Marginal Likelihoods for Distributed Estimation of Graphical Model Parameters

(Invited Paper)

Zhaoshi Meng, Dennis Wei, Alfred O. Hero III
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, Michigan 48109-2122

Ami Wiesel
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem 91904, Israel

***Abstract*—This paper considers the estimation of graphical model parameters with distributed data collection and computation. We first discuss the use and limitations of well-known distributed methods for marginal inference in the context of parameter estimation. We then describe an alternative framework for distributed parameter estimation based on maximizing marginal likelihoods. Each node independently estimates local parameters through solving a low-dimensional convex optimization with data collected from its local neighborhood. The local estimates are then combined into a global estimate without iterative message-passing. We provide an asymptotic analysis of the proposed estimator, deriving in particular its rate of convergence. Numerical experiments validate the rate of convergence and demonstrate performance equivalent to the centralized maximum likelihood estimator.**

I. Introduction

Graphical models play a prominent role in distributed statistical inference. Their parsimonious structure allows for efficient and distributed inference of marginal distributions using well-known and well-studied message-passing algorithms such as belief propagation [1]. Less well-studied however in the distributed context is the equally important task of estimating the parameters of a graphical model from data. The goal of this work is to develop similarly distributed methods for model parameter estimation.

This paper focuses on Gaussian graphical models (GGM) with known graph structure, i.e, the pattern of edges is known. Our approach can also be extended to general graphical models, including discrete distributions. In the Gaussian case, parameter estimation essentially reduces to (inverse) covariance estimation, and knowledge of the edge pattern imposes sparsity constraints on the inverse covariance matrix, also known as the concentration or precision matrix. While the resulting maximum likelihood (ML) parameter estimation problem is a convex optimization, centralized algorithms as in [2] become impractical in large networks where data collection and computational resources are decentralized and communication is also constrained.

A natural approach toward distributed parameter estimation is to leverage the methods for distributed marginal inference mentioned above, such as (loopy) belief propagation and its extensions. The idea is to replace the objective function and gradient in the ML estimation problem with approximations that can be computed through iterative message-passing. In Section II, we elaborate on the use of distributed marginal inference techniques for parameter estimation and discuss their limitations. In particular, loopy belief propagation (LBP) may fail to converge or give good marginal estimates in many cases, and when it does converge, the resulting parameter estimates may be biased because of the required approximations.

In Section III, we describe an alternative framework for distributed estimation of GGM parameters based on *marginal* likelihoods, first introduced in [3]. This framework generalizes some previous work based on pseudo-likelihoods [4], [5]. A marginal likelihood is associated with each node and its surrounding neighborhood. Local parameter estimates are obtained by independently maximizing convex relaxations of these marginal likelihoods given local data from the neighborhoods. The local estimates are then combined in a non-iterative fashion to produce the global parameter estimate. We characterize the asymptotic behavior of the proposed estimator, in particular its rate of convergence to the true parameter in terms of mean squared error (MSE). Numerical experiments in Section IV confirm the rate of convergence and demonstrate that a version of the proposed estimator with two-hop local neighborhoods can match the performance of the much more expensive centralized ML estimator. With respect to [3], the main contributions of the current paper are the discussion of the inadequacy of distributed marginal inference for parameter estimation in Section II, the derivation of the asymptotic rate of convergence in Section III-C and its numerical validation in Section IV.

II. Background

We begin by providing background on graphical models and their statistical inference. We refer the reader to [1] for a detailed treatment.

A. Graphical models

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.

If the vector $\mathbf{x}$ follows a multivariate Gaussian distribution, the corresponding model is called a Gaussian graphical model (GGM). 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:

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

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

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

This research was supported in part by ARO grant W911NF-11-1-0391 and Israel Science Foundation Grant No. 786/11.

B. Margianl Inference for GGMs

Given a joint distribution as in (1), a fundamental task is to infer the marginal distribution of a subset of variables, which involves marginalization of the remaining variables, possibly conditioned on observations of certain variables. In the discrete case, the computational cost of exact inference in generally structured graphical models grows exponentially in the graph *treewidth*. Therefore, exact inference is only considered tractable for graphs with small numbers of nodes or with special structures, such as trees and "thin" low-treewidth graphs. In the Gaussian case, marginal inference amounts to estimating the mean parameters, i.e. the covariance matrix $\boldsymbol{\Sigma} = \mathbf{J}^{-1}$. The cost of this global matrix inversion is cubic in the number of variables in general graphs.

