Collaborative sequential state estimation under unknown heterogeneous noise covariance matrices

Kamil Dedecius and Ondřej Tichý

***Abstract*—We study the problem of distributed sequential estimation of common states and measurement noise covariance matrices of hidden Markov models by networks of collaborating nodes. We adopt a realistic assumption that the true covariance matrices are possibly different (heterogeneous) across the network. This setting is frequent in many distributed real-world systems where the sensors (e.g., radars) are deployed in a spatially anisotropic environment, or where the networks may consist of sensors with different measuring principles (e.g., using different wavelengths). Our solution is rooted in the variational Bayesian paradigm. In order to improve the estimation performance, the measurements and the posterior estimates are communicated among adjacent neighbors within one network hop distance using the information diffusion strategy. The resulting adaptive algorithm selects neighbors with compatible information to prevent degradation of estimates.**

***Index Terms*—Diffusion network, diffusion strategy, state estimation, Kalman filtering, variational Bayesian methods.**

I. Introduction

DISTRIBUTED inference of stochastic model parameters and states in networks of collaborating nodes (agents) has attracted tremendous interest in the last years. The vast spectrum of possible applications ranges from sensor networks, target tracking systems, and social networks up to the highly progressive phenomenon of the Internet of Things (IoT) [1], [2].

The network-based inference algorithms may be categorized with respect to several criteria. The mode of communication gives rise to the consensus, incremental, and diffusion strategies [3]. The incremental strategy relies on a Hamiltonian cycle passing through every node of the network, through which the individual estimates are gradually refined. The drawback is evident: each node and link of the network is a potential single point of failure, and recovery from failure – arranging a new cycle – is an NP-hard problem. This is prohibitive for larger networks with dynamic topology [4]. The consensus and diffusion strategies enjoy attractive robustness, as each network node collaborates with its adjacent neighbors, mostly within one network hop distance. The consensus-based algorithms mostly rely on several intermediate iterations among the nodes, while the diffusion algorithms typically involve only two communication steps. In the *adaptation* step, each node assimilates neighbors' measurements into its own statistical knowledge about the inferred variables. The *combination* step then serves for the exchange and incorporation of the local estimates. One or both of these steps may be used, in the latter case in the adapt-then-combine (ATC) or the reversed combine-then-adapt (CTA) order. As the ATC diffusion strategy was shown to outperform the CTA strategy [3], [5], this paper focuses on ATC. Sometimes, the nodes are allowed to communicate with one randomly or deterministically selected neighboring node, possibly with intermediate iterations. These so-called *gossip* or *random-exchange* algorithms are either included in the consensus or the diffusion strategy [6]–[8], or perceived as a separate group [9].

If networks are used for observing and modeling stochastic processes, we often face the issue of spatial anisotropy of observing conditions, e.g., different noise distributions. Similarly, the devices that could be deployed in such networks may use different measuring principles which limits the degree of collaboration among them. We focus on this kind of problems. In particular, we consider the inference of state-space models with unknown and possibly heterogeneous observation noise covariances. The aim is to exploit the spatial or principal diversity of the signals in order to improve the global network modeling performance.

In non-distributed settings, the inference of unknown covariance matrices of the state-space models has a long history, dating back to the seminal work [10]. In order to remove the computational burden, the recent solutions are mostly based on the variational Bayesian methods [11], and comprise the variational Bayesian Kalman filter [12], [13] and its extensions to the cases of unknown process noise covariance [14], [15]. To present, their *distributed* counterparts have been largely based on the consensus communication strategies, e.g., [16], [17] and the references therein. The problem of unknown measurement noise covariances has been tackled rather sparingly. The combination of the nonlinear cubature Kalman filter integrating the variational Bayesian estimation of the *global* measurement noise covariance matrix was proposed in [18]. In [19], the authors devise a distributed consensus linear filter based on the H_∞ filtering and interacting multiple model algorithm, naturally with the advantages and drawbacks of the underlying H_∞ approach. The diffusion strategy-based state estimation algorithms started with the basic diffusion Kalman filter [20]. Its combination step involved only the point estimates of the state variable, leaving the associated covariance matrices intact. The work [21] modifies this step by involving the covariance intersection procedure. The solution

Manuscript received XXX; revised XXX. First published XXX; current version published XXX. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Subhro Das. The work of O. Tichý was supported by the Czech Science Foundation, grant no. GA20-27939S.

K. Dedecius and O. Tichý are with the Institute of Information Theory and Automation, Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 08 Prague 8, Czech Republic (email: dedecius,otichy@utia.cas.cz).

Digital Object Identifier XXX/XXX

coincides with the special case of the sequential Bayesian estimator [5]. Despite the multitude of algorithms focused on advanced aspects like the communication effectiveness or robustness, the authors of the present paper are not aware of any analytically tractable method allowing simultaneous diffusion-based estimation of states and measurement noise covariances in general, let alone in the mentioned heterogeneity case. In the particle filtering realm, there exists a distributed solution [23] providing joint state and parameter estimation with reduced communication overhead, but it does not closely focus on the problem of shared covariance matrices. The present paper aims at filling the gap in the field of analytically tractable algorithms.

To summarize, we are interested in scenarios where the network nodes observe a stochastic process and employ state-space models for its description. It is assumed that the nodes' measurements are subject to noise that may statistically vary from node to node due to their spatial locations or different measuring principles. That is, there are clusters of agents that are interested in estimating different parameter vectors, where part of the vector is common to all nodes while other part differs among clusters (for more information about inference over clustered networks see [24]–[26]). Besides the traditional task of sequential (online) filtration of states they face the need for estimation of the associated measurement noise covariance matrices, where interference among clusters may degrade estimation quality. Inspired by the principles of the variational Bayesian estimation in networks [22] and variational non-distributed Kalman filtration [12], [13], we propose an efficient algorithm for joint inference of states and covariances with information diffusion. It is described from the perspective of a single network node aware solely of its adjacent neighbors, but lacking any knowledge of their compatibility regarding the noise properties. The determination of this compatibility is a part of the solution. The nodes do not rely on fixed network topology. The resulting algorithm thus enables the nodes to be correctly clustered and to run sequential inference with improved performance through inter-cluster collaboration.

The paper is organized as follows: Section II formulates the studied problem. Section III summarizes the Bayesian inference theory under conjugate prior distributions that is extensively used in the sequel. In Section IV, the novel distributed Bayesian filter is proposed. In particular, the local prediction step, the measurement update step with diffusion adaptation, and the combination step fusing the estimates are devised. Section V discusses the filter properties, the performance is subsequently studied on simulation examples in Section VI. Finally, Section VII concludes the work.

II. Problem statement

Let us assume a network represented by a connected undirected graph $(\mathcal{I}, \mathcal{E})$ consisting of a set of nodes $\mathcal{I} = \{1, \ldots, |\mathcal{I}|\}$ where $|\mathcal{I}|$ denotes the cardinality of $\mathcal{I}$. The set of edges $\mathcal{E}$ defines the network topology, i.e., the communication paths among nodes. Each node $i \in \mathcal{I}$ is allowed to communicate only with its *adjacent* neighbors that form its closed neighborhood $\mathcal{I}^{(i)}$ (i belongs to $\mathcal{I}^{(i)}$ too). At each discrete time instant $k = 1, 2, \ldots$, we limit the exchange of any available information (the measurements and/or the estimates) between two nodes to be performed at most once. This type of communication is known as the diffusion strategy [2], [3].

In our examined problem, the network nodes $i \in \mathcal{I}$ acquire univariate or multivariate noisy measurements $y_k^{(i)} \in \mathbb{R}^n$, $n \geq 1$, of a hidden Markov process

$$x_k = A_k x_{k-1} + B_k u_k + w_k, \tag{1a}$$

$$y_k^{(i)} = H_k x_k + v_k^{(i)}, \tag{1b}$$

where the state variable $x_k \in \mathbb{R}^p$, $p \geq 1$, is common to all the nodes, u_k – if exists – is a globally known control variable, A_k, B_k and H_k are known matrices of compatible dimensions, and $w_k \in \mathbb{R}^p$ is an independent process noise

$$w_k \sim \mathcal{N}(0, Q), \qquad Q \in \mathbb{R}^{p \times p}.$$

The measurement noise variable $v_k^{(i)} \in \mathbb{R}^n$ is an independent identically distributed zero-centered variable that is potentially heterogeneous with respect to i,

$$v_k^{(i)} \sim \mathcal{N}(0, R^{(i)}), \qquad R^{(i)} \in \mathbb{R}^{n \times n}.$$

The covariance matrices $R^{(i)}$ are unknown.

This formulation of the model (1) coincides the viewpoint of the node i that will be followed in the sequel. The global situation is as follows: There are several different covariance matrices $R_1, \ldots, R_L$ where $1 \leq L \leq |\mathcal{I}|$. For each node i the covariance matrix $R^{(i)}$ is equal to one of those matrices. The nodes that share the same covariance matrix $R_l, l \in \{1, \ldots, L\}$ are called *R-compatible* and constitute the set $\mathcal{I}_{R_l} \subseteq \mathcal{I}$. Since the nodes communicate only within their closed neighborhoods, we furthermore define the *R-compatible neighborhood* $\mathcal{I}_R^{(i)}$ that consists of those i's neighbors $j \in \mathcal{I}^{(i)}$ that belong to the same $\mathcal{I}_{R_l}$ as i, that is, their $R^{(j)} = R^{(i)}$ for all $j \in \mathcal{I}_R^{(i)}$. Besides the ignorance of own $R^{(i)}$, the nodes are not aware which of their neighbors are R-compatible.

The described situation, illustrated in Figure 1, is frequent in the cases where a single phenomenon (e.g., a flying target) with a state x_k (e.g., representing its position, velocity and acceleration) is observed by several instruments with different measuring principles, e.g., radars with different wavelengths, lidars etc. Then, the measurement noise distribution may be common for the instruments using the same measuring technology, and thus belonging to the same set $\mathcal{I}_{R_l}$.

The goal is to perform online collaborative filtering of the states x_k and estimation of the local $R^{(i)}$. In addition to own measurements $y_k^{(i)}$, the filtering algorithm should consider the neighbors' measurements $y_k^{(j)}$, as well as the neighbors estimates of x_k and possibly $R^{(j)}$. A procedure for the identification of the R-compatible neighbors and incorporation of the relevant information is to be devised.

III. Preliminaries on Bayesian inference

This section briefly summarizes the fundamentals of the Bayesian inference with conjugate prior distributions providing posterior estimates in the closed form. The theory, although being well-known, is worth reviewing for its extensive use in the subsequent exposition.

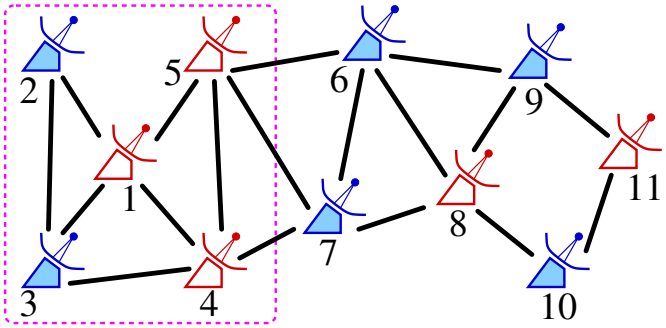

Fig. 1. Illustrative example of a diffusion network with two principally different instruments with two different measurement noise covariance matrices. The red (unshaded) nodes form the R-compatible set $\mathcal{I}_{R_1} = \{1, 4, 5, 8, 11\}$, and the blue (shaded) nodes form $\mathcal{I}_{R_2} = \mathcal{I} \setminus \mathcal{I}_1$. The closed neighborhood of node 1 is $\mathcal{I}^{(1)} = \{1, 2, 3, 4, 5\}$ and its R-compatible neighborhood is $\mathcal{I}_R^{(1)} = \{1, 4, 5\}$. However, the nodes are not aware of this configuration a priori.

