FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of master's thesis

Title:	Bayesian filtering of state-space models with unknown covariance matrices
Student:	Bc. Tomáš Vlk
Supervisor:	Ing. Kamil Dedecius, Ph.D.
Study program:	Informatics
Branch / specialization:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	until the end of summer semester 2021/2022

Instructions

Bayesian sequential estimation of unknown states of the state-spacemodels - also known as filtering - is generally a well-established discipline. Various types of the Kalman filters are predominantly used if the process and measurement noise variables are independent and identically (normally) distributed, and their covariance matrices are known. However, this knowledge is not always available. The aim of the thesis is focus on this issue and to propose a method that allows simultaneous filtering of the state variable and estimation of the observation noise covariance matrix. Inspiration can be found in [1].

–

[1] Y. Huang, Y. Zhang, Z. Wu, N. Li, and J. Chambers, "A Novel Adaptive Kalman Filter With Inaccurate Process and Measurement Noise Covariance Matrices," IEEE Trans. Automat. Contr., vol. 63, no. 2, pp. 594-601, Feb. 2018.

Electronically approved by Ing. Karel Klouda, Ph.D. on 18 January 2021 in Prague.

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Master's thesis

Bayesian filtering of state-space models with unknown covariance matrices

Bc. Tomáš Vlk

Department of Applied Mathematics
Supervisor: Ing. Kamil Dedecius, Ph.D.

May 5, 2021

Acknowledgements

First and foremost I have to express my gratitude to my supervisor Ing. Kamil Dedecius Ph.D. for his guidance and invaluable suggestion on the topis, as well as his willingness and patience and all the valuable time that he spent answering my relentless questions, which there were many.

Second, I must thank my family, for their support throughout my studies, without which I could not have made it.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as a school work under the provisions of Article 60 (1) of the Act.

In Prague on May 5, 2021 . . . . . .

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis

Vlk, Tomáš. *Bayesian filtering of state-space models with unknown covariance matrices.* Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2021.

Abstrakt

Tato závěrečná práce se věnuje problému distribuovaného Baysovského sekvenčního odhadu neznámých stavů stavových modelů s neznámými kovariačními maticemi šumu procesu i měření. Tento problém je velmi častý v reálných případech, kde specifické informace o kovariačních maticích šumu pro jednotlivé senzory nemusí být dostupné. Řešení navržené v této práci je postavené na teorii variačního Bayese, ta je využitá jak k odhadu stavů, tak i k odhadu kovariační matice šumu měření. Z důvodu zlepšení sdílíme jak měření, tak i posteriorní odhady mezi sousedními uzly v síti. Práce zároveň ukazuje způsob optimalizace kovariační matice procesního šumu.

Klíčová slova Difuzní sítě, difuzní strategie, odhad stavů, Kalman filtrace, variační Bayesovské metody

Abstract

This thesis explores the problem of distributed Bayesian sequential estimation of unknown state-spacemodels with unknown processes and measurement noise covariance matrices. This is a frequent problem in real-world scenarios, where the information about noise covariance matrices for specific sensors may not be available. The solution proposed in this thesis is built upon the variational Bayesian paradigm, which is used for the estimation of the states, as well as the unknown measurement noise covariance matrix. From performance improvements, the measurements and posterior estimates are shared between the adjacent node in the network. It also shows a way of optimizing the process noise covariance matrix.

Keywords Diffusion network, diffusion strategy, state estimation, Kalman filtering, variational Bayesian methods

Contents

List of Figures

List of Tables

Introduction

The Bayesian filtering of state-space models with the unknown process and measurement noise covariance matrices is a problem where we aim to estimate a variable in time, whose measurements are affected by noise and are therefore imprecise. This problem has many real-world use cases, ranging from medicine all to way to telecommunications. Since the sensors that can provide measurements are getting increasingly cheaper, many devices have more than one way to measure something. Hence the prospect of filtering in a distributed setting is more and more appealing.

This thesis is based upon the Article Variational diffusion Kalman filtering with unknown process and measurement noise covariance matrices by Tomáš Vlk and Kamil Dedecius, which will be sent to the IEEE Transactions on Signal Processing.

Goals of the thesis

The primary goals of this thesis could be summarized in the following way:

- Explore the issue of filtering when we do not know the exact process or measurement noise covariance matrices.
- Propose a method that would allow simultaneous filtering of the state variable and estimation of the observed noise covariance matrix.

Personal motivation

My first encounter with filtering was when I had the opportunity to work on tracking surrounding cars using measurements from a car-mounted camera. I was tasked with implementing the filtering of the signals provided and my colleagues at the time introduced me to the Kalman filter. Ever since I have had a great interest in filtering topics and the Kalman filter in particular. Hence I

wanted to discover more and expand my knowledge to this background. I had the incredible opportunity to attend astonishing subjects that were focused on this topic, namely the Bayesian machine learning and the Statistical analysis of time series, where I could expand my overview of the topic as well as fill in some gaps in my understanding of the subject. After all of this, I still wanted to know what is the state-of-the-art in this field and I felt that choosing it for a topic of my thesis was the ideal choice. Thankfully I must say, that I have never had any form of regrets about this decision.

Structure of the work

We have chosen to arrange this work into four main chapters. Chapter 1, is used for the introduction of necessary concepts and ideas, we build upon in the later parts of the thesis. We then proceed to show some of the current state-of-the-art methods in this particular field and provide some comparison with our work in Chapter 2. Our proposed solution to the thesis goal is shown in Chapter 3, where we first show our approach to non-distributed setting, and then in a later section of the same chapter, we show how to adapt it for distributed setting. Finally, in Chapter 4 we show the results of our proposed method and compare them with other state-of-the-art approaches.

Chapter 1

Background

In the following chapter, we provide a short introduction to some basic concepts and prerequisites needed for understanding the later parts of this work. This will range through various topics, at first focusing on the basics of Bayesian inference. Later, we will build upon those ideas with the concept of the Kalman filter and show an example of its usage. Last but not least, we will show how to solve the issue of intractable posteriors by using variational Bayesian inference.

1.1 Bayesian inference

In this section, we explain the concept of Bayesian inference that will be needed for further understanding of this work. Most of this chapter is inspired by [1].

1.1.1 Bayes' theorem

First the most basic concept, the Bayes' theorem. This is one of the fundamental concepts in statistics, since the late 18 century, when it was discovered by Thomas Bayes, even though the publication was published posthumously. The exact formulation of the Bayes' theorem is the following,

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}, \tag{1.1}$$

where A and B are events, $p(B) \neq 0$, $p(A|B)$ is a conditional probability of event A given event B and $p(A)$, $p(B)$ is probability of event A, B respectively.

For our case of the Bayesian interference, we will call the terms of the Bayes' theorem in the following way:

- A could be any hypothesis that could be affected by the data B.
- $p(A)$ is the prior probability of the hypothesis A, before the data B, that is the current evidence of development is observed.

- $p(A|B)$ is the posterior probability of A given B.
- $p(B|A)$ is a probability of B given A, usually called likelihood.
- $p(B)$ is mostly called marginal probability.

However, for us, the most useful variant of the Bayes' theorem will be the following. Assume that y and θ are random variables with the following probability density functions[1] $f(y|\theta)$ and $f(\theta)$. Then we can express them by using the Bayes' theorem,

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}, \tag{1.2}$$

where the $f(\theta|y)$ is a posterior density of θ, $f(\theta)$ is the prior density of θ, $f(y)$ is marginal density of observations and finally the $f(y|\theta)$ is the likelihood of observations. Since $f(y)$ purpose is only as a normalizing constant, hence we can write rewrite the Equation 1.2 in proportional form in the following way,

$$f(\theta|y) \propto f(y|\theta)f(\theta). \tag{1.3}$$

1.1.2 Conjugate priors

Based on the Bayes' theorem shown in Section 1.1.1, we would like to have the posterior distribution in the same family of probability distributions as the prior distribution. Unfortunately, this behavior is not guaranteed nor common. For this to happen, both posterior and prior distributions must be what are called conjugate distributions. Then the prior distribution will be called conjugate prior.

For our case, we will use distributions from the exponential family, as established in [2], that can be defined in the following way,

Definition 1.1.1 (Exponential family)**.** A family $\{F_\theta\}$ of distributions of a random variable y_t parameterized by a scalar or multivariate parameter θ is said to form an exponential family if the probability density function can be written in the form

$$f(y|\theta) = \exp\{\eta(\theta)^T T_\theta(y) - B(\theta)\}h(y), \tag{1.4}$$

where $\eta(\theta)$ is natural parameter, $T_\theta(y)$ is the sufficient statistic encompassing all information necessary for the estimation of θ, $B(\theta)$ is the log-normalizing function, and $h(x)$ is the base measure.

From [3] we know that the conjugate prior distribution of exponential family distribution is crucial for tractability of the Bayesian update. Hence we will define the conjugate prior distribution for θ.

[1] Usually abbreviated as pdf

Definition 1.1.2 (Conjugate prior distribution for θ)**.** Let us suppose that we have $f(y|\theta)$, that is an exponential family distribution in a way defined by Definition 1.1.1. Assuming that $\pi(\theta)$ is conjugate to $f(y|\theta)$, then probability density function of $f(y|\theta)$ has to have the following form,

$$\pi(\theta) = \exp\{\eta(\theta)^T \Xi_\theta^- - \nu^- B(\theta)\} g(\theta), \tag{1.5}$$

where $g(\theta)$ is a known function, $B(\theta)$ coincides with the log-normalization function of $f(y|\theta)$, the hyperparameter Ξ_θ^- has the same dimensions as $T_\theta(y)$. There is a possibility of not needing the hyperparameter $\nu^- > 0$, if $B(\theta) = 1$ for every θ. There is also a slight possibility of parameter $\nu_- > 0$ being absorbed by Ξ_θ^-.

Just for the sake of completeness, we have shown some of the other conjugate priors for the most common distributions in Table 1.1.

Posterior	Model parameters	Conjugate prior
Binomial	p probability	Beta
Poisson	λ (rate)	Gamma
Multinomial	p (probability vector), k (number of categories)	Dirichlet
Normal with known variance σ^2	μ (mean)	Normal
Multivariate normal with known covariance matrix Σ	μ (mean vector)	Multivariate normal
Multivariate normal with known mean μ	Σ (covariance matrix)	Inverse-Wishart
Multivariate normal	μ(mean vector), Σ (covariance matrix)	normal-inverse-Wishart

Table 1.1: Distributions and their conjugate priors

1.1.3 Example

In this example, we will show how to estimate the mean μ of the normal distribution with a known variance σ^2 using the conjugate prior normal distribution. It aims to show the application of the Bayes' theorem and the conjugate priors, as shown in Section 1.1.1 and 1.1.2 respectively.

We will assume that we have a scalar data named Y_t that gets an update in discrete steps $t = 1, 2, \ldots$ and is approximately modeled by

$$Y_t \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+, \tag{1.6}$$

where μ is the mean, and σ^2 is the variance. As stated previously, we will estimate the unknown μ given that σ^2 is known.

To use the Bayesian update, we have to rewrite the model probability density function[2] as

$$\begin{aligned} f(y_t|\theta) = f(y_t|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(y_t-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(y_t^2 - 2y_t\mu + \mu^2)\right\}. \end{aligned} \tag{1.7}$$

As established in Section 1.1.2, the conjugate prior for the normal distribution with a known variance σ^2 is again the normal distribution. Therefore we will model μ by the normal distribution with a mean m and a variance s^2,

$$\mu \sim \mathcal{N}(m_{t-1}, s_{t-1}^2), \qquad m \in \mathbb{R}, s^2 \in \mathbb{R}^+. \tag{1.8}$$

The index $t-1$ means that in the current step labeled as t, the most up to date information we have is from step $t-1$ therefore those measurements are from the previous step, and we will assimilate them to the current step t.

As a next step, we have to rewrite the pdf of the prior normal as well as the normal data model into compatible forms. First, we will rewrite the data model. It will be characterized by parameter η, the sufficient statistic $T(y)$, and a normalizing function $g(\mu)$. The transformation is following:

$$\begin{aligned} f(y_t|\theta) = f(y_t|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(y_t-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(y_t^2 - 2y_t\mu + \mu^2)\right\} \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{y_t^2}{2\sigma^2}\right\}}_{h(y_t)} \cdot \underbrace{1}_{g(\mu)} \exp\left\{ \underbrace{\begin{bmatrix} \mu \\ -\frac{\mu^2}{2} \end{bmatrix}^\intercal \begin{bmatrix} \frac{y_t}{\sigma^2} \\ \frac{1}{\sigma^2} \end{bmatrix}}_{\eta(\mu)^\intercal T(y_t)} \right\}. \end{aligned} \tag{1.9}$$

Now we also have to rewrite the prior normal into a compatible form. It

[2] Usually abbreviated as PDF

can be done in the following way:

$$\begin{aligned}
\pi(\mu|\xi_{t-1},\nu_{t-1}) &= \pi(\mu|m_{t-1},s^2_{t-1}) \\
&= \frac{1}{\sqrt{2\pi s^2_{t-1}}}\exp\left\{\frac{-1}{2s^2_{t-1}}(\mu-m_{t-1})^2\right\} \\
&= \frac{1}{\sqrt{2\pi s^2_{t-1}}}\exp\left\{\frac{-1}{2s^2_{t-1}}(\mu^2-2\mu m_{t-1}+m^2_{t-1})\right\} \\
&= \underbrace{\frac{1}{\sqrt{2\pi s^2_{t-1}}}\cdot\exp\left\{-\frac{m^2_{t-1}}{2s^2_{t-1}}\right\}}_{q(\xi_{t-1},\nu_{t-1})}\cdot\underbrace{1}_{g(\mu)^{\nu_{t-1}}}\exp\left\{\underbrace{\begin{bmatrix}\mu\\-\frac{\mu^2}{2}\end{bmatrix}^\intercal\begin{bmatrix}\frac{m_{t-1}}{s^2_{t-1}}\\\frac{1}{s^2_{t-1}}\end{bmatrix}}_{\eta(\mu)^\intercal\xi_{t-1}}\right\}.
\end{aligned} \tag{1.10}$$

Next, we can use the Bayes' theorem, shown in Section 1.1.1, to update the hyperparameters of the prior normal (ξ,ν).

$$\begin{aligned}
\xi_t &= \xi_{t-1}+T(y_t), \\
\nu_t &= \nu_{t-1}+1.
\end{aligned} \tag{1.11}$$

Due to $g(\nu)=1$ in this case, it can be omitted from our calculations. Because of that, the posterior hyperparameter ξ_t will be equivalent to:

$$\begin{bmatrix}\frac{m_t}{s^2_t}\\\frac{1}{s^2_t}\end{bmatrix} = \begin{bmatrix}\frac{m_{t-1}}{s^2_{t-1}}\\\frac{1}{s^2_{t-1}}\end{bmatrix} + \begin{bmatrix}\frac{y_t}{\sigma^2}\\\frac{1}{\sigma^2}\end{bmatrix}. \tag{1.12}$$

Now we can relatively simply rewrite it back to the form of a direct update to the prior normal hyperparameters, m_{t-1} and s^2_{t-1}. The final result should look like this:

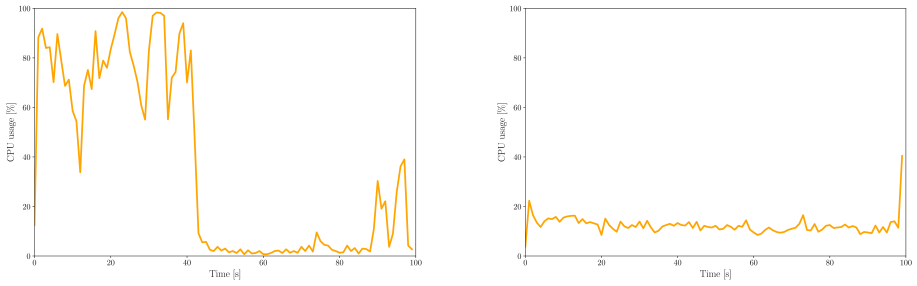
$$\begin{aligned}
s^2_t &= \left(\frac{1}{s^2_{t-1}}+\frac{1}{\sigma^2}\right)^{-1}, \\
m_t &= \left(\frac{m_{t-1}}{s^2_{t-1}}+\frac{y_t}{\sigma^2}\right)\cdot s^2_t = \frac{\sigma^2 m_{t-1}+s^2_{t-1}y_t}{s^2_{t-1}+\sigma^2}.
\end{aligned} \tag{1.13}$$

For us to demonstrate its functionality, we have estimated a mean utilization of CPU[3], both under intense load as well as under a mild load. The percentual load of CPU was logged every second for the duration of 100 seconds. Representation of those measurements can be seen in Figure 1.1.

