Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Tomáš Vlk, Kamil Dedecius

Diffusion Kalman filtering under unknown process and measurement noise covariance matrices

No. 2395 October 17, 2022

ÚTIA AV ČR, P.O.Box 18, 182 08 Prague, Czech Republic
Tel: +420 266052570, Fax: +420 266052068, Url: http://www.utia.cas.cz,
E-mail: dedecius@utia.cas.cz

Abstract

The state-of-the-art algorithms for Kalman filtering in agent networks with information diffusion impose the requirement of well-defined state-space models. In particular, they assume that both the process and measurement noise covariance matrices are known and properly set. This is a relatively strong assumption in the signal processing domain. By design, the Kalman filters are rather sensitive to its violation, which may potentially lead to their divergence. In this paper we propose a novel distributed filtering algorithm with an increased robustness under unknown process and measurement noise covariance matrices. It is formulated as a Bayesian variational message passing procedure for simultaneous analytically tractable inference of states and measurement noise covariance matrices. The ignorance of the process noise covariance matrix is overcome by a cheap hypotheses-testing procedure and an optimization of a relevant prior distribution. The solution is robust to slow variations of the covariances and has a very low number of tuning parameters. The simulation results demonstrate that the estimation performance is close to the standard diffusion Kalman filter employing a well-defined model.

1 Introduction

Distributed inference of parameters and states of probabilistic models from streaming data is an important task in the modern world. It finds applications in many fields including navigation, target localization, collaborative spectral sensing, biomedical signal processing, social networks, or the Internet of Things [1, 2, 3]. Unlike the centralized solutions, the distributed strategies enjoy high robustness to the agent/link failures, and still achieve an excellent estimation performance [4, 5].

In the scope of sequential (online) estimation strategies where the agents communicate with their neighbors within one hop distance, two principally different branches of solutions can be distinguished. First, the consensus strategies where the agents usually exchange information about measurements and estimates in a couple of iterations per time instant. The agents aim at reaching a consensus in estimates [5, 6, 7, 8]. The second branch of solutions comprises the diffusion strategies that remove the intermediate iterations [9, 10, 11, 12, 13, 14]. They consist of two one-shot information exchange steps: (i) the *adaptation step* where the agents share their measurements that are subsequently incorporated into agents' own estimates, and (ii) the *combination step* during which the agents acquire and fuse the neighbors' estimates. According to the presence and ordering of the steps, the adapt-then-combine (ATC), combine-then-adapt (CTA), and the single-phase algorithms can be constructed [14]. It was shown that the ATC algorithms provide superior performance [2, 11, 15].

In this paper, we focus on sequential distributed inference of the discrete-time linear state-space models in diffusion networks. The first solution – the diffusion Kalman filter – is due to Cattivelli and Sayed [10]. Their combination step fuses the point estimates of the states using an averaging procedure. In order to reflect the associated uncertainty about these estimates, a modification involving the covariance intersection method was proposed later [16]. A Bayesian solution to the sequential estimation problems that employs a Kullback-Leibler-optimal combination procedure and yields the diffusion Kalman filter as a special case was proposed lately [15]. Since then, various modifications of the diffusion Kalman filters were proposed. Inspired by partial diffusion least means squares [17] and recursive least squares [18], a partial diffusion Kalman filter [19] was developed in order to reduce the communication burden in networks with limited communication resources. Its performance and modifications

for noisy links were studied recently [20], and the tuning procedures for the combination rules were proposed [21].

We focus on the important case of *imperfectly specified models*. This topic has been attracting a lot of research interest. Indeed, the standard Kalman filter – and hence its distributed variants – suffer significant robustness issues that may degrade the performance, or even cause total divergence of the filter [22]. However, there is only a modest number of algorithms suited specifically for the distributed settings. For instance, Zorzi *et al.*[3] suggest that the uncertainty about the nominal state-space model can be studied based on the assumption that the actual model belongs to a ball in the Kullback-Leibler topology. The ball radius then reflects the mismatch modeling budget. Another earlier work, though in a somewhat different setting, consists in a sensitivity minimization approach[23]. The case of linear state-space models with unknown measurement noise covariance matrix that is possibly spatially heterogeneous across the network was investigated by the second author. It resulted in a distributed variational message passing algorithm for a simultaneous inference of both states and the unknown covariance matrix[24]. Still, the topic of completely unknown model covariance matrices has been neglected. Inspired by the variational Bayesian adaptive Kalman filter (VBAKF) [25], we propose a novel algorithm that should fill this substantial gap. Its novelty is three-fold:

- The VBAKF filter is recast into a message-passing procedure. The probabilistic models and relevant prior distributions possess compatible forms that allow for a straightforward information exchange and fusion.
- A computationally cheap hypotheses testing-based procedure for the selection of the best available 'crude' estimate of the process noise covariance matrix is proposed.
- A corresponding diffusion filter relying on the adapt-then-combine (ATC) strategy is designed.

We remark that there exists a recent variant of the VBAKF filter that exploits compound mixture distributions[26]. Its aim is to remove the sensitivity of the original VBAKF to the 'crude' estimate of the process noise covariance matrix. We solve this particular issue by the mentioned computationally cheap hypotheses testing procedure.

The paper is organized as follows: In Section 2, we formulate the problem in detail and on the background of the classical Kalman filter. The following Section 3 derives the variational filter for the case of non-collaborating agents; its modification for the diffusion networks follows in Section 4. Section 5 demonstrates the performance of the proposed algorithms on three simulated examples. Finally, Section 6 concludes the work. For the sake of easier reading, the definitions of the relevant probability density functions and point estimators are deferred to the Appendix and numbered A.1, A.2 etc.

2 Problem statement

Let us consider a non-hierarchical network consisting of $I \in \mathbf{N}$ agents modeling a common process of interest. These agents, labeled by ordinal numbers $i \in \mathcal{I} \equiv \{1, \ldots, I\}$ are connected by bi-directional communication links yielding a mesh network – a connected undirected graph. The communication among agents is limited to 1 link distance, i.e., each agent $i \in \mathcal{I}$

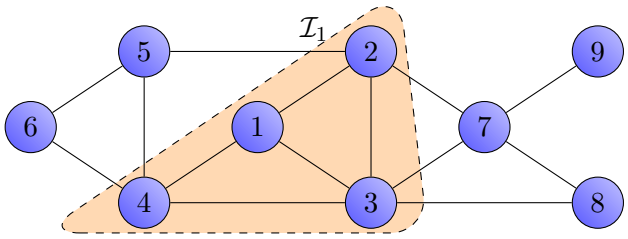

Figure 1: Network of $I = 9$ agents that form the set $\mathcal{I} = \{1, \ldots, 9\}$. If we focus on the agent $i = 1$, then its (closed) neighborhood $\mathcal{I}_1 = \{1, 2, 3, 4\}$ consists of neighbors within one hop distance. The agent is allowed to exchange information solely with $j \in \mathcal{I}_1$.

can perform information exchanges only with its adjacent neighbors $j \in \mathcal{I}$ that form i's closed neighborhood $\mathcal{I}_i$. The agent i belongs to $\mathcal{I}_i$ too. Figure 1 illustrates the principles.

The agents $i \in \mathcal{I}$ independently acquire discrete-time observations $y_{i,t} \in \mathbb{R}^m, t = 1, 2, \ldots$ of a stochastic process described by a linear state-space model

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

$$y_{i,t} = H_t x_t + \varepsilon_{i,t}, \tag{1b}$$

where the process equation (1a) expresses the evolution of the hidden state vector $x_t \in \mathbb{R}^n$ based on the transformation of its previous value by a known matrix $A_t \in \mathbb{R}^{n \times n}$, the known input u_t (if exists), and a matrix B_t of compatible dimensions. The random process noise variable $\omega_t \in \mathbb{R}^n$ makes the (hidden Markov) process stochastic. The measurement equation (1b) then connects x_t with the local measurements $y_{i,t}$ through a known matrix $H_t \in \mathbb{R}^{m \times n}$ and a local random measurement noise $\varepsilon_{i,t} \in \mathbb{R}^m$. The noise variables ω_t and $\varepsilon_{i,t}$ are assumed to be zero-centered independent Gaussian variables with the covariance matrices Q_t and R_t, respectively. This gives rise to the probabilistic form of (1a) and (1b):

$$x_t \sim \mathcal{N}(A_t x_{t-1} + B_t u_t, Q_t), \tag{2a}$$

$$y_{i,t} \sim \mathcal{N}(H_t x_t, R_t), \tag{2b}$$

whose probability densities will be denoted by $p(x_t|x_{t-1}, u_t)$ and $f(y_{i,t}|x_t)$, respectively.

Under a fully specified model (2), the Bayesian approach to the *non-collaborative* inference estimates x_t by means of a Gaussian prior distribution centered at $\hat{x}^+_{i,t-1} \in \mathbb{R}^n$ and with a covariance matrix $P^+_{i,t-1} \in \mathbb{R}^{n \times n}$. That is,

$$x_t|\Delta_{i,t-1} \sim \mathcal{N}(\hat{x}^+_{i,t-1}, P^+_{i,t-1}), \tag{3}$$

where $\Delta_{i,t-1}$ comprises all prior information available for the local estimation of x_t. It primarily consists of the local (or possibly shared) measurements and inputs from the beginning up to time $t-1$. The prior probability density $\pi_i(x_{t-1}|\Delta_{i,t-1})$ is traditionally updated in two Kalman filtering steps [27]:

1. Prediction The mean vector is predicted from $\hat{x}^+_{i,t-1}$ and the covariance matrix $P^+_{i,t-1}$ is appropriately scaled according to the process model (2a),

$$\pi_i(x_t|\Delta_{i,t-1}, u_t) = \int \pi_i(x_{t-1}|\Delta_{i,t-1}) p(x_t|x_{t-1}, u_t) dx_{t-1}. \tag{4}$$

This yields the predicted prior Gaussian distribution

$$\mathcal{N}(\hat{x}_{i,t}^{-}, P_{i,t}^{-}) = \mathcal{N}(A_t \hat{x}_{i,t}^{-}, A_t P_{i,t}^{-} A_t^{\intercal} + Q_t). \tag{5}$$

2. Measurement update The new measurement $y_{i,t}$ is assimilated into the predicted prior Gaussian distribution via the Bayes' theorem,

$$\begin{aligned} \pi_i(x_t|\Delta_t) &= \pi_i(x_t|\Delta_{i,t-1}, u_t, y_{i,t}) \\ &\propto \pi_i(x_t|\Delta_{i,t-1}, u_t) f(y_{i,t}|x_t). \end{aligned} \tag{6}$$

This update is analytically tractable and yields the posterior Gaussian distribution $\mathcal{N}(\hat{x}_{i,t}^{+}, P_{i,t}^{+})$ with

$$\begin{aligned} P_{i,t}^{+} &= \left[\left(P_{i,t}^{-} \right)^{-1} + H_t^{\intercal} R_t^{-1} H_t \right]^{-1}, \\ \hat{x}_{i,t}^{+} &= P_{i,t}^{+} \left[\left(P_{i,t}^{-} \right)^{-1} \hat{x}_{i,t}^{-} + H_t^{\intercal} R_t^{-1} y_{i,t} \right]. \end{aligned} \tag{7}$$

This and alternative expressions can be found, e.g., in Simon's excellent book on state estimation [28].

