Distributed Estimation in Wireless Sensor Networks via Variational Message Passing

Yanbing Zhang and Huaiyu Dai
Department of Electrical and Computer Engineering
NC State University
Raleigh, N.C. 27695
{yzhang, Huaiyu_Dai}@ncsu.edu

Abstract – **In this paper, a variational message passing framework is proposed for Markov random fields. Analogous to the traditional belief propagation algorithm, variational message passing is performed by only exchanging messages between adjacent nodes in a graph and updating local estimations, but with more energy and computation saving achieved. Explicit forms for distributions in the exponential family are derived and applied to a distributed estimation problem in wireless sensor networks. Furthermore, structured variational methods are explored to improve the estimation accuracy, whose performance is elaborated in a Gaussian Markov random field, by both theoretical analysis and simulation results. To our best knowledge, this is the first work to explicitly apply the structured variational approach in wireless sensor networks.**

Keywords: **Variational method, mean field, sensor networks**

I. INTRODUCTION

Distributed and energy efficient in nature, message passing algorithms, like belief propagation (BP) [1], are attractive for wireless sensor applications [2]. However, BP and other relevant algorithms are known not always to converge in general graphs and less powerful in handling continuous variables. Although sampling methods can be used for approximation, they are computationally intensive, stochastic in nature and difficult to analyze.

Variational methods [3], which have been used in physics, artificial intelligence, image processing, etc. for years, involve the minimization of the Kullback–Leibler (KL) divergence between the target probability distribution and some "simpler" variational distribution. They can be applied to a variety of graphical models, whether acyclic or cyclic, discrete or continuous. Actually, BP can be regarded as a special case of variational methods with respect to optimizing the "Bethe free energy" [4]. However, the application of variational methods in wireless sensor networks (WSN) is still largely unexplored [5].

Probabilistic inference lies in the core of a wide range of sensor applications, like localization, tracking and time synchronization. In general, the objective of distributed inference is to compute the posterior probability, given a set of observations and some underlying graph model. Variational message passing (VMP) has been proposed in [7] for directed Bayesian networks, where many advantages over conventional inference methods were observed. In this paper, we derive a variational message passing framework for Markov random fields (MRF), which is considered as more suitable for modeling sensor networks. In particular, we formulate explicit message passing rules for distributions in the exponential family, which covers a large class of probabilistic models, and show that the proposed framework can be applied on the distributed time synchronization problem in wireless sensor networks.

Inference on a flat network enjoys the benefits of uniformity in message formation and modeling convenience. However, higher accuracy and faster convergence can be achieved by organizing the whole network into a more structured one. Clustering technique has been utilized on improving network performance, lifetime and scalability [9]. Meanwhile, exploiting tractable substructure in graphs is also proven an effective way to improve accuracy of variational methods [10][11]. However, structured variational approaches are mainly studied in the artificial intelligence area, and little consideration is given on their applications in real networks. The graph partition algorithms (e.g.[12]) in this context may not be suitable for wireless network application either, due to complexity concerns. This motivates us to further investigate how to exploit substructures of networks to improve variational methods in real systems, where a new distributed clustering algorithm may also be necessary. To our best knowledge, this is the first work to explicitly apply the structured variational approach in wireless networks.

The rest part of this paper is organized as follows. Section II formulates the general variational method and derives its message passing form for MRF, which is also exemplified by a distributed time synchronization application in sensor networks. The application of structured variational methods is discussed in Section III, where an accessorial distributed clustering scheme is also proposed. Section IV presents some supporting simulation results. Finally, concluding remarks are given in Section V.

II. VARIATIONAL MESSAGE PASSING IN MRF

A. Variational Method and Mean Field Approach

Assume $X = [X_1, ..., X_N]^T$ is an N-dimension random vector and $y = [y_1, ..., y_M]^T$ is the observation (M is not necessarily equal to N, as the following example shows). The basic idea of variation methods to estimate the posterior probability $P(X \mid y)$ is approximating it by a variational probability $Q(X)$ through the means of minimizing the KL divergence

This work was supported in part by the National Science Foundation under Grant CCF-0515164 and CNS-0721815.

$$KL(Q\|P)=\int_X Q(X)\log\frac{Q(X)}{P(X\mid y)}dX=-\mathbf{H}(Q)-\langle\log P(X\mid y)\rangle_{Q(X)}, \quad (1)$$

where $\mathbf{H}(Q)$ is the entropy of $Q(X)$ and $\langle\ \rangle_{Q(X)}$ refers to expectation with respect to $Q(X)$. Therefore the optimal marginal should be

$$Q^*(X)=\arg\max_{Q(X)}\left\{\langle\log P(X\mid y)\rangle_{Q(X)}+\mathbf{H}(Q)\right\}. \quad (2)$$

