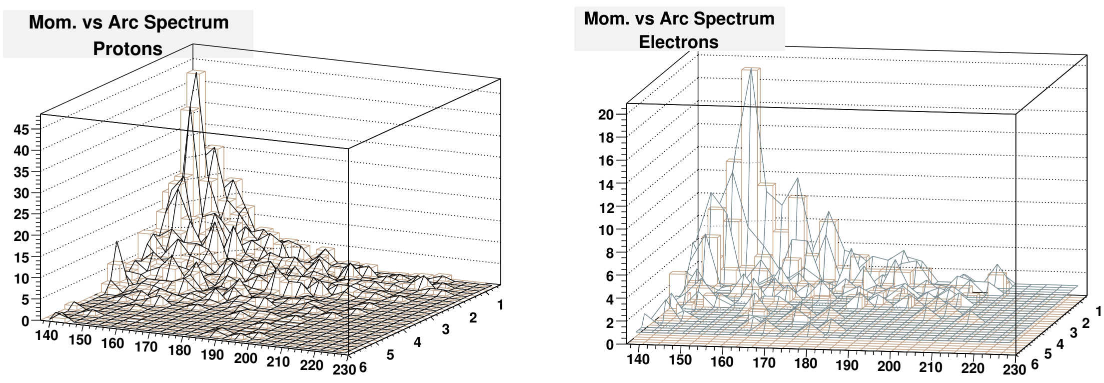

Figure 7: Momentum versus Arclength distributions from Pythia MC (histograms) and the fitter (surface): left Protons, right Electrons.

main requirement for the *sPlot* technique to actually work is that the kinematical variables we wish to plot be uncorrelated with the *pdf*'s that describe the response of the *PID*. In our case we do not need to worry because any dependence of the *pdf*'s from kinematical variables is $frozen$ by our strategy to perform a separate fit in each momentum and arclength bin.

To test this feature, we use a Pythia Montecarlo sample of $B^- \rightarrow D^0\pi^-$, such a sample will provide a reasonable description of all the kinematical variables associated with a given particle and the correlations among them. We then fit the composition of the particles produced in the vicinity of the B^- meson. A parametric simulation of dE/dx and TOF is used.

We observe that momentum and arclength can be replaced by P_T and η with no loss of information. The advantage is that η distributions for different particle types are very similar. This in shown in Fig.8 where η distributions for different particle types are overlaid. No significant difference is observed and this allows us to factorize the η distribution and eliminate it from the Likelihood. This is a big simplification since it allows us to avoid $2-$dimensional binning and consider P_T as the only unknown variable.

Tab.7 reports the total fraction of each particle type, and Fig.9 shows the P_T distributions resulting from the fit.

	π	K	P	e
True fraction	0.828	0.112	0.0557	0.0039
Fitted fraction	0.829 ± 0.002	0.111 ± 0.001	0.0557 ± 0.0006	0.0044 ± 0.0003

Table 1: Pythia Monte Carlo fractions (true fractions) and fitter result

Using the particle fractions and the covariance matrices estimated in each P_T bin we can produce the *sPlots* of other kinematical variables.

