Marginal Likelihoods for Distributed Parameter Estimation of Gaussian Graphical Models

Zhaoshi Meng, *Student Member, IEEE*, Dennis Wei, *Member, IEEE*, Ami Wiesel, *Member, IEEE*, and Alfred O. Hero, III, *Fellow, IEEE*

Abstract—We consider distributed estimation of the inverse covariance matrix, also called the concentration or precision matrix, in Gaussian graphical models. Traditional centralized estimation often requires global inference of the covariance matrix, which can be computationally intensive in large dimensions. Approximate inference based on message-passing algorithms, on the other hand, can lead to unstable and biased estimation in loopy graphical models. Here, we propose a general framework for distributed estimation based on a maximum marginal likelihood (MML) approach. This approach computes local parameter estimates by maximizing marginal likelihoods defined with respect to data collected from local neighborhoods. Due to the non-convexity of the MML problem, we introduce and solve a convex relaxation. The local estimates are then combined into a global estimate without the need for iterative message-passing between neighborhoods. The proposed algorithm is naturally parallelizable and computationally efficient, thereby making it suitable for high-dimensional problems. In the classical regime where the number of variables p is fixed and the number of samples T increases to infinity, the proposed estimator is shown to be asymptotically consistent and to improve monotonically as the local neighborhood size increases. In the high-dimensional scaling regime where both p and T increase to infinity, the convergence rate to the true parameters is derived and is seen to be comparable to centralized maximum-likelihood estimation. Extensive numerical experiments demonstrate the improved performance of the two-hop version of the proposed estimator, which suffices to almost close the gap to the centralized maximum likelihood estimator at a reduced computational cost.

Index Terms—Distributed estimation, Gaussian graphical models, structured covariance.

I. INTRODUCTION

GRAPHICAL models provide a principled framework for compactly characterizing dependencies among many random variables, represented as nodes in a network [3], [4].

Their sparse structure allows for efficient and distributed inference using message-passing algorithms such as loopy belief propagation (LBP), which makes them especially well-suited to large networks, such as sensor, social, and biological networks [5]–[7]. Less well-studied, however, is the equally important task of distributed estimation of the parameters of a graphical model from data. The goal of this work is to develop and analyze distributed methods for model parameter estimation.

In this paper we focus on Gaussian graphical models (GGM) with known graph structure, i.e., the pattern of edges is known. Our approach can also be extended to more general graphical models, including discrete distributions. For GGMs, 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 GGM maximum likelihood (ML) parameter estimation problem is a convex optimization, solving it exactly for generally structured networks using centralized algorithms as in [8]–[10] becomes impractical in large real-world networks where data collection and computational resources are limited.

A natural approach toward distributed parameter estimation is to leverage methods for distributed marginal inference, such as LBP and its extensions. The idea is to replace the objective function and its gradient in the ML estimation problem with approximations that can be computed through iterative message-passing. However, in many cases LBP may fail to converge or give good marginal estimates, and when it does converge, the resulting parameter estimates may be biased [11], [12].

Another direction for distributed estimation is to consider a surrogate objective that decomposes into smaller problems that are locally parameterized. Then a distributed ML algorithm estimates the local parameters by processing local data with limited message passing. Some recent efforts along this direction [5], [6] have considered a pseudo-likelihood framework for exponential family distributions.

This paper proposes a general framework for distributed estimation based on *marginal* likelihoods, as contrasted with pseudo-likelihoods. Each node collects data within its extended neighborhood and independently forms a local estimate by maximizing a marginal likelihood. To deal with the non-convexity of the maximum marginal likelihood (MML) estimation problem, we formulate a convex relaxation of the problem. The resulting distributed estimator is computationally efficient, and involves minimal message passing.

We analyze the mean squared error (MSE) of the proposed distributed estimator in both the classical asymptotic regime

(fixed number of parameters p and increasing number of samples $T \to \infty$), and also the high-dimensional regime where both p and T increase to infinity ($p, T \to \infty$). In the classical regime, the distributed estimator is shown to be asymptotically consistent. Furthermore, the asymptotic error improves monotonically as the local neighborhood size increases. In the high-dimensional regime, we show that under certain conditions and proper scaling between p and T, the proposed estimator achieves a comparable statistical convergence rate to the (more expensive) global ML estimator.

Our analytical results are supported by extensive numerical experiments on both synthetic and real-world data sets. In particular, we show that two-hop local information is sufficient for the proposed distributed estimator to match the performance of the more expensive centralized ML estimator. The proposed estimator also improves significantly upon existing distributed estimators [5], [6]. In terms of computation, the complexity of our estimator increases at most linearly with p in most cases and can be further reduced through parallelization. In the case of a physical network implementation, the near-absence of message passing and long-distance communication is also an advantage.

We emphasize that the problem we consider is different from covariance selection [10], [13]–[15], in which the graph topology is not known *a priori* and must be estimated in addition to the parameters. To test our assumption of known graph structure, we also study the robustness of the proposed estimators against small model (i.e., structure) mismatch. Both theoretical analysis and numerical results show that the proposed distributed estimator is as robust as the centralized ML estimator.

The algorithm and some preliminary experimental results in the current paper were first presented in [1]. During the preparation of this extended version, a related and independent work [16] has come to our attention. The authors of [16] consider a distributed learning algorithm for general Markov random fields which works on local unions of cliques, generalizing nodes and edges in the Gaussian case. Asymptotic consistency is also discussed. After submission of this extended version, an anonymous reviewer made us aware of an earlier arXiv posting [17] by Massam and Wang that considered a version of our algorithm [1] for discrete graphical models. Focusing on the discrete case, they established novel asymptotic theory that parallels the theory developed in the present paper for the Gaussian case.

The outline of the paper is as follows. In Section II, we give a brief review of graphical models, centralized ML parameter estimation, and the difficulty of parameter estimation using traditional marginal inference techniques. In Section III, we propose a general approach to distributed estimation based on marginal likelihoods. In Section IV, we provide extensive analysis of the convergence rates and robustness of the proposed estimator. Section V discusses the computational complexity and implementation advantages of the estimator. Numerical experiments are presented in Section VI and the paper concludes in Section VII.

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., $\langle \mathbf{A}, \mathbf{B} \rangle = \text{trace}(\mathbf{AB}) = \sum_{i,j} \mathbf{A}_{i,j} \mathbf{B}_{i,j}$. We distinguish the following two norms for matrices: the induced $\ell_\infty / \ell_\infty$ norm $\|\mathbf{A}\|_\infty := \max_{i=1,\ldots,p} \sum_{j=1}^{p} |\mathbf{A}_{i,j}|$, and the element-wise ℓ_∞ norm $\|\mathbf{A}\|_\infty := \max_{i,j=1,\ldots,p} |\mathbf{A}_{i,j}|$. $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the maximum and minimum eigenvalues of matrix $\mathbf{A}$, respectively.

II. BACKGROUND

We begin by providing background on graphical models and their statistical inference. We refer the reader to [3], [4] for a detailed treatment.

A. Gaussian 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:

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

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

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

B. Maximum Likelihood Parameter Estimation for GGMs

Estimating the parameters of a graphical model from sample data is the first step for many applications. For Gaussian graphical models this reduces to estimating the non-zero elements of the concentration matrix $\mathbf{J}$ (including the diagonal elements). Defining

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

as the index set for these non-zero elements, the centralized global maximum likelihood (GML) estimation problem can be formulated as [3]:

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

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 (4) is a convex log-determinant-regularized semidefinite program ($\log \det$-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) [4], chordally-embedded Newton's method [9], etc. The standard gradient descent algorithm for solving problem (4) has the following update rule at each iteration:

$$\begin{aligned}
\widehat{\mathbf{J}}_{i,j}^{(t+1)} &\leftarrow \widehat{\mathbf{J}}_{i,j}^{(t)} - \gamma \cdot \nabla \ell(\widehat{\mathbf{J}}^{(t)})_{i,j} \\
&= \begin{cases} \widehat{\mathbf{J}}_{i,j}^{(t)} - \gamma \cdot \left(2\widehat{\boldsymbol{\Sigma}}_{i,j} - 2(\widehat{\mathbf{J}}^{(t)})_{i,j}^{-1}\right), & i \neq j \\ \widehat{\mathbf{J}}_{i,j}^{(t)} - \gamma \cdot \left(\widehat{\boldsymbol{\Sigma}}_{i,j} - (\widehat{\mathbf{J}}^{(t)})_{i,j}^{-1}\right), & i = j \end{cases}
\end{aligned} \quad (5)$$

where $\ell(\mathbf{J})$ is the GML objective function and $\nabla \ell(\mathbf{J})$ denotes its gradient, γ is the step-size, and we have used the facts $\frac{\partial \log \det \mathbf{X}}{\partial \mathbf{X}_{i,j}} = 2(\mathbf{X}^{-1})_{i,j}$ for $i \neq j$ and $\frac{\partial \log \det \mathbf{X}}{\partial \mathbf{X}_{i,i}} = (\mathbf{X}^{-1})_{i,i}$ for symmetric matrices [18]. The obvious difficulty is the global matrix inversion involved in computing the gradient at each step, whose computational cost is cubic in the number of variables for generally structured models.

Given the expense of the matrix inversion in (5), an alternative is to consider distributed message-passing algorithms, such as loopy belief propagation (LBP), an iterative message-passing algorithm for inference of marginal distributions. When applied to tree-structured graphs, LBP yields exact marginals. Unfortunately, this does not hold for loopy graphs in general [19]. For Gaussian models, many *sufficient* conditions exist for Gaussian LBP to converge, such as diagonal dominance, walk-summablility, pairwise normalizability, etc. [11]. However, when these sufficient conditions do not hold, Gaussian LBP can be divergent, or it may converge to degenerate, unnormalized marginal distributions. A recent work [20] uses the method of multipliers to improve the convergence behavior of Gaussian LBP for some less ill-conditioned models. However, even if LBP converges, its final estimate is not guaranteed to be consistent. For discrete graphical models, the authors of [12] show that many models are in principle not learnable through LBP, which implies that an estimator based on LBP inference is inevitably biased for a subset of models. Similar drawbacks also hold when using other approximate inference techniques, for example, tree-reweighted BP [21]. The above difficulties of parameter estimation using traditional marginal inference techniques motivate us to consider a different distributed framework for parameter estimation, as introduced in the next section.

