DIFFUSION ESTIMATION OF MIXTURE MODELS WITH LOCAL AND GLOBAL PARAMETERS

Kamil Dedecius and Vladimíra Sečkárová

Institute of Information Theory and Automation
Czech Academy of Sciences
Pod Vodárenskou věží 1143/4, 182 08 Prague 8, Czech Republic

ABSTRACT

The state-of-art methods for distributed estimation of mixtures assume the existence of a common mixture model. In many practical situations, this assumption may be too restrictive, as a subset of parameters may be purely local, e.g., if the numbers of observable components differ across the network. To reflect this issue, we propose a new online Bayesian method for simultaneous estimation of local parameters, and diffusion estimation of global parameters. The algorithm consists of two steps. First, the nodes perform local estimation from own observations by means of factorized prior/posterior distributions. Second, a diffusion optimization step is used to merge the nodes' global parameters estimates. A simulation example demonstrates improved performance in estimation of both parameters sets.

Index Terms— Diffusion estimation, distributed estimation, exponential family, mixture models, message passing.

1. INTRODUCTION

Distributed estimation of parameters of stochastic models has attracted a significant attention in the last two decades, particularly due to a rapid development of cheap ad-hoc wireless sensor networks consisting of nodes endowed with sensing, data processing and communication capabilities. Their applications range from localisation, target tracking, intrusion detection, dictionary learning etc. [1].

According to the communication strategies among sources, several groups of methods can be recognized. First, the incremental algorithms, passing and processing information in a Hamiltonian cyclic path. These algorithms are prone to failures, as each node and link are single-points of failure, and recovery is an NP-hard problem [2]. This issue can be solved by the diffusion and consensus strategies. The former perform only a limited amount of communication among neighboring nodes, typically represented by an exchange of observations during the adaptation step and an exchange of estimates during the combination step [1, 3]. The standard consensus strategies involved intermediate steps to reach consensus, but the recent *running* consensus algorithms removes this overhead, e.g. [4].

In this paper, we specifically focus on sequential estimation of mixture models. Under (roughly) common distribution of observations $y_{i,t}$ across the network, there exist several methods for collaborative estimation of mixtures. One branch of solutions is based on expectation-maximization (EM) algorithm. For instance, Gu proposed a version of a decentralized EM algorithm with a consensus step evaluating global network statistics between the standard local E and M steps [5]. A similar approach, called distributed expectation-maximization, was proposed by Safarinejadian, Menhai and Karrari [6], where, between the E and M steps, a distributed averaging approach is used to diffuse local sufficient statistics to neighboring nodes and estimate global sufficient statistics in each node. In the M-step, each node updates parameters of the normal mixture model using the estimated global sufficient statistics. A fully diffusion-oriented EM algorithm was proposed by Pereira, Pages-Zamora, and Lopez-Valcarce in [7]. Another branch of solutions, stemming from the Bayesian theory, is mostly based on variational inference. The distributed variational Bayesian algorithm (DVBA) of Safarinejadian, Menhaj and Karrari allows distributed estimation of normal mixture models [8]. However, DVBA is an incremental algorithm. In order to remove the robustness issue of this topology, the authors later proposed a peer-to-peer DVBA algorithm, where the nodes communicate with their neighbors [9]. Recently, Dedecius, Reichl and Djurić proposed a distributed quasi-Bayesian algorithm with point-estimated component indicator, applicable to a wide class of mixture models with component distributions from the exponential family and suitable for online estimation [10].

With the exception of component weights in [5], the listed algorithms rely on the assumption of (at least roughly) common model parameters. However, the parameters may differ across the network, e.g., if the nodes observe different targets maneuvering in a formation. Then, the collaboration may be based on an assumption of correlation or similarity of inferred parameters [11]. These *multitask* diffusion algorithms are mostly LMS-based, e.g. [11, 12] with a few modifications reflecting, for instance, sparsity [13]. Some more general recent algorithms allow unsupervised determination of neighbors belonging to the same cluster [11, 15, 16]. These algorithms allow static or sequential collaborative estimation of local and global parameters, and parameters shared by clusters of nodes.

