Distributed covariance estimation in Gaussian graphical models

Ami Wiesel and Alfred O. Hero III

***Abstract*—We consider distributed estimation of the inverse covariance matrix in Gaussian graphical models. These models factorize the multivariate distribution and allow for efficient distributed signal processing methods such as belief propagation (BP). The classical maximum likelihood approach to this covariance estimation problem, or potential function estimation in BP terminology, requires centralized computing and is computationally intensive. This motivates suboptimal distributed alternatives that tradeoff accuracy for communication cost. A natural solution is for each node to perform estimation of its local covariance with respect to its neighbors. The local maximum likelihood estimator is asymptotically consistent but suboptimal, i.e., it does not minimize mean squared estimation (MSE) error. We propose to improve the MSE performance by introducing additional symmetry constraints using averaging and pseudo-likelihood estimation approaches. We compute the proposed estimates using message passing protocols, which can be efficiently implemented in large scale graphical models with many nodes. We illustrate the advantages of our proposed methods using numerical experiments with synthetic data as well as real world data from a wireless sensor network.**

***Index Terms*—Covariance estimation, graphical models, distributed signal processing.**

I. Introduction

Covariance estimation is a classical and fundamental problem in statistical signal processing. Many applications, ranging from array processing to functional genomics, rely on accurately estimated covariance matrices [1], [2]. Recent interest in large scale inference of graphical models with small sample sizes has caused the topic to rise to prominence once again. A natural approach to this ”large p small n” problem is to incorporate additional prior knowledge on the structure, patterning, and/or sparsity of the covariance matrix or its inverse. Graphical models provide such prior information by representing conditional dependencies between variables by edges in an undirected graph. When the graph is sparse and the variables are jointly Gaussian, the graphical model imposes sparsity on the inverse covariance, variously called the information, concentration or precision matrix. The resulting graphical model is represented by a network called the concentration graph. Computationally efficient implementation of statistical inference algorithms, such as as belief propagation (BP), can be implemented on this network.

Graphical models are attractive since inference can be performed as local decentralized computations with message passing [3]–[5]. Decentralization of computation of an accurate inverse covariance estimator can facilitate high dimensional data analysis applications such as anomaly detection in wide area sensor networks, spatial correlation in images, and analysis of high throughput gene expression arrays. When the topology (local dependency) of the graphical model matches the topology (local data passage) of internode communication, near optimal estimation performance can be achieved at significantly reduced computational cost as compared to the global centralized approach. In applications such as those mentioned above, often there is a good match between local dependency and local data passage, e.g., in geographically distributed networks of sensing devices. The premise of this paper is that the model topology and communication topology are matched. Such an assumption is common in other decentralized formulations of networked estimation, e.g, BP via message passing in imaging and networks. BP has been successfully applied when the underlying graph is a tree and more recently it has been applied to arbitrary topologies [6]–[10]. A crucial step underlying Bayesian inference is to learn the parameters of the distribution, also known as the potential functions. In the Gaussian case, this step corresponds to covariance estimation, or more precisely, inverse covariance estimation.

The time-tested approach to covariance estimation in Gaussian graphical models (GGM) is maximum likelihood (ML) [11]–[13]. This approach is consistent and asymptotically optimal in terms of minimizing mean squared error (MSE). When the underlying graph is a tree, the ML estimate has a simple closed form solution which requires little communication between nodes. This closed form solution can also be generalized to chordal graphs using junction trees. Low complexity numerical solutions for near chordal graphs have been recently proposed in [14]. In arbitrary topologies, finding the ML estimate requires solving a difficult high dimensional convex optimization problem. For small graphs, i.e., graphs with few vertices, general purpose optimization toolboxes [15] or iterative proportional fitting (IPF) [11] can be applied to solve this optimization problem. While a distributed version of IPF can be derived by implementing each of its iterations via BP methods [16], [17], such approaches are too computationally intensive to be practical for large concentration graphs. In some scenarios, approximate estimation may be accomplished using low rank approximations [18]. In general, distributed covariance estimation via message passing remains a difficult task and suboptimal approaches are necessary.

In this work, we propose alternative distributed estimation methods that approximate the global ML solution by trading

A preliminary version of this work has been presented in the 6th IEEE Sensor Array and Multichannel Signal Processing Workshop, SAM-2010. A. Wiesel is with the School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel. A. O. Hero III is with the Department of Electrical Engineering and Computer Science, University of Michigan. Emails: amiw@cs.huji.ac.il and hero@umich.edu. The work of A. Wiesel was supported by a Marie Curie Outgoing International Fellowship within the 7th European Community Framework Programme. This research was partially supported by AFOSR grant FA9550-06-1-0324.

estimator accuracy for lower communication costs. All of the proposed methods are based on aggregating local estimates of parts of the inverse covariance matrix. The most natural approach to distributed estimation would have each node generate a local estimate of its covariance with its neighbors in the graph. This simple approach yields a closed form solution which requires no message passing. However, while such local approaches yield asymptotically consistent estimates, they suffer from higher asymptotic MSE than the global ML estimator. In fact, for finite samples the overall estimate is not symmetric. The two distributed aggregation methods proposed in this paper enforce symmetry via message passing and therefore reduce MSE. The first method uses simple symmetric averaging of the local estimates, whereas the second is based on a pseudo-likelihood technique [19]–[24]. We implement the latter via an Alternating Direction Method of Multipliers that has shown promising performance in similar distributed signal processing problems [25]–[31]. We demonstrate the advantages of the proposed methods using synthetic simulations as well as real world data from a wireless sensor network [43], [44].

To avoid confusion, we emphasize that we consider the problem of covariance estimation with a known graphical model, i.e., the conditional independence graph topology is known a priori. A related problem which recently attracted considerable attention is covariance selection in which the graphical model is unknown and is detected based on the measurements. The latter is clearly a much more difficult task aimed at different applications, e.g. topology discovery. More details on covariance selection can be found in [12], [24], [32]–[35] and references within.

The outline of the paper is as follows. In Section II we briefly review the basics of GGM and formulate the distributed covariance estimation problem. In Section III, we review the global ML estimator and define the three distributed local ML estimators. In Section IV we discuss the exact and asymptotic performance analysis of the estimators. In Section V we demonstrate the advantages of our proposed estimators using synthetic simulations and experiments with real world data. Concluding remarks are given in Section VI.

The following notation is used. Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars. We use indices in the subscript $[\mathbf{x}]_a$ or $[\mathbf{x}]_{a,b}$ to denote sub-vectors or sub-matrices, respectively, and $[\mathbf{X}]_{a,:}$ denotes the sub-matrix formed by the a'th rows in $\mathbf{X}$. The superscripts $(\cdot)^T$ and $(\cdot)^{-1}$ denote the transpose and matrix inverse, respectively. For sets a and b, the set difference operator is denoted by $a \setminus b$, and cardinality is denoted as $|a|$. The operator $\|\mathbf{X}\|$ denotes the Frobenius norm of a matrix $\mathbf{X}$, namely $\|\mathbf{X}\|^2 = \mathrm{Tr}(\mathbf{X}^T\mathbf{X})$, $\mathbf{X} \succeq \mathbf{0}$ means that $\mathbf{X}$ is positive semidefinite and $\mathbf{X} \succ \mathbf{0}$ means that $\mathbf{X}$ is positive definite.

II. Problem formulation

In this section, we briefly review the formulation of the GGM. For more details the reader is referred to [11]. We then formulate the distributed covariance estimation problem addressed in this paper.

Graphical models are intuitive characterizations of conditional independence structures exhibited by variables with a joint distribution $p(\mathbf{x}_1, \cdots, \mathbf{x}_p)$. Specifically, define an undirected graph $\mathcal{G} = (V, E)$ with a set of nodes $V = \{1, \cdots, p\}$ connected by undirected edges $E = \left\{(i_1, j_1), \cdots \left(i_{|E|}, j_{|E|}\right)\right\}$, where we use the convention that each node is connected to itself, i.e., $(i,i) \in E$ for all $i \in V$. We define N_i as the set of neighbors of the i'th node, $i = 1, \cdots, p$, i.e., $N_i = \{j | (i,j) \in E, j \neq i\}$.

