MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Process-Noise Adaptive Particle Filtering with Dependent Process and Measurement Noise

Berntorp, K.; Di Cairano, S.

TR2016-148 December 12, 2016

Abstract

Knowledge of the noise distributions is typically key for reliable state estimation. However, in many applications only the measurement noise can be determined a priori, since only this correspond to measurable quantities. Moreover, modeling of physical systems often leads to nonlinear state-space models with dependent noise sources. Here, we design a computationally efficient marginalized particle filter for jointly estimating the state trajectory and the parameters of the process noise, assuming dependent noise sources. Our approach relies on marginalization and subsequent update of the sufficient statistics of the process-noise parameters. Results and comparisons for a benchmark example indicate that our method gives clear improvements

IEEE Annual Conference on Decision and Control (CDC)

Process-Noise Adaptive Particle Filtering with Dependent Process and Measurement Noise

Karl Berntorp[1] and Stefano Di Cairano[1]

***Abstract*— Knowledge of the noise distributions is typically key for reliable state estimation. However, in many applications only the measurement noise can be determined a priori, since only this correspond to measurable quantities. Moreover, modeling of physical systems often leads to nonlinear state-space models with dependent noise sources. Here, we design a computationally efficient marginalized particle filter for jointly estimating the state trajectory and the parameters of the process noise, assuming dependent noise sources. Our approach relies on marginalization and subsequent update of the sufficient statistics of the process-noise parameters. Results and comparisons for a benchmark example indicate that our method gives clear improvements.**

I. Introduction

Particle filtering is a sampling-based technique for solving the nonlinear filtering problem. The particle filter (PF) numerically approximates the posterior density function of the state given the measurement history, by generating random state trajectories and assigning a weight to them according to how well they predict the observations. PFs are often based on the discrete-time state-space model

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{w}_k), \\ \boldsymbol{y}_k &= \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{e}_k), \end{aligned} \tag{1}$$

where $\boldsymbol{x}_k \in \mathbb{R}^{n_x}$ is the state, $\boldsymbol{y}_k \in \mathbb{R}^{n_y}$ is the measurement, and $\boldsymbol{u}_k$ is the known input. The characteristics of the process noise $\boldsymbol{w}_k$ and measurement noise $\boldsymbol{e}_k$ are often treated as known and independent of each other, see for example [1], [2]. However, the noise sources often depend on each other; for example, the discretization of a continuous-time model typically introduces a noise dependence, even if the original system has independent noise [3].

In this paper, we develop a marginalization-based PF [3] for jointly estimating the state and the unknown, possibly time varying, parameters of the Gaussian process noise, when the process noise and the Gaussian measurement noise are dependent. Employing marginalization is crucial to reduce the dimensionality and to increase performance of PFs [4]. To allow for marginalization, our approach relies on moment matching and propagation of the sufficient statistics of the noise parameters. We assume that the structure of the dependence between $\boldsymbol{w}_k$ and $\boldsymbol{e}_k$ is known and that the measurement noise can be decomposed into a dependent and independent part, respectively, with the characteristics of the independent part known a priori. Dependence between $\boldsymbol{w}_k$ and $\boldsymbol{e}_k$ frequently arises in engineering applications, such as in inertial navigation, target tracking, or automotive applications [3], [5]–[7], often as a consequence of discretization of a continuous-time system. The noise dependence typically appears when constructing the model (1) and is therefore known, which motivates our assumption.

A common approach for solving the joint state and parameter estimation is to augment the state vector [8]. Besides increasing the state dimension, which is undesirable in general and particularly for PFs [3], it also requires artificial dynamics of the parameter evolution. Kalman-type filters have been proposed for estimating the unknown statistics of the process and/or measurement noise [9]–[11]. Dependency between $\boldsymbol{w}_{k-1}$ and $\boldsymbol{e}_k$, when the statistics of both noise sources are unknown, has been considered in a PF framework in [12]. However, assuming that both noise sources are unknown might have implications on observability and identifiability of the model. Furthermore, in many applications the measurement noise can be determined a priori, for example from sensor specifications. Another related work is [13], which considers estimation of independent noise sources in a similar framework as in this paper. In [14], estimation of covariances is considered, [15] addresses static parameter estimation for the exponential family, [16] focuses on abruptly changing parameters, and [17] introduces particle learning for estimation of static parameters.

II. Problem Formulation and Preliminaries

We consider a structured version of (1) in the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k) + \boldsymbol{g}(\boldsymbol{x}_k, \boldsymbol{u}_k)\boldsymbol{w}_k, \tag{2a}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k) + \bar{\boldsymbol{e}}_k. \tag{2b}$$

In the following, we assume that $\boldsymbol{u}_k = \boldsymbol{0}$ and we define $\boldsymbol{f}_k := \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k)$ for a function $\boldsymbol{f}$. The measurement noise in (2b) can be decomposed as $\bar{\boldsymbol{e}}_k = \bar{\boldsymbol{g}}_k \boldsymbol{w}_k + \boldsymbol{e}_k$, where $\boldsymbol{w}_k \in \mathbb{R}^{n_w}$ and $\boldsymbol{e}_k \in \mathbb{R}^{n_e}$, and we set $d = n_w + n_e$. Both $\boldsymbol{w}_k$ and $\boldsymbol{e}_k$ are Gaussian distributed as $\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{\mu}_{w,k}, \boldsymbol{Q}_k)$ and $\boldsymbol{e}_k \sim \mathcal{N}(\boldsymbol{\mu}_{e,k}, \boldsymbol{R}_k)$, respectively, where the unknown quantities are defined by $\boldsymbol{\theta}_k := \{\boldsymbol{\mu}_{w,k}, \boldsymbol{Q}_k\}$. The joint Gaussian is written as $\bar{\boldsymbol{w}}_k = \begin{bmatrix} \boldsymbol{w}_k^{\mathrm{T}} & \bar{\boldsymbol{e}}_k^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \sim \mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$, where

