Track fitting with ambiguities and noise: a study of elastic tracking and non-linear filters

R. Frühwirth [a] and A. Strandlie [b]

[a] *Institute for High Energy Physics of the Austrian Academy of Sciences, Vienna, Austria*

[b] *Department of Physics, University of Oslo, Norway and Department of Electrical Engineering and Science, Gjøvik College, Norway*

We present a study of track fitting in the presence of ambiguous measurements and noise. We consider four methods for solving the resulting assignment problem, two elastic arm algorithms and two non-linear filters, including a novel development, the Deterministic Annealing Filter. We describe their basic features and investigate their relationships with each other and with other popular estimators, in particular with the EM algorithm. The performance of the methods is optimized by means of several simulation experiments. We study the influence of the annealing schedule on the performance and show that the application of advanced minimization methods is required in order to obtain reliable estimates of the track parameters. We compare the relative efficiencies and the computational costs of the four methods both under ideal conditions and with noise. A final experiment under realistic conditions focusses on the robustness of the proposed approach.

1 Introduction

Track reconstruction in the LHC experiments will be faced by high track densities and consequently high occupancy of the detector modules, particularly in the inner tracking devices. In addition, there will be a considerable amount of noise hits in the detector, both random noise uncorrelated with the passage of a track, and track correlated noise like crosstalk, delta rays, and cluster decays. As a consequence, it is unlikely that the track finding algorithms, which have to be fast, will be able to resolve entirely the problem of assigning the detector hits to the track candidates. In our opinion it is most likely that the final solution of this combinatorial problem will be deferred to the track fit which is designed to use the available information in a statistically optimal fashion. The track fit will thus be confronted with several competing hits at

Preprint submitted to Elsevier Preprint *3 May 1999*

some or even most of the layers of the tracking device. (We assume that a track detector can be represented as a collection of shells or layers each of which may contribute one or several measurements. It is of course possible that a track misses a layer or that a layer does not respond although it is hit by the track.) If the device is a drift detector the problem is aggravated by the inherent ambiguity of the measurements.

Several methods have been proposed in the literature in order to cope with this assignment problem. They can be classified in two main classes: elastic tracking and non-linear filters. Elastic tracking comes in two main varieties. The first one is inspired by neural networks and features least-squares estimation of the track parameters concurrently with solving the combinatorial assignment problem by an annealing algorithm [1,2]. We will show below that this approach is also closely related to the EM algorithm well-known in statistics [3], and in particular to a generalized version, the EM algorithm with deterministic annealing (DAEM), which has been proposed recently [4]. The second variety of elastic tracking is more closely related to the Radon and the Hough transform [5]. The estimator is neither least-squares nor maximum likelihood, and we will show below that its statistical properties are clearly inferior to the least-squares estimator.

The second class comprises filters which are non-linear extensions of the standard linear Kalman filter. Recently it has been proposed to use the Gaussian-sum filter (GSF, see [6]) for solving the assignment problem concurrently with the estimation of the track parameters [7]. It is one of the crucial features of the GSF that prior information on the measurements can be optimally used as long as it can be modelled by mixtures of Gaussians. If the assignment problem is caused only by the left-right ambiguity of the measurements this is straightforward, and the GSF is a competitive alternative to the elastic tracking approach. If the main source of competing hits is noise, either electronic noise or hits from other tracks, finding a suitable prior model is much more difficult, and the GSF has to be supplemented by ad-hoc assumptions on the frequency and the distribution of noise hits. For this reason we have developed a related, but more robust approach, the Deterministic Annealing Filter (DAF). This is related to the DAEM algorithm and also to methods used in the tracking of single and multiple targets. It is similar to the Probabilistic Data Association Filter (PDAF) used for multitarget tracking in clutter [8], but has the additional feature of an annealing process. It can also be regarded as a simplified GSF with an additional validation feature which eliminates hits which are not compatible with the predicted track position.

The paper is organized as follows. Section 2 reviews the elastic tracking methods and describes further developments required to approach statistically optimal performance. Section 3 gives a brief summary of the Gaussian-sum filter and introduces a novel method, the Deterministic Annealing Filter. In Section 4 we discuss the relation of the methods with each other and with other algorithms proposed in the relevant literature, in particular with the EM algorithm. We then proceed to study the methods on simulated data. In Section 5 we first describe briefly the ATLAS transition radiation tracker (TRT) used in this study. Then we present the results of three simulation experiments. The first experiment concentrates on the statistical properties of the estimators, and we mainly report results on the precision of the various methods and their computational cost. The second experiment focusses on the robustness, i.e. the performance in the presence of noise. Finally, in the third experiment we present results from the application to track candidates produced by the TRT track finder. The paper is concluded by Section 6 containing some final remarks and a brief outlook to further research.

2 Elastic tracking algorithms

2.1 Elastic Arms or Deformable Templates

The Elastic Arms algorithm (EAA) has been developed by Ohlsson, Peterson and Yuille [1]. The algorithm aims to concurrently resolve the problem of finding correct hit-to-track assignments together with fitting the selected points to the respective tracks. This is done by defining a suitable energy function, which is basically a sum of squared distances between the hits in the detector and the arms. Each term in the sum is switched on or off, depending on the state of a binary hit-to-arm assignment variable. The global minimum of this energy function with respect to the track parameters and the assignment variables gives the solution to the track reconstruction problem. In order to avoid the problem of ending up in a local minimum during the minimization procedure, one first requires the configurations of the system to obey the Boltzmann distribution of statistical mechanics by invoking a temperature parameter T. A marginal probability distribution is then obtained by summing over all allowable configurations of the assignment variables, and this in turn defines an effective energy. The strategy is then to minimize this effective energy at successively smaller temperatures and obtain the final results in the limit $T \to 0$. The effect of the annealing procedure is to lead the search for the global minimum into the correct region of parameter space as the temperature gets so low that the original structure of the energy landscape shows up. The risk of ending up in a local minimum in the final minimization step is thus greatly reduced, and this is indeed a very powerful feature of the algorithm.

The ability of the Elastic Arms algorithm to efficiently resolve the problem of correct hit-to-track assignments has been demonstrated several times earlier. In this paper, however, we will assume that a separate pattern recognition procedure has been applied first. We therefore intend to use the EAA to obtain the final fit of the tracks, and the pattern recognition problem is here to discern the correct track points from the mirror hits and noise. The scenario is thus to optimally fit a track to a subset of points of a track candidate, the track candidate consisting of track points and possibly mirror points and noise. This procedure is applied to each track candidate independently. It has to be noted that the standard Elastic Arms algorithm applied to a global tracking problem is doing effectively the same thing as the temperature gets very low. In this phase the arms have settled in the vicinity of the tracks, and the final fit depends only on points being very close to the arms.

In order to obtain optimal results from the EAA we have to slightly modify the formalism. The energy function of Lindström [2] implies competition in the final fit only between a hit and its corresponding mirror hit. Since there may well be several hits in the same detector layer, a more plausible solution is to let all hits in the same layer compete in the fit. The energy function therefore becomes

$$E\left(\{S_k, s_{ik}\}, \boldsymbol{p}\right) = \sum_k \left[S_k \left(\sum_{i=1}^{n_k} s_{ik} M_{ik} \right) + \lambda \left(S_k - 1 \right)^2 \right], \tag{1}$$

where the sum over k denotes sum over layers, n_k is the number of measurements (including mirror hits) in layer k, and M_{ik} is the squared distance from point i in layer k to the one arm under consideration. There is now no sum over arms, since we treat the arms independently. The track parameters of the arm are denoted by the vector $\boldsymbol{p}$.

The assignment variables S_k and s_{ik} are either 1 or 0, and the s_{ik}'s are subject to the constraint $\sum_{i=1}^{n_k} s_{ik} = 1$. The S_k's denote whether the measurements in layer k as a whole are assigned to the arm or not, while the s_{ik}'s play the same role for the individual measurements inside the layer. For instance, $S_k = 0$ means that the measurements in layer k are assigned to noise, irrespective of the values of the s_{ik}'s. The constraint on the s_{ik}'s tells us that only one of the points in the layer is allowed to contribute to the energy. This constraint is applied during the derivation of the marginal probability density, and its effect is seen in the limit $T \to 0$. At non-zero temperatures, however, all possible configurations are allowed. We will therefore observe that all points in a layer are assigned to the track, each with a probability depending on the distance from the arm to the point and the locations of the other points.

Following the same strategy as in [2] in deriving the marginal probability density, the effective energy will in our case be

$$E_{\text{eff}} = -\frac{1}{\beta} \sum_k \log \left(n_k \cdot e^{-\beta\lambda} + \sum_{i=1}^{n_k} e^{-\beta M_{ik}} \right), \tag{2}$$

with $\beta = 1/T$. From the form of the effective energy it is clear that the "natural" temperature is $T_{\rm n} = 2\sigma^2$, where σ^2 is the variance of the observation error. If there is only one measurement per layer and $\beta\lambda$ is very large, $-\beta E_{\rm eff}$ is equal to the log-likelihood function plus a constant, and minimizing the effective energy at $T_{\rm n}$ gives the maximum likelihood estimate of $\boldsymbol{p}$.

As $T \to 0$ or $\beta \to \infty$, the arguments of the logarithm become very simple, because the term with the largest exponent will totally dominate the sum inside the logarithm in this limit. The energy function then reduces to

$$E_{\rm eff} = \sum_k \min\left(\{M_{ik}\}, \lambda\right), \tag{3}$$

which is a sum of squared distances and possibly λ's. The expression $\{M_{ik}\}$ denotes the set of squared distances from the points in layer k to the arm. The effect is that for each layer only the point closest to the arm contributes to the final fit. If all points in the layer are further away from the arm than $\sqrt{\lambda}$, none of the points will contribute to the fit. The quantity $\sqrt{\lambda}$ is seen to define a cutoff in the sense that only points within this distance of the arm are able to have a pull on the track. The fit will therefore be very robust, since outliers further away from the track than $\sqrt{\lambda}$ have no effect on the estimates at all.

