Distributed Learning of Gaussian Graphical Models via Marginal Likelihoods

Zhaoshi Meng* **Dennis Wei*** **Ami Wiesel†** **Alfred O. Hero III***
*University of Michigan
{mengzs,dlwei,hero}@eecs.umich.edu
†The Hebrew University of Jerusalem
ami.wiesel@huji.ac.il

Abstract

We consider distributed estimation of the inverse covariance matrix, also called the concentration matrix, in Gaussian graphical models. Traditional centralized estimation often requires iterative and expensive global inference and is therefore difficult in large distributed networks. In this paper, we propose a general framework for distributed estimation based on a maximum marginal likelihood (MML) approach. Each node independently computes a local estimate by maximizing a marginal likelihood defined with respect to data collected from its local neighborhood. Due to the non-convexity of the MML problem, we derive and consider solving a convex relaxation. The local estimates are then combined into a global estimate without the need for iterative message-passing between neighborhoods. We prove that this relaxed MML estimator is asymptotically consistent. Through numerical experiments on several synthetic and real-world data sets, we demonstrate that the two-hop version of the proposed estimator is significantly better than the one-hop version, and nearly closes the gap to the centralized maximum likelihood estimator in many situations.

1 Introduction

Graphical models provide a principled framework for compactly representing dependencies among random variables (Lauritzen, 1996; Wainwright & Jordan, 2008). The ability to perform distributed and efficient inference makes them especially well-suited to large networks, *e.g.*, sensor networks, social networks, gene regulatory networks, and smart grids (Liu & Ihler, 2012; Meng et al., 2012; Wiesel & Hero, 2012). Estimating graphical model parameters is therefore an important first step in enabling these applications.

For Gaussian graphical models (GGM), parameter estimation essentially reduces to covariance estimation and is further simplified by the graph structure: the pattern of edges in the graph corresponds to the sparsity pattern of the inverse covariance matrix, also known as the concentration or precision matrix. When the sparsity pattern is known *a priori*, the traditional technique for inverse covariance estimation is maximum likelihood (ML), resulting in a convex optimization that can be solved either by generic solvers or specialized algorithms (Banerjee et al., 2006; Dahl et al., 2008; Friedman et al., 2008). However, solving the exact ML estimation problem requires either expensive data communication (e.g., to a fusion center) and centralized computation/storage, or in a distributed setting, iterative message-passing in loopy graphs to compute global derivatives (Koller & Friedman, 2009). Neither of those strategies are feasible in large, generally structured and distributed networks, such as a sensor network where the corresponding graph could be very loopy and the computational resources are highly decentralized.

Given the limitations of centralized solutions, researchers have begun proposing distributed methods for graphical model learning, where agents in the network estimate local parameters by processing local data with limited communication between agents (Liu & Ihler, 2012; Wiesel & Hero, 2012). In addition to decentralizing computation and communication across the network, distributed solutions are naturally robust against localized attacks or failures, and are also more protective of data privacy. These advantages make distributed algorithms particularly appealing for network applications.

This paper proposes a general framework for distributed estimation based on *marginal* likelihoods.

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

Each node collects data within its neighborhood and independently forms a local estimate by maximizing a marginal likelihood. We show that this marginal likelihood approach is sufficient in principle to determine the global parameter. Notably, the formation of our global estimate from local estimates does not require iterative consensus or message-passing as in Liu & Ihler (2012); Wiesel & Hero (2012). Due to the non-convexity of maximum marginal likelihood (MML) estimation, we derive and consider solving a convex relaxation of the problem. We prove that the relaxed MML estimator is asymptotically consistent. We then focus on two specific cases in which the local neighborhoods correspond to either one or two communication hops through the network. Extensive experiments show that the two-hop version of the proposed estimator is significantly better than the one-hop version and nearly closes the gap to the centralized ML estimator in many situations.

The outline of the paper is as follows. In Section 2, we give a brief review of graphical models and centralized ML estimation. In Section 3, we propose a general approach to distributed estimation based on marginal likelihoods, show asymptotic consistency, and discuss two important specific cases. Numerical experiments are presented in Section 4 and the paper concludes in Section 5.

1.1 Related work

Recent works by Liu & Ihler (2012); Wiesel & Hero (2012) have discussed a similar but less general framework for distributed estimation in which several pseudo-likelihood-based estimators for GGMs and general exponential families are proposed. In Section 3.4, we show that the estimators in Wiesel & Hero (2012) can be seen as a special case of the proposed relaxed MML estimator, and a significant performance improvement can be obtained through a reasonable increase in the computation and communication cost.

Our distributed estimator is formed through solving multiple relaxed MML estimation problems by local agents. The structure of each local MML problem is related to the latent graphical model considered in Chandrasekaran et al. (2012). However, the assumption in Chandrasekaran et al. (2012) is that the number of latent variables is relatively small, thus contributing a low-rank term, whereas in our case, all variables outside the local neighborhood can be seen as latent variables.

Another line of work (Friedman et al., 2008; Johnson et al., 2011; Ravikumar et al., 2011; Rothman et al., 2008) focuses on the problem of covariance selection, where the graph topology is not known *a priori* and must be estimated in addition to the parameters. This appears to be a harder problem than ours since we assume the structure is known, but some insights regarding distributed algorithms in particular can perhaps be shared.