Let $\mathbf{x}$ be a length p zero mean random vector, called the vector of node states, whose elements are indexed by the nodes in V. The vector $\mathbf{x}$ satisfies the Markov property with respect to $\mathcal{G}$, if for any pair of non-adjacent nodes the corresponding pair of elements in $\mathbf{x}$ are conditionally independent given the remaining elements:

$$\begin{aligned} p\left(\mathbf{x}_i, \mathbf{x}_j | \mathbf{x}_{V \setminus i,j}\right) &= p\left(\mathbf{x}_i | \mathbf{x}_{V \setminus i,j}\right) p\left(\mathbf{x}_j | \mathbf{x}_{V \setminus i,j}\right) \\ &\qquad \text{for all} \qquad \{i,j\} \notin E. \end{aligned} \quad (1)$$

This conditional factorization property induces simplifications in optimal prediction over the nodes of the graph. For example, the optimal minimum mean squared error (MMSE) predictor of $\mathbf{x}_i$ given $\mathbf{x}_{V \setminus i}$ reduces to the form

$$\begin{aligned} \hat{\mathbf{x}}_i\left(\mathbf{x}_{V \setminus i}\right) &= \mathrm{E}\left\{\mathbf{x}_i | \mathbf{x}_{V \setminus i}\right\} \\ &= \mathrm{E}\left\{\mathbf{x}_i | \mathbf{x}_{N_i}\right\}, \end{aligned} \quad (2)$$

which is a low dimensional function of the vector of node states when the number of neighbors of node i is small.

The class of GGMs are graphical models over the multivariate Gaussian distribution. This distribution is appealing due to the fact that it is completely characterized by its mean $\boldsymbol{\eta}$ and covariance $\boldsymbol{\Sigma} \succ \mathbf{0}$:

$$\frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \mathbf{e}^{-\frac{(\mathbf{x}-\boldsymbol{\eta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\eta})}{2}}, \quad (3)$$

and the fact that the optimal MMSE predictors are linear functions of the vector of node states. In our context, it will be more natural to use the canonical parameters $\mathbf{J} = \boldsymbol{\Sigma}^{-1} \succ \mathbf{0}$ and $\mathbf{h} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ which lead to the following representation

$$\mathcal{N}(\mathbf{x}; \mathbf{h}, \mathbf{J}) = \frac{e^{-\frac{1}{2}\mathbf{h}^T\mathbf{J}^{-1}\mathbf{h}}}{(2\pi)^{p/2} |\mathbf{J}|^{-\frac{1}{2}}} \mathbf{e}^{-\frac{1}{2}\mathbf{x}^T\mathbf{J}\mathbf{x} + \mathbf{x}^T\mathbf{h}}, \quad (4)$$

in which the exponent exhibits a linear dependency on the parameters $\mathbf{h}$ and $\mathbf{J}$. This representation is natural since the graph $\mathcal{G}$ is directly related to the sparsity of $\mathbf{J}$. Indeed, simple algebraic manipulations reveal that specializing the factorization property (1) to this Gaussian case yields

$$[\mathbf{J}]_{i,j} = 0 \qquad \text{for all} \qquad \{i,j\} \notin E. \quad (5)$$

This sparsity of $\mathbf{J}$ simplifies statistical inference. The MMSE predictor in (2) has a simple linear form which depends only on the neighboring nodes

$$\begin{aligned} \hat{\mathbf{x}}_i\left(\mathbf{x}_{V \setminus i}\right) &= -\frac{\mathbf{J}_{i,V\setminus i}\mathbf{x}_{V\setminus i}}{\mathbf{J}_{i,i}} \\ &= -\frac{\mathbf{J}_{i,N_i}\mathbf{x}_{N_i}}{\mathbf{J}_{i,i}}. \end{aligned} \quad (6)$$

This relation forms the main building block for Gaussian BP methods.

The global centralized GGM covariance estimation problem can be formulated as follows. Let $\mathbf{x}$ be a zero mean Gaussian random vector, with inverse covariance matrix $\mathbf{J}$. Given T independent and identically distributed (i.i.d.) realizations of node states $\mathbf{x}$, denoted by $\{\mathbf{x}[t]\}_{t=1}^{T}$, and knowledge of the conditional independence structure through the topology of $\mathcal{G} = (V, E)$, the goal is to estimate the inverse covariance $\mathbf{J}$.

In this paper, we consider a distributed version of the covariance estimation problem. Specifically, we associate with each of the random variables $\mathbf{x}_i$ a node in a network. We assume that the topology of communications links between nodes and their neighbors matches the topology of $\mathcal{G}$. Each node i has access only to $\{\mathbf{x}_{[i\ N_i]}[t]\}_{t=1}^{T}$. Using this data, along with message passing with its neighbors, each node tries to estimate its local information parameters defined as the submatrix $\mathbf{J}_{i,[i,N_i]}$ of the inverse covariance matrix $\mathbf{J}$.

Our definition of local information parameters stems from the fact that, in view of (6), $\mathbf{J}_{i,[i,N_i]}$ is all that is required for minimum MSE prediction of the node state $\mathbf{x}_i$ given all other node states. Knowledge of the easier-to-estimate local covariance matrix $\mathbf{\Sigma}_{[i,N_i],[i,N_i]}$ involves more parameters, yet not enough to specify the predictor coefficients in (6). In BP terminology, $\mathbf{J}_{i,[i,N_i]}$ is more directly related to the Gaussian potential functions.

Thus, in our framework each node tries to estimate its own local information parameters $\mathbf{J}_{i,[i,N_i]}$ with the help of its neighbors. Notationally, we collect the local estimates in one global matrix, namely $\hat{\mathbf{J}}$, whose $\{i, [i\ N_i]\}$ elements are the local estimates in the i'th node. Note that $\hat{\mathbf{J}}_{i,j}$ and $\hat{\mathbf{J}}_{j,i}$ are both estimators of $\mathbf{J}_{i,j} = \mathbf{J}_{j,i}$ but may be different since each is estimated by a different node. Reduction of the effects of this asymmetry in the local estimates is a principal contribution of this paper.

III. Estimators

A. Global maximum likelihood

The classical approach to covariance estimation is based on the ML principle. The estimate is chosen as the parameter that maximizes the log-likelihood function. The ML estimator of the inverse covariance matrix is [11]

$$\begin{aligned}\hat{\mathbf{J}}^{\mathrm{ML}} &= \arg\max_{\mathbf{J}\in\mathcal{J}} \sum_{t=1}^{T} \log p\left(\mathbf{x}\left[t\right];\mathbf{J}\right) \\ &= \arg\min_{\mathbf{J}\in\mathcal{J}} \sum_{t=1}^{T} \frac{1}{2}\mathbf{x}^{T}[t]\mathbf{J}\mathbf{x}[t] - \frac{T}{2}\log\left|\mathbf{J}\right| \\ &= \arg\min_{\mathbf{J}\in\mathcal{J}} \mathrm{Tr}\left\{\mathbf{S}\mathbf{J}\right\} - \log\left|\mathbf{J}\right|, \qquad (7)\end{aligned}$$

where the feasible set is defined as

$$\mathcal{J} = \left\{\mathbf{J} : \mathbf{J} \succ \mathbf{0},\ \mathbf{J}_{i,j} = 0,\ \forall\ (i,j) \notin E\right\}, \qquad (8)$$

and the sample covariance $\mathbf{S}$ is given by

$$\mathbf{S} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}[t]\mathbf{x}^{T}[t]. \qquad (9)$$

The global ML optimization problem in (7) is a convex optimization problem. It can be solved in a centralized manner but does not lend itself to a natural distributed implementation.

B. Local maximum likelihood

A natural alternative to the global ML strategy is the local ML method. This estimator, denoted by LOC, aggregates p decoupled ML estimators implemented independently at each of the nodes. Each node belongs to its local network of $1 + |N_i|$ nodes. The marginal local distribution is a Gaussian distribution

$$p\left(\mathbf{x}_{[i\ N_i]};\mathbf{J}^i\right) = \mathcal{N}\left(\mathbf{x}_{[i\ N_i]};\mathbf{0},\mathbf{J}^i\right), \qquad (10)$$

where the local information matrices are given by

$$\begin{aligned}\mathbf{J}^i &= \left[\left[\mathbf{J}^{-1}\right]_{[i\ N_i],[i\ N_i]}\right]^{-1} \\ &= \mathbf{J}_{[i\ N_i],[i\ N_i]} - \left[\begin{array}{cc} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{N_i,R_i}\mathbf{J}^{-1}_{R_i,R_i}\mathbf{J}_{R_i,N_i}\end{array}\right], \qquad (11)\end{aligned}$$

where $R_i = V \setminus [i, N_i]$. In general, $\mathbf{J}^i$ is not equivalent to $\mathbf{J}_{[i\ N_i],[i\ N_i]}$, but it is easy to see that its first row is exactly the elements we are interested in, namely $\mathbf{J}_{i,[i\ N_i]}$. Thus, a natural approach to distributed covariance estimation is to let each node independently estimate its own $\mathbf{J}^i$. Assuming that the nodes generate local ML estimates independently we obtain:

$$\begin{aligned}\hat{\mathbf{J}}^i &= \arg\max_{\mathbf{J}^i\succeq\mathbf{0}} \sum_{t=1}^{T} \log p\left(\mathbf{x}_{[i\ N_i]}\left[t\right];\mathbf{J}^i\right) \\ &= \arg\min_{\mathbf{J}^i\succeq\mathbf{0}} \sum_{t=1}^{T} \frac{1}{2}\mathbf{x}^{T}_{[i\ N_i]}\left[t\right]\mathbf{J}^i\mathbf{x}_{[i\ N_i]}\left[t\right] - \frac{T}{2}\log\left|\mathbf{J}^i\right| \\ &= \arg\min_{\mathbf{J}^i\succeq\mathbf{0}} \mathrm{Tr}\left\{\mathbf{S}_{[i\ N_i],[i\ N_i]}\mathbf{J}^i\right\} - \log\left|\mathbf{J}^i\right| \\ &= \left(\mathbf{S}_{[i\ N_i],[i\ N_i]}\right)^{-1}, \qquad (12)\end{aligned}$$

where $\mathbf{S}$ is defined in (9) and we assume $T \geq |N_i|$ for all i so that the inverse exists with probability one. Given our definition of local covariance information, the i'th node utilizes only the first row of $\hat{\mathbf{J}}^i$ and defines

$$\hat{\mathbf{J}}^{\mathrm{LOC}}_{i,[i\ N_i]} = \hat{\mathbf{J}}^i_{1,:} = \left[\left(\mathbf{S}_{[i\ N_i],[i\ N_i]}\right)^{-1}\right]_{1,:}. \qquad (13)$$

Together, $\hat{\mathbf{J}}^{\mathrm{LOC}}$ is the $p \times p$ matrix having the first rows of $\hat{\mathbf{J}}^i$ for $i = 1, \cdots, p$ in its $(i, [i\ N_i])$ positions, and having zero padding elsewhere. While $\hat{\mathbf{J}}^{\mathrm{LOC}}$ is a more simply computed matrix than $\hat{\mathbf{J}}^{ML}$, it is not necessarily symmetric or positive definite.

In the next sections, we consider distributed estimators which provide an appealing tradeoff between LOC and ML. The goal is to improve estimator accuracy through message passing protocols between adjacent nodes, entailing only a small increase in computation. The basic principle behind this improvement is to exploit the known symmetry in the inverse covariance.

C. Average estimator

The simplest method to enforce symmetry in the inverse covariance estimate is to perform collaborative local averaging. Each pair of neighboring nodes (i,j) exchange their values of $\hat{\mathbf{J}}^{\mathrm{LOC}}_{i,j}$ and $\hat{\mathbf{J}}^{\mathrm{LOC}}_{j,i}$ and modify their estimate to the average value:

$$\hat{\mathbf{J}}^{\mathrm{AVE}}_{i,j} = \begin{cases} \hat{\mathbf{J}}^{\mathrm{LOC}}_{i,j} & i = j \\ \frac{1}{2}\hat{\mathbf{J}}^{\mathrm{LOC}}_{i,j} + \frac{1}{2}\hat{\mathbf{J}}^{\mathrm{LOC}}_{j,i} & i \neq j, (i,j) \in E \\ 0 & (i,j) \notin E. \end{cases} \tag{14}$$

This simple estimator, denoted by AVE, can be easily implemented in a distributed manner by passing two messages per edge.

D. Pseudo-likelihood estimator

While symmetrization by collaborative local averaging is simple, better performance can be achieved by using pseudo-likelihood optimization with symmetry constraints. To motivate this approach, we first re-derive the LOC estimator of Sec. III-B as a solution to a convex optimization problem and then propose to enhance its accuracy by imposing additional symmetry constraints.

The pseudo-likelihood approximation to the ML estimator approximates the intractable likelihood using a surrogate function, called a pseudo-likelihood or composite-likelihood function [19]–[24]. Specifically, we define the conditional estimator of the global information matrix as the solution to

$$\hat{\mathbf{J}}^{\mathrm{COND}} = \arg\max_{\mathbf{J}} \sum_{t=1}^{T} \log \tilde{p}\left(\mathbf{x}[t];\mathbf{J}\right), \tag{15}$$

where we have defined the log-pseudo-likelihood as a weighted sum of log-conditional likelihoods

$$\log \tilde{p}\left(\mathbf{x}[t];\mathbf{J}\right) = \sum_{i=1}^{p} w_i \log p(\mathbf{x}_i[t]|\mathbf{x}_{V\setminus i}[t];\mathbf{J}), \tag{16}$$

where w_i for $i = 1, \cdots, p$ are weight coefficients. By choosing $w_i = 1$ for all i, (20) becomes a standard (unweighted) pseudo-likelihood estimator. The generalization allows the estimator to taper the influence of different regions in the graph, for example the weights can be functions of the number of neighbors associated with each node. In Section V we provide a promising data-dependent choice for these weights.

It is easy to see why this approximate objective function has attractive computational advantages as compared to the global likelihood. Due to the structure of the GGM, we obtain

$$\begin{aligned} p(\mathbf{x}_i[t]|\mathbf{x}_{V\setminus i}[t];\mathbf{J}) &= p(\mathbf{x}_i[t]|\mathbf{x}_{N_i}[t];\mathbf{J}_{i,[i\ N_i]}) \\ &= \mathcal{N}\left(\mathbf{x}_i[t]; -\mathbf{J}_{i,N_i}\mathbf{x}_{N_i}, \mathbf{J}_{i,i}\right), \end{aligned} \tag{17}$$

where we have invoked conditional independence to obtain the first equality and the properties of the conditional Gaussian distribution to obtain the second equality. Note that the i'th conditional distribution is fully characterized by $\mathbf{J}_{i,[i,N_i]}$, which is the parameter of interest. Thus, the problem in (15) decouples into p independent problems

$$\max_{\mathbf{J}_{i,[i\ N_i]}} w_i \sum_{t=1}^{T} \log p\left(\mathbf{x}_i[t]|\mathbf{x}_{N_i}[t];\mathbf{J}_{i,[i\ N_i]}\right), \tag{18}$$

for $i = 1, \cdots, p$. Each of these problems can be solved in closed form. In fact, we obtain the following result.

Proposition 1: The local estimator $\hat{\mathbf{J}}^{\mathrm{LOC}}$ in (13) and the conditional estimator $\hat{\mathbf{J}}^{\mathrm{COND}}$ in (15) are identical.

Proof: We need to show that (13) is the solution to (18) for each i. The objective is a convex function of $\mathbf{J}_{i,[i\ N_i]}$ and can be simplified to (19) below. Solving for $\mathbf{J}_{i,N_i}$ yields

$$\mathbf{J}_{i,N_i} = -\mathbf{J}_{i,i}\mathbf{S}_{i,N_i}\mathbf{S}^{-1}_{N_i,N_i}. \tag{20}$$

Plugging this into the objective and solving for $\mathbf{J}_{i,i}$ results in

$$\mathbf{J}_{i,i} = \frac{1}{\mathbf{S}_{i,i} - \mathbf{S}_{i,N_i}\mathbf{S}^{-1}_{N_i,N_i}\mathbf{S}_{N_i,i}}. \tag{21}$$

Using the matrix inversion formula for partitioned matrices, it is easy to verify that (20)-(21) are exactly the first row of the inverse of $\mathbf{S}_{[i,N_i],[i\ N_i]}$ as expressed in (13). ■