We would like to contribute to this topic from the probabilistically consistent and versatile Bayesian viewpoint, allowing abstract formulation of inference tasks. In particular, we propose a new method for collaborative estimation of mixture models with local and global parameter sets. The configuration of parameters is very general, the mixtures may differ in the number of components, their weights and the components may have different parameters, too. The method is generic and applicable to a wide class of mixture models. Under certain conditions, it simplifies to an analytically tractable mean-field variational method. The resulting algorithm consists of two phases. First, the local estimation, where the nodes perform inference of both local and global parameters sets locally from own observations. Second, the diffusion optimization step, where the nodes exchange the distributions of global parameters with their neighbors within 1 hop distance. If the mixture components are exponential

K. Dedecius is supported by the Czech Science foundation, grant No. 14-06678P. V. Sečkárová is supported by the Czech Science Foundation, grant No. 13-13502S.

family distributions, the whole algorithm can be reduced to variational message passing (e.g., [17]) extended by message-passing diffusion optimization.

2. PROBLEM STATEMENT

Let us consider a network — a directed or undirected graph of nodes $\mathcal{I} = \{1, \ldots, I\}$, connected by a set of edges determining the graph topology. Each node $i \in \mathcal{I}$ acquires observations $y_{i,t}$ where $t = 1, 2, \ldots$ is a discrete time index. These observations can be modelled by mixture models with a common (*global*) subset of parameters. Each node i also communicates with its adjacent neighbors within 1 hop distance, forming its *neighborhood* $\mathcal{I}_i$. We emphasize that $i \in \mathcal{I}_i$, too. The adopted communication strategy is *diffusion* [1, 3], however, compared to the ordinary diffusion algorithms, the adaptation step for an observations exchange among nodes is not applicable here for their potentially different distributions. The combination step merging the parameters estimates is preserved (in a special form).

The incoming observations $y_{i,t}$ of node i follow a mixture distribution, i.e., a convex combination of K_i probability distributions called components. Their probability densities are denoted $p_{i,k}(y_{i,t}|\theta_{i,k})$, where $\theta_{i,k}$ are component parameters and $k = 1, \ldots, K_i$ are component indices. Denoting by Θ_i all unknown parameters of the considered mixture, we have

$$p_i(y_{i,t}|\Theta_i) = \sum_{k=1}^{K_i} \phi_{i,k} p_{i,k}\left(y_{i,t}|\theta_{i,k}\right), \tag{1}$$

where $\phi_i = [\phi_{i,1}, \ldots, \phi_{i,K}]$ is a vector of component weights, that is, positive real numbers taking values in the $(K_i - 1)$-dimensional probabilistic simplex. Note, that Θ_i may be any combination of ϕ_i, θ_i or K_i. As a minor simplification, we assume the local numbers of components K_i to be known *a priori*.

We face the problem of sequential collaborative estimation of nodes' mixtures parameters with the assumption that there exists a *global* subset of parameters Θ^G, that is common to all network nodes, and a local subset that is its local complement to Θ_i,

$$\Theta^G = \bigcap_{i \in \mathcal{I}} \Theta_i, \qquad \text{and} \qquad \Theta_i^L = \Theta_i \setminus \Theta^G.$$

The goal is to exploit the existing network in order to collaboratively arrive at better estimates of Θ^G under modest communication requirements. In addition, we conjecture that an improvement in the estimation of Θ^G will induce an improved estimation of Θ_i^L.

