LEARNING GAUSSIAN GRAPHICAL MODELS WITH DIFFERING PAIRWISE SAMPLE SIZES

Lili Zheng and Genevera I. Allen[†]

Department of Electrical and Computer Engineering
Rice University, USA

ABSTRACT

High-dimensional Gaussian graphical models have been a powerful tool for learning connections or interaction patterns among a large number of variables. While most prior work focuses on the case when all variables are measured simultaneously, we consider a practically common but new setting when no simultaneous measurement of all variables is available and there are differing numbers of measurements on each pair of nodes. This occurs in estimating gene expression networks from single-cell sequencing data, functional connectivity from neuronal recordings, and sensor networks. Incorporating a projection step in the estimation of the covariance matrix, we develop a convex method for learning the neighborhood of any node. We provide theoretical guarantees for the proposed method, suggesting that the exact recovery of the neighborhood of any node a hinges on sufficient joint measurements for the pairs of nodes that involve a's neighbor(s). We also present some consequences of the theory and numerical experiments on special graphs.

***Index Terms*—** Graphical Model, Neighborhood Recovery, Missing Data, Non-uniform Measurements, Sample Complexity

1. INTRODUCTION

Over the past decades, Gaussian graphical models have been extensively used and studied in a number of applications for its capability of capturing conditional independence relationships, such as learning gene expression networks [1, 2], functional connectivity in neuroscience [3, 4, 5], etc. Despite the abundant literature, most works address this problem assuming all variables (e.g., expression of all genes, fluorescence traces of all neurons) are measured simultaneously, which is unlikely in many applications. For instance, consider the frequent dropouts in single-cell sequencing data [6], multiple sessions of neural recording or neural imaging where one needs to stitch together the graph [7], and sensor networks where the simultaneous measurements of a large number of sensors can be extremely expensive [8, 9]. Furthermore, the number of joint measurements of different pairs/collections of variables may also differ: for instance, neurons that are spatially closer to each other would be likely measured together more than neurons far away from each other [7].

Here arise two significant open questions: (i) How should we estimate a sparse Gaussian graphical model with differing pairwise sample sizes? This is related to some prior work on covariance or graphical model estimation with missing data or size-constrained measurements [9, 10, 11, 8], while our problem setting is more general in that there is nearly no constraint on the missing pattern. (ii) Moreover, if it is of particular interest to accurately estimate a certain part of the graph, e.g., the neighborhood of a particular gene, can we tolerate low sample sizes for some pairs of variables? These two questions are non-trivial: most existing methods for Gaussian graphical model hinge on the fact that the sample covariance can serve as a good estimate for the full covariance matrix, and hence leads to the exact recovery of the graph [12]; while under our setting the sample covariance does not exist; one can only come up with an estimate for each pairwise covariance separately, which can lead to a non-positive semi-definite estimate of the full covariance, bringing *optimization challenges*; the estimation errors for different entries of the covariance would also differ due to varying sample sizes, presenting *challenges in statistical analysis*.

In this work, we address these two open questions by first proposing a feasible convex method and then providing theoretical guarantees to characterize the conditions for successful recovery of part of the graph. Our theoretical results have interesting implications in graph auditing: it can help understand whether certain estimated edges can be trusted or not given varying pairwise sample sizes. It may also provide insights for developing active learning algorithms for graphical models.

1.1. Contributions

Our main contributions are summarized as follows:

- We propose a neighborhood selection algorithm suited for

*The authors acknowledge support from NSF DMS-1554821, NSF NeuroNex-1707400, and NIH 1R01GM140468.

[†]This author also has joint affiliations with the Department of Statistics, Department of Computer Science at Rice University, and the Jan and Dan Duncan Neurological Research Institute at Baylor College of Medicine.

our problem setting. In particular, our method is built upon the idea of neighborhood lasso [13] and a novel weighted ℓ_∞-norm projection step for estimating the covariance matrix, where the latter addresses the challenge brought by the non-simultaneous measurements of different nodes: an entry-wise estimate for the covariance matrix may not be positive semi-definite under this setting.

- We provide theoretical guarantees for the successful recovery of certain neighborhoods with only sample size requirements on certain pairs of nodes. Our work is the first attempt to characterize the conditions for successful recovery of particular neighborhoods without assuming all node pairs to be measured accurately.
- We also perform numerical experiments on chain graphs and block graphs for validating the theoretical results.

