CHAPTER

BAYESIAN APPROACH TO COLLABORATIVE INFERENCE IN NETWORKS OF AGENTS

4

Kamil Dedecius*, Petar M. Djurić†

*Department of Adaptive Systems, Institute of Information Theory and Automation, The Czech Academy of Sciences, Prague, Czech Republic** *Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY, United States†*

4.1 INTRODUCTION

With the rapid development of devices with high computational performance, the probabilistically consistent and versatile Bayesian methods have become a popular standard in many applications of signal processing [1]. Their main difference in comparison to traditional approaches consists in the representation of the unknown variables of interest. They are described by probability distributions whose location statistics (the mean, mode, or median) express the probable locations of these variables while the dispersion statistics of the distribution (e.g., the variance) quantify the associated uncertainty. An important facet of Bayesian theory is the universality of its methods. According to Bayes' theorem, one updates the initial knowledge of a considered variable, represented by a prior distribution, by following the same generic steps regardless of whether the underlying task is linear or nonlinear regression, filtering of state-space model parameters, or estimation of hierarchical models.

In the last decade, signal processing has faced a host of new challenges related to the fast evolution of spatially distributed systems with components—termed agents, sensors, nodes, or vertices and here referred to as agents—that have relatively high sensing and computational performance, and may communicate with other agents of the network. The applications of these systems range from environment monitoring, disaster relief management, source localization, and precision agriculture to medicine [2–5]. The first algorithms for processing of data acquired by agents were centralized. More specifically, there the agents locally sense the relevant data and send them to a *fusion center*, responsible for evaluation of (nearly) all necessary computations. Subsequently, the results are sent back to the agents, if necessary. In this setting, the fusion center exploits all the data present in the network and thus reaches the best possible estimation performance. The price for this is the high communication and computation demands and the lack of redundancy, making the centralized algorithms prone to failures [6].

Cooperative and Graph Signal Processing. https://doi.org/10.1016/B978-0-12-813677-5.00004-3

Table 4.1 Summary of Notation

Notation	Description
$\mathcal{K} = \{1, \ldots, K\}$	set of agents
K	total number of agents
$k \in \mathcal{K}$	agent index
$\mathfrak{N}_k$	neighborhood of agent k
$\lvert\mathfrak{N}_k\rvert$	cardinality of the set $\mathfrak{N}_k$
$t = 1, 2, \ldots$	discrete time index
$y_{t,k}$	observation of kth agent at time t
$z_{t,k}$	explanatory variable observed by agent k at time t
θ	model parameter
$\widehat{\theta}$	model parameter point estimate
$f(y_{t,k}\lvert z_{t,k}, \theta)$	probability density of observations of agent k
$\pi_k(\theta)$	probability density of θ of agent k
$\eta = \eta(\theta)$	natural parameter
$T(\cdot)$	sufficient statistic
$\xi_{t,k}, \nu_{t,k}$	hyperparameters of agent k at time t
$\zeta_{t,k}$	information of agent k at time t
$a_{kj} \in [0, 1]$	weight assigned by agent k to agent j
$\mathbb{E}[\cdot]$	expectation operator
$\mathcal{D}(\cdot\lvert\lvert\cdot)$	Kullback-Leibler divergence
$\tilde{\pi}_k(\cdot)$	combined posterior density of agent k
$\tilde{\xi}_{t,k}, \tilde{\nu}_{t,k}$	hyperparameters of $\tilde{\pi}_k(\cdot)$
Tr	trace operator
det	determinant
$\mathcal{N}(\mu, \Sigma)$	normal distribution with mean μ and covariance Σ

In order to remove this drawback, fully distributed processing settings have been proposed. First, incremental algorithms have been studied [7–13], where information is passed from agent to agent in a cyclic Hamiltonian path connecting the whole network. Although this removes the need for a fusion center and alleviates the communication and computational burden, the reliability of the system is not improved as each agent and link are single points of failure. A recovery from such a failure by constructing a new path is an NP-hard problem [4]. Then *consensus* [14–20] and *diffusion* strategies [21–26] have been introduced, where the agents share information with their neighbors within a one-hop distance. Both strategies offer significantly more robust solutions. The consensus strategies aim at a general agreement in the value of the estimated variables of interest while the diffusion strategies put emphasis on a local improvement of the estimation quality of each agent. Therefore, while the diffusion algorithms intrinsically exploit a single time scale both for sensing and collaborative data processing, the consensus algorithms usually need multiple iterations between two time instants. This chapter focuses primarily on diffusion strategies.

The existing diffusion algorithms mostly extend their nondistributed counterparts to distributed settings. A majority of them exploit the least squares criterion and its variants, e.g., the least-mean-squares methods [22,27–29], the recursive least-squares methods [21,26], and the Kalman filter [30,31]. There are also algorithms for distributed expectation maximization-based inference of mixture models [32–34] and distributed particle filtering, e.g., in [35,36].

Quite surprisingly, most of the algorithms for distributed inference are *independently* developed from the original nondistributed ones, e.g., the already-mentioned classical least-mean-squares method, recursive least-squares method, and Kalman filter. However, because there is a common underlying principle in Bayesian estimation, a unifying framework has been recently proposed [37]. Within this framework, one can develop methods that can be applied to a wide class of inference tasks with minimal modifications. For instance, the existing recursive least-squares method [21] or the Kalman filter [31] are special cases when particular models and prior distributions are used. In [34], a quasi-Bayesian algorithm for sequential estimation of mixture models was introduced. The components of the models considered there belong to the exponential family of distributions. In [38], a diffusion approximate Bayesian computation method was presented. The method extends the particle filtering principles to cases of unknown or intractable models. The foundations of these approaches are described in the present chapter.

4.2 BAYESIAN INFERENCE OVER NETWORKS

We consider a network to be represented by a connected undirected graph consisting of a set of vertices termed agents. The agents are interconnected by a set of edges, which determine the network topology. The set of agents is denoted by $\mathcal{K} = \{1, 2, \ldots, K\}$, where K is the number of agents in the network. (A summary of notation for this chapter is presented in Table 4.1.) Each agent $k \in \mathcal{K}$ may communicate only with agents in its close neighborhood $\mathfrak{N}_k$, here defined as a set of agents within one-hop distance (note that $k \in \mathfrak{N}_k$). The agents independently observe outcomes $y_{t,k}$ of a common discrete-time stochastic process $\{Y_t; t = 1, 2, \ldots\}$ with an unknown parameter θ, and a known explanatory variable $z_{t,k}$, if it exists. For instance, $z_{t,k}$ may be a regressor.

