Distributed covariance estimation in Gaussian graphical models

Ami Wiesel
School of Computer Science and Engineering
Hebrew University in Jerusalem, Israel
Email: amiw@cs.huji.ac.il

Alfred O. Hero III
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, Michigan, USA
Email: hero@umich.edu

***Abstract*—We consider distributed covariance estimation in Gaussian graphical models. A typical motivation is learning the potential functions for inference via belief propagation in large scale networks. The classical approach based on a centralized maximum likelihood principle is infeasible, and suboptimal distributed alternatives which tradeoff performance with communication costs are required. We begin with a natural solution where each node performs independent estimation of its local covariance with its neighbors. We show that these local solutions are consistent, and can be interpreted as a pseudo-likelihood method. Based on this interpretation, we propose to enhance the performance by introducing additional symmetry constraints. We enforce these using the methodology of the Alternating Direction Method of Multipliers. This results in a flexible message passing protocol between neighboring nodes which can be implemented in large scale networks.**

I. Introduction

Covariance estimation in Gaussian distributions is a classical and fundamental problem in statistical signal processing. Recent interest in large scale inference using small sample sizes has caused the topic to rise to prominence once again. A natural approach in these settings is to incorporate additional prior knowledge in the form of structure and/or sparsity in order to ensure stable estimation. Graphical models provide a method of representing conditional independence structure using graphs. In the Gaussian case, this structure leads to sparsity in the inverse covariance and allows for efficient implementation of statistical inference algorithms, e.g., belief propagation (BP).

From a distributed signal processing perspective, graphical models are attractive as inference can be preformed as decentralized computations in large scale networks [1], [2]. In many applications, the topology of the network can be associated with a statistical graphical model, and this can be exploited to distribute statistical analysis. For example, Bayesian inference can be implemented using message passing. This has been shown when the underlying graph is a tree and recently in arbitrary topologies [3]. A crucial step in Bayesian inference is learning the model parameters from data. In the Gaussian case, this step corresponds to covariance estimation where the BP potential functions are learned from data.

The classical approach to covariance estimation in Gaussian graphical models (GGM) is based on the maximum likelihood (ML) principle [4], [5]. This approach is consistent and asymptotically optimal in terms of mean squared error (MSE). When the underlying graph is a tree, the ML estimate has a simple closed form solution which requires little communication between nodes. Unfortunately, in arbitrary topologies, finding the ML estimate requires the solution to a difficult high dimensional convex optimization problem. It can be solved using general purpose optimization toolboxes or using the classical iterative proportional fitting (IPF) method [4]. A distributed version of IPF can be achieved by implementing each of its iterations via BP or its variants [6], but this approach can be too computationally intensive. In some scenarios, approximate estimation may be achieved using low rank assumptions [7]. In general, distributed covariance estimation via message passing is still a difficult task and suboptimal approaches are in order.

In this work, we propose alternative distributed estimation methods that approximate global ML estimators by trading high performance for lower communication costs. The most natural distributed approach to covariance estimation is to force each node to perform independent estimation of its local covariance. This simple approach has a closed form solution which requires no message passing. We show that it is consistent, and can be interpreted as a pseudo or composite likelihood method (See [8]–[10] and references within). Based on this interpretation, we propose to improve it by introducing additional symmetry constraints. We enforce these constraints using the methodology of the Alternating Direction Method of Multipliers [11]. This results in a flexible message passing protocol between neighboring nodes.

The outline of the paper is as follows. In Section II we briefly review the basics of GGM and formulate the distributed covariance estimation problem. In Section III, we describe the classical global estimator, derive a competing local estimator and then propose a distributed estimator which provides a promising tradeoff between them. Finally, in Section IV we demonstrate the performance of the estimators using a numerical example.

