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

Approximate Noise-Adaptive Filtering Using Student-t Distributions

Berntorp, K.; Di Cairano, S.

TR2018-088 July 13, 2018

Abstract

We present an adaptive method for Bayesian filtering of linear state-space models with unknown noise statistics. The proposed method makes use of separation of the state and parameter posterior at each time step recursively for subsequent approximate inference. The filter exploits properties of the inverse-Wishart and the Student-t distributions and has relations to recent results from outlier-robust filtering. The method is well suited to platforms with limited computational resources because of its simplicity. Simulation results show that the proposed method can correctly estimate the measurementnoise statistics under large initial errors, in addition to being robust to outliers in the measurement and process noise.

American Control Conference (ACC)

Approximate Noise-Adaptive Filtering Using Student-t Distributions

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

***Abstract*— We present an adaptive method for Bayesian filtering of linear state-space models with unknown noise statistics. The proposed method makes use of separation of the state and parameter posterior at each time step recursively for subsequent approximate inference. The filter exploits properties of the inverse-Wishart and the Student-t distributions and has relations to recent results from outlier-robust filtering. The method is well suited to platforms with limited computational resources because of its simplicity. Simulation results show that the proposed method can correctly estimate the measurement-noise statistics under large initial errors, in addition to being robust to outliers in the measurement and process noise.**

I. Introduction

The Kalman filter (KF) is the standard tool for state estimation in linear state-space models [1]. It is the best linear unbiased filter in the minimum-variance sense, and for Gaussian noise it is the optimal Bayesian filter [2]. The classical formulation of the KF assumes that the noise processes are Gaussian and have known mean and covariance, which can be severely limiting. Model uncertainties and possible data outliers affect the performance of the KF, and in many practical cases the model parameters are unknown, or at least uncertain. For instance, in navigation systems where inertial sensing and/or GPS is used [3], the noise statistics often have temporal dependence that cannot be determined a priori. Other examples are changing noise statistics due to linearization errors in approximated nonlinear models, environment dependent sensor statistics, and outliers in unreliable sensors that the Gaussian distribution handles poorly because of its low probability mass in the tails. The noise parameters determine the reliability of the different parts of the model and are therefore of particular importance for the filter performance. However, manual tuning of the noise parameters, as is often done in practice, can be a challenging, time consuming, and tedious task.

This paper develops a computationally efficient Bayesian approach for joint estimation of the state and the parameters of the noise for linear state-space models. The formal solution to the joint state and noise-parameter filtering problem is intractable and approximate solutions are necessary. Our proposed method is based on intermediate Student-t approximations of the state posterior, to account for the uncertainty of the noise parameters. We use the conjugacy of the Normal-inverse Wishart (NiW) distribution to the Gaussian likelihood [4], and exploit that the posterior predictive distribution of the NiW is Student-t distributed [5]. This connects the parameter and state estimates and motivates the intermediate Student-t approximations of the state posterior. The resulting filtering equations for the state are similar to the KF, except for a nonlinear dependence on the measurement. Our method has limited computational demands and is therefore well suited for embedded implementations.

One of the early works on adaptive KF for noise identification is [6], and an example of variational Bayesian (VB) methods for inference on systems with uncertain parameters can be found in [7]. These methods can be extended to the nonlinear setting in the spirit of the extended KF (EKF) and the unscented KF (UKF). Another alternative to handle nonlinearities is to resort to particle filtering (PF) [8], [9]. PF can achieve arbitrary precision, but at the price of a higher computational cost than the KF-type filters. Approaches based on variational approximations and the Student-t distribution can be found in [10], [11]. Our proposed method is similar to the Student-t filtering method in [12] in that we make use of intermediate Student-t approximations of the state posterior to achieve a readily implementable algorithm. Because of its relations to outlier-robust filtering, our proposed algorithm is expected to avoid some of the robustness issues with the Gaussian-noise assumption [12], which we verify in our numerical validation. We make a simplifying approximation by separating the state and parameter posterior, similar to what has previously been proposed in VB approaches (see, e.g., [7]). An important distinction is that we propagate the state posterior using Student-t distributions, which makes for a natural connection with the NiW distribution, whose predictive density is a Student-t.

Notation: With $p(\boldsymbol{x}_k|\boldsymbol{y}_{0:k})$, we mean the posterior density function of the state $\boldsymbol{x}_k$ at time index k conditioned on the measurement sequence $\boldsymbol{y}_{0:k} := \{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_k\}$. Throughout, for a vector $\boldsymbol{x}$, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates that $\boldsymbol{x}$ is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $|\boldsymbol{\Sigma}|$ is the determinant of the matrix $\boldsymbol{\Sigma}$. The notation $\mathrm{St}(\boldsymbol{\mu}, \Upsilon, \nu)$ means the multivariate Student-t distribution with mean $\boldsymbol{\mu}$, scaling Υ, and ν degrees of freedom. The notation $\mathrm{NiW}(\gamma, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)$ denotes the NiW distribution with statistics (hyperparameters) summarized in $S := (\gamma, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)$. Similarly, $\mathrm{iW}(\boldsymbol{\Lambda}, \nu)$ means the iW distribution with hyperparameters $S := (\boldsymbol{\Lambda}, \nu)$. The notation $\hat{\boldsymbol{z}}_{k|m}$ denotes the estimate of $\boldsymbol{z}$ at time index k given measurements up to time index m.