For the sake of prediction, filtering, and smoothing, the agents employ a probabilistic model in the form of a probability density $f(y_{t,k}|z_{t,k}, \theta)$, or $f(y_{t,k}|\theta)$ if $z_{t,k}$ is not assumed. The value of θ remains unknown, but its reliable estimation is of main interest in the rest of the chapter. The Bayesian approach to estimating θ proceeds by updating the prior distributions of θ at time t, $\pi_k(\theta|y_{0:t-1,k}, z_{0:t-1,k})$, where $y_{0:t-1,k} \equiv \{y_{0,k}, \ldots, y_{t-1,k}\}$ and $z_{0:t-1,k} \equiv \{z_{0,k}, \ldots, z_{t-1,k}\}$, by using the new observation $y_{t,k}$ via Bayes' theorem,

$$\pi_k(\theta|y_{0:t,k}, z_{0:t,k}) \propto f(y_{t,k}|z_{t,k}, \theta)\pi_k(\theta|y_{0:t-1,k}, z_{0:t-1,k}). \tag{4.1}$$

Note that in writing this equation we assume that the observations are independent given the explanatory variables $z_{t,k}$ and the parameter θ. In this chapter, we assume that the reader is familiar with the principles of Bayesian inference, and skip technical details of derivation of posterior distributions and their computations via Monte Carlo methods, variational approaches, and the like. There is a vast literature on this topic, e.g., [39,40]. Next, we describe a prominent case where the posterior distribution is analytically tractable and which will be used in the sequel.

Suppose that the model $f(y_{t,k}|z_{t,k},\theta)$ belongs to the exponential family of distributions, i.e., that it can be written in the form [41]

$$f(y_{t,k}|z_{t,k},\theta) = h(y_{t,k},z_{t,k})g(\theta)\exp\left\{\eta^{\mathsf{T}}(\theta)\,T(y_{t,k},z_{t,k})\right\}, \tag{4.2}$$

where $h(y_{t,k},z_{t,k})$ is a known function, $g(\theta)$ is a normalizing log-partition function, $\eta(\theta)$ is a natural parameter, and $T(y_{t,k},z_{t,k})$ is a sufficient statistic that completely summarizes all the information about θ contained in $y_{t,k}$ and $z_{t,k}$. Now, we assume that the prior distribution can be written in the *conjugate* form

$$\begin{aligned}\pi_k(\theta|y_{0:t-1,k},z_{0:t-1,k}) &= \pi_k(\theta|\xi_{t-1,k},\nu_{t-1,k})\\ &= q(\xi_{t-1,k},\nu_{t-1,k})g(\theta)^{\nu_{t-1,k}}\exp\left\{\eta^{\mathsf{T}}(\theta)\,\xi_{t-1,k}\right\},\end{aligned} \tag{4.3}$$

where $\xi_{t-1,k}$ and $\nu_{t-1,k}$ are the prior *hyperparameters*. The former is of the same dimension as $T(y_{t,k},z_{t,k})$ and the latter is a scalar whereas $q(\xi_{t-1,k},\nu_{t-1,k})$ is a known function. Then Bayes' theorem (4.1) updates the prior hyperparameters according to [37,41]

$$\xi_{t,k} = \xi_{t-1,k} + T(y_{t,k},z_{t,k}), \tag{4.4}$$

$$\nu_{t,k} = \nu_{t-1,k} + 1. \tag{4.5}$$

Naturally, it is possible to write

$$\pi_k(\theta|y_{0:t,k},z_{0:t,k}) = \pi_k(\theta|\xi_{t,k},\nu_{t,k}). \tag{4.6}$$

We point out here that in sequential processing at time t, $\pi_k(\theta|y_{0:t,k},z_{0:t,k})$ is the posterior of θ. This distribution is also the prior of θ for the processing that takes place at time instant $t+1$.

4.2.1 STRATEGIES FOR INFERENCE OVER NETWORKS

There are various types of settings of inference over networks. First and foremost, a question of paramount importance is whether the models and their parameters are the same for all the collaborating agents. If the model parameters have a physical interpretation, then their homogeneity is usually guaranteed. However, if $f_k(\theta|\cdot)$ are black box models that may have different structures, problems arise. For the sake of simplicity, we adopt the assumption that the models are the same for all agents, and that they are all interested in the same θ. The Bayesian treatment of inhomogeneous parameters is studied in [42].

The next question is what kind of information may be shared among network agents. If there are (virtually) no limitations in communication resources, the agents may share their observations $y_{t,k}$ and explanatory variables $z_{t,k}$ (the sharing may be in the form of sufficient statistics $T(y_{t,k},z_{t,k})$), and estimates $\hat{\theta}_k$ potentially accompanied by related statistical properties such as covariance matrices. If the Bayesian approach to inference is employed, then the most legitimate way is to share the posterior distributions $\pi_k(\theta|\cdot)$, or their hyperparameters (e.g., $\xi_{t,k}$ and $\nu_{t,k}$) whenever possible.

We discriminate among three possible strategies:

1. Incorporation of neighbors' (and own) measurements. This step is often called *adaptation* (A) in the literature.
2. Incorporation of *estimates* provided by neighbors. This is known as *combination* (C). Unlike in adaptation, a combination criterion is required that ensures the result to be as close to the original estimates as possible.

3. Incorporation of both measurements and estimates of neighbors (i.e., implementation of both adaptation and combination). There are two strategies, and they are known as ATC (adapt-then-combine) or alternatively CTA (combine-then-adapt). It has been proved that ATC outperforms CTA in terms of estimation quality [4,37].

4.2.2 SHARING OF MEASUREMENTS OR STATISTICS

Let us fix an agent $k \in \mathcal{K}$ and assume that at time t it has access to the neighbors' observations $y_{t,j}$ and explanatory variables $z_{t,j}$, where $j \in \mathfrak{N}_k$. Alternatively, these observations may be surrogated by the sufficient statistics $T(y_{t,j}, z_{t,j})$. Then agent k can improve its knowledge about θ by incorporating them in the same way as its own sufficient statistic. If $\pi_k(\theta|\zeta_{t-1,k})$ is the k's prior distribution of θ at time t, where $\zeta_{t-1,k}$ stands for all the information available to agent k by time $t-1$, including any previously shared information, the distributed variant of the Bayes' theorem (4.1) reads

$$\pi_k(\theta|\zeta_{t,k}) \propto \pi_k(\theta|\zeta_{t-1,k}) \prod_{j\in\mathfrak{N}_k} f(y_{t,j}|z_{t,j},\theta), \tag{4.7}$$

where we assumed that the observations are conditionally independent. If the models $f(y_{t,j}|z_{t,j},\theta)$ belong to the exponential family of distributions and the prior distribution $\pi_k(\theta|\zeta_{t-1,k})$ is a conjugate distribution, then Bayes' theorem reduces to the update of the k's hyperparameters according to (see Eqs. (4.4) and (4.5)),

$$\xi_{t,k} = \xi_{t-1,k} + \sum_{j\in\mathfrak{N}_k} T(y_{t,j}, z_{t,j}), \tag{4.8}$$

$$\nu_{t,k} = \nu_{t-1,k} + |\mathfrak{N}_k|, \tag{4.9}$$

where $|\mathfrak{N}_k|$ is the cardinality of the set $\mathfrak{N}_k$, i.e., the number of neighbors plus one.