The formulas (4) and (6) constitute the Bayesian interpretation of the celebrated Kalman filter. Their direct use is limited to the cases where the probabilistic form of the state-space model (2) is fully known. If R_t is unknown, the variational Kalman filter [24, 29, 30] can be used. However, if Q_t is unknown too, the situation gets severely complicated. An effective way towards a computationally cheap filter termed the variational Bayesian adaptive Kalman filter (VBAKF) was recently proposed by Huang et al. [25].[1] In the following section, we modify it to a robust message passing algorithm. The message passing formulation is a remarkable result allowing to very straightforwardly extend the filter to the networked environment with collaboration among agents. This extension is described in the subsequent Section 4.

3 Variational Kalman filtering

The exposition in this section is made from the viewpoint of an isolated agent $i \in \mathcal{I}$ for easier understanding. The subsequent Section 4 then casts the results into distributed setting.

During the sequential estimation of x_t, the generic Kalman filter requires a fully specified state-space model (2) with known covariance matrices Q_t and R_t used in (5) and (7), respectively. The ignorance of Q_t impairs the prediction of $\mathcal{N}(\hat{x}_{i,t}^{-}, P_{i,t}^{-})$, and the ignorance of R_t prevents the subsequent Bayesian update yielding $\mathcal{N}(\hat{x}_{i,t}^{+}, P_{i,t}^{+})$. We aim at solving this issue by (i) a simultaneous inference of x_t and R_t, (ii) an optimization of P_t, and (iii) selecting the best available hypothesized value of Q_t. Let us represent all these unknowns by a vector θ_t as follows:

$$\theta_t = [x_t, R_t, P_t], \tag{8}$$

where we assume vectorizations of the positive definite covariance matrices R_t and P_t, but avoid the vec operators for easier reading. Since the optimization of Q_t will be based on a hypotheses testing procedure, we do not include it in θ_t.

[1] For alternative solutions, the reader is referred to the cited paper.

The estimation of θ_t during the measurement update proceeds with the prior distribution $\pi_i(\theta_t|\Delta_{i,t-1}, u_t)$. Unlike to the standard Kalman filter, there is no convenient distribution for θ_t that would yield an analytically tractable posterior distribution. Therefore, we suggest to proceed with the variational approximation proposed by Huang et al. [25] who extend the earlier work of Särkka and Hartikainen [30]. The true posterior distribution

$$\pi_i(\theta_t|\Delta_{i,t}) \propto \pi_i(\theta_t|\Delta_{i,t-1}, u_t)\, f(y_{i,t}|x_t) \tag{9}$$

is replaced by approximating variational factors (distributions)

$$\pi_i(\theta_t|\Delta_{i,t}) \approx \rho_i(\theta_t) \equiv \rho_i(x_t)\, \rho_i(R_t)\, \rho_i(P_t). \tag{10}$$

We emphasize that Formula (10) sticks with the frequent (Bayesian) custom to identify the distributions by the argument. That is, the factors on the right-hand side are different distributions. This approximation releasing the dependence among the elements of θ_t is the cornerstone of the mean-field variational Bayesian inference [31]. Now, we seek the hyperparameters of the factors that guarantee that the mutual Kullback-Leibler divergence $\mathcal{D}[\rho_i(\theta_t)||\pi_i(\theta_t|\Delta_{i,t})]$ is minimal. By definition of the Kullback-Leibler divergence it follows that

$$\begin{aligned}
\mathcal{D}[\rho_i(\theta_t)||\pi_i(\theta_t|\Delta_{i,t})] &= \mathbb{E}_{\rho_i(\theta_t)}\left[\log \frac{\rho_i(\theta_t)}{\pi_i(\theta_t|\Delta_{i,t})}\right] \\
&= \mathbb{E}_{\rho_i(\theta_t)}\left[\log \frac{\rho_i(\theta_t) \times \int f(y_{i,t}|\theta_t)\pi_i(\theta_t|\Delta_{i,t-1}, u_t)\mathrm{d}\theta_t}{f(y_{i,t}|\theta_t)\pi_i(\theta_t|\Delta_{i,t-1}, u_t)}\right] \\
&= \mathbb{E}_{\rho_i(\theta_t)}\left[\log \frac{\rho_i(\theta_t)}{f(y_{i,t}|\theta_t)\pi_i(\theta_t|\Delta_{i,t-1}, u_t)}\right] + \log f(y_{i,t}|\Delta_{i,t-1}, u_t) \\
&= -\mathcal{L}[\rho_i(\theta_t)] + \log f(y_{i,t}|\Delta_{i,t-1}, u_t),
\end{aligned} \tag{11}$$

where we exploited the Bayes' theorem

$$\pi_i(\theta_t|\Delta_{i,t}) = \frac{f(y_{i,t}|\theta_t)\pi_i(\theta_t|\Delta_{i,t-1}, u_t)}{\int f(y_{i,t}|\theta_t)\pi_i(\theta_t|\Delta_{i,t-1}, u_t)\mathrm{d}\theta_t}. \tag{12}$$

We also take into account the absence of θ_t in the log-evidence

$$f(y_{i,t}|\Delta_{i,t}, u_t) = \int f(y_{i,t}|\theta_t)\pi_i(\theta_t|\Delta_{i,t-1}, u_t)\mathrm{d}\theta_t, \tag{13}$$

which leaves its logarithm intact under the expectation operator in Formula (11). The therm $\mathcal{L}[\rho_i(\theta_t)]$ is known as the evidence lower bound (ELBO) [31]. From the last row in (11) it is apparent that it bounds the log-evidence $\log f(y_{i,t}|\Delta_{i,t}, u_t)$ – if the divergence $\mathcal{D}[\rho_i(\theta_t)||\pi_i(\theta_t|\Delta_{i,t})]$ were 0, the ELBO would be equal to $\log f(y_{i,t}|\Delta_{i,t}, u_t)$.

The log-evidence $\log f(y_{i,t}|\Delta_{i,t}, u_t)$ does not involve θ_t. Hence it is fixed and the minimization of the Kullback-Leibler divergence is achieved by maximization of the negative ELBO

$-\mathcal{L}(\theta_t)$. Let us expand it in terms of individual variables:

$$\begin{aligned}
-\mathcal{L}(\theta_t) &= \mathbb{E}_{\rho_i(\theta_t)}\left[\log \frac{\rho_i(\theta_t)}{f(y_{i,t}|\Delta_{i,t-1}, u_t)\pi_i(\theta_t|\Delta_{i,t-1})}\right] \\
&= \mathbb{E}_{\rho_i(x_t)}\left[\log \frac{\rho_i(x_t)}{\exp\left\{\mathbb{E}_{\rho_i(R_t,P_t)}\left[\log f(y_{i,t}, \theta_t|\Delta_{i,t-1}, u_t)\right]\right\}}\right] + c_x \\
&= \mathbb{E}_{\rho_i(R_t)}\left[\log \frac{\rho_i(R_t)}{\exp\left\{\mathbb{E}_{\rho_i(x_t,P_t)}\left[\log f(y_{i,t}, \theta_t|\Delta_{i,t-1}, u_t)\right]\right\}}\right] + c_R \\
&= \mathbb{E}_{\rho_i(P_t)}\left[\log \frac{\rho_i(P_t)}{\exp\left\{\mathbb{E}_{\rho_i(x_t,R_t)}\left[\log f(y_{i,t}, \theta_t|\Delta_{i,t-1}, u_t)\right]\right\}}\right] + c_P, \qquad (14)
\end{aligned}$$

where c_x, c_R, and c_P represent terms that are independent of x_t, R_t, and P_t, respectively. Obviously, for each individual element of θ_t we have one Kullback-Leibler divergence that attains minimum if

$$\rho_i(x_t) \propto \exp\left\{\mathbb{E}_{\rho_i(R_t,P_t)}\left[\log f(y_{i,t}, \theta_t|\Delta_{i,t-1}, u_t)\right]\right\}, \qquad (15a)$$

$$\rho_i(R_t) \propto \exp\left\{\mathbb{E}_{\rho_i(x_t,P_t)}\left[\log f(y_{i,t}, \theta_t|\Delta_{i,t-1}, u_t)\right]\right\}, \qquad (15b)$$

$$\rho_i(P_t) \propto \exp\left\{\mathbb{E}_{\rho_i(x_t,R_t)}\left[\log f(y_{i,t}, \theta_t|\Delta_{i,t-1}, u_t)\right]\right\}. \qquad (15c)$$

The maximization of the negative ELBO is achievable by means of a coordinate-ascent variational inference (CAVI) algorithm iteratively adjusting the factors [32, 33]. At each iteration, the factors (15a), (15b), and (15c) employ point estimates of the other elements of θ_t while being updated. Since the ELBO is generally a non-convex objective function, the CAVI algorithm only guarantees convergence to the local optimum [34]. The linear measurement model (2b) and an appropriate choice of the prior factors (10) yield an unimodal posterior distribution, hence the convergence to the global optimum is achieved.

A remarkable property of Formulas (15) is their analytical tractability if $f(y_{i,t}|\theta_t)$ is a probability density function of an exponential family distribution, and convenient conjugate factors are used for the estimation of x_t, R_t, and P_t.

Let us generally define the exponential family and the conjugate prior distributions.

Definition 1 (Exponential family)**.** *A family $\{F_\vartheta\}$ of distributions of a random variable y parameterized by a scalar or multivariate parameter ϑ is said to form an exponential family if the probability density functions can be written in the form*

$$f(y|\vartheta) = \exp\left\{\eta(\vartheta)^\intercal T_\vartheta(y) - B(\vartheta)\right\} h(y), \qquad (16)$$

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

Definition A.1 in the Appendix shows that the Gaussian measurement model for $y_{i,t}$ parameterized by either x_t or R_t is an exponential family distribution. Similarly, the Gaussian model for x_t given P_t is an exponential family distribution too, cf. Definition A.2.

Definition 2 (Conjugate prior distribution for ϑ)**.** *Assume that $f(y|\vartheta)$ is an exponential family distribution according to Definition 1. Then, a distribution $\pi(\vartheta)$ is said to be conjugate to $f(y|\vartheta)$, if its probability density function has the form*

$$\pi(\vartheta) = \exp\left\{\eta(\vartheta)^\intercal \Xi_\vartheta^- - \nu^- B(\vartheta)\right\} g(\vartheta), \tag{17}$$

where $g(\vartheta)$ is a known function, $B(\vartheta)$ coincides with the log-normalization function of $f(y|\vartheta)$, and the hyperparameter Ξ_ϑ^- has the same dimension as $T_\vartheta(y)$. The hyperparameter $\nu^- > 0$ may be absent if $B(\vartheta) = 1$ for all ϑ, or it can sometimes be absorbed into Ξ_ϑ^-.

Obviously, a multiplication of the exponential family distribution $f(y|\vartheta)$ by a conjugate prior $\pi(\vartheta)$ yields a distribution of the same type as the prior,

$$\begin{aligned}\pi(\vartheta|y) &\propto f(y|\vartheta)\,\pi(\vartheta)\\ &\propto \exp\left\{\eta(\vartheta)^\intercal\left(\Xi_\vartheta^- + T_\vartheta(y)\right) - \left(\nu^- + 1\right) B(\vartheta)\right\}\\ &\propto \exp\left\{\eta(\vartheta)^\intercal \Xi_\vartheta^+ - \nu^+ B(\vartheta)\right\},\end{aligned} \tag{18}$$

where the updated hyperparameters read

$$\Xi_\vartheta^+ = \Xi_\vartheta^- + T_\vartheta(y), \tag{19}$$

$$\nu^+ = \nu^- + 1. \tag{20}$$

(The indices $^-$ and $^+$ denote the prior and the posterior hyperparameters, respectively, and will be used throughout the paper.) In Bayesian modeling, this principle is the cornerstone of the exact (non-approximate) sequential inference [35]. In the scope of the CAVI-based methods, it provides a way towards analytically tractable approximate inference as follows: Let us focus on the factors in (15) and assume that for each particular factor and its variable, the remaining elements of θ_t on the right-hand side are replaced by their point estimates. If the measurement model $f(y_{i,t}|\theta_t)$ is an exponential family distribution, and the prior distribution of the considered element of θ_t is conjugate to it, then the CAVI-based update of the variational factor exploits exactly the procedure (18). This procedure is successively applied to all the factors in (15) in several iterations.