Notation. Boldface upper case letters denote matrices and boldface lower case letters denote column vectors. Sets of single indices are denoted by calligraphic upper case letters. The cardinality of a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$ and the difference of two sets is denoted as $\mathcal{A}\backslash\mathcal{B}$. Following common notation, $\mathbf{A}_{\mathcal{M},\mathcal{N}}$ represents a submatrix of $\mathbf{A}$ with rows indexed by $\mathcal{M}$ and columns indexed by $\mathcal{N}$. We also make reference to irregular sets of index pairs such as the edge set E of a graph, for which we use standard upper case letters. $\mathbf{A}_E$ then refers to the vector of entries of $\mathbf{A}$ indexed by E. The standard inner product between two symmetric matrices is denoted as $\langle \mathbf{A}, \mathbf{B} \rangle$, i.e., $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{trace}(\mathbf{AB}) = \sum_{i,j} \mathbf{A}_{i,j}\mathbf{B}_{i,j}$.

2 Background

2.1 Graphical models

We refer the reader to Lauritzen (1996) for a detailed treatment of graphical models. We consider a p-dimensional random vector $\mathbf{x}$ following a graphical model with respect to an undirected graph $\mathcal{G} = (V, E)$, where $V = \{1, \ldots, p\}$ is a set of nodes corresponding to elements of $\mathbf{x}$ and E is a set of edges connecting nodes. The vector $\mathbf{x}$ satisfies the Markov property with respect to $\mathcal{G}$ if for any pair of nonadjacent nodes in $\mathcal{G}$, the corresponding pair of variables in $\mathbf{x}$ are conditionally independent given the remaining variables.

For Gaussian graphical models (GGM), the vector $\mathbf{x}$ follows a multivariate Gaussian distribution. We assume without loss of generality that $\mathbf{x}$ has zero mean. Then the probability density function can be written in canonical form in terms of the concentration matrix $\mathbf{J}$ as follows:

$$p(\mathbf{x}; \mathbf{J}) = (2\pi)^{-p/2} (\det \mathbf{J})^{1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{J} \mathbf{x}\right). \quad (1)$$

The Markov property manifests itself in a simple way through the sparsity pattern of $\mathbf{J}$:

$$\mathbf{J}_{i,j} = 0 \text{ for all } (i,j) \notin E. \quad (2)$$

This property leads to efficient inference algorithms for GGMs, and as we will show, it can also be exploited for distributed parameter learning.

2.2 Centralized estimation in GGMs

Maximum likelihood estimation (MLE) is a classical approach to estimating model parameters from data.

For Gaussian graphical models, the problem reduces to estimating the non-zero elements of the concentration matrix $\mathbf{J}$. These elements are indexed by $\widetilde{E}$, the union of the edges and pairs corresponding to diagonal elements,

$$\widetilde{E} := E \cup \{(i,i)\}_{i=1}^{p}. \tag{3}$$

When all the data are accessible, the centralized global maximum likelihood (GML) estimation problem can be formulated as Lauritzen (1996)

$$\begin{aligned} \widehat{\mathbf{J}}^{\mathrm{GML}} = \arg\min_{\mathbf{J}} \ & \langle \widehat{\mathbf{\Sigma}}, \mathbf{J} \rangle - \log\det \mathbf{J} \\ \text{s.t.} \ \ & \mathbf{J}_{j,k} = 0 \quad \forall\, (j,k) \notin \widetilde{E}, \\ & \mathbf{J} \succeq \mathbf{0}. \end{aligned} \tag{4}$$

where $\widehat{\mathbf{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}(t)\mathbf{x}(t)^T$ is the sample covariance matrix and $\mathbf{x}(1), \ldots, \mathbf{x}(T)$ are i.i.d. samples of $\mathbf{x}$. Since the GML problem (4) is convex in $\mathbf{J}_{\widetilde{E}}$, efficient gradient-based algorithms can be applied, many of which have specialized implementations on graphs, *e.g.* iterative proportional fitting (IPF) (Wainwright & Jordan, 2008), chordally-embedded Newton's method (Dahl et al., 2008), and an iterative regression method introduced in Friedman et al. (2009). However, as mentioned before, the main drawback of these methods is the computational and communication complexity when implemented in networks.

3 Distributed Estimation in GGMs

We now consider a general approach to distributed parameter estimation motivated by network applications. Each random variable $\mathbf{x}_i$ is associated with a node in the network. The topology of the graph $\mathcal{G}$, which encodes statistical dependences, is assumed to coincide with the topology of internode communication. Each node collects all the data samples from within a neighborhood and computes a local parameter estimate. A global estimate of $\mathbf{J}$ is then formed by combining these local estimates. The key components in this framework are the choice of neighborhood, the local estimation method, and the method for combining estimates.

3.1 Marginal Likelihood Maximization

In this paper, we propose to estimate local parameters by maximizing *marginal likelihood* functions. For each node i, define the index set for its *immediate neighbors* as

$$\mathcal{I}_i := \{j \mid (i,j) \in E\}, \tag{5}$$

and consider a neighborhood indexed by a set $\mathcal{N}_i$ containing the nodes within a certain number of communication hops from node i in the network. By definition, the set $\mathcal{N}_i$ includes at least the node i itself and its immediate neighbors $\mathcal{I}_i$. Let $\mathbf{K}$ denote the concentration matrix corresponding to the marginal distribution over the variables $\{\mathbf{x}_j, j \in \mathcal{N}_i\}$ in the neighborhood, and let

$$\mathbf{S}^i := \widehat{\mathbf{\Sigma}}_{\mathcal{N}_i,\mathcal{N}_i} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{\mathcal{N}_i}(t)\mathbf{x}_{\mathcal{N}_i}(t)^T$$

be the local sample covariance matrix. Then the maximum marginal likelihood (MML) estimation problem in neighborhood $\mathcal{N}_i$ can be formulated as a joint optimization of $\mathbf{K}$ and $\mathbf{J}$:

$$\begin{aligned} \widehat{\mathbf{K}}^{i,\mathrm{MML}} = \arg\min_{\mathbf{K},\mathbf{J}} \ & \langle \mathbf{S}^i, \mathbf{K} \rangle - \log\det \mathbf{K} \\ \text{s.t.} \ \ & \mathbf{K} = \left[\left(\mathbf{J}^{-1} \right)_{\mathcal{N}_i,\mathcal{N}_i} \right]^{-1}, \\ & \mathbf{J}_{j,k} = 0 \quad \forall\, (j,k) \notin \widetilde{E}, \\ & \mathbf{J} \succeq \mathbf{0}, \end{aligned} \tag{6}$$

where the first constraint represents the marginalization relationship between $\mathbf{K}$ and $\mathbf{J}$ and the second line of constraints reflects the sparsity of the global parameter.