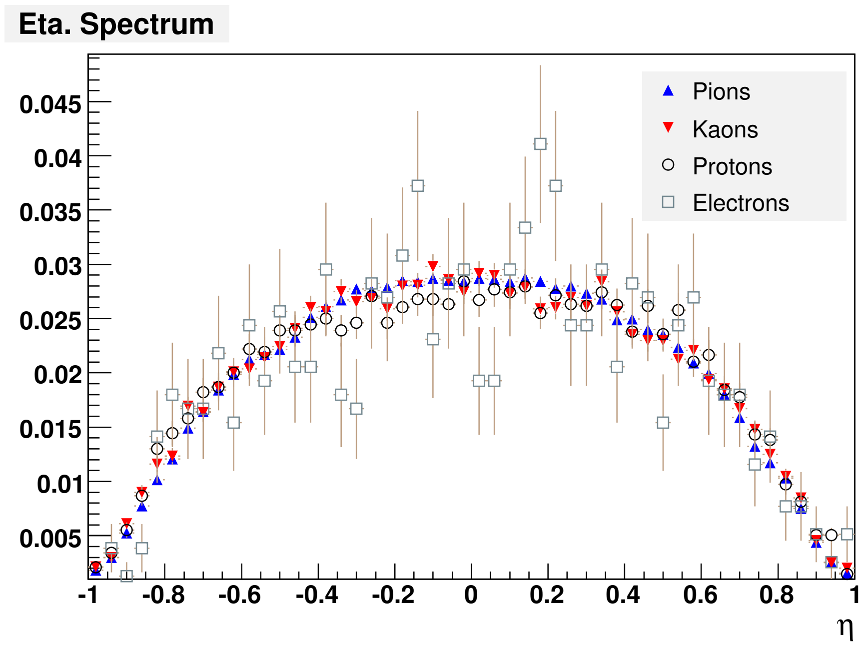

Figure 8: Overlay of Pythia η distributions for each particle type.

Figures 10 − 14 show examples of distributions of a number of kinematical variables obtained in this way. In all cases a very good agreement is observed between the *true* distributions and the corresponding fit.

8 Conclusions and outlook

We have tested, with a toy Montecarlo, a statistical method to estimate the composition of a mixed sample of particle types using the information from PID detectors. We have also shown that the distributions of kinematical variables of different particle species can be disentangled with remarkable precision. The method uses a likelihood function that correctly contains the *pdf*'s of all the relevant event observables and is therefore immune to the *Punzi effect*. This approach looks very promising for the determination of distributions of different types of particles produced, for example, in conjunction with a B meson in a completely data-driven mode, without making any prior assumption on their functional forms. We are planning to apply the principles described in this note to real data shortly and we expect the information obtained in this way to be very valuable especially for the optimization of flavor tagging algorithms.

