ELSEVIER

Computer Physics Communications 85 (1995) 189–199

Computer Physics
Communications

Track fitting with long-tailed noise: a Bayesian approach

Rudolf Frühwirth *

Institut für Hochenergiephysik der Österreichischen Akademie der Wissenschaften, Nikolsdorfer Gasse 18, A-1050 Wien, Austria

Received 27 July 1994

Abstract

If the measurement noise in a linear dynamic system is non-Gaussian, the optimal linear filter (Kalman filter) is not necessarily the one with minimum variance. We describe a non-linear filter, based on a Bayesian approach, which performs better than the linear filter. The relative efficiency of the non-linear filter in the context of track reconstruction is determined in a simulation study. As the filter presupposes a Gaussian mixture model of the measurement noise, we address the problem of approximating the distribution of the measurement errors by a Gaussian mixture. We also study the performance of the filter on some types of long-tailed distributions other than Gaussian mixtures. Finally, the filter is extended to cope with long-tailed process noise, for example a Gaussian mixture model of multiple scattering.

1. Introduction

Most track reconstruction algorithms currently in use are based on linear least-squares estimators. A typical example is the Kalman filter plus smoother, now widely used for track and vertex reconstruction in collider experiments. The justification for using a linear least-squares estimator is twofold. First, it is assumed that the track model, if not already linear, can be approximated by a linear function, i.e. a first order Taylor expansion, in a sufficiently large neighbourhood of the measurements. The quality of this linear approximation can in most cases be improved by a careful choice of the expansion point. Second, it is assumed that the distribution of the measurement errors is Gaussian, or at least very close to Gaussian. If both of these assumptions hold, the linear least-squares estimator with the correct weight matrix not only is the best linear estimator, but is efficient. In this case no non-linear estimator can do better.

In this note we want to investigate the case where the second assumption fails. In real detectors the measurement errors are hardly ever Gaussian, and there is virtually always a tail of outlying observations. In principle there are two approaches to the treatment of these outliers [1]. In the first approach one tries to identify and eliminate the outliers. This is an iterative and lengthy procedure, which is not guaranteed to be unambiguous, particularly if there are several outliers. If one wants to avoid this, there is a second approach: one tries to accommodate the outliers by a modified robust estimator. For example, outlying measurements can simply be

* E-mail: fruhwirth@hephy.oeaw.ac.at.

SSDI 0010-4655(94)00121-9

downweighted (M-estimator). More sophisticated approaches make use of prior information or assumptions on the distribution of the outliers. In the following the observation distribution will be modelled by a Gaussian mixture with two components, the first one describing the "regular" measurements or the "core", the second one describing the outliers or the "tails". It is natural to assume that the variance of the second component is larger than the variance of the first one, and that outliers are relatively rare with respect to the regular measurements. This model is in general well-suited to the observations in an actual detector.

A linear least-squares estimator uses only the first and second moment (mean and variance) of the observation distribution; therefore it is blind to the specific form of the distribution. An estimator taking this form into account is therefore necessarily non-linear. Non-linear estimators are unpopular for several reasons. Usually they need more computing time, and their asymptotic properties are less well known. Nevertheless they can have a smaller variance than the optimal linear estimator if the observation distribution is sufficiently non-Gaussian, for instance a Gaussian mixture as described above. In Section 2 we present a robust non-linear modification of the Kalman filter which is based on a proposal of Guttman and Peña [2]. In Section 3 its efficiency in the context of track fitting is evaluated on a sample of simulated tracks in an idealized tracking detector. This robust filter presupposes a Gaussian mixture model of the observation errors. Section 4 deals with the sensitivity of the filter to the prior assumptions on the mixture parameters and presents a method of estimating the mixture parameters from a sample of tracks. In Section 5 we extend the investigation of the robust filter to some other types of long-tailed distributions. Finally, in Section 6 we describe how the robust filter can be adapted to long-tailed process noise, i.e. to a non-Gaussian model of multiple scattering.

2. A robust non-linear filter

Our starting point is a linear track model suitable for estimation of the track parameters by the Kalman filter [3]. In most cases the linear track model is actually a first order Taylor approximation to a non-linear model. The model is specified by a set of system equations and a set of measurement equations.

System equations:

$$\boldsymbol{x}_k = \boldsymbol{F}_k\,\boldsymbol{x}_{k-1} + \boldsymbol{c}_k + \boldsymbol{\omega}_k,$$
$$\mathsf{E}(\boldsymbol{\omega}_k) = \boldsymbol{0}, \quad \mathsf{cov}(\boldsymbol{\omega}_k) = \boldsymbol{Q}_k, \quad k = 1,\dots,n.$$

Measurement equations:

$$\boldsymbol{m}_k = \boldsymbol{H}_k\,\boldsymbol{x}_k + \boldsymbol{d}_k + \boldsymbol{\epsilon}_k,$$
$$\mathsf{E}(\boldsymbol{\epsilon}_k) = \boldsymbol{0}, \quad \mathsf{cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{V}_k = \boldsymbol{G}_k{}^{-1}, \quad k = 1,\dots,n.$$

Here, $\boldsymbol{x}_k$ denotes the state vector of the five track parameters at measurement surface k, i.e. the intersection point, the track direction, and the curvature. The linear model is described by the system matrix $\boldsymbol{F}_k$ and the constant term $\boldsymbol{c}_k$. The process noise between surface $k-1$ and k is denoted by $\boldsymbol{\omega}_k$. If energy loss is neglected, it is the sum of the integrated continuous multiple scattering plus all discrete scattering between surface $k-1$ and k. The measurements in surface k are denoted by $\boldsymbol{m}_k$, and the associated observation error by $\boldsymbol{\epsilon}_k$. The linear function which maps the state vector $\boldsymbol{x}_k$ on the measurement vector $\boldsymbol{m}_k$ is defined by the matrix $\boldsymbol{H}_k$ and the constant term $\boldsymbol{d}_k$.