The following definition is central to the statistical estimation theory, both the Bayesian and frequentists' one. It introduces the crucial concept of the sufficient statistic [27].

Definition 1 (Exponential family of distributions)**.** *A class of distributions of a uni- or multivariate random variable z parameterized by a vector θ is called the exponential family if the probability density function of z given θ can be written in the (non-unique) form*

$$f(z|\theta) = h(z)\exp\{\eta_\theta^\intercal T_\theta(z) - A(\eta_\theta)\}, \tag{2}$$

where $\eta_\theta \equiv \eta(\theta)$ is the natural parameter, $T_\theta(z)$ is the sufficient statistic carrying all the information contained in z about θ necessary for its estimation, $A(\eta_\theta)$ is the log-normalizing function and $h(z)$ is the base-measure function of z independent of θ. The family is called canonical if $\eta_\theta = \theta$, and curved if $\dim(\eta_\theta) > \dim(\theta)$.

The Bayesian theory estimates θ by virtue of the prior distribution summarizing all available knowledge about θ before assimilation of new measurements. If this distribution is *conjugate* to the model, then the assimilation results in the analytically tractable posterior distribution of the same functional type as the prior distribution [28]. This property is essential for sequential estimation from streaming data. The existence of the conjugate prior distributions is guaranteed for the exponential family models [29].

Definition 2 (Conjugate prior distribution)**.** *A prior distribution $\pi(\theta|\xi_\theta^-, \nu_\theta^-)$ is said to be conjugate to the model distribution $f(z|\theta)$ belonging to the exponential family, if its probability density function can be written in the form*

$$\pi(\theta|\xi_\theta^-, \nu_\theta^-) = l(\theta)\exp\{\eta_\theta^\intercal \xi_\theta^- - \nu_\theta^- A(\eta_\theta)\}, \tag{3}$$

where $\eta_\theta \equiv \eta(\theta)$ is the natural parameter of $f(z|\theta)$, $A(\eta_\theta)$ is the log-normalizing function of $f(z|\theta)$, $\dim(\xi_\theta^-) = \dim(T_\theta(z))$, and $\nu_\theta^- \in \mathbb{R}^+$. The parameters ξ_θ^- and ν_θ^- are called the hyperparameters.

Similarly to η_θ, the prior hyperparameters ξ_θ^- and ν_θ^- are transformed versions of the "standard" parameters of the concrete distributions. Sometimes, ν_θ is not used, or alternatively it may be absorbed in ξ_θ as its element.

Lemma 1 (Bayesian update)**.** *Assume a random variable z modeled by an exponential family distribution $f(z|\theta)$. Let $\pi(\theta|\xi_\theta^-, \nu_\theta^-)$ be the prior distribution for θ that is conjugate to the model $f(z|\theta)$. Assume $m \geq 0$ realizations of z denoted $z^{(1)}, \ldots, z^{(m)}$. Then, the Bayesian update yields the posterior distribution of θ given $\xi_\theta^+, \nu_\theta^+$ and $z^{(1)}, \ldots, z^{(m)}$ in the form*

$$\pi(\theta|\xi_\theta^+, \nu_\theta^+, z^{(1)}, \ldots, z^{(m)}) \propto \pi(\theta|\xi_\theta^-, \nu_\theta^-)\prod_{j=1}^{m} f(z^{(j)}|\theta), \tag{4}$$

where the posterior hyperparameters are given by

$$\xi_\theta^+ = \xi_\theta^- + \sum_{j=1}^{m} T_\theta(z^{(j)}),$$
$$\nu_\theta^+ = \nu_\theta^- + m.$$

The proof is given in Appendix A.

IV. Proposed Diffusion Filter

The state-space model (1) can be represented in the probabilistic form by the probability density functions

$$g(x_k|x_{k-1}) \equiv \mathcal{N}(A_k x_{k-1} + B_k u_k, Q), \tag{5}$$
$$f_i(y_k^{(i)}|x_k) \equiv \mathcal{N}(H_k x_k, R^{(i)}). \tag{6}$$

Consistently with the previous section, we denote by $\theta_k^{(i)}$ the vector of inferred variables,

$$\theta_k^{(i)} = [\![x_k, R^{(i)}]\!] \equiv \left[x_k^\intercal, \left(\mathrm{vec}(R^{(i)})\right)^\intercal\right]^\intercal, \tag{7}$$

For the sake of easier reading, we will stick with the double-bracket notation in the sequel. From the perspective of a node $i \in \mathcal{I}$, the Bayesian sequential estimation of $\theta_k^{(i)}$ proceeds with the prior distribution $\pi_i(\theta_k^{(i)}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_{k-1}^{(i)})$ where $\tilde{y}_{k-1}^{(i)}$ and $\tilde{u}_{k-1}^{(i)}$ represent *all* the information about the measurements and control variables up to time $k-1$ that is available to the ith node.

Let us focus on the evolution of the prior/posterior distributions at the ith node. The key steps of virtually any variant of the diffusion Kalman filter should be the following:

1) *Local prediction step*: The nodes perform the standard prediction, i.e., transition from the last posterior distribution from time $k-1$ to the prior distribution at the current time k using the state evolution equation (1a),

$$\pi_i(\theta_{k-1}^{(i)}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_{k-1}^{(i)}) \to \pi_i(\theta_k^{(i)}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)}). \tag{8}$$

2) *Measurement update step*: The nodes update the prior distribution (8) by their local measurements, or by the compatible neighbors' measurements. The latter is called the *diffusion adaptation*,

$$\pi_i(\theta_k^{(i)}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)}) \to \pi_i(\theta_k^{(i)}|\tilde{y}_k^{(i)}, \tilde{u}_k^{(i)}). \tag{9}$$

3) *Combination step*: The nodes share the posterior distributions with their compatible neighbors. These distributions are combined in order to improve the estimation performance,

$$\bigoplus\nolimits_{\forall j\in\mathcal{I}^{(i)}} \pi_j(\theta_k^{(j)}|\tilde{y}_k^{(j)},\tilde{u}_k^{(j)}) \to \bar{\pi}_i(\theta_k^{(i)}|Y_k^{(i)},U_k^{(i)}), \tag{10}$$

where $\bigoplus$ symbolizes a convenient combination operator, $Y_k^{(i)}$ and $U_k^{(i)}$ portray the combined measurements and control variables, respectively. The resulting distribution serves as the prior $\pi_i(\theta_{k-1}^{(i)}|\tilde{y}_{k-1}^{(i)},\tilde{u}_{k-1}^{(i)})$ for the next local prediction step in (8).

These three steps will be elaborated in the sequel. Since the Bayesian update – Lemma 1 – will be exploited in the measurement update step, we will start with it in order to decide the convenient form of the prior distribution, as this is a critical point. The local prediction step and the combination step will follow. Finally, Algorithm 1 will summarize the initialization and the steps of the resulting diffusion filter.

A. Measurement update step

During the measurement update step, the nodes either assimilate their own measurements,

$$\pi_i(\theta_k^{(i)}|\tilde{y}_k^{(i)},\tilde{u}_k^{(i)}) \propto f_i(y_k^{(i)}|\theta_k^{(i)})\pi_i(\theta_k^{(i)}|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)}), \tag{11}$$

or the measurements of their R-compatible neighbors,

$$\pi_i(\theta_k^{(i)}|\tilde{y}_k^{(i)},\tilde{u}_k^{(i)}) \propto \pi_i(\theta_k^{(i)}|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)}) \prod_{j\in\mathcal{I}_R^{(i)}} f_i(y_k^{(j)}|\theta_k^{(j)}), \tag{12}$$

where $\theta_k^{(j)}=\theta_k^{(i)}$ as expected. The latter version is a variant of the Bayesian *diffusion adaptation* [5].

The measurement model (6) is the normal distribution centered at $H_k x_k$ and with the node-specific covariance matrix $R^{(i)}$. The probability density function at the node i is

$$f_i(y_k|\theta_k^{(i)}) = (2\pi)^{-\frac{n}{2}}|R^{(i)}|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(y_k^{(i)}-H_k x_k)^\intercal (R^{(i)})^{-1}(y_k^{(i)}-H_k x_k)\right\}. \tag{13}$$

If the unknown $\theta_k^{(i)}$ were identically just the state variable x_k, the normal distribution

$$\pi_i(x_k|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)}) \equiv \mathcal{N}(\hat{x}_k^{(i),-},P_k^{(i),-}) \tag{14}$$

would be the convenient prior distribution for the closed-form sequential estimation of x_k according to Lemma 1 [5]. Under $\theta_k^{(i)}=[\![x_k,R^{(i)}]\!]$, the situation gets more complicated as there is no convenient alternative. However, there is a way around the problem. The model (1) asserts that the hidden states x_k and the measurement noise covariance matrix $R^{(i)}$ are mutually independent. This allows to construct the joint prior distribution (9) for x_k and $R^{(i)}$ as a product of two independent priors,

$$\pi_i(\theta_k^{(i)}|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)}) = \pi_i(x_k,R^{(i)}|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)}) \tag{15}$$

$$= \pi_i(R^{(i)}|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)})\,\pi_i(x_k|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)}). \tag{16}$$

Still, no such conjugate prior for the joint inference of both x_k and $R^{(i)}$ exists. But there is a conjugate prior for x_k given known $R^{(i)}$, namely the normal prior distribution used in the diffusion Kalman filter [5], and a conjugate prior for $R^{(i)}$ given known x_k. Both these facts allow to sequentially estimate $\theta_k^{(i)}=[\![x_k,R^{(i)}]\!]$ by means of the variational message passing (VMP) [30].