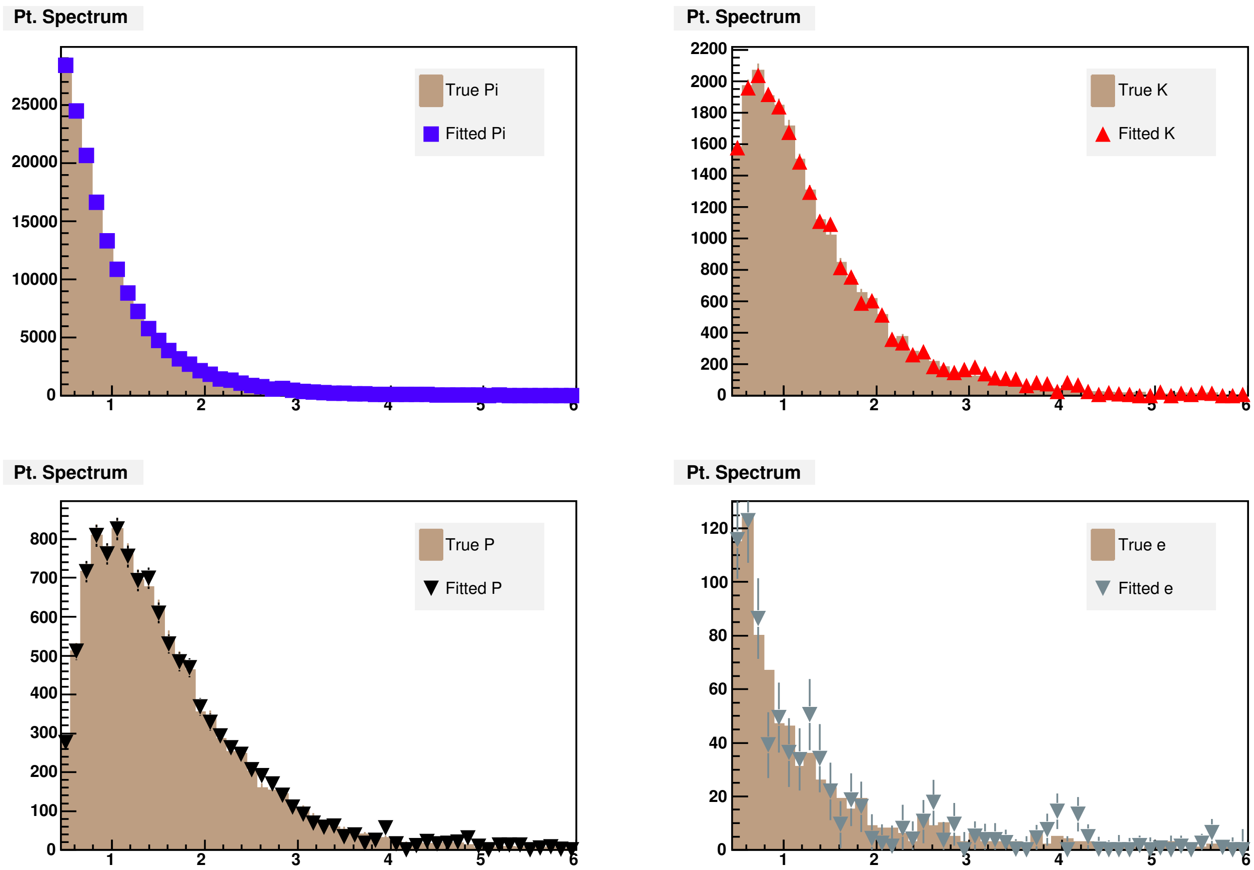

Figure 9: Overlay of the MC P_T distributions (filled histograms) and the result from the fit (dots).

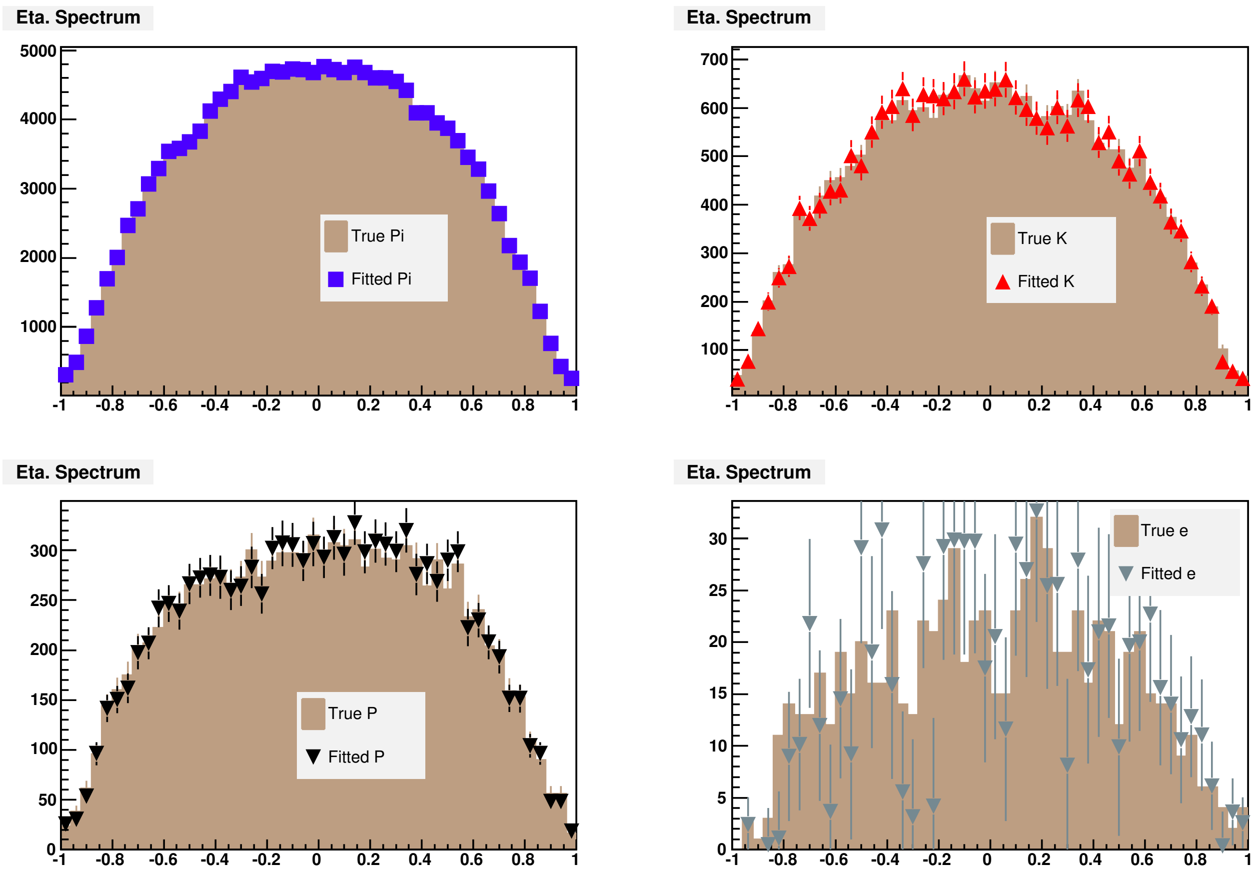

Figure 10: Pythia MC η distributions overlaid to the corresponding data *sPlot*.

