DELPHI Collaboration

DELPHI 2002-0 XXXX 000
8 March, 2002

Exact optimization of likelihood analyses in a unified Bayesian-Frequentist Monte Carlo Framework: The "k2-filter"

Niels Jørgen Kjær
CERN, CH - 1211 Geneva 23, Switzerland
e-mail: Niels.Jorgen.Kjaer@cern.ch
Jorgen D'Hondt
IIHE (ULB-VUB), Pleinlaan 2, BE-1050 Brussels, Belgium
e-mail: Jorgen.D'Hondt@cern.ch

Abstract

This paper presents a generalization of the Kalman filter process control to a situation where the stochastic parts are not analytically computable and only available as Monte Carlo generated events.

In a Monte Carlo framework where the exact Matrix elements, $P_i(x)$, are known on an event-by-event basis the maximum likelihood technique provides the optimal analysis. To $\mathcal{O}(1/n)$, where n is the number of events the Fisher information can be computed as:

$$\mathcal{I} = 1/\sigma^2 = \sum_i g_i^2, g_i = \frac{d\ln(P_i(x))}{dx}|x_0$$

g_i is thus exactly computable from the generation information. When the measurement result, $E(x)$, is close to the reference point, x_0, g_i contains all the needed information for the estimation of $E(x)$. On the analysis level this paper shows that the error is given by:

$$\mathcal{I} = 1/\sigma^2 = \sum_i [g_i^2 - (a_i - g_i)^2],$$

where $a_i = \frac{d\ln(L_i(x))}{dx}|x_0$ is the information on the analysis level. Therefore, any likelihood analysis is optimized by maximizing $\mathcal{I}$. The optimization is explicitly linear in a_i and g_i, which leads to solutions of simple linear equations as in other Kalman filter cases.

The optimization procedure assumes an initial analysis. Any set of parameters used in the initial analysis can be optimized and the analysis output can be optimized as a function of any observable simultaneously extracting all information content of the observables.

The analyses can be optimized for any value of x_0 through reweighting of the Monte Carlo events, leading to the final likelihoods being integrals of the type $L_i(x) = \exp(\int a_i(x_0)dx_0)$.

Generalizing the optimization procedure to simultaneous measurements of several observables is shown to give complications which rise only linearly as a function of the number of observables.

The Bayesian optimization procedure itself is controlled using a Frequentist's approach, which is simply applying the Jackknife in case of uncorrelated events. This procedure is shown to give exact estimations of any potential overtraining in the optimization procedure. Finally, several examples and recommendations for practical use are given.

Contributed paper to "Advanced Statistical Techniques in Particle Physics",
Grey College, Durham, 18 - 22 March 2002

1 Introduction

In contemporary high energy physics the analyses are becoming more and more sophisticated in the search for better precision. The central tool in most physics analyses is the Monte Carlo (MC) generator. MC is not only used as an integrator for determining expectation values, but also for determining analysis performance. This is true when the analyses are statistically limited. When significant systematics are involved the picture becomes more complex.

Most MC generators work generally in an sequential way:

1. Initial state radiation (and parton density functions for hadron colliders).
2. Electro-weak (or QCD) hard scattering processes.
3. Final state radiation and possible interferences.
4. Fragmentation, hadronization, and decays.
5. Particle interactions and measurements in the detectors.

The important point is that these generators are capable of generating independent events with unit weight, which can be compared directly with the data. The unit weights are obtained using complicated integration techniques, but at the same time the absolute probabilities, i.e. the matrix element squares are also available for a given event.

The absolute event probabilities are not of much relevance compared to their dependency of the quantities we wish to measure in an analysis. The list of quantities for which this probability information is available is very long. This paper concentrates on a few related to the measurement of m_{W}[1][1] at LEP2. Here, the matrix element squared is dominated by a double Breit-Wigner from the two W decays. When the value of the nominal m_{W} is changed from $m_{\mathrm{W}}^{\mathrm{gen}}$ to $m_{\mathrm{W}}^{\mathrm{new}}$ each event with generated W masses m_1 and m_2 changes its weight according to:

$$w = \frac{P(m_{\mathrm{W}}^{\mathrm{new}})}{P(m_{\mathrm{W}}^{\mathrm{gen}})} \simeq \frac{\mathrm{BW}(m_1|m_{\mathrm{W}}^{\mathrm{new}}) \cdot \mathrm{BW}(m_2|m_{\mathrm{W}}^{\mathrm{new}})}{\mathrm{BW}(m_1|m_{\mathrm{W}}^{\mathrm{gen}}) \cdot \mathrm{BW}(m_2|m_{\mathrm{W}}^{\mathrm{gen}})}. \tag{1}$$

In practice these matrix element squares are available as functions with the full numerical precision of the generator.

