Marginalized Adaptive Particle Filtering for Non-linear Models with Unknown Time-varying Noise Parameters

Emre Özkan, Václav Šmídl, Saikat Saha, Christian Lundquist and Fredrik Gustafsson

Abstract

Knowledge of the noise distribution is typically crucial for the state estimation of general state-space models. However, properties of the noise process are often unknown in the majority of practical applications. The distribution of the noise may also be non-stationary or state dependent and that prevents the use of off-line tuning methods. For linear Gaussian models, Adaptive Kalman filters (AKF) estimate unknown parameters in the noise distributions jointly with the state. The same problem for the particle filtering is less studied. We provide a Bayesian solution for the estimation of the noise distributions in the exponential family, leading to a marginalized adaptive particle filter (MAPF) where the noise parameters are updated using finite dimensional sufficient statistics for each particle. The time evolution model for the noise parameters is defined implicitly as a Kullback-Leibler norm constraint on the time variability, leading to an exponential forgetting mechanism operating on the sufficient statistics. Many existing methods are based on the standard approach of augmenting the state with the unknown variables and attempting to solve the resulting filtering problem. The MAPF is significantly more computationally efficient than a comparable particle filter that runs on the full augmented state. Further, the MAPF can handle sensor and actuator offsets as unknown means in the noise distributions, avoiding the standard approach of augmenting the state with such offsets. We illustrate the MAPF on first a standard example, and then on a tire radius estimation problem on real data.

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

1 Introduction

Systems with unknown and potentially time-varying noise statistics are common in many applications, and a lot of effort was invested into estimation of the noise properties. Estimation of the covariance matrices for the Kalman filter was addressed in [20], where 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., [14],[17]. A correlation based adaptive Kalman filter for noise identification using the weighted least squares criterion has been proposed in [22], while an asymptotic (in time) maximum likelihood estimate has been proposed in [19]. On the other hand, the Bayesian approach has been used, for example, in [16] and [29]. In [16], the non-stationary noise statistics are estimated using the so called IMM method, while an adaptive Kalman filter based on variational Bayesian methods is used in [29]. An adaptive sequential estimation with unknown noise statistics has been proposed in [21]. Estimation of a state dependent covariance matrix using the marginalized particle filter approach has been considered by [32], where the covariance matrix is treated as an additional state, for which a state transition equation has been defined. Many of the parameter estimation methods in particle filtering rely on state augmentation technique e.g., [18],[28]. Such approaches have two main disadvantages. One is the increase in the state dimension which should be avoided in particle filters as they suffer from the curse of dimensionality. The second is the error accumulation in case of static parameters estimation as addressed in [1].

In this paper, we are concerned with a more general case of non-stationary noise characteristics belonging to the exponential family. Specifically, we focus on systems with slowly varying parameters, where the term “slowly varying” is defined as a constraint on Kullback-Leibler divergence rather than an explicit random-walk model. We show that under such a constraint, explicit parame-
The approach is closely related to the published results for the estimation of stationary noise parameters using marginalized particle filters e.g. by [6],[2] and [28],[3].

Since the transition model is in the exponential family, we assume that the prior on \( \theta_t \) is in the form conjugate to (1), i.e.

\[
p(\theta_t | e_{1:t-1}) = \frac{1}{\gamma(V_{t|t-1}, \nu_{t|t-1})} \exp(\eta(\theta_t)V_{t|t-1} - \nu_{t|t-1}\phi(\theta_t)).
\]

where \( V_{t|t-1} \) is a vector of sufficient statistics and \( \nu_{t|t-1} \) is a scalar counter of the effective number of samples in the statistics. The normalization factor \( \gamma(V_{t|t-1}, \nu_{t|t-1}) \) is uniquely determined by the statistics \( V_{t|t-1} \) and \( \nu_{t|t-1} \).

Then, the posterior density \( p(\theta_t | e_{1:t}) \) is in the form (2) with statistics

\[
\begin{align*}
V_{t|t} &= V_{t|t-1} + \tau(e_t), \\
\nu_{t|t} &= \nu_{t|t-1} + 1,
\end{align*}
\]

The result is convenient for recursive evaluation of sufficient statistics starting from a prior defined by \( V_0, \nu_0 \).

The predictive distribution of \( e_t \) is then

\[
p(e_t | e_{1:t-1}) = \int p(e_t | \theta_t)p(\theta_t | e_{1:t-1})d\theta_t = \frac{\gamma(V_{t|t}, \nu_{t|t})}{\gamma(V_{t|t-1}, \nu_{t|t-1})} p(e_t).
\]

2.2 Time Update in Exponential Family

Bayesian estimation of non-stationary parameters \( \theta_t \) requires formalization of the parameter evolution model \( p(\theta_{t+1} | \theta_t) \). The predictive density of the parameter \( \theta_{t+1} \) is obtained by marginalization

\[
p(\theta_{t+1} | e_{1:t}) = \int p(\theta_{t+1} | \theta_t)p(\theta_t | e_{1:t})d\theta_t.
\]

Since the transition model \( p(\theta_{t+1} | \theta_t) \) is unknown, we seek an estimate of the marginal \( p(\theta_{t+1} | e_{1:t}) \) among many possible transition models. To restrict the space of all possible models, we implicitly limit the change in the prediction density in time by the Kullback-Leibler distance constraint

\[
\text{KL}(p(\theta_{t+1} | e_{1:t}) || p_{\text{const}}(\theta_{t+1} | e_{1:t})) \leq \kappa,
\]
where KL is the Kullback-Leibler divergence defined as

$$\text{KL}(p_1||p_2) = \int_{-\infty}^{\infty} p_1(x) \log\left(\frac{p_1(x)}{p_2(x)}\right) dx,$$  

(7)

$0 \leq \kappa < \infty$, is a known constant and $p_{\text{const}}$ corresponds to the predictive density in case the parameters do not change in time,

$$p_{\text{const}}(\theta_{t+1}|e_{1:t}) = \int \delta(\theta_{t+1} - \theta_t)p(\theta_t|e_{1:t})d\theta_t,$$  

(8)

where $\delta()$ is the Dirac delta function. In other words, equation (8) gives the predictive density for the case of time-invariant parameters. The interpretation of (6) is that, we obtain an implicit definition of a class of transition models $p(\theta_{t+1}|\theta_t)$ giving predictive densities $p(\theta_{t+1}|e_{1:t})$ which are close to $p_{\text{const}}(\theta_t|\theta_t)$, where the closeness is measured in the Kullback-Leibler sense. A deeper discussion is provided in Section 2.3.