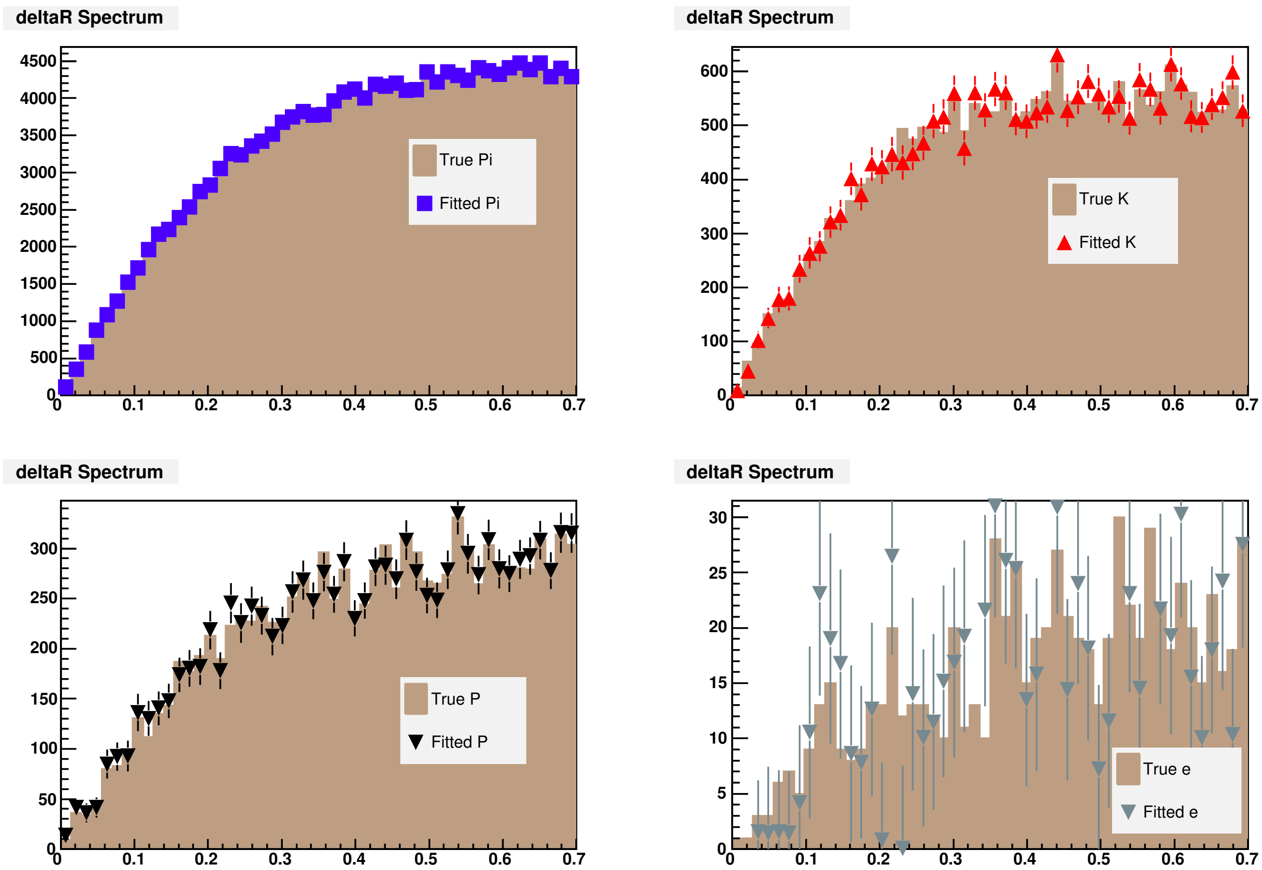

Figure 11: Pythia MC ΔR distributions overlaid to the corresponding data *sPlot*.

