A DIFFUSION-BASED DISTRIBUTED EM ALGORITHM FOR DENSITY ESTIMATION IN WIRELESS SENSOR NETWORKS

Silvana Silva Pereira*, Alba Pagès-Zamora* and Roberto López-Valcarce†,
*SPCOM Group, Universitat Politècnica de Catalunya-Barcelona Tech, Spain.
†GPSC, Universidade de Vigo, Spain.

ABSTRACT

Distributed implementations of the Expectation-Maximization (EM) algorithm reported in literature have been proposed for applications to solve specific problems. In general, a primary requirement to derive a distributed solution is that the structure of the centralized version enables the computation involving global information in a distributed fashion. This paper treats the problem of distributed estimation of Gaussian densities by means of the EM algorithm in wireless sensor networks using diffusion strategies, where the information is gradually diffused across the network for the computation of the global functions. The low-complexity implementation presented here is based on a two time scale operation for information averaging and diffusion. The convergence to a fixed point of the centralized solution has been studied and the appealing results motivates our choice for this model. Numerical examples provided show that the performance of the distributed EM is, in practice, equal to that of the centralized scheme.

Index Terms— Distributed algorithms, expectation-maximization algorithms, Gaussian distribution, unsupervised learning, wireless sensor networks.

1. INTRODUCTION

Density estimation belongs to the general class of unsupervised learning problems and is widely studied in fields like data mining and machine learning, particularly in applications like pattern classification and pattern recognition for image or speech analysis, but also for clustering. In presence of hidden variables, the EM algorithm provides a means to iteratively compute the maximum likelihood (ML) estimator when the data follows an exponential distribution [1]. Starting from an initial set of estimates, the method alternates between an expectation (E) step, where the expected log-likelihood function of the observations is evaluated using the current estimates, and a maximization (M) step, where the log-likelihood function of the E-step is maximized with respect to the estimates. Whereas the E-step performs computations using local information only, the M-step performs computations using global information. A distributed implementation of the algorithm for a wireless sensor network (WSN) entails therefore a reformulation of the computations such that they can be performed locally at each node. Related contributions to the literature propose distributed EM implementations where the global sufficient statistics are computed using for instance aggregation [2–5], or using a consensus-based scheme [6–8]. Aggregation strategies require the assignment of routing paths or junction trees within the network, whereas consensus type strategies result in an increase in the total number of iterations until convergence, since a consensus algorithm [9, 10] is fully executed at each M-step. Based on diffusion strategies [11–13], a distributed EM is proposed in [14] where the authors use the results from [15] to show that their implementation is a Robbins-Monro stochastic approximation to the centralized EM approach. A similar scheme is proposed in [16] for tracking applications using particle filtering, and a diffusion adaptation scheme is proposed in [17] for learning in Gaussian mixture models (GMM).

In this paper we present a novel Diffusion-Based Distributed EM (DB-DEM) algorithm for density estimation and classification in WSNs. The method is based on a distributed algorithm derived for ML estimation in presence of unreliable observations [18]. Whereas the problem of unreliable observations in [18] could be seen as a particular case of a GMM, we address here the general case of density estimation under the assumption of GMM’s. The main idea behind the DB-DEM algorithm is that the diffusion of the information across the network is embedded in the iterative update of the parameter estimates. Therefore, the M-step could be seen as a two time scale operation where a faster term for information diffusion is combined with a slower term for information averaging, and differing therefore from [14–17] which assume a single time scale operation. As opposed to most diffusion adaptive techniques which consider constant step-sizes, e.g. [16, 17], we consider here time-varying step-sizes.

The work of S.S. Pereira and A. Pagès-Zamora is supported by the Spanish Government, ERDF funds (CONSOLIDER-INGENIO 2010 CSD2008-00010 COMONSENS and TEC2010-19171 MOSAIC) and the Catalan Government (2009SGR-01236 AGAUR).