II. Problem Definition

This paper considers adaptive Bayesian inference for discrete-time linear state-space models

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}_k \boldsymbol{x}_k + \boldsymbol{w}_k, \tag{1a}$$

$$\boldsymbol{y}_k = \boldsymbol{C}_k \boldsymbol{x}_k + \boldsymbol{e}_k, \tag{1b}$$

[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

where $\boldsymbol{x}_k \in \mathbb{R}^n$ is the state at time step k and $\boldsymbol{y}_k \in \mathbb{R}^m$ is the measurement. The model is specified by the state-transition matrix $\boldsymbol{A}_k \in \mathbb{R}^{n\times n}$, measurement matrix $\boldsymbol{C}_k \in \mathbb{R}^{n\times m}$, and the noise sources $\boldsymbol{w}_k$, $\boldsymbol{e}_k$. The process noise $\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{\mu}_{w,k}, \boldsymbol{\Sigma}_{\boldsymbol{w},k})$ is Gaussian distributed with mean $\boldsymbol{\mu}_{\boldsymbol{w},k}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{w}_k,k}$. Similarly, the measurement noise $\boldsymbol{e}_k \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{e},k}, \boldsymbol{\Sigma}_{\boldsymbol{e},k})$ is Gaussian distributed with mean $\boldsymbol{\mu}_{\boldsymbol{e},k}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{e},k}$. The noise sources are individually independent, but $\boldsymbol{w}_k$ and $\boldsymbol{e}_k$ can be dependent. Dependence between $\boldsymbol{w}_k$ and $\boldsymbol{e}_k$ frequently arises in engineering applications, such as inertial navigation, target tracking, or automotive applications [13], [14], often as a consequence of discretization of a continuous-time system.

Denote the parameters of the process and measurement noise with $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, where $\boldsymbol{\mu}_k = [\boldsymbol{\mu}_{w,k} \;\; \boldsymbol{\mu}_{e,k}]^{\mathrm{T}}$ and where $\boldsymbol{\Sigma}_k$ is the (at least partially) unknown covariance matrix

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{w},k} & \boldsymbol{\Sigma}_{\boldsymbol{we},k} \\ \boldsymbol{\Sigma}_{\boldsymbol{ew},k} & \boldsymbol{\Sigma}_{\boldsymbol{e},k} \end{bmatrix}. \tag{2}$$

In its most general formulation, all entities in $\boldsymbol{\theta}_k$ are unknown.

The Bayesian filtering for model (1) with unknown $\boldsymbol{\theta}_k$ amounts to computing $p(\boldsymbol{x}_k, \boldsymbol{\theta}_k | \boldsymbol{y}_{0:k})$. Recursive expressions for the filtering problem consist of a prediction step [2]

$$p(\boldsymbol{x}_k, \boldsymbol{\theta}_k | \boldsymbol{y}_{0:k-1}) = \int p(\boldsymbol{x}_k, \boldsymbol{\theta}_k | \boldsymbol{x}_{k-1}, \boldsymbol{\theta}_{k-1}) \cdot p(\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_{k-1} | \boldsymbol{y}_{0:k-1}) \,\mathrm{d}\boldsymbol{x}_{k-1} \mathrm{d}\boldsymbol{\theta}_{k-1} \tag{3}$$

and a measurement update according to Bayes' rule yielding the filtering posterior

$$p(\boldsymbol{x}_k, \boldsymbol{\theta}_k | \boldsymbol{y}_{0:k}) = \frac{p(\boldsymbol{y}_k, \boldsymbol{x}_k, \boldsymbol{\theta}_k | \boldsymbol{y}_{0:k-1})}{p(\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1})}, \tag{4}$$

where $p(\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1})$ is a normalization constant,

$$p(\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1}) = \int p(\boldsymbol{y}_k | \boldsymbol{x}_k, \boldsymbol{\theta}_k) p(\boldsymbol{x}_k, \boldsymbol{\theta}_k | \boldsymbol{y}_{0:k-1}) \,\mathrm{d}\boldsymbol{x}_k \mathrm{d}\boldsymbol{\theta}_k. \tag{5}$$

The integrations involved in (3)–(5) are in general not analytically tractable, and hence may require a significant amount of computations that may not be available in embedded computing platforms. In the following, we perform approximate inference by using properties of the Student-t distribution and suitable heuristics of the parameter distribution.

III. Noise-Adaptive Filtering by Intermediate Student-t Approximations

The key assumption in our approach is that we can approximate the conditional distribution $p(\boldsymbol{x}_k, \boldsymbol{\theta}_k | \boldsymbol{y}_{0:k})$ in (4) at each time step k as the product of a Student-t distribution and an iW distribution,[1]

$$p(\boldsymbol{x}_k, \boldsymbol{\theta}_k | \boldsymbol{y}_{0:k}) \approx \mathrm{St}(\boldsymbol{x}_k | \hat{\boldsymbol{x}}_{k|k}, \boldsymbol{P}_{k|k}, \nu'_k) \cdot \mathrm{iW}(\boldsymbol{\theta}_k | \boldsymbol{\Lambda}_{k|k}, \nu_{k|k}). \tag{6}$$

For later convenience, we first give some preliminaries and useful results.

[1] In cases where also the mean vector is unknown, we replace the iW with the NiW distribution.

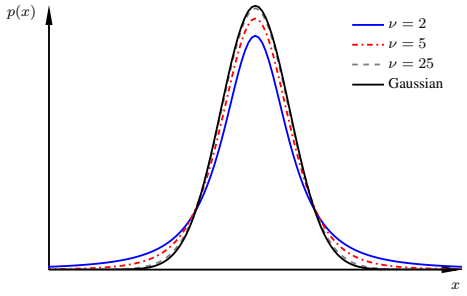

Fig. 1. Illustration of how the Student-t approaches a Gaussian as the degrees of freedom increase.