II. Problem formulation

Traditionally, the multivariate Gaussian distribution is represented by its mean $\boldsymbol{\eta}$ and covariance $\boldsymbol{\Sigma}$. In graphical models, it is more natural to use the canonical parameters $\mathbf{J} = \boldsymbol{\Sigma}^{-1}$ and $\mathbf{h} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ which lead to the following representation

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

978-1-4244-8977-0/10/$26.00 ©2010 IEEE

Graphical models are intuitive characterizations of conditional independence structures within distributions [4]. An undirected graph $\mathcal{G} = (V, E)$ is a set of nodes V connected by undirected edges[1] E. Let $\mathbf{x}$ be a zero mean random vector whose elements are indexed by the nodes in V. The vector $\mathbf{x}$ satisfies the Markov property with respect to $\mathcal{G}$ - if for any pair of non-adjacent nodes the corresponding pair of elements in $\mathbf{x}$ are conditionally independent of the remaining elements

$$p\left(\mathbf{x}_i, \mathbf{x}_j | \mathbf{x}_{V\setminus i,j}\right) = p\left(\mathbf{x}_i | \mathbf{x}_{V\setminus i,j}\right) p\left(\mathbf{x}_j | \mathbf{x}_{V\setminus i,j}\right), \qquad (2)$$

for all $\{i, j\} \notin E$. Applying (2) to the Gaussian distribution in (1), yields

$$[\mathbf{J}]_{i,j} = 0 \qquad \text{for all} \qquad \{i, j\} \notin E. \qquad (3)$$

This property is the core of GGM: the concentration matrix $\mathbf{J}$ has a sparsity pattern which follows the topology of the conditional independence graph. The sparsity allows for efficient inference. For example, simple algebraic manipulations establish that the minimum min squared error (MMSE) estimator reduces to

$$\mathrm{E}\left\{\mathbf{x}_i | \mathbf{x}_{V\setminus i}\right\} = \mathbf{\Sigma}_{i,V\setminus i}\left[\mathbf{\Sigma}_{V\setminus i,V\setminus i}\right]^{-1}\mathbf{x}_{V\setminus i} = -\frac{\mathbf{J}_{i,N_i}\mathbf{x}_{N_i}}{\mathbf{J}_{i,i}}, \quad (4)$$

that depends only on the neighboring nodes N_i. This principle is the main building block of BP methods [3].

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

In this paper, we consider a distributed version of the covariance estimation problem. Specifically, we associate with each of the random variables $\mathbf{x}_i$ a node in a network. We assume that the nodes are physically separated and have communication links to their neighbors in the graph $\mathcal{G} = (V, E)$. Each node i has access only to $\{\mathbf{x}_{[i\ N_i]}[t]\}_{t=1}^{T}$ where N_i is the set of indices corresponding to the neighbors of i. Using this data along with message passing with its neighbors, each node tries to estimate its local covariance information defined by $\mathbf{J}_{i,[i,N_i]}$.

Our definition of local covariance information stems from the fact that $\mathbf{J}_{i,[i,N_i]}$ is sufficient for inference as shown in (4), whereas knowledge of the intuitive local covariance matrix $\mathbf{\Sigma}_{[i,N_i],[i,N_i]}$ involves more parameters and is insufficient. Thus, our framework is that each node tries to estimate its own $\mathbf{J}_{i,[i,N_i]}$. For simplicity, we collect these in one global matrix, namely $\hat{\mathbf{J}}$, whose $\{i, [i\ N_i]\}$ elements are the local estimates in the i'th node. Note that $\hat{\mathbf{J}}_{i,j}$ and $\hat{\mathbf{J}}_{i,j}$ are both estimators of $\mathbf{J}_{i,j} = \mathbf{J}_{j,i}$ but may be different since each is estimated by a different node.

There are different criteria for measuring the performance of the estimator $\hat{\mathbf{J}}$. First, we will use the standard MSE of estimators of $\mathbf{\Sigma} = \mathbf{J}^{-1}$ and of $\mathbf{J}$ based on the Frobenius norm. Second, assuming that the overall goal is to implement the estimator in (4), we will use the Bayesian MSE for evaluating the estimator of $\mathbf{x}_S$ given $\mathbf{x}_{V\setminus S}$ where S is a subset of indices and the true $\mathbf{J}$ is replaced by a plug in estimate $\hat{\mathbf{J}}$.

