# Factorized Estimation of Partially Shared Parameters in Diffusion Networks

Kamil Dedecius and Vladimíra Sečkárová

***Abstract*—Collaborative estimation of partially common parameters over ad hoc diffusion networks where the nodes directly communicate with their neighbors is a challenging task. The problem complexity is significantly high under the lack of knowledge which parameters are shared and among which network nodes. In this paper, we propose an adaptive framework suitable for this task. It is abstractly formulated in the Bayesian and information-theoretic paradigms and, therefore, versatile and easily applicable to a relatively wide class of models. If the observation models belong to the exponential family and the same functional types of prior probability distributions are used for estimation of the shared parameters, the method reduces to an analytically tractable variational algorithm extended by a procedure that passes messages among network nodes. A simulation example demonstrates that the collaboration improves estimation performance of both the shared and strictly local parameters, compared with the noncollaborative scenario.**

***Index Terms*—Diffusion network, diffusion estimation, adaptation, heterogeneous parameters, multitask networks.**

## I. Introduction

COLLABORATIVE estimation of parameters and states of stochastic models over networks of interconnected nodes has attracted a tremendous interest in the last two decades, particularly due to the rapid development of cheap ad-hoc wireless networks with computationally powerful devices endowed with sensing, data processing and communication capabilities. The application fields of these networks are rapidly growing, and involve sensor networks, precision agriculture, environment monitoring, disaster relief management, military and civil surveillance, medical applications, nuclear hazard assessment, power network monitoring, spectrum sensing in cognitive radio networks and many others [1]–[7].

The existing decentralized strategies can be categorized into several groups according to their communication and data processing philosophy. The incremental strategies pass the shared information from one node to another in the Hamiltonian cycle [9]–[11]. Although this is relatively communication efficient, the robustness of these solutions is limited, because each node and link represents a single point of failure, and the recovery from a failure – a calculation of a new Hamiltonian cycle – is an NP-hard problem [12]. The diffusion strategies [13]–[26] and the consensus-based strategies [27]–[30] rely on networks with node degrees (the number of links incident to the node) mostly higher than one and offer much higher robustness, adaptivity and scalability. In the diffusion strategies, the shared information gradually diffuses through the whole network by local communication among adjacent nodes (neighbors) within one hop distance. Mostly, two communication steps are used: an adaptation step exchanging and incorporating neighbors' observations into the local statistical knowledge, and a combination step, that merges neighbors' estimates. The *consensus* strategies adopt a similar communication scheme but usually with multiple intermediate iterations increasing the communication requirements. Recognizing this as a potential drawback, several strategies alleviating the communication burden were proposed as well, e.g., the running consensus [31], or consensus + innovations algorithms where the local parameter estimates result from the combination of the neighbors' information and the locally sensed new information [32]–[34].

The diffusion strategies are mostly based on the least squares paradigm, in particular, the least mean squares (LMS) [19], [22] with numerous modifications, e.g., for sparse models [23], the recursive least-squares (RLS) [18] and its modified versions [13], [24], and the Kalman filters [17], [25]. Other diffusion algorithms consider, for instance, inference of mixtures [26], [35]. The common features of these algorithms are their single-problem orientation and independent derivation from the traditional methods. Recognizing this as a drawback, the first author recently proposed an abstract unifying approach extending his previous results [21], [36], rooted in the Bayesian paradigm and providing straightforward solution for a wide class of possible modeling tasks including the linear and logistic regression, Poisson process estimation etc., [20].

Recently, a focus has been given to the problem of collaborative estimation of parameters and signals that are inhomogeneous over the network. This becomes particularly useful in multiple target tracking, classification with multiple models, cooperative spectrum sensing in cognitive radio networks and other machine learning applications [7], [37], and in node-specific signal estimation applications where each node aims to estimate samples of a desired signal [8]. The consensus

Manuscript received December 30, 2016; revised May 22, 2017 and June 28, 2017; accepted June 28, 2017. Date of publication July 11, 2017; date of current version July 31, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Subhrakanti Dey. The work of K. Dedecius was supported by the Czech Science Foundation under Grant 14-06678P. The work of V. Sečkárová was supported by the Czech Science Foundation under Grant 16-09848S. *(Corresponding author: Kamil Dedecius.)*

The authors are with the Institute of Information Theory and Automation, Czech Academy of Sciences, Prague 18208, Czech Republic (e-mail: dedecius@utia.cas.cz; seckarov@utia.cas.cz).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2017.2725226

strategies allowing intermediate iterations for estimation of local and global parameters originally dominated this field, e.g. [38]. However, the class of diffusion solutions is rapidly growing. In [7] the authors reformulate their originally incremental LMS algorithm [39], providing node-specific estimation of parameters in networks with local and global parameters, and parameters shared by clusters of nodes. A diffusion LMS-based algorithm for estimation of similar parameters across the (so-called multitask) networks was proposed in [40], and extended to cases where the parameters are identical in node clusters and similar between neighboring clusters [41]. In [42], [43] a diffusion LMS-based algorithm is used to estimate time-varying parameters whose space-varying (node-specific) nature is characterized by a set of basis functions, such as sinusoids, B-splines or shifted Chebyshevs polynomials. With only a few exceptions, e.g. [37], [44], [45], the existing solutions assume the a priori knowledge of cluster memberships.

In this paper we extend our previous results published in a conference paper [46] devoted to static and sequential estimation of mixture parameters that are heterogeneous over the *non-clustered* network, hence either strictly local, or global. The present paper deals with the following important extensions: First, it focuses on virtually any types of models – not only mixtures – with emphasis on conditionally conjugate prior distributions. That is, unlike the referenced state-of-the-art algorithms, the proposed method is not restricted to any particular model type. Second, the parameters may be strictly local, global, or shared by a cluster of network nodes. It is not assumed that the nodes have any prior knowledge of the clusters. The clusters are formed per parameter and may differ for different parameters. The framework is formulated abstractly using the Bayesian and information-theoretic paradigms, which allows its application to a wide class of models, and besides parameter heterogeneity admits model heterogeneity. We adopt the variational approach to inference (see, e.g., [47], [48]) for its analytical tractability under exponentially distributed models, but other inferential techniques (e.g., the Markov chain Monte Carlo) may be used, too. Finally we remark that the present paper extends our earlier results [20] proposing a generic Bayesian diffusion framework suitable for estimation of *global*, i.e., homogeneous parameters.

The paper is organized as follows: Section II formulates the considered problem and discusses the departures from the ordinary diffusion strategies. Section III is devoted to the basic Bayesian estimation principles, that are extensively used in the rest of the text, particularly in Section IV proposing the diffusion algorithm for estimation under parameters inhomogeneity. The following Section V contains a theoretical application of the proposed procedures to a toy problem of collaborative estimation of possibly inhomogeneous normal model parameters. A simulation example is given in Section VI. Section VII concludes the work.

## II. Problem Statement

We consider a *diffusion* network [49] represented by an undirected or directed graph with a set of nodes $\mathcal{I} = \{1, \ldots, I\}$, interconnected by a set of edges determining the network topology. Every node $i \in \mathcal{I}$ may communicate exclusively with its neighbors within one hop distance. They and the node $i$ form a closed neighborhood $\mathcal{I}_i \subseteq \mathcal{I}$. There is only one information exchange between two adjacent nodes, either in a single direction if the edge is directed, or in both directions otherwise. This setting implies that the information may only gradually *diffuse* over the network as time proceeds.

Each node $i \in \mathcal{I}$ observes a (possibly local) discrete-time random process $\{Y_{i,t}; t = 1, 2, \ldots\}$ and acquires its outcomes – noisy observations $y_{i,t}$, possibly determined by known explanatory variables $x_{i,t}$, e.g. the regressors. The nodes model $y_{i,t}$ given $x_{i,t}$ (if present) using a user-defined local parametric statistical model – a probability density function

$$p_i(y_{i,t}|x_{i,t}, \Theta_i), \qquad \Theta_i \subseteq \Theta, \tag{1}$$

where the global parameter set $\Theta$ encompasses all parameters present in the network.

*Remark 1:* Strictly speaking, the Bayesian paradigm admits uncertainty about the parametric model $p_i(y_{i,t}|x_{i,t}, \Theta_i)$. That is, the true observations-generating model may be different, but the user's aim is to adopt a convenient model $p_i(y_{i,t}|x_{i,t}, \Theta_i)$ that is as close to the true model as possible. The model selection theory is very complex and far beyond the scope of this paper. More on it can be found, e.g., in the survey papers [50] and [51].

The role of the global parameter set $\Theta$ is purely conceptual, the nodes' interests – and inference objective – lie only in the elements that are present in their own models, hence $\Theta_i$. From the assumption of relationships among the modeled phenomena it follows that the sets $\Theta_i$ may overlap for different $i \in \mathcal{I}$. For instance, a global parameter may be contained in each $\Theta_i$, and/or there may be parameters common only to subsets of nodes. The aim of the nodes is to arrive at (in a sense) the best estimates of own parameters $\Theta_i$ by exploiting any information about them available from the network.

In this paper, we strictly adopt the perspective of individual nodes, whose knowledge of the network is limited to their closed neighborhoods only. This approach, corresponding to most realistic situations present in the nature, computer and social networks etc., leads to the following definition of the cluster:

*Definition 1 (Cluster):* Fix a node $i \in \mathcal{I}$. For any $\theta \in \Theta_i$ the cluster $\mathcal{I}_i^\theta$ is a set of nodes $j \in \mathcal{I}_i$ such that $\theta \in \Theta_j$.