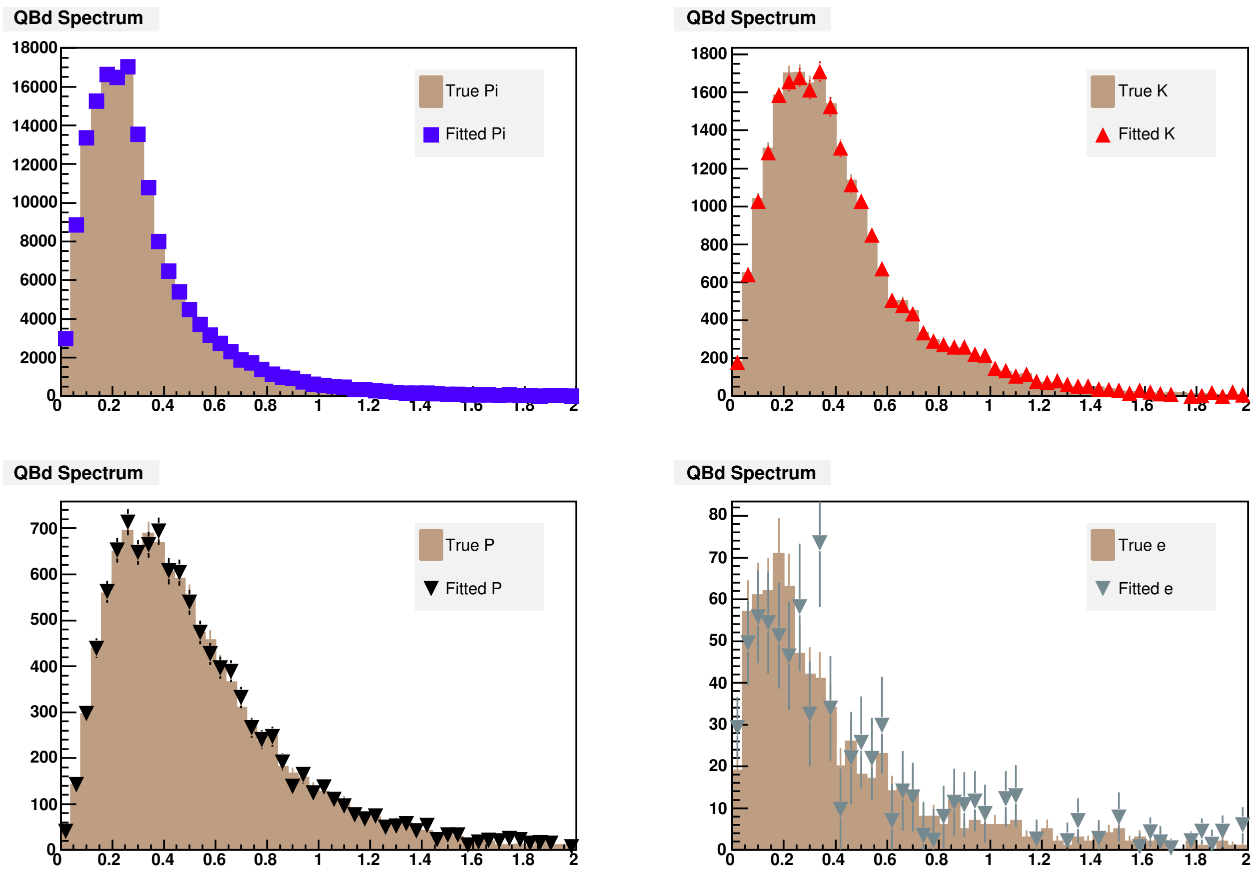

Figure 12: Pythia MC $M(B\pi) - M(B) - M(\pi)$ distributions overlaid to the corresponding data *sPlot*.