III. DISTRIBUTED ESTIMATION IN GGMS

Our framework avoids the weakness of LBP and other message passing approaches to distributed estimation of GGMs. The proposed distributed algorithm collects all the data samples from within each 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 with a simple, single pass aggregation rule.

A. Marginal Likelihood Maximization

We consider estimating local parameters by maximizing *marginal likelihood* functions in neighborhoods around each node. Define the index set for *immediate neighbors* of node i as

$$\mathcal{I}_i := \{j \mid (i,j) \in E\}, \quad (6)$$

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 $\widehat{\boldsymbol{\Sigma}}_{\mathcal{N}_i, \mathcal{N}_i} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{\mathcal{N}_i}(t)\mathbf{x}_{\mathcal{N}_i}(t)^T$ be the marginal sample covariance matrix. The maximum marginal likelihood (MML) estimation problem in neighborhood $\mathcal{N}_i$ can be formulated as:

$$\begin{aligned}
\widehat{\mathbf{K}}^{i,\mathrm{MML}} = \underset{\mathbf{K}, \mathbf{J}}{\arg\min} \; &\langle \widehat{\boldsymbol{\Sigma}}_{\mathcal{N}_i, \mathcal{N}_i}, \mathbf{K} \rangle - \log \det \mathbf{K} \\
\text{s.t.} \quad \mathbf{K} = &\left[(\mathbf{J}^{-1})_{\mathcal{N}_i, \mathcal{N}_i} \right]^{-1}, \\
\mathbf{J}_{j,k} = &\, 0 \quad \forall \, (j,k) \notin \widetilde{E}, \\
\mathbf{J} \succeq &\, \mathbf{0}, \quad (7)
\end{aligned}$$

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. We index the nodes in the MML problem (7) in the same way as in the GML problem (4). (For example, if $\mathcal{N}_1 = \{1, 3, 6\}$, the rows and columns of $\mathbf{K}$ are indexed by $\{1, 3, 6\}$ and not re-indexed to $\{1, 2, 3\}$.)

The difficulty with direct application of MML is that problem (7) 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 next a convex relaxation of the MML estimation problem.

B. Convex Relaxation of MML

We apply the Schur complement identity to the marginalization constraint in (7), yielding

$$\mathbf{K} = \mathbf{J}_{\mathcal{N}, \mathcal{N}} - \mathbf{J}_{\mathcal{N}, \mathcal{N}^C} \cdot \left[\mathbf{J}_{\mathcal{N}^C, \mathcal{N}^C} \right]^{-1} \cdot \mathbf{J}_{\mathcal{N}^C, \mathcal{N}}, \quad (8)$$

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 $\mathcal{N}^C$,

$$\mathcal{B} := \{j \mid j \in \mathcal{N} \text{ and } \mathcal{I}_j \cap \mathcal{N}^C \neq \emptyset\}. \quad (9)$$

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 Fig. 1(a) and (b). Due to the Markov property, we have
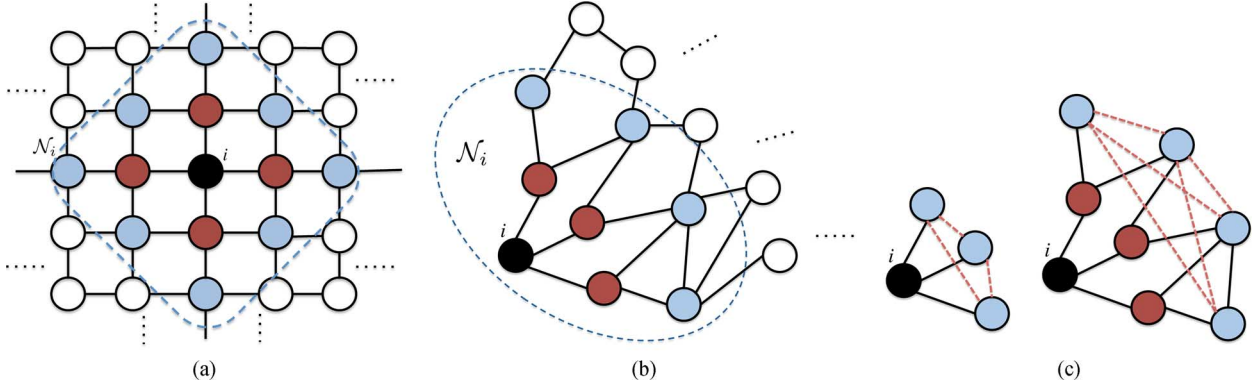
Fig. 1. Illustration of defined sets in the proposed local relaxation of MML. In (a) and (b) we show two different graphs, in which the two-hop neighborhood $\mathcal{N}$ for node i is indicated with dashed contours. The buffer set variables $\mathbf{x}_\mathcal{B}$ and the protected set variables $\mathbf{x}_\mathcal{P}$ (excluding node i itself) are colored blue and red, respectively. For the graph in (b), we illustrate the one-hop and two-hop local relaxations in (c). The dashed red lines in (c) denote the fill-in edges due to relaxation. (a) 2D lattice and two-hop neighborhood $\mathcal{N}_i$. (b) A general graph and two-hop neighborhood $\mathcal{N}_i$. (c) Local relaxations (one-hop (left) and two-hop (right)). Dotted lines denote fill-in edges.

$\mathbf{J}_{\mathcal{P},\mathcal{N}^C} = \mathbf{0}$. Decomposing $\mathcal{N}$ into $\mathcal{B}$ and $\mathcal{P}$ then reveals the sparsity pattern of $\mathbf{K}$ using (8):

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

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

and hence

$$\mathbf{K}_{\mathcal{P},\mathcal{P}} = \mathbf{J}_{\mathcal{P},\mathcal{P}}, \ \ \mathbf{K}_{\mathcal{P},\mathcal{B}} = \mathbf{J}_{\mathcal{P},\mathcal{B}}, \tag{10}$$

$$\mathbf{K}_{\mathcal{B},\mathcal{B}} = \mathbf{J}_{\mathcal{B},\mathcal{B}} - \mathbf{J}_{\mathcal{B},\mathcal{N}^C} \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \mathbf{J}_{\mathcal{N}^C,\mathcal{B}}. \tag{11}$$

An important observation from (10) 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 from $\mathbf{J}_{\mathcal{B},\mathcal{B}}$ due to the fill-in term, *i.e.*, the second term in (11).

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 (11) as

$$E^{\mathrm{Prot}} := \widetilde{E} \cap \left\{\{\mathcal{P} \times \mathcal{P}\} \cup \{\mathcal{P} \times \mathcal{B}\} \cup \{\mathcal{B} \times \mathcal{P}\}\right\}, \tag{12}$$

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 (11), resulting in a *relaxed edge set* R (see Fig. 1(c) for illustrations):

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

In light of (10) and (11), any feasible marginal concentration matrix $\mathbf{K}$ for the MML estimation problem (7) is guaranteed to be supported only on the set R. Therefore we can relax the feasible set and formulate the following relaxation of (7) at each node i, called the relaxed MML (RMML) problem:

$$\widehat{\mathbf{K}}^{i,\mathrm{Relax}} = \arg\min_{\mathbf{K}} \langle \widehat{\mathbf{\Sigma}}_{\mathcal{N}_i,\mathcal{N}_i}, \mathbf{K} \rangle - \log \det \mathbf{K}$$

$$\text{s.t.} \ \ \mathbf{K}_{j,k} = 0 \ \ \forall\, (j,k) \notin R$$

$$\mathbf{K} \succeq \mathbf{0}. \tag{14}$$

The above RMML problem is a convex optimization with respect to $\mathbf{K}$ and has the same form as the global MLE problem (4) but over matrices of much lower dimension.

After solving the RMML 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

$$L_i := \{(j,k) \in \widetilde{E} \mid j = i\}, \tag{15}$$

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 (10), when there are no sampling errors, i.e., $T \to \infty$, the marginal and global concentration matrices are guaranteed to share the same parameters in L_i. Therefore our global estimate of $\mathbf{J}$ is formed by concatenating local solutions of (14):

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

The proposed RMML framework is very general and applies to many possible choices of local neighborhoods, which include, e.g., nearest neighbors, second-order nearest neighbors, or, in general, k-th order nearest neighbors of a node i. In the following subsections, we consider one- and two-hop neighborhoods. The absence of sampling errors is still assumed, i.e., $T \to \infty$.

C. Case I: One-Hop Estimator

We first consider a first-order (i.e., one-hop) neighborhood consisting of node i and its immediate neighbors $\mathcal{I}_i$, i.e., $\mathcal{N}_i = \{i\} \cup \mathcal{I}_i$. Generically in the worst case where the immediate neighbors are all buffer nodes, we have $\mathcal{B}_i = \mathcal{I}_i$, and $\mathcal{P}_i = \{i\}$. The fill-in term in (11) 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 leftmost graph of Fig. 1(c) for an illustration): $R_i^{\mathrm{1hop}} = \mathcal{N}_i \times \mathcal{N}_i$.

The solution to the relaxed MML problem (14) for the first-order neighborhood is simply the inverse of the local sample covariance,

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

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 in [6]. As shown in [6], this estimator is also equivalent to the pseudolikelihood estimator [22] without symmetry constraints, and the covariance selection procedure in [10] when the graph is known.

D. Case II: Two-Hop Estimator

We next consider a second-order neighborhood (two-hop), $\mathcal{N}_i$ that includes nodes that are reachable from node i within two hops. In this setting, the worst-case protected set is given by $\mathcal{P}_i = \{i\} \cup \mathcal{I}_i$ 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 Fig. 1(c) for an illustration).

Unlike in the one-hop case, the two-hop problem (14) does not admit a general closed-form solution. However, as mentioned before, (14) can be solved using efficient algorithms for semidefinite programming. A global estimate is obtained as before by combining row parameter estimates (16).

E. Symmetrization of RMML Estimator

When $\widehat{\mathbf{\Sigma}}$ is estimated from finite sample sizes, 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 [5], [6]. In this work however, we find that a single round of naive local averaging along edges is sufficient to ensure convergence to the true parameters, and also to yield a good approximation to the global MLE. Specifically, the local average is given by

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

which is the only message passing required. This message passing is single pass, unlike LBP which requires several iterations (if it converges at all). In the one-hop case, the resulting symmetric estimator coincides with the AVE estimator proposed in [6].