Using the VMP approach, also known as the variational mean-field Bayesian approximation [11], we seek the best available approximation of the posterior distribution $\pi_i(\theta_k^{(i)}|\cdot)$ in (12) by another *tractable* distribution $\tilde{\pi}_i(\theta_k^{(i)}|\cdot)$. The result should minimize the mutual Kullback-Leibler divergence

$$\mathcal{D}\Big[\tilde{\pi}_i(\theta_k^{(i)}|\cdot)\Big|\Big|\pi_i(\theta_k^{(i)}|\cdot)\Big] = \int \tilde{\pi}_i(\theta_k^{(i)}|\cdot)\log\frac{\tilde{\pi}_i(\theta_k^{(i)}|\cdot)}{\pi_i(\theta_k^{(i)}|\cdot)}\mathrm{d}\theta_k^{(i)} = \sum_{j\in\mathcal{I}_R^{(i)}} \log f_i(y_k^{(j)}|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)}) + \mathcal{L}(\theta_k^{(i)}). \tag{17}$$

The sum involves the distributions of $y_k^{(j)}$ with $\theta_k^{(i)}$ integrated out, hence it is fixed in the divergence. The minimization thus involves the term $\mathcal{L}(\theta_k^{(i)})$ called the evidence lower bound (ELBO) or the negative variational free energy [11]. The factorization $\tilde{\pi}_i(\theta_k^{(i)}|\cdot)=\tilde{\pi}_i(x_k|\cdot)\,\tilde{\pi}_i(R^{(i)}|\cdot)$ and simple rearrangements show that

$$\begin{aligned}\mathcal{L}(\theta_k^{(i)}) &= \int \tilde{\pi}_i(\theta_k^{(i)}|\cdot)\log\frac{\tilde{\pi}_i(\theta_k^{(i)}|\cdot)}{p_{i,k}\Big(\theta_k^{(i)},\{y_k^{(j)}\}_{j\in\mathcal{I}_R^{(i)}}\Big)}\mathrm{d}\theta_k^{(i)}\\ &= \int\tilde{\pi}_i(x_k|\cdot)\log\frac{\tilde{\pi}_i(x_k|\cdot)\,\mathrm{d}x_k}{\exp\Big\{\mathbb{E}_{R^{(i)}}\Big[\log p_{i,k}\Big(\theta_k^{(i)},\{y_k^{(j)}\}_{j\in\mathcal{I}_R^{(i)}}\Big)\Big]\Big\}}+c_1\\ &= \int\tilde{\pi}_i(R^{(i)}|\cdot)\log\frac{\tilde{\pi}_i(R^{(i)}|\cdot)\,\mathrm{d}R^{(i)}}{\exp\Big\{\mathbb{E}_{x_k}\Big[\log p_{i,k}\Big(\theta_k^{(i)},\{y_k^{(j)}\}_{j\in\mathcal{I}_R^{(i)}}\Big)\Big]\Big\}}+c_2\end{aligned} \tag{18}$$

where, for easier reading,

$$p_{i,k}\Big(\theta_k^{(i)},\{y_k^{(j)}\}_{j\in\mathcal{I}_R^{(i)}}\Big) = \pi_i(\theta_k^{(i)}|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)}) \prod_{j\in\mathcal{I}_R^{(i)}} f_i(y_k^{(j)}|\theta_k^{(j)}), \tag{19}$$

and c_1, c_2 are constants independent of x_k and $R^{(i)}$, respectively. The last two integrals in (18) are the Kullback-Leibler divergences whose minimization yields mutually related variational distributions

$$\begin{aligned}\tilde{\pi}_i(x_k|\tilde{y}_k^{(i)},\tilde{u}_k^{(i)}) &\propto \exp\Big\{\mathbb{E}_{R^{(i)}}\Big[\log p_{i,k}\Big(\theta_k^{(i)},\{y_k^{(j)}\}_{j\in\mathcal{I}_R^{(i)}}\Big)\Big]\Big\},\\ \tilde{\pi}_i(R^{(i)}|\tilde{y}_k^{(i)},\tilde{u}_k^{(i)}) &\propto \exp\Big\{\mathbb{E}_{x_k}\Big[\log p_{i,k}\Big(\theta_k^{(i)},\{y_k^{(j)}\}_{j\in\mathcal{I}_R^{(i)}}\Big)\Big]\Big\},\end{aligned} \tag{20}$$

where the expectations are taken with respect to the subscripted variable. If we investigate the equations and recall that the prior distributions $\pi_i(x_k|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)})$ and $\pi_i(R^{(i)}|\tilde{y}_{k-1}^{(i)},\tilde{u}_k^{(i)})$ are conjugate to the measurement model with the other variable fixed, the measurement update step is clearly the Bayesian update (Lemma 1), possibly with some terms in the sufficient statistics replaced by their expectations. The circular dependency between x_k and $R^{(i)}$ in (20) then

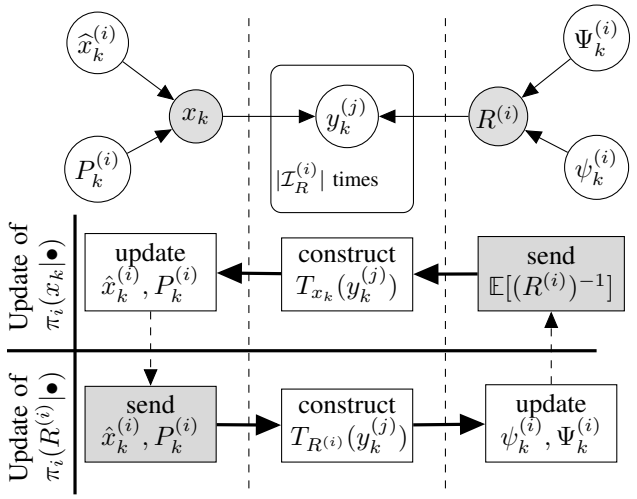

Fig. 2. Graphical model describing the measurement update step at the node i. The state variable is modeled by the normal distribution $\mathcal{N}(\hat{x}_k^{(i)}, P_k^{(i)})$, and the measurement noise covariance matrix $R^{(i)}$ by the inverse-Wishart distribution $i\mathcal{W}(\psi_k^{(i)}, \Psi_k^{(i)})$ (superscripts $+/-$ are omitted). During the measurement update step, the measurements of the R-compatible neighbors $j \in \mathcal{I}_R^{(i)}$ are assimilated. The chain of the variational steps starts from one of the shaded rectangles.

calls for an iterative algorithm similar to the expectation-maximization (EM), which is essentially a sequence of the Bayesian updates:

1) To update $\tilde{\pi}_i(x_k|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)})$, use the sufficient statistics $T_{x_k}(y_k^{(i)})$ with $R^{(i)}$-related terms replaced by their expectations following from $\tilde{\pi}_i(R^{(i)}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)})$.
2) To update $\tilde{\pi}_i(R^{(i)}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)})$, use the sufficient statistics $T_{R^{(j)}}(y_k^{(j)})$ with x_k-related terms replaced by their expectations following from $\tilde{\pi}_i(x_k|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)})$.

Algorithm of this sort is guaranteed to converge [31]. The graphical model depicted in Fig. 2 summarizes the measurement update step. The details of each substep follow.

1) Estimation of x_k: The variational Bayesian estimation of x_k proceeds with the model (2), or equivalently (6), where the parameter $R^{(i)}$ is replaced by its point estimate in the sense similar to the plug-in principle [32]. Only then it is possible to reveal the prior distribution for x_k that is conjugate to the model. We rewrite the probability density function (13) to the exponential family form corresponding to Definition 1 (for the sake of better reading, we avoid vectorizations):

$$f_i(y_k|x_k, \mathbb{E}[R^{(i)}]) \propto \exp\left\{-\frac{1}{2}\operatorname{Tr}\Big(\underbrace{\begin{bmatrix}-1\\ x_k\end{bmatrix}\begin{bmatrix}-1\\ x_k\end{bmatrix}^{\intercal}}_{\eta_{x_k}} \underbrace{\begin{bmatrix}y_k^{(i),\intercal}\\ H_k^{\intercal}\end{bmatrix}\mathbb{E}\Big[(R^{(i)})^{-1}\Big]\begin{bmatrix}y_k^{(i),\intercal}\\ H_k^{\intercal}\end{bmatrix}^{\intercal}}_{T_{x_k}(y_k^{(i)})}\Big)\right\}. \tag{21}$$

The role of the conjugate prior distribution plays the normal distribution centered at $\hat{x}_k^{(i),-}$ and with the covariance matrix $P_k^{(i),-}$,

$$\pi_i(x_k|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)}) \equiv \mathcal{N}(x_k|\hat{x}_k^{(i),-}, P_k^{(i),-}). \tag{22}$$

Its probability density function can be rewritten into the form compatible to (21) prescribed by Definition 2:

$$\begin{aligned}\pi_i(x_k|\hat{x}_k^{(i),-}, P_k^{(i),-}) &= (2\pi)^{-\frac{p}{2}}|P_k^{(i),-}|^{-\frac{1}{2}} \\ &\times \exp\left\{-\frac{1}{2}(\hat{x}_k^{(i),-} - x_k)^{\intercal}(P_k^{(i),-})^{-1}(\hat{x}_k^{(i),-} - x_k)\right\} \qquad (23)\\ &\propto \exp\left\{-\frac{1}{2}\operatorname{Tr}\Big(\underbrace{\begin{bmatrix}-1\\ x_k\end{bmatrix}\begin{bmatrix}-1\\ x_k\end{bmatrix}^{\intercal}}_{\eta_{x_k}} \underbrace{\begin{bmatrix}(\hat{x}_k^{(i),-})^{\intercal}\\ I\end{bmatrix}(P_k^{(i),-})^{-1}\begin{bmatrix}(\hat{x}_k^{(i),-})^{\intercal}\\ I\end{bmatrix}^{\intercal}}_{\xi_{x_k,k}^{(i),-}}\Big)\right\}. \qquad (24)\end{aligned}$$

The hyperparameter $\nu_\theta^{(j)}$, Eq. (3), is not necessary for the estimation of x_k.

The measurement update of the x_k-estimate by the diffusion adaptation (12) is the Bayesian update formulated in Lemma 1. The posterior hyperparameter $\xi_{x_k,k}^{(i),+}$ is given by

$$\xi_{x_k,k}^{(i),+} = \xi_{x_k,k}^{(i),-} + \prod_{j\in\mathcal{I}_R^{(i)}} T_{x_k}(y_k^{(j)}). \tag{25}$$

The local update (11) is a special case.

In order to derive the posterior hyperparameters $x_k^{(i),+}$ and $P_k^{(i),+}$ we rewrite the sufficient statistic and the hyperparameter into the following block-matrix form

$$T_{x_k}(y_k^{(j)}) = \left[\begin{array}{c|c} y_k^{(j),\intercal}\mathbb{E}\big[(R^{(i)})^{-1}\big]\, y_k^{(j)} & y_k^{(j),\intercal}\mathbb{E}\big[(R^{(i)})^{-1}\big]\, H_k \\ \hline H_k^{\intercal}\mathbb{E}\big[(R^{(i)})^{-1}\big]\, y_k^{(j)} & H_k^{\intercal}\mathbb{E}\big[(R^{(i)})^{-1}\big]\, H_k \end{array}\right], \tag{26}$$

$$\xi_{x_k k}^{-,(i)} = \left[\begin{array}{c|c} \hat{x}_k^{(i),-,\intercal}(P_k^{(i),-})^{-1}\hat{x}_k^{(i),-} & (P_k^{(i),-})^{-1}\hat{x}_k^{(i),-} \\ \hline (P_k^{(i),-})^{-1}\hat{x}_k^{(i),-} & (P_k^{(i),-})^{-1} \end{array}\right]. \tag{27}$$

Then from the update (25) it immediately follows that

$$P_k^{(i),+} = \left[\Big(P_k^{(i),-}\Big)^{-1} + \Big|\mathcal{I}_R^{(i)}\Big|\, H_k^{\intercal}\,\mathbb{E}\Big[(R^{(i)})^{-1}\Big]\, H_k\right]^{-1}, \tag{28}$$

$$x_k^{(i),+} = P_k^{(i),+}\left[\Big(P_k^{(i),-}\Big)^{-1}\hat{x}_k^{(i),-} + H_k^{\intercal}\mathbb{E}\Big[(R^{(i)})^{-1}\Big]\sum_{j\in\mathcal{I}_R^{(i)}} y_k^{(j)}\right], \tag{29}$$

where $|\mathcal{I}_R^{(i)}|$ denotes the cardinality of the R-compatible neighborhood of the node i. The equations are the distributed counterparts of the standard and information Kalman filter equations, see Appendix B.