4.2.3 MERGING OF BAYESIAN ESTIMATORS

Now we fix again an agent $k \in \mathcal{K}$ and assume that the network agents updated their prior distributions by either own or shared measurements, and that their posterior distributions $\pi_j(\theta|\zeta_{t,j})$ are shared with the neighbors. That is, agent k has access to the set $\{\pi_j(\theta|\zeta_{t,j}); j \in \mathfrak{N}_k\}$. Each member of this set may be assigned a nonnegative weight $a_{kj} \leq 1$ expressing the (subjective) probability that the related posterior is true at the moment. The weights thus take values from a corresponding probability simplex and sum to unity. For simplicity, we assume that the weights a_{kj} are constant and either uniform, or selected according to a convenient rule, e.g., [5, Chap. 8]. A model-based Bayesian treatment of the weights can be found in [37].

Keeping all the posterior distributions in the set would, however, quickly lead to an explosion of its size. In order to prevent this situation, we aim to combine the individual posterior distributions to a single distribution $\tilde{\pi}_k(\theta|\cdot)$, which best expresses the information in all of them. The Bayesian theory advocates the use of the Kullback-Leibler divergence, $\mathcal{D}(\cdot||\cdot)$, as a proper dissimilarity (loss) measure [41]. A theoretically consistent combination step is equivalent to seeking the minimizer of

$$\begin{aligned}\sum_{j\in\mathfrak{N}_k} a_{kj}\mathcal{D}\left(\tilde{\pi}_k\left(\theta|\cdot\right)\|\pi_j\left(\theta|\cdot\right)\right) &= \sum_{j\in\mathfrak{N}_k} a_{kj}\mathbb{E}_{\tilde{\pi}_k}\left[\log\frac{\tilde{\pi}_k\left(\theta|\cdot\right)}{\pi_j\left(\theta|\cdot\right)}\right]\\ &= \mathbb{E}_{\tilde{\pi}_k}\left[\log\frac{\tilde{\pi}_k(\theta|\cdot)}{\frac{\prod_{j\in\mathfrak{N}_k}\left[\pi_j(\theta|\cdot)\right]^{a_{kj}}}{\int\prod_{j\in\mathfrak{N}_k}\left[\pi_j(\theta|\cdot)\right]^{a_{kj}}d\theta}}\right] - \log\int\prod_{j\in\mathfrak{N}_k}\left[\pi_j(\theta|\cdot)\right]^{a_{kj}}d\theta\\ &= \mathcal{D}\left(\tilde{\pi}_k(\theta|\cdot)\middle\|c\prod_{j\in\mathfrak{N}_k}\left[\pi_j(\theta|\cdot)\right]^{a_{kj}}\right) + \text{const.},\end{aligned} \tag{4.10}$$

where c is a proportionality constant assuring that the result is a valid density. The first equality in Eq. (4.10) is due to the definition of the Kullback-Leibler divergence, and the second follows from easy algebra. Because the Kullback-Leibler divergence is minimal if the arguments are equal, the minimizing density has the form

$$\tilde{\pi}_k\left(\theta|\cdot\right) \propto \prod_{j\in\mathfrak{N}_k}\left[\pi_j(\theta|\cdot)\right]^{a_{kj}}. \tag{4.11}$$

There is a notable property of this combination rule. If the posterior distributions belong to the exponential family, then Eq. (4.11) provides an analytically tractable way for obtaining the posterior $\tilde{\pi}_k(\theta|\cdot) = \tilde{\pi}_k(\theta|\tilde{\xi}_{t,k}, \tilde{\nu}_{t,k})$. Then its hyperparameters are given by

$$\tilde{\xi}_{t,k} = \sum_{j\in\mathfrak{N}_k} a_{kj}\xi_{t,k}, \quad \text{and} \quad \tilde{\nu}_{t,k} = \sum_{j\in\mathfrak{N}_k} a_{kj}\nu_{t,k}. \tag{4.12}$$

The above equations suggest that the hyperparameters of the resulting distribution of the kth agent are obtained by a linear combination of the hyperparameters of the individual distributions, $\xi_{t,j}$ and $\nu_{t,j}$, $j \in \mathfrak{N}_k$. Recall from Eq. (4.5) that $\xi_{t,j}$ and $\nu_{t,j}$ aggregate the agents' observations.

A natural question is whether it is possible to proceed with swapped arguments of the divergence. We proceed by writing,

$$\begin{aligned}\sum_{j\in\mathfrak{N}_k} a_{kj}\mathcal{D}\left(\pi_j\left(\theta|\cdot\right)\middle\|\tilde{\pi}_k\left(\theta|\cdot\right)\right) &= \sum_{j\in\mathfrak{N}_k} a_{kj}\mathbb{E}_{\pi_j}\left[\log\frac{\pi_j\left(\theta|\cdot\right)}{\tilde{\pi}_k\left(\theta|\cdot\right)}\right]\\ &= \sum_{j\in\mathfrak{N}_k} a_{kj}\mathbb{E}_{\pi_j}\left[\log\pi_j(\theta|\cdot)\right]\\ &\quad - \mathbb{E}_{\sum_{j\in\mathfrak{N}_k} a_{kj}\pi_j}\left[\log\sum_{j\in\mathfrak{N}_k} a_{kj}\pi_j(\theta|\cdot)\right]\\ &\quad + \mathbb{E}_{\sum_{j\in\mathfrak{N}_k} a_{kj}\pi_j}\left[\log\frac{\sum_{j\in\mathfrak{N}_k} a_{kj}\pi_j\left(\theta|\cdot\right)}{\tilde{\pi}_k\left(\theta|\cdot\right)}\right] \qquad (4.13)\\ &= \mathcal{D}\left(\sum_{j\in\mathfrak{N}_k} a_{kj}\pi_j(\theta|\cdot)\middle\|\tilde{\pi}_k(\theta|\cdot)\right) + \text{const.} \qquad (4.14)\end{aligned}$$

The first two terms in Eq. (4.13) do not depend on $\tilde{\pi}_k(\theta|\cdot)$, and the minimum of the divergence is achieved by minimization of the last term, which leads to

$$\tilde{\pi}_k(\theta|\cdot) = \sum_{j \in \mathfrak{N}_k} a_{kj} \pi_j(\theta|\cdot). \tag{4.15}$$

The result of the minimization is a mixture of the neighbors' posterior distributions. The number of components is equal to $|\mathfrak{N}_k|$, and with time it will explode, unless a suitable component merging/pruning procedure is implemented, e.g., [43].