The work of R. López-Valcarce is supported by the Spanish Government, ERDF funds (TEC2010-21245-C02-02/TCM DYNACS, CONSOLIDER-INGENIO 2010 CSD2008-00010 COMONSENS) and the Galician Government (CN 2012/260 AtlantTIC).
in order to converge to the desired values. The advantage of the DB-DEM with respect to the consensus-based EM algorithm in [6–8] is a significant reduction in the total number of iterations, since only one averaging operation is performed at each M-step. This reduction in the number of iterations can be translated into energy savings, a critical issue especially in large-scale deployments. Although not included due to lack of space, the convergence of the DB-DEM algorithm to a fixed point of the centralized EM solution has been studied for different choices of the control parameters. Numerical examples provided here show that the performance of the DB-DEM is, in practice, equal to that of the centralized scheme. The paper is organized as follows. In section 2 we describe the observation model and derive the expressions for the centralized EM algorithm. Section 3 presents the DB-DEM for density estimation in WSNs under the assumption of GMM’s. Simulations results and conclusions are presented in sections 4 and 5 respectively.

2. PROBLEM FORMULATION

2.1. Observation Model

Consider a set of \( N \) independent observations indexed with \( i = 1, \cdots, N \). A D-dimensional observation vector \( y_i \) is formed with variables drawn from a set of \( M \) Gaussian densities with probability distribution \( \mathcal{N}(y_i | \mu_m, \Sigma_m) \), where \( \mu_m \in \mathbb{R}^{D \times 1} \) is the vector of means and \( \Sigma_m \in \mathbb{R}^{D \times D} \) is the covariance matrix \( \forall m \in \{1, \cdots, M\} \). Let \( z \in \mathbb{R}^{N \times 1} \) be an indicator vector with the \( i^{th} \) entry \( z_i = m \) if the \( i^{th} \) observation belongs to the \( m^{th} \) density. The probability density function (pdf) for \( y_i \) is given by

\[
 f(y_i) = \sum_{m=1}^{M} \pi_m \mathcal{N}(y_i | \mu_m, \Sigma_m) \tag{1}
\]

where \( \pi_m \) is the prior probability of observing the \( m^{th} \) Gaussian, i.e., \( \pi_m = \Pr(z_i = m) \) for all \( i \) with \( \sum_{m=1}^{M} \pi_m = 1 \). We regard the vector \( y = [y_1^T \cdots y_N^T]^T \) as the incomplete observation and \( \{y, z\} \) as the complete observation. Since \( z \) is unknown, we treat it as a random variable (r.v.). The set of unknown parameters to be estimated is therefore the tuple \( \theta = \{\pi, \mu, \Sigma\} \), where \( \pi = \{\pi_1, \cdots, \pi_M\} \), \( \mu = \{\mu_1, \cdots, \mu_M\} \), and \( \Sigma = \{\Sigma_1, \cdots, \Sigma_M\} \).

2.2. Centralized EM Algorithm

The ML estimator has the desirable properties of being unbiased and asymptotically efficient as the number of samples goes to infinity, and the EM algorithm is a numerical method to iteratively compute the ML estimates in the presence of incomplete observations. Under mild conditions, it is guaranteed to converge to a -local- maximum of the likelihood function [19]. Assuming a centralized approach in which the vector of observations \( y \) is available, at time \( t \) one performs the following:

1. **E-step**: given an estimate \( \hat{\theta}^{t-1} = \{\hat{\pi}^{t-1}, \hat{\mu}^{t-1}, \hat{\Sigma}^{t-1}\} \), where \( \hat{\pi} = \{\hat{\pi}_1, \cdots, \hat{\pi}_M\} \), \( \hat{\mu} = \{\hat{\mu}_1, \cdots, \hat{\mu}_M\} \), \( \hat{\Sigma} = \{\hat{\Sigma}_1, \cdots, \hat{\Sigma}_M\} \), compute the conditional expectation

\[
 Q(\hat{\theta} ; \hat{\theta}^{t-1}) = E_y \left[ \log f(y, z | \hat{\theta}) | \hat{\theta}^{t-1}, y \right] \tag{2}
\]

where \( \hat{\theta} \) denotes a trial value of \( \theta \).