3. SEQUENTIAL DIFFUSION ESTIMATION

The ordinary sequential Bayesian inference of unknown parameters Θ_i relies on a joint prior distribution

$$\pi_i(\Theta_i|y_{i,0:t-1}) = \pi_i\left(\Theta_i^L, \Theta_i^G \middle| y_{i,0:t-1}\right),$$

quantifying the current statistical knowledge about Θ_i^L and Θ^G up to time $t-1$, which is based on the previous observations $y_{i,1}, \ldots, y_{i,t-1}$, and the pseudo-observations $y_{i,0}$ determining any initial knowledge, e.g. from historical data or an expert's opinion. The incorporation of new observation $y_{i,t}$ is performed by virtue of the Bayes' theorem,

$$\pi_i(\Theta_i|y_{i,0:t}) = \frac{\pi_i(\Theta_i|y_{i,0:t-1}) p_i(y_{i,t}|\Theta_i)}{\int \pi_i(\Theta_i|y_{i,0:t-1}) p_i(y_{i,t}|\Theta_i) d\Theta_i}. \tag{2}$$

Unfortunately, this rigorous Bayesian approach to inference of Θ_i is impractical not only computationally, but also from aspect of the distributed estimation. In order to combine information from several sources (nodes), the distribution of Θ^G must have the same form across the network, conditionally independent of any local parameters Θ_i^L. That means,

$$\begin{aligned}\pi_i(\Theta_i|y_{i,0:t}) &= \pi_i\left(\Theta_i^L, \Theta^G \middle| y_{i,0:t}\right) \\ &\approx \rho_i\left(\Theta_i^L \middle| y_{i,0:t}\right) \rho_i\left(\Theta^G \middle| y_{i,0:t}\right).\end{aligned} \tag{3}$$

In other words, the true distribution of Θ_i^L and Θ^G at each node $i \in \mathcal{I}$ must be replaced by two lower-dimensional distributions $\rho_i(\Theta_i^L|\cdot)$ and $\rho_i(\Theta^G|\cdot)$, the latter with the same functional form for all i, and whose product is *as close to the original distribution as possible (in the Kullback-Leibler sense)*. Sometimes, this factorization is natural, e.g., in mixture estimation, where the component weights may be modelled as independent on component parameters. More on possible factorizations can be found in [18].

The second step is the diffusion optimization of the distribution of Θ^G. From node $i \in \mathcal{I}$ viewpoint, it is possible to view the neighbors' distributions $\rho_j(\Theta^G|\cdot)$ as *hypotheses* about Θ^G. The goal is to optimally merge these hypotheses into a single distribution $\tilde{\rho}_i(\Theta^G|\cdot)$ that is as close to individual hypotheses as possible.

Consistently with the Bayesian theory, both steps will exploit the Kullback-Leibler divergence as the measure of proximity of probability distributions. For two probability densities $f(x)$ and $g(x)$, the latter absolutely continuous with respect to the former, the Kullback-Leibler divergence is defined as

$$\mathcal{D}\left[f(x)||g(x)\right] = \mathbb{E}_{f(x)}\left[\log \frac{f(x)}{g(x)}\right]. \tag{4}$$

This divergence is a premetric: it is nonnegative and equal to zero if $f = g$ almost everywhere, but it is neither symmetric, nor does it satisfy the triangle inequality.

For the sake of simplicity, we adopt the assumption that the nodes know which parameters are local and which are global. This situation occurs, e.g., if different types of sensors are used for taking observations. The proposed algorithms is as follows:

Algorithm formulation

We aim to design a distributed algorithm for online mixtures estimation performing the following two steps:

1. Local estimation:
At each node $i \in \mathcal{I}$, locally estimate parameters Θ_i^L and Θ^G in a Kullback-Leibler-optimal factorized form

$$\mathcal{D}\left[\rho_i\left(\Theta_i^L\middle|\cdot\right)\rho_i\left(\Theta^G\middle|\cdot\right)\middle|\middle|\pi_i\left(\Theta_i^L, \Theta^G\middle|\cdot\right)\right] \to \min.$$