A. Preliminaries

For further details about the results on Student-t distributions that we outline here, see, for example, [15]. A Student-t density of a random variable $\boldsymbol{x}$ is characterized by the mean $\hat{\boldsymbol{x}}$, the scale matrix Υ, and the scalar degrees of freedom ν, where a lower value of ν results in a heavier tail of the distribution. As the number of degrees of freedom increase, the Student-t approaches the Gaussian distribution (Fig. 1). The probability density function of a Student-t is [12]

$$\mathrm{St}(\boldsymbol{x} | \hat{\boldsymbol{x}}, \Upsilon, \nu) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{(\pi\nu)^{n/2}} \frac{1}{\sqrt{\det(\Upsilon)}} \cdot \left(1 + \frac{1}{\nu}(\boldsymbol{x} - \hat{\boldsymbol{x}})^{\mathrm{T}} \Upsilon^{-1} (\boldsymbol{x} - \hat{\boldsymbol{x}})\right)^{-\frac{n+\nu}{2}}, \tag{7}$$

where the covariance matrix $\mathrm{cov}(\boldsymbol{x})$ is given by

$$\mathrm{cov}(\boldsymbol{x}) = \frac{\nu}{\nu - 2}\Upsilon.$$

For linear transformations of Student-t distributions, the mean and scale matrix are transformed similar to the parameters in the Gaussian case. Two useful properties of conditional Student-t distributions used in this paper are that for partitioned vectors $\boldsymbol{x} = [\boldsymbol{x}_1^{\mathrm{T}} \;\; \boldsymbol{x}_2^{\mathrm{T}}]^{\mathrm{T}}$ where $\boldsymbol{x}_1 \in \mathbb{R}^{n_1}$, $\boldsymbol{x}_2 \in \mathbb{R}^{n_2}$ with the joint density

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathrm{St}\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \middle| \begin{bmatrix} \hat{\boldsymbol{x}}_1 \\ \hat{\boldsymbol{x}}_2 \end{bmatrix}, \begin{bmatrix} \Upsilon_1 & \Upsilon_{12} \\ \Upsilon_{21} & \Upsilon_2 \end{bmatrix}, \nu\right), \tag{8}$$

the marginal density is given by

$$p(\boldsymbol{x}_1) = \mathrm{St}(\boldsymbol{x}_1 | \hat{\boldsymbol{x}}_1, \Upsilon_1, \nu), \tag{9}$$

and the conditional density is given by

$$p(\boldsymbol{x}_1 | \boldsymbol{x}_2) = \mathrm{St}(\boldsymbol{x}_1 | \hat{\boldsymbol{x}}_{1|2}, \Upsilon'_{1|2}, \nu_{1|2}), \tag{10}$$

where

$$\hat{\boldsymbol{x}}_{1|2} = \hat{\boldsymbol{x}}_1 + \Upsilon_{12}\Upsilon_2^{-1}(\boldsymbol{x}_2 - \hat{\boldsymbol{x}}_2), \tag{11a}$$

$$\Upsilon_{1|2} = \Upsilon_1 - \Upsilon_{12}\Upsilon_2^{-1}\Upsilon_{21}, \tag{11b}$$

$$\Upsilon'_{1|2} = \frac{\nu + (\boldsymbol{x}_2 - \hat{\boldsymbol{x}}_2)\Upsilon_2^{-1}(\boldsymbol{x}_2 - \hat{\boldsymbol{x}}_2)^{\mathrm{T}}}{\nu + n_2}\Upsilon_{1|2}, \tag{11c}$$

$$\nu_{1|2} = \nu + n_2. \tag{11d}$$

Note the similarity of (11a) and (11b) with the KF update equations. The NiW distribution as used in this work is defined through a hierarchical distribution according to [5]

$$\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_k, \boldsymbol{\Sigma}_k), \tag{12a}$$

$$\begin{aligned}\boldsymbol{\Sigma}_k &\sim \text{iW}(\nu_k, \boldsymbol{\Lambda}_k) \\ &\propto |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}(\nu_k+d+1)} e^{\left(-\frac{1}{2}\text{tr}(\boldsymbol{\Lambda}_k\boldsymbol{\Sigma}_k^{-1})\right)},\end{aligned} \tag{12b}$$

where d is the dimension of the data and where the definition of the iW distribution is given by (12b).

There are several reasons for approximating the conditional distribution $p(\boldsymbol{x}_k, \boldsymbol{\theta}_k|\boldsymbol{y}_{0:k})$ according to (6). For instance, the iW (NiW) distribution is the conjugate prior to the Gaussian likelihood. That is, for an iW prior $p(\boldsymbol{\theta})$ and a Gaussian likelihood $p(\bar{\boldsymbol{w}}|\boldsymbol{\theta})$, the posterior

$$p(\boldsymbol{\theta}|\bar{\boldsymbol{w}}) \propto p(\bar{\boldsymbol{w}}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{13}$$

is also NiW distributed [5]. Leveraging the conjugacy of the NiW relative to a Gaussian likelihood leads to several algorithmic simplifications. In (12) the unknown $\boldsymbol{\Sigma}_k$ appears on the right-hand side. However, due to the conjugacy of the iW to the Gaussian likelihood it is possible to marginalize out $\boldsymbol{\Sigma}_k$. The one-step prediction of the sufficient statistics can be calculated as [16], [17]

$$\gamma_{k+1} = \frac{1}{\lambda}\gamma_k, \tag{14a}$$

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k, \tag{14b}$$

$$\nu_{k+1} = \lambda\nu_k, \tag{14c}$$

$$\boldsymbol{\Lambda}_{k+1} = \lambda\boldsymbol{\Lambda}_k, \tag{14d}$$

where $\lambda \in [0, 1]$ produces exponential forgetting and determines how fast the parameters are allowed to change. In the case of iW priors, (14a) and (14b) can be ignored. Furthermore, for an (N)iW prior, the predictive distribution of the data $\bar{\boldsymbol{w}}$ is Student-t distributed with $\nu_k - d + 1$ degrees of freedom and scale matrix

$$\tilde{\boldsymbol{\Lambda}}_k = \frac{1}{\nu_k - d + 1}\boldsymbol{\Lambda}_k.$$

This property is important since it forms the basis for why our proposed method makes the assumption of Student-t distributed state posterior.