A global estimate of $\mathbf{J}$ can be formed by combining optimal solutions to problem (6) for each node i. We consider a simple concatenation approach whose justification will be given in Section 3.2 below. Denote by L_i the set of index pairs corresponding to the non-zero entries in the ith row of $\mathbf{J}$, i.e.,

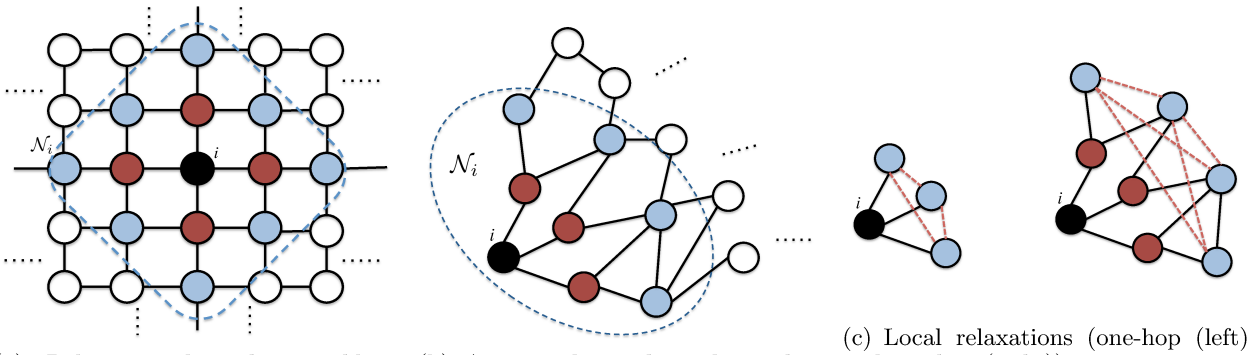
$$L_i := \{(j,k) \in \widetilde{E} \mid j = i\}. \tag{7}$$

Note that the self-edge pair (i,i) is included in L_i given the definition of $\widetilde{E}$ in (3). We refer to the parameters indexed by L_i as the *row parameters* for node i. The global estimate is constructed by extracting the row parameters from each local estimate and concatenating them:

$$\widehat{\mathbf{J}}^{\mathrm{MML}}_{L_i} = \widehat{\mathbf{K}}^{i,\mathrm{MML}}_{L_i}, \quad \text{for } i = 1, \ldots, p. \tag{8}$$

Note that this construction of the global estimate $\mathbf{J}^{\mathrm{MML}}$ (and also the other surrogate estimator proposed below) does not guarantee symmetry, since it is not a major concern of this work. However, as will shown in Section 3.3, the symmetry is naturally recovered in the asymptotic limit; also it can be imposed in the non-asymptotic case through a simple strategy (see Section 3.6).

The difficulty with the MML approach is that problem (6) is in general a non-convex optimization. The non-convexity arises from the coupling of the nonlinear marginalization constraint linking $\mathbf{K}$ to $\mathbf{J}$ and the sparsity constraints on $\mathbf{J}$. In the next subsection, we derive a convex relaxation of the MML estimation problem as a surrogate.



(a) 2D lattice and two-hop neighborhood (b) A general graph and two-hop neighborhood (c) Local relaxations (one-hop (left) and two-hop (right))

Figure 1: Illustration of defined sets and local relaxation. In (a) and (b) we show two different graphs, in which the two-hop neighborhood $\mathcal{N}$ for node i is indicated with dashed contours. The buffer set variables $\mathbf{x}_\mathcal{B}$ and the protected set variables $\mathbf{x}_\mathcal{P}$ (excluding node i itself) are colored blue and red, respectively. For the graph in (b), we illustrate the one-hop and two-hop local relaxation problems in (c). The dashed red lines in (c) denote the fill-in edges due to relaxation.

3.2 Convex Relaxation of Marginal Likelihood Maximization

In the remainder of the section, we suppress the node index i whenever possible for clarity. To derive a convex relaxation of the MML estimation problem (6), we apply the Schur complement identity to the marginalization constraint in (6), yielding

$$\mathbf{K} = \mathbf{J}_{\mathcal{N},\mathcal{N}} - \mathbf{J}_{\mathcal{N},\mathcal{N}^C} \cdot \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \cdot \mathbf{J}_{\mathcal{N}^C,\mathcal{N}}, \tag{9}$$

where $\mathcal{N}^C$ is the complementary set to $\mathcal{N}$. Define the *buffer set* $\mathcal{B} \subset \mathcal{N}$ as the set of all variables in $\mathcal{N}$ that have immediate neighbors in the complement $\mathcal{N}^C$,

$$\mathcal{B} := \{j \mid j \in \mathcal{N} \text{ and } \mathcal{I}_j \cap \mathcal{N}^C \neq \emptyset\}. \tag{10}$$