Let us now concentrate on the individual factors. The Gaussianity of $y_{i,t}$ and the properties of x_t, R_t, and P_t advocate to adopt the Gaussian and the inverse-Wishart distributions as follows:

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

To avoid confusion in the interdependence with respect to P_t, the (intermediate) posterior estimate of P_t from the inverse-Wishart factor that enters the prior factor of x_t is denoted as $\hat{P}_{i,t}^\star$, and similarly are denoted the related posterior hyperparameters $\Psi_{i,t}^\star$ and $\psi_{i,t}^\star$.

The detailed descriptions of the probability density functions $f(y_{i,t}|\theta_t)$, $\rho_i(x_t)$, $\rho_i(R_t)$, and $\rho_i(P_t)$ in all required forms convey Definitions A.1—A.4 in the Appendix. The sufficient statistics, conjugate hyperparameters, and point estimates are given there. From these definitions it follows that the Gaussian distribution of x_t is conjugate to the measurement model

$f(y_{i,t}|\theta_t)$ with R_t fixed. The inverse-Wishart distribution of P_t is conjugate to the distribution of x_t, and the inverse-Wishart distribution of R_t is conjugate to $f(y_{i,t}|\theta_t)$ with x_t fixed.

With the help of Definitions 1 and 2, the particular CAVI updates (15) follow easily. Below, their details are given for each variational factor in terms of the expected sufficient statistics entering the updates of the relevant hyperparameters, and the resulting point estimates. The iterations of the CAVI algorithm are denoted by $d = 1, \ldots, D$. For the initial iteration $d = 1$, we set $\hat{P}_{i,t}^{+,0} = \hat{P}_{i,t}^{-}$ and $\hat{x}_{i,t}^{+,(0)} = \hat{x}_{i,t}^{-}$. After the last iteration D, the variational factors $\rho_i(R_t)$ and $\rho_i(x_t)$ serve as the posterior distributions, whose hyperparameters enter the subsequent prediction step at the next instant $t+1$. To simplify notation, we write the outer product $(a-b)(a-b)^\intercal$ as $(a-b)(\bullet)^\intercal$.

Update of $\rho_i(P_t) \equiv i\mathcal{W}(\Psi_{i,t}^{-}, \psi_{i,t}^{-})$:

- Expectation of the sufficient statistic $T_{P_t}(x_t)$:

$$\mathbb{E}_{\rho_i(x_t)}^{(d)}[T_{P_t}(x_t)] = \begin{bmatrix} \hat{P}_{i,t}^{+,(d-1)} + \left(\hat{x}_{i,t}^{+,(d-1)} - \hat{x}_{i,t}^{-}\right)\left(\bullet\right)^\intercal \\ 1 \end{bmatrix}. \tag{21}$$

- Update of the hyperparameter $\Xi_{P_t,i}^{-}$:

$$\begin{aligned} \Xi_{P_t,i}^{\star,(d)} &= \Xi_{P_t,i}^{-} + \mathbb{E}_{\rho_i(x_t)}^{(d)}[T_{P_t}(x_t)] \\ &= \begin{bmatrix} \Psi_{i,t}^{-} \\ \psi_{i,t}^{-} + n + 1 \end{bmatrix} + \mathbb{E}_{\rho_i(x_t)}^{(d)}[T_{P_t}(x_t)] \\ &= \begin{bmatrix} \Psi_{i,t}^{\star,(d)} \\ \psi_{i,t}^{\star,(d)} + n + 1 \end{bmatrix}. \end{aligned} \tag{22}$$

- Point estimates:

$$\hat{P}_{i,t}^{\star,(d)} = \mathbb{E}_{\rho_i(P_t)}^{(d)}[P_t] = \frac{\Psi_{i,t}^{\star,(d)}}{\psi_{i,t}^{\star,(d)} - n - 1}, \tag{23}$$

$$\left(\hat{P}_{i,t}^{\star,(d)}\right)^{-1} = \mathbb{E}_{\rho_i(P_t)}^{(d)}[P_t^{-1}] = \psi_{i,t}^{\star,(d)}\left(\Psi_{i,t}^{\star,(d)}\right)^{-1}. \tag{24}$$

Remark The symbol $\star$ is used to denote the intermediate value of the estimate that enters $\rho_i(x_t)$ for the subsequent measurement update.

Update of $\rho_i(R_t) \equiv i\mathcal{W}(\Phi_{i,t}^{-}, \phi_{i,t}^{-})$:

- Expectation of the sufficient statistic $T_{R_t}(y_{i,t})$:

$$\mathbb{E}_{\rho_i(x_t,P_t)}^{(d)}[T_{R_t}(y_{i,t})] = \begin{bmatrix} \left(y_{i,t} - H_t\hat{x}_{i,t}^{+,(d-1)}\right)\left(\bullet\right)^\intercal + H_t\hat{P}_{i,t}^{+,(d-1)}H_t^\intercal \\ 1 \end{bmatrix} \tag{25}$$

- Update of the hyperparameter $\Xi^{-}_{R_t,i}$:

$$\begin{aligned}\Xi^{+,(d)}_{R_t,i} &= \Xi^{-}_{R_t,i} + \mathbb{E}^{(d)}_{\rho_i(x_t,P_t)}[T_{R_t}(y_{i,t})] \\ &= \begin{bmatrix} \Phi^{-}_{i,t} \\ \phi^{-}_{i,t} + n + 1 \end{bmatrix} + \mathbb{E}^{(d)}_{\rho_i(x_t,P_t)}[T_{R_t}(y_{i,t})] \\ &= \begin{bmatrix} \Phi^{+,(d)}_{i,t} \\ \phi^{+,(d)}_{i,t} + n + 1 \end{bmatrix}. \end{aligned} \tag{26}$$

- Point estimates:

$$\hat{R}^{+,(d)}_{i,t} = \mathbb{E}^{(d)}_{\rho_i(R_t)}[R_t] = \frac{\Phi^{+,(d)}_{i,t}}{\phi^{+,(d)}_{i,t} - m - 1}, \tag{27}$$

$$\left(\hat{R}^{+,(d)}_{i,t}\right)^{-1} = \mathbb{E}^{(d)}_{\rho_i(R_t)}[R_t^{-1}] = \phi^{+,(d)}_{i,t} \left(\Phi^{+,(d)}_{i,t}\right)^{-1}. \tag{28}$$

Update of $\rho_i(x_t) \equiv \mathcal{N}(\hat{x}^{-}_{i,t}, \hat{P}^{\star,(d)}_{i,t})$:

- Expectation of the sufficient statistic $T_{x_t}(y_{i,t})$:

$$\mathbb{E}^{(d)}_{\rho_i(R_t,P_t)}[T_{x_t}(y_{i,t})] = \begin{bmatrix} y^{\intercal}_{i,t} \\ H^{\intercal}_t \end{bmatrix} \left(\hat{R}^{+,(d)}_{i,t}\right)^{-1} \begin{bmatrix} y^{\intercal}_{i,t} \\ H^{\intercal}_t \end{bmatrix}^{\intercal}. \tag{29}$$

- Update of the hyperparameter $\Xi^{-}_{x_t,i}$:

$$\begin{aligned}\Xi^{+,(d)}_{x_t,i} &= \Xi^{-}_{x_t,i} + \mathbb{E}^{(d)}_{\rho_i(R_t,P_t)}[T_{x_t}(y_{i,t})] \\ &= \begin{bmatrix} (\hat{x}^{-}_{i,t})^{\intercal} \\ I \end{bmatrix} \left(\hat{P}^{\star,(d)}_{i,t}\right)^{-1} \begin{bmatrix} (\hat{x}^{-}_{i,t})^{\intercal} \\ I \end{bmatrix}^{\intercal} + \mathbb{E}^{(d)}_{\rho_i(x_t,P_t)}[T_{R_t}(y_{i,t})] \\ &= \begin{bmatrix} (\hat{x}^{+,(d)}_{i,t})^{\intercal} \\ I \end{bmatrix} \left(\hat{P}^{+,(d)}_{i,t}\right)^{-1} \begin{bmatrix} (\hat{x}^{+,(d)}_{i,t})^{\intercal} \\ I \end{bmatrix}^{\intercal}. \end{aligned} \tag{30}$$

- Point estimates:

$$\hat{P}^{+,(d)}_{i,t} = \left[\left(\hat{P}^{\star,(d)}_{i,t}\right)^{-1} + H^{\intercal}_t \left(\hat{R}^{+,(d)}_{i,t}\right)^{-1} H_t\right]^{-1}, \tag{31}$$

$$\hat{x}^{+,(d)}_{i,t} = \hat{P}^{+,(d)}_{i,t} \left[\left(\hat{P}^{\star,(d)}_{i,t}\right)^{-1} \hat{x}^{-,(d)}_{i,t} + H^{\intercal}_t \left(\hat{R}^{+,(d)}_{i,t}\right)^{-1} y_{i,t}\right]. \tag{32}$$

The estimates of the inverse covariance matrices (24) and (28) should be used in place of the inverses of the estimates. However, the difference vanishes with growing $\phi^{+}_{i,t}$ and $\psi^{+}_{i,t}$. The formulas (31) and (32) easily follow from (30). It is worth to notice their similarity to the Kalman filter formulas (7).

Remarks While it is apparent how the updates of the inverse-Wishart factors (22) and (26) transform the original hyperparameters $(\Psi^-_{i,t}, \psi^-_{i,t})$ and $(\Phi^-_{i,t}, \phi^-_{i,t})$, the calculation of the (generally preferred) Gaussian hyperparameters $(\hat{x}^+_{i,t}, \hat{P}^+_{i,t})$ from $\Xi^+_{x_t,i}$ requires a bit of easy algebra. The result follows from the block-matrix form of (29) and (30), and from Lemma A.1 in the Appendix. We suggest that sticking with the hyperparameters $\Xi^{+,(d)}_{x_t,i}$ and evaluation of the related mean $\hat{x}^{+,(d)}_{i,t}$ and covariance matrix $\hat{P}^{+,(d)}_{i,t}$ only when necessary has several notable advantages. First, they are directly used in the diffusion combination step. Second, the associated algebra during the updates and combinations is computationally more stable. And third, additional stability can be easily achieved by decompositions of $\Xi^{\cdot}_{x_t,i}$, e.g., in the Cholesky sense [36].

The overall structure of the message passing algorithm shows Figure 2.

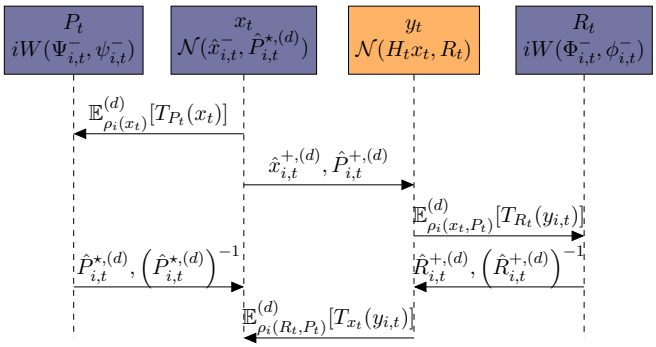

Figure 2: Scheme of the message passing algorithm. The vertical ordering of messages corresponds to the order of the CAVI steps in Algorithm 1.

