Journal of Machine Learning Research 13 (2012) 2293-2337 Submitted 7/11; Revised 4/12; Published 8/12

High-Dimensional Gaussian Graphical Model Selection: Walk Summability and Local Separation Criterion

Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
Electrical Engineering and Computer Science
University of California, Irvine
Irvine, CA 92697

Vincent Y. F. Tan TANYFV@I2R.A-STAR.EDU.SG
Data Mining Department
Institute for Infocomm Research
Singapore
Electrical and Computer Engineering, National University of Singapore

Furong Huang FURONGH@UCI.EDU
Electrical Engineering and Computer Science
University of California, Irvine
Irvine, CA 92697

Alan S. Willsky WILLSKY@MIT.EDU
Stochastic Systems Group
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Editor: Martin Wainwright

Abstract

We consider the problem of high-dimensional Gaussian graphical model selection. We identify a set of graphs for which an efficient estimation algorithm exists, and this algorithm is based on thresholding of empirical conditional covariances. Under a set of transparent conditions, we establish structural consistency (or *sparsistency*) for the proposed algorithm, when the number of samples $n = \Omega(J_{\min}^{-2} \log p)$, where p is the number of variables and $J_{\min}$ is the minimum (absolute) edge potential of the graphical model. The sufficient conditions for sparsistency are based on the notion of *walk-summability* of the model and the presence of sparse *local vertex separators* in the underlying graph. We also derive novel non-asymptotic necessary conditions on the number of samples required for sparsistency.

Keywords: Gaussian graphical model selection, high-dimensional learning, local-separation property, walk-summability, necessary conditions for model selection

1. Introduction

Probabilistic graphical models offer a powerful formalism for representing high-dimensional distributions succinctly. In an undirected graphical model, the conditional independence relationships among the variables are represented in the form of an undirected graph. Learning graphical models using its observed samples is an important task, and involves both structure and parameter estima-

tion. While there are many techniques for parameter estimation (e.g., expectation maximization), structure estimation is arguably more challenging. High-dimensional structure estimation is NP-hard for general models (Karger and Srebro, 2001; Bogdanov et al., 2008) and moreover, the number of samples available for learning is typically much smaller than the number of dimensions (or variables).

The complexity of structure estimation depends crucially on the underlying graph structure. Chow and Liu (1968) established that structure estimation in tree models reduces to a maximum weight spanning tree problem and is thus computationally efficient. However, a general characterization of graph families for which structure estimation is tractable has so far been lacking. In this paper, we present such a characterization based on the so-called *local separation* property in graphs. It turns out that a wide variety of (random) graphs satisfy this property (with probability tending to one) including large girth graphs, the Erdős-Rényi random graphs (Bollobás, 1985) and the power-law graphs (Chung and Lu, 2006), as well as graphs with short cycles such as the small-world graphs (Watts and Strogatz, 1998) and other hybrid/augmented graphs (Chung and Lu, 2006, Ch. 12). The small world and augmented graphs are especially relevant for modeling data from social networks. Note that these graphs can simultaneously possess many short cycles as well as large node degrees (growing with the number of nodes), and thus, we can incorporate a wide class of graphs for high-dimensional estimation.

Successful structure estimation also relies on certain assumptions on the parameters of the model, and these assumptions are tied to the specific algorithm employed. For instance, for convex-relaxation approaches (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2011), the assumptions are based on certain *incoherence* conditions on the model, which are hard to interpret as well as verify in general. In this paper, we present a set of transparent conditions for Gaussian graphical model selection based on *walk-sum* analysis (Malioutov et al., 2006). Walk-sum analysis has been previously employed to analyze the performance of loopy belief propagation (LBP) and its variants in Gaussian graphical models. In this paper, we demonstrate that walk-summability also turns out to be a natural criterion for efficient structure estimation, thereby reinforcing its importance in characterizing the tractability of Gaussian graphical models.

1.1 Summary of Results

Our main contributions in this work are threefold. We propose a simple local algorithm for Gaussian graphical model selection, termed as conditional covariance threshold test (CMIT) based on a set of conditional covariance thresholding tests. Second, we derive sample complexity results for our algorithm to achieve structural consistency (or sparsistency). Third, we prove a novel non-asymptotic lower bound on the sample complexity required by any learning algorithm to succeed. We now elaborate on these contributions.

Our structure learning procedure is known as the Conditional Covariance Test[1] (CMIT) and is outlined in Algorithm 1. Let $\mathsf{CMIT}(\mathbf{x}^n; \xi_{n,p}, \eta)$ be the output edge set from CMIT given n i.i.d. samples $\mathbf{x}^n$, a threshold $\xi_{n,p}$ (that depends on both p and n) and a constant $\eta \in \mathbb{N}$, which is related to the local vertex separation property (described later). The conditional covariance test proceeds

1. An analogous test is employed for Ising model selection in Anandkumar et al. (2012b) based on conditional mutual information. We later note that conditional mutual information test has slightly worse sample complexity for learning Gaussian models.

Algorithm 1 Algorithm $\mathsf{CMIT}(\mathbf{x}^n; \xi_{n,p}, \eta)$ for structure learning using samples $\mathbf{x}^n$.

Initialize $\widehat{G}_p^n = (V, \emptyset)$.
For each $i, j \in V$, if

$$\min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} |\widehat{\Sigma}(i,j|S)| > \xi_{n,p}, \tag{1}$$

then add (i,j) to $\widehat{G}_p^n$.
Output: $\widehat{G}_p^n$.

in the following manner. First, the empirical absolute conditional covariances[2] are computed as follows:

$$\widehat{\Sigma}(i,j|S) := \widehat{\Sigma}(i,j) - \widehat{\Sigma}(i,S)\widehat{\Sigma}^{-1}(S,S)\widehat{\Sigma}(S,j),$$

where $\widehat{\Sigma}(\cdot,\cdot)$ are the respective empirical variances. Note that $\widehat{\Sigma}^{-1}(S,S)$ exists when the number of samples satisfies $n > |S|$ (which is the regime under consideration). The conditional covariance is thus computed for each node pair $(i,j) \in V^2$ and the conditioning set which achieves the minimum is found, over all subsets of cardinality at most η; if the minimum value exceeds the threshold $\xi_{n,p}$, then the node pair is declared as an edge. See Algorithm 1 for details.

The computational complexity of the algorithm is $O(p^{\eta+2})$, which is efficient for small η. For the so-called *walk-summable* Gaussian graphical models, the parameter η can be interpreted as an upper bound on the size of local vertex separators in the underlying graph. Many graph families have small η and as such, are amenable to computationally efficient structure estimation by our algorithm. These include Erdős-Rényi random graphs, power-law graphs and small-world graphs, as discussed previously.

We establish that the proposed algorithm has a sample complexity of $n = \Omega(J_{\min}^{-2} \log p)$, where p is the number of nodes (variables) and $J_{\min}$ is the minimum (absolute) edge potential in the model. As expected, the sample complexity improves when $J_{\min}$ is large, that is, the model has strong edge potentials. However, as we shall see, $J_{\min}$ cannot be arbitrarily large for the model to be walk-summable. We derive the minimum sample complexity for various graph families and show that this minimum is attained when $J_{\min}$ takes the maximum possible value.

We also develop novel techniques to obtain necessary conditions for consistent structure estimation of Erdős-Rényi random graphs and other ensembles with non-uniform distribution of graphs. We obtain non-asymptotic bounds on the number of samples n in terms of the expected degree and the number of nodes of the model. The techniques employed are information-theoretic in nature (Cover and Thomas, 2006). We cast the learning problem as a source-coding problem and develop necessary conditions which combine the use of Fano's inequality with the so-called asymptotic equipartition property.

Our sufficient conditions for structural consistency are based on walk-summability. This characterization is novel to the best of our knowledge. Previously, walk-summable models have been extensively studied in the context of inference in Gaussian graphical models. As a by-product of our analysis, we also establish the correctness of loopy belief propagation for walk-summable Gaussian graphical models Markov on locally tree-like graphs (see Section 5 for details). This suggests

2. Alternatively, conditional independence can be tested via sample partial correlations which can be computed via regression or recursion. See Kalisch and Bühlmann (2007) for details.

that walk-summability is a fundamental criterion for tractable learning and inference in Gaussian graphical models.

1.2 Related Work

Given that structure learning of general graphical models is NP-hard (Karger and Srebro, 2001; Bogdanov et al., 2008), the focus has been on characterizing classes of models on which learning is tractable. The seminal work of Chow and Liu (1968) provided an efficient implementation of maximum-likelihood structure estimation for tree models via a maximum weighted spanning tree algorithm. Error-exponent analysis of the Chow-Liu algorithm was studied (Tan et al., 2011a, 2010) and extensions to general forest models were considered by Tan et al. (2011b) and Liu et al. (2011). Learning trees with latent (hidden) variables (Choi et al., 2011) have also been studied recently.

For graphical models Markov on general graphs, alternative approaches are required for structure estimation. A recent paradigm for structure estimation is based on convex relaxation, where an estimate is obtained via convex optimization which incorporates an ℓ_1-based penalty term to encourage sparsity. For Gaussian graphical models, such approaches have been considered in Meinshausen and Bühlmann (2006) and Ravikumar et al. (2011) and d'Aspremont et al. (2008), and the sample complexity of the proposed algorithms have been analyzed. A major disadvantage in using convex-relaxation methods is that the incoherence conditions required for consistent estimation are hard to interpret and it is not straightforward to characterize the class of models satisfying these conditions.

An alternative to the convex-relaxation approach is the use of simple greedy local algorithms for structure learning. The conditions required for consistent estimation are typically more transparent, albeit somewhat restrictive. Bresler et al. (2008) propose an algorithm for structure learning of general graphical models Markov on bounded-degree graphs, based on a series of conditional-independence tests. Abbeel et al. (2006) propose an algorithm, similar in spirit, for learning factor graphs with bounded degree. Spirtes and Meek (1995), Cheng et al. (2002), Kalisch and Bühlmann (2007) and Xie and Geng (2008) propose conditional-independence tests for learning Bayesian networks on directed acyclic graphs (DAG). Netrapalli et al. (2010) proposed a faster greedy algorithm, based on conditional entropy, for graphs with large girth and bounded degree. However, all the works (Bresler et al., 2008; Abbeel et al., 2006; Spirtes and Meek, 1995; Cheng et al., 2002; Netrapalli et al., 2010) require the maximum degree in the graph to be bounded ($\Delta = O(1)$) which is restrictive. We allow for graphs where the maximum degree can grow with the number of nodes. Moreover, we establish a natural tradeoff between the maximum degree and other parameters of the graph (e.g., girth) required for consistent structure estimation.