Example: Covariance intersection

A nice example of the Kullback-Leibler optimal combination (4.11) is the merging of normal densities, which belong to the exponential family of distributions. Assume that the posterior of $\theta \in \mathbb{R}^n$ is $\pi_j(\theta|\cdot) \sim \mathcal{N}(\mu_{t,j}, \Sigma_{t,j})$. If we drop the time indices for notational simplicity and rewrite the density in the form (4.2), we obtain

$$\begin{aligned} \pi_j\left(\theta|\mu_j, \Sigma_j\right) &= (2\pi)^{-\frac{n}{2}} (\det \Sigma_j)^{-\frac{1}{2}} e^{-\frac{1}{2}(\theta-\mu_j)^\intercal \Sigma_j^{-1}(\theta-\mu_j)} \\ &\propto \exp\left\{ \mathrm{Tr}\left(\begin{bmatrix} \mu_j^\intercal \Sigma_j^{-1} \\ -\frac{1}{2}\Sigma_j^{-1} \end{bmatrix}^\intercal \begin{bmatrix} \theta^\intercal \\ \theta\theta^\intercal \end{bmatrix} \right) - \frac{1}{2}\mu_j^\intercal \Sigma_j^{-1} \mu_j \right\}, \end{aligned} \tag{4.16}$$

where the natural parameter and the sufficient statistic have the form

$$\eta_j = \begin{bmatrix} \mu_j^\intercal \Sigma_j^{-1} \\ -\frac{1}{2}\Sigma_j^{-1} \end{bmatrix} \quad \text{and} \quad T(\theta) = \begin{bmatrix} \theta^\intercal \\ \theta\theta^\intercal \end{bmatrix}.$$

The combination of all the Gaussian distributions will produce another Gaussian. From Eq. (4.11) it follows that

$$\tilde{\eta}_k = \begin{bmatrix} \tilde{\mu}_k^\intercal \tilde{\Sigma}_k^{-1} \\ -\frac{1}{2}\tilde{\Sigma}_k^{-1} \end{bmatrix} = \sum_{j \in \mathfrak{N}_k} a_{kj} \eta_j = \sum_{j \in \mathfrak{N}_k} a_{kj} \begin{bmatrix} \mu_j^\intercal \Sigma_j^{-1} \\ -\frac{1}{2}\Sigma_j^{-1} \end{bmatrix}, \tag{4.17}$$

and a little algebra yields the resulting mean vector and covariance matrix of the resulting Gaussian,

$$\tilde{\mu}_{t,k} = \tilde{\Sigma}_{t,k} \left(\sum_{j \in \mathfrak{N}_k} a_{kj} \Sigma_{t,j}^{-1} \mu_{t,j} \right) \quad \text{and} \quad \tilde{\Sigma}_{t,k} = \left[\sum_{j \in \mathfrak{N}_k} a_{kj} \Sigma_{t,j}^{-1} \right]^{-1}. \tag{4.18}$$

This result is known as *covariance intersection*. It is worth noting that the same rule applies to any other distribution from the exponential family.

Which combination algorithm?

Both the presented algorithms that combine the posterior estimates are optimal in the Kullback-Leibler sense, yet they may lead to significantly different results. The obvious question is which algorithm should be used in a given situation. As mentioned above, the algorithm (4.11) should be used in situations where model and parameter homogeneity are assumed. Further, its analytical tractability under conjugate priors is a very attractive feature. On the other hand, the second algorithm (4.15) is better in situations where the agents use different models and/or parameters. Then, the algorithm

provides a (mixture) density that better fits the regions where the neighbors' densities are large enough (for more details, see [42]). As pointed out, the algorithm requires an additional procedure for pruning and merging of the components to prevent the mixture from a rapid growth of its number of components. A specific situation arises if (sequential) Monte Carlo methods are used for estimation or filtering, and the posterior distribution is approximated by samples. Sampling from the mixture (4.15) is equivalent to gathering a relevant number of samples from the neighbors. As this may be communication-intensive, the posteriors may be approximated by a normal mixture at each agent as in the Gaussian particle filter [44–46]. The combination rule (4.15) yields again a normal mixture from which sampling is trivial.

Example

Let us briefly investigate the properties of the two combination algorithms on a simple example. Assume that there are two normal posterior distributions available to agent 1,

$$\begin{aligned} \pi_1(\theta|\cdot) &= \mathcal{N}(0,1) \quad \text{with} \quad a_{11} = 0.5, \\ \pi_2(\theta|\cdot) &= \mathcal{N}(1,1) \quad \text{with} \quad a_{12} = 0.5. \end{aligned}$$

It is straightforward to prove that the combination rule (4.11) yields a normal distribution,

$$\tilde{\pi}_1(\theta|\cdot) = \mathcal{N}(0.5,1),$$

which is a good compromise between the two original distributions. The combination rule (4.15) yields a mixture

$$\tilde{\pi}_1(\theta|\cdot) = 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(1,1), \tag{4.19}$$

which preserves the available information about the location of the mean values, but at the cost of higher complexity.

Now, assume two normal distributions that differ only in the variance,

$$\begin{aligned} \pi_1(\theta|\cdot) &= \mathcal{N}(0,100) \quad \text{with} \quad a_{11} = 0.5, \\ \pi_2(\theta|\cdot) &= \mathcal{N}(0,1) \quad \text{with} \quad a_{12} = 0.5. \end{aligned}$$

The combination rule (4.11) produces a normal distribution

$$\tilde{\pi}_1(\theta|\cdot) = \mathcal{N}(0,1.98),$$

and the second combination rule (4.15) leads to a mixture

$$\tilde{\pi}_1(\theta|\cdot) = 0.5\mathcal{N}(0,100) + 0.5\mathcal{N}(0,1). \tag{4.20}$$

If we assume that agent 1 joined the network with an initial prior while the second agent already performed several updates and has a good knowledge of θ, we see that the former combination rule significantly improves the distribution of agent 1. If agent 2 uses $a_{21} = 0.5$ and $a_{22} = 0.5$ (has the same beliefs as agent 1), the combined distribution of this agent would be the same as that of agent 1.

Hence, a combination of a distribution that reflects high-ignorance with one with low ignorance, does not much affect the latter.

4.3 EXAMPLE: DIFFUSION KALMAN FILTER

The first diffusion Kalman filter was proposed in [30]. Later in [31], it was improved with a covariance intersection-based procedure, which was applied in the combination step. Below we derive the diffusion Kalman filter from a Bayesian viewpoint. More on the diffusion Kalman filter but derived from a different perspective can be found in the chapter on Distributed Kalman and Particle Filtering.

To begin, we assume a hidden Markov model of the form[1]

$$x_t|x_{t-1}, z_t \sim \mathcal{N}\left(A_t x_{t-1} + B_t z_t, Q_t\right), \tag{4.21}$$

$$y_t|x_t \sim \mathcal{N}\left(H_t x_t, R_t\right), \tag{4.22}$$

where x_t is an n-dimensional state vector, y_t is an l-dimensional observation vector, z_t is a known input vector of length m, A_t, B_t, and H_t are matrices of compatible dimensions, Q_t and R_t are $n \times n$ and $\ell \times \ell$ state and observation covariance matrices, respectively.