We have then estimated the mean of the utilization of CPU in time, using a normal distribution with known variance. Graphs of this estimation can be seen in Figure 1.2. It can be easily seen, in Figure 1.2a, that the estimation has

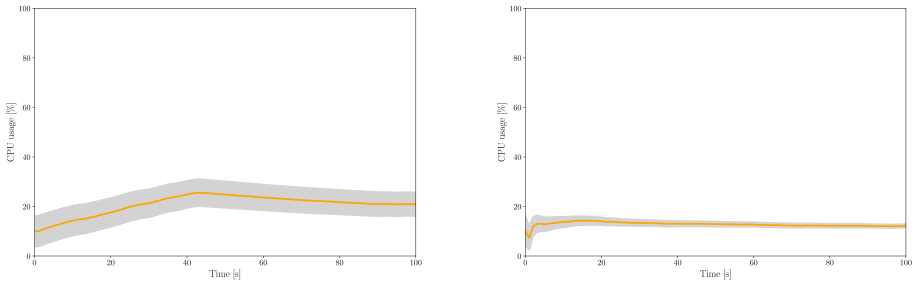
[3]Central processing unit

troubles with the estimation of highly dynamic measurements, where the mean estimation does not react fast enough. On the other hand, the mean estimation for idle CPU is very good, and even the interval around our estimate gets better with further progress of the estimation.



(a) (b)

Figure 1.1: (a) CPU load under heavy impact application (b) CPU load under passive conditions



(a) (b)

Figure 1.2: (a) Estimate of mean CPU load under heavy use (b) Estimate of mean CPU load under mild use

1.2 Kalman filter

For this section, we have used a similar approach and notation as in [4], and [5].

1.2.1 Introduction

Let us assume that we have a state model that is time-invariant and is comprised of the following equations:

$$x_t = Ax_{t-1} + Bu_t + w_t, \tag{1.14}$$
$$y_t = Hx_t + \varepsilon_t, \tag{1.15}$$

where x_t and x_{t-1} are the state vectors, we will attempt to estimate, u_t is a control vector, that is known, both w_t and ε_t are the state noise and measurement noise respectively. Last but not least, A, B and H are matrices of compatible shape.

Let us also assume that both w_t and ε_t are independent and from Normal distribution centered at zero, therefore:

$$\begin{aligned} w_t &\sim \mathcal{N}(0, Q), \\ \varepsilon &\sim \mathcal{N}(0, R). \end{aligned} \tag{1.16}$$

The fact that both noise distributions are centered in zero is profoundly important since if it was not the case, we would end up with a systematic error in our estimation.

The problem as defined above can be solved using the Kalman filter. Kalman filter is one of the most well-known sequential estimators, usually called filters, due to its use in the navigation for the Apollo program. Since then, it has become one of the most commonly used filters in all possible types of industry, from telecommunications through medical instruments all the way to the automotive industry.

The typical usage of the Kalman filter is in improving the sensory measurements. Since the Kalman filter is independent of the way that the specific sensor operates, it can work the same way regardless of the type of sensor that supplies the data. However, the Kalman filter in its natural form has one major limitation, that it is limited to linear systems. There are variants of the Kalman filter that remove these limitations, but they are out of scope for this thesis.

1.2.2 Derivation of the Kalman filter

Now we can build upon the state model, defined in Section 1.2.1. From normality, we can deduce that:

$$x_t \sim \mathcal{N}(Ax_t + Bu_t, Q), \quad \text{with the density equal to } p(x_t|x_{t-1}, u_t) \tag{1.17}$$
$$y_t \sim \mathcal{N}(Hx_t, R), \quad \text{with the density equal to } f(y_t|x_t) \tag{1.18}$$

The last thing that we will need is a prior distribution for x_t. As a consequence of model y_t being Normal, we known that correct conjugate prior will again be the normal distribution, this was shown in Section 1.1.2. Hence

we will select prior distribution for x_t as Normal distribution, with mean x_{t-1}^+ and covariance matrix P_{t-1}^+.

$$\pi(x_t|y_{0:t-1}, u_{0:t-1}) = \mathcal{N}(x_{t-1}^+, P_{t-1}^+) \tag{1.19}$$

The Kalman filter, like many other filters, runs in two main steps, usually called the prediction step and the update step, respectively.The prediction step, as its name suggests, predicts the next state vector using the filter's statistical knowledge. Afterward, the update step takes measurement obtained from the sensor and corrects the prediction. Two main things are important to realize. First, that if there is no noise, there is no need for the update step, since the prediction should be already correct, and there will be no need to correct it. Second, even though usually every prediction step is followed by an update step, this does not have to be the case. Prediction can run multiple times without being followed by any update step, e.g. if there is a drop-out in measurements. However, the longer we do not have updates, the more uncertain our prediction becomes.

Now let us derive the prediction step. As stated previously, in the prediction step we want to approximate the time development of $x_{t-1} \to x_t$. We will combine the prior distribution with the evolution model to obtain the posterior distribution.

$$\pi(x_t|y_{0:t-1}, u_{0:t}) = \int p(x_t|x_{t-1}, u_t), \pi(x_{t-1}|y_{0:t-1}, u_{0:t-1}) dx_{t-1}. \tag{1.20}$$

Since we are multiplying two Normal distributions, we will again get a Normal distribution $\mathcal{N}(x_t^-, P_t^+)$, with hyperparameters equal to:

$$\begin{aligned} x_t^- &= Ax_{t-1}^+ + Bu_t, \\ P_t^- &= AP_{t-1}^+ A^T + Q. \end{aligned} \tag{1.21}$$

Now that we have successfully predicted the x_t, we have to derive the update step based on measurement y_t that we have received. For this, the Bayes' theorem will be used,

$$\pi(x_t|y_{0:t}, u_{0:t}) \propto f(y_t|x_t)\pi(x_t|y_{0:t-1}, u_{0:t}). \tag{1.22}$$

In order to make the derivation easier, we will rewrite both model and prior distribution into an exponential distribution family. In this form, we can do the Bayesian update simply by adding up the hyperparameters and

sufficient statistics. Model and prior will be converted to the following form:

$$
\begin{aligned}
f(y_t|x_t) &\propto \exp\left\{-\frac{1}{2}(y_t - Hx_t)^T R^{-1}(y_t - Hx_t)\right\} \\
&= \exp\left\{\mathrm{Tr}\left(-\frac{1}{2}\underbrace{\begin{bmatrix}-1\\ x_t\end{bmatrix}\begin{bmatrix}-1\\ x_t\end{bmatrix}^T}_{\eta}\underbrace{\begin{bmatrix}y_t^T\\ H^T\end{bmatrix}R^{-1}\begin{bmatrix}y_t^T\\ H^T\end{bmatrix}^T}_{T(y_t)}\right)\right\}, \\
\pi(x_t|y_{0:t-1}, u_{0:t}) &\propto \exp\left\{-\frac{1}{2}(x_t - x_t^-)^T (P_t^-)^{-1}(x_t - x_t^-)\right\} \\
&= \exp\left\{\mathrm{Tr}\left(-\frac{1}{2}\underbrace{\begin{bmatrix}-1\\ x_t\end{bmatrix}\begin{bmatrix}-1\\ x_t\end{bmatrix}^T}_{\eta}\underbrace{\begin{bmatrix}(x_t^-)^T\\ I\end{bmatrix}(P_t^-)^{-1}\begin{bmatrix}(x_t^-)^T\\ I\end{bmatrix}^T}_{\xi_t}\right)\right\},
\end{aligned}
\tag{1.23}
$$

where I is an identity matrix of matching dimensions. We can now add up hyperparameters and sufficient statistics as:

$$
\begin{aligned}
\xi_t &= \xi_{t-1} + T(y_t) \\
&= \begin{bmatrix}(x_t^-)^T(P_t^-)^{-1}x_t^- + y_t^T R^{-1} y_t, & (x_t^-)^T(P_t^-)^{-1} + y_t^T R^{-1} H \\ (P_t^-)^{-1}(x_t^-)^T + H^T R^{-1} y_t & (P_t^-)^{-1} + H^T R^{-1} H\end{bmatrix}.
\end{aligned}
\tag{1.24}
$$

Now we can easily derive the posterior distribution hyperparameters.

$$
\begin{aligned}
P_t^+ &= (\xi_{t;[2,2]})^{-1} \\
&= [(P_t^-)^{-1} + H^T R^{-1} H]^{-1} \\
&= (I - K_t H) P_t^-, \\
x_t^+ &= (\xi_{t;[2,2]})^{-1}\xi_{t;[2,1]} \\
&= P_t^+[(P_t^-)^{-1}(x_t^-)^T + H^T R^{-1} y_t] \\
&= x_t^- + P_t^+ H^T R^{-1}(y_t - Hx_t^-),
\end{aligned}
\tag{1.25}
$$

where

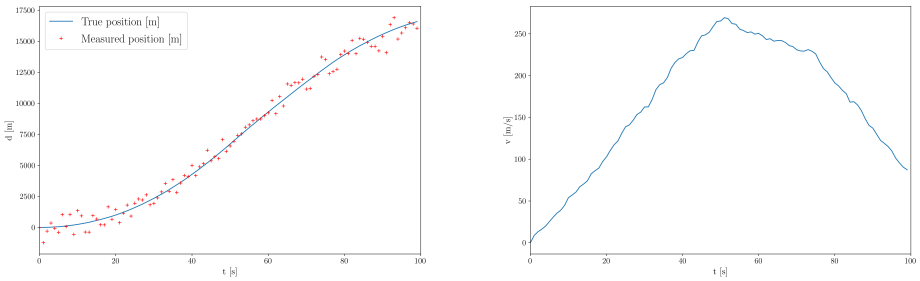
$$K_t = P_t^- H^T (R + H P_t^- H^T),$$

is usually called Kalman gain. The Kalman gain describes the relationship between the measurements and the current state estimate therefore it can be used to "tune in" the desired behavior of the filter. The higher the gain is, the more weight will the recent measurements have, and therefore the filter will follow them much more closely, and vice versa when we lower the gain.

1.2.3 Example

In order to show the functionality of the Kalman filter, we have added one example.

We have simulated a trajectory of a vehicle with variable acceleration. For the sake of simplicity, is the vehicle moving in a single dimension[4] and any form of friction is not taken into account. We will estimate the position of the vehicle, hence we have simulated inaccurate measurements of it. Length of the simulated trajectory is 100 seconds. Both, simulated position and velocity can be seen in Figure 1.3.



(a) (b)

Figure 1.3: (a) Simulated position (b) Simulated velocity

We have used the Kalman filter as explained in Section 1.2.2. Therefore we have supplied the filter with corresponding matrices A, B, H, R, and Q. As explained previously, the prediction step is followed by the update step, but in this case, we have limited the update step to only be taken between steps 30 and 100. The results of estimates of trajectory and velocity can be seen in Figure 1.4.

1.3 Variational Bayesian inference

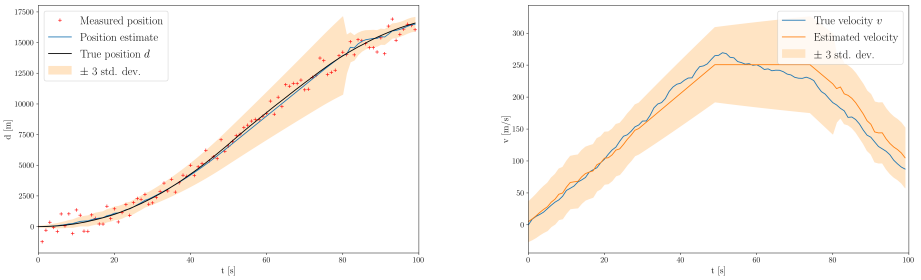
This section introduces the topic of variational Bayesian inference. It is mostly inspired by [1] and [6].

1.3.1 Variational inference

As was already apparent from Section 1.1.2, there are cases, where the estimate will be intractable, due to non-existent conjugate priors. We can solve this issue utilizing the variational inference.

Originally introduced in the 18th century by Euler, Lagrange, and others in form of the calculus of variations. In standard calculus, where we can

[4]Example of such vehicle could be a train.



(a) (b)

Figure 1.4: (a) Estimate of position (b) Estimate of velocity

look upon a function as a mapping between an input value of a variable and the corresponding value of the function. On the other hand, the calculus of variations resembles mapping that takes a function as an input and returns the value of the functional. Calculus of variations specifically uses variations, which expresses the difference in the value of functional in response to the changes of the input function. This behavior is especially useful in the case of optimization problems, where we can obtain the solution by exploring the possible input functions and finding one that maximizes or minimizes our functional.

Now let us take a look at how this will be useful for inference. Imagine that we have a fully Bayesian model that has all parameters given by prior distributions. Latent variables and parameters[5] will be labeled by Z, and observed variables will be labeled by X. If we have a set of N i.i.d.[6] data, where $X = \{x_1, \ldots, x_N\}$ and $Z = \{z_1, \ldots, z_N\}$. We have the joined distribution $p(X, Z)$ specified by our model, and we would like to find the approximation of both the posterior distribution $p(Z|X)$ and model evidence $p(X)$. We can decompose the log marginal probability[7] by

$$\begin{aligned} \ln p(x) &= \mathcal{L}(q) + KL(q||p), \\ \mathcal{L}(q) &= \int q(Z) \ln \left\{ \frac{p(X,Z)}{q(Z)} \right\} dZ, \\ \mathrm{KL}(q||p) &= -\int q(Z) \ln \left\{ \frac{p(Z|X)}{q(Z)} \right\} dZ, \end{aligned} \tag{1.26}$$

where $\mathcal{L}(q)$ is a lower bound and KL is Kullback–Leibler divergence[8], which measures the difference between one probability distribution and a reference

[5] Latent variables are not directly observed, but rather inferred from other observed variables

[6] independent, identically distributed

[7] Similar approach as in case of expectation-maximization algorithm, as shown in [7]

[8] Sometimes called relative entropy

probability distribution.

Naturally, we want to maximize $\mathcal{L}(q)$ with respect to the distribution $q(Z)$ since it is equivalent to minimization of KL divergence. Obviously, since we want to do this when our desired posterior is intractable, we have to restrict $q(Z)$. Otherwise, the ideal $q(Z)$ would be $p(Z|X)$, which we have established as intractable. Therefore we have to restrict the $q(Z)$ to such a family of distribution that both minimizes the KL divergence so that we have a "good enough" approximation of the real distribution and is also tractable.

The main question now is how to achieve this restriction. For this, we can divide the elements of Z into M disjoint groups, which will be denoted as Z_i, where $i \in [1, \ldots, M]$. Then we can suppose that the q distribution is factorized with respect to those groups as

$$q(Z) = \prod_{i=1}^{N} q_i(Z_i). \tag{1.27}$$

Now we have to find the distribution from $q(Z)$, for which the $\mathcal{L}(q)$ will be the largest. We can achieve this by a free form optimization of $\mathcal{L}(q)$ with respect to the distributions $q_i(Z_i)$. So we will substitute (1.27) into (1.26) and dissect out the dependence on one factors $q_j(Z_j)$.

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i(Z_i) \left\{ \ln p(X, Z) - \sum_i \ln q_i(Z_i) \right\} dZ \\
&= \int q_j(Z_j) \left\{ \int \ln p(X, Z) \prod_{i \neq j} q_i(Z_i) dZ_i \right\} dZ_j - \int q_i(Z_i) \ln q_j(Z_j) dZ_j + c \\
&= \int q_j(Z_j) \ln \tilde{p}(X, Z_j) dZ_j - \int q_j(Z_j) \ln q_j dZ_j + c,
\end{aligned} \tag{1.28}$$

where c is a constant and we have defined a new distribution $\tilde{p}(X, Z_j)$ by the relation

$$\ln \tilde{p}(X, Z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + c, \tag{1.29}$$

where c again denotes a constant.

If we take a closer look on an Equation (1.28), we can see it is a negative KL divergence between $q_j(Z_j)$ and $\tilde{p}(X, Z_j)$. So if we its maximization is equivalent to the minimization of the KL divergence and therefore the minimum will be when $q_j(Z_j) = \tilde{p}(X, Z_j)$. Using this notion, we can get to the optimal solution $q_j^*(Z_j)$ by,

$$\ln q_j^*(Z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + c, \tag{1.30}$$

where

$$\mathbb{E}_{i \neq j}[\ln p(X, Z)] = \int \ln p(X, Z) \prod_{i \neq j} q_i(Z_i) dZ_i. \tag{1.31}$$

Now we finally have all the necessary steps to variational inference, however, this form is not ideal for us and hence we will explore a different form in Section 1.3.2.

1.3.2 Message passing

Even though Variational Inference, shown in Section 1.3.1, can approximate the posterior distribution, the form that it takes is cumbersome and not ideal for our use case. In the ideal case, we would like to have it similar to the Kalman filter update in the form of adding hyperparameters and sufficient statistics, as shown in Section 1.2.2.

However, there exists an algorithm that uses our desired approach and, it is called Variational message passing and was shown in [6]. Let us now explain the basics of their approach.

We will again take a factorized distribution in the same form as shown in Equation (1.27). From Section 1.3.1 we know that the optimal form of the jth factor is

$$\ln q_j^*(Z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + c. \tag{1.32}$$

If we now assume that the model has a form of Bayesian network and we can express its joined distribution $p(X)$ in terms of individual nodes conditional distributions. If we label nodes by i, it will look like following,

$$p(X) = \prod_i p(X_i | pa_i), \tag{1.33}$$

where X_i are variables of node i and pa_i are variables of parent nodes of node i.