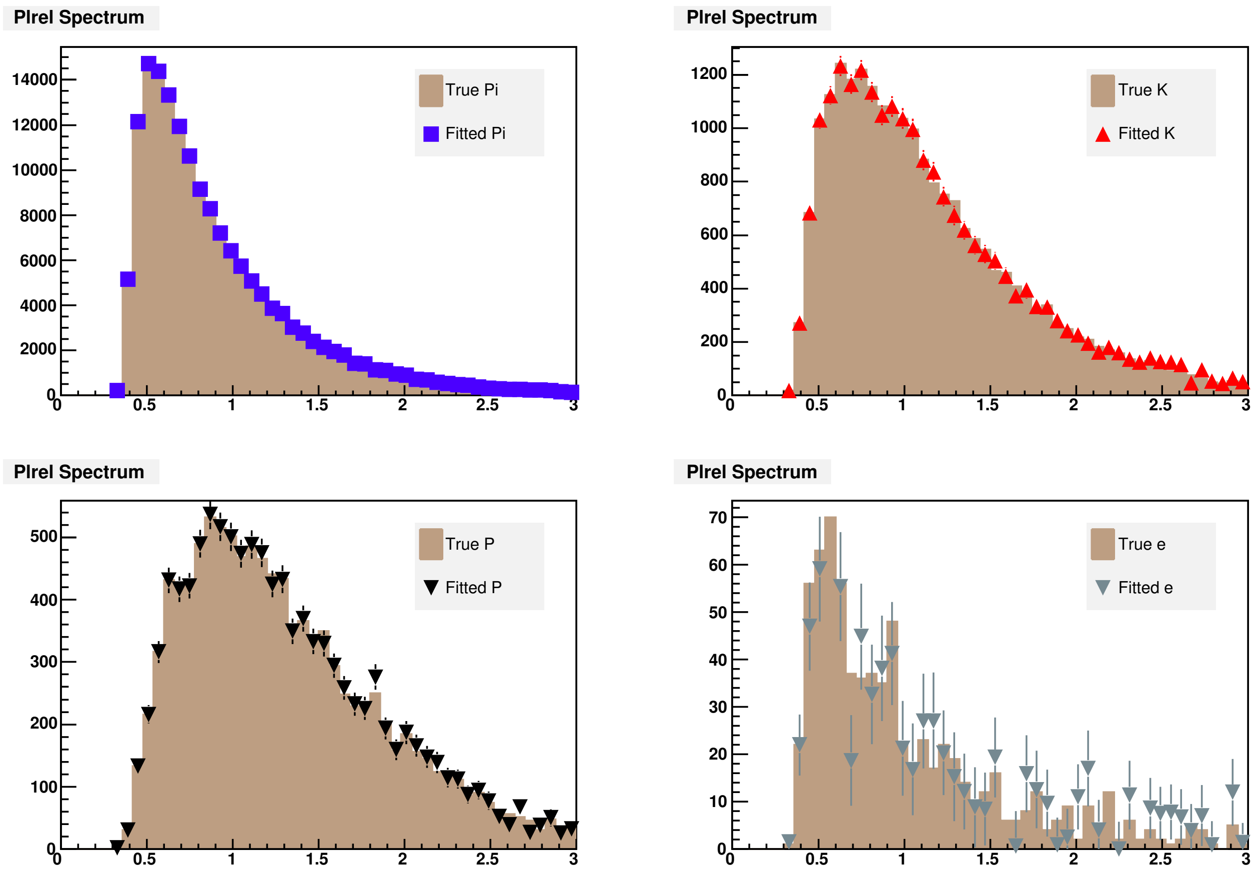

Figure 13: Pythia MC P_{Lrel} distributions overlaid to the corresponding data *sPlot*.