IV. ANALYSIS

A. Asymptotic Analysis: Classical Fixed-Dimensional Regime

First we analyze the proposed distributed RMML estimator in the classical asymptotic regime, where the number of variables p is fixed while the number of samples T goes to infinity. Let $\mathbf{J}^*$ and $\mathbf{\Sigma}^*$ denote the true precision and covariance matrices, respectively. The following theorem states the asymptotic consistency of the RMML estimator $\widehat{\mathbf{J}}^{\text{Relax}}$ and characterizes its asymptotic mean squared error:

Theorem 1 (Asymptotic MSE): The relaxed MML estimator $\widehat{\mathbf{J}}^{\text{Relax}}$ is asymptotically consistent, and its mean squared (Frobenius) error satisfies

$$T \cdot \mathbb{E}\|\widehat{\mathbf{J}}^{\text{Relax}} - \mathbf{J}^*\|_F^2 \xrightarrow{T \to \infty} \sum_{i=1}^p \sum_{j \in L_i} [\text{diag}\left(\mathbf{F}_i^{-1}\right)]_j, \tag{19}$$

where T is the number of samples, $\text{diag}(\cdot)$ denotes the diagonal of a matrix, and $\mathbf{F}_i$ is the Fisher information matrix of the relaxed MML problem in the ith neighborhood (14), which takes the following form:

$$(\mathbf{F}_i)_{(m,n),(l,k)}$$
$$= \begin{cases} 2\mathbf{\Sigma}_{m,l}^{*}{}^2, & m = n \text{ and } l = k \\ 2\mathbf{\Sigma}_{m,k}^* \mathbf{\Sigma}_{l,n}^*, & m = n, l \neq k \text{ or } m \neq n, l = k \\ \mathbf{\Sigma}_{m,k}^* \mathbf{\Sigma}_{n,l}^*, & \text{otherwise.} \end{cases} \tag{20}$$

The above result can be derived by applying classical asymptotic theory [23] to each local RMML problem (14), which is a well-defined M-estimation problem. Then the asymptotic behavior of the global RMML estimate follows by aggregation. The detailed proof of Theorem 1 is provided in Appendix A.

While Theorem 1 ensures the consistency of RMML estimators with arbitrary local neighborhoods (as long as the row parameters are included), the following theorem guarantees that, in the asymptotic limit, larger neighborhoods always yield reduced estimation variance:

Theorem 2 (Monotonicity of Asymptotic MSE): Let $\widehat{\mathbf{J}}^{\text{Relax},k\text{-hop}}$ be the RMML estimate obtained from k-hop local neighborhoods. When the number of samples $T \to \infty$, for $k = 1, 2, \ldots 1$, we have

$$\mathbb{E}\|\widehat{\mathbf{J}}^{\text{Relax},k\text{-hop}} - \mathbf{J}^*\|_F^2 \geq \mathbb{E}\|\widehat{\mathbf{J}}^{\text{Relax},(k+1)\text{-hop}} - \mathbf{J}^*\|_F^2 \tag{21}$$
$$\geq \mathbb{E}\|\widehat{\mathbf{J}}^{\text{GML}} - \mathbf{J}^*\|_F^2. \tag{22}$$

While Theorem 2 is stated for Gaussian graphical models, it was first proven for the case of discrete graphical models by Massam and Wang in [17]. As pointed out by the authors of [17], their proof can be easily extended to the Gaussian case. For completeness, we include our own proof of Theorem 2 in Appendix B. The two proofs follow parallel lines of argument.

In Section VI, we present numerical results that verify Theorem 2 not only in the large-sample regime but also when the sample size T is comparable to or smaller than p. In particular, it will be seen that the difference between $k = 1$ and $k = 2$ hops is most significant while the difference between $k = 2$ and the GML estimator (and by extension $k > 2$ and GML) is much smaller.

B. Asymptotic Analysis: High-Dimensional Regime

Theorems 1 and 2 characterize the classical asymptotic behavior of the RMML estimator. In this subsection we analyze the high-dimensional convergence rate of the RMML estimator,

which can be applied to settings where both the number of variables p and the number of samples T increase to infinity, i.e., $p, T \to \infty$. Such problems arise in high-dimensional applications, and have attracted much attention in modern statistics [10], [13], [14]. We will show that under very mild conditions, the proposed RMML estimator enjoys a sharp MSE convergence rate to the true parameter, which is almost the same as the more expensive global ML estimator.

Similar to [13], [14], we first assume that the maximum eigenvalue of $\mathbf{J}^*$ is bounded from above:

$$\lambda_{\max}(\mathbf{J}^*) \le \overline{\kappa} < \infty. \tag{23}$$

Recall that R_i defines the relaxed edge set in the ith local neighborhood. Let $\overline{R}$ denote the maximum cardinality among all local relaxed edge sets, i.e.,

$$\overline{R} := \max_{i=1,\ldots,p} |R_i|, \tag{24}$$

and let r denote the sum of the cardinalities of all local relaxed edge sets:

$$r := \sum_{i=1}^{p} |R_i|. \tag{25}$$

Also denote $\overline{\sigma} := \max_{i=1,\ldots,p} \mathbf{\Sigma}^*_{i,i}$ as the maximum variance.

The following theorem states an upper bound on the estimation error rate in the high-dimensional regime.

Theorem 3 (High-Dimensional MSE): Assume the number of samples T satisfies

$$T \ge C^2 c_1 \log p, \tag{26}$$

for $c_1 = 6400 \overline{\sigma}^2 / \min^2 \left\{ \frac{1}{9 \overline{\kappa} \sqrt{\overline{R}}}, 40 \overline{\sigma} \right\}$ and an arbitrary constant $C \ge 1$. Then

$$\|\widehat{\mathbf{J}}^{\mathrm{Relax}} - \mathbf{J}^*\|_F \le 720 C \cdot \overline{\kappa}^2 \overline{\sigma} \sqrt{\frac{r \log p}{T}}, \tag{27}$$

with probability greater than $1 - 4/p^{2(C^2-1)}$.

Proof of Theorem 3 can be found in Appendix C.

Remarks:

1) It is interesting to compare the result in Theorem 3 with the standard convergence rate for the GML estimator (e.g., [13], [14], [24]). Theorem 3 assumes a very mild condition (26) on the sample size, which is less restrictive than the requirement $O(p \log p)$ shown in [14] in the high dimensional regime, and is comparable to those obtained in [13], [24] when the local neighborhood size increases more slowly than p, i.e., $\overline{R} = o(p)$. However, we emphasize that, unlike some of the literature, we assume the graph structure is known.

2) The error bound in Theorem 3 is (up to a constant) slightly more pessimistic than the rate $O(\sqrt{p \log p / T})$ shown in [13], [14] by the additional factor of $r/p = \frac{\sum_{i=1}^{p} |R_i|}{p}$, which is roughly the average cardinality of local neighborhoods. Again, when the local neighborhood size increases more slowly than p in the high-dimensional regime, this additional factor becomes relatively insignificant.

3) The mild sample size requirement is partly due to our distributed framework, under which the stochastic deviation is smaller since a smaller set of parameters needs to be considered for each local RMML problem. However, the additional parameters introduced by convex relaxation and the aggregation of local estimation errors result in the additional factor r/p mentioned above. This demonstrates the trade-off due to the desire for distributed, convex optimization in the proposed framework.

C. Robustness Against Model Mismatch

One of the premises of the estimation framework we consider in this paper is that the true structure of the graph is known. However, this assumption could be violated in practice. In this section, we investigate the robustness of the estimators against small structure mismatch. Our specific interest is in the bias due to model mismatch and hence we focus on the infinite sample regime.

We first consider the GML problem. The GML estimator effectively provides a mapping from the edge elements of moment (covariance) parameters $\widehat{\mathbf{\Sigma}}_{\widetilde{E}}$ to the canonical (concentration) parameters $\widehat{\mathbf{J}}^{\mathrm{GML}}_{\widetilde{E}}$. We denote this mapping as $\mathcal{M}(\cdot; \widetilde{E})$, i.e., $\widehat{\mathbf{J}}^{\mathrm{GML}}_{\widetilde{E}} = \mathcal{M}(\widehat{\mathbf{\Sigma}}; \widetilde{E})$. This mapping is specified implicitly by the optimality condition:

$$\widehat{\mathbf{\Sigma}}_{\widetilde{E}} - \left(\left(\widehat{\mathbf{J}}^{\mathrm{GML}} \right)^{-1} \right)_{\widetilde{E}} = \mathbf{0}. \tag{28}$$

Due to a property of minimal exponential families, $\mathcal{M}(\cdot; \widetilde{E})$ exists and is unique provided that covariance matrix $\widehat{\mathbf{\Sigma}}$ is positive definite [4]. Also by the implicit function theorem, $\mathcal{M}(\cdot; \widetilde{E})$ is differentiable and thus continuous.

Consider a perturbed concentration matrix $\widetilde{\mathbf{J}}^*$ which has uniformly bounded perturbations on the non-edge entries with respect to the nominal parameter $\mathbf{J}^*$:

$$\widetilde{\mathbf{J}}^* = \mathbf{J}^* + \Delta \mathbf{J}, \tag{29}$$

where $\Delta \mathbf{J}$ is supported only on $\widetilde{E}^C$. We assume the perturbation is small enough, such that the perturbed matrix is still positive definite. Denote the corresponding covariance matrix as $\widetilde{\mathbf{\Sigma}} = (\widetilde{\mathbf{J}}^*)^{-1}$. Then the bias of the GML estimator due to model perturbation can be obtained by a first-order perturbation analysis of the GML mapping defined above.

Let $\Gamma = \mathbf{\Sigma} \otimes \mathbf{\Sigma}$ denote the Hessian of the GML problem (4) with no sparsity constraints, which is also related to the Jacobian of (28) with respect to $\widehat{\mathbf{J}}^{\mathrm{GML}}_{\widetilde{E}}$. We have

$$
\begin{aligned}
\widetilde{\mathbf{J}}^{\mathrm{GML}}_{\widetilde{E}} &= \mathcal{M}(\widetilde{\mathbf{\Sigma}}; \widetilde{E}) \\
&= \mathcal{M}((\widetilde{\mathbf{J}}^*)^{-1}; \widetilde{E}) \\
&= \mathcal{M}((\mathbf{J}^* + \Delta \mathbf{J})^{-1}; \widetilde{E}) \\
&= \mathcal{M}(\mathbf{J}^{*-1} + \Gamma_{\widetilde{E}, \widetilde{E}^C} \Delta \mathbf{J}_{\widetilde{E}^C} + O(\|\Delta \mathbf{J}\|^2); \widetilde{E}) \\
&= \mathcal{M}(\mathbf{\Sigma}^*; \widetilde{E}) + (\Gamma_{\widetilde{E}, \widetilde{E}})^{-1} \Gamma_{\widetilde{E}, \widetilde{E}^C} \Delta \mathbf{J}_{\widetilde{E}^C} + O(\|\Delta \mathbf{J}\|^2),
\end{aligned}
$$

where in the second-to-last relation we have used the first-order approximation of matrix inversion, and the last identity is due to the implicit function theorem applied to the optimality condition

(28). Also note that $\mathbf{J}^* = \mathcal{M}(\mathbf{\Sigma}^*; \widetilde{E})$ due to consistency of the GML estimator.