One simple and widely adopted form for variational distributions is to assume all variables are independent:

$$Q(X)=\prod_{i=1}^{N}Q_i(X_i), \quad (3)$$

which is also referred to as the mean field (MF) approach. With this setting, it yields

$$\begin{aligned}&\langle\log P(X\mid y)\rangle_{Q(X)}+\mathbf{H}(Q)\\&=\int_X\prod_i Q_i(X_i)\log P(X\mid y)dX-\int_X\prod_i Q_i(X_i)\sum_i\log Q_i(X_i)dX \quad (4)\\&=\int_{X_k}Q_k(X_k)\langle\log P(X\mid y)\rangle_{\prod_{i\neq k}Q_i(X_i)}dX_k+\mathbf{H}(Q_k)+\sum_{i\neq k}\mathbf{H}(Q_i).\end{aligned}$$

It is easy to verify that maximizing this objective function with respect to the marginal $Q_k(X_k)$ gives the standard Gibbs' distribution [3]

$$\log Q_k^*=\langle\log P(X\mid y)\rangle_{\prod_{i\neq k}Q_i(X_i)}+\text{const}. \quad (5)$$

B. Variational Message Passing in Pairwise MRF

Consisted of homogeneous nodes, Markov random fields (MRF) exhibit modeling convenience for sensor networks mostly. A pairwise MRF[1] is an undirected graph (V,E) with maximum cliques of size 2, where each node $i\in V$ is associated with a random variable X_i. Define for each node a local potential function $\phi_i(X_i)$, and for each edge $(i,j)\in E$ a compatibility function $\psi_{ij}(X_i,X_j)$, the Hammersley-Clifford theorem [1] dictates that the posterior probability can be represented as the product of the potential functions and compatibility functions:

$$P(X\mid y)=\frac{1}{Z}\prod_{i\in V}\phi_i(X_i)\prod_{(i,j)\in E}\psi_{i,j}(X_i,X_j), \quad (6)$$

where Z is a normalization factor (called partition function in physics).

The exponential family includes a variety of widely used distributions in practice, like Bernoulli, Gaussian and Poisson [6]. In the following, we assume that both the potential functions $\{\phi_i\}$ and the compatibility functions $\{\psi_{ij}\}$ belong to the exponential family, i.e. they take the form

$$\phi_i(X_i,y)\propto\exp\left\{\boldsymbol{\theta}_i^T\boldsymbol{\eta}_i(X_i)+g_i(\boldsymbol{\theta}_i)\right\}, \quad (7)$$

$$\psi_{ij}(X_i,X_j)\propto\exp\left\{\boldsymbol{\theta}_{ij}^T\boldsymbol{\eta}_{ij}(X_i,X_j)+g_{ij}(\boldsymbol{\theta}_{ij})\right\}, \quad (8)$$

where $\boldsymbol{\theta}$'s and $\boldsymbol{\eta}$'s are referred to as the natural parameters and sufficient statistics, respectively.

To facilitate variational message passing, rearrange $\boldsymbol{\theta}_{ij}^T\boldsymbol{\eta}_{ij}(X_i,X_j)$ as a function of X_i:

$$\boldsymbol{\theta}_{ij}^T\boldsymbol{\eta}_{ij}(X_i,X_j)=(\boldsymbol{\theta}'_{ij})^T\boldsymbol{\eta}'_{ij}(X_i). \quad (9)$$

Note $\boldsymbol{\theta}'_{ij}$ may contain the term X_j. Let $\tilde{\boldsymbol{\eta}}_i(X_i)$ be the *union* of sufficient statistics $\boldsymbol{\eta}_i(X_i)$ and $\boldsymbol{\eta}'_{ij}(X_i)$, then the product of parameters and statistics in ϕ_i and ψ_{ij} can be always expressed in term of $\tilde{\boldsymbol{\eta}}_i(X_i)$ (by inserting zeros in the appropriate positions in the parameter vectors if needed), i.e.,

$$\boldsymbol{\theta}_i^T\boldsymbol{\eta}_i(X_i)=\tilde{\boldsymbol{\theta}}_i^T\tilde{\boldsymbol{\eta}}_i(X_i), \quad (10)$$

$$\boldsymbol{\theta}_{ij}^T\boldsymbol{\eta}_{ij}(X_i,X_j)=(\boldsymbol{\theta}'_{ij})^T\boldsymbol{\eta}'_{ij}(X_i)=\tilde{\boldsymbol{\theta}}_{ij}^T\tilde{\boldsymbol{\eta}}_i(X_i). \quad (11)$$

We call $\tilde{\boldsymbol{\theta}}_i$ and $\tilde{\boldsymbol{\theta}}_{ij}$ the extended natural parameters and $\tilde{\boldsymbol{\eta}}_i(X_i)$ the extended sufficient statistics.