Necessary conditions for consistent graphical model selection provide a lower bound on sample complexity and have been explored before by Santhanam and Wainwright (2008) and Wang et al. (2010). These works consider graphs drawn uniformly from the class of bounded degree graphs and establish that $n = \Omega(\Delta^k \log p)$ samples are required for consistent structure estimation, in an p-node graph with maximum degree Δ, where k is typically a small positive integer. However, a direct application of these methods yield poor lower bounds if the ensemble of graphs has a highly non-uniform distribution. This is the case with the ensemble of Erdős-Rényi random graphs (Bollobás, 1985). Necessary conditions for structure estimation of Erdős-Rényi random graphs were derived for Ising models by Anandkumar et al. (2012b) based on an information-theoretic covering argument. However, this approach is not directly applicable to the Gaussian setting. We present a novel approach for obtaining necessary conditions for Gaussian graphical model selection based on

the notion of *typicality*. We characterize the set of typical graphs for the Erdős-Rényi ensemble and derive a modified form of Fano's inequality and obtain a non-asymptotic lower bound on sample complexity involving the average degree and the number of nodes.

We briefly also point to a large body of work on high-dimensional covariance selection under different notions of sparsity. Note that the assumption of a Gaussian graphical model Markov on a sparse graph is one such formulation. Other notions of sparsity include Gaussian models with sparse covariance matrices, or having a banded Cholesky factorization. Also, note that many works consider covariance estimation instead of selection and in general, estimation guarantees can be obtained under less stringent conditions. See Lam and Fan (2009), Rothman et al. (2008), Huang et al. (2006) and Bickel and Levina (2008) for details.

1.3 Paper Outline

The paper is organized as follows. We introduce the system model in Section 2. We prove the main result of our paper regarding the structural consistency of conditional covariance thresholding test in Section 3. We prove necessary conditions for model selection in Section 4. In Section 5, we analyze the performance of loopy belief propagation in Gaussian graphical models. Section 7 concludes the paper. Proofs and additional discussion are provided in the appendix.

2. Preliminaries and System Model

We now provide an overview of Gaussian graphical models and the problem of structure learning given samples from the model.

2.1 Gaussian Graphical Models

A Gaussian graphical model is a family of jointly Gaussian distributions which factor in accordance to a given graph. Given a graph $G = (V,E)$, with $V = \{1,\ldots,p\}$, consider a vector of Gaussian random variables $\mathbf{X} = [X_1, X_2, \ldots, X_p]^T$, where each node $i \in V$ is associated with a scalar Gaussian random variable X_i. A Gaussian graphical model Markov on G has a probability density function (pdf) that may be parameterized as

$$f_{\mathbf{X}}(\mathbf{x}) \propto \exp\left[-\frac{1}{2}\mathbf{x}^T\mathbf{J}_G\mathbf{x} + \mathbf{h}^T\mathbf{x}\right], \tag{2}$$

where $\mathbf{J}_G$ is a positive-definite symmetric matrix whose sparsity pattern corresponds to that of the graph G. More precisely,

$$J_G(i,j) = 0 \iff (i,j) \notin G.$$

The matrix $\mathbf{J}_G$ is known as the potential or information matrix, the non-zero entries $J(i,j)$ as the edge potentials, and the vector $\mathbf{h}$ as the potential vector. A model is said to be *attractive* if $J_{i,j} \leq 0$ for all $i \neq j$. The form of parameterization in (2) is known as the information form and is related to the standard mean-covariance parameterization of the Gaussian distribution as

$$\mu = \mathbf{J}^{-1}\mathbf{h}, \quad \mathbf{\Sigma} = \mathbf{J}^{-1},$$

where $\mu := \mathbb{E}[\mathbf{X}]$ is the mean vector and $\mathbf{\Sigma} := \mathbb{E}[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T]$ is the covariance matrix.

We say that a jointly Gaussian random vector $\mathbf{X}$ with joint pdf $f(\mathbf{x})$ satisfies local Markov property with respect to a graph G if

$$f(x_i|\mathbf{x}_{\mathcal{N}(i)}) = f(x_i|\mathbf{x}_{V\setminus i})$$

holds for all nodes $i \in V$, where $\mathcal{N}(i)$ denotes the set of neighbors of node $i \in V$ and, $V \setminus i$ denotes the set of all nodes excluding i. More generally, we say that $\mathbf{X}$ satisfies the global Markov property, if for all disjoint sets $A, B \subset V$, we have

$$f(\mathbf{x}_A, \mathbf{x}_B|\mathbf{x}_S) = f(\mathbf{x}_A|\mathbf{x}_S) f(\mathbf{x}_B|\mathbf{x}_S).$$

where set S is a *separator*[3] of A and B The local and global Markov properties are equivalent for non-degenerate Gaussian distributions (Lauritzen, 1996).

Our results on structure learning depend on the precision matrix $\mathbf{J}$. Let

$$J_{\min} := \min_{(i,j)\in G} |J(i,j)|,\ J_{\max} := \max_{(i,j)\in G} |J(i,j)|,\ D_{\min} := \min_i J(i,i).$$

Intuitively, models with edge potentials which are "too small" or "too large" are harder to learn than those with comparable potentials. Since we consider the high-dimensional case where the number of variables p grows, we allow the bounds $J_{\min}$, $J_{\max}$, and $D_{\min}$ to potentially scale with p.

The *partial correlation coefficient* between variables X_i and X_j, for $i \neq j$, measures their conditional covariance given all other variables. These are computed by normalizing the off-diagonal values of the information matrix, that is,

$$R(i,j) := \frac{\Sigma(i,j|V\setminus\{i,j\})}{\sqrt{\Sigma(i,i|V\setminus\{i,j\})\Sigma(j,j|V\setminus\{i,j\})}} = -\frac{J(i,j)}{\sqrt{J(i,i)J(j,j)}}. \tag{3}$$

For all $i \in V$, set $R(i,i) = 0$. We henceforth refer to $\mathbf{R}$ as the partial correlation matrix.

An important sub-class of Gaussian graphical models of the form in (19) are the *walk-summable* models (Malioutov et al., 2006). A Gaussian model is said to be α-walk summable if

$$\|\overline{\mathbf{R}}\| \leq \alpha < 1,$$

where $\overline{\mathbf{R}} := [|R(i,j)|]$ denotes the entry-wise absolute value of the partial correlation matrix $\mathbf{R}$ and $\|\cdot\|$ denotes the spectral or 2-norm of the matrix, which for symmetric matrices, is given by the maximum absolute eigenvalue.

In other words, walk-summability means that an attractive model formed by taking the absolute values of the partial correlation matrix of the Gaussian graphical model is also valid (i.e., the corresponding potential matrix is positive definite). This immediately implies that attractive models form a sub-class of walk-summable models. For detailed discussion on walk-summability, see Section A.1.

2.2 Tractable Graph Families

We consider the class of Gaussian graphical models Markov on a graph G_p belonging to some ensemble $\mathcal{G}(p)$ of graphs with p nodes. We consider the high-dimensional learning regime, where both

3. A set $S \subset V$ is a separator for sets A and B if the removal of nodes in S partitions A and B into distinct components.

p and the number of samples n grow simultaneously; typically, the growth of p is much faster than that of n. We emphasize that in our formulation the graph ensemble $\mathcal{G}(p)$ can either be deterministic or random—in the latter, we also specify a probability measure over the set of graphs in $\mathcal{G}(p)$. In the setting where $\mathcal{G}(p)$ is a random-graph ensemble, let $P_{\mathbf{X},G}$ denote the joint probability distribution of the variables $\mathbf{X}$ and the graph $G \sim \mathcal{G}(p)$, and let $f_{\mathbf{X}|G}$ denote the conditional (Gaussian) density of the variables Markov on the given graph G. Let P_G denote the probability distribution of graph G drawn from a random ensemble $\mathcal{G}(p)$. We use the term *almost every* (a.e.) graph G satisfies a certain property $\mathcal{Q}$ if

$$\lim_{p\to\infty} P_G[G \text{ satisfies } \mathcal{Q}] = 1.$$

In other words, the property $\mathcal{Q}$ holds asymptotically almost surely[4] (a.a.s.) with respect to the random-graph ensemble $\mathcal{G}(p)$. Our conditions and theoretical guarantees will be based on this notion for random graph ensembles. Intuitively, this means that graphs that have a vanishing probability of occurrence as $p \to \infty$ are ignored.

We now characterize the ensemble of graphs amenable for consistent structure estimation under our formulation. To this end, we define the concept of local separation in graphs. See Fig. 1 for an illustration. For $\gamma \in \mathbb{N}$, let $B_\gamma(i;G)$ denote the set of vertices within distance γ from i with respect to graph G. Let $H_{\gamma,i} := G(B_\gamma(i))$ denote the subgraph of G spanned by $B_\gamma(i;G)$, but in addition, we retain the nodes not in $B_\gamma(i)$ (and remove the corresponding edges). Thus, the number of vertices in $H_{\gamma,i}$ is p.

Definition 1 (γ-Local Separator) *Given a graph G, a* γ-local separator $S_\gamma(i,j)$ *between i and j, for $(i,j) \notin G$, is a* minimal *vertex separator*[5] *with respect to the subgraph $H_{\gamma,i}$. In addition, the parameter γ is referred to as the* path threshold *for local separation.*

In other words, the γ-local separator $S_\gamma(i,j)$ separates nodes i and j with respect to paths in G of length at most γ. We now characterize the ensemble of graphs based on the size of local separators.

Definition 2 ((η,γ)-Local Separation Property) *An ensemble of graphs satisfies (η,γ)-local separation property if for a.e. G_p in the ensemble,*

$$\max_{(i,j)\notin G_p} |S_\gamma(i,j)| \leq \eta. \tag{4}$$

We denote such a graph ensemble by $\mathcal{G}(p;\eta,\gamma)$.

In Section 3, we propose an efficient algorithm for graphical model selection when the underlying graph belongs to a graph ensemble $\mathcal{G}(p;\eta,\gamma)$ with sparse local separators (i.e., small η, for η defined in (4)). We will see that the computational complexity of our proposed algorithm scales as $O(p^{\eta+2})$. We now provide examples of several graph families satisfying (4).

4. Note that the term a.a.s. does not apply to deterministic graph ensembles $\mathcal{G}(p)$ where no randomness is assumed, and in this setting, we assume that the property $\mathcal{Q}$ holds for every graph in the ensemble.
5. A minimal separator is a separator of smallest cardinality.

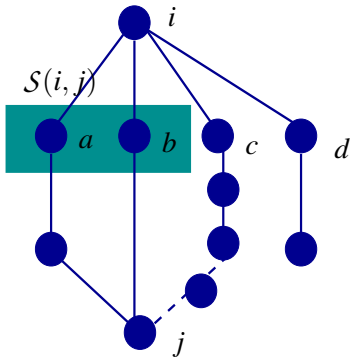


Figure 1: Illustration of l-local separator set $\mathcal{S}(i,j;G,l)$ for the graph shown above with $l=4$. Note that $\mathcal{N}(i)=\{a,b,c,d\}$ is the neighborhood of i and the l-local separator set $\mathcal{S}(i,j;G,l)=\{a,b\}\subset\mathcal{N}(i;G)$. This is because the path along c connecting i and j has a length greater than l and hence node $c\notin\mathcal{S}(i,j;G,l)$.