The observation model (4.22) is normal, so we rewrite it by using the exponential family form (4.2),

$$\begin{aligned} f(y_t|x_t) &\propto \exp\left\{-\frac{1}{2}(y_t - H_t x_t)^\mathsf{T} R_t^{-1}(y_t - H_t x_t)\right\} \\ &= \exp\left\{\mathrm{Tr}\Bigg(\underbrace{-\frac{1}{2}\begin{bmatrix}-1\\ x_t\end{bmatrix}\begin{bmatrix}-1\\ x_t\end{bmatrix}^\mathsf{T}}_{\eta(x_t)}\underbrace{\begin{bmatrix}y_t^\mathsf{T}\\ H_t^\mathsf{T}\end{bmatrix} R_t^{-1}\begin{bmatrix}y_t^\mathsf{T}\\ H_t^\mathsf{T}\end{bmatrix}^\mathsf{T}}_{T(y_t)}\Bigg)\right\}. \end{aligned} \tag{4.23}$$

Because the model (4.22) is normal, it is advantageous to set the prior distribution to normal too, as it is conjugate to the model and hence the posterior will be analytically tractable,

$$\pi(x_t|y_{0:t-1}, z_{0:t-1}) = \mathcal{N}(x_t^-, P_t^-), \qquad x_t^- \in \mathbb{R}^n, P_t^- \in \mathbb{R}^{n\times n}$$

$$\begin{aligned} &\propto \exp\left\{-\frac{1}{2}(x_t - x_t^-)^\mathsf{T}(P_t^-)^{-1}(x_t - x_t^-)\right\} \\ &= \exp\left\{\mathrm{Tr}\Bigg(\underbrace{-\frac{1}{2}\begin{bmatrix}-1\\ x_t\end{bmatrix}\begin{bmatrix}-1\\ x_t\end{bmatrix}^\mathsf{T}}_{\eta(x_t)}\underbrace{\begin{bmatrix}(x_t^-)^\mathsf{T}\\ I\end{bmatrix}(P_t^-)^{-1}\begin{bmatrix}(x_t^-)^\mathsf{T}\\ I\end{bmatrix}^\mathsf{T}}_{\xi_t}\Bigg)\right\}, \end{aligned} \tag{4.24}$$

where I is an identity matrix of appropriate dimensions. The minus superscripts denote parameters of the Gaussian where the measurements from time t have not been incorporated yet, whereas the plus signs signify that they have been used. Also, the state transition $x_{t-1} \to x_t$ amounts to updating $\mathcal{N}(x_{t-1}^+, P_{t-1}^+) \to \mathcal{N}(x_t^-, P_t^-)$ according to Eq. (4.21).

Next, we return to the implementation of the Kalman filter over a network, where all the agents use the same model. The matrices A_t, B_t, and Q_t are the same for all the agents whereas $H_{t,k}$, and $R_{t,k}$ are

[1] We temporarily drop the agent's indices for simplicity. Also, instead of θ, the unknowns of interest here are the vectors x_t.

distinctive as are the variables $z_{t,k}$. The adaptation step of the diffusion Kalman filter is similar to the ordinary Kalman filter update. First, the state transition given by Eq. (4.21) is performed as usual,

$$\begin{aligned}\pi_k(x_{t,k}|\zeta_{t-1,k}) &= \int \pi(x_{t,k}|x_{t-1}, z_t)\pi_k(x_{t-1}|\zeta_{t-1,k})dx_{t-1} \\ &= \mathcal{N}_k\left(A_t x^+_{t-1,k} + B_t z_{t,k}, A_t P^+_{t-1,k} A_t^\intercal + Q_t\right) \\ &= \mathcal{N}_k(x^-_{t,k}, P^-_{t,k}),\end{aligned} \tag{4.25}$$

and followed by the Bayesian update by the *neighbors'* measurements

$$\begin{aligned}\pi_k(x_t|\zeta_{t,k}) &\propto \pi_k(x_{t,k}|\zeta_{t-1,k}) \prod_{j\in\mathfrak{N}_k} f(y_{t,j}|x_{t,j}) \\ &= \mathcal{N}_k(x^+_{t,k}, P^+_{t,k}).\end{aligned} \tag{4.26}$$

Taking the exponential family form (4.23) and the conjugate form (4.24) into account, we see that

$$\xi_{t,k} = \xi_{t-1,k} + \sum_{j\in\mathfrak{N}_k} \begin{bmatrix} y_{t,j} \\ H_{t,j}\end{bmatrix} R^{-1}_{t,j} \begin{bmatrix} y_{t,j} \\ H_{t,j}\end{bmatrix}^\intercal, \tag{4.27}$$

$$v_{t,k} = v_{t-1,k} + |\mathfrak{N}_k|. \tag{4.28}$$

Simple algebra reveals the recursions

$$P^+_{t,k} = \left[(P^-_{t,k})^{-1} + \sum_{j\in\mathfrak{N}_k} H^\intercal_{t,j} R^{-1}_{t,j} H_{t,j}\right]^{-1}, \tag{4.29}$$

$$x^+_{t,k} = x^-_{t,k} + P^+_{t,k} \sum_{j\in\mathfrak{N}_k} H^\intercal_{t,j} R^{-1}_{t,j}\left(y_{t,j} - H_{t,j} x^-_{t,k}\right). \tag{4.30}$$

The above two equations describe the adaptation step.

The combination step operates directly with $\xi_{t,j}$ according to Eq. (4.11). Application of Eq. (4.18) from Section 4.2.3 shows that

$$\tilde{\pi}_k(x_{t,k}|\zeta_{t,k}) = \prod_{j\in\mathfrak{N}_k}\left[\mathcal{N}_j(x^+_{t,j}, P^+_{t,j})\right]^{a_{kj}} = \mathcal{N}_k(\tilde{x}^+_{t,k}, \tilde{P}^+_{t,k}) \tag{4.31}$$

with the hyperparameters

$$\tilde{P}^+_{t,k} = \left[\sum_{j\in\mathfrak{N}_k} a_{kj}\left(P^+_{t,j}\right)^{-1}\right]^{-1}, \tag{4.32}$$

$$\tilde{x}^+_{t,k} = \tilde{P}^+_{t,k} \sum_{j\in\mathfrak{N}_k} a_{kj}\left(P^+_{t,j}\right)^{-1} x^+_{t,j}. \tag{4.33}$$

The algorithm is summarized in Algorithm 4.1.

Algorithm 4.1 DIFFUSION KALMAN FILTER WITH ADAPT-THEN-COMBINE (ATC) STRATEGY

Initialize agents $k = 1, 2, \ldots, K$ with the prior densities $\pi_k(\theta|\zeta_{k,0})$. Set the weights a_{kj}. For $t = 1, 2, \ldots$ and each agent k do:

Kalman prediction:

- Update the prior densities $\mathcal{N}_k(x^+_{t-1,k}, P^+_{t-1,k}) \rightarrow \mathcal{N}_k(x^-_{t,k}, P^-_{t,k})$, Eq. (4.25).