Fig. 1 provides a graphical explanation of the definition. We remark that the following cases may occur:

- A parameter $\theta \in \Theta$ is global, i.e., common to all nodes $i \in \mathcal{I}$. Then, $\mathcal{I}_i^\theta = \mathcal{I}_i$ for all $i \in \mathcal{I}$.
- A parameter $\theta \in \Theta_i$ is strictly local to $i \in \mathcal{I}$. Then $\mathcal{I}_i^\theta$ is a singleton $\{i\}$.

Two clusters $\mathcal{I}_i^\theta$ and $\mathcal{I}_i^{\theta'}$ naturally may be identical for two different $\theta, \theta' \in \Theta_i$. Since this paper does not impose any limitations regarding the parameters/clusters configuration, it covers virtually any type of clustered multitask diffusion networks [41].

## III. Preliminaries on Bayesian Estimation

This section briefly reviews the basic principles of the Bayesian parameter inference that will be extensively exploited

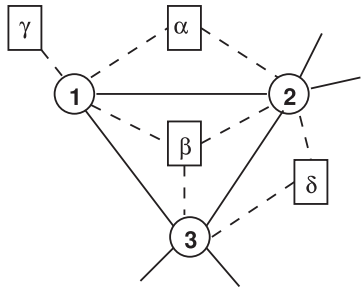


Fig. 1. Example of a network consisting of three nodes $\mathcal{I} = \{1, 2, 3\}$ with parameters $\Theta = \{\alpha, \beta, \gamma, \delta\}$. There are three clusters identified in node 1: $\mathcal{I}_1^{\alpha} = \{1, 2\}, \mathcal{I}_1^{\beta} = \{1, 2, 3\}$ and $\mathcal{I}_1^{\gamma} = \{1\}$. The set $\Theta_1 = \{\alpha, \beta, \gamma\}$.

in the rest of the paper. Let us focus on a single node, say $i \in \mathcal{I}$. The ordinary noncollaborative Bayesian estimation of unknown parameters $\Theta_i$ from the observed variables $y_{i,t}$ and the explanatory variables $x_{i,t}$ proceeds with a joint prior distribution $\pi_i\left(\Theta_i | x_{i,0:t-1}, y_{i,0:t-1}\right)$ where

$$x_{i,0:t-1} = \{x_{i,0}, \ldots, x_{i,t-1}\},$$
$$y_{i,0:t-1} = \{y_{i,0}, \ldots, y_{i,t-1}\}$$

denote the past data. Specifically, $x_{i,0}$ and $y_{i,0}$ stand for pseudo-observations carried by the initial prior distribution, and summarizing the initial knowledge obtained, e.g., from historical data or an expert's opinion. The prior distribution is sequentially updated by acquired $y_{i,t}$ and $x_{i,t}$ by virtue of the Bayes' theorem,

$$\pi_i\left(\Theta_i | x_{i,0:t}, y_{i,0:t}\right) \propto p_i(y_{i,t} | x_{i,t}, \Theta_i) \pi_i\left(\Theta_i | x_{i,0:t-1}, y_{i,0:t-1}\right) \tag{2}$$

where $\propto$ stands for proportionality, i.e., equality up to a normalizing constant.

The analytical tractability of the posterior distribution under rigorous Bayesian approach (2) is practically limited to models belonging to the exponential family. If the model $p_i(y_{i,t} | x_{i,t}, \Theta_i)$ belongs to the exponential family of distributions, and a conjugate prior distribution is used for parameter estimation, the posterior distribution $\pi_i(\Theta_i | \cdot)$ is analytically tractable [52]. The corresponding definitions are as follows:

*Definition 2 (Exponential family distribution):* An exponential family distribution of a random variable $y_{i,t}$ conditioned on $x_{i,t}$ and with a parameter $\Theta_i$ is a distribution whose probability density function can be written in the form

$$p(y_{i,t} | x_{i,t}, \Theta_i) = \exp\left\{\eta_i^{\intercal} T_{i,t} - A(\eta_i) + k(x_{i,t}, y_{i,t})\right\} \tag{3}$$

where $\eta_i = \eta(\Theta_i)$ is the natural parameter of the exponential family distribution, $T_{i,t} \equiv T(x_{i,t}, y_{i,t})$ is a sufficient statistic encompassing all information provided by $x_{i,t}$ and $y_{i,t}$ about the parameter $\Theta_i$, and

$$A(\eta_i) = \log \int \exp\left\{\eta_i^{\intercal} T_{i,t} + k(x_{i,t}, y_{i,t})\right\} \mathrm{d}x_{i,t} \mathrm{d}y_{i,t}$$

where the integral is over the space of $x_{i,t}$ and $y_{i,t}$, and $A(\eta_i)$ is a known log-partition function normalizing the density, and $k(x_{i,t}, y_{i,t})$ is a known function independent of the parameter.

The exponential family encompasses many important distributions, e.g., the normal, Poisson, multinomial, gamma, or the categorical distribution. The form (3) is not unique, as the natural parameter and the sufficient statistic may be multiplied by any constant and its reciprocal, respectively.

*Definition 3 (Conjugate prior distribution for $\Theta_i$):* Assume that a random variable $y_{i,t}$ is conditioned by a variable $x_{i,t}$ and obeys an exponential family distribution — model — with a parameter $\Theta_i$. A prior distribution for $\Theta_i$ conjugate to the model is characterized by conjugate hyperparameters $\Xi_{i,t-1}$ and $\Upsilon_{i,t-1}$ and has the probability density of the form

$$\pi_i(\Theta_i | \Xi_{i,t-1}, \Upsilon_{i,t-1}) = \exp\left\{\eta_i^{\intercal} \Xi_{i,t-1} + \Upsilon_{i,t-1} A(\eta_i) + l(\Xi_{i,t-1}, \Upsilon_{i,t-1})\right\}, \tag{4}$$

where $\Xi_{i,t-1}$ has the same size as $T_{i,t-1}$, $\Upsilon_{i,t-1} \in \mathbb{R}_+$, and $l(\Xi_{i,t-1}, \Upsilon_{i,t-1})$ is a known function.

We remind that the parameterizations of the probability densities are not unique in general, but there always exist bijective mappings among them. For instance, the univariate normal distribution may be parameterized by a mean and either a variance or precision. Therefore, to avoid misunderstanding, the hyperparameters of the conjugate prior density compatible with the model in the exponential family form (3) will be called *conjugate hyperparameters*.

The use of conjugate prior distribution reduces the Bayesian update (2) to two simple summations, as shows the next lemma.

*Lemma 1 (Bayesian update under conjugacy):* Assume that a random variable $y_{i,t}$ conditioned by a variable $x_{i,t}$ has an exponential family distribution with a parameter $\Theta_i$, estimated by means of a conjugate prior distribution. The Bayesian update (2) is then equivalent to the update of conjugate hyperparameters

$$\Xi_{i,t} = \Xi_{i,t-1} + T_{i,t},$$
$$\Upsilon_{i,t} = \Upsilon_{i,t-1} + 1. \tag{5}$$

The proof of Lemma 1 is trivial, as the Bayesian update is a multiplication of the model probability density (3) and the prior density (4), followed by a normalization ensuring that the result is again a probability density [52]. We point out that it is possible to perform the update (5) on a batch of data of a positive length $\tau \leq t$, summarized by $T_{\tau} + \ldots + T_{i,t}$, and with a conjugate prior $\pi_i(\Theta_i | x_{i,0:\tau-1}, y_{i,0:\tau-1})$.

A conjugate prior distribution for the whole $\Theta_i$ is rather an exception than a rule. Still, there are many cases where a conjugate prior distribution exist for certain elements of $\Theta_i$ provided that the rest is treated as known, e.g., by plugging their point estimates into the model. The ongoing section deals with this situation and sheds light on its use in distributed inference.

## IV. Collaborative Estimation by Diffusion

The heterogeneity of parameters $\Theta_i$ across the network imposes the requirement of *information separation*: each element of $\Theta_i$ has to be interpreted independently and with a probabilistic description compatible over the network, i.e., the same functional type of probability distribution. The separation prevents any inadvertent impairment of estimates $\Theta_i \setminus \{\theta\}$ when

collaborating on the estimate of $\theta \in \Theta_i$. The requirement of the same functional type of probability distribution guarantees straightforwardness of information merging. Both assumptions can be relaxed in certain cases. For instance, the normal mean vector and covariance matrix can be estimated together using a normal-inverse-Wishart distribution if their homogeneity in a cluster is assumed; different probability distributions for a particular parameter can be approximated by a common type of distribution, etc. However, we leave this beyond the scope of this paper.

We note that the factorized estimation that we adopt here is frequently used in inference for computational reasons, e.g., in connection with the Markov chain Monte Carlo methods, where working with the joint distribution of all parameters would lead to enormous computational demands. In our case, the factorization provides easy and theoretically justified rules for combination of posterior estimates.

Let us now formulate the proposed framework. It will consist of two steps:

1) *Factorized local estimation* – we establish the factorized local estimation, evaluating the posterior distribution from the batch of data and the local prior information.
2) *Diffusion optimization* – an alternative of the diffusion combination step [49], where the information about $\theta \in \Theta_i$ is combined.

