SAMPLE COMPLEXITY OF JOINT STRUCTURE LEARNING

Saurabh Sihag and Ali Tajer

Electrical, Computer, and Systems Engineering Department
Rensselaer Polytechnic Institute

ABSTRACT

This paper considers the problem of jointly recovering the structures of two graphical models with unknown edge structures. It is assumed that both graphs have the same number of nodes and a known *subset* of nodes have identical structures in both graphs. The classes of Ising models and Gaussian models are considered. For Ising models, the objective is to recover the connectivity of both graphs under an approximate recovery criterion. For Gaussian models, the objectives of edge structure recovery and inverse covariance estimation are considered. Information-theoretic bounds on the sample complexity for bounded probability of error under the aforementioned criteria are established and compared with the corresponding bounds on the sample complexity for recovering the graphs independently.

***Index Terms*—** Graphical models, information-theoretic bounds, joint model selection, structural similarity.

1. INTRODUCTION

Conditional dependence among multiple random variables can be structurally modeled by graphical models, in which the random variables form the nodes of the graphs and their interdependence is captured by the edges among them [1, 2]. Graph-based models have widespread applications in many domains, e.g., computer vision [3], genetics [4–6], social networks [7], and power systems [8]. In this paper, we consider the problem of joint model selection of a pair of graphs with partial structural similarity using samples from their joint distributions, where the focus is on Gaussian and Ising models. The problem of model selection consists of edge structure recovery for Ising and Gaussian models and also inverse covariance matrix estimation for Gaussian models.

Graphical models with partially similar structures are effective for modeling inference problems in various domains such as physical infrastructures [9], biological networks [4], and behavioral analysis [10]. In such domains, the data is generated by multiple networks (modeled by graphs), in which there exists a partial structure common to all networks, while each network has also a unique partial structure. In such applications, the data collected from different graphs has redundancy of information that can be leveraged for jointly analyzing the data from all the relevant models. Motivated by this premise, we analyze algorithm-independent information-theoretic bounds on the sample complexity for joint model selection of a pair of partially similar graphical models that belong to the classes of Ising and Gaussian models. Furthermore, we also analyze inverse covariance estimation for a pair of graphs from the class of Gaussian models.

This research was supported in part by the U. S. National Science Foundation under the CAREER Award ECCS-1554482, the grant ECCS-1455228, and the grant DMS-1737976.

1.1. Related Work

The problem of graphical model selection is feasible under certain restrictions on the graph structure, e.g., sparsity [11–14]. Such restrictions on the graphical models can be analyzed by studying graphs with bounded degree and number of edges. Information-theoretic bounds on the sample complexity for model selection of single graphs in various classes of Ising models and Gaussian models have been studied in [15–18]. In [15] and [16], necessary conditions on the sample complexity for the exact recovery of sub-classes of Ising models are established. In [17], the problem of graphical model selection is investigated for Ising models under the criterion of approximate recovery, i.e., at most a fixed number of errors are tolerated in the estimated graph structure. Necessary conditions for set-based graph model selection, i.e., a set of graphs that potentially contains the true graph is identified by the graph estimator, are characterized in [19].

Joint graphical model inference has been investigated in [4, 5, 10, 20–25] for graphs that may be structurally similar. In [5] and [20–23], optimization techniques are used for joint inference of Gaussian graphical models. In [4] and [24], Bayesian frameworks are developed for joint model inference. The aforementioned papers investigate various empirical frameworks for joint graph inference when different models may be partially similar. In this paper, we focus on characterizing the sample complexity for joint model selection of a pair of graphs with same number of nodes that have the same graph structure within a known cluster of nodes.