2. PROBLEM SETTING

Notations: For any matrix $A \in \mathbb{R}^{p_1 \times p_2}$, we define $\|A\|_\infty = \max_{i,j} |A_{ij}|$, $|||A|||_\infty = \sup_{u \in \mathbb{R}^{p_2}, \|u\|_\infty = 1} \|Au\|_\infty$.

Consider a sparse Gaussian graphical model:

$$x \sim \mathcal{N}(0, \Sigma^*), \quad \Theta^* = (\Sigma^*)^{-1}, \tag{1}$$

where $\Theta^* \in \mathbb{R}^{p \times p}$ is the sparse precision matrix. The graph structure is dictated by the nonzero patterns in Θ^*:

$$\mathcal{G} = (V, E), \quad V = [p], \quad E = \{(i,j) : \Theta^*_{ij} \neq 0\}.$$

We aim to estimate the graph structure from the following observations: For each pair of nodes (i,j), we observe $n_{ij} > 0$ independent samples $\{(x_i^{(k)}, x_j^{(k)})\}_{k \in O_{ij}}$ with index set O_{ij} of size n_{ij}. Due to the symmetry, $O_{ij} = O_{ji}$ and $n_{ij} = n_{ji}$. We also define $O_{ii} = \cup_{j \neq i} O_{ij}$ and $n_{ii} = |O_{ii}|$. Other than sizes, we do not impose any constraints on O_{ij}: O_{i_1,j_1} and O_{i_2,j_2} can have arbitrary overlap for any i_1, i_2, j_1, j_2.

2.1. Estimator

To estimate the graph $\mathcal{G}$, we consider an algorithm based on the idea of neighborhood lasso [13]. Different from other graphical model estimation methods such as graphical lasso, the neighborhood lasso algorithm is separable w.r.t. the neighborhoods of different nodes, and hence provides us with some convenience in analysis.

One key quantity required by our algorithm is a finite sample estimate of the true covariance matrix Σ^*. We first define a naive estimate $\widehat{\Sigma}$ as follows: for each entry (i,j), given observations $\{(x_i^{(k)}, x_j^{(k)})\}_{k \in O_{ij}}$, $\widehat{\Sigma}_{ij} = \frac{1}{n_{ij}} \sum_{k \in O_{ij}} x_i^{(k)} x_j^{(k)}$. However, one drawback of $\widehat{\Sigma}$ is its possibly negative eigenvalues, which could lead to both optimization and statistical issues in our algorithm. To ensure non-negative eigenvalues and preserve the entry-wise error bounds for $\widehat{\Sigma}_{ij} - \Sigma^*_{ij}$, we project $\widehat{\Sigma}$ upon the positive semi-definite cone w.r.t. the weighted ℓ_∞ norm:

$$\widetilde{\Sigma} = \arg\min_{\Sigma \succ 0} \max_{i,j} \sqrt{n_{ij}} |\Sigma_{ij} - \widehat{\Sigma}_{ij}|. \tag{2}$$

(2) is a variant of the ℓ_∞ norm projection, the strategy previously exploited to address another non-convex problem in [14]. The projection problem (2) can be solved by the ADMM. The following proposition suggests that $\widetilde{\Sigma}$ enjoys similar entry-wise error bounds to $\widehat{\Sigma}$.

Proposition 1 (Entry-wise error bounds for $\widetilde{\Sigma}$)**.** *With probability at least $1 - p^{-c}$, for $1 \leq i, j \leq p$,*

$$|\widetilde{\Sigma}_{ij} - \Sigma^*_{ij}| \leq C \|\Sigma^*\|_\infty \sqrt{\frac{\log p}{n_{ij}}}.$$

Neighborhood Estimation: Given $\widetilde{\Sigma}$, to recover the neighborhood of any node $a \in [p]$, we perform neighborhood selection for a using a variant of the Lasso:

$$\widehat{\theta}^{(a)} = \arg\min_{\theta \in \mathbb{R}^p, \theta_a = 0} \frac{1}{2} \theta^\top \widetilde{\Sigma} \theta - \widetilde{\Sigma}_{a,:} \theta + \sum_{j=1}^{p} \lambda_j |\theta_j|, \tag{3}$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)^\top \in \mathbb{R}^p$ is a vector of tuning parameters, with each entry λ_j corresponding to a potential edge connecting node j and a. The solution $\widehat{\theta}^{(a)}$ serves as an estimate for

$$\theta^{(a)*} = \arg\min_{\theta \in \mathbb{R}^p, \theta_a = 0} \frac{1}{2} \theta^\top \Sigma^* \theta - \Sigma^*_{a,:} \theta,$$

which satisfies $\theta^{(a)*}_a = 0$ and $\theta^{(a)*}_{\setminus a} = (\Sigma^*_{\setminus a, \setminus a})^{-1} \Sigma^*_{\setminus a, a}$. The support set of $\theta^{(a)*}$ equals the true neighborhood of node a, and thus we estimate it by the support set of $\widehat{\theta}^{(a)}$. We will show the support recovery consistency of this estimator in Section 3.1, with sample size requirements on certain pairs of nodes.