2) Estimation of $R^{(i)}$: In order to reveal the functional form of $\pi_i(R^{(i)}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)})$, we now need to rewrite the measurement model (13) into the form parameterized by $R^{(i)}$ with all x_k-related terms replaced by their expected values. The form is

$$f_i(y_k|\mathbb{E}[x_k], R^{(i)}) \propto \exp\left\{-\frac{1}{2}\operatorname{Tr}\Big(\underbrace{\begin{bmatrix}(R^{(i)})^{-1}\\ \ln|R^{(i)}|\end{bmatrix}}_{\eta_{R^{(i)}}}\underbrace{\begin{bmatrix}\mathbb{E}\Big[(y_k^{(i)} - H_k x_k)(y_k^{(i)} - H_k x_k)^{\intercal}\Big]\\ 1\end{bmatrix}^{\intercal}}_{T_{R^{(i)}}(y_k^{(i)})}\Big)\right\}, \tag{30}$$

where

$$\mathbb{E}[(y_k^{(j)} - H_k x_k)(y_k^{(j)} - H_k x_k)^\intercal] = (y_k^{(j)} - H_k \hat{x}_k^{(i),+})(y_k^{(j)} - H_k \hat{x}_k^{(i),+})^\intercal + H_k P_k^{(i),+} H_k^\intercal. \quad (31)$$

The conjugate prior distribution for the inference of $R^{(i)} \in \mathbb{R}^{n\times n}$ is the inverse Wishart distribution $i\mathcal{W}(\psi_k^{(i),-}, \Psi_k^{(i),-})$ with the hyperparameters $\psi_k^{(i),-} \in \mathbb{R}^+$ and $\Psi_k^{(i),-} \in \mathbb{R}^{n\times n}$, and the probability density function

$$\pi_i(R^{(i)}|\psi_k^{(i),-}, \Psi_k^{(i),-}) = \frac{|\Psi_k^{(i),-}|^{\frac{\psi_k^{(i),-}}{2}}}{2^{\frac{n\psi_k^{(i),-}}{2}}\Gamma_n\left(\frac{\psi_k^{(i),-}}{2}\right)} \times |R^{(i)}|^{-\frac{\psi_k^{(i),-}+n+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{Tr}\left(\Psi_k^{(i),-}(R^{(i),-})^{-1}\right)\right\} \propto \left\{-\frac{1}{2}\operatorname{Tr}\left(\underbrace{\begin{bmatrix} R^{(i)-1} \\ \ln|R^{(i)}| \end{bmatrix}^\intercal}_{\eta_{R^{(i)}}} \underbrace{\begin{bmatrix} \Psi_k^{(i),-} \\ \psi_k^{(i),-}+n+1 \end{bmatrix}}_{\xi_{R^{(i)},k}^{(i),-}}\right)\right\}, \quad (32)$$

where $\Gamma_n(\cdot)$ is the multivariate gamma function.

The variational measurement update with diffusion adaptation then proceeds with the sufficient statistics $T_{R^{(j)}}(y_k^{(j)})$ and the prior hyperparameters $\xi_{R^{(i)}}^{(i),-}$,

$$\xi_{R^{(i)},k}^{(i),+} = \xi_{R^{(i)},k}^{(i),-} + \sum_{j\in\mathcal{I}_R^{(i)}} T_{R^{(j)}}(y_k^{(j)}), \quad (33)$$

where

$$T_{R^{(j)}}(y_k^{(j)}) = \begin{bmatrix} \mathbb{E}\left[(y_k^{(j)} - H_k x_k)(y_k^{(j)} - H_k x_k)^\intercal\right] \\ 1 \end{bmatrix}, \quad (34)$$

$$\xi_{R^{(i)},k}^{(i),-} = \begin{bmatrix} \Psi_k^{(i),-} \\ \psi_k^{(i),-}+n+1 \end{bmatrix}. \quad (35)$$

The posterior inverse-Wishart hyperparameters $\Psi_k^{(i),+}$ and $\psi_k^{(i),+}$ thus read

$$\Psi_k^{(i),+} = \Psi_k^{(i),-} + |\mathcal{I}_R^{(i)}| H_k P_k^{(i),+} H_k^\intercal + \sum_{j\in\mathcal{I}_R^{(i)}} (y_k^{(j)} - H_k \hat{x}_k^{(i),+})(y_k^{(j)} - H_k \hat{x}_k^{(i),+})^\intercal, \quad (36)$$

$$\psi_k^{(i),+} = \psi_k^{(i),-} + |\mathcal{I}_R^{(i)}|. \quad (37)$$

Finally, the posterior expectations of $R^{(i)}$ and $(R^{(i)})^{-1}$ are

$$\mathbb{E}[R^{(i)}] = (\psi_k^{(i),+} - n - 1)^{-1}\Psi_k^{(i),+}, \quad (38)$$

$$\mathbb{E}\left[\left(R^{(i)}\right)^{-1}\right] = \psi_k^{(i),+}\left(\Psi_k^{(i),+}\right)^{-1}, \quad (39)$$

respectively. We emphasize that the estimation of x_k in Section IV-A1 requires the latter one. A frequent mistake in the literature dealing with the estimation of covariance matrices (say R) is assuming that $\mathbb{E}[R^{-1}]$ coincides with $(\mathbb{E}[R])^{-1}$, which is incorrect as the inverse is not a linear transformation. However, the hyperparameter ψ_k is virtually a counter of the number of data, hence $(\mathbb{E}[R])^{-1}$ tends relatively quickly to $\mathbb{E}[R^{-1}]$ with increasing k.

3) Summary of the measurement update step with diffusion adaptation: The variational inference-based measurement update step has an iterative character schematically depicted in Fig. 2. It consists of the following routines, described from the perspective of the node i:

(i) For each neighbor $j \in \mathcal{I}_R^{(i)}$, prepare the sufficient statistic $T_{x_k}(y_k^{(j)})$ – Equation (26) – with the point estimate $\mathbb{E}[(R^{(i)})^{-1}]$ obtained from the inverse-Wishart distribution according to (39).

(ii) Update the hyperparameter $\xi_{x_k,k}^{(i),-}$ defined by (27) by its summation with the sufficient statistics from the previous step according to (25). The result is the intermediate variational distribution $\mathcal{N}(x_k^{(i),+}, P_k^{(i),+})$ with hyperparameters given by (28) and (29), respectively.

(iii) Using the results of the previous step, prepare the sufficient statistic $T_{R^{(j)}}(y_k^{(j)})$ – Formula (34).

(iv) Update the hyperparameter $\xi_{R^{(i)}}^{(i),-}$ defined by (35) by its summation with the sufficient statistics from the previous step according to (33). This results in the intermediate variational distribution $i\mathcal{W}(\psi_k^{(i),+}, \Psi_k^{(i),+})$ with the expected values (38) and (39), respectively.

(v) Repeat the steps (i)–(iv) for the preset number of variational iterations.

If the diffusion adaptation is not used, the nodes assimilate only their local measurements.

B. Local prediction step

The local prediction step (8) performs the forward shift $x_{k-1} \to x_k$ according to the state-evolution model (1a). Since the estimate of x_{k-1} represents the marginal distribution $\pi(x_{k-1}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_{k-1}^{(i)})$, the local prediction step amounts to

$$\pi(x_k|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)}) = \int \underbrace{g(x_k|x_{k-1}, u_k)}_{\mathcal{N}(A_k x_{k-1}+B_k u_k, Q_k)} \underbrace{\pi_i(x_{k-1}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_{k-1}^{(i)})}_{\mathcal{N}(x_{k-1}^{(i),+}, P_{k-1}^{(i),+})} dx_{k-1}. \quad (40)$$

This straightforwardly yields the normal distribution with the transformed mean and covariance matrix

$$\hat{x}_k^{(i),-} = A_k \hat{x}_{k-1}^{(i),+} + B_k u_k,$$
$$P_k^{(i),-} = A_k P_{k-1}^{(i),+} A_k^\intercal + Q_k. \quad (41)$$

The observation noise covariance matrix $R^{(i)}$ is constant, however, due to the distributed nature of its estimation, it may be advantageous to slightly increase the uncertainty about the estimate in order to suppress any accidentally incorporated incompatible information. We suggest to exploit the exponential forgetting [33],

$$\pi_i(R^{(i)}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_k^{(i)}) = \Big[\underbrace{\pi_i(R^{(i)}|\tilde{y}_{k-1}^{(i)}, \tilde{u}_{k-1}^{(i)})}_{i\mathcal{W}(\psi_{k-1}^{(i),+}, \Psi_{k-1}^{(i),+})}\Big]^\lambda, \quad (42)$$

where $\lambda \in [0, 1]$ is the forgetting factor, usually not lower than 0.95. This procedure results in

$$\xi_{R^{(i)},k}^{(i),-} = \begin{bmatrix} \Psi_k^{(i),-} \\ \psi_k^{(i),-}+n+1 \end{bmatrix} = \lambda \begin{bmatrix} \Psi_{k-1}^{(i),+} \\ \psi_{k-1}^{(i),+}+n+1 \end{bmatrix}, \quad (43)$$

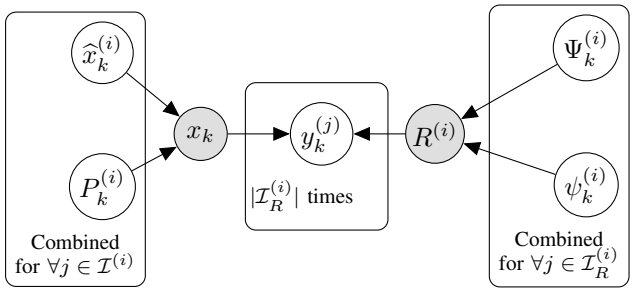

Fig. 3. Graphical model of the proposed filter from the perspective of a node i. The plate of $y_k^{(j)}$ carries the measurements of the R-compatible neighbors. They are assimilated in the adaptation step. x_k is a global variable and the local parameters $\hat{x}_k^{(i)}$ and $P_k^{(i)}$ serve for its estimation. $R^{(i)}$ is common only to neighbors $\mathcal{I}_R^{(i)} \subseteq \mathcal{I}^{(i)}$. $\Psi_k^{(i)}$ and $\psi_k^{(i)}$ serve for its estimation. While the hyperparameters for the global x_k are combined with all the neighbors $j \in \mathcal{I}^{(i)}$, the hyperparameters for $R^{(i)}$ are combined only with the R-compatible neighbors.

from which it is easy to identify the 'predicted' hyperparameters $\Psi_k^{(i),-}$ and $\psi_k^{(i),-}$.

C. Combination step

During the measurement update step each network node $i \in \mathcal{I}$ assimilates measurements $y_k^{(j)}$ provided by its R-compatible neighbors $j \in \mathcal{I}_R^{(i)}$. This gradually corrects the local statistical knowledge about the inferred variable $\theta_k^{(i)} = [\![x_k, R^{(i)}]\!]$ summarized by the posterior distribution $\tilde{\pi}_i(\theta_k^{(i)}|\tilde{y}_k^{(i)}, \tilde{u}_k^{(i)})$. The second opportunity to refine this knowledge is to take the neighbors' posterior distributions $\tilde{\pi}_j(\theta_k^{(i)}|\tilde{y}_k^{(j)}, \tilde{u}_k^{(j)})$ into account. This could effectively bring more measurements into i's knowledge. Moreover, as x_k is global and independent of $R^{(i)}$, and both variables are represented by separate probability density functions, we could incorporate the information about x_k provided by the possibly larger neighborhood $\mathcal{I}^{(i)}$. Figure 3 presents the resulting graphical model.