2.2.1 EXAMPLE 1: BOUNDED-DEGREE

We now show that the local-separation property holds for a rich class of graphs. Any (deterministic or random) ensemble of degree-bounded graphs $\mathcal{G}_{\text{Deg}}(p,\Delta)$ satisfies (η,γ)-local separation property with $\eta=\Delta$ and arbitrary $\gamma\in\mathbb{N}$. If we do not impose any further constraints on $\mathcal{G}_{\text{Deg}}$, the computational complexity of our proposed algorithm scales as $O(p^{\Delta+2})$ (see also Bresler et al., 2008 where the computational complexity is comparable). Thus, when Δ is large, our proposed algorithm and the one in Bresler et al. (2008) are computationally intensive. Our goal in this paper is to relax the usual bounded-degree assumption and to consider ensembles of graphs $\mathcal{G}(p)$ whose maximum degrees may grow with the number of nodes p. To this end, we discuss other structural constraints which can lead to graphs with sparse local separators.

2.2.2 EXAMPLE 2: BOUNDED LOCAL PATHS

Another sufficient condition[6] for the (η,γ)-local separation property in Definition 2 to hold is that there are at most η paths of length at most γ in G between any two nodes (henceforth, termed as the (η,γ)-*local paths property*). In other words, there are at most $\eta-1$ number of overlapping[7] cycles of length smaller than 2γ.

In particular, a special case of the local-paths property described above is the so-called girth property. The *girth* of a graph is the length of the shortest cycle. Thus, a graph with girth g satisfies (η,γ)-local separation property with $\eta=1$ and $\gamma=g/2$. Let $\mathcal{G}_{\text{Girth}}(p;g)$ denote the ensemble of graphs with girth at most g. There are many graph constructions which lead to large girth. For example, the bipartite Ramanujan graph (Chung, 1997, p. 107) and the random Cayley graphs (Gamburd et al., 2009) have large girths.

6. For any graph satisfying (η,γ)-local separation property, the number of vertex-disjoint paths of length at most γ between any two non-neighbors is bounded above by η, by appealing to Menger's theorem for bounded path lengths (Lovász et al., 1978). However, the property of local paths that we describe above is a stronger notion than having sparse local separators and we consider all distinct paths of length at most γ and not just vertex disjoint paths in the formulation.
7. Two cycles are said to overlap if they have common vertices.

The girth condition can be weakened to allow for a small number of short cycles, while not allowing for typical node neighborhoods to contain short cycles. Such graphs are termed as *locally tree-like*. For instance, the ensemble of Erdős-Rényi graphs $\mathcal{G}_{\mathrm{ER}}(p, c/p)$, where an edge between any node pair appears with a probability c/p, independent of other node pairs, is locally tree-like. The parameter c may grow with p, albeit at a controlled rate for tractable structure learning. We make this more precise in Example 3 in Section 3.1. The proof of the following result may be found in Anandkumar et al. (2012a).

Proposition 3 (Random Graphs are Locally Tree-Like) *The ensemble of Erdős-Rényi graphs* $\mathcal{G}_{\mathrm{ER}}(p, c/p)$ *satisfies the* (η, γ)*-local separation property in* (4) *with*

$$\eta = 2,\ \gamma \leq \frac{\log p}{4 \log c}. \tag{5}$$

Thus, there are at most two paths of length smaller than γ between any two nodes in Erdős-Rényi graphs a.a.s, or equivalently, there are no overlapping cycles of length smaller than 2γ a.a.s. Similar observations apply for the more general *scale-free* or *power-law* graphs (Chung and Lu, 2006; Dommers et al., 2010). Along similar lines, the ensemble of Δ-random regular graphs, denoted by $\mathcal{G}_{\mathrm{Reg}}(p, \Delta)$, which is the uniform ensemble of regular graphs with degree Δ has no overlapping cycles of length at most $\Theta(\log_{\Delta-1} p)$ a.a.s. (McKay et al., 2004, Lemma 1).

2.2.3 EXAMPLE 3: SMALL-WORLD GRAPHS

The previous two examples showed local separation holds under two different conditions: bounded maximum degree and bounded number of local paths. The former class of graphs can have short cycles but the maximum degree needs to be constant, while the latter class of graphs can have a large maximum degree but the number of overlapping short cycles needs to be small. We now provide instances which incorporate both these features: large degrees and short cycles, and yet satisfy the local separation property.

The class of hybrid graphs or augmented graphs (Chung and Lu, 2006, Ch. 12) consists of graphs which are the union of two graphs: a "local" graph having short cycles and a "global" graph having small average distances. Since the hybrid graph is the union of these local and global graphs, it has both large degrees and short cycles. The simplest model $\mathcal{G}_{\mathrm{Watts}}(p, d, c/p)$, first studied by Watts and Strogatz (1998), consists of the union of a d-dimensional grid and an Erdős-Rényi random graph with parameter c. It is easily seen that a.e. graph $G \sim \mathcal{G}_{\mathrm{Watts}}(p, d, c/p)$ satisfies (η, γ)-local separation property in (4), with

$$\eta = d + 2,\ \gamma \leq \frac{\log p}{4 \log c}.$$

Similar observations apply for more general hybrid graphs studied in Chung and Lu (2006, Ch. 12).

Thus, we see that a wide range of graphs satisfy the property of having sparse local separators, and that it is possible for graphs with large degrees as well as many short cycles to have this property.

2.2.4 COUNTER-EXAMPLE: DENSE GRAPHS

While the above examples illustrate that a large class of graphs satisfy the local separation criterion, there indeed exist graphs which do not satisfy it. Such graphs tend to be "dense", that is, the

number of edges scales super-linearly in the number of nodes. For instance, the Erdős-Rényi graphs $\mathcal{G}_{\mathrm{ER}}(p, c/p)$ in the dense regime, where the average degree scales as $c = \Omega(p^2)$. In this regime, the node degrees as well as the number of short cycles grow with p. However, there is no simple decomposition into a local and a global graph with desirable properties, as in the previous example of small world graphs. Thus, the size of the local separators also grows with p in this case. Such graphs are hard instances for our framework.

3. Guarantees for Conditional Covariance Thresholding

We now characterize conditions under which the underlying Markov structure can be recovered successfully under conditional covariance thresholding.

3.1 Assumptions

(A1) *Sample Scaling Requirements:* We consider the asymptotic setting where both the number of variables (nodes) p and the number of samples n tend to infinity. We assume that the parameters $(n, p, J_{\min})$ scale in the following fashion:[8]

$$n = \Omega(J_{\min}^{-2} \log p). \tag{6}$$

We require that the number of nodes $p \to \infty$ to exploit the local separation properties of the class of graphs under consideration.

(A2) α-*Walk-summability:* The Gaussian graphical model Markov on $G_p \sim \mathcal{G}(p)$ is α-walk summable a.a.s., that is,

$$\|\overline{\mathbf{R}}_{G_p}\| \leq \alpha < 1, \quad \text{a.e. } G_p \sim \mathcal{G}(p), \tag{7}$$

where α is a constant (i.e., not a function of p), $\overline{\mathbf{R}} := [|R(i,j)|]$ is the entry-wise absolute value of the partial correlation matrix $\mathbf{R}$ and $\|\cdot\|$ denotes the spectral norm.

(A3) *Local-Separation Property:* We assume that the ensemble of graphs $\mathcal{G}(p; \eta, \gamma)$ satisfies the (η, γ)-local separation property with η, γ satisfying:

$$\eta = O(1), \; J_{\min} D_{\min}^{-1} \alpha^{-\gamma} = \omega(1), \tag{8}$$

where α is given by (7) and $D_{\min} := \min_i J(i,i)$ is the minimum diagonal entry of the potential matrix $\mathbf{J}$.

(A4) *Condition on Edge-Potentials:* The minimum absolute edge potential of an α-walk summable Gaussian graphical model satisfies

$$D_{\min}(1-\alpha) \min_{(i,j) \in G_p} \frac{J(i,j)}{K(i,j)} > 1 + \delta, \tag{9}$$

for almost every $G_p \sim \mathcal{G}(p)$, for some $\delta > 0$ (not depending on p) and[9]

$$K(i,j) := \|\mathbf{J}(V \setminus \{i,j\}, \{i,j\})\|^2,$$

8. The notations $\omega(\cdot)$, $\Omega(\cdot)$ refer to asymptotics as the number of variables $p \to \infty$.
9. Here and in the sequel, for $A, B \subset V$, we use the notation $\mathbf{J}(A,B)$ to denote the sub-matrix of $\mathbf{J}$ indexed by rows in A and columns in B.

is the spectral norm of the submatrix of the potential matrix $\mathbf{J}$, and $D_{\min} := \min_i J(i,i)$ is the minimum diagonal entry of $\mathbf{J}$. Intuitively, (9) limits the extent of non-homogeneity in the model and the extent of overlap of neighborhoods. Moreover, this assumption is not required for consistent graphical model selection when the model is attractive ($J_{i,j} \leq 0$ for $i \neq j$).[10]

(A5) *Choice of threshold* $\xi_{n,p}$*:* The threshold $\xi_{n,p}$ for graph estimation under CMIT algorithm is chosen as a function of the number of nodes p, the number of samples n, and the minimum edge potential $J_{\min}$ as follows:

$$\xi_{n,p} = O(J_{\min}),\ \xi_{n,p} = \omega\left(\frac{\alpha^\gamma}{D_{\min}}\right),\ \xi_{n,p} = \Omega\left(\sqrt{\frac{\log p}{n}}\right), \tag{10}$$

where α is given by (7), $D_{\min} := \min_i J(i,i)$ is the minimum diagonal entry of the potential matrix $\mathbf{J}$, and γ is the path-threshold (4) for the (η,γ)-local separation property to hold.

Assumption (A1) stipulates how n, p and $J_{\min}$ should scale for consistent graphical model selection, that is, the sample complexity. The sample size n needs to be sufficiently large with respect to the number of variables p in the model for consistent structure reconstruction. Assumptions (A2) and (A4) impose constraints on the model parameters. Assumption (A3) restricts the class of graphs under consideration. To the best of our knowledge, all previous works dealing with graphical model selection, for example, Meinshausen and Bühlmann (2006), Ravikumar et al. (2011), also impose some conditions for consistent graphical model selection. Assumption (A5) is with regard to the choice of a suitable threshold $\xi_{n,p}$ for thresholding conditional covariances. In the sequel, we compare the conditions for consistent recovery after presenting our main theorem.