B. Approximation of State Posterior by Student-t

The starting point for updating the state estimate is a Student-t assumption (7) of the state filtering distribution as

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{0:k}) = \text{St}(\boldsymbol{x}_k|\hat{\boldsymbol{x}}_{k|k}, \boldsymbol{P}_{k|k}, \nu'_k). \tag{15}$$

Time Update: To obtain the prediction density $p(\boldsymbol{x}_{k+1}|\boldsymbol{y}_{0:k})$ of the state, we write

$$p(\boldsymbol{x}_{k+1}|\boldsymbol{y}_{0:k}) = \int p(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k|\boldsymbol{y}_{0:k})p(\boldsymbol{x}_k|\boldsymbol{y}_{0:k})\,\mathrm{d}\boldsymbol{x}_k. \tag{16}$$

The joint density $p(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k|\boldsymbol{y}_{0:k})$ is a product of Student-t densities. If we assume that the joint density can be written as a joint Student-t density with a common degree of freedom η_k, for uncorrelated noise processes we obtain

$$\begin{aligned}&p(\boldsymbol{x}_k, \boldsymbol{w}_k|\boldsymbol{y}_{0:k}) \approx \\ &\quad \text{St}\left(\begin{bmatrix}\boldsymbol{x}_k \\ \boldsymbol{w}_k\end{bmatrix}\middle|\begin{bmatrix}\hat{\boldsymbol{x}}_{k|k} \\ \hat{\boldsymbol{\mu}}_{\boldsymbol{w}_k,k}\end{bmatrix}, \begin{bmatrix}\boldsymbol{P}_{k|k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\boldsymbol{w},k}\end{bmatrix}, \eta_k\right),\end{aligned} \tag{17}$$

from which it follows by rules of linear transformations of Student-t distributed vectors that

$$\begin{aligned}&p(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k|\boldsymbol{y}_{0:k}) \approx \\ &\text{St}\left(\begin{bmatrix}\boldsymbol{x}_k \\ \boldsymbol{x}_{k+1}\end{bmatrix}\middle|\begin{bmatrix}\hat{\boldsymbol{x}}_{k|k} \\ \hat{\boldsymbol{x}}_{k+1|k}\end{bmatrix}, \begin{bmatrix}\boldsymbol{P}_{k|k} & \boldsymbol{P}_{k|k}\boldsymbol{A}_k^{\mathrm{T}} \\ \boldsymbol{A}_k\boldsymbol{P}_{k|k} & \boldsymbol{P}_{k+1|k}\end{bmatrix}, \eta_k\right),\end{aligned} \tag{18}$$

which leads to the filter updates

$$\hat{\boldsymbol{x}}_{k+1|k} = \boldsymbol{A}_k\hat{\boldsymbol{x}}_{k|k}, \tag{19a}$$

$$\boldsymbol{P}_{k+1|k} = \boldsymbol{A}_k\boldsymbol{P}_{k|k}\boldsymbol{A}_k^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{w},k}, \tag{19b}$$

of the mean and scale matrix. Instead, for dependent noise processes it follows from (10), (11) that the time update is

$$\hat{\boldsymbol{x}}_{k+1|k} = \boldsymbol{A}_k\hat{\boldsymbol{x}}_{k|k} + \hat{\boldsymbol{\mu}}_{w,k} + \boldsymbol{\Sigma}_{\boldsymbol{we},k}\boldsymbol{\Sigma}_{\boldsymbol{e},k}^{-1}(\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k}), \tag{20a}$$

$$\begin{aligned}\boldsymbol{P}'_{k+1|k} &= \bar{\boldsymbol{A}}_k\boldsymbol{P}_{k|k}\bar{\boldsymbol{A}}_k^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{w},k} \\ &\quad - \boldsymbol{\Sigma}_{\boldsymbol{we},k}\boldsymbol{\Sigma}_{\boldsymbol{e},k}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{ew},k},\end{aligned} \tag{20b}$$

$$\boldsymbol{P}_{k+1|k} = \frac{\eta_k + (\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k})\boldsymbol{\Sigma}_{\boldsymbol{e},k}^{-1}(\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k})^{\mathrm{T}}}{\eta_k + n_2}\boldsymbol{P}'_{k+1|k}, \tag{20c}$$

$$\eta'_k = \eta_k + m, \tag{20d}$$

where $\bar{\boldsymbol{A}}_k = \boldsymbol{A}_k - \boldsymbol{\Sigma}_{\boldsymbol{we},k}\boldsymbol{\Sigma}_{\boldsymbol{e},k}^{-1}\boldsymbol{C}_k$. Hence, (19) ((20) for dependent noise) provides the parameters of

$$p(\boldsymbol{x}_{k+1}|\boldsymbol{y}_{0:k}) = \text{St}(\boldsymbol{x}_k|\hat{\boldsymbol{x}}_{k|k}, \boldsymbol{P}_{k+1|k}, \eta'_k). \tag{21}$$

The time-update equations (19) and (20) are very similar to the KF time updates for independent and dependent noise processes, respectively. However, the scale matrix update (20c) also contains a factor that is quadratically dependent on the measurement.