In earlier work several different minimization methods have been suggested. For the search direction, for instance gradient descent [1] and use of the Hessian matrix [2] have been proposed. To our knowledge, however, little effort has been made in the line search part of the minimization procedure, that is, in assessing how far in parameter space along the search direction one should go. In this work we use ordinary gradient descent during the annealing, but at the final temperature we propose to use the Davidon-Fletcher-Powell algorithm [9] to find the search direction. This algorithm is an example of a quasi-Newton method and basically finds the search direction by approximating the inverse Hessian using only the first derivatives, and no matrix inversions are required. The line search is performed by fitting a polynomial to the one-dimensional function that arises when the energy function is evaluated along the search direction, and the parameter vector at the minimum point of this polynomial is chosen as the starting value for the next search. This procedure is repeated iteratively until convergence. We will see later that the application of such a procedure during the final minimization step is mandatory in order to find the global minimum with a satisfactory accuracy.

2.2 Elastic Tracking

The Elastic Tracking algorithm (ETA) has been proposed by Gyulassy and Harlander [5]. The basic idea is to interpret the classical Radon transform as an interaction energy between a template track and the hits in the detector. The parameters of the template track giving the minimum interaction energy define the solution of the problem. In order to avoid local minima of the energy function the minimization is organized as an iterated search, similar to the annealing described in the previous subsection.

The interaction energy R_V is defined in terms of a sum over interactions between each hit and the track. The interaction between each hit and the track is governed by a potential V such that

$$R_V(\boldsymbol{p}, I) = -\sum_k \sum_{i=1}^{n_k} V(M_{ik}, I). \tag{4}$$

The quantity $\boldsymbol{p}$ symbolizes the parameter vector, I is the iteration number, and the sum over k is again a sum over all layers. In [5] it was proposed to use a Lorentzian potential of the form

$$V(M_{ik}, I) = \frac{w^2(I)}{M_{ik} + w^2(I)}. \tag{5}$$

The quantity $w(I)$ governs the width of the potential; actually it is equal to the half width at half maximum. It should be chosen quite large in the first iteration in order to smooth out the energy surface; it is then gradually lowered until the final value is reached. In [5] it was claimed that the natural asymptotic value of w is given by the standard deviation of the measurement error. It can be shown, however, that this choice does not give the best possible resolution. If there are no mirror hits and no noise, the resolution is clearly best for large w, because then the potential is nearly linear in M_{ik}:

$$\frac{w^2}{M_{ik} + w^2} \approx 1 - \frac{M_{ik}}{w^2}, \quad \text{for } M_{ik} \ll w^2,$$

and minimizing the energy in Equation (4) is tantamount to a least-squares estimation. In the presence of mirror hits and/or noise there is a dilemma. The width should be large in order to have near-optimal performance, but it cannot be very large because then wrong hits would get too much weight. A good compromise between these conflicting requirements can be expected to be at about three to four standard deviations. We will show in Subsection 5.2 that this simple reasoning gives a fairly good description of the actual behaviour.

The minimization of the potential in each iteration is usually done by a gradient descent method. We will show that also in this case a more sophisticated minimization method like the Davidon-Fletcher-Powell algorithm gives far better results.

The Lorentzian potential has the required feature that it approaches zero as the distance between the hit and the track increases, but it does so very slowly.

This suggests a potential problem insofar that outliers might after all have a significant effect on the estimates. We therefore propose to use a Gaussian potential instead, of the form

$$V\left(M_{ik}, I\right) = \exp\left(-\frac{M_{ik}}{2w^2(I)}\right). \tag{6}$$

This potential decreases much more rapidly than the Lorentzian potential and should therefore be more robust. This situation is exactly the opposite of the one encountered in least-squares or maximum likelihood estimation, where a Gaussian model is much more sensitive to outliers than a long-tailed one.

3 Non-linear filters

The linear or Kalman filter is now widely used for the estimation of the parameters (initial values) of track candidates delivered by the track finder. It is well known that it has a certain number of advantages over least-squares estimation in a global linear model:

- Process noise like multiple Coulomb scattering or energy loss can be treated locally; no long-range correlations of the observations arise.
- The estimated state vector closely follows the actual path of the particle; material effects can be evaluated more precisely, and the quality of the linear approximation is better.
- In combination with the smoother, optimal estimates of the track parameters can be obtained at any point along the track; this can be used for the detection of outliers or the resolution of ambiguities.

Application of the Kalman filter requires that the assignment problem has been entirely resolved by the preceding track finder. If this is not the case, one can envisage to run the filter on every possible assignment and to select the "best" one, for instance in terms of the p-value of the chi-square statistic. We will call this approach a combinatorial Kalman filter (CKF). Apart from being computationally expensive, we will show below that the CKF is in general not the best solution.

In the Kalman filter, the innovations, i.e. the residuals of the observations with respect to the predicted state vector, are used only in the update of the state vector, and only in a linear fashion. If the filter is to solve the assignment problem by giving different weights to competing hits, the innovations have to be used in a non-linear fashion, both in the update of the state vector and in the update of the covariance matrix. An example of such a non-linear filter is the Gaussian-sum filter (GSF) developed by Kitagawa [10,11]. The potentially large computational cost of the GSF and a certain lack of robustness have led us to the development of a faster and more robust method which we call the Deterministic Annealing Filter (DAF). It is an iterated filter with the possibility of including an annealing phase. The two filters will be described in the following two subsections.

3.1 The Gaussian-sum filter

The Gaussian-sum filter (GSF) is a relatively straightforward extension of the standard Kalman filter, its main distinction being that all densities involved are allowed to be mixtures or sums of Gaussians instead of single Gaussians. The application of the GSF to the treatment of non-Gaussian noise in track fitting has been explored in a couple of publications [6,12]. It has been shown there that it is a useful tool for the treatment of long-tailed measurement noise and long-tailed or non-Gaussian process noise. More recently, it has been proposed to use the GSF for solving the assignment problem in a track detector with ambiguities [7]. This approach can be further generalized to cope with an arbitrary number of competing hits in any measurement layer of the detector.

Let $\boldsymbol{x}_k$ denote the state vector at layer k, i.e. the collection of the parameters required to specify the track. Five parameters are required for a track fit in space, and three for a fit in a projection. In general, the state vector cannot be observed directly. We assume that an observation $\boldsymbol{m}_k$ in layer k is a linear function of the state plus some measurement error:

$$\boldsymbol{m}_k = \boldsymbol{H}_k\,\boldsymbol{x}_k + \boldsymbol{\epsilon}_k; \quad \mathrm{E}(\boldsymbol{\epsilon}_k) = 0,\ \mathrm{cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{V}_k. \tag{7}$$

Under the further assumption of $\boldsymbol{\epsilon}_k$ being distributed according to a Gaussian we can write down the observation density of $\boldsymbol{m}_k$ conditional on the state:

$$p_m(\boldsymbol{m}_k|\boldsymbol{x}_k) = \varphi(\boldsymbol{m}_k; \boldsymbol{H}_k\,\boldsymbol{x}_k, \boldsymbol{V}_k), \tag{8}$$

where $\varphi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{V})$ is a multivariate Gaussian probability density function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$.

The predicted distribution of the state vector $\boldsymbol{x}_k$ can be assumed to be a Gaussian sum or mixture with N_{k-1} components:

$$\pi(\boldsymbol{x}_k) = \sum_{j=1}^{N_{k-1}} \pi_k^j\,\varphi(\boldsymbol{x}_k; \boldsymbol{x}_{k|k-1}^j, \boldsymbol{C}_{k|k-1}^j), \qquad \sum_{j=1}^{N_{k-1}} \pi_k^j = 1. \tag{9}$$

This is the prior distribution of the state vector $\boldsymbol{x}_k$. Using Bayes's theorem on

$p_m(\boldsymbol{m}_k|\boldsymbol{x}_k)$ and $\pi(\boldsymbol{x}_k)$ gives the posterior distribution of $\boldsymbol{x}_k$:

$$p(\boldsymbol{x}_k) = \frac{p_m(\boldsymbol{m}_k|\boldsymbol{x}_k) \cdot \pi(\boldsymbol{x}_k)}{\int p_m(\boldsymbol{m}_k|\boldsymbol{x}_k) \cdot \pi(\boldsymbol{x}_k)\, d\boldsymbol{x}_k}. \tag{10}$$

It is not difficult to show that this posterior is again a Gaussian sum:

$$p(\boldsymbol{x}_k) = \sum_{j=1}^{N_{k-1}} q_k^j \, \varphi(\boldsymbol{x}_k; \boldsymbol{x}_{k|k}^j, \boldsymbol{C}_{k|k}^j), \tag{11}$$

where the mean vector and the covariance matrix $\{\boldsymbol{x}_{k|k}^j, \boldsymbol{C}_{k|k}^j\}$ of component j is obtained by a Kalman filter from the observation $\{\boldsymbol{m}_k, \boldsymbol{V}_k\}$ and the prediction $\{\boldsymbol{x}_{k|k-1}^j, \boldsymbol{C}_{k|k-1}^j\}$ [6]. The posterior weights are given by:

$$q_k^j \propto \pi_k^j \, \varphi(\boldsymbol{m}_k; \boldsymbol{H}_k \boldsymbol{x}_{k|k-1}^j, \boldsymbol{V}_k + \boldsymbol{H}_k \boldsymbol{C}_{k|k-1}^j \boldsymbol{H}_k^T). \tag{12}$$

The constant of proportionality is determined by the requirement that the sum of all q_k^j is equal to 1.