The formulas for the computation of the predicted, filtered, and smoothed least-squares estimates of the state vector are well-known and can be found in the literature [3], along with the formulas for the corresponding covariance matrices and χ^2-statistics.

We now assume that the distribution of the observation error $\boldsymbol{\epsilon}_k$ can be modelled by a Gaussian mixture:

$$f(\boldsymbol{\epsilon}_k) = p_k{}^{(0)} \cdot \varphi(\boldsymbol{\epsilon}_k; \mathbf{0}, \boldsymbol{V}_k{}^{(0)}) + p_k{}^{(1)} \cdot \varphi(\boldsymbol{\epsilon}_k; \mathbf{0}, \boldsymbol{V}_k{}^{(1)}), \quad p_k{}^{(0)} + p_k{}^{(1)} = 1,$$

where $\varphi(\cdot; \boldsymbol{\mu}, \boldsymbol{V})$ is a multivariate Gaussian p.d.f. with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$. $\boldsymbol{V}_k{}^{(0)} = (\boldsymbol{G}_k{}^{(0)})^{-1}$ is the covariance matrix of the regular measurements, $\boldsymbol{V}_k{}^{(1)} = (\boldsymbol{G}_k{}^{(1)})^{-1}$ is the covariance matrix of the outliers. It is reasonable to postulate that $p_k{}^{(0)} > p_k{}^{(1)}$ and $\boldsymbol{V}_k{}^{(1)} > \boldsymbol{V}_k{}^{(0)}$, although this is not essential to the method. The covariance matrix of $\boldsymbol{\epsilon}_k$ is given by

$$\boldsymbol{V}_k = p_k{}^{(0)} \cdot \boldsymbol{V}_k{}^{(0)} + p_k{}^{(1)} \cdot \boldsymbol{V}_k{}^{(1)}.$$

In order to derive the robust filter step, the distribution of the predicted estimate is approximated by a normal distribution with mean $\tilde{\boldsymbol{x}}_k^{k-1}$ and covariance matrix $\boldsymbol{C}_k^{k-1}$, as in the case of the Kalman filter. The posterior distribution of the estimate $\boldsymbol{x}_k$ can then be computed by means of Bayes' theorem [2]:

$$f(\boldsymbol{x}_k | \boldsymbol{m}_1, \ldots, \boldsymbol{m}_k) = \sum_{i=0}^{1} q_k{}^{(i)} \cdot \varphi(\boldsymbol{x}_k; \tilde{\boldsymbol{x}}_k{}^{(i)}, \boldsymbol{C}_k{}^{(i)}),$$

with:

$$\tilde{\boldsymbol{x}}_k{}^{(i)} = \tilde{\boldsymbol{x}}_k^{k-1} + \boldsymbol{C}_k^{k-1} \boldsymbol{H}_k{}^T \boldsymbol{W}_k{}^{(i)} \boldsymbol{r}_k^{k-1}, \qquad \boldsymbol{r}_k^{k-1} = \boldsymbol{m}_k - \boldsymbol{d}_k - \boldsymbol{H}_k \tilde{\boldsymbol{x}}_k^{k-1},$$
$$\boldsymbol{W}_k{}^{(i)} = (\boldsymbol{V}_k{}^{(i)} + \boldsymbol{H}_k \boldsymbol{C}_k^{k-1} \boldsymbol{H}_k{}^T)^{-1}, \qquad \boldsymbol{C}_k{}^{(i)} = [(\boldsymbol{C}_k^{k-1})^{-1} + \boldsymbol{H}_k{}^T \boldsymbol{G}_k{}^{(i)} \boldsymbol{H}_k]^{-1}.$$

The coefficients $q_k{}^{(i)}$ can be interpreted as the posterior probabilities of the measurement $\boldsymbol{m}_k$ being an outlier or not:

$$q_k{}^{(0)} = \left[1 + \frac{p_k{}^{(1)}}{p_k{}^{(0)}} \frac{|\boldsymbol{W}_k{}^{(1)}|}{|\boldsymbol{W}_k{}^{(0)}|} \exp\left(\tfrac{1}{2} \boldsymbol{r}_k^{k-1T} \boldsymbol{D}_k \boldsymbol{r}_k^{k-1} \right) \right]^{-1}, \qquad q_k{}^{(1)} = 1 - q_k{}^{(0)},$$

with:

$$\boldsymbol{D}_k = \boldsymbol{W}_k{}^{(0)} - \boldsymbol{W}_k{}^{(1)}.$$

The final estimate $\tilde{\boldsymbol{x}}_k$ and its covariance matrix $\boldsymbol{C}_k$ are obtained as the mean and the covariance matrix of the posterior distribution of $\boldsymbol{x}_k$. The update of the state vector turns out to be a weighted sum of two Kalman filters, the weights being $q_k{}^{(0)}$ and $q_k{}^{(1)}$ [2]:

