A pitfall in the use of extended likelihood for fitting fractions of pure samples in a mixed sample

A. Nappi
University of Perugia and INFN, Sezione di Perugia

Abstract
The paper elucidates, with an analytic example, a subtle mistake in the application of the extended likelihood method to the problem of determining the fractions of pure samples in a mixed sample from the shape of the distribution of a random variable. This mistake, which affects two widely used software packages, leads to a misestimate of the errors.

Introduction
In particle physics experiments it is often necessary to determine the fractions of several types of events contributing to a measured sample, on the basis of the shape of the distribution of a (possibly multi-dimensional) random variable. The formulation of this problem with the maximum likelihood method is discussed in textbooks about statistical data analysis. Using the total number of events in an additional term, involving the expectation value of this number (extended likelihood), does not add any information for this particular problem, but leads to a more symmetric analytic formulation of the problem, with equivalent results\(^1\).

In this paper I want to bring attention to a misunderstanding of the role of the total number of events in the extended likelihood approach, which may lead to a subtle mistake in the error calculations. This is the case, for example, of HMCMLL\(^2\), and TFractionFitter\(^3\), two widely used software packages that treat the case\(^4\) where the probability density functions (pdf) are not specified analytically, but sampled by a Monte Carlo calculation.

The purpose of the paper is to spread awareness on this possible mistake, which may cause incorrect physics results. The approach will be rather didactic and the emphasis will be in trying to elucidate the origin of the mistake in an example where the relevant arguments are not obscured by the complications in the algebra. The case considered is the one of binned data with the pdf of the different components specified analytically, but the conclusions are easily extended to more realistic cases where the pdf are sampled by Monte-Carlo calculations. The well understood issue of the relation between the two, seemingly alternative, approaches of considering the histograms poissonian or multinomial variables, is also discussed, for completeness, in the first section.

1. The maximum likelihood approach
We shall assume that there is a mixed sample under study, containing unknown fractions of two or more pure samples. For the latter, the pdf of a (possibly multi-dimensional) variable is fully specified.

Let \( q_s^{(k)} \) be the binned pdf of the pure sample \( s \) (i.e. the probability that an event of the pure sample \( s \) falls in bin \( k \)) and \( q_k \) the binned pdf of the mixed sample. The model is specified by

\(^1\) See, e.g., G.Cowan, Statistical Data Analysis, Oxford Science Publications 1998, sections 6.9, 6.10
\(^2\) http://www.hep.man.ac.uk/u/roger/hfrac.f
\(^3\) http://root.cern.ch/root/doc/RootDoc.html
\[ q_k = \sum_{s=1}^{S} p_s q_k^{(s)} \]  
where \( p_s \) are the fractions that one wishes to determine, constrained by

\[ \sum_{s=1}^{S} p_s = 1 . \]  

Since the total number of events in the mixed sample does not carry any information relevant to the problem, it can be treated as a fixed number (i.e. not a random variable), in the formulation of the likelihood. Thus, if we call \( n_k \) the observed numbers of events in bin \( k \) in the mixed sample, the distribution of the \( n_k \)'s is the multinomial distribution

\[ P(n_k) = \frac{N!}{\prod_{k=1}^{K} n_k!} \left( \prod_{k=1}^{K} q_s^{n_k} \right) \]  

and the problem can be formulated as a maximization of

\[ \ln L = \sum_{k=1}^{K} n_k \ln \left( \sum_{s=1}^{S} p_s q_k^{(s)} \right) \]  

with respect to the \( p_s \), subject to the constraint \( \sum_{s=1}^{S} p_s = 1 . \)

The fact that the total number of events \( N \) does not carry any information on the \( p_s \) can be seen explicitly. The probability distribution of the observations \( N, n_k \) can be written in terms of the poissonian probability to observe \( N \) events times the probability of observing \( n_k \) events in the individual bins, conditional to the hypothesis that the total number of events in all bins is \( N \)

\[ P(n_k, N) = P(n_k|N)P(N) = \frac{N!}{\prod_{k=1}^{K} n_k!} \left( \prod_{k=1}^{K} q_s^{n_k} \right) \frac{N^N e^{-N}}{N!} \]  

containing the additional parameter \( \nu \), the expectation value of the total number of events.