Following the principle of maximum entropy, we choose to approximate (5) by a distribution $\hat{p}(\theta_{t+1}|e_{1:t})$ that has the maximum entropy of all distributions satisfying (6). Since most of our applications is using continuous distributions, we will use the “continuous” generalization of entropy by Jaynes, [8], where the entropy is defined with respect to an invariant measure of entropy, $u(x)$:

$$H(p) = -\int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{u(x)}\right) dx.$$  

(9)

The straightforward generalization (known as differential entropy) is revealed for $u(x) = 1$. If the invariant measure integrates to unity, i.e. $p_u(x) = u(x)$, (9) becomes equivalent to the relative entropy (7).

**Theorem 1 (Maximum entropy under KL divergence constraint)** For a given $p_{\text{const}}(\theta_{t+1}|e_{1:t})$, the probability distribution

$$\hat{p}(\theta_{t+1}|e_{1:t}, \lambda_t) \propto p_{\text{const}}(\theta_{t+1}|e_{1:t})^{\lambda_t} u(\theta_{t+1})^{1-\lambda_t},$$  

(10)

has maximum entropy of all densities $p(\theta_{t+1})$ defined on the same support as $p_{\text{const}}(\theta_{t+1}|e_{1:t})$ which satisfies (6) for a given value of $\kappa$ and $u(\theta_{t+1})$. The forgetting factor $\lambda_t$ has two possible values: $\lambda_t = 0$ if there exists $p_u(\theta_{t+1}) \propto u(\theta_{t+1})$ and $\text{KL}(p_u(\cdot)||p_{\text{const}}(\cdot)) < \kappa$, otherwise it is a solution to the equation

$$\text{KL}(\hat{p}(\theta_{t+1}|e_{1:t}, \lambda_t)||p_{\text{const}}(\theta_{t+1}|e_{1:t})) = \kappa.$$  

(11)

**Proof:** outlined in [13] and elaborated in detail in Appendix A.1 for discrete densities.

The Theorem states that if the true parameter evolution model is in the class (6) the time update in (10) will not underestimate the uncertainty by maximizing the entropy.

Note that, for the special case of stationary parameters, $\kappa = 0$, (11) yields $\lambda = 1$. For sudden changes of the parameter, $\kappa \to \infty$, $\lambda \to 0$ and the invariant measure $p_u(\theta_{t+1})$ has the role of the prior density.

The solution (10) is particularly advantageous in the exponential family, since (10) preserves the exponential form with statistics

$$\nu_{t+1|t} = \lambda \nu_{t|t} + (1 - \lambda)\nu_u,$$  

(12a)

$$V_{t+1|t} = \lambda V_{t|t} + (1 - \lambda)V_u,$$  

(12b)

where we assume that the invariant measure is also in the exponential form (2) with statistics $\nu_u, V_u$.

**Example 2 (Normal distributed parameters)** Consider a scalar time-varying parameter $\mu_t$ with Normal distributed posterior

$$p(\mu_t|e_{1:t}) = \mathcal{N}(\hat{\mu}_{t|t}, \hat{\sigma}_{t|t}^2).$$  

(13)

The forgetting operator (10) with invariant measure $u(\mu_{t+1}) = 1$ yields

$$p(\mu_{t+1}|e_{1:t}) = \mathcal{N}(\hat{\mu}_{t|t}, \frac{1}{\lambda}\hat{\sigma}_{t|t}^2),$$  

(14)

which is again Normal with $\hat{\mu}_{t+1|t} = \hat{\mu}_{t|t}$, and $\hat{\sigma}_{t+1|t}^2 = \frac{1}{\lambda}\hat{\sigma}_{t|t}^2$. Since the KL divergence between two Normal distributions is

$$\text{KL}(p(\mu_{t+1}|e_{1:t})||p(\mu_{t|t})) = \frac{(\hat{\mu}_{t+1|t} - \hat{\mu}_{t|t})^2}{2\hat{\sigma}_{t|t}^2} + \frac{1}{2} \left(\frac{\sigma_{t+1|t}^2}{\hat{\sigma}_{t|t}^2} - 1 - \ln \frac{\sigma_{t+1|t}^2}{\hat{\sigma}_{t|t}^2}\right),$$

equation (11) has form

$$\frac{1}{2} \left(\frac{1}{\lambda} - 1 - \ln \frac{1}{\lambda}\right) = \kappa.$$  

(15)

Thus, it is independent of the statistics $\mu$ and $\sigma^2$ and can be solved numerically. For example, the solutions of (15) for $\kappa = 1$ and $\kappa = 0.01$ are $\lambda = 0.222$ and $\lambda = 0.824$, respectively. Note that (14) is also a result of marginalization (5) for the parameter evolution model

$$p(\mu_{t+1}|\mu_t) = \mathcal{N}(\mu_t, (\frac{1}{\lambda} - 1)\sigma_{t|t}^2).$$  

(16)

Hence, the exponential forgetting is equivalent to standard Bayesian filtering with transition model (16), [25].
Maximum entropy principle guarantees that if the true parameter evolution model is in the class (6) the estimation procedure will not under estimate the uncertainty.