Based on the optimization formulation of the local estimator, we can enforce the known symmetry constraint on the inverse covariance and define the pseudo-likelihood maximization (PML) estimator:

$$\hat{\mathbf{J}}^{\mathrm{PML}} = \arg\max_{\mathbf{J}_{i,j}=\mathbf{J}_{j,i}} \sum_{t=1}^{T} \log \tilde{p}\left(\mathbf{x}[t];\mathbf{J}\right), \tag{22}$$

which is identical to (15) except for the additional symmetry conditions $\mathbf{J}_{i,j} = \mathbf{J}_{j,i}$. The objective function in (22) is separable and the only thing that would prevent a straightforward distributed implementation is the added complexity due to the constraints. We propose to decouple the symmetry constraints in (22) using the Alternating Direction Method of Multipliers [25]–[31]. We define the augmented Lagrangian[1]

$$\begin{aligned} \mathcal{L}\left(\hat{\mathbf{J}},\overline{\mathbf{J}};\mathbf{M}\right) = \sum_{i=1}^{p}\Bigg[& \frac{w_i}{T}\sum_{t=1}^{T} \log p\left(\mathbf{x}_i\left[t\right]|\mathbf{x}_{N_i}\left[t\right];\hat{\mathbf{J}}_{i,[i\ N_i]}\right) \\ & + \sum_{j\in N_i}\left(\mathbf{M}_{i,j}\left(\hat{\mathbf{J}}_{i,j}-\overline{\mathbf{J}}_{i,j}\right) - \frac{c}{2}\left(\hat{\mathbf{J}}_{i,j}-\overline{\mathbf{J}}_{i,j}\right)^2\right)\Bigg], \end{aligned} \tag{23}$$

where $\hat{\mathbf{J}}$ is the PML estimate (whose superscript is omitted for clarity), $\overline{\mathbf{J}}$ is a symmetric auxiliary matrix, $\mathbf{M}$ is a matrix of dual multipliers and c is a positive scalar parameter. In standard dual decomposition methods, a saddle point of the augmented Lagrangian $\mathcal{L}\left(\hat{\mathbf{J}},\overline{\mathbf{J}};\mathbf{M}\right)$ is found by iteratively solving for the primal variables $\hat{\mathbf{J}}$ and $\overline{\mathbf{J}}$ with fixed dual variables $\mathbf{M}$ and then updating $\mathbf{M}$. Due to the coupling in the quadratic term, solving the primal problem is difficult. Remarkably, the method is guaranteed to converge to the global solution even if we update $\mathbf{M}$ with suboptimal primal values (see [25] and references within). This leads to the following iterations

$$\overline{\mathbf{J}}_{i,j \text{ and } j,i} \leftarrow \frac{\hat{\mathbf{J}}_{i,j}+\hat{\mathbf{J}}_{j,i}}{2} - \frac{\mathbf{M}_{i,j}+\mathbf{M}_{j,i}}{2c}; \tag{24}$$

$$\hat{\mathbf{J}}_{i,[i\ N_i]} \leftarrow \arg\max_{\hat{\mathbf{J}}_{i,[i\ N_i]}} g_i\left(\hat{\mathbf{J}}_{i,[i\ N_i]}\right); \tag{25}$$

$$\mathbf{M}_{i,j} \leftarrow \mathbf{M}_{i,j} + c\left(\hat{\mathbf{J}}_{i,j}-\overline{\mathbf{J}}_{i,j}\right), \tag{26}$$

[1] For simplicity, we scale the original objective by a factor of $1/T$.

$$\begin{aligned} w_i \sum_{t=1}^{T} \log p\left(\mathbf{x}_i[t] | \mathbf{x}_{N_i}[t]; \mathbf{J}_{i,[i\ N_i]}\right) &= w_i \sum_{t=1}^{T} \left[-\frac{\mathbf{J}_{i,i}\mathbf{x}_i^2[t]}{2} - \mathbf{x}_i[t]\mathbf{J}_{i,N_i}\mathbf{x}_{N_i}[t] - \frac{\mathbf{J}_{i,N_i}\mathbf{x}_{N_i}[t]\mathbf{x}_{N_i}^T[t]\mathbf{J}_{i,N_i}^T}{2\mathbf{J}_{i,i}} + \frac{T}{2}\log|\mathbf{J}_{i,i}| \right] \\ &= w_i T \left[-\frac{\mathbf{J}_{i,i}\mathbf{S}_{i,i}}{2} - \mathbf{J}_{i,N_i}\mathbf{S}_{N_i,i} - \frac{\mathbf{J}_{i,N_i}\mathbf{S}_{N_i,N_i}\mathbf{J}_{i,N_i}^T}{2\mathbf{J}_{i,i}} + \frac{1}{2}\log|\mathbf{J}_{i,i}| \right]. \end{aligned} \tag{19}$$

$$g_i\left(\hat{\mathbf{J}}_{i,[i\ N_i]}\right) = w_i \left[-\frac{1}{2}\hat{\mathbf{J}}_{i,i}\mathbf{S}_{i,i} - \hat{\mathbf{J}}_{i,N_i}\mathbf{S}_{i,N_i}^T - \frac{\hat{\mathbf{J}}_{i,N_i}\mathbf{S}_{N_i,N_i}\hat{\mathbf{J}}_{i,N_i}^T}{2\hat{\mathbf{J}}_{i,i}} + \frac{1}{2}\log\left|\hat{\mathbf{J}}_{i,i}\right| \right] + \sum_{j\in N_i} \mathbf{M}_{i,j}\hat{\mathbf{J}}_{i,j} - \frac{c}{2}\hat{\mathbf{J}}_{i,j}^2 - c\hat{\mathbf{J}}_{i,j}\overline{\mathbf{J}}_{i,j}. \tag{27}$$

for all $i \in V$ and $j \in N_i$, where $g_i\left(\hat{\mathbf{J}}_{i,[i\ N_i]}\right)$ is defined in (27) below. This algorithm involves local computations and message passing through the dual multipliers in $\mathbf{M}$.

The most computationally intensive operations in the distributed algorithm (24)-(27) are the p optimizations in (25). We now show that each of these p sub-problems can be reduced to a simple line search. For this purpose we first solve for $\hat{\mathbf{J}}_{i,N_i}$. Taking the derivative of (27) with respect to $\hat{\mathbf{J}}_{i,N_i}$ and equating to zero yields

$$\hat{\mathbf{J}}_{i,N_i} = -\left(\frac{w_i\mathbf{S}_{N_i,N_i}}{\hat{\mathbf{J}}_{i,i}} + c\mathbf{I}\right)^{-1}\mathbf{s}_i, \tag{28}$$

where

$$\mathbf{s}_i = w_i\mathbf{S}_{N_i,i} - \mathbf{M}_{N_i,i} - c\overline{\mathbf{J}}_{N_i,i}. \tag{29}$$

Plugging $\hat{\mathbf{J}}_{i,N_i}$ back into the objective (27) yields a line search with respect to $\hat{\mathbf{J}}_{i,i}$:

$$\max_{\hat{\mathbf{J}}_{i,i}} -w_i\hat{\mathbf{J}}_{i,i}\mathbf{S}_{i,i} + \mathbf{s}_i^T\left(\frac{w_i\mathbf{S}_{N_i,N_i}}{\hat{\mathbf{J}}_{i,i}} + c\mathbf{I}\right)^{-1}\mathbf{s}_i + w_i\log\left|\hat{\mathbf{J}}_{i,i}\right|. \tag{30}$$

This line search is unimodal since the original problem was jointly convex. Therefore, it can be efficiently solved using a bisection method. In the special case in which $\mathbf{M} = \mathbf{0}$ and $c = 0$, the line search has a simple closed form solution which coincides with $\hat{\mathbf{J}}_{i,i}^{\mathrm{LOC}}$ as given in (20)-(21).

E. Generalizations

PML exploits the known symmetry in $\mathbf{J}$ but may produce an estimator which is not positive definite. Performance can be improved by adding the positive definite constraint to (22). However, we are not aware of simple distributed methods for enforcing it. Moreover, as detailed in Section IV, this modification will not affect the asymptotic performance of PML.

Another interesting direction would be to try to improve the summation aggregation of likelihoods. The pseudo-likelihood is a weighted sum of conditional distributions of $\mathbf{x}_i$. It can be improved by summing over conditional distributions of $\mathbf{x}_{S_i}$ where S_i are overlapping subsets of indices. For example, S_i can be chosen as the pairs of adjacent nodes. By appropriately choosing the cardinalities of S_i we can achieve a flexible accuracy vs. complexity tradeoff. Similar extensions are discussed in [21].

IV. Performance analysis

In this section, we analyze the performance of the different estimators discussed in Sec. III.

For notational convenience, we will parameterize the various sparse matrices using vectors. We will use two different parameterizations: $\boldsymbol{\theta}_s$ for the non-zero elements (associated with edges) of a symmetric matrix $\mathbf{J}$, and $\boldsymbol{\theta}_{ns}$ for the non-zero elements of a non-symmetric matrix $\mathbf{J}$. The difference is that the symmetric version models both $\mathbf{J}_{i,j}$ and $\mathbf{J}_{j,i}$ for $i \neq j$ using a single element in $\boldsymbol{\theta}_s$, whereas the non-symmetric version models them using two different elements in $\boldsymbol{\theta}_{ns}$. These notations with appropriate superscripts hold for both the true inverse covariance and its estimates.