### A. Factorized Local Estimation

During the factorized local estimation step, the true but inconvenient and often intractable posterior distribution $\pi_i(\Theta_i|\cdot)$ is approximated by a factorized distribution

$$\rho_{i,t}(\Theta_i) = \prod_{\theta \in \Theta_i} \rho_{i,t}^{\theta}(\theta), \tag{6}$$

that is as close to $\pi_i(\Theta_i|\cdot)$ in the Kullback-Leibler divergence sense as possible. That is, we aim to minimize

$$\begin{aligned}\mathcal{D}\left[\rho_{i,t}(\Theta_i)\middle|\middle|\pi_i(\Theta_i|\cdot)\right] &= \mathbb{E}_{\rho_{i,t}(\Theta_i)}\left[\log\frac{\rho_{i,t}(\Theta_i)}{\pi_i(\Theta_i|\cdot)}\right] \\ &= \log p_i(y_{i,1:t}|x_{i,1:t}) \\ &\quad + \underbrace{\mathbb{E}_{\rho_{i,t}(\Theta_i)}\left[\log\frac{\rho_{i,t}(\Theta_i)}{p_i(y_{i,1:t},\Theta_i|x_{i,1:t})}\right]}_{-\mathcal{L}[\rho_{i,t}(\Theta_i)]}.\end{aligned} \tag{7}$$

The term $\log p_i(y_{i,1:t}|x_{i,1:t})$ is the logarithm of the posterior predictive distribution

$$p_i(y_{i,1:t}|x_{i,1:t}) = \int p_i(y_{i,1:t}|x_{i,1:t},\Theta_i)\pi_i(\Theta_i|x_0,y_0)\mathrm{d}\Theta_i,$$

(the integral is over the space of $\Theta_i$), and $\mathcal{L}[\cdot]$ is called the free energy.

*Remark 2:* We remark that although our motivation is different, the final formulation and the criterion (7) coincides with the variational inference [47], [48]. Therefore we shall exploit its attractive properties in the proposed factorized local estimation step.

Since the log-evidence in (7) is fixed for observed data, the optimization of (7) is performed via minimization of the negative free energy $-\mathcal{L}[\rho_{i,t}(\Theta_i)]$. To this end, we rewrite the negative free energy in terms of a single element $\theta \in \Theta_i$,

$$\begin{aligned}&-\mathcal{L}\left[\rho_{i,t}\left(\Theta_i\right)\right] \\ &= \int \prod_{\theta\in\Theta_i}\rho_{i,t}^{\theta}(\theta)\sum_{\theta\in\Theta_i}\log\rho_{i,t}^{\theta}(\theta)\mathrm{d}\Theta_i \\ &- \int \prod_{\theta\in\Theta_i}\rho_{i,t}^{\theta}(\theta)\log p_i(y_{i,1:t},\Theta_i|x_{i,1:t})\mathrm{d}\Theta_i \\ &= \int\rho_{i,t}^{\theta}(\theta)\log\frac{\rho_{i,t}^{\theta}(\theta)\mathrm{d}\theta}{\exp\left\{\mathbb{E}_{\rho_{i,t}(\Theta_i\setminus\{\theta\})}\left[\log p(y_{i,1:t},\Theta_i|x_{i,1:t})\right]\right\}}+k \\ &= \mathcal{D}\left[\rho_{i,t}^{\theta}(\theta)\middle|\middle|\exp\left\{\mathbb{E}_{\rho_{i,t}(\Theta_i\setminus\{\theta\})}\left[\log p\left(y_{i,1:t},\Theta_i|x_{i,1:t}\right)\right]\right\}\right]+k\end{aligned} \tag{8}$$

where $k$ stands for a constant independent of $\theta$ and

$$\rho_{i,t}(\Theta_i\setminus\{\theta\}) = \prod_{\vartheta\in\Theta_i\setminus\{\theta\}}\rho_{i,t}^{\vartheta}(\vartheta)$$

is the product of distributions of all elements of $\Theta_i$ excluding $\theta$. Obviously, setting

$$\rho_{i,t}^{\theta}(\theta) \leftarrow \frac{\exp\left\{\mathbb{E}_{\rho_{i,t}(\Theta_i\setminus\{\theta\})}\left[\log p(y_{i,1:t},\Theta_i|x_{i,1:t})\right]\right\}}{\int\exp\left\{\mathbb{E}_{\rho_{i,t}(\Theta_i\setminus\{\theta\})}\left[\log p(y_{i,1:t},\Theta_i|x_{i,1:t})\right]\right\}\mathrm{d}\theta} \tag{9}$$

with the integral over the space of $\theta$ minimizes the Kullback-Leibler divergence (8) in the free energy under fixed $\theta \in \Theta_i$. It can be shown by means of the calculus of variations that cycling through each element $\theta \in \Theta_i$ and revising the corresponding estimates minimize the divergence between $\rho_{i,t}(\Theta_i)$ and $\pi_i(\Theta_i|\cdot)$. The convergence towards the target density is guaranteed by convexity of the free energy in each of the factors $\rho_{i,t}^{\theta}(\theta)$ [53]. However, the independence assumptions underlying the variational approximations may influence asymptotic properties of the estimator, see, e.g. [54], [55] and [56, Chap. 1.3.6] and the references therein.

A prominent case occurs if the observed variable $y_{i,t}$ obeys a model (distribution) belonging to the exponential family introduced by Definition 2. From the numerator in Equation (9) it is apparent that using the point estimates of the parameters in $\Theta_i\setminus\{\theta\}$ reduces the set of free parameters to $\theta$ only. If in this situation $\rho_{i,t}^{\theta}(\theta)$ is conjugate, then the factor can be updated via the Bayes' theorem (2) in terms of the factors' hyperparameters, (5) [57]. Another element of $\Theta_i$ is selected in the next iteration and the situation is repeated.

This algorithm can be used for sequential factorized local estimation with a batch of data of a fixed or variable size. The last posterior distribution $\rho_{i,t}(\Theta_i)$ serves as the prior for the next time instant as usually. Although the variational estimation of the posterior distribution may be inaccurate in the early stage of the online learning it gradually becomes accurate as learning proceeds [58]. Therefore, it can be practical to start with a higher number of local iterations to speed up this convergence, and to decrease it later down to one iteration between two diffusion steps.

### B. Diffusion Optimization

The aim of the diffusion optimization step is to combine the information provided by the posterior distributions $\rho_{j,t}(\Theta_j)$ of the neighboring nodes $j \in \mathcal{I}_i$. In particular, this means to acquire the factors $\rho^{\theta}_{j,t}(\theta)$ of the elements $\theta \in \Theta_i$ from the clusters $\mathcal{I}^{\theta}_i$. The assumptions adopted in the previous section guarantee that $\rho^{\theta}_{j,t}(\theta)$ have the same functional form and are fully characterized by the hyperparameters $\Xi^{\theta}_{j,t}$ and $\Upsilon^{\theta}_{j,t}$, see Definition 3.

From the $i$'s viewpoint, these distributions can be seen as hypotheses about the true value of the considered parameter $\theta$. In order to reflect the credibility of $j \in \mathcal{I}^{\theta}_i$, node $i$ employs nonnegative weights $c^{\theta}_{ij}$ summing to unity. A weight $c^{\theta}_{ij}$ may be interpreted as the level of belief of node $i$ in information about $\theta$ provided by neighboring node $j \in \mathcal{I}_i$. It may be based, e.g., on the amount of data, evidence, or reflect the reliability of the neighbor. Alternatively, uniform weights $c^{\theta}_{ij} = |\mathcal{I}_i|^{-1}$ can be used. There are several ways towards static and dynamic selection rules for $c^{\theta}_{ij}$, e.g., [20], [49], hence we leave the topic beyond the scope of the present paper.

Now assume that the node $i$ has acquired the factor set $\{\rho^{\theta}_{j,t}(\theta)\}_{j\in\mathcal{I}^{\theta}_i}$. Working with it would be impractical for memory and computational reasons. Instead, we reduce $\{\rho^{\theta}_{j,t}(\theta)\}_{j\in\mathcal{I}^{\theta}_i}$ to a single distribution $\widetilde{\rho}^{\theta}_{i,t}(\theta)$ that best represents the original set in the sense of a convenient information divergence measure. The Bayesian theory advocates the use of the Kullback-Leibler divergence, also known as the relative entropy, as this measure [59]. The following proposition stemming from our earlier result [20] for the case of globally-valid parameters is applied here to the case of parameters shared by a cluster of nodes.

*Proposition 1:* Fix a node $i \in \mathcal{I}$ and assume that it has acquired approximate posterior distributions $\rho^{\theta}_{j,t}(\theta)$ of a parameter $\theta \in \Theta_i$ shared by a cluster of its neighbors $\mathcal{I}^{\theta}_i \subseteq \mathcal{I}_i$. Furthermore assume that $\rho^{\theta}_{j,t}(\theta)$ have the same functional form and are therefore characterized by the conjugate hyperparameters $\Xi^{\theta}_{j,t}$ and $\Upsilon^{\theta}_{j,t}$ of compatible dimensions. Then, the distribution $\widetilde{\rho}^{\theta}_{i,t}(\theta)$ that best represents all $\rho^{\theta}_{j,t}(\theta), j \in \mathcal{I}_i$ in the sense of the minimum weighted arithmetic average of the Kullback-Leibler divergences $\mathcal{D}\left[\bar{\rho}^{\theta}_{i,t}(\theta)\middle|\middle|\rho^{\theta}_{j,t}(\theta)\right]$ with weights $c^{\theta}_{ij} \in [0,1]$ summing to unity is given by

$$\widetilde{\rho}_{i,\theta}(\theta) = \arg\min_{\bar{\rho}^{\theta}_{i,t}\in\mathcal{R}} \sum_{j\in\mathcal{I}^{\theta}_i} c^{\theta}_{ij}\mathcal{D}\left[\bar{\rho}^{\theta}_{i,t}(\theta)\middle|\middle|\rho^{\theta}_{j,t}(\theta)\right] \propto \prod_{j\in\mathcal{I}^{\theta}_i}\left[\rho^{\theta}_{j,t}(\theta)\right]^{c^{\theta}_{ij}}, \tag{10}$$

where $\mathcal{R}$ is the set of all admissible distributions with the same functional form as $\rho^{\theta}_{j,t}$. The optimal distribution $\widetilde{\rho}_{i,\theta}(\theta)$ is a (weighted) geometric mean of the neighbors' posterior distributions and inherits their functional form. Its conjugate hyperparameters are

$$\widetilde{\Xi}^{\theta}_{i,t} = \sum_{j\in\mathcal{I}^{\theta}_i} c^{\theta}_{ij}\Xi^{\theta}_{j,t} \qquad \widetilde{\Upsilon}^{\theta}_{i,t} = \sum_{j\in\mathcal{I}^{\theta}_i} c^{\theta}_{ij}\Upsilon^{\theta}_{j,t} \tag{11}$$

