Sequential Estimation of Mixtures in Diffusion Networks

Kamil Dedecius, *Member, IEEE*, Jan Reichl, and Petar M. Djurić, *Fellow, IEEE*

***Abstract*—The letter studies the problem of *sequential* estimation of mixtures in diffusion networks whose nodes communicate only with their adjacent neighbors. The adopted quasi-Bayesian approach yields a probabilistically consistent and computationally non-intensive and fast method, applicable to a wide class of mixture models with unknown component parameters and weights. Moreover, if conjugate priors are used for inferring the component parameters, the solution attains a closed analytic form.**

***Index Terms*—Diffusion, distributed parameter estimation, sensor networks, sequential mixture estimation.**

I. Introduction

THE problem of distributed estimation of unknown parameters has attracted considerable attention recently, particularly due to continued interest in wireless sensor networks (WSNs). Many of these networks have ad-hoc topologies with easy deployment and removal of nodes endowed with sensing, data processing and communication capabilities. The applications of WSNs range from localisation, field monitoring and target tracking to distributed noise cancellation [1].

This letter focuses on the estimation of mixture models over networks. Most of the approaches to this problem are based on expectation-maximization (EM) algorithms. The decentralized solution of Gu [2] consists of a distributed EM algorithm where each node first calculates local summary statistics in the E-step as usual. This is followed by a consensus step to arrive at global statistics and by a subsequent standard M-step. A similar approach with averaging was proposed by Safarinejadian [3]. The potentially prohibitive communication overhead of the standard multi-iterative consensus step can be alleviated by diffusion, where the nodes exchange local information only once [4].

The referred algorithms were limited to the Gaussian mixture models (GMM). Towfic *et al.* pointed this out and proposed a diffusion adaptation algorithm for general mixture models, where the M-step is solved by an adaptive diffusion process rather than a closed form optimization [6]. Most recently, Pereira *et al.* [7] proposed another diffusion-type estimator, where the propagation of information across the network was embedded in the iterative update of the parameters, where a faster term for information diffusion was combined with a slower term for information averaging. With a sufficient number of iterations, their solution attains equal mean square error as that of the centralized EM algorithm. However, it is again GMM-oriented.

All these estimation algorithms assume *offline cases*, where all the measurements are already available and can serve for iterative optimization of the underlying mixture parameters. However, this may be an obstacle in many *online* sequential cases with the need for instantaneous estimates from sequentially arriving measurements (where the computationally demanding Monte Carlo methods dominate). With the exception of Towfic's method [6], another restriction is that all the methods are for Gaussian mixture models.

The novelty of this letter consists in removing both restrictions. We propose a diffusion-oriented method exploiting the sequential quasi-Bayesian mixture estimation of Smith and Makov [8], [9], further developed, among others, by Titterington *et al.* [10] and Kárný *et al.* [11] (whose variant with dependent variable is considered in Section II). The Bayesian paradigm allows a probabilistically consistent and *abstract* (model independent) formulation of the method, making it applicable to a wide class of mixture models. Furthermore, we show that if the mixture components belong to the exponential family of distributions and the parameters have conjugate priors, then the method is fully analytically tractable.

II. Sequential Quasi-Bayes Mixture Estimation

Let us consider a finite mixture model for the observed real possibly multivariate random variable y_t in the form

$$p(y_t|x_t,\boldsymbol{\phi},\boldsymbol{\theta}) = \sum_{k=1}^{K} \phi_k p_k(y_t|x_t,\theta_k), \qquad t = 1,2,\dots \quad (1)$$

where the components $p_k(y_t|x_t,\theta_k)$ are conditional probability density functions (pdfs) of y_t given (if present) an observed random possibly multivariate real variable x_t, e.g., the regressor, and an unknown vector of mutually independent model parameters $\boldsymbol{\theta} = [\theta_1,\dots,\theta_K]$. We assume that the number of components K is known. The unknown nonnegative component weights ϕ_k form a Dirichlet-distributed random vector $\boldsymbol{\phi} = [\phi_1,\dots,\phi_K]$ taking values in the unit real K-simplex.