Therefore the maximum element-wise bias with respect to the new model can be bounded as follows (disregarding higher-order terms):

$$
\begin{aligned}
&\|\widetilde{\mathbf{J}}^{\mathrm{GML}} - \widetilde{\mathbf{J}}^*\|_\infty \\
&\leq \|\widetilde{\mathbf{J}}^{\mathrm{GML}} - \mathbf{J}^*\|_\infty + \|\widetilde{\mathbf{J}}^* - \mathbf{J}^*\|_\infty \\
&\lesssim \left\|\left|\left(\Gamma_{\widetilde{E}, \widetilde{E}}\right)^{-1} \Gamma_{\widetilde{E}, \widetilde{E}^C}\right|\right\|_\infty \|\Delta\mathbf{J}_{\widetilde{E}^C}\|_\infty + \|\Delta\mathbf{J}_{\widetilde{E}^C}\|_\infty, \quad (30)
\end{aligned}
$$

where we recall $\|\cdot\|$ is the induced ∞/∞ matrix norm.

The second term in the last display is the inevitable bias due to model mismatch, while the first term captures the additional bias attributable to the GML estimator under model perturbation. The additional bias depends on $\left\|\left|\left(\Gamma_{\widetilde{E}, \widetilde{E}}\right)^{-1} \Gamma_{\widetilde{E}, \widetilde{E}^C}\right|\right\|_\infty$, which is intuitively related to the level of *incoherence* between the edge and non-edge elements in the Hessian of the GML problem (4). Similar incoherence quantities have been shown to play a crucial role in the literature on variable selection [25] (e.g., Lasso) and structure estimation in Gaussian graphical models [13]. Therefore the smaller this incoherence parameter is, the more robust the GML estimator will be.

Since each local problem in RMML estimation has the same structure as the GML problem, we can apply similar analysis to each local neighborhood. The resulting bound on the bias of the RMML estimator is dependent on similar incoherence parameters but defined with respect to relaxed edge sets in the local neighborhoods. We conjecture that these local incoherence parameters are comparable to, if not smaller than, the global incoherence. Hence the robustness of the distributed RMML estimator is expected to be comparable to the GML estimator. While our conjecture is not formally proven in this paper, it is positively supported by the numerical experiments in Section VI.

V. Computational Complexity and Implementation

In this section we discuss the computational complexity of the proposed RMML approach and some implementation issues. First we note that each local RMML problem has the same structure as the centralized ML problem, which is a $\log\det$-regularized semidefinite program ($\log\det$-SDP). Therefore many well-developed solvers and efficient specialized algorithms can be used. Furthermore, due to the distributed nature of the RMML approach, the local problems can all be solved in parallel before the final one-step averaging. The combination of lower dimensionality in the local problems and parallelization can significantly reduce the total run time compared with centralized algorithms.

In terms of algorithms, we find the iterative regression method introduced in [26] is very efficient for sparse graphs. This algorithm iteratively performs linear regressions of each node variable against its immediate neighbors until global convergence. However, the major drawback of this algorithm is the need to maintain global parameters, which prevents direct parallelization and also makes implementation difficult in distributed networks (as discussed below).

The computational advantage of the proposed RMML algorithm becomes more obvious when the number of variables p increases to large numbers. Assuming that the local neighborhood dimensions increase more slowly than p, such as with K-NN graphs and lattice graphs, the total complexity of the RMML estimator scales linearly in p, independent of the algorithm used to solve the local problems. The run time increases even more slowly if the overall algorithm can be parallelized. In contrast, for the centralized algorithms, the dependence of complexity on p is at least linear and is much faster for denser graphs and/or if generic $\log\det$-SDP solvers are used.

Another advantage of the proposed RMML algorithm is that it is highly suitable for network applications due to its minimal requirement for message passing which reduces communication cost. In sharp contrast, many centralized algorithms, such as the iterative regression algorithm mentioned above, require centralized storage and iterative updating of a large number of variables, which in turn requires expensive communication among non-adjacent nodes in the network.

VI. Experiments

In this section, we evaluate the proposed RMML estimator and compare it with the centralized and other distributed estimators in the literature. All methods have been coded in Matlab routines that will be available at the reproducible research web page[1]. We focus on the one-hop and two-hop versions of the RMML estimator (denoted as RelaxMML-1hop and RelaxMML-2hop, respectively). Other estimators considered in this section are:

- The centralized GML estimator, denoted as GML in the legends;
- The *LOCAL* and *AVE* estimators from [6], 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 [6] and [5] and denoted as PML-ADMM. We use the weights $\left[\hat{\mathbf{J}}_{i,i}^{LOC}\right]^2$ as in [6].

We first verify the classical asymptotic rates for the proposed estimators predicted by Theorems 1 and 2 (see Fig. 2) using 10,000 randomized runs sampled from a four-nearest-neighbor Gaussian graphical model with $p = 20$ nodes distributed uniformly in space over the unit square. 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 Euclidean distance between the ith and jth nodes. The empirical normalized mean squared errors (MSE), defined as $\frac{\|\hat{\mathbf{J}} - \mathbf{J}\|_F^2}{\|\mathbf{J}\|_F^2}$, are computed from Monte Carlo samples, and they are compared with the theoretical bounds predicted by Theorem 1. Fig. 2 illustrates the tightness of these bounds. It is also worth noting that the bound for the two-hop RMML estimator is much lower than that of the one-hop estimator, as predicted by Theorem 2. The two-hop bound approximates the bound for the GML estimator closely, suggesting that RMML
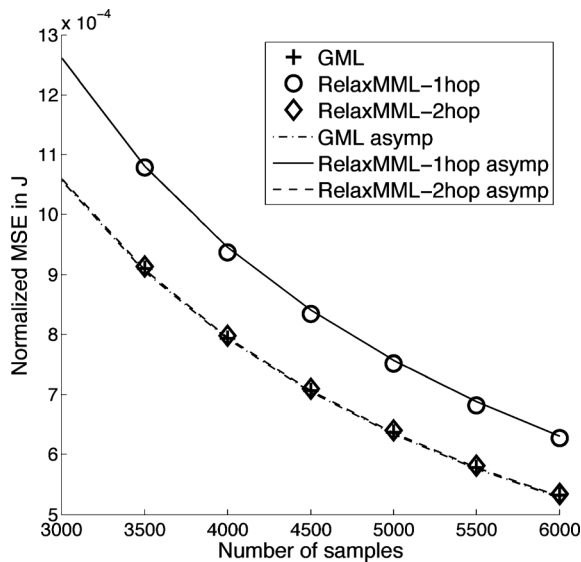
---

[1] http://tbayes.eecs.umich.edu/rrpapers

Fig. 2. Asymptotic normalized MSE for K-NN graphs ($p = 20$, $K = 4$). The curves denote the theoretical asymptotic limits, whereas the symbols denote the empirical normalized MSE over 10,000 runs.

estimators are nearly asymptotically efficient. The asymptotic bounds for RMML estimators with larger neighborhoods follow the monotonicity relation in Theorem 2, however the differences are too small to visually identify, and hence are omitted from the plot.

Next we evaluate the non-asymptotic MSE performance of the proposed estimator, and compare it with the other estimators on both synthetic and real-world data sets. For synthetic data sets, we consider three classes of graphs that are motivated by real-world applications. For each class we follow similar experiment settings as in [6]. Specifically, we randomly generate 20 topologies and associated sparse 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 MSEs are averaged over all 200 experiments, and are reported in Fig. 3. An illustration of the graph topology is shown in the top-right corner of each plot. The classes of graphs we consider are:

- **K-NN graphs** (Fig. 3(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 Euclidean distance between the ith and jth nodes. Finally we add a small value to the diagonal to ensure positive definiteness.

- **Lattice graphs** (Fig. 3(b)): A lattice graph is appropriate for networks with regular spatial correlations, e.g., images that are Markov random fields. 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** (Fig. 3(c)): Small-world graphs have been proposed for social networks, biological networks, etc., where most nodes have few immediate neighbors but can be reached from any other node through a small number of hops [28]. We generate graphical models structured as random small-world networks using the Watts-Strogatz mechanism [28] 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 to ensure that $\mathbf{J}$ is positive definite.

The MSE curves shown in Fig. 3 match our theoretical predictions in Section IV-B, and they also demonstrate the superior performance of the proposed RMML estimator. In particular, for the graphs that have relatively small two-hop neighborhoods, namely the K-NN graphs and the lattice grids, the MSE of the proposed two-hop relaxed MML estimator almost coincides with the MSE of the global MLE. On the other hand, for 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. These graphs are known to be harder to learn through distributed algorithms. The two-hop relaxed MML estimator still outperforms the other distributed algorithms by a large margin.

Next, we apply the estimators to a real-world sensor network. The IntelLab dataset [27] 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 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 Fig. 3(d), the proposed two-hop relaxed MML estimator still gives a very tight approximation to the centralized GML estimator and its advantage over other distributed estimators is obvious.