2. GRAPH MODEL

Consider a pair of undirected graphs, denoted by $\mathcal{G}_1 \triangleq (V_1, E_1)$ and $\mathcal{G}_2 \triangleq (V_2, E_2)$, where $V_i \triangleq \{1, \ldots, p\}$ and $E_i \subseteq V_i \times V_i$ are the set of p vertices and the set of edges, respectively, in graph $\mathcal{G}_i$, for $i \in \{1, 2\}$. We denote the edge between nodes $u, v \in E_i$, by $E_i^{u,v}$. Each node $j \in V_i$ in graph $\mathcal{G}_i$ is associated with a random variable denoted by X_i^j, and the joint probability density function (pdf) of the random vector $X_i \triangleq [X_i^1, \ldots X_i^p]$ is denoted by $f_i(\cdot)$, for $i \in \{1, 2\}$. In this paper, we call X_i as one graph sample. We collect n graph samples from each graph to perform joint model selection of $\mathcal{G}_1$ and $\mathcal{G}_2$. The collection of n samples from graph $\mathcal{G}_i$ is denoted by $\mathcal{X}_i^n$. Given $\mathcal{X}_1^n$ and $\mathcal{X}_2^n$, the graph decoder

$\hat{\mathcal{G}}(\mathcal{X}_1^n, \mathcal{X}_2^n) \triangleq \{\hat{\mathcal{G}}_1, \hat{\mathcal{G}}_2\}$ provides the estimates $\hat{\mathcal{G}}_1$ and $\hat{\mathcal{G}}_2$ for $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively.

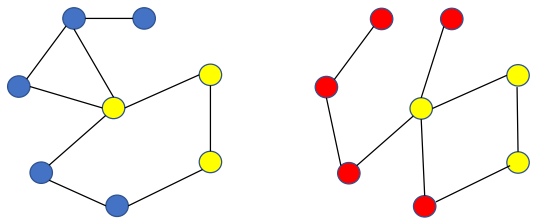

Fig. 1. Two graphs with partially similar structures. Yellow nodes in both graphs have the same internal edge structure.

In this paper, we assume that the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are structurally identical within a pre-specified cluster of nodes, i.e., both $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same internal sub-graph within a set of nodes $V_c \subseteq V_1, V_2$. An example of this setting is illustrated in Fig. 1. We establish information-theoretic bounds on the sample complexity of approximate joint model selection of the two graphs for Ising models, and that for exact joint model selection and inverse covariance matrix estimation for Gaussian models.

3. PROBLEM FORMULATION

We formalize the notation for structurally similar graphs in the following definition.

Definition 1. *A pair of graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ is said to be $\zeta-$similar if both graphs have the same internal graphical structure within a cluster of nodes of size $\lfloor \zeta p \rfloor$, for some $\zeta \in (0,1)$.*

For both $\mathcal{G}_1$ and $\mathcal{G}_2$, the edge structures between any pair of nodes with at least one node not in V_c are assumed to be structurally independent of each other. For graph $\mathcal{G}_i$, the degree of a node $u \in V_i$ is denoted by d_u^i, which captures the number of nodes in the immediate neighborhood of u (i.e., the nodes that are directly connected to u by an edge). Given the family of graphical models that $\mathcal{G}_1$ and $\mathcal{G}_2$ belong to, the pdfs f_1 and f_2 represent the graphical structure of $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively.

3.1. Ising Model

For a graph $\mathcal{G}_i$ in the family of Ising models, each node $u \in V_i$ is associated with a binary random variable $X_i^u \in \{-1, 1\}$. The pdf of the random vector X_i associated with $\mathcal{G}_i$ is given by

$$f_i(X_i) = \frac{1}{Z_i} \exp\left(\sum_{u,v \in V_i} \lambda_i^{uv} X_i^u X_i^v \right) , \qquad (1)$$

where

$$\lambda_i^{uv} \triangleq \begin{cases} \lambda, & \text{if } E_i^{u,v} \in E_i \\ 0, & \text{otherwise} \end{cases} , \qquad (2)$$

for $\lambda > 0$, and Z_i is the partition function given by

$$Z_i \triangleq \sum_{X_i \in \{-1,1\}^p} \exp\left(\sum_{u,v \in V_i} \lambda_i^{uv} X_i^u X_i^v \right) . \qquad (3)$$

Note that the parameter λ in (2) controls the dependence among the nodes in the graph. In [15], it is shown that recovering the graph structure from the data becomes more difficult as λ approaches 0 or grows to infinity.