Manuscript received July 01, 2014; revised August 27, 2014; accepted August 27, 2014. Date of publication September 05, 2014; date of current version September 10, 2014. The work of K. Dedecius and J. Reichl was supported by the Czech Science Foundation under Grant 14-06678P. The work of P. M. Djurić was supported by the National Science Foundation under AwardsCCF-1320626 and ECCS-1346854. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Michael Rabbat.

K. Dedecius and J. Reichl are with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Prague 8, Czech Republic (e-mail: dedecius@utia.cas.cz; reichja3@fjfi.cvut.cz).

P. M. Djurić is with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794-2350 USA (e-mail: petar.djuric@stonybrook.edu).

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/LSP.2014.2353652

At each time $t = 1, 2, \ldots$ the measurement y_t is generated by a single active component $k_t \in \{1, \ldots, K\}$, selected with probability ϕ_k, that is,

$$p(y_t, k_t|x_t, \boldsymbol{\phi}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \left[\phi_k p_k(y_t|x_t, \theta_k)\right]^{\delta_k(k_t)}, \quad (2)$$

where the Dirac delta $\delta_k(k_t) = 1$ if $k_t = k$ and 0 otherwise is the component membership indicator. In other words, the random vector $[\delta_1(k_t), \ldots, \delta_K(k_t)]$ obeys the multinomial distribution $\mathcal{M}(1, \boldsymbol{\phi})$ with 1 trial and probabilities given by $\boldsymbol{\phi}$. Notice that (1) can be recovered by the marginalization, i.e., summation over $k_t = 1, \ldots, K$.

The recursive Bayesian estimation of mutually independent unknown components parameters $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]$ and mixture weights $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_K]$ relies on the joint prior pdf

$$\pi_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\phi}|X_{t-1}, Y_{t-1}, K_{t-1}) = \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|X_{t-1}, Y_{t-1}, K_{t-1}) \times \pi_{\boldsymbol{\phi}}(\boldsymbol{\phi}|X_{t-1}, Y_{t-1}, K_{t-1}), \quad (3)$$

where $X_{t-1} = \{x_1, \ldots, x_{t-1}\}$; the same applies to Y_{t-1} and K_{t-1}. The Dirichlet pdf of the component weights with hyperparameters $\boldsymbol{\kappa}_{t-1} = [\kappa_{1,t-1}, \ldots, \kappa_{K,t-1}]$ reads

$$\pi_{\boldsymbol{\phi}}(\boldsymbol{\phi}|X_{t-1}, Y_{t-1}, K_{t-1}) \propto \prod_{k=1}^{K} \phi_k^{\kappa_{k,t-1}-1}, \quad (4)$$

and the joint pdf of component parameters is

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|X_{t-1}, Y_{t-1}, K_{t-1}) \propto \prod_{k=1}^{K} \pi_{\theta_k}(\theta_k|X_{t-1}, Y_{t-1}, K_{t-1}) \quad (5)$$

due to the conditional independence of the θ_ks.

The Bayesian estimation consists in incorporation of new data x_t and y_t into the prior (3),

$$\begin{aligned}
&\pi_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\phi}|X_t, Y_t, K_t) \\
&\propto p(y_t, k_t|x_t, \boldsymbol{\theta}, \boldsymbol{\phi}) \pi_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\phi}|X_{t-1}, Y_{t-1}, K_{t-1}) && (6)\\
&= \prod_{k=1}^{K} \phi_k^{\delta_k(k_t)} \phi_k^{\kappa_{k,t-1}-1} && (7)\\
&\times \prod_{k=1}^{K} \left[p_k(y_t|x_t, \theta_k)\right]^{\delta_k(k_t)} \pi_{\theta_k}(\theta_k|X_{t-1}, Y_{t-1}, K_{t-1}), && (8)
\end{aligned}$$

which leads to two separate updates due to the problem formulation. First, the multinomial vector $[\delta_1(k_t), \ldots, \delta_K(k_t)]$ updates the conjugate Dirichlet distribution hyperparameters $\boldsymbol{\kappa}_{t-1}$. Second, the pair x_t, y_t updates the prior distribution of the active component k_t. The first update is analytical, whereas the second one is only under certain conditions (Section II-A).