3.1.1 EXAMPLE 1: DEGREE-BOUNDED ENSEMBLES

To gain a better understanding of conditions (A1)–(A5), consider the ensemble of graphs $\mathcal{G}_{\mathrm{Deg}}(p;\Delta)$ with bounded degree $\Delta \in \mathbb{N}$. It can be established that for the walk-summability condition in (A2) to hold,[11] we require that for normalized precision matrices ($J(i,i) = 1$),

$$J_{\max} = O\left(\frac{1}{\Delta}\right).$$

See Section A.2 for detailed discussion. When the minimum potential achieves the bound ($J_{\min} = \Theta(1/\Delta)$), a sufficient condition for (A3) to hold is given by

$$\Delta\alpha^\gamma = o(1), \tag{11}$$

where γ is the path threshold for the local-separation property to hold according to Definition 2. Intuitively, we require a larger path threshold γ, as the degree bound Δ on the graph ensemble increases.

Note that (11) allows for the degree bound Δ to grow with the number of nodes as long as the path threshold γ also grows appropriately. For example, if the maximum degree scales as $\Delta = O(\mathrm{poly}(\log p))$ and the path-threshold scales as $\gamma = O(\log\log p)$, then (11) is satisfied. This implies that graphs with fairly large degrees and short cycles can be recovered successfully using our algorithm.

10. The assumption (A5) rules out the possibility that the neighbors are marginally independent. See Section B.3 for details.

11. We can provide improved bounds for random-graph ensembles. See Section A.2 for details.

3.1.2 EXAMPLE 2: GIRTH-BOUNDED ENSEMBLES

The condition in (11) can be specialized for the ensemble of girth-bounded graphs $\mathcal{G}_{\text{Girth}}(p;g)$ in a straightforward manner as

$$\Delta\alpha^{\frac{g}{2}} = o(1), \tag{12}$$

where g corresponds to the *girth* of the graphs in the ensemble. The condition in (12) demonstrates a natural tradeoff between the girth and the maximum degree; graphs with large degrees can be learned efficiently if their girths are large. Indeed, in the extreme case of trees which have infinite girth, in accordance with (12), there is no constraint on node degrees for successful recovery and recall that the Chow-Liu algorithm (Chow and Liu, 1968) is an efficient method for model selection on tree distributions.

3.1.3 EXAMPLE 3: ERDŐS-RÉNYI AND SMALL-WORLD ENSEMBLES

We can also conclude that a.e. Erdős-Rényi graph $G \sim \mathcal{G}_{\text{ER}}(p, c/p)$ satisfies (8) when $c = O(\text{poly}(\log p))$ under the best-possible scaling of $J_{\min}$ subject to the walk-summability constraint in (7) (i.e., $J_{\min}$ achieves the upper bound).

This is because it can be shown that $J_{\min} = O(1/\sqrt{\Delta})$ for walk-summability in (7) to hold. See Section A.2 for details. Noting that a.a.s., the maximum degree Δ for $G \sim \mathcal{G}_{\text{ER}}(p, c/p)$ satisfies

$$\Delta = O\left(\frac{\log p \log c}{\log\log p}\right),$$

from Bollobás (1985, Ex. 3.6) and $\gamma = O(\frac{\log p}{\log c})$ from (5). Thus, the Erdős-Rényi graphs are amenable to successful recovery when the average degree $c = O(\text{poly}(\log p))$. Similarly, for the small-world ensemble $\mathcal{G}_{\text{Watts}}(p, d, c/p)$, when $d = O(1)$ and $c = O(\text{poly}(\log p))$, the graphs are amenable for consistent estimation.

3.2 Consistency of Conditional Covariance Thresholding

Assuming (A1)–(A5), we now state our main result. The proof of this result and the auxiliary lemmata for the proof can be found in Sections B and Section C.

Theorem 4 (Structural consistency of CMIT) *For structure learning of Gaussian graphical models Markov on a graph $G_p \sim \mathcal{G}(p;\eta,\gamma)$, $\mathsf{CMIT}(\mathbf{x}^n;\xi_{n,p},\eta)$ is consistent for a.e. graph G_p. In other words,*

$$\lim_{\substack{n,p\to\infty \\ n=\Omega(J_{\min}^{-2}\log p)}} P\left[\mathsf{CMIT}\left(\{\mathbf{x}^n\};\xi_{n,p},\eta\right) \neq G_p\right] = 0$$

Remarks:

1. *Consistency guarantee:* The CMIT algorithm consistently recovers the structure of Gaussian graphical models asymptotically, with probability tending to one, where the probability measure is with respect to both the random graph (drawn from the ensemble $\mathcal{G}(p;\eta,\gamma)$ and the samples (drawn from $\prod_{i=1}^{n} f(\mathbf{x}_i|G)$).

2. *Analysis of sample complexity:* The above result states that the sample complexity for the CMIT ($n = \Omega(J_{\min}^{-2} \log p)$), which improves when the minimum edge potential $J_{\min}$ is large.[12] This is intuitive since the edges have stronger potentials in this case. On the other hand, $J_{\min}$ cannot be arbitrarily large since the α-walk-summability assumption in (7) imposes an upper bound on $J_{\min}$. The minimum sample complexity (over different parameter settings) is attained when $J_{\min}$ achieves this upper bound. See Section A.2 for details. For example, for any degree-bounded graph ensemble $\mathcal{G}(p,\Delta)$ with maximum degree Δ, the minimum sample complexity is $n = \Omega(\Delta^2 \log p)$, that is, when $J_{\min} = \Theta(1/\Delta)$, while for Erdős-Rényi random graphs, the minimum sample complexity can be improved to $n = \Omega(\Delta \log p)$, that is, when $J_{\min} = \Theta(1/\sqrt{\Delta})$.

3. *Comparison with Ravikumar et al. (2011):* The work by Ravikumar et al. (2011) employs an ℓ_1-penalized likelihood estimator for structure estimation in Gaussian graphical models. Under the so-called incoherence conditions, the sample complexity is $n = \Omega((\Delta^2 + J_{\min}^{-2}) \log p)$. Our sample complexity in (6) is the same in terms of its dependence on $J_{\min}$, and there is no explicit dependence on the maximum degree Δ. Moreover, we have a transparent sufficient condition in terms of α-walk-summability in (7), which directly imposes scaling conditions on $J_{\min}$. It is an open question if the models satisfying incoherence conditions are walk-summable or viceversa. However, for random graph models, we can obtain better guarantees in terms of average degrees while the incoherence conditions are based on maximum degree in the graph. We also present experimental comparison between this method and our developed method in Section 6.

4. *Comparison with Meinshausen and Bühlmann (2006):* The work by Meinshausen and Bühlmann (2006) considers ℓ_1-penalized linear regression for neighborhood selection of Gaussian graphical models and establish a sample complexity of $n = \Omega((\Delta + J_{\min}^{-2}) \log p)$. We note that our guarantees allow for graphs which do not necessarily satisfy the conditions imposed by Meinshausen and Bühlmann (2006). For instance, the assumption of neighborhood stability (assumption 6 in Meinshausen and Bühlmann, 2006) is hard to verify in general, and the relaxation of this assumption corresponds to the class of models with diagonally-dominant covariance matrices. Note that the class of Gaussian graphical models with diagonally-dominant covariance matrices forms a strict sub-class of walk-summable models, and thus satisfies assumption (A2) for the theorem to hold. Thus, Theorem 4 applies to a larger class of Gaussian graphical models compared to Meinshausen and Bühlmann (2006). Furthermore, the conditions for successful recovery in Theorem 4 are arguably more transparent.

5. *Local vs. Global Conditions for Success:* The conditions required for the success of our methods as well as the ℓ_1 penalized MLE of Ravikumar et al. (2011) are global, meaning that the entire model (i.e., all the parameters) need to satisfy the specified conditions for recovering the entire graph. It does not appear straightforward to characterize local conditions for successful recovery under our formulation, that is, when our algorithm may succeed in recovering some parts of the graph, but not others. On the other hand, the ℓ_1 penalized neighborhood selection method of Meinshausen and Bühlmann (2006) provides a separate

12. Note that the sample complexity also implicitly depends on walk-summability parameter α through (8).

set of conditions for recovery of each neighborhood in the graph. However, as discussed above, these conditions appear stronger and more opaque than our conditions.

6. *Comparison with Ising models:* Our above result for learning Gaussian graphical models is analogous to structure estimation of Ising models subject to an upper bound on the edge potentials (Anandkumar et al., 2012b), and we characterize such a regime as a *conditional uniqueness* regime. Thus, walk-summability is the analogous condition for Gaussian models.

Proof Outline: We first analyze the scenario when exact statistics are available. (i) We establish that for any two non-neighbors $(i, j) \notin G$, the minimum conditional covariance in (1) (based on exact statistics) does not exceed the threshold $\xi_{n,p}$. (ii) Similarly, we also establish that the conditional covariance in (1) exceeds the threshold $\xi_{n,p}$ for all neighbors $(i, j) \in G$. (iii) We then extend these results to empirical versions using concentration bounds.

3.2.1 PERFORMANCE OF CONDITIONAL MUTUAL INFORMATION TEST

We now employ the conditional mutual information test, analyzed in Anandkumar et al. (2012b) for Ising models, and note that it has slightly worse sample complexity than using conditional covariances. Using the threshold $\xi_{n,p}$ defined in (10), the conditional mutual information test CMIT is given by the threshold test

$$\min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,p}^2,$$

and node pairs (i, j) exceeding the threshold are added to the estimate $\widehat{G}_p^n$. Assuming (A1)–(A5), we have the following result.

Theorem 5 (Structural consistency of CMIT) *For structure learning of the Gaussian graphical model on a graph $G_p \sim \mathcal{G}(p; \eta, \gamma)$, $\mathsf{CMIT}(\mathbf{x}^n; \xi_{n,p}, \eta)$ is consistent for a.e. graph G_p. In other words,*

$$\lim_{\substack{n,p \to \infty \\ n = \Omega(J_{\min}^{-4} \log p)}} P\left[\mathsf{CMIT}\left(\{\mathbf{x}^n\}; \xi_{n,p}, \eta\right) \neq G_p\right] = 0$$

The proof of this theorem is provided in Section C.3.

Remarks:

1. For Gaussian random variables, conditional covariances and conditional mutual information are equivalent tests for conditional independence. However, from above results, we note that there is a difference in the sample complexity for the two tests. The sample complexity of CMIT is $n = \Omega(J_{\min}^{-4} \log p)$ in contrast to $n = \Omega(J_{\min}^{-2} \log p)$ for CMIT. This is due to faster decay of conditional mutual information on the edges compared to the decay of conditional covariances. Thus, conditional covariances are more efficient for Gaussian graphical model selection compared to conditional mutual information.

4. Necessary Conditions for Model Selection

In the previous sections, we proposed and analyzed efficient algorithms for learning the structure of Gaussian graphical models Markov on graph ensembles satisfying local-separation property. In this section, we study the problem of deriving *necessary* conditions for consistent structure learning.

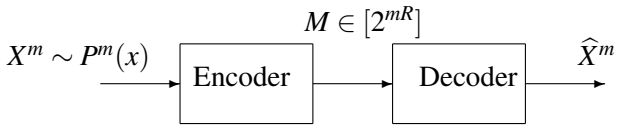


Figure 2: The canonical source coding problem. See Chapter 3 in Cover and Thomas (2006).