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{w,k} \\ \boldsymbol{\mu}_{e,k} + \bar{\boldsymbol{g}}_k \boldsymbol{\mu}_{w,k} \end{bmatrix}, \tag{3a}$$

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{Q}_k & \boldsymbol{Q}_k \bar{\boldsymbol{g}}_k^{\mathrm{T}} \\ \bar{\boldsymbol{g}}_k \boldsymbol{Q} & \bar{\boldsymbol{g}}_k \boldsymbol{Q}_k \bar{\boldsymbol{g}}_k^{\mathrm{T}} + \boldsymbol{R}_k \end{bmatrix}. \tag{3b}$$

The noise processes are assumed to be individually independent. With $p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k})$, we mean the posterior density function of the state trajectory from time index 0 to time index

[1]Karl Berntorp and Stefano Di Cairano are with Mitsubishi Electric Research Laboratories (MERL), 02139 Cambridge, MA, USA. Email:{karl.o.berntorp,dicairano}@ieee.org

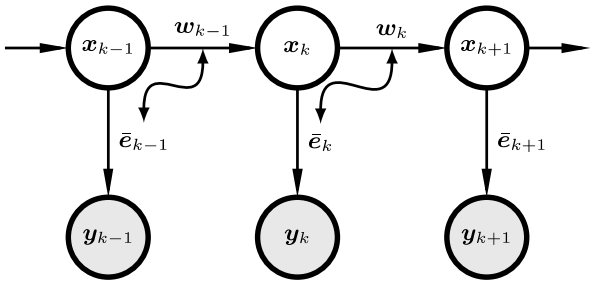

Fig. 1. An illustration of the dependence structure of the noise processes.

k given the measurement sequence $\boldsymbol{y}_{0:k} := \{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_k\}$. The notation $\mathrm{St}(\boldsymbol{\mu}, \Upsilon, \nu)$ reads the multivariate Student-t distribution with mean μ, scaling Υ, and ν degrees of freedom. Similarly, $\mathrm{NiW}(\gamma, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)$ indicates the Normal-inverse-Wishart distribution with statistics (hyperparameters) $S := (\gamma, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)$. The notation $\hat{\boldsymbol{z}}_{k|m}$ means the estimate of $\boldsymbol{z}$ at time index k given measurements up to time index m, and $|\boldsymbol{\Sigma}|$ is the determinant of the matrix $\boldsymbol{\Sigma}$.

Problem 1: Given (2) and dependent Gaussian noise between $\boldsymbol{w}_k$ and $\bar{\boldsymbol{e}}_k$ as in (3), see Fig. 1, where the unknown parameters $\boldsymbol{\theta}$ may be time varying, recursively estimate

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{0:k}), \tag{4a}$$
$$p(\boldsymbol{\theta}_k|\boldsymbol{y}_{0:k}). \tag{4b}$$

We approach Problem 1 by recursively approximating the joint posterior $p(\boldsymbol{\theta}_k, \boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k})$. We rely on the factorization of the joint posterior into conditional densities as

$$p(\boldsymbol{\theta}_k, \boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k}) = p(\boldsymbol{\theta}_k|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k})p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k}). \tag{5}$$

The second term on the right-hand side of (5) is approximated with a PF. To approximate the distribution of the parameters in (5), we rely on updating the sufficient statistics of the parameters. This is possible because we condition on the state trajectory. In what follows, we will go through the necessary steps to efficiently solve Problem 1.

III. Particle Filtering with Dependent Noise Processes

PFs [1], [12] estimate the density $p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k})$ by a set of N weighted trajectories as

$$p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k}) \approx \sum_{i=1}^{N} \omega_k^i \delta(\boldsymbol{x}_{0:k} - \boldsymbol{x}_{0:k}^i). \tag{6}$$

Here, $\delta(\cdot)$ is the Dirac delta function and ω_k^i is the associated importance weight for the ith particle $\boldsymbol{x}_{0:k}^i$. The posterior density can be computed recursively as

$$p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k}) \propto p(\boldsymbol{y}_k|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k-1})p(\boldsymbol{x}_k|\boldsymbol{x}_{0:k-1}, \boldsymbol{y}_{0:k-1}) \cdot p(\boldsymbol{x}_{0:k-1}|\boldsymbol{y}_{0:k-1}). \tag{7}$$

Since it is hard to obtain samples from (6) directly, sampling is done from a tractable, user-designed *proposal distribution*, with the general form $q(\boldsymbol{x}_k|\boldsymbol{x}_{0:k-1}, \boldsymbol{y}_{0:k})$. Inserting (6) into (7) and accounting for the proposal, importance weight ω_k^i is obtained as [12]

$$\omega_k^i \propto \omega_{k-1}^i \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_{0:k}^i, \boldsymbol{y}_{0:k-1})p(\boldsymbol{x}_k^i|\boldsymbol{x}_{0:k-1}^i, \boldsymbol{y}_{0:k-1})}{q(\boldsymbol{x}_k^i|\boldsymbol{x}_{0:k-1}^i, \boldsymbol{y}_{0:k})}. \tag{8}$$

Setting $q(\boldsymbol{x}_k|\boldsymbol{x}_{0:k-1}, \boldsymbol{y}_{0:k}) = p(\boldsymbol{x}_k|\boldsymbol{x}_{0:k-1}, \boldsymbol{y}_{0:k-1})$ reduces the weight update to

$$\omega_k^i \propto \omega_{k-1}^i p(\boldsymbol{y}_k|\boldsymbol{x}_{0:k}^i, \boldsymbol{y}_{0:k-1}^i). \tag{9}$$

When the process noise $\boldsymbol{w}$ and measurement noise $\boldsymbol{e}$ are independent (i.e., $p(\boldsymbol{w}_k, \boldsymbol{e}_m) = p(\boldsymbol{w}_k)p(\boldsymbol{e}_m)$), (8) transforms to the well-known update equation

$$\omega_k^i \propto \omega_{k-1}^i \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k^i)p(\boldsymbol{x}_k^i|\boldsymbol{x}_{k-1}^i)}{q(\boldsymbol{x}_k^i|\boldsymbol{x}_{k-1}^i, \boldsymbol{y}_k)}.$$