The ignorance which of the components $k = 1, \ldots, K$ is the active one (k_t) makes the update purely conceptual. Smith and Makov [8], [9] propose a quasi-Bayesian estimation, replacing all the $\delta_k(k_t)$s by their expected values

$$\begin{aligned}
\widehat{\delta}_k(k_t) &= \mathbb{E}[\delta_k(k_t)|X_t, Y_t, K_t] \\
&\propto \mathbb{E}[\phi_k|X_{t-1}, Y_{t-1}, K_{t-1}] p_k(y_t|x_t, X_{t-1}, Y_{t-1}, K_{t-1}) \\
&\propto \frac{\kappa_{k,t-1}}{\sum_{m=1}^{K} \kappa_{m,t-1}} p_k(y_t|x_t, X_{t-1}, Y_{t-1}, K_{t-1}), \quad (9)
\end{aligned}$$

where the predictive likelihoods of y_t read

$$\begin{aligned}
&p_k(y_t|X_t, Y_{t-1}, K_{t-1}) \\
&= \int p_k(y_t|x_t, \theta_k) \pi_{\theta_k}(\theta_k|X_{t-1}, Y_{t-1}, K_{t-1}) d\theta_k. \quad (10)
\end{aligned}$$

In essence, (10) tests how the kth component fits the actual y_t; this measure is weighted by the prior component probabilities in (9) and normalized. The resulting estimates $\widehat{\delta}_k(k_t)$ are then used in place of $\delta_k(k_t)$ in (7) and (8). That is, for $k = 1, \ldots, K$, the parameters of the posterior of $\boldsymbol{\phi}$ are

$$\kappa_{k,t} = \kappa_{k,t-1} + \widehat{\delta}_k(k_t), \quad (11)$$

and the posterior pdf of the respective component parameters θ_k is given by the weighted Bayesian update

$$\begin{aligned}
\pi_{\theta_k}(\theta_k|X_t, Y_t) &\propto \left[p_k(y_t|x_t, \theta_k)\right]^{\widehat{\delta}_k(k_t)} \\
&\times \pi_{\theta_k}(\theta_k|X_{t-1}, Y_{t-1}, K_{t-1}). \quad (12)
\end{aligned}$$

A. Exponential Family Components

Analytical tractability of (12) is guaranteed if the mixture components $p_k(y_t|x_t, \theta_k)$ are exponential family distributions and the conjugate prior distributions for θ_k are used. That is, the components could be written in the form

$$p_k(y_t|x_t, \theta_k) = \exp[\langle \eta_k, T(x_t, y_t) \rangle - \Psi(\eta_k) - k(x_t, y_t)],$$

where $\langle \cdot, \cdot \rangle$ is the inner product, $T(x_t, y_t)$ is the sufficient statistic, $\eta_k \equiv \eta(\theta_k)$ is the natural parameter, $\Psi(\eta_k)$ is the normalizing log-partition function and $k(x_t, y_t)$ is the link function. The conjugate prior, often belonging to the exponential family too, can be written in the conjugate form

$$\begin{aligned}
&\pi_{\theta_k}(\theta_k|\xi_{k,t-1}, \nu_{k,t-1}) \\
&= \exp[\langle \eta_k, \xi_{k,t-1} \rangle - \nu_{k,t-1} \Psi(\eta_k) - l(\xi_{k,t-1}, \nu_{k,t-1})],
\end{aligned}$$

where the hyperparameters $\nu_{k,t-1} \in \mathbb{R}_+$ and $\xi_{k,t-1}$ has the same shape as $T(x_t, y_t)$. The posterior pdf (12) of the kth component in (8) is fully determined by the hyperparameters

$$\begin{aligned}
\xi_{k,t} &= \xi_{k,t-1} + \widehat{\delta}_k(k_t) T(x_t, y_t), \\
\nu_{k,t} &= \nu_{k,t-1} + \widehat{\delta}_k(k_t)
\end{aligned}.$$