If these generated masses are precisely observable, the optimal way to perform an analysis is simply the likelihood technique. In case of a significant number of events the central limit theorem can be invoked and the optimal estimator becomes the maximum likelihood (ML), i.e. the value of m_{W} for which

$$P(m_{\mathrm{W}}) = \prod_{i=1}^{n} \frac{P_i(m_{\mathrm{W}})}{P_i(m_{\mathrm{W}}^{\mathrm{gen}})} \tag{2}$$

is maximal, where the product is over all events, i.

For later purposes it is simpler to work in a linear framework where:

$$\ln P(m_{\mathrm{W}}) = \sum_{i=1}^{n} \ln\left(\frac{P_i(m_{\mathrm{W}})}{P_i(m_{\mathrm{W}}^{\mathrm{gen}})}\right), \tag{3}$$

[1]Subsequently m_{W} is used for the measured quantity, but can be replaced by any other.

which can be expanded to give:

$$\ln P(m_{\rm W}) \simeq \sum_{i=1}^{n} 1 + \frac{\partial \ln(P_i(m_{\rm W}))}{\partial m_{\rm W}}|_{m_{\rm W}^{\rm gen}}(m_{\rm W} - m_{\rm W}^{\rm gen}) + \frac{1}{2}\frac{\partial^2 \ln(P_i(m_{\rm W}))}{\partial m_{\rm W}^2}|_{m_{\rm W}^{\rm gen}}(m_{\rm W} - m_{\rm W}^{\rm gen})^2 \tag{4}$$

Defining $g_i = \frac{\partial \ln(P_i)}{\partial m_{\rm W}}|_{m_{\rm W}^{\rm gen}}$ and $g_{i2} = -\frac{\partial^2 \ln(P_i)}{\partial m_{\rm W}^2}|_{m_{\rm W}^{\rm gen}}$[2] gives the ML $m_{\rm W}$ estimator as:

$$m_{\rm W} - m_{\rm W}^{\rm gen} = \frac{\sum_i g_i}{\sum_i g_{i2}}. \tag{5}$$

Eq. 5 is only valid in the limit $m_{\rm W} - m_{\rm W}^{\rm gen} \to 0$. In order to achieve this the MC can be "reweighted" to a new value of $m_{\rm W}^{\rm gen} = m_{\rm W}$, and the procedure iterated to give the final result. The result using events with weights, w_i[3], is simply:

$$m_{\rm W} - m_{\rm W}^{\rm gen} = \frac{\sum_i w_i g_i}{\sum_i w_i g_{i2}}. \tag{6}$$

The MC reweighting scheme can also be used to do a ML analysis if the generator information is not directly accessible. When only a set of observables $\vec{o}_i$ is available for each event, i, the likelihood can be estimated as the predicted density of events in the observable space $\vec{o}$:

$$L_i(m_{\rm W}) = \sum_j w_j(m_{\rm W})\delta(\vec{o}_i - \vec{o}_j). \tag{7}$$

In principle this procedure is an elegant way of extracting the analysis information from the generator information. In practice it is a delicate procedure to define the requirements on $\delta(\vec{o}_i - \vec{o}_j)$. Therefore most LEP2 analyses[2] have been restricted to a few observables. Another interesting point to observe is that a reweighting analysis is not automatically unbiased if the events selected by the analysis depend on the value of $m_{\rm W}$. This is however easily corrected for by dividing the individual event likelihoods L_i by the sum of the likelihoods of all the selected MC events.

Returning to the linear ML estimator in Eq. 5 it is easy to determine the statistical properties of such analysis. The standard error is given by:

$$\mathcal{I} = \frac{1}{\sigma^2} = \sum_i g_{i2}. \tag{8}$$

In the limit $\mathcal{O}(1/n)$ this error can also be computed directly from the variance of g_i, which is just the Fisher information:

$$\mathcal{I} = \sum_i g_i^2, \tag{9}$$

which shows that in a linear approximation all the ML analysis properties can be extracted from the distribution of g_i. Since the value of g_i also determines how much the information from a single event changes the result, we will denote this the "action" of event, i, on the generator level. For a ML analysis the similar quantity, a_i, can be defined on the analysis level via the expansion:

$$\ln L(m_{\rm W}) \simeq \ln L(m_{\rm W}^{\rm gen}) + \sum_i w_i[-a_i(m_{\rm W} - m_{\rm W}^{\rm gen}) + \frac{1}{2}a_{i2}(m_{\rm W} - m_{\rm W}^{\rm gen})^2] \tag{10}$$

[2]The sign convention is chosen so that events with positive g_i will increase the measurement
[3]Weights will be only included in the following formulae, when it is not trivial how to apply them

It is the purpose of this paper to present strategies to find the optimal determination of a_i as a function of event observables $\vec{o}_i$. This optimization is done using the information on the generator level g_i. Since all information is explicitly linearized an obvious answer to the optimization is the Kalman filter (KF)[3]. The KF is a process which is used to optimize processes. This Bayesian algorithm basically differentiates the process with respect to the parameters steering the process and uses the variation to estimate an optimum. Since it works in linear approximations this leads to the solution being obtained from simple matrix equations.