We denote the family of Ising models by $\mathcal{I}$, and the family of $\zeta-$similar pairs of Ising models by $\mathcal{I}^{\zeta}$. Furthermore, we consider a restricted subclass of Ising models, given by $\mathcal{I}_{d,k,\eta,\gamma}^{\zeta,\theta}$, that consists of pairs of graphs with each graph having at most k number of edges and at most $\lfloor \theta k \rfloor$, for some $\theta \in (0,1)$, edges in the cluster with common structure, each node in the graph having a degree of at most d, and the paths of length γ or less between any two non-connected nodes in the graph can be blocked by blocking at most η number of nodes. For convenience in notations, throughout the paper we use the shorthand $\bar{\mathcal{I}}$ to refer to $\mathcal{I}_{d,k,\eta,\gamma}^{\zeta,\theta}$. Restrictions on the maximum degree and the number of edges in the graph are motivated by practical applications in which the models are sparse. Restriction on the number of edges in the graph in the shared cluster is motivated by the fact that the size of the shared cluster determines the maximum number of edges allowed within it, which may be significantly less than the total number of edges allowed in the graphs. Restrictions on the paths between any two disconnected nodes are of interest as existing literature suggests polynomial time recovery of the graphs in several cases [26].

3.2. Gaussian Model

In this paper, we assume that the mean of the pdf associated with Gaussian model is a zero vector. Hence, for a graph $\mathcal{G}_i$, the joint distribution of X_i with inverse covariance matrix Σ_i is given by

$$f_i(X_i) = \frac{1}{\sqrt{(2\pi)^p \mathsf{det}(\Sigma_i^{-1})}} \exp\left(\frac{1}{2} X_i^{\mathsf{T}} \Sigma_i X_i \right) . \qquad (4)$$

Note that the off diagonal elements of Σ_i reflect the edge structure of the graph $\mathcal{G}_i$, i.e., the element at a coordinate (u,v) in Σ_i, given by $\Sigma_i(u,v)$, is non-zero if and only if $E_i^{u,v} \in E_i$.

We denote the class of Gaussian models by $\mathcal{G}$ and the class of $\zeta-$similar Gaussian graphical models by $\mathcal{G}^{\zeta}$. The recovery of the Gaussian model is contingent upon the matrix elements of the inverse-covariance matrix [18]. Therefore, for a graph $\mathcal{G}_i$, we also define

$$\lambda_i^* \triangleq \min_{u,v \in V_i} \frac{|\, \Sigma_i(u,v) \,|}{\sqrt{\Sigma_i(u,u)\Sigma_i(v,v)}} , \qquad (5)$$

which reflects the scale-invariant minimum value of the matrix Σ_i.

In this paper, we consider the following sub-class of Gaussian model. A pair of graphical models $\mathcal{G}_1$ and $\mathcal{G}_2$ belong to the class $\mathcal{G}_d^{\zeta}(\lambda)$ if and only if they are $\zeta-$similar, have a de-

gree of at most d, and satisfy $\min\{\lambda_1^*, \lambda_2^*\} \geq \lambda$.

3.3. Recovery Criteria

Given the collection of samples $\mathcal{X}_1^n$ and $\mathcal{X}_2^n$, the aim of the graph decoder $\hat{\mathcal{G}}(\mathcal{X}_1^n, \mathcal{X}_2^n)$ is to form estimates for the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$. We first provide the graph recovery criteria for Ising models and Gaussian models, which have been adopted in the existing literature for recovering single graphs [17, 18].

3.3.1. Ising Model

For Ising models, we adopt an *approximate* graph recovery criteria, that is, we tolerate at most a pre-specified number of erroneous decisions, denoted by $q \geq 0$, about the edges in the recovery of each graph. The probability of error in the approximate graph recovery over the class $\bar{\mathcal{I}}$ is defined as

$$\mathsf{P}_{\bar{\mathcal{I}}}^{q} \triangleq \max_{\mathcal{G}_1, \mathcal{G}_2 \in \bar{\mathcal{I}}} \mathbb{P}\left[\min_{i \in \{1,2\}} \{|E_i \Delta \hat{E}_i|\} \geq q\right] , \tag{6}$$

where $|E_i \Delta \hat{E}_i|$ is the edit distance between E_i and $\hat{E}_i$ given by $|E_i \Delta \hat{E}_i| \triangleq |(E_i \backslash \hat{E}_i) \cup (\hat{E}_i \backslash E_i)|$. Note that $|E_i \Delta \hat{E}_i|$ represents the number of modifications to be made in the edge structure to transform $\mathcal{G}_i$ to $\hat{\mathcal{G}}_i$, and q characterizes the mismatch between the estimated and true graphs.