III. Distributed Quasi-Bayes Estimation

The novel method for decentralized distributed estimation of general mixtures adopts the diffusion strategy for internodal communication [12]–[15]. Formally, the network consists of N spatially distributed nodes, each of them modelling an identical process of interest using the same functional form of the mixture model (1). At every discrete time instant t, the nodes $i = 1, \ldots, N$ take measurements $x_t^{[i]}$ and $y_t^{[i]}$. Each node i can exchange certain information (defined below) with its adjacent neighbors, forming a neighborhood $\mathcal{N}_i$ of cardinality $|\mathcal{N}_i|$. The node i belongs to $\mathcal{N}_i$, too.

We adopt the adapt-then-combine (ATC) diffusion strategy, consisting of an *adaptation phase*, incorporating neighborhood's measurements, and a *combination phase*, merging the neighborhood's estimates. Its superiority to several other major strategies was demonstrated by Cattivelli and Sayed [14] and Tu and Sayed [5].

The following Bayesian formulation of the distributed mixture estimation is theoretically independent of the component model and thus interesting *per se*. For the purpose of this letter we make the following two assumptions: (i) the mixture components are ordered in the same way in all the network nodes, i.e., the orderings of elements of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is identical across the network; (ii) the combination phase assumes that all the nodes have adequate statistical knowledge of all the inferred parameters. The first assumption, keeping the notation uncluttered, can be abandoned by node-specific indexing of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Its application can be achieved by starting from the same initial prior at all the nodes. The second assumption preserves the number of components after the combine step; a short discussion of its purpose and relaxation (a part of the ongoing research) is in the next section.

A. Adaptation Phase

In the adaptation phase, a node i gathers the neighborhood's measurements $x_t^{[j]}$ and $y_t^{[j]}$ ($j \in \mathcal{N}_i$), either raw or in the form of sufficient statistics, and infers the parameters of interest from them and its own prior statistical knowledge. The information provided by the network nodes may have different reliability due to the observation noise or nodes/links failures, which is reflected by the weights $c_{ij} \in [0,1]$ that i assigns to $j \in \mathcal{N}_i$. The resulting diffusion counterpart of the Bayesian update (6) at node i then reads

$$\pi^{[i]}_{\boldsymbol{\theta},\boldsymbol{\phi}}\left(\boldsymbol{\theta},\boldsymbol{\phi}|\mathcal{H}_t^{[i]}\right) \propto \pi^{[i]}_{\boldsymbol{\theta},\boldsymbol{\phi}}\left(\boldsymbol{\theta},\boldsymbol{\phi}|\mathcal{H}_{t-1}^{[i]}\right) \prod_{j\in\mathcal{N}_i}\left[p^{[j]}\left(y_t^{[j]},k_t\middle|x_t^{[j]},\boldsymbol{\theta},\boldsymbol{\phi}\right)\right]^{c_{ij}} \tag{13}$$

$$= \prod_{k=1}^{K}(\phi_k)^{\kappa^{[i]}_{k,t-1}-1}\prod_{j\in\mathcal{N}_i}(\phi_k)^{c_{ij}\widehat{\delta}^{[j]}_k(k_t)} \tag{14}$$

$$\times \prod_{k=1}^{K}\pi^{[i]}_{\theta_k}\left(\theta_k|\mathcal{H}_{t-1}^{[i]}\right)\prod_{j\in\mathcal{N}_i}\left[p^{[j]}_k\left(y_t^{[j]}\middle|x_t^{[j]},\theta\right)\right]^{c_{ij}\widehat{\delta}^{[j]}_k(k_t)}, \tag{15}$$