2. **M-step**: obtain the estimate for the next iteration as

\[
 \hat{\theta}^t = \arg \max_{\theta} Q(\hat{\theta} ; \theta^{t-1}).
\]

Note that we can write \( Q(\hat{\theta} ; \theta^{t-1}) = \sum_{i=1}^{N} Q_i(\hat{\theta} ; \theta^{t-1}) \) due to independence of the pairs \( \{z_i, y_i, \forall i\} \), where

\[
 Q_i(\hat{\theta} ; \theta^{t-1}) = E_{z_i} \left[ \log f(y_i, z_i | \hat{\theta}) | \theta^{t-1}, y_i \right]
\]

Therefore,

\[
 Q_i(\hat{\theta} ; \theta^{t-1}) = \sum_{m=1}^{M} \hat{\pi}_{m,i} \log \left[ \hat{\pi}_m \mathcal{N}(y_i | \hat{\mu}_m, \Sigma_m) \right]
\]

where

\[
 \hat{\pi}_{m,i} = \Pr(z_i = m | \theta^{t-1}, y_i) = \frac{\hat{\pi}_m \mathcal{N}(y_i | \hat{\mu}_m, \Sigma_m)}{\sum_{l=1}^{M} \hat{\pi}_{l,i} \mathcal{N}(y_i | \hat{\mu}_l, \Sigma_l)} \tag{4}
\]

is the (a posteriori) probability of having observed the \( m^{th} \) Gaussian density given \( y_i \) and the estimates from the previous iteration. Hence, the expression in (2) is equal to

\[
 \sum_{m=1}^{M} \sum_{i=1}^{N} \hat{\pi}_{m,i} \log \hat{\pi}_m + \sum_{m=1}^{M} \sum_{i=1}^{N} \hat{\pi}_{m,i} \log \mathcal{N}(y_i | \hat{\mu}_m, \Sigma_m). \tag{5}
\]

The estimator for the a priori probability for the \( m^{th} \) density is obtained maximizing the expression above with respect to \( \hat{\pi}_m \) subject to the constraint \( \sum_{m=1}^{M} \hat{\pi}_m = 1 \), whereas the estimators for the mean and the covariance matrix are obtained maximizing it with respect to \( \hat{\mu}_m \) and \( \Sigma_m \), i.e.,

\[
 \hat{\pi}_m^t = \frac{1}{N} \sum_{i=1}^{N} \hat{\pi}_{m,i} \tag{6a}
\]

\[
 \hat{\mu}_m^t = \frac{\sum_{i=1}^{N} y_i \hat{\pi}_{m,i}}{\sum_{i=1}^{N} \hat{\pi}_{m,i}} \tag{6b}
\]

\[
 \hat{\Sigma}_m^t = \frac{\sum_{i=1}^{N} (y_i - \hat{\mu}_m^t)(y_i - \hat{\mu}_m^t)^T \hat{\pi}_{m,i}}{\sum_{i=1}^{N} \hat{\pi}_{m,i}} \tag{6c}
\]

It is well known that the EM algorithm is guaranteed to converge to a local maximum of the likelihood function, but it is
sensitive to the initialization of the parameters. Therefore, in order to start up the iteration, a suitable initializer is needed. Notice that the a posteriori probabilities in (4) require knowledge of local information only, whereas the estimates in (6) require knowledge of global information. Therefore, a distributed implementation of the EM algorithm entails the computation of (6) in a distributed fashion.

3. DIFFUSION-BASED EM ALGORITHM

Based on the distributed implementation of the EM algorithm in [18], we propose a scheme where the summations among all observations in (6) are computed via interaction. Let us assume a WSN composed of \( N \) nodes where the communications for each node are restricted to a small neighborhood. The information flow among the nodes is described by means of an undirected graph \( G = \{ V, E \} \), where \( V \) is the set of vertices (nodes) and \( E \) is the set of edges (links) \( e_{ij} \in V \} \), \( \{ \} \). The set of neighbors of node \( i \) is denoted as \( \mathcal{N}_i = \{ j \in V : e_{ij} \in E \} \) for all \( i = 1, \cdots, N \).