It can be shown that the optimal approximation Q_i^* is also a member of the exponential family with sufficient statistics $\tilde{\boldsymbol{\eta}}_i(X_i)$ and natural parameter vector

$$\boldsymbol{\theta}_i^*=\left\langle\tilde{\boldsymbol{\theta}}_i\right\rangle_{\prod_{k\neq i}Q_k(X_k)}+\sum_{j\in\Gamma_i}\left\langle\tilde{\boldsymbol{\theta}}_{ij}\right\rangle_{\prod_{k\neq i}Q_k(X_k)}, \quad (12)$$

where Γ_i stands for the neighboring node set of node *i*. Note $\left\langle\tilde{\boldsymbol{\theta}}_i\right\rangle_{\prod_{k\neq i}Q_k(X_k)}=\tilde{\boldsymbol{\theta}}_i$ and $\left\langle\tilde{\boldsymbol{\theta}}_{ij}\right\rangle_{\prod_{k\neq i}Q_k(X_k)}=\left\langle\tilde{\boldsymbol{\theta}}_{ij}\right\rangle_{Q_j(X_j)}$, We have the variational message passing rules in MRF for the exponential family as:

- Messaging passing: $m_{j\to i}^{(n)}=\left\langle\tilde{\boldsymbol{\theta}}_{ij}\right\rangle_{Q_j^{(n-1)}(X_j)}$;
- Parameter updating: $\boldsymbol{\theta}_i^{(n)}=\tilde{\boldsymbol{\theta}}_i+\sum_{j\in\Gamma_i}m_{j\to i}^{(n)}$.

In other words, the message is the expectation of the extended parameter in the corresponding edge compatibility function. The node sums up all the messages from its neighbors and the extended parameter in its own potential function to get the updated parameter vector of the optimal variational distribution. Notice its similarity with belief propagation (in the logarithm scale).

C. Example-Distributed Time Synchronization in Wireless Sensor Networks

We consider a distributed time synchronization problem as an example. Suppose in a sensor network, a pair of time differences is observed on each edge $(i,j)\in E$:

$$\begin{aligned}y_{ij}&=x_i-x_j+\varepsilon_{ij}\\y_{ji}&=x_j-x_i+\varepsilon_{ji}\end{aligned}, \quad (13)$$

where $x_i, i\in V$, is the clock offset of node *i*, which we want to estimate. Stacking up the above equations yields

$$Y=HX+\varepsilon, \quad (14)$$

where X is a $|V|\times 1$ state vector, Y is a $2|E|\times 1$ observation vector, H is a $2|E|\times|V|$ matrix determined by network topology, and ε is the noise vector.

[1] MRF with higher order cliques can always be converted into an equivalent pairwise MRF.

Let $Z \sim \mathcal{N}(\mu, \Sigma)$ be a Gaussian random vector with mean μ and positive definite covariance Σ. One can define a new set of parameters (ϑ, Λ) by $\vartheta = \Sigma^{-1}\mu$, $\Lambda = \Sigma^{-1}$, and alternatively denote $Z \sim \mathcal{N}^{-1}(\vartheta, \Lambda)$. Using these notations, we dictate

$$P(X) \sim \mathcal{N}(\mu, \Sigma) \sim \mathcal{N}^{-1}(\Sigma^{-1}\mu, \Sigma^{-1}),\ P(\varepsilon) \sim \mathcal{N}(0, R), \quad (15)$$

where μ and $\Sigma = [\Sigma_{ij}]_{|V|\times|V|}$ respectively represent the mean vector and covariance matrix of clock offset and $R = \sigma^2 I_{2|E|\times 2|E|}$ is the covariance of noise. Define matrix

$$U = \Sigma^{-1} + H^T R^{-1} H, \quad (16)$$

then we have

$$P(X \mid y) \sim \mathcal{N}(U^{-1}H^T R^{-1} y, U^{-1}) \sim \mathcal{N}^{-1}(H^T R^{-1} y, U). \quad (17)$$

The mean filed approach imposes that the approximate distribution takes the form

$$Q(X) = \prod_{i\in V} Q_i(x_i) \propto K \prod_{i\in V} \exp\left(\frac{-(x_i - q_i)^2}{2(q_{ii} - q_i^2)}\right), \quad (18)$$

where q_i and q_{ii} are the first and second moments of $Q_i(x_i)$. Noticing that $H = [B^T \ -B^T]^T$, with B being the incidence matrix of the underlying network graph and satisfying

$$(BB^T)_{ij} = \begin{cases} |\Gamma_i| & \text{if } i = j \\ -1 & \text{if } j \in \Gamma_i \\ 0 & \text{otherwise,} \end{cases} \quad (19)$$

we can get the extended natural parameters of the node potential and compatibility function as

$$\tilde{\boldsymbol{\theta}}_i = \begin{bmatrix} 0 \\ -\frac{1}{2}\left(\Sigma_{ii} + \frac{2|\Gamma_i|}{\sigma^2}\right) \end{bmatrix} \text{ and } \tilde{\boldsymbol{\theta}}_{ij} = \begin{bmatrix} \frac{y_{ij} - y_{ji} + 2X_j}{\sigma^2} \\ 0 \end{bmatrix}. \quad (20)$$