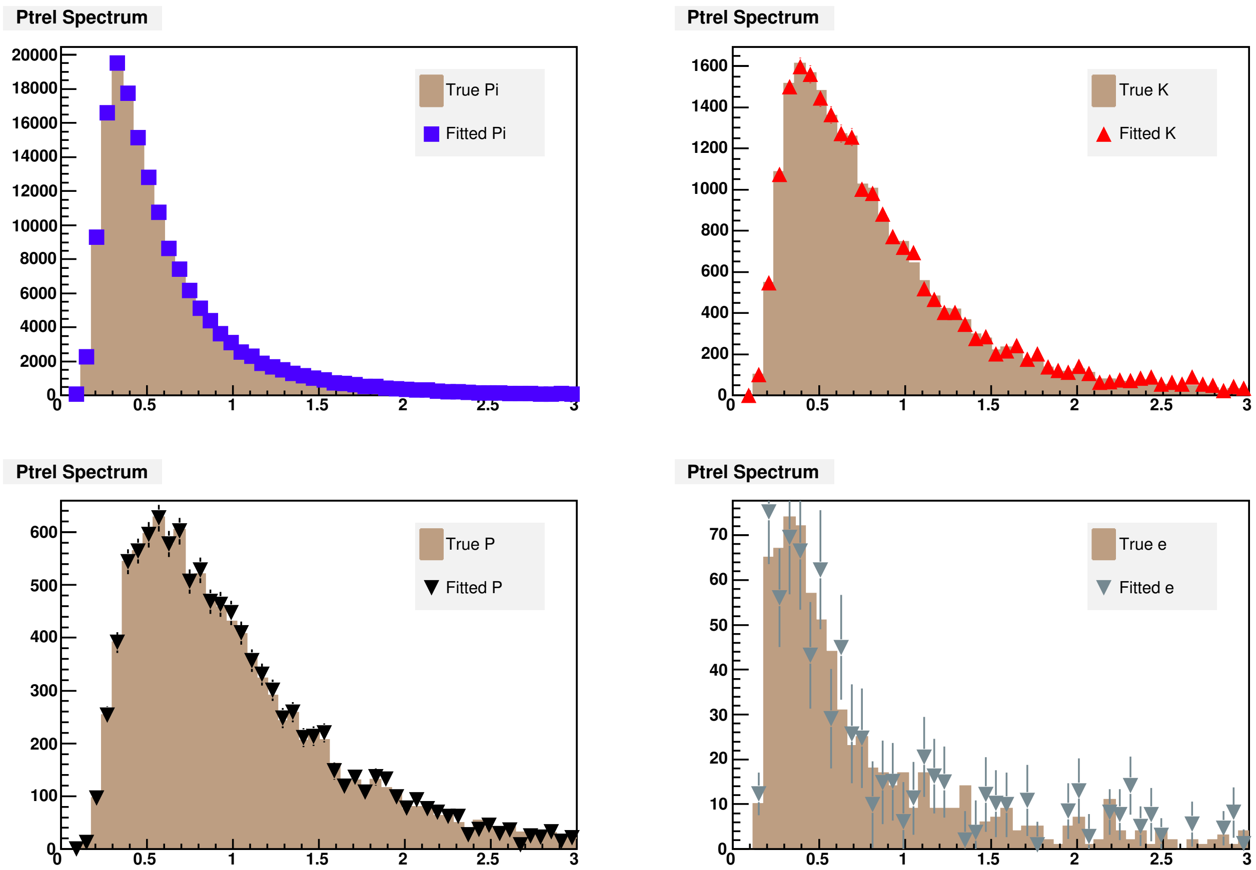

Figure 14: Pythia MC P_{Trel} distributions overlaid to the corresponding data *sPlot*.

References

[1] G. Punzi, *[physics/0401045]*. 2

[2] P. Catastini, G. Punzi, Proceedings of PHYSTAT05: Statistical Problems in Particle Physics, Astrophysics and Cosmology, Oxford, England, United Kingdom, 12-15 Sep 2005 *[arXiv:physics/0605130]*. 2, 3

[3] M., Jones, J. Kroll, D. Usynin, *CDF/DOC/BOTTOM/CDFR/7810*. 3

[4] P. E. Condon and P. L. Cowell, *Phys. Rev. D* **9**, *2558 (1974)*. 18, 19

[5] J. B. Gay *et al.* [Amsterdam-CERN-Nijmegen-Oxford Collaboration], *Phys. Lett. B* **63**, *220 (1976)*. 19

[6] D. Buskulic *et al.* [ALEPH Collaboration], *Phys. Lett. B* **307**, *209 (1993)*.
P. D. Acton *et al.* [OPAL Collaboration], *Z. Phys. C* **59**, *1 (1993)*. 19

[7] B. Aubert *et al.* [BABAR Collaboration], *Phys. Rev. D* **69**, *031101 (2004) [arXiv:hep-ex/0310050]*. 19

[8] G. D'Agostini, *Nucl. Instrum. Meth. A* **362**, *487 (1995)*. 19

[9] M. Pivk and F. R. Le Diberder, *Nucl. Instrum. Meth. A* **555**, *356 (2005) [arXiv:physics/0402083]*. 8

9 Appendix

An iterative procedure to estimate particle fractions and the corresponding momentum distributions

We need to find a procedure to maximize the likelihood function (6) with the constraint (7).

We will briefly review here a statistical method known as *Channel Likelihood* [4] and explain how it can be applied to our problem.

Suppose we have a sample of N events that is a mix of M classes, or types of events. Each event is described by a set of discriminating variables that we will indicate with the vector $\vec{R}$. We also know the distributions of the $\vec{R}$ variables for each class of events that will be indicated as $pdf(\vec{R})_j$ where $j = 1, 2, ..., M$.