Aggregating Neighborhood Estimates: If one is only interested in the neighborhood of a particular node a, then the estimator (3) suffices; However, when people are interested in the neighborhoods of multiple nodes or even the full graph, it is desired to aggregate the neighborhood estimates of all nodes together: if node i is an estimated neighbor of node j, then j should also be an estimated neighbor of i. Two typical solutions that have been used in the literature are the AND rule and OR rule [13]. Under the AND rule, we estimate estimate the edge set E by

$$\widehat{E}^{\text{AND}} = \{(i,j) : \widehat{\theta}^{(i)}_j \neq 0 \text{ and } \widehat{\theta}^{(j)}_i \neq 0\}; \tag{4}$$

while under the OR rule, the estimated edge set is

$$\widehat{E}^{\text{OR}} = \{(i,j) : \widehat{\theta}^{(i)}_j \neq 0 \text{ or } \widehat{\theta}^{(j)}_i \neq 0\}. \tag{5}$$

We will discuss theoretical consequences of these two rules in Section 3 as well.

3. THEORETICAL GUARANTEES

In this section, we first focus on the estimator (3), establishing conditions for the successful recovery of node a's neighborhood using the support of $\widehat{\theta}^{(a)}$. Section 3.2 discusses some implications of this result subsequently.

3.1. Main Result: Neighborhood Selection Consistency

Let $\mathcal{N}_a$ be the true neighborhood of node a: $\mathcal{N}_a = \{j \neq a : \Theta^*_{ja} \neq 0\}$, and d_a be its degree: $d_a = |\mathcal{N}_a|$. Also define $\overline{\mathcal{N}}_a = \{a\} \cup \mathcal{N}_a$. We first state the following mutual incoherence assumption that has been widely used in the literature for Lasso [15]:

Assumption 1 (Mutual incoherence condition)**.**

$$\left|\left|\left|\Sigma^*_{(\overline{\mathcal{N}}_a)^c,\mathcal{N}_a}(\Sigma^*_{\mathcal{N}_a,\mathcal{N}_a})^{-1}\right|\right|\right|_\infty \leq 1-\gamma, \tag{6}$$

for some $0 < \gamma \leq 1$.

As will be revealed by Theorem 1, our sample size requirements for certain pair of nodes depend on the following quantities:

$$\kappa_1^{(a)} = \|\theta^{(a)*}\|_1, \quad \kappa_2^{(a)} = \left|\left|\left|(\Sigma^*_{\mathcal{N}_a,\mathcal{N}_a})^{-1}\right|\right|\right|_\infty,$$
$$\kappa_3^{(a)} = \left|\left|\left|\Sigma^*_{(\overline{\mathcal{N}}_a)^c,\mathcal{N}_a}\right|\right|\right|_\infty, \quad \theta^{(a)}_{\min} = \min_{\theta^*_j \neq 0} |\theta^*_j|.$$

Smaller $\kappa_i^{(a)}$'s and larger $\theta^{(a)}_{\min}$ would be favorable. For ease of notation, we will omit the superscript of $\kappa_i^{(a)}$ in the following.

Theorem 1. *Consider the model setting described in Section 2 and the estimator* $\widehat{\theta}^{(a)}$ *defined in* (3)*. Assume that the pairwise sample size* $n_{ij} \geq 1$ *for all pairs* $1 \leq i \neq j \leq p$ *and Assumption 1 holds. If each entry of the tuning parameter* $\lambda \in \mathbb{R}^p$ *satisfies*

$$\lambda_j = \max\{C_1 \frac{\kappa_1 + 1}{\gamma}\|\Sigma^*\|_\infty \max_{i \neq j}\sqrt{\frac{\log p}{n_{ij}}}, \frac{\theta^{(a)*}_{\min}}{(4+\gamma)\kappa_2}\},$$

and for $i \in \mathcal{N}_a$, $j \neq i$,

$$n_{ij} \geq \frac{C_2\|\Sigma^*\|^2_\infty \kappa_2^2}{\gamma^2}\left[(\kappa_3^2+1)d_a^2 + \frac{(\gamma+4)^2(\kappa_1+1)^2}{(\theta^{(a)}_{\min})^2}\right]\log p, \tag{7}$$

then $\{j : \widehat{\theta}^{(a)}_j \neq 0\} = \mathcal{N}_a$ *with probability at least* $1 - p^{-c}$ *for some absolute constants* $c, C_1, C_2 > 0$.