An open research question is how to determine $\kappa_t$ or, alternatively, the forgetting factor $\lambda_t$ since the relation between these two is rather tight as demonstrated in Examples 2 and 3. Research results on the choice of forgetting factor for many particular cases are available, e.g. [23,31]. However, in many practical applications, the forgetting factor is chosen to be constant and manually tuned. We follow this approach in this paper and show that this approach yields good results both in simulations and real data. The results for different choices on the constant forgetting factor are illustrated in section 5.2. Bayesian treatment of $\lambda_t$ is also possible but outside the scope of this paper.

2.4 Invariant Measure

In this paper, we use the invariant measure mainly as a technical mean to derive the main results. In practical applications, we choose $u(\cdot)$ as close to the uniform measure as possible, as demonstrated in Example 2. However, it may be used as a regularization term in recursive Bayesian estimation. Its benefits and dangers are discussed in Appendix A.2.

3 Joint Estimation of State and Noise Parameters

Consider the following nonlinear discrete time state space model relating a hidden state $x_t$ to the observation $y_t$

$$x_t = f_t(x_{t-1}, u_{t-1}) + g_t(x_{t-1}, u_{t-1}) v_t,$$  \hspace{1cm} (19a)

$$y_t = h_t(x_t, u_t) + d_t(x_t, u_t) w_t.$$  \hspace{1cm} (19b)

Here, $t$ denotes the time index. $f(\cdot)$, $h(\cdot)$, $d(\cdot)$ and $g(\cdot)$ are possibly nonlinear functions of the state vector $x$ and the input $u$. In order to avoid the degenerated case of perfect noise-free observations, we will assume that $d(\cdot)$ is invertible. On the other hand, $g(\cdot)$ is not assumed invertible, since most motion models in practice, including those with integrators, lead to noninvertible $g(\cdot)$. We define the noise vector $e_t \triangleq [v_t^T, w_t^T]^T$ as a realization from a distribution which belongs to the exponential family (1) with unknown time-varying parameter $\theta_t$.

We are concerned with the evaluation of the joint density $p(x_t, \theta_t|y_{0:t})$. Following the concept of marginalized particle filtering, we decompose the joint posterior density into conditional densities as follows:

$$p(x_{0:t}, \theta_t|y_{0:t}) = p(\theta_t|x_{0:t}, y_{0:t}) p(x_{0:t}|y_{0:t}),$$  \hspace{1cm} (20)
where we choose to approximate \( p(x_{0:t}|y_{0:t}) \) by an empirical density

\[
p(x_{0:t}|y_{0:t}) \approx \sum_{i=1}^{n} \omega^{(i)}_t \delta(x_{0:t} - x^{(i)}_0),
\]

with sample trajectories \( x^{(i)}_0 \) and weights \( \omega^{(i)}_t \). Such a decomposition will result in a particle approximation of the state density and analytical expressions for the conditional density of the parameters \( p(\theta_t|x_{0:t}, y_{0:t}) \).

Two key ideas will help us in deriving the recursive equations. First, for a given value of \( (x_{0:t}, y_{0:t}) \), the conditional density \( p(\theta_t|x_{0:t}, y_{0:t}) \) can be considered as the posterior density of the parameters and can be computed by a measurement update of the noise distribution parameters. Second, since we have the analytical expression \( p(\theta_t|x_{0:t-1}, y_{0:t-1}) \) from the previous time instant, the unknown parameter \( \theta_t \) can be integrated out when computing the recursive expressions for the marginal density of the state \( p(x_{0:t}|y_{0:t}) \). The latter will be explained together with the weight update equation later.

Under the approximation (21), the first part of (20) needs to be evaluated only at points \( x^{(i)}_0 \). Note that, for a known value of \( x^{(i)}_t \), (19a)–(19b) can be transformed into

\[
\epsilon^{(i)}_t = e(x^{(i)}_t, y_t) = \left[ g^{(i)}_t(x^{(i)}_{t-1}, u_{t-1})|x^{(i)}_t - f_t(x^{(i)}_{t-1}, u_{t-1})| d^{(i)}_t(x^{(i)}_t, u_t)|y_t - h_t(x^{(i)}_t, u_t) \right].
\]

(22)

where \( g^{(i)}_t(x^{(i)}_{t-1}, u_{t-1}) \) stands for the pseudo-inverse of \( g_t(x^{(i)}_{t-1}, u_{t-1}) \). Then, \( p(\theta_t|x_{0:t}, y_{0:t}) = p(\theta_t|x^{(i)}_t) \) and the results from Section 2 can be readily applied.

The joint density (20) is

\[
p(x_{0:t}, \theta_t|y_{0:t}) \approx \sum_{i=1}^{n} \omega^{(i)}_t p(\theta_t|V^{(i)}_t, \nu^{(i)}_t) \delta(x_{0:t} - x^{(i)}_0),
\]

(23)

where the statistics \( \omega^{(i)}_t, V^{(i)}_t, \nu^{(i)}_t, x^{(i)}_0 \) are evaluated as follows.