$$\tilde{\boldsymbol{x}}_k = \tilde{\boldsymbol{x}}_k^{k-1} + \boldsymbol{C}_k^{k-1} \boldsymbol{H}_k{}^T (q_k{}^{(0)} \boldsymbol{W}_k{}^{(0)} + q_k{}^{(1)} \boldsymbol{W}_k{}^{(1)}) \boldsymbol{r}_k^{k-1},$$
$$\boldsymbol{C}_k = \boldsymbol{C}_k^{k-1} - \boldsymbol{C}_k^{k-1} \boldsymbol{H}_k{}^T (q_k{}^{(0)} \boldsymbol{W}_k{}^{(0)} + q_k{}^{(1)} \boldsymbol{W}_k{}^{(1)} - \boldsymbol{S}_k) \boldsymbol{H}_k \boldsymbol{C}_k^{k-1},$$
$$\boldsymbol{S}_k = q_k{}^{(0)} q_k{}^{(1)} \boldsymbol{D}_k \boldsymbol{r}_k^{k-1} \boldsymbol{r}_k^{k-1T} \boldsymbol{D}_k.$$

The posterior distribution of $\boldsymbol{x}_k$, being a mixture of two Gaussians with different means, is asymmetric; therefore $\tilde{\boldsymbol{x}}_k$ is in general not identical to the maximum-likelihood or posterior mode estimate. If $q_k{}^{(1)} = 0$ or, a fortiori, $p_k{}^{(1)} = 0$, the robust filter reduces to the Kalman filter.

The prior p.d.f. of the residual $\boldsymbol{r}_k^{k-1}$ is a mixture of two Gaussians with zero mean:

$$f(\boldsymbol{r}_k^{k-1}) = \sum_{i=0}^{1} p_k{}^{(i)} \cdot \varphi(\boldsymbol{r}_k^{k-1}; \mathbf{0}, \boldsymbol{V}_k{}^{(i)} + \boldsymbol{H}_k \boldsymbol{C}_k^{k-1} \boldsymbol{H}_k{}^T).$$

Therefore $\boldsymbol{R}_k^{k-1}$, the covariance matrix of $\boldsymbol{r}_k^{k-1}$, is given by:

$$R_k^{k-1} = p_k{}^{(0)} V_k{}^{(0)} + p_k{}^{(1)} V_k{}^{(1)} + H_k C_k^{k-1} H_k{}^T = V_k + H_k C_k^{k-1} H_k{}^T.$$

The filtered residual is given by:

$$r_k = m_k - d_k - H_k \tilde{x}_k = [I - H_k C_k^{k-1} H_k{}^T (q_k{}^{(0)} W_k{}^{(0)} + q_k{}^{(1)} W_k{}^{(1)})] r_k^{k-1}.$$

We compute the derivative of r_k w.r.t. r_k^{k-1}:

$$\partial r_k / \partial r_k^{k-1} = I - H_k J_k, \qquad J_k = C_k^{k-1} H_k{}^T (q_k{}^{(0)} W_k{}^{(0)} + q_k{}^{(1)} W_k{}^{(1)} - S_k).$$

J_k can be considered as a modified gain matrix which reduces to the Kalman gain matrix if $q_k{}^{(1)} = 0$. Using J_k, we compute R_k, the covariance matrix of r_k, by linear error propagation:

$$R_k = (I - H_k J_k) R_k^{k-1} (I - J_k{}^T H_k{}^T).$$

Finally, we obtain a generalized χ^2-statistic of the filter step. We observe that its χ^2-increment can be computed as in the standard case:

$$\chi^2_{k,F} = r_k{}^T G_k r_k + (\tilde{x}_k - \tilde{x}_k^{k-1})^T (C_k^{k-1})^{-1} (\tilde{x}_k - \tilde{x}_k^{k-1}).$$

The weight matrix $G_k = V_k{}^{-1}$ can be computed using either the prior or the posterior probabilities, yielding two different statistics. Neither of them is, of course, actually χ^2-distributed.

If the prior distribution of x_k^{k-1} is Gaussian, the posterior distribution of x_k is a mixture of two Gaussians. An exact prediction in the subsequent filter step would result in the posterior of x_{k+1} being a mixture of four Gaussians, and continuing in this way would yield an exponentially increasing number of components. In order to keep the filter simple, the posterior distribution of x_k is approximated in each step by a single Gaussian with mean $\tilde{x}_k$ and covariance matrix C_k. It has been shown that this approximation is optimal in the sense that it minimizes the Kullback-Liebler distance between the two density functions [4]. The smoother is not affected by the robustification and remains unchanged.

3. The relative efficiency of the non-linear filter

In order to evaluate the possible gain in efficiency by using the robust filter we have conducted a simulation study in an idealized track detector. In order to simplify the study the detector is assumed to be homogeneous in the sense that the distribution of the observation error is the same in every measurement surface, irrespective of the track parameters. Typically this is the case in a central track detector like a TPC or a silicon tracker. The detector is rotationally symmetric w.r.t. the z-axis and consists of 12 cylindrical measurement surfaces at radii $R = 30, 35, \ldots, 80, 85$ cm. In every surface two co-ordinates are measured, $R\Phi$ and z, where Φ is the azimuth of the crossing point. The standard deviation of the measurement is assumed to be 0.2 mm for $R\Phi$ and 0.5 mm for z, which are typical values in actual detectors at high energy colliders. The correlation between the measurements is set to zero. The magnetic field is assumed to be homogeneous and parallel to z, resulting in a helical track model. We have used a standard sample of 10000 tracks with radii between 300 and 3000 cm, corresponding roughly to a p_T between 1 and 10 GeV at a field of 1.1 Tesla.