For the class of degree-bounded graphs $\mathcal{G}_{\mathrm{Deg}}(p,\Delta)$, necessary conditions on sample complexity have been characterized before (Wang et al., 2010) by considering a certain (limited) set of ensembles. However, a naïve application of such bounds (based on Fano's inequality (Cover and Thomas, 2006, Ch. 2)) turns out to be too weak for the class of Erdős-Rényi graphs $\mathcal{G}_{\mathrm{ER}}(p,c/p)$, where the average degree[13] c is much smaller than the maximum degree.

We now provide necessary conditions on the sample complexity for recovery of Erdős-Rényi graphs. Our information-theoretic techniques may also be applicable to other ensembles of random graphs. This is a promising avenue for future work.

4.1 Setup

We now describe the problem more formally. A graph G is drawn from the ensemble of Erdős-Rényi graphs $G \sim \mathcal{G}_{\mathrm{ER}}(p,c/p)$. The learner is also provided with n conditionally i.i.d. samples $\mathbf{X}^n := (\mathbf{X}_1,\ldots,\mathbf{X}_n) \in (\mathcal{X}^p)^n$ (where $\mathcal{X}=\mathbb{R}$) drawn from the conditional (Gaussian) product probability density function (pdf) $\prod_{i=1}^n f(\mathbf{x}_i|G)$. The task is then to estimate G, a random quantity. The estimate is denoted as $\widehat{G} := \widehat{G}(\mathbf{X}^n)$. It is desired to derive tight necessary conditions on n (as a function of c and p) so that the *probability of error*

$$P_e^{(p)} := P(\widehat{G} \neq G) \to 0 \tag{13}$$

as the number of nodes p tends to infinity. Note that the probability measure P in (13) is associated to *both* the realization of the random graph G and the samples $\mathbf{X}^n$.

The task is reminiscent of source coding (or compression), a problem of central importance in information theory (Cover and Thomas, 2006)—we would like to derive fundamental limits associated to the problem of reconstructing the source G given a compressed version of it $\mathbf{X}^n$ ($\mathbf{X}^n$ is also analogous to the "message"). However, note the important distinction; while in source coding, the source coder can design both the encoder *and* the decoder, our problem mandates that the code is fixed by the conditional probability density $f(\mathbf{x}|G)$. We are only allowed to design the decoder. See comparisons in Figs. 2 and 3.

4.2 Necessary Conditions for Exact Recovery

To derive the necessary condition for learning Gaussian graphical models Markov on sparse Erdős-Rényi graphs $G \sim \mathcal{G}_{\mathrm{ER}}(p,c/p)$, we assume that the strict walk-summability condition with parameter α, according to (7). We are then able to demonstrate the following:

Theorem 6 (Weak Converse for Gaussian Models) *For a walk-summable Gaussian graphical model satisfying* (7) *with parameter* α*, for almost every graph* $G \sim \mathcal{G}_{\mathrm{ER}}(p,c/p)$ *as* $p \to \infty$*, in order*

13. The techniques in this section are applicable when the average degree (c) of $\mathcal{G}_{\mathrm{ER}}(p,c/p)$ ensemble is a function of p, for example, $c = O(\mathrm{poly}(\log p))$.

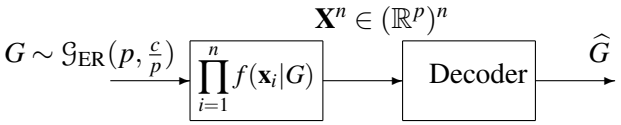


Figure 3: The estimation problem is analogous to source coding: the "source" is $G \sim \mathcal{G}_{\mathrm{ER}}(p, \frac{c}{p})$, the "message" is $\mathbf{X}^n \in (\mathbb{R}^p)^n$ and the "decoded source" is $\widehat{G}$. We are asking what the minimum "rate" (analogous to the number of samples n) are required so that $\widehat{G} = G$ with high probability.

for $P_e^{(p)} \to 0$, we require that

$$n \geq \frac{2}{p \log_2 \left[2\pi e \left(\frac{1}{1-\alpha}+1\right)\right]} \binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right) \tag{14}$$

for all p sufficiently large.

The proof is provided in Section D.1. By expanding the binary entropy function, it is easy to see that the statement in (14) can be weakened to the necessary condition:

$$n \geq \frac{c \log_2 p}{\log_2 \left[2\pi e \left(\frac{1}{1-\alpha}+1\right)\right]}.$$

The above condition does not involve any asymptotic notation, and also demonstrates the dependence of the sample complexity on p, c and α transparently. Finally, the dependence on α can be explained as follows: any α-walk-summable model is also β-walk-summable for all $\beta > \alpha$. Thus, the class of β-walk-summable models contains the class of α-walk-summable models. This results in a looser bound in (14) for larger α.

4.3 Necessary Conditions for Recovery with Distortion

In this section, we generalize Theorem 6 to the case where we only require estimation of the underlying graph up to a certain edit distance: an error is declared if and only if the estimated graph $\widehat{G}$ exceeds an edit distance (or distortion) D of the true graph. The *edit distance* $d : \mathfrak{G}_p \times \mathfrak{G}_p \to \mathbb{N} \cup \{0\}$ between two undirected graphs $G = (V, E)$ and $G = (V, E')$ is defined as $d(G, G') := |E \triangle E'|$, where $\triangle$ denotes the symmetric difference between the edge sets E and E'. The edit distance can be regarded as a distortion measure between two graphs.

Given an positive integer D, known as the *distortion*, suppose we declare an error if and only if $d(G, G') > D$, then the probability of error is redefined as

$$P_e^{(p)} := P(d(G, \widehat{G}(\mathbf{X}^n)) > D). \tag{15}$$

We derive necessary conditions on n (as a function of p and c) such that the probability of error (15) goes to zero as $p \to \infty$. To ease notation, we define the ratio

$$\beta := D / \binom{p}{2}. \tag{16}$$

Note that β may be a function of p. We do not attempt to make this dependence explicit. The following corollary is based on an idea propounded by Kim et al. (2008) among others.

Corollary 7 (Weak Converse for Discrete Models With Distortion) *For $P_e^{(p)} \to 0$, we must have*

$$n \geq \frac{2}{p \log_2 \left[2\pi e \left(\frac{1}{1-\alpha}+1\right)\right]} \binom{p}{2} \left[H_{\mathrm{b}}\left(\frac{c}{p}\right) - H_{\mathrm{b}}(\beta)\right] \tag{17}$$

for all p sufficiently large.

The proof of this corollary is provided in Section D.7. Note that for (17) to be a useful bound, we need $\beta < c/p$ which translates to an allowed distortion $D < cp/2$. We observe from (17) that because the error criterion has been relaxed, the required number of samples is also reduced from the corresponding lower bound in (14).

4.4 Proof Techniques

Our analysis tools for deriving necessary conditions for Gaussian graphical model selection are information-theoretic in nature. A common and natural tool to derive necessary conditions (also called converses) is to resort to Fano's inequality (Cover and Thomas, 2006, Chapter 2), which (lower) bounds the probability of error $P_e^{(p)}$ as a function of the *equivocation* or *conditional entropy* $H(G|\mathbf{X}^n)$ and the size of the set of all graphs with p nodes. However, a direct and naïve application Fano's inequality results in a trivial lower bound as the set of all graphs, which can be realized by $\mathcal{G}_{\mathrm{ER}}(p, c/p)$ is, loosely speaking, "too large".

To ameliorate such a problem, we employ another information-theoretic notion, known as *typicality*. A *typical set* is, roughly speaking, a set that has small cardinality and yet has high probability as $p \to \infty$. For example, the probability of a set of length-m sequences is of the order $\approx 2^{mH}$ (where H is the entropy rate of the source) and hence those sequences with probability close to this value are called *typical*. In our context, given a graph G, we define the $\bar{d}(G)$ to be the ratio of the number of edges of G to the total number of nodes p. Let $\mathfrak{G}_p$ denote the set of all graphs with p nodes. For a fixed $\varepsilon > 0$, we define the following set of graphs:

$$\mathcal{T}_\varepsilon^{(p)} := \left\{ G \in \mathfrak{G}_p : \left| \frac{\bar{d}(G)}{c} - \frac{1}{2} \right| \leq \frac{\varepsilon}{2} \right\}.$$

The set $\mathcal{T}_\varepsilon^{(p)}$ is known as the ε-*typical set of graphs*. Every graph $G \in \mathcal{T}_\varepsilon^{(p)}$ has an average number of edges that is $\frac{c}{2}\varepsilon$-close in the Erdős-Rényi ensemble. Note that typicality ideas are usually used to derive *sufficient* conditions in information theory (Cover and Thomas, 2006) (*achievability* in information-theoretic parlance); our use of *both* typicality for graphical model selection as well as Fano's inequality to derive converse statements seems novel. Indeed, the proof of the converse of the source coding theorem in Cover and Thomas (2006, Chapter 3) uses only Fano's inequality. We now summarize the properties of the typical set.

Lemma 8 (Properties of $\mathcal{T}_\varepsilon^{(p)}$) *The ε-typical set of graphs has the following properties:*

1. *$P(\mathcal{T}_\varepsilon^{(p)}) \to 1$ as $p \to \infty$.*
2. *For all $G \in \mathcal{T}_\varepsilon^{(p)}$, we have*[14]

$$\exp_2\left[-\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right)(1+\varepsilon)\right] \leq P(G) \leq \exp_2\left[-\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right)\right].$$

14. We use the notation $\exp_2(\cdot)$ to mean $2^{(\cdot)}$.

3. *The cardinality of the* ε*-typical set can be bounded as*

$$(1-\varepsilon)\exp_2\left[\binom{p}{2}H_{\mathrm{b}}\left(\frac{c}{p}\right)\right] \leq |\mathcal{T}_{\varepsilon}^{(p)}| \leq \exp_2\left[\binom{p}{2}H_{\mathrm{b}}\left(\frac{c}{p}\right)(1+\varepsilon)\right]$$

for all p *sufficiently large.*

The proof of this lemma can be found in Section D.2. Parts 1 and 3 of Lemma 8 respectively say that the set of typical graphs has high probability and has very small cardinality relative to the number of graphs with p nodes $|\mathfrak{G}_p| = \exp_2(\binom{p}{2})$. Part 2 of Lemma 8 is known as the *asymptotic equipartition property*: the graphs in the typical set are almost uniformly distributed.