This can be used to formulate an extended likelihood, function of the parameters \( p_s \) (\( S-1 \) parameters because of the constraint) and of \( \nu \):

\[ \ln L = \sum_{k=1}^{K} n_k \ln \left( \sum_{s=1}^{S} p_s q_k^{(s)} \right) + N \ln \nu - \nu \]  

It is useful to stress, at this point, that there are \( S \) parameters to be determined: \( S-1 \) are the fractions that can be fixed independently and one is the expectation value of the total number of events. These parameters are not mixed in the likelihood function (only the \( p_s \) appear in the first term and only \( \nu \) appears in the last two terms) One obtains the trivial result that the \( p_s \) estimates are those that would be determined with the multinomial approach, whereas the maximum likelihood estimate of \( \nu \) is given by \( \hat{\nu} = N . \)

The extended likelihood can be put in a form more useful for applications, using the fact that formula (4) implies that the numbers of events in each bin can be considered independent poissonian variables. By trivial algebra it can be shown, in fact, that

\[ P(n_k, N) = P(n_k|N)P(N) = \frac{N!}{\prod_{k=1}^{K} n_k!} \left( \prod_{k=1}^{K} q_s^{n_k} \right) \frac{N^N e^{-N}}{N!} = \prod_{k=1}^{K} \frac{\left( \sum_{s=1}^{S} \nu p_s q_k^{(s)} \right)^{n_k} e^{-\sum_{s=1}^{S} \nu p_s q_k^{(s)}}}{n_k!} = \prod_{k=1}^{K} \frac{\left( \nu_k \right)^{n_k} e^{-\nu_k}}{n_k!} \]  

with \( \nu_k = \sum_{s=1}^{S} \nu p_s q_k^{(s)} \)  

\[ (6). \]
When the problem is formulated in this way, the likelihood function takes the form

\[ \ln L = \sum_{k=1}^{K} \left[ n_k \ln \nu_k - \nu_k \right] \quad \text{with} \quad \nu_k = \sum_v v_p q_k^{(v)}, \]

which depends on \( S \) independent parameters, that can be labeled, e.g.,

\[ v_p^{(v)} = v_p, \]

representing the expected numbers of events from sample \( s \) in the mixed sample. These parameters are no longer constrained explicitly, since they represent the numbers observed for a given running time and therefore can fluctuate independently.

Apparently no track is left of the constraint (2). However, the previous discussion shows that the likelihood defined by formulae (7) and (8) can be obtained by a one to one transformation of variables from the likelihood (5), namely

\[
\begin{align*}
\nu_1^{(v)} &= v_1 p_1 \\
\nu_2^{(v)} &= v_2 p_2 \\
\vdots &\quad \Rightarrow \quad \ldots \\
\nu_{S-1}^{(v)} &= v_{S-1} p_{S-1} \\
\nu_S^{(v)} &= v_S \left[ 1 - \sum_{i=1}^{S-1} p_i \right] \\
\nu &= \sum_{i=1}^{S} \nu_S^{(v)}
\end{align*}
\]

and therefore the values of the \( v_S^{(v)} \) that maximize it, can be expressed just replacing in the last formulae the parameters \( v, p_s \) that maximize (5). This implies that the maximization of (7) with respect to the parameters (8), automatically satisfies the normalization condition \( \sum_i \nu_i^{(v)} = N \). This could be proven by direct inspection.

### 2. The errors

If the problem is formulated with the multinomial approach, the covariance matrix of the \( S-1 \) selected \( p_s \) can be estimated, as usual, as the inverse of the \( (S-1) \times (S-1) \) matrix of second derivatives of the likelihood function

\[
\left[ -\frac{\partial^2 \ln L}{\partial p_i \partial p_m} \right]^{-1}
\]

Note the asymmetric treatment of the \( p_s \), due to the fact that one of them can be computed as a function of all others using constraint (2). One can, of course, augment the error matrix with an additional row and column, using error propagation on

\[ p_S = 1 - \sum_{i=1}^{S-1} p_i \]

but the error matrix thus obtained is singular.
If the problem is formulated with the poissonian approach, with the likelihood function (7), (8), the parameters estimated are not the fractions but the expected numbers of events\(^(*)\). One can estimate the $S \times S$ covariance matrix of the estimate $\hat{\nu}^{(i)}$ as

$$
\left[ - \frac{\partial^2 \ln L}{\partial \nu^{(i)} \partial \nu^{(m)}} \right]^{-1}
$$

The fractions can be computed a posteriori as

$$
p_i = \frac{\nu_i^{(i)}}{\sum_{s=1}^{S} \nu_s^{(i)}}
$$

and their covariance matrix using error propagation. Note that, if one does that, again the covariance matrix is singular, since definitions (9) are redundant.

### 3. The Pitfall

Since the estimates of the $\nu^{(i)}$ satisfy

$$\sum_{s=1}^{S} \nu^{(i)} = N,$$

one could be tempted to rewrite the likelihood (7) as a function of the $p_s$, setting $\nu = N$ in (8). One could then perform the minimization directly in terms of the $p_s$ (S parameters!) and obtain the error matrix directly, with no need to perform error propagation on the basis of (9). Doing that, one obtains the correct estimates for the $p_s$, but the error matrix is wrong. This will be shown below in an explicit example, but it is already apparent from the fact that this $S \times S$ error matrix is not singular.

