Marginalized Particle Filters for Bayesian Estimation of Gaussian Noise Parameters

Saikat Saha, Emre Özkan, Fredrik Gustafsson
Department of Electrical Engineering
Linköping University
Linköping, Sweden
{saha, emre, fredrik}@isy.liu.se

Václav Šmíd
Institute of Information Theory and Automation
Prague, Czech Republic
smidl@utia.cas.cz

Abstract – The particle filter provides a general solution to the nonlinear filtering problem with arbitrary accuracy. However, the curse of dimensionality prevents its application in cases where the state dimensionality is high. Further, estimation of stationary parameters is a known challenge in a particle filter framework. We suggest a marginalization approach for the case of unknown noise distribution parameters that avoid both aforementioned problem. First, the standard approach of augmenting the state vector with sensor offsets and scale factors is avoided, so the state dimension is not increased. Second, the mean and covariance of both process and measurement noises are represented with parametric distributions, whose statistics are updated adaptively and analytically using the concept of conjugate prior distributions. The resulting marginalized particle filter is applied to and illustrated with a standard example from literature.

Keywords: Unknown Noise Statistics, Adaptive Filtering, Marginalized Particle Filter, Bayesian Conjugate prior

1 Introduction

State space models are widely used in many engineering applications. Depending on the nature of the problem, these models could involve simple linear equations or complex nonlinearities. Estimating the unknown state based on the available measurements is an important and a well studied subject in the literature. Most of the estimation algorithms rely on the prior knowledge of the model and its parameters. In many scenarios, the model parameters, especially the noise/disturbance parameters, might not be known a priori and should be estimated on the run. This problem is referred to as noise adaptive filtering in the literature. The solution is typically given by the joint estimation of noise parameters together with the dynamic state. One very common approach is to augment the state vector with unknown parameters and redefine the problem as a filtering problem. This approach has readily been applied in the particle filtering context [16]. Such an approach has some major disadvantages as it requires artificial dynamics for the static parameters and it leads to an increase in the state dimension which is not preferable for particle filters. Many alternative approaches have also been proposed to circumvent such issues. In [3], the different approaches have been systematically classified into the following categories: Bayesian, maximum likelihood, correlation and covariance matching. Traditionally the problem has been addressed for linear systems (see e.g., [1],[8]). A correlation based adaptive Kalman filter for noise identification using the weighted least squares criterion has been proposed in [2], while an asymptotic (in time) maximum likelihood estimate has been proposed in [5]. On the other hand, the Bayesian approach has been used, for example, in [6] and [7]. In [6], the nonstationary noise statistics are estimated using the so called IMM method, while an adaptive Kalman filter based on variational Bayesian methods is used in [7]. An adaptive sequential estimation with unknown noise statistics has been proposed in [4]. Estimation of state dependent covariance matrix using the marginalized particle filter approach has been considered by [9]. Here the covariance matrix is treated as additional state, for which a state transition equation has been defined.

In this article, we propose an efficient method in a Bayesian framework for approximating the joint density of the unknown parameters and the state based on the particle filters and marginalization concepts [10],[11]. Analytical substructures in the joint distribution of the state and the model parameters are important in applying the marginalization idea. We assume suitable prior distributions for the unknown noise parameters. Conditional on the particle filter output for the state, we define analytical posterior distribution for the unknown noise parameters and propagate the hyper-parameters of the posterior recursively. Among the previous studies, [14] and [15] are the most related ones to our work. The system considered in [15] is a specific model for a
binary output and it is partially linear. The approach in [14] reflects a more general framework but only non-informative prior is used and it aims at estimating the state rather than the unknown parameters. In both studies, the noise sequences are assumed to have zero mean. The methodology we describe in this article defines a more general framework and it is applicable to noise sequences with unknown mean. In [13], estimating the unknown model and the noise parameters is aimed at by using the marginalization idea. Sampling for the unknowns is required in the update steps. In the method we propose, we integrate out the unknown parameters to compute the relevant distributions which results in a more efficient algorithm. Our experiments show that the proposed method is capable of estimating the unknown parameters of the measurement noise as well as the process noise even for highly nonlinear models.