Using the variational message passing we derived in Section II.B, the natural parameter of the optimal variational distribution is updated by

$$\begin{aligned} \boldsymbol{\theta}_i^{(n)} &= \tilde{\boldsymbol{\theta}}_i + \sum_{j\in\Gamma_i} \left\langle \tilde{\boldsymbol{\theta}}_{ij} \right\rangle_{Q_j^{(n-1)}(X_j)} \\ &= \begin{bmatrix} 0 \\ -\frac{1}{2}\left(\Sigma_{ii} + \frac{2|\Gamma_i|}{\sigma^2}\right) \end{bmatrix} + \sum_{j\in\Gamma_i} \left\langle \begin{bmatrix} \frac{y_{ij} - y_{ji} + 2X_j}{\sigma^2} \\ 0 \end{bmatrix} \right\rangle_{Q_j^{(n-1)}(X_j)} \\ &= \begin{bmatrix} \sum_{j\in\Gamma_i} (y_{ij} - y_{ji} + 2q_j^{(n-1)})/\sigma^2 \\ -\frac{1}{2}\left(\Sigma_{ii} + \frac{2|\Gamma_i|}{\sigma^2}\right) \end{bmatrix} \end{aligned}. \quad (21)$$

This leads to the variational iteration formula for distributed time synchronization:

$$q_i^{(n)} = \frac{1}{\sigma^2\Sigma_{ii} + 2|\Gamma_i|} \sum_{j\in\Gamma_i} \left(y_{ij} - y_{ji} + 2q_j^{(n-1)}\right), \quad (22)$$

Which, when $\Sigma_{ii} = 0$, coincides with the result obtained by Giridhar and Kumar through a least-squares approach in [8].

The iteration formula can be represented in matrix form as

$$\mathbf{q}^{(n)} = (\sigma^2\Sigma + 2\Gamma)^{-1}(2\Gamma - H^T H)\mathbf{q}^{(n-1)} + (\sigma^2\Sigma + 2\Gamma)^{-1} H^T y, \quad (23)$$

where $\Gamma = diag\{|\Gamma_i|\}$. Denote matrix

$$K = (\sigma^2\Sigma + 2\Gamma)^{-1}(2\Gamma - H^T H) = 2 \times \mathrm{diag}\left\{(\sigma\Sigma_{ii} + 2|\Gamma_i|)^{-1}\right\} \times A, \quad (24)$$

where A is the adjacency matrix of the network graph. If $\sigma > 0$ and $\Sigma_{ii} > 0$, the spectral radius of K satisfies

$$\rho(K) \le \max_i \sum_{j=1}^{n} K_{ij} = \max_i \frac{2}{\sigma^2\Sigma_{ii} + 2|\Gamma_i|} |\Gamma_i| < 1. \quad (25)$$

Therefore the iteration converges.

III. DISTRIBUTED INFERENCE IN STRUCTURED NETWORKS

A. Structured Mean Field

The above discussion reveals that the naive mean field approach is especially attractive for its computational simplicity. However, it may not yield sufficient accuracy and fast convergence due to the independence restriction on distributions. A natural idea for improvement is to integrate exact probabilistic inference algorithms with the naive mean field method. In other words, if we could identify some tractable substructures (clusters) within the large graph model, then these substructures could be handled by exact methods. This is referred to as the structured mean field (SMF) method [3].

Given s clusters identified by some means, denoted by $C_1, ..., C_s$, the variational distribution Q factors across the clusters as

$$Q = \prod_{i=1}^{s} Q_i(X_{C_i}). \quad (26)$$

The update equations resulting from the structured approximation are exactly analogous to the naive mean field method. The intuition behind is that we can interpret the structured mean filed method as a mean field approach over "mega variables" X_{C_i}. Within a cluster, any exact or more accurate inference algorithms can be used, like junction tree or (loopy) belief propagation. We can expect, as the cluster size (the number of nodes in one cluster) increases, exact algorithms are performed on more nodes, which should result in a better approximation, while this also inevitably increases the computation burden. So it's important to choose an appropriate cluster size and associated structure to balance approximate accuracy and computation complexity.

B. Distributed Clustering Algorithm

Identifying substructures is a non-trivial problem, for which very few automated solutions have been proposed [12]. In the following, we describe a simple distributed algorithm to construct a clustered structure within the large network, which will be used as a basis for structured variational message passing.