[1]We use the convention that each node is connected to itself, i.e., $\{i, i\} \in E\ \forall\ i \in V$.

III. Estimators

In this section, we review the computationally intensive global ML approach to covariance estimation in GGM. We describe a local ML approach which is completely decoupled, but does not exploit global information. Finally, we propose a distributed solution which improves the local approach through a flexible increase in computational complexity.

A. Global estimator

The classical covariance estimation method is based on a global ML technique [4]

$$\begin{aligned}\hat{\mathbf{J}}^{global} &= \arg\max_{\mathbf{J}\in\mathcal{J}} \sum_{t=1}^{T} \log p\left(\mathbf{x}\left[t\right]; \mathbf{J}\right) \qquad (5)\\ &= \arg\max_{\mathbf{J}\in\mathcal{J}} \log|\mathbf{J}| - \mathrm{Tr}\left\{\mathbf{S}\mathbf{J}\right\},\end{aligned}$$

where $\mathbf{S}$ is the sample covariance

$$\mathbf{S} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}[t]\mathbf{x}^T[t], \qquad (6)$$

and the feasible set defined by the conditional independence structure is

$$\mathcal{J} = \left\{\mathbf{J} \succeq \mathbf{0}, \quad \mathbf{J}_{i,j} = 0 \quad \{i, j\} \notin E\right\}. \qquad (7)$$

This global estimator is consistent in the sense that $\hat{\mathbf{J}}^{global}$ converges to $\mathbf{J}$ asymptotically in T. It is efficient and its MSE asymptotically achieves the Cramer Rao bound.

The problem in (5) is a convex optimization problem. It can be solved using general purpose convex optimization techniques or special purpose methods such as iterative proportional fitting [4]. However, it is difficult to implement these procedures in a distributed manner, motivating suboptimal approaches.

B. Local estimator

A natural alternative to the global solution is the local ML method. This estimator aggregates p decoupled ML estimators implemented independently at each of the nodes. Locally, each node belongs to a network of $1 + |N_i|$ nodes. Due to properties of the multivariate normal distribution, the marginal distribution of $\mathbf{x}_{[i\ N_i]}$ is also Gaussian with zero mean and an inverse covariance equal to

$$\mathbf{J}^i = \left[\left[\mathbf{J}^{-1}\right]_{[i\ N_i],[i\ N_i]}\right]^{-1}. \qquad (8)$$

In general, $\mathbf{J}^i$ is not equivalent to $\mathbf{J}_{[i\ N_i],[i\ N_i]}$, but using the sparsity and the inversion formula for partitioned matrices, it is easy to show that the elements we are interested in, namely $\{i, [i\ N_i]\}$, are identical. Thus, a natural approach to

distributed covariance estimation is to let each node independently estimate its $\mathbf{J}^i$. Assuming that the nodes are not aware of the conditional independence relations between their neighbors, the local estimators are simply the local sample covariances:

$$\hat{\mathbf{J}}^i = \left(\mathbf{S}_{[i\ N_i],[i\ N_i]}\right)^{-1}, \tag{9}$$

where $\mathbf{S}$ is defined in (6). Together, the local estimator for $\mathbf{J}$ is constructed by placing each of these elements in its global position:

$$\hat{\mathbf{J}}^{local}_{i,[i\ N_i]} = \hat{\mathbf{J}}^i_{i,[i\ N_i]} = \left[\left(\mathbf{S}_{[i\ N_i],[i\ N_i]}\right)^{-1}\right]_{i,[i\ N_i]}. \tag{10}$$

The local estimates are all locally consistent, i.e., $\hat{\mathbf{J}}^i \rightarrow \mathbf{J}^i$ as $T \rightarrow \infty$. Together with the fact that the local and global inverse covariances are identical in their $\{i, [i\ N_i]\}$ positions, we obtain global consistency:

$$\hat{\mathbf{J}}^{local} \rightarrow \mathbf{J} \qquad \text{as} \qquad T \rightarrow \infty. \tag{11}$$

C. Distributed estimator

We now propose a modification to the local estimators which allows for improved performance by message passing. The basic principle behind this improvement is to exploit the symmetry in the inverse covariance. The local estimators do not utilize this inherent property and can therefore be improved. Symmetry can be enforced using simple averaging between neighboring nodes but this will result in suboptimal local likelihood values. Instead, we now propose a rigorous approach using an alternative formulation.