3.1 Prediction (time update)

In the standard Kalman filter, the prediction step transforms the recent posterior estimate $x^+_{i,t-1}$ and its covariance $P^+_{i,t-1}$ to the prior estimate $x^-_{i,t}$ with $P^-_{i,t}$ for the current time t. It relies on the known process model (2a) used in (4), and results in the predicted estimates (5) that enter the subsequent measurement update step. In our task, we face the uncertainty about all elements of $\theta_t = [x_t, R_t, P_t]$, but the process model (2a) applies only to the estimate of x_t and the related estimate of P_t. We need to devise a convenient prediction step taking the ignorance of Q_t into account, plus surrogate the missing evolution model for R_t.

Let us first focus on the variational factor $\rho_i(x_t) \equiv \mathcal{N}(\hat{x}^+_{i,t-1}, \hat{P}^+_{i,t-1})$. Its transformation using the process model coincides with Formula (5) with the difference that the true covariance $P^+_{i,t-1}$ is replaced by its estimate $\hat{P}^+_{i,t-1}$. Keeping in mind that the process noise covariance matrix Q_t inflates the uncertainty about the estimates $\hat{x}^-_{i,t}$, it is possible to compensate the ignorance of Q_t by its crude (nominal) estimate $\hat{Q}_{i,t}$ providing a convenient degree of this inflation. Then, (4) yields a predicted Gaussian distribution with

$$\begin{aligned} \hat{x}^-_{i,t} &= A_t \hat{x}^+_{i,t-1} + B_t u_t, \\ \hat{P}^-_{i,t} &= A_t \hat{P}^+_{i,t-1} A_t^\intercal + \hat{Q}_{i,t}. \end{aligned} \tag{33}$$

Our procedure for selecting $\hat{Q}_{i,t}$ describes the following Section 3.2.

Next, we focus on R_t. A direct prediction of $\hat{R}_{i,t}$ from $\hat{R}_{i,t-1}$ is impossible due to the ignorance of an appropriate evolution model. Under slow time-variability of R_t, the first-choice procedure is the exponential forgetting [27],

$$\rho_i(R_t) = [\rho_i(R_{t-1})]^{\alpha_R}, \quad \alpha_R \in [0,1]. \tag{34}$$

The forgetting factor α_R is usually not less than 0.95. In terms of $\Xi^-_{R_t,i}$ this corresponds to

$$\Xi^-_{R_t,i} = \alpha_R \, \Xi^+_{R_{t-1},i}, \tag{35}$$

which effectively affects the amount of information in the distribution and increases the uncertainty about R_t. The resulting less concentrated prior distribution is more responsive to new information about the actual value of R_t.

Finally, we have to construct a new factor $\rho_i(P_t)$. Naturally, its expectation should be equal to the predicted value (33),

$$\mathbb{E}_{\rho_i(P_t)}[P_t] = \frac{\Psi^-_{i,t}}{\psi^-_{i,t} - n - 1} = \hat{P}^-_{i,t}, \tag{36}$$

but similarly to the case of R_t we want to admit some degree of uncertainty about this value. Since $\phi^+_{i,t-1}$ has the interpretation of the data counter, cf. (22), we set

$$\begin{aligned} \psi^-_{i,t} &= \psi^+_{i,t-1}, \\ \Psi^-_{i,t} &= (\psi^-_{i,t} - n - 1)\hat{P}^-_{i,t}. \end{aligned} \tag{37}$$

That is, with the increasing amount of incorporated data, the uncertainty about P_t decreases.

3.2 Optimization of $\hat{Q}_{i,t}$

The prediction step (33) employs a crude estimate $\hat{Q}_{i,t}$ of the process noise covariance matrix Q_t. While it is possible to use a single heuristic value as proposed by Huang et al. [25], we conjecture that a computationally cheap procedure for selecting a convenient value from a set of candidates $\hat{Q}^{(1)}_{i,t}, \ldots, \hat{Q}^{(C)}_{i,t}$ could increase the estimation performance. Moreover, this would open a way towards a real tuning of $\hat{Q}_{i,t}$, but we leave this behind the scope of the paper.

Our standpoint is that if the state-space model (2) were correctly and completely specified, then the prior predictive distribution

$$f(y_{i,t}|\Delta_{i,t-1}, u_t) = \int f(y_{i,t}|\theta_t)\pi_i(\theta_t|\Delta_{i,t-1}, u_t)d\theta_t \tag{38}$$

would well explain the measurements $y_{i,t}$. Somewhat similarly, if we use the plug-in principle [37] and replace the true R_t by its point estimate $\hat{R}^-_{i,t}$, then the role of $\pi_i(\theta_t|\Delta_{i,t-1}, u_t)$ will play the factor $\rho_i(x_t)$, and

$$\begin{aligned} f(y_{i,t}|\Delta_{i,t-1}, u_t) &= \int \mathcal{N}(H_t x_t, \hat{R}^-_{i,t}) \times \mathcal{N}(\hat{x}^-_{i,t}, \hat{P}^-_{i,t})dx_t \\ &= \int \mathcal{N}(H_t x_t, \hat{R}^-_{i,t}) \times \mathcal{N}(\hat{x}^-_{i,t}, A_t\hat{P}^+_{i,t-1}A_t^\intercal + Q_t)dx_t \\ &= \mathcal{N}\left(H_t\hat{x}^-_{i,t}, \hat{R}^-_{i,t} + H_t(A_t\hat{P}^+_{i,t-1}A_t^\intercal + Q_t)H_t^\intercal\right). \end{aligned} \tag{39}$$

Thus the better estimate $\hat{Q}_{i,t}$ is used in place of the true Q_t, the higher value of the predictive probability density function for a particular measurement $y_{i,t}$ will be. From this viewpoint, the formula (39) can play the role of the prior predictive likelihood for $\hat{Q}_{i,t}$. If we define a set of C hypothetical (candidate) matrices

$$\mathcal{Q}_{i,t} = \{\hat{Q}_{i,t}^{(1)}, \ldots, \hat{Q}_{i,t}^{(C)}\}, \tag{40}$$

and plug them into (39), we obtain C hypothetical predictive densities. They differ in their predictive performance with respect to the available measurement $y_{i,t}$, which allows to discriminate among them in the sense of the *Bayesian* hypotheses testing [35]. To conclude, we aim at selecting that element of $\mathcal{Q}_{i,t}$ that maximizes the predictive performance, i.e., the (logarithm of the) predictive density

$$\hat{Q}_{i,t} = \underset{\tilde{Q}_t \in \mathcal{Q}_{i,t}}{\arg\max} \log \mathcal{N}\left(y_{i,t} \middle| H_t \hat{x}_{i,t}^-, R(\tilde{Q}_t)\right), \tag{41}$$

where the hypothesized covariance matrices

$$R(\tilde{Q}_t) = \hat{R}_{i,t}^- + H_t \left(A_t \hat{P}_{i,t-1}^+ A_t^\intercal + \tilde{Q}_t\right) H_t^\intercal. \tag{42}$$

As we will see later, this approach is very favorable in the distributed setting due to the increased number of available measurements.

An interesting research direction would be to tune the elements of the set $\mathcal{Q}_{i,t}$, e.g., by sampling from the neighborhood of the 'best' candidate values. However, we leave this open for a potential future research.

4 Collaborative filtering with information diffusion

In this section, we recast the developed isolated sequential estimation procedure to a networked environment equipped with the information diffusion strategy [2]. The strategy consists of two steps:

1. *The adaptation step* where each agent $i \in \mathcal{I}$ receives the measurements $y_{j,t}$ of its neighbors $j \in \mathcal{I}_i$. These measurements are locally assimilated into i's statistical knowledge by means of the Bayesian update.

2. *The combination step* where each agent $i \in \mathcal{I}$ integrates the posterior estimates of θ_t of neighbors $j \in \mathcal{I}_i$. The result serves as the i's final knowledge of θ_t.

In the sequel we assume that each agent has locally performed the prediction step presented in Section 3.1 and hence the prior variational factors $\rho_i(x_t), \rho_i(R_t)$, and $\rho_i(P_t)$ are at disposal.

4.1 Adaptation step: Collaborative measurement update

The role of the adaptation step is closely connected with the large-number behavior of the considered estimators. By taking advantage of the increased number of available independent and identically distributed (iid) measurements $y_{j,t}, j \in \mathcal{I}_i$, we aim at acceleration of the

Algorithm 1 Local variational Kalman filtering under unknown Q_t and R_t

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. Set the number D of CAVI iterations.

For $t = 1, 2, \ldots$ and measurements $y_{i,t}$ do:
Prediction:

1. Predict $\rho_i(R_t)$: Evaluate $\Xi^-_{R_t,i} = \alpha_R \Xi^+_{R_{t-1},i}$, Formula (35).
2. Select $\hat{Q}_{i,t}$, Formula (42).
3. Predict $\rho_i(x_t)$: Evaluate $\hat{x}^-_{i,t}$ and $\hat{P}^-_{i,t}$, Formula (33).
4. Predict $\rho_i(P_t)$: Evaluate (37).

Measurement update:
For $d = 1, \ldots, D$ do:

1. Update $\rho_i(P_t)$:
   - Calculate $\mathbb{E}^{(d)}_{\rho_i(x_t)}[T_{P_t}(x_t)]$, Formula (21).
   - Update $\Xi^-_{P_t,i}$, Formula (22)
2. Update $\rho_i(R_t)$:
   - Calculate $\mathbb{E}^{(d)}_{\rho_i(x_t,P_t)}[T_{R_t}(y_t)]$, Formula (25).
   - Update $\Xi^-_{R_t,i}$, Formula (26)
3. Update $\rho_i(x_t)$:
   - Prepare estimates $\hat{P}^{\star,(d)}_{i,t}$, $\hat{R}^{(d)}_{i,t}$, and their inverses, Formulas (23),(24), (27), and (28).
   - Calculate $\mathbb{E}^{(d)}_{\rho_i(P_t,R_t)}[T_{x_t}(y_t)]$, Formula (29).
   - Update $\Xi^-_{x_t,i}$, Formula (30)

convergence of the estimates of θ_t to the true values. To this end, we extend the measurement update step (6),

$$\begin{aligned}\pi_i(\theta_t|\Delta_{i,t}) &= \pi_i\left(\theta_t|\Delta_{i,t-1}, u_t, \{y_{j,t}\}_{j\in\mathcal{I}_i}\right)\\ &\propto \pi_i(\theta_t|\Delta_{i,t-1}, u_t)\prod_{j\in\mathcal{I}_i} f(y_{j,t}|\theta_t),\end{aligned} \tag{43}$$

where $\Delta_{i,t}$ now encompasses more information about θ_t. The probability density functions $f(y_{j,t}|\theta_i)$ are identically parameterized due to the iid nature of $y_{j,t}$. Recall that these (identical) distributions belong to the exponential family, hence they locally have the expected sufficient statistics $\mathbb{E}^{(d)}_{\rho_i(x_t,P_t)}[T_{R_t}(y_{j,t})]$ or $\mathbb{E}^{(d)}_{\rho_i(R_t,P_t)}[T_{x_t}(y_{j,t})]$, respectively, cf. Formulas (25), (29), and the definitions in the Appendix. Therefore, we may express the product (43) as