The application of the KF presented in this paper is more involved. The first Kalman step is already done in this introduction where a_i and g_i are the changes of the global ML analysis adding event i. The expression of what is optimal is derived in Section 2. The next step is to take an existing analysis and optimize all its parameters which is described in Section 3.1. The analysis can also be optimized directly on event observables which is detailed in Section 3.2. The simple extension to a non-linear likelihood is described in Section 3.4, while more general non-linear properties are briefly discussed in Section 3.7. Finally, this paper gives a few practical examples in Section 4.

An important point in connection with the use of KF techniques is to control the optimization. This is in the literature normally referred to as "Robustness" of the KF. In this paper the KF is controlled by requiring that only statistically significant updates to the process are allowed. The significance is estimated from the variance using frequentist methods on the generated MC samples. In a linear framework the variance estimator is the classical one, while it is the Jackknife estimator in non-linear cases. This unification of Bayesian and Frequentist approaches in one analysis is a feature which seems to be possible and consistent uniquely in a MC framework.

2 Derivation of the analysis performance

In investigating the performance of an analysis it is of use to perform a differential reweighting of the events:

$$\frac{\partial w_i}{\partial m_{\mathrm{W}}}|_{m_{\mathrm{W}}=m_{\mathrm{W}}^{\mathrm{gen}}} = -g_i w_i \tag{11}$$

On the full MC sample a number of useful quantities can be derived:

$$w_t = \sum_i w_i,\ \ T_1 = \sum_i w_i a_i,\ \ T_2 = \sum_i w_i a_{i2}, \tag{12}$$
$$T_3 = \sum_i w_i g_i a_i,\ \ T_4 = \sum_i w_i g_i a_{i2},\ \ T_5 = \sum_i w_i {a_i}^2$$

Then all the relevant properties of the analysis can be computed. Firstly, the bias of the analysis which should be zero:

$$\mathtt{BIAS} = T_1/T_2 \tag{13}$$

Secondly, the slope of a calibration curve, which is computed as minus the derivative of the bias w.r.t. the value of the generated mass which should be unity:

$$\mathtt{SLOP} = T_3/T_2 - T_4 \cdot T_1/T_2^2 \tag{14}$$

The error predicted by the analysis is given by:

$$\mathtt{PERR} = 1/\sqrt{T_2}, \tag{15}$$

while in reality it gives an r.m.s. of the pull, defined as the measured value minus the generated one divided by the predicted error, which should be unity and computed as:

$$\texttt{PULL} = \sqrt{T_5/T_2} \tag{16}$$

Correcting the predicted error to the true error gives:

$$\texttt{ERROR} = \texttt{PERR} \cdot \texttt{PULL}/\texttt{SLOP}, \tag{17}$$

or:

$$\texttt{ERROR}^2 = \frac{T_5 T_2^2}{(T_3 T_2 - T_4 T_1)^2} \tag{18}$$

When the analysis is unbiased $T_1 = 0$, while $T_2 = T_3 = T_5$ when the analysis is fully calibrated with both SLOP and PULL equalling unity. For the fully calibrated analysis the information then becomes $\mathcal{I} = T_2 = T_3 = T_5$. This is made most transparent when written as:

$$\mathcal{I} = \sum_i w_i (g_i^2 - (a_i - g_i)^2) \tag{19}$$

Eq. 19 serves as the master formula for the subsequent analysis optimization. Not unlike the minimum variance bound the r.h.s. of Eq. 19 is in reality a limit for the information, which is only reached when the above mentioned constraints are fulfilled.

2.1 The basic analysis constraints

The two analysis constraints, $T_1 = 0$ and $T_3 = T_5$, are most conveniently written as:

$$\begin{aligned} \sum_i w_i a_i &= 0 \\ \sum_i w_i a_i \cdot (a_i - g_i) &= 0 \end{aligned} \tag{20}$$

The second constraint is not linear in a_i, which means that including this constraint in an optimization procedure will force the algorithm to become iterative. In this case the simplest way of implementing the constraints is an iterative constrained fit using Lagrange multipliers.

Corrections to the analysis can be constructed in such a way that the bias on and scale of a_i automatically become zero and unity. This can be done by including two parameters, α and β, in the expression:

$$a_i' = \alpha a_i + \beta, \tag{21}$$

and minimize $\sum_i w_i (a_i' - g_i)^2$ w.r.t. α and β.

Either of the two approaches are valid and the best suited for a problem should be adapted.

3 Implementations of the optimization

There is an infinite number of different ways to optimize an analysis using Eq. 19. A natural way would be to apply a standard neural network and perform the optimization using $\sum_i w_i(a_i - g_i)^2$ as the penalty function. But this does not take advantage of the fact that linearity is already built into the actions, a_i and g_i.