Now we can substitute Equation (1.33) into (1.32) and simplyfy it into the following form,

$$\ln q_j^*(Z_j) = \mathbb{E}_{i \neq j}[\ln p(Z_j, pa_j)] + \sum_{k \in ch_j} \mathbb{E}_{i \neq j}[\ln p(X_k | pa_k)] + c, \tag{1.34}$$

where ch_j denotes all children of node j. Hence, in order to evaluate q_j^*, we only need input from nodes that are in Markov blanket[9] of Z_j. In this case, Markov blanket of Z_j consists of its parents pa_j, children ch_j and last but not least, its coparents[10]. Visualization of this Markov blanket can be seen in Figure 1.5.

[9] Markov blanket marks a subset of variables that are sufficient to infer the desired random variable

[10] Coparents of node Z are nodes that have at least one child, that is also a child of node Z

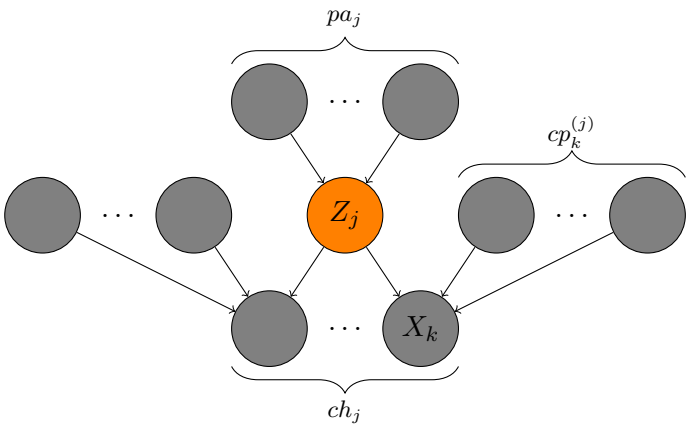


Figure 1.5: Markov blanket for a node H_j

How specifically will the form of Equation (1.34) look is determined by the conditional distribution of the model. It is known that for ideal simplification of variational update equations, the distribution of variables conditioned on their parents should be from exponential family and also be conjugate with their parents' distributions. We have defined the exponential family in Definition 1.1.1 in Section 1.1.2.

We can equivalently rewrite the definition from Definition 1.1.1 into its, in this case more useful variation, in the form of

$$f(y|\theta) = \exp\{\eta(\theta)^T T_\theta(y) + A(\theta) + B(y)\}, \tag{1.35}$$

for this particular case, we will substitute y by an X, that we use for labeling the particular node. Hence the equation will be equal to,

$$f(X|\theta) = \exp\{\eta(\theta)^T T_\theta(X) + A(\theta) + B(X)\}, \tag{1.36}$$

where we will call $\eta(\theta)$ as natural parameter vector and $A(\theta)$ is normalization function.

Since we know $\eta(\theta)$, we can deduce the expectation of the natural statistic vector with respect to distribution. First, we will rewrite (1.36) in the following way,

$$f(X|\theta) = \exp[\eta^T T_\theta(X) + \tilde{A}(\theta) + B(X)], \tag{1.37}$$

in which $\tilde{A}$ is a reparameterisation of A in terms of ϑ. If we integrate over the X and then differentiate the result over ϑ, we can see that the expectation of natural statistics vector $T_\vartheta(X)$ is,

$$\mathbb{E}[T_\theta(X)] = -\frac{d\tilde{A}(\theta)}{d\theta}. \tag{1.38}$$

Now that we know how to compute the expectation of $T_\theta(X)$, we need to figure out how to optimize our variational distribution when the model is conjugate-exponential. We will show this on a case when we are optimizing a factor $q(Z_j)$ of node Z_j, whose children are X_k hence it is equivalent to the graph shown in Figure 1.5.

We will start by rewriting Equation (1.36), for Z_j into

$$\ln f(Z_j|pa_j) = \theta_j(pa_j)^T T_{\theta,j}(Z_j) + A_j(pa_j) + B_j(Z_j). \tag{1.39}$$

Next, we can also write a logarithm of the conditional probability of node X_k given its parents, which are also of distributions from the exponential family. It will look in the following way,

$$\ln f(X_k|Z_j, cp_k^j) = \theta_j(Z_j|cp_k^j)^T T_{\theta,X_k}(X_k) + A_{X_k}(Z_j, cp_k^j) + B_{X_k}(X_k), \tag{1.40}$$

where cp_k^j are coparents of Z_j, that are also parents of of node X_k.

In order to use the conjugacy, we have to have both Equations (1.39) and (1.40) in the same functional form with respect to Z_j. Hence we can simplify it in terms of $T_{\vartheta,Z_j}(Z_j)$ by defining functions ϑ_{XZ} and λ in the following way,

$$\ln f(X_k|Z_j, cp_k^j) = \theta_{XZ} + T_{\theta,Z_j}(Z_j) + \lambda(X_k, cp_k^j). \tag{1.41}$$

If we take Equation (1.34) for a node Z_j, we can calculated in terms of $\mathbb{E}[T_\theta]$ for each node in Markov blanket of Z_j. After substitution we get and some simplification we can get to

$$\begin{aligned} \ln q^*_{Z_j}(Z_j) = &\left[\mathbb{E}_{i\neq j}[\theta_{Z_j}(pa_{Z_j})] + \sum_{k\in ch_{Z_j}} \mathbb{E}_{i\neq j}[\theta_{XZ}(X_k, cp_k)]\right]^T \\ &T_{\theta,Z_j}(Z_j) + B_{Z_j}(Z_j) + c. \end{aligned} \tag{1.42}$$

Naturally, the $q^*_{Z_j}$ is also from the exponential family of distributions and, its natural parameter vector $\theta^*_{Z_j}$ will be the following

$$\theta^*_{Z_j} = \mathbb{E}[\theta_{Z_j}(pa_{Z_j})] + \sum_{k\in ch_{Z_j}} \mathbb{E}[\theta_{XZ}(X_k, cp_k)], \tag{1.43}$$

where all expectations are with respect to q.

It is known, that based on Equation (1.40) and (1.41), the $\ln f(X_k|cp_{Z_j})$ is a multi-linear function.Expanding on this, the θ_{Z_j} and the θ_{XZ} must also be a multi-linear functions. Therefore we can reparameterise them in the following way,

$$\begin{aligned} \tilde{\theta}_{Z_j}(\{\mathbb{E}[T_{\theta,i}]\}_{i\in pa_{Z_j}}) &= \mathbb{E}[\theta_{Z_j}(pa_{Z_j})], \\ \tilde{\theta}_{XZ}(\mathbb{E}[T_{\theta,k}], \{\mathbb{E}[T_{\theta,j}]\}_{j\in cp_k}) &= \mathbb{E}[\theta_{XZ}(X_k, cp_k)]. \end{aligned} \tag{1.44}$$

Lastly, we need to compute the expectation of $T_{\theta,X}$, however this can simply be done in the same way as in Equation (1.38).

Now we just have to specify what should message between nodes contain and between which node should they be passed. First is the message from a parent node, in our case Z_j, to a child node, for us X_k, contains only the expectation of natural statistic vector of q. Hence it is

$$m_{Z_j \to X_k} = \mathbb{E}[T_{\theta,Z_j}]. \tag{1.45}$$

Message from child node X_k to parent Z_j is slightly more complicated since it provides aggregated information from all coparents of Z_j and logically node X_k has to await the information from all those nodes. The final message has the following form

$$m_{X_k \to Z_j} = \tilde{\theta}_{XZ}(\mathbb{E}[T_{\theta,X_k}], \{m_{i \to X_k}\}_{i \in cp_k^j}). \tag{1.46}$$

Once a node Z_j obtain all messages, that means both from all its parents, as well as all its children, it can compute its posterior distribution $q^*_{Z_j}$, by updating the natural parameter vector $\theta^*_{Z_j}$. It is computed in the following way,

$$\theta^*_{Z_j} = \tilde{\theta}_{Z_j}(\{m_{i \to Z_j}\}_{i \in pa_{Z_j}}) + \sum_{j \in ch_{Z_j}} m_{j \to Z_j}. \tag{1.47}$$

If we sum it all up, we can write the whole message passing algorithm as shown in Algorithm 1, where G labels all nodes.

Algorithm 1: Variational message passing

1 Initialize each factor distribution q_j by initialising the corresponding moment vector $\mathbb{E}[T_{\theta,j}(X_j)]$.
2 **foreach** $X_j \in G$ **do**
3 Retrieved messages from all parent and children nodes, as shown in Equations (1.45) and (1.46).
4 Update natural parameter vector θ^*_j, as shown in Equation (1.47).
5 Update the moment vector $\mathbb{E}[T_{\theta,j}(X_j)]$.
6 **end**
7 Calculate the new value of lower bound $\mathcal{L}$(q).
8 If the increase of lower bound is small enough or a specific number of iterations have been reached, stop. Otherwise, return to line 2.

Chapter 2

Current state-of-the-art

In this chapter, we will show the current state-of-the-art approaches to the problem that we aim to solve in this thesis. We mainly focus on the following three articles [8], [9] and [10].

2.1 Recursive Noise Adaptive Kalman Filtering by Variational Bayesian Approximations

One of the influential papers in Bayesian filtering is work by [8].

This paper expands the idea of Kalman filter, shown in Section 1.2 and variational Bayesian inference, Section 1.3, with the idea of estimation on an model with unknown measurement noise covariance matrix R_t. Hence the state model will be following,

$$x_t = A_t x_{t-1} + q_t,$$
$$y_t = H_t x_t + r_t,$$

where $q_t \sim \mathcal{N}(0, Q_t)$ is Gaussian process noise, A_t and H_t are matrices of correct dimension and state x_t is n-dimensional vector, while measurement y_t is a d-dimensional vector. The key is the fact that $r_t \sim \mathcal{N}(O, R_t)$, where R_t is unknown and is therefore estimated.

In the paper, they derive a way of estimating the unknown R_t by using the variational inference and by using the correct prior distribution. They demonstrate the performance of their approach on an illustrative case of stochastic resonator model, with very promising results.

However, this paper encompasses only a part of the scope of this thesis. First of all, their approach does not solve the issue when the matrix Q_t is also unknown. On top of that, it does not develop such an approach for a distributed scenario, but it concentrates on a more traditional approach of using a single estimator.

2.2 Collaborative sequential state estimation under unknown heterogeneous noise covariance matrices

A paper that expands the ideas from paper [8], is a new publication [9].The authors of this paper extend the idea of working with a state-space model that has an unknown measurement noise covariance matrix $R^{(i)}$, to a distributed scenario, meaning multiple independent models[11] are cooperating in the estimation with one another.

They assume that the nodes observe the following state-space model,

$$\begin{aligned} x_t &= A_t x_{t-1} + B_t u_t + w_t, \\ y_t^{(i)} &= H_t x_t + v_t^{(i)}, \end{aligned}$$

where the state variable x_t is p-dimensional vector and $p \geq 1$ and is shared among all nodes, u_t is a control variable, that is also known by all nodes, A_t, B_t and H_t are matrices of matching dimensions. As expected, w_t is an independent process noise, $w_t \sim \mathcal{N}(0, Q)$, where Q is a known process noise covariance matrix. Finally, $v_t^{(i)}$ is an i.i.d. zero-centered variable $v_t^{(i)} \sim \mathcal{N}(0, R^{(i)})$, where the $R^{(i)}$ is obviously unknown.

Since the whole method is distributed, they have also defined a network of nodes, as an undirected graph $G(\mathcal{I}, \mathcal{E})$, that has a set of nodes $\mathcal{I} = \{1, \ldots, |\mathcal{I}|\}$ and set of edges $\mathcal{E}$ that creates the network topology. Nodes are limited in their communication to only their neighborhood $\mathcal{I}^{(i)}$, which is comprised of their adjacent nodes.

Based on those assumptions, they expand the distribution scenario with the fact that the measurement noise covariance matrix $R^{(i)}$ does not have to be the same in the whole network G, but there can be multiple different R matrices, such as $R_1, \ldots, R_L$, where $1 \leq L \leq |\mathcal{L}|$. Then each node will have one of those matrices assigned as his $R^{(i)}$ matrix.

Throughout the paper, they develop a solution to this problem and also show its performance on two different examples of tracking 2-dimensional trajectory. One with the same $R^{(i)}$ for all nodes, and the second assumes two distinct groups of nodes with different $R^{(i)}$.

This paper shares some similarities with the approach taken in this work, but there is still some difference in a few of its components. Mainly the need for known process covariance matrix Q. However, their work is distributed, just as this work.

[11] Sometimes called nodes

2.3 A novel adaptive Kalman filter with inaccurate process and measurement noise covariance matrices

One of the recent novelties in the field of Bayesian filtering was introduced in paper [10]. The authors of this paper propose a solution to an issue with an unknown process noise covariance matrix, as well as a measurement noise covariance matrix.

They assume the following state-space model,

$$x_t = F_t x_{t-1} + w_t,$$
$$z_t = H_t x_t + v_t,$$

where x_t is n-dimensional state vector, z_t is m-dimensional measurement vector, F_t and H_t are matrices of compatible dimensions. At last, w_t and v_t are respectively Gaussian process and measurement noise vectors, hence $w_t \sim \mathcal{N}(0, Q_t)$ and $v_t \sim \mathcal{N}(0, R_t)$, where both Q_t, as well as R_t are unknown.

It can be easily seen that their paper aims to solve a very similar issue as this work, but there are still some significant differences. The first one is that they do not use the message-passing approach, as shown in Section 1.3.2. The second is that they do not take into account the possibility of having it in a distributed setting but rather show it only for a case of the single-node estimator.

Chapter 3

Analysis and design

In this chapter, we provide a complete formal specification of the problem that we aim to solve, as well as a description of our approach. In order to provide better readability, we have split the explanation of our approach into two main parts, being the explanation on a non-distributed setting and a part where we expand that by all things necessary for distributed setting.

3.1 Problem statement

We consider a network of I collaborating agents that are connected by edges in a representation of an undirected graph. Labeling of agents is done according to their ordinal number $i \in \mathcal{I} = \{1, \ldots, I\}$ and every agent is allowed to have bidirectional sharing of information with his adjacent neighbors $j \in \mathcal{I}$, within a maximum distance of one hop. All agents that fulfil those conditions for agent i will be called neighbourhood $\mathcal{I}^{(i)}$.

Each agent independently observes a stochastic process determined by the following linear state-space model,

$$x_t = A_t x_{t-1} + B_t u_t + \omega_t, \tag{3.1}$$

$$y_{i,t} = H_t x_t + \epsilon_{i,t}, \tag{3.2}$$

where the Equation (3.1) shows the development of hidden state vector $x_t \in \mathbb{R}^n$ build upon is combination of its previous value with known matrix $A_t \in \mathbb{R}^{n \times n}$, the known input u_t, a matrix B_t that have a compatible dimensions, last but not least the random process noise $\omega_{i,t}$. On the other hand the Equation (3.2) adapts x_t to the local measurement $y_{i,t} \in \mathbb{R}^m$ via a known matrix $H_t \in \mathbb{R}^{m \times n}$ and local random measurement noise variable $\epsilon_{i,t} \in \mathbb{R}^m$. We also have the following assumption about $\omega_{i,t}$ and $\epsilon_{i,t}$,

$$\begin{aligned} \omega_{i,t} &\sim \mathcal{N}(0, Q_t), \\ \epsilon_{i,t} &\sim \mathcal{N}(0, R_t). \end{aligned} \tag{3.3}$$

Based on this we can rewrite Equations (3.1) and (3.2) into their probabilistic forms as,

$$x_t \sim \mathcal{N}(A_t x_{t-1} + B_t u_t, Q_t), \tag{3.4}$$
$$y_{i,t} \sim \mathcal{N}(H_t x_t, R_t), \tag{3.5}$$

hence their densities will be equivalent to,

$$\begin{aligned} &p(x_t|x_{t-1}, u_t), \\ &f(y_{i,t}|x_t). \end{aligned} \tag{3.6}$$

Then for Bayesian approach to non-collaborative inference, we estimate x_t based on the model specified by Equations (3.4) and (3.5), that has Gaussian distribution, with mean equal to $\hat{x}_{t-1}^{+} \in \mathbb{R}^n$ and covariance matrix $P_{t-1}^{+} \in \mathbb{R}^{n \times n}$, as a prior,

$$x_t|y_{0:t-1}, u_{0:t-1} \sim \mathcal{N}(\hat{x}_{t-1}^{+}, P_{t-1}^{+}), \tag{3.7}$$

where $y_{0:t-1}$ and $u_{0:t-1}$ embraces all information about y_τ and u_τ up to this point, hence for $\tau = 0$ up to $\tau = t-1$.