$$\begin{aligned}\prod_{j\in\mathcal{I}_i} f(y_{j,t}|\theta_t) &\propto \prod_{j\in\mathcal{I}_i}\exp\left\{\eta_{x_t}^{\intercal}\cdot T_{x_t}\left(y_{j,t}\right)\right\}\\ &\propto \exp\left\{\eta_{x_t}^{\intercal}\cdot\sum_{j\in\mathcal{I}_i} T_{x_t}\left(y_{j,t}\right)\right\},\end{aligned} \tag{44}$$

with

$$\mathbb{E}^{(d)}_{\rho_i(R_t,P_t)}\left[\sum_{j\in\mathcal{I}_i} T_{x_t}(y_{j,t})\right] = \sum_{j\in\mathcal{I}_i}\begin{bmatrix} y_{j,t}{}^{\intercal}\\ H_t^{\intercal}\end{bmatrix}\left(\hat{R}_{i,t}^{+,(d)}\right)^{-1}\begin{bmatrix} y_{j,t}^{\intercal}\\ H_t^{\intercal}\end{bmatrix}^{\intercal} \tag{45}$$

for the variational message to $\rho_i(x_t)$. By analogy, the variational message for $\rho_i(R_t)$ follows from

$$\prod_{j\in\mathcal{I}_i} f_j(y_{j,t}|\theta_t) \propto \exp\left\{\eta_{R_t}^{\intercal}\cdot\sum_{j\in\mathcal{I}_i} T_{R_t}\left(y_{j,t}\right)\right\} \tag{46}$$

and equals to

$$\mathbb{E}^{(d)}_{\rho_i(x_t,P_t)}\left[\sum_{j\in\mathcal{I}_i} T_{R_t}(y_{j,t})\right] = \sum_{j\in\mathcal{I}_i}\begin{bmatrix}\left(y_{j,t} - H_t\hat{x}_{i,t}^{+,(d-1)}\right)\begin{pmatrix}\bullet\\ 1\end{pmatrix}^{\intercal} + H_t\hat{P}_{i,t}^{+,(d-1)}H_t^{\intercal}\end{bmatrix}. \tag{47}$$

These expected sufficient statistics replace their single-measurement counterparts (29) and (25) in the CAVI updates (26) and (30). The local estimates of x_t, P_t, and R_t result from the same formulas and respective elements of the hyperparameters $\Xi^{\star,(d)}_{P_t,i}$, $\Xi^{+,(d)}_{R_t,i}$, and $\Xi^{+,(d)}_{x_t,i}$. It also follows that the diffusion variants of the local estimates (31) and (32) are

$$\hat{P}_{i,t}^{+,(d)} = \left[\left(\hat{P}_{i,t}^{\star,(d)}\right)^{-1} + |\mathcal{I}_i|\, H_{i,t}^{\intercal}\left(\hat{R}_{i,t}^{+,(d)}\right)^{-1} H_{i,t}\right]^{-1}, \tag{48}$$

$$\hat{x}_{i,t}^{+,(d)} = \hat{P}_{i,t}^{+,(d)}\left[\left(\hat{P}_{i,t}^{\star,(d)}\right)^{-1}\hat{x}_{i,t}^{-,(d)} + H_{i,t}^{\intercal}\left(\hat{R}_{i,t}^{+,(d)}\right)^{-1}\sum_{j\in\mathcal{I}_i} y_{j,t}\right]. \tag{49}$$

The impact of multiple measurements on the estimator is apparent.

The communication requirements of the adaptation step are connected with the exchange of the m-dimensional vectors $y_{j,t}$. That is, the agent i with the neighborhood $\mathcal{I}_i$ has to acquire $[\mathrm{card}(\mathcal{I}_i) - 1]\, m$ real numbers.

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

The local optimization of $\hat{Q}_{i,t}$ over the set $\mathcal{Q}_{i,t}$ described in Section 3.2 should naturally profit from the neighbors' information, i.e., $y_{j,t}, j \in \mathcal{I}_i$. For the their iid nature, the joint predictive density is a product of individual densities,

$$f\left(\{y_{j,t}\}_{j\in\mathcal{I}_i}|\Delta_{i,t-1}, u_t\right) = \prod_{j\in\mathcal{I}_i} f(y_{j,t}|\Delta_{i,t-1}, u_t), \tag{50}$$

which allows to modify (42) as follows:

$$\begin{aligned}\hat{Q}_{i,t} &= \arg\max_{\tilde{Q}_t\in\mathcal{Q}_{i,t}} \log \prod_{j\in\mathcal{I}_i} \mathcal{N}\left(y_{j,t}|H_t\hat{x}^-_{i,t}, R(\tilde{Q}_t)\right) \\ &= \arg\max_{\tilde{Q}_t\in\mathcal{Q}_{i,t}} \sum_{j\in\mathcal{I}_i} \log\mathcal{N}\left(y_{j,t}|H_t\hat{x}^-_{i,t}, R(\tilde{Q}_t)\right)\end{aligned}$$

with the hypothesized covariance matrices

$$R(\tilde{Q}_t) = \hat{R}^-_{i,t} + H_t\left[A_t\hat{P}^+_{i,t-1}A_t^\intercal + \tilde{Q}_t\right]H_t^\intercal. \tag{51}$$

4.3 Combination step

During the combination step, each agent $i \in \mathcal{I}$ acquires the posterior estimates from its neighbors $j \in \mathcal{I}_i$. In our realm, these estimates are completely expressed by the variational factors $\rho_j(x_t)$ and $\rho_j(R_t)$. Their repetitive fusion makes the information gradually *diffuse* through the network. Recall that the information about θ_t is represented by the hyperparameters $\Xi^+_{x_t,j}$ and $\Xi^+_{R_t,j}$ which accumulate – linearly combine – the relevant sufficient statistics. A combination step fusing them using certain linear operations should therefore be optimal in some sense. The most straightforward rules read

$$\widetilde{\rho}_i(x_t) \propto \exp\left\{\eta^\intercal_{x_t}\cdot\widetilde{\Xi}^+_{x_t,i}\right\} = \exp\left\{\eta^\intercal_{x_t}\cdot\frac{1}{|\mathcal{I}_i|}\sum\Xi^+_{x_t,j}\right\}, \tag{52}$$

for the estimates of x_t, and

$$\widetilde{\rho}_i(R_t) \propto \exp\left\{\eta^\intercal_{R_t}\cdot\widetilde{\Xi}^+_{R_t,i}\right\} = \exp\left\{\eta^\intercal_{R_t}\cdot\frac{1}{|\mathcal{I}_i|}\sum\Xi^+_{R_t,j}\right\}, \tag{53}$$

for the estimates of R_t. The rules (52) and (53) enjoy several attractive properties that fully advocate their use:

- *Tractability, numerical stability:* Unlike some other combination rules, (52) and (53) yield hyperparameters of the same type as the the original hyperparameters. That is, the functional form of the distribution is preserved, which allows to immediately use the resulting $\widetilde{\rho}_i(x_t)$ and $\widetilde{\rho}_i(R_t)$ in the subsequent prediction step at time t. Moreover, the convex combination is not prone to any numerical difficulties.

- *Compliance with the Bayesian information processing:* The convex combination of hyperparameters of the conjugate prior distributions is due to the linearity equivalent to the (weighted) Bayesian updating.

- *Robustness to data incest:* A typical problem in distributed data processing is the data incest [38, 39], where some information (measurements, estimates) is replicated and repeatedly enters the information processing procedures. While it could be useful to increase the amount of information in (52) and (53) by avoiding the factor $1/|\mathcal{I}_i|$ in one-shot static estimation, in sequential estimation this approach would duplicate the information each time step. Moreover, the proposed rules (52) and (53) would be safe even if all hyperparameters were identical. The second author demonstrated this in his recent work [24].

- *Kullback-Leibler optimality:* The fusion rules (52) and (53) are Kullback-Leibler optimal in the sense

$$\widetilde{\rho}_i(x_t) = \arg\min_{\bar{\rho}_i(x_t)} \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \mathcal{D}\left[\bar{\rho}_i(x_t) || \rho_i(x_t)\right], \tag{54}$$

  (analogously for $\rho_i(R_t)$). The second author studied this optimality recently [15]. Since the Kullback-Leibler divergence is a special case of several information divergence families, e.g., the α and γ divergences [40], the optimality can be perceived in a wider sense. It can also be shown that the result is optimal in the Bregman and functional Bregman sense [41].

- *Covariance intersection:* The Kullback-Leibler-optimal fusion (54) applied to Gaussian distributions yields a result known as the covariance intersection [42]. For completeness, the combined estimates are

$$\begin{aligned} \tilde{\hat{P}}^+_{i,t} &= \left[\frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \left(\hat{P}^+_{j,t}\right)^{-1}\right]^{-1}, \\ \tilde{\hat{x}}^+_{i,t} &= \tilde{\hat{P}}^+_{i,t} \left[\frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \left(\hat{P}^+_{j,t}\right)^{-1} \hat{x}^+_{j,t}\right], \end{aligned} \tag{55}$$

  which is obvious from (52) and (30).

The minimal communication requirements of the combination step are connected with the transmission of the hyperparameters $\Xi^+_{x_t,j}$ and $\Xi^+_{R_t,j}$. The former can be reconstructed from the vectors $\hat{x}^+_{j,t}$ consisting of n elements and the symmetric $n \times n$ matrix $\hat{P}^+_{j,t}$ that can be represented by $\frac{n}{2}(n+1)$ positive real numbers. The latter consists of the positive real scalar $\phi^+_{j,t}$ and the symmetric $m \times m$ matrix $\Phi^+_{j,t}$ that can be represented by $\frac{m}{2}(m+1)$ positive real numbers. For the neighborhood $\mathcal{I}_i$ and keeping in mind that $i \in \mathcal{I}_i$, we have

$$[\mathrm{card}(\mathcal{I}_i) - 1]\left[n + \frac{n}{2}(n+1) + 1 + \frac{m}{2}(m+1)\right] = [\mathrm{card}(\mathcal{I}_i) - 1]\,\frac{n^2 + 3n + m^2 + m + 2}{2} \tag{56}$$

real numbers to be acquired by the agent i.