2 Problem Definition
Consider the following nonlinear discrete time state space model relating a hidden state \(x_t\) to the observation \(y_t\):

\[
x_t = f_t(x_{t-1}) + v_t
\]
\[
y_t = h_t(x_t) + w_t
\]

Here \(t\) denotes the time index. \(f()\) and \(h()\) are possibly nonlinear functions of the state vector \(x_t\). \(v_t\) and \(w_t\) are mutually independent Gaussian noise sequences with unknown mean and covariances.

\[
v_t \overset{i.i.d.}{\sim} \mathcal{N}(\mu_v, \Sigma_v),
\]
\[
w_t \overset{i.i.d.}{\sim} \mathcal{N}(\mu_w, \Sigma_w).
\]

The means and the covariances of the noise sequences are unknown and denoted by \(\theta\).

\[
\theta \triangleq [\theta^v, \theta^w] \triangleq [\mu_v, \Sigma_v, \mu_w, \Sigma_w].
\]

The typical problem here is to infer sequentially the unobserved state \(x_t\) together with the unknown noise statistics based on a set of observation \(y_{0:t}\). This problem appears quite naturally in many practical applications of interests where the exact knowledge of the noises are unavailable. The main difficulty arises from the fact that the estimation of the hidden state also depends on the unknown noise parameters \(\theta\). We aim to address this problem by estimating sequentially the joint density of the unknown noise parameters \((\mu_v, \Sigma_v, \mu_w, \Sigma_w)\) and the state sequence \(x_{0:t}\) given the set of measurements \(y_{0:t}\).

3 Methodology
The method we present here heavily relies on the marginalization concept. We make use of the conjugate priors\(^1\) for the unknown parameters such that, it is sufficient to keep only the hyper-parameters of the posterior distribution for each particle in order to express the joint distribution of the state and the unknown noise parameters. Moreover, within the same approach, it is possible to integrate out the unknown parameters and derive the marginal density for the state easily.

3.1 Posterior distribution for the conjugate prior
For multivariate Normal data with unknown mean \(\mu\) and covariance \(\Sigma\), a Normal-inverse-Wishart distribution defines a conjugate prior. Let us denote it as \([\mu_w, \sigma_w] \sim \text{NiW}(k_0, \mu_0, \nu_0, \Lambda_0)\). Assuming Normal-inverse-Wishart distribution with parameters, \((k_0, \mu_0, \nu_0, \Lambda_0)\) defines a hierarchical Bayesian model given below:

\[
z \sim \mathcal{N}(\mu, \Sigma)
\]
\[
\mu | \Sigma \sim \mathcal{N}(\mu_0, \Sigma/k_0)
\]
\[
\Sigma \sim \text{iW}(\nu_0, \Lambda_0)
\]

where \(\text{iW}(\cdot)\) denotes Inverse Wishart distribution. The joint density of \((\mu, \Sigma)\) is of the form

\[
p(\mu, \Sigma) = \text{NiW}(k_0, \mu_0, \nu_0, \Lambda_0)
\]
\[
= \frac{1}{c} |\Sigma|^{-((\nu_0+d)/2)+1}
\]
\[
\times \exp\left(-\frac{1}{2} \text{tr}(\Lambda_0\Sigma^{-1}) - \frac{k_0}{2}(\mu - \mu_0)^T\Sigma^{-1}(\mu - \mu_0)\right),
\]

where

\[
c = \frac{2^{\nu_0d/2} \Gamma_d(\nu_0/2)(2\pi)^{d/2}}{|\Lambda_0|^{\nu_0/2}}.
\]

The parameters \(\mu_0\) and \(k_0\) define the prior mean and the number of prior measurements, while \(\nu_0\) and \(\Lambda_0\) define the degrees of freedom and the scale matrix for the inverse-Wishart distribution. Notice that the mean and the covariance are dependent. A larger covariance results in a larger variance on \(\mu\) whereas a smaller covariance will pull the mean towards \(\mu_0\). Further, the Gaussian distribution for \(z\) becomes a \(t\)-distribution for a \(\text{NiW}\) prior,

\[
p(z|\mu, \nu, \Lambda) = \text{t}_{\nu - d + 1}(\mu, \frac{(k + 1)}{k \nu - d + 1}\Lambda)
\]