where $\mathcal{H}_{t-1}^{[i]}$ represents all prior knowledge of the node i. $\widehat{\delta}^{[j]}_k(k_t)$ measures the probability of the jth node's measurements belonging to the kth component with respect to the ith node's knowledge of θ_k,

$$\widehat{\delta}^{[j]}_k(k_t) \propto \mathbb{E}^{[i]}\left[\phi_k|\mathcal{H}_{t-1}^{[i]}\right]p^{[j]}_k\left(y_t^{[j]}\middle|x_t^{[j]},\mathcal{H}_{t-1}^{[i]}\right) \propto \frac{\kappa^{[i]}_{k,t-1}}{\sum_{m=1}^{K}\kappa^{[i]}_{m,t-1}}p^{[j]}_k\left(y_t^{[j]}\middle|x_t^{[j]},\mathcal{H}_{t-1}^{[i]}\right), \tag{16}$$

and the predictive likelihood

$$p^{[j]}_k\left(y_t^{[j]}\middle|x_t^{[j]},\mathcal{H}_{t-1}^{[i]}\right) = \int p^{[j]}_k\left(y_t^{[j]}\middle|x_t^{[j]},\theta_k\right)\pi^{[i]}_{\theta_k}\left(\theta_k|\mathcal{H}_{t-1}^{[i]}\right)d\theta_k. \tag{17}$$

Under normal operating conditions with reliable nodes ($c_{ij}=1$ for all $j\in\mathcal{N}_i$), the diffusion adaptation phase is a sequence of $|\mathcal{N}_i|$ consecutive ordinary quasi-Bayesian updates.

The resulting Dirichlet posterior hyperparameters after the adaptation phase read

$$\kappa^{[i]}_{k,t} = \kappa^{[i]}_{k,t-1} + \sum_{j\in\mathcal{N}_i}c_{ij}\widehat{\delta}^{[j]}_k(k_t). \tag{18}$$

If the mixture components come from an exponential family of distributions and the parameters have conjugate priors, the parameters of the posterior read

$$\xi^{[i]}_{k,t} = \xi^{[i]}_{k,t-1} + \sum_{j\in\mathcal{N}_i}c_{ij}\widehat{\delta}^{[j]}_k(k_t)T\left(x_t^{[j]},y_t^{[j]}\right), \tag{19}$$

$$\nu^{[i]}_{k,t} = \nu^{[i]}_{k,t-1} + \sum_{j\in\mathcal{N}_i}c_{ij}\widehat{\delta}^{[j]}_k(k_t). \tag{20}$$

B. Combination Phase

During the combination phase the nodes exchange and merge estimates, which in the Bayesian context are the posterior pdfs (13). Denote by $a_{ij}\in[0,1]$ the weights of the posteriors from the nodes $j\in\mathcal{N}_i$ and assume, that

$$\sum_{j\in\mathcal{N}_i}a_{ij} = 1. \tag{21}$$

These weights may be interpreted as probabilities that the information from the respective nodes is true. Besides, the nodes/links reliabilities, a_{ij} may also reflect the statistical properties of the posterior pdfs of the nodes, e.g., their variances and kurtoses. This may be of interest in ad-hoc networks, where new "unlearned" nodes enter the neighborhood.

We adopt the Kullback-Leibler optimal fusion [16] of the posterior pdfs $\pi^{[j]}_{\boldsymbol{\theta}},\boldsymbol{\phi}, j\in\mathcal{N}_i$ to a single pdf $\pi^{[i]}_{\boldsymbol{\theta}},\boldsymbol{\phi}$ with the criterion

$$\sum_{j\in\mathcal{N}_i}a_{ij}D\left(\tilde{\pi}_{\boldsymbol{\theta}^{[i]}},\boldsymbol{\phi}\|\pi^{[j]}_{\boldsymbol{\theta},\boldsymbol{\phi}}\right)\to\min, \tag{22}$$

which results in the geometric mean of the involved pdfs

$$\tilde{\pi}^{[i]}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta},\boldsymbol{\phi}|\cdot) \propto \prod_{j\in\mathcal{N}_i}\left[\pi^{[j]}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta},\boldsymbol{\phi}|\cdot)\right]^{a_{ij}}. \tag{23}$$

The choice of the Kullback-Leibler divergence is not arbitrary: it is a consistent information-theoretic relative entropy measure, the only one simultaneously belonging to the most important divergence classes [17]. The result (23) holds true for the whole class of the Bregman divergences and has appealing computational properties for the exponential family distributions [18] (which will become evident shortly).