*Proof:* Using the definition of the Kullback-Leibler divergence we obtain

$$\sum_{j\in\mathcal{I}^{\theta}_i} c^{\theta}_{ij}\mathcal{D}\left[\bar{\rho}^{\theta}_{i,t}(\theta)\middle|\middle|\rho^{\theta}_{j,t}(\theta)\right] = \sum_{j\in\mathcal{I}^{\theta}_i} c^{\theta}_{ij}\mathbb{E}\left[\log\frac{\bar{\rho}^{\theta}_{i,t}(\theta)}{\rho^{\theta}_{j,t}(\theta)}\right]$$

$$= \mathbb{E}\left[\sum_{j\in\mathcal{I}^{\theta}_i} c^{\theta}_{ij}\log\frac{\bar{\rho}^{\theta}_{i,t}(\theta)}{\rho^{\theta}_{j,t}(\theta)}\right] = \mathbb{E}\left[\log\frac{\bar{\rho}^{\theta}_{i,t}(\theta)}{\prod_{j\in\mathcal{I}^{\theta}_i}\left[\rho^{\theta}_{j,t}(\theta)\right]^{c^{\theta}_{ij}}}\right]$$

$$= \mathcal{D}\left(\bar{\rho}^{\theta}_{i,t}(\theta)\middle|\middle|k\prod_{j\in\mathcal{I}^{\theta}_i}\left[\rho^{\theta}_{j,t}(\theta)\right]^{c^{\theta}_{ij}}\right),$$

where $k$ is a proportionality constant. Minimum of the Kullback-Leibler divergence is reached under equality of its arguments. That is, the resulting distribution of $\theta$ is a (weighted) geometric mean of the neighbors' posterior distributions, from which (11) follows. ■

### C. Determination of Compatible Neighbors

With only a few exceptions [14], [37] the state-of-the-art algorithms assume that the nodes have the full knowledge of clusters and parameter structures a priori. In [37], the authors propose a hypotheses testing procedure assessing whether two neighboring nodes belong to the same cluster. Their analysis adopts the assumption that the asymptotic normality of the normalized error sequence advocates the approximation of the marginal distribution of estimators by the normal distribution. This idea is valid from the large-sample behavior viewpoint, but may be inappropriate in relatively small-sample cases. In [14] an adjustment of combination weights that considers the closeness of the local estimate to the neighboring estimates and the local slope of the LMS-based cost function is proposed. The drawback of these solutions is their solely LMS-oriented design, limiting their applicability to many other tasks, e.g., estimation of covariance matrices.

Still, the underlying idea that the neighbors' estimates within an acceptable tolerance around own estimates may describe the same parameters is universal. Therefore, we suggest to exploit it in a conservative way. We construct a cluster of $i$'s neighbors $j \in \mathcal{I}_i$ whose estimates $\hat{\theta}_{j,t}$ lie within a predefined tolerance $h^{\theta} > 0$ around its own estimate $\hat{\theta}_{i,t}$. That is, they form a set

$$\mathcal{I}^{\theta}_i = \{j \in \mathcal{I}_i : d(\hat{\theta}_{j,t}, \hat{\theta}_{i,t}) \leq h^{\theta}\},$$

where $d(\cdot,\cdot)$ is a convenient metric.

For instance, the distance of the location parameters like the means of the normal distributions may be measured using a classical Euclidean 2-norm. A very appealing dissimilarity measure for the covariance matrices is the Jensen-Bregman LogDet divergence $d_{\mathrm{JBLD}}$ proposed in [67]. If $A$ and $B$ are two symmetric positive definite matrices, then

$$d_{\mathrm{JBLD}}(A,B) = \log\det\left(\frac{A+B}{2}\right) - \frac{1}{2}\log\det(AB).$$

Besides the computational simplicity (no matrix inversion), this divergence inherits the attractive properties of the symmetrized Bregman divergences [68]. In particular, it satisfies the

nonnegativity, symmetry and triangle inequality requirements imposed on metrics.[1]

This conservativeness has several advantages. First, the solution is very universal, and may be straightforwardly used with virtually any models and parameters – univariate or multivariate, location or scale parameters etc. Second, there usually exists an a priori available information about the (acceptable) closeness of estimated parameters. And even a very small tolerance does not compromise the estimator, as it only prevents the neighbors' information from incorporation.

A more elaborate approach inspired by [68] could consist of clustering of neighbors densities, e.g., using the K-means algorithm with suitable divergence measures. This is postponed to further research.

### D. Discussion of Estimator Properties

A remarkable property of Proposition 1 is its close relation with the Bayesian update (Lemma 1). This can be seen from comparison of Equations (5) and (11). Namely, Equation (11) combines neighbors' hyperparameters resulting from the Bayesian update (5), that contain their observations $y_{j,\cdot}$ and possibly explanatory variables $x_{j,\cdot}$ accumulated in the sufficient statistics $\Xi^{\theta}_{j,\cdot}$ and $\Upsilon^{\theta}_{j,\cdot}$. That is, Equation (11) is effectively a *weighted Bayesian update*, studied, e.g., in [62], [63]. This becomes evident particularly if we consider identical prior hyperparameters. Here, the weights $c^{\theta}_{ij}$ suppress knowledge duplication, also known as the data incest [64].

The asymptotic properties, discussed in the previous paper [20] focused on estimation of models with global parameters, are valid for the estimation of local, global and shared parameters too. In particular, the consistency of Bayesian estimators [60], [61] guarantees the consistency of the diffusion estimator under very mild conditions, and the diffusion estimators asymptotically converge to the centralized estimator. However, a difficulty is connected with the factorization and free-energy-based optimization. The relative accuracy of variational inference is still not fully understood. It is known, that variational inference generally underestimates the variance of the posterior density, which is a consequence of its objective function. Still, this is acceptable in many practical tasks [55], [65]. Until recently, the analyses of the variational methods were highly dependent on the concrete algorithm for the specific model and the choice of the prior distribution, and considered the direct minimization of the variational free energy over the expected sufficient statistics. In order to overcome this drawback, an alternative approach inspired by local variational approximations was adopted in [66]. The resulting analyses are more general and independent of the concrete algorithms for the specific models, however, they are quite technical.

## V. Illustrative Example

This example demonstrates the application straightforwardness of the proposed method. Let us for simplicity assume two connected nodes $i \in \mathcal{I} = \{1, 2\}$ that model own normally distributed observations $y_{i,t} \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ with probability density functions

$$p\left(y_{i,t}|\mu_i,\sigma_i^2\right) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left\{-\frac{1}{2\sigma_i^2}(y_{i,t}-\mu_i)^2\right\} = \exp\left\{\begin{bmatrix}\frac{\mu_i}{\sigma_i^2}\\ -\frac{1}{2\sigma_i^2}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}y_{i,t}\\ y_{i,t}^2\end{bmatrix} - \frac{\mu_i^2}{2\sigma_i^2} - \frac{1}{2}\log\left(2\pi\sigma_i^2\right)\right\}.$$

Let us summarize the window of observations $y_{i,1}, \ldots, y_{i,t}$ by a joint density and rewrite it into the forms useful for the factorized local estimation, separating the individual parameters into a natural statistic vectors (c.f. Definition 2):

$$p\left(y_{i,1:t}|\mu_i,\sigma_i^2\right) = \prod_{\tau=1}^{t} p(y_{i,\tau}|\mu_i,\sigma_i^2)$$

$$= \prod_{\tau=1}^{t}\exp\left\{\begin{bmatrix}\frac{\mu_i}{\sigma_i^2}\\ -\frac{1}{2\sigma_i^2}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}y_{i,\tau}\\ y_{i,\tau}^2\end{bmatrix} - \frac{\mu_i^2}{2\sigma_i^2} - \frac{1}{2}\log\left(2\pi\sigma_i^2\right)\right\} \quad (12)$$

$$= \exp\left\{\sum_{\tau=1}^{t}\left(\begin{bmatrix}\frac{\mu_i}{\sigma_i^2}\\ -\frac{1}{2\sigma_i^2}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}y_{i,\tau}\\ y_{i,\tau}^2\end{bmatrix}\right) - \frac{t}{2\sigma_i^2}\mu_i^2 - \frac{t}{2}\log\left(2\pi\sigma_i^2\right)\right\}$$

$$= \exp\left\{\begin{bmatrix}\frac{\mu_i}{\sigma_i^2}\\ -\frac{1}{2\sigma_i^2}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\sum_{\tau=1}^{t} y_{i,\tau}\\ \sum_{\tau=1}^{t} y_{i,\tau}^2\end{bmatrix} + K_1\right\} \quad (13)$$