More delicate clustering schemes suitable for structured inference in wireless networks are open for the further investigation.

First, we assume sensor nodes are aware of their own degrees and neighbors' degrees information. This can be achieved through local broadcasting. Then every node sets its probability of becoming a cluster head (CH) as follows:

$$CH_{prob} = E_{prob}\frac{E_{residual}}{E_{\max}} + D_{prob}\frac{D}{D_{\max}}, \quad (27)$$

where $E_{residual}$ and $E_{\max}$ denote its current residual energy and the maximum energy level, D and $D_{\max}$ are its own degree and the maximum degree of itself and its neighbors. E_{prob} and D_{prob} represents the selection weights for energy and degree, constrained by $E_{prob} + D_{prob} = 1$. The underlying rationale is that nodes with larger degrees are generally more correlated with others, therefore more suitable to be CHs. Taking residual energy into consideration tries to avoid that larger-degree nodes are quickly depleted of energy supply. Once the CHs are selected, they broadcast cluster formation messages to their neighbors, bearing the maximum hop (Max_Hop) and the correlation threshold (ϕ_{th}) information. Max_Hop constrains the maximum hops from a child node to the CH, while ϕ_{th} is used to determine how likely that a node belongs to a cluster. A node hearing this broadcast will join the cluster if its current hop is less than Max_Hop and the edge potential function is greater than ϕ_{th}. The clustered nodes will continue forward this message until Max_Hop reached. At last, nodes that don't hear any messages from neighbors or withhold from joining other clusters within T seconds are forced to be CHs.

This algorithm is fully distributed, and can flexibly control the cluster size (by Max_Hop) and cluster concentration (by ϕ_{th}).

C. *Distributed Inference in Structured Gaussian Markov Random Field*

We elaborate the structured variational method in a Gaussian Markov random field, which is a widely adopted model in theoretical studies and many practical situations.

Assume X is a Gaussian Markov random field and each node is only associated to a spatial component X_i of it. Each node makes a noisy linear observation

$$y_i = H_i x_i + \varepsilon_i,\ i = 1, \cdots, N, \quad (28)$$

where channel gain matrix H_i is assumed known, and noise ε_i is Gaussian with zero mean and variance $R_i = \sigma^2$.

The prior marginal (assumed the same for all nodes) and joint correlation functions are formulated as

$$p_i(x_i) = \mathcal{N}^{-1}(\mu_S/\sigma_S^2, 1/\sigma_S^2), \quad (29)$$

$$p_{ij}(x_i, x_j) = \mathcal{N}^{-1}(\Upsilon_{ij}\mu_S[1,1]^T, \Upsilon_{ij}),\ j \in \Gamma_i, \quad (30)$$

where

$$\Upsilon_{ij} = \frac{1}{\sigma_S^2(1-\Sigma_{ij}^2)}\begin{bmatrix} 1 & -\Sigma_{ij} \\ -\Sigma_{ij} & 1 \end{bmatrix}. \quad (31)$$

Belief propagation is used for intra-cluster inference, where the messages and node beliefs are all Gaussian distributed. Assume the message $m_{ij}^{(n)}$ and the belief $b_i^{(n)}$ are parameterized by

$$m_{ij}^{(n)}(x) \sim \mathcal{N}^{-1}(\mu_{ij}^{(n)}, V_{ij}^{(n)}) \text{ and } b_i^{(n)}(x) \sim \mathcal{N}^{-1}(q_i^{(n)}, W_i^{(n)}),$$

we have obtained the updating rules as [14]

$$\begin{cases} \mu_{ij}^{(n)} = \dfrac{\Sigma_{ij}\left(\mu_i + \sum_{k\in\Gamma_i\setminus\{j\}}\mu_{ki}^{(n-1)} + \mu_S/\sigma_S^2(1+\Sigma_{ij})\right)}{1+\sigma_S^2(1-\Sigma_{ij}^2)\left(V_i + \sum_{k\in\Gamma_i\setminus\{j\}}V_{ki}^{(n-1)}\right)} + \dfrac{\mu_S}{\sigma_S^2(1+\Sigma_{ij})} \\ V_{ij}^{(n)} = \dfrac{V_i + \sum_{k\in\Gamma_i\setminus\{j\}}V_{ki}^{(n-1)} + 1/\sigma_S^2}{1+\sigma_S^2(1-\Sigma_{ij}^2)\left(V_i + \sum_{k\in\Gamma_i\setminus\{j\}}V_{ki}^{(n-1)}\right)}, \end{cases} \quad (32)$$

with

$$\begin{aligned} \mu_i &= H_i^T R_i^{-1} y_i + (1-|\Gamma_i|)\mu_S/\sigma_S^2 \\ V_i &= H_i^T R_i^{-1} H_i + (1-|\Gamma_i|)/\sigma_S^2, \end{aligned} \quad (33)$$

and

$$\begin{cases} q_i^{(n)} = \mu_i + \sum_{k\in N(i)}\mu_{ki}^{(n)} \\ W_i^{(n)} = V_i + \sum_{k\in N(i)}V_{ki}^{(n)}. \end{cases} \quad (34)$$