2. Diffusion optimization:
At each node $i \in \mathcal{I}$, find the Kullback-Leibler-optimally merged distribution $\widetilde{\rho}(\Theta^G|\cdot)$ using the neighbors' distribution $\rho_j(\Theta^G|\cdot)$ as hypotheses about Θ^G,

$$\widetilde{\rho}_i(\Theta^G|\cdot) = \arg\min_{\rho_i^*} |\mathcal{I}_i|^{-1} \sum_{j \in \mathcal{I}_i} \mathcal{D}\left[\rho_i^*\left(\Theta^G\middle|\cdot\right)\middle|\middle|\rho_j\left(\Theta^G\middle|\cdot\right)\right]. \tag{5}$$

3.1. Local estimation

Local estimation asserts approximation of the intractable true density $\pi_i\left(\Theta_i^L, \Theta^G|\cdot\right)$ by individual lower-dimensional densities $\rho_i\left(\Theta_i^L|\cdot\right)$ and $\rho_i\left(\Theta^G|\cdot\right)$ as in (5). That is, the goal is to perform minimization of the Kullback-Leibler divergence

$$\begin{aligned}
&\mathcal{D}\left[\rho_i\left(\Theta_i^L, \Theta^G|\cdot\right)\middle|\middle|\pi_i\left(\Theta_i^L, \Theta^G|y_{i,0:t}\right)\right] \\
&= \mathbb{E}_{\rho_i(\Theta_i^L, \Theta^G|\cdot)}\left[\log\frac{\rho_i\left(\Theta_i^L, \Theta^G|\cdot\right)}{\pi_i\left(\Theta_i^L, \Theta^G|y_{i,0:t}\right)}\right] \\
&= \mathbb{E}_{\rho_i(\Theta_i^L|\cdot)}\left[\log\rho_i\left(\Theta_i^L|\cdot\right)\right] + \mathbb{E}_{\rho_i(\Theta^G|\cdot)}\left[\log\rho_i\left(\Theta^G|\cdot\right)\right] \\
&- \mathbb{E}_{\rho_i(\Theta_i^L|\cdot)}\mathbb{E}_{\rho_i(\Theta^G|\cdot)}\left[\log\pi_i\left(\Theta_i^L, \Theta^G\middle|\cdot\right)\right]. \qquad (6)
\end{aligned}$$

It is straightforward to see that the minimization of (6) with respect to Θ_i^L and Θ^G leads to minimizations of expected Kullback-Leibler divergences with lower-dimensional densities. Because the minimum is reached under equal arguments of the Kullback-Leibler divergence, the solution is given by the system

$$\begin{aligned}
\rho_i\left(\Theta^G|\cdot\right) &= c_L \exp\left\{\mathbb{E}_{\rho_i(\Theta_i^L|\cdot)}\left[\log\pi_i\left(\Theta_i^L, \Theta^G\middle|\cdot\right)\right]\right\}, \\
\rho_i\left(\Theta_i^L|\cdot\right) &= c_G \exp\left\{\mathbb{E}_{\rho_i(\Theta^G|\cdot)}\left[\log\pi_i\left(\Theta_i^L, \Theta^G\middle|\cdot\right)\right]\right\},
\end{aligned}$$

where c_L and c_G are normalization constants.

The presented factorization is closely related to the mean-field variational Bayesian inference [19, 20]. Indeed, it is possible to further factorize the particular densities $\rho_i(\Theta_i^L|\cdot)$ and/or $\rho_i(\Theta^G|\cdot)$ and proceed with the ordinary variational Bayes method, sequentially seeking a consistent solution by revising the lower-dimensional densities in a circular way until a convergence criterion is met. The mean-field variational algorithms are guaranteed to converge [21]. In the sequential estimation framework, usually a few iterations are performed at each time step, using previous observations stored in memory. The posterior distributions from the previous time step then serve as the prior distributions. More elaborated methods for online variational inference suitable for large data can be found in [22, 23].