We have evaluated the efficiency of the robust filter relative to the optimal linear filter systematically for a wide range of Gaussian mixture distributions of the observation error. The total variance of the observation error was the same in every case, corresponding to the standard deviations quoted above. The efficiency of the estimator is measured by the generalized variance of the five estimated track parameters $\tilde{x} = (R\Phi, z, \vartheta, \varphi, 1/r)$, i.e. the determinant of the sample covariance matrix C of $(\tilde{x} - x_{true})$:

Table 1
Variance of the robust estimate relative to the linear estimate

p =	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
$\varrho = 1.0$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\varrho = 1.5$	1.00	0.994	0.979	0.969	0.962	0.957	0.959	0.960	0.960	0.959	0.960
$\varrho = 2.0$	1.00	0.931	0.854	0.812	0.785	0.777	0.789	0.803	0.819	0.824	0.838
$\varrho = 2.5$	1.00	0.806	0.662	0.599	0.573	0.566	0.591	0.626	0.668	0.684	0.718
$\varrho = 3.0$	1.00	0.661	0.492	0.431	0.412	0.413	0.446	0.495	0.557	0.583	0.635
$\varrho = 3.5$	1.00	0.523	0.363	0.326	0.312	0.328	0.366	0.417	0.495	0.524	0.585
$\varrho = 4.0$	1.00	0.406	0.269	0.255	0.246	0.273	0.317	0.378	0.467	0.498	0.568
$\varrho = 4.5$	1.00	0.310	0.207	0.213	0.214	0.243	0.301	0.364	0.468	0.501	0.568
$\varrho = 5.0$	1.00	0.271	0.178	0.190	0.202	0.234	0.301	0.359	0.482	0.521	0.588

$$C = \mathrm{E}[(\tilde{x} - x_{true})(\tilde{x} - x_{true})^T] - [\mathrm{E}(\tilde{x} - x_{true})][\mathrm{E}(\tilde{x} - x_{true})]^T,$$

where the expectation operator denotes the sample average.

As the detector is homogeneous the Gaussian mixture distribution can be specified by two global parameters p and ϱ, where $p \leq \frac{1}{2}$ is the probability of an outlier and $\varrho \geq 1$ is the ratio of standard deviations σ_1/σ_0. If the variance of the observation error is denoted by σ^2, then

$$\sigma_0^2 = \sigma^2/(1 - p + p\varrho^2), \quad \sigma_1^2 = \sigma^2\varrho^2/(1 - p + p\varrho^2).$$

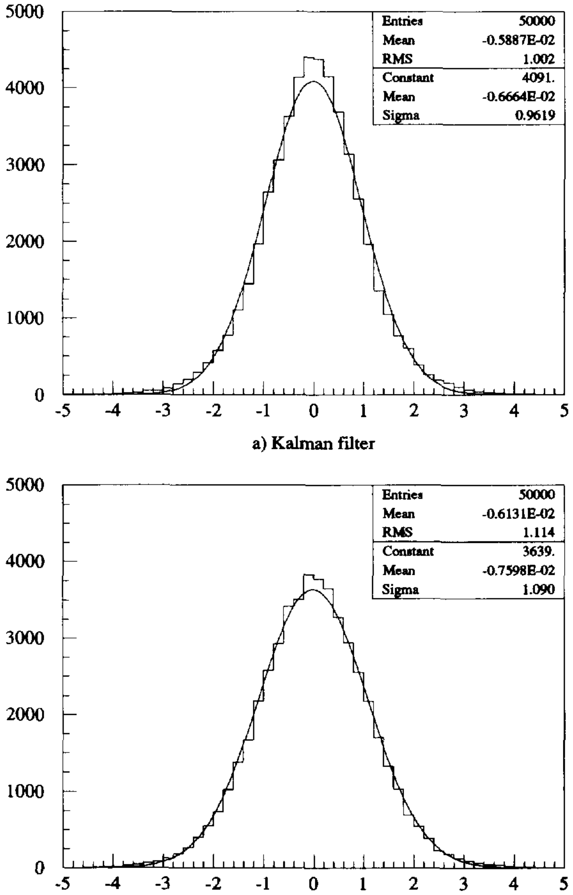
For the sake of simplicity, the same values of p and ϱ are chosen for both $R\Phi$- and z-measurements.

Table 1 shows the inverse relative efficiency η of the robust filter, i.e. the generalized variance of the robust estimate divided by the generalized variance of the optimal linear estimate. Note that both $p = 0$ and $\varrho = 1$ yield the optimal linear filter. The prior probabilities and the respective variances of regular measurements and outliers required by the robust filter have been set to the true values used in the simulation of the measurements. The issue of estimating these quantities from the data will be addressed in the next section.

It is fortunate that the robust filter yields the largest gain in efficiency for relatively small outlier probabilities between 10% and 20%, i.e. precisely in the range which is the most relevant for applications to track reconstruction. In the realistic case of 20% contamination and threefold standard deviation of the tails the generalized variance of the robust estimate is about 40% of the generalized variance of the optimal linear estimate, a non-negligible gain of information.

Fig. 1 shows the normalized differences of the estimated and the true values of the track parameters for both the Kalman filter and the robust filter, for $p = 0.2$ and $\varrho = 3$. In addition, a Gaussian has been fitted to both frequency distributions. In the case of the Kalman filter, the distribution of the normalized differences is correct as far as the first two moments are concerned, although the shape is not a perfect Gaussian. This is also reflected in the fact that the standard deviation of the fitted Gaussian is significantly smaller than 1. With the robust filter, the shape is nearly Gaussian, the fitted standard deviation being closer to the r.m.s. of the observed frequency distribution. However, the r.m.s. is about 10% too large, indicating that the covariance matrix of the estimate is somewhat too small. Clearly, the robust filter, like the linear filter, is unbiased.