First, \( x^{(i)}_t \) are sampled from a proposal density \( q(x_t|x^{(i)}_{0:t-1}, y_{0:t-1}) \). Second, for the known value \( x^{(i)}_0 \), the conditional density \( p(\theta_t|x^{(i)}_0, y_{0:t}) \) is updated using the mapping (22) to \( \epsilon^{(i)}_t \) and the statistics \( V^{(i)}_{t|t}, \nu^{(i)}_{t|t} \) are updated using (3). Finally, the update equation for the weights \( \omega^{(i)}_t \) can be derived using the marginal density \( p(x_{0:t}|y_{0:t}) \) from (20). Since,

\[
p(x_{0:t}|y_{0:t}) \propto p(y_t, x_t|x_{0:t-1}, y_{0:t-1})p(x_{0:t-1}|y_{0:t-1}),
\]

(24)

substituting (21) into (24) in place of \( p(x_{0:t}|y_{0:t}) \) and \( p(x_{0:t-1}|y_{0:t-1}) \) yields

\[
\omega^{(i)}_t \propto p(y_t, x_t|x^{(i)}_{0:t-1}, y_{0:t-1}) w^{(i)}_{t-1},
\]

where

\[
p(y_t, x_t|x_{0:t-1}, y_{0:t-1}) = \int p(y_t, x_t|\theta_t, x_{0:t-1}, y_{0:t-1})p(\theta_t|x_{0:t-1}, y_{0:t-1})d\theta_t,
\]

(25)

is the marginal predictive distribution of \( x_t, y_t \). This marginal distribution is computed by integrating out the unknown parameters which leads to the predictive distribution of \( x_t, y_t \), and consequently \( \epsilon_t \) via (22). Notice that the predictive distribution \( p(\epsilon_t|x_{0:t-1}) \) is readily available for the exponential family in the form of (4). The predictor (25) can be obtained using the lemma on transformation of variables in probability density functions:

\[
p(y_t, x_t|x_{0:t-1}, y_{0:t-1}) = |J(x_t, y_t)|p(e(x_t, y_t)|V^{(i)}_{t|t-1}, \nu^{(i)}_{t|t-1}).
\]

(26)

where \( J(x_t, y_t) \) is the Jacobian of the transformation (22) and \( p(e_t|\cdot) \) is given by (4).

The final algorithm is summarized in Algorithm 1.

Remark 4 (Stationary parameters) Note that estimation of stationary parameters can be obtained as a special case of the above approach for \( \kappa = 0 \) in (6). Then, the only solution of (11) is \( \lambda = 1 \) reducing update of sufficient statistics (3) to the form of [28]. As pointed out e.g. by [4] the stationary case suffers from the path degeneracy problem. Here, we note that for a sequence of \( \lambda_t < 1, \forall t \), the posterior density \( p(\theta_t|x_t, y_t, 1) \) and thus \( p(y_t, x_t|x_{0:t-1}, y_{0:t-1}) \) satisfies the exponential forgetting property [5]. Therefore, the path degeneracy problem is less severe in this case.

4 Special case of Gaussian Noise

In this section, we specialize Algorithm 1 to the practically important case of normal distributed noises.

4.1 Likelihood and Conjugate Prior

For multivariate normal distribution of \( e_t \) with unknown mean \( \mu_t \) and covariance \( \Sigma_t \), a normal-inverse-Wishart
Algorithm 1 Marginalized adaptive particle filter for non-linear model with time-varying noise parameters.

Initialization:
For each particle i = 1,..., N do
  - Sample \( x_0^{(i)} \sim p_0(x_0) \),
  - Set initial weights \( \omega_0^{(i)} = \frac{1}{N} \),
  - Set initial noise statistics \( [v_0, V_0] \) for each particle,

Iterations:
For \( t = 1, 2, \ldots \) do
  - For each particle i = 1,..., N do
    - perform the time update of the statistics \( V_{t-1}^{(i)}, \hat{x}_{t-1}^{(i)} \) using (12),
    - sample \( x_t^{(i)} \sim q(x_t^{(i)}|x_{t-1}^{(i)}, y_{0:t}) \),
    - set \( e_t^{(i)} = c(x_t^{(i)}, y_t) \),
    - compute the predictive likelihood \( p(y_t, x_t^{(i)}|x_{t-1}^{(i)}, y_{0:t-1}) \) using (26),
    - update the weights:
      \[
      \tilde{\omega}_t^{(i)} = \omega_{t-1}^{(i)} \frac{p(y_t, x_t^{(i)}|x_{t-1}^{(i)}, y_{0:t-1})}{q(x_t^{(i)}|x_{t-1}^{(i)}, y_{0:t-1})}
      \]
    - perform the measurement update of the statistics \( V_{t}^{(i)}, \hat{x}_{t}^{(i)} \), using (3).
  - Normalize the weights, \( \omega_t^{(i)} = \frac{\tilde{\omega}_t^{(i)}}{\sum_{i=1}^{N} \tilde{\omega}_t^{(i)}} \).
  - Compute \( N_{\text{eff}} = \sum_{i=1}^{N} \omega_t^{(i)^2} \).
  - If \( N_{\text{eff}} \leq \eta \), resample the particles. Copy the corresponding statistics and set \( \omega_t^{(i)} = 1/N \).

The quantities \( \mu_t, \Sigma_t \) are conditionally independent, \( \mu_t | \Sigma_t \sim N(\mu_{t-1}, \Sigma_{t-1}) \)

\[
\begin{align*}
\mu_t | \Sigma_t \sim & N(\mu_{t-1}, \Sigma_{t-1}), \\
\Sigma_t \sim & \text{iW}(\nu_t, \Lambda_t), \\
\Sigma_t \sim & \text{iW}(\nu_t, \Lambda_t) \\
\propto & |\Sigma_t|^{-\frac{1}{2}}(|\nu_t| + d + 1) \exp(-\frac{1}{2} \text{tr}((\Lambda_t) \Sigma_t^{-1})),
\end{align*}
\]

where the statistics of the predictive distribution are

\[
\begin{align*}
\gamma_t | t-1 & = \frac{1}{1 + \gamma_{t-1}}, \\
\hat{\mu}_t | t-1 & = \hat{\mu}_{t-1} + \gamma_t | t-1 | e_t - \hat{\mu}_{t-1}, \\
\nu_t | t-1 & = \nu_{t-1} + 1, \\
\Lambda_t | t-1 & = \Lambda_{t-1} + \frac{1}{1 + \gamma_{t-1}}(\hat{\mu}_{t-1} - e_t)(\hat{\mu}_{t-1} - e_t)', 
\end{align*}
\]