In the next section, the described local estimation via factorized densities will be extended by diffusion optimization of the posterior distribution of Θ^G. Then, it will be shown, that if the distributions of the mixture components belong to the exponential family, the proposed method is analytically tractable as a message passing algorithm.

3.2. Diffusion optimization

Each node $i \in \mathcal{I}$ has access to densities $\rho_j(\Theta^G|\cdot)$ of its neighbors $j \in \mathcal{I}_i$, representing hypotheses about the true parameter Θ^G. Instead of working with the whole system of individual densities, we want to replace them by a single density $\tilde{\rho}_i(\Theta^G|\cdot)$. In particular, we seek an element from the set of all admissible densities minimizing the average divergence to all individual densities,

$$\begin{aligned}
\tilde{\rho}_i(\Theta^G|\cdot) &= \arg\min_{\rho_i^*} |\mathcal{I}_i|^{-1} \sum_{j\in\mathcal{I}_i} \mathcal{D}\left[\rho_i^*(\Theta^G|\cdot)\middle|\middle|\rho_j(\Theta^G|\cdot)\right] \\
&= \arg\min_{\rho_i^*} \mathcal{D}\left[\rho_i^*(\Theta^G|\cdot)\middle|\middle|\prod_{j\in\mathcal{I}_i}\left[\rho_j(\Theta^G|\cdot)\right]^{1/|\mathcal{I}_i|}\right] \\
&= \prod_{j\in\mathcal{I}_i}\left[\rho_j(\,\Theta^G|\,\cdot\,)\right]^{1/|\mathcal{I}_i|}. \qquad (7)
\end{aligned}$$

The resulting Kullback-Leibler-optimal distribution $\tilde{\rho}_i(\Theta^G|\cdot)$ is thus a geometric average of neighbors' distributions. In the Bayesian diffusion framework, this merging coincides with the *combine* step, preferably performed on neighbors' posterior distributions [25]. However, in our case, the local estimation method is iterative, and (7) may be used between any subsequent iterations to speed up convergence. This promising topic is postponed to further research.

Generally, the variational estimation of the posterior distribution is inaccurate in the early stage of the online learning and gradually becomes accurate as learning proceeds [24]. Therefore, it is reasonable 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. The proposed collaboration by fusion of posterior distributions can be seen as a weighted Bayesian learning that contributes to the inference process (this will become more apparent in the ongoing section).

4. COMPONENTS FROM THE EXPONENTIAL FAMILY

If the mixture components belong to the exponential family of distributions, the local estimation can take the form of a message passing algorithm. In general, any random variable y_t has an exponential family distribution with a parameter ϑ, if its probability density function can be written in the form

$$p(y_t|\vartheta) = \exp\left[\eta \cdot T_y(y_t) - A(\eta) - k(y_t)\right], \qquad (8)$$

where $\eta \equiv \eta(\vartheta)$ is a natural parameter, $T_y(y_t)$ is a sufficient statistic, $A(\eta)$ is a log-partition (normalizing) function and $k(y_t)$ is a function of y_t. The particular form is not unique. The Bayesian estimation of ϑ is analytically tractable if the prior distribution of ϑ is conjugate, i.e., parameterized by hyperparameters ξ_{t-1} of the same size as $T_y(y_t)$ and real scalar ν_{t-1}, and with a density of the form

$$\begin{aligned}
\pi_\vartheta(\vartheta|\xi_{t-1}, \nu_{t-1}) = \exp\left[\eta \cdot \xi_{t-1} - \nu_{t-1}A(\eta)\right] \\
- \exp\left[h(\xi_{t-1}, \nu_{t-1})\right], \qquad (9)
\end{aligned}$$

where $A(\eta)$ is the same function as in the exponential family distribution and $h(\xi_{t-1}, \nu_{t-1})$ is a known function. The posterior distribution of ϑ is then given by updated hyperparameters

$$\xi_t = \xi_{t-1} + T_y(y_t), \quad \text{and} \quad \nu_t = \nu_{t-1} + 1. \qquad (10)$$