Remark 1: The degree of freedom η_k in (18) can be chosen in several ways. A simple way that preserves the heaviest tails (and hence makes the algorithm more robust to outliers) is to choose $\eta_k = \min(\nu_k, \nu'_k)$.

Measurement Update: For the measurement update, the Student-t distribution (21) must be combined with $p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{y}_{0:k-1})$ by the expansion

$$p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{y}_{0:k-1}) = \int p(\boldsymbol{y}_k|\boldsymbol{\theta}_k, \boldsymbol{x}_k)p(\boldsymbol{\theta}_k|\boldsymbol{x}_k, \boldsymbol{y}_{0:k-1})\,\mathrm{d}\boldsymbol{\theta}_k. \tag{22}$$

Since the integrand of (22) by assumption is a product of a Gaussian distribution and an (N)iW distribution, $p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{y}_{0:k-1})$ is Student-t distributed. Hence, the measurement update, similar to the time update, consists of the

product of two Student-t densities. By approximating the joint density between $\boldsymbol{x}_k$ and $\boldsymbol{e}_k$ as a joint Student-t density

$$p(\boldsymbol{x}_k, \boldsymbol{e}_k|\boldsymbol{y}_{0:k-1}) \approx \mathrm{St}\left(\begin{bmatrix}\boldsymbol{x}_k\\ \boldsymbol{e}_k\end{bmatrix}\middle|\middle|\begin{bmatrix}\hat{\boldsymbol{x}}_{k|k-1}\\ \hat{\boldsymbol{\mu}}_k\end{bmatrix}, \begin{bmatrix}\boldsymbol{P}_{k|k-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\boldsymbol{e},k}\end{bmatrix}, \eta_k''\right), \quad (23)$$

we obtain the joint density in Bayes' rule (4) as

$$p(\boldsymbol{x}_k, \boldsymbol{y}_k|\boldsymbol{y}_{0:k-1}) \approx \mathrm{St}\left(\begin{bmatrix}\boldsymbol{x}_k\\ \boldsymbol{y}_k\end{bmatrix}\middle|\middle|\begin{bmatrix}\hat{\boldsymbol{x}}_{k|k-1}\\ \hat{\boldsymbol{y}}_k\end{bmatrix}, \begin{bmatrix}\boldsymbol{P}_{k|k-1} & \boldsymbol{P}_{k|k-1}\boldsymbol{C}_k^{\mathrm{T}}\\ \boldsymbol{C}_k\boldsymbol{P}_{k|k-1} & \boldsymbol{S}_k\end{bmatrix}, \eta_k''\right). \quad (24)$$

Utilizing the results on conditional densities (10), (11), the measurement update step becomes

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \hat{\boldsymbol{\mu}}_{\boldsymbol{e},k} + \boldsymbol{K}_k(\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1}), \quad (25a)$$

$$\boldsymbol{P}'_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}_k^{\mathrm{T}}, \quad (25b)$$

$$\boldsymbol{P}_{k|k} = \frac{\eta'_k + (\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1})\boldsymbol{S}_k^{-1}(\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1})^{\mathrm{T}}}{\eta'_k + m}\boldsymbol{P}'_{k|k}, \quad (25c)$$

$$\nu'_k = \eta''_k + m, \quad (25d)$$

with $\boldsymbol{S}_k = \boldsymbol{C}_k\boldsymbol{P}_{k|k-1}\boldsymbol{C}_k^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{e},k}$. Thus, (25) gives the state filtering posterior (15).

Similar to the prediction step, the parameter η_k'' in (23) needs to be chosen. One possibility is $\eta_k'' = \nu_k - d + 1$. This choice is intuitive since the predictive distribution of the (N)iW is a Student-t with degrees of freedom $\nu_k - d + 1$, where ν_k increases with time. In both the time update and measurement update, limiting the degree of freedom gives increased robustness in the algorithm. However, it also prevents the algorithm from approaching a Gaussian distribution. With $\eta_k'' = \nu_k - d + 1$ and for a forgetting factor $\lambda < 1$, the degree of freedom will converge to an equilibrium. At equilibrium, $\nu_k = \nu_{k-1}$ and if (14c) is used for predicting the degree of freedom it follows that $\lim_{k\to\infty} \nu_k = 1/(1-\lambda)$.

C. Parameter Update

For the update of the parameter distribution, we employ fixed-point iterations [7], [11]. The measurement noise parameters can be updated as

$$\gamma_{k|k} = \frac{\gamma_k}{1+\gamma_k}, \quad (26a)$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{e},k|k} = \hat{\boldsymbol{\mu}}_{\boldsymbol{e},k} + \gamma_k\boldsymbol{e}_k, \quad (26b)$$

$$\nu_{k|k} = \nu_k + 1, \quad (26c)$$

$$\boldsymbol{\Lambda}_{\boldsymbol{e},k|k} = \boldsymbol{\Lambda}_{\boldsymbol{e},k} + \boldsymbol{C}_k\boldsymbol{P}_{k|k}\boldsymbol{C}_k^{\mathrm{T}} + \frac{1}{1+\gamma_k}\boldsymbol{e}_k\boldsymbol{e}_k^{\mathrm{T}}, \quad (26d)$$

$$\boldsymbol{e}_k = \boldsymbol{y}_k - \boldsymbol{C}_k\hat{\boldsymbol{x}}_{k|k} - \hat{\boldsymbol{\mu}}_{\boldsymbol{e},k} \quad (26e)$$

where (26a) and (26b) are ignored and $\gamma_k = 0$ if the iW is used as a parameter prior. For unknown process-noise parameters, the VB update of the covariance involves [18]

$$\boldsymbol{\Lambda}_{k|k} = \boldsymbol{\Lambda}_k + \mathbb{E}((\boldsymbol{x}_k - \boldsymbol{A}_k\boldsymbol{x}_{k-1})(\boldsymbol{x}_k - \boldsymbol{A}_k\boldsymbol{x}_{k-1})^{\mathrm{T}}). \quad (27)$$