In order to derive a Bayes-compatible combination step, let us review and modify the Bayes' update – Lemma 1. For simplicity, assume now that the parameter θ is common, i.e., all the nodes are compatible, and posses the same prior probability density $\pi_i(\theta|\xi_\theta^{(i),-}, \nu_\theta^{(i),-})$. The geometrically averaged update in the form

$$\pi_i(\theta|\xi_\theta^{(i),+}, \nu_\theta^{(i),+}) \propto \pi_i(\theta|\xi_\theta^{(i),-}, \nu_\theta^{(i),-}) \prod_{j\in\mathcal{I}^{(i)}} [f(z^{(j)}|\theta)]^{1/|\mathcal{I}^{(i)}|} \tag{44}$$

then virtually amounts to an arithmetically averaged measurement update:

$$\xi_\theta^{(i),+} = \xi_\theta^{(i),-} + \frac{1}{|\mathcal{I}^{(i)}|} \sum_{j\in\mathcal{I}^{(i)}} T_\theta(z^{(j)}),$$
$$\nu_\theta^{(i),+} = \nu_\theta^{(i),-} + 1.$$

The result is hence the same as if we combine only the posterior distributions (without the diffusion adaptation),

$$\pi_i(\theta|\xi_\theta^{(i),+}, \nu_\theta^{(i),+}) \propto \prod_{j\in\mathcal{I}^{(i)}} \left[\pi_j(\theta|\xi_\theta^{(i),-}, \nu_\theta^{(i),-}) f(z^{(j)}|\theta)\right]^{1/|\mathcal{I}^{(i)}|}, \tag{45}$$

that is,

$$\bar{\xi}_\theta^{(i),+} = \frac{1}{|\mathcal{I}^{(i)}|} \sum_{j\in\mathcal{I}^{(i)}} \xi_\theta^{(j),+},$$
$$\bar{\nu}_\theta^{(i),+} = \frac{1}{|\mathcal{I}^{(i)}|} \sum_{j\in\mathcal{I}^{(i)}} \nu_\theta^{(j),+}, \tag{46}$$

where the bar symbol is used for elements after the combination step. Now, if the neighbors incorporated only i's measurement, i.e., $z^{(j)} = z^{(i)}$ for all $j \in \mathcal{I}^{(i)}$, this combination rule has an information averaging property, which means that it is *repeated-measurement-safe*. Obviously, this property is valid if the prior distributions are not identical and if the diffusion adaptation is used. The preservation of the functional form of distribution is an attractive property allowing to reuse the combined posterior distribution as the prior at the subsequent time step $k+1$. In [5] an alternative derivation of the combination rule (46) is presented. It assumes a Kullback-Leibler-optimal fusion.

The combination of the x_k-estimates can be performed over the whole neighborhood, as the state variable is global,

$$\bar{\xi}_{x_k,k}^{(i),+} = \frac{1}{|\mathcal{I}^{(i)}|} \sum_{j\in\mathcal{I}^{(i)}} \xi_{x_k,k}^{(j)}. \tag{47}$$

An inspection of $\xi_{x_k,k}^{(i),+}$ in (27) reveals that the mean and the covariance of the resulting normal distribution are

$$\bar{P}_k^{(i),+} = \left[\frac{1}{|\mathcal{I}^{(i)}|} \sum_{j\in\mathcal{I}^{(i)}} \left(P_k^{(j),+}\right)^{-1}\right]^{-1},$$
$$\bar{\hat{x}}_k^{(i),+} = \bar{P}_k^{(i),+} \left[\frac{1}{|\mathcal{I}^{(i)}|} \sum_{j\in\mathcal{I}^{(i)}} \left(P_k^{(j),+}\right)^{-1} \hat{x}_k^{(i),+}\right]. \tag{48}$$

We remark that this result is equivalently the covariance intersection [34]. The above-given reasoning sheds an alternative view on it.

The estimates of $R^{(i)}$ can be combined only within the R-compatible neighborhoods. The rule (46) yields

$$\bar{\xi}_{R^{(i)},k}^{(i),+} = \frac{1}{|\mathcal{I}_R^{(i)}|} \sum_{j\in\mathcal{I}_R^{(i)}} \xi_{R^{(j)},k}^{(j),+}. \tag{49}$$

From the definition of $\xi_{R^{(i)},k}^{(i),+}$ – Equation (35) – it easily follows that the combined inverse-Wishart distribution is characterized by the hyperparameters

$$\bar{\Psi}_k^{(i),+} = \frac{1}{|\mathcal{I}_R^{(i)}|} \sum_{j\in\mathcal{I}_R^{(i)}} \Psi_k^{(i),+},$$
$$\bar{\psi}_k^{(i),+} = \frac{1}{|\mathcal{I}_R^{(i)}|} \sum_{j\in\mathcal{I}_R^{(i)}} \psi_k^{(i),+}. \tag{50}$$

The resulting distribution enters the local prediction step (Section IV-B) as the prior for $R^{(i)}$ at the time instant $k+1$.

The proposed combination rule coincides with the philosophy 'let data speak for themselves' [33] in the sense that the impact of each density is weighted by the amount of its associated uncertainty, c.f. formula (48), and no external

information enters the combination procedure. Naturally, the user may want to perform *weighted* averaging in place of the uniform averaging (46). In [5], a general Bayesian procedure for weighted combination along with the optimization of the combining weights is proposed that may be adopted here.

D. Identification of compatible neighbors

While the state variable x_k is global, i.e., common for all network nodes, the (normal) measurement noise $v_k^{(i)}$ is generally heterogeneous. Its statistical properties are common only to the R-*compatible* nodes, e.g., with the same measuring principles, or close (geo)spatial locations. This complicates both the adaptation step, where the nodes may incorporate only the measurements of the R-compatible neighbors with the same observation model (1b), and the combination step, which may enter only the posterior distributions of $R^{(i)}$ of the R-compatible neighbors. A reliable identification of these neighbors is hence a task of paramount importance.

In order to identify the R-compatible neighbors, we propose to examine the similarity of the point estimates of $R^{(i)}$. This strategy is very robust and allows to fine-tune the bounds of compatibility. There are several measures of similarity of two covariance matrices, e.g., based on the Mahalanobis distance [35], comparison of eigenstructures [36]–[39], or the family of the logarithm-determinant divergences excellently reviewed and extended in [40]. We suggest to stick with the Jensen-Bregman divergence [40]–[42] for its low computational demands and attractive theoretical properties. For two covariance matrices R_1, R_2 it is defined by

$$d(R_1, R_2) = \log\det\left(\frac{R_1+R_2}{2}\right) - \frac{1}{2}\log\det(R_1 R_2). \quad (51)$$

The Jensen-Bregman divergence has very attractive properties, the three most useful for our task are [42]:

1) Nonnegativity: $d(R_1, R_2) \geq 0$ with equality if and only if $R_1 = R_2$.
2) Symmetry: $d(R_1, R_2) = d(R_2, R_1)$.
3) Affine invariance – for two compatible matrices A, B, it holds $d(AR_1B, AR_2B) = d(R_1, R_2)$.

The following example demonstrates how to find the bound for the determination of the R-compatible neighbors based on the (dis)similarity of the diagonal covariance matrices. In the case of non-diagonal matrices, the reasoning needs to involve certain heuristic.

Example: Assume a fixed node $i \in \mathcal{I}$ with a diagonal covariance matrix $R^{(i)} \in \mathbb{R}^{n\times n}$. We aim at the construction of the R-compatible neighborhood consisting of those neighbors $j \in \mathcal{I}^{(i)}$, whose estimates of the covariance matrices $R^{(j)}$ are closer to $R^{(i)}$ than $a^2R^{(i)}$, where a is some positive real number. From (51) it follows that the bound of the Jensen-Bregman divergence is given by

$$\begin{aligned} d(R^{(i)}, a^2R^{(i)}) &= \log\left|\frac{R^{(i)}+a^2R^{(i)}}{2}\right| - \frac{1}{2}\log|R^{(i)}\cdot a^2R^{(i)}| \\ &= \log\left|\frac{(a^2+1)R^{(i)}}{2}\right| - \frac{1}{2}\log|a^2(R^{(i)})^2| \\ &= \log\left(\frac{a^2+1}{2a}\right)^n = \delta. \end{aligned} \quad (52)$$

The set of R-compatible neighbors is then

$$\mathcal{I}_R^{(i)} = \{j \in \mathcal{I}^{(i)} : d(R^{(i)}, R^{(j)}) \leq \delta\}. \quad (53)$$

Naturally, we employ the estimates of the corresponding covariance matrices.

Algorithm 1 summarizes the resulting diffusion filter.

Algorithm 1 STATE FILTERING WITH INFORMATION DIFFUSION UNDER HETEROGENEOUS NOISE

At each node $i \in \mathcal{I}$ set the prior densities

$$\pi_i(x_k, R^{(i)}|\tilde{y}_0^{(i)}, \tilde{u}_0^{(i)}) = \pi_i(x_k|\tilde{y}_0^{(i)}, \tilde{u}_0^{(i)})\pi_i(R^{(i)}|\tilde{y}_0^{(i)}, \tilde{u}_0^{(i)})$$

in the form of

- the normal distribution
$$\pi_i(x_k|\tilde{y}_0^{(i)}, \tilde{u}_0^{(i)}) = \mathcal{N}(\hat{x}_0^{(i),+}, P_0^{(i),+}),$$
- the inverse-Wishart distribution
$$\pi_i(R^{(i)}\tilde{y}_0^{(i)}, \tilde{u}_0^{(i)}) = i\mathcal{W}(\psi_0^{(i),+}, \Psi_0^{(i),+}).$$

Initialize the sets of R-compatible neighbors $\mathcal{I}_R^{(i)} \equiv \{i\}$. Set the forgetting factor $\lambda \in [0, 1]$ and the number of variational Bayes iterations V. Set the boundary distance $\delta > 0$ for the identification of the R-compatible neighbors.

For $k = 1, 2, \ldots$ and each node i do:
Local prediction step:
1) Predict $\hat{x}_k^{(i),-}$ and $P_k^{(i),-}$, Equation (41).
2) Predict $\xi_{R^{(i)},k}^{(i),-}$, Equation (43)

Measurement update step with diffusion adaptation:
1) Acquire measurements $y_k^{(j)}$ of neighbors $j \in \mathcal{I}_R^{(i)}$.
2) For $v = 1, \ldots, V$ do:
   i) Prepare the suff. statistics $T_{x_k}(y_k^{(j)})$, Eq. (26),
   ii) Update $\xi_{x_k,k}^{(i)}$, Eq. (25),
   iii) Prepare the suff. statistic $T_{R^{(i)}}(y_k^{(j)})$, Eq. (34),
   iv) Update $\xi_{R^{(i)},k}^{(i)}$, Eq. (33).

Combination step:
1) Get the posterior probability density functions of neighbors $j \in \mathcal{I}^{(i)}$.
2) Combine the posterior densities for x_k, Eq. (47) or (48).
3) Calculate the point estimates of $R^{(j)}$, Eq. (38), and determine $\mathcal{I}_R^{(i)}$, Eq. (53).
4) Combine the posterior densities for $R^{(i)}$, Eq. (49) or (50).

V. DISCUSSION OF THE FILTER PROPERTIES

It is well known that under mild conditions the Bayesian posterior estimates are consistent, i.e., they converge to the true parameter with the increasing number of measurements [43]. Simultaneously, the posterior consistency guarantees that the incoming observations have to gradually dominate the role of the prior distribution in inference. Albeit this topic may be of a considerable interest, it is far behind the scope of the paper and the reader is referred, e.g., to [44]–[46] and the

references therein. We can stick with the widely recognized notion of the Bayesian information processing optimality.