Kalman update:

1. Acquire observations $y_{t,j}$ of neighbors $j \in \mathfrak{N}_k$.
2. *Adaptation:* Perform Kalman adaptation according to Eq. (4.26) by updating the hyperparameters via Eq. (4.27).
3. Obtain posterior densities $\pi_j(x_{t,j}|\zeta_{t,j})$ of neighbors $j \in \mathfrak{N}_k$ by acquiring their respective hyperparameters $\xi_{t,j}$, $\nu_{t,j}$.
4. *Combination:* Combine posterior densities according by implementing Eqs. (4.32) and (4.33).

4.3.1 SIMULATION EXAMPLE

We illustrate the performance of the diffusion Kalman filter with a two-dimensional tracking problem. The matrices of the model were time-invariant, and they were defined as follows:

$$A = \begin{bmatrix} 1 & 0 & \Delta & 0 \\ 0 & 1 & 0 & \Delta \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad Q = q \cdot \begin{bmatrix} \frac{\Delta^3}{3} & 0 & \frac{\Delta^2}{2} & 0 \\ 0 & \frac{\Delta^3}{3} & 0 & \frac{\Delta^2}{2} \\ \frac{\Delta^2}{2} & 0 & \Delta & 0 \\ 0 & \frac{\Delta^2}{2} & 0 & \Delta \end{bmatrix},$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad R = r^2 \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

where $\Delta = 0.1$, $q = 5.0$, $r = 0.1k$, and $k = 1, \ldots, 15$ is the agent's number. That is, the agents had different observation noise covariance matrices. The simulation was started from the origin of the coordinate system. The state vector elements represent the position of the target in the plane and its velocity components.

Figs. 4.1 and 4.2 show the topology of the network and the simulated trajectory with noisy observations of agents 1 and 15, respectively. The Kalman filters were initialized with $P^+_{0,k} = 1000 I_{4\times 4}$ and zero vectors $x^+_{0,k}$. The combination weights a_{kj} were uniform and constant. Four strategies were tested: (1) no cooperation (nocoop), (2) adaptation only (A-only), (3) combination only (C-only), and (4) adapt then combine (ATC). Fig. 4.3 depicts the box plots of the mean square errors (MSEs) of the estimation of all four elements of the state vector. We see that the collaboration among agents improved the estimation performance. Further, the performance of the C-only strategy was superior to that of the A-only strategy.

4.4 CONCLUSION

The Bayesian approach to the inference of unknown parameters of probabilistic models has numerous attractive features. One of the most prominent is its wide applicability. Further, regardless of whether

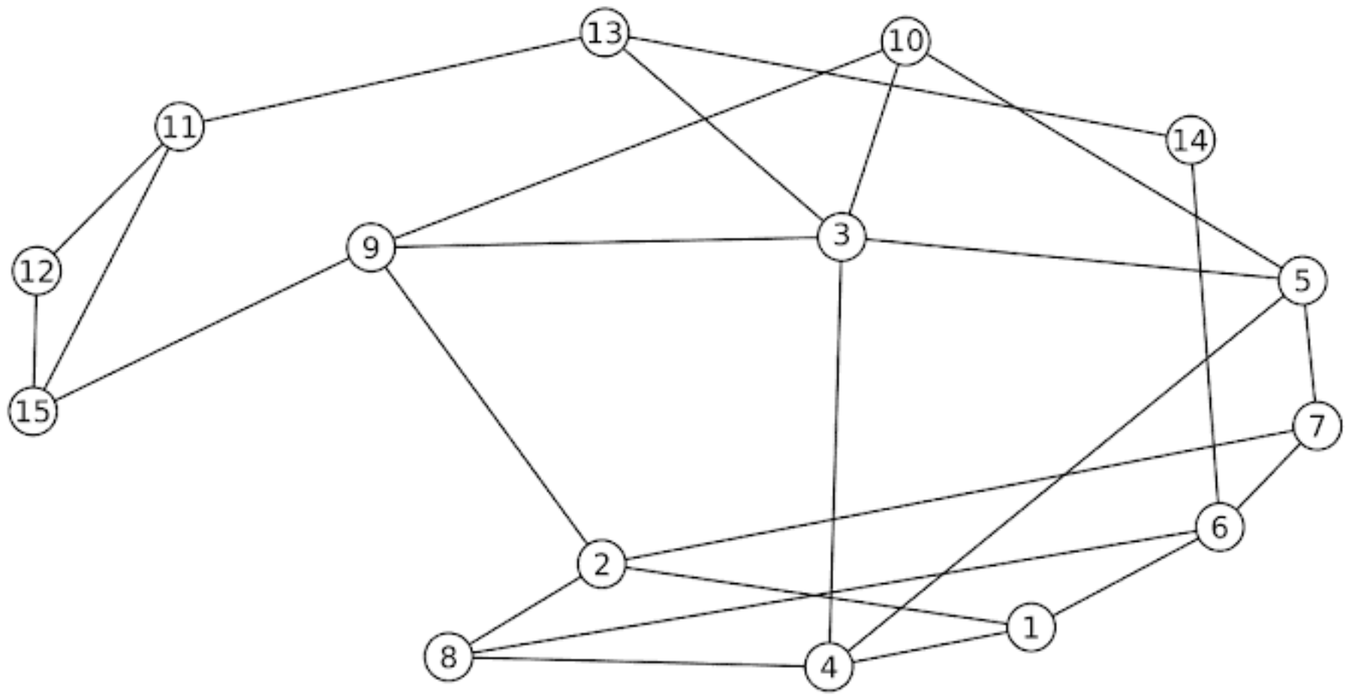

FIG. 4.1

Network layout.

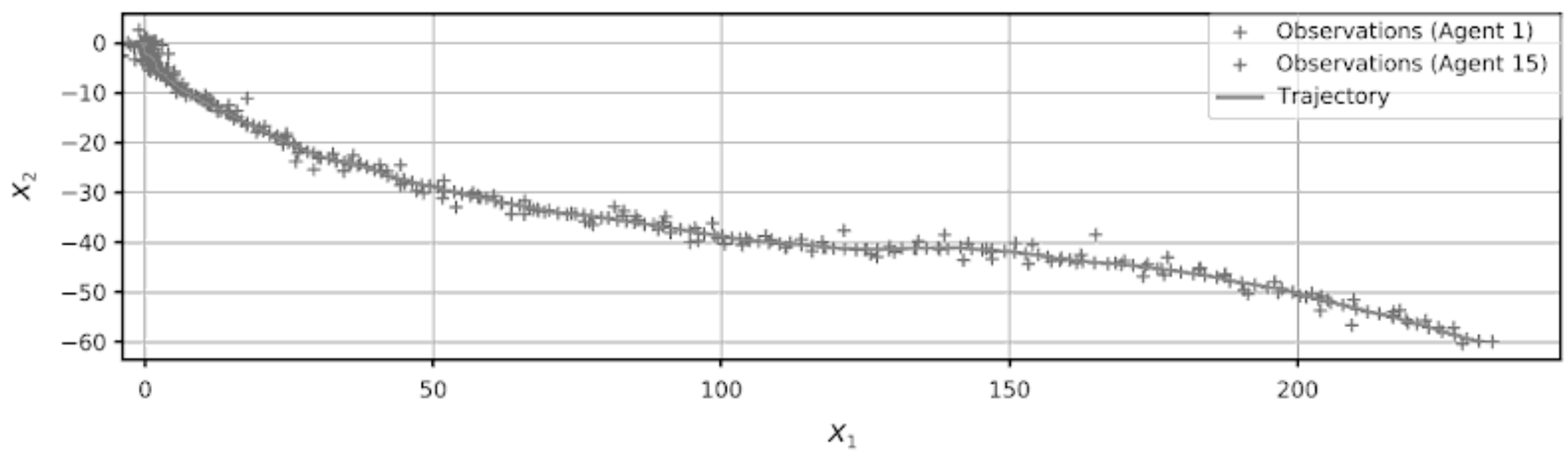

FIG. 4.2

Simulated trajectory and observations of agents 1 and 15.