Due to the expensive cost of centralized marginal inference, distributed message-passing algorithms, such as loopy belief propagation (LBP), are of particular interest. It can be shown that LBP can be seen as an iterative fixed point algorithm for finding stationary points of the so-called Bethe free energy. For Gaussian models, many *sufficient* conditions exist for Gaussian LBP to converge, such as diagonal dominance, walk-summablility, pairwise normalizablility, etc. [6]. However, when these sufficient conditions do not hold, the Bethe free energy can be proven to be unbounded from below in many settings [7], which leads to divergent Gaussian LBP, or convergence to degenerate, unnormalized marginal distributions. A recent work [8] proposes to use the method of multipliers to improve the convergence behavior of Gaussian LBP when the free energy is well-behaved. However, the unboundedness of the Bethe free energy in continuous models remains a difficult problem for inference. As we will discuss in Section II-D, this difficulty also prevents the direct application of many well-studied message-passing algorithms for the task of parameter estimation.

C. Maximum Likelihood Parameter Estimation for GGMs

A different and equally important task is to learn the parameters of a graphical model from sample data. For Gaussian graphical models this 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, $\widetilde{E} := E \cup \{(i,i)\}_{i=1}^{p}$.

When all the data are accessible, the centralized global maximum likelihood (GML) estimation problem can be formulated as [1]

$$\begin{aligned} \widehat{\mathbf{J}}^{\text{GML}} &= \underset{\mathbf{J}\succeq\mathbf{0}}{\arg\min}\ \langle \widehat{\boldsymbol{\Sigma}}, \mathbf{J}\rangle - \log\det \mathbf{J} \\ \text{s.t.}\ \ & \mathbf{J}_{j,k} = 0 \quad \forall\, (j,k) \notin \widetilde{E}. \end{aligned} \tag{3}$$

where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}(t)\mathbf{x}(t)^T$ is the sample covariance matrix and $\mathbf{x}(1), \ldots, \mathbf{x}(T)$ are i.i.d. samples of $\mathbf{x}$. The GML problem (3) is a convex semidefinite program (SDP) with respect to $\mathbf{J}_{\widetilde{E}}$ and various gradient-based algorithms can be applied to solve this problem, many of which have specialized implementations on graphs, *e.g.* iterative proportional fitting (IPF) [1], chordally-embedded Newton's method [2], etc. However, as we will discuss in more details in the following section, the main drawback of these methods is the computational and communication complexity when implemented in a distributed network setting.

D. Diffculty of Distributed Estimation via LBP

The standard gradient descent algorithm for solving problem (3) has the following update rule at each iteration:

$$\begin{aligned} \widehat{\mathbf{J}}^{(t+1)}_{\widetilde{E}} &\leftarrow \widehat{\mathbf{J}}^{(t)}_{\widetilde{E}} + \gamma \cdot \nabla \ell(\widehat{\mathbf{J}}^{(t)})_{\widetilde{E}} \\ &= \widehat{\mathbf{J}}^{(t)}_{\widetilde{E}} + \gamma \cdot (\widehat{\boldsymbol{\Sigma}}_{\widetilde{E}} - (\widehat{\mathbf{J}}^{(t)})^{-1}_{\widetilde{E}}), \end{aligned} \tag{4}$$

where γ is the step-size. The obvious difficulty is the global matrix inversion involved in computing the gradient at each step, which is intractable in a general distributed network setting.

To obtain a distributed method for estimating GGM parameters, it is natural to consider distributed marginal inference techniques, such as LBP, for approximating the gradient in (4). Essentially this approach optimizes the Bethe surrogate likelihood function [1]. Unfortunately, such parameter learning based on approximate marginal inference does not guarantee a convergent algorithm, and more importantly, it in general yields a *biased* parameter estimator. As mentioned in Section II-B, when the GGM does not satisfy the conditions that ensure convergent and stable LBP inference, the computation of the gradient can be infeasible or result in incorrect values. Even if the overall estimation algorithm is convergent, [9] shows that many MRF models are in principle not learnable through LBP inference, which implies that the estimator is inevitably biased. Similar results also hold when using many other approximate inference techniques, for example, tree-reweighted BP [10].