The difference set between $\mathcal{N}$ and $\mathcal{B}$ is referred to as the *protected set* $\mathcal{P} := \mathcal{N} \backslash \mathcal{B}$. The buffer and protected sets are illustrated in Figure 1(a) and 1(b). Due to the Markov property, we have $\mathbf{J}_{\mathcal{P},\mathcal{N}^C} = \mathbf{0}$. Decomposing $\mathcal{N}$ into $\mathcal{B}$ and $\mathcal{P}$ then reveals the sparsity pattern of $\mathbf{K}$ from (9):

$$\begin{aligned}\mathbf{K} &= \mathbf{J}_{\mathcal{N},\mathcal{N}} - \begin{bmatrix} \mathbf{0} \\ \mathbf{J}_{\mathcal{B},\mathcal{N}^C} \end{bmatrix} \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \begin{bmatrix} \mathbf{0}, & \mathbf{J}_{\mathcal{N}^C,\mathcal{B}} \end{bmatrix} \\ &= \mathbf{J}_{\mathcal{N},\mathcal{N}} - \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{\mathcal{B},\mathcal{N}^C} \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \mathbf{J}_{\mathcal{N}^C,\mathcal{B}} \end{bmatrix}\end{aligned}$$

and hence

$$\mathbf{K}_{\mathcal{P},\mathcal{P}} = \mathbf{J}_{\mathcal{P},\mathcal{P}}, \quad \mathbf{K}_{\mathcal{P},\mathcal{B}} = \mathbf{J}_{\mathcal{P},\mathcal{B}}, \tag{11}$$

$$\mathbf{K}_{\mathcal{B},\mathcal{B}} = \mathbf{J}_{\mathcal{B},\mathcal{B}} - \mathbf{J}_{\mathcal{B},\mathcal{N}^C} \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \mathbf{J}_{\mathcal{N}^C,\mathcal{B}}. \tag{12}$$

An important observation from (11) is that the sparsity pattern of $\mathbf{J}_{\mathcal{N},\mathcal{N}}$ is entirely preserved in the rows and columns indexed by the protected set $\mathcal{P}$. In particular, because node i itself is always protected, the marginal and global concentration matrices share the same values of the row parameters indexed by L, which motivates our strategy for fusing local estimates in (8). On the other hand, the sparsity pattern in the "buffer submatrix" $\mathbf{K}_{\mathcal{B},\mathcal{B}}$ is in general modified due to the fill-in term, *i.e.*, the second term in (12).

Based on these observations, we now specify a relaxed set of constraints on the marginal concentration matrix $\mathbf{K}$. First denote the set of all local edges that are not affected by the fill-in term in (12) as

$$E^{\text{Prot}} := \widetilde{E} \cap \{\{\mathcal{P} \times \mathcal{P}\} \cup \{\mathcal{P} \times \mathcal{B}\} \cup \{\mathcal{B} \times \mathcal{P}\}\},$$

where the superscript stands for "protected". We then add to E^{Prot} all index pairs $\mathcal{B} \times \mathcal{B}$ that could potentially be affected by fill-in in (12), resulting in a *relaxed edge set* R (see Figure 1(c) for illustrations):

$$R = E^{\text{Prot}} \cup \{\mathcal{B} \times \mathcal{B}\}. \tag{13}$$

In light of (11) and (12), any feasible marginal concentration matrix $\mathbf{K}$ for the MML estimation problem (6) is guaranteed to be supported only on the set R. Therefore the following constraints specify a set that contains the feasible set of the MML problem (6):

$$\mathbf{K} \succeq \mathbf{0} \text{ and } \mathbf{K}_{j,k} = 0 \quad \forall\, (j,k) \notin R. \tag{14}$$

Using the constraints in (14), we formulate the following relaxation of the original MML estimation problem (6) at each node i:

$$\begin{aligned}\widehat{\mathbf{K}}^{i,\text{Relax}} = \underset{\mathbf{K}}{\arg\min}\ & \langle \mathbf{S}^i, \mathbf{K} \rangle - \log\det \mathbf{K} \\ \text{s.t.}\ \ & \mathbf{K}_{j,k} = 0 \quad \forall\, (j,k) \notin R, \\ & \mathbf{K} \succeq \mathbf{0}.\end{aligned} \tag{15}$$

The above relaxed MML problem is a convex optimization with respect to $\mathbf{K}_R$ and has the same form as the global MLE problem (4) but with much lower dimensions in general.

Since the original MML estimation problem (6) is difficult to solve, we propose to solve the relaxed MML estimation problem (15) as a surrogate to estimate the local parameters. Then a global estimate of the concentration matrix can be similarly obtained by extracting and concatenating row parameters as in (8):

$$\widehat{\mathbf{J}}_{L_i}^{\text{Relax}} = \widehat{\mathbf{K}}_{L_i}^{i,\text{Relax}}, \quad \text{for } i = 1, \ldots, p. \tag{16}$$

3.3 Asymptotic Consistency

In this section, we suppress the index i for matrices $\mathbf{S}$ and $\mathbf{K}^{\text{Relax}}$ for clarity. We prove that the relaxed MML estimator defined in (15) and (16) is asymptotically consistent. To do so, we first establish the following lemma by making use of the continuity of the mapping from sample covariance $\mathbf{S}$ to local estimate $\widehat{\mathbf{K}}^{\text{Relax}}$ in (15) (proof omitted).