A more appealing way is to benefit from the linearity by only computing corrections to an ongoing analysis. These corrections will normally be small and in themselves linear. If the corrections become large this is in itself not a problem, since it shows exactly where large improvements of the analysis can happen. These large analysis corrections should then be understood and evaluated separately in order to control systematic effects. This section describes two ways of doing this correction optimization. The first one assumes an already developed analysis, and the second constructs the analysis by expansions in the observable space.

3.1 Corrections to analyses

Assume a complicated analyses to begin with. This can consist of thousands lines of code and many parameters, $\vec{p}$[4], which have either been established from first principle or been estimated from the Monte Carlo itself. One can then test if all these parameters are optimal by varying the analysis changing the parameters one by one and then recompute all the results. This gives linear estimators for the analysis dependency on the parameters:

$$a_i(\vec{p} + \Delta\vec{p}) = a_i(\vec{p}) + \frac{\partial a_i(\vec{p})}{\partial \vec{p}} \Delta\vec{p} \tag{22}$$

The information to be optimized is written as:

$$\mathcal{I} = \sum_i w_i (g_i^2 - (a_i(\vec{p}) + \frac{\partial a_i(\vec{p})}{\partial \vec{p}} \Delta\vec{p} - g_i)^2), \tag{23}$$

with its derivatives:

$$-\frac{1}{2}\frac{\partial \mathcal{I}}{\partial \Delta\vec{p}} = \left(\sum_i w_i \frac{\partial a_i(\vec{p})}{\partial \vec{p}} \cdot \frac{\partial a_i(\vec{p})}{\partial \vec{p}} \right) \Delta\vec{p} + \sum_i w_i \frac{\partial a_i(\vec{p})}{\partial \vec{p}} \cdot (a_i(\vec{p}) - g_i), \tag{24}$$

from which optimal corrections can be computed as:

$$\Delta\vec{p} = \left(\sum_i w_i \frac{\partial a_i(\vec{p})}{\partial \vec{p}} \cdot \frac{\partial a_i(\vec{p})}{\partial \vec{p}} \right)^{-1} \sum_i w_i \frac{\partial a_i(\vec{p})}{\partial \vec{p}} \cdot (g_i - a_i(\vec{p})) \tag{25}$$

In the case of additional (non-linear) constraints, the solution becomes iterative. Since the expected corrections are small this procedure becomes very stable and can accomodate a very sizable number of parameters for simultaneous optimization. Using the Jackknife and $\mathbf{V} = (\sum_i w_i \frac{\partial a_i(\vec{p})}{\partial \vec{p}} \cdot \frac{\partial a_i(\vec{p})}{\partial \vec{p}})^{-1}$ gives the covariance expression:

$$\sigma_{jk}^2 = \sum_l \sum_m \mathbf{V}_{jl} \mathbf{V}_{km} \sum_i w_i \frac{\partial a_i(\vec{p})}{\partial p_j} \frac{\partial a_i(\vec{p})}{\partial p_k} \cdot (g_i - a_i(\vec{p}))^2, \tag{26}$$

which is very useful to estimate the significance of the correction to the parameter given by: $|\Delta p_j|/\sigma_j$

[4]This can in principle be any constant number or even parameterizations of functions in the analysis code.

3.2 Building a whole analysis

Another way to construct an optimal analysis is to start from scratch. The analysis is simply $a_i = 0$. Then search directions in observable space where there is the most information. If o_1 is one such direction the optimal analysis in that direction is obtained by constructing:

$$a_i(o_1) =< g_i > (o_1). \tag{27}$$

Then proceed by investigating another direction in observable space and compute how the first analysis should be corrected as a function of this observable to provide an optimal analysis with this new information. The new analysis becomes $a_i = a_i(o_1) + a_i(o_2)$, which is optimal for:

$$a_i(o_2) =< g_i - a_i(o_1) > (o_2). \tag{28}$$

This suggests an iterative algorithm which produces optimized analyses which take into account linear correlations between event observables. When all significant directions in observable space have been used it can then be tested if there remain higher order correlations which contain still more information. The most important correlations are typically found between observables and the actions themselves. The reason for this is that the actions already contain the information of all the other directions in observable space. It is therefore most relevant to observe if there are possible corrections like:

$$\Delta a_i(a_i, o_k) =< g_i - a_i > (a_i, o_k). \tag{29}$$

3.3 Combining different analyses

As seen from the above it is now simple to optimally combine two different ML analyses with this framework. If the results of the two analyses are given by $a_i(1)$ and $a_i(2)$ they project optimal directions in the observable space and the combined analysis is simply:

$$a_i(a_i(1), a_i(2)) =< g_i > (a_i(1), a_i(2)). \tag{30}$$

If one analysis, $a_i(1)$, is dominating it is more useful to make a correction on this analysis rather then creating a new combined one:

$$\Delta a_i(1)(a_i(1), a_i(2)) =< g_i - a_i(1) > (a_i(1), a_i(2)), \tag{31}$$

which has the advantage that the statistical significance of correcting analysis $a_i(1)$ can be computed directly.