$$= \exp\left\{\begin{bmatrix}\mu_i\\ \mu_i^2\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\frac{1}{\sigma_i^2}\sum_{\tau=1}^{t} y_{i,\tau}\\ -\frac{t}{2\sigma_i^2}\end{bmatrix} + K_2\right\} \quad (14)$$

$$= \exp\left\{\begin{bmatrix}\frac{1}{\sigma_i^2}\\ \log\sigma_i^2\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}-\frac{1}{2}\sum_{\tau=1}^{t}(y_{i,\tau}-\mu_i)^2\\ -\frac{t}{2}\end{bmatrix} + K_3\right\} \quad (15)$$

where (12) is a product of the normal densities, and

$$K_1 = -\frac{t}{2\sigma_i^2}\mu_i^2 - \frac{t}{2}\log\left(2\pi\sigma_i^2\right),$$

$$K_2 = -\frac{1}{2\sigma_i^2}\sum_{\tau=1}^{t} y_{i,\tau}^2 - \frac{t}{2}\log\left(2\pi\sigma_i^2\right),$$

$$K_3 = -\frac{t}{2}\log\left(2\pi\right).$$

A careful investigation of these forms reveals the possibility to employ the normal prior distribution $\mathcal{N}(m_{i,t-1}, s^2_{i,t-1})$ conjugate to the model (14) provided $\sigma_i^2$ is replaced by its point estimate. That is,

$$\pi^{\mu_i}_{i,t-1}(\mu_i|y_{i,t-1}) = \pi^{\mu_i}_{i,t-1}(\mu_i|m_{i,t-1}, s^2_{i,t-1}) = \frac{1}{\sqrt{2\pi s^2_{i,t-1}}}\exp\left\{-\frac{1}{2s^2_{i,t-1}}(\mu_i - m_{i,t-1})^2\right\}$$

$$= \exp\left\{\begin{bmatrix}\mu_i\\ \mu_i^2\end{bmatrix}^{\mathsf{T}}\underbrace{\begin{bmatrix}\frac{m_{i,t-1}}{s^2_{i,t-1}}\\ -\frac{1}{2s^2_{i,t-1}}\end{bmatrix}}_{\Xi^{\mu_i}_{i,t-1}} + \cdots\right\},$$

[1] The Euclidean 2-norm and the Jensen-Bregman LogDet divergence are used in the simulation examples in Section VI.

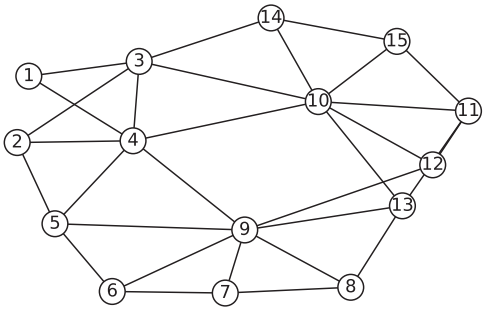


Fig. 2. Topology of network consisting of 15 nodes.

**Algorithm 1:** Factorized estimation of heterogeneous parameters in diffusion networks.

The nodes $i = 1, \dots, I$ employing user-defined models $p_i(y_{i,t}|x_{i,t}, \Theta_i)$ are initialized with the factorized prior densities $\rho_{i,0}(\Theta_i) = \prod_{\theta \in \Theta_i} \rho_{i,0}^{\theta}(\theta)$ reflecting the initial knowledge of $\Theta_i$. Set the number of variational iterations $n \in \mathbb{N}$. Set the tolerances $h^{\theta}$. For $t = 1, 2, \dots$ and each node $i$ do:

*Factorized local estimation:*

1) Get observations $y_{i,t}$ and (if present) explanatory variables $x_{i,t}$.
2) Perform $n$ variational iterations, Equation (9).

*Diffusion optimization:*

For each $\theta \in \Theta_i$:

1) Get the posterior densities $\rho_{j,t}^{\theta}(\theta)$ from $j \in \mathcal{I}_i^{\theta}$ and extract their conjugate hyperparameters $\Xi_{j,t}^{\theta}$ and $\Upsilon_{j,t}^{\theta}$.
2) Compute the combined posterior hyperparameters $\widetilde{\Xi}_{i,t}^{\theta}$ and $\widetilde{\Upsilon}_{i,t}^{\theta}$ according to Proposition 1, Equation (11).

Set the resulting posterior densities as the prior densities for the next time step.

where $\Xi_{i,t-1}^{\mu_i}$ is the conjugate hyperparameter, see Definition 3 and the remark below it. The expected value serving as the point estimate is $\widehat{\mu_i} = m_{i,t-1}$.

Analogously, it is possible to identify the inverse-gamma distribution $i\mathcal{G}(a_{i,t-1}, b_{i,t-1})$ as the conjugate prior distribution for the estimation of $\sigma_i^2$ with $\mu_i$ replaced by its point estimate in (15),

$$\begin{aligned}\pi_{i,t-1}^{\sigma_i^2}\left(\sigma_i^2|y_{i,t-1}\right) &= \pi_{i,t-1}^{\sigma_i^2}\left(\sigma_i^2|a_{i,t-1}, b_{i,t-1}\right)\\ &= \frac{b_{i,t-1}^{a_{i,t-1}}}{\Gamma(a_{i,t-1})}\left(\sigma_i^2\right)^{-a_{i,t-1}-1}\exp\left\{-\frac{b_{i,t-1}}{\sigma_i^2}\right\}\\ &= \exp\left\{\begin{bmatrix}\frac{1}{\sigma_i^2}\\ \log\sigma_i^2\end{bmatrix}^{\mathsf{T}}\underbrace{\begin{bmatrix}-b_{i,t-1}\\ -a_{i,t-1}-1\end{bmatrix}}_{\Xi_{i,t-1}^{\sigma_i^2}} + \dots\right\},\end{aligned}$$

where $\Xi_{i,t-1}^{\sigma_i^2}$ is the conjugate hyperparameter of the prior inverse-gamma distribution for $\sigma_i^2$. The expected value serving as the point estimate is $\widehat{\sigma_i^2} = b_{i,t-1}/(a_{i,t-1} - 1)$.

With the knowledge of compatible forms of distributions, it is now possible to perform the factorized estimation of $\mu_i$ and $\sigma_i^2$ according to Algorithm 1. It can be summarized as follows:

**Factorized Local Estimation**

**At each node** $i$:

1) Acquire $y_{i,t}$. Initialize the prior distributions using the conjugate hyperparameters from the previous time step $t-1$.
2) Perform the preset number $n$ of iterations, replacing the unknown parameters $\mu_i$ and $\sigma_i^2$ in the sufficient statistics by their point estimates – mean values of the respective distributions, i.e., $\widehat{\mu_i}^{(n-1)}$ and $(\widehat{\sigma_i^2})^{(n-1)}$:

$$\Xi_{i,t}^{\mu_i(n)} \leftarrow \Xi_{i,t}^{\mu_i(n-1)} + \begin{bmatrix}\frac{1}{\left(\widehat{\sigma_i^2}\right)^{(n-1)}}\sum_{\tau=1}^{t} y_{i,\tau}\\ -\frac{t}{2\left(\widehat{\sigma_i^2}\right)^{(n-1)}}\end{bmatrix},$$

$$\Xi_{i,t}^{\sigma_i^2(n)} \leftarrow \Xi_{i,t}^{\sigma_i^2(n-1)} + \begin{bmatrix}-\frac{1}{2}\sum_{\tau=1}^{t}\left(y_{i,\tau} - \widehat{\mu}_i^{(n-1)}\right)^2\\ -\frac{t}{2}\end{bmatrix}.$$

3) Assign $\Xi_{i,t}^{\mu_i} \leftarrow \Xi_{i,t}^{\mu_i(n)}$ and $\Xi_{i,t}^{\sigma_i^2} \leftarrow \Xi_{i,t}^{\sigma_i^2(n)}$.

**Diffusion Optimization**

**At each node** $i$**:**

1) Acquire posterior conjugate hyperparameters $\Xi_{i,t}^{\mu_i}$ and/or $\Xi_{i,t}^{\sigma_i^2}$ from the neighbors $j \in \mathcal{I}_i$.
2) Perform Kullback-Leibler optimal merging of pdfs used for common parameters according to Proposition 1, Equation (11). If the mean values are common, i.e. $\mu_1 = \mu_2 = \mu$, then

$$\widetilde{\Xi}_{1,t}^{\mu} = c_{11}^{\mu}\Xi_{1,t}^{\mu} + c_{12}^{\mu}\Xi_{2,t}^{\mu}, \text{ and } \widetilde{\Xi}_{2,t}^{\mu} = c_{21}^{\mu}\Xi_{1,t}^{\mu} + c_{22}^{\mu}\Xi_{2,t}^{\mu}.$$

If the variances are common, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$\widetilde{\Xi}_{1,t}^{\sigma^2} = c_{11}^{\sigma^2}\Xi_{1,t}^{\sigma^2} + c_{12}^{\sigma^2}\Xi_{2,t}^{\sigma^2}, \text{ and } \widetilde{\Xi}_{2,t}^{\sigma^2} = c_{21}^{\sigma^2}\Xi_{1,t}^{\sigma^2} + c_{22}^{\sigma^2}\Xi_{2,t}^{\sigma^2}.$$

Note that $\Upsilon_{\cdot}^{\cdot}$ are absorbed into $\Xi_{\cdot}^{\cdot}$.

## VI. Simulation Examples

### A. Homogeneous Parameters

The purpose of the first example is to demonstrate that the proposed algorithm yields results convergent to its centralized counterpart, where all observations are processed in a single dedicated network node, and performs better than a non-cooperative scenario. The network consists of 15 nodes, its topology is depicted in Fig. 2.