5. Implications on Loopy Belief Propagation

An active area of research in the graphical model community is that of inference—that is, the task of computing node marginals (or MAP estimates) through efficient distributed algorithms. The simplest of these algorithms is the belief propagation[15] (BP) algorithm, where messages are passed among the neighbors of the graph of the model. It is known that belief propagation (and max-product) is exact on tree models, meaning that correct marginals are computed at all the nodes (Pearl, 1988). On the other hand on general graphs, the generalized version of BP, known as loopy belief propagation (LBP), may not converge and even if it does, the marginals may not be correct. Motivated by the twin problems of convergence and correctness, there has been extensive work on characterizing LBP's performance for different models. As a by-product of our previous analysis on graphical model selection, we now show the asymptotic correctness of LBP on walk-summable Gaussian models when the underlying graph is locally tree-like.

5.1 Background

The belief propagation (BP) algorithm is a distributed algorithm where messages (or beliefs) are passed among the neighbors to draw inferences at the nodes of a graphical model. The computation of node marginals through naïve variable elimination (or Gaussian elimination in the Gaussian setting) is prohibitively expensive. However, if the graph is sparse (consists of few edges), the computation of node marginals may be sped up dramatically by exploiting the graph structure and using distributed algorithms to parallelize the computations.

For the sake of completeness, we now recall the basic steps in LBP, specific to Gaussian graphical models. Given a message schedule which specifies how messages are exchanged, each node j receives information from each of its neighbors (according to the graph), where the message, $m_{i\to j}^t(x_j)$, from i to j, in t^{th} iteration is parameterized as

$$m_{i\to j}^t(x_j) := \exp\left[-\frac{1}{2}\Delta J_{i\to j}^t x_j^2 + \Delta h_{i\to j}^t x_j\right].$$

Each node i prepares message $m_{i\to j}^t(x_j)$ by collecting messages from neighbors of the previous iteration (under parallel iterations), and computing

$$\hat{J}_{i\setminus j}(t) = J(i,i) + \sum_{k\in\mathcal{N}(i)\setminus j}\Delta J_{k\to i}^{t-1}, \quad \hat{h}_{i\setminus j}(t) = h(i) + \sum_{k\in\mathcal{N}(i)\setminus j}\Delta h_{k\to i}(t),$$

15. The variant of the belief propagation algorithm which computes the MAP estimates is known as the max-product algorithm.

where

$$\Delta J^t_{i\to j} = -J(j,i)\hat{J}^{-1}_{i\setminus j}(t)J(j,i), \quad \Delta h^t_{i\to j} = -J(j,i)\hat{J}^{-1}_{i\setminus j}(t)\hat{h}_{k\to i}(t).$$

5.2 Results

Let $\Sigma_{\mathrm{LBP}}(i,i)$ denote the variance at node i at the LBP fixed point.[16] Without loss of generality, we consider the normalized version of the precision matrix

$$\mathbf{J} = \mathbf{I} - \mathbf{R},$$

which can always be obtained from a general precision matrix via normalization. We can then renormalize the variances, computed via LBP, to obtain the variances corresponding to the unnormalized precision matrix.

We consider the following ensemble of locally-tree like graphs. Consider the event that the neighborhood of a node i has no cycles up to graph distance γ, given by

$$\Gamma(i;\gamma,G) := \{B_\gamma(i;G) \text{ does not contain any cycles}\}.$$

We assume a random graph ensemble $\mathcal{G}(p)$ such that for a given node $i \in V$, we have

$$P[\Gamma^c(i;\gamma,G)] = o(1). \tag{18}$$

Proposition 9 (Correctness of LBP) *Given an α-walk-summable Gaussian graphical model on a.e. locally tree-like graph $G \sim \mathcal{G}(p;\gamma)$ with parameter γ satisfying* (18)*, we have*

$$|\Sigma_G(i,i) - \Sigma_{\mathrm{LBP}}(i,i)| \overset{a.a.s.}{=} O(\max(\alpha^\gamma, P[\Gamma^c(i;\gamma,G)])).$$

The proof is given in Section B.4.

Remarks:

1. The class of Erdős-Rényi random graphs, $G \sim \mathcal{G}_{\mathrm{ER}}(p,c/p)$ satisfies (18), with $\gamma = O(\log p/\log c)$ for a node $i \in V$ chosen uniformly at random.
2. Recall that the class of random regular graphs $G \sim \mathcal{G}_{\mathrm{Reg}}(p,\Delta)$ have a girth of $O(\log_{\Delta-1} p)$. Thus, for any node $i \in V$, (18) holds with $\gamma = O(\log_{\Delta-1} p)$.

6. Experiments

In this section we present some experimental results on real and synthetic data. We implement the proposed CMIT method as well the convex relaxation methods, namely, the ℓ_1 penalized maximum likelihood estimate (MLE) (Ravikumar et al., 2011) and ℓ_1 penalized neighborhood selection (Meinshausen and Bühlmann, 2006). We measure the performance of methods using the notion of the edit distance between the true and estimated graphs (for synthetic data). We also compare the penalized likelihood scores of the estimated models using the notion of Bayesian information criterion (BIC) for both synthetic and real data. We implement the proposed CMIT method in MATLAB and the convex relaxation methods using the YALMIP package.[17] We also used CONTEST[18] to generate the synthetic graphs. The data sets, software code and results are available at `http://newport.eecs.uci.edu/anandkumar`.

16. Convergence of LBP on walk-summable models has been established by Malioutov et al. (2006).
17. YALMIP is available at `http://users.isy.liu.se/johanl/yalmip/pmwiki.php?n=Main.Download`.
18. CONTEST is at `http://www.maths.strath.ac.uk/research/groups/numerical_analysis/contest`.

6.1 Data Sets

Synthetic data: In order to evaluate the performance of our algorithm in terms of error in reconstructing the graph structure, we generated samples from Gaussian graphical models for different graphs. These include a single cycle graph with $p = 80$ nodes, an Erdős-Rényi (ER) graph $G \sim \mathcal{G}_{\mathrm{ER}}(p, c/p)$ with average degree $c = 1.2$ and Watts-Strogatz model $\mathcal{G}_{\mathrm{Watts}}(p, d, c/p)$ with degree of local graph $d = 2$ and average degree of the global graph $c = 1.2$. Given the graph structure G, we generate the potential matrix $\mathbf{J}_G$ whose sparsity pattern corresponds to that of G. We set the diagonal elements in $\mathbf{J}_G$ to unity and nonzero off-diagonal entries are picked uniformly[19] from $[0, 0.1]$. We set the potential vector $\mathbf{h}$ to be $\mathbf{0}$ without loss of generality. We let the number of samples be $n \in \{10^2, 5\times 10^2, 10^3, 5\times 10^3, 10^4\}$. Note that for synthetic data, we know η, the size of local separators for non-neighboring node pairs in the graph, and we incorporate it in the implementation of the CMIT algorithm. We present edit distance results for CMIT and the above mentioned convex relaxation methods for different thresholds and regularization parameters.[20]

Foreign exchange data: We consider monthly trends of foreign exchange rates[21] of 19 currencies[22] with respect to the US dollar from 10/1/1983 to 02/1/2012. We evaluate the BIC score for models estimated using our algorithm under different thresholds $\xi_{n,p}$ and different sizes of the local separator sets η, and compared it with the convex relaxation methods under different regularization parameters.

6.2 Performance Criteria

The BIC score has been extensively used to enable tradeoff between data fitting and model complexity (Schwarz, 1978). We use a modified version of the BIC score proposed for high-dimensional data sets (Foygel and Drton, 2010) as the performance criterion for model fitting. Given n samples $\mathbf{x}^n := [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, and parameters θ,

$$\mathrm{BIC}(\mathbf{x}^n; \theta) := \sum_{k=1}^{n} \log f(\mathbf{x}_k; \theta) - 0.5|E| \log n - 2|E| \log p,$$

where $|E|$ is the number of edges in the Markov graph and $\boldsymbol{\theta}$ is the set of parameters characterizing the model. It has been observed elsewhere (Liu et al., 2009) that the BIC score tends to overselect the edges leading to dense graphs, and thus, we impose a hard threshold on the number of edges (both for our method and for convex relaxation methods), and select the model with the best BIC score. For the foreign exchange data, we limited the number of edges to 100, while for synthetic data, we limited it to 100 for cycle and Erdős-Rényi (ER) graphs and to 200 for the Watts-Strogatz model. We note that alternatively, the thresholds/regularization parameters can be selected via cross validation or other mechanisms, see Liu et al. (2009) for details.

19. The choice of parameters and graphs result in valid models in our experiments, that is, the potential matrix is positive definite.
20. For the convex relaxation methods in Ravikumar et al. (2011) and Meinshausen and Bühlmann (2006), the regularization parameter denotes the weight associated with the ℓ_1 term.
21. Data set available at `http://research.stlouisfed.org/fred2/categories/15/downloaddata`.
22. The European countries which switched to Euro are not considered in our analysis.

Graph	n	CMIT	ℓ_1 MLE	ℓ_1 Nbd
Cycle	10^2	1.0000	1.0000	1.0000
ER	10^2	1.0000	1.0000	1.0000
WS	10^2	1.0000	1.0000	1.0000
Cycle	10^3	0.95	0.9875	0.9000
ER	10^3	0.6825	1.1087	1.0000
WS	10^3	0.8580	0.9520	0.8063
Cycle	10^4	0.4125	0.3875	0.3625
ER	10^4	0.3273	0.3469	0.5435
WS	10^4	0.3252	0.3313	0.2688

Table 1: Normalized edit distance under CMIT, ℓ_1 penalized MLE and ℓ_1 penalized neighborhood selection on synthetic data from graphs listed above, where n denotes the number of samples.

6.3 Experimental Outcomes

Synthetic data: We compare the performance of our method CMIT with convex relaxation methods for synthetic data as described earlier. We evaluate the normalized edit distance (normalized with respect to the number of edges), since we know the ground truth for synthetic data and present the results in Table 1 for CMIT, ℓ_1 penalized MLE and ℓ_1 penalized neighborhood selection. methods. The results are also presented in figures 7a, 7b and 7c. We note that CMIT has better edit distance performance and BIC scores compared to ℓ_1 penalized MLE in most cases, and similar performance compared to the ℓ_1 penalized neighborhood selection.

Foreign exchange data: We evaluate the BIC scores under our algorithm CMIT with different values of η (the constraint on the size of subsets used for conditioning)[23] and threshold $\xi_{n,p}$. We present the results in Table 2, where for each value of η, we present the threshold $\xi_{n,p}$ which achieves the best BIC score under the sparsity constraint. We also present the regularization parameters for convex relaxation methods with the best BIC. The estimated graphs are shown in figures 4 and 5. We note that while CMIT distributes the edges fairly uniformly across the nodes, the ℓ_1 penalized MLE tends to cluster all the edges together between the "dominant" variables leading to a densely connected component and several isolated nodes. We observe from the reconstructed graphs that geography plays a crucial role in the foreign exchange trends. In Fig.4 recovered using the CMIT method, we note that among Asian countries India and S. Korea are high degree nodes and are connected to countries which are geographically close (e.g., Sri Lanka for India, and Australia, Thailand, Taiwan and China for S. Korea). On the other hand, the ℓ_1 method outputs a densely connected graph where such geographical relationships are missing. Thus, we see that in the experiments, the proposed CMIT method tends to enforce local sparsity in the graph, while the ℓ_1 method of Ravikumar et al. (2011) enforces global sparsity, and tends to cluster the edges together. On the other hand, the ℓ_1 penalized neighborhood selection (Meinshausen and Bühlmann, 2006) is better than the MLE in distributing the edges across all the nodes, but carries this out to a lesser extent than our method.