Lemma 1. *The mapping from the sample covariance* $\mathbf{S}$ *to the concentration matrix* $\widehat{\mathbf{K}}^{Relax}$ *through solving the relaxed MML problem* (15) *is a continuous mapping provided that* $\mathbf{S}$ *is positive definite.*

Theorem 1. *The relaxed MML estimator* $\widehat{\mathbf{J}}^{Relax}$ *is asymptotically consistent.*

Proof of Theorem 1. In the asymptotic limit, the local sample covariance matrix $\mathbf{S}$ converges to the corresponding sub-matrix of the true covariance matrix, $\mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}}$, with probability one and is therefore positive definite. Let $\mathbf{K}^* = (\mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}})^{-1}$ be the corresponding true marginal concentration matrix. We first show that $\mathbf{K}^*$ is the optimal solution to the relaxed MML estimation problem (15) when $\mathbf{S} = \mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}}$. To see this, note that the true global concentration matrix $\mathbf{J}^*$ conforms to the sparsity pattern specified by $\widetilde{E}$, and therefore the marginal concentration matrix $\mathbf{K}^*$ is feasible for (6). Furthermore, from relations (11)-(12) and our construction of the edge set R, $\mathbf{K}^*$ also satisfies the sparsity constraints of the relaxed problem (15). It is straightforward to show that $\mathbf{K}^*$ satisfies the optimality conditions when $\mathbf{S} = \mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}}$. Then by strict convexity, $\mathbf{K}^*$ is the unique optimum of the relaxed MML problem when $\mathbf{S} = \mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}}$.

It now follows from Lemma 1 and the continuous mapping theorem (Van der Vaart, 2000) that the relaxed MML estimates $\widehat{\mathbf{K}}^{\text{Relax}}$ converge asymptotically to $\mathbf{K}^*$ with probability one. Lastly, due to the absence of fill-in in (11), the construction of the global estimate $\widehat{\mathbf{J}}^{\text{Relax}}$ in (16) also yields a consistent estimate of $\mathbf{J}^*$. □

It is interesting to note that the true marginal concentration matrix $\mathbf{K}^*$ is also the optimal solution to the MML estimation problem (6) when $\mathbf{S} = \mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}}$. This follows because the feasible region for the unrelaxed MML problem (with respect to $\mathbf{K}$) is contained in the feasible region for the relaxed problem (15). However, convergence of the MML estimates $\widehat{\mathbf{K}}^{\text{MML}}$ to $\mathbf{K}^*$ is less clear because the feasible region in the unrelaxed problem is not known to be convex.

Theorem 1 implies that an asymptotically consistent estimate of the global concentration matrix can be constructed by solving p local and convex estimation problems without any message-passing between neighborhoods. The highly localized nature of the proposed approach stands in sharp contrast to the centralized MLE, which requires either iterative global message-passing in triangulated graphs with potentially large tree-width, or tree-like approximations to avoid non-convergence in loopy graphs (Wainwright et al., 2003, 2005). Even in the case of chordal or decomposable GGMs, previous methods for estimating the concentration matrix require non-trivial message-passing between cliques, with cost that is quadratic in the dimensions of the separators, see *e.g.*, (Wiesel et al., 2010).

While Theorem 1 claims asymptotic consistency for all choices of local neighborhoods containing $\{i, \mathcal{I}_i\}$, different choices will yield estimators with different properties in the finite sample case, and each chosen local neighborhood might result in different convergence rates. In principle, larger neighborhoods would allow each node to access more data and hence increase its information for estimating its local parameters. In the extreme case, if each node has access to all the data in the network, the local estimate is equal to the global MLE. On the other hand, larger neighborhoods require more parameters to be estimated, thus tending to result in slower convergence rate towards the true parameters. Also the convex relaxation we propose removes the sparsity constraints on the buffer sub-matrix of the local concentration matrix. This relaxation would be expected to propagate and influence the parameter estimates in the protected edge set E^{Prot}, and in particular the row parameters that we extract. Larger local neighborhoods would result in larger protected edge sets, which might be expected to decrease the effect of the relaxation on the row parameters of interest.

In sum, the optimal choice of local neighborhoods is unclear due to the above trade-off. In the next subsections, we focus on two cases of the relaxed MML estimator with local neighborhood diameters corresponding to one and two communication hops through the network. The non-asymptotic behavior of the two estimators is studied through numerical experiments in

Section 4.

3.4 Case I: One-hop Estimator

We first consider a one-hop neighborhood consisting of node i and its immediate neighbors $\mathcal{I}_i$, i.e., $\mathcal{N}_i = \{i, \mathcal{I}_i\}$. Generically in the worst case where all its immediate neighbors are in buffer set, we have $\mathcal{B}_i = \mathcal{I}_i$, and $\mathcal{P}_i = \{i\}$. The fill-in term in (12) affects the submatrix $\mathbf{K}_{\mathcal{I}_i,\mathcal{I}_i}$, leaving only the first row and column untouched. In this case, since i is by definition connected to all elements in $\mathcal{I}_i$, the relaxed edge set R_i defined in (13) includes all possible pairs (see Figure 1(c) for an illustration): $R_i^{\text{1hop}} = \mathcal{N}_i \times \mathcal{N}_i$.

The solution to the relaxed MML problem (15), where no sparsity is imposed, is simply the inverse of the local sample covariance (assuming enough samples for invertibility),

$$\widehat{\mathbf{K}}^{i,\text{1hop}} = \left(\widehat{\mathbf{\Sigma}}_{\mathcal{N}_i,\mathcal{N}_i}\right)^{-1}. \tag{17}$$

The global estimate is obtained by combining the local one-hop estimates as in (16).

In the one-hop case, the proposed relaxed MML estimator reduces to the *LOC* estimator proposed in Wiesel & Hero (2012).