In order to be able to compare the results, the parameters have to be identical for all network nodes (the inhomogeneous case is treated in the next example). The observations are generated from a three-component normal mixture model

$$y_{i,t} \sim \sum_{k=1}^{3} w_k \mathcal{N}(\mu_k, \Sigma_k), \tag{16}$$

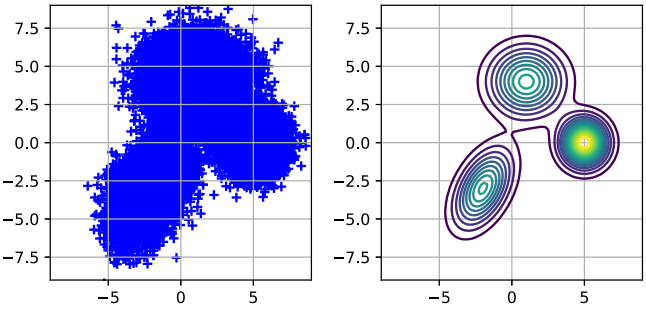

Fig. 3. Left: Samples drawn from the mixture (16). Right: Contour plot of the mixture components.

where $i \in \mathcal{I} = \{1, \ldots, 15\}$ is the node index, the component weights

$$[w_1, w_2, w_3] = [0.3, 0.3, 0.4],$$

and the normal components are

$$\mathcal{N}(\mu_1, \Sigma_1) = \mathcal{N}\left(\begin{bmatrix} -2 \\ -3 \end{bmatrix}, \begin{bmatrix} 1.5 & 0.9 \\ 0.9 & 2.5 \end{bmatrix}\right),$$

$$\mathcal{N}(\mu_2, \Sigma_2) = \mathcal{N}\left(\begin{bmatrix} 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right),$$

$$\mathcal{N}(\mu_3, \Sigma_3) = \mathcal{N}\left(\begin{bmatrix} 1 \\ 4 \end{bmatrix}, \begin{bmatrix} 2.5 & 0 \\ 0 & 2.0 \end{bmatrix}\right).$$

For each node of the network, 1500 samples are drawn from the mixture. The results below are averaged over 50 independent simulation runs. Fig. 3 depicts the samples of a randomly chosen run and the contours of the components. The nodes and the center start with identical prior distributions

$$\begin{aligned}
\mu_{1,i} &\sim \mathcal{N}([-5, -5]^\mathsf{T}, 10 \cdot I_{2\times 2}), \\
\mu_{2,i} &\sim \mathcal{N}([5, 0]^\mathsf{T}, 10 \cdot I_{2\times 2}), \\
\mu_{3,i} &\sim \mathcal{N}([5, 5]^\mathsf{T}, 10 \cdot I_{2\times 2}), \\
\Sigma_{1,i} &\sim i\mathcal{W}(0.1 \cdot I_{2\times 2}, 5), \\
\Sigma_{2,i} &\sim i\mathcal{W}(0.1 \cdot I_{2\times 2}, 5), \\
\Sigma_{3,i} &\sim i\mathcal{W}(0.1 \cdot I_{2\times 2}, 5), \\
w_i &\sim \mathcal{D}ir(1, 1, 1),
\end{aligned}$$

where $i\mathcal{W}$ and $\mathcal{D}ir$ denote the inverse-Wishart and the Dirichlet distributions, respectively. The combination weights $c_{ij}$ are uniform.

The inference exploits a floating window of 50 observations and starts after it is filled. The time is reset to $t = 1$ for comprehensibility. The center infers from $50|\mathcal{I}| = 750$ observations at each $t$. The tolerances are 1.0 both for the mean vectors and covariance matrices estimation (see Section IV-C). All three scenarios employ 15 local variational iterations each time step.

Figs. 4, 5, and 6 show the estimation performance in terms of the root mean squared errors (RMSE). In the diffusion (*diff*) and non-cooperative (*nocoop*) scenarios, the RMSE values are averaged over the network. The estimation of mean vectors $\mu_1, \mu_2,$ and $\mu_3$ is significantly better in the diffusion and centralized (*central*) scenarios. The results of the proposed diffusion algorithm are close to the centralized scenario. The estimation performance of the component weights vector $w = [w_1, w_2, w_3]$ is similar in the non-cooperative and diffusion algorithms, as there is no collaboration among nodes for this variable. Naturally, the centralized algorithm estimates $w$ best for its significantly larger amount of data.

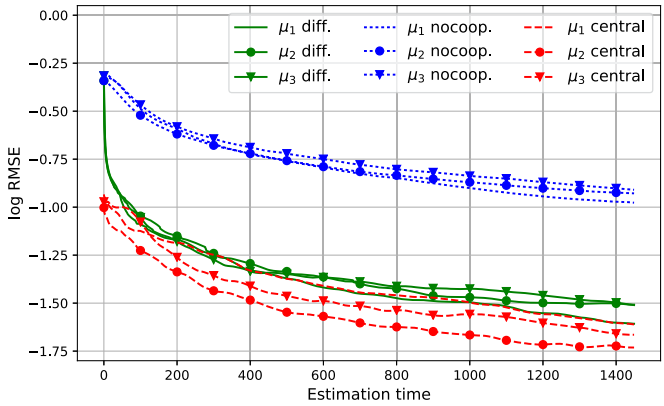


Fig. 4. Decimal logarithm of RMSE of the mean vectors estimates.

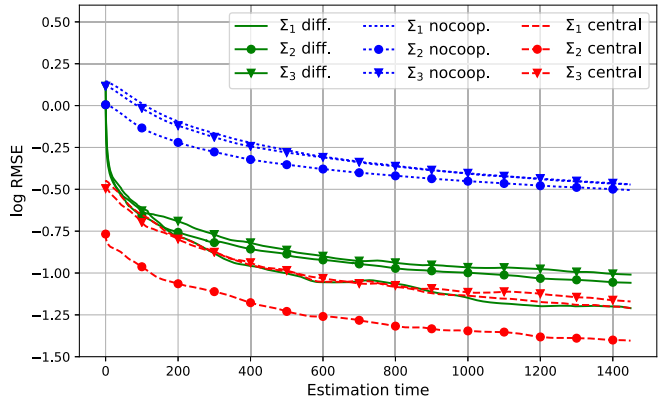


Fig. 5. Decimal logarithm of RMSE of the covariance matrices estimates.

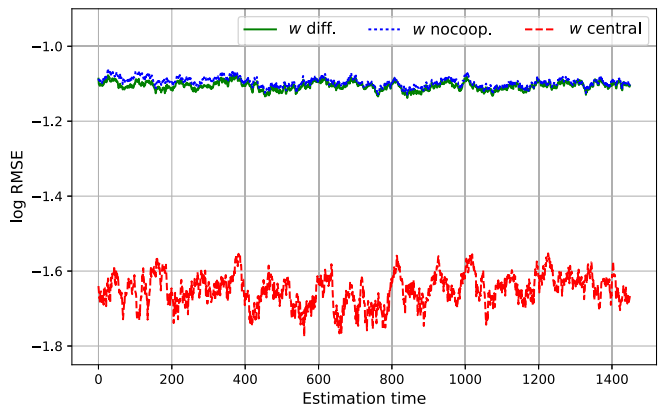


Fig. 6. Decimal logarithm of RMSE of the component weights vector estimates.

To summarize, the proposed diffusion algorithm provides estimation performance between the centralized approach and the non-cooperative scenario as expected. The collaboration among nodes significantly improves estimation of common parameters.

### B. Heterogeneous Normal Mixtures

This example demonstrates the diffusion estimation of a normal mixture model with heterogeneous parameters across the network. The network has the same topology as in the previous example (Fig. 2). The observations $y_{i,t}$ are generated from two-component normal mixture models of the form

$$y_{i,t} \sim w_{1,i}\mathcal{N}(\mu_1, \Sigma_{1,i}) + w_{2,i}\mathcal{N}(\mu_{2,i}, \Sigma_{2,i})$$

where $w_{1,i} + w_{2,i} = 1$. The first component mean $\mu_1 = [-10, -10]^\intercal$ is global for the whole network. The second component mean $\mu_{2,i}$ is randomly drawn from the set $\{[-2, 2]^\intercal, [2, 2]^\intercal\}$. Similarly, the covariance matrices are independently drawn from the set $\{\Sigma_a, \Sigma_b\}$, whose members are initially randomly drawn from the Wishart distributions:

$$\Sigma_a \sim \mathcal{W}\left(\begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}, 500\right),$$

$$\Sigma_b \sim \mathcal{W}\left(\begin{bmatrix} 0.03 & 0 \\ 0 & 0.03 \end{bmatrix}, 500\right).$$

The component weights $w_i = [w_{1,i}, w_{2,i}]$ are randomly sampled from the Dirichlet (or equivalently beta or uniform) distribution

$$[w_{1,i}, w_{2,i}] \sim \mathcal{D}ir(100, 100).$$

To summarize, there is one global parameter $\mu_1$, a set of strictly local parameters $w_i$, and a set of potentially shared parameters $\mu_{2,i}, \Sigma_{1,i}$ and $\Sigma_{2,i}$. The nodes are not a priori aware which parameter do they share with their neighbors. The results below are averaged over 50 independent simulation runs, each exploits 1500 data samples per node. The diffusion algorithm employs 3 variational iterations each time step, while the no-cooperative scenario performs with 10 iterations in order to demonstrate the superior performance of the proposed method. The combination weights $c_{ij}$ are uniform, the tolerances are 1.0 for the mean vectors and 0.05 for the covariance matrices estimation.