3.3.2. Gaussian Model

As discussed earlier, the Gaussian models are characterized by the inverse covariance matrix and the non-zero off-diagonal elements represent the edges between the nodes. Note that the edge structure can be determined by estimating the support of the inverse covariance matrix (i.e., the non-zero off-diagonal elements). We list the two recovery criteria used for Gaussian models as follows.

1. *Exact Recovery*: Under this criterion, we aim to perform joint model selection for the two graphs to recover the edge structure of both graphs exactly based on the given collection of graph samples, $\mathcal{X}_1^n$ and $\mathcal{X}_2^n$. For Gaussian models, exact recovery is equivalent to estimating the corresponding support sets of the inverse covariance matrices Σ_1 and Σ_2. We define $\mathsf{P}_{\mathcal{G}}$ as the probability of error in exact recovery over the class $\mathcal{G}_d^{\zeta}(\lambda)$, i.e.,

$$\mathsf{P}_{\mathcal{G}} \triangleq \max_{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{G}_d^{\zeta}(\lambda)} \mathbb{P}[\hat{\mathcal{G}}(\mathcal{X}_1^n, \mathcal{X}_2^n) \neq (\mathcal{G}_1, \mathcal{G}_2)] . \tag{7}$$

2. *Inverse Covariance Matrix Estimation:* Under this criterion, the graph decoder aims to estimate the inverse covariance matrices Σ_1 and Σ_2. Note that estimating the numerical values of the inverse covariance matrix and estimating the support set of the inverse covariance matrix are fundamentally different tasks [18]. Define $\hat{\Sigma}_i$ as the estimate of Σ_i. Then, the maximal probability of error in the inverse covariance matrix estimation is defined as

$$\tilde{\mathsf{P}}_{\mathcal{G}}(\delta) \triangleq \max_{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{G}_d^{\zeta}(\lambda)} \mathbb{P}\left[\min_{i \in \{1,2\}} \|\hat{\Sigma}_i - \Sigma_i\|_{\infty} < \delta/2\right] , \tag{8}$$

where $\|\cdot\|_{\infty}$ denotes the ℓ_{∞}-norm.

3.4. Comparison with Existing Works

When graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are recovered jointly under the aforementioned criteria, a total of $2n$ graph samples are used, for which we establish performance guarantees on the probability of error for the corresponding criterion. Let n_{s} be the number of samples necessary for recovering a single graph under the different criteria with the same performance guarantees. Such sample complexities have been analyzed in [17] and [18]. To analyze the difference between the two settings, we define $D \triangleq 2(n_{\mathrm{s}} - n)$.

4. MAIN RESULTS

In this section, we provide the necessary conditions on the sample size n for any graph decoder to recover a pair of structurally similar graphs, and compare them with the existing results for single graphs. Note that the necessary conditions provided in this paper are algorithm-independent and therefore, provide benchmarks for the sample complexity analysis of any designed algorithm.