Now its probability density $\pi_i(x_t|y_{i,0:t-1}, u_{i,0:t-1})$ can be updated in two Kalman filtering steps, shown in Section 1.2,

1. Prediction
   The new value of mean $\hat{x}_t^{-}$ is predicted, while the covariance matrix P_t^{-} is scaled,

   $$\pi_i(x_t|y_{i,0:t-1}, u_{0:t}) = \int \pi_i(x_t|y_{i,0:t-1}, u_{0:t-1}) p(x_t|x_{t-1}, u_t) dx_{t-1}, \tag{3.8}$$

   hence we will end up with the predicted prior Gaussian distribution

   $$\mathcal{N}_i(\hat{x}_{i,t}^{-}, P_{i,t}^{-}) = \mathcal{N}(A_t x_{i,t}^{-}, A_t P_{i,t}^{-} A_t^T + Q_t). \tag{3.9}$$

2. Update
   We can combine the new measurement $y_t^{(i)}$ with predicted prior Gaussian distribution using Bayes' theorem,

   $$\pi_i(x_t|y_{i,0:t}, u_{0:t}) \propto \pi_i(x_t|y_{i,0:t-1}, u_{0:t-1}) f(y_{i,t}|x_t). \tag{3.10}$$

   This will lead to a tractable posterior Gaussian distribution $\mathcal{N}(\hat{x}_{i,t}^{+}, P_{i,t}^{+})$, with

   $$\begin{aligned} P_{i,t}^{+} &= \left[(P_{i,t}^{-})^{-1} + H_t^T R_t^{-1} H_t \right]^{-1}, \\ \hat{x}_{i,t}^{+} &= P_{i,t}^{+} \left[(P_{i,t}^{-})^{-1} \hat{x}_{i,t}^{-} + H_t^T R_t^{-1} y_{i,t} \right]. \end{aligned} \tag{3.11}$$

   .

As explained in Section 1.2, the Kalman filter in this form has certain limitations, one being the necessity of full knowledge of the state-space model probability form, as in Equations (3.4) and (3.5). There are quite a few known solutions for case of missing R_t, such as [8], shown in Section 2.1, or [11]. However in our case we do not know both R_t, as well as Q_t. In that case the solutions get far more complicated. One solution was shown in [10], we have provided a short summary in Section 2.3. On top of this we have added the complication of distributed setting, similar to one shown in [9] and shortly explained in 2.2.

3.2 Variational Kalman filtering

We use the following section to show how to apply the variational approach to the problem described in Section 3.1. For the sake of brevity, the whole section will be from the point of a single agent, hence there will be no need for node indices.

As established in Section 1.2, generic Kalman filter needs for sequential estimation of x_t the full knowledge of the state model (3.4) and (3.5), with known matrices Q_t and R_t. Since Q_t is unknown the prediction of $\mathcal{N}(\hat{x}_t^-, P_t^-)$ is not possible. On top of that, the missing R_t prevents the Bayesian update yielding $\mathcal{N}(\hat{x}_t^+, P_t^+)$. We have developed an approach, that aims to solve this issue by a simultaneous inference of x_t and R_t together with optimization of P_t and selection of best hypothesise value of Q_t. Let us now introduce a vector θ_t, that contains all the unknown variables in the following way,

$$\theta_t = [x_t, R_t, P_t]. \tag{3.12}$$

In order to achieve better reading experience, we will not vectorize either R_t or P_t, Q_t is also not included in vector θ_t since is optimization is based on hypothesis testing procedure.

As usual, the estimation of θ_t proceeds with prior distribution $\pi(\theta_t|y_{0:t-1}, u_{0:t-1})$, but unlike in standard Kalman filter, we do not have any distribution of this type, that would produce an analytically tractable posterior distribution. One of the approaches to solving this issue was shown in [10], which in turn expands on ideas by [11]. We will use a similar approach. First let us start with replacing the true posterior distribution $\pi(\theta_t|y_{0:t-1}, u_{0:t-1})$ by variational factors,

$$\pi(\theta_t|y_{0:t-1}, u_{0:t-1}) \approx \rho(\theta_t) \equiv \rho(x_t)\rho(R_t)\rho(P_t). \tag{3.13}$$

Next step is to seek hyperparameters of the factors, such that they minimize the mutual Kullbach-Leibler divergence $\mathcal{D}[\rho(\theta_t)||\pi(\theta_t|y_{0:t}, u_{0:t})]$. There-

fore

$$\begin{aligned}\mathcal{D}[\rho(\theta_t)||\pi(\theta_t|\cdot)] &= \mathbb{E}_{\rho(\theta_t)}\left[\log\frac{\rho(\theta_t}{\pi(\theta_t|\cdot)}\right] \\ &= \mathbb{E}_{\rho(\theta_t)}\left[\log\frac{\rho(\theta_t)}{f(y_t|\cdot)\pi(\theta_t|\cdot)}\right] + \mathbb{E}[\log f(y_t|\cdot)] \\ &= -\mathcal{L}[\rho(\theta_t)] + \log f(y_t|\cdot),\end{aligned} \tag{3.14}$$

where we have exploited the Bayes' theorem

$$\pi_t(\theta_t|\cdot) = \frac{f(y_t|\cdot)\pi_i(\theta_t|\cdot)}{\int f(y_t|\cdot)\pi_i(\theta_t|\cdot)d\theta_t}, \tag{3.15}$$

and the absence of θ_t in log-evidence

$$f(y_t|\cdot) = \int f(y_t|\cdot)\pi_i(\theta_t|\cdot)d\theta_t, \tag{3.16}$$

leaving its logarithm intact under by the expectation operator. The term $\mathcal{L}[\rho(\theta_t)]$ is known as evidence lower bound[12]. If we take a closer look on the last line of Equation (3.14), we can see that it bounds the log-evidence $\log f(y_t|\cdot)$, hence if the divergence $\mathcal{D}[\rho(\theta_t)||\pi(\theta_t|\cdot)]$ was equal to zero, the ELBO will be equivalent to $\log f(y_t|\cdot)$.

As the log-evidence $\log f(y_t|\cdot)$ is static, the minimization of Kullbach-Leibler divergence is simplified to maximization of ELBO. This was already discussed in Section 1.3.1.

3.2.1 Update by measurements

If we once again take a look at Equation (3.14), specifically ELBO, we can investigate its properties further. First, we can expand it in terms of its individual variables,

$$\begin{aligned}-\mathcal{L}(\theta_t) &= \mathbb{E}_{\rho(\theta_t)}\left[\log\frac{\rho(\theta_t)}{f(y_t|\cdot)\pi(\theta_t|\cdot)}\right] \\ &= \mathbb{E}_{\rho(xt)}\left[\log\frac{\rho(x_t)}{\exp\{\mathbb{E}_{\rho(R_t,P_t)}[\log f(y_t|\cdot)\pi(\theta_t|\cdot)]\}}\right] + c_x \\ &= \mathbb{E}_{\rho(R_t)}\left[\log\frac{\rho(R_t)}{\exp\{\mathbb{E}_{\rho(x_t,P_t)}[\log f(y_t|\cdot)\pi(\theta_t|\cdot)]\}}\right] + c_R \\ &= \mathbb{E}_{\rho(P_t)}\left[\log\frac{\rho(P_t)}{\exp\{\mathbb{E}_{\rho(x_t,R_t)}[\log f(y_t|\cdot)\pi(\theta_t|\cdot)]\}}\right] + c_P,\end{aligned} \tag{3.17}$$

[12]Usually abbreviated as ELBO

where c_x, c_R and c_P are terms that are independent of x_t, R_t and P_t respectively. It can be clearly seen, that every element of θ_t, has is own KL divergence, that will reach its minimum if

$$\begin{aligned}\rho(x_t) &\propto \exp\{\mathbb{E}_{\rho(R_t,P_t)}[\log f(y_t|\cdot)\pi(\theta_t|\cdot)]\},\\ \rho(R_t) &\propto \exp\{\mathbb{E}_{\rho(x_t,P_t)}[\log f(y_t|\cdot)\pi(\theta_t|\cdot)]\},\\ \rho(P_t) &\propto \exp\{\mathbb{E}_{\rho(x_t,R_t)}[\log f(y_t|\cdot)\pi(\theta_t|\cdot)]\}.\end{aligned} \tag{3.18}$$

One of the remarkable properties of Equation (3.18) is the fact, that if $f(y_t|\theta_t)$ is a probability density function from an exponential family of distributions and we use convenient conjugate priors for estimation of x_t, R_t and P_t, then the left-hand side factors of Equation (3.18) will be tractable. If these prerequisites are fulfilled, the ELBO optimization can be achieved by means of coordinate-ascent variational inference[13], shown for example in [1]. It will use the individual factors from Equation (3.18) to employ point estimates of the other elements of θ_t during its update at each iteration. It has been shown, for example in [12], that ELBO does not have to be a convex objective function, hence CAVI can only guarantee a local optimum.

We will show that the Gaussian measurement model for y_t parameterized by either x_t or R_t is an exponential family distribution, in form already defined by Definition 1.1.1. It will look the following way.

Definition 3.2.1. If we assume that an m-dimensional random vector $y_t \sim \mathcal{N}(H_t x_t, R_t)$, where $x_t \in \mathbb{R}^n$, $H_t \in \mathbb{R}^{m\times n}$ and the positive definite covariance matrix $R_t \in \mathbb{R}^{m\times m}$. Based on this parametrization, the probability density function will be in the following form,

$$\begin{aligned} f_i(y_t|x_t,R_t) &= (2\pi)^{-\frac{m}{2}}|R_t|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(y_t-H_tx_t)^TR_t^{-1}(y_t-H_tx_t)\right\}\\ &\propto \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\underbrace{\begin{bmatrix}-1\\ x_t\end{bmatrix}\begin{bmatrix}-1\\ x_t\end{bmatrix}^T}_{\eta_{x_t}}\underbrace{\begin{bmatrix}y_t^T\\ H_t^T\end{bmatrix}R_t\begin{bmatrix}y_t^T\\ H_t^T\end{bmatrix}^T}_{T_{x_t}(y_t)}\right)\right\}\\ &\propto \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\underbrace{\begin{bmatrix}(R_t)^{-1}\\ \ln|R_t|\end{bmatrix}}_{\eta_R}\underbrace{\begin{bmatrix}(y_t-H_tx_t)(y_t-H_tx_t)^T\\ 1\end{bmatrix}}_{T_R(y_t)}\right)\right\}. \end{aligned} \tag{3.19}$$

In a similar fashion, we can show that the Gaussian model for x_t given P_t is also an exponential family distribution.

[13] Usually abbreviated as CAVI

Definition 3.2.2. Let us assume, that we have an n-dimensional variable $x_t \sim \mathcal{N}(\hat{x}_t, P_t)$, where $\hat{x}_t$ is the mean vector of length n and P_t is a $n \times n$ positive definite covariance matrix. Given that our assumptions are correct, the probability density function can be written in the following form,

$$\begin{aligned}\pi_x(x_t|\hat{x}_t, P_t) &= (2\pi)^{-\frac{n}{2}}|P_t|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\hat{x}_t - x_t)^T(P_t)^{-1}(\hat{x}_t - x_t)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\underbrace{\begin{bmatrix}-1\\ x_t\end{bmatrix}\begin{bmatrix}-1\\ x_t\end{bmatrix}^T}_{\eta_{x_t}}\underbrace{\begin{bmatrix}(\hat{x}_t)^T\\ I\end{bmatrix}(P_t)^{-1}\begin{bmatrix}(\hat{x}_t)^T\\ I\end{bmatrix}^T}_{\Xi_{x_t,k}}\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\underbrace{\begin{bmatrix}P_t^{-1}\\ \ln|P_t|\end{bmatrix}^T}_{\eta_{P_t}}\underbrace{\begin{bmatrix}(x_t - \hat{x}_t)(x_t - \hat{x}_t)^T\\ 1\end{bmatrix}}_{T_{P_t}(x_t)}\right)\right\}.\end{aligned} \tag{3.20}$$

From the definitions established in Section 1.1.2 is clear, that multiplication of the exponential family distribution $f(y|\vartheta)$ with its conjugate prior $\pi(\vartheta)$ yields a distribution of the same type as prior. Therefore

$$\begin{aligned}\pi(\vartheta|y) &\propto f(y|\vartheta)\pi(\vartheta) \\ &\propto \exp\{\eta(\vartheta)^T(\Xi_\vartheta^- + T_\vartheta(y)) - (\nu^- + 1)B(\vartheta)\} \\ &\propto \exp\{\eta(\vartheta)^T\Xi_\vartheta^+ - \nu^+B(\vartheta)\},\end{aligned} \tag{3.21}$$

where

$$\Xi_\vartheta^+ = \Xi_\vartheta^- + T_\vartheta(y), \tag{3.22}$$

$$\nu^+ = v^- + 1, \tag{3.23}$$

where we use the $^-$ and $^+$ to denote the prior and posterior hyperparameter, respectively[14]. This principle of conjugate priors is the key for an exact sequential approximation, shown for example in Section 1.2. However, in our case of variational inference, it can be also used for achieving an analytically tractable approximate inference. This can be done in the following way.

First, let us focus on the individual factors in Equation (3.18) and assume for each factor, that the remaining elements of θ_t are replaced by their point estimates. Then if the measurement model $f(y|\cdot)$ is an exponential family distribution and the prior distribution of the considered element of θ_t that is

[14]We will continue to use this notation throughout the thesis

conjugate to it. Finally, we can update the variational factor by using the principles of Bayesian updating, as shown in Equation (3.21).

Based on the individual properties of x_t, R_t and P_t, we can select matching distributions as variational factors, in form of the Gaussian and the inverse-Wishart. In order to avoid confusion, we will denote the posterior estimate of P_t from the inverse-Wishart factor that enters the prior factor of x_t as $\hat{P}_t^*$, and we will do the same for the related posterior hyperparameters. Both prior and posterior distributions will be shown in Table 3.1.

Variable		Prior distribution		Posterior distribution
x_t	$\sim$	$\mathcal{N}(\hat{x}_{i,t}^-, \hat{P}_{i,t}^*)$	$\rightarrow$	$\mathcal{N}(\hat{x}_{i,t}^+, \hat{P}_{i,t}^+)$
P_t	$\sim$	$i\mathcal{W}(\Psi_{i,t}^-, \psi_{i,t}^-)$	$\rightarrow$	$i\mathcal{W}(\Psi_{i,t}^*, \psi_{i,t}^\star)$
R_t	$\sim$	$i\mathcal{W}(\Phi_{i,t}^-, \phi_{i,t}^-)$	$\rightarrow$	$i\mathcal{W}(\Phi_{i,t}^+, \phi_{i,t}^+)$

Table 3.1: Priors and posteriors

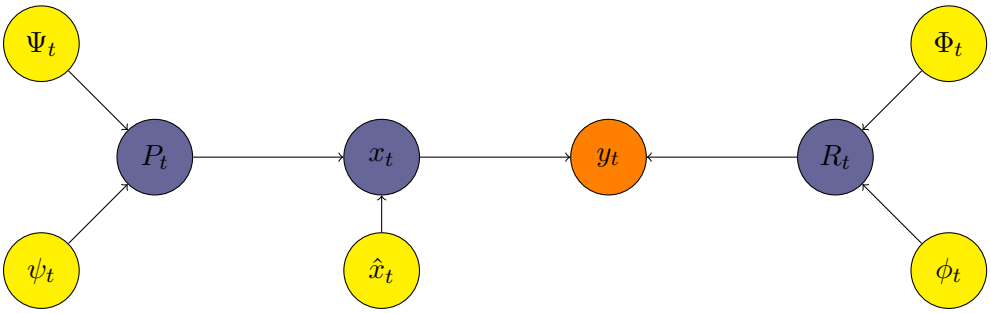


Figure 3.1: Graphical model of message passing algorithm

The graphical representation of this model is shown in Figure 3.1. We can now define how will the factors $\rho(R_t)$ and $\rho(P_t)$ look. We will start with $\rho(R_t)$.

Definition 3.2.3. Let us suppose, that we have a positive definite measurement noise covariance matrix $R_t \in \mathbb{R}^{m\times m}$. Then a convenient model for estimation is the inverse-Wishart distribution, $R_t \sim i\mathcal{W}(\Psi_t, \psi_t)$ with hyperparameters Ψ_t $\mathbb{R}^{m\times m}$ and $\psi_t > 0$. Its probability density function can be

written in the following form,

$$\pi_R(R_t|\psi_t,\Psi_t) = \frac{|\Psi_t|^{\frac{\psi_t}{2}}}{2^{\frac{n\psi_t}{2}}\Gamma_n(\frac{\psi_t}{2})} \times |R_t|^{-\frac{\psi_t+n+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{Tr}(\Psi_t(R)^{-1})\right\}$$
$$\propto \left\{ -\frac{1}{2}\mathrm{Tr}\left(\underbrace{\begin{bmatrix} R_t^{-1} \\ \ln|R_t| \end{bmatrix}^T}_{\eta_{R_t}} \underbrace{\begin{bmatrix} \Psi_t \\ \psi_t + m + 1 \end{bmatrix}}_{\xi_{R_t}} \right) \right\}, \tag{3.24}$$

where $\Gamma_n(\cdot)$ is a multivariate gamma function and the expected values of R_t and R_t^{-1} are the following,

$$\mathbb{E}[R_t] = \hat{R}_t = \frac{\Psi_t}{\psi_t - m - 1}, \tag{3.25}$$
$$\mathbb{E}[R_t^{-1}] = \hat{R}_t^{-1} = \psi_t \Psi_t^{-1}. \tag{3.26}$$

Next for the form of $\rho(P_t)$.