However, for dependence as in Fig. 1, with known noise parameters, the factors in the numerator in (8) equal

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{0:k-1}, \boldsymbol{y}_{0:k-1}) = p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}), \tag{10a}$$
$$p(\boldsymbol{y}_k|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k-1}) = p(\boldsymbol{y}_k|\boldsymbol{x}_k). \tag{10b}$$

From (10b), knowing $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ provides full knowledge of $\boldsymbol{e}_k$. Hence, (10a) can be characterized by the conditional distribution $p(\boldsymbol{w}_{k-1}|\boldsymbol{e}_{k-1})$. When the parameters of the noise processes are known, the noise processes can be decorrelated using Gram-Schmidt orthogonalization [12]. However, for unknown process noise, such an approach is not possible.

IV. Joint State and Noise-Parameter Estimation

Regardless whether (8) or (9) is used in the PF weight update, we need to compute

$$p(\boldsymbol{y}_k|\boldsymbol{x}_{0:k}^i, \boldsymbol{y}_{0:k-1}), \tag{11a}$$
$$p(\boldsymbol{x}_{k+1}^i|\boldsymbol{x}_{0:k}^i, \boldsymbol{y}_{0:k}), \tag{11b}$$

where we have time shifted (11b) to be consistent with (2). The unknown process-noise parameters affect both the prediction and measurement update steps according to (2), so both the prediction step and the weight update in the PF will depend on the parameter estimates.

To solve Problem 1, we will first describe how to compute the first term on the right-hand side in (5), assuming that we know the state trajectory. This is then followed by a procedure for computing (11), which allows us to compute the posterior of the state trajectory in (5).

A. Parameter Estimation

According to (11), knowing both the state and measurement trajectory leads to full knowledge about $\bar{\boldsymbol{w}}_{0:k}$. Hence, the posterior for the noise parameters in (5) can be rewritten using Bayes' rule as

$$p(\boldsymbol{\theta}_k|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k}) = p(\boldsymbol{\theta}_k|\bar{\boldsymbol{w}}_{0:k}) \propto p(\bar{\boldsymbol{w}}_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\bar{\boldsymbol{w}}_{0:k-1}). \tag{12}$$

The posterior of the parameters in (12) is composed of a Gaussian likelihood $p(\bar{\boldsymbol{w}}_k|\boldsymbol{\theta}_k)$ and a prior $p(\boldsymbol{\theta}_k|\bar{\boldsymbol{w}}_{0:k-1})$, similar to the measurement update in the PF. Therefore, we can utilize conjugate priors.

Definition 1: Given a likelihood, the conjugate prior is the prior distribution such that the prior and posterior are in the same family of distributions.

Lemma 1 provides an explicit expression of the conjugate prior for Gaussian likelihoods [18].

Lemma 1: For multivariate Normal data $\bar{\boldsymbol{w}} \in \mathbb{R}^d$ with unknown mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, a Normal-inverse-Wishart distribution defines the conjugate prior $p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) := \mathrm{NiW}(\gamma_{k|k}, \hat{\boldsymbol{\mu}}_{k|k}, \boldsymbol{\Lambda}_{k|k}, \nu_{k|k})$, through the hierarchical model

$$\begin{aligned}\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{k|k}, \gamma_{k|k}\boldsymbol{\Sigma}_k),\\ \boldsymbol{\Sigma}_k &\sim \mathrm{iW}(\nu_{k|k}, \boldsymbol{\Lambda}_{k|k})\\ &\propto |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}(\nu_{k|k}+d+1)} e^{\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{k|k}\boldsymbol{\Sigma}_k^{-1})\right.},\end{aligned}$$

where $\mathrm{tr}(\cdot)$ is the trace operator.

The computation of the statistics in Lemma 1 can be done in several ways. In this work, we use the statistics $S_{k|k} := (\gamma_{k|k}, \hat{\boldsymbol{\mu}}_{k|k}, \boldsymbol{\Lambda}_{k|k}, \nu_{k|k})$ (see [13] and references therein)

$$\gamma_{k|k} = \frac{\gamma_{k|k-1}}{1+\gamma_{k|k-1}}, \tag{13a}$$

$$\hat{\boldsymbol{\mu}}_{k|k} = \hat{\boldsymbol{\mu}}_{k|k-1} + \gamma_{k|k}\boldsymbol{z}_k, \tag{13b}$$

$$\nu_{k|k} = \nu_{k|k-1} + 1, \tag{13c}$$

$$\boldsymbol{\Lambda}_{k|k} = \boldsymbol{\Lambda}_{k|k-1} + \frac{1}{1+\gamma_{t|t-1}}\boldsymbol{z}_k\boldsymbol{z}_k^{\mathrm{T}}, \tag{13d}$$

$$\boldsymbol{z}_k = \bar{\boldsymbol{w}}_k - \hat{\boldsymbol{\mu}}_{k|k-1}. \tag{13e}$$

For slowly time-varying parameters, the prediction step consists of

$$\begin{aligned}\gamma_{k|k-1} &= \frac{1}{\lambda}\gamma_{k-1|k-1},\\ \hat{\boldsymbol{\mu}}_{k|k-1} &= \hat{\boldsymbol{\mu}}_{k-1|k-1},\\ \nu_{k|k-1} &= \lambda\nu_{k-1|k-1},\\ \boldsymbol{\Lambda}_{k|k-1} &= \lambda\boldsymbol{\Lambda}_{k-1|k-1},\end{aligned} \tag{14}$$

where $\lambda \in [0,1]$ provides exponential forgetting. Furthermore, for a Normal-inverse-Wishart prior, the predictive distribution of the data $\bar{\boldsymbol{w}}$ is a Student-t,

$$\mathrm{St}\left(\hat{\boldsymbol{\mu}}_{k|k-1}, \frac{1+\gamma_{k|k-1}}{\nu_{k|k-1}-d+1}\boldsymbol{\Lambda}_{k|k-1}, \nu_{k|k-1}-d+1\right).$$