If there are several, say n_k, competing observations $\boldsymbol{m}_k^i$ in layer k, we can form n_k hypotheses H_i, where H_i asserts that $\boldsymbol{m}_k^i$ is the observation to be assigned to the track. A priori all H_i are equally likely:

$$p_k^i = P(H_i) = 1/n_k, \; i = 1, \ldots, n_k. \tag{13}$$

The posterior density of $\boldsymbol{x}_k$ conditional on H_i is then equal to

$$p(\boldsymbol{x}_k|H_i) = \sum_{j=1}^{N_{k-1}} q_k^j \, \varphi(\boldsymbol{x}_k; \boldsymbol{x}_{k|k}^{ij}, \boldsymbol{C}_{k|k}^{ij}), \tag{14}$$

where the mean vector and the covariance matrix $\{\boldsymbol{x}_{k|k}^{ij}, \boldsymbol{C}_{k|k}^{ij}\}$ of component j is obtained by a Kalman filter from the observation $\{\boldsymbol{m}_k^i, \boldsymbol{V}_k\}$ and the prediction $\{\boldsymbol{x}_{k|k-1}^j, \boldsymbol{C}_{k|k-1}^j\}$. Finally the total posterior density is obtained by summing over all hypotheses:

$$p(\boldsymbol{x}_k) = \sum_{i=1}^{n_k} p(\boldsymbol{x}_k|H_i) \cdot P(H_i) = \sum_{i=1}^{n_k} \sum_{j=1}^{N_{k-1}} p_k^i q_k^j \, \varphi(\boldsymbol{x}_k; \boldsymbol{x}_{k|k}^{ij}, \boldsymbol{C}_{k|k}^{ij}). \tag{15}$$

This posterior is then propagated to the next layer and used again as the prior in the next filter step. At the end of the track, the final estimate $\boldsymbol{x}_{n|n}$ and its covariance matrix $\boldsymbol{C}_{n|n}$ are obtained as the mean and the covariance matrix of the posterior distribution $p(\boldsymbol{x}_n)$.

If there are competing measurements in most of the layers, the number of components in the mixture distribution of the state vector rises very quickly. This is particularly serious in the initial phase of the filter during which the information is not sufficient to select the correct assignments. In subsequent steps the components corresponding to wrong solutions get small posterior weights and can be suppressed.

Apart from its potentially high computational complexity the GSF has another serious shortcoming: there is no built-in protection against wrong hits which have no competing good hit. In the GSF a single hit in a layer always gets a posterior weight equal to one, i.e. it is always used with its full weight. This is particularly damaging in the initial phase where there is no way of detecting such wrong hits. This problem has been solved in [7] by modeling suspicious hits by a mixture of two Gaussians. The first Gaussian has a variance corresponding to the observation error and represents the hypothesis that the hit is correct; the second one has a larger width and represents the hypothesis that the hit is wrong. Apart from this approach being rather ad-hoc, it is not clear how the width of the second component and its mixture weight can be reliably determined. There is a solution which is entirely in the spirit of the GSF, by allowing in each layer the additional hypothesis that no hit is associated with the track. However, there still is the difficulty in determining the prior probability of this hypothesis, and the additional hypothesis results in a further increase of the computational cost.

3.2 The Deterministic Annealing Filter

The obvious way of guarding against wrong hits is a selection procedure based on the innovations, i.e. the predicted residuals of the hits. The problem of insufficient information in the initial phase of the filter can be overcome by adopting an iterative procedure. After a first pass of filter plus smoother the track position can be predicted in every layer, using information from all the other layers. Based on these predictions, the assignment probabilities of all competing hits can be computed in every layer. If this probability falls below a certain threshold, the hit is suppressed during the next iteration. (It is, however, not excluded that the hit is used again in a later iteration.) The assignment probabilities of the remaining hits are normalized to 1 and used as the weights in the next iteration of the filter. If all hits are eliminated the sum of the assignment probabilities is equal to zero and no normalization is possible. This requires some modification of the filter update formulas, effectively allowing the weights to sum to any number in the interval $[0, 1]$.

The filter itself is a Kalman filter with reweighted observations. The propagation part is identical to the standard case and therefore much faster than with the GSF; the prescription for the update of the state vector at layer k is a Kalman filter with prediction $\boldsymbol{x}_{k|k-1}$ and observations $\{\boldsymbol{m}_k^i, i = 1, \ldots, n_k\}$. If p_k^i is the assignment probability of observation $\boldsymbol{m}_k^i$, its weight matrix is

$p_k^i \boldsymbol{V}_k^{-1}$. This leads to the following update of the state vector:

$$\boldsymbol{x}_{k|k} = \boldsymbol{x}_{k|k-1} + \boldsymbol{K}_k \sum_{i=1}^{n_k} p_k^i (\boldsymbol{m}_k^i - \boldsymbol{H}_k \boldsymbol{x}_{k|k-1}). \tag{16}$$

$\boldsymbol{K}_k$ is the Kalman gain matrix, which has to be written in terms of inverse covariance or weight matrices if we want to allow for zero weights:

$$\boldsymbol{K}_k = (\boldsymbol{C}_{k|k-1}^{-1} + p_k \boldsymbol{H}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{H}_k)^{-1} \boldsymbol{H}_k^T \boldsymbol{V}_k^{-1}, \tag{17}$$

where p_k is the sum over all weights p_k^i. As in the standard Kalman filter, the covariance matrix $\boldsymbol{C}_{k|k}$ of the updated estimate $\boldsymbol{x}_{k|k}$ is given by the first factor of the gain matrix:

$$\boldsymbol{C}_{k|k} = (\boldsymbol{C}_{k|k-1}^{-1} + p_k \boldsymbol{H}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{H}_k)^{-1}. \tag{18}$$

This updating procedure is correct when the weights are considered as constants. We will see in the next section that this assumption can be theoretically justified.

After completion of the filter a second filter, called the backward filter, is run in the opposite direction, using the same weights as the forward filter. By taking a weighted mean of the predictions of both filters at every layer, we obtain a prediction $\boldsymbol{x}_{k|n}^*$ using all hits except the ones at layer k, along with its covariance matrix $\boldsymbol{C}_{k|n}^*$. (The asterisk indicates that the information from layer k is not used in this prediction). Based on this prediction and its covariance matrix, the assignment probabilities of the hits are then computed in the following way:

$$p_k^i \propto \varphi(\boldsymbol{m}_k^i; \boldsymbol{H}_k \boldsymbol{x}_{k|n}^*, \boldsymbol{V}_k + \boldsymbol{H}_k \boldsymbol{C}_{k|n}^* \boldsymbol{H}_k^T). \tag{19}$$

The assignment probabilities exceeding the threshold are normalized to 1 and used again as weights in the next iteration, and so on. The iteration is stopped when the assignment probabilities settle to their final value. An even simpler formula for the assignment probabilities can be obtained if the covariance matrix $\boldsymbol{C}_{k|n}^*$ of the smoothed estimate is neglected in Equation (19). Again it will be shown in the next section that this procedure has a theoretical basis.

It is not excluded that the final assignment probabilities, and consequently the final estimate, depend on the starting value of the iteration. In order to avoid this we propose to apply an annealing procedure very similar to the one described in Subsection 2.2. This means that the variances $\boldsymbol{V}_k$ of the observations now depend on the iteration number I. If we start with values $\boldsymbol{V}_k(1)$ well above the nominal variance of the observation error, the assignment probabilities after the first iteration will be close to $1/n_k$, ensuring that no hit is prematurely excluded from the filter. The variance is then gradually lowered, until the nominal value is attained. Finally, a few iterations at the nominal value are required so that the assignment probabilities can settle to their final value. If one is sure that the starting value of the iteration is close to the true track, the annealing can be dispensed with, and only a few iterations at the nominal variance are necessary. If, for instance, the starting value of the iteration is determined by an estimator with high break-down point such as the LMS estimator [13], it is guaranteed that the initial track is close to the majority of the hits, and annealing is very likely unnecessary.

It should be noted that this procedure results in a "soft" or "fuzzy" assignment of the hits to the track candidate. It may be tempting to continue the annealing in order to force the assignment probabilities to values very close to 0 or 1, in other words to obtain a "hard" assignment. In this case the filter formulas (Equations (16)–(18)) approach the standard Kalman filter formulas. We will, however, show below that nothing is gained by this.

4 Discussion of the methods

4.1 Comparison between different elastic tracking algorithms

If there is only one measurement in each layer, the energy function of the Elastic Arms algorithm (EAA) can be interpreted in terms of a potential, in the same spirit as for the Elastic Tracking algorithm (ETA). In the $T \to 0$ limit, the potential for the EAA is defined by

$$V\left(M_{ik}\right) = \begin{cases} M_{ik} \text{ if } M_{ik} < \lambda \\ \lambda \quad \text{otherwise} \end{cases} . \tag{20}$$

This follows directly from the zero-temperature expression of the energy function given in Equation (3). The potential rises quadratically as a function of distance until it reaches the cutoff $\sqrt{\lambda}$, and then it abruptly flattens out. For low but non-zero temperatures the potential is slightly modified around the cutoff region. Figure 1 shows the potential of the EAA at two temperatures: (a) at $T = 0$ and (b) at the natural temperature $T_{\mathrm{n}} = 2\sigma^2$, where σ^2 is the variance of the measurement error.

The quadratic form of the potential clearly shows us that the EAA is doing a least-squares fit at very low temperatures. This is not the case for the ETA. The Lorentzian and the Gaussian potential are plotted in Figure 1 (c) and (d). We have chosen the width w of the potentials such that the curvature at $x = 0$ is equal to the curvature of the quadratic potential of the EAA (a). Since this gives locally the best approximation to the quadratic potential it is reasonable to assume that this choice of w gives the best precision. This implies that the final width of the Lorentzian should be equal to the desired cutoff $\sqrt{\lambda}$, while for the Gaussian it should be equal to the cutoff divided by $\sqrt{2}$. In the following section we will compare this rule-of-thumb with simulation results. Figure 1 also shows that the Gaussian potential is a better global approximation to the EAA potential than the Lorentzian. It can therefore be expected that the Gaussian gives better results.