Definition 3.2.4. Assume a positive definitive process noise covariance matrix $P_t \in \mathbb{R}^{n\times n}$. Then the best model for its estimations is going to be an inverse-Wishart distribution $P_t \sim i\mathcal{W}(\Phi_t, \phi_t)$ with hyperparameters $\Phi_t \in \mathbb{R}^{n\times n}$ and $\phi_t > 0$. Therefore we can write its probability density function in the following way,

$$\pi(P_t|\phi_t,\Phi_t) \propto \left\{ -\frac{1}{2}\mathrm{Tr}\left(\underbrace{\begin{bmatrix} P_t^{-1} \\ \ln|P_t| \end{bmatrix}^T}_{\eta_{P_t}} \underbrace{\begin{bmatrix} \Phi_t \\ \phi_t + n + 1 \end{bmatrix}}_{\xi_{P_t}} \right) \right\}. \tag{3.27}$$

Finally, the expected values of P_t and P_t^{-1} will be,

$$\mathbb{E}[P_t] = \hat{P}_t = \frac{\Phi_t}{\phi_t - n - 1}, \tag{3.28}$$
$$\mathbb{E}[P_t^{-1}] = \hat{P}_t^{-1} = \phi_t \Phi_t^{-1}. \tag{3.29}$$

We can see that the Gaussian distribution of x_t is conjugate to the measurement model $f(y_t|\theta_t)$ with fixed R_t. The inverse-Wishart distribution of P_t is then conjugate to the distribution of x_t and the inverse-Wishart distribution of R_t is conjugate to the measurement model $f(y_t|\theta_t)$ with fixed x_t.

Using the Definitions 1.1.1 and 1.1.2 we can derive the CAVI updates, as shown in Equation (3.18). The exact form of the sufficient statistics will be shown. We will substitute the true parameter values with point estimates.

The iterations of CAVI algorithm will be denoted by $d = 1, \ldots, D$. In the first iteration $d = 1$, we set $\hat{P}_t^{+,(0)} = \hat{P}_t^-$ and $\hat{x}_t^{+,(0)} = \hat{x}_t^-$. Then after the last iteration D, we use the variational factors $\rho_i(R_t)$ and $\rho_i(x_t)$ as the posterior distributions, whose hyperparameters enter the subsequent prediction step at the next instance $t + 1$.

Now we will show how to do the update of $\rho_i(P_t) \equiv i\mathcal{W}(\Psi_{i,t}^-, \psi_{i,t}^-)$.First, we will show the expectation of the sufficient statistic $T_{P_t}(x_t)$.

$$\mathbb{E}_{\rho_i(x_t)}^{(d)}[T_{P_t}(x_t)] = \begin{bmatrix} \hat{P}_{i,t}^{+,(d-1)} + (\hat{x}_{i,t}^{+,(d-1)} - \hat{x}_{i,t}^-)(\hat{x}_{i,t}^{+,(d-1)} - \hat{x}_{i,t}^-)^T \\ 1 \end{bmatrix}. \tag{3.30}$$

Next, we have to update the hyperparameter $\Xi_{P_t,i}^-$

$$\begin{aligned} \Xi_{P_t,i}^{*,(d)} &= \Xi_{P_t,i}^- + \mathbb{E}_{\rho_i(x_t)}^{(d)}[T_{P_t}(x_t)] \\ &= \begin{bmatrix} \Psi_{i,t}^- \\ \psi_{i,t}^- + n + 1 \end{bmatrix} + \mathbb{E}_{\rho_i(x_t)}^{(d)}[T_{P_t}(x_t)] \\ &= \begin{bmatrix} \Psi_{i,t}^{*,(d)} \\ \psi_{i,t}^{*,(d)} + n + 1 \end{bmatrix}. \end{aligned} \tag{3.31}$$

Last but not least, we have the point estimate,

$$\hat{P}_{i,t}^{*,(d)} = \mathbb{E}_{\rho_i(P_t)}^{(d)}[P_t^*] = \frac{\Psi_{i,t}^{(d)}}{\psi_{i,t}^{(d)} - n - 1}, \tag{3.32}$$

$$(\hat{P}_{i,t}^{*,(d)})^{-1} = \mathbb{E}_{\rho_i(P_t)}^{(d)}[(P_t^*)^{-1}] = \psi_{i,t}^{*,(d)}(\Psi_{i,t}^{*,(d)})^{-1}. \tag{3.33}$$

Where we use the symbol $*$ to label the intermediate value, that will be used in the measurement update of $\rho_i(x_t)$. In order to proceed, we have to show how to do the update of $\rho_i(R_t) \equiv i\mathcal{W}(\Phi_{i,t}^-, \phi_{i,t}^-)$. Lets start by writing down the expectation of the sufficient statistic $T_{R_t,i}(y_{i,t})$.

$$\mathbb{E}_{\rho_i(x_t,P_t)}^{(d)}[T_{R_t}(y_{i,t})] = \begin{bmatrix} (y_{i,t} - H_t\hat{x}_{i,t}^{+,(d-1)})(y_{i,t} - H_t\hat{x}_{i,t}^{+,(d-1)})^T + H_t\hat{P}_{i,t}^{+,(d-1)}H_t^T \\ 1 \end{bmatrix}. \tag{3.34}$$

Similarly, we can also write down the update of the hyperparameter $\Xi_{R_t,i}^-$.

$$\begin{aligned} \Xi_{R_t,i}^{+,(d)} &= \Xi_{R_t,i}^- + \mathbb{E}_{\rho_i(x_t,P_t)}[T_{R_t}(y_{i,t})] \\ &= \begin{bmatrix} \Phi_{i,t}^- \\ \phi_{i,t}^- + n + 1 \end{bmatrix} + \mathbb{E}_{\rho_i(x_t,P_t)}[T_{R_t}(y_{i,t})] \\ &= \begin{bmatrix} \Phi_{i,t}^{(d)} \\ \phi_{i,t}^{(d)} + n + 1 \end{bmatrix}. \end{aligned} \tag{3.35}$$

Finally, we can also write down individual point estimates.

$$\hat{R}_{i,t}^{(d)} = \mathbb{E}_{\rho_i(R_t)}^{(d)}[R_t] = \frac{\Phi_{i,t}^{(d)}}{\phi_{i,t}^{(d)} - n - 1}, \tag{3.36}$$

$$(\hat{R}_{i,t}^{(d)})^{-1} = \mathbb{E}_{\rho_i(R_t)}^{(d)}[R_t^{-1}] = \phi_{i,t}^{(d)}(\Phi_{i,t}^{(d)})^{-1}. \tag{3.37}$$

Now we miss a last update, being the update of $\rho_i(x_t) \equiv \mathcal{N}(\hat{x}_{i,t}^{-}, \hat{P}_{i,t}^{*,(d)})$. Let us again start with expectation of sufficient statistic $T_{x_t}(y_{i,t})$, that will be following.

$$\mathbb{E}_{\rho_i(R_t,P_t)}[T_{x_t}(y_{i,t})] = \begin{bmatrix} y_{i,t}^T \\ H_t^T \end{bmatrix} (\hat{R}_t^{(d)})^{-1} \begin{bmatrix} y_{i,t}^T \\ H_t^T \end{bmatrix}^T. \tag{3.38}$$

We proceed with the update of the hyperparameter $\Xi_{x_t,i}^{-}$.

$$\begin{aligned} \Xi_{x_t,i}^{+,(d)} &= \Xi_{x_t,i}^{-} + \mathbb{E}_{\rho_i(R_t,P_t)}[T_{x_t}(y_{i,t})] \\ &= \begin{bmatrix} (\hat{x}_{i,t}^{-})^T \\ I \end{bmatrix} (\hat{P}_{i,t}^{*,(d)})^{-1} \begin{bmatrix} (\hat{x}_{i,t}^{-})^T \\ I \end{bmatrix} + \mathbb{E}_{\rho_i(R_t,P_t)}[T_{x_t}(y_{i,t})] \\ &= \begin{bmatrix} (\hat{x}_{i,t}^{+,(d)})^T \\ I \end{bmatrix} (P_{i,t}^{+,(d)})^{-1} \begin{bmatrix} (\hat{x}_{i,t}^{+,(d)})^T \\ I \end{bmatrix}. \end{aligned} \tag{3.39}$$

In the end, we reach the point estimates.

$$\hat{P}_{i,t}^{+,(d)} = [(\hat{P}_{i,t}^{*,(d)})^{-1} + H_t^T \hat{R}_{i,t}^{-1} H_t] \tag{3.40}$$

$$\hat{x}_{i,t}^{+,(d)} = \hat{P}_{i,t}^{+,(d)}[\hat{P}_{i,t}^{*,(d)})^{-1}\hat{x}_{i,t}^{-,(d)} + H_t^T(\hat{R}_{i,t}^{(d)})^{-1}y_{i,t}] \tag{3.41}$$

Finally, we have shown how to update all our factors with the new measurements and how to form their respective point estimates. For better understanding, we have provided the visualization of the complete model of message passing in Figure 3.2.

3.2.2 Prediction

In the following Section we will show how to run the prediction on our proposed filter. As shown in Section 1.2, in case of standard Kalman filter the prediction step transforms the posterior estimate $x_{i,t-1}^{+}$ and its covariance matrix $P_{i,t-1}^{+}$ to their respective prior estimate $x_{i,t}^{-}$ with $P_{i,t}^{-}$ for the time t.

However in our case, we need have the issue of uncertainty about all elements of $\theta_t = [x_t, R_t, P_t]$, but the process model, shown in Equation (3.4), applies only to the estimate of x_t and the related estimate of P_t. Even though

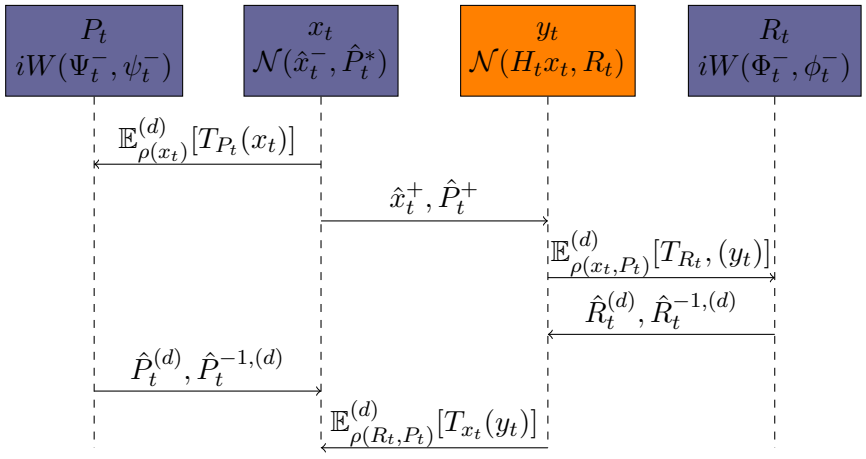


Figure 3.2: Scheme of message passing algorithm

we have issues with the ignorance of Q_t, as well as evolution model for R_t, the prediction step should still prepare the prior distributions for the subsequent measurement update step, as described in Section 3.2.1.

First, we will focus on the variational factor $\rho_i(x_t) \equiv \mathcal{N}(\hat{x}^+_{i,t-1}, \hat{P}^+_{i,t-1})$. If we take a closer look at Equation (3.9), we can see that the only difference in our case is our lack of true covariance matrix $P^+_{i,t-1}$, hence we replace it with our estimate $\hat{P}^+_{i,t-1}$. Last problem is that we do not know the true process noise covariance matrix Q_t, it is obvious that Q_t has an impact on the uncertainty of estimation of x_t, therefore we use a crude estimate $\hat{Q}_{i,t}$ in order to compensate for this. Using Equation (3.8), we can form the predicted Gaussian distribution as,

$$\begin{aligned} \hat{x}^-_{i,t} &= A_t \hat{x}^+_{i,t-1} + B_t u_t, \\ \hat{P}^-_{i,t} &= A_t \hat{P}^+_{i,t-1} A_t^T + \hat{Q}_{i,t}. \end{aligned} \tag{3.42}$$

The question about how to select the optimal $\hat{Q}_{i,t}$ will be discussed in Section 3.2.3. Following this, we need to predict the R_t. Unfortunately, since we ignore the evolution model, we have no way to directly predict $\hat{R}_{i,t}$ from $\hat{R}_{i,t-1}$. Our solution proposed solution to this issue is, given that the R_t has a slow variability in time, then it can be solved using an exponential forgetting with a factor $\alpha_R \in [0, 1]$, as shown [13].

$$\rho_i(R_t) = [\rho_i(R_t)]^{\alpha_R}, \tag{3.43}$$

In typical scenarios the forgetting factor would not be set lower than 0.95. If we rewrite the Equation (3.43) in terms of hyperparameter $\Xi^-_{R_{i,t}}$, we get

$$\Xi^-_{R_{i,t}} = \alpha_R \Xi^+_{R_{i,t-1}}. \tag{3.44}$$

This directly affects the amount of information contained in the distribution and increases the uncertainty of R_t, but it helps us to get the concentration of the distribution closer to the true R_t.

Now, all that remains to be shown is the construction of a new factor $\rho_i(P_t)$. We know that its expectation should equal the predicted value, shown in Equation (3.42).

$$\mathbb{E}_{\rho_i(P_t)}[P_t] = \frac{\Psi_{i,t}^-}{\psi_{i,t}^- - n - 1} = \hat{P}_{i,t}^-. \tag{3.45}$$

Just as in the case of R_t, we have a desire to admin some uncertainty about its value. We know, that $\psi_{i,t-1}^+$ can be interpreted as data counter, hence we set $\psi_{i,t}^-$ and $\Psi_{i,t}^-$ in the following way,

$$\begin{aligned} \psi_{i,t}^- &= \psi_{i,t-1}^+, \\ \Psi_{i,t}^- &= (\psi_{i,t}^- - n - 1)\hat{P}_{i,t}^-. \end{aligned} \tag{3.46}$$

With this approach, the uncertainty of about P_t will decrease when more data are incorporated.

3.2.3 $\hat{Q}_{i,t}$ optimization

As we have stated in Section 3.2.2, in the prediction step Equation (3.42) we use a crude estimate $\hat{Q}_{i,t}$ of the process noise covariance matrix Q_t. Now we will show our approach to chossing the best estimate.

Of course, the simplest approach is to use a single heuristic value, this approach is taken in Article [10], but we have instead decided to use a relatively computationally cheap method of testing a set of C candidates, $\hat{Q}_{i,t}^{(1)}, ;\hat{Q}_{i,t}^{(C)}$ and searching for one that increases the estimation stability the most.

We know that if the space-model, as established in Equations (3.4) and (3.5), was correct and fully known, then the prior predictive distribution of form,

$$f(y_{i,t}|y_{0:t-1}, u_{0:t}) = \int f(y_{i,t}|\theta_t)\pi_i(\theta_t|y_{0:t-1}, u_{0:t})d\theta_t, \tag{3.47}$$

can nicely explain the measurements $y_{i,t}$. Using the plug-in principle we can substitute the true R_t with its point estimate, hence the role of $\pi_i(\theta_t|y_{0:t-1}, u_{0:t})$ will be taken over by $\rho_i(x_t)$. Then we can rewrite $f(y_{i,t}|y_{0:t-1}, u_{0:t})$ into the following form.

$$\begin{aligned} f(y_{i,t}|y_{0:t-1}, u_{0:t}) &= \int \mathcal{N}(H_t x_t, \hat{R}_{i,t}^-) \times \mathcal{N}(\hat{x}_{i,t}^-, \hat{R}_{i,t}^-)dx_t \\ &= \int \mathcal{N}(H_t x_t, \hat{R}_{i,t}^-) \times \mathcal{N}(\hat{x}_{i,t}^-, A_t\hat{P}_{i,t-1}^+A_t^T + Q_t)dx_t \\ &= \mathcal{N}(H_t\hat{x}_{i,t}^-, \hat{R}_{i,t}^- + A_t\hat{P}_{i,t-1}^+A_t^T + Q_t). \end{aligned} \tag{3.48}$$

It can be seen that the better of an estimate $\hat{Q}_{i,t}$ we use instead of the true Q_t, the higher the value of the predictive probability density function for a particular measurement y_t will be. We can get even more use of this, in the case of distributed setting, since there are a lot more measurements to take in. Using this principle, we can tune the $\hat{Q}_{i,t}$, hence if we have C candidate matrices

$$\mathcal{Q}_{i,t} = \{\hat{Q}_{i,t}^{(1)}, \ldots, \hat{Q}_{i,t}^{(C)}\}, \tag{3.49}$$

we can test their impact on the predictive formula, shown in Equation 3.48 by inserting them into the formula. Afterward, we can simply select the one that maximizes the prior predictive probability density function value, in the form of

$$\hat{Q}_{i,t} = \underset{\tilde{Q}_t \in \mathcal{Q}_{i,t}}{\operatorname{argmax}} \log \mathcal{N}(y_{i,t} | H_t \hat{x}_{i,t}^-, \hat{R}_{i,t}^- + A_t \hat{P}_{i,t-1}^+ A_t^T + \tilde{Q}_t). \tag{3.50}$$

As has already been said, this approach is especially useful for the case of distributed setting, which we will show in Section 3.3. Finally, we can write the whole algorithm down, in Algorithm 2.