A. Global maximum likelihood

The global estimator is an ML estimator for a smooth regular family and is therefore asymptotically consistent and efficient. Using the symmetric vector parameterization, we obtain

$$\hat{\boldsymbol{\theta}}_s^{\mathrm{ML}} - \boldsymbol{\theta}_s \rightarrow \mathbf{0}. \tag{31}$$

The consistency rate is weakly dependent on p as long as the number of maximal neighbors is small. When the topology of the graph is unknown, the rate in the covariance selection problem depends only logarithmically on p [33]–[35], [37]. Clearly, the rate will be better or equal to logarithmic in the easier covariance estimation problem treated here where the graph topology is known.

The asymptotic error covariance follows the well known ML analysis:

$$T\mathrm{E}\left\{\left(\hat{\boldsymbol{\theta}}_s^{\mathrm{ML}} - \boldsymbol{\theta}_s\right)\left(\hat{\boldsymbol{\theta}}_s^{\mathrm{ML}} - \boldsymbol{\theta}_s\right)^T\right\} \rightarrow \mathbf{F}^{-1}, \tag{32}$$

where $\mathbf{F}$ is the Fisher Information Matrix (FIM). The element of the FIM associated with the information between $\mathbf{J}_{i_1,j_1}$ and $\mathbf{J}_{i_2,j_2}$ can be easily derived as [36]

$$\begin{aligned} &\mathbf{F}\left(\mathbf{J}_{i_1,j_1}, \mathbf{J}_{i_2,j_2}\right) \\ &= \begin{cases} \frac{1}{2}\boldsymbol{\Sigma}_{i_1,i_2}^2 & i_1 = j_1,\ i_2 = j_2 \\ \boldsymbol{\Sigma}_{i_1,i_2}\boldsymbol{\Sigma}_{j_2,i_1} & \begin{array}{c} i_1 = j_1, i_2 \neq j_2 \\ \text{or } i_1 \neq j_1, i_2 = j_2 \end{array} \\ \boldsymbol{\Sigma}_{j_1,i_2}\boldsymbol{\Sigma}_{j_2,i_1} + \boldsymbol{\Sigma}_{i_1,i_2}\boldsymbol{\Sigma}_{j_2,j_1} & i_1 = j_1,\ i_2 \neq j_2. \end{cases} \end{aligned} \tag{33}$$

B. Local maximum likelihood

The local estimator (13) is a simple aggregation of local ML estimators. Each of these follows a scaled inverse Wishart distribution whose moments have expressions given in [38]. Therefore, its exact non-asymptotic total MSE in terms of Frobenius norm can be obtained in closed form:

$$\begin{aligned}\mathrm{E}\left\{\left\|\hat{\mathbf{J}}^{\mathrm{LOC}}-\mathbf{J}\right\|^2\right\} &= \sum_{i=1}^{p}\sum_{j\in[i\ N_i]}\mathrm{E}\left\{\left(\hat{\mathbf{J}}^{\mathrm{LOC}}_{i,j}-\mathbf{J}_{i,j}\right)^2\right\}\\ &=\sum_{i=1}^{p}\sum_{k=1}^{1+|N_i|}\mathrm{E}\left\{\left(\hat{\mathbf{J}}^{i}_{1,k}-\mathbf{J}^{i}_{1,k}\right)^2\right\}\\ &=\sum_{i=1}^{p}\sum_{k=1}^{1+|N_i|}\mathrm{bias}^2\left(\mathbf{J}^{i}_{1,k}\right)+\mathrm{var}\left(\hat{\mathbf{J}}^{i}_{1,k}\right).\end{aligned} \tag{34}$$

Assuming that $T > |N_i| + 1$, the biases and the variances satisfy

$$\mathrm{bias}\left(\mathbf{J}^{i}_{1,k}\right)=\left(1-\frac{T}{T-|N_i|-2}\right)\mathbf{J}^{i}_{1,k}, \tag{35}$$

$$\mathrm{var}\left(\hat{\mathbf{J}}^{i}_{1,k}\right)=\frac{\left(T-|N_i|\right)\left[\mathbf{J}^{i}_{1,k}\right]^2+\left(T-|N_i|-2\right)\mathbf{J}^{i}_{1,1}\mathbf{J}^{i}_{k,k}}{\left(T-|N_i|-1\right)\left(T-|N_i|-2\right)^2\left(T-|N_i|-4\right)/T^2}. \tag{36}$$

As expected, these results show that the MSE goes to zero as $T \to \infty$, i.e. the estimator is consistent, and that the bias and variance will be small when $T \gg |N_i|$ for all i.

The results above characterize the sum of mean squared errors over all the elements in $\mathbf{J}$, but do not address the correlation between these errors. For this purpose we turn to asymptotic error analysis. As proven in Prop. 1, the local estimator is an M-estimator, i.e., a solution to a function maximization or minimization problem, for which there are known asymptotic analysis results. Using the non-symmetric vector parameterization, we rewrite (15) as

$$\hat{\boldsymbol{\theta}}^{\mathrm{LOC}}_{ns}=\arg\max_{\theta_{ns}}\sum_{t=1}^{T}\log\tilde{p}\left(\mathbf{x}[t];\mathbf{J}\left(\boldsymbol{\theta}_{ns}\right)\right). \tag{37}$$

Under technical conditions, the asymptotic error covariance satisfies Huber's sandwich formula [39, Section 6.3]:

$$T\mathrm{E}\left\{\left(\hat{\boldsymbol{\theta}}^{\mathrm{LOC}}_{ns}-\boldsymbol{\theta}_{ns}\right)\left(\hat{\boldsymbol{\theta}}^{\mathrm{LOC}}_{ns}-\boldsymbol{\theta}_{ns}\right)^T\right\}\to\mathbf{H}_{ns}^{-1}\mathbf{G}_{ns}\mathbf{H}_{ns}^{-1}, \tag{38}$$

where

$$\begin{aligned}\mathbf{G}_{ns} &= \mathrm{E}\left\{\frac{\partial\log\tilde{p}\left(\mathbf{x}[t];\mathbf{J}\left(\boldsymbol{\theta}_{ns}\right)\right)}{\partial\boldsymbol{\theta}_{ns}}\frac{\partial^T\log\tilde{p}\left(\mathbf{x}[t];\mathbf{J}\left(\boldsymbol{\theta}_{ns}\right)\right)}{\partial\boldsymbol{\theta}_{ns}}\right\}\\ \mathbf{H}_{ns} &= \mathrm{E}\left\{\frac{\partial^2\log\tilde{p}\left(\mathbf{x}[t];\mathbf{J}\left(\boldsymbol{\theta}_{ns}\right)\right)}{\partial\boldsymbol{\theta}_{ns}\partial\boldsymbol{\theta}_{ns}^T}\right\}.\end{aligned} \tag{39}$$

Due to space limitations, we omit the details on these straight forward computations. We do note that the log-pseudo-likelihood is linear in the sample covariance so that $\mathbf{G}_{ns}$ and $\mathbf{H}_{ns}$ are linear combinations of

$$\mathrm{E}\left\{\mathbf{S}_{i,j}\right\}=\boldsymbol{\Sigma}_{i,j}, \tag{40}$$

and

$$\begin{aligned}&\mathrm{cov}\left(\mathbf{S}_{i_1,j_1},\mathbf{S}_{i_2,j_2}\right)\\ &=\begin{cases}\frac{2}{T}\boldsymbol{\Sigma}^2_{i_1,i_2} & i_1=j_1, i_2=j_2\\ \frac{2}{T}\boldsymbol{\Sigma}_{i_2,i_1}\boldsymbol{\Sigma}_{j_2,j_1} & \begin{array}{c}i_1=j_1, i_2\neq j_2\\ \text{or } i_1\neq j_1, i_2=j_2\end{array}\\ \frac{1}{T}\boldsymbol{\Sigma}_{i_1,i_2}\boldsymbol{\Sigma}_{j_1,j_2}+\frac{1}{T}\boldsymbol{\Sigma}_{i_1,j_2}\boldsymbol{\Sigma}_{j_1,i_2} & i_1\neq j_1, i_2\neq j_2.\end{cases}\end{aligned} \tag{41}$$