For the case of more than one measurement in a layer the EAA and the ETA clearly differ. The EAA is, as $T \to 0$, still doing a least-squares fit including only the measurement nearest to the arm. The other measurements in the same layer are not participating in the fit. The ETA, however, does not have any competition between the different hits in the same layer. The pull which each hit has on the track is not affected by the positions of the other measurements. Since there is rarely more than one hit per layer originating from the track to be fitted, the fitting algorithm should try to select the correct hit, include it in the fit and discard the rest of the hits. This behaviour is implicit in the EAA algorithm. One can therefore expect the ETA estimates to be influenced more strongly by mirror hits and noise than the EAA estimates. The results of the subsequent section will show that this is indeed the case.

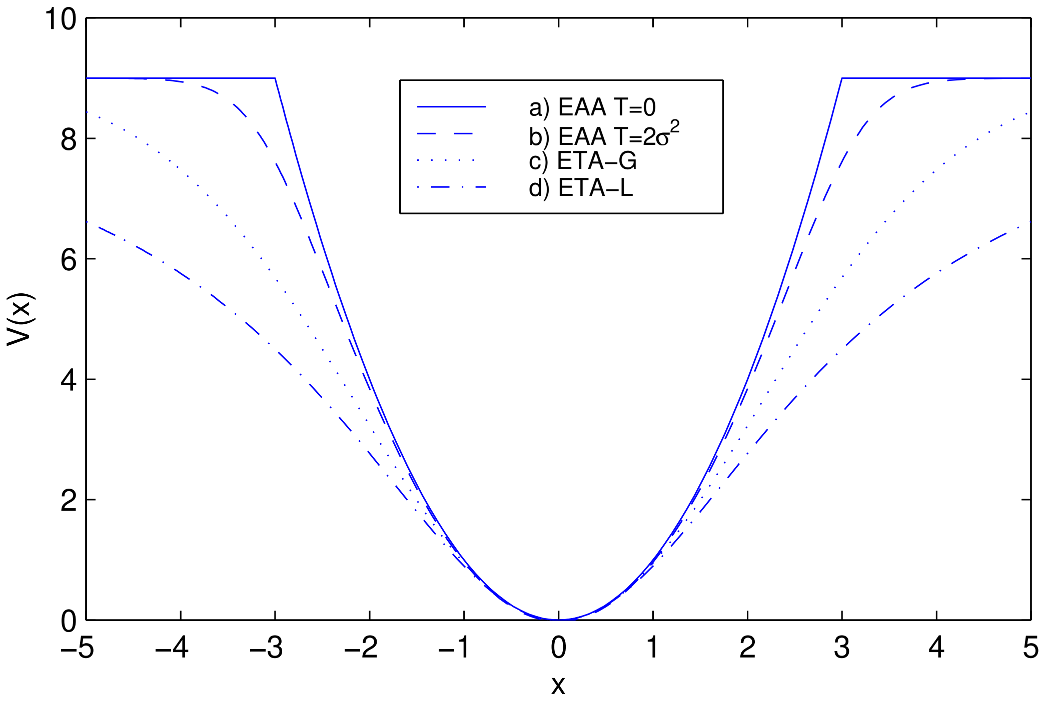

Fig. 1. The potential V as a function of distance x, expressed in standard deviations of the measurement error. The cut $\sqrt{\lambda}$ is set at three standard deviations. a) EAA at $T = 0$. Note the abrupt change in the potential's behaviour at $|x| = 3$. b) EAA at $T = 2\sigma^2$. This temperature corresponds to the width of the measurement error. c) The Gaussian potential of the ETA with the same curvature at $x = 0$. d) The Lorentzian potential of the ETA with the same curvature at $x = 0$.

It has been noted earlier [1] that the EM algorithm is an alternative way of performing the minimization of the energy in the EAA algorithm. This is because under the assumption of Gaussian observation errors the maximum likelihood estimate coincides with the least-squares estimate. In fact, the convergence theorem by Dempster, Laird and Rubin [3] guarantees that the EM algorithm will minimize the effective energy (Equation (2)). Each iteration of the EM algorithm can be divided into two major steps: an expectation step and a minimization step. The expectation step calculates the average of the original energy function (Equation (1)) over the assignment variables, conditional on the current value $\boldsymbol{p}'$ of the track parameters . This expectation defines the function

$$\begin{aligned} Q\left(\boldsymbol{p}|\boldsymbol{p}'\right) &= \sum_{\{S_k, s_{ik}\}} E\left(\{S_k, s_{ik}\}, \boldsymbol{p}\right) \cdot P\left(\{S_k, s_{ik}\} | \boldsymbol{p}'\right) \\ &= \sum_{\{S_k, s_{ik}\}} E\left(\{S_k, s_{ik}\}, \boldsymbol{p}\right) \cdot \frac{P\left(\{S_k, s_{ik}\}, \boldsymbol{p}'\right)}{P_M\left(\boldsymbol{p}'\right)}, \end{aligned} \tag{21}$$

where the sum over $\{S_k, s_{ik}\}$ means sum over all allowable configurations and $P_M\left(\boldsymbol{p}'\right)$ is the marginal probability density function. The prime on the parameter vector emphasises the fact that $\boldsymbol{p}'$ is a fixed quantity, namely the current estimate of the parameters, whereas $\boldsymbol{p}$ is variable. By writing out the energy function and the probabilities explicitely one can perform the summation. The result is

$$\begin{aligned} Q\left(\boldsymbol{p}|\boldsymbol{p}'\right) &= \sum_k \left[\sum_{i=1}^{n_k} M_{ik} \cdot \frac{e^{-\beta M'_{ik}}}{n_k e^{-\beta\lambda} + \sum_{j=1}^{n_k} e^{-\beta M'_{jk}}} + \lambda \cdot \frac{n_k e^{-\beta\lambda}}{n_k e^{-\beta\lambda} + \sum_{j=1}^{n_k} e^{-\beta M'_{jk}}}\right] \\ &= \sum_k \left[\sum_{i=1}^{n_k} M_{ik} p'_{ik} + \lambda p'_{0k}\right], \end{aligned} \tag{22}$$

where the primed quantities again refer to an arm with parameter vector $\boldsymbol{p}'$. The p'_{ik} is to be interpreted as the probability that the hit i in layer k is assigned to the arm, while p'_{0k} defines the probability that none of the measurements is assigned to the arm. The minimization step is then to minimize the Q function with respect to the parameters $\boldsymbol{p}$. Since the last term in the above expression is independent of $\boldsymbol{p}$, we will obtain exactly the same result if we minimize

$$Q\left(\boldsymbol{p}|\boldsymbol{p}'\right) = \sum_k \sum_{i=1}^{n_k} M_{ik} p'_{ik} \tag{23}$$

with respect to $\boldsymbol{p}$. The new value of the parameter vector is used to update the probabilities, and the Q function is again minimized. This procedure is repeated until convergence. We see that the EM algorithm in this case is nothing but an iteratively reweighted least-squares procedure, and the weights are given by the assignment probabilities.

If there is non-negligible process noise, for instance multiple Coulomb scattering in the detector material, it is of advantage to replace the global least-squares fit by a filter plus smoother (see Section 3). The main reason is that the initial track parameters $\boldsymbol{p}$ do not capture any longer the entire information about the trajectory: the smoothed estimate at a given layer k is a better estimate of the track parameters at layer k than the extrapolation of the initial track parameters to layer k. The M step now consists of a pass of the filter and of the smoother, the measurements being weighted by the p_{ik} determined in the previous iteration. In the E step the weights are recalculated, using the distance of the hits from the smoothed track parameters. The natural scale of β is $\beta = 1/(2V_k)$, where V_k is the error variance of the observation which here is assumed to be one-dimensional (see Equation (7)). Using this value of β one arrives at a procedure which is almost identical to the DAF described in Subsection 3.2; merely the weights are slightly different. If there is only a single measurement in layer k, the weight used by the DAF is equal to 1 if it is above the threshold, and equal to 0 if it is below the threshold, resulting in a "hard" cut. With the EAA, the weight is a continuous function of the distance between the measurement and the track position, resulting in a "soft" cut even if there is no competition. At the position of the cut the weight is equal to 0.5. This is illustrated in Figure 2.

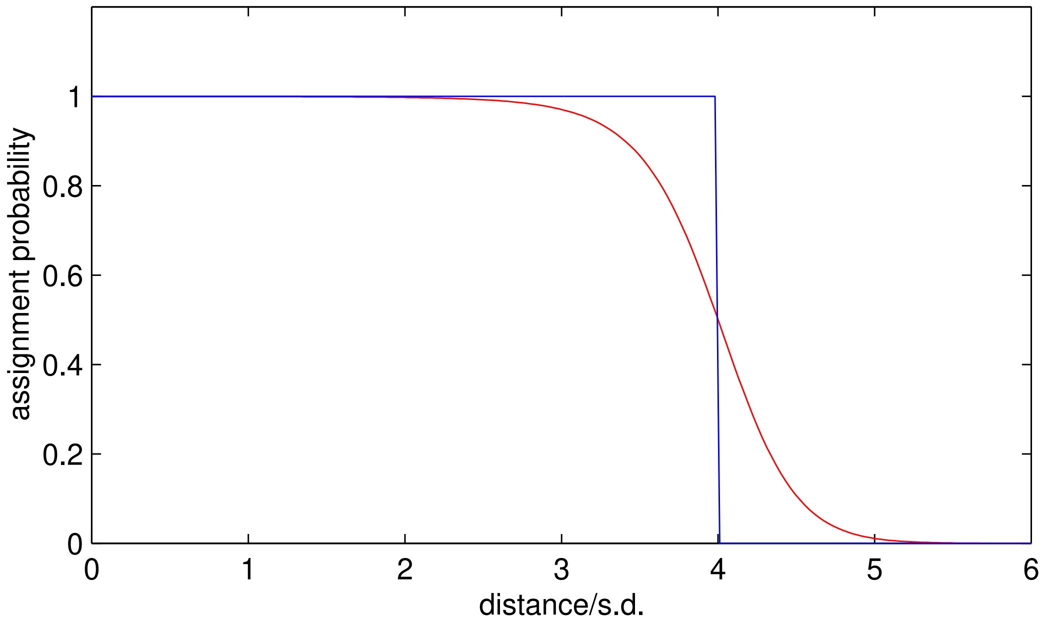

Fig. 2. The association probability of a single hit as a function of its distance from the predicted track position, measured in standard deviations of the measurement error. The step function is used by the DAF, the continuous function is used by the EAA at $\beta = 1/(2V_k)$. The cut is set at four standard deviations.