These equations are derived in [24] and their relation to exponential family is discussed in [12]. The predictive distribution of \( e_t \) (4) becomes a multivariate Student-t density with \( \nu_{t-1} - d + 1 \) degrees of freedom

\[
p(e_t | \nu_{t-1}, V_{t-1}) = \text{St} \left( \hat{\mu}_{t-1}, \Lambda_{t-1}, \nu_{t-1} - d + 1 \right)
\]

\[
\propto 1 + (e_t - \hat{\mu}_{t-1}) \Lambda_{t-1}^{-1} (e_t - \hat{\mu}_{t-1})^{-\frac{1}{2}} |
u_{t-1} - d + 1|.
\]

The first two moments of (30) are

\[
\mathbb{E}(e_t) = \mu_{t-1}, \quad \text{Var}(e_t) = \frac{1}{\nu_{t-1} - d + 1} \Lambda_{t-1}.
\]

A special case of the MAPF algorithm for the model (19a)–(19b) with independent noise \( v_t \) and \( w_t \), and without transformations \( d_k \) and \( g_k \) is described in Algorithm 2. Due to the noise independence, their posterior distributions are conditionally independent, with statistics \( S_{v,t} = \{ \hat{\mu}_v, \gamma_v, \Lambda_v, \nu_v \} \) and \( S_{w,t} = \{ \hat{\mu}_w, \gamma_w, \Lambda_w, \nu_w \} \). The predictive distribution (25) then simplifies to a product of multivariate Student-t predictors

\[
p(x_t | S_{v,t-1}) = \text{St} \left( f(x_{t-1}), \hat{\mu}_v, \gamma_v, \Lambda_v, \nu_v, d_v + 1 \right)
\]

\[
p(y_t | S_{w,t-1}) = \text{St} \left( h(x_t), \hat{\mu}_w, \gamma_w, \Lambda_w, \nu_w, d_w + 1 \right)
\]

where we have used (26) with unit Jacobian. The proposal distribution \( q(x_t | x_{t-1}, y_{1:t}) \) is chosen as the predictor (31) which simplifies evaluation of the weights \( \omega_t^{(i)} \), see Algorithm 2.
Algorithm 2 Marginalized adaptive particle filter for non-linear model with Gaussian noise with time-varying parameters.

Initialization:
For each particle \( i = 1, \ldots, N \) do
- Sample \( x_0^{(i)} \) from (31).
- Set initial weights \( \omega_0^{(i)} = \frac{1}{N} \).
- Set initial noise statistics \( S_v,0, S_w,0 \) for each particle.

Iterations:
For \( t = 1, 2, \ldots \) do
- For each particle \( i = 1, \ldots, N \) do
  - perform the time update of the statistics \( S_{v,t}, S_{w,t} \), using (29),
  - sample \( x_t^{(i)} \) from (31),
  - update the weights \( \omega_t^{(i)} \)
    \[ \omega_t^{(i)} = p(y_t|S_{w,t})\omega_{t-1}^{(i)}, \]
  - perform the measurement update of the statistics \( S_{v,t}, S_{w,t} \), using (28).
- Normalize the weights, \( \omega_t^{(i)} = \frac{\omega_t^{(i)}}{\sum_{i=1}^{N} \omega_t^{(i)}} \).
- If \( N_{eff} \leq \eta \), resample the particles. Copy the corresponding statistics and set \( \omega_t^{(i)} = 1/N \).

5 Experimental Results

5.1 Illustrative example

In this section we illustrate the performance of the proposed marginalized particle filter algorithm and compare it with the state augmentation approach. We use the following benchmark scalar nonlinear time series model for the illustrations:

\[
x_t = \frac{x_{t-1}}{2} + \frac{25x_{t-1}}{1 + x_{t-1}^2} + 8\cos(1.2t) + v_t, \quad (33)
\]

\[
y_t = \frac{x_t^2}{20} + w_t, \quad v_t \perp w_t, \quad t = 1, 2, \ldots \quad (34)
\]

where \( v_t \sim N(\mu_v, \Sigma_v) \) and \( w_t \sim N(\mu_w, \Sigma_w) \). Both the mean and the variance of the measurement and process noise sequences are unknown and time varying. The true parameters of the noises are initially set to an arbitrary choice of values: \( \mu_v,0 = 1, \Sigma_v,0 = 2, \mu_w,0 = 3, \Sigma_w,0 = 4 \) and the final values are set to \( \mu_v,4000 = 2, \Sigma_v,4000 = 4, \mu_w,4000 = 1, \Sigma_w,4000 = 7 \); see Figure 2. In the following, we first give a brief description of the state augmentation method and later describe the MAPF method.