For this purpose, we re-derive $\hat{\mathbf{J}}^{local}$ using global quantities as sub-blocks of $\mathbf{J}$, rather than local quantities as $\mathbf{J}^i$. Consider the conditional distribution of $\mathbf{x}_i$ given $\mathbf{x}_{N_i}$. Using the properties of the Gaussian conditional distribution, $\mathbf{x}_i$ given $\mathbf{x}_{N_i}$ is jointly Gaussian with canonical parameters $-\mathbf{J}_{i,N_i}\mathbf{x}_{N_i}$ and $\mathbf{J}_{i,i}$, and is therefore fully characterized by our parameter of interest $\mathbf{J}_{i,[i,N_i]}$. The i'th conditional ML estimate is the solution to

$$\hat{\mathbf{J}}^{local}_{i,[i\ N_i]} = \arg\max_{\mathbf{J}_{i,[i\ N_i]}} \sum_{t=1}^{T} \log p\left(\mathbf{x}_i[t] | \mathbf{x}_{N_i}[t]; \mathbf{J}_{i,[i\ N_i]}\right), \tag{12}$$

and it can be shown to be identical to the local estimate in (10). Moreover, the p independent conditional ML estimates can be conveniently expressed as a single (yet fully decoupled) optimization problem:

$$\hat{\mathbf{J}}^{local} = \arg\max_{\mathbf{J}} \sum_{i=1}^{p}\sum_{t=1}^{T} \log p\left(\mathbf{x}_i[t] | \mathbf{x}_{N_i}[t]; \mathbf{J}_{i,[i\ N_i]}\right). \tag{13}$$

This formulation is known in the statistical literature as pseudo-likelihood estimation and is known to provide a good tradeoff between performance and computational complexity (See [8]–[10] and references within).

Based on the pseudo-likelihood formulation, we can now rigorously exploit the known symmetry in the inverse covariance and propose the distributed estimator:

$$\hat{\mathbf{J}}^{dist} = \arg\max_{\mathbf{J}_{i,j}=\mathbf{J}_{j,i}} \sum_{i=1}^{p}\sum_{t=1}^{T} \log p\left(\mathbf{x}_i[t] | \mathbf{x}_{N_i}[t]; \mathbf{J}_{i,[i\ N_i]}\right), \tag{14}$$

where we have added simple symmetry conditions $\mathbf{J}_{i,j} = \mathbf{J}_{j,i}$. The objective of (14) is separable and the only coupling constraints which prevent a distributed implementation are the constraints. Following [11], we propose to distribute the solution using the Alternating Direction Method of Multipliers. We define the augmented Lagrangian

$$\mathcal{L}\left(\hat{\mathbf{J}}, \overline{\mathbf{J}}; \mathbf{M}\right) = \sum_{i=1}^{p}\left[\sum_{t=1}^{T} \log p\left(\mathbf{x}_i[t] | \mathbf{x}_{N_i}[t]; \hat{\mathbf{J}}_{i,[i\ N_i]}\right) + \sum_{j\in N_i} \mathbf{M}_{i,j}\left(\hat{\mathbf{J}}_{i,j} - \overline{\mathbf{J}}_{i,j}\right) - \frac{c}{2}\left(\hat{\mathbf{J}}_{i,j} - \overline{\mathbf{J}}_{i,j}\right)^2\right], \tag{15}$$