In a clustered graph, the posterior probability could be reformulated as

$$P(X \mid y) = \frac{1}{Z}\prod_{C_i\in C}\phi_{C_i}(X_{C_i})\prod_{C_j\in\Gamma_{C_i}}\psi_{C_i,C_j}(X_{C_i}, X_{C_j}), \quad (35)$$

where $C = \{C_i\}$ denotes the entire cluster set. The Markov blanket (MB) of cluster C_i, $MB(C_i)$ is the set of all nodes outside of C_i but connected to some nodes in C_i, and the clusters that intersect with $MB(C_i)$ are called the Markov blanket clusters (MBC) of C_i, denoted by Γ_{C_i}. Fig. 1 gives a conceptual illustration for MB and MBC.

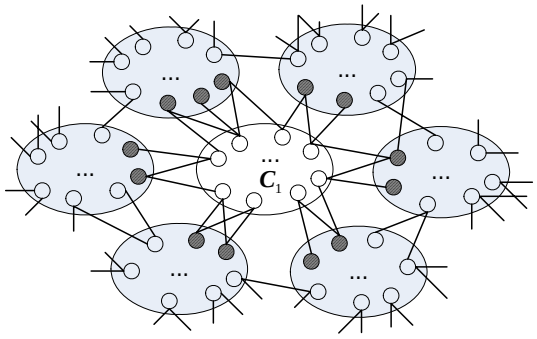

Fig. 1 The Markov blanket (shaded nodes) for Cluster C_1. The shaded clusters constitute Γ_{C_1} (reproduced Figure 1. in [13])

The extended natural parameters of the cluster potential and cluster compatibility functions are

$$\boldsymbol{\theta}_{C_i} = [y_{C_i}/2\sigma^2] \text{ and } \tilde{\boldsymbol{\theta}}_{C_iC_j} = \left\{[-\Sigma_{ij}X_j/2], i \in C_i\, j \in \Gamma_{C_i}\right\}. \quad (36)$$

Inter-cluster updating in this setting, by applying mean field equations, equivalently yields

$$y_{i,new} = y_{i,old} - \sum_{j \in MB(C_i)} \Sigma_{ij} m_j \ , \ i \in C_i \ . \qquad (37)$$

That is, the nodes use the estimates of the neighbors in their Markov blanket to get "new" observations, and use these "new" observations for the next round intra-cluster inference.

For the structured variational method, we have the following theorem:

Theorem 1: In a Gaussian MRF, the structured variational method using GBP as the intra-cluster inference algorithm converges.

Proof: See appendix.

IV. SIMULATION RESULTS

In this section, we present some numerical results to support the previous analysis, by comparing the performance of BP, MF and SMF on the same network. A snapshot of the network topology is shown in Fig.2, where 150 sensor nodes are randomly distributed in a unit plane and clusters are formed by using a simplified version of our proposed clustering algorithm. Nodes belonging to different clusters are indicated by distinct markers, while the cluster headers are plotted by shaded triangles.

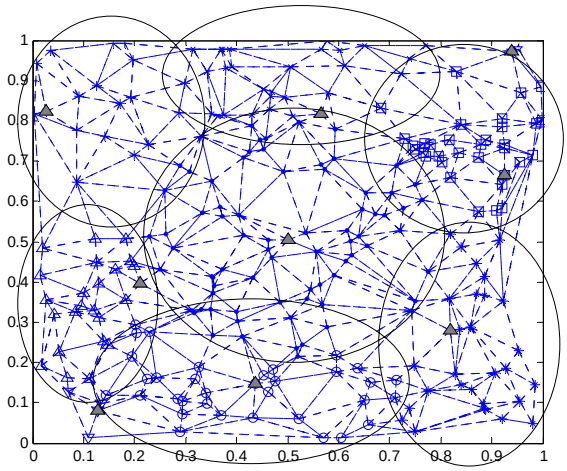

Fig. 2 Snapshot of a clustered network with 150 nodes

We adopt a similar simulation setting as in [16], where the true x is estimated from the noisy y in a Gaussian MRF, as shown in (28) . For simplicity, the channel gain H_i is assumed to be 1. All the other parameters take the values in [16]. The estimation mean square error (MSE) is used as the comparison metric, defined by

$$MSE = \left\| \hat{x}_m - \hat{x} \right\|_2 / \left\| \hat{x} \right\|_2 , \qquad (38)$$

with $\hat{x}_m$ denoting the estimation after mth iterations, and $\hat{x}$ being the exact estimation (the MMSE solution).