Before coming to that example, it will be useful to show that the multinomial and the poissonian approach give consistent results, instead. I will perform an explicit calculation for the case $S=2$, where there is only one non-trivial parameter, e.g. $p_1$.

For the multinomial case

$$\ln L = \sum_{k=1}^{K} n_k \ln \left[ p_i q_k^{(i)} + (1 - p_i) q_k^{(2)} \right]$$

and the variance of $p_1$ is

$$\sigma_{\nu}^2 (p_1) = - \frac{1}{\partial^2 \ln L / \partial p_1^2} = \sum_{k=1}^{K} \frac{n_k (q_k^{(i)} - q_k^{(2)})^2}{\sum_{k=1}^{K} \left[ p_i q_k^{(i)} + (1 - p_i) q_k^{(2)} \right]^2}$$

For the poissonian approach

$$\ln L = \sum_{k=1}^{K} [n_k \ln (\nu^{(i)} q_k^{(i)} + \nu^{(2)} q_k^{(2)}) - (\nu^{(i)} q_k^{(i)} + \nu^{(2)} q_k^{(2)})]$$

The covariance matrix of $p_1$ and $p_2$, using the error matrix of $\hat{\nu}^{(i)}$ and $\hat{\nu}^{(2)}$ and performing error propagation on the basis of formula (9), is given by

\(^(*)\) In the likelihood (7) \(\nu\) and $p_s$ always appear in the combination $\nu p_s$, and it is not possible to determine them separately.
\[
\left[ \frac{\partial p_1}{\partial v^{(1)}}, \frac{\partial p_1}{\partial v^{(2)}} \right] \times \left[ -\frac{\partial^2 \ln L}{\partial v^{(1)^2}} - \frac{\partial^2 \ln L}{\partial v^{(1)} \partial v^{(2)}} \right]^{-1} \left[ \frac{\partial p_1}{\partial v^{(1)}}, \frac{\partial p_2}{\partial v^{(2)}} \right] = \left[ \frac{1}{\sum_{k=1}^{K} \frac{n_k q_k^{(1)^2}}{p_i q_k^{(1)} + (1 - p_i) q_k^{(2)^2}}} \right]^2 \times \left[ \begin{array}{cc} N & -N \\ -N & N \end{array} \right]
\]

This error matrix shows the expected features: the errors on \( p_1 \) and \( p_2 \) are the same (as they should since, \( p_2 = 1 - p_1 \)) and their correlation coefficient is 100%. For purpose of comparison with (10), we can compute the diagonal element in the asymptotic (\( N \to \infty \)) approximation, where \( n_k = N \left( p_i q_k^{(1)} + p_2 q_k^{(2)} \right) \).

For the multinomial case, this allows to rewrite (10) as
\[
\sigma_M^2(p_i) = \frac{1}{\sum_{k=1}^{K} \frac{n_k q_k^{(1)^2}}{p_i q_k^{(1)} + (1 - p_i) q_k^{(2)^2}}} \cdot \frac{1}{N} \frac{1}{\sum_{k=1}^{K} \frac{q_k^{(1)^2}}{p_i q_k^{(1)} + (1 - p_i) q_k^{(2)^2}}} \left( \frac{p_i q_k^{(1)}}{p_i q_k^{(1)} + (1 - p_i) q_k^{(2)^2}} \right)^2 - \left( \frac{p_i q_k^{(1)}}{p_i q_k^{(1)} + (1 - p_i) q_k^{(2)^2}} \right)^2 \sum_{k=1}^{K} \frac{n_k q_k^{(1) q_k^{(2)}}}{p_i q_k^{(1)} + (1 - p_i) q_k^{(2)^2}} \right]^2 \times \left[ \begin{array}{cc} N & -N \\ -N & N \end{array} \right]
\]

For the poissonian case
\[
\sigma_p^2(p_i) = \frac{1}{N} \left( \frac{1}{\sum_{k=1}^{K} \frac{q_k^{(1)^2}}{p_i q_k^{(1)} + (1 - p_i) q_k^{(2)^2}}} \right)^2 - \left( \frac{q_k^{(1)}}{p_i q_k^{(1)} + (1 - p_i) q_k^{(2)^2}} \right)^2 \sum_{k=1}^{K} \frac{q_k^{(1) q_k^{(2)}}}{p_i q_k^{(1)} + (1 - p_i) q_k^{(2)^2}} \right]^2 \times \left[ \begin{array}{cc} N & -N \\ -N & N \end{array} \right]
\]