- Augmented State PF: In this approach, a new state vector \( \tilde{x}_t \) is defined by augmenting the model state with the unknown parameters. Artificial dynamics are used to account for the change of the parameters in time. The augmented state vector is defined as follows

\[
\tilde{x}_t \triangleq [x_t, \mu_v, t, \mu_w, t, \Sigma_v, t, \Sigma_w, t]^T \quad (35)
\]

In our simulations, the unknown means are propagated by a Gaussian random walk.

\[
p(\mu_v|\mu_v,t-1) = N(\mu_v,t-1, \sigma^2_{\mu_v}) \quad (36a)
\]

\[
p(\mu_w|\mu_w,t-1) = N(\mu_w,t-1, \sigma^2_{\mu_w}) \quad (36b)
\]

where the standard deviation of the random walk is set to 5 percent of the average value of the true parameters. The following Markovian model with Inverse-Gamma distribution is used to propagate the unknown covariances.

\[
p(\Sigma_v|\Sigma_v,t-1) = i\Gamma(\alpha_{\Sigma_v}, \beta_{\Sigma_v}) \quad (37a)
\]

\[
p(\Sigma_w|\Sigma_w,t-1) = i\Gamma(\alpha_{\Sigma_w}, \beta_{\Sigma_w}) \quad (37b)
\]

The parameters \( \alpha \) and \( \beta \) are chosen such that the mean value is preserved and the standard deviation is equal to 5 percent of the previous value of the parameter.

\[
E(\Sigma_v, t|\Sigma_v, t-1) = \Sigma_v, t-1 \quad (38a)
\]

\[
E(\Sigma_w, t|\Sigma_w, t-1) = \Sigma_w, t-1 \quad (38b)
\]

\[
\text{Std}(\Sigma_v, t|\Sigma_v, t-1) = 0.05 \Sigma_v, t-1 \quad (38c)
\]

\[
\text{Std}(\Sigma_w, t|\Sigma_w, t-1) = 0.05 \Sigma_w, t-1 \quad (38d)
\]

- MAPF: For the marginalized adaptive particle filter, a NIW distribution is used as the prior. The initial parameters \((\gamma_0, \mu_0, \nu_0, \Lambda_0)\) are set to \( \phi^0 = [0.2, 1.5, 27] \) and \( \phi^0 = [0.2, 3, 5, 9] \) for the measurement and process noises respectively so that the initial conditions match with the augmented state PF. The exponential forgetting factor \( \lambda \) is chosen as 0.98 by considering the average RMS error over 100 MC runs for different values of \( \lambda \).

In order to make a fair comparison, we set the initial values of the unknown parameters the same for both methods. Both algorithms start from the initial values of parameters being equal to \( \mu_v = 3, \Sigma_v = 3, \mu_w = 1, \Sigma_w = 9 \); see Figures 2 and 3. Moreover, in order to avoid mistuning of the Augmented PF algorithm, we have made multiple tests on the step size of the random walk and have chosen the value which produced the minimum average RMS error on the state estimates over 100 MC runs. Among the values 1 to 10 percent, 5 percent provided the best tuning. The performance is not over sensitive to the step size unless it is chosen as the extreme values. Hence, a finer grid was not needed.

In 100 MC runs, the effects of increasing the number of particles is also examined. In Figures 2 and 3, the estimation performances of the two methods are shown...
for the case where both algorithms use 500 particles. The standard deviation of the estimates based on different MC runs are also depicted on top of the estimates in the same figures. The MAPF method produces estimates with smaller covariance in comparison with the Augmented PF approach. Another comparison is made in order to illustrate the effects of changing the number of particles on both algorithms. The MC runs are repeated for 50, 100, 200, 500 and 1000 particles and the average RMS errors of the state estimate are compared.

In Figure 4, the average RMS state estimation errors are plotted with respect to different number of particles for both methods. The same curve for Oracle particle filter (the particle filter which uses the true values of the parameters) is also plotted. The performance gain by marginalization can be observed more explicitly in this plot. As an example, in order to achieve the performance of the MAPF method which uses 100 particles, one needs to use 500 particles in the state augmentation method. On the other hand, for a fixed number of particles, one can get lower RMS error with MAPF method especially when the number of particles is kept low. Similar results are obtained for the estimated parameters. In Figure 5, the average RMS error of the measurement noise variance estimate is shown as an example. The average runtime of a single MC run of the two methods on a PC are given in Table 1. As can be seen from the table, the computation time of the MAPF is only slightly higher than that of the augmented PF and the algorithms are of the same computational complexity. Hence a lower RMS error can be achieved for a fixed amount of available computational power using MAPF.

Table 1
Average runtime of the algorithms in seconds:

<table>
<thead>
<tr>
<th># of Particles</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPF</td>
<td>0.95</td>
<td>1.24</td>
<td>1.89</td>
<td>4.71</td>
<td>12.96</td>
</tr>
<tr>
<td>Augmented</td>
<td>0.87</td>
<td>1.04</td>
<td>1.67</td>
<td>4.40</td>
<td>12.88</td>
</tr>
</tbody>
</table>

5.2 Forgetting Factor

In this section we illustrate the effects of changing the forgetting factor. For this purpose we present a single run of the algorithm for different values of λ. In Figures 6 and 7 the estimation results are shown for λ = 0.98 and λ = 0.995 respectively. 500 particles are used in the algorithm. As can be seen from the figures, the variance of the estimates are larger for smaller λ and smoother estimates are obtained for larger λ. On the other hand the smaller forgetting factor can track faster changes in the parameters whereas a larger value of the forgetting factor will produce a slower response.

6 Application to Odometry

In this section we test the proposed algorithm on real data. An odometry application is investigated. Odometry is the term used for dead reckoning the rotational speeds of two wheels on the same axle of a wheeled vehicle. It is used in a large range of robotics applications, as well as in some vehicle navigation systems. As in all dead-reckoning, sensor offsets generate a drift over time that can be quite substantial. For odometry, the main reason for the drift is due to unknown wheel radii. Therefore, all odometric applications use some kind of absolute reference sensor to correct the drift. For open air conditions, the global positioning system (GPS) is the perfect complement. For indoor applications, markers or beacons are usually placed in the environment. The raw signals are the angular velocities of the wheels which can be measured by the ABS sensors in cars or wheel encoders in ground robots.
The angular velocities can be converted to virtual measurements of the absolute longitudinal velocity and yaw rate assuming a front wheel driven vehicle with slip-free motion of the rear wheels, as described in Chapter 13 and 14 of [7],

\begin{align}
\dot{\vartheta}_{\text{virt}} &= \omega_3 r + \omega_4 r \\
\dot{\psi}_{\text{virt}} &= \frac{\omega_3 r - \omega_4 r}{B},
\end{align}