III. Distributed Estimation for GGMs based on Marginal Likelihoods

Given the difficulties of plugging-in well-established distributed marginal inference techniques for the task of parameter estimation, we describe a novel approach to address this problem motivated by many network applications. We assume that the topology of the graph $\mathcal{G}$, which encodes statistical dependences, coincides 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 the parameter (e.g. precision matrix $\mathbf{J}$) is then formed by combining these local estimates.

A. Marginal Likelihood Maximization

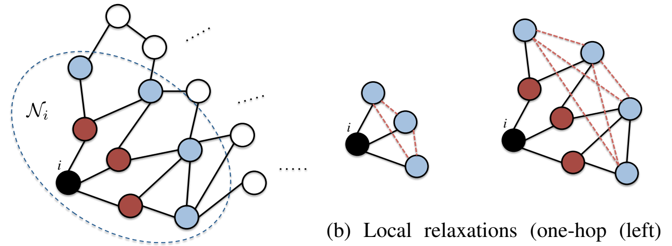
We consider estimating local parameters by maximizing *marginal likelihood* functions in neighborhoods around each node i. Define the index set for *immediate neighbors* of node i as $\mathcal{I}_i := \{j \mid (i,j) \in E\}$, and consider a neighborhood indexed by a set $\mathcal{N}_i$ containing at 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 $\mathbf{S}^i := \widehat{\boldsymbol{\Sigma}}_{\mathcal{N}_i,\mathcal{N}_i} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{\mathcal{N}_i}(t)\mathbf{x}_{\mathcal{N}_i}(t)^T$ be the marginal sample covariance matrix.

The maximum marginal likelihood (MML) estimation problem in neighborhood $\mathcal{N}_i$ can be formulated as:

$$\begin{aligned} \widehat{\mathbf{K}}^{i,\text{MML}} &= \underset{\mathbf{K},\mathbf{J}\succeq\mathbf{0}}{\arg\min}\ \langle \mathbf{S}^i, \mathbf{K}\rangle - \log\det \mathbf{K} \\ \text{s.t.}\ \ & \mathbf{K} = \left[\left(\mathbf{J}^{-1}\right)_{\mathcal{N}_i,\mathcal{N}_i}\right]^{-1}, \\ & \mathbf{J}_{j,k} = 0 \quad \forall\, (j,k) \notin \widetilde{E}, \end{aligned} \tag{5}$$

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

The difficulty with the MML approach is that problem (5) is in general a non-convex optimization with respect to $\mathbf{K}$ and $\mathbf{J}$. The non-convexity arises from the coupling of the nonlinear marginalization constraint linking $\mathbf{K}$ to $\mathbf{J}$ and the sparsity constraints on $\mathbf{J}$. As a surrogate, we derive and consider a convex relaxation of the MML estimation problem.



(a) A general graph and two-hop neighborhood

(b) Local relaxations (one-hop (left) and two-hop (right))

Fig. 1. Illustration of defined sets and local relaxation. In (a) we indicate the two-hop neighborhood $\mathcal{N}$ for node i with a dashed contour. 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. We illustrate the one-hop and two-hop local relaxations in (b). The dashed red lines in (b) denote the fill-in edges due to relaxation. These illustrations also appear in [3].

B. Convex Relaxation of MML

We apply the Schur complement identity to the marginalization constraint in (5), yielding

$$\mathbf{K} = \mathbf{J}_{\mathcal{N},\mathcal{N}} - \mathbf{J}_{\mathcal{N},\mathcal{N}^C} \cdot \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \cdot \mathbf{J}_{\mathcal{N}^C,\mathcal{N}}, \tag{6}$$

where $\mathcal{N}^C$ is the complementary set to $\mathcal{N}$, and we have dropped the subscript i to simplify notation. 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 $\mathcal{N}^C$,

$$\mathcal{B} := \{j \mid j \in \mathcal{N} \text{ and } \mathcal{I}_j \cap \mathcal{N}^C \neq \emptyset\}. \tag{7}$$