C. Average estimator

The AVE estimator in (14) is a simple linear transformation of the LOC estimator. This transformation can only reduce the mean squared Frobenius error. Indeed, due to the symmetry of $\mathbf{J}$, we have

$$\begin{aligned}&\left(\hat{\mathbf{J}}^{\mathrm{LOC}}_{i,j}-\mathbf{J}_{i,j}\right)^2+\left(\hat{\mathbf{J}}^{\mathrm{LOC}}_{j,i}-\mathbf{J}_{j,i}\right)^2\\ &\quad=\frac{1}{2}\left(\hat{\mathbf{J}}^{\mathrm{LOC}}_{i,j}-\hat{\mathbf{J}}^{\mathrm{LOC}}_{j,i}\right)^2+2\left(\hat{\mathbf{J}}^{\mathrm{AVE}}_{i,j}-\mathbf{J}_{i,j}\right)^2\\ &\quad\geq 2\left(\hat{\mathbf{J}}^{\mathrm{AVE}}_{i,j}-\mathbf{J}_{i,j}\right)^2.\end{aligned} \tag{42}$$

Summing over all $i \neq j$, $(i, j) \in E$ indices and adding the diagonal errors leads to

$$\mathrm{E}\left\{\left\|\hat{\mathbf{J}}^{\mathrm{AVE}}-\mathbf{J}\right\|^2\right\}\leq\mathrm{E}\left\{\left\|\hat{\mathbf{J}}^{\mathrm{LOC}}-\mathbf{J}\right\|^2\right\}. \tag{43}$$

This inequality can also be derived by noting that AVE is the orthogonal projection of LOC onto the convex set of symmetric matrices, and using the projection onto convex sets theorem (see [13] for more details). We emphasize that this result is non-asymptotic and holds for finite samples.

A more precise (yet asymptotic) characterization of the errors of AVE can be obtained via those of LOC. Using their vector parameterizations, the estimators are related as

$$\hat{\boldsymbol{\theta}}^{\mathrm{AVE}}_{s}=\mathbf{A}\hat{\boldsymbol{\theta}}^{\mathrm{LOC}}_{ns} \tag{44}$$

where $\mathbf{A}$ is a matrix that averages the elements in $\hat{\boldsymbol{\theta}}^{local}_{ns}$ associated with the same off diagonal elements. The true inverse covariance is symmetric and also satisfies $\boldsymbol{\theta}_s = \mathbf{A}\boldsymbol{\theta}_{ns}$. Therefore, the covariance of the error can be expressed as

$$\begin{aligned}&\mathrm{E}\left\{\left(\hat{\boldsymbol{\theta}}^{\mathrm{AVE}}_{s}-\boldsymbol{\theta}_s\right)\left(\hat{\boldsymbol{\theta}}^{\mathrm{AVE}}_{s}-\boldsymbol{\theta}_s\right)^T\right\}\\ &\quad=\mathbf{A}\mathrm{E}\left\{\left(\hat{\boldsymbol{\theta}}^{\mathrm{LOC}}_{ns}-\boldsymbol{\theta}_{ns}\right)\left(\hat{\boldsymbol{\theta}}^{\mathrm{LOC}}_{ns}-\boldsymbol{\theta}_{ns}\right)^T\right\}\mathbf{A}^T,\end{aligned} \tag{45}$$

where the limit of the inner covariance is provided in (38).

D. Pseudo-likelihood estimator

Asymptotic performance analysis of the PML in (22) is similar to that of LOC. PML is an M-estimator defined as

$$\hat{\boldsymbol{\theta}}^{\mathrm{PML}}_{ns}=\arg\max_{\theta_{s}}\sum_{t=1}^{T}\log\tilde{p}\left(\mathbf{x}[t];\mathbf{J}\left(\boldsymbol{\theta}_{s}\right)\right). \tag{46}$$

where the only difference from (37) is that PML uses the symmetric vector parameterization. Thus, similarly to (38), its asymptotic error covariance is given by

$$T\mathrm{E}\left\{\left(\hat{\boldsymbol{\theta}}^{\mathrm{PML}}_{s}-\boldsymbol{\theta}_{s}\right)\left(\hat{\boldsymbol{\theta}}^{\mathrm{PML}}_{s}-\boldsymbol{\theta}_{s}\right)^T\right\}\to\mathbf{H}_{s}^{-1}\mathbf{G}_{s}\mathbf{H}_{s}^{-1}, \tag{47}$$

where

$$\begin{aligned}\mathbf{G}_s &= \mathrm{E}\left\{\frac{\partial \log \tilde{p}\left(\mathbf{x}[t];\mathbf{J}\left(\boldsymbol{\theta}_s\right)\right)}{\partial \boldsymbol{\theta}_s}\frac{\partial^T \log \tilde{p}\left(\mathbf{x}[t];\mathbf{J}\left(\boldsymbol{\theta}_s\right)\right)}{\partial \boldsymbol{\theta}_s}\right\}\\ \mathbf{H}_s &= \mathrm{E}\left\{\frac{\partial^2 \log \tilde{p}\left(\mathbf{x}[t];\mathbf{J}\left(\boldsymbol{\theta}_s\right)\right)}{\partial \boldsymbol{\theta}_s \partial \boldsymbol{\theta}_s^T}\right\}. \qquad (48)\end{aligned}$$

Unlike AVE which never degrades the Frobenius MSE, for PML this is not guaranteed. The reason is that the pseudo-likelihood and likelihood objectives are not directly related to MSE. For large sample size, the global estimator is optimal in terms of MSE and additional linear constraints can only lower the error^2. Unfortunately, this may not be true for PML which is sub-optimal even under asymptotic conditions. Numerical examples in representative settings provided in the next section suggest that PML is usually better than AVE both in terms of finite sample performance as well as its asymptotic errors.

E. Positive definiteness

As discussed in Sec. III-E, performance may be improved by adding a positive definiteness constraint to (22). However, assuming that the true inverse covariance is strictly positive definite and in the interior of the feasible set, it is well known that relaxing the inequality constraints cannot affect the asymptotic performance of M-estimators (or the Cramer Rao performance bound [40]). A similar phenomenon in the context of covariance estimation was recently obtained in [41]. On the other hand, for finite sample size, incorporation of positive definiteness constraints can likely improve performance. This is a worthwhile goal for future study.

V. Numerical results

In this section, we demonstrate the performance advantages of the distributed estimators using synthetic simulations and an experiment with real world data.

In the first synthetic simulation, we randomly generate a networks of $p = 500$ sensors whose locations are uniformly distributed over the unit square. We connect each sensor (node) to its four nearest neighbors. We then compute the values of the inverse covariance as follows: $\mathbf{J}_{i,j} = 0$ if nodes i and j are not connected, and $\mathbf{J}_{i,j} = e^{-0.7 d_{i,j}}$ where $d_{i,j}$ is their distance, otherwise. We add an arbitrary small value to the diagonal elements in order to guarantee that the true covariance matrix be positive definite. We keep the network fixed throughout the simulation. The topology and its associated graphical model are illustrated in Fig. 2. We then perform 200 experiments in which we generate T independent realizations of $\mathbf{x}$ with $p = 500$ and estimate $\mathbf{J}$ using the four estimators. PML is implemented using the message passing protocol described above with $w_i = 1$ for all i, $c = 0.05$ and 20 iterations of (24)-(26). We report the average performance using three performance measures:

- Fig. 2: normalized MSE in the inverse covariance:

$$\frac{\left\|\hat{\mathbf{J}} - \mathbf{J}\right\|^2}{\left\|\mathbf{J}\right\|^2}. \qquad (49)$$

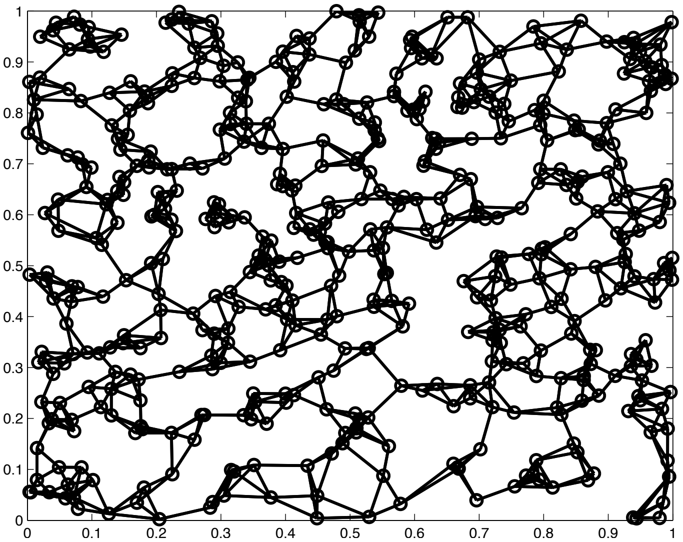

Fig. 1. Topology of the random network.