Suppose now that the predictive distribution $p(\boldsymbol{\theta}_k|\bar{\boldsymbol{w}}_{0:k-1})$ in (12) is Normal-inverse-Wishart. Hence,

$$\begin{aligned}p(\boldsymbol{\theta}_k|\bar{\boldsymbol{w}}_{0:k-1}) =\\ \mathrm{NiW}(\gamma_{k|k-1}, \hat{\boldsymbol{\mu}}_{k|k-1}, \boldsymbol{\Lambda}_{k|k-1}, \nu_{k|k-1}).\end{aligned} \tag{15}$$

Using (12), (15), and Lemma 1, results in that the posterior is also Normal-inverse Wishart,

$$p(\boldsymbol{\theta}_k|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k}) = \mathrm{NiW}(\gamma_{k|k}, \hat{\boldsymbol{\mu}}_{k|k}, \boldsymbol{\Lambda}_{k|k}, \nu_{k|k}). \tag{16}$$

Note that only the parameters of $\boldsymbol{w}_k$ are unknown, so (13) and (14) are only applied to the process noise, that is, $\bar{\boldsymbol{w}} \in \mathbb{R}^{n_w}$ in Lemma 1, which decreases the dimensionality.

Finally, to find (4b) in Problem 1, we marginalize out the state trajectory as

$$\begin{aligned}p(\boldsymbol{\theta}_k|\boldsymbol{y}_{0:k}) &= \int p(\boldsymbol{\theta}_k|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k})p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k})\mathrm{d}\boldsymbol{x}_{0:k}\\ &\approx \sum_{i=1}^{N}\omega_k^i p(\boldsymbol{\theta}_k|\boldsymbol{x}_{0:k}^i, \boldsymbol{y}_{0:k}),\end{aligned} \tag{17}$$

which has complexity $\mathcal{O}(N)$, where $p(\boldsymbol{\theta}_k|\boldsymbol{x}_{0:k}^i, \boldsymbol{y}_{0:k})$ is given by (16). The unknown parameters can be extracted from (17); for instance, the minimum mean-square estimate of $\boldsymbol{\mu}_{w,k}$ is

$$\hat{\boldsymbol{\mu}}_{w,k} = \sum_{i=1}^{N}\omega_k^i\hat{\boldsymbol{\mu}}_{w,k|k}^i.$$

B. State Estimation

To estimate the state, we need to determine expressions for (11). Lemma 2 provides the basic version of a useful result on transformation of variables in densities.

Lemma 2: Let X be a random variable with probability density function $p(\boldsymbol{x})$. Let $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ be one-to-one for which $\boldsymbol{g}^{-1}(\boldsymbol{y})$ exists with a continuous Jacobian $J(\boldsymbol{y}) = \frac{\partial \boldsymbol{g}^{-1}}{\partial \boldsymbol{y}}$. Then, the random variable $Y = \boldsymbol{g}(X)$ has the probability density function

$$p(\boldsymbol{y}) = |J(\boldsymbol{y})|p(\boldsymbol{g}^{-1}(\boldsymbol{y})).$$

Proof: See [19]. ∎

To compute (11a), we first note that from (2) and the noise dependence, knowledge of $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ characterizes $\bar{\boldsymbol{e}}_k$. Using Lemma 2, adapted to our scenario, with $|J(\boldsymbol{y}_k)| = 1$,

$$\begin{aligned}p(\boldsymbol{y}_k|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k-1}) &= p(\bar{\boldsymbol{e}}_k(\boldsymbol{y}_k, \boldsymbol{x}_k)|\boldsymbol{x}_{0:k-1}, \boldsymbol{y}_{0:k-1})\\ &= p(\bar{\boldsymbol{e}}_k(\boldsymbol{y}_k, \boldsymbol{x}_k)|\bar{\boldsymbol{e}}_{0:k-1}).\end{aligned} \tag{18}$$

We marginalize out the noise parameters using the law of total probability as

$$\begin{aligned}p(\boldsymbol{y}_k|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k-1}) = \int p(\boldsymbol{y}_k|\boldsymbol{\theta}_k, \boldsymbol{x}_k)\\ \cdot p(\boldsymbol{\theta}_k|\boldsymbol{x}_{0:k-1}, \boldsymbol{y}_{0:k-1})\,\mathrm{d}\boldsymbol{\theta}_k.\end{aligned} \tag{19}$$

If we for a moment concentrate on $\bar{\boldsymbol{g}}_k\boldsymbol{w}_k$ and ignore $\boldsymbol{e}_k$ in $p(\boldsymbol{y}_k|\boldsymbol{\theta}_k, \boldsymbol{x}_k)$, (19) is the integral of the product of a Gaussian distribution and a Normal-inverse-Wishart distribution. Hence, $p(\bar{\boldsymbol{e}}_k(\boldsymbol{y}_k, \boldsymbol{x}_k)|\bar{\boldsymbol{e}}_{0:k-1})$ in (18) is a Student-t distribution [18]

$$p(\bar{\boldsymbol{e}}_k(\boldsymbol{y}_k, \boldsymbol{x}_k)|\bar{\boldsymbol{e}}_{0:k-1}) = \mathrm{St}(\hat{\boldsymbol{\mu}}_{k|k-1}, \tilde{\boldsymbol{\Lambda}}_{k|k-1}, \tilde{\nu}_{k|k-1}), \tag{20}$$

with $\bar{\nu}_{k|k-1} = \nu_{k|k-1} - n_e + 1$ and mean and scaling as

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{k|k-1} &= \bar{\boldsymbol{g}}_k\hat{\boldsymbol{\mu}}_{w,k|k-1},\\ \tilde{\boldsymbol{\Lambda}}_{k|k-1} &= \frac{1+\gamma_{k|k-1}}{\tilde{\nu}_{k|k-1}}\bar{\boldsymbol{g}}_k\boldsymbol{\Lambda}_{w,k|k-1}\bar{\boldsymbol{g}}_k^{\mathrm{T}}.\end{aligned}$$

However, in our case, $\bar{\boldsymbol{e}}_k$ is partially known through the Gaussian $\boldsymbol{e}_k$, which implies that (19) is a mixture of a Gaussian and a Student-t, whose density is computed as an infinite series [20]. To obtain an algorithm suitable for online implementations, we rely on Approximation 1.