Remark 1. *Theorem 1 suggests that, even when many pairs of nodes are only measured few times together, as long as the tuning parameters are chosen carefully w.r.t. the pairwise sample sizes, we are still able to recover the neighborhood of node* a*. In particular, we only need to collect sufficient samples for a pair of nodes if at least one node in this pair is a neighbor of* a*. This is not a trivial result, since the estimator* (3) *has to perform a selection from its neighbors and a large number of non-neighbors which are not measured well.*

Remark 2. *Similar to the proof of the support recovery of the Lasso [15], our proof of Theorem 1 is built on a primal-dual witness argument but requires some careful analysis for differing pairwise sample sizes. For the pairs of nodes that involve neighbors of* a*, the pairwise sample size requirement is* $n_{ij} = \Omega((d_a^2 + (\theta^{(a)}_{\min})^{-2})\log p)$*, stronger than the established result (*$\Omega((d_a + (\theta^{(a)}_{\min})^{-2})\log p)$*) for Lasso [15]. This difference is due to the lack of simultaneous measurements of all nodes.*

3.2. Consequences for Aggregated Graph Estimate and Special Examples

As discussed in Section 2.1, combining the neighborhood estimates for different nodes is necessary when multiple neighborhoods or the full graph is of interest. Here we present some corollaries of Theorem 1 that illustrate the consequences of aggregating neighborhood estimates using the AND/OR rule.

Corollary 1 (Guarantee for any specific edge)**.** *Consider the model setting detailed in Section 2 and an arbitrary node pair* (a, b) *for* $1 \leq a \neq b \leq p$*. Assume* $n_{ij} \geq 1$ *for all* $i \neq j$*, Assumption 1 holds for both node* a *and* b*, and* λ *is as specified in Theorem 1. Then as long as one of the following two sample size requirements is satisfied: (i)* $n_{ij} = \Omega(\max\{d_a^2, \beta^{(a)*}_{\min}\}\log p)$ *for* $i \in \mathcal{N}_a$, $1 \leq j \leq p$*; (ii)* $n_{ij} = \Omega(\max\{d_b^2, \beta^{(b)*}_{\min}\}\log p)$ *for* $i \in \mathcal{N}_b$, $1 \leq j \leq p$*, we have the following:*

- *If we estimate the edge set by* $\widehat{E}^{\mathrm{AND}}$ *defined in* (4)*, then* $(i,j) \notin \widehat{E}^{\mathrm{AND}}$ *if* $(i,j) \notin E$ *(no false positive);*
- *If we estimate the edge set by* $\widehat{E}^{\mathrm{OR}}$ *defined in* (5)*, then* $(i,j) \in \widehat{E}^{\mathrm{OR}}$ *if* $(i,j) \in E$ *(no false negative).*

Corollary 1 suggests that by employing the AND (OR) rule, to ensure no false positive (false negative) of one particular non-edge (edge), the sample size requirements can be weaker if measuring the neighbors of the node with smaller degree. To understand further the implication of Theorem 1 and Corollary 1, we consider three special graphs in the following corollary.

Corollary 2 (Pairwise Measurements for Particular Graphs)**.** *Consider the model described in Section 2 and the edge set estimates* $\widehat{E}^{\mathrm{AND}}$ *and* $\widehat{E}^{\mathrm{OR}}$ *defined in* (4) *and* (5)*. If Assumption 1 holds for all nodes,* λ *is as specified in Theorem 1, then with appropriately designed measurements strategy, we need the following amount of pairwise measurements (*$\sum_{i<j} n_{ij}$*) to ensure exact recovery for neighborhoods in certain graphs:*

1. *For the neighborhood of* $k > 0$ *consecutive nodes of a chain graph, exact recovery by* $\widehat{E}^{\mathrm{AND}}$ *requires* $\Omega(kp\log p)$ *pairwise measurements.*
2. *For a block of size* k *in a block graph, exact recovery by* $\widehat{E}^{\mathrm{AND}}$ *requires* $\Omega(k^3 p\log p)$ *pairwise measurements.*