The difference set between $\mathcal{N}$ and $\mathcal{B}$ is referred to as the *protected set* $\mathcal{P}$. The buffer and protected sets are illustrated in Figure 1a. 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 (6):

$$\mathbf{K}_{\mathcal{P},\mathcal{P}} = \mathbf{J}_{\mathcal{P},\mathcal{P}}, \;\; \mathbf{K}_{\mathcal{P},\mathcal{B}} = \mathbf{J}_{\mathcal{P},\mathcal{B}}, \tag{8}$$

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

An important observation from (8) is that, in the rows and columns indexed by the protected set $\mathcal{P}$, the sparsity pattern of $\mathbf{J}_{\mathcal{N},\mathcal{N}}$ is entirely preserved and the local parameters are equal to the global ones. 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 (9).

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 (9) 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 (9), resulting in a *relaxed edge set* R (see Figure 1b for illustrations):

$$R = E^{\text{Prot}} \cup \{\mathcal{B}\times\mathcal{B}\}. \tag{10}$$

In light of (8) and (9), any feasible marginal concentration matrix $\mathbf{K}$ for the MML estimation problem (5) is guaranteed to be supported only on the set R. Therefore we can relax the feasible set and formulate the following relaxation of the original MML estimation problem (5) at each node i:

$$\begin{aligned} \widehat{\mathbf{K}}^{i,\text{Relax}} &= \underset{\mathbf{K}\succeq\mathbf{0}}{\arg\min}\ \langle \mathbf{S}^i, \mathbf{K}\rangle - \log\det\mathbf{K} \\ \text{s.t.}\ \ & \mathbf{K}_{j,k} = 0 \quad \forall\,(j,k)\notin R. \end{aligned} \tag{11}$$

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 (3) but with much lower dimensions in general.

After solving the relaxed MML estimation problems as surrogates to estimate local parameters, a global estimate of the concentration matrix can then be constructed by extracting a subset of parameters from each local estimate and concatenating them. Specifically, we extract the local parameter estimates indexed by $L_i := \{(j,k) \in \widetilde{E} \mid j = i\}$, *i.e.*, the non-zero entries in the ith row of $\mathbf{J}$. We refer to the parameters indexed by L_i as the *row parameters* for node i. From (8), the marginal and global concentration matrices are guaranteed to share the same parameters in L_i. Therefore our final global estimate of $\mathbf{J}$ is formed by fusing local estimates from solving each local problem (11):

$$\widehat{\mathbf{J}}^{\text{Relax}}_{L_i} = \widehat{\mathbf{K}}^{i,\text{Relax}}_{L_i}, \quad \text{for } i = 1,\ldots,p. \tag{12}$$

Note that this construction of the global estimate $\widehat{\mathbf{J}}^{\text{Relax}}$ does not guarantee symmetry. However, symmetrization can be done through simple local averaging along each edge. Experiments show that this single-step averaging is sufficient to achieve good performance in most situations.

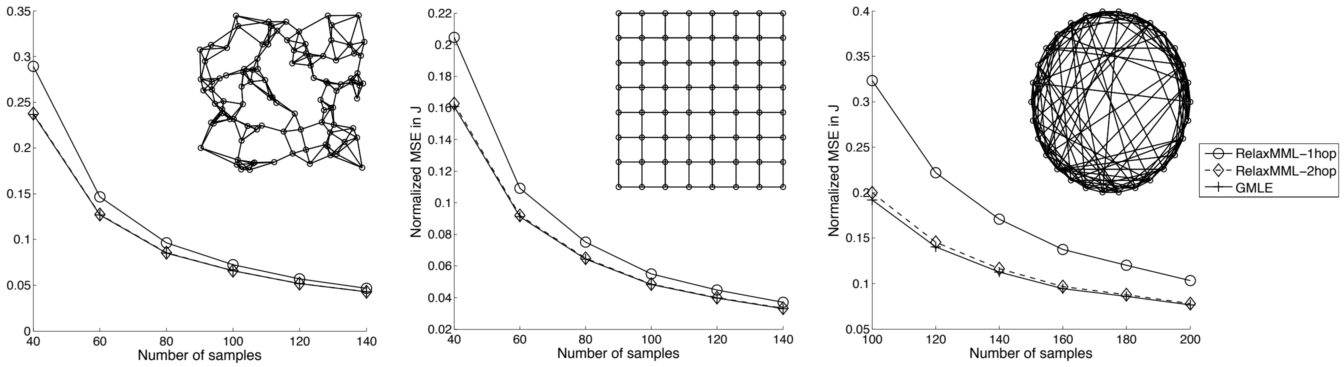
C. Asymptotic Analysis

The following theorem states the asymptotic behavior of the proposed relaxed MML estimator (12).