Algorithm 2: Local variational Kalman filtering under unknown Q_t and R_t

1 Initialization: Set the hyperparameters of the initial prior densities $\rho_i(x_t)$, $\rho_i(P_t)$ and $\rho_i(R_t)$. Set the forgetting factor α_R. Prepare a set $\mathcal{Q}_{i,t}$ of candidate process noise covariance matrices and set the number D of CAVI iterations.

2 **foreach** $t \in [1, 2, \ldots]$ *and measurements* $y_{i,t}$ **do**

Prediction:

1. Predict $\rho_i(R_t)$: Evaluate $\xi_{R,t}^- = \alpha_R \xi_{R,t-1}$, Equation (3.44).
2. Select $\hat{Q}_{i,t}$, Equation (3.50).
3. Predict $\rho_i(x_t)$: Evaluate $\hat{x}_{i,t}^-$ and $\hat{P}_{i,t}^-$, Equation (3.42).
4. Predict $\rho_i(P_t)$, Equation (3.45).

Kalman update:

foreach $d \in [1, \ldots, D]$ **do**

1. Update $\rho_i(P_t)$:
   - Calculate $\mathbb{E}^{(d)}_{\rho_i(x_t)}[T_{P_t}(x_t)]$, Equation (3.30).
   - Update $\Xi^-_{P_t,i}$, Equation (3.31).
2. Update $\rho_i(R_t)$:
   - Calculate $\mathbb{E}^{(d)}_{\rho_i(x_t,P_t)}[T_{R_t}(y_t)]$, Equation (3.34).
   - Update $\Xi^-_{R_i,i}$, Equation (3.35).
3. Update $\rho_i(x_t)$:
   - Prepare estimates $\hat{P}^{*,(d)}_{i,t}$, $\hat{R}^{(d)}_{i,t}$ and their inverses, Equations (3.28), (3.25), (3.29) and (3.26).
   - Calculate $\mathbb{E}_{\rho_i(R_t,P_t)}[T_{x_t}(y_t)]$, Equation (3.38).
   - Update $\Xi^-_{x_i,i}$, Equation (3.39).

end

3 **end**

3.3 Collaborative filtering with information diffusion

We will devote this Section to modification of the developed sequential estimation procedure, shown in Section 3.2, to a distributed environment using information diffusion strategy, as shown in [14].

The diffusion strategy has the following two steps:

1. **The adaptation step**
   Each agent $i \in \mathcal{I}$ receives the measurements $y_{j,t}$ from all of his neighbours $j \in \mathcal{I}_i$. We can then include those measurements into i's node statistical knowledge using Bayesian update. Shown in Section 3.3.1.

2. **The combination step**
   Each agent $i \in \mathcal{I}$ incorporates the posterior estimate of $\theta_t = [x_t, R_t, P_t]$ of all of his neighbours $j \in \mathcal{I}_i$ into his own statistical knowledge. Shown in Section 3.3.3.

We also show how the optimization of $\hat{Q}_{i,t}$ for node i can be improved by using the measurements of $y_{j,t}$ that are available from the neighbours $j \in \mathcal{I}_i$.

3.3.1 Adaptation step

As we have already stated, one of the steps that are needed in a distributed setting is the adaptation step. Where we solve the issues of incorporating the measurements of $y_{j,t}$ from j node into i-th node statistical knowledge.

The position of the adaptation step is closely intervened with the law of large numbers behavior of the considered estimator, precisely with the speedup of its convergence to the true value. Hence we aim to achieve this, by having each node $i \in \mathcal{I}$ incorporating his neighbours measurements $y_{j,t}, j \in \mathcal{I}_i$ into his own probability distribution, as in Equation (3.10). We can write in the following form,

$$\pi_i(\theta_t|\overline{y}_{0:t}, u_{0:t}) \propto \pi_i(\theta_t|\overline{y}_{0:t-1}, u_{0:t}) \prod_{j\in\mathcal{I}_i} f(y_{j,t}|\theta_t), \tag{3.51}$$

where $\overline{y}_{i,0:t}$ contains all information about the measurements, that are known by the node i. We know that $y_{i,t}$ are i.i.d.,hence the probability density functions $f(y_{j,t}|\theta_i)$ are identically parameterized. From Section 3.2 we know that the distribution belongs to the exponential family with expected sufficient statistics $\mathbb{E}_{\rho_i(R_t,P_t)}[T_{x_t}(y_{j,t})]$ or $\mathbb{E}_{\rho_i(x_t,P_t)}[T_{R_t}(y_{j,t})]$ respectively.

$$\begin{aligned}\prod_{j\in\mathcal{I}_i} f(y_{j,t}|\theta_t) &\propto \prod_{j\in\mathcal{I}_i} \exp\{\eta_{x_t}^T T_{x_t}(y_{j,t})\} \\ &\propto \exp\left\{\eta_{x_t}^T \sum_{j\in\mathcal{I}_i} T_{x_t}(y_{j,t})\right\},\end{aligned} \tag{3.52}$$

with

$$\mathbb{E}_{\rho_i(R_t,P_t)}\left[\sum_{j\in\mathcal{I}_i} T_{x_t}(y_{j,t})\right] = \sum_{j\in\mathcal{I}_i}\begin{bmatrix} y_{j,t}^T \\ H_t^T \end{bmatrix}(\hat{R}_{i,t}^{(d)})^{-1}\begin{bmatrix} y_{j,t}^T \\ H_t^T \end{bmatrix}, \tag{3.53}$$

for the variational message to $\rho_i(x_t)$. Now we can do the same for the variational message to $\rho_i(R_t)$ in the following way,

$$\prod_{j\in\mathcal{I}_i} f_j(y_{j,t}|\theta_j) \propto \exp\left\{\eta_{R_t}^T \sum_{j\in\mathcal{I}_i} T_{R_t}(y_{j,t})\right\}, \tag{3.54}$$

and that will equal to

$$\mathbb{E}^{(d)}_{\rho_i(x_t,P_t)}\left[\sum_{j\in\mathcal{I}_i} T_{R_t}(y_{j,t})\right] = \sum_{j\in\mathcal{I}_i}\begin{bmatrix}(y_{j,t} - H_t\hat{x}_{i,t}^{+,(d-1)})(\bullet)^T + H_t\hat{P}_{i,t}^{+,(d-1)}H_t^T \\ 1\end{bmatrix}, \tag{3.55}$$

where $(a-b)(\bullet)^T$ is used to simplify the notation of the outer product. Now we can use the expected sufficient statistics, shown in Equations (3.53) and (3.55), instead of their single-measurement counterparts in CAVI, shown in Equations (3.38) and (3.34). The local estimates of P_t, R_t and x_t use same equations with their respective hyperparameters $\Xi^{*,(d)}_{P_t,i}$, $\Xi^{+,(d)}_{R_t,i}$ and $\Xi^{+,(d)}_{x_t,i}$. Finally we can deduce the following,

$$\hat{P}_{i,t}^{+,(d)} = \left[(\hat{P}_{i,t}^{*,(d)})^{-1} + |\mathcal{I}_i|H_{i,t}^T(\hat{R}_{i,t}^{(d)})^{-1}H_{i,t}\right]^{-1}, \tag{3.56}$$

$$\hat{x}_{i,t}^{+,(d)} = \hat{P}_{i,t}^{+,(d)}\left[(\hat{P}_{i,t}^{*,(d)})^{-1}\hat{x}_{i,t}^{-,(d)} + H_{i,t}^T(\hat{R}_{i,t}^{(d)})^{-1}\sum_{i\in\mathcal{I}_i} y_{j,t}\right]. \tag{3.57}$$

3.3.2 Optimization of $\hat{Q}_{i,t}$ in distributed setting

Just as we have established in Section 3.2.3, the local optimization of $\hat{Q}_{i,t}$ over the set $\mathcal{Q}_{i,t}$ can benefit from the distributed scenario in form of the additional observations of $y_{j,t}$ as they are provided from neighbours $j \in \mathcal{I}_i$. We know that the observations are i.i.d. and their joint predictive density is just of the product of their individual densities, hence we can modify the Equation (3.50) in the following way,

$$\begin{aligned}\hat{Q}_{i,t} &= \underset{\hat{Q}_t\in\mathcal{Q}_{i,t}}{\operatorname{argmax}} \log\prod_{j\in\mathcal{I}_i}\mathcal{N}(y_{j,t}|H_t\hat{x}_{i,t}^-\hat{R}_{i,t}^- + A_t\hat{P}_{i,t-1}^+A_t^T + \tilde{Q}_t) \\ &= \underset{\hat{Q}_t\in\mathcal{Q}_{i,t}}{\operatorname{argmax}} \sum_{j\in\mathcal{I}_i}\log\mathcal{N}(y_{j,t}|H_t\hat{x}_{i,t}^-\hat{R}_{i,t}^- + A_t\hat{P}_{i,t-1}^+A_t^T + \tilde{Q}_t).\end{aligned} \tag{3.58}$$

3.3.3 Combination step

The last step that has to be done in order to adapt the method to the distributed setting is the combination step. In this step we propose the idea of how to combine the posterior estimate $\theta_t = [x_t, R_t, P_t]$ of neighbours of node i into its statistical knowledge. Hence we will work with the following.

Each agent $i \in \mathcal{I}$ acquires the posterior estimates from its neighbours $j \in \mathcal{I}_i$. For our approach, the posterior estimates are completely expressed by the variational factors $\rho_j(x_t)$ and $\rho_j(R_t)$.

We know that all the information on inferred variables is embraced inside the hyperparameters $\Xi^+_{x_i,j}$ and $\Xi^+_{R_i,j}$, that have accumulated their respective sufficient statistics, hence merging done in combination step should be based upon them.

One of the very interesting approaches is an averaging of those hyperparameters. To be exact, for x_t it means,

$$\tilde{\rho}_i(x_t) \propto \exp\left\{\eta_{x_t}^T \tilde{\Xi}^+_{x_t,i}\right\} = \exp\left\{\eta_{x_t}^T \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \Xi^+_{x_t,j}\right\}, \tag{3.59}$$

and for R_t it will be,

$$\tilde{\rho}_i(R_t) \propto \exp\left\{\eta_{R_t}^T \tilde{\Xi}^+_{R_t,i}\right\} = \exp\left\{\eta_{R_t}^T \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \Xi^+_{R_t,j}\right\}. \tag{3.60}$$

If we analyse the Equations (3.59) and (3.60), we can see that it has a few very appealing properties. Namely:

- **Tractability and numerical stability:** One of the major advantages of this approach compared to other combination rules is that both Equations (3.59) and (3.60) yield the hyperparameters of the same type. Thanks to that, we can immediately use them in the subsequent prediction step, on top of that, this approach does not tend to produce any numerical issues during its computation.

- **Complience with the Bayesian information processing:** To a keen eye, it may become obvious, that the convex combination of hyperparameters of the conjugate prior distributions is equivalent to the weighted Bayesian updating. It becomes even more apparent when we compare the model of multiple measurements, shown in Equation (3.54), with the combined density, as shown in Equation (3.59).

- **Robustness to data incest:** One of the most common problems that can be encountered in a distributed setting is data incest, meaning the issue of encountering the same information multiple times and therefore the same information enters the information processing procedure more

than once. This was explored, for example in Article [15]. Our way to combat this issue is by introducing the factor $\frac{1}{|\mathcal{I}_i|}$ into the Equations (3.59) and (3.60).

- **Kullback-Leibler optimality:** We can show that both fusion rules, shown in Equations (3.59) and (3.60) are KL optimal in the following sense

$$\tilde{\rho}_i(x_t) = \underset{\tilde{\rho}_i(x_t)}{\operatorname{argmin}} \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \mathcal{D}[\overline{\rho}_i(x_t) || \rho_i(x_t)], \tag{3.61}$$

  for case of $\rho_i(R_t)$ the approach will be the same. This was explored in [16].

- **Covariance intersection:** It can be seen, that the KL optimal fusion, as shown in Equation (3.61), when applied to the Gaussian probability density provides a result known as covariance intersection, this was shown in [17] and further explored in [16]. From this, we know that the combined estimates will be

$$\begin{aligned} \tilde{\hat{P}}^{+}_{i,t} &= \left[\frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \left(\hat{P}^{+}_{j,t} \right)^{-1} \right]^{-1}, \\ \overline{\hat{x}}^{+}_{i,t} &= \tilde{\hat{P}}^{+}_{i,t} \left[\frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} (\hat{P}^{+}_{j,t})^{-1} \hat{x}^{+}_{i,t} \right]. \end{aligned} \tag{3.62}$$

Chapter 4

Results evaluation

We have devoted this chapter to present the results of our approach. It will be divided into two main parts, one where the true measurement noise covariance matrix R is static in time and the second where the measurement noise covariance matrix R develops throughout time.

We average all data over 60 independent runs that ran on 60 completely different simulated data. Most scenarios had a length of 1000 samples, with one exception having 1500 samples. All our samples represent simulated 2-dimensional trajectory. The state-space model of this trajectory has the following form,

$$x_k = Ax_{k-1} + w_k, \tag{4.1}$$
$$y_k^{(i)} = Hx_k + v_k^{(i)}, \tag{4.2}$$

where $x_k \in \mathbb{R}^4$ is the unknown state space model and consists of location coordinates $x_{1,k}$ and $x_{2,k}$ for our two dimensional case, as well as, their respective velocities $x_{3,k}$ and $x_{4,k}$. Initial value of x_0 is $[0, 0, 0, 0]^T$. The measurement vector $y_k \in \mathbb{R}^2$ contains only coordinates. We also need to specify the necessary matrices, hence

$$\begin{aligned} A &= \begin{bmatrix} 1 & 0 & \Delta_k & 0 \\ 0 & 1 & 0 & \Delta_k \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \\ H &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \end{aligned} \tag{4.3}$$

where Δ_k is the time difference, where it applies $\Delta_k = 1$. Finally the the i.i.d.

noise variable $w_k \sim \mathcal{N}(0, Q)$, where the true covariance matrix is equal to,

$$Q = \frac{1}{2}\begin{bmatrix} \frac{\Delta_k^3}{3} & 0\frac{\Delta_k^2}{2} & 0 & \\ 0 & \frac{\Delta_k^3}{3} & 0 & \frac{\Delta_k^2}{2} \\ \frac{\Delta_k^2}{2} & 0 & \Delta_k & 0 \\ 0 & \frac{\Delta_k^2}{2} & 0 & \Delta_k \end{bmatrix} \tag{4.4}$$

and measurement noise $v_k^{(i)} \sim \mathcal{N}(0, R^{(i)})$ is also i.i.d.. Since the covariance matrix $R^{(i)}$ differs between scenarios, we will specify them individually, when necessary. The true trajectory is shared amongst all nodes, but each node has its own noisy measurements, different from all other nodes. An example of such trajectory with one variant of noisy measurements can be seen in Figure 4.1.

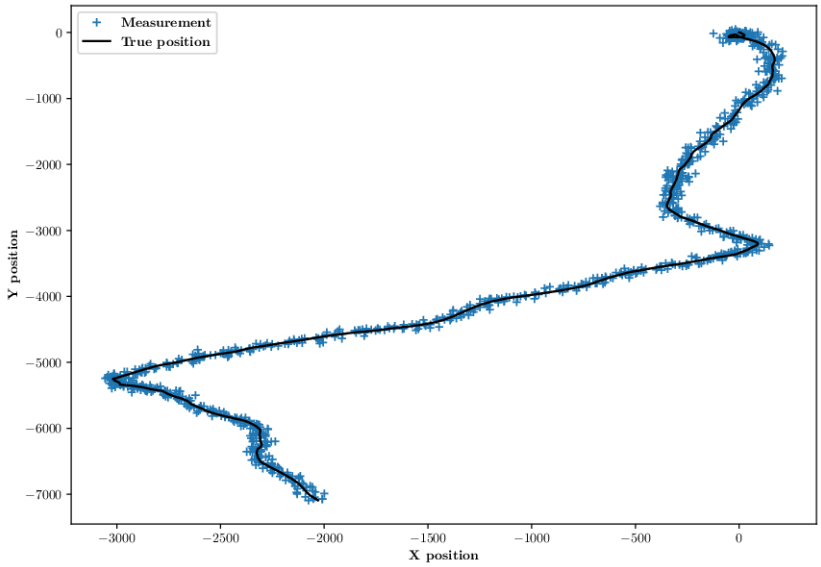


Figure 4.1: Example of true trajectory with noisy measurements

We have chosen a network that consists of $|\mathcal{I}| = 15$ nodes and is the same for all scenarios. Its topology can be seen in Figure 4.2. Some of the initial factors for all nodes are shared throughout all scenarios, namely initial $R^{(i)}$ is for all nodes represented by inverse-Wishard distribution $i\mathcal{W}(4, 100 * I_{[2\times2]})$, initial P is set to $i\mathcal{W}(10, 100 \cdot I_{[4\times4]})$ and lastly initial x_t is zero centered with covariance matrix equal to $100 \cdot I_{[4\times4]}$, wherein all cases I represents an identity matrix. Finally the forget factor for R is set to 0.99 and at each time k, we run $V = 5$ iterations of variational algorithm.