(39a) 

(39b)

where $\omega_3$ and $\omega_4$ are the angular velocities of the rear left and the rear right wheels respectively and $r$ is the nominal value of the wheel radii; see Figure 8 for the notation. The actual wheel radii are unknown and may differ from their nominal value in practice,

\[ r_3 = r + \delta_3 \]

(40a)
where are the wheel radii of the rear left and the rear right wheels respectively. The actual velocity and yaw rate of the vehicle differ from the virtual measurements, according to
\[
\begin{align*}
\vartheta &= \frac{\omega_3 r_3 + \omega_4 r_4}{2} \quad (41a) \\
\dot{\psi} &= \frac{\omega_3 r_3 - \omega_4 r_4}{B}. \quad (41b)
\end{align*}
\]

We model the error in the wheel radii with a noise term which is subject to change in time,
\[
\begin{pmatrix} r_3(t) \\ r_4(t) \end{pmatrix} = \begin{pmatrix} r \\ r \end{pmatrix} + w_r(t), \quad (42)
\]

where is assumed to be Gaussian
\[
w_r(t) \sim \mathcal{N} \left( \begin{pmatrix} \delta_3 \\ \delta_4 \end{pmatrix}, \begin{pmatrix} \Sigma_3 & 0 \\ 0 & \Sigma_4 \end{pmatrix} \right). \quad (43)
\]

Substituting (42) in equations (41a) and (41b) results in
\[
\begin{pmatrix} \dot{\vartheta} \\ \dot{\psi} \end{pmatrix} = \begin{pmatrix} \dot{\vartheta}_{\text{virt}} \\ \dot{\psi}_{\text{virt}} \end{pmatrix} + \begin{pmatrix} \frac{\omega_3}{T} & \frac{\omega_3}{B} \\ \frac{\omega_4}{B} & \frac{\omega_4}{B} \end{pmatrix} w_r. \quad (44)
\]

The odometric dead reckoning can be formulated using the following discrete time model by defining the state vector as the planar position and the heading angle:
\[
x_t = \begin{pmatrix} X_t \\ Y_t \\ \psi_t \end{pmatrix}, \quad x_{t+1} = x_t + \begin{pmatrix} T \vartheta_t \cos(\psi_t(t)) \\ T \vartheta_t \sin(\psi_t(t)) \\ T \dot{\psi}_t \end{pmatrix}. \quad (45)
\]

Plugging in the observed speed and yaw rate gives the following dynamic model with the process noise
\[
\begin{align*}
X_{t+1} &= X_t + \left( \dot{\vartheta}_{\text{virt}}(t) + \left[ \frac{\omega_3(t)}{2} \frac{\omega_4(t)}{2} \right] w_r(t) \right) T \cos(\psi_t), \\
Y_{t+1} &= Y_t + \left( \dot{\psi}_{\text{virt}}(t) + \left[ \frac{\omega_3(t)}{B} \frac{\omega_4(t)}{B} \right] w_r(t) \right) T \sin(\psi_t), \\
\psi_{t+1} &= \psi_t + \left( \dot{\psi}_{\text{virt}}(t) + \left[ \frac{\omega_3(t)}{B} \frac{\omega_4(t)}{B} \right] w_r(t) \right) T. \quad (46a-c)
\end{align*}
\]

Note that the virtual measurements in (39a) and (39b) of speed and yaw rate are computed from the rotational speeds. Here, the rotational speeds are considered as inputs rather than measurements. This is in accordance to all navigation systems where inertial measurements are dead-reckoned in similar ways. This formulation is in accordance with the general state space model given in equations (19a) and (19b) where the GPS measurements are used as the reference measurements
\[
\begin{pmatrix} x_{t,\text{GPS}} \\ y_{t,\text{GPS}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} x_t + v_t. \quad (47)
\]

### 6.2 Experiments

In the experiments, two sets of data are collected with a passenger car in the urban area of Linköping. The car is equipped with standard vehicle sensors, such as wheel speed sensors, and a GPS receiver. We estimate the tire radii as well as the trajectory via the GPS and the virtual velocity and yaw rate measurements online. The trajectory followed by the car is plotted in Figure 9. Two runs are completed with different tire pressure settings for the rear wheels. In the first setup, the tire pressure of the rear left (RL) wheel is reduced to 1.5 bar where as the tire pressure of the rear right (RR) wheel is kept at its nominal value of 2.8 bar. In the second setup, the tire pressure of the rear left wheel is kept at 2.8 bar and the tire pressure of the rear right wheel is reduced to 1.4 bar. The estimated tire radii in both experiments are plotted in Figures 10 and 11. The true tire radii difference is calculated by computing the effective tire radii using the data collected during a long and straight segment of the road. The true tire radius differences are approximately 1.5 mm and 1.9 mm in the two experiments in the Figures 10 and 11, respectively. As can be observed from the figures, the mean value of the tire radii in the upper plots can be estimated within sub-millimeter accuracy by the algorithm. Note that the covariances of the tire radii bias are larger for the tires with reduced pressure than the ones with nominal pressure. This can be explained by the increased vibration amplitude of a soft tire. The estimated trajectory in one run is also plotted in Figure 9. The estimated trajectory matches the GPS and road map successfully in both runs.