Hence, the process noise statistics can be updated by

$$\gamma_{k|k} = \frac{\gamma_k}{1+\gamma_k}, \quad (28a)$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{w},k|k} = \hat{\boldsymbol{\mu}}_{\boldsymbol{w},k} + \gamma_k\boldsymbol{z}_k, \quad (28b)$$

$$\nu_{k|k} = \nu_k + 1, \quad (28c)$$

$$\boldsymbol{\Lambda}_{\boldsymbol{w},k|k} = \boldsymbol{\Lambda}_{\boldsymbol{w},k} + \boldsymbol{P}_{k|k} - \boldsymbol{A}_k\boldsymbol{P}_{k|k-1}\boldsymbol{A}_k^{\mathrm{T}} + \frac{1}{1+\gamma_k}\boldsymbol{z}_k\boldsymbol{z}_k^{\mathrm{T}}, \quad (28d)$$

$$\boldsymbol{z}_k = \hat{\boldsymbol{x}}_{k|k} - \hat{\boldsymbol{x}}_{k|k-1} - \hat{\boldsymbol{\mu}}_{\boldsymbol{w},k} \quad (28e)$$

where the first two equations are ignored and $\gamma_k = 0$ for known mean. If the noise processes are dependent, the prediction of the state and scale matrix are done using (20). The case for fully unknown parameters can be updated similarly but is omitted due to lack of space. One iteration of the algorithm is summarized in Algorithm 1 for the special case of unknown measurement noise covariance and independent noise, and with a predetermined number J (typically $J \approx 1$) of fixed-point iterations of the parameter update.

Algorithm 1 Pseudo-code of the estimation algorithm

1: Predict state posterior using (19).
2: Predict noise statistics using (14c), (14d).
3: Set $\hat{\boldsymbol{y}}_k = \boldsymbol{C}_k\boldsymbol{x}_{k|k-1}$, $\hat{\boldsymbol{x}}_{k|k}^{(0)} = \hat{\boldsymbol{x}}_{k|k-1}$, $\boldsymbol{P}_{k|k}^{(0)} = \boldsymbol{P}_{k|k-1}$, $\nu_{k|k} = \nu_k + 1$, $\boldsymbol{\Sigma}_{\boldsymbol{e},k}^{(0)} = \boldsymbol{\Lambda}_{\boldsymbol{e},k}/(\nu_{k|k} - m - 1)$.
4: **for** $j \leftarrow 1$ **to** J **do**
5: Update state posterior using (25) resulting in $\boldsymbol{x}_{k|k}^{(j)}, \boldsymbol{P}_{k|k}^{(j)}$.
6: Update parameters using (26d) with $\boldsymbol{x}_{k|k}^{(j)}, \boldsymbol{P}_{k|k}^{(j)}$.
7: Set $\boldsymbol{\Sigma}_{\boldsymbol{e},k}^{(j)} = \boldsymbol{\Lambda}_{\boldsymbol{e},k|k}^{(j)}/(\nu_{k|k} - m - 1)$.
8: **end for**
9: Set $\hat{\boldsymbol{x}}_{k|k} = \boldsymbol{x}_{k|k}^{(J)}$, $\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k}^{(J)}, \boldsymbol{\Sigma}_{\boldsymbol{e},k|k} = \boldsymbol{\Lambda}_{\boldsymbol{e},k|k}^{(J)}/(\nu_{k|k} - m - 1)$.

IV. Numerical Results

We evaluate the proposed noise-adaptive Student-t filter on a tracking example using a generic motion model, with outlier-corrupted measurement and process noise. We consider the problem of tracking an object moving in the Cartesian two-dimensional plane while estimating the unknown measurement-noise covariance. The state vector consists of the position and velocity vector of the object. The object moves according to a constant-velocity model [19]. The sampling time is $T_s = 1$ s and the simulation lasts for $T = 4000$ s. The motion model is linear

$$\boldsymbol{x}_{k+1} = \mathrm{diag}(\boldsymbol{F}, \boldsymbol{F})\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{w}_k, \quad \boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}), \quad (29)$$

where $\mathrm{diag}(\cdot)$ is a diagonal matrix with the arguments on the diagonal, $\boldsymbol{Q} = \mathrm{diag}(\sigma_w^2, \sigma_w^2)$, and

$$\boldsymbol{F} = \begin{bmatrix}1 & T_s\\ 0 & 1\end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix}T_s^2/2 & 0\\ T_s & 0\\ 0 & T_s^2/2\\ 0 & T_s\end{bmatrix}.$$

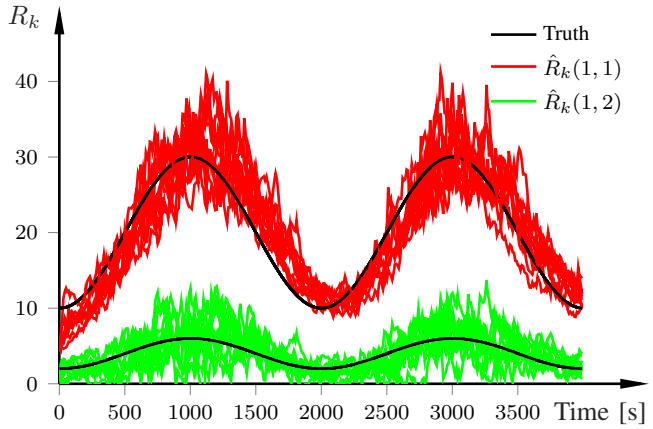

Fig. 2. Estimated values of the first row of the measurement covariance matrix for 10 Monte-Carlo simulations, using a forgetting factor $\lambda = 0.98$.