In all scenarios, we compare with the following algorithms:

- **NOCOOP:**
  There is no cooperation at all between nodes, they work only with their

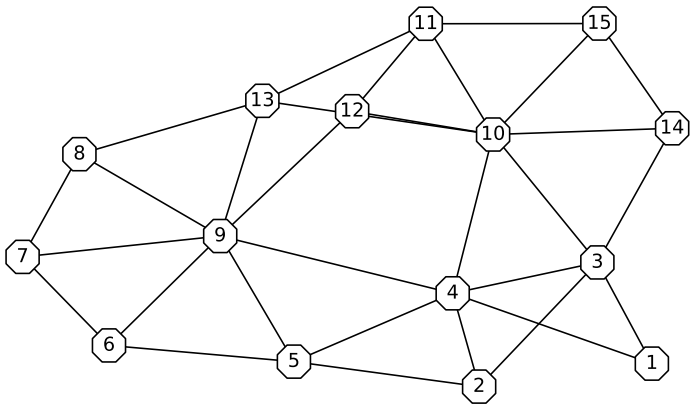


Figure 4.2: Topology of the network

own measurements, and therefore, all of their estimates are based solely on them.

- **ATC:**
  The adapt-then-combine strategy, where measurements and posterior estimates are shared between compatible nodes. As shown in [9].

- **ATCOMP:**
  The fusion center scenario, where all information is procesed by single node.

- **ATCSBW:**
  Our proposed method expands the adapt-then-combine strategy by the optimization of $\hat{Q}_{i,t}$.

For all scenarios, all nodes share the same amount of $\hat{Q}$, from which they can choose during the optimization of $\hat{Q}_{i,t}$, as shown in Section 3.2.3. They available matrices are the following $i \cdot I_{[4\times4]}$, where $i \in [0, |\mathcal{I}|]$ and I is an identity matrix. Each node has its initial $\hat{Q}$ set to one of those options, namely to $i \cdot I_{[4\times4]}$, where i is equal to the index of the node.

4.1 Static R

First, we will look at cases where we have a static R in time. Here we have two examples that differ in time at which they can optimize their $\hat{Q}_{i,t}$. In one variant, the optimization can begin immediately, while in the other, it has to

wait for 15 steps. Both scenarios have trajectories of length 1000. The static true $R^{(i)}$ is in both cases equal to

$$R^{(i)} = \begin{bmatrix} 900 & 0 \\ 0 & 900 \end{bmatrix}. \tag{4.5}$$

4.1.1 $\hat{Q}$ optimization from beginning

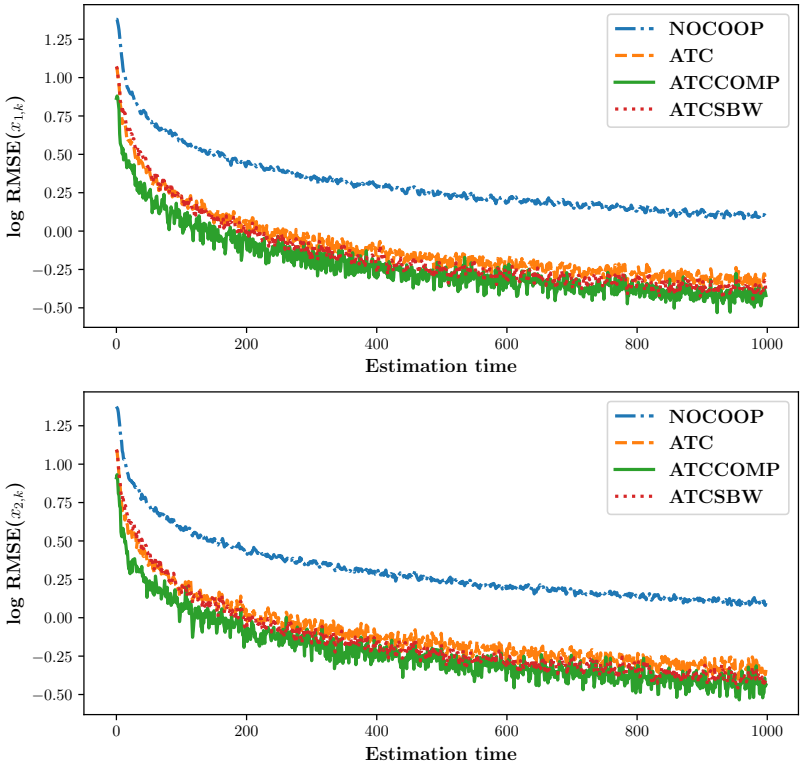


Figure 4.3: Decimal logarithm of average RMSE of state estimates

Now, we will take a closer look on a case where the optimization of $\hat{Q}_{i,t}$ is allowed from the start. In Figures 4.3 and 4.4 is root mean square error[15] of the states $x_{1,k}$ and $x_{2,k}$, as well as of the measurement noise covariance matrix R. All values are averaged over the network. We can see that our approach tends to perform, between **ATC** and **ATCCOMP**. This behaviour can be seen in both estimation of states $x_{1,k}$ and $x_{2,k}$, as seen in Figure 4.3, as well as for estimation of the measurement noise covariance matrix R, as shown in Figure 4.4.

[15]Usually abbreviated as RMSE.

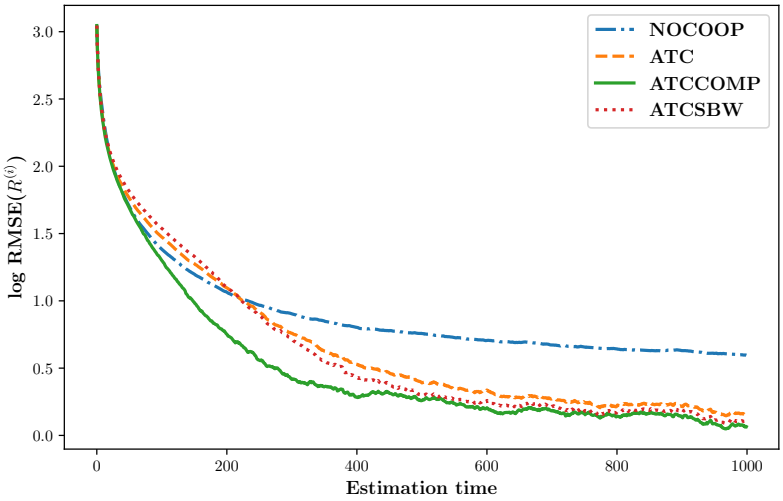


Figure 4.4: Decimal logarithm of average RMSE of measurements noise covariance matrix estimate

We can see that **ATCCOMP** is still better, which is as expected. This is especially noticeable in the case of the estimation of the matrix R, where the **ATCCOMP** converges much faster than our method, but in the end, our estimation is very close to it. A quite similar development can be seen also for the state estimation, Figure 4.3. However due to the higher volatility of the estimates is the exact difference between our approach, **ATC** and **ATCCOMP** more blurry and harder to pinpoint exactly. But there is still larger convergence with the **ATCCOMP** method, which vanishes as the estimates develop.

Finally, we have also provided a visualization of the differences between the individual node $\hat{Q}$ matrices and the true Q matrix. This can be seen in Figure 4.5. Since the nodes can optimize their $\hat{Q}$ right from the start, they tend to vary more, as can be seen at the beginning, where they all converge to a $\hat{Q}$, that is quite far from the true Q, but after a few more steps, all converge to the closest option to the true Q they have available.

4.1.2 $\hat{Q}$ optimization after first 15 steps

We will now take a look at the other scenario with the fixed matrix R. For this one, we have limited the nodes in terms of their ability to optimize their $\hat{Q}$ matrices, in particular, they cannot optimize their $\hat{Q}$ in any way until prediction step 15 is reached.

The estimation of states and of the measurement noise covariance matrix R can be seen in Figure 4.6 and 4.7 respectively. It can be seen that their development is very similar to the previous scenario, as shown in Section 4.1.1.

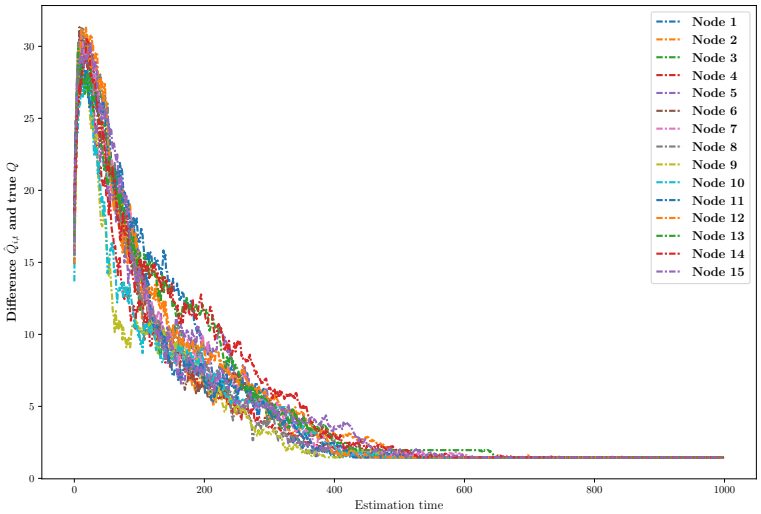


Figure 4.5: Difference of $\hat{Q}_{i,t}$ and true Q for individual nodes

This fact is not surprising since the only difference between the two scenarios is in their $\hat{Q}$ and nodes, in the beginning, tend to optimize their $\hat{Q}$ matrices into values further from the true Q matrix, rather than closer. This is very similar to the behavior that was already seen in previous scenarios. Hence the difference in state and R estimation is not major.

However, the interesting difference could be in the difference between individual nodes $\hat{Q}$ matrices and the true Q matrix. This can be seen in Figure 4.8. Since we have not allowed any optimization of $\hat{Q}$ until the fifteenth step, we can see that they still at first converge to the worse option, rather than the better one. But their option is not as bad as in the previous scenario, and they similarly converge to the optimal value they have at their disposal.

4.2 Varying R

We have devised the other set of scenarios to test the performance when we have the true measurement noise covariance matrix R variable in time. We show two variants of this, one with only increasing R and the other one with increasing and decreasing R. They are shown in Section 4.2.1 and 4.2.2 respectively. In both scenarios, we have allowed the nodes to run the optimization of $\hat{Q}$ right from the start.

4.2.1 Only increasing R

First we will take a look on a scenario, where the true R matrix is only increasing. The precise development of $R_{[0,0]}$ can be seen in Figure 4.9, but

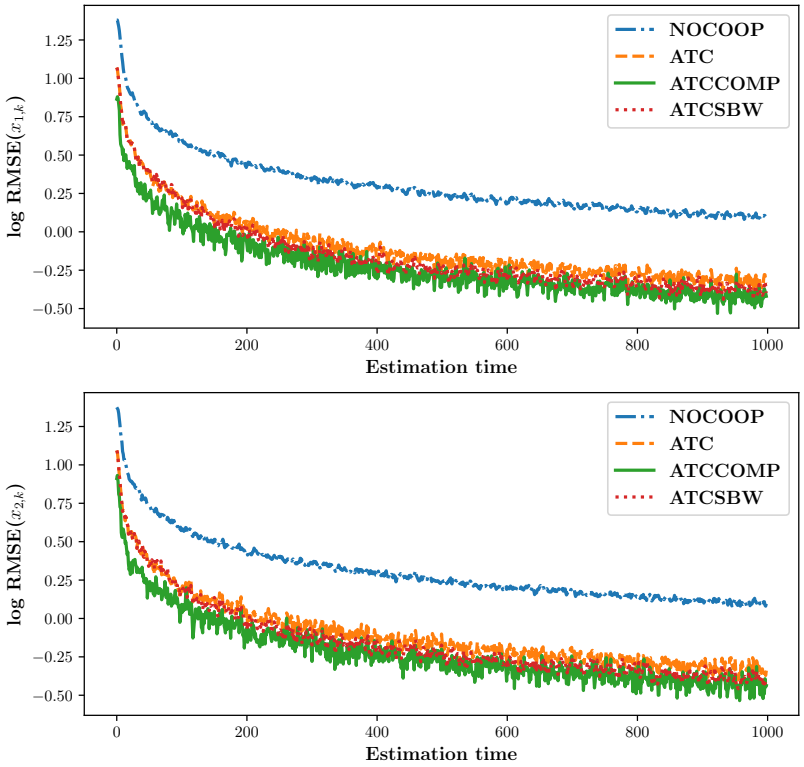


Figure 4.6: Decimal logarithm of average RMSE of state estimates

the development for $R_{[1,1]}$ is exactly the same and both $R_{[1,0]}$ and $R_{[0,1]}$ are always equal to zero.

Now we can take a look at specific of the estimation of both states and R, as shown in Figure 4.10 and 4.11. The estimation of states is very similar to the previous scenarios, which is a very good sign since the estimate did not deteriorate in any severe way, even if the R changes through time.

What is even more interesting is the estimation of R in Figure 4.11. Our interest lies in this, particularly due to the fact that we have varying R in time. However, it shows truly promising results, as our estimate of R is very close, especially in later steps, to the estimate done by **ATCCOMP**.

Just for the sake of completeness, we show the difference between $\hat{Q}$ matrices of individual nodes and the true Q in Figure 4.12. It can be easily seen that the optimization of $\hat{Q}$ is not heavily affected by the increasing R, and its performance is almost identical to the one shown in Section 4.1.1.

4.2.2 Increasing and decreasing R

Finally, for the last scenarios we have set R to at first increase and then decrease to the previous value. In order to accommodate this change in more

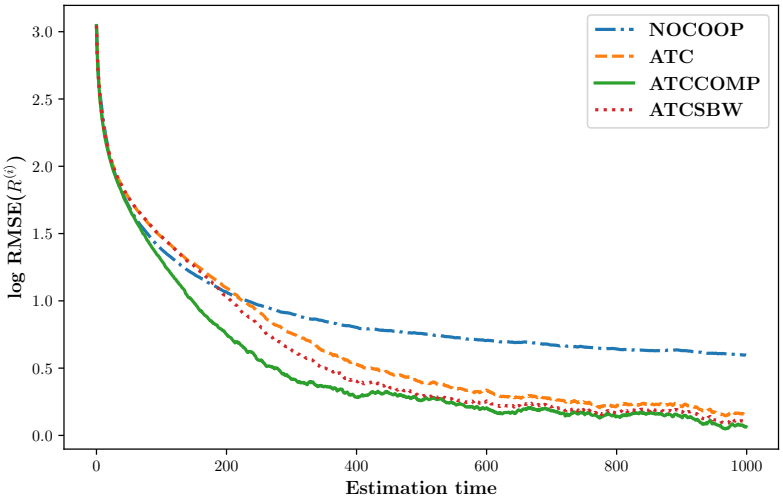


Figure 4.7: Decimal logarithm of average RMSE of measurements noise covariance matrix estimate

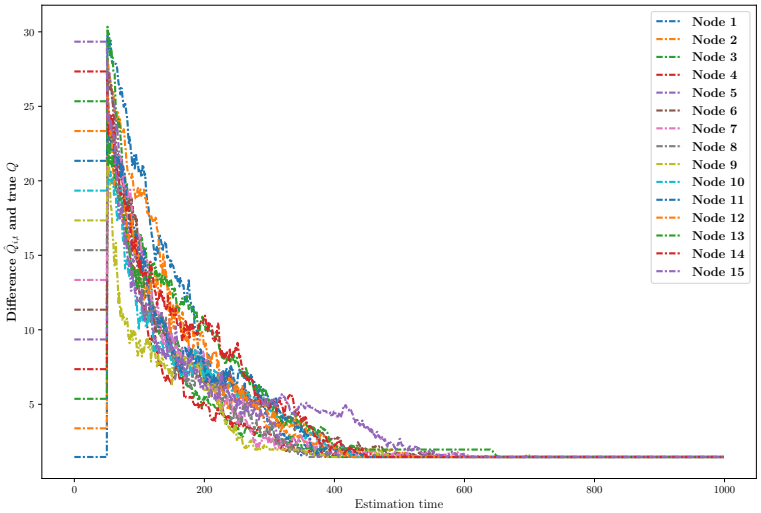


Figure 4.8: Difference of $\hat{Q}_{i,t}$ and true Q for individual nodes

procedural matter, we have extended this prediction from 1000 steps to a 1500 steps. As in Section 4.2.2, we show the development of $R_{[0,0]}$ in Figure 4.13 and just as in last time, the $R_{[1,1]}$ follows the same development and $R_{[1,0]}$ and $R_{[0,1]}$ is always zero.