The estimation proceeds with the prior distributions

$$\begin{aligned}
\mu_{1,i} &\sim \mathcal{N}([-15, -15]^\intercal, 10 \cdot I_{2\times 2}), \\
\mu_{2,i} &\sim \mathcal{N}([0, 0]^\intercal, 10 \cdot I_{2\times 2}), \\
\Sigma_{1,i} &\sim i\mathcal{W}(0.1 \cdot I_{2\times 2}, 5), \\
\Sigma_{2,i} &\sim i\mathcal{W}(0.1 \cdot I_{2\times 2}, 5), \\
w_i &\sim \mathcal{D}ir(1, 1),
\end{aligned}$$

where $i\mathcal{W}$ and $\mathcal{D}ir$ denote the inverse-Wishart and the Dirichlet distributions, respectively. The data window for online estimation is 50 observations.

Unfortunately, the authors are not aware of any method suitable for estimation under so complex heterogeneity conditions. This prevented them from including a comparative study.

Figures 7, 8, and 9 depict the RMSE values averaged over the network for all estimated parameters. The proposed algorithm provides more precise estimates of the global and shared parameters, and a slight improvement in estimation quality of the component weights $w_i = [w_{i,1}, w_{i,2}]$. Moreover, this improvement is achieved with a lower computational requirements: 3 versus 10 local variational iterations used by the diffusion and no-cooperation algorithms, respectively. The final estimates of three selected nodes resulting from a randomly chosen experiment run are given in Table I.

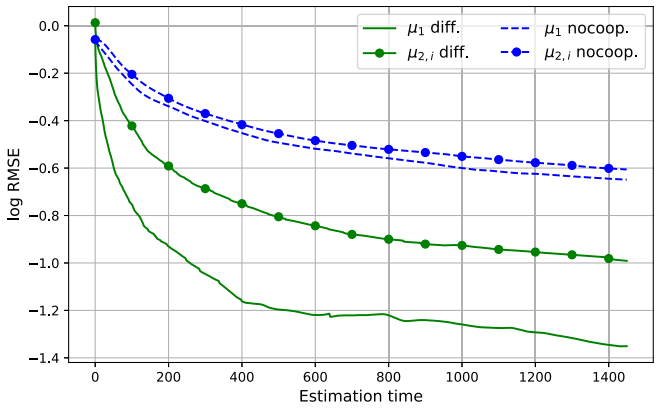


Fig. 7. Decimal logarithm of RMSE of the mean vector estimates.

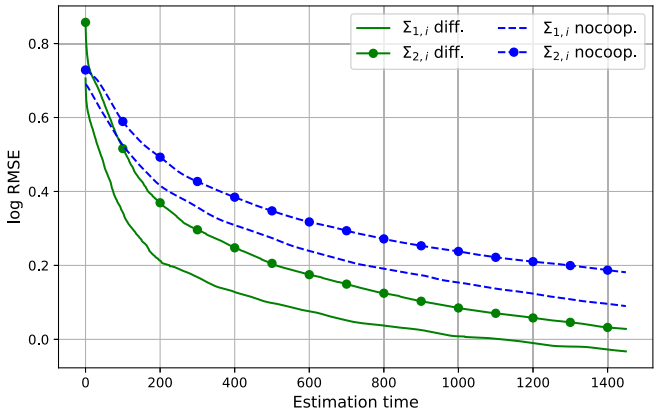


Fig. 8. Decimal logarithm of RMSE of the covariance matrices estimates.

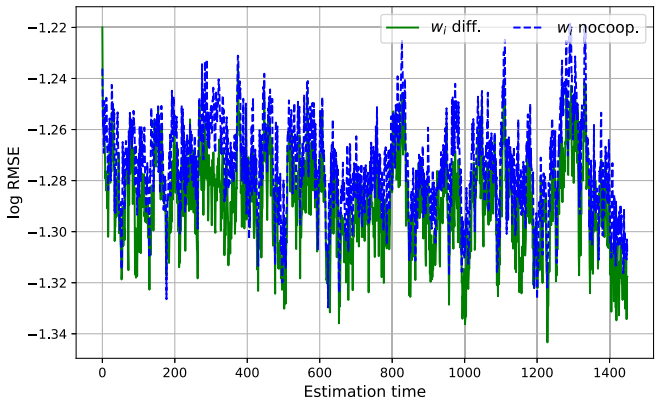


Fig. 9. Decimal logarithm of RMSE of the component weights estimates.

TABLE I
TRUE PARAMETER VALUES AND THEIR FINAL ESTIMATES AT NODES 1, 6 AND 9 FOR BOTH THE DIFFUSION AND THE NON-COOPERATIVE SCENARIOS, RESPECTIVELY. A RANDOMLY CHOSEN EXPERIMENT RUN

| | True | Diffusion | No cooperation |
|---|---|---|---|
| $\mu_{1,1}$ | [-10, -10] | [-9.972, -10.002] | [-10.091, -10.201] |
| $\mu_{2,1}$ | [-2, -2] | [-2.079, -1.989] | [-2.138, -2.035] |
| $\Sigma_{1,1}$ | $\begin{bmatrix} 5.220 & -0.095 \\ -0.095 & 4.897 \end{bmatrix}$ | $\begin{bmatrix} 5.083 & -0.072 \\ -0.073 & 5.084 \end{bmatrix}$ | $\begin{bmatrix} 5.131 & -0.467 \\ -0.467 & 4.805 \end{bmatrix}$ |
| $\Sigma_{2,1}$ | $\begin{bmatrix} 15.672 & 1.068 \\ 1.068 & 15.469 \end{bmatrix}$ | $\begin{bmatrix} 15.629 & 1.595 \\ 1.595 & 15.802 \end{bmatrix}$ | $\begin{bmatrix} 16.650 & 2.256 \\ 2.256 & 16.375 \end{bmatrix}$ |
| $w_1$ | [0.546, 0.454] | [0.554, 0.446] | [0.538, 0.461] |
| $\mu_{1,6}$ | [-10, -10] | [-9.972, -10.002] | [-10.092, -10.201] |
| $\mu_{2,6}$ | [-2, -2] | [-2.077, -1.987] | [-2.816, -2.805] |
| $\Sigma_{1,6}$ | $\begin{bmatrix} 15.672 & 1.068 \\ 1.068 & 15.469 \end{bmatrix}$ | $\begin{bmatrix} 16.314 & 0.948 \\ 0.948 & 17.604 \end{bmatrix}$ | $\begin{bmatrix} 12.767 & -1.769 \\ -1.769 & 14.587 \end{bmatrix}$ |
| $\Sigma_{2,6}$ | $\begin{bmatrix} 15.672 & 1.068 \\ 1.068 & 15.469 \end{bmatrix}$ | $\begin{bmatrix} 14.782 & 0.685 \\ 0.685 & 15.586 \end{bmatrix}$ | $\begin{bmatrix} 16.959 & 2.752 \\ 2.752 & 18.51 \end{bmatrix}$ |
| $w_6$ | [0.459, 0.541] | [0.431, 0.569] | [0.374, 0.626] |
| $\mu_{1,9}$ | [-10, -10] | [-9.972, -10.003] | [-9.92, -10.15] |
| $\mu_{2,9}$ | [2, 2] | [2.048, 1.99 ] | [2.001, 2.167] |
| $\Sigma_{1,9}$ | $\begin{bmatrix} 5.22 & -0.095 \\ -0.095 & 4.897 \end{bmatrix}$ | $\begin{bmatrix} 5.072 & -0.071 \\ -0.071 & 5.087 \end{bmatrix}$ | $\begin{bmatrix} 5.245 & -0.019 \\ -0.019 & 5.425 \end{bmatrix}$ |
| $\Sigma_{2,9}$ | $\begin{bmatrix} 5.22 & -0.095 \\ -0.095 & 4.897 \end{bmatrix}$ | $\begin{bmatrix} 5.355 & -0.158 \\ -0.158 & 4.527 \end{bmatrix}$ | $\begin{bmatrix} 5.201 & -0.081 \\ -0.081 & 4.676 \end{bmatrix}$ |
| $w_9$ | [0.489, 0.511] | [0.574, 0.426] | [0.574, 0.426] |

## VII. CONCLUSION

In this paper we developed a distributed framework for static and sequential estimation of parameters that are heterogeneous over the network. It is formulated in the Bayesian and information-theoretic realm, and allows its relatively straightforward application to a wider class of models from the exponential family. The future work should focus on possible accelerations of factorized local estimation via intermediate cooperations among neighbors, optimal communication scheduling, extensions to arbitrary models and prior distributions (e.g. [69], [70]), and investigation of yet more complex situations of intrinsically incompatible information and heterogeneous inferential approaches.

## REFERENCES


[1] I. Akyildiz and M. C. Vuran *Wireless Sensor Networks*. Hoboken, NJ, USA: Wiley, 2010.

[2] H. Chen, C. K. Tse, and J. Feng, "Impact of topology on performance and energy efficiency in wireless sensor networks for source extraction," *IEEE Trans. Parall. Distrib. Syst.*, vol. 20, no. 6, pp. 886–897, Jun. 2009.

[3] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

[4] J. Liu, M. Chu, and J. E. Reich, "Multitarget tracking in distributed sensor networks," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 36–46, May 2007.

[5] R. Olfati-Saber, J. Alex Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proc. IEEE*, vol. 95, no. 1, pp. 215–233, Jan. 2007.

[6] F. Pasqualetti, R. Carli, and F. Bullo, "Distributed estimation via iterative projections with application to power network monitoring," *Automatica*, vol. 48, no. 5, pp. 747–758, 2012.