Approximation 1: $p(\bar{e}_k(\boldsymbol{y}_k, \boldsymbol{x}_k)|\bar{e}_{0:k-1})$ is distributed according to (20) with mean $\hat{\boldsymbol{\mu}}_{\bar{e},k|k-1}$ and scale $\tilde{\boldsymbol{\Lambda}}_{\bar{e},k|k-1}$ as

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{\bar{e},k|k-1} &= \bar{\boldsymbol{g}}_k \hat{\boldsymbol{\mu}}_{w,k|k-1} + \boldsymbol{\mu}_{e,k},\\ \tilde{\boldsymbol{\Lambda}}_{\bar{e},k|k-1} &= \frac{1+\gamma_{k|k-1}}{\tilde{\nu}_{k|k-1}} \bar{\boldsymbol{g}}_k \boldsymbol{\Lambda}_{w,k|k-1} \bar{\boldsymbol{g}}_k^{\mathrm{T}} + \frac{\tilde{\nu}_{k|k-1}-2}{\tilde{\nu}_{k|k-1}} \boldsymbol{R}.\end{aligned} \tag{21}$$

Approximation 1 implies moment matching of a known Gaussian with a Student-t and can be interpreted as a robustification of the measurement noise, by choosing the smallest common degree of freedom [21]. Proposition 1 justifies the approximation.

Proposition 1: As $k \to \infty$, (20) converges to a Gaussian with precision determined by the forgetting factor λ.

Proof: From (13) and (14), $\lim_{k\to\infty} \nu_{k|k} = 1/(1-\lambda)$. The result immediately follows from that $\lim_{\nu\to\infty} \mathrm{St}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ and that the sum of Gaussians is a Gaussian. ■

We compute (11a) by replacing the mean in (20) with $\boldsymbol{h}_k + \hat{\boldsymbol{\mu}}_{\bar{e},k|k-1}$. To compute (11b), from Lemma 2,

$$\begin{aligned}p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k}) &\propto p(\boldsymbol{g}_k^{-\dagger}(\boldsymbol{x}_{k+1} - \boldsymbol{f}_k)|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{0:k})\\ &= p(\boldsymbol{g}_k^{-\dagger}(\boldsymbol{x}_{k+1} - \boldsymbol{f}_k)|\bar{e}_{0:k})\\ &= p(\boldsymbol{w}_k(\boldsymbol{x}_{k+1})|\bar{e}_{0:k}),\end{aligned} \tag{22}$$

where $\boldsymbol{g}_k^{-\dagger}$ is the pseudo-inverse of $\boldsymbol{g}_k$. By marginalizing out the noise parameters in (11b) and combining with (22),

$$p(\boldsymbol{w}_k(\boldsymbol{x}_{k+1})|\bar{e}_{0:k}) = \mathrm{St}(\hat{\boldsymbol{\mu}}_k^*, \tilde{\boldsymbol{\Lambda}}_k^*, \nu_k^*). \tag{23}$$

Theorem 1 provides the hyperparameters in (23).

Theorem 1: The hyperparameters in (23) are given by

$$\begin{aligned}\hat{\boldsymbol{\mu}}_k^* &= \hat{\boldsymbol{\mu}}_{w,k|k-1} + \boldsymbol{d}_k \boldsymbol{\Lambda}_{w,k|k-1} \tilde{\boldsymbol{\Lambda}}_{\bar{e},k|k-1}^{-1} \boldsymbol{z}_k,\\ \tilde{\boldsymbol{\Lambda}}_k^* &= \frac{\nu_{k|k-1} - d + 1 + \boldsymbol{z}_k \tilde{\boldsymbol{\Lambda}}_{\bar{e},k|k-1}^{-1} \boldsymbol{z}_k^{\mathrm{T}}}{\nu_{k|k-1} - n_e + 1} \Big(\boldsymbol{\Lambda}_{w,k|k-1}\\ &\quad - \boldsymbol{d}_k \boldsymbol{\Lambda}_{w,k|k-1} \tilde{\boldsymbol{\Lambda}}_{\bar{e},k|k-1}^{-1} \boldsymbol{\Lambda}_{w,k|k-1}^{\mathrm{T}} \boldsymbol{d}_k^{\mathrm{T}} \Big),\\ \nu_k^* &= \nu_{k|k-1} - n_w + 1,\\ \boldsymbol{z}_k &= \bar{\boldsymbol{e}}_k - \hat{\boldsymbol{\mu}}_{\bar{e},k|k-1}.\end{aligned} \tag{24}$$

Proof: First, given Approximation 1, the joint predictive distribution of $\boldsymbol{w}_k$ and $\bar{\boldsymbol{e}}_k$ is a Student-t. Now, for $\boldsymbol{x}_1 \in \mathbb{R}^{d_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{d_2}$ jointly distributed according to a Student-t,

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathrm{St}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{12}^{\mathrm{T}} & \boldsymbol{\Lambda}_{22} \end{bmatrix}, \nu \right),$$

by using the factorization $p(\boldsymbol{x}_1|\boldsymbol{x}_2) = p(\boldsymbol{x}_1, \boldsymbol{x}_2)/p(\boldsymbol{x}_2)$, it can be shown [22] that the conditional density is given by