3.5 Case II: Two-hop Estimator

We now consider the two-hop version of the estimator, where $\mathcal{N}_i$ now includes nodes that are reachable from node i within two routing hops. In this setting, the worst-case protected set is given by $\mathcal{P}_i = \{i, \mathcal{I}_i\}$ and the buffer set $\mathcal{B}_i = \mathcal{N}_i \backslash \mathcal{P}_i$ consists of all nodes that are exactly two hops away from the ith node. Hence $\mathcal{B}_i$ can be thought of as the set of *second-hop* nodes. In the two-hop case, the protected edge set E^{Prot} includes not only edges between node i and its immediate first-hop neighbors, but also edges between first-hop neighbors and between first- and second-hop neighbors (see Figure 1(c) for an illustration).

Unlike in the one-hop case, the two-hop problem does not admit a closed-form solution in general. A global estimate is obtained as before by combining row parameter estimates (16).

As the problem dimension p increases, if the numbers of non-zero variables in the relaxed local problems, $|R_i|$, are bounded by a constant or a lower-order term compared to $|\widetilde{E}|$, the number of non-zero variables in the global problem, then the cost of solving the two-hop relaxed MML problem is only marginally higher than the cost of the one-hop problem, while both grow more slowly or not at all compared to the cost of the global MLE. In this large-scale setting, as will be seen in Section 4, the two-hop estimator achieves significantly improved performance relative to the one-hop estimator with a modest increase in cost.

3.6 Local Averaging

In practice with finite samples, the local estimates from the relaxed MML problems are not perfectly consistent with each other. For example, $\widehat{\mathbf{J}}_{i,j}^{\text{Relax}}$, which comes from node i's local estimate, may not agree with $\widehat{\mathbf{J}}_{j,i}^{\text{Relax}}$, which comes from node j's local estimate. Therefore the resulting global estimate $\widehat{\mathbf{J}}^{\text{Relax}}$ in (16) is not guaranteed to be symmetric.

A common way of addressing these discrepancies is to use iterative consensus methods as in Liu & Ihler (2012); Wiesel & Hero (2012). In this work however, we find that a single round of naive local averaging along edges is sufficient to yield a good approximation to the global MLE. Specifically, the local average is given by

$$\widehat{\mathbf{J}}_{i,j}^{\text{Relax}} \leftarrow \frac{1}{2}(\widehat{\mathbf{J}}_{i,j}^{\text{Relax}} + \widehat{\mathbf{J}}_{j,i}^{\text{Relax}}), \quad (i,j) \in E, \tag{18}$$

an inexpensive operation in terms of computation and communication. In the one-hop case, the resulting symmetric estimator coincides with the AVE estimator proposed in Wiesel & Hero (2012).

4 Experiments

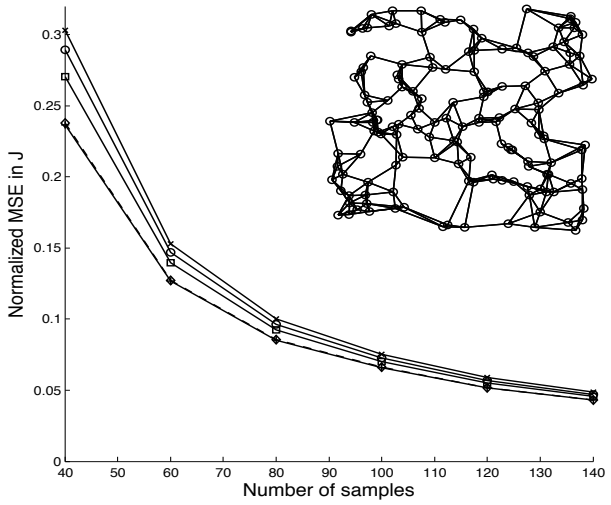
In this section, we evaluate and compare the two-hop version of the relaxed MML estimator with local averaging (denoted as `RelaxedMML 2 hops`) with the following estimators:

- The global ML estimator, denoted as `MLE` in the legend;
- The *LOCAL* and *AVE* estimators from Wiesel & Hero (2012), denoted as `LOC` and `AVE`. They coincide with the asymmetric and symmetric versions respectively of the *one-hop* relaxed MML estimator;
- The weighted maximum pseudo-likelihood estimator using Alternating Direction Method of Multipliers (ADMM) consensus, proposed in Wiesel & Hero (2012) and Liu & Ihler (2012) and denoted as `PML-ADMM`. We use the weights $\left[\hat{\mathbf{J}}_{i,i}^{LOC}\right]^2$, which are found to be successful in Wiesel & Hero (2012).

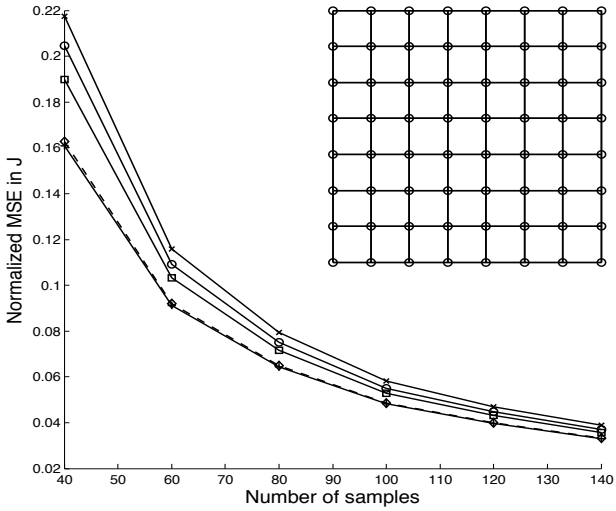
We evaluate these estimators on both synthetic data and a real-world sensor network data set.