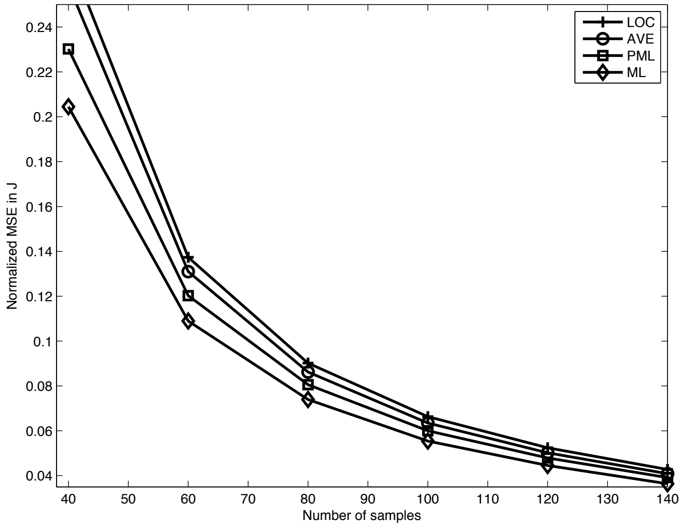

Fig. 2. Normalized MSE in $\mathbf{J}$ as a function of T.

- Fig. 3: normalized MSE in the covariance:

$$\frac{\left\|\hat{\mathbf{J}}^{-1} - \mathbf{J}^{-1}\right\|^2}{\left\|\mathbf{J}^{-1}\right\|^2}. \qquad (50)$$

- Fig. 4: normalized MSE in predicting $\mathbf{x}_s$ given $\mathbf{x}_{V\setminus s}$ where $s = \{1, \cdots, 100\}$ assuming the MMSE predictor in (6) and replacing the unknown $\mathbf{J}$ with its estimate $\hat{\mathbf{J}}$:

$$\frac{\mathrm{Tr}\left\{\left[\mathbf{I} - \hat{\mathbf{J}}_{ss}^{-1}\hat{\mathbf{J}}_{s,V\setminus s}\right]\mathbf{J}^{-1}\left[\mathbf{I} - \hat{\mathbf{J}}_{ss}^{-1}\hat{\mathbf{J}}_{s,V\setminus s}\right]^T\right\}}{\mathrm{Tr}\left\{\left[\mathbf{J}^{-1}\right]_{s,s}\right\}} \qquad (51)$$

The graphs illustrate the advantages of the symmetry enforcing estimators which succeed in closing about half of the performance gap between the local and global estimators. As expected, the PML estimator outperforms the naive AVE estimator.

[2] Asymptotically, ML estimation attains the CRB whereas linearly constrained ML estimation attains the constrained CRB which is always lower or equal to the CRB [40].

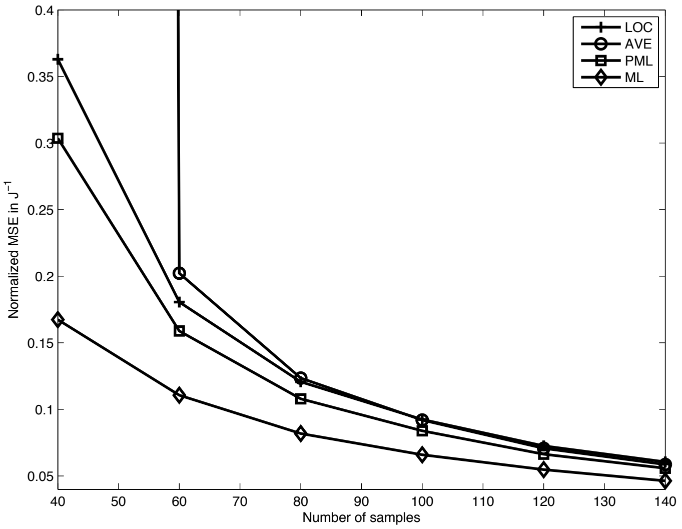

Fig. 3. Normalized MSE in $\mathbf{J}^{-1}$ as a function of T.

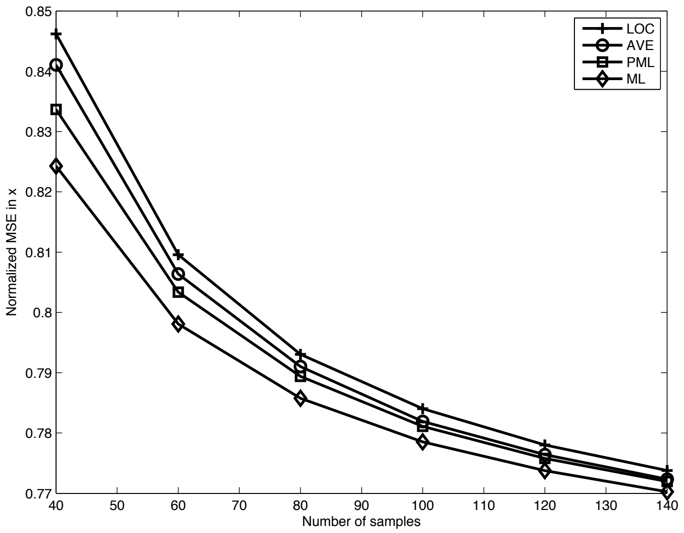

Fig. 4. Normalized MSE in $\mathbf{x}$ as a function of T.

In our next synthetic simulation, we demonstrate and verify the asymptotic performance analysis presented in Section IV. For this purpose, we choose a simple GGM which consists of a four node loop, i.e., four nodes $V = \{1, \cdots, 4\}$ with edges $(1,2)$, $(2,3)$, $(3,4)$ and $(4,1)$. We let the diagonal elements in $\mathbf{J}$ be equal to 2.01 and let the non-zero off-diagonal elements be equal to 1. We then simulate LOC, AVE, PML and ML and compare their averages errors over 10000 independent experiments to the asymptotic errors given in (38), (45), (47), and (32), respectively. The results of this simulation are presented in Fig. 5 and demonstrate the tightness of the analysis in the small error regime. As explained in Section IV, AVE is provably better than LOC, and, while PML has no provable performance improvement, it does perform even better in this specific setting (and many other settings we have experimented with).

Next, we present numerical results obtained from real world data. Following [43], [44], we use the *Lab* dataset. It contains temperature information from a sensor network of 54 motes deployed in the Intel Berkeley Research lab between February 28th and April 5th, 2004. This dataset is known to be very difficult. It has lots of missing data, noise and failed sensors. We preprocess the data as follows: we use 1800 consecutive samples per sensor, we interpolate missing readings (NaN) and we de-trend the data using a local rectangular window of 10 samples. Next, we compute the sample covariance and invert it to get the ground truth inverse covariance matrix $\mathbf{J}$. We derive a graphical model with a sparsity level of 70% by thresholding this matrix. Based on this graphical model and repeated sampling of n samples out of the 1800, we try to estimate $\mathbf{J}$ using the different proposed estimators. In our first attempt, LOC, AVE and ML performed as expected. However, the PML method failed to improve the performance of LOC and AVE. We suspected that the reason is the inhomogeneous nature of the sensors. Thus, we repeated the experiment again with a weighted PML using $w_i = \left[\hat{\mathbf{J}}^{\mathrm{LOC}}_{i,i}\right]^2$ for all i. This resulted in significantly better performance of PML. The average normalized squared Frobenius errors over 200 random re-samplings are reported in Fig. 6. The errors satisfy the expected order and illustrate the advantages of the proposed techniques.

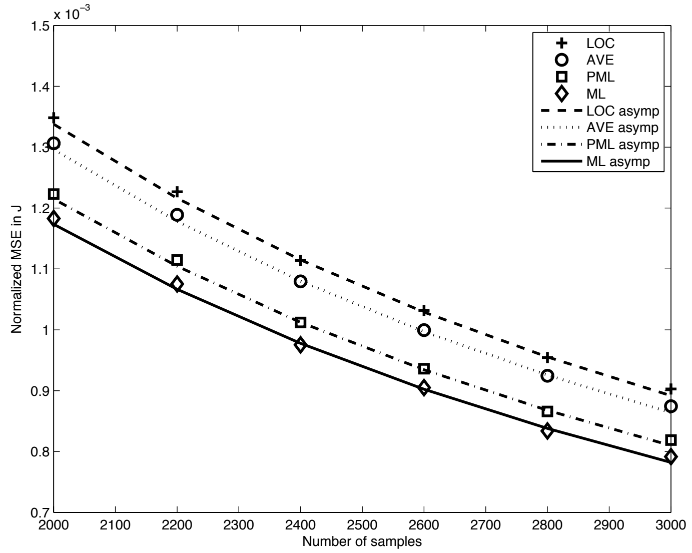

Fig. 5. Normalized MSE in a simple four node loop GGM. The curves denote the asymptotic limits, whereas the symbols present the estimated performance in a Monte Carlo simulation.