In general, the weights in the EAA do not add up to 1, because for every measurement there is a non-zero probability that it is not the correct one, even if it is very close to the track. If there are several competing measurements in a layer, and if one of them is reasonably close to the track, the weights of the EAA and of the DAF are virtually indistinguishable. It is of course perfectly possible to run the DAF with the weights as computed by the EAA, and this is what has actually been done in the simulation experiments described in the next section.

If multiple scattering is not negligible it is perfectly straightforward to incorporate it into the DAF (or into the GSF), whereas it is not obvious how to do this in the energy function of the EAA. In this case we therefore recommend using the filter rather than the EAA. Otherwise it is largely a matter of choice, although we will show below that a sophisticated and therefore time-consuming minimization method is required if the EAA is to deliver the best possible precision. In the absence of process noise the iterated re-weighted least-squares estimate is the fastest approach.

Both the EAA and the DAF offer the possibility of an annealing phase, in order to overcome the problem of ending up in a local minimum of the energy function or LS objective function. A version of the EM algorithm with deterministic annealing has been proposed recently [4]. It is easy to see that this algorithm is exactly the same as the EAA and the DAF with annealing. It should be noted that it is not necessary to actually find the minimum of the energy function or of the LS objective function at every value of β: In our implementation of the EAA only a single step of gradient descent is made at every β except the final one. Similarly, in the DAF only one EM iteration is carried out except at the final value of β.

Although developed independently, the EAA/DAF is also closely related to the probabilistic data association filter (PDAF) used for the tracking of mobile targets in a noisy environment [8]. The PDAF is rather similar to the GSF, but explicitely allows for the hypothesis that the target has not been observed, i.e. that all observations inside a "validation gate" are noise. Being designed to run in real time the PDAF foresees neither iterations nor annealing.

5 Results of simulation experiments

5.1 The ATLAS TRT

The ATLAS Inner Detector Transition Radiation Tracker (TRT) consists of two different parts: the barrel TRT and the end-cap TRT. Both parts of the detector are built up by drift tubes, so-called "straws". In the barrel the straws are placed in cylindrical layers, each straw being parallell to the beam axis. All straws inside one layer have the same distance to the beam axis. In the end-caps the straws are perpendicular to the beam axis and radially positioned. Only the barrel part of the TRT has been used in this study.

The barrel TRT consists of 75 layers with straws, the layers ranging from a radius of 56 cm to a radius of 106 cm. The distance between the layers is about 6.8 mm. The z range of the barrel layers is from $z = -75$ cm to $z = 75$ cm. The straws in the innermost 9 layers are active only for $|z| > 40$ cm. This reduction of the active region is done in order to reduce the occupancy in the innermost layers. All straws are divided into two halves with independent readout at $z = 0$; apart from that there is no z-information. In each layer the distance between the straws is about 6.8 mm, and each drift tube has a diameter of 4 mm. In total there are about 50000 straws in the barrel TRT. The geometry and the construction of all parts of the ATLAS Inner Detector is described in [14]; for more details the reader is referred to this report.

The simulated observations are given in polar coordinates; they consist of the layer number k, the layer radius R_k, the polar angle Φ of the sense wire of the straw, the absolute value of the drift distance, and the sign of z. The observations are therefore ambiguous. In the analysis we also use the track label attached to each observation.

5.2 Results with perfect tracks

For a precise evaluation of the statistical properties of the four methods presented in Sections 2 and 3 we have first simulated a sample of 9800 "perfect" tracks, without noise and neglecting all material effects in the detector. Only the measurement error has been simulated, with a nominal standard deviation of 0.25 mm. The correct solution of the left-right ambiguity is known and has been used in the analysis of the results, but not in the reconstruction. Figure 3 shows two characteristic properties of the sample: a) the distribution of the transverse momentum, and b) the distribution of the number of tubes which have been hit, with an average of 36.1 and an r.m.s. of 2.6.

Initialization

All estimation methods used in our simulation experiments require the knowledge of at least approximate initial track parameters. In the filter algorithms (GSF, DAF) they are used as the reference track, i.e. the point in parameter space which serves as the expansion point of the linear approximation of the track model. In the iterative algorithms (DAF, EAA, ETA) they are the starting point of the iteration.

The initial track parameters have been obtained in the following way. In the barrel part of the TRT the projection of a track onto the (x, y)-plane is a circle. If it passes through the origin the circle can be described by two parameters, its (signed) radius ρ and the angle ψ between the tangent to the track at the origin and the x-axis. In polar coordinates the track is approximately a straight line. This follows immediately from the fact that the polar angle Φ of the intercept of the track with a measurement layer of radius R is given by:

$$\Phi(R) = \psi + \arcsin(R/2\rho) \approx \psi + R/2\rho, \tag{24}$$

where the approximation is valid for $|\rho| \gg R$, in our case for transverse momenta larger than about 2 GeV/c. The initial track parameters are obtained by a least-squares fit of all observations to a straight line. In the subsequent track fit the track is not required to pass through the origin and is therefore described by a set of three track parameters, for instance ρ, ψ, and the (signed) distance from the origin at the point of closest approach.

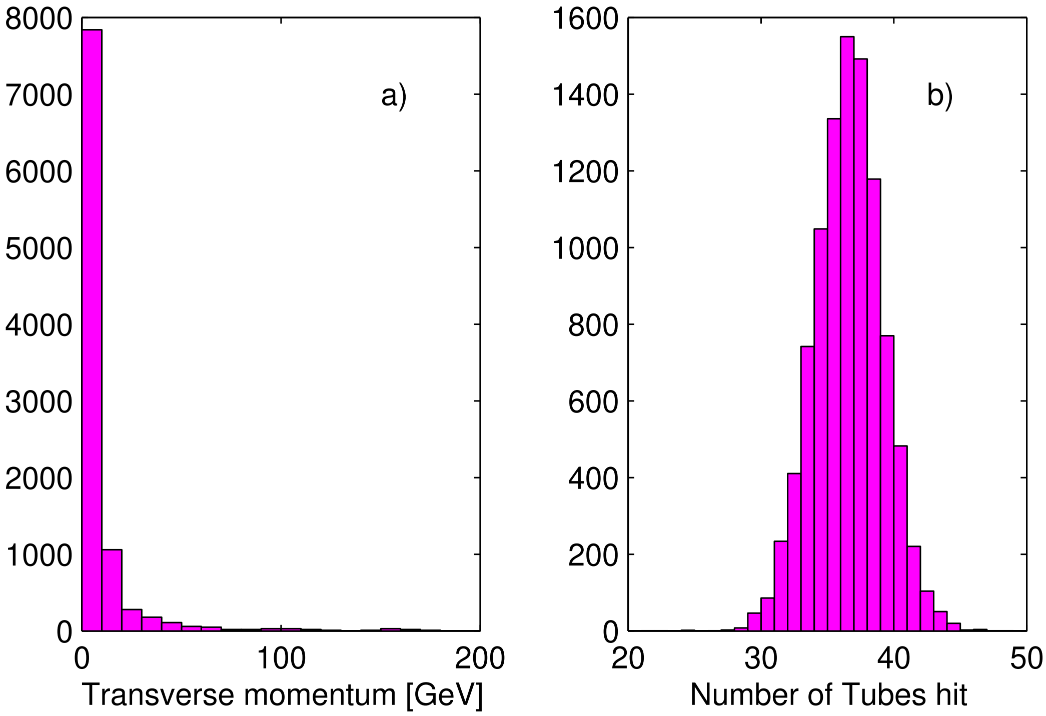

Fig. 3. Frequency distribution of a) the transverse momentum of the tracks and b) the number of tubes hit by a track

Tracks without mirror hits

The precision of the estimators is assessed by the generalized variance V of the residuals of the estimated track parameters with respect to the true values, i.e. by the determinant of the matrix of the second moments about zero of these residuals. Our baseline is the performance on tracks without mirror hits where there is no assignment problem. The point of reference is the generalized variance of the DAF with a cutoff $\sqrt{\lambda}$ at four standard deviations of the measurement error. In order to have full compatibility the EAA weights (see Equation (22)) have been used in the DAF.

If there are no mirror hits the ETA should perform best if the final width w of the potential is large. This is illustrated in Figure 4, showing the relative generalized variance as a function of w, the latter being measured in standard deviations of the measurement error. A width larger than four (Gaussian potential) to six (Lorentzian potential) standard deviations gives the optimal performance. The Davidon-Fletcher-Powell algorithm [9] was used for minimizing the energy function; with a simple gradient descent the results are slightly worse for the Gaussian potential and much worse for the Lorentzian (see also Table 1).