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathrm{St}(\boldsymbol{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Lambda}_{1|2}, \nu_{1|2}), \tag{25}$$

where

$$\begin{aligned}\nu_{1|2} &= \nu + d_2,\\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Lambda}_{12} \boldsymbol{\Lambda}_{22}^{-1} (\boldsymbol{x}_2 - \boldsymbol{\mu}_2),\\ \boldsymbol{\Lambda}_{12} &= \frac{\nu + (\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \boldsymbol{\Lambda}_{22}^{-1} (\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^{\mathrm{T}}}{\nu + d_2} \Big(\boldsymbol{\Lambda}_{11}\\ &\quad - \boldsymbol{\Lambda}_{12} \boldsymbol{\Lambda}_{22}^{-1} \boldsymbol{\Lambda}_{12}^{\mathrm{T}} \Big).\end{aligned}$$

Then, the hyperparameters (24) follow from (25) by utilizing (20), (21), yielding

$$\begin{aligned}\nu &= \nu_{k|k-1} - d + 1, \quad d_2 = n_e,\\ \boldsymbol{\mu}_1 &= \boldsymbol{\mu}_{w,k|k-1},\\ \boldsymbol{\Lambda}_{11} &= \boldsymbol{\Lambda}_{w,k|k-1},\\ \boldsymbol{\Lambda}_{12} &= \bar{\boldsymbol{g}}_k \boldsymbol{\Lambda}_{w,k|k-1},\\ \boldsymbol{\Lambda}_{22} &= \tilde{\boldsymbol{\Lambda}}_{\bar{e},k|k-1},\\ \boldsymbol{x}_2 - \boldsymbol{\mu}_2 &= \bar{\boldsymbol{e}}_k - \hat{\boldsymbol{\mu}}_{\bar{e},k|k-1}.\end{aligned}$$

■

Next, (11b) is found by replacing the mean in (23) with $\boldsymbol{f}_k + \hat{\boldsymbol{\mu}}_k^*$. To summarize, (11a) and (11b), and therefore (8) (or (9) if (11b) is used directly as proposal density) can be computed by evaluation of (20), (21), and (23).

To obtain an approximation of the filtering distribution (4a) in Problem 1, we extract the last state to obtain

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{0:k}) \approx \sum_{i=1}^{N} \omega_k^i \delta(\boldsymbol{x}_k - \boldsymbol{x}_k^i). \tag{26}$$

Remark 1: Both (26) and (17) overlook a potential path-degeneracy problem, but taking into account different paths leads to an algorithm that is intractable in many online implementations, which is the focus here. Furthermore, for sufficient mixing in the dynamic model (2), errors in the state are forgotten exponentially in time and ensures convergence of (26) as $N \to \infty$ [1]. For (17), the use of exponential forgetting suppresses the path-degeneracy problem, which causes issues for estimation of static parameters [13].

C. Summary of the Algorithm

If (11b) is used as proposal the algorithm simplifies, since samples are first generated from (23), and then used in the dynamic model (2a) to create $\boldsymbol{x}_{k+1}^i$. Furthermore, the samples can be used directly in (13e) to update the statistics $S_{k|k}$. If a general proposal is used, we need to use (22) to update the statistics $S_{k|k}$, which increases complexity. The algorithm is summarized in Algorithm 1 for the simplified setting, with (11b) as proposal.

V. Numerical Evaluation

We use a common benchmark problem [3], [6], [13], consisting of scalar nonlinear dynamics with a squared measurement relation,

$$x_{k+1} = \frac{x_k}{2} + \frac{25 x_k}{1 + x_k^2} + 8\cos(1.2k) + w_k, \tag{27a}$$

$$y_k = \frac{x_k^2}{20} + e_k. \tag{27b}$$

The noise sources are dependent and Gaussian distributed, with initial parameters set as $\mu_{w,0} = 1$, $\Sigma_{w,0} = 4$, $\mu_{e,0} = 0$, $\Sigma_{e,0} = 3$, and $\Sigma_{we,0} = 0.5\Sigma_{w,0}$. We execute 100 Monte-Carlo simulations and use the time average of the root-mean-square error (RMSE). Because of the quadratic term in (27b), the resulting filtering posterior is in general bimodal [6], limiting the use of linear/linearized filters.

Algorithm 1 Filtering with Dependent Noise

Initialize: Set $\{\boldsymbol{x}_0^i\}_{i=1}^N \sim p_0(\boldsymbol{x}_0)$, $\{\omega_0^i\}_{i=1}^N = 1/N$, $\{S_0^i\}_{i=1}^N = \{\gamma_0^i, \boldsymbol{\mu}_{w,0}^i, \boldsymbol{\Lambda}_{w,0}^i, \nu_0^i\}$

for $k \leftarrow 0$ to T **do**
  for $i \in \{1, \ldots, N\}$ **do**
    Update weight $\bar{\omega}_k^i$ using (20) and (21):

$$\bar{\omega}_k^i = \omega_{k-1}^i p(\boldsymbol{y}_k | \bar{\boldsymbol{e}}_{0:k-1})$$

    Update noise statistics $S_{k|k}^i$ using (13).
  end for
  Normalize weights as $\omega_k^i = \bar{\omega}_k^i / (\sum_{i=1}^N \bar{\omega}_k^i)$.
  Compute $N_{\text{eff}} = 1/(\sum_{i=1}^N (\omega_k^i)^2)$
  if $N_{\text{eff}} \leq N_{\text{thr}}$ **then**
    Resample particles and copy the corresponding statistics. Set $\{\omega_k^i\}_{i=1}^N = 1/N$.
  end if
  Approximate state posterior with (26).
  Approximate parameter posterior with (17).
  for $i \in \{1, \ldots, N\}$ **do**
    Predict noise statistics $S_{k+1|k}^i$ using (14).
    Sample $\boldsymbol{w}_k^i$ from (23) using (24).
    Predict state $\boldsymbol{x}_{k+1}^i$ using (2a).
  end for
end for