We further assume that the network is connected such that there is a path between any pair of nodes \( \{ i, j \} \in \mathcal{V} \). Consider then a weight matrix \( W \in \mathbb{R}^{N \times N} \) with a nonzero \( \{ij\}^{\text{th}} \) entry \( W_{ij} \) only if \( j \in \mathcal{N}_i \) and \( i \in \mathcal{N}_j \). \( W \) is assumed symmetric and satisfies

\[
1^T W = 1^T, \quad W 1 = 1, \quad \rho(W - \frac{11^T}{N}) < 1, \]

where \( 1 \) is an all-ones vector and \( \rho(\cdot) \) denotes spectral radius [9].

Further, for any node \( i \) and any time instant \( k \geq 0 \) define the following vector of length \( L = M(1+D+D^2)+1 \)

\[
f(i, k) = [1, \hat{s}_{1,i}^k, \cdots, \hat{s}_{M,i}^k, \hat{y}_i^T, \hat{y}_i^T, \cdots, \hat{y}_i^T, \text{vec}(y_i^T), \cdots, \text{vec}(y_i^T)]^T \quad \text{(7a)}
\]

where \( \text{vec}(\cdot) \) denotes the operation of vectorizing a matrix. Note that the entries from 2 to \( M+1 \) of \( f(i, k) \) are the estimates at time \( k \) of the a posteriori probabilities, the following \( D \times M \) entries correspond to the product of the a posteriori probabilities and the observation vector, whereas the last \( D^2 \times M \) entries correspond to the product of the a posteriori probabilities and the entries of \( y_i^T \).

3. M-Step: For \( l = 1, \cdots, L \) compute:

\[
\phi_l(i, k) = (1 - \beta_k) \sum_{j=1}^N W_{ij} \phi_l(j, k-1) + \alpha_k \sum_{j=1}^N W_{ij} f_l(j, k) \quad \text{(8)}
\]

and update the parameters as follows:

\[
\hat{\phi}_m^k = \phi_{m+1}(i, k), \quad \hat{\mu}_m^k = \left( \phi_{m+1}(i, k) \right)^T, \quad \hat{\Sigma}_m^k = \text{vec}^{-1}_D \left( \frac{\phi_{m+1}(i, k)}{\phi_{m+1}(i, k)} - \hat{\mu}_m^k [\hat{\mu}_m^k]^T \right)
\]

where \( \Delta \) and \( \Delta \Sigma \) define the intervals of subindices to choose the entries for the computations. Finally, note also that for \( k = 1 \), the first term on the right-hand side of (8) is zero, so the initial condition \( \phi_1(i, 0) \) is irrelevant. Based on convergence results not included here due to lack of space, we claim that the DB-DEM converges to a fixed point of the centralized EM solution for a proper initialization of the parameters, provided that assumption \( i) \) holds and assuming the step-sizes \( \alpha_k \) and \( \beta_k \) in (9).

4. Repeat steps 2 and 3 until convergence.

### Table I: DB-DEM Estimator

<table>
<thead>
<tr>
<th>( l )</th>
<th>( \phi_m^k )</th>
<th>( \mu_m^k )</th>
<th>( \Sigma_m^k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \frac{\phi_{m+1}(i, k)}{\phi_{m+1}(i, k)} )</td>
<td>( \frac{\phi_{m+1}(i, k)}{\phi_{m+1}(i, k)} )</td>
<td>( \frac{\phi_{m+1}(i, k)}{\phi_{m+1}(i, k)} )</td>
</tr>
</tbody>
</table>