The relative generalized variance $V_{\rm rel}$ of all methods is shown in Table 1. As mentioned above, the results of the elastic algorithms (EAA and ETA) strongly depend on the minimization algorithm. In the table we show the results for a

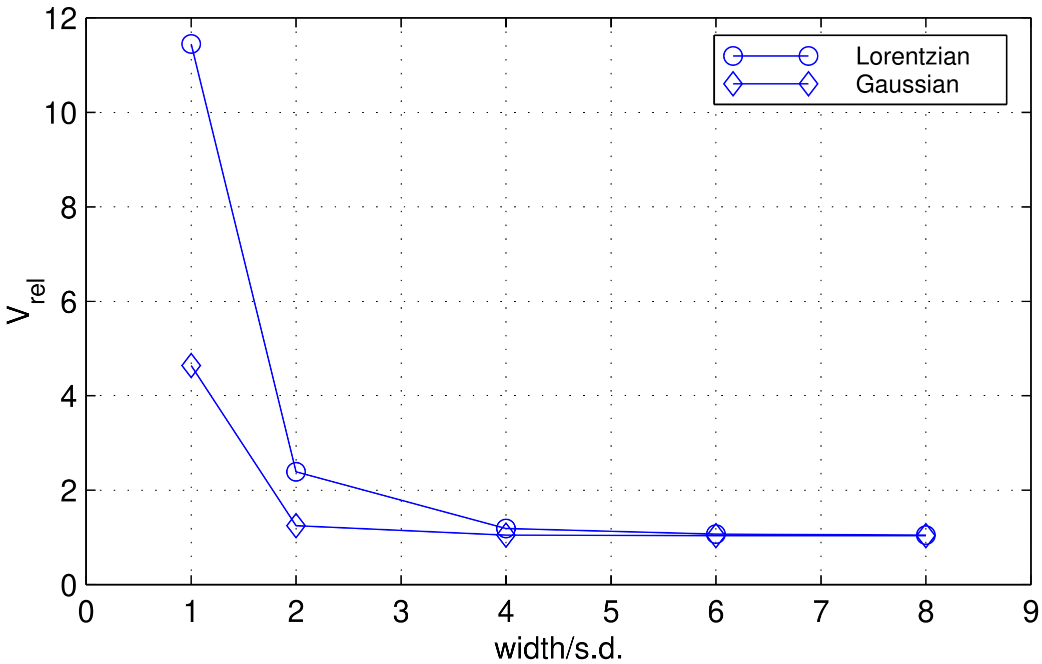

Fig. 4. The relative generalized variance of the ETA for tracks without mirror hits, as a function of the width w divided by the standard deviation of the measurement error.

simple gradient descent (GD) and for the Davidon-Fletcher-Powell algorithm (DFP). The ETA results have been obtained with a width w of four (ETA-G) and six (ETA-L) times the standard deviation of the measurement error. We also show the results of a plain least-squares estimator, implemented as a Kalman filter (KF). The last column of Table 1 shows the relative computing time t_{rel}. In order to have a meaningful comparison the annealing schedules are the same in all cases, with the exception of the GSF which does not require annealing. The timing refers to an implementation in MATLAB, an interpreter language; for compiler languages the behaviour may turn out to be different.

We can draw several conclusions from these results:

- If the EAA and the ETA are to achieve full precision, simple gradient descent is not sufficient. The Davidon-Fletcher-Powell algorithm which has been used here gives a satisfactory result.
- The Kalman filter, the DAF, the GSF, the EAA, and the ETA with a sufficiently wide potential give virtually the same result. This is no surprise as in the absence of mirror hits they all amount to a least-squares estimation.
- The Kalman filter is clearly much faster than all the other methods as it does not involve any iterations. Although the GSF never has to deal with more than one component the overhead in the code makes it no more than two times faster than the DAF. The annealing schedule of the DAF foresees six iterations, four of them at the final "temperature". As each iteration requires the computation of two filters (forward and backward filter) plus their weighted mean, we may expect it to be slower than the plain Kalman filter by a factor somewhat larger than 12. This is in good agreement with our experimental results. The elastic methods with the DFP minimization are somewhat slower than the DAF, by about 50 percent. About 80 percent of the time is spent in the search for the minimum energy at the final temperature. This is indispensable for a satisfactory precision.

Method	V_{rel}	t_{rel}
DAF	1.00	1.00
GSF	1.00	0.44
EAA with GD	8.83	1.49
EAA with DFP	1.03	1.63
ETA-G with GD	1.79	0.73
ETA-G with DFP	1.05	1.41
ETA-L with GD	125.6	0.65
ETA-L with DFP	1.07	1.67
KF	1.00	0.07

Table 1
The relative generalized variance for tracks without mirror hits.

Tracks with mirror hits

We now add the mirror hits to the tracks. Immediately there arise several questions:

– What is the best final value of w in the annealing schedule of the ETA?
– What is the best cutoff $\sqrt{\lambda}$ in the weights of the EAA and the DAF?
– What is the best final value of β in the annealing schedule of the EAA and the DAF?

In order to answer the first question we have run the ETA algorithm with different final values of the potential width w. It has turned out that even the DFP algorithm has difficulties in locating the global minimum of the energy function in some cases, particularly for small values of w. In order to give meaningful results we have eliminated these tracks by cutting off the tails in the residual distribution. The results after trimming are shown in Figure 5. Both types of potentials show roughly the same behaviour, although on a different scale. With the Lorentzian potential the minimum is at about three standard deviations; with the Gaussian potential it is between 1.5 and two standard deviations. The ratio is close to the expected factor of $\sqrt{2}$ (see Subsection 4.1). Without trimming the precision is worse by at least a factor of ten. The problem can be cured by slower cooling, with accordingly higher computational costs.

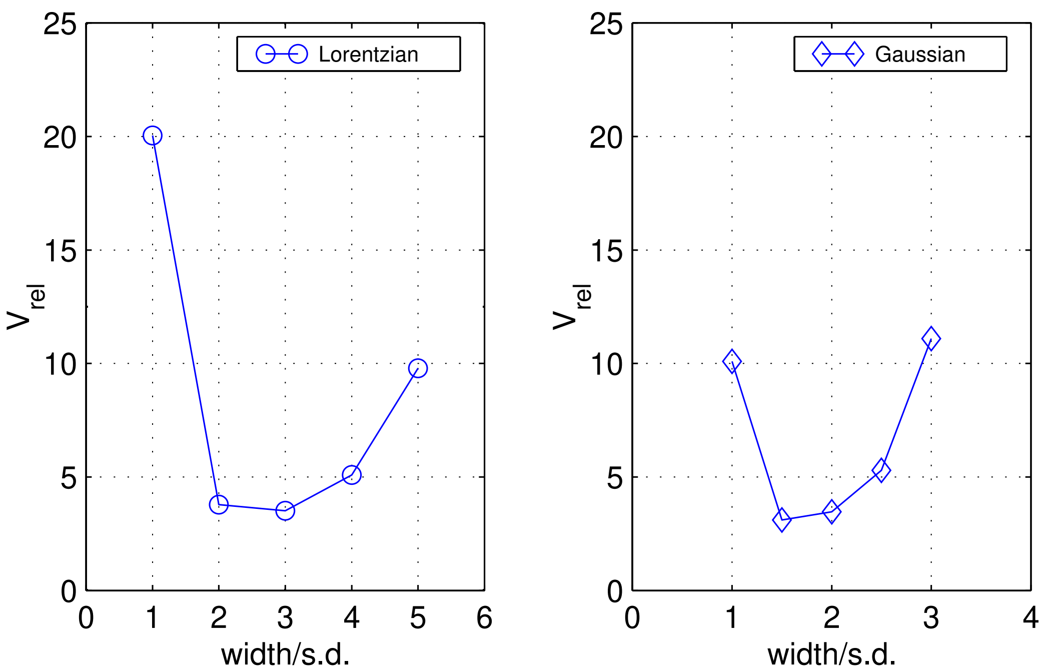

Fig. 5. The relative generalized variance of the ETA for tracks with mirror hits, as a function of the final width of the potential divided by the standard deviation of the measurement error.

The answer to the second question is pretty obvious: as the correct hit is always there, the cut should not be too tight in order not to lose it. On the other hand, the mirror hit will normally have low weight compared to the correct hit, so it does not matter if the cut is too loose. This is borne out by our simulation results. Figure 6 shows the relative generalized variance of the DAF as a function of the cutoff, the latter being given in terms of standard deviations of the measurement error. The relative variance drops from about 1.93 at a cutoff of 2.5 standard deviations to about 1.54 at four standard deviations; this is the optimum precision that can be obtained if the mirror hits are included, but no additional noise. Note that on average the r.m.s width of the estimates increases by only 7.5 percent if the mirror hits are included.

The last question about the final value of β can be answered by looking at Table 2. It shows the generalized variance of the DAF and of the EAA for two final values of the annealing schedule. The "nominal" value of beta corresponds to the standard deviation of the measurement error, whereas the "frozen" value is higher by a factor of 9. In all cases the cutoff has been set at four standard deviations. The results of the ETA have been obtained with a final potential width of 1.5 standard deviations with the Gaussian potential and three standard deviations with the Lorentzian potential. We also show the results of the GSF; the first one uses all components for the computation of the estimates, while the second one uses only the best (most probable) component. The latter is equivalent to a combinatorial Kalman filter (see Section 3) which explores all possible combinations and chooses the best one.

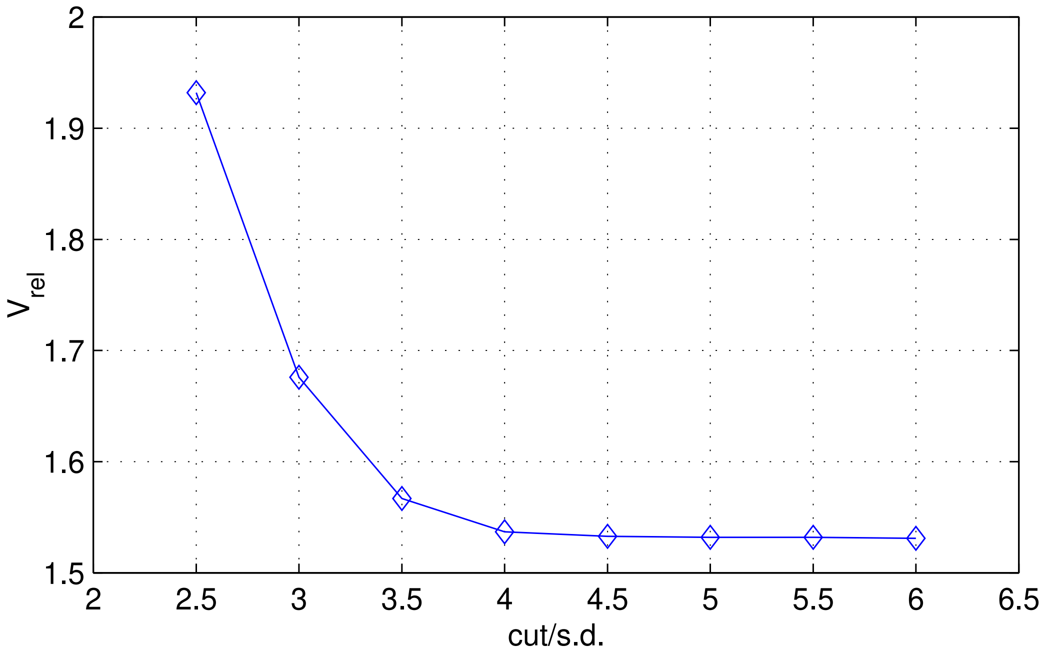

Fig. 6. The relative generalized variance of the DAF for tracks with mirror hits, as a function of the cutoff divided by the standard deviation of the measurement error.