The measurements are generated by

$$\boldsymbol{y}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \boldsymbol{x}_k + \boldsymbol{e}_k. \tag{30}$$

The measurement noise is nominally distributed as $\boldsymbol{e}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_k)$, where

$$\boldsymbol{R}_k = \sigma_e^2 \left(2 - \cos\left(\frac{4\pi k}{T} \right) \right) \cdot \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}. \tag{31}$$

The nominal noise parameters are given by $\sigma_w^2 = 3, \sigma_e^2 = 2$. The prior for the initial state is assumed Gaussian with mean and covariance given by $\boldsymbol{x}_0 = [0\ 5\ 0\ 5]^{\mathrm{T}}$, $\boldsymbol{P}_0 = \mathrm{diag}([30^2\ 30^2\ 30^2\ 30^2]^{\mathrm{T}})$, the initial degrees of freedom and scale matrix are $\nu_0 = 5$, $\boldsymbol{\Lambda}_0 = (\nu_0 - 3)\boldsymbol{R}_0$, and we use $J = 1$ number of fixed-point iterations. Fig. 2 displays the estimates of the first row of the measurement covariance matrix for 10 Monte-Carlo simulations, $\lambda = 0.98$, and with the time-varying noise parameters (31).

To assess the performance of the proposed filter against other filters, we simulate three different scenarios, each with 1000 Monte-Carlo simulations. In the first scenario, we execute the filters using the nominal noise parameters given by (31). In the second scenario, we assume that the measurement noise is corrupted by outliers. We simulate an outlier as generated by a Gaussian distribution with the covariance matrix $100\boldsymbol{R}_k$. In this scenario, the measurement noise is generated from

$$\boldsymbol{e}_k = \begin{cases} \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_k) & \text{with probability } 0.995 \\ \mathcal{N}(\boldsymbol{0}, 100\boldsymbol{R}_k) & \text{with probability } 0.005 \end{cases}$$

In the third scenario we also assume that the process noise is subject to outliers, in which case the process noise is generated from

$$\boldsymbol{w}_k = \begin{cases} \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}) & \text{with probability } 0.995 \\ \mathcal{N}(\boldsymbol{0}, 100\boldsymbol{Q}) & \text{with probability } 0.005 \end{cases}$$

The third scenario is relevant in situations when sensor measurements are used as inputs to the motion model.

We compare the proposed Algorithm 1 against

- KF1: A KF that knows the nominal covariance matrices of both the process and measurement noise, but does not know anything about the outliers.
- KF2: a KF that knows the nominal covariance matrices and the instances at which outliers occur and can adjust its process and measurement noise accordingly. This filter solves the Bayesian filtering recursions and is the best available algorithm, but infeasible to implement in practice because of its assumption of known outliers. This filter serves as benchmark for the given problem.
- ST: The student-t filter in [12], which knows the nominal covariance (31) but is not aware of the outliers, similar to KF1. This method is designed to be robust to outliers but assumes known noise covariances, and should therefore always perform better than our proposed filter when outliers are present.
- VB: The adaptive KF proposed in [20], which does online covariance estimation. This filter assumes a Gaussian state posterior at each step.
- ST-A: The adaptive student-t filter proposed in this paper and given by Algorithm 1, which is designed to be robust to outliers and does not assume knowledge of the covariances, but rather estimates them concurrently.

In our proposed ST-A, we set the number of degrees of freedom in the prediction step (18) to $\eta_k = \nu_0 = 5$. We compare the different filters in terms of the average of the time-averaged root-mean square error (ARMSE) of the position estimates over all Monte-Carlo simulations, where the RMSE for Monte-Carlo simulation j is given by

$$\mathrm{RMSE}(j) = \sqrt{\left(\frac{1}{T+1} \sum_{k=0}^{T} \| \boldsymbol{C}(\hat{\boldsymbol{x}}_{k|k}^j - \boldsymbol{x}_k^j) \|^2 \right)}, \tag{32}$$

in which $\hat{\boldsymbol{x}}_{k|k}^j$ denotes the filtered state estimate (mean) for Monte-Carlo simulation j at time k. The RMSE for the covariance matrix is taken as the square root of the average Frobenius norm square normalized by the number of elements [17],

$$E_R(j) = \left(\frac{1}{m^2(T+1)} \sum_{k=0}^{T} \mathrm{Tr}\left((\hat{\boldsymbol{R}}_{k|k}^j - \boldsymbol{R}_k^j)^2 \right) \right)^{1/4}. \tag{33}$$

Tables I–III show the three different scenarios The first three filters already know the covariance matrices and the corresponding error terms are not given in the tables. Without outliers (Table I) KF1 and KF2 (which are the same in this scenario) perform best, as expected. When comparing the proposed ST-A (Algorithm 1) with VB, our method performs slightly better, although the difference is small. It is interesting to note that without outliers, the proposed ST-A performs better than ST in terms of ARMSE, despite that ST knows the covariance matrices. The reason is that ST retains a small degree of freedom throughout, whereas our method only retains the small degree of freedom in the prediction step. Hence, as time evolves the degrees of freedom of the measurement update step in our method will converge to a value determined by the forgetting factor (c.f.