23. The BIC score for $\eta = 0$ is too low and we do not present it in our results. This implies that there is dependence between the variables.

Thres.(CMIT)	η	LL-train$\times 10^7$	LL-test$\times 10^7$	BIC-train$\times 10^7$	BIC-test$\times 10^7$	$\lvert E\rvert$
0.5	1	-2.9521	-7.4441	-2.9522	-7.4442	23
0.5	2	-3.2541	-8.5923	-3.2541	-8.5923	8
0.01	3	-2.9669	-7.3773	-2.9670	-7.3774	19
0.001	**4**	**-2.9653**	**-7.3674**	**-2.9654**	**-7.3675**	**25**
0.0005	5	-3.2901	-8.8396	-3.3068	-8.8397	24
0.0005	6	-3.2921	-8.8466	-3.2921	-8.8467	18
Thres.(ℓ_1 MLE)	–	LL-train$\times 10^7$	LL-test$\times 10^7$	BIC-train$\times 10^7$	BIC-test$\times 10^7$	$\lvert E\rvert$
6.5803	–	-2.5831	-6.3167	-2.5832	-6.3167	28
Thres.(ℓ_1 Nbd)	–	LL-train$\times 10^7$	LL-test$\times 10^7$	BIC-train$\times 10^7$	BIC-test$\times 10^7$	$\lvert E\rvert$
13.1606	–	-2.7971	-6.9630	-2.7972	-6.9631	26

Table 2: Experimental outcome for CMIT, ℓ_1 penalized MLE and ℓ_1 penalized neighborhood selection for different thresholds/regularization parameters and size of conditioning sets η for foreign exchange data. $|E|$ denotes the number of edges.

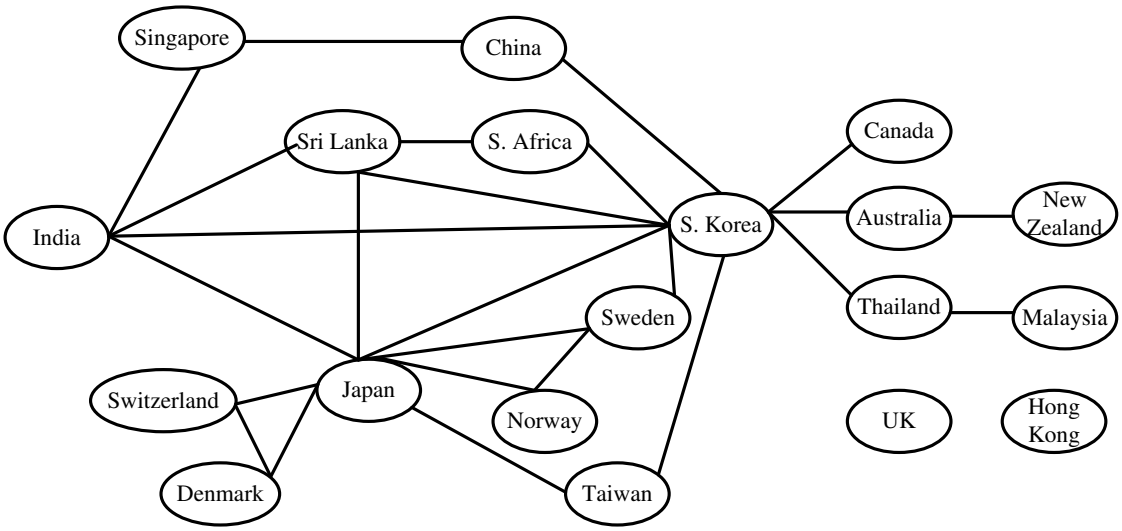


Figure 4: Graph estimate under CMIT algorithm for foreign exchange data set for $\eta = 4$, see Table 2.

7. Conclusion

In this paper, we adopted a novel and a unified paradigm for graphical model selection. We presented a simple local algorithm for structure estimation with low computational and sample complexities under a set of mild and transparent conditions. This algorithm succeeds on a wide range of graph ensembles such as the Erdős-Rényi ensemble, small-world networks etc. We also employed novel information-theoretic techniques for establishing necessary conditions for graphical model selection.

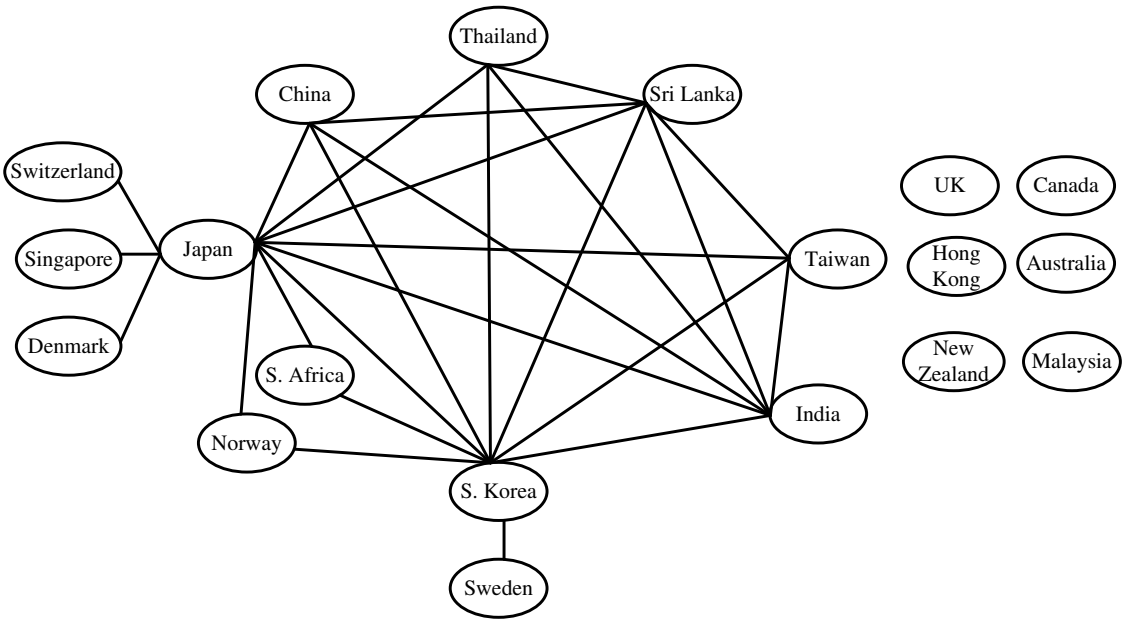


Figure 5: Graph estimate under ℓ_1 penalized MLE for foreign exchange data set. See Table 2.

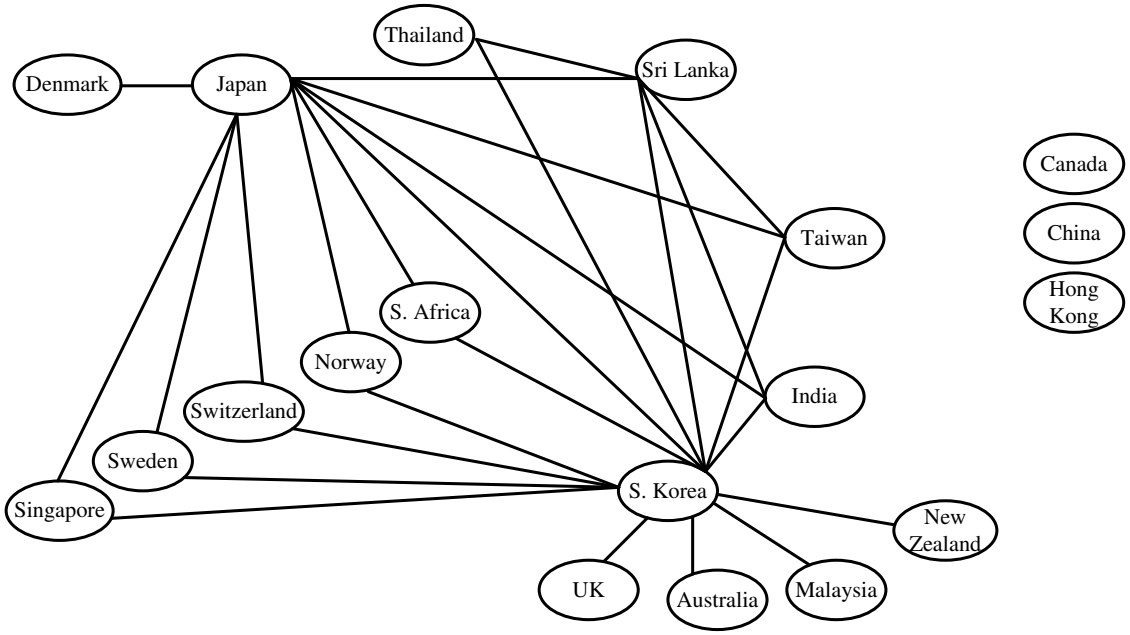
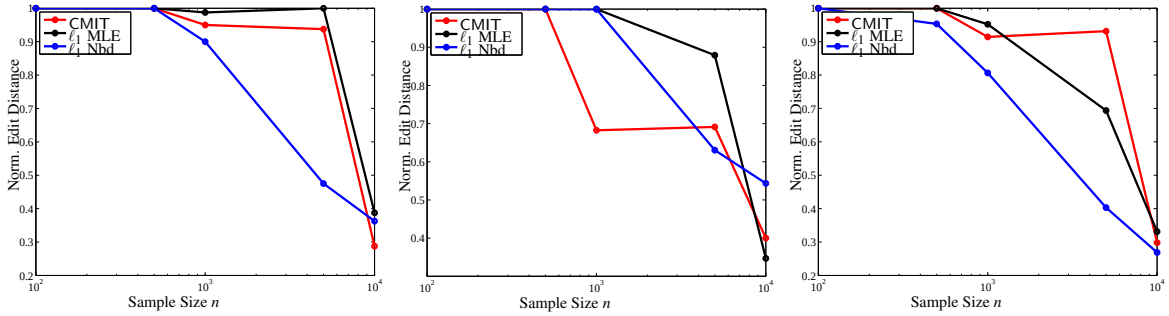


Figure 6: Graph estimate under ℓ_1 penalized neighborhood selection method for foreign exchange data set. See Table 2.



(a) Cycle (b) Erdös-Rényi (c) Watts-Strogatz

Figure 7: CMIT, ℓ_1 penalized MLE and ℓ_1 penalized neighborhood selection methods.