We investigate the robustness of the centralized and distributed estimators in the presence of model mismatch. The nominal precision matrix $\mathbf{J}$ corresponds to a four-nearest-neighbor graphical model with $p = 200$ as in the previous experiments. We add ± 0.1 random perturbations to the non-edge components of the nominal precision matrix (also with minimal diagonal loading to ensure positive definiteness), then generate samples from the perturbed model. The different estimation algorithms are applied assuming the nominal graph structure and the resulting MSEs are plotted with respect to
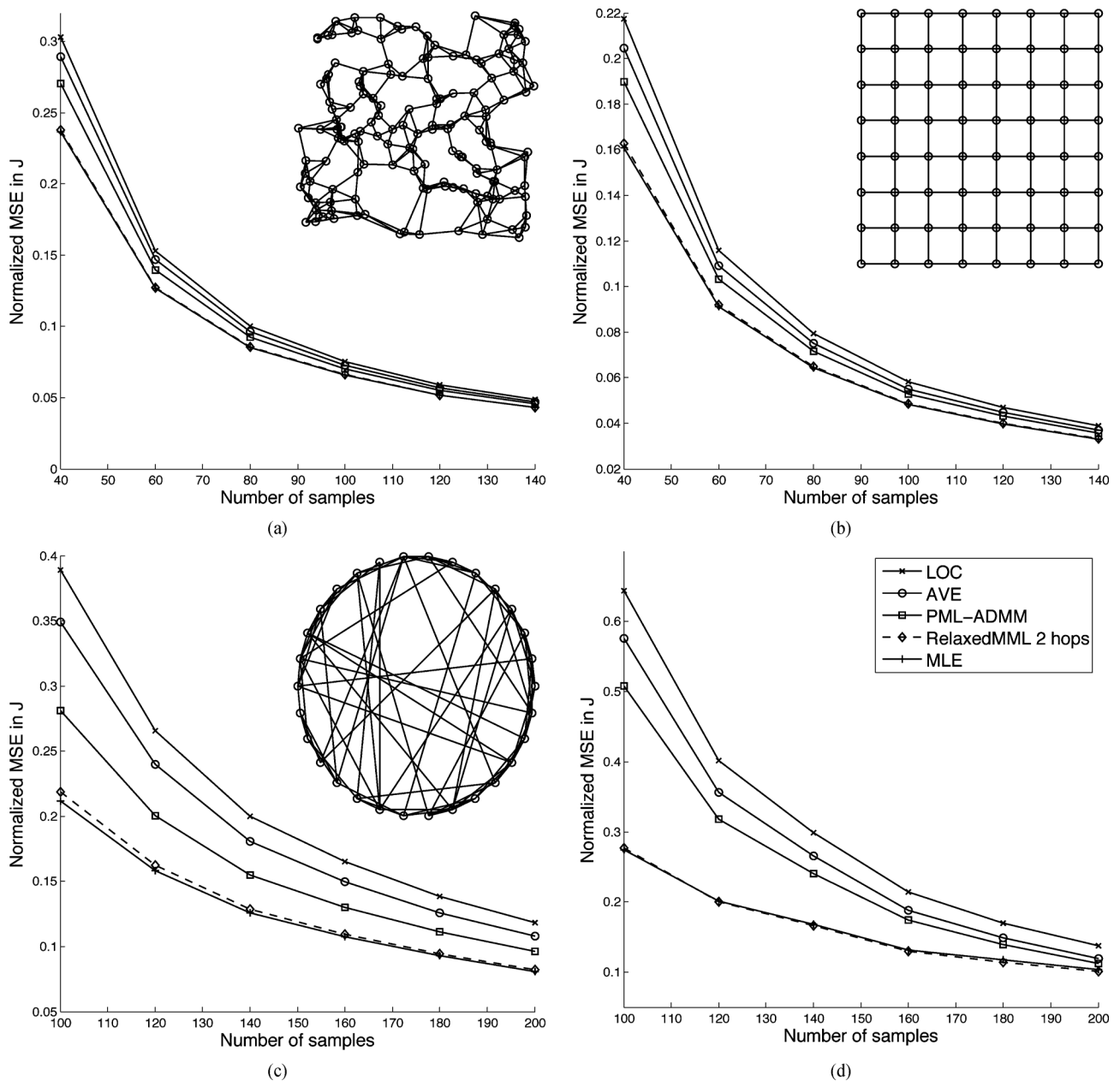
Fig. 3. Normalized MSE in the concentration matrix estimates for different graphical models. The legend in Fig. 3(d) applies to all plots. The proposed 2-hop relaxed maximum marginal likelihood (RMML) estimator clearly improves upon existing distributed estimators and nearly closes the gap to the centralized maximum likelihood estimator. (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 [27] ($p = 50$).

the nominal model. The MSEs of all estimators (using samples from both the original and perturbed models, respectively) are reported in Fig. 4. All errors are averaged across 50 randomized experiments. As can be seen, the model mismatch leads to estimation bias for both centralized and distributed estimators. The magnitudes of the model mismatch bias for all estimators are comparable, as predicted by the theoretical analysis in Section IV-C. These experiment results confirm the robustness of the proposed distributed algorithm.

We next turn to computational comparisons. In the following experiments, we illustrate the computational gain of our distributed estimator over the centralized one through two runtime comparisons performed in Matlab. Our main focus is on the rela-

tive scaling of the runtime with respect to the number of nodes p for different estimators. We consider two algorithms for solving both the centralized GML problem and the local RMML problems. The first is an interior point algorithm implemented in the solver logdetPPA [29], which is specially designed for solving $\log \det$-SDPs. The second algorithm is the iterative regression approach in [26] for solving the covariance selection problem [10] with known structure. In both experiments, the graphical model is a four-NN graph with similar parameter settings as before. We compare the total runtime of the GML estimator and that of different versions of RMML estimators. For the RMML estimators, we implement a sequential and a parallel version using the parfor function in Matlab. The results are reported in
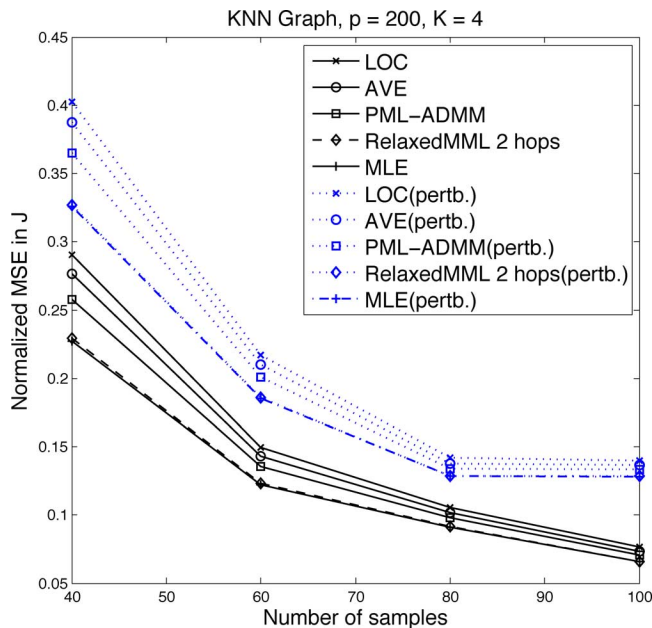
Fig. 4. Robustness of estimators under model mismatch. All errors are obtained from K-NN ($p = 200$, $K = 4$) graphs and averaged over 50 experiments. For the perturbed models, ± 0.1 is added to the non-edge components of the nominal precision matrix. The proposed distributed RMML estimator is as robust as the GML estimator.

Fig. 5. As expected, the runtime of the GML estimator is at least linear in p and the generic solver appears to be much more expensive than the iterative regression algorithm for this particular task. The total cost of the RMML estimator without parallelization is also linear in p, and is slightly higher than the GML estimator. However, when four-core parallelization is used, the run time is approximately reduced by a factor of four, resulting in lower computational complexity after $p > 500$.

It is expected that with a higher degree of parallelization, the run time of the proposed RMML estimator will continue to decrease almost linearly with the number of cores. As discussed in Section V, all local RMML problems can be solved in parallel without the need for any iterative message-passing. Therefore the communication overhead is minimal, consisting of the final concatenation and symmetrization steps (16) and (18).

VII. CONCLUSION

We have proposed a distributed MML framework for estimating the concentration matrix in Gaussian graphical models. The proposed method solves a convex relaxation of a marginal likelihood maximization problem independently in each local neighborhood. 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 statistically consistent and computationally efficient. In particular, we have shown that the statistical convergence rate of our estimator is comparable to that of the more expensive centralized maximum likelihood estimator. Likewise in numerical experiments, a two-hop version of the distributed estimator is seen to be sufficient to attain centralized performance. Its improved performance relative to existing distributed estimators is also illustrated.

APPENDIX A
PROOF OF THEOREM 1

Proof: Consider the following set of sparse positive semidefinite matrices with respect to a non-zero element set R:

$$\mathcal{K}^R := \{\mathbf{K} \mid \mathbf{K} \succeq 0, \mathbf{K}_{(j,k)} = 0, \quad \forall (j,k) \notin R\}$$

We first note that, when R is taken to be the relaxed edge set of a neighborhood as defined in (13), 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 (10) and (11). Therefore the proposed relaxed MML problem (14) 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 [23]. In particular, the variances of the errors achieve the diagonal elements of the inverse Fisher information matrix $\mathbf{F}$ defined in (20) (see [30] for the derivation). 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 $\widehat{\mathbf{J}}^{\mathrm{Relax}}$. ∎

APPENDIX B
PROOF OF THEOREM 2

Proof: We first consider the case of $k = 1$, i.e., we compare the asymptotic variances of the one-hop and two-hop RMML estimators. Subsequently we generalize the arguments to $k > 1$ and to the global ML estimator. Suppressing the index i for local neighborhoods, let $\mathcal{B}^j, \mathcal{N}^j$ be the sets of buffer and all nodes (i.e., variables) with respect to the j-hop neighborhood, respectively ($j = 1, 2$).

Next we define some set notation for edge parameters. Let $E^j = \widetilde{E} \cap (\mathcal{N}^j \times \mathcal{N}^j)$ denote the subset of edges in $\widetilde{E}$ with both endpoints in $\mathcal{N}^j$. Let B^j be the set of all possible edges connecting j-hop buffer nodes, i.e., $B^j := \mathcal{B}^j \times \mathcal{B}^j$. Recall from (15) that L denotes the set of row parameters, which is defined as $L = E^1 \backslash B^1$. Finally note that the (j-hop) relaxed edge sets defined in (13) are related to the above two sets as $R^j := E^j \cup B^j, j = 1, 2$.