3.4 Optimizing globally

Until now, the analysis has only been considered locally around $m_{\mathrm{W}} = m_{\mathrm{W}}^{\mathrm{gen}}$. The MC events can be reweighted according to the information $P_i(m_{\mathrm{W}}^{\mathrm{gen}})$ to any other value of $m_{\mathrm{W}}^{\mathrm{gen}}$, and reoptimized at that point. Then the global likelihood simply becomes:

$$L_i(m_{\mathrm{W}}) = \exp\left(\int^{m_{\mathrm{W}}} a_i(m_{\mathrm{W}}^{\mathrm{gen}}) dm_{\mathrm{W}}^{\mathrm{gen}}\right) \tag{32}$$

Typical analyses have a tendency to be linear w.r.t. $m_{\mathrm{W}}^{\mathrm{gen}}$, which facilitates the construction of a optimization grid. In typical analyses it is sufficient just to have two points of optimization. In other cases, like near phase space boundaries, a more complicated grid will be needed.

3.5 Optimizing correlated measurements

One of the most appealing features of the event-by-event analysis is that it automatically takes into account all correlations between the measurements. In the case of two parameter fits, x and y, two analysis actions are needed, a_i and b_i, with the generator information given by: g_i and h_i. Then the total information simply becomes:

$$\mathcal{I}_x = \sum_i w_i(g_i^2 - (a_i - g_i)^2), \quad \mathcal{I}_y = \sum_i w_i(h_i^2 - (b_i - h_i)^2), \tag{33}$$

which can be optimized independently.

In the case of global optimization, it is no longer possible in a unique way to simultaneously optimize the analysis parameters. A choice has to be made whether to optimize to a combination of $\mathcal{I}_x$ and $\mathcal{I}_y$, or choose to have two independent optimizations, which is the most performant and easiest to handle.

Since the analyses can become independent the complications of a multi-dimensional measurement is just linear in number of parameters. One complication remains if the measurement has to be global in many parameters. In this case care has to be taken in the choice of the global grid.

3.6 Applying additional constraints

The two constraints used in the standard analysis can easily be generalized to include other constraints on the analysis. Typical examples are independency of parameters known on the generator level and independence of special event samples, typically backgrounds. When including background events in the optimization they are assigned $g_i = 0$. The analysis can then be made independent of the level of background by substituting the bias constraint,

$$\sum_{i \in s,b} w_i a_i = 0, \tag{34}$$

by two different constraints:

$$\sum_{i \in s} w_i a_i = 0 \quad \wedge \quad \sum_{i \in b} w_i a_i = 0. \tag{35}$$

3.7 Non-linear optimization

There is no obvious reason why this MC analysis optimization cannot be extended to non-linear problems. The procedure can be expanded to higher orders in the expansion parameters, and specific treatments can be invoked for special non-linear problems. However, such a discussion is beyond the scope of this paper.

3.8 Systematic errors and finite Monte Carlo samples

Using procedures like the k2-filter there are two significant challenges to make the final analysis optimal:

- Systematic errors in the MC
- Limited MC statistics

There might be features in the MC which do not describe the data, and which in the MC model have sensitivity to the measurement. In that case there is no way to identify this problem using only the procedure described in this paper. Either the analysis must by hand be constructed using observables that are under control w.r.t. systematics, or the systematics can be estimated within the MC model. The latter approach is addressed in [4].

There is one substantial advantage of the k2-filter approach w.r.t. systematics. The final analysis is not a "black box" process providing one number, but consists of a set of informations and corrections. Each iterative step in the k2-filter yields a contribution to the information, $\mathcal{I}$, and a change of the final number. All these corrections have numerical values when analyzing the data. Therefore the k2-filter provides useful information on the data-MC agreement at each step in terms of the change of the final measured value. This means that each individual step of the analysis can be verified individually. This provides a very useful tool to estimate potential systematics from the data itself.

Assuming that the MC model systematics can be kept under control there still remains one significant problem with algorithms based on finite MC training. This is commonly referred to as "overtraining". It happens when there are:

- Too few generated MC events.
- Too many free parameters in the optimization.
- Single generated events which have anomalous values of some observables, also known as outlayers.
- $\mathcal{I}_a \ll \mathcal{I}_g$.

All these four effects are correlated, and lead to a potential overtraining of the analysis. Overtraining means that the analysis is estimated to be closer to the generator than it actually is. The overtraining can be evaluated by using independent MC samples to perform the training and the overtraining evaluation. This is however a delicate procedure which does not take advantage of the full MC statistics in an optimal way.

Using MC samples with infinite statistics clearly solves all the above problems. For Monte Carlo samples with a fixed number of events, the other three problems have to be addressed.