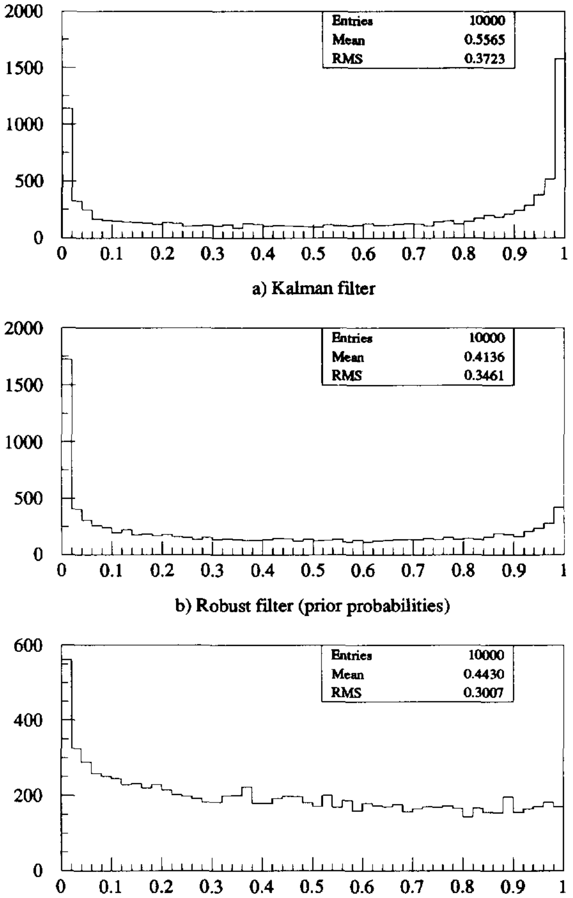
The distribution of the χ^2-probability of both filters is shown in Fig. 2. The probability of the Kalman filter (a) displays the U-shape characteristic for data contaminated with outliers. The average χ^2 is, however, correct, as it is bound to be with the linear estimator. With the robust filter, the χ^2 can be computed in two ways, either with the prior probabilities (b) or with the posterior probabilities (c). Whereas (b) looks similar to (a), the distribution in (c) is much closer to a uniform distribution, although there is still a spike at very small probabilities, the average χ^2 being slightly too large (20.6 instead of 19.0).



a) Kalman filter

b) Robust filter

Fig. 1.



a) Kalman filter

b) Robust filter (prior probabilities)

c) Robust filter (posterior probabilities)

Fig. 2.

Fig. 1. Normalized residuals of the estimated track parameters (p=0.2, ϱ=3).

Fig. 2. Probability transform of the total χ^2 (p=0.2, ϱ=3).

4. Determination of mixture parameters

The results of the preceding section have been obtained by plugging the correct mixture model into the robust filter. This is of course possible only with simulated data. In a real-world application the model has to be determined from a selected subsample of tracks, possibly from a calibration experiment. We now turn to the problem of estimating the mixture parameters from such a sample of tracks. The first question which arises in this context is the following: How sensitive is the filter to wrong assumptions on the mixture parameters? In order to find an answer we have simulated a sample of tracks with $p = 0.2$ and $\varrho = 3$. The sample was then reconstructed with different prior values of p and ϱ. Table 2 shows the generalized variance of the estimate relative to the estimate with the correct model.

The table shows that – at least in this example – the filter is not very sensitive to the prior assumptions on the mixture model. It is somewhat unexpected that some of the entries are smaller than 1, implying that the correct model does not yield the estimator with the smallest variance. For values of ϱ smaller than the true

Table 2
Variance of the robust estimate relative to the correct model (p=0.2, ϱ=3)

p =	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
$\varrho = 1.0$	2.43	2.43	2.43	2.43	2.43	2.43	2.43	2.43	2.43	2.43	2.43
$\varrho = 1.5$	2.43	1.87	1.69	1.59	1.52	1.48	1.45	1.44	1.44	1.46	1.48
$\varrho = 2.0$	2.43	1.50	1.27	1.15	1.08	1.03	1.01	1.00	1.00	1.02	1.04
$\varrho = 2.5$	2.43	1.36	1.13	1.02	0.96	0.93	0.92	0.92	0.92	0.93	0.95
$\varrho = 3.0$	2.43	1.32	1.11	1.03	*1.00*	1.00	1.01	1.01	1.02	1.02	1.02
$\varrho = 3.5$	2.43	1.32	1.16	1.16	1.19	1.22	1.24	1.25	1.25	1.23	1.20
$\varrho = 4.0$	2.43	1.36	1.32	1.41	1.50	1.59	1.64	1.66	1.62	1.58	1.48
$\varrho = 4.5$	2.43	1.44	1.63	1.85	2.08	2.25	2.33	2.30	2.19	2.03	1.86
$\varrho = 5.0$	2.43	1.58	2.08	2.57	3.07	3.35	3.38	3.23	2.97	2.66	2.34

one, the relative variance is remarkably flat, indicating little sensitivity to the choice of p. In contrast, if ϱ is larger than the true value we observe a curious, doubly peaked functional dependence on p, which is difficult to explain by a straightforward argument. The smallest relative variance is obtained if the width of the outlier distribution is slightly underestimated and the frequency of outliers is overestimated.