4.1 Synthetic data

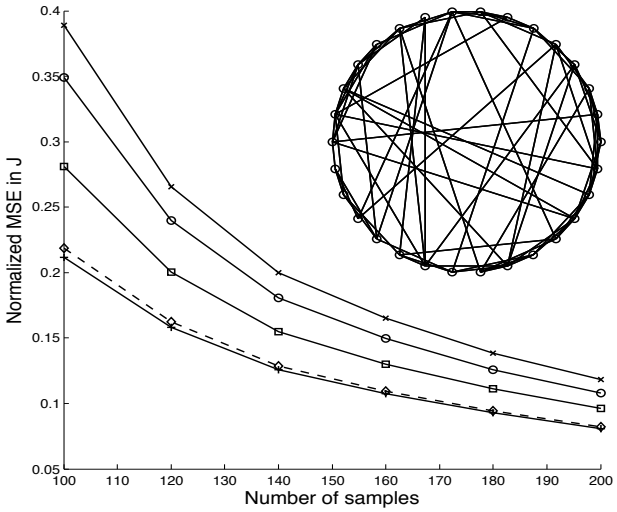
In our experiments, we consider three classes of graphs that are motivated by real-world applications. For



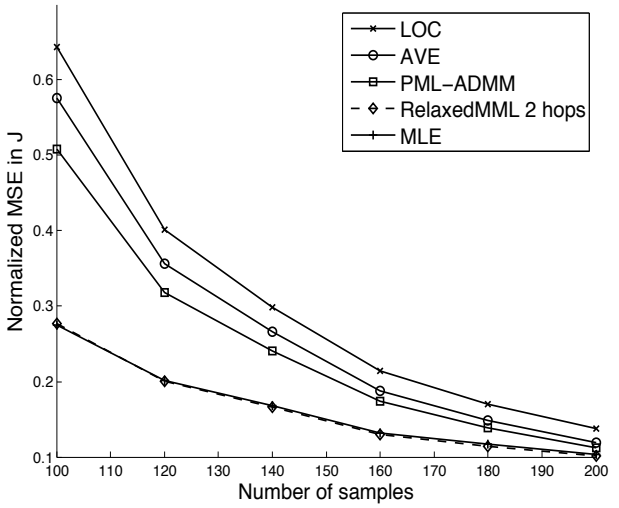
(a) Normalized MSE for K-NN graphs ($p = 500, K = 4$)



(b) Normalized MSE for lattice graphs ($p = 20 \times 20 = 400$, $\mu = 0.5$, $\sigma^2 = 0.2$)



(c) Normalized MSE for small-world graphs ($p = 100$, $K = 20$, $\beta = 0.5$)



(d) Normalized MSE for IntelLab Sensor Network Data set (Guestrin et al., 2004) ($p = 50$)

Figure 2: Normalized MSE in the concentration matrix estimates for different networks. The legend in Figure 2(d) applies to all plots.

each class we follow similar experiment settings as in Wiesel & Hero (2012). Specifically, we randomly generate 20 topologies and concentration matrices $\mathbf{J}$, and for each $\mathbf{J}$, we perform 10 experiments in which random samples are drawn from the distribution and the concentration matrix is estimated from the samples. The normalized mean squared errors (MSE), defined as $\frac{\|\widehat{\mathbf{J}}-\mathbf{J}\|_F^2}{\|\mathbf{J}\|_F^2}$, and averaged over all 200 experiments, are reported in Figure 2. An illustration of the graph topology is shown in the top-right corner of each plot. In Figure 2(a) and 2(b), the nodes are positioned according to the physical layout of the network, while in Figure 2(c), we show a common rendering for a small-world graph. The classes of graphs we consider are:

- **K-NN graphs** (Figure 2(a)): A K-nearest neighbor graph is a straightforward model for real-world networks whose measurements have correlations that depend on pairwise Euclidean distances, *e.g.*, sensor networks. For these experiments, we randomly generate $p = 500$ nodes uniformly over the unit square. Each node is then connected to its K-nearest neighbors, where $K = 4$. The concentration matrix is initialized as $\mathbf{J}_{i,j} = \pm \exp(-0.5 \cdot d_{i,j})$ with random sign, where $d_{i,j}$ is the distance between the ith and jth nodes. Finally we add a small value to the di-

agonal to ensure positive definiteness.

- **Lattice graphs** (Figure 2(b)): A lattice graph is appropriate for networks with regular spatial correlations. We generate a square lattice graph with $p = 20 \times 20 = 400$ nodes and edge weights generated as $\mathbf{J}_{i,j} = \min\{w, 1\}$, where w is a normally distributed random variable with mean 0.5 and variance 0.2. A small value is added to the diagonal to ensure positive definiteness.
- **Small-world graphs** (Figure 2(c)): A small-world graph is a useful model for social networks, where most nodes have few immediate neighbors but can be reached from any other node through a small number of hops. We use the Watts-Strogatz model (Watts & Strogatz, 1998) to generate random small-world networks, with $p = 100$, $K(\text{mean degree}) = 20$, and parameter $\beta = 0.5$. Under this particular setting, a large fraction of nodes have large second-hop neighborhoods with dimension close to p. In general we expect the second-hop neighborhood to scale linearly with respect to p. We choose the edge weights to be uniformly distributed and also add a small diagonal loading.

From the experiment results in Figure 2, we can make the following comments. For the graphs that have relatively small two-hop neighborhoods, such as the K-NN graphs and the lattice grids, the proposed two-hop relaxed MML estimator almost coincides with the global MLE and outperforms other distributed estimators significantly. On the other hand, for graphs such as the small-world networks, the dimensions of the two-hop neighborhoods grow as fast as p. In this case, a noticeable gap emerges between the global MLE and the two-hop relaxed MML estimator. However, these graphs are known to be harder to learn through distributed algorithms (Liu & Ihler, 2011) and therefore the competing estimators' performance also degrades. The two-hop relaxed MML estimator still outperforms by a large margin.