Careful investigation of (23) reveals its principal similarity with Equations (3) with factors (4) and (5),

$$\tilde{\pi}^{[i]}_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\cdot) \propto \prod_{j\in\mathcal{N}_i}\left[\pi^{[j]}_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\cdot)\right]^{a_{ij}} \tag{24}$$

$$\tilde{\pi}^{[i]}_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\cdot) \propto \prod_{j\in\mathcal{N}_i}\left[\pi^{[j]}_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\cdot)\right]^{a_{ij}}. \tag{25}$$

Sensu stricto, the Kullback-Leibler optimal fusion of the posterior pdfs at node i yields a parameterisation of a mixture with $K\cdot|\mathcal{N}_i|$ components, significantly overlapping if all the neighboring nodes have adequate statistical knowledge of the parameters. The assumption of the components' identical ordering allows to easily merge the posterior pdfs via (23) and obtain the hyperparameters of the distribution of $\boldsymbol{\phi}$ in the form

$$\tilde{\kappa}^{[i]}_{k,t} = \sum_{j\in\mathcal{N}_i}a_{ij}\kappa^{[j]}_{k,t}, \tag{26}$$

and under conjugacy, the hyperparameters of the distribution of $\boldsymbol{\theta}$ as

$$\tilde{\xi}_{k,t}^{[i]} = \sum_{j \in \mathcal{N}_i} a_{ij} \xi_{k,t}^{[j]}, \tag{27}$$

$$\tilde{\nu}_{k,t}^{[i]} = \sum_{j \in \mathcal{N}_i} a_{ij} \nu_{k,t}^{[j]}. \tag{28}$$

We conjecture that abandoning the requirement of the constant number of components resulting from (23), and inclusion of a method for merging of nearly matching components could (i) significantly improve the tracking capabilities and (ii) eliminate the need of the weights a_{ij}. This is a part of future research.

C. Brief Discussion of Estimator Properties

The limited length of the letter prevents us from thorough analysis of our estimator. In general, assuming appropriate weights, the adaptation phase is equivalent to the ordinary quasi-Bayesian estimation (Section II) from a wider data set. The convergence of the basic method is established for known components [8] and for known weights [9]. The combination phase yields a convex combination of the posterior hyperparameters and hence results in a shrinkage estimator, ideally equal to the centroid of the set of posteriors with respect to the chosen divergence (similarly to clustering [18]).

Algorithm 1 Diffusion Quasi-Bayes Estimation

The nodes $i = 1, \ldots, N$ are initialized with the prior pdfs $\pi_{\boldsymbol{\theta},\boldsymbol{\phi}}^{[i]}$. The weights a_{ij} and c_{ij} are set. For $t = 1, 2, \ldots$ and each node i do:

Adaptation:

1) Gather measurements $x_t^{[j]}, y_t^{[j]}, j \in \mathcal{N}_i$.

2) Calculate the component membership indicator $\widehat{\delta}_k^{[j]}$, Equation (16), using the predictive likelihood (17).

3) Update the prior of $\boldsymbol{\phi}$ using $\widehat{\delta}_k^{[j]}$, Eq. (18).

4) Update the prior of $\boldsymbol{\theta}$ using $\widehat{\delta}_k^{[j]}$ and $x_t^{[j]}, y_t^{[j]}$, Eq. (15), or under conjugacy directly (19) and (20).