Our method of estimating the mixture model parameters is based on the observation that the moments of the χ^2-statistic of the linear filter can be calculated explicitely. We propose to estimate σ^2, p and ϱ from the first three moments. Although the functional dependence of the moments on σ^2, p and ϱ cannot be inverted analytically, it can be inverted numerically, thus yielding the desired estimator.

Actually, the linear filter is equivalent to a linear regression. The regression model can be written in the following form:

$$\boldsymbol{y} = \boldsymbol{A} \cdot \boldsymbol{x} + \boldsymbol{b} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is the vector of measurements, $\boldsymbol{A}$ and $\boldsymbol{b}$ specify the linear model, and $\boldsymbol{x}$ is the vector of track parameters. If multiple scattering is neglected, the covariance matrix of $\boldsymbol{\epsilon}$ is diagonal. As the measurements of $R\Phi$ and z are uncorrelated, we set up two separate regression models, one for $(R\Phi, \varphi, 1/r)$, one for $(z, \vartheta, 1/r)$. Thus the dimension of $\boldsymbol{y}$ is $n = 12$, and the dimension of $\boldsymbol{x}$ is $m = 3$. $\boldsymbol{A}$ can be assumed to have rank m. The distribution of $\boldsymbol{\epsilon}$ is in both cases a Gaussian mixture of $2^{12} = 4096$ components. Each component is described by a multi-index $I \in \{0, 1\}^n$. If $I_j = 0$, y_j is a regular measurement, otherwise it is an outlier. The covariance matrix of the component with index I is given by

$$V_I = \mathrm{diag}(\sigma^2_{I_1}, \ldots, \sigma^2_{I_n}),$$

where σ_0^2 is the variance of the regular measurements and σ_1^2 is the variance of the outliers. Under the assumption that the occurence of outliers is independent in different measurements, the proportion p_I of component I in the mixture turns out to be

$$p_I = (1-p)^{n-\Sigma I} \cdot p^{\Sigma I}, \quad \Sigma I = \sum_{j=1}^{n} I_j.$$

The χ^2-statistic of the regression can be written as

$$\chi^2 = (\boldsymbol{y} - \boldsymbol{b})^T \boldsymbol{B} (\boldsymbol{y} - \boldsymbol{b}), \quad \boldsymbol{B} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T.$$

$\boldsymbol{B}$ is symmetric and idempotent, and $\mathrm{tr}(\boldsymbol{B}) = n - m$. Let us now consider the distribution of χ^2 conditional on a fixed component I. The conditional cumulants of χ^2 are given by [5, p. 357]:

Table 3
True versus estimated parameters of the Gaussian mixture model

True values			Estimated values			
p	ϱ	η	$\sigma_{R\Phi}$	p	ϱ	η
0.15	2.25	0.705	0.201	0.13	2.28	0.709
0.15	2.75	0.507	0.201	0.13	2.77	0.511
0.15	3.25	0.372	0.201	0.14	3.29	0.375
0.25	2.25	0.668	0.201	0.23	2.25	0.671
0.25	2.75	0.481	0.201	0.24	2.73	0.480
0.25	3.25	0.365	0.201	0.25	3.25	0.365

$$\kappa_s^I = 2^{s-1}(s-1)!\,\mathrm{tr}[(S_I \boldsymbol{B} S_I)^s], \quad s \geq 1,$$

where S_I is a square root of V_I. Next, we compute the conditional moments μ_s^I about 0 up to order 3 via the relations [5, p. 69]:

$$\mu_1^I = \kappa_1^I, \qquad \mu_2^I = \kappa_2^I + (\kappa_1^I)^2, \qquad \mu_3^I = \kappa_3^I + 3\kappa_2^I\kappa_1^I + (\kappa_1^I)^3.$$

It is easily seen that the moments around 0 of the unconditional distribution of χ^2 are mixtures of the conditional moments, the proportion of μ_s^I being equal to p_I:

$$\mu_s = \sum_I p_I \mu_s^I, \quad s = 1, 2, 3.$$

In particular, the expectation of χ^2 is given by

$$\mathrm{E}(\chi^2) = \sum_I p_I\,\mathrm{tr}(S_I \boldsymbol{B} S_I) = \sigma^2(n-m),$$

a well-known result. From this relation we can estimate the total variance of $\boldsymbol{\epsilon}$.

There remain two more parameters, p and ϱ, to be determined from the second and third moment of the distribution of χ^2. To this end we have tabulated the standard deviation and the skewness of χ^2 for $p = 0.5(0.02)1.0$ and $\varrho = 1.0(0.5)5.0$. The inverse function, i.e. p and ϱ as a function of standard deviation and skewness of the χ^2-distribution of the sample, can be approximated most easily by training a neural network of the multi-layer perceptron type, which is in fact an universal approximator [6]. We have used the JETNET package [7] to implement the inverse function on a 2-layer perceptron with 2 inputs, 2 outputs, and 20 hidden neurons. The hidden layer has logistic activation, the output layer is linear. Table 3 shows the estimated mixture model parameters for some selected values of p and ϱ, which are not in the table. Also shown is the resulting η-value of the robust estimator, i.e. its variance divided by the variance of the linear estimator. The same procedure can be repeated independently for z (not shown).

Although we have no proof that the estimate of p and ϱ is consistent or unbiased, the results indicate that the estimation procedure works well and leads to prior models which are only marginally worse than the ones obtained by using the true values. Only a small sample of about 10000 tracks is required to this purpose. In order to minimize multiple scattering, high-energy tracks should be selected.