- Rare outlayer events are easy to identify, and can be controlled by setting the value of an anomalous observable to a reasonable value without influencing the optimization performance.
- $\mathcal{I}_a \ll \mathcal{I}_g$ means that there is much more information on the generator level than on the analysis level and the optimization becomes unstable. In order to still have useful information for the data analysis it must be required that $n_{\mathrm{MC}}/\mathcal{I}_g > n_{\mathrm{data}}/\mathcal{I}_a$, where n_{MC} and n_{data} are the numbers of MC and data events. If this is not the case it is necessary to alter the MC generation strategy as described in Section 3.9.

Assuming that the above problems have been solved only the problem with too many parameters remains. Within the k2-filter framework it is straight forward to compute the potential overtraining from the full MC sample. Assume that one correction using one parameter is proposed, changing the action from a_i to a'_i. Then

$$\mathcal{I}' - \mathcal{I} = \sum_i w_i[-a'_i(a'_i - 2g_i) + a_i(a_i - 2g_i)] \tag{36}$$

When the change is small the variance of $\mathcal{I}' - \mathcal{I}$ can be computed using the Jackknife to be:

$$\mathtt{VAR}(\mathcal{I}' - \mathcal{I}) = \sum_i w_i^2[-a'_i(a'_i - 2g_i) + a_i(a_i - 2g_i)]^2, \tag{37}$$

and the significance, $\mathtt{S}$, estimated as

$$\mathtt{S} = \frac{\mathcal{I}' - \mathcal{I}}{\sqrt{\mathtt{VAR}(\mathcal{I}' - \mathcal{I})}} \tag{38}$$

The expected improvement is then $\mathcal{I}' - \mathcal{I}$, while the real improvement, which can be estimated using larger independent MC samples, is in average smaller and can be estimated[5] to be reduced in average by a factor $\mathtt{S}^2/(1 + \mathtt{S}^2)$. As a good practice, only corrections with a significance $\mathtt{S} > 3$ should be applied. Notice that such an approach also reduces the influence of outlayers, which will be insignificant when estimated by the Jackknife.

In general it is recommended that the optimization should be performed according to what is significant and not according to what gives the largest improvement in the information $\mathcal{I}$.

3.9 Monte Carlo samples without reweighting

Until now, it was assumed that normal MC reweighting can be performed. However, there are three classes of processes which do not comply to this assumption.

- $\mathcal{I}_g \gg \mathcal{I}_a$. This occurs when there is “too much” information on the generator level. Take as an example a hypothetical “precision” measurement of the Higgs boson mass for $m_{\mathrm{H}} = 116$ GeV/c^2. In this case the Higgs boson width is only a few MeV/c^2, which leads to big fluctuations in g_i not reflected in information of how to obtain the optimal a_i.

- The event matrix element squares cannot be computed as a function of the parameter of interest. An example of such a situation is the tuning of MC QCD parameters using the LEP1 Z^0 data. The events cannot be reweighted according to a matrix element, because the whole complicated parton shower is governed by these parameters in a very complex way. The problem was to estimate 5 parameters using all the event shape information of the data. The solution was to make a grid in the 5 dimensions and generate events at each point. The grid was so large that one could not use fully simulated events and had to compare unfolded data to the generators. The solution described below can be applied to such problems and gives a good motivation for the rather complicated procedure.

- There is no obvious parameter which can be measured or reweighted. This is the case when an analysis is performed to estimate which of two models is the most correct.

[5] in a Bayesian calculation

The last problem is actually the easiest, since it is solvable by the amplitude method. An amplitude parameter, $\mathcal{A}$, is introduced to infer which of two models is the most likely. In this linear implementation each event from one model obtains a matrix element squared equal to $\mathcal{A}$, while the events generated by the other model have $1-\mathcal{A}$. This approach is forced to assume a linearity, which is not present in the first two problems.

The solution to the first two problems is more involved. First notice that the two problems are basically the same. Events can only be generated with essentially fixed numbers for the MC generator parameter. A "dummy" generator function can be introduced to regularize the problem. The likelihoods on the generator and analysis level described above are then complimented by two new likelihoods:

- The dummy analysis likelihood, which is the analysis likelihoods convoluted with the dummy likelihood.
- The dummy generator likelihood, which is the generator likelihood convoluted with the dummy likelihood.

If the dummy analysis information is then optimized to the dummy generator information, the analysis information is automatically optimal w.r.t. the real generator information. It is also possible to optimize the analysis information to the dummy generator information, and after which the dummy likelihood is deconvoluted to obtain the analysis likelihood. It is however good practice to avoid introducing deconvolution algorithms whenever it is possible.

The choice of the dummy generator function should be such that:

- $\mathcal{I}_d > \mathcal{I}_a$, which insures that the optimization can proceed with linear approximations.
- $n_{\mathrm{MC}}/\mathcal{I}_d > n_{\mathrm{data}}/\mathcal{I}_a$, which insures that there is enough MC information to optimize the analysis.