Theorem 1. *The relaxed MML estimator $\widehat{\mathbf{J}}^{Relax}$ is asymptotically consistent and normal, with an asymptotic variance (i.e. mean squared Frobenius error) of $\frac{1}{T}\cdot\sum_{i=1}^{p}\sum_{j\in L_i}[diag\left(\mathbf{F}_i^{-1}\right)]_j$, where T is the number of samples, $diag(\cdot)$ denotes the diagonal of a matrix, and $\mathbf{F}_i$ is the Fisher information matrix of the relaxed MML problem in the ith neighborhood* (11)*, which takes the following form [5]:*

$$(\mathbf{F}_i)_{(m,n),(l,k)} = \begin{cases} 2\cdot\mathbf{\Sigma}^2_{m,l}, & m = n \text{ and } l = k \\ 2\cdot\mathbf{\Sigma}_{m,k}\cdot\mathbf{\Sigma}_{l,n}, & m = n, l \neq k \text{ or } m \neq n, l = k \\ \mathbf{\Sigma}_{m,k}\cdot\mathbf{\Sigma}_{n,l}, & o.w.. \end{cases} \tag{13}$$

Proof: (abbreviated) Consider the following set of sparse positive semidefinite matrices with respect to a non-zero element set R: $\mathcal{K}^R := \{\mathbf{K} \mid \mathbf{K} \succeq 0, \mathbf{K}_{(j,k)} = 0, \forall (j,k) \notin R\}$. We first note that, when R is taken to be the relaxed edge set of a neighborhood as defined in (10), then the true marginal concentration matrix corresponding to the neighborhood, $\mathbf{K}^* = (\mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}})^{-1}$, must belong to the set $\mathcal{K}^R$. This can be seen from the fact that the true global concentration matrix $\mathbf{J}^*$ conforms to the sparsity pattern specified by $\widetilde{E}$ and from relations (8) and (9). Therefore the proposed relaxed MML problem (11) is equivalent to a standard ML problem with respect to a GGM distribution parameterized by matrix $\mathbf{K} \in \mathcal{K}^R$, with $\mathbf{K}^*$ being the population parameter. Then the asymptotic consistency, normality and efficiency of the proposed relaxed MML estimator (with respect to the local problem) all follow from the standard asymptotic analysis of the ML estimator [11]. In particular, the variances of the errors achieve their Cramer-Rao bounds, which are the diagonal elements of the inverse Fisher information matrix $\mathbf{F}$ defined in (13). Finally by extracting and summing the variances corresponding to the row parameters, we obtain the expression for the asymptotic mean squared Frobenius error of the proposed global estimator $\widehat{\mathbf{J}}^{\text{Relax}}$, as stated in the theorem. ■



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

(b) Normalized MSE for lattice graphs ($p = 20 \times 20 = 400$)

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

Fig. 2. Numerical experiments. The true concentration matrix is initialized as: (a) $\mathbf{J}_{i,j} = \pm\exp(-0.5 \cdot d_{i,j})$, $d_{i,j}$ is the Euclidean distance between two uniformly distributed points in the unit square; (b) $\mathbf{J}_{i.j} = \min\{w, 1\}$, $w \sim \mathcal{N}(0.5, 0.2)$; (c) Watts-Strogatz model with K(mean degree) $= 20$, and parameter $\beta = 0.5$. Diagonal loading is added for all models to ensure positive definiteness. The legend in (c) applies to all plots. These plots also appear in [3].