3. *For the full star graph, exact recovery by $\widehat{E}^{\mathrm{OR}}$ requires $\Omega(p \log p)$ pairwise measurements; exact recovery by $\widehat{E}^{\mathrm{AND}}$ requires $\Omega(p^4 \log p)$ pairwise measurements.*

Remark 3. *Corollary 2 suggests that the AND rule behaves drastically different from the OR rule for the star graph. This is due to that the neighborhood selection for the hub node is extremely hard due to its high degree $(p-1)$, but is much easier for the spokes. The OR rule $\widehat{E}^{\mathrm{OR}}$ trusts the edges selected by all the spokes and hence works well. While for the AND rule, many edges selected by the spokes might not be selected by the hub node and do not enter the edge set estimate $\widehat{E}^{\mathrm{AND}}$.*

Remark 4. *Here we provide a comparison between Corollary 2 and the implications of previous theoretical guarantees on neighborhood selection. When all nodes are measured simultaneously and we directly apply the original neighborhood selection algorithm [13], then the exact recovery of the graph requires $\Omega(d_{\max} p^2 \log p)$ pairwise measurements, which takes the value of $\Omega(p^2 \log p)$ for chain graph; $\Omega(kp^2 \log p)$ for block graph with block sizes bounded by k; $\Omega(p^3 \log p)$ for star graph. These values can be significantly larger than those in Corollary 2 if we are only interested in the neighborhoods of a few nodes; and also for the full graph recovery of the star graph with the OR rule.*

4. NUMERICAL EXPERIMENTS

In this section, we conduct some numerical experiments to validate our theoretical results. Specifically, we consider chain graphs and block graphs with number of nodes $p = 50$ and $p = 100$; the graph structure dictates the non-zero patterns of Θ^*, and the non-zero entries are generated from uniform distribution $\mathcal{U}(0.6, 0.8)$. We set the diagonal entries of Θ^* to ensure $\lambda_{\min}(\Theta^*) \geq 0.5$, avoiding singularity. For each graph, we consider the exact recovery of the neighborhoods of certain nodes, and assign the pairwise sample sizes accordingly in the following three cases. For each of the three cases, we define a set $\mathcal{M}$ of node pairs that are required to be measured well by our theoretical results, and for $i \neq j$, we set the pairwise sample size $n_{ij} = 1$ for $(i,j) \notin \mathcal{M}$, $n_{ij} = N$ for $(i,j) \in \mathcal{M}$ with various values of N.

1. For a chain graph, consider the exact recovery of the neighborhood of a node a with degree 1 (an endpoint); let $\mathcal{M} = \{(i,j) : i \in \mathcal{N}_a \text{ or } j \in \mathcal{N}_a\}$.
2. For the chain graph, we also consider the exact recovery of the neighborhood of a node b with degree 2. In this case, let $\mathcal{M} = \{(i,j) : i \in \mathcal{N}_b \text{ or } j \in \mathcal{N}_b\}$.
3. For a block graph with block size 2, consider the exact recovery of the neighborhoods of two nodes a and b in the same block, when the AND rule is applied to obtain a full graph estimate. Let $\mathcal{M} = \{(i,j) : i = a, b, \text{ or } j = a, b\}$.

Given the specified pairwise sample sizes $n_{ij}, i < j$, we generate i.i.d. samples x_k, $k = 1, \ldots, \max_{i<j} n_{ij}$ from (1), and then for each node pair (i,j), we randomly select n_{ij} samples to obtain $\{(x_i^{(k)}, x_j^{(k)}\}_{k \in O_{ij}}$. As discussed in Section 2, O_{ii} is constructed as $\cup_{j \neq i} O_{ij}$ and $n_{ii} = |O_{ii}|$. Figure 1 shows the probability of successful recovery of the target neighborhoods under the three settings described above, where each point is averaged over 50 trials. We can see that successful recovery of certain neighborhoods is indeed possible as long as appropriate node pairs ($(i,j) \in \mathcal{M}$) are measured well. In addition, the comparison between the second and the first settings in Figure 1 suggests that more samples are needed to recover the neighborhoods of nodes with higher degrees. The solid lines and dashed lines are close to each other, validating our theoretical scaling of pairwise sample sizes w.r.t. p.