Fig. 3 compares the convergence rate of these three approaches, where the number of clusters is set to 65 and 15 respectively for SMF. The results verify that BP converges faster than the other two (alternatively is more accurate with iteration number fixed). But we also observe that even when the network is divided into very small clusters (65 clusters, corresponding to 2-3 nodes per cluster on average), SMF can significantly speed up the convergence. A fairly large cluster size (15 clusters, corresponding to 10 nodes per cluster on average) achieves almost indistinguishable performance to that of BP.

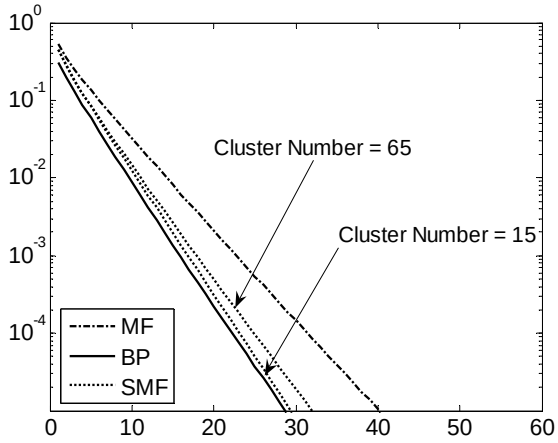

Fig. 3 Mean square error of estimation versus iteration number

Another important factor for sensor applications is the energy consumption. This is compared for the three algorithms in Fig. 4, where we assume for simplicity that the communication energy per message is proportional to the square of the distance between two communicating nodes, and the computation energy is neglected. It is evident that SMF consumes least energy to obtain the same estimation accuracy, which testifies its superiority.

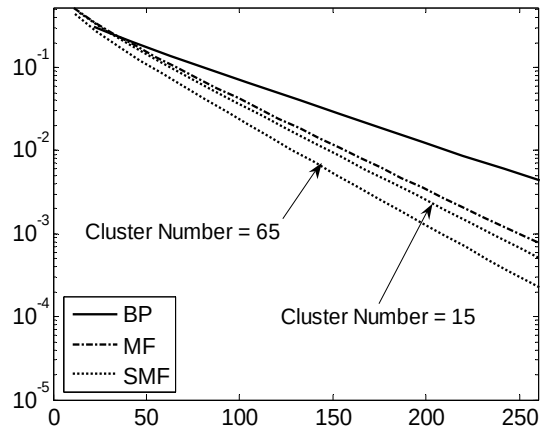

Fig. 4 Mean square error of estimation versus communication energy

Fig. 3 and Fig. 4 jointly indicate that larger cluster size helps improve estimation accuracy, while smaller cluster size leads to better energy efficiency. Therefore an appropriate cluster size should be selected depending on applications.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we develop a general variational message passing framework for Markov random fields. Particularly, its explicit forms for distributions in the exponential family are derived and applied to distributed time synchronization in wireless sensor

networks. Structured variational methods are also explored to further improve the performance. Our future work includes exploring its application in more scenarios (e.g. non-linear estimations) and more in-depth analysis.

REFERENCES

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[2] M. Cetin, L. Chen, J. Fisher, A. Ihler, R. Moses, M. J. Wainwright and A. S. Willsky. "Distributed fusion in sensor networks," *IEEE Signal Processing Magazine*, vol. 23, no. 4, 2006.

[3] T. Jaakkola, "Tutorial on variational approximation methods," In *Advanced mean field methods: theory and practice*. MIT Press, 2000.

[4] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Generalized belief propagation", *Advances in Neural Information Processing Systems* (NIPS), vol. 13, December 2000.

[5] S. Mukherjee and H. Kargupta., "Distributed Probabilistic Inferencing in Sensor Networks using Variational Approximation," *Journal of Parallel and Distributed Computing*, Volume 68, Issue 1, January 2008, Pages 78-92

[6] M. J. Wainwright, and M. I. Jordan, "Graphical models, exponential families, and variational inference," UC Berkeley, Dept. of Statistics, Technical Report 649. September, 2003.

[7] J. Winn, "Variational message passing and its applications," Ph.D. Thesis, Department of Physics, University of Cambridge, 2003.

[8] A. Giridhar and P. R. Kumar, "Distributed Clock Synchronization over Wireless Networks: Algorithms and Analysis," *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, 2006.

[9] O. Younis, M. Krunz, and S. Ramasubramanian, "Node Clustering in Wireless Sensor Networks: Recent Developments and Deployment Challenges," *IEEE Network*, Vol. 20, issue 3, pp. 20-25, May 2006

[10] L. Saul and M. Jordan, "Exploiting tractable substructures in intractable networks," In *Advances in Neural Information Processing Systems*, Vol. 8. The MIT Press, 1996.