If $\mathcal{I}_d \simeq \mathcal{I}_a$, the analysis can no longer easily be optimized at one single value of the "dummy" reweighting parameter. In that case the optimization has to be made purely to the convoluted likelihoods like:

$$\mathcal{I} = \int dx_0 d_i(x_0)(ad_i(x_0) - gd_i(x_0))^2, \tag{39}$$

the details of which lies beyond the scope of this paper.

4 Practical examples

4.1 Analysing the DELPHI m_{W} analysis

The DELPHI m_{W} measurement[1] in the 4-quark final state has evolved over several years becoming a complicated Bayesian convolution analysis. It has many ingredients of which some are:

- Event selection unbiased w.r.t. m_{W}.
- Event purity estimation unbiased w.r.t. m_{W}.

- Several alternative jet clustering algorithms to improve the knowledge of jet ambiguities and additional gluon jets.
- Several constrained kinematical fits to investigate different hypotheses.
- A combination of the above information to provide likelihoods for the experimental mass information, the so-called "ideogram".
- A convolution with the theoretical prediction with m_{W} as a free parameter.

The analysis provides for each event the analysis action, a_i, estimated for $m_{\mathrm{W}} = 80.35$ GeV/c^2. On a sample of 125000 MC events it was then tested if the analysis was optimal w.r.t. the following observables:

- The a_i of the analysis.
- The a_{i2} of the analysis.
- The χ^2 of one constrained fit.
- The observed multiplicity and energy in the event.
- The jet resolution parameters, y_{23}, y_{34}, y_{45}.
- Information of possible ISR photons from the constrained fit.
- The estimated purity of the event.

From this optimization test, ten corrections came out as being significant with $\mathtt{S} > 3$. However the total improvement to $\mathcal{I}$ was only 2%, which is 1% improvement in the statistical error. Due to the increased complexity it was decided not to include this improvement in the final analysis. This lack of improvement is not necessarily negative. The positive aspect is that it shows that the measurement was already correctly optimized.

In addition to the observables the parameters of the analysis were also tuned. It was attempted to optimize the parameters used in the constrained fit, but this did basically not lead to any significant improvement.

The m_{W} analysis uses three different jet algorithms in parallel assuming equal statistical weights for the final result. Two additional versions of the analysis were performed with different jet clustering weights and the analysis was optimized with the procedure described in section 3.1. The outcome of optimization was that the optimal weights should be $(0.45, 0.36, 0.19)$. The numbers have uncertainties of about 0.02.

4.2 Constructing an analysis from scratch

The next example is from a measurement of a model dependent colour reconnection parameter also using W mass information. The SK1 model [5] predicts final state interactions between two hadronically decaying W bosons. This model predicts shifts in the observed value of m_{W} and other event observables. The model has a nearly free parameter κ, which predicts the probability for the phenomenon to occur as a function of an event property which is computable from the string configuration on the generator level.

This measurement of κ is not to be considered a colour reconnection measurement but purely a model estimation, to be used for further evaluations. Each event is generated in two versions, one normal and one reconnected. The probability of each of the versions is then computable for each individual event.

In order to estimate κ one can use observables related to the m_{W} measurement. In particular it is expected that an alternative m_{W} estimator using a different jet algorithm (cone) leads to a different bias at a certain fixed value of kappa, compared to the standard (std) m_{W} estimation. In order to check the sensitivity, four event m_{W} observables were extracted: $a_1(\mathrm{std})$, $a_1(\mathrm{std}) - a_1(\mathrm{cone})$, $a_2(\mathrm{std})$, and $a_2(\mathrm{std}) - a_2(\mathrm{cone})$. In order not to be sensitive to the value of m_{W} itself a normal analysis would just look on the most sensitive observable: $a_1(\mathrm{std}) - a_1(\mathrm{cone})$. The k2-filter can be applied to build the analysis as expansions in the 4-dimensional observable space, and using independence of m_{W} as a constraint in the optimization will still insure the same property as the classical approach. As an example Figure 1 shows the sensitivity to the normal $a_1(\mathrm{std}) - a_1(\mathrm{cone})$ observable and the sensitivity to additional information like the width estimator, $a_2(\mathrm{std})$, in the right hand plot. Combining all the information optimally creates a full analysis, which provides a final sensitivity to κ which has nearly a factor 2 smaller error compared to the usual one dimensional analysis.