4.1. Ising Models

We provide the results for approximate recovery of $\zeta-$similar graphs in the class $\bar{\mathcal{I}}$, and discuss the scaling behavior of the gap between the results in this section and the existing results for single graphs. To describe the results for this setting, we denote the binary entropy function by

$$h(\theta) \triangleq -\theta \log \theta - (1-\theta)\log(1-\theta), \text{ for } \theta \in (0,1) . \tag{9}$$

Theorem 1 (Class $\bar{\mathcal{I}}$ with $k \leq p/4$)**.** *Consider a pair of $\zeta-$similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the class $\bar{\mathcal{I}}$ with parameters $k \leq p/4$ and $\eta \leq \lfloor \frac{d}{2} \rfloor$. For any graph decoder $\hat{\mathcal{G}} : \mathcal{X}_1^n \times \mathcal{X}_2^n \to \bar{\mathcal{I}}$ that tolerates the distortion $q = \lfloor \beta \frac{\lfloor (c\eta-1)^2 k}{2c\eta(2\eta+m(\gamma+1))} \rfloor$ for some $m \in \{0, \ldots, \lfloor \frac{d}{2} \rfloor - \eta\}$, $\beta \in (0, \frac{1}{2})$ and $c \in (1/\eta, 1]$, and achieves $\mathsf{P}_{\bar{\mathcal{I}}}^{q} \leq \delta$, the sample complexity satisfies*

$$n \geq \max\{A_1, A_2\}(1 - \delta - o(1)) , \tag{10}$$

where we have defined

$$A_1 \triangleq \frac{2(1-\theta)k\log(p(1-\zeta)) + k\theta\log(\zeta p) - 2q\log p}{k\lambda\tanh\lambda} , \tag{11}$$

$$A_2 \triangleq \frac{\left(1 + (\cosh 2\lambda)^{(1-c)\eta - 1}\left(\frac{1+(\tanh\lambda)^{\gamma+1}}{1-(\tanh\lambda)^{\gamma+1}}\right)^m\right)}{2\lambda c\eta} \times ((1-\zeta/2)\log 2 - h(\beta)) . \tag{12}$$

To gain more insight from Theorem 1, we note that the terms A_1 and A_2 have different scaling behavior in λ, p, and d. Therefore, it is imperative to explore different regimes of variartion of parameters that characterize the sample complexity.

1. $\lambda = \omega(\min\{1/\sqrt{\eta}, 1/m^{\frac{1}{\gamma+1}}\}$: In this regime, we can verify that the term A_2 grows exponentially in $\lambda^2\eta$ and $\lambda^{\gamma+1}m$, and dominates the sample complexity. By comparing A_2 with the corresponding result for a single graph, we observe that D scales exponentially in $\lambda^2\eta$ and $\lambda^{\gamma+1}m$.

2. $\lambda = O(\min\{1/\sqrt{\eta}, 1/m^{\frac{1}{\gamma+1}}\}$: In this regime, the term A_1 dominates the sample complexity and as $\lambda \to 0$, $\tanh\lambda$ scales according to $O(\lambda)$ and therefore, A_1 scales according to $\Omega(\max\{\eta, m^{\frac{2}{\gamma+1}}\}\log p)$. Also, by comparing with the corresponding result for a single graph, we conclude that D scales at the same rate as the sample complexity, i.e., $\Omega(\max\{\eta, m^{\frac{2}{\gamma+1}}\}\log p)$.

4.2. Gaussian Models

In this section, we consider the class of Gaussian graphical models and provide information-theoretic bounds on the sample size n for recovering $\zeta-$similar graphs.

4.2.1. Exact Recovery

Theorem 2 (Class $\mathcal{G}_d^\zeta(\lambda)$)**.** *For a pair of $\zeta-$similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{G}_d^\zeta(\lambda)$ with $\lambda \in [0, \frac{1}{2}]$, for any graph decoder $\hat{\mathcal{G}} : \mathcal{X}_1^n \times \mathcal{X}_2^n \to \mathcal{G}_d^\zeta(\lambda)$ that achieves $\mathsf{P}_\mathcal{G} \xrightarrow{n\to\infty} 0$, the sample size n for each graph satisfies*

$$n > \max\{B_1, B_2\}(1 - \delta - o(1)) \,, \tag{13}$$

where we have defined

$$B_1 \triangleq \max\left\{\frac{\log\binom{\zeta p - d}{2} - 1}{8\lambda^2}, \frac{\log\binom{p-\zeta p-d}{2} - 1}{4\lambda^2}\right\} \,, \tag{14}$$

$$B_2 \triangleq \max\left\{\frac{\log\binom{\zeta p}{d} - 1}{\log\left(1 + \frac{d\lambda}{1-\lambda}\right) - \frac{d\lambda}{1+(d-1)\lambda}}, \frac{\log\binom{p-\zeta p}{d} - 1}{\frac{1}{2}\log\left(1 + \frac{d\lambda}{1-\lambda}\right) - \frac{d\lambda}{1+(d-1)\lambda}}\right\} . \tag{15}$$