In the local estimation step, the component densities $p_{i,k}(y_{i,t}|\theta_{i,k})$ from Equation (1) and the convenient prior distributions for $\theta_{i,k}$ are rewritten to compatible forms according to (8) and (9). Similarly, the component indicators for $y_{i,t}$, determining which component generated the actual $y_{i,t}$, are modelled by multinomial distributions with the probabilities (i.e., component weights $\phi_{i,k}$) provided by the conjugate Dirichlet distribution. The local estimation algorithm then iterates by passing messages – expectations – from the prior distributions to the components and multinomial indicator models, which, in turn, return messages containing the sufficient statistics updating the prior hyperparameters in a sense (10). This is known as the variational message passing [17].

From (9) and the diffusion optimization step (7) it follows, that merging of posterior distributions takes the form

$$\tilde{\xi}_{i,t} = \frac{1}{|\mathcal{I}_i|}\sum_{j\in\mathcal{I}_i}\xi_{j,t}, \quad \text{and} \quad \tilde{\nu}_{i,t} = \frac{1}{|\mathcal{I}_i|}\sum_{j\in\mathcal{I}_i}\nu_{j,t}.$$

Remind, that it can be performed between any two subsequent iterations of the message passing algorithm, e.g. to speed up convergence, or after the local estimation step to save communication resources. Also remark its principal similarity with the Bayesian update under conjugacy (10) – this is where the cornerstone of this merging lies.

5. SIMULATION EXAMPLE

This example demonstrates the sequential diffusion estimation of a normal mixture model by a network consisting of 16 nodes. Its topology is depicted in Figure 1. The mixture has the form

$$y_{i,t}|\Theta_i^L,\Theta^G \sim \sum_{k=1}^{K_i}\phi_{i,k}\mathcal{N}(\mu_{i,k},\Sigma_{i,k}),$$

where $\mu_{i,k}$ and $\Sigma_{i,k}$ are the mean vectors and covariance matrices. The first component observed by all network nodes has parameters

$$\mu_1=\begin{bmatrix}-5\\-5\end{bmatrix},\qquad \Sigma_{1,i}=\begin{bmatrix}1.5+\varepsilon_i & 0.9\\ 0.9 & 2.0+\varepsilon_i'\end{bmatrix},$$

where $\varepsilon_i,\varepsilon_i' \sim \mathcal{E}xp(0.5)$ are i.i.d. random samples from the exponential distribution. The second component is observed by nodes $i\in\{7,\dots,16\}$ only. Their parameters are

$$\mu_{2,i}=\begin{bmatrix}0\\0\end{bmatrix},\qquad \Sigma_{2,i}=\begin{bmatrix}2.0+\varepsilon_i & 0.8\\ 0.8 & 3.0+\varepsilon_i'\end{bmatrix},$$

where $\varepsilon_i,\varepsilon_i' \sim \mathcal{E}xp(0.5)$ are i.i.d. random samples from the exponential distribution. The component weights in nodes $i\in\{7,\dots,16\}$ are $\phi_i=[0.7,0.3]$. The parameters sets are

$$\Theta^G=\{\mu_1\},\qquad \text{and}\qquad \Theta_i^L=\{\mu_{2,i},\Sigma_{1,i},\Sigma_{2,i},\pi_i\}.$$

(Note that there is a potential for cooperation in estimation of $\mu_{2,i}$ and ϕ_i.) The estimation starts from the initial normal ($\mathcal{N}$), inverse-Wishart ($i\mathcal{W}$) and Dirichlet ($\mathcal{D}ir$) prior distributions

$$\begin{aligned}\mu_1 &\sim \mathcal{N}\left([-7,-7]^\intercal, 15\cdot I_{[2\times 2]}\right),\\ \mu_2 &\sim \mathcal{N}\left([7,7]^\intercal, 15\cdot I_{[2\times 2]}\right),\\ \Sigma_1,\Sigma_2 &\sim i\mathcal{W}\left(5, 0.1\cdot I_{[2\times 2]}\right),\\ \pi &\sim \mathcal{D}ir([1,1]).\end{aligned}$$

The single-component nodes provide their information to all neighbors, but incorporate $\rho_i(\Theta^G|\cdot)$ only from other single-component nodes. 1000 observations were generated for each node, the estimation starts when the first 100 observations are received. At each time t, 5 iterations of the local estimation step are performed, then, the diffusion optimization follows.