Now we can take a look at how good our estimation of states and matrix R. We show them in Figures 4.14 and 4.15 respectively. Remarkably, we can see that the estimation of states does not suffer any apparent penalty even though the true R matrix varies severely in time. Our estimation of states is

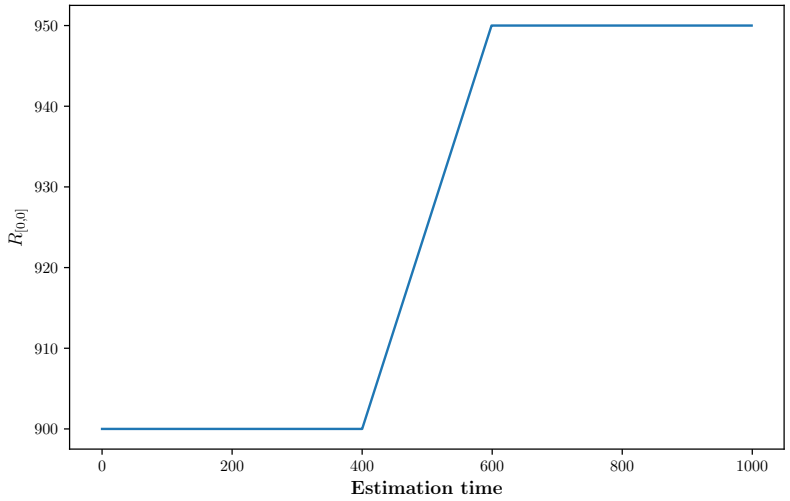


Figure 4.9: Development of R

still very similar to the **ATCCOMP**.

Just as in the previous scenario, we have a profound interest in the estimation of the true R since it is what changes in this particular scenario. This can be seen Figure 4.15. Astonishingly, the performance is still really good and comes very close to the performance of **ATCCOMP**. Just as in all previous cases, the **ATCCOMP** converges much faster, but our method converges quite fast too and, in the end, has basically the same as of **ATCCOMP**.

Finally, to round everything up and hold the same visualizations for all scenarios, we again show the difference between the individual nodes $\hat{Q}$ matrices and the true Q matrix in Figure 4.16. We can see that the development of $\hat{Q}$ matrices is again very close to both previous scenarios, as shown in Section 4.1.1 and 4.2.1. Their behavior is so similar that their difference could be labeled as negligible.

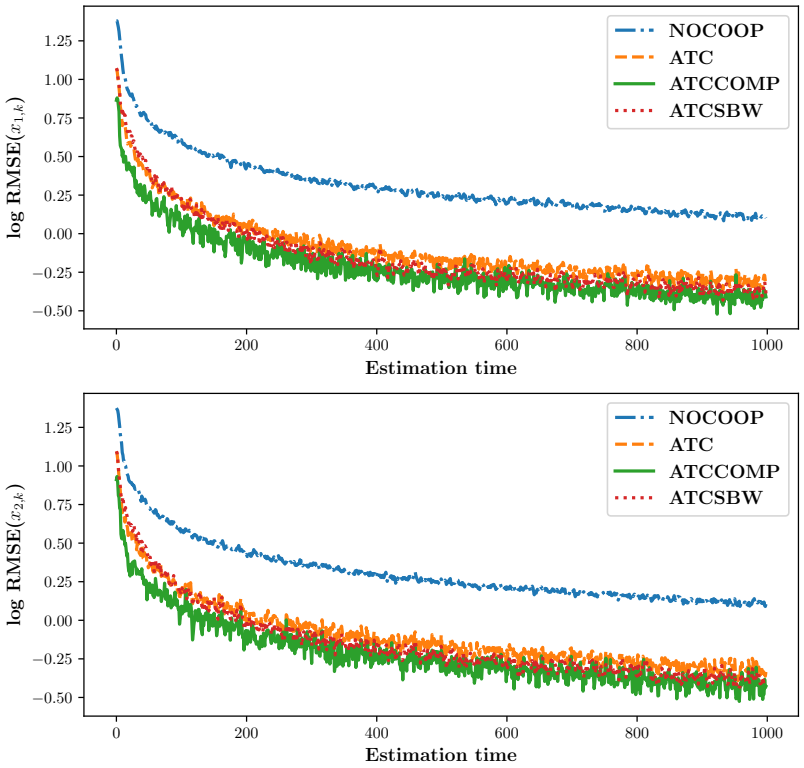


Figure 4.10: Decimal logarithm of average RMSE of state estimates

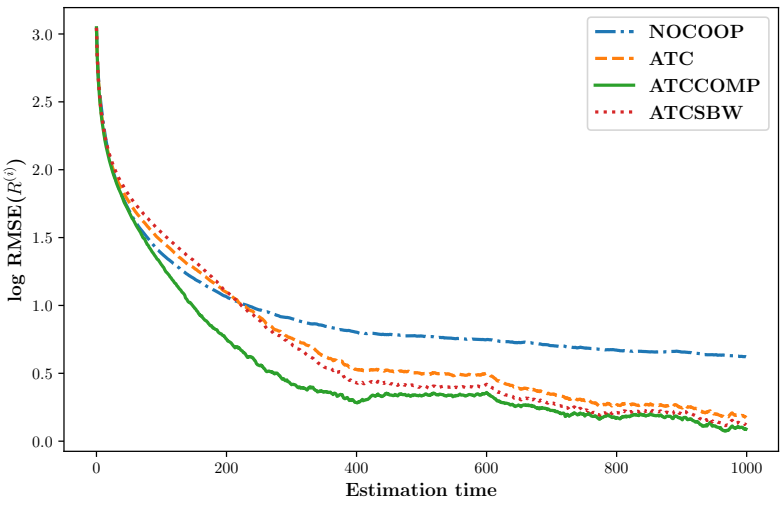


Figure 4.11: Decimal logarithm of average RMSE of measurements noise covariance matrix estimate

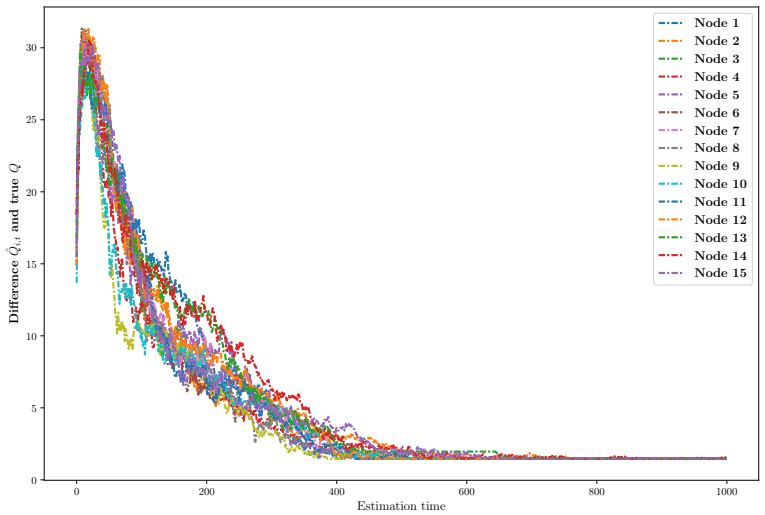


Figure 4.12: Difference of $\hat{Q}_{i,t}$ and true Q for individual nodes

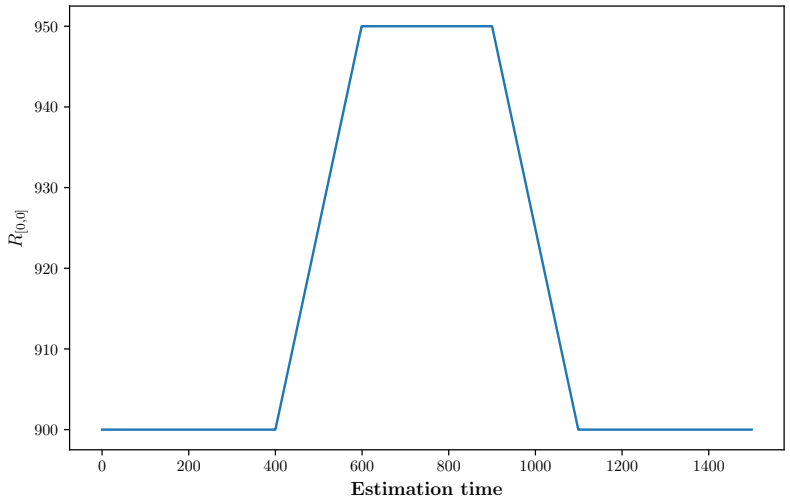


Figure 4.13: Development of R

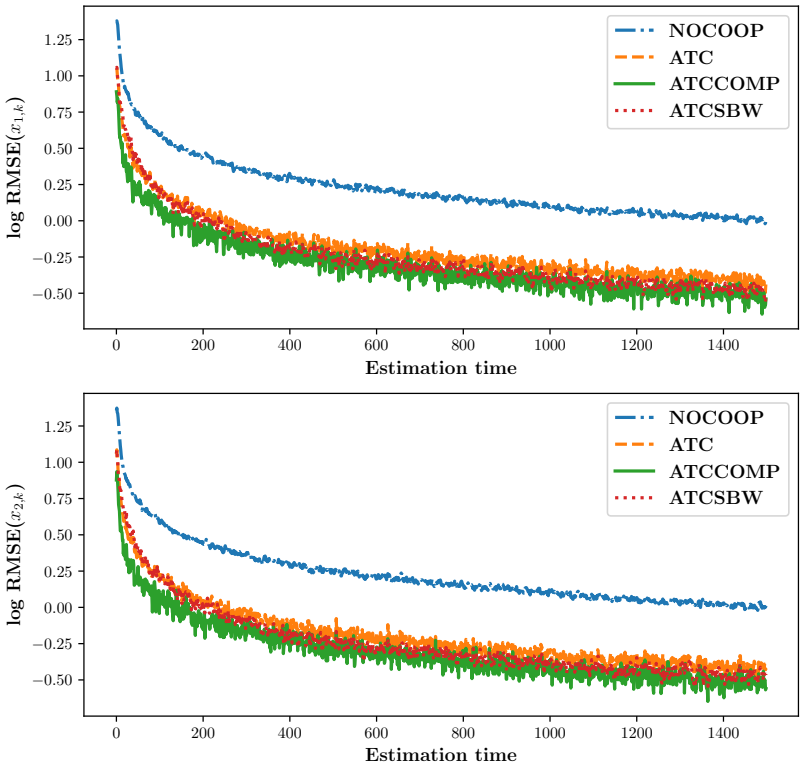


Figure 4.14: Decimal logarithm of average RMSE of state estimates

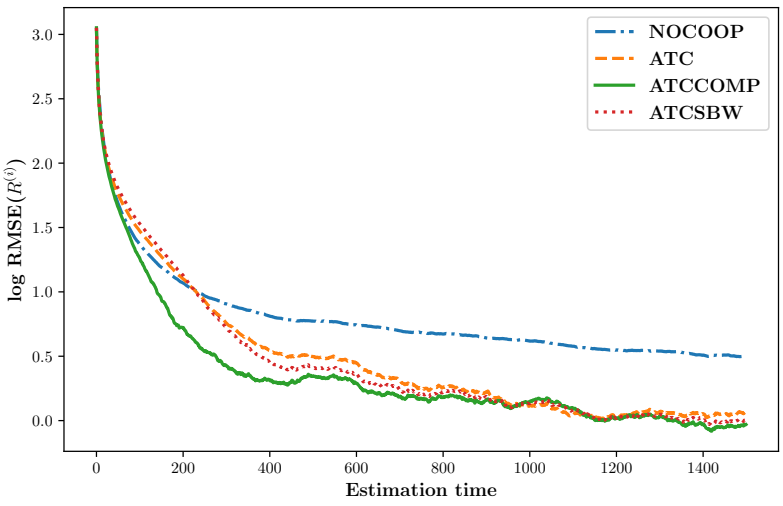


Figure 4.15: Decimal logarithm of average RMSE of measurements noise covariance matrix estimate

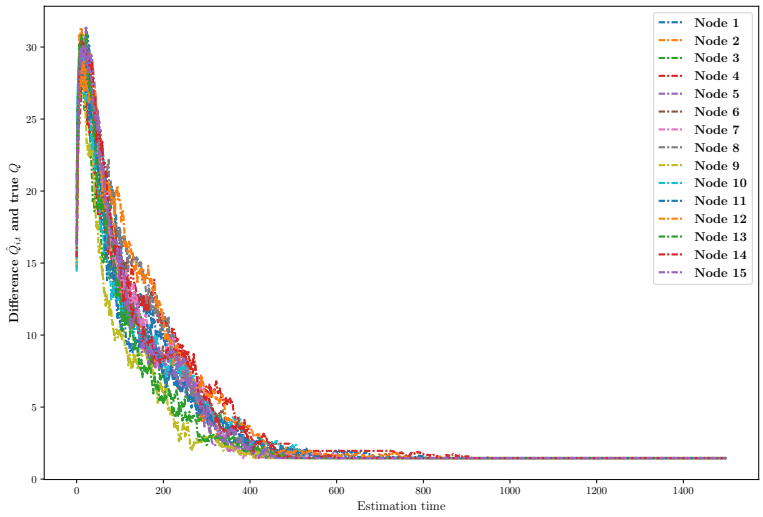


Figure 4.16: Difference of $\hat{Q}_{i,t}$ and true Q for individual nodes

Conclusion

We use this final chapter to summarize what has been done and also what kind of enhancements could be done in the future.

Thesis summary

The goal of this thesis was set to explore the field of Bayesian sequential estimation of unknown states of the state-spacemodels, with unknown covariance matrices for both process and measurement noise, and propose a method that could achieve such goals.

The necessary prerequisites are explained in Chapter 1. Following that, we have used Chapter 2 for a brief exploration of the current state-of-the-art in this field. We have primarily shown approaches that are similar to our approach and pointed on some similarities, as well as differences between theirs and our approach. Then we have devoted the Chapter 3 to elaborate on our method, which we have proposed to achieve the goal of this thesis. We first establish the method in a local setting and then expand this definition to accommodate necessary aspects for the distributed setting. Finally, Chapter 4 is used to discuss the result of our proposed method and comparison with other state-of-the-art methods.

Future works

As we have already stated, this thesis's primary topic is in the development of a filtering method for the case with unknown covariance matrices of both process and measurement noise and its adaptation for the distributed setting. But we still believe that there is a potential for the future expansion of this work, in particular in the following topics.

- **Filtering under nonlinear state-space models**
  One of the intensive research topics is the issue of nonlinear state-space

models, for example, [11]. We can adapt our approach to accommodate this approach as well.

- **Use of normal inverse-Wishart**
  Another possible improvement could be the use of compound distribution. In our case, we can replace the normal distribution that we use to model x and the inverse-Wishart distribution, that we use for modeling of matrix P, with one normal inverse-Wishart distribution.

- **Tuning of $\hat{Q}_{i,t}$**
  The last expansion that we will propose here is the opportunity to improve the optimization of $\hat{Q}_{i,t}$. One of the possibilities for this improvement would, for example, be sampling from the neighborhood of the candidate values. Another example would be sharing of $\hat{Q}_{i,t}$ between a set of neighbor nodes.

Bibliography

[1] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[2] B. O. Koopman, "On distributions admitting a sufficient statistic," *Transactions of the American Mathematical society*, vol. 39, no. 3, pp. 399–409, 1936.

[3] H. Raiffa and R. Schlaifer, "Applied statistical decision theory." Tech. Rep., 1961.

[4] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, [1st ed.] ed. Cambridge: The Mit Press, 2006.

[5] K. Dedecius, "Bayesian machine learning - kalman filter," [online], 2018, accessed: 2021-04-04. [Online]. Available: https://github.com/kamil-dedecius/bml/tree/master/prednasky

[6] J. Winn, C. M. Bishop, and T. Jaakkola, "Variational message passing." *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[7] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[8] S. Sarkka and A. Nummenmaa, "Recursive noise adaptive kalman filtering by variational bayesian approximations," *IEEE Transactions on Automatic control*, vol. 54, no. 3, pp. 596–600, 2009.

[9] K. Dedecius and O. Tichỳ, "Collaborative sequential state estimation under unknown heterogeneous noise covariance matrices," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5365–5378, 2020.

[10] Y. Huang, Y. Zhang, Z. Wu, N. Li, and J. Chambers, "A novel adaptive kalman filter with inaccurate process and measurement noise covariance matrices," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 594–601, 2017.

[11] S. Särkkä and J. Hartikainen, "Non-linear noise adaptive kalman filtering via variational bayes," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2013, pp. 1–6.

[12] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[13] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System identification*. Elsevier, 1981, pp. 239–304.

[14] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. ARTICLE, pp. 311–801, 2014.

[15] S. McLaughlin, V. Krishnamurthy, and S. Challa, "Managing data incest in a distributed sensor network," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 5. IEEE, 2003, pp. V–269.

[16] K. Dedecius and P. M. Djurić, "Sequential estimation and diffusion of information over networks: A bayesian approach with exponential family of distributions," *IEEE Transactions on Signal Processing*, vol. 65, no. 7, pp. 1795–1809, 2016.

[17] S. J. Julier and J. K. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations," in *Proceedings of the 1997 American Control Conference (Cat. No. 97CH36041)*, vol. 4. IEEE, 1997, pp. 2369–2373.

Appendix A

Acronyms

pdf Probability density function

i.i.d. Independent identically distributed

ELBO Evidence lower bound

KL divergence Kullbach-Leibler divergence

CAVI Coordinate-ascent variational inference

RMSE Root mean square error

Appendix B

Contents of enclosed SD card

- `readme.txt` ..... the file with SD card contents description
- `implementation` ..... the directory of code of the implementation
- `text` ..... the directory of source codes
  - `thesis` ..... the directory of LaTeX source codes of the thesis
  - `thesis.pdf` ..... the thesis text in PDF format