While the idea of combining the posterior estimates of $R^{(i)}$ within the R-compatible neighborhood is clear, a question may arise whether the combination of *all* neighbors' posterior distributions of x_k does not adversely affect its estimation performance. The answer lies in the asymptotic properties of the Bayesian inference: with increasing number of observations, the posterior estimates converge to the true parameter, in our case to the x_k global for the whole network. Furthermore, the posterior estimate of x_k at each node i is inherently connected with the covariance matrix $P_k^{(i),+}$ that quantifies the amount of uncertainty connected with it. The source of this uncertainty is the combination of the initial uncertainty at $k = 0$ that vanishes with time, and the uncertainty due to the measurement noise, see Equation (28). The combination rule (48) performs penalization of the estimates with respect to this uncertainty, suppressing the influence of the less credible estimates.

The potential difficulty connected with the proposed method lies in the variational Bayesian procedure present in the measurement update step. The statistical properties of the variational procedures are not well understood. Their thorough analyses are usually technical and mostly single-problem oriented. A generalization in this respect provides the local variational approximation: the variational Bayes for the latent models can be interpreted as its application [47].

Let us review the theoretical drawbacks of the variational Bayesian methods discussed in [48] and see, how the proposed algorithm counteracts them:

1) *The variational methods do not provide guarantees of producing (asymptotically) exact posterior distributions, they only seek for distributions closest to the target with respect to the optimization criterion (the Kullback-Leibler divergence). In particular, the sought distribution is located at the mode (or one of them).* This is not a considerable problem in our task as all the involved distributions are unimodal.
2) *The variational inference releases statistical dependence among inferred quantities in order to be analytically tractable, hence it cannot capture their correlations.* Fortunately, in our case the elements of $\theta_k^{(i)}$ are statistically independent by nature, see Section IV-A.
3) *The variational inference underestimates the posterior (co)variances.* In the proposed algorithm, the local prediction step counteracts this issue by inflating the covariance of $x_k^{(i),-}$ by means of the Kalman prediction step, as well as the variance of the $R^{(i)}$ estimate by forgetting.

The combination procedure devised in Section IV-C is shown to be Bayes-compatible and only sensitive to an appropriate determination of the R-compatible neighborhood. A conservative setting of the bound δ (Section IV-D) effectively prevents the incompatible nodes from joining the neighborhood. Moreover, even if the value of δ becomes prohibitive, the nodes still process own information, see the simulation examples in Section VI.

TABLE I
SUMMARY OF DOWNLINK COMMUNICATION FROM THE VIEWPOINT OF NODE i: THE NUMBER OF NEIGHBORS INVOLVED IN COMMUNICATION, SHARED VARIABLES AND THEIR SIZES, AND THE TOTAL DOWNLINK COMMUNICATION COSTS.

	Adaptation	**Combination** $\pi_j(x_k\|\cdot)$		**Combination** $\pi_j(R^{(j)}\|\cdot)$	
no. of neighbors	$\|\mathcal{I}_R^{(i)}\|-1$	$\|\mathcal{I}^{(i)}\|-1$		$\|\mathcal{I}^{(i)}\|-1$	
shared variables	$y_k^{(j)}$	$\hat{x}_k^{(j),+}$	$P_k^{(j),+}$	$\Psi_k^{(j),+}$	$\psi_k^{(j),+}$
var. size	n	p	$p \times p$	$n \times n$	1
comm. cost	$(\|\mathcal{I}_R^{(i)}\|\text{-}1)n$	$(\|\mathcal{I}^{(i)}\|\text{-}1)p$	$(\|\mathcal{I}^{(i)}\|\text{-}1)$ $\times \frac{p(p+1)}{2}$	$(\|\mathcal{I}^{(i)}\|\text{-}1)$ $\times \frac{n(n+1)}{2}$	$(\|\mathcal{I}^{(i)}\|\text{-}1)$

A pertinent question is whether the filter requires the observation matrices H_k to be equal for all the network nodes. The answer is negative as long as these $H_k^{(i)}$ would refer to the same (global) state variable x_k. The local observation matrices $H_k^{(i)}$ would simply replace the global H_k in the sufficient statistics (26) and (34). This would have only a minor impact on the subsequent equations. Namely, the summation symbol would replace the multiplier $|\mathcal{I}_R^{(i)}|$ in (28) and (36), and the summation symbol in (29) would move before $H_k^{(i)}$. This immediately opens the way towards nonlinear filters with *additive* normal noise, where the respective matrices arise from the Taylor-type linearizations. Due to the limited extent of the paper we leave this topic for future research.

The communication costs – the number of real numbers that need to be obtained from the neighbors of a node i at each time step k – is summarized in Table I. There is a huge potential for reduction based on the particular application, e.g., by scheduling the combination steps.

VI. EXAMPLES

The performance of the proposed method was assessed in the following two examples. The first example assumes common states x_k and common noise covariance matrix $R^{(i)} = R$ for all $i \in \mathcal{I}$. The aim is to prove that the method provides a better estimation quality than the noncooperative scenario where the nodes do not collaborate at all, and that the quality is close to the case where a fusion center processes *all* available information. The second example assumes common states x_k, but two different measurement covariance matrices dividing $\mathcal{I}$ into two disjoint R-compatible sets $\mathcal{I}_{R_1}$ and $\mathcal{I}_{R_2}$. The nodes are randomly assigned to these sets, however, they have *no* knowledge of this partitioning nor compatibility. This simulates situations where two different instruments are used to measure the same phenomenon, and the nodes are not aware of any mutual compatibility. The goal is to show that collaboration with identified compatible neighbors leads to a significant improvement of the estimation performance.

The results of each example are averaged over 100 independent runs, i.e., over 100 completely different simulated data.

In both examples the data represent 1000 samples of a simulated 2D trajectory. The state-space model has the form

$$x_k = Ax_{k-1} + w_k, \tag{54}$$

$$y_k^{(i)} = Hx_k + v_k^{(i)}, \tag{55}$$

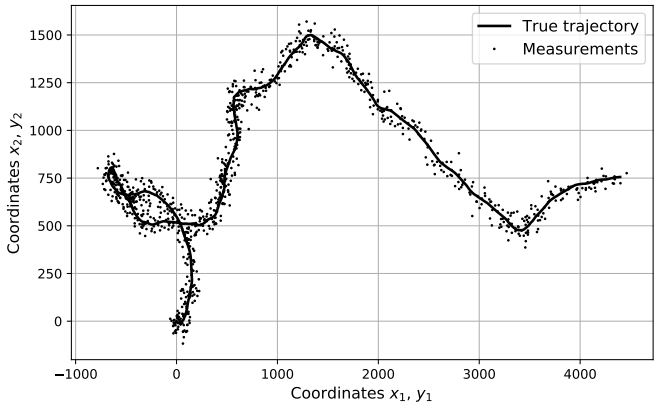

Fig. 4. Example of a true trajectory and noisy measurements of one randomly chosen network node.

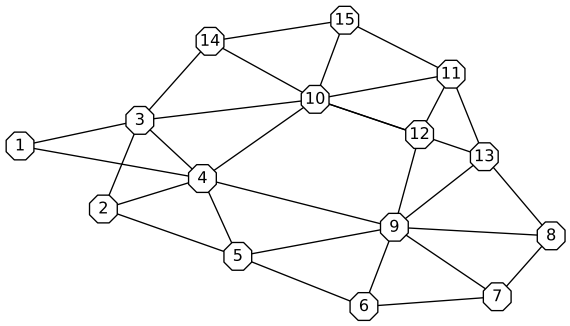

Fig. 5. Network topology.

where $x_k \in \mathbb{R}^4$ is the unknown state vector of location coordinates $x_{1,k}$ and $x_{2,k}$, and associated velocities $x_{3,k}$ and $x_{4,k}$, $x_0 = [0, 0, 0, 0]^\intercal$. The measurement vector of observed coordinates is $y_k \in \mathbb{R}^2$. The matrices

$$A = \begin{bmatrix} 1 & 0 & \Delta_k & 0 \\ 0 & 1 & 0 & \Delta_k \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \tag{56}$$

where the time difference $\Delta_k = 1$. The independent identically distributed noise variable $w_k \sim \mathcal{N}(0, Q)$ with the covariance matrix

$$Q = \frac{1}{2} \begin{bmatrix} \frac{\Delta_k^3}{3} & 0 & \frac{\Delta_k^2}{2} & 0 \\ 0 & \frac{\Delta_k^3}{3} & 0 & \frac{\Delta_k^2}{2} \\ \frac{\Delta_k^2}{2} & 0 & \Delta_k & 0 \\ 0 & \frac{\Delta_k^2}{2} & 0 & \Delta_k \end{bmatrix}. \tag{57}$$

The measurement noise $v_k^{(i)} \sim \mathcal{N}(0, R^{(i)})$ is independent and identically distributed. Its covariance matrices are defined in the examples. Fig. 4 depicts one of generated trajectories. The true trajectory is common, while the measurements are node-specific.

The network consists of $|\mathcal{I}| = 15$ nodes. Its topology is depicted in Fig. 5. All network nodes have the same initial prior distributions. Namely, the prior distribution for $R^{(i)}$ is the inverse-Wishart distribution $i\mathcal{W}(4, 100 \cdot I_{[2\times 2]})$, the normal prior distribution for x_k is zero-centered and with the covariance matrix $100 \cdot I_{[4\times 4]}$ where I is the identity matrix. The forgetting factor for the estimation of $R^{(i)}$ is 0.99. At each time k, $V = 5$ iterations of the variational algorithm are run. The neighbors are declared to have the same measurement noise covariance if the Jensen-Bregman divergence of the estimates is less than 0.005. From (52) it follows that this value is very conservative. Under collaboration, the nodes may share the posterior distributions of the state x_k and, after the detection of compatible neighbors, the posterior distributions of $R^{(i)}$. In the adapt-then-combine (ATC) algorithm, the compatible nodes share their raw measurements too.

We emphasize that the model is the constant velocity model, where the velocity is driven (and hence modeled) solely by the additive noise term. Therefore, we focus on the estimates of the location coordinates $x_{1,k}$ and $x_{2,k}$ only.

A. Example 1: Common x_k and $R^{(i)} = R$ for all nodes

The first example demonstrates the ability of the proposed method to gradually detect and collaborate with compatible neighboring nodes. In this case, the whole network shares the same model with the measurement noise covariance matrix $R = 40^2 \cdot I_{2\times 2}$. Four scenarios are compared: (i) NOCOOP, where the nodes do not cooperate at all and evaluate their estimates based on own measurements, (ii) C – the combination-only scenario, i.e., the reduced ATC scenario where the nodes do not share the measurements but only the posterior estimates, (iii) ATC – the adapt-then-combine strategy, where the compatible nodes share the posterior estimates and measurements, and finally (iv) FC, the fusion center scenario where a single node processes all available information.

Figures 6 and 7 depict the RMSE evolution of the estimates of the states $x_{1,k}$ and $x_{2,k}$, and the measurement noise covariance matrix R, respectively. The values are averaged over the network. The proposed algorithm provides the estimation quality of x_k between the non-cooperative scenario and the FC scenario. The estimation of R yields – particularly in the ATC scenario – the estimation quality very close to the fusion center (whose convergence is naturally much faster). In both cases the two-stage ATC algorithm performs slightly better than the combination-only (C) algorithm, of course at the price of higher computational and communication burden. To summarize, the nodes progressively detected the neighbors with a compatible information and started to collaborate with them, which resulted in an improvement of the estimation quality.