We compare the results in Theorem 2 with that in [18]. Note that the terms B_1 and B_2 have different scaling behavior with respect to λ, and therefore it is imperative to characterize the scaling behavior of all the terms to characterize the gain in sample complexity under joint model selection over model selection of a single graph.

1. $\lambda = \Theta(\frac{1}{d})$: In this regime, the sum of edge weights in the neighborhood of every node remains bounded. Also, B_1 and B_2 scale differently with λ. The terms B_1 scale according to $\Omega(d^2\log(\max\{\zeta, 1-\zeta\}p - d))$. Also, the first term in B_1 is dominant when $\zeta \to 1$. The corresponding result for a single graph from [18] is $\frac{1}{4\lambda^2}\left(\log\binom{p-d}{2} - 1\right)$. Therefore, under the regime when B_1 dominates the sample complexity, we have $D = \Omega(d^2\log(\max\{\zeta, 1-\zeta\}p - d))$.

2. $\lambda = O(1)$ and $\lambda \in [0, 1/2]$: In this regime, the sample complexity is dominated by B_2 which scales as $\Omega\left(\frac{d\log(p/d)}{\log(1+d\lambda)}\right)$. By comparison with the corresponding result for recovering a single graph, we have $D = \Omega\left(\frac{d\log(p/d)}{\log(1+d\lambda)}\right)$ in this regime.

4.2.2. Inverse Covariance Matrix Recovery

Theorem 3. *For a pair of $\zeta-$similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{G}_d^\zeta(\lambda)$, if there exists a graph decoder such that $\tilde{\mathsf{P}}_\mathcal{G} \leq 1/2$, then the sample complexity satisfies*

$$n > \max\left\{\frac{\log\left(\frac{\zeta pd}{4}\right) - 2}{8\delta^2}, \frac{\log\left(\frac{(1-\zeta)pd}{4}\right) - 2}{4\delta^2}\right\} . \tag{16}$$

The bound in (16) captures the sample complexity corresponding to inverse covariance matrix estimation of $\zeta-$similar pair of graphs. In [18], the corresponding result for a single graph is $\frac{\log\left(\frac{pd}{4}\right)-2}{4\delta^2}$. It can be shown that $D = \frac{1}{4\delta^2}\log\left(\frac{pd}{4\zeta}\right)$ when $\frac{1}{8\delta^2}\left(\log\left(\frac{\zeta pd}{4}\right) - 2\right)$ dominates, i.e., the shared cluster of the graph pair consists of more edges than the non shared cluster. When $\frac{1}{4\delta^2}\left(\log\left(\frac{(1-\zeta)pd}{4}\right) - 2\right)$ dominates, $D = \frac{1}{2\delta^2}\log(1-\zeta)$. Clearly, for fixed ζ, the variation scales with δ and becomes significantly large as $\delta \to 0$. However, when the shared cluster is denser compared to the non-shared cluster of the graph pair, the bound on the sample complexity scales with $\Omega(d^2\log(pd))$ when $\delta = O(1/d)$. Also, D scales according to $\Omega(d^2)$ when $\delta = O(1/d)$.

5. CONCLUSION

In this paper, we have analyzed the problem of joint model selection of partially similar graphical models in the path-restricted, edge and degree bounded sub-class of Ising models and the degree bounded sub-class of Gaussian models. For Ising models, we have characterized the information-theoretic bounds on the sample complexity for approximate recovery of the graph structures. For Gaussian models, we have established the information-theoretic bounds on the sample complexity for exact recovery of the graph structures and inverse covariance matrix estimation. We have also investigated the scaling behavior of the difference between the sample complexity for joint model selection and that for single graphs from the results in existing literature.