As a performance measure, we employ MSE averaged over the network (denoted by AMSE). The evolution of AMSEs under cooperation and no-cooperation are depicted in Fig. 2 for μ_1 and $\mu_{2,i}$, Fig. 3 for $\Sigma_{1,i}$ and $\Sigma_{2,i}$ and Fig. 4 for π_i. We conclude, that the proposed method leads to a significant improvement in estimation of μ_1. Moreover, it simultaneously helps the estimation of other parameters.

Since the memory length for $y_{i,t}$ would be limited in practice, we also performed a simulation with a queue-type memory for 100 observations. The results were very similar to those presented here, the method performed very slightly worse in terms of AMSE.

Finally, we remark that [5] provides an algorithm that allows distributed mixture estimation with inhomogeneous component weights across the network, but assumes identical components.

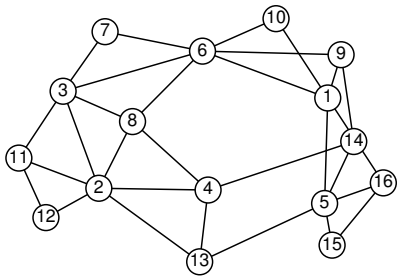

Fig. 1. Topology of the diffusion network.

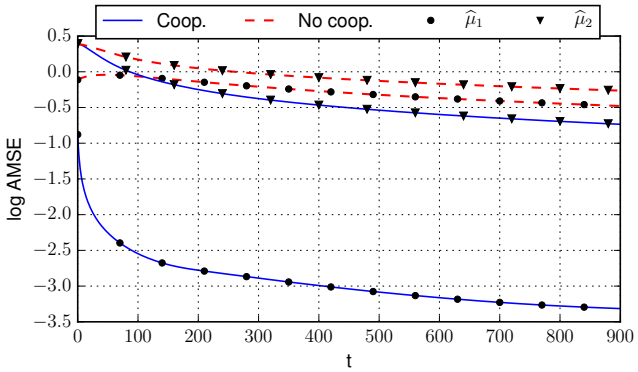

Fig. 2. Evolution of logarithm of AMSE of μ_1 and $\mu_{2,i}$ under cooperation (Coop.) and no cooperation (No coop.).

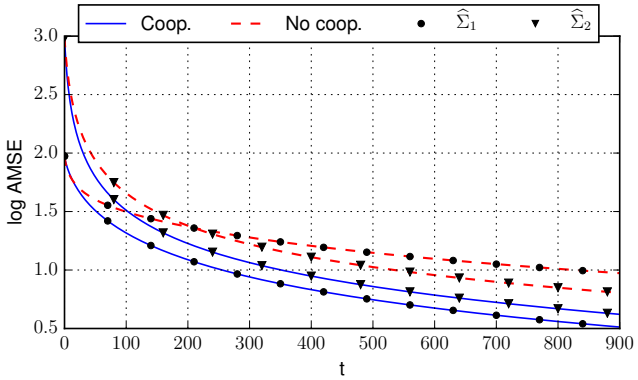

Fig. 3. Evolution of logarithm of AMSE of $\Sigma_{1,i}$ and $\Sigma_{2,i}$ under cooperation (Coop.) and no cooperation (No coop.).