Our goal is to estimate the number of events of each class N_j, in the initial sample N, with the constraint:

$$\sum_{j=1}^{M} N_j = N \tag{10}$$

We can define the *relative* probability of event i having come from class j:

$$N_j pdf(\vec{R}_i)_j \tag{11}$$

and the normalization for the event i:

$$\sum_{j=1}^{M} N_j pdf(\vec{R}_i)_j \tag{12}$$

Combining (11) and (12) we define the weight w_{ij} of event i for the class j:

$$w_{ij} = \frac{N_j pdf(\vec{R}_i)_j}{\sum\limits_{k=1}^{M} N_k pdf(\vec{R}_i)_k} \tag{13}$$

that is the probability for event i to come from channel j.

Notice that $\sum\limits_{j=1}^{M} w_{ij} = 1$ for any given event i.

We can now write the N_j as:

$$N_j = \sum_{i=1}^{N} w_{ij} \tag{14}$$

Starting from a first guess of the parameters, N_j^0 we can define an iterative procedure expressed by:

$$N_j^n = \sum_{i=1}^{N} \frac{N_j^{n-1} pdf(\vec{R}_i)_j}{\sum\limits_{k=1}^{M} N_k^{n-1} pdf(\vec{R}_i)_k} \tag{15}$$

where, at every step n of the iteration, we estimate the N_j^n that will be used as input to step $n+1$.

The procedure converges to the set of N_j that are the solutions of the equations:

$$N_j = \sum_{i=1}^{N} \frac{N_j pdf(\vec{R}_i)_j}{\sum\limits_{k=1}^{M} N_k pdf(\vec{R}_i)_k} \tag{16}$$

or equivalently:

$$1 = \sum_{i=1}^{N} \frac{pdf(\vec{R}_i)_j}{\sum\limits_{k=1}^{M} N_k pdf(\vec{R}_i)_k} \tag{17}$$

After the solution has been found, we can use the w_{ij} to weight events when making plots of a variable belonging to $\vec{R}$.

In [4] it is also shown that the equation system (17) is the same that we are supposed to solve in order to maximize the likelihood with respect to the parameters N_j:

$$Log(L) = \mathcal{L} = \sum_{i=0}^{N} Log(\sum_{j=1}^{M} N_j pdf(\vec{R}_i)_j) \tag{18}$$

with the constraint:

$$\sum_{j=1}^{M} N_j = N \tag{19}$$

Notice also that , from the equivalence between CL and Maximum Likelihood, once the N_j are estimated, to evaluate the statistical error of our parameters we can write the inverse of the covariance matrix V_{jl}^{-1} of the fit as:

$$V_{jl}^{-1} = \frac{\partial(-\mathcal{L})}{\partial N_j \partial N_l} = \sum_{i=0}^{N} \frac{pdf(\vec{R}_i)_j pdf(\vec{R}_i)_l}{\left(\sum\limits_{k=1}^{M} N_k pdf(\vec{R}_i)_k\right)^2} \tag{20}$$

The Channel Likelihood method has been used in several analyses, see for example [5], [6] and more recently [7]. In [8] a similar method is used for unfolding experimental distribution to get the best estimate of the true ones.

It turn out that this iterative process is a very efficient and robust way to perform our fir in each momentum bin.

Using the same notation introduced earlier, N_j is the number of particles of type j where $j = \pi, K, P, e$. We then define the iteration in each momentum bin Δmom as:

$$N_{j,\Delta mom}^{n+1} = \sum_{i=1}^{N} \frac{N_{j,\Delta mom}^{n} pdf(pid_i|mom_i)_j}{\sum\limits_{k=\pi,K,P,e} N_{k,\Delta mom}^{n} pdf(pid_i|mom_i)_k} \tag{21}$$

Once we obtain the convergence of the iterative process, we can write, for each particle type:

$$f_j = \frac{N_j}{N} = \frac{1}{N} \sum_{\Delta mom} N_{j,\Delta mom} \tag{22}$$

observing that each bin content is fitted independently, the corresponding statistical uncertainty is:

$$\sigma(f_j) = \frac{\sqrt{\sigma^2(N_j)}}{N} = \frac{1}{N} \sqrt{\sum_{\Delta mom} \sigma^2(N_{j,\Delta mom})} \tag{23}$$

where $\sigma^2(N_{j,\Delta mom})$ in the variance of the estimator of the number of particles of type j in the momentum interval Δmom.