We compare Algorithm 1 with a state-augmented PF (AUGPF) and the method in [13]. All of these methods have computational complexity $\mathcal{O}(N)$. The method in [13] assumes independent noise sources. Hence, the comparison shows the benefits of accounting for noise dependence in this particular problem. In AUGPF, the state vector is augmented to also include the unknown parameters of the process noise. The unknown mean is modeled as a Gaussian random walk, and the inverse-Gamma distribution is used to propagate the unknown variance, $p(\Sigma_{w,k}, \Sigma_{w,k-1}) = i\Gamma(\alpha, \beta)$. The deviation of the random walk is set to 3% of the value of the true parameter, and α, β are set such that the mean value is held constant and the deviation of the distribution is 3% of the previous value [13]. These values were set as a tradeoff between small average error, fast convergence, and reasonably small variability of the estimates over 100 Monte-Carlo executions. All algorithms are initialized with the same values. The mean of the initial state and the initial covariance, respectively, are $x_0 = 5, P_0 = 5$, which is also used to generate the ground truth. All algorithms assume that the initial values of the mean and variance of the process noise are $\mu_{w,0} = 3, \Sigma_{w,0} = 9$.

Figs. 2–4 show the resulting parameter estimates for 500 particles. Clearly, Algorithm 1 (Fig. 2) provides more accurate and less biased estimates than when not accounting for the dependence (Fig. 4). The standard deviation over the executions, shown by the size of the vertical bars (2σ) in the figures, indicate that the variability between different Monte-Carlo executions are much larger for AUGPF (Fig. 3).

Fig. 5 provides a comparison of the time-averaged RMSE

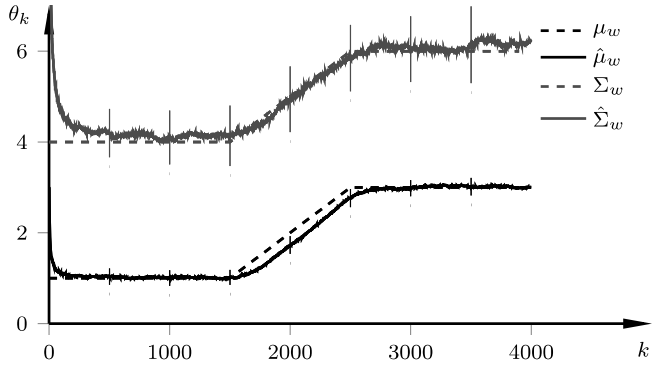

Fig. 2. Estimated mean and standard deviation of the process noise using Algorithm 1, averaged over 100 Monte-Carlo simulations using 500 particles, with $\lambda = 0.99$. The vertical bars indicate the 2 standard deviations of the estimates.

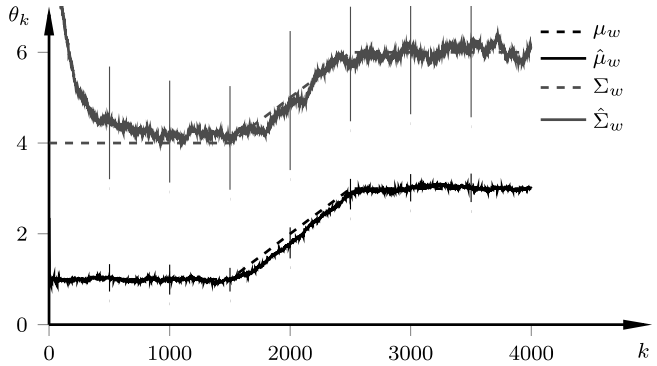

Fig. 3. Estimated mean and standard deviation of the process noise using AUGPF, averaged over 100 Monte-Carlo executions using 500 particles.

of the estimated standard deviation for different number of particles. There is a clear performance increase in the variance estimation by accounting for the dependence. The conclusions for the mean value in Fig. 6 are similar. Both Fig. 5 and the state error in Fig. 6 indicate that irrespective of the number of particles, Algorithm 1 performs best.

Table I provides the corresponding average execution time for one Monte-Carlo trial in a MATLAB implementation. Care should be taken when interpreting the execution time, since it is highly implementation and program-language dependent. Nevertheless, to get the same time-averaged state

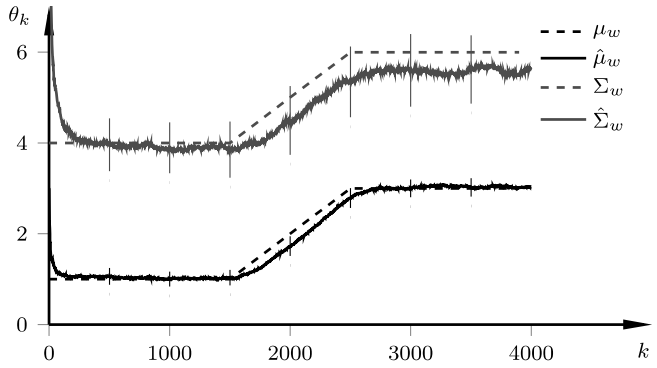

Fig. 4. Estimated mean and standard deviation of the process noise using the approach in [13], averaged over 100 Monte-Carlo executions using 500 particles, with $\lambda = 0.99$.

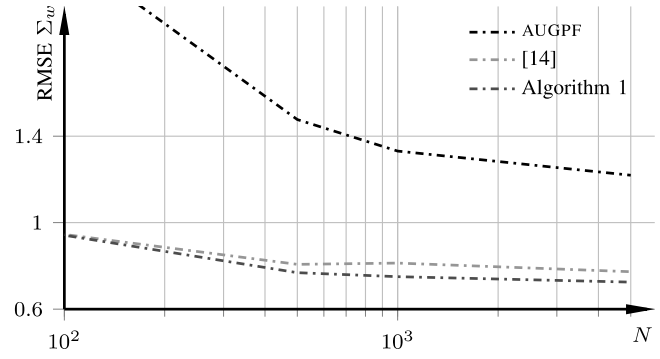

Fig. 5. Time-averaged RMSE of the standard deviation of the process noise versus the number of particles, averaged over 100 Monte-Carlo executions. Values are obtained using 100, 500, 1000, 5000 particles.