We augment the two-hop neighborhood graph by adding all edges among one-hop buffer nodes and among two-hop buffer nodes that are not already in E^2 (see Fig. 6 for an illustration). This augmented edge set is denoted as $\overline{E^2} := E^2 \cup B^1 \cup B^2$. After this augmentation, the one-hop buffer clique B^1 separates the two-hop neighborhood graph into two components and a non-overlapping decomposition follows:

$$\overline{E^2} = [\underbrace{L, B^1}_{C_1 = R^1}, L, \overbrace{B^1, (E^2 \backslash E^1) \cup B^2}^{C_2}], \qquad (31)$$

where we define two subsets C_1 and C_2. The augmented two-hop neighborhood graph is therefore decomposed by $(C_1 \backslash B^1, B^1, C_2 \backslash B^1)$ [3, Def. 2.1].
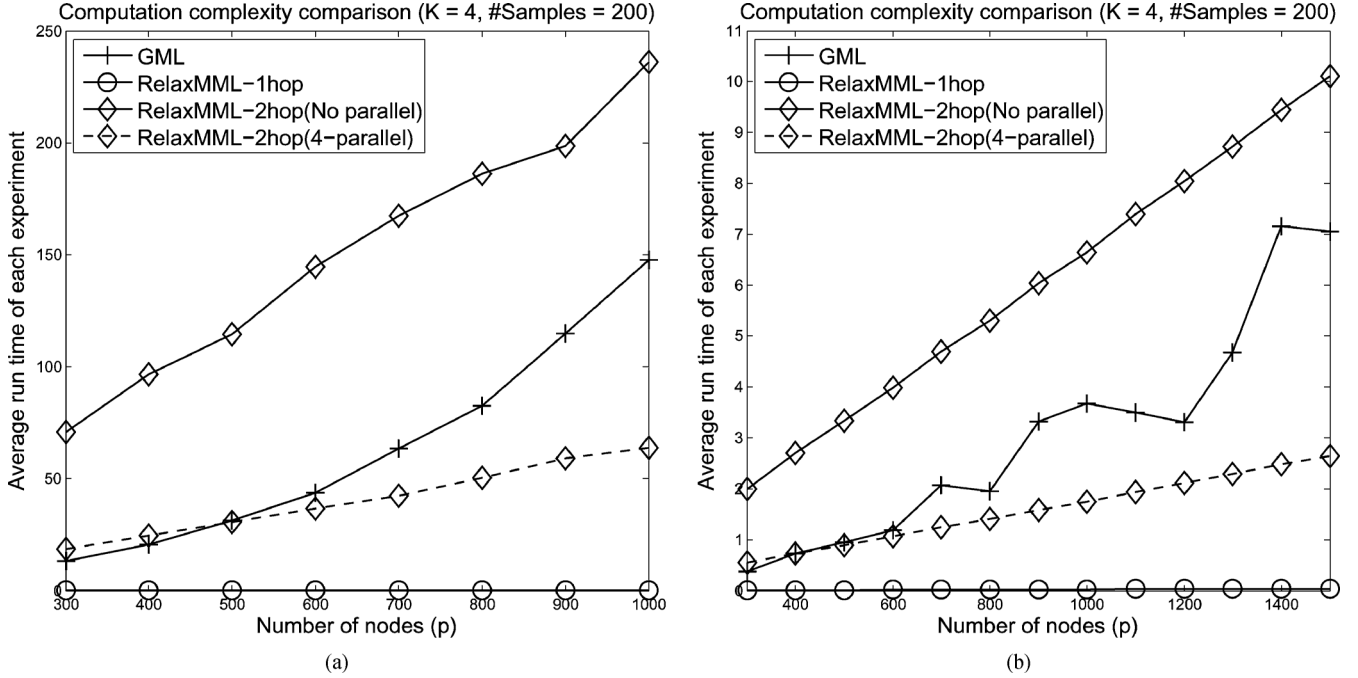
Fig. 5. Run time comparisons for GML and RMML estimators. In panel (a) the logdetPPA solver is used, and in panel (b) the iterative regression algorithm is used. In both figures, solid lines denote the runtime scaling of the sequential version of the algorithm, while the dashed lines denote runtime scaling for a parallelized version with four cores. In both figures, the runtime of the GML estimator is super-linear in p, while the RMML estimator exhibits linear scaling in p, and the runtime is further reduced by a factor approximately equal to the number of cores used. All experiments are implemented in a Matlab environment. (a) Run time comparison using logdetPPA solver. (b) Run time comparison using iterative regression algorithm.
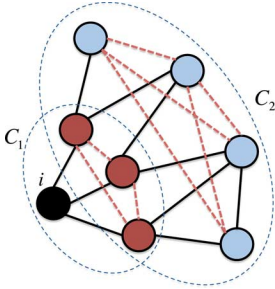

Fig. 6. Illustration of the graph augmentation in the proof of Theorem 2. Dashed red lines indicate the added edges, and dashed blue contours indicate the sets C_1 and C_2, which intersect at the one-hop separator clique formed by red nodes.

Similar to Theorem 1, the asymptotic error covariance matrix of the RMML estimator for the augmented two-hop neighborhood is the inverse of corresponding Fisher information matrix (FIM), denoted as $\overline{\mathbf{F}}$. By Proposition 5.8 in [3], the decomposability of the augmented graph leads to the following decomposition of the inverse of FIM:

$$\overline{\mathbf{F}}^{-1} = \left[(\overline{\mathbf{F}}_{C_1,C_1})^{-1}\right]^0 + \left[(\overline{\mathbf{F}}_{C_2,C_2})^{-1}\right]^0 - \left[(\overline{\mathbf{F}}_{B^1,B^1})^{-1}\right]^0,$$

where $[\cdot]^0$ appropriately zero-pads its argument to conform to the dimensions of $\overline{\mathbf{F}}^{-1}$.

Restricting this relation to the row parameters L, we have

$$\overline{\mathbf{F}}^{-1}_{L,L} = (\overline{\mathbf{F}}_{C_1,C_1})^{-1}_{L,L}, \tag{32}$$

since the row parameters are only contained in C_1. Noting that set C_1 is equivalent to the one-hop relaxed edge set R^1, then

$$\overline{\mathbf{F}}^{-1}_{L,L} = (\overline{\mathbf{F}}_{C_1,C_1})^{-1}_{L,L} = (\overline{\mathbf{F}}_{R^1,R^1})^{-1}_{L,L}. \tag{33}$$

Therefore, from Theorem 1 we have that the asymptotic mean squared error of the RMML estimator using the augmented graph is the same as that of the one-hop RMML estimator.

On the other hand, the augmented edge set $\overline{E^2}$ is different from the relaxed edge set R^2 only in the one-hop buffer clique B^1. Therefore another possible decomposition of the augmented edge set is (after re-ordering):

$$\overline{E^2} = [\underbrace{L, E^1\backslash L, (E^2\backslash E^1) \cup B^2}_{R^2}, \underbrace{B^1\backslash E^1}_{D}], \tag{34}$$

where we define the difference set as D. Then using a property of Schur complements of positive semidefinite matrices, the variance matrix corresponding to R^2 (i.e., the non-zero pattern of the two-hop RMML estimator) satisfies

$$\overline{\mathbf{F}}^{-1}_{R^2,R^2} = \left(\overline{\mathbf{F}}_{R^2,R^2} - \overline{\mathbf{F}}_{R^2,D}(\overline{\mathbf{F}}_{D,D})^{-1}\overline{\mathbf{F}}_{D,R^2}\right)^{-1} \succeq \left(\overline{\mathbf{F}}_{R^2,R^2}\right)^{-1}. \tag{35}$$

Restricting this relation to the submatrix indexed by set L, we have

$$\overline{\mathbf{F}}^{-1}_{L,L} \succeq \left(\overline{\mathbf{F}}_{R^2,R^2}\right)^{-1}_{L,L}. \tag{36}$$

Now combining (33), (36) and Theorem 1, we can conclude that the asymptotic variance of the one-hop RMML estimator (i.e., the mean squared error) is larger than that of the two-hop estimator.

Similar arguments can be established for comparing the asymptotic variances of the two-hop RMML and the GML estimators, which shows that the asymptotic variance of RMML estimator is larger than that of the GML estimator. The above proof can be easily generalized to arbitrary k-hop neighborhoods. ∎

APPENDIX C
PROOF OF THEOREM 3

The key ingredient in proving Theorem 3 is the following lemma, which provides a bound for the error of the RMML estimator $\widehat{\mathbf{K}}^{\mathrm{Relax}}$ (14) in a given local neighborhood (the neighborhood index i is suppressed). Let $\boldsymbol{\Sigma}^*$ be the true global covariance matrix, and $\mathbf{K}^*$ be the true marginal precision matrix corresponding to the given neighborhood.

Lemma 1: For a given local neighborhood $\mathcal{N}$, if

$$\left\| \left(\widehat{\boldsymbol{\Sigma}}^{(T)} - \boldsymbol{\Sigma}^* \right)_R \right\|_\infty \leq \epsilon_\Sigma \leq \frac{1}{9\overline{\kappa}\sqrt{|R|}}, \qquad (37)$$

we have

$$\left\| \widehat{\mathbf{K}}^{\mathrm{Relax}} - \mathbf{K}^* \right\|_F \leq 9\overline{\kappa}^2 \epsilon_\Sigma \sqrt{|R|}. \qquad (38)$$

The proof of Lemma 1 is given in Appendix D. The above lemma is deterministic in nature. To ensure that assumption (37) is satisfied with high probability when the sample covariance $\widehat{\boldsymbol{\Sigma}}^{(T)}$ is random, we make use of the following concentration result for Gaussian random variables by Ravikumar *et al.* [13]:

Lemma 2: For a p-dimensional Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}^*$, the sample covariance matrix obtained from T samples $\widehat{\boldsymbol{\Sigma}}^{(T)}$ satisfies

$$P\left\{ |\widehat{\boldsymbol{\Sigma}}_{i,j}^{(T)} - \boldsymbol{\Sigma}_{i,j}^*| > \epsilon \right\} \leq 4 \exp\left(-\frac{T \cdot \epsilon^2}{3200\overline{\sigma}^2} \right), \qquad (39)$$

for all $\epsilon \in (0, 40\overline{\sigma})$, *where* $\overline{\sigma} := \max_{i=1,\dots,p} \boldsymbol{\Sigma}_{i,i}^*$.

Now we are ready to prove Theorem 3.