Method	$V_{\rm rel}$	$t_{\rm rel}$
DAF nominal	1.54	1.21
DAF frozen	1.74	1.41
GSF all	1.59	7.04
GSF best	1.78	7.04
EAA nominal	1.56	2.12
EAA frozen	1.71	2.44
ETA-G with DFP	3.11	2.38
ETA-L with DFP	3.51	2.87
KF	~ 1500	0.08

Table 2
The relative generalized variance for tracks with mirror hits.

We make the following observations:

- The Kalman filter is strongly influenced by the mirror hits, resulting in a very large relative variance.
- The equivalence of the DAF and of the EAA is again illustrated by the nearly identical results.
- It does not pay to cool the DAF and the EAA below the standard deviation of the measurements; actually the precision of the estimates gets slightly worse.
- The GSF estimate using all components is comparable to the results obtained with the DAF and the EAA. The estimate using only the most probable component is slightly worse, comparable to the "frozen" version of the DAF and EAA.
- The ETA is worse than the other methods by a factor of about two, even after trimming. The precision with the Gaussian potential is somewhat better than with the Lorentzian.
- Both the DAF and the EAA are slowed down by the mirror hits, although the effect is not dramatic. The ETA is hardly slower, but it has to be kept in mind that the cooling is faster than it should be. The GSF is now very slow because of a much larger number of components, about 22 on average.

A point in favour of the DAF is the fact that the covariance matrix of the estimated track parameters is immediately available after the last iteration. This raises the question how well the actual spread of the estimates is described by this matrix. Our simulation experiment shows that with mirror hits the standardized residuals of the estimated track parameters have a standard deviation of about 1.08 to 1.09, indicating a fairly good but not perfect agreement between the actual spread and the elements of the covariance matrix, the latter being too small by about 20 percent.

Tracks with mirror hits and noise hits

We have concluded our investigation of the perfect tracks by studying the robustness of the estimators. To this end we have contaminated the perfect tracks with noise in the following way. In each tube the correct drift distance is replaced by a random drift distance with probability $p_{\text{noise}} = 0.1$. The random drift distance is drawn from a uniform distribution between 0 and the tube radius, i.e. 2 mm. Therefore in some of the tubes the correct hit (and its mirror hit) is missing; there are two wrong hits instead.

Because of this, the search for the optimal cutoff in the DAF and EAA becomes very important. Figure 7 shows the relative variance of the DAF as a function of the cut, the latter being given again in terms of standard deviations of the measurement error. A parabola fit to the results indicates that the optimal cutoff is at about 2.8 standard deviations. This value depends of course on the frequency and on the distribution of the noise hits. In any case, a cutoff at 3 standard deviations seems to be a reasonable compromise between minimizing the loss of correct hits and limiting the influence of noise hits. The relative generalized variance at the optimal cutoff is about 4; this means that the standard deviations of the estimated track parameters are larger by only about 25 percent as compared to the baseline (no noise, no mirror hits).

For the sake of comparison we have processed the sample with artificial noise also with the GSF, the Kalman filter (KF), and the ETA with the Davidon-

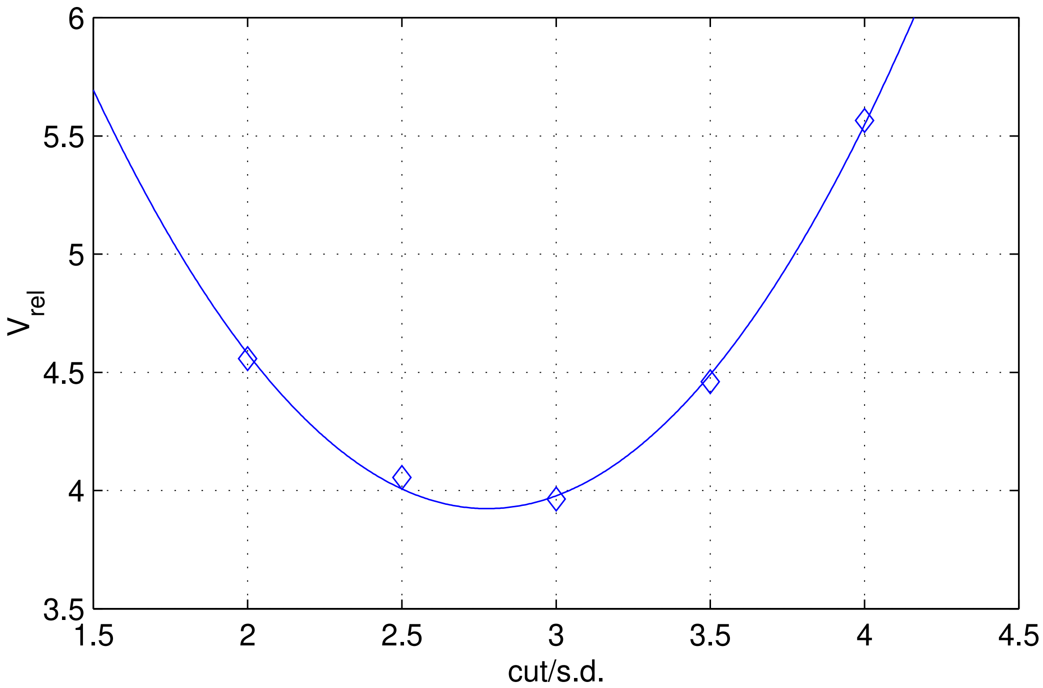

Fig. 7. The relative generalized variance of the DAF for tracks with noise hits, as a function of the cutoff divided by the standard deviation of the measurement error.

Fletcher-Powell minimization. The results are summarized in Table 3. We observe that the performance of the GSF has deteriorated considerably, because of its inherent lack of robustness. The precision of the ETA is again worse than the DAF, by a factor of about 1.5. As predicted, the Gaussian potential shows somewhat less susceptibility to mirror hits and noise than the Lorentzian. Note that again the tails in the residual distribution have been cut off. Without trimming the results are worse by more than a factor of ten. The timing of the algorithms is virtually the same as without noise.

Method	V_{rel}	t_{rel}
DAF	3.96	1.19
GSF	27.33	6.86
ETA-G with DFP	5.77	2.72
ETA-L with DFP	6.56	2.89
LS	~ 1600	0.08

Table 3
The relative generalized variance for tracks with mirror hits and noise hits.

5.3 Results with tracks from pattern recognition

Our investigation has been concluded by studying the performance of the DAF on tracks from pattern recognition. The data sample was a set of 100 simulated $H \longrightarrow b\bar{b}$ events without pile-up. In this sample the average number of tracks is about 66, the average number of tubes with a hit is about 2500, and the average number of tubes hit per track is about 38. About 4.4% of the hits are noise.

The initial track finding is performed by a modified Hough transform procedure. A detailed description of the algorithm can be found in [15]. The aim is to look for high-energetic tracks that come from the origin and go entirely through the barrel part of the TRT. The track search is therefore inititated in the outermost layers of the barrel. For each of the points in these layers one picks out a set of points in the mid-region of the barrel. This set consists of all points that can belong to a track with a transverse momentum larger than 1 GeV/c, coming from the origin and going through the current point in the outermost layers. For each of the points in this set it is checked whether this hypothesis is correct or not. This is done by counting the number of points in a circular band around the hypothesized track. If this number is small, no track is found. If the number is large, we assume that a track is found, and the set of points inside the circular band is accepted as a track candidate.

It can be noted that this procedure differs slightly from a combinatorial Hough transform. Such a Hough transform picks pairs of points in the same way as the method described above, but instead of checking whether the pairs of points really belong to a track, one fills up an accumulator array in parameter space and finally looks for peaks in this space.

The comparison of the estimated track parameters with the truth values is now less straightforward than it was with the perfect tracks. First, a reconstructed track may comprise observations form several simulated tracks. The truth values are now defined as the true track parameters of the simulated track contributing the majority of the observations. Second, the truth values are available only at the vertex and not at the entry of the TRT. The comparison is therefore affected by multiple scattering and secondary interactions in the material in front of the TRT. For this reason we have been forced to restrict our analysis to a subsample of tracks for which the truth values are reasonably close to the actual track, i.e. within 1 mm of at least one of the two ambiguous hits in the first layer hit by the track. This leaves us with a restricted sample of 1687 tracks, out of 2508 produced by the track finding.

Most of the track candidates are contaminated by hits belonging to other tracks; these should be counted as noise, in addition to occasional genuine noise hits. The fraction of noise hits in the track candidates in the restricted sample is shown in Figure 8a. The average noise content is close to 13%. It should be noted that track candidates with a noise content of more than 50% are eliminated right after the track finding.

In the next step we have classified the track candidates according to their noise content. Table 4 shows the range of the four classes, the number of tracks in each class, and the relative precision of the estimated track parameters, as measured by the generalized variance of the residuals with respect to the truth values. The baseline is now the class with the lowest noise content. Due to the material effects described earlier a direct comparison with the perfect tracks is not meaningful. All results have been obtained with the DAF, using the same cooling schedule and the same cutoff as before.