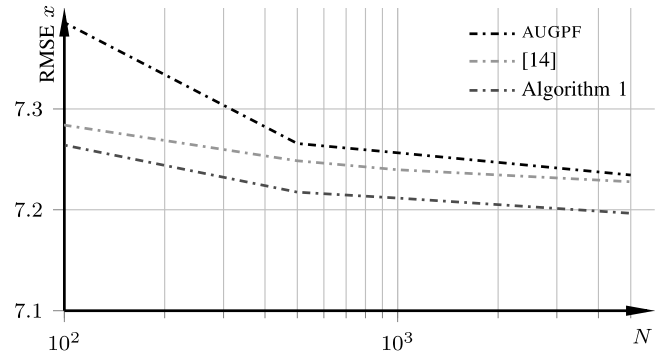

Fig. 6. Time-averaged RMSE of the state versus the number of particles, averaged over 100 Monte-Carlo executions. Values are obtained using 100, 500, 1000, 5000 particles.

RMSE (Fig. 6) as when using 500 particles in Algorithm 1, AUGPF and the method in [13] need well more than 5000 particles, corresponding to an execution time of 2–3.5 times that of Algorithm 1. Thus, the increase in execution time is more than compensated for by the performance increase, in this example.

VI. Conclusion

We developed a PF-based method for jointly estimating the state and learning the parameters of the underlying Gaussian process noise. The method is applicable to nonlinear state-space models where the parameters of the process noise are unknown and there is a dependence between the process noise and the measurement noise. As pointed out, this is a situation that arises naturally in many applications.

The proposed method relies on conjugate priors and moment matching to obtain a computationally efficient marginalized particle filter, and by using exponential forgetting in the parameter prediction, the problem of path dependence is significantly suppressed. A numerical study demonstrated the significance of the approach.

TABLE I

Average execution times in seconds for one Monte-Carlo execution in Matlab for different number of particles.

N	100	500	1000	5000
AUGPF	1.5	2.3	3.2	9.2
[13]	2.8	4	5.4	15
Algorithm 1	3	4.4	6	16.8

References

[1] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," in *Handbook of Nonlinear Filtering*, D. Crisan and B. Rozovsky, Eds. Oxford University Press, 2009.

[2] F. Gustafsson, "Particle filter theory and practice with positioning applications," *IEEE Aerosp. Electron. Syst. Mag*, vol. 25, no. 7, pp. 53–82, 2010.

[3] ——, *Statistical Sensor Fusion*. Lund, Sweden: Utbildningshuset/Studentlitteratur, 2010.

[4] T. B. Schön, F. Gustafsson, and P.-J. Nordlund, "Marginalized particle filters for mixed linear nonlinear state-space models," *IEEE Trans. Signal Processing*, vol. 53, pp. 2279–2289, 2005.

[5] F. Gustafsson and S. Saha, "Particle filtering with dependent noise," in *13th Int. Conf. Information Fusion*, Edinburgh, UK, July 2010.

[6] K. Berntorp, "Particle filtering and optimal control for vehicles and robots," Ph.D. dissertation, Department of Automatic Control, Lund University, Sweden, May 2014.

[7] ——, "Joint wheel-slip and vehicle-motion estimation based on inertial, GPS, and wheel-speed sensors," *IEEE Trans. Contr. Syst. Technol.*, vol. 24, no. 3, pp. 1020–1027, 2016.

[8] J. Liu and M. West, "Combined parameter and state estimation in simulation-based filtering," in *Sequential Monte Carlo Methods in Practice*, ser. Statistics for Engineering and Information Science, A. Doucet, N. de Freitas, and N. Gordon, Eds. Springer New York, 2001, pp. 197–223.

[9] R. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Trans. Automat. Contr.*, vol. 15, no. 2, pp. 175–184, 1970.

[10] S. Särkkä and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *IEEE Trans. Automat. Contr.*, vol. 54, no. 3, pp. 596–600, 2009.

[11] X. Rong Li and Y. Bar-Shalom, "A recursive multiple model approach to noise identification," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 30, no. 3, pp. 671–684, 1994.

[12] S. Saha and F. Gustafsson, "Particle filtering with dependent noise processes," *IEEE Trans. Signal Processing*, vol. 60, no. 9, pp. 4497–4508, 2012.

[13] E. Özkan, V. Šmídl, S. Saha, C. Lundquist, and F. Gustafsson, "Marginalized adaptive particle filtering for nonlinear models with unknown time-varying noise parameters," *Automatica*, vol. 49, no. 6, pp. 1566–1575, 2013.

[14] P. Djuric and J. Miguez, "Sequential particle filtering in the presence of additive Gaussian noise with unknown parameters," in *IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Orlando, Fl, May 2002.

[15] G. Storvik, "Particle filters for state-space models with the presence of unknown static parameters," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 281–289, 2002.

[16] C. Nemeth, P. Fearnhead, and L. Mihaylova, "Sequential Monte Carlo methods for state and parameter estimation in abruptly changing environments," *IEEE Trans. Signal Processing*, vol. 62, no. 5, pp. 1245–1255, 2014.

[17] C. Carvalho, M. S. Johannes, H. F. Lopes, and N. Polson, "Particle learning and smoothing," *Statistical Science*, vol. 25, no. 1, pp. 88–106, 2010.

[18] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," UBC, Tech. Rep., 2007.

[19] C. R. Rao, *Linear Statistical Inference and its Applications*. Wiley, 2001.

[20] C. Berg and C. Vignat, "On the density of the sum of two independent Student t-random vectors," *Statistics & Probability Letters*, vol. 80, no. 13, pp. 1043–1055, 2010.

[21] M. Roth, E. Özkan, and F. Gustafsson, "A Student's t filter for heavy tailed process and measurement noise," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013.

[22] M. Roth, "On the multivariate t distribution," Linköping University, Tech. Rep. 3059, 2012.