5 Examples

This section contains three simulation examples. The first one assesses the performance of the local (non-collaborative) filter corresponding to Algorithm 1 and compares it with the

Algorithm 2 VARIATIONAL KALMAN FILTERING UNDER UNKNOWN Q_t AND R_t FOR NETWORKS WITH INFORMATION DIFFUSION

Initialization: At each agent $i \in \mathcal{I}$, 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. Set the number D of CAVI iterations.
For $t = 1, 2, \ldots$ and and each agent $i \in \mathcal{I}$ do:
Adaptation step:

1. Acquire measurements $y_{j,t}$ of neighbors $j \in \mathcal{I}_i$.
2. *Prediction:*
   (a) Predict $\rho_i(R_t)$: Evaluate $\Xi^-_{R_t,i} = \alpha_R \Xi^+_{R_{t-1},i}$, Formula (35).
   (b) Select $\hat{Q}_{i,t}$, Formula (51).
   (c) Predict $\rho_i(x_t)$: Evaluate $\hat{x}^-_{i,t}$ and $\hat{P}^-_{i,t}$, Formula (33).
   (d) Predict $\rho_i(P_t)$: Evaluate (37).
3. *Measurement update:*
   For $d = 1, \ldots, D$ do:
   (a) Update $\rho_i(P_t)$:
      - Calculate $\mathbb{E}^{(d)}_{\rho_i(x_t)}[T_{P_t}(x_t)]$, Formula (21).
      - Update $\Xi^-_{P_t,i}$, Formula (22)
   (b) Update $\rho_i(R_t)$:
      - Calculate $\mathbb{E}^{(d)}_{\rho_i(x_t,P_t)}[\sum_{j\in\mathcal{I}_i} T_{R_t}(y_{j,t})]$, Formula (47).
      - Update $\Xi^-_{R_t,i}$, Formula (26)
   (c) Update $\rho_i(x_t)$:
      - Prepare estimates $\hat{P}^{\star,(d)}_{i,t}$, $\hat{R}^{(d)}_{i,t}$, and their inverses, Formulas (23), (24), (27), and (28).
      - Calculate $\mathbb{E}^{(d)}_{\rho_i(P_t,R_t)}[\sum_{j\in\mathcal{I}_i} T_{x_t}(y_{j,t})]$, Formula (45).
      - Update $\Xi^-_{x_t,i}$, Formula (30)

Combination step

1. Acquire posterior hyperparameters $\Xi^+_{x_t,j}$ and $\Xi^+_{R_t,j}$ of neighbors $j \in \mathcal{I}_i$.
2. Calculate the combined posterior hyperparameters $\widetilde{\Xi}^+_{x_t,j}$ and $\widetilde{\Xi}^+_{R_t,j}$, Formulas (52) and (53).

performance of the standard Kalman filter requiring a fully specified model. Examples 2 and 3 assume a network of 15 agents and evaluate a comparative study of four scenarios: ATC – the adapt-then-combine variational filter (Algorithm 2), NOCOOP – the no-cooperation scenario assuming isolated agents, FC – the fusion center scenario where a dedicated agent processes all measurements using Algorithm 1, and finally KF – the adapt-then-combine diffusion Kalman filter [15, 16] proceeding with a full knowledge of the state-space model.

The examples assume a 2D target tracking from trajectories simulated by a constant velocity model corresponding to (1). The simulation period is $\tau = 1$ time step. The matrices $A \equiv A_t$ and $H \equiv H_t$ are

$$A = \begin{bmatrix} 1 & 0 & \tau & 0 \\ 0 & 1 & 0 & \tau \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \tag{57}$$

the variables B_t and u_t are absent in the model. The independent zero-mean process noise ω_t is characterized by its (constant) covariance matrix

$$Q_t = 0.5 \begin{bmatrix} \frac{\tau^3}{3} & 0 & \frac{\tau^2}{2} & 0 \\ 0 & \frac{\tau^3}{3} & 0 & \frac{\tau^2}{2} \\ \frac{\tau^2}{2} & 0 & \tau & 0 \\ 0 & \frac{\tau^2}{2} & 0 & \tau \end{bmatrix}, \tag{58}$$

and except for Example 3, the independent measurement noise $\varepsilon_{i,t}$ is zero-centered Gaussian noise with the (constant) covariance matrix

$$R_t = \begin{bmatrix} 100^2 & 0 \\ 0 & 100^2 \end{bmatrix}. \tag{59}$$

The variational filters are always initialized with

$$x_0 \sim \mathcal{N}(\hat{x}_{i,t}^{-}, \hat{P}_{i,t}^{\star}) \equiv \mathcal{N}\left([0,0,0,0]^{\intercal}, 100 \cdot I_{[4\times 4]}\right), \tag{60}$$

$$P_0 \sim i\mathcal{W}(\Psi_{i,t}^{-}, \psi_{i,t}^{-}) \equiv i\mathcal{W}(100 I_{[4\times 4]}, 10), \tag{61}$$

$$R_0 \sim i\mathcal{W}(\Phi_{i,t}^{-}, \phi_{i,t}^{-}) \equiv i\mathcal{W}(100 I_{[2\times 2]}, 4), \tag{62}$$

where I is the identity matrix of indicated dimensions. All examples assume

$$\mathcal{Q}_{i,t} = \{0.01 I_{[4\times 4]}, 0.1 I_{[4\times 4]}, I_{[4\times 4]}, 10 I_{[4\times 4]}, 100 I_{[4\times 4]}\}.$$

At each time step t, the variational algorithms run $D = 4$ CAVI iterations. Our investigation has shown that $D \geq 3$ yields almost identical results. The initial setting of the standard Kalman filter assumes the same Gaussian prior for x_t as the variational filters, and proceeds with true Q_t and R_t. We focus on the estimation quality of the first two elements of x_t representing the the target coordinates, and on the estimates of R_t.

5.1 Example 1: Comparison with the standard Kalman filter

The first example studies the performance of the basic variational filter summarized in Algorithm 1. For this purpose, we independently simulate 200 trajectories initiated from

$x_0 = [500, 500, 0, 0]^\intercal$ that are observed by a single agent. The simulations are 600 time steps long. Figure 3 shows one randomly selected trajectory. The variational filter is run with two different values of the forgetting factor α_R, namely 1.0 and 0.99 (lower values are not used as R_t is constant). The estimates resulting from these two settings and from the standard Kalman filter exploiting the true values of the covariance matrices Q_t and R_t are compared.

Figure 4 shows the detail of the first 200 points of the trajectory from Figure 3 decomposed to target coordinates $x_{1,t}$ and $x_{2,t}$. As the filters are initiated from [0,0], there is a short adapting period after which they closely follow the true trajectory. Apparently, a small amount of forgetting ($\alpha_R = 0.99$) does not noticeably influence the state estimation. We attribute this behavior to the optimization of $\hat{P}_{i,t}$ and the procedure of selecting $\hat{Q}_{i,t}$. Except for the initial stabilization period, the evolution of the state estimates is very close to that of the Kalman filter. Figure 5 which is also related to the same trajectory depicts the estimates of the diagonal elements of R_t. Here, the forgetting factor evidently plays a role.

The performance in terms of RMSE averaged over all 200 independent trajectories depict Figures 6 and 7. They support the claims from the previous paragraph.

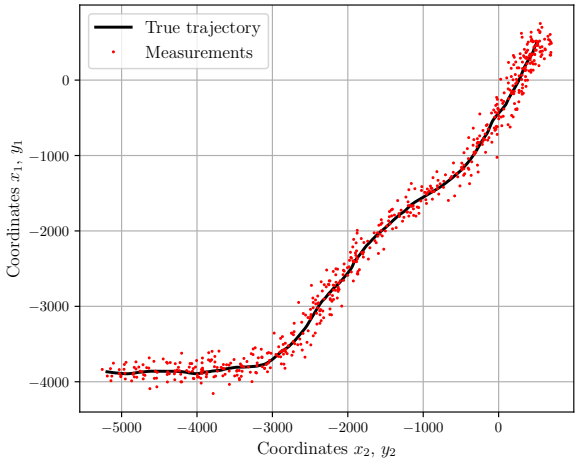

Figure 3: Example 1: A randomly selected example of generated trajectory.

5.2 Example 2: Distributed estimation under constant R_t

The second example assumes a network of 15 agents depicted in Figure 8. The agents independently observe measurements of target trajectories starting from $x_0 = [0, 0, 0, 0]^\intercal$ and with a constant $R_t = 100^2 I_{[2\times 2]}$. The horizon of simulation is 600 time steps. The initial settings are as described earlier, the forgetting factor $\alpha_R = 0.99$.

Four scenarios are compared: (ATC) – the proposed adapt-then-combine variational filter, (FC) – the fusion center-based scenario where all measurements are processed at a dedicated node using Algorithm 1, (NOCOOP) where the agents do not cooperate at all, and (KF) standing for the adapt-then-combine diffusion Kalman filter [15, 16].

Figures 9 and 10 provide the target coordinate estimates RMSE and the R_t estimates RMSE averaged over the network and 300 independent runs (trajectories). It is possible to conclude that the proposed method significantly accelerates the convergence of estimates

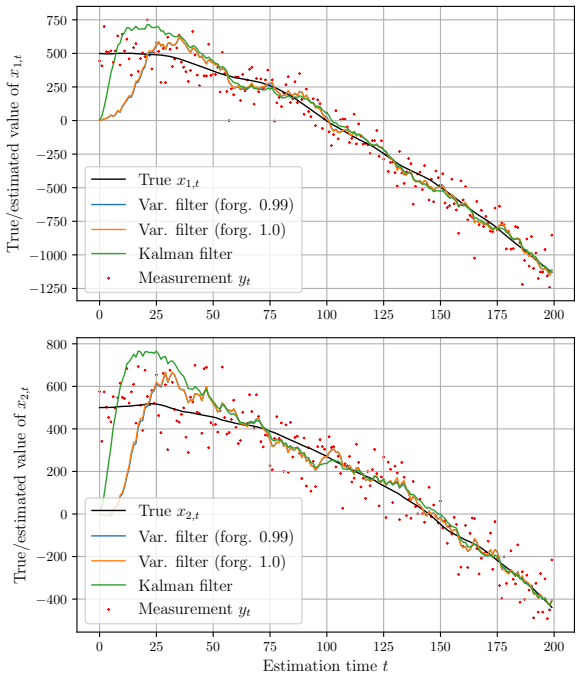

Figure 4: Example 1: Trajectory presented in Fig. 3 – a detail of its true evolution and estimates (200 time steps). The forgetting factor α_R does not noticeably influence the state estimates.

to the true values. The performance of the proposed ATC variational filter is close to the performance of the fusion center (FC) both in the estimation of the states and measurement noise covariance matrix. Moreover, both the ATC filter and the fusion center-based variational estimation provide only slightly worse results than the diffusion Kalman filter (KF) employing known covariance matrices.

5.3 Example 3: Distributed estimation under time-varying R_t

The third example demonstrates the behavior of the estimator under a time-varying measurement noise covariance matrix R_t and three different values of α_R, namely 0.99, 0.975, and 0.95. The diagonal elements of R_t evolve according to Figure 11 (top). The simulation horizon is 1400 time steps. Otherwise, the settings of the agents and the trajectory-generating mechanism coincide with the previous example. The results are averaged over 300 independent experiment runs, too.

Besides the evolution of the diagonal elements of R_t, Figure 11 depicts the average RMSE of both target coordinates associated with three different settings of the forgetting factor α_R. As in Example 1 it is possible to state that its impact on the state estimation is very low. The estimation performance of the scenarios ATC and FC is close to the diffusion Kalman filter (KF) operating with known Q_t and R_t.

Figure 12 depicts the RMSE of the measurement noise covariance matrix estimates. Naturally, the value of α_R plays an important role in this example. The ATC and FC scenarios still attain significantly better estimation performance than NOCOOP, but there is a pronounced sensitivity to the time-varying nature of R_t under higher α_R. The reason behind this behavior

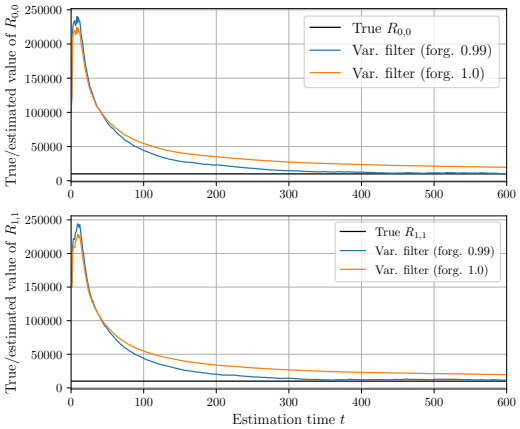

Figure 5: Example 1: The diagonal elements of R_t and its estimates for the selected trajectory. The forgetting factor α_R may accelerate the convergence.

is that the prior/posterior distributions encompass significantly more information about the (previous) process behavior. Unless this information is adequately suppressed by forgetting, it prevents to flexible response to changes in the measurements-generating process.