VI. Conclusion

In this paper we considered the problem of distributed inverse covariance estimation in GGM. For large dimensional problems with many nodes in the graph, implementing the global ML estimator has high computation and communication costs. As a lower cost alternative we proposed a natural local ML approach that, while asymptotically consistent, produces inverse covariance estimates that are not symmetric matrices. To overcome this lack of symmetry we introduced two modified estimators that enforce symmetry by averaging and by pseudo-likelihood optimization, respectively. These methods

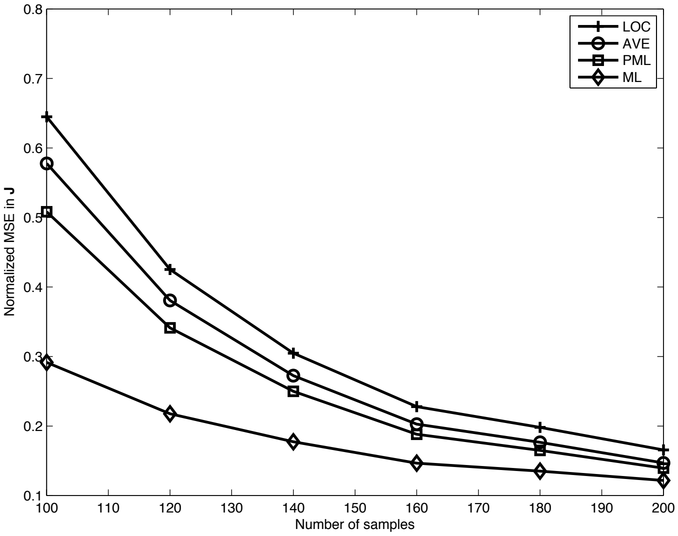

Fig. 6. Normalized MSE in $\mathbf{J}$ as a function of T in 'Lab' dataset. The curve labeled PML corresponds to the weighted PML estimator.

enforce the known symmetry in the inverse covariance through simple message passing protocols. We demonstrate their advantages using synthetic simulations as well as real world data from a wireless sensor network.

Acknowledgment

A. Wiesel would like to acknowledge discussions with Venkat Chandrasekaran. The authors would also like to thank the anonymous reviewers for their constructive comments which significantly improved the paper and led to the weighted PML estimator.

References

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, Jul 1996.

[2] E. Dougherty, A. Datta, and C. Sima, "Research issues in genomic signal processing," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 46–68, Nov. 2005.

[3] M. Cetin, L. Chen, J. W. Fisher, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. S. Willsky, "Distributed fusion in sensor networks: A graphical models perspective," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 42– 55, July 2006.

[4] A. Wiesel and A. O. Hero, "Decomposable principal component analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4369 –4377, nov. 2009.

[5] V. Delouille, R. Neelamani, and R. Baraniuk, "Robust distributed estimation using the embedded subgraphs algorithm," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 2998–3010, 2006.

[6] Y. Weiss and W. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *Neural Computation*, vol. 13, no. 10, pp. 2173–2200, 2001.

[7] V. Chandrasekaran, J. Johnson, and A. Willsky, "Estimation in Gaussian graphical models using tractable subgraphs: A walk-sum analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, p. 1916, 2008.

[8] J. Johnson, D. Malioutov, and A. Willsky, "Walk-sum interpretation and analysis of Gaussian belief propagation," *Advances in Neural Information Processing Systems*, vol. 18, p. 579, 2006.

[9] E. Sudderth, M. Wainwright, and A. Willsky, "Embedded trees: estimation of Gaussian processes on graphs with cycles," *IEEE Transactions on Signal Processing*, vol. 52, no. 11, pp. 3136–3150, Nov. 2004.

[10] D. M. Malioutov, J. Johnson, and A. Willsky, "Walk-sums and belief propagation in Gaussian graphical models," *Journal of Machine Learning Research*, vol. 7, pp. 2031–2064, Oct. 2006.

[11] S. L. Lauritzen, *Graphical models*. New York: Oxford Statistical Science Series, 1996, vol. 17.

[12] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, pp. 157–175, 1972.

[13] A. Wiesel, Y. C. Eldar, and A. O. Hero, "Covariance estimation in decomposable Gaussian graphical models," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1482 –1492, march 2010.

[14] J. Dahl, L. Vandenberghe, and V. Roychowdhury, "Covariance selection for nonchordal graphs via chordal embedding," *Optimization Methods and Software*, vol. 23, no. 4, pp. 501–520, 2008.

[15] L. Vandenberghe, S. Boyd, and S. P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 19, no. 2, pp. 499–533, 1998.

[16] Y. Teh and M. Welling, "On improving the efficiency of the iterative proportional fitting procedure," in *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, vol. 9, 2003.

[17] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[18] D. Malioutov, J. Johnson, M. J. Choi, and A. Willsky, "Low-rank variance approximation in GMRF models: Single and multiscale approaches," *Signal Processing, IEEE Transactions on*, vol. 56, no. 10, pp. 4621 –4634, oct. 2008.

[19] P. Liang and M. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 584–591.

[20] J. Besag, "Statistical analysis of non-lattice data," *The statistician*, pp. 179–195, 1975.

[21] J. V. Dillon and G. Lebanon, "Statistical and Computational Tradeoffs in Stochastic Composite Likelihood," *Arxiv preprint arXiv:1003.0691v1*, 2010.

[22] B. Lindsay, "Composite likelihood methods," *Contemporary Mathematics*, vol. 80, no. 1, pp. 221–39, 1988.

[23] M. Wainwright, "Estimating the Wrong Graphical Model: Benefits in the Computation-Limited Setting," *The Journal of Machine Learning Research*, vol. 7, pp. 1829 – 1859, 2006.

[24] G. Rocha, P. Zhao, and B. Yu, "A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice)," *Arxiv preprint arXiv:0807.3734*, 2008.

[25] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," working paper on line, Stanford, Univ, Tech. Rep., 2010.

[26] D. Bertsekas and J. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Athena Scientific, 1997.

[27] Q. Ling and Z. Tian, "Decentralized sparse signal recovery for compressive sleeping wireless sensor networks," *IEEE Trans. on Signal Processing*, vol. PP, no. 99, pp. 1 –1, 2010.

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

[29] I. D. Schizas, G. B. Giannakis, and Z. Q. Luo, "Distributed estimation using reduced-dimensionality sensor observations," *IEEE Trans. on Signal Processing*, vol. 55, no. 8, pp. 4284–4299, Aug. 2007.

[30] I. Schizas, A. Ribeiro, and G. Giannakis, "Consensus in ad hoc WSNs with noisy links-part I: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[31] G. Mateos, J. Bazerque, and G. Giannakis, "Distributed sparse linear regression," *Signal Processing, IEEE Transactions on*, vol. 58, no. 10, pp. 5262 –5276, 2010.

[32] N. Meinshausen and P. Buhlmann, "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, pp. 1436–1462, 2006.

[33] A. Rothman, P. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[34] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing L1-penalized log-determinant divergence," *Advances in Neural Information Processing Systems*, 2008.

[35] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *The Annals of Statistics*, vol. 37, no. 6B, pp. 4254–4278, 2009.

[36] J. Johnson, "Fisher information in Gaussian graphical models," *unpublished technical note*, 2006.

[37] P. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[38] S. Press, “Applied multivariate analysis: using bayesian and frequentist methods of inference.” 1982.

[39] P. Huber and E. Ronchetti, *Robust statistics*. John Wiley and Sons Inc, 2009.

[40] J. Gorman and A. Hero, “Lower bounds for parametric estimation with constraints,” *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1285–1301, 1990.

[41] K. Werner, M. Jansson, and P. Stoica, “On estimation of covariance matrices with kronecker product structure,” *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 478–491, Feb. 2008.

[42] H. White, “Maximum likelihood estimation of misspecified models,” *Econometrica*, vol. 50, no. 1, pp. pp. 1–25, 1982. [Online]. Available: http://www.jstor.org/stable/1912526

[43] C. Guestrin, P. Bodik, R. Thibaux, M. Paskin, and S. Madden, “Distributed regression: an efficient framework for modeling sensor network data,” in *Proceedings of the 3rd international symposium on Information processing in sensor networks*. ACM, 2004, pp. 1–10.

[44] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong, “Model-driven data acquisition in sensor networks,” in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 2004, pp. 588–599.