3.2 Recursive updates of the conjugate prior
Suppose we observe a set of \(m\) observations, \(\{z_i\}_{i=1}^m\), from a multivariate Gaussian distribution for which

\(^1\) A family of prior distributions is conjugate to a particular likelihood function if the posterior distribution belongs to the same family as the prior.
we assumed a normal-inverse-Wishart prior for the unknown mean and variance. Via conjugacy, the posterior distribution of the unknown parameters is again a normal-inverse-Wishart distribution with the updated hyper-parameters. The hyper-parameters of the posterior distribution are updated as follows [12],

\[ \mu_{m+1} = \frac{k_0}{k_0 + m} \mu_0 + \frac{m}{k_0 + m} \tilde{z}_m \] (13a)
\[ \Lambda_{m+1} = v_0 \Lambda_0 + S_m + \frac{k_0 m}{k_0 + m} (\tilde{z}_m - \mu_0)(\tilde{z}_m - \mu_0)^T \] (13b)
\[ k_{m+1} = k_0 + m \] (13c)
\[ v_{m+1} = v_0 + m \] (13d)

where
\[ \tilde{z}_m = \frac{1}{m} \sum_{t=1}^{m} z_t, \] (13e)
\[ S_m = \sum_{t=1}^{m} (z_t - \bar{z})(z_t - \bar{z})^T. \] (13f)

3.3 Marginalization in nonlinear filtering

Let us define NiW priors for the unknown process noise and the measurement noise sequences of the system defined by (1) and (2). Let \( \Phi_0 = [\phi_0^w \phi_0^v] \) denote the initial hyper-parameters describing the prior distributions (\( \phi_0^w = [(k_0^w, \mu_0^w, v_0^w, \Lambda_0^w)] \) for the process noise and \( \phi_0^v = [(k_0^v, \mu_0^v, v_0^v, \Lambda_0^v)] \) for the measurement noise). Our aim is to approximate the joint density for \( p(x_{0:t}, \theta | y_{0:t}) \) and allow marginalization if possible. The joint distribution of the states and the unknown parameters can be decomposed into conditional distributions:

\[ p(x_{0:t}, \theta | y_{0:t}) = p(\theta | x_{0:t}, y_{0:t})p(x_{0:t} | y_{0:t}). \] (14)

Suppose we approximate the distribution \( p(x_{0:t} | y_{0:t}) \) by a set of \( N \) particles and their weights as

\[ p(x_{0:t} | y_{0:t}) \approx \sum_{i=1}^{N} \omega^{(i)} \delta_{x_{0:t}}(\cdot). \] (15)

For each particle we can compute analytical expressions for the posterior distribution of the unknown parameters. Notice that given the state trajectory \( x_{0:t} \) and the measurements \( y_{0:t} \), the measurement and process noise parameters become independent. Hence the hyper-parameter update for the posterior distributions can be done separately. The posteriors follow normal-inverse-Wishart distribution and the hyper-parameters are updated according to the equations (13a)-(13d). The following expressions are substituted with the pseudo measurement \( z_t \) in the update equations.

\[ z_t^v = x_t^{(i)} - f_t(x_t^{(i-1)}) \quad \text{for the process noise update} \] (16)
\[ z_t^w = y_t - h_t(x_t^{(i)}) \quad \text{for the measurement noise update}. \] (17)

Using the sequential importance sampling scheme for propagating the particle approximation (15) leads to the standard weight update equation:

\[ \omega^{(i)}_t = \omega^{(i)}_{t-1} \frac{p(y_t | x_t^{(i)})p(x_t^{(i)} | x_{t-1}^{(i)})}{q(x_t^{(i)} | x_{t-1}^{(i)}, y_t)}, \] (18)

where \( q(\cdot) \) is the importance distribution from which we sample \( x_t^{(i)} \).