Proof: (Theorem 3) Given the condition (26) on T, we have

$$C\sqrt{\frac{3200\overline{\sigma}^2 \log p^2}{T}} \leq 40\overline{\sigma}. \qquad (40)$$

Then applying Lemma 2 and the union bound, we have

$$P\left\{ \left\| \left(\widehat{\boldsymbol{\Sigma}}^{(T)} - \boldsymbol{\Sigma}^* \right)_{R_i} \right\|_\infty \leq C\sqrt{\frac{3200\overline{\sigma}^2 \log p^2}{T}} \right\}$$

$$\geq P\left\{ \left\| \widehat{\boldsymbol{\Sigma}}^{(T)} - \boldsymbol{\Sigma}^* \right\|_\infty \leq C\sqrt{\frac{3200\overline{\sigma}^2 \log p^2}{T}} \right\}$$

$$\geq 1 - \frac{4}{p^{2(C^2-1)}}. \qquad (41)$$

Conditioned on the event in (41), condition (26) also guarantees that (37) holds for all local neighborhoods. Then the total Frobenius error in the global estimate $\widehat{\mathbf{J}}^{\mathrm{Relax}}$ can be bounded by Lemma 1:

$$\|\widehat{\mathbf{J}}^{\mathrm{Relax}} - \mathbf{J}^*\|_F \overset{(i)}{=} \left(\sum_{i=1}^p \|(\widehat{\mathbf{J}}^{\mathrm{Relax}} - \mathbf{J}^*)_{L_i}\|_F^2 \right)^{1/2}$$

$$\overset{(ii)}{=} \left(\sum_{i=1}^p \|(\widehat{\mathbf{K}}^{\mathrm{Relax}} - \mathbf{K}^*)_{L_i}\|_F^2 \right)^{1/2}$$

$$\overset{(\mathrm{Lem.\ 1})}{\leq} \left(\sum_{i=1}^p \left(9\overline{\kappa}^2 C\sqrt{\frac{3200\overline{\sigma}^2 |R_i| \log p^2}{T}} \right)^2 \right)^{1/2}$$

$$\leq 720 C \cdot \overline{\kappa}^2 \overline{\sigma} \sqrt{\frac{r \log p}{T}},$$

where identity (i) is due to the fact that the global estimator is a concatenation of non-overlapping row parameter sets (see (15) for definition of L_i's), equality (ii) is due to our construction of $\widehat{\mathbf{J}}^{\mathrm{Relax}}$ from $\widehat{\mathbf{K}}^{\mathrm{Relax}}$ (see (16)), and the fact that row parameters are always protected. ∎

APPENDIX D
PROOF OF LEMMA 1

Proof: The main idea of this proof is inspired by [14]. The difference is that we focus on the local RMML problem, rather than the global ML problem (which is studied in [14]). Define the marginal likelihood function for a local neighborhood $\mathcal{N}$ as $\mathcal{L}(\mathbf{K}) = \langle \widehat{\boldsymbol{\Sigma}}_{\mathcal{N},\mathcal{N}}^{(T)}, \mathbf{K} \rangle - \log \det(\mathbf{K})$, where we super-script the sample covariance to emphasize that it is obtained from T samples.

Recall $\mathbf{K}^* := \left(\boldsymbol{\Sigma}_{\mathcal{N},\mathcal{N}}^* \right)^{-1}$ is the local marginal precision matrix. Define the shorthand notation for the local RMML estimate as $\widehat{\mathbf{K}} := \widehat{\mathbf{K}}^{\mathrm{Relax}}$.

Consider the function $\mathcal{Q}(\Delta) := \mathcal{L}(\mathbf{K}^* + \Delta) - \mathcal{L}(\mathbf{K}^*)$, where Δ respects the sparsity structure of the RMML problem, i.e., $\Delta_{R^C} = \mathbf{0}$ and $\Delta = \Delta^T$. Let $0 < \delta \leq \overline{\kappa}$ be a given radius, define the following set

$$\mathcal{C}(\delta) := \{\Delta \mid \Delta_{R^C} = \mathbf{0}, \Delta = \Delta^T, \|\Delta\|_F = \delta\}, \qquad (42)$$

where R is the local relaxed edge set. Note that $\mathcal{C}(\delta)$ defines a sphere, not a ball.

Note that $\mathcal{Q}(\Delta)$ is a convex function of Δ. By construction we have $\mathcal{Q}(\mathbf{0}) = 0$, and the optimality of $\widehat{\mathbf{K}}^{\mathrm{Relax}}$ implies that $\mathcal{Q}(\widehat{\Delta}) \leq \mathcal{Q}(\mathbf{0}) = 0$, where we define $\widehat{\Delta} := \widehat{\mathbf{K}} - \mathbf{K}^*$. Then if we can establish that

$$\inf_{\Delta \in \mathcal{C}(\delta)} \mathcal{Q}(\Delta) > 0,$$

then the optimal error matrix $\widehat{\Delta}$ must lie inside the sphere defined by $\mathcal{C}(\delta)$ by convexity of $\mathcal{Q}$, implying that $\|\widehat{\Delta}\|_F \leq \delta$. Now it suffices to find a suitable radius $\delta > 0$ such that $\mathcal{Q}(\Delta)$ is lower-bounded from zero for all $\Delta \in \mathcal{C}(\delta)$.

Since

$$\mathcal{Q}(\Delta) = \mathcal{L}(\mathbf{K}^* + \Delta) - \mathcal{L}(\mathbf{K}^*)$$
$$= \langle \widehat{\boldsymbol{\Sigma}}_{\mathcal{N},\mathcal{N}}^{(T)}, \Delta \rangle - (\log \det(\mathbf{K}^* + \Delta) - \log \det(\mathbf{K}^*)).$$

Similar to [14], we make use of the Taylor's theorem for the $\log \det(\cdot)$ function

$$\log \det(\mathbf{K}^* + \Delta) - \log \det(\mathbf{K}^*) = \langle (\mathbf{K}^*)^{-1}, \Delta \rangle$$
$$- \overrightarrow{\Delta}^T \left[\int_0^1 (1-t)(\mathbf{K}^* + t\Delta)^{-1} \otimes (\mathbf{K}^* + t\Delta)^{-1} dt \right] \overrightarrow{\Delta}, \qquad (43)$$

where $\otimes$ denotes the Kronecker product, and $\overrightarrow{\Delta}$ is the properly vectorized form of matrix Δ.

Using this identity, we have

$$
\mathcal{Q}(\Delta) = \underbrace{\langle \widehat{\boldsymbol{\Sigma}}^{(T)}_{\mathcal{N},\mathcal{N}} - (\mathbf{K}^*)^{-1}, \Delta \rangle}_{T_1}
$$
$$
+ \underbrace{\overrightarrow{\Delta}^T \left[\int_0^1 (1-t)(\mathbf{K}^* + t\Delta)^{-1} \otimes (\mathbf{K}^* + t\Delta)^{-1} dt \right] \overrightarrow{\Delta}}_{T_2}.
$$

$$(44)$$

Next we bound T_1 and T_2 defined above separately.

For T_1, notice that the difference matrix Δ is non-zero only in a restricted set R, therefore it reduces to a lower-dimensional inner product:

$$
\begin{aligned}
|T_1| &= |\langle (\widehat{\boldsymbol{\Sigma}}^{(T)} - \boldsymbol{\Sigma}^*)_R, \Delta_R \rangle| \\
&\overset{(i)}{\leq} \|(\widehat{\boldsymbol{\Sigma}}^{(T)} - \boldsymbol{\Sigma}^*)_R\|_\infty \cdot \|\Delta_R\|_1 \\
&\overset{\text{Eq.(37)}}{\leq} \epsilon_\Sigma \cdot \sqrt{|R|} \cdot \|\Delta\|_F,
\end{aligned}
$$

$$(45)$$

where (i) is due to the duality between norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$.

For T_2, we follow similar derivations as in [14]:

$$
\begin{aligned}
T_2 &\geq \|\Delta\|_F^2 \\
&\quad \cdot \lambda_{\min}\left(\int_0^1 (1-t)(\mathbf{K}^* + t\Delta)^{-1} \otimes (\mathbf{K}^* + t\Delta)^{-1} dt \right) \\
&\overset{(i)}{\geq} \|\Delta\|_F^2 \int_0^1 (1-t)\lambda_{\min}^2\left((\mathbf{K}^* + t\Delta)^{-1} \right) dt \\
&\geq \frac{1}{2}\|\Delta\|_F^2 \min_{0 \leq t \leq 1} \lambda_{\min}^2\left((\mathbf{K}^* + t\Delta)^{-1} \right) \\
&\overset{(ii)}{\geq} \frac{1}{2}\|\Delta\|_F^2 \min_{\|\widetilde{\Delta}\|_F \leq \delta} \lambda_{\min}^2\left((\mathbf{K}^* + \widetilde{\Delta})^{-1} \right) \\
&\geq \frac{1}{2}\|\Delta\|_F^2 \min_{\|\widetilde{\Delta}\|_F \leq \delta} \|\mathbf{K}^* + \widetilde{\Delta}\|_2^{-2} \\
&\geq \frac{1}{2}\|\Delta\|_F^2 \min_{\|\widetilde{\Delta}\|_F \leq \delta} (\|\mathbf{K}^*\|_2 + \|\widetilde{\Delta}\|_2)^{-2} \\
&\overset{(iii)}{\geq} \frac{1}{2}\|\Delta\|_F^2 \min_{\|\widetilde{\Delta}\|_F \leq \delta} (\overline{\kappa} + \|\widetilde{\Delta}\|_F)^{-2} \geq \frac{1}{8\overline{\kappa}^2}\|\Delta\|_F^2,
\end{aligned}
$$

where (i) follows the eigenvalue property of Kronecker product, (ii) is due to the fact that $\Delta \in \mathcal{C}(\delta)$, (iii) is due to the interlacing property of eigenvalues of sub-matrices

$$
\|\mathbf{K}^*\|_2 = \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma}^*_{\mathcal{N},\mathcal{N}})} \leq \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma}^*)} = \|\mathbf{J}^*\|_2 = \overline{\kappa}, \quad (46)
$$

The last inequality is due to construction, i.e., $\delta \leq \overline{\kappa}$.