Computational complexity. In the relaxed MML approach, each local problem has the same structure as the centralized ML problem for which there are many efficient algorithms. Furthermore, the local problems can all be solved in parallel before the final one-step averaging. In our experiments, we used the graphical LASSO with known structure (Friedman et al., 2009) for solving both the centralized ML problem and the local problems in the distributed algorithm. For moderate p (number of variables) as in the examples presented here, the total run time of the distributed algorithm with no parallelization is comparable to the centralized one. As p grows however, we would expect the complexity of the distributed algorithm to scale at most linearly with p (assuming that the local neighborhood dimension scales more slowly, such as with K-NN graphs and lattice). The growth is even slower if the algorithm can be parallelized. For the centralized algorithm, the dependence of complexity on p is expected to be at least linear (the growth is much faster when generic solvers are used), and centralized storage/communication is required.

4.2 Real-world data

Finally we apply the estimators to a real-world data set to evaluate their performance and robustness. The *Lab* dataset (Guestrin et al., 2004) contains temperature information from a sensor network of 54 nodes deployed in the Intel Berkeley Research lab between February 28 and April 5, 2004. This dataset is known to be very difficult with many missing data, noise and failed sensors. We select 50 sensors with relatively stable and regular measurements. To obtain a target concentration matrix, we use 1800 consecutive samples per sensor, interpolate the missing or failed readings and de-trend the data using a local rectangular window of 10 samples. Next, we compute the sample covariance and invert it to obtain a sample concentration matrix. This concentration matrix is then thresholded to yield a ground truth graphical model with a sparsity level of 70% zeros. Using knowledge of the sparsity and sampling from the original 1800 samples, we estimate the concentration matrix using the same estimators as before. As shown in Figure 2(d), the proposed two-hop relaxed MML estimator still gives a very tight approximation to the global MLE and its advantage over other distributed estimators is obvious.

5 Conclusion

We have proposed a distributed MML framework for estimating the concentration matrix of Gaussian graphical models. The proposed method independently solves convex relaxations of marginal likelihood maximization problems in local neighborhoods. A global estimate is then obtained by combining the local estimates via a single round of local averaging. The proposed estimator is shown to be asymptotically consistent and computationally efficient. Its improved performance relative to existing distributed estimators is illustrated through extensive experiments.

Acknowledgement

This research was supported in part by ARO grant W911NF-11-1-0391 and Israel Science Foundation Grant No. 786/11. The authors would like to thank the anonymous reviewers for their valuable comments.

References

Banerjee, O., El Ghaoui, L., d'Aspremont, A., and Natsoulis, G. Convex optimization techniques for fitting sparse Gaussian graphical models. In *ACM International Conference Proceeding Series*, volume 148, pp. 89–96. Citeseer, 2006.

Chandrasekaran, V., Parrilo, P.A., and Willsky, A.S. Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40(4), 2012.

Dahl, J., Vandenberghe, L., and Roychowdhury, V. Covariance selection for non-chordal graphs via chordal embedding. *Optimization Methods and Software*, 23(4):501–520, 2008.

Friedman, J., Hastie, T., and Tibshirani, R. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag New York, 2009.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Guestrin, C., Bodik, P., Thibaux, R., Paskin, M., and Madden, S. Distributed regression: an efficient framework for modeling sensor network data. In *Information Processing in Sensor Networks, 2004. IPSN 2004. Third International Symposium on*, pp. 1–10. IEEE, 2004.

Johnson, C.C., Jalali, A., and Ravikumar, P. High-dimensional sparse inverse covariance estimation using greedy methods. *arXiv preprint arXiv:1112.6411*, 2011.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Lauritzen, S.L. *Graphical models*, volume 17. Oxford University Press, USA, 1996.

Liu, Q. and Ihler, A. Learning scale free networks by reweighted ℓ_1 regularization. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Liu, Qiang and Ihler, Alexander. Distributed parameter estimation via pseudo-likelihood. *International Conference on Machine Learning (ICML)*, June 2012.

Meng, Z., Wiesel, A., and Hero, A.O. Distributed principal component analysis on networks via directed graphical models. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 2877–2880. IEEE, 2012.

Ravikumar, P., Wainwright, M.J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Rothman, A.J., Bickel, P.J., Levina, E., and Zhu, J. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Van der Vaart, A.W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wainwright, M.J. and Jordan, M.I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1 (1-2):1–305, 2008.

Wainwright, M.J., Jaakkola, T.S., and Willsky, A.S. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In *Workshop on Artificial Intelligence and Statistics*, volume 21, pp. 97. Society for Artificial Intelligence and Statistics Np, 2003.

Wainwright, M.J., Jaakkola, T.S., and Willsky, A.S. A new class of upper bounds on the log partition function. *Information Theory, IEEE Transactions on*, 51(7):2313–2335, 2005.

Watts, D. and Strogatz, S. The small world problem. *Collective Dynamics of Small-World Networks*, 393: 440–442, 1998.

Wiesel, A. and Hero, A.O. Distributed covariance estimation in Gaussian graphical models. *Signal Processing, IEEE Transactions on*, 60(1):211–220, 2012.

Wiesel, A., Eldar, Y.C., and Hero, A.O. Covariance estimation in decomposable Gaussian graphical models. *Signal Processing, IEEE Transactions on*, 58 (3):1482–1492, 2010.