6. REFERENCES

[1] S. L. Lauritzen, *Graphical models*. Clarendon Press, May 1996, vol. 17.

[2] J. Pearl, *Causality: models, reasoning, and inference*. Oxford: Cambridge University Press, 2009.

[3] C. S. Won and H. Derin, "Unsupervised segmentation of noisy and textured images using Markov random fields," *CVGIP: Graphical models and image processing*, vol. 54, no. 4, pp. 308–328, 1992.

[4] X. Chen, F. J. Slack, and H. Zhao, "Joint analysis of expression profiles from multiple cancers improves the identification of microRNA–gene interactions," *Bioinformatics*, vol. 29, no. 17, pp. 2137–2145, 2013.

[5] J. Fang, L. S. Dongdong, S. Charles, Z. Xu, V. D. Calhoun, and Y.-P. Wang, "Joint sparse canonical correlation analysis for detecting differential imaging genetics modules," *Bioinformatics*, vol. 32, no. 15, pp. 3480–3488, 2016.

[6] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, "Sparse graphical models for exploring gene expression data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 196–212, 2004.

[7] Y. Jacob, L. Denoyer, and P. Gallinari, "Learning latent representations of nodes for classifying in heterogeneous social networks," in *Proc. ACM international Conference on Web Search and Data Mining*, New York, Feb. 2014, pp. 373–382.

[8] K. Dvijotham, M. Chertkov, P. V. Hentenryck, M. Vuffray, and S. Misra, "Graphical models for optimal power flow," *Constraints*, vol. 22, no. 1, pp. 24–49, 2017.

[9] M. Rabbat, R. Nowak, and M. Coates, "Network tomography and the identification of shared infrastructure," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2002, pp. 34–38.

[10] J. Guo, J. Cheng, E. Levina, G. Michailidis, and J. Zhu, "Estimating heterogeneous graphical models for discrete data with an application to roll call voting," *The Annals of Applied Statistics*, vol. 9, no. 2, pp. 821 – 848, Jun. 2015.

[11] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

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

[13] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *The Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, Jun. 2010.

[14] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *Journal of Machine learning research*, vol. 9, pp. 485–516, Jun. 2008.

[15] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions." *IEEE Trans. Information Theory*, vol. 58, no. 7, pp. 4117–4134, May 2012.

[16] R. Tandon, K. Shanmugam, P. K. Ravikumar, and A. G. Dimakis, "On the information theoretic limits of learning Ising models," in *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2014, pp. 2303–2311.

[17] J. Scarlett and V. Cevher, "On the difficulty of selecting Ising models with approximate recovery," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 625–638, Dec. 2016.

[18] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *Proc. IEEE International Symposium on Information Theory*, Austin, Texas, Jun. 2010.

[19] D. Vats and J. M. Moura, "Necessary conditions for consistent set-based graphical model selection," in *Proc. IEEE International Symposium on Information Theory*, Saint-Petersburg, Russia, Jul. 2011, pp. 303–307.

[20] J. Guo, E. Levina, G. Michailidis, and J. Zhu, "Joint estimation of multiple graphical models," *Biometrika*, vol. 98, no. 1, pp. 1–15, 2011.

[21] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, Mar. 2014.

[22] K. Mohan, P. London, M. Fazel, D. Witten, and S.-I. Lee, "Node-based learning of multiple Gaussian graphical models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 445–488, 2014.

[23] S. Yang, Z. Lu, X. Shen, P. Wonka, and J. Ye, "Fused multiple graphical lasso," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 916–943, 2015.

[24] C. B. Peterson, F. C. Stingo, and M. Vannucci, "Bayesian inference of multiple Gaussian graphical models," *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 159–174, 2015.

[25] H. Qiu, F. Han, H. Liu, and B. Caffo, "Joint estimation of multiple graphical models from high-dimensional time series," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 2, pp. 487–504, 2016.

[26] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky, "High-dimensional structure estimation in Ising models: Local separation criterion," *The Annals of Statistics*, vol. 40, no. 3, pp. 1346–1375, Jun. 2012.