Now $\mathcal{Q}(\Delta)$ can be bounded by

$$
\mathcal{Q}(\Delta) \geq -\epsilon_\Sigma \cdot \sqrt{|R|} \cdot \|\Delta\|_F + \frac{1}{8\overline{\kappa}^2}\|\Delta\|_F^2 \quad (47)
$$

$$
= \|\Delta\|_F \left(\frac{1}{8\overline{\kappa}^2}\|\Delta\|_F - \epsilon_\Sigma \cdot \sqrt{|R|} \right). \quad (48)
$$

The proof is complete if the RHS can be lower bounded away from zero. It can be verified that with the choice of ϵ_Σ as in (37), letting $\delta = 9\overline{\kappa}^2 \epsilon_\Sigma \sqrt{|R|}$ suffices. Therefore $\|\widehat{\Delta}\|_F \leq \delta = 9\overline{\kappa}^2 \epsilon_\Sigma \sqrt{|R|}$. ∎

REFERENCES

[1] Z. Meng, D. Wei, A. Hero, and A. Wiesel, "Distributed learning of Gaussian graphical models via marginal likelihoods," in *Proc. 16th Int. Conf. Artif. Intell. Statist.*, 2013, pp. 39–47.

[2] Z. Meng, D. Wei, A. Hero, and A. Wiesel, "Marginal likelihoods for distributed estimation of graphical model parameters," in *Proc. IEEE 5th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Dec. 2013, pp. 73–76.

[3] S. Lauritzen, *Graphical models*. New York, NY, USA: Oxford Univ. Press, 1996, vol. 17.

[4] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1-2, pp. 1–305, 2008.

[5] Q. Liu and A. T. Ihler, "Distributed parameter estimation via pseudo-likelihood," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1487–1494.

[6] A. Wiesel and A. Hero, "Distributed covariance estimation in Gaussian graphical models," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 211–220, 2012.

[7] Z. Meng, A. Wiesel, and A. Hero, "Distributed principal component analysis on networks via directed graphical models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 2877–2880.

[8] O. Banerjee, L. El Ghaoui, A. d'Aspremont, and G. Natsoulis, "Convex optimization techniques for fitting sparse Gaussian graphical models," in *Proc. ACM Int. Conf. Proc. Ser.*, Citeseer, 2006, vol. 148, pp. 89–96.

[9] J. Dahl, L. Vandenberghe, and V. Roychowdhury, "Covariance selection for non-chordal graphs via chordal embedding," *Optim. Methods Softw.*, vol. 23, no. 4, pp. 501–520, 2008.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical Lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[11] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Walk-sums and belief propagation in Gaussian graphical models," *J. Mach. Learn. Res.h*, vol. 7, pp. 2031–2064, 2006.

[12] U. Heinemann and A. Globerson, "What cannot be learned with Bethe approximations," ArXiv preprint, 2012, arXiv:1202.3731 [Online]. Available: http://arxiv.org/abs/1202.3731

[13] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Statist.*, vol. 5, pp. 935–980, 2011.

[14] A. Rothman, P. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electron. J. Statist.*, vol. 2, pp. 494–515, 2008.

[15] C. C. Johnson, A. Jalali, and P. D. Ravikumar, "High-dimensional sparse inverse covariance estimation using greedy methods," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 574–582.

[16] Y. D. Mizrahi, N. de Freitas, and L. Tenorio, "Efficient learning of practical Markov random fields with exact inference," ArXiv preprint, 2013, arXiv:1308.6342 [Online]. Available: http://arxiv.org/abs/1308.6342v3

[17] H. Massam and N. Wang, "Distributed parameter estimation of discrete hierarchical models via marginal likelihoods," ArXiv preprint, 2013, arXiv:1310.5666 [Online]. Available: http://arxiv.org/abs/1310.5666

[18] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Tech. Rep., 2012 [Online]. Available: http://homepages.dcc.ufmg.br/~assuncao/EstatCC/aulas/matrixcookbook.pdf

[19] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proc. 15th Conf. Uncertainty Artif. Intell.*, 1999, pp. 467–475.

[20] J. Pacheco and E. B. Sudderth, "Minimization of continuous Bethe approximations: A positive variation," in *Adv. Neural Inf. Process. Syst.*, 2012, pp. 2573–2581.

[21] M. J. Wainwright, "Estimating the wrong graphical model: Benefits in the computation-limited setting," *J. Mach. Learn. Res.*, vol. 7, pp. 1829–1859, 2006.

[22] P. Liang and M. I. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 584–591.

[23] A. Van der Vaart, *Asymptotic Statistics*. Cambridge, U.K.: Cambridge Univ. Press, 2000, vol. 3.

[24] M. Wainwright, "Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[25] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.

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

[27] C. Guestrin, P. Bodik, R. Thibaux, M. Paskin, and S. Madden, "Distributed regression: An efficient framework for modeling sensor network data," in *Proc. 3rd Int. Symp. Inf. Process. Sensor Netw. (IPSN)*, 2004, pp. 1–10.

[28] D. Watts and S. Strogatz, "The small world problem," *Collect. Dynam. Small-World Netw.*, vol. 393, pp. 440–442, 1998.

[29] C. Wang, D. Sun, and K.-C. Toh, "Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm," *SIAM J. Optim.*, vol. 20, no. 6, pp. 2994–3013, 2010.

[30] J. Johnson, "Fisher information in Gaussian graphical models," MIT, Tech. Note, Sep. 2006 [Online]. Available: http://ssg.mit.edu/~jasonj/gauss-fisher-note-sep06.pdf

Ami Wiesel (M'07) received the B.Sc. and M.Sc. degrees in Electrical Engineering from Tel-Aviv University, Tel-Aviv, Israel, in 2000 and 2002, respectively, and the Ph.D. degree in Electrical Engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 2007. He was a post-doctoral fellow with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, in 2007–2009. Since January 2010, he has been a faculty member at the Rachel and Selim Benin School of Computer Science and Engineering at the Hebrew University of Jerusalem, Israel.

Dr. Wiesel was a recipient of the Young Author Best Paper Award for a 2006 paper in the IEEE Transactions on Signal Processing and a Student Paper Award for a 2005 Workshop on Signal Processing Advances in Wireless Communications (SPAWC) paper. He was awarded the Weinstein Study Prize in 2002, the Intel Award in 2005, the Viterbi Fellowship in 2005 and 2007, and the Marie Curie Fellowship in 2008.

Zhaoshi Meng (S'13) received the B.S. degree in Electrical Engineering in 2009 from Tsinghua University, Beijing, China, and the M.S. degree in Electrical Engineering in 2011 from University of Michigan, Ann Arbor, MI, USA. He is currently a Ph.D. student in the Electrical Engineering and Computer Science department at the University of Michigan. His research interests include machine learning, applied statistics, optimization and signal processing.

Zhaoshi Meng has co-authored a Notable Paper at the 2013 International Conference on Artificial Intelligence and Statistics (AISTATS), a Best Student Paper at the 2013 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), and the Best Paper at the 2014 International Conference on Machine Learning (ICML).
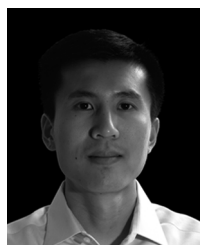
Dennis Wei (S'09–M'11) received S.B. degrees in electrical engineering and in physics in 2006, the M.Eng. degree in electrical engineering in 2007, and the Ph.D. degree in electrical engineering in 2011, all from the Massachusetts Institute of Technology, Cambridge, MA, USA.

He is currently a Research Staff Member at the Thomas J. Watson Research Center, IBM Research, Yorktown Heights, NY, USA. From 2011 to 2013, he was a Post-Doctoral Research Fellow in the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, MI, USA. His research interests lie broadly in signal processing, machine learning, optimization, and statistics. Areas of focus include adaptive sensing/sampling, low-dimensional models, applications of optimization, and sparse filter design.

Dr. Wei received a Notable Paper Award at the 2013 International Conference on Artificial Intelligence and Statistics (AISTATS) and co-authored a Best Student Paper at the 2013 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP). He is a recipient of the William Asbjornsen Albert Memorial Fellowship at MIT and a Siebel Scholar. He is a member of Phi Beta Kappa, Sigma Xi, Eta Kappa Nu, and Sigma Pi Sigma.

Alfred O. Hero, III (S'79–M'84–SM'96–F'98) received the B.S. degree (*summa cum laude*) from Boston University (1980) and the Ph.D. degree from Princeton University (1984), both in Electrical Engineering. Since 1984 he has been with the University of Michigan, Ann Arbor, where he is the R. Jamison and Betty Williams Professor of Engineering. His primary appointment is in the Department of Electrical Engineering and Computer Science and he also has appointments, by courtesy, in the Department of Biomedical Engineering and the Department of Statistics. From 2008 to 2013 he held the Digiteo Chaire d'Excellence, sponsored by Digiteo Research Park in Paris, located at the Ecole Superieure d'Electricite, Gif-sur-Yvette, France. He has held other visiting positions at LIDS Massachusetts Institute of Technology (2006), Boston University (2006), I3S University of Nice, Sophia-Antipolis, France (2001), Ecole Normale Supérieure de Lyon (1999), Ecole Nationale Supérieure des Télécommunications, Paris (1999), Lucent Bell Laboratories (1999), Scientific Research Labs of the Ford Motor Company, Dearborn, Michigan (1993), Ecole Nationale Superieure des Techniques Avancees (ENSTA), Ecole Superieure d'Electricite, Paris (1990), and M.I.T. Lincoln Laboratory (1987–1989). His recent research interests are in statistical signal processing, machine learning and the analysis of high dimensional spatio-temporal data. Of particular interest are applications to networks, including social networks, multi-modal sensing and tracking, database indexing and retrieval, imaging, biomedical signal processing, and biomolecular signal processing.

Alfred Hero is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE). He received the University of Michigan Distinguished Faculty Achievement Award (2011). He has been plenary and keynote speaker at several workshops and conferences. He has received several best paper awards including: an IEEE Signal Processing Society Best Paper Award (1998), a Best Original Paper Award from the Journal of Flow Cytometry (2008), a Best Magazine Paper Award from the IEEE Signal Processing Society (2010), a SPIE Best Student Paper Award (2011), an IEEE ICASSP Best Student Paper Award (2011), an AISTATS Notable Paper Award (2013), and an IEEE ICIP Best Paper Award (2013). He received an IEEE Signal Processing Society Meritorious Service Award (1998), an IEEE Third Millenium Medal (2000), an IEEE Signal Processing Society Distinguished Lecturership (2002), and an IEEE Signal Processing Society Technical Achievement Award (2014). He was President of the IEEE Signal Processing Society (2006–2007). He was a member of the IEEE TAB Society Review Committee (2009), the IEEE Awards Committee (2010–2011), and served on the Board of Directors of the IEEE (2009–2011) as Director of Division IX (Signals and Applications). He presently serves on the IEEE TAB Nominations and Appointments Committee.