Finally, we compare the state estimation performance with the generic diffusion Kalman filter (denoted by diffKF) *requiring known measurement covariance matrices* [5], [21]. It exploits the adapt-than-combine strategy. Instead of plotting its performance in Figure 6, the results are compared only to the proposed ATC filter in Figure 8, because after approximately 150 steps the filters attain very similar average RMSE (only 500 time steps are depicted to show the difference in detail). We attribute the initial dissimilarity to the period where the proposed filter had insufficient knowledge of the (estimated) measurement noise covariance matrix.

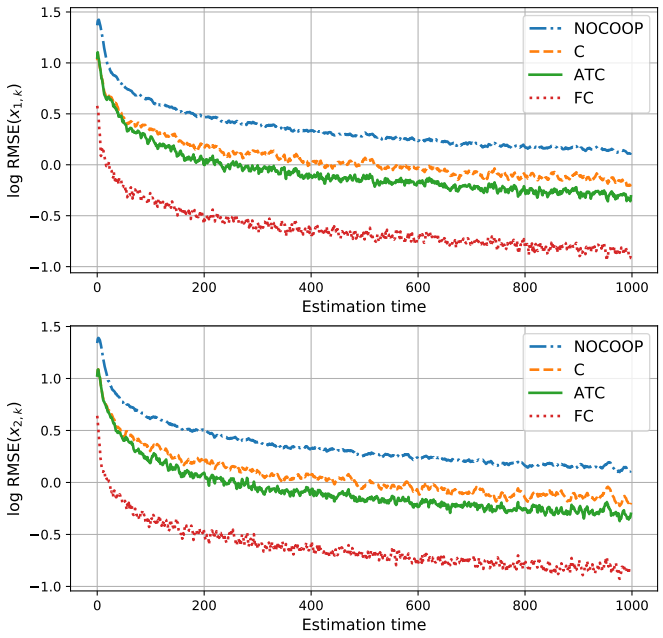

Fig. 6. Decimal logarithm of average RMSE of state estimates (Example 1).

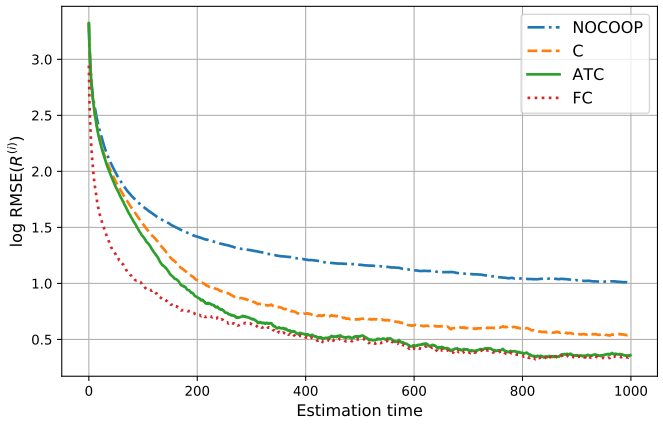

Fig. 7. Decimal logarithm of average RMSE of measurements noise covariance estimates (Example 1).

B. Example 2: Common x_k, heterogeneous $R^{(i)}$s

The second example demonstrates the case of heterogeneous measurement noise covariances. The network of 15 nodes (Fig. 5) observes the trajectory corrupted by a zero-centered normal noise with the covariance matrix either $R_1 = 30^2 I_{2\times 2}$ or $R_2 = 40^2 I_{2\times 2}$, respectively. In the 100 experiment runs, the covariance matrices are randomly and with equal probabilities assigned to individual nodes during the data simulation stage. In order to initiate collaboration, the nodes have to identify their R-compatible neighbors first.

Three scenarios are studied: (i) NOCOOP, where the nodes do not collaborate at all and evaluate their estimates solely from locally measured data, (ii) C – the combination-only scenario where the nodes no not share their measurements but only the posterior distributions, and (iii) ATC, where both the adaptation and combination steps are used. The fusion center

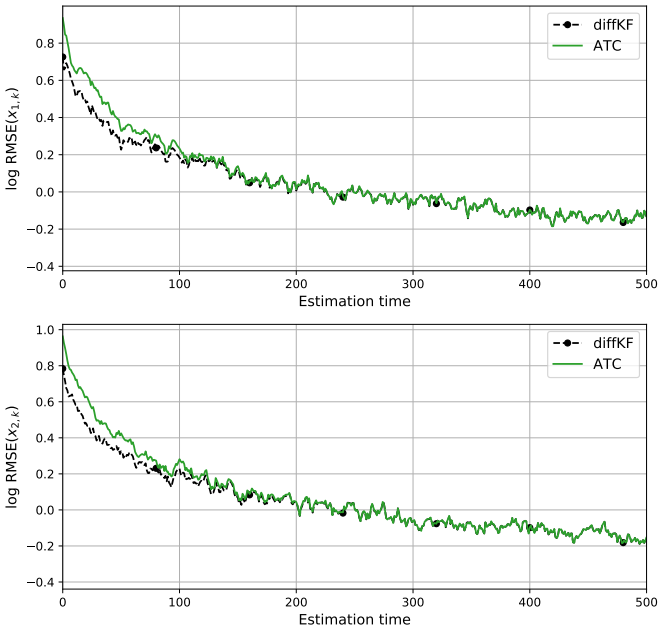

Fig. 8. Decimal logarithm of average RMSE of state estimates (Example 1). Comparison with the diffusion Kalman filter [5], [21], denoted by 'diffKF'. Only the first 500 steps are depicted.

scenario is not studied, because the underlying estimation algorithm is not directly suitable for the mixture cases.

The RMSE evolutions averaged over 100 independent simulation runs are depicted in Figures 9 and 10 for the estimates of both $x_{1,k}$ and $x_{2,k}$, and $R^{(i)}$, respectively. They are consistent with the previous example – as the nodes start to collaborate, their estimation quality improves. The ATC algorithm where both the measurements and posterior estimates are shared performs better than the combination-only (C) scenario, of course at the price of slightly higher communication overhead.

VII. Conclusion

In this paper we proposed a new algorithm for sequential (online) distributed estimation of the state-space models with unknown and heterogeneous measurement noise covariance matrices. The algorithm assumes that the states are common to all network nodes and their estimates can be directly shared, while the covariances may differ. The nodes are not aware of the global situation. After the detection of neighbors with sufficiently similar covariance estimates, the relevant information – covariance estimates and possibly the raw measurements – are incorporated by the nodes into their local knowledge about the inferred variables. The algorithm is suitable for the linear state-space models, but the principles equivalently apply to the nonlinear models with a Taylor-type linearization. The future work should focus on filtration under unknown hidden process noise and under time-varying covariance matrices.

Appendix A
Proof of Lemma 1

Assume that the model of z is an exponential family distribution (Def. 1) and the conjugate prior (Def. 2) is used

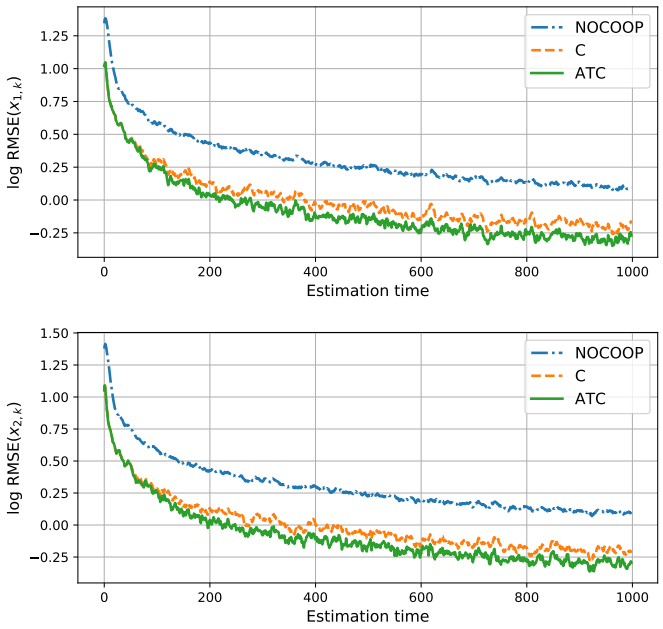

Fig. 9. Decimal logarithm of average RMSE of state estimates (Example 2).

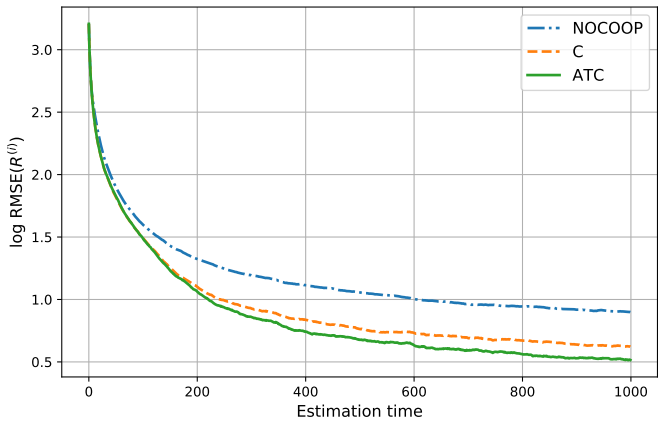

Fig. 10. Decimal logarithm of average RMSE of measurements noise covariance estimates (Example 2).

for the estimation of its parameters. By Lemma 1,

$$\begin{aligned}\pi(\theta|\xi_\theta^+,\nu_\theta^+) &\propto \pi(\theta|\xi_\theta^-,\nu_\theta^-)\prod_{j=1}^m f(z^{(j)}|\theta)\\ &\propto \exp\{\eta_\theta^\intercal\xi_\theta^- - \nu_\theta^- A(\eta_\theta)\}\prod_{j=1}^m \exp\{\eta_\theta^\intercal T_\theta(z^{(j)}) - A(\eta_\theta)\}\\ &\propto \exp\left\{\eta_\theta^\intercal\left(\xi_\theta^- + \sum_{j=1}^m T_\theta(z^{(j)})\right) - (\nu_\theta^- + m)A(\eta_\theta)\right\}\\ &\propto \exp\{\eta_\theta^\intercal\xi_\theta^+ - \nu_\theta^+ A(\eta_\theta)\},\end{aligned}$$

where

$$\begin{aligned}\xi_\theta^+ &= \xi_\theta^- + \sum_{j=1}^m T_\theta(z^{(j)}),\\ \nu_\theta^+ &= \nu_\theta^- + m.\end{aligned}$$

The functional form of the posterior is thus the same as the prior distribution, and the subsequent normalization provides the proper posterior probability density function.

Appendix B
Alternative formulation of $x_k^{(i),-}$ update step

This appendix formulates the Kalman update of $x_k^{(i),-}$ – Formulas (28) and (29) – in alternative forms involving the Kalman gain.