This method of estimating the parameters of the mixture model has, however, obvious limitations: it works only in a homogeneous detector in which multiple scattering can be neglected. If the prior probability of an outlier or the width of the outlier distribution varies within the detector, there is a profusion of parameters to be determined, and the procedure breaks down.

Table 4
Relative variance of the robust estimate for exponential and t-distributed tails

Error distribution	True model		Estimated model			Best model		
	p	ϱ	p	ϱ	η	p	ϱ	η
Gaussian with exp. tails	0.2	2.0	0.04	3.33	0.719	0.16	2.2	0.660
Gaussian with exp. tails	0.2	3.0	0.04	4.31	0.469	0.24	2.6	0.323
Gaussian with t_7-tails	0.2	2.0	0.02	3.30	0.794	0.18	2.0	0.725
Gaussian with t_7-tails	0.2	3.0	0.03	4.02	0.559	0.26	2.6	0.367

5. Application to other long-tailed distributions

We have shown in the preceding sections that the non-linear filter is robust with respect to outliers generated by a Gaussian mixture model. This is not surprising as the prior density which is entered into Bayes' theorem is precisely of this form. Next we want to investigate to which extent the robustness extends to other types of long-tailed distributions. We consider two types of such distributions, a mixture of a Gaussian with a "Student's" t, the tail of which is a rational function, and a mixture of a Gaussian with a double exponential. The mixture is again described by two global parameters p and ϱ, as in the case of a mixture of two Gaussians. The resulting distribution is scaled such that the standard deviation is equal to the values quoted above (0.2 mm in $R\Phi$ and 0.5 mm in z). In all cases the Gaussian mixture model required by the filter is determined from the simulated sample by the method outlined in the preceding section.

We have simulated four samples of tracks, two with tails according to a t-distribution with seven degrees of freedom, and two with double exponential tails. The tails of the t_7-distribution decay like x^{-4}, whereas the tails of the double exponential decay like e^{-x}. Table 4 summarizes the performance of the robust filter. The central three columns contain the values obtained by plugging in the mixture model estimated from the data, whereas the last three columns show the best values obtained by a search in the (p, ϱ)-plane.

Obviously the estimates of the prior model parameters are rather poor. This is not surprising as the method has been developed under the assumption of a Gaussian mixture distribution of the observation errors. Even so, the robust filter still does better than the linear filter, shown by the values of η which are smaller than 1. If one uses the best values, the relative efficiency of the estimate is comparable to the case of Gaussian mixture errors. The robustness of the filter therefore is not confined to Gaussian tails.

6. Extension to non-Gaussian process noise

The robust filter can be modified to cope with non-Gaussian process noise. The primary source of process noise in track fitting is multiple Coulomb scattering. Note that for electrons also energy loss is serious and must be considered as a stochastic process. The distribution of the energy loss is however asymmetric and very skew, so that a Gaussian mixture is not a very suitable model. On the other hand, multiple scattering usually is considered as being Gaussian and treated as such, for lack of an alternative. The distribution of the projected deflection angle is indeed close to a Gaussian by virtue of the central limit theorem, provided that the scatterer is sufficiently thick and that all single scatters are sufficiently small to allow linear superposition. However, it seems to be a fact that rare processes like nuclear scattering (for hadrons) and hard single Coulomb scattering add some tails to the Gaussian core which we can try to take into account by the robust filter. The robustification could also be useful for the treatment of scattering in very thin layers to which the central limit theorem cannot be applied.

First, we assume a Gaussian mixture model of the process noise:

$$f(\boldsymbol{\omega}_k) = p_k{}^{(0)} \cdot \varphi(\boldsymbol{\omega}_k; \boldsymbol{0}, \boldsymbol{Q}_k{}^{(0)}) + p_k{}^{(1)} \cdot \varphi(\boldsymbol{\omega}_k; \boldsymbol{0}, \boldsymbol{Q}_k{}^{(1)}), \quad p_k{}^{(0)} + p_k{}^{(1)} = 1.$$

It is to be expected that $\boldsymbol{Q}_k{}^{(1)} > \boldsymbol{Q}_k{}^{(0)}$ and $p_k{}^{(0)} \gg p_k{}^{(1)}$. Now the distribution of $\boldsymbol{x}_k^{k-1}$ is a mixture of two Gaussians:

$$f(\boldsymbol{x}_k^{k-1}|\boldsymbol{m}_1,\ldots,\boldsymbol{m}_{k-1}) = \sum_{i=0}^{1} p_k{}^{(i)} \cdot \varphi(\boldsymbol{x}_k^{k-1}; \tilde{\boldsymbol{x}}_k^{k-1}, \boldsymbol{C}_k^{k-1(i)}),$$

with:

$$\boldsymbol{C}_k^{k-1(i)} = \boldsymbol{F}_k \boldsymbol{C}_{k-1} \boldsymbol{F}_k{}^T + \boldsymbol{Q}_k{}^{(i)}.$$