6 Conclusion

The state-of-the-art distributed Kalman filters naturally inherit the robustness issues of the standard Kalman filters. In particular, if the model is not fully known and well-specified, the estimation performance is degraded, or the filters may completely diverge. Seeing this as an important issue, we propose a novel algorithm that aims at relaxing the assumption of known noise moments, namely the measurement and process noise covariance matrices. The measurement noise covariance is inferred simultaneously with the unknown state variables. The problem of unknown process noise covariance matrix is circumvented by a exploiting a set of convenient candidate matrices and a Bayesian hypotheses testing procedure. The main advantage of the resulting filter is its analytically tractable form and an excellent performance that is close to the diffusion Kalman filter operating with a complete knowledge of the system model. The filter has a low number of tuning parameters. The experimental results indicate its robustness to slowly time-varying measurement noise covariance matrix.

The future work will focus on nonlinear state-space models and potentially spatial heterogeneity of the measurement noise distribution across the network. Also, the local or collaborative optimization of process noise covariance matrix (or its candidates) could be beneficial and deserves our attention.

A (Properties of probability distributions)

The appendix defines selected properties of probability distributions used in the paper. In particular, the natural parameter vectors, the sufficient statistics, and the point estimators (the expected values) are given. The multivariate elements of the sufficient statistics and conjugate hyperparameters are assumed to be vectorized. Motivated by easier reading, we

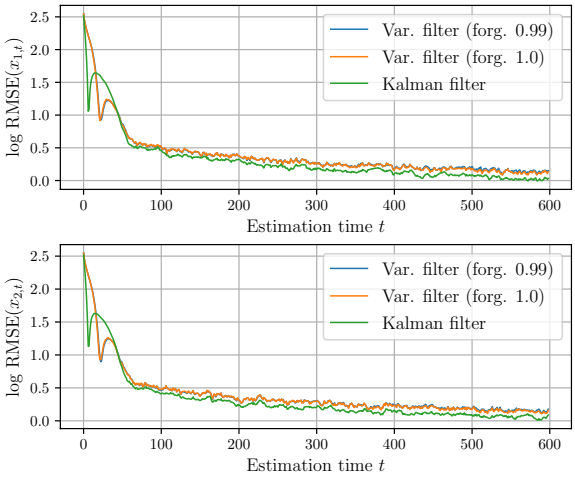

Figure 6: Example 1: Decimal logarithm of RMSE of state estimates averaged over 200 independent experiment runs. Apparently, the performance of both variational filter settings are almost identical, and it quickly gets close to the standard Kalman filter.

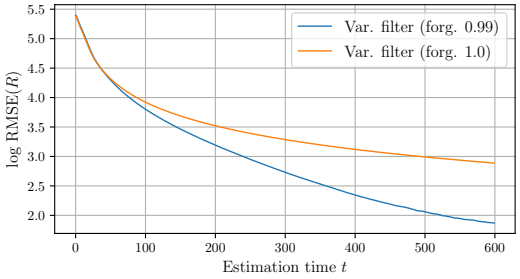

Figure 7: Example 1: Decimal logarithm of RMSE of R_t estimates averaged over 200 independent experiment runs.

omit the vec operators from the formulas.

Definition A.1. *Assume 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 $R_t \in \mathbb{R}^{m\times m}$. With respect to its parameterisation, the probability density function has the following forms*

$$
\begin{aligned}
f_i(y_t|x_t,R_t) &= (2\pi)^{-\frac{m}{2}}|R_t|^{-\frac{1}{2}}\exp\left\{\frac{-1}{2}(y_t-H_tx_t)^\intercal R_t^{-1}(y_t-H_tx_t)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\operatorname{Tr}\left(\underbrace{\begin{bmatrix}-1\\ x_t\end{bmatrix}\begin{bmatrix}-1\\ x_t\end{bmatrix}^\intercal}_{\eta_{x_t}}\underbrace{\begin{bmatrix}y_t^\intercal\\ H_t^\intercal\end{bmatrix}R_t\begin{bmatrix}y_t^\intercal\\ H_t^\intercal\end{bmatrix}^\intercal}_{T_{x_t}(y_t)}\right)\right\} \qquad (63)\\
&\propto \exp\left\{-\frac{1}{2}\operatorname{Tr}\left(\underbrace{\begin{bmatrix}(R_t^{-1})^\intercal\\ \ln|R_t|\end{bmatrix}}_{\eta_R}\underbrace{\begin{bmatrix}(y_t-H_tx_t)(y_t-H_tx_t)^\intercal\\ 1\end{bmatrix}}_{T_R(y_t)}\right)\right\}. \qquad (64)
\end{aligned}
$$

In order to estimate the state variable x_t, we exploit the Gaussian distribution as the

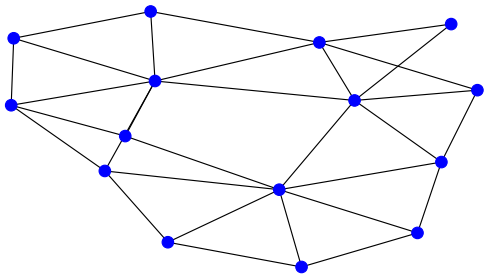

Figure 8: Network used in Examples 2 and 3.

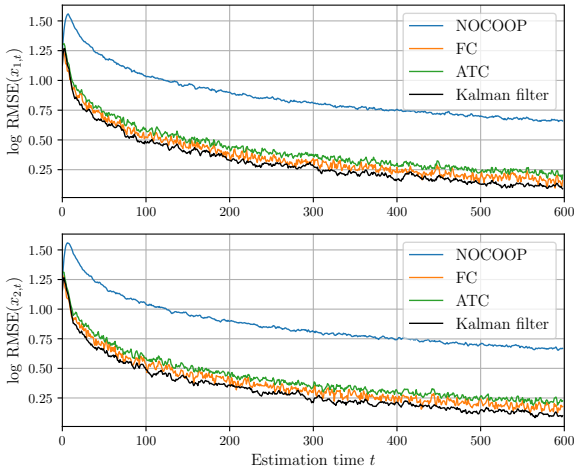

Figure 9: Example 2: Decimal logarithm of RMSE of state estimates averaged over 300 independent experiment runs.

prior. It also serves for the inference of P_t.

Definition A.2. *Assume an n-dimensional random 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 the $n \times n$ positive definite covariance matrix. The probability density function can be written in the forms*

$$\pi_x(x_t|\hat{x}_t, P_t) = (2\pi)^{-\frac{n}{2}} |P_t|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\hat{x}_t - x_t)^\intercal (P_t)^{-1}(\hat{x}_t - x_t)\right\} \tag{65}$$

$$\propto \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left(\underbrace{\begin{bmatrix} -1 \\ x_t \end{bmatrix} \begin{bmatrix} -1 \\ x_t \end{bmatrix}^\intercal}_{\eta_{x_t}} \underbrace{\begin{bmatrix} \hat{x}_t^\intercal \\ I \end{bmatrix} P_t^{-1} \begin{bmatrix} \hat{x}_t^\intercal \\ I \end{bmatrix}^\intercal}_{\Xi_{x_t}} \right) \right\} \tag{66}$$

$$\propto \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left(\underbrace{\begin{bmatrix} P_t^{-1} \\ \ln|P_t| \end{bmatrix}^\intercal}_{\eta_{P_t}} \underbrace{\begin{bmatrix} (x_t - \hat{x}_t)(x_t - \hat{x}_t)^\intercal \\ 1 \end{bmatrix}}_{T_{P_t}(x_t)} \right) \right\}. \tag{67}$$

Lemma A.1. *Assume the conjugate form of the Gaussian probability density* (66) *with a*

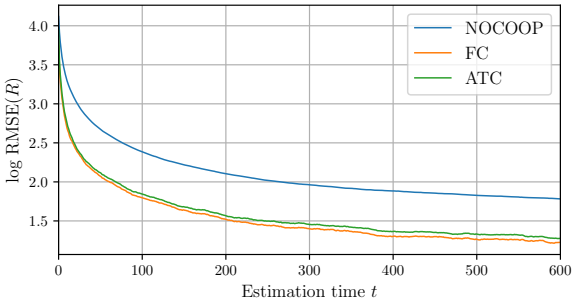

Figure 10: Example 2: Decimal logarithm of RMSE of the measurement noise covariance matrix estimates averaged over 300 independent experiment runs.

hyperparameter Ξ_{x_t} *written in the block-matrix form*

$$\Xi_{x_t} = \begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \end{bmatrix}_{x_t} = \begin{bmatrix} \hat{x}_t^\intercal P_t^{-1} \hat{x}_t & \hat{x}_t^\intercal P_t^{-1} \\ P_t^{-1} \hat{x}_t & P_t^{-1} \end{bmatrix}. \tag{68}$$

The transformation $\Xi_{x_t} \to (\hat{x}_t, P_t)$ *is as follows:*

$$\begin{aligned} P_t &= (\xi_{22})^{-1}, \\ \hat{x}_t &= \xi_{22}^{-1} \xi_{21} = P_t \xi_{21}. \end{aligned} \tag{69}$$

The proof is trivial.

The inference of the measurement noise covariance matrix R_t exploits the inverse-Wishart distribution. Its definition follows.

Definition A.3. *Assume a positive definite measurement noise covariance matrix* $R_t \in \mathbb{R}^{m \times m}$. *A convenient model for its estimation is the inverse-Wishart distribution,* $R_t \sim i\mathcal{W}(\Phi_t, \phi_t)$ *with parameters* $\Phi_t \in \mathbb{R}^{m \times m}$ *and* $\phi_t > 0$. *Its probability density function can be written in the forms*

$$\begin{aligned} \pi_R(R_t | \phi_t, \Phi_t) &= \frac{|\Phi_t|^{\frac{\phi_t}{2}}}{2^{\frac{n\phi_t}{2}} \Gamma_n\left(\frac{\phi_t}{2}\right)} |R_t|^{-\frac{\phi_t + n + 1}{2}} \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left(\Phi_t (R)^{-1}\right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left(\underbrace{\begin{bmatrix} R_t^{-1} \\ \ln |R_t| \end{bmatrix}^\intercal}_{\eta_{R_t}} \underbrace{\begin{bmatrix} \Phi_t \\ \phi_t + m + 1 \end{bmatrix}}_{\xi_{R_t}} \right) \right\}, \end{aligned} \tag{70}$$

where $\Gamma_n(\cdot)$ *is the multivariate gamma function. The expected values of* R_t *and* R_t^{-1} *read*

$$\mathbb{E}[R_t] = \hat{R}_t = \frac{\Phi_t}{\phi_t - m - 1}, \tag{71}$$

$$\mathbb{E}[R_t^{-1}] = \hat{R}_t^{-1} = \phi_t \Phi_t^{-1}. \tag{72}$$

The inference of P_t exploits the inverse-Wishart distribution too. The following definition – given for completeness – is analogous to Definition A.3.