First, let us focus on the covariance update (28) and rewrite it using the celebrated matrix inversion lemma:

$$\begin{aligned}P_k^{(i),+} &= \left[\left(P_k^{(i),-}\right)^{-1} + \left|\mathcal{I}_R^{(i)}\right| H_k^\intercal\,\mathbb{E}\left[(R^{(i)})^{-1}\right] H_k\right]^{-1} \qquad (58)\\ &= P_k^{(i),-} - |\mathcal{I}_R^{(i)}|P_k^{(i),-}H_k^\intercal\\ &\quad\times\left[|\mathcal{I}_R^{(i)}|H_kP_k^{(i),-}H_k^\intercal + \mathbb{E}\left[(R^{(i)})^{-1}\right]^{-1}\right]^{-1} H_kP_k^{(i),-}\\ &= \left[I - |\mathcal{I}_R^{(i)}|K_k^{(i)}H_k\right]P_k^{(i),-}, \qquad (59)\end{aligned}$$

where $|\mathcal{I}_R^{(i)}|$ denotes the cardinality of the R-compatible neighborhood of the node i, I is the identity matrix of compatible size, and

$$K_k^{(i)} = P_k^{(i),-}H_k^\intercal\left[|\mathcal{I}_R^{(i)}|H_kP_k^{(i),-}H_k^\intercal + \mathbb{E}\left[(R^{(i)})^{-1}\right]^{-1}\right]^{-1} \qquad (60)$$

is the Kalman gain under update by multiple measurements. In order to obtain its alternative formulation, we premultiply it on the right-hand side by $P_k^{(i),+}\left(P_k^{(i),+}\right)^{-1}$ which is equal to the identity matrix, substitute the inverse of (58) for $\left(P_k^{(i),+}\right)^{-1}$, and rearrange terms:

$$\begin{aligned}K_k^{(i)} &= P_k^{(i),+}\left(P_k^{(i),+}\right)^{-1}K_k^{(i)}\\ &= P_k^{(i),+}\left[\left(P_k^{(i),-}\right)^{-1} + \left|\mathcal{I}_R^{(i)}\right|H_k^\intercal\,\mathbb{E}\left[(R^{(i)})^{-1}\right]H_k\right]\\ &\quad\times P_k^{(i),-}H_k^\intercal\left[|\mathcal{I}_R^{(i)}|H_kP_k^{(i),-}H_k^\intercal + \mathbb{E}\left[(R^{(i)})^{-1}\right]^{-1}\right]^{-1}\\ &= P_k^{(i),+}H_k^\intercal\left[I + \left|\mathcal{I}_R^{(i)}\right|\mathbb{E}\left[(R^{(i)})^{-1}\right]H_kP_k^{(i),-}H_k^\intercal\right]\\ &\quad\times\left[|\mathcal{I}_R^{(i)}|H_kP_k^{(i),-}H_k^\intercal + \mathbb{E}\left[(R^{(i)})^{-1}\right]^{-1}\right]^{-1}. \qquad (61)\end{aligned}$$

Now, we bring the expectation out to the left side, and the formula simplifies as follows:

$$\begin{aligned}K_k^{(i)} &= P_k^{(i),+}H_k^\intercal\,\mathbb{E}\left[(R^{(i)})^{-1}\right]\\ &\quad\times\left[\mathbb{E}\left[(R^{(i)})^{-1}\right]^{-1} + \left|\mathcal{I}_R^{(i)}\right|H_kP_k^{(i),-}H_k^\intercal\right]\\ &\quad\times\left[\mathbb{E}\left[(R^{(i)})^{-1}\right]^{-1} + |\mathcal{I}_R^{(i)}|H_kP_k^{(i),-}H_k^\intercal\right]^{-1}\\ &= P_k^{(i),+}H_k^\intercal\,\mathbb{E}\left[(R^{(i)})^{-1}\right]. \qquad (62)\end{aligned}$$

The result is the well-known formula for the Kalman gain.

The update of the estimate $x_k^{(i),-}$ prescribed by Formula (29) can be rewritten as follows:

$$\hat{x}_k^{(i),+} = P_k^{(i),+}\left[\left(P_k^{(i),-}\right)^{-1}\hat{x}_k^{(i),-} + H_k^{\intercal}\mathbb{E}\left[(R^{(i)})^{-1}\right]\sum_{j\in\mathcal{I}_R^{(i)}} y_k^{(j)}\right]$$
$$= P_k^{(i),+}\left(P_k^{(i),-}\right)^{-1}\hat{x}_k^{(i),-} + P_k^{(i),+}H_k^{\intercal}\mathbb{E}\left[(R^{(i)})^{-1}\right]\sum_{j\in\mathcal{I}_R^{(i)}} y_k^{(j)}.$$

Now, we substitute (59) for $P_k^{(i),+}$ in the first summand, then (62) for $P_k^{(i),+}H_k^{\intercal}\mathbb{E}\left[(R^{(i)})^{-1}\right]$ in the second, and rearrange. This yields the Kalman update formula

$$x_k^{(i),+} = \hat{x}_k^{(i),-} - K_k^{(i)}\sum_{j\in\mathcal{I}_R^{(i)}}\left(y_k^{(j)} - H_k\hat{x}_k^{(i),-}\right). \qquad (63)$$

The formulas (58) – (63) are the counterparts of the standard Kalman filter formulas summarized, e.g., in [49, Chap. 5].

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, Mar. 2002.

[2] A. H. Sayed, "Diffusion Adaptation over Networks," in *Academic Press Library in Signal Processing*, R. Chellapa and S. Theodoridis, Eds. Academic Press, Elsevier, May 2014, vol. 3, pp. 323–454.

[3] ——, "Adaptation, Learning, and Optimization over Networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 310–801, 2014.

[4] X. Zhao and A. H. Sayed, "Distributed clustering and learning over networks," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3285–3300, Jul. 2015.

[5] K. Dedecius and P. M. Djurić, "Sequential estimation and diffusion of information over networks: A Bayesian approach with exponential family of distributions," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1795–1809, Apr. 2017.

[6] S. Kar and J. M. F. Moura, "Gossip and distributed Kalman filtering: Weak consensus under weak detectability," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1766–1784, Apr. 2011.

[7] M. G. S. Bruno and S. S. Dias, "Collaborative emitter tracking using Rao-Blackwellized random exchange diffusion particle filtering," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, pp.1-18, Feb. 2014.

[8] M. G. Bruno and S. S. Dias, "A Bayesian interpretation of distributed diffusion filtering algorithms [Lecture Notes]," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 118–123, May 2018.

[9] S. He, H.-S. Shin, S. Xu, and A. Tsourdos, "Distributed estimation over a low-cost sensor network: A review of state-of-the-art," *Information Fusion*, vol. 54, pp. 21–43, Feb. 2020.

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

[11] T. Jaakkola and M. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, no. 1, pp. 25–37, Jan. 2000.

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

[13] S. Särkkä and J. Hartikainen, "Non-linear noise adaptive Kalman filtering via variational Bayes," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2013.

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

[15] 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. Autom. Control*, vol. 63, no. 2, pp. 594–601, Feb. 2018.

[16] G. Battistelli, L. Chisci, G. Mugnai, A. Farina, and A. Graziano, "Consensus-based linear and nonlinear filtering," *IEEE Trans. Autom. Control*, vol. 60, no. 5, pp. 1410–1415, 2015.

[17] S. P. Talebi and S. Werner, "Distributed Kalman filtering and control through embedded average consensus information fusion," *IEEE Trans. Autom. Control*, vol. 64, no. 10, pp. 4396–4403, 2019.

[18] K. Shen, Z. Jing, and P. Dong, "A consensus nonlinear filter with measurement uncertainty in distributed sensor networks," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1631–1635, 2017.

[19] Y. Yu, "Consensus-based distributed linear filter for target tracking with uncertain noise statistics," *IEEE Sens. J.*, vol. 17, no. 15, pp. 4875–4885, 2017.

[20] F. Cattivelli and A. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 2069–2084, 2010.

[21] J. Hu, L. Xie, and C. Zhang, "Diffusion Kalman filtering based on covariance intersection," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 891–902, Feb. 2012.

[22] K. Dedecius and V. Sečkárová, "Factorized estimation of partially shared parameters in diffusion networks," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5153–5163, Oct. 2017.

[23] S. S. Dias and M. G. S. Bruno, "Cooperative target tracking using decentralized particle filtering and RSS sensors," *IEEE Trans. Signal Process.*, vol. 61, no. 14, pp. 3632–3646, 2013.

[24] S. Khawatmi, A. H. Sayed, and A. M. Zoubir, "Decentralized clustering and linking by networked agents," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3526–3537, 2017.

[25] J. Plata-Chaves, A. Bertrand, M. Moonen, S. Theodoridis, and A. M. Zoubir, "Heterogeneous and multitask wireless sensor networks-algorithms, applications, and challenges," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 3, pp. 450–465, 2017.

[26] R. Nassif, S. Vlaski, C. Richard, J. Chen, and A. H. Sayed, "Multitask learning over graphs: An approach for distributed, streaming machine learning," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 14–25, May 2020.

[27] B. O. Koopman, "On distributions admitting a sufficient statistic," *Transactions of the American Mathematical Society*, vol. 39, no. 3, pp. 399–499, May 1936.

[28] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory (Harvard Business School Publications)*. Harvard University Press, Jan. 1961.

[29] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis, Second Ed.*. Chapman & Hall/CRC, Jul. 2003.

[30] J. Winn and C. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, pp. 661–694, 2005.

[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[32] R. L. Smith, "Bayesian and frequentist approaches to parametric predictive inference," *Bayesian Statistics*, vol. 6, pp. 589–612, 1999.

[33] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System Identification*, P. Eykhoff, Ed. Oxford, U.K.: Pergamon Press, 1981, pp. 239–304.

[34] S. J. Julier and J. K. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations," in *Proc. 1997 American Control Conference*, vol. 4, 1997, pp. 2369–2373.

[35] B. Flury, "Some relations between the comparison of covariance matrices and principal component analysis," *Computational Statistics & Data Analysis*, vol. 1, pp. 97–109, Mar. 1983.

[36] ——, *Common Principal Components & Related Multivariate Models*. Wiley, 1988.

[37] W. Förstner and B. Moonen, "A metric for covariance matrices," in *Geodesy-The Challenge of the 3rd Millennium*. Springer, 2003, pp. 299–309.

[38] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek, "Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels," in *Proc. 61st Vehicular Technology Conference*. 2005, pp. 136–140.

[39] C. Garcia, "A simple procedure for the comparison of covariance matrices." *BMC Evolutionary Biology*, vol. 12, no. 1, p. 222, 2012.

[40] A. Cichocki, S. Cruces, and S.-I. Amari, "Log-determinant divergences revisited: Alpha-beta and gamma log-det divergences," *Entropy*, vol. 17, no. 5, pp. 2988–3034, May 2014.

[41] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, "Jensen-Bregman LogDet divergence with application to efficient similarity search for covariance matrices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2161–2174, Sep 2013.

[42] S. Sra, "Positive definite matrices and the S-divergence," *Proceedings of the American Mathematical Society*, vol. 144, no. 7, pp. 2787–2797, Oct. 2015.

[43] J. L. Doob, "Application of the theory of martingales," *Le Calcul des Probabilites et ses Applications*, pp. 23–27, 1949.
[44] S. Walker, "New approaches to Bayesian consistency," *The Annals of Statistics*, vol. 32, no. 5, pp. 2028–2043, Oct. 2004.
[45] T. Choi and R. V. Ramamoorthi, "Remarks on consistency of posterior distributions," in *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*. Institute of Mathematical Statistics, 2008, vol. 3, pp. 170–186.
[46] B. J. Kleijn and Y. Y. Zhao, "Criteria for posterior consistency and convergence at a rate," *Electronic Journal of Statistics*, vol. 13, no. 2, pp. 4709–4742, 2019.
[47] K. Watanabe, "An alternative view of variational Bayes and asymptotic approximations of free energy," *Machine Learning*, vol. 86, no. 2, pp. 273–293, Feb. 2012.
[48] 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, Apr. 2017.
[49] D. Simon, *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Wiley, 2006.