3.4 Likelihood marginalization

In order to compute the likelihood \( p(y_t | x_t^{(i)}) \), we can utilize the posterior distribution of the unknown parameters that we computed for each particle. One important advantage of using conjugate priors reveals itself here as it is possible to integrate out unknown noise parameters as they follow normal-inverse-Wishart distribution.

\[ p(y_t | x_t) = \int p(y_t | \theta^w, x_{0:t})p(\theta^w | x_{0:t})d\theta^w. \] (19)

In accordance with the notations described in equations (13a)-(13d), the resulting predictive distribution is a multivariate Student-t distribution as follows from (12),

\[ p(z_t | z_{1:t-1}, k, \mu, v, \Lambda) = t_{v_t-\nu+1}(\mu, \frac{(k_t + 1)}{k_t(v_t - \nu + 1)} \Lambda_t) \] (20)

where \( t_{\nu}(\mu, \lambda) \) is the student-t distribution with \( \nu \) degrees of freedom, located at \( \mu \) with scale parameter \( \lambda \). The likelihood can be computed using the above expression together with (17).

3.5 State prediction

In most of the cases it is not possible to sample from the optimal importance distribution. The state transition density \( p(x_t | x_{t-1}) \) can be used as the importance distribution. Once again the unknown process noise can be integrated out.

\[ p(x_t | x_{0:t-1}) = \int p(x_t | \theta^w, x_{0:t-1})p(\theta^w | x_{0:t-1})d\theta^w. \] (21)

The resulting predictive distribution is a multivariate Student-t distribution similar to (12).
3.6 Posterior distribution for the noise parameters

The marginal posterior density of the unknown parameters can be computed by integrating out the states in the joint density.

\[ p(\theta|y_{1:t}) = \int p(\theta|x_{0:t}, y_{1:t})p(x_{0:t}|y_{1:t})dx_{0:t} \approx \sum_{i=1}^{N} \omega_{t}^{(i)} p(\theta|x_{t}^{(i)}, y_{1:t}). \] (22)

Then the estimate of the unknown parameters could be computed according to a chosen criterion. As an example, according to minimum mean square error (MMSE) criterion, the noise variance estimate at time \( t \) could be computed as

\[ \hat{\Sigma}_t = \frac{\sum_{i=1}^{N} \omega_{t}^{(i)} \Lambda_{t}^{(i)}}{v_t - d - 1}, \] (23)

where the weights are inherited from the particles.

3.7 Marginalized particle filter

In the proposed method, each particle keeps its own estimate for the parameters of the unknown process noises and measurement noise. In the importance sampling step, the particles use their own posterior distribution of the unknown parameters. The weight update of the particles is made according to the measurement likelihood. It is our expectation that the particles, keeping the unknown parameters which best explain/fit to the observed measurement sequence will survive in time. The methodology followed here is described in the next paragraph as a pseudo code.

4 Simulations

We use the following benchmark scalar nonlinear time series model for our illustrations:

\[ x_t = \frac{x_{t-1}}{2} + \frac{25}{1 + x_{t-1}^2} + 8 \cos(1.2t) + v_t, \] (24)

\[ y_t = \frac{x_t^2}{20} + w_t, \quad v_t \perp w_t, \quad t = 1, 2, \ldots \] (25)

where \( v_t \sim N(0, \Sigma_v) \) and \( w_t \sim N(0, \Sigma_w) \) and both \( \Sigma_v \) and \( \Sigma_w \) are unknown. For simulated data, we use \( \Sigma_v = 10 \) and \( \Sigma_w = 1 \).

In the first subsection, we assume that only the variances of the noises are unknown, whereas in the second subsection, we consider both the means and the variances to be unknown.