where $\hat{\mathbf{J}}$ is the estimated inverse covariance, $\overline{\mathbf{J}}$ is a symmetric auxiliary matrix, $\mathbf{M}$ is a matrix of dual multipliers and c is a positive scalar parameter. In standard dual decomposition methods, a saddle point of the augmented Lagrangian $\mathcal{L}\left(\hat{\mathbf{J}}, \overline{\mathbf{J}}; \mathbf{M}\right)$ is found by iteratively solving for the primal variables $\hat{\mathbf{J}}$ and $\overline{\mathbf{J}}$ with fixed dual variables $\mathbf{M}$ and then updating $\mathbf{M}$. Due to the coupling in the quadratic term, solving the primal problem is difficult. Remarkably, the method is guaranteed to converge to the global solution even if we update $\mathbf{M}$ with suboptimal primal values, and use the following iterations:

$$\overline{\mathbf{J}}_{i,j \text{ and } j,i} \leftarrow \frac{\hat{\mathbf{J}}_{i,j} + \hat{\mathbf{J}}_{j,i}}{2} - \frac{\mathbf{M}_{i,j} + \mathbf{M}_{j,i}}{2c}; \tag{16}$$

$$\hat{\mathbf{J}}_{i,[i\ N_i]} \leftarrow \arg\max_{\hat{\mathbf{J}}_{i,[i\ N_i]}} g_i\left(\hat{\mathbf{J}}_{i,[i\ N_i]}\right); \tag{17}$$

$$\mathbf{M}_{i,j} \leftarrow \mathbf{M}_{i,j} + c\left(\hat{\mathbf{J}}_{i,j} - \overline{\mathbf{J}}_{i,j}\right), \tag{18}$$

for all $i \in V$ and $j \in N_i$, where

$$g_i\left(\hat{\mathbf{J}}_{i,[i\ N_i]}\right) = \sum_{t=1}^{T} \log p\left(\mathbf{x}_i[t] | \mathbf{x}_{N_i}[t]; \hat{\mathbf{J}}_{i,[i\ N_i]}\right) + \sum_{j\in N_i} \mathbf{M}_{i,j}\hat{\mathbf{J}}_{i,j} - \frac{c}{2}\hat{\mathbf{J}}^2_{i,j} + c\hat{\mathbf{J}}_{i,j}\overline{\mathbf{J}}_{i,j}. \tag{19}$$

This algorithm involves local computations and message passing through the dual multipliers in $\mathbf{M}$.

The most computationally intensive tasks of this distributed algorithm are the solutions to the local penalized likelihoods in (17). We now show that each of these sub-problems can be reduced to a simple line search. For this purpose, we first solve for $\hat{\mathbf{J}}_{i,N_i}$. Taking the derivative with respect to $\hat{\mathbf{J}}_{i,N_i}$ and equating to zero yields

$$\hat{\mathbf{J}}_{i,N_i} = -\left(\frac{1}{\hat{\mathbf{J}}_{i,i}}\mathbf{S}_{N_i,N_i} + c\mathbf{I}\right)^{-1}\left(\mathbf{S}_{N_i,i} - \mathbf{m}_i - c\overline{\mathbf{J}}_{N_i,i}\right), \tag{20}$$

where $\mathbf{m}_i$ is vector with the neighboring dual variables. Plugging $\hat{\mathbf{J}}_{i,N_i}$ back into the objective yields a line search with respect to $\hat{\mathbf{J}}_{i,i}$:

$$\max_{\hat{\mathbf{J}}_{i,i}} -\hat{\mathbf{J}}_{i,i}\mathbf{S}_{i,i} + \mathbf{s}_i^T \left(\frac{\mathbf{S}_{N_i,N_i}}{\hat{\mathbf{J}}_{i,i}} + c\mathbf{I} \right)^{-1} \mathbf{s}_i + \log \left| \hat{\mathbf{J}}_{i,i} \right|, \quad (21)$$

where $\mathbf{s}_i = \mathbf{S}_{N_i,i} - \mathbf{m}_i - c\overline{\mathbf{J}}_{N_i,i}$. This line search is unimodal since the original problem was jointly convex. Therefore, it can be efficiently solved using a bisection method. In the special case in which $\mathbf{M} = \mathbf{0}$ and $c = 0$, the line search has a simple closed form solution which coincides with $\hat{\mathbf{J}}_{i,i}^{local}$.