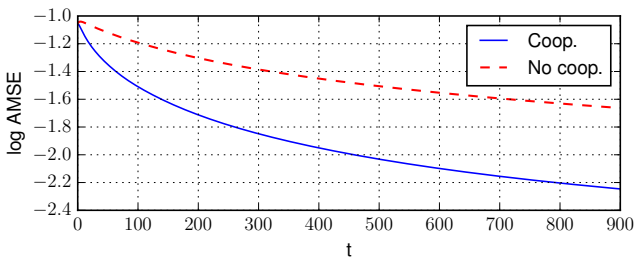

Fig. 4. Evolution of logarithm of AMSE of π_i under cooperation (Coop.) and no cooperation (No coop.).

6. REFERENCES

[1] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[2] C. H. Papadimitriou, *Computational Complexity*, Addison Wesley, 1994.

[3] A. H Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[4] P. Braca, S. Marano, and V. Matta, "Running consensus in wireless sensor networks," in *11th International Conference on Information Fusion*, June 2008, pp. 1–6.

[5] D. Gu, "Distributed EM Algorithm for Gaussian Mixtures in Sensor Networks," *IEEE Transactions on Neural Networks*, vol. 19, no. 7, pp. 1154–1166, July 2008.

[6] B. Safarinejadian, M. B. Menhaj, and M. Karrari, "Distributed unsupervised Gaussian mixture learning for density estimation in sensor networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 9, pp. 2250–2260, Sept. 2010.

[7] S. S. Pereira, A. Pages-Zamora, and R. Lopez-Valcarce, "A diffusion-based distributed EM algorithm for density estimation in wireless sensor networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. May 2013, pp. 4449–4453.

[8] B. Safarinejadian, M.B. Menhaj, and M. Karrari, "Distributed variational Bayesian algorithms for Gaussian mixtures in sensor networks," *Signal Processing*, vol. 90, no. 4, pp. 1197–1208, Apr. 2010.

[9] B. Safarinejadian and M. B. Menhaj, "Distributed density estimation in sensor networks based on variational approximations," *International Journal of Systems Science*, Apr. 2011.

[10] K. Dedecius, J. Reichl, and P.M. Djurić, "Sequential estimation of mixtures in diffusion networks," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 197–201, 2015.

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

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

[13] R. Nassif, C. Richard, A. Ferrari, and A.H. Sayed. Multitask diffusion LMS with sparsity-based regularization. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3207–3211, 2015.

[14] J. Chen, C. Richard, A. O. Hero, and A. H. Sayed, "Diffusion LMS for multitask problems with overlapping hypothesis subspaces," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept. 2014, pp. 1–6, IEEE.

[15] X. Zhao and A. H. Sayed, "Distributed clustering and learning over networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 13, pp. 3285–3300, July 2015.

[16] S. Khawatmi, A.M. Zoubir, and A.H. Sayed. Decentralized clustering over adaptive networks. *Proc. 2015 European Signal Processing Conference*, 2696–2700, 2015.

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

[18] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[19] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[20] T. S. Jaakkola, "Tutorial on variational approximation methods," in *Advanced Mean Field Methods: Theory and Practice*, 2000, pp. 129–159.

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

[22] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent Dirichlet allocation," in *Advances in Neural Information Processing Systems 23*, pp. 856–864. Curran Associates, Inc., 2010.

[23] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M.I. Jordan, "Streaming variational Bayes," in *Advances in Neural Information Processing Systems 26*, pp. 1727–1735. Curran Associates, Inc., 2013.

[24] M. Sato, "Online model selection based on the variational Bayes," *Neural Computation*, vol. 13, no. 7, pp. 1649–1681, 2001.

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

[26] D. A. Knowles and T. Minka, "Non-conjugate variational message passing for multinomial and binary regression," in *Advances in Neural Information Processing Systems 24*, pp. 1701–1709. Curran Associates, Inc., 2011.

[27] M. P. Wand, "Fully simplified multivariate normal updates in non-conjugate variational message passing," *Journal of Machine Learning Research*, vol. 15, pp. 1351–1369, 2014.