We note that in class 2, up to a noise fraction of 15%, the effect of the noise hits is negligible; this confirms the excellent robustness of the algorithm. We

Class	Noise fraction f_{n}	Number of tracks	V_{rel}
1	$0.00 \leq f_{\mathrm{n}} < 0.05$	559	1.00
2	$0.05 \leq f_{\mathrm{n}} < 0.15$	526	1.06
3	$0.15 \leq f_{\mathrm{n}} < 0.25$	305	4.09
4	$0.25 \leq f_{\mathrm{n}} \leq 0.50$	297	16.48

Table 4
The relative generalized variance of the DAF for pattern recognition tracks.

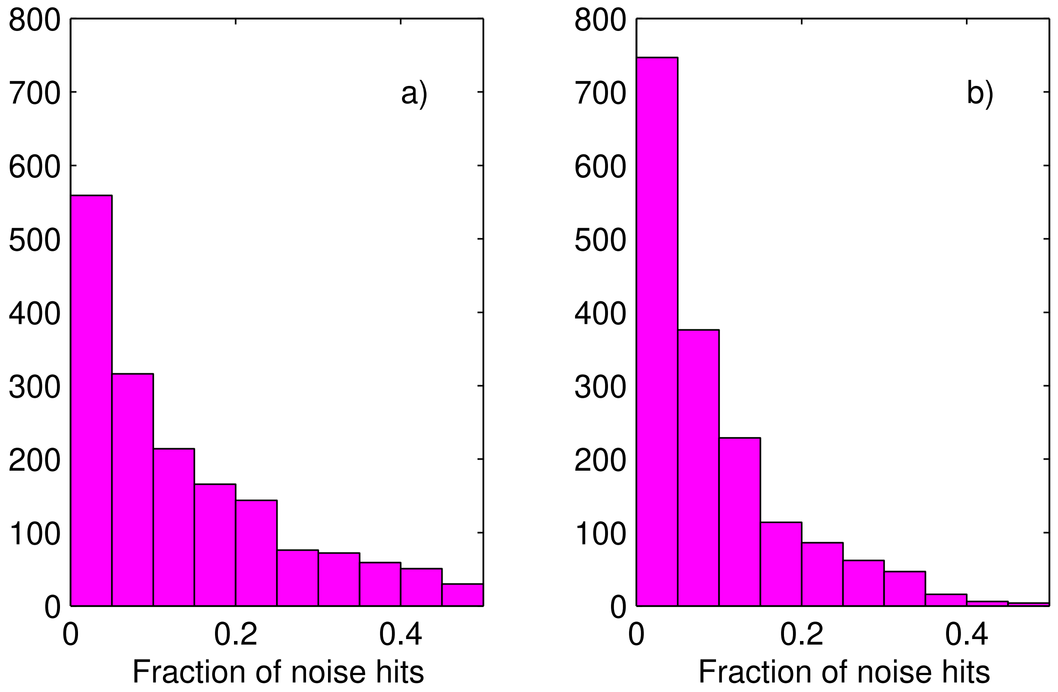

Fig. 8. Frequency distribution of the fraction of noise hits in the track candidates in the restricted sample a) before filtering and b) after filtering.

can also compute the noise fraction in the fitted tracks, by counting each hit with its final weight in the DAF. The frequency distribution of the noise content obtained in this way is shown in Figure 8b. The average noise content is now 8.7%, about two thirds of the value before the filter. This means that on average about one third of the noise hits are suppressed by the DAF. At a first glance this seems to be a rather modest performance; it should be noted, however, that all noise hits assigned to the candidate are necessarily very close to the “true” track as defined by the majority hits.

6 Conclusions and outlook

Track fitting in the presence of ambiguities and noise is basically a problem with incomplete data: it is not known a priori which of several competing hits in a layer of the tracking detector is the correct one. It is therefore hardly surprising that the EM algorithm — in various disguises — is a highly efficient and robust method for tackling this problem. A potential drawback of the EM algorithm, the dependence on the starting point of the iteration, is avoided by combining it with the idea of deterministic annealing. We have considered two implementations of this concept: the Elastic Arms Algorithm (EAA) and the Deterministic Annealing Filter (DAF). The EAA requires the numerical minimization of a non-quadratic objective (energy) function. This is a delicate task which we have solved by applying advanced functions from the MATLAB Optimization Toolbox [16]. We also have shown that a less sophisticated method like gradient descent does not give the required performance. In addition, the most accurate estimates are not found in the limit $T \to 0$ but rather at the natural value of T which corresponds to the variance of the observation error ($T = 2\sigma^2$). The DAF, on the other hand, is an iterated Kalman filter/smoother, with some additional code for computing the assignment probabilities. In terms of computing time the two methods are comparable, at least in our MATLAB implementation. The DAF, however, has two additional points in its favour: it is entirely straightforward to introduce process noise like multiple Coulomb scattering, and the covariance matrix of the estimated track parameters is immediately available.

The other methods considered in this study are not competitive with the EM algorithm. The Elastic Tracking Algorithm (ETA) is less precise, being neither a maximum likelihood nor a least-squares estimator, and less robust, as it lacks proper competition between hits. In addition, the minimization of the energy function is even more difficult than with the EAA. The Gaussian-sum filter (GSF) also suffers from a lack of robustness which can be cured in principle but only at the expense of considerable additional computation.

In this study we have restricted our attention to individual track candidates. We have not considered the case where several tracks are competing for the same hit. This situation can arise for instance in very narrow jets. In the original version of the EAA as proposed in [1] this possibility is already foreseen. If this general EAA is translated into an EM algorithm, one arrives at a scheme in which several iterated filters run in parallel, similar to the PMHT algorithm developed for the tracking of multiple targets [17]. We believe that such an approach could give significant improvements in the reconstruction of narrow jets and is certainly worth to be studied in detail.

It would also be interesting to see how the PMHT performs on the same kind of data that has been used in this study. A necessary modification of the plain PMHT would then be to adjust the assignment probabilities in order to obtain proper competition between the hits. A generalization of the algorithm to include a deterministic annealing scheme will probably also turn out to be vital. These topics are currently under study and will be the theme of a subsequent report.

References

[1] M. Ohlsson, C. Peterson and A. Yuille, Track finding with deformable templates — the elastic arms approach. *Computer Physics Communications* 71 (1992) 77.

[2] M. Lindström, Track reconstruction in the ATLAS detector using elastic arms. *Nuclear Instruments and Methods* A357 (1995) 129.

[3] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B* 39 (1977) 1.

[4] N. Ueda and R. Nakano, Deterministic annealing EM algorithm. *Neural Networks* 11 (1998) 271.

[5] M. Gyulassy and M. Harlander, Elastic tracking and neural network algorithms for complex pattern recognition. *Computer Physics Communications* 66 (1991) 31.

[6] R. Frühwirth, Track fitting with non-Gaussian noise. *Computer Physics Communications* 100 (1997) 1.

[7] R. Frühwirth and A. Strandlie, Track finding and fitting with the Gaussian-sum filter. *Proceedings of the Conference on Computing in High Energy Physics (CHEP'98)*, Chicago, September 1998.

[8] Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association.* Academic Press, Orlando, 1988.

[9] W. Press et al., *Numerical Recipes in C.* Cambridge University Press, Cambridge, 1988.

[10] G. Kitagawa, Non-Gaussian seasonal adjustment. *Computers & Mathematics with Applications* 18 (1989) 503.

[11] G. Kitagawa, The two-filter formula for smoothing and an inplementation of the Gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics* 46 (1994) 605.

[12] R. Frühwirth and S. Frühwirth-Schnatter, On the treatment of energy loss in track fitting. *Computer Physics Communications* 110 (1998) 80.

[13] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection.* John Wiley & Sons, New York, 1987.

[14] *ATLAS Inner Detector Technical Design Report.* CERN/LHCC 97-16, CERN, Geneva, 1997.

[15] B. Lillekjendlie and A. Strandlie, Particle tracking in the ATLAS Inner Detector. *Proceedings of the Conference on Computing in High Energy Physics (CHEP'98)*, Chicago, September 1998.

[16] The MathWorks, Inc., *MATLAB Optimization Toolbox User's Guide.* Version 5, 1996.

[17] R.L. Streit and T.E. Luginbuhl, *Probabilistic Multi-Hypothesis Tracking.* NUWC-NPT Technical Report 10 428, Naval Undersea Warfare Center, Newport, Rhode Island, 1995.

Figure captions

Figure 1:
The potential V as a function of distance x, expressed in standard deviations of the measurement error. The cut $\sqrt{\lambda}$ is set at three standard deviations. a) EAA at $T = 0$. Note the abrupt change in the potential's behaviour at $|x| = 3$. b) EAA at $T = 2\sigma^2$. This temperature corresponds to the width of the measurement error. c) The Gaussian potential of the ETA with the same curvature at $x = 0$. d) The Lorentzian potential of the ETA with the same curvature at $x = 0$.

Figure 2:
The association probability of a single hit as a function of its distance from the predicted track position, measured in standard deviations of the measurement error. The step function is used by the DAF, the continuous function is used by the EAA at $\beta = 1/(2V_k)$. The cut is set at four standard deviations.

Figure 3:
Frequency distribution of a) the transverse momentum of the tracks and b) the number of tubes hit by a track

Figure 4:
The relative generalized variance of the ETA for tracks without mirror hits, as a function of the width w divided by the standard deviation of the measurement error.

Figure 5:
The relative generalized variance of the ETA for tracks with mirror hits, as a function of the final width of the potential divided by the standard deviation of the measurement error.

Figure 6:
The relative generalized variance of the DAF for tracks with mirror hits, as a function of the cutoff divided by the standard deviation of the measurement error.

Figure 7:
The relative generalized variance of the DAF for tracks with noise hits, as a function of the cutoff divided by the standard deviation of the measurement error.

Figure 8:
Frequency distribution of the fraction of noise hits in the track candidates in the restricted sample a) before filtering and b) after filtering.

Table captions

Table 1:
The relative generalized variance for tracks without mirror hits.

Table 2:
The relative generalized variance for tracks with mirror hits.

Table 3:
The relative generalized variance for tracks with mirror hits and noise hits.

Table 4:
The relative generalized variance of the DAF for pattern recognition tracks.