If the distribution of $\boldsymbol{m}_k$ is assumed to be Gaussian with covariance matrix $\boldsymbol{V}_k$, one obtains the following posterior p.d.f. of $\boldsymbol{x}_k$:

$$f(\boldsymbol{x}_k|\boldsymbol{m}_1,\ldots,\boldsymbol{m}_k) = \sum_{i=0}^{1} q_k{}^{(i)} \cdot \varphi(\boldsymbol{x}_k; \tilde{\boldsymbol{x}}_k{}^{(i)}, \boldsymbol{C}_k{}^{(i)}),$$

with:

$$\tilde{\boldsymbol{x}}_k{}^{(i)} = \tilde{\boldsymbol{x}}_k^{k-1} + \boldsymbol{K}_k{}^{(i)} \boldsymbol{r}_k^{k-1}, \qquad \boldsymbol{K}_k{}^{(i)} = \boldsymbol{C}_k^{k-1(i)} \boldsymbol{H}_k{}^T \boldsymbol{W}_k{}^{(i)},$$
$$\boldsymbol{W}_k{}^{(i)} = (\boldsymbol{V}_k + \boldsymbol{H}_k \boldsymbol{C}_k^{k-1(i)} \boldsymbol{H}_k{}^T)^{-1}, \qquad \boldsymbol{C}_k{}^{(i)} = (\boldsymbol{I} - \boldsymbol{K}_k{}^{(i)} \boldsymbol{H}_k) \boldsymbol{C}_k^{k-1(i)}.$$

The coefficients $q_k{}^{(i)}$ are now interpreted as the posterior probabilities of the scattering event being in the core or in the tail of the distribution. Formally they are the same as in Section 2:

$$q_k{}^{(0)} = \left[1 + \frac{p_k{}^{(1)}}{p_k{}^{(0)}} \frac{|\boldsymbol{W}_k{}^{(1)}|}{|\boldsymbol{W}_k{}^{(0)}|} \exp\left(\tfrac{1}{2} \boldsymbol{r}_k^{k-1T} \boldsymbol{D}_k \boldsymbol{r}_k^{k-1}\right)\right]^{-1}, \qquad q_k{}^{(1)} = 1 - q_k{}^{(0)},$$

with:

$$\boldsymbol{D}_k = \boldsymbol{W}_k{}^{(0)} - \boldsymbol{W}_k{}^{(1)}.$$

The filter is again a weighted sum of two Kalman filters:

$$\tilde{\boldsymbol{x}}_k = \tilde{\boldsymbol{x}}_k^{k-1} + (q_k{}^{(0)} \boldsymbol{K}_k{}^{(0)} + q_k{}^{(1)} \boldsymbol{K}_k{}^{(1)}) \boldsymbol{r}_k^{k-1},$$
$$\boldsymbol{C}_k = q_k{}^{(0)} \boldsymbol{C}_k{}^{(0)} + q_k{}^{(1)} \boldsymbol{C}_k{}^{(1)} + q_k{}^{(0)} q_k{}^{(1)} (\boldsymbol{K}_k{}^{(0)} - \boldsymbol{K}_k{}^{(1)}) \boldsymbol{r}_k^{k-1} \boldsymbol{r}_k^{k-1T} (\boldsymbol{K}_k{}^{(0)} - \boldsymbol{K}_k{}^{(1)})^T.$$

We are not aware of any Gaussian mixture model of multiple Coulomb scattering plus nuclear and hard single Coulomb scattering which is based on theoretical considerations or experimental data. The only way to proceed at the moment seems to be to make some reasonable prior assumptions on the tails and to tune these with real tracks.

7. Conclusions

We have shown that in a linear model with symmetric long-tailed measurement noise a non-linear robust filter based on a Bayesian approach has smaller variance than the optimal linear filter (Kalman filter). If the distribution of the observation error is a Gaussian mixture of a narrow core and long tails, the parameters of the mixture model required for the non-linear filter can be determined with sufficient precision from the higher moments of the χ^2-statistic of the linear filter. Only a relatively small subsample of tracks is required to this end. For other types of long-tailed error distributions this process is more difficult. If the number of parameters which have to be determined is small the problem can be solved by a search in the parameter space. Due to this limitation, the method seems to be best suited to track fitting in a homogeneous detector, for example a TPC or a silicon tracker. The robust filter is actually a linear combination of two Kalman filters, and both the prediction and the smoothing steps are the same as with the linear filter. Therefore the implementation is straightforward, and the speed is comparable to that of the Kalman filter.

We also have shown how the filter can be modified to deal with long-tailed process noise. A necessary prerequisite, however, is a Gaussian mixture model of multiple Coulomb scattering and other types of scattering processes, like nuclear scattering and hard single Coulomb scattering. Such a model, based either on theoretical considerations or on experimental measurements, would allow us to take into account the actual distribution of rare processes, not only their mean-squared properties, as is necessarily the case with a linear filter. As long as such a model is not available one has to resort to a heuristic procedure. The distribution of the Gaussian core is relatively well known, so it is probably sufficient to start with some reasonable assumptions on the tails and to tune the mixture proportions and the width of the tails with real tracks.

Acknowledgements

I want to thank M. Regler and W. Mitaroff for many useful comments on the manuscript.

References

[1] R.J. Beckman and R.D. Cook, Technometrics 25 (1983) 119.

[2] I. Guttman and D. Peña, Robust Kalman Filtering and its Applications, Technical Report No. 1, Dept. of Statistics, University of Toronto (1985).

[3] R. Frühwirth, Nucl. Instrum. Methods A 262 (1987) 444.

[4] D. Peña and I. Guttman, Optimal Collapsing of Mixture Distributions in Robust Recursive Estimation (unpublished).

[5] M.G. Kendall and A. Stuart, The Adavanced Theory of Statistics, Vol. 1, 3rd ed. (Charles Griffin, London, 1969).

[6] K. Hornik, Neural Networks 4 (1991) 251.

[7] C. Peterson, T. Rögnvaldsson and L. Lönnblad, Comput. Phys. Commun. 81 (1994) 185.