Positive definiteness: The distributed estimator exploits the known symmetry in $\mathbf{J}$ but may not produce a positive definite estimator. Performance can be improved by adding this constraint to (14). However, we are not aware of simple distributed methods for enforcing it. Moreover, it is well known that inequality constraints which do not reduce the dimensionality of the unknown parameter do not change the asymptotic performance of M estimators (or the Cramer Rao performance bound [12]). For the case of finite sample size, incorporation of positive definiteness constraints can improve performance and is worthwhile for future study.

Generalizations: Another interesting direction would be to try to improve the summation aggregation of likelihoods. The sum of conditional distributions of $\mathbf{x}_i$, could be replaced by a sum over conditional distributions of $\mathbf{x}_{S_i}$ where S_i are overlapping subsets of indices. The sum could also be replaced by weighted sum according to the number of neighbors each node has or other criteria.

IV. Numerical example

In this section, we demonstrate the performance advantage of the distributed estimator using a numerical example. We simulate a network of $p = 50$ sensors whose locations are uniformly distributed over the unit square. We generate a graph by connecting each sensor (node) to its four nearest neighbors. We then compute the values of the inverse covariance as follows: $\mathbf{J}_{i,j} = 0$ if nodes i and j are connected, and $\mathbf{J}_{i,j} = e^{-0.7d_{i,j}}$ where $d_{i,j}$ is their distance, otherwise. We add an arbitrary value to the diagonal elements in order to guarantee that the matrix is positive definite. After computing $\mathbf{J}$ we keep it fixed throughout all the Monte Carlo simulations. In each of the 500 experiments, we generate independent realizations of $\mathbf{x}$ and estimate $\mathbf{J}$ using the three estimators. The distributed estimator is implemented using the message passing protocol described above with $c = 0.1$ and 5 iterations. In Fig. 1, we report the normalized MSEs in $\mathbf{J}$, $\mathbf{J}^{-1}$ and $\mathbf{x}_{1,\cdots,p/2}$ as detailed in Section II. The performance advantage of the distributed estimator in comparison to the local estimator is evident in all three performance measures. The MSES in $\mathbf{J}$ and in $\mathbf{x}$ show similar behavior, and strengthen our intuition that $\mathbf{J}$, rather than $\mathbf{J}^{-1}$, is the natural parameter required for inference. Examining the iteration count of the distributed estimator reveals that almost all of the performance gain is obtained within the first 2-3 iterations, corresponding to a small amount of message passing.

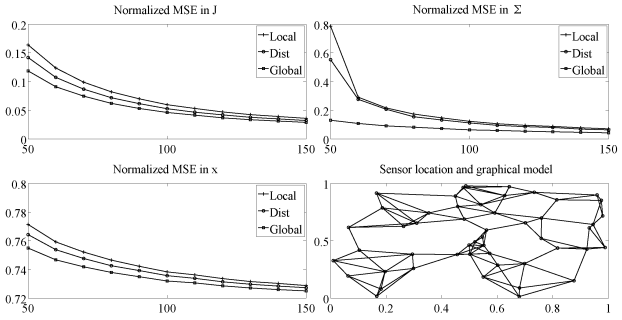

Fig. 1. MSEs as a function of number of samples and the graphical model.

Acknowledgment

A. Wiesel would like to acknowledge fruitful discussions with Venkat Chandrasekaran. The work of A. Wiesel was supported by a Marie Curie Outgoing International Fellowship within the 7th European Community Framework Programme. This research was partially supported by AFOSR grant FA9550-06-1-0324.

References

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

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

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

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

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

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

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

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

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

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

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

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