### 7 Conclusions

A new Bayesian solution of the noise adaptive filtering problem is presented in this article. The algorithm is based on particle filtering, and it can be applied to a large class of nonlinear state space models. The algorithm makes use of marginalization and conjugate priors, such that analytic posterior distributions of the noise parameters is obtained, which makes the implementation simple and efficient. We employ the maximum entropy approach in computing the posterior distribution of the noise parameters where the parameters are assumed to be slowly varying but the evolution of the parameters is unknown. The solution utilizes the exponential forgetting factor which prevents the accumulation of error in
the sufficient statistics of the noise. Performance of the algorithm is tested on a highly non-linear benchmark models and in an odometry application using real data.

8 Acknowledgement

The authors gratefully acknowledge fundings from the Swedish Research Council VR in the Linnaeus Center CADICS. Both E. Özkan and S. Saha are supported by grants from the Linnaeus Center CADICS. V. Smidl is supported by grant GA CR 102/08/P250. The authors would also like to specifically thank Kristoffer Lundahl, at the vehicular systems division of Linköpings University, for contributing with the application example.

A Appendix

A.1 Proof of Theorem 1

The proof is described for discrete densities for simplicity. Proof of the continuous version is technically more complex but completely analogous using the infinite dimensional setting of Karush-Kuhn-Tucker conditions [30].

Consider a distribution \( p_{\text{const}} \equiv p_{\text{const}}(\theta_{t+1}|e_{1:t}) \) and a measure \( u \equiv u(\theta_{t+1}|e_{1:t}) \) to be defined on a discrete set of parameters \( \theta_{t+1} \in \{\theta_1, \ldots, \theta_m\} \). Maximization of entropy of a distribution \( p^* \equiv p^*(\theta_{t+1}|e_{1:t}) \) is then an m-dimensional optimization problem in \( p^*_t, i = 1, \ldots, m, \)

\[
p^*_t = \arg\max(-\sum p^*_t \log \frac{p^*_t}{u_t}),
\]

\[
KL(p^*||p_{\text{const}}) \leq \kappa,
\]

\[
\sum_{i=1}^{m} p^*_t = 1.
\]

Using the definition of KL divergence, the Lagrangian of the optimization problem is:

\[
\sum_i p^*_t \ln \frac{p^*_t}{u_t} + \mu(\sum_i p^*_t \ln \frac{p^*_t}{p_{\text{const},t}} - \kappa) + \lambda(\sum_i p^*_t - 1) = 0,
\]

yielding a set of Karush-Kuhn-Tucker conditions:

\[
(\ln p^*_t - \ln u_t) + 1 + \mu(\ln p^*_t - \ln p_{\text{const},t} + 1) + \lambda = 0,
\]

\[
\sum_i p^*_i (\ln p^*_i - \ln p_{\text{const},i}) \leq \kappa,
\]

\[
\mu(\sum_i p^*_i (\ln p^*_i - \ln p_{\text{const},i} - \kappa) = 0.
\]
\[ \sum p_1^* = 1, \quad \mu \geq 0. \quad \text{(A.4)} \]

From (A.1) it follows that
\[ p_1^* \propto u_i \frac{1}{p_{\text{const},i}}. \quad \text{(A.5)} \]

The conditions (A.3) are satisfied if: (i) \( \mu = 0, p_1^* = p_u \propto u_i \) and \( KL(p_u||p_{\text{const}}) \leq \kappa \), or (ii) \( KL(p_u||p_{\text{const}}) > \kappa \), \( \mu > 0 \), in which case \( p^* \) is at the boundary
\[ KL(p^*||p_{\text{const}}) = \kappa. \quad \text{(A.6)} \]

An analytical solution for (A.6) is not available, however, it is a smooth function in \( \mu \), for \( \mu \to \infty \), \( KL(p^*||p_{\text{const}}) \to 0 \) and for \( \mu \to 0 \), \( KL(p^*||p_{\text{const}}) > \kappa \). Hence, there exists a value \( \mu^* \) such that (A.6) holds. The equality (10) corresponds to (A.5) under substitution \( \lambda_t \mu = \mu/(1 + \mu) \). Since entropy is a convex function and the Slater regularity condition is trivially satisfied for \( p^* = p_{\text{const}} \), (10) is the global maximum of the entropy.

### A.2 Invariant measure

Over the classical formulation of forgetting in [10], the entropy formulation has an additional degree of freedom in the choice of the invariant measure \( u(\cdot) \). This element is equivalent to the alternative distribution of decision theoretic approach [15], which compares several of its possible choices. In this text, we focus on the original formulation of [8], in which the main purpose of the invariant measure is to preserve invariance of the entropy under the change of coordinates. However, it should be as uninformative as possible. Hence, its choice is governed by the same rules that apply to uninformative prior distributions [9,11]. This was the case in Examples 2 and 3, where the Jeffrey's invariant measures for location and scale parameters were used, respectively. In cases where prior information is available, as a prior distribution \( p_0(\theta_{t+1}) \propto u(\theta_{t+1}) \), it can be used as the invariant measure. Note that the influence of this choice on the posterior can be significant. To see that, consider a stationary \( \lambda_t = \lambda \) and a constant \( u(\theta_{t+1}) = u(\theta_t) = \ldots = p_u(|V_u, \nu_u) \). Recursive substitution of (5) into Bayes rule yields:

\[ \hat{p}(\theta | e_{1,t}, \lambda_t) \propto p(\theta_t | e_{1,t-1}, \lambda_t) p(e_t | \theta_t) \]
\[ \propto u(\theta_t) \prod_{\tau=1}^t p(e_{\tau | \theta_{\tau}}) \lambda^{t-\tau}. \]

Hence, \( u(\theta_t) \) can be interpreted as a prior for estimation of a stationary parameter \( \theta_t \) on an exponential window of the measurements with the effective number of records \( 1/(1 - \lambda) \). The posterior may become prior dominated especially for small values of \( \lambda \).

### References