one deals with linear or nonlinear regression, state-space models, hierarchical models, or any other model type, Bayesian inference relies on the same principles. Unlike in classical (frequentist) statistics, the estimate is represented by a posterior distribution, quantifying not only its expected location but also the uncertainty associated with it. The representation of estimates by posterior distributions remains a cornerstone of the use of Bayesian principles in networked systems, where agents collaborate to improve their own estimates. The main idea is that less-informed agents improve their knowledge while well-informed ones do not reduce it.

There are many open problems that could be investigated. For instance, the determination of combination weights is one, although several methods for choosing them have already been proposed [37,47]. The described combination methods are a small sample from the set of possible approaches, too. For instance, in [48], yet another approach was proposed where the agents fuse received information from neighbors by using mixtures with weights proportional to predictive distributions obtained from the posteriors of the respective agents. Furthermore, the topic of heterogeneous models and/or parameters has attained a huge interest in the last years, but the only Bayesian treatment the

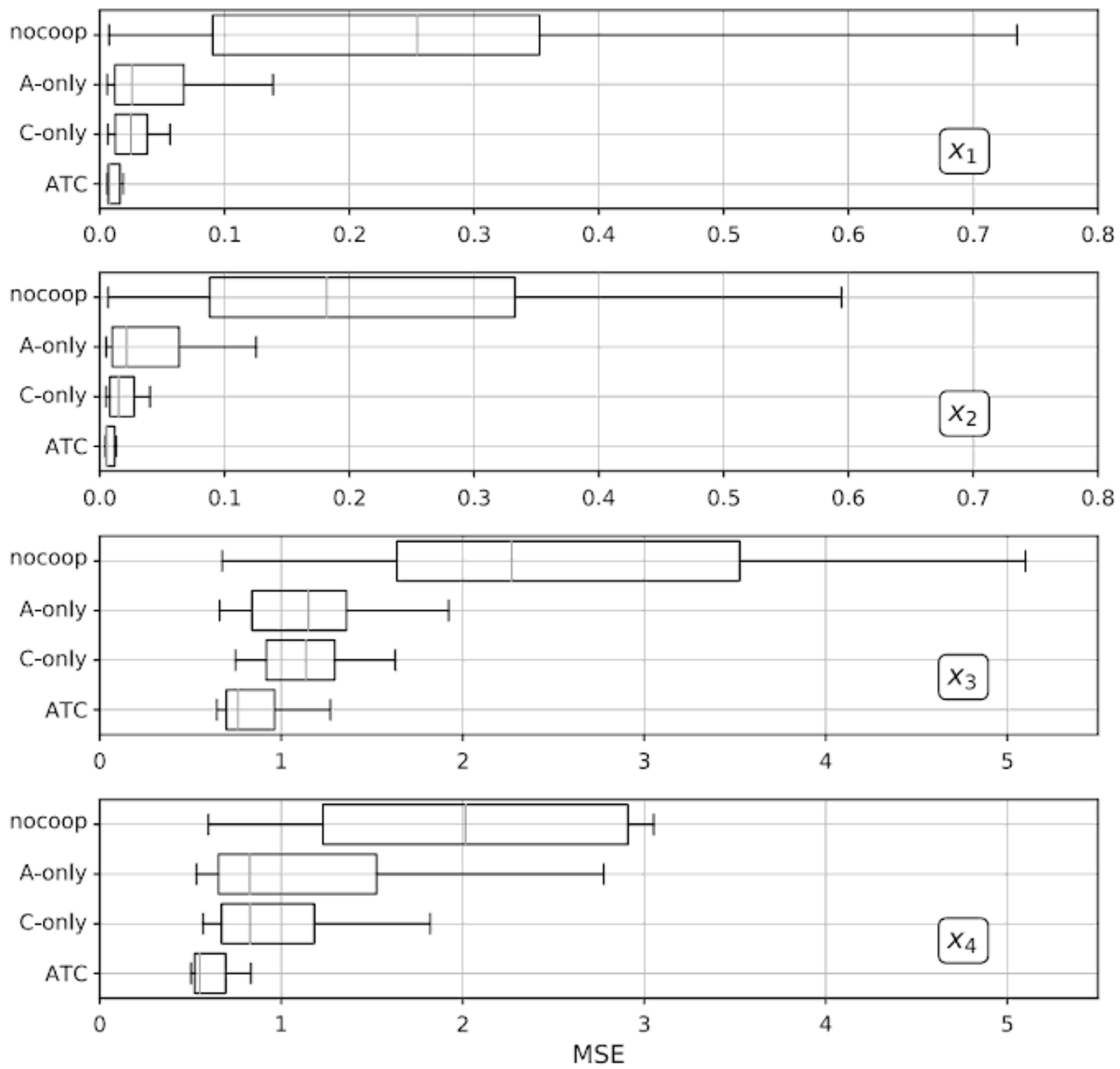

FIG. 4.3

MSEs of various strategies.

authors are aware of was proposed in [42]. Naturally, the robustness to failures and the communication and computational limitations of the agents form other interesting topics for research on collaborative inference in networks of agents.

ACKNOWLEDGMENTS

The work of K. Dedecius was supported by the Czech Science Foundation, project No. 16-09848S. The work of P.M. Djurić was supported by NSF under Award CCF-1618999.

REFERENCES

[1] Candy JV. Bayesian signal processing: classical, modern, and particle filtering methods. John Wiley & Sons; 2016.

[2] Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E. A survey on sensor networks. IEEE Commun Mag 2002;40(8):102–14.