TABLE I

UNKNOWN MEASUREMENT NOISE COVARIANCE WITHOUT OUTLIERS. THE TIME-AVERAGED RMSES AND THE RESPECTIVE STANDARD DEVIATIONS ARE SHOWN.

Filter	ARMSE	AE_R
KF1	3.671 ± 0.064	
KF2	3.671 ± 0.064	
ST	3.851 ± 0.076	
VB	3.685 ± 0.061	1.727 ± 0.060
ST-A	3.680 ± 0.062	1.679 ± 0.068

TABLE II

UNKNOWN MEASUREMENT NOISE COVARIANCE WITH MEASUREMENT NOISE OUTLIERS. THE TIME-AVERAGED RMSES AND THE RESPECTIVE STANDARD DEVIATIONS ARE SHOWN.

Filter	ARMSE	AE_R
KF1	4.511 ± 0.264	
KF2	3.702 ± 0.065	
ST	4.281 ± 0.151	
VB	5.997 ± 1.526	7.649 ± 3.562
ST-A	4.448 ± 0.246	3.450 ± 0.556

Sec. III-B), which better approximates the true (Gaussian) noise distribution than ST.

For the case of measurement outliers (Table II), the Student-t filter ST in [12] performs better than KF1. Similarly, the proposed method ST-A outperforms VB thanks to its propagation of Student-t distributions instead of the Gaussian assumption in VB. For instance, the RMSE of the covariance estimates are decreased by more than 50% when comparing ST-A with VB. Despite not knowing the measurement covariance and hence needing to estimate it, ST-A achieves performance very close to ST, which knows the nominal covariance matrices. Moreover, the proposed ST-A performs better than KF1 even though KF1 knows the nominal covariance matrices. The inherent robustness to disturbances in our proposed method is further pronounced when also incorporating process-noise outliers, whose results can be seen in Table III. With outliers in both noise sources, the position RMSE is decreased by more than 30% when comparing ST-A and VB.

V. CONCLUSION

We presented an approximate method for noise-adaptive filtering robust to outliers in the noise processes. By leveraging properties of the predictive distributions of the NiW (iW) distribution, we can model the joint state and parameter posterior at each step as the product of a Student-t distribution and an NiW distribution, which leads to analytic, although approximate, expressions. Since we propagate the state posterior using a Student-t approximation, the algorithm simplifies due to the connection to the NiW distribution. A Monte-Carlo evaluation on a benchmark example showed improvements over a recent adaptive filter, as well as robustness of the method to outliers. The results were obtained using synthetic data. It is future work to evaluate and refine the method for applications on real-world data.

TABLE III

UNKNOWN MEASUREMENT NOISE COVARIANCE WITH MEASUREMENT AND PROCESS NOISE OUTLIERS. THE TIME-AVERAGED RMSES AND THE RESPECTIVE STANDARD DEVIATIONS ARE SHOWN.

Filter	ARMSE	AE_R
KF1	5.101 ± 0.321	
KF2	3.718 ± 0.066	
ST	4.596 ± 0.188	
VB	6.949 ± 2.196	8.954 ± 5.144
ST-A	4.993 ± 0.292	4.387 ± 0.565

REFERENCES

[1] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall, 2000.

[2] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge University Press, 2013, vol. 3.

[3] C. Hide, T. Moore, and M. Smith, "Adaptive Kalman filtering for low-cost INS/GPS," *J. Navigation*, vol. 56, no. 1, pp. 143–152, 2003.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. NJ, USA: Springer-Verlag New York, 2006.

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

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

[7] 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.

[8] 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.

[9] K. Berntorp and S. Di Cairano, "Process-noise adaptive particle filtering with dependent process and measurement noise," in *IEEE Int. Conf. Decision and Control*, Las Vegas, NV, Dec. 2016.

[10] R. Piche, S. Särkkä, and J. Hartikainen, "Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate student-t distribution," in *IEEE Int. Workshop Machine Learning for Signal Processing*, Santander, Spain, Sep. 2012.

[11] G. Agamennoni, J. I. Nieto, and E. M. Nebot, "Approximate inference in state-space models with heavy-tailed noise," *IEEE Trans. Signal Processing*, vol. 60, no. 10, 2012.

[12] M. Roth, T. Ardeshiri, E. Özkan, and F. Gustafsson, "Robust Bayesian filtering and smoothing using student's t distribution," *ArXiv e-prints*, Mar. 2017.

[13] K. Berntorp, "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.

[14] K. Berntorp and S. Di Cairano, "Tire-stiffness and vehicle-state estimation based on noise-adaptive particle filtering," *IEEE Trans. Contr. Syst. Technol.*, vol. PP, no. 99, pp. 1–15, 2018, in press.

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

[16] V. Peterka, "Bayesian approach to system identification," in *in Trends and Progress in System Identification, P. Eykhoff, Ed*. Citeseer, 1981.

[17] T. Ardeshiri, E. Özkan, U. Orguner, and F. Gustafsson, "Approximate Bayesian smoothing with unknown process and measurement noise covariances," *IEEE Signal Processing Lett.*, vol. 22, no. 12, pp. 2450–2454, 2015.

[18] ——, *Variational iterations for smoothing with unknown process and measurement noise covariances*. Linköping University Electronic Press, 2015.

[19] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[20] S. Särkkä and J. Hartikainen, "Non-linear noise adaptive Kalman filtering via variational Bayes," in *IEEE Int. Workshop Machine Learning for Signal Processing*, Southampton, UK, Sep. 2013.