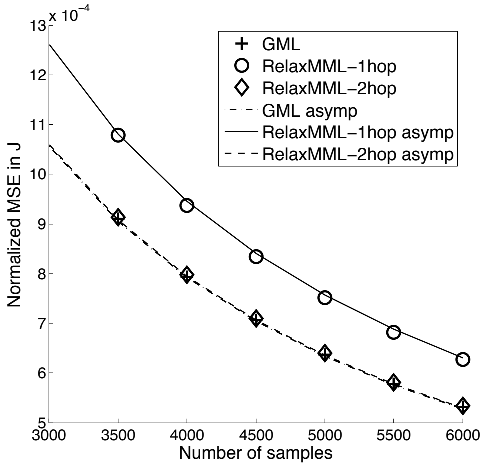


Fig. 3. Asymptotic normalized MSE for K-NN graphs ($p = 20, K = 4$, $\mathbf{J}_{i,j}$ chosen as in Fig. 2a). The curves denote the theoretical asymptotic limits, whereas the symbols denote the empirical normalized MSE over 10,000 runs.

IV. Experiments

In this section, we provide numerical results to demonstrate the performance and verify the asymptotic analysis of the proposed relaxed MML estimator. We define the local neighborhoods ($\mathcal{N}_i$) based on a fixed number of communication hops in the network from the centering node i. We consider one-hop and two-hop neighborhoods, and compare them with the centralized global ML estimator (GML).

We first verify the asymptotic analysis of the MSE performance of the proposed estimator (see Fig. 3) using 10,000 randomized runs sampled from a four-nearest-neighbor graphical model with $p = 20$ nodes. The empirical normalized MSEs $\frac{\|\hat{\mathbf{J}}-\mathbf{J}\|_F^2}{\|\mathbf{J}\|_F^2}$ are computed and compared with theoretical bounds provided by Theorem 1. The tightness of the bounds is easily seen. The bound for the two-hop relaxed MML estimator approximates the bound for the GML estimator closely, which indicates its statistical efficiency.

Lastly we evaluate and compare the non-asymptotic MSE performance of the proposed estimator and GML estimator on randomly generated K-NN, 2-D grid, and small-world graphs. The results are shown in Fig. 2a-2c. Please refer to the caption for parameter settings. Results for all types of graphs consistently demonstrate the improvement of the two-hop estimator over the one-hop one, and also the closeness between the two-hop estimator and the much more expensive centralized GML estimator.

V. Conclusion

We have presented a distributed MML framework for estimating the concentration matrix of a Gaussian graphical model, avoiding the limitations of distributed marginal inference methods such as belief propagation. We have derived the asymptotic properties of the estimator, in particular its rate of MSE convergence, and demonstrated empirical performance equivalent to the centralized ML estimator.

References

[1] M. Wainwright and M. Jordan, “Graphical models, exponential families, and variational inference,” *Foundations and Trends® in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[2] J. Dahl, L. Vandenberghe, and V. Roychowdhury, “Covariance selection for non-chordal graphs via chordal embedding,” *Optimization Methods and Software*, vol. 23(4), pp. 501–520, 2008.

[3] Z. Meng, D. Wei, A. Wiesel, and A. Hero III, “Distributed learning of Gaussian graphical models via marginal likelihoods,” *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

[4] Q. Liu and A. Ihler, “Distributed parameter estimation via pseudo-likelihood,” *International Conference on Machine Learning (ICML)*, Jun. 2012.

[5] A. Wiesel and A. Hero, “Distributed covariance estimation in Gaussian graphical models,” *Signal Processing, IEEE Transactions on*, vol. 60, no. 1, pp. 211–220, 2012.

[6] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, “Walk-sums and belief propagation in Gaussian graphical models,” *The Journal of Machine Learning Research*, vol. 7, pp. 2031–2064, 2006.

[7] B. Cseke and T. Heskes, “Properties of Bethe free energies and message passing in Gaussian models,” *Journal of Artificial Intelligence Research*, vol. 41, no. 2, pp. 1–24, 2011.

[8] J. Pacheco and E. B. Sudderth, “Minimization of continuous Bethe approximations: A positive variation,” in *Advances in Neural Information Processing Systems*, 2012, pp. 2573–2581.

[9] U. Heinemann and A. Globerson, “What cannot be learned with Bethe approximations,” *arXiv preprint arXiv:1202.3731*, 2012.

[10] M. J. Wainwright, “Estimating the wrong graphical model: Benefits in the computation-limited setting,” *The Journal of Machine Learning Research*, vol. 7, pp. 1829–1859, 2006.

[11] A. Van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 2000, vol. 3.