Combination:

1) Gather posterior pdfs $\pi_{\boldsymbol{\theta},\boldsymbol{\phi}}^{[j]} = \pi_{\boldsymbol{\theta}}^{[j]} \cdot \pi_{\boldsymbol{\phi}}^{[j]}, j \in \mathcal{N}_i$, and calculate $\tilde{\pi}_{\boldsymbol{\theta},\boldsymbol{\phi}}^{[i]}$ using (25) for $\boldsymbol{\phi}$ and (24) for $\boldsymbol{\theta}$, under conjugacy directly (27) and (28).

IV. Simulation Results

We assume a randomly generated diffusion network of 25 nodes whose objective is to estimate the coefficients of a mixture of two linear models given by

$$y_t^{[i]} = x_t^{\mathsf{T}} \beta_k + \varepsilon_t^{[i]}, \qquad k \in \{1, 2\}, t = 1, \ldots, 600,$$

where $y_t^{[i]}$ is a scalar measurement taken by a node i, $x_t \in \mathbb{R}^5$ is a row regression vector generated from $\mathcal{N}(0, I)$ where I is a 5×5 identity matrix, $\varepsilon_t^{[i]} \sim \mathcal{N}(0, 0.05i)$ and $\beta_k \in \mathbb{R}^5$ are two column vectors of regression coefficients, $\beta_1 = [0.3, 0, 0.9, 0.3, 1]^{\mathsf{T}}$ and $\beta_2 = [0.3, 1, -0.9, 0.8, -1.5]^{\mathsf{T}}$, respectively. The data generation mostly exploits β_1, and the switch to β_2 occurs for short periods only. These are depicted by gray bands in Fig. 1. The task is the online estimation of β_1 and β_2, while $\boldsymbol{\phi}$ is not of direct interest due to the problem setting, thought it still plays an important role in the estimation process. For simplicity, the weights $c_{ij} = 1$ and $a_{ij} = 1/|\mathcal{N}_i|$ when $j \in \mathcal{N}_i$. Of course, their setting reflecting the noise properties is likely to improve the estimation.

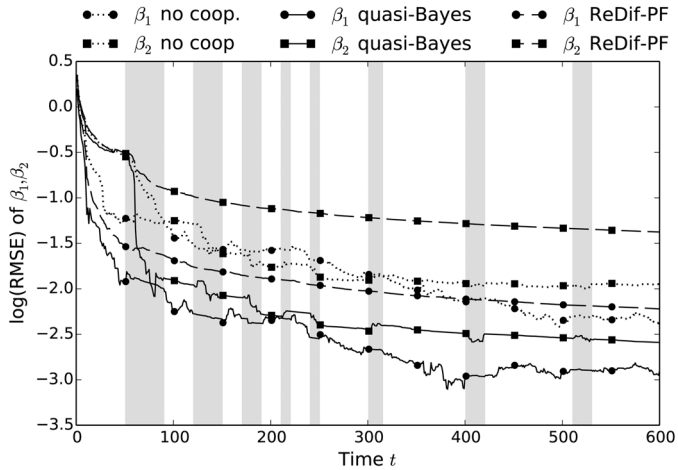

Fig. 1. Evolution of the network log (RMSE) in time.

Besides the no-cooperation quasi-Bayesian estimation, the proposed method is compared with the recent random exchange diffusion particle filter (ReDif-PF), proposed by Bruno and Dias [19] with the multimodality-preserving technique of Vermaak *et al.* [20] employed. The ReDif-PF posterior distribution is represented by a Gaussian mixture model (GMM), significantly reducing the network communication requirements and preventing the particle depletion problem [21]. In the simulation, the posterior is approximated by one normal density per component.

All methods start from the same initial normal priors/proposals with means being rounded values of β_1 and β_2 and with diagonal variances with elements 0.05. The time evolution of the root mean square error $\mathrm{RMSE}(\beta_k^{[i]}) = \sqrt{\frac{1}{t}\sum_{\tau=1}^{t} \left(\widehat{\beta}_{k,\tau}^{[i]} - \beta_k\right)^2}$ averaged over the network is depicted in Fig. 1. The results of ReDif-PF are obtained with 1000 samples per component and averaged over 100 runs.