Although I have not been able to prove that \( \sigma_M = \sigma_p \) identically\(^\dagger\), I could not detect any numerically significant difference between the two, in many numerical exercises that I performed. Conversely, it is very easy to prove that the poissonian approach, using the \( p \)'s as variables, is wrong, as far as errors are concerned. In that case, one has
\[
\ln L = \sum_{k=1}^{K} \left[ n_k \ln(N p_k q_k^{(1)} + N p_2 q_k^{(2)}) - N \left( p_i q_k^{(1)} + p_2 q_k^{(2)} \right) \right]
\]

where \( p_1 \) and \( p_2 \) are now parameters to be determined independently, since they are just another name for \( v^{(1)} \), \( v^{(2)} \). Their error matrix is given by

\(^\dagger\) One would not expect a formal identity, since it is not obvious that approximations like, e.g., the one implicit in the propagation of errors, have the same effects in the two cases.
This matrix does not satisfy the conditions $\sigma^2(p_1) = \sigma^2(p_2)$, nor the condition of 100% correlation between $p_1$ and $p_2$. It can also be seen that, at least for some values of $p_1$, the use of these formulae lead to a gross misestimate of the errors. Using the index $N$ to identify results obtained from the likelihood (12), in the asymptotic approximation for the $n_k$, one obtains

\[
\sigma^2_N(p_1) = \frac{1}{N} \sum_{k=1}^{K} \left[ \frac{q_k^{(2)} }{ p_1 q_k^{(1)} + (1-p_1) q_k^{(2)} } \right]^2
\]

I computed numeric values for a particular example, where the pdf's of the pure samples are both linear functions of a random variable $x$ contained in $0 \leq x \leq 1$

\[
f^{(1)}(x) = 2x
\]

\[
f^{(2)}(x) = 2(1-x)
\]

The results are shown in Figure 1 as a function of the value of $p_1$.

**Figure 1**

Comparison of the error computed with the incorrect poissonian approach, $\sigma_N(p_1)$, and the multinomial approach, $\sigma_M(p_1)$, in the numerical example considered. The quantity plotted is the ratio of the two errors. By incorrect poissonian approach I mean the one where the expectation values of the numbers of events in the likelihood are written as the measured total number of events times the fractions to be determined.
The fact that likelihood (12) gives the right estimates of the fractions, but a wrong estimate for the errors needs some explanation. If one wants to interpret (12) as a likelihood function, the symbols $p_s$ that appear in it cannot be interpreted directly as the event fractions, but represent just a linear change of variables

$$\nu^{(i)} = Np_s$$

in the expression of the correct likelihood (11). Such a change of variables is legitimate, whatever the value of the arbitrary constant $N$ is. The values of the $p_s$ that maximize (12), call them $\hat{p}_s$, are related to those of of the $\nu^{(i)}$ that maximize (11), $\hat{\nu}^{(i)}$, by

$$\hat{p}_s = \frac{\hat{\nu}^{(i)}}{N}$$

and cannot be interpreted directly as event fractions. Instead, these will be obtained from formula (9), which, in terms of the $\hat{p}_s$ reads

$$p_s = \frac{\hat{p}_s}{\sum_{s=1}^{S} \hat{p}_s}.$$

In the special case where the arbitrary constant $N$ is the measured total number of events, we can use the previous result that, at the likelihood maximum,

$$\sum_{i=1}^{S} \hat{\nu}^{(i)} = N$$

and, as a consequence, $\sum_{i=1}^{S} \hat{p}_s = 1$.

However these relations are valid only at the likelihood maximum and therefore it is wrong to assume them for the error calculations that concerns the behaviour of the likelihood away from the maximum, where the $\nu_i$ are unconstrained parameters.

4. Conclusions

Inconsistencies observed in the errors provided by the HMCMLL and TFractionFitter packages are not related to the formulation of the likelihood, but to an incorrect replacement, in the likelihood function, of expected numbers of events with the total measured number of events times the expected fractions. The extended likelihood approach, based on the use of the Poisson distribution will give the right result if the event fractions are computed by formulae (9).

The results provided by these packages are valid for what concerns the estimates of the event fractions, but are incorrect for what concerns the errors, because they are based on the assumption that the normalization condition for the parameters incorrectly interpreted as event fractions, which holds only at the likelihood maximum, is valid everywhere. As a practical remark, the correct errors can be computed from the covariance matrix provided by these packages, applying error propagation to the formula

$$p_s = \frac{\hat{p}_s}{\sum_{s=1}^{S} \hat{p}_s}.$$

Note, however, that the full covariance matrix of the $\hat{p}_s$ must be used in this.

5. Acknowledgements

I wish to thank S. Goy Lopez and M. Raggi for bringing this problem to my attention and for useful discussions.