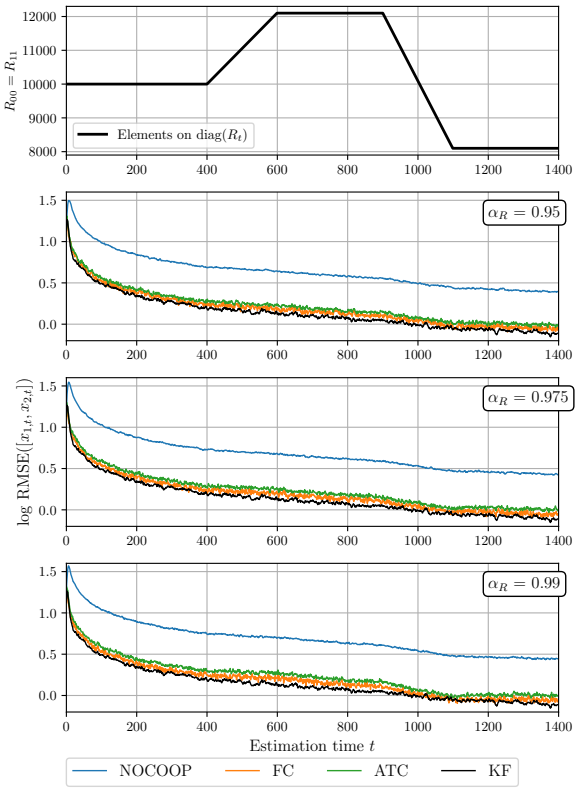

Figure 11: Example 3: Decimal logarithm of RMSE of state estimates averaged over 300 independent experiment runs. The evolution of the diagonal elements of R_t are given on the top.

Definition A.4. *Assume a positive definite covariance matrix $P_t \in \mathbb{R}^{n\times n}$. A convenient model for its estimation is the inverse-Wishart distribution, $P_t \sim i\mathcal{W}(\Psi_t, \psi_t)$ with parameters $\Psi_t \in \mathbb{R}^{n\times n}$ and $\psi_t > 0$. Its probability density function can be written in the forms*

$$\pi(P_t|\psi_t, \Psi_t) \propto \exp\left\{ -\frac{1}{2}\operatorname{Tr}\left(\underbrace{\begin{bmatrix} P_t^{-1} \\ \ln|P_t| \end{bmatrix}^{\intercal}}_{\eta_{P_t}} \underbrace{\begin{bmatrix} \Psi_t \\ \psi_t + n + 1 \end{bmatrix}}_{\xi_{P_t}} \right) \right\}. \tag{73}$$

The expected values of P_t and P_t^{-1} are

$$\mathbb{E}[P_t] = \hat{P}_t = \frac{\Psi_t}{\psi_t - n - 1}, \tag{74}$$

$$\mathbb{E}[P_t^{-1}] = \hat{P}_t^{-1} = \psi_t \Psi_t^{-1}. \tag{75}$$

References

[1] Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E. Wireless sensor networks: A survey. *Comput Netw.* 2002; 38(4):393–422.

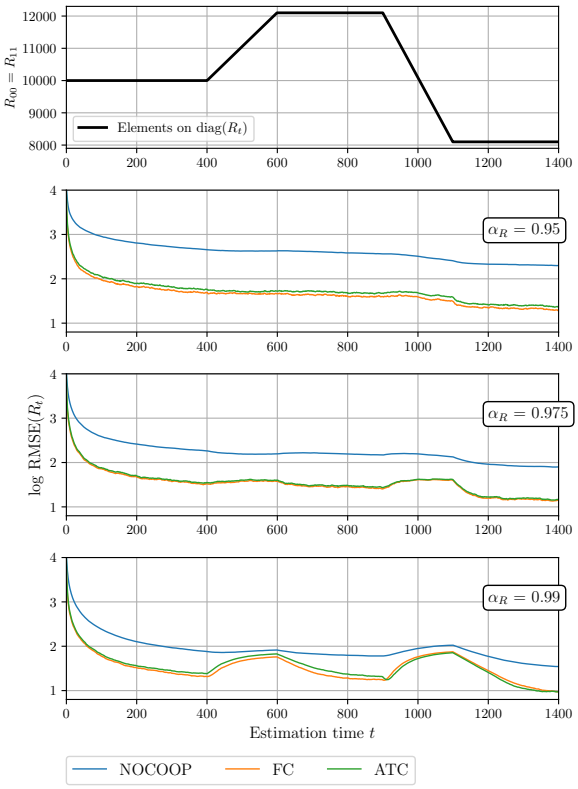

Figure 12: Example 3: Decimal logarithm of RMSE of the measurement noise covariance matrix estimates averaged over 300 independent experiment runs. The evolution of the diagonal elements of R_t are given on the top.

[2] Sayed AH. Adaptation, learning, and optimization over networks. *Foundations and Trends in Machine Learning*. 2014; 7(4-5):311–801.

[3] Zorzi M. Distributed Kalman filtering under model uncertainty. *IEEE T Control Netw Sys*. 2020; 7(2):990–1001.

[4] Sayed AH. Adaptive networks. *Proc IEEE*. 2014; 102(4):460–497.

[5] Djurić PM, Richard C. *Cooperative and Graph Signal Processing*. Cambridge, MA: Academic Press; 2018. ISBN 978-0-12-813677-5.

[6] Tsitsiklis JN, Bertsekas DP, Athans M. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. In Proc. American Control Conference, Jun. 6–8, 1984; San Diego, CA. (pp. 484–489).

[7] Braca P, Marano S, Matta V. Running consensus in wireless sensor networks. In Proc. 11th International Conferences on Information Fusion, Jun. 30– Jul. 03, 2008; Cologne, Germany.

[8] Kar S, Moura JMF. Asymptotically efficient distributed estimation with exponential family statistics. *IEEE T Inform Theory*. 2014; 60(8):4811–4831.

[9] Cattivelli FS, Lopes CG, Sayed AH. Diffusion recursive least-squares for distributed estimation over adaptive networks. *IEEE T Signal Proces.* 2008; 56(5):1865–1877.

[10] Cattivelli FS, Sayed AH. Diffusion strategies for distributed Kalman filtering and smoothing. *IEEE T Automat Contr.* 2010; 55(9):2069–2084.

[11] Cattivelli FS, Sayed AH. Diffusion LMS strategies for distributed estimation. *IEEE T Signal Proces..* 2010; 58(3):1035–1048.

[12] Chen J, Sayed AH. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE T Signal Proces.* 2012; 60(8):4289–4305.

[13] Tu S-Y, Sayed AH. Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks. *IEEE T Signal Proces.* 2012; 60(12):6217–6234.

[14] Sayed AH. Diffusion Adaptation over Networks. In *Academic Press Library in Signal Processing*. Academic Press, 2014. 3:323–454.

[15] Dedecius K, Djurić PM. Sequential estimation and diffusion of information over networks: A Bayesian approach with exponential family of distributions. *IEEE T Signal Proces.* 2017; 65(7):1795–1809.

[16] Hu J, Xie L, Zhang C. Diffusion Kalman filtering based on covariance intersection. *IEEE T Signal Proces.* 2012; 60(2):891–902.

[17] Arablouei R, Werner S, Huang Y-F, Dogancay K. Distributed least mean-square estimation with partial diffusion. *IEEE T Signal Proces.* 2014; 62(2):472–484.

[18] Arablouei R, Dogancay K, Werner S, Huang Y-F. Adaptive distributed estimation based on recursive least-squares and partial diffusion. *IEEE T Signal Proces.* 2014; 62(14):3510–3522.

[19] Vahidpour V, Rastegarnia A, Khalili A, Sanei S. Partial diffusion Kalman filtering for distributed state estimation in multiagent networks. *IEEE T Neur Net Lear.* 2019; 30(12):3839–3846.

[20] Vahidpour V, Rastegarnia A, Latifi M, Khalili A, Sanei A. Performance analysis of distributed Kalman filtering with partial diffusion over noisy network. *IEEE T Aero Elec Sys.* 2020; 56(3):1767–1782.

[21] Khalili A, Vahidpour V, Rastegarnia A, Bazzi WM, Sanei S. Partial diffusion Kalman filter with adaptive combiners. *IEEE T Aero Elec Sys.* 2020; 57(3):1972–1980.

[22] Shmaliy YS, Lehmann F, Zhao S, Ahn CK. Comparing robustness of the Kalman, H_∞, and UFIR Filters. *IEEE T Signal Proces..* 2018; 66(13):3447–3458.

[23] Zhou T. Coordinated one-step optimal distributed state prediction for a networked dynamical system. *IEEE T Automat Contr.* 2013; 58(11):2756–2771.

[24] Dedecius K, Tichý O. Collaborative sequential state estimation under unknown heterogeneous noise covariance matrices. *IEEE T Signal Proces.* 2020; 68:5365–5378.

[25] Huang Y, Zhang Y, Wu Z, Li N, Chambers J. A novel adaptive Kalman filter with inaccurate process and measurement noise covariance matrices. *IEEE T Automat Contr.* 2018; 63(2):594–601.

[26] Huang Y, Zhang Y, Shi P, Chambers J. Variational adaptive Kalman filter with Gaussian-inverse-Wishart mixture distribution. *IEEE T Automat Contr.* 2021; 66(4):1786–1793.

[27] Peterka V. Bayesian approach to system identification. In *Trends and Progress in System Identification.* Oxford, U.K.: Pergamon Press; 1981 (pp. 239–304).

[28] Simon D. *Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches.* Hoboken, NJ: Wiley; 2006. ISBN 0-471-70858-5.

[29] Särkka S, Nummenmaa A. Recursive noise adaptive Kalman filtering by variational Bayesian approximations. *IEEE T Automat Contr.* 2009; 54(3):596–600.

[30] Särkka S, Hartikainen J. Non-linear noise adaptive Kalman filtering via variational Bayes. In Proc. IEEE Int Works Mach (MLSP 2013); Sept. 22–25, 2013; Southampton, UK.

[31] Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. *Mach Learn.* 1999; 37(2):183–233.

[32] Winn J, Bishop CM. Variational message passing. *J Mach Learn Res.* 2005; 6:661–694.

[33] Bishop CM. *Pattern Recognition and Machine Learning*, Cambridge, UK: Springer; 2006. ISBN 0-387-31073-8.

[34] Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. *J Am Stat Assoc.* 2017; 112(518):859–877.

[35] Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* Boca Raton, FL: Chapman & Hall/CRC, 2003. ISBN 1-58488-388-X.

[36] Kárný M, Böhm J, Guy TV, Jirsa L, Nagy I, Nedoma P, Tesař L. *Optimized Bayesian Dynamic Advising: Theory and Algorithms.* London: Springer; 2006. ISBN 1-85233-928-4.

[37] Smith RL, Bayesian and frequentist approaches to parametric predictive inference. In *Bayesian Statistics.* Oxford University press; 1999; 6:589–612.

[38] Khaleghi B, Khamis A, Karray FO, Razavi SN. Multisensor data fusion: A review of the state-of-the-art. *Inform Fusion.* 2013; 14(1):28–44.

[39] Krishnamurthy V, Poor HV. A tutorial on interactive sensing in social networks. *IEEE T Comp Soc Syst.* 2014; 1(1):3–21.

[40] Cichocki A, Amari S-I. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy.* 2010; 12:1532–1568.

[41] Dedecius K, Information fusion with functional Bregman divergence. Czech Acad. Sci., Institute of Information Theory and Automation, Prague, Tech. Rep., 2015. Available: http://library.utia.cas.cz/separaty/2015/AS/dedecius-0438485.pdf

[42] Julier SJ, Uhlmann JK. A non-divergent estimation algorithm in the presence of unknown correlations. In Proc. 1997 American Control Conference; June 4–6, 1997; Albuquerque. (pp. 2369–2373).