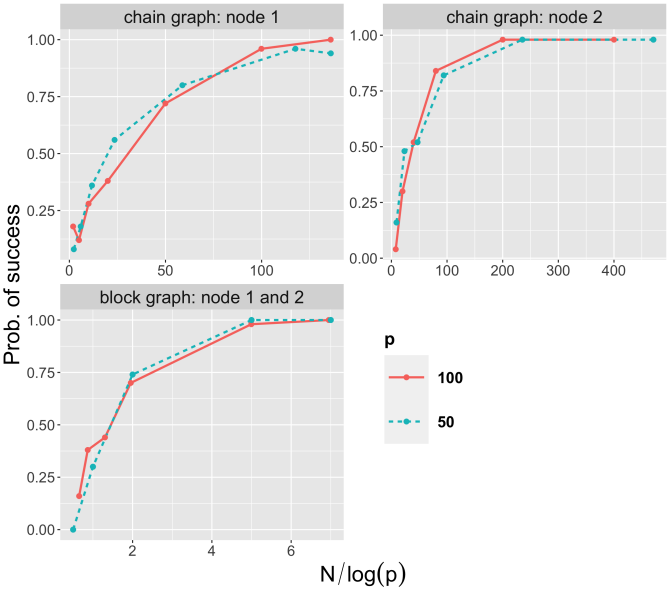

Fig. 1. Probability of successful recovery for the neighborhoods of (i) node with degree 1 in chain graphs, (ii) node with degree 2 in chain graphs and (iii) two nodes in the same block of block graphs.

5. CONCLUSION

This work considers learning Gaussian graphical models from differing number of joint measurements of each pair. We propose a convex method for learning the neighborhood of any given node, which can be further combined by the AND/OR rule to obtain a full graph estimate. The proposed method is proved to achieve exact recovery of the neighborhood of any node a, as long as sufficient measurements are obtained for any pairs of nodes that involve a's neighbors. We also validate our theory with numerical experiments on some special graphs. Our theoretical results can help with graph auditing, and may also provide insights for designing sampling strategy, or developing new active learning algorithms.

References

[1] Genevera I Allen and Zhandong Liu, "A local poisson graphical model for inferring networks from sequencing data," *IEEE transactions on nanobioscience*, vol. 12, no. 3, pp. 189–198, 2013.

[2] Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R Nevins, Guang Yao, and Mike West, "Sparse graphical models for exploring gene expression data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 196–212, 2004.

[3] Minjie Wang and Genevera I Allen, "Thresholded graphical lasso adjusts for latent variables: Application to functional neural connectivity," *arXiv preprint arXiv:2104.06389*, 2021.

[4] Andersen Chang, Tianyi Yao, and Genevera I Allen, "Graphical models and dynamic latent factors for modeling functional brain connectivity," in *2019 IEEE Data Science Workshop (DSW)*. IEEE, 2019, pp. 57–63.

[5] Dimitri Yatsenko, Krešimir Josić, Alexander S Ecker, Emmanouil Froudarakis, R James Cotton, and Andreas S Tolias, "Improved estimation and interpretation of correlations in neural circuits," *PLoS computational biology*, vol. 11, no. 3, pp. e1004083, 2015.

[6] Luqin Gan, Giuseppe Vinci, and Genevera I Allen, "Correlation imputation in single cell rna-seq using auxiliary information and ensemble learning," in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2020, pp. 1–6.

[7] Giuseppe Vinci, Gautam Dasarathy, and Genevera I Allen, "Graph quilting: graphical model selection from partially observed covariances," *arXiv preprint arXiv:1912.05573*, 2019.

[8] Gautamd Dasarathy, Aarti Singh, Maria-Florina Balcan, and Jong H Park, "Active learning algorithms for graphical model selection," in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 1356–1364.

[9] Jonathan Scarlett and Volkan Cevher, "Lower bounds on active learning for graphical model selection," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 55–64.

[10] Eduardo Pavez and Antonio Ortega, "Active covariance estimation by random sub-sampling of variables," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4034–4038.

[11] Gautam Dasarathy, "Gaussian graphical model selection from size constrained measurements," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 1302–1306.

[12] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[13] Nicolai Meinshausen and Peter Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[14] Abhirup Datta and Hui Zou, "Cocolasso for high-dimensional error-in-variables regression," *The Annals of Statistics*, vol. 45, no. 6, pp. 2400–2426, 2017.

[15] Martin J Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using $\backslash\ell_1$-constrained quadratic programming (lasso)," *IEEE transactions on information theory*, vol. 55, no. 5, pp. 2183–2202, 2009.