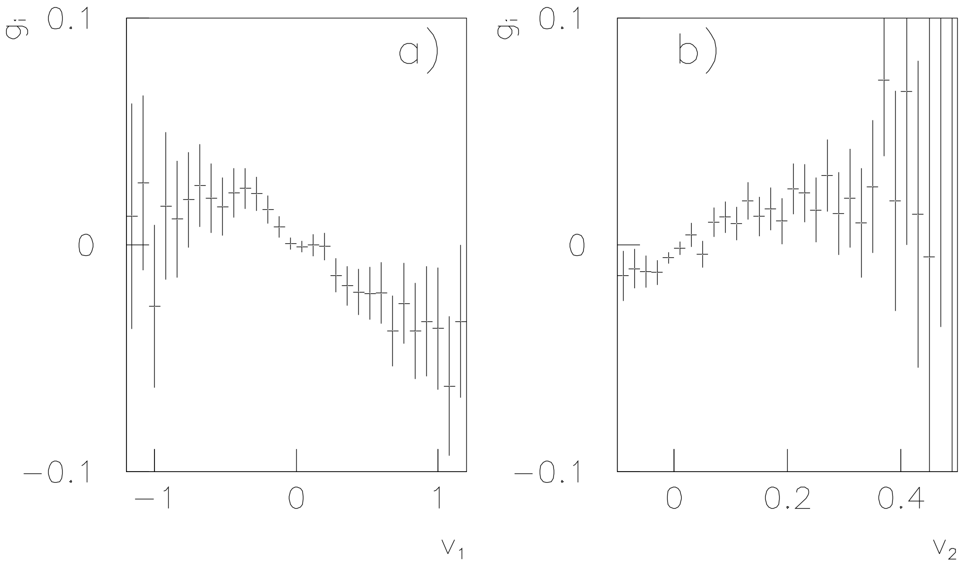

Figure 1: The mean value of the action on κ evaluated at $\kappa = 0.66$ are shown for two event observables. In a) the observable is the difference between the cone and the std m_{W} estimator, and b) is the std width estimator.

4.3 Searches and event classification

Even though this paper concentrates on measurements, it will be briefly mentioned that the methods can to a large extent also be used in search and event classification analyses. In a search for a hypothetical signal, s, a candidate event, i, has the linear action:

$$a_i = \ln(1 + s_i/b_i), \tag{40}$$

where s_i/b_i is the estimated likelihood ratio for the individual event. In more general event classification analyses the expression becomes symmetric:

$$a_i = \ln(s_i/b_i). \tag{41}$$

These actions have the linear properties on the analysis level corresponding to the measurement actions described in this paper. However, they do not have a corresponding linear action on the generator level. Therefore, the actions cannot be compared on an event-by-event level, but only in a statistical sense. Assume a given classification analysis, then the question can be asked if it is optimal. This question can be posed as a function of any direction in the observable space, $\vec{o}$. If o is subdivided in bins o_j, it can be asked whether the information that an event is in bin j should change the estimated action. In the linear approximation it is equivalent to asking if replacing a_i by $a_i + c_j$ in bin j, gives a more optimal estimation of a_i. This can be computed in a Bayesian framework from likelihoods:

$$P = \prod_{i \in s} P_i(s) \prod_{i \in b} P_i(b) \tag{42}$$

This likelihood is a function of c_j for events belong to bin j:

$$\ln(P(c_j)) = \sum_{i \in s, i \in j} - \ln(1 + \exp(a_i + c_j)) + \sum_{i \in b, i \in j} [a_i + c_j - \ln(1 + \exp(a_i + c_j))] \tag{43}$$

The value of c_j can be estimated from this likelihood expression together with its statistical error.

Basically most of the tools developed for the measurements can then be applied to the classification analyses.

5 Conclusion

This paper has reviewed a method to optimize MC based analysis. The proposed k2-filter algorithm is basically founded on the linear Kalman filter approach. The practical implementation was described in detail. Even though it might seem involved the final optimizing code is typically extremely compact, consisting of computing sums of quantities, and solving one set of normally linear equations. When applying the method to analysis optimization a change of paradigm is needed compared to classical analysis methods used in high energy physics. Since the k2-filter works only with derivatives it becomes very "smart", but the interpretation of the results become less intuitive. Therefore, some care has to be used in order to insure that the results make sense.

The k2-filter was shown to be applicable not only for linear 1-dimensional measurements. Integrating the derivative of a locally linear likelihood provides a non-linear global likelihood, which is also easily generalized to multi-dimensional measurements.

The problem of "overtraining" was addressed in this paper, and it was shown how to control and estimate the problem, which arises when only a limited sample of MC events are available.

Finally, some examples were shown together with recommendation on how to use the k2-filter in practice.

References

[1] DELPHI Collaboration, Phys. Lett. B511 (2001) 159.

[2] ALEPH Collaboration, Eur. Phys. J. C17 (2000) 241.
L3 Collaboration, Phys. Lett. B454 (1999) 386.
OPAL Collaboration, Phys. Lett. B507 (2001) 29.

[3] R.E. Kalman: A New Approach to Linear Filtering and Prediction Problems
Transactions of the ASME-Journal of Basic Engineering, 82 (Series D): 35-45, 1960

[4] N. J. Kjær
Optimal statistical treatment of parameterized systematic errors in a unified Bayesian-Frequentist Monte Carlo Framework.
Contributed paper to "Advanced Statistical Techniques in Particle Physics",
Grey College, Durham, 18 - 22 March 2002

[5] T. Sjöstrand and V. Khoze
On colour rearrangement in hadronic W^+W^- events
Zeit. Phys. C62 (1994) 281-309.