[3] Liu Y, Li C, Tang WK, Zhang Z. Distributed estimation over complex networks. Information Sciences 2012;197:91–104.
[4] Sayed AH, et al. Adaptation, learning, and optimization over networks. In: Foundations and trends in machine learning, vol. 7(4-5); 2014. p. 311–801.
[5] Sayed AH. Diffusion adaptation over networks. In: Chellapa R, Theodoridis S, editors. Academic Press Library in Signal Processing, vol. 3. Academic Press, Elsevier; 2014. p. 323–454.
[6] Li W, Wang Z, Yuan Y, Guo L. Particle filtering with applications in networked systems: a survey. Complex Intell Syst 2016;2(4):293–315.
[7] Athans M, Tsitsiklis JN. Convergence and asymptotic agreement in distributed decision problems. IEEE Trans. Autom. Control 1982;21:692–701.
[8] Tsitsiklis J, Bertsekas DP, Athans M. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. IEEE Trans Autom Control 1986;31(9):803–12.
[9] Bertsekas DP. A new class of incremental gradient methods for least squares problems. SIAM J Optim 1997;7(4):913–26.
[10] Nedic A, Bertsekas DP. Incremental subgradient methods for nondifferentiable optimization. SIAM J Optim 2001;12(1):109–38.
[11] Rabbat MG, Nowak RD. Quantized incremental algorithms for distributed optimization. IEEE J Sel Areas Commun 2005;23(4):798–808.
[12] Lopes CG, Sayed AH. Incremental adaptive strategies over distributed networks. IEEE Trans Signal Process 2007;55(8):4064–77.
[13] Plata-Chaves J, Bogdanovic N, Berberidis K. Distributed incremental-based RLS for node-specific parameter estimation over adaptive networks. In: Proceedings of the 21st European signal processing conference (EUSIPCO); 2013. p. 1–5.
[14] DeGroot MH. Reaching a consensus. J Am Stat Assoc 1974;69(345):118–21.
[15] Olfati-Saber R, Murray RM. Consensus problems in networks of agents with switching topology and time-delays. IEEE Trans Autom Control 2004;49(9):1520–33.
[16] Olfati-Saber R, Fax JA, Murray RM. Consensus and cooperation in networked multi-agent systems. Proc IEEE 2007;95(1):215–33.
[17] Schizas ID, Giannakis GB, Luo ZQ. Distributed estimation using reduced-dimensionality sensor observations. IEEE Trans Signal Process 2007;55(8):4284–99.
[18] Schizas ID, Mateos G, Giannakis GB. Distributed LMS for consensus-based in-network adaptive processing. IEEE Trans Signal Process 2009;57(6):2365–82.
[19] Guldogan MB. Consensus Bernoulli filter for distributed detection and tracking using multi-static Doppler shifts. IEEE Signal Process Lett 2014;21(6):672–6.
[20] Hlinka O, Hlawatsch F, Djurić PM. Distributed particle filtering in agent networks: a survey, classification, and comparison. IEEE Signal Process Mag 2013;30(1):61–81.
[21] Cattivelli FS, Lopes CG, Sayed AH. Diffusion recursive least-squares for distributed estimation over adaptive networks. IEEE Trans Signal Process 2008;56(5):1865–77.
[22] Cattivelli FS, Sayed AH. Diffusion LMS strategies for distributed estimation. IEEE Trans Signal Process 2010;58(3):1035–48.
[23] Bertrand A, Moonen M, Sayed AH. Diffusion bias-compensated RLS estimation over adaptive networks. IEEE Trans Signal Process 2011;59(11):5212–24.
[24] Zhao X, Tu SY, Sayed AH. Diffusion adaptation over networks under imperfect information exchange and non-stationary data. IEEE Trans Signal Process 2012;60(7):3460–75.
[25] Dedecius K, Sečkárová V. Dynamic diffusion estimation in exponential family models. IEEE Signal Process Lett 2013;20(11):1114–7.
[26] Arablouei R, Dogancay K, Werner S, Huang YF. Adaptive distributed estimation based on recursive least-squares and partial diffusion. IEEE Trans Signal Process 2014;62(14):3510–22.

[27] Lopes CG, Sayed AH. Diffusion least-mean squares over adaptive networks: formulation and performance analysis. IEEE Trans Signal Process 2008;56(7):3122–36.
[28] Chen J, Richard C, Hero AO, Sayed AH. Diffusion LMS for multitask problems with overlapping hypothesis subspaces. In: Proceedings of the IEEE international workshop on machine learning for signal processing; 2014. p. 1–6.
[29] Plata-Chaves J, Bogdanović N, Berberidis K. Distributed diffusion-based LMS for node-specific adaptive parameter estimation. IEEE Trans Signal Process 2015;63(13):3448–60.
[30] Cattivelli FS, Sayed AH. Diffusion strategies for distributed Kalman filtering and smoothing. IEEE Trans Autom Control 2010;55(9):2069–84.
[31] Hu J, Xie L, Zhang C. Diffusion Kalman filtering based on covariance intersection. IEEE Trans Signal Process 2012;60(2):891–902.
[32] Towfic ZJ, Chen J, Sayed AH. Collaborative learning of mixture models using diffusion adaptation. In: Proceedings of the IEEE international workshop on machine learning for signal processing; 2011. p. 1–6.
[33] Pereira SS, Pages-Zamora A, López-Valcarce R. A diffusion-based distributed EM algorithm for density estimation in wireless sensor networks. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing; 2013. p. 4449–53.
[34] Dedecius K, Reichl J, Djurić PM. Sequential estimation of mixtures in diffusion networks. IEEE Signal Process Lett 2015;22(2):197–201.
[35] Bruno MG, Dias SS. Collaborative emitter tracking using Rao-Blackwellized random exchange diffusion particle filtering. EURASIP J Adv Signal Process 2014;2014(1):19.
[36] Dedecius K. Adaptive approximate filtering of state-space models. In: Proceedings of the European signal processing conference (EUSIPCO); 2015. p. 2236–40.
[37] Dedecius K, Djurić PM. Sequential estimation and diffusion of information over networks: a Bayesian approach with exponential family of distributions. IEEE Trans Signal Process 2017;65(7):1795–809.
[38] Dedecius K, Djurić PM. Diffusion filtration with approximate Bayesian computation. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing; 2015. p. 3207–11.
[39] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Chapman & Hall/CRC; 2003.
[40] Robert C. The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media; 2007.
[41] Bishop CM. Pattern recognition and machine learning. New York, NY: Springer; 2006.
[42] Dedecius K, Sečkárová V. Factorized estimation of partially shared parameters in diffusion networks. IEEE Trans Signal Process 2017;65(19):5153–63.
[43] Frühwirth-Schnatter S. Finite mixture and Markov switching models. Springer Science & Business Media; 2006.
[44] Kotecha JH, Djurić PM. Gaussian particle filtering. IEEE Trans Signal Process 2003;51(10):2592–601.
[45] Kotecha JH, Djurić PM. Gaussian sum particle filtering. IEEE Trans Signal Process 2003;51(10):2602–12.
[46] Hlinka O, Sluciak O, Hlawatsch F, Djurić PM, Rupp M. Likelihood consensus and its application to distributed particle filtering. IEEE Trans Signal Process 2012;60(8):4334–49.
[47] Takahashi N, Yamada I, Sayed AH. Diffusion least-mean squares with adaptive combiners: formulation and performance analysis. IEEE Trans Signal Process 2010;58(9):4795–810.
[48] Djurić PM, Dedecius K. Bayesian estimation of unknown parameters over networks. In: Proceedings of the 24th European signal processing conference (EUSIPCO); 2016. p. 1508–12.