[11] W. Wiegerinck, "Variational approximations between mean field theory and the junction tree algorithm," In *Proceeding of 16th Annual Conference on Uncertainty in Artificial Intelligence* (UAI), 2000.

[12] E.P. Xing, M.I Jordan and S. Russell, "Graph partition strategies for generalized mean field inference," In *Proc. of 20th Annual Conference on Uncertainty in Artificial Intelligence* (UAI), 2004.

[13] E.P. Xing, M.I. Jordan, and S. Russell, "A generalized mean field algorithm for variational inference in exponential families," In *Proceeding of 19th Annual Conference on Uncertainty in Artificial Intelligence* (UAI), 2003.

[14] H. Dai and Y. Zhang, "Consensus estimation via belief propagation," *2007 Conference on Information Sciences and Systems* (CISS), The Johns Hopkins University, Baltimore, MD, March 2007.

[15] C. C. Moallemi and B. Van Roy, "Consensus Propagation," *IEEE Transactions on Information Theory*, Vol. 52, No. 11, pp. 4753-4766, 2006.

[16] V. Delouille, R. Neelamani, R. Baraniuk, "Robust Distributed Estimation using the Embedded Subgraphs Algorithm," *IEEE Trans. Signal Processing*, Vol. 54, No.8, Aug. 2006.

APPENDIX- PROOF OF THEOREM 1

Combining inter- and intra-cluster updates, we can see the inverse message variance updating in belief propagation is independent with mean field approach. So we can follow the work in [15]. Define the function

$$F_{ij}(V)=\frac{V_i+\sum_{k\in\Gamma_i\setminus\{j\}}V_{ki}+1/\sigma_S^2}{1+\sigma_S^2(1-\Sigma_{ij}^2)\left(V_i+\sum_{k\in\Gamma_i\setminus\{j\}}V_{ki}\right)}. \tag{39}$$

First we have

Lemma 1: $F_{ij}(V)$ is continuous, monotonic and bounded for $V>0$.

Proof: Define function

$$f(x)=\frac{\alpha+x}{\beta+\gamma x}=\frac{1}{\gamma}+\left(\alpha-\frac{\beta}{\gamma}\right)\frac{1}{\beta+\gamma x}, \tag{40}$$

where $\alpha=V_i+1/\sigma_S^2$, $\beta=1+\sigma_S^2(1-\rho_{ij}^2)V_i$, $\gamma=\sigma_S^2(1-\Sigma_{ij}^2)$. Since $1/\gamma>0$ and $\alpha-\beta/\gamma=-\Sigma_{ij}^2/[\sigma_S^2(1-\Sigma_{ij}^2)]<0$, it's straightforward to verify that $f(x)$ is continuous, strictly increasing and bounded by $(0,1/\gamma)$. This leads to the convergence of the inverse variance iteration.

Now assuming the inverse variance has already converged, we can analyze the convergence of message mean.

Lemma 2: Define function

$$G_{ij}(\mu,V)=\frac{\Sigma_{ij}\left(\mu_i+\sum_{k\in\Gamma_i\setminus\{j\}}\mu_{ki}^{n-1}-\Sigma_{ij}\mu_S/\sigma_S^2(1-\Sigma_{ij}^2)\right)}{1+\sigma_S^2(1-\Sigma_{ij}^2)\left(V_i+\sum_{k\in\Gamma_i\setminus\{j\}}V_{ki}^{n-1}\right)}+\frac{\Sigma_{ij}\mu_S}{\sigma_S^2(1-\Sigma_{ij}^2)} \tag{41}$$

then $G_{ij}(\mu,V)$ is a contraction mapping with respect to μ, i.e. for any μ and μ', there exists $\eta\in(0,1)$ such that

$$\left\|G_{ij}(\mu,V)-G_{ij}(\mu',V)\right\|<\eta\left\|\mu-\mu'\right\|. \tag{42}$$

Proof: From the definition, we can choose

$$\eta=\max_{\substack{k\in\Gamma_i\setminus\{j\}\\(i,j)\in E}}\frac{\Sigma_{ij}}{1+\sigma_S^2(1-\Sigma_{ij}^2)\left(V_i+\sum_{k\in N(i)\setminus\{j\}}V_{ki}^{n-1}\right)}<1, \tag{43}$$

which satisfies the contraction mapping condition.

Now taking inter-cluster updating into account, the change in observations (37) will only reflect on μ_i, which will be cancelled out in $G_{ij}(\mu,V)-G_{ij}(\mu',V)$. So Lemma 2 still can be applied. That is, the overall iteration converges.

This concludes the proof of Theorem 1.