4.1 Unknown Variances

We note that when the mean of the Gaussian noise is known, the conjugate prior for the covariance matrix is given by Inverse Wishart distribution (iW), which for the scalar case reduces to Inverse Gamma distribution (iG) [12]. At time step \( t = 0 \), we set the priors as \( p(x_0) = N(0, 5) \), \( p(\Sigma_v) = iG(\alpha_0, \beta_0) \) and \( p(\Sigma_w) = iG(\lambda_0, \delta_0) \). We take \( \alpha_0 = 2.01 \), \( \beta_0 = 5 \), \( \lambda_0 = 2.01 \) and \( \delta_0 = 5 \). The hyper-parameters update equations for the iG are shown in the appendix A. In the particle filtering step, the state transition density is taken as the proposal and the resampling is done whenever the effective sample size falls below the one-third of the original sample size. The expressions for the mean and variance of the posterior are shown in appendix B. Realizations of the estimates of \( \Sigma_v \) and \( \Sigma_w \) with particle size \( N = 500 \) and \( N = 5000 \) are respectively shown in Figures 1–2. We observe that the estimation procedure works quite well. Next, keeping the particle size \( N = 5000 \), we repeat the estimates over 100 Monte Carlo runs. The Monte Carlo average of the estimates of \( \Sigma_v \) and \( \Sigma_w \) are shown in Figures 3–4. Here, the

Algorithm:

- **Initialization:**
  - For each particle \( i = 1, \ldots, N \) do
    - Sample \( x_0^{(i)} \sim p_0(x_0) \)
    - Set initial weights \( \omega_0^{(i)} = \frac{1}{N} \)
    - Set initial noise hyper-parameters \([\phi_0^v, \phi_0^w] \) corresponding to each particle

- **Iterations:**
  - For \( t = 1, 2, \ldots \) do
    - For each particle \( i = 1, \ldots, N \) do
      * Sample \( x_t^{(i)} \sim q(x_t^{(i)}|y_t, x_{t-1}^{(i)}) \)
      - For \( i = 1, \ldots, N \), update the weights
        \[ \tilde{\omega}_t^{(i)} = \frac{\omega_t^{(i)} p(y_t|x_t^{(i)}) p(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{t-1}^{(i)}, y_t)} \]
    - Update hyper-parameters of the process noise, using the pseudo measurement \( z_t^w = x_t^{(i)} - f_t(x_{t-1}^{(i)}) \) (Equations (13a)-(13d)).
    - Update hyper-parameters of the measurement noise, using the pseudo measurement \( z_t^w = y_t - g_t(x_t^{(i)}) \) (Equations (13a)-(13d)).
    - Normalize weights, \( \omega_t^{(i)} = \frac{\tilde{\omega}_t^{(i)}}{\sum_{i=1}^{N} \tilde{\omega}_t^{(i)}} \).
    - Compute \( N_{\text{eff}} = \frac{1}{\sum_{i=1}^{N} (\tilde{\omega}_t^{(i)})^2} \).
      * If \( N_{\text{eff}} \leq \eta \), Resample the particles. Copy the corresponding hyper-parameters and set \( \omega_t^{(i)} = 1/N \).
estimates appear to be slightly biased. We also compute the root mean squared error (RMSE) estimates of $\Sigma_v$ and $\Sigma_w$ at each time step $t$ (over $M = 100$ Monte Carlo runs) given by $\left( \frac{1}{M} \sum_{j=1}^{M} (\hat{z}_j^t - z_j^t)^2 \right)^{1/2}$. Here $z_j^t$ is the true parameter for time $t$ in the $j$'th run and $\hat{z}_j^t$ is the corresponding estimate. The results are shown in Figures 5–6. Next, for a typical realization with 5000 particles, we also plot the posterior densities $p(\Sigma_v | y_1:T)$ and $p(\Sigma_w | y_1:T)$ at final time $T = 1000$. This is shown in Figure 7. Subsequently, we compute the mean and the maximum a posteriori (MAP) estimates of both $\Sigma_v$ and $\Sigma_w$ at final time step $T = 1000$. Here, MAP is obtained by maximizing the argument of the respective posterior density. Each posterior is given by the weighted mixture of inverse Gamma densities. We maximize the posterior numerically using Matlab’s fminunc command, with the starting value taken to be the corresponding
4.2 Unknown Means and Variances