The results clearly show that the quasi-Bayesian estimation with cooperation outperforms the other two algorithms. We make a few other comments. After a short learning, the quasi-Bayesian methods react relatively sensitively to periods where the respective components are active. ReDif-PF suffers from an inappropriate initial proposal for β_2, whereas the quasi-Bayesian methods do not. The communication requirements of both cooperative methods are the same for the adaptation phase. The exchange of posteriors is $|\mathcal{N}_i|$-times more costly in the quasi-Bayesian method, since ReDif-PF communicates them with a single randomly chosen neighbor, which adopts them as the new proposal. This strategy of the ReDif-PF may lead to performance degradation. The computation requirements of quasi-Bayes are significantly lower because there are no iterations and no particles to be communicated.

REFERENCES

[1] 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.

[2] D. Gu, "Distributed EM algorithm for Gaussian mixtures in sensor networks," *IEEE Trans. Neural Netw.*, vol. 19, no. 7, pp. 1154–1166, Jul. 2008.

[3] B. Safarinejadian, M. B. Menhaj, and M. Karrari, "Distributed unsupervised Gaussian mixture learning for density estimation in sensor networks," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 9, pp. 2250–2260, Sep. 2010.

[4] Y. Weng, W. Xiao, and L. Xie, "Diffusion-based EM algorithm for distributed estimation of Gaussian mixtures in wireless sensor networks," *Sensors*, vol. 11, no. 6, pp. 6297–316, Jan. 2011.

[5] S.-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6217–6234, 2012.

[6] Z. J. Towfic, J. Chen, and A. H. Sayed, "Collaborative learning of mixture models using diffusion adaptation," in *Proc. 2011 IEEE Int. Workshop on Machine Learning for Signal Processing*, Sep. 2011, pp. 1–6.

[7] S. S. Pereira, R. Lopez-Valcarce, and A. Pages-Zamora, "A diffusion-based EM algorithm for distributed estimation in unreliable sensor networks," *IEEE Signal Process. Lett.*, vol. 20, no. 6, pp. 595–598, Jun. 2013.

[8] A. F. M. Smith and U. E. Makov, "A quasi-Bayes sequential procedure for mixtures," *J. Roy. Statist. Soc. B (Methodological)*, vol. 40, no. 1, pp. 106–112, Jan. 1978.

[9] A. Smith and U. Makov, "Unsupervised learning for signal versus noise (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 27, no. 4, pp. 498–500, Jul. 1981.

[10] D. M. Titterington, A. F. M. Smith, and U. M. Makov, *Statistical Analysis of Finite Mixture Distributions*. Chichester, UK: Wiley, 1985.

[11] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař, *Optimized Bayesian Dynamic Advising: Theory and Algorithms (Advanced Information and Knowledge Processing)*. London, U.K.: Springer, 2006.

[12] 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.

[13] 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.

[14] 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.

[15] A. H. Sayed, "Adaptive networks," in *Proc. IEEE*, Apr. 2014, vol. 102, no. 4, pp. 460–497.

[16] 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.

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

[18] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2882–2904, Jun. 2009.

[19] M. G. S. Bruno and S. S. Dias, "Collaborative emitter tracking using Rao-Blackwellized random exchange diffusion particle filtering," *EURASIP J. Adv. Signal Process.*, vol. 19, pp. 1–18, 2014.

[20] J. Vermaak, A. Doucet, and P. Pérez, "Maintaining multi-modality through mixture tracking," in *Proc. 9th IEEE Int. Conf. Computer Vision*, Oct. 2003, vol. 2, pp. 1110–1116.

[21] R. van der Merwe and E. Wan, "Gaussian mixture sigma-point particle filters for sequential probabilistic inference in dynamic state-space models," in *Proc. 2003 IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '03). 2003*, vol. 6, pp. 701–704.