Remark that the computation of \( \phi_l(i, k) \) in (8) at each node \( i \), at time \( k \) and for all \( l = 1, \cdots, L \), entails communication with neighboring nodes only, since \( \phi_l(j, k) \) is nonzero for \( j \in N_i \) only. The step-size parameter \( \alpha_k \) in (8) controls the rate of information diffusion across the network, whereas the parameter \( \beta_k \) controls the rate of information averaging. Note that \( \Delta \) and \( \Delta \Sigma \) define the intervals of subindices to choose the entries used for the computations. Finally, note also that for \( k = 1 \), the first term on the right-hand side of (8) is zero, so the initial condition \( \phi_1(i, 0) \) is irrelevant. Based on convergence results not included here due to lack of space, we claim that the DB-DEM converges to a fixed point of the centralized EM solution for a proper initialization of the parameters, provided that assumption \( i) \) holds and assuming the step-sizes \( \alpha_k \) and \( \beta_k \) in (9). In fact, the convergence to a fixed point of the centralized solution has been studied for different values of the control parameters \( \alpha_k \) and \( \beta_k \), showing that an attractive feature of the DB-DEM is its robustness with respect to this choice, as it has been verified by the numerical results.

4. SIMULATION RESULTS

The analytical results are supported with computer simulations of a WSN composed of \( N = 50 \) nodes randomly deployed on a unit square with connectivity radius \( r_c = 0.2 \). The step-size parameter \( \beta_k = 1/k^2 \) is chosen with \( \delta = 0.85 \), and we assume mixture model of \( M = 2 \) Gaussians. For simplicity, we consider \( D = 1 \), i.e., the 1-dimensional case with \( \mu_1 = 2, \mu_2 = 4, \sigma_1 = 0.4, \sigma_2 = 0.2, \pi_1 = 0.3, \pi_1 = 0.7 \) where
the nodes initialize the estimates equally with $\mu_0^1 = 0$, $\mu_0^2 = 5$, $\sigma_0^1 = 0.5$, $\sigma_0^2 = 0.3$ and $\pi_0^1 = \pi_0^2 = 0.5$. The DB-DEM algorithm is run assuming a Metropolis weight matrix with entries defined as

$$W_{ij} = \left\{ \begin{array}{ll} \frac{1}{1 + \max\{|N_i|, |N_j|\}} & j \in N_i \\ \frac{1}{1 - \sum_{k \in N_i} W_{ik}} & i = j \\ 0 & \text{otherwise} \end{array} \right. $$

where $|\cdot|$ denotes cardinality [21]. In summary, the weights assigned to the data arriving from neighboring nodes are decided locally. Fig. 1 depicts the evolution of the mean estimates for each node as a function of the iteration index $k$ for one random realization, whereas Fig. 2 depicts the evolution of $\hat{\zeta}_{1,i}$ and $\hat{\zeta}_{2,i}$. Clearly, the nodes are able to estimate the parameters for each density. In particular, for classification applications where the nodes make a decision based on the a posteriori probabilities, we observe from the simulations that a correct classification can be made after approximately 50 iterations. Finally, Fig. 3 depicts the log-likelihood function $Q(\hat{\theta}^{k-1}; \hat{\theta}^{k-1})$ computed as in (5) for the DB-DEM estimator (solid line) along with the curves for the centralized EM solution (dashed-dotted line) and a consensus-based EM (dashed line) [6–8]. In order to make a fair comparison, and since the curves for the DB-DEM and for the centralized EM have converged after approximately 100 and 10 iterations respectively, we run a consensus-based EM with 10 averaging iterations at each M-step. Clearly, the truncation of the averaging computation after only 10 iterations is not enough to estimate the parameters correctly. In summary, a higher number of averaging iterations per M-step is required by the consensus-based scheme to obtain better results.

5. CONCLUSIONS

The problem of distributed density estimation and classification under the assumption of Gaussian mixtures models has been addressed in this paper. A distributed expectation-maximization estimator has been derived for wireless sensor networks using diffusion strategies, where the information diffusion across the network is embedded in the iterative update of parameter estimates. The advantage with respect to prior art relies in the simplicity of the iterative algorithm including the two time scale M-step, a reduction in the number of iterations with respect to consensus-based schemes, the robustness with respect to the choice of step-sizes, and the existing convergence analysis supporting the model. Numerical examples provided show that the performance of the distributed EM is, in practice, equal to that of the centralized EM solution.
6. REFERENCES