Here, we investigate the case where both the mean and the variance of the noise sequences are unknown. The same system defined by the equations (24)-(25) is used once more. The true parameters of the noises are set to: \(\mu_v = 3, \sigma^2_v = 4, \mu_w = 1, \sigma^2_w = 6\). NiW distribution is used as the prior and the initial hyper-parameters are set to \(\phi_w^0 = [(k_w^0, \mu_w^0, \nu_w^0, \Lambda_w^0)] = [(5, 0, 5, 10)]\) for the process noise and \(\phi_v^0 = [(k_v^0, \mu_v^0, \nu_v^0, \Lambda_v^0)] = [(5, 0, 5, 10)]\) for the measurement noise. In Figure 10, the estimates for the measurement and the process noise covariances and the means are depicted together.

5 Conclusions and Discussions

A new method for estimation of unknown noise parameters in general state space models is presented in this article. The method is defined in Bayesian framework where we define conjugate priors for the unknown noise parameters. We also make use of the marginalization idea in order to keep the algorithm implementation simple and efficient, with analytic posterior distributions for the noise parameters. The methodology described here is generic and can be extended and generalized in several ways: (i) A larger class of noises from the exponential family with suitable conjugate priors can be used. (ii) The independence assumption on the process noise and the measurement noise sequences can be relaxed. (iii) The principle of exponential forgetting
can be applied to allow for time varying noise characteristics. Such modifications are left as future work.

6 Acknowledgement

Both S. Saha and E. Özkan are supported by grants from the Linnaeus Center for Control, Autonomy, and Decision-making in Complex Systems (CADICS).

A Updating hyper-parameters of Inverse Gamma distribution

Suppose the prior $p(\Sigma) \sim \text{iG}(a_0, b_0)$. Note that iG is a special case of iW for scalar variables. If the observations $X = \{x_1, \ldots, x_n\}$ are independent Gaussian variables drawn from $N(\mu, \Sigma)$ distribution, then conjugacy implies that the posterior distribution $p(\Sigma | X)$ is also inverse Gamma distribution, i.e. $p(\Sigma | X) \sim \text{iG}(a_n, b_n)$, where the hyper parameters can be updated as

$$a_n = a_0 + \frac{n}{2} - a_{n-1} + \frac{1}{2}, \quad (26a)$$

$$b_n = b_0 + \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 = b_{n-1} + \frac{(x_n - \mu)^2}{2}. \quad (26b)$$

B Mean and variance for the posterior of $\Sigma$

The posterior of $\Sigma$ at time step $t$ is computed as

$$p(\Sigma | y_{1:t}) = \int p(\Sigma | x_{0:t}, y_{1:t}) p(x_{0:t} | y_{1:t}) dx_{0:t} \approx \sum_{i=1}^N p(\Sigma | x_{0:t}^{(i)}, y_{1:t}) w_t^{(i)}, \quad (27)$$

where $p(\Sigma | x_{0:t}^{(i)}, y_{1:t}) \sim \text{iG}(a_t^{(i)}, b_t^{(i)})$ with mean and variance as

$$E(\Sigma | x_{0:t}^{(i)}, y_{1:t}) = \frac{b_t^{(i)}}{(a_t^{(i)} - 1)} \quad (for \ a_t^{(i)} > 1) \quad (28)$$

$$\text{Var}(\Sigma | x_{0:t}^{(i)}, y_{1:t}) = \frac{(b_t^{(i)})^2}{(a_t^{(i)} - 1)^2(a_t^{(i)} - 2)} \quad (for \ a_t^{(i)} > 2). \quad (29)$$

Then the mean and variance for the posterior of $\Sigma$ are given by

$$E(\Sigma | y_{1:t}) \approx \sum_{i=1}^N E(\Sigma | x_{0:t}^{(i)}, y_{1:t}) w_t^{(i)} \quad (30)$$

$$\text{Var}(\Sigma | y_{1:t}) \approx \sum_{i=1}^N \left\{ \text{Var}(\Sigma | x_{0:t}^{(i)}, y_{1:t}) + \left\{ \text{E}(\Sigma | x_{0:t}^{(i)}, y_{1:t}) - E(\Sigma | y_{1:t}) \right\}^2 \right\} \quad (31)$$

References