[7] J. Plata-Chaves, N. Bogdanović, and K. Berberidis, "Distributed diffusion-based LMS for node-specific adaptive parameter estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3448–3460, Jul. 2015.

[8] A. Hassani, A. Bertrand, and M. Moonen, "GEVD-Based low-rank approximation for distributed adaptive node-specific signal estimation in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2557–2572, May 2016.

[9] C. G. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4064–4077, Aug. 2007.

[10] D. P. Bertsekas, "A new class of incremental gradient methods for least squares problems," *SIAM J. Optim.*, vol. 7, no. 4, pp. 913–926, 1997.

[11] J. N. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," in *Proc. 21st IEEE Conf. Decis. Control*, 1982, pp. 692–701.

[12] C. H. Papadimitriou, *Computational Complexity*. Hoboken, NJ, USA: Wiley, 2003.

[13] A. Bertrand, M. Moonen, and A. H. Sayed, "Diffusion bias-compensated RLS estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5212–5224, Nov. 2011.

[14] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS over multitask networks," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2733–2748, Jun. 2015.

[15] S. Chouvadras, K. Slavakis, and S. Theodoridis, "Adaptive robust distributed learning in diffusion sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4692–4707, Oct. 2011.

[16] N. Takahashi, I. Yamada, and A. H. Sayed, "Diffusion least-mean squares with adaptive combiners: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4795–4810, Sep. 2010.

[17] F. S. Cattivelli and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 2069–2084, Sep. 2010.

[18] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion recursive least-squares for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1865–1877, May 2008.

[19] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[20] 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.

[21] K. Dedecius and V. Sečkárová, "Dynamic diffusion estimation in exponential family models," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1114–1117, Nov. 2013.

[22] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, Jul. 2008.

[23] P. DiLorenzo and A. H. Sayed, "Sparse distributed learning based on diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1419–1433, Mar. 2013.

[24] R. Arablouei, K. Dogancay, S. Werner, and Y.-F. Huang, "Adaptive distributed estimation based on recursive least-squares and partial diffusion," *IEEE Trans. Signal Process.*, vol. 62, no. 14, pp. 3510–3522, Jul. 2014.

[25] 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.

[26] Z. J. Towfic, J. Chen, and A. H. Sayed, "Collaborative learning of mixture models using diffusion adaptation," in *Proc. 2011 IEEE Int. Workshop Mach. Learn. Signal Process.*, 2011, pp. 1–6.

[27] M. H. DeGroot, "Reaching a consensus," *J. Amer. Statist. Assoc.*, vol. 69, no. 345, pp. 118–127, 1974.

[28] I. D. Schizas, G. Mateos, and G. B. Giannakis, "Distributed LMS for consensus-based in-network adaptive processing," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2365–2382, Jun. 2009.

[29] O. Hlinka, F. Hlawatsch, and P. M. Djurić, "Consensus-based distributed particle filtering with distributed proposal adaptation," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3029–3041, Jun. 2014.

[30] A. Bertrand and M. Moonen, "Consensus-based distributed total least squares estimation in ad hoc wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2320–2330, May 2011.

[31] P. Braca, S. Marano, and V. Matta, "Running consensus in wireless sensor networks," in *Proc. 11th Int. Conf. Inf. Fusion*, Jun. 2008, pp. 1–6.

[32] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 674–690, Aug. 2011.

[33] S. Kar and J. M. F. Moura, "Consensus + innovations distributed inference over networks: Cooperation and sensing in networked systems," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 99–109, May 2013.

[34] S. Kar and J. M. F. Moura, "Asymptotically efficient distributed estimation with exponential family statistics," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4811–4831, Aug. 2014.

[35] S. S. Pereira, A. Pages-Zamora, and R. Lopez-Valcarce, "A diffusion-based distributed EM algorithm for density estimation in wireless sensor networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 4449–4453.
[36] K. Dedecius, J. Reichl, and P. M. Djurić, "Sequential estimation of mixtures in diffusion networks," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 197–201, Feb. 2015.
[37] 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.
[38] V. Kekatos and G. B. Giannakis, "Distributed robust power system state estimation," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1617–1626, May 2013.
[39] N. Bogdanović, J. Plata-Chaves, and K. Berberidis, "Distributed incremental-based LMS for node-specific adaptive parameter estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5382–5397, Oct. 2014.
[40] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS over multitask networks," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2733–2748, Jun. 2015.
[41] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4129–4144, Aug. 2014.
[42] R. Abdolee, B. Champagne, and A. H. Sayed, "Diffusion LMS for source and process estimation in sensor networks," in *Proc. 2012 IEEE Statist. Signal Process. Workshop*, 2012, pp. 165–168.
[43] R. Abdolee, B. Champagne, and A. H. Sayed, "Estimation of space-time varying parameters using a diffusion LMS algorithm," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 403–418, Jan. 2014.
[44] S. Khawatmi, A. M. Zoubir, and A. H. Sayed, "Decentralized clustering over adaptive networks," in *Proc. Eur. Signal Process. Conf.*, 2015, 2696–2700.
[45] 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.
[46] K. Dedecius and V. Sečkárová, "Diffusion estimation of mixture models with local and global parameters," in *Proc. IEEE Statist. Signal Process. Workshop*, 2016, pp. 1–6.
[47] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.
[48] T. S. Jaakkola, "Tutorial on variational approximation methods," in *Advanced Mean Field Methods: Theory and Practice*. Cambridge, MA, USA: MIT Press, 2000, pp. 129–159.
[49] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.
[50] E. I. George and M. Clyde, "Model uncertainty," *Statist. Sci.*, vol. 19, no. 1, pp. 81–94, 2004.
[51] R. Dutta, M. Bogdan, and J. K. Ghosh, "Model selection and multiple testing—A Bayesian and empirical Bayes overview and some new results," *J. Indian Statist. Assoc.*, vol. 50, pp. 105–142, 2012.
[52] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*. Cambridge, MA, USA: Harvard Univ. Press, Jan. 1961.
[53] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[54] B. Wang and D. M. Titterington, "Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model," *Bayesian Anal.*, vol. 1, no. 3, pp. 625–650, 2006.
[55] D. M. Blei, A. Kucukelbir, and J. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Statist. Assoc.*, to be published.
[56] K. L. Mengersen, C. P. Robert, and D. M. Titterington, *Mixtures: Estimation and Applications*. Hoboken, NJ, USA: Wiley, 2011.
[57] J. Winn and C. M. Bishop, "Variational message passing," *J. Mach. Learn. Res.*, vol. 6, pp. 661–694, 2005.
[58] M. Sato, "Online model selection based on the variational Bayes," *Neural Comput.*, vol. 13, no. 7, pp. 1649–1681, 2001.
[59] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. Hoboken, NJ, USA: Wiley, 1994.
[60] I. A. Ibragimov and R. Z. Hasminskii, "Asymptotic behavior of some statistical estimators II. Limit theorems for the a posteriori density and Bayes' estimators," *Theory Probab. Appl.*, vol. 18, no. 1, pp. 76–91, 1973.
[61] S. Ghosal, J. K. Ghosh, and T. Samanta, "On convergence of posterior distributions," *Ann. Statist.*, vol. 23, no. 6, pp. 2145–2152, 1995.
[62] X. Wang, "Approximating Bayesian inference by weighted likelihood," *Can. J. Statist.*, vol. 34, no. 2, pp. 279–298, 2006.
[63] C. Agostinelli and L. Greco, "A weighted strategy to handle likelihood uncertainty in Bayesian inference," *Comput. Statist.*, vol. 28, no. 1, pp. 319–339, Jan. 2012.
[64] S. McLaughlin, V. Krishnamurthy, and S. Challa, "Managing data incest in a distributed sensor network," in *Proc. IEEE Conf. Acoust., Speech, Signal Process.*, 2003, vol. 5, pp. 269–272.
[65] C. You, J. T. Ormerod, and S. Müller, "On bariational Bayes estimation and variational information criteria for linear regression models," *Aust. N. Z. J. Statist.*, vol. 56, no. 1, pp. 73–87, Mar. 2014.
[66] K. Watanabe, "An alternative view of variational Bayes and asymptotic approximations of free energy," *Mach. Learn.*, vol. 86, no. 2, pp. 273–293, Feb. 2012.
[67] 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.
[68] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2882–2904, Jun. 2009.
[69] D. A. Knowles and T. Minka, "Non-conjugate variational message passing for multinomial and binary regression," in *Proc. Adv. Neural Inf. Process. Syst., 24*, 2011, pp. 1701–1709.
[70] M. P. Wand, "Fully simplified multivariate normal updates in non-conjugate variational message passing," *J. Mach. Learn. Res.*, vol. 15, pp. 1351–1369, 2014.

**Kamil Dedecius** received the Ph.D. degree in engineering informatics from Czech Technical University, Prague, Czech Republic, in 2010. Since 2010, he has been a Postdoc and a Research Assistant with the Institute of Information Theory and Automation, Czech Academy of Sciences. His primary research interests include mainly Bayesian probability and statistics, in particular the estimation theory and its application in signal processing. Since 2013, he has been focusing on the theory of fully distributed estimation in diffusion networks. His work has been recognized by the 2015 Otto Wichterle Award.

**Vladimíra Sečkárová** received the Ph.D. degree in probability and mathematical statistics from the Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, in 2015. Since 2015, she has been a Postdoc with the Institute of Information Theory and Automation, Czech Academy of Sciences. Her primary research interests include mainly Bayesian probability and statistics, and information theory.