Acknowledgments

An abridged version of this paper appeared in the Proceedings of NIPS 2011. The first author is supported in part by the setup funds at UCI and the AFOSR Award FA9550-10-1-0310, the second author is supported by A*STAR, Singapore and the third author is supported in part by AFOSR under Grant FA9550-08-1-1080. The authors thank Venkat Chandrasekaran (UC Berkeley) for discussions on walk-summable models, Elchanan Mossel (UC Berkeley) for discussions on the necessary conditions for model selection and Divyanshu Vats (U. Minn.) for extensive comments. The authors thank the Associate Editor Martin Wainwright (Berkeley) and the anonymous reviewers for comments which significantly improved this manuscript.

Appendix A. Walk-summable Gaussian Graphical Models

We first provide an overview of the notion of walk-summability for Gaussian graphical models.

A.1 Background on Walk-Summability

We now recap the properties of walk-summable Gaussian graphical models, as given by (7). For details, see Malioutov et al. (2006). For simplicity, we first assume that the diagonal of the potential matrix $\mathbf{J}$ is normalized ($J(i,i) = 1$ for all $i \in V$). We remove this assumption and consider general unnormalized precision matrices in Section B.2. Consider splitting the matrix $\mathbf{J}$ into the identity matrix and the partial correlation matrix $\mathbf{R}$, defined in (3):

$$\mathbf{J} = \mathbf{I} - \mathbf{R}. \tag{19}$$

The covariance matrix $\mathbf{\Sigma}$ of the graphical model in (19) can be decomposed as

$$\mathbf{\Sigma} = \mathbf{J}^{-1} = (\mathbf{I} - \mathbf{R})^{-1} = \sum_{k=0}^{\infty} \mathbf{R}^k, \quad \|\mathbf{R}\| < 1, \tag{20}$$

using Neumann power series for the matrix inverse. Note that we require that $\|\mathbf{R}\| < 1$ for (20) to hold, which is implied by walk-summability in (7) (since $\|\mathbf{R}\| \leq \|\overline{\mathbf{R}}\|$).

We now relate the matrix power $\mathbf{R}^l$ to walks on graph G. A walk $\mathbf{w}$ of length $l \geq 0$ on graph G is a sequence of nodes $\mathbf{w} := (w_0, w_1, \ldots, w_l)$ traversed on the graph G, that is, $(w_k, w_{k+1}) \in G$. Let $|\mathbf{w}|$

denote the length of the walk. Given matrix $\mathbf{R}_G$ supported on graph G, let the weight of the walk be

$$\phi(\mathbf{w}) := \prod_{k=1}^{|\mathbf{w}|} R(w_{k-1}, w_k).$$

The elements of the matrix power $\mathbf{R}^l$ are given by

$$R^l(i,j) = \sum_{\mathbf{w}: i \xrightarrow{l} j} \phi(\mathbf{w}), \tag{21}$$

where $i \xrightarrow{l} j$ denotes the set of walks from i to j of length l. For this reason, we henceforth refer to $\mathbf{R}$ as the *walk matrix*.

Let $i \to j$ denote all the walks between i and j. Under the walk-summability condition in (7), we have convergence of $\sum_{\mathbf{w}: i \to j} \phi(\mathbf{w})$, irrespective of the order in which the walks are collected, and this is equal to the covariance $\Sigma(i,j)$.

In Section A.3, we relate walk-summability in (7) to the notion of correlation decay, where the effect of faraway nodes on covariances can be controlled and the local-separation property of the graphs under consideration can be exploited.

A.2 Sufficient Conditions for Walk-summability

We now provide sufficient conditions and suitable parameterization for walk-summability in (7) to hold. The adjacency matrix $\mathbf{A}_G$ of a graph G with maximum degree Δ_G satisfies

$$\lambda_{\max}(\mathbf{A}_G) \leq \Delta_G,$$

since it is dominated by a Δ-regular graph which has maximum eigenvalue of Δ_G. From Perron-Frobenius theorem, for adjacency matrix $\mathbf{A}_G$, we have $\lambda_{\max}(\mathbf{A}_G) = \|\mathbf{A}_G\|$, where $\|\mathbf{A}_G\|$ is the spectral radius of $\mathbf{A}_G$. Thus, for $\overline{\mathbf{R}}_G$ supported on graph G, we have

$$\alpha := \|\overline{\mathbf{R}}_G\| = O\left(J_{\max}\Delta\right),$$

where $J_{\max} := \max_{i,j} |R(i,j)|$. This implies that

$$J_{\max} = O\left(\frac{1}{\Delta}\right)$$

to have $\alpha < 1$, which is the requirement for walk-summability.

When the graph G is a Erdős-Rényi random graph, $G \sim \mathcal{G}_{\mathrm{ER}}(p, c/p)$, we can provide better bounds. When $G \sim \mathcal{G}_{\mathrm{ER}}(p, c/p)$, we have Krivelevich and Sudakov (2003), that

$$\lambda_{\max}(\mathbf{A}_G) = (1 + o(1)) \max(\sqrt{\Delta_G}, c),$$

where Δ_G is the maximum degree and $\mathbf{A}_G$ is the adjacency matrix. Thus, in this case, when $c = O(1)$, we require that

$$J_{\max} = O\left(\sqrt{\frac{1}{\Delta}}\right),$$

for walk-summability ($\alpha < 1$). Note that when $c = O(\mathrm{poly}(\log p))$, w.h.p. $\Delta_{G_p} = \Theta(\log p / \log\log p)$ (Bollobás, 1985, Ex. 3.6).

A.3 Implications of Walk-Summability

Recall that $\mathbf{\Sigma}_G$ denotes the covariance matrix for Gaussian graphical model on graph G and that $\mathbf{J}_G = \mathbf{\Sigma}_G^{-1}$ with $\mathbf{J}_G = \mathbf{I} - \mathbf{R}_G$ in (19). We now relate the walk-summability condition in (7) to correlation decay in the model. In other words, under walk-summability, we can show that the effect of faraway nodes on covariances decays with distance, as made precise in Lemma 10.

Let $B_\gamma(i)$ denote the set of nodes within γ hops from node i in graph G. Denote $H_{\gamma;ij} := G(B_\gamma(i) \cap B_\gamma(j))$ as the induced subgraph of G over the intersection of γ-hop neighborhoods at i and j and retaining the nodes in $V \setminus \{B_\gamma(i) \cap B_\gamma(j)\}$. Thus, $H_{\gamma;ij}$ has the same number of nodes as G. We first make the following simple observation: the (i,j) element in the γ^{th} power of walk matrix, $R_G^\gamma(i,j)$, is given by walks of length γ between i and j on graph G and thus, depends only on subgraph[24] $H_{\gamma;ij}$, see (21). This enables us to quantify the effect of nodes outside $B_\gamma(i) \cap B_\gamma(j)$ on the covariance $\Sigma_G(i,j)$.

Define a new walk matrix $\mathbf{R}_{H_{\gamma;ij}}$ such that

$$R_{H_{\gamma;ij}}(a,b) = \begin{cases} R_G(a,b), & a,b \in B_\gamma(i) \cap B_\gamma(j), \\ 0, & \text{o.w.} \end{cases}$$

In other words, $\mathbf{R}_{H_{\gamma;ij}}$ is formed by considering the Gaussian graphical model over graph $H_{\gamma;ij}$. Let $\mathbf{\Sigma}_{H_{\gamma;ij}}$ denote the corresponding covariance matrix.[25]

Lemma 10 (Covariance Bounds Under Walk-summability) *For any walk-summable Gaussian graphical model* ($\alpha := \|\overline{\mathbf{R}}_G\| < 1$)*, we have*[26]

$$\max_{i,j} |\Sigma_G(i,j) - \Sigma_{H_{\gamma;ij}}(i,j)| \leq \alpha^\gamma \frac{2\alpha}{1-\alpha} = O(\alpha^\gamma). \tag{22}$$

Thus, for walk-summable Gaussian graphical models, we have $\alpha := \|\overline{\mathbf{R}}_G\| < 1$, implying that the error in (22) in approximating the covariance by local neighborhood decays exponentially with distance. Parts of the proof below are inspired by Dumitriu and Pal (2009).

Proof: Using the power-series in (20), we can write the covariance matrix as

$$\mathbf{\Sigma}_G = \sum_{k=0}^{\gamma} \mathbf{R}_G^k + \mathbf{E}_G,$$

where the error matrix $\mathbf{E}_G$ has spectral radius

$$\|\mathbf{E}_G\| \leq \frac{\|\mathbf{R}_G\|^{\gamma+1}}{1 - \|\mathbf{R}_G\|},$$

from (20). Thus,[27] for any $i, j \in V$,

$$|\Sigma_G(i,j) - \sum_{k=0}^{\gamma} R_G^k(i,j)| \leq \frac{\|\mathbf{R}_G\|^{\gamma+1}}{1 - \|\mathbf{R}_G\|}. \tag{23}$$

24. Note that $R^\gamma(i,j) = 0$ if $B_\gamma(i) \cap B_\gamma(j) = \emptyset$.
25. When $B_\gamma(i) \cap B_\gamma(j) = \emptyset$ meaning that graph distance between i and j is more than γ, we obtain $\mathbf{\Sigma}_{H_{\gamma;ij}} = \mathbf{I}$.
26. The bound in (22) also holds if $H_{\gamma;ij}$ is replaced with any of its supergraphs.
27. For any matrix $\mathbf{A}$, we have $\max_{i,j} |A(i,j)| \leq \|\mathbf{A}\|$.

Similarly, we have

$$|\Sigma_{H_{\gamma;ij}}(i,j) - \sum_{k=0}^{\gamma} R^k_{H_{\gamma;ij}}(i,j)| \leq \frac{\|\mathbf{R}_{H_{\gamma;ij}}\|^{\gamma+1}}{1-\|\mathbf{R}_{H_{\gamma;ij}}\|} \stackrel{(a)}{\leq} \frac{\|\overline{\mathbf{R}}_G\|^{\gamma+1}}{1-\|\overline{\mathbf{R}}_G\|}, \tag{24}$$

where for inequality (a), we use the fact that

$$\|\mathbf{R}_{H_{\gamma;ij}}\| \leq \|\overline{\mathbf{R}}_{H_{\gamma;ij}}\| \leq \|\overline{\mathbf{R}}_G\|,$$

since $H_{\gamma;ij}$ is a subgraph[28] of G.

Combining (23) and (24), using the triangle inequality, we obtain (22). □

We also make some simple observations about conditional covariances in walk-summable models. Recall that $\overline{\mathbf{R}}_G$ denotes matrix with absolute values of $\mathbf{R}_G$, and $\mathbf{R}_G$ is the walk matrix over graph G. Also recall that the α-walk summability condition in (7), is $\|\overline{\mathbf{R}}_G\| \leq \alpha < 1$.

Proposition 11 (Conditional Covariances under Walk-Summability) *Given a walk-summable Gaussian graphical model, for any $i, j \in V$ and $S \subset V$ with $i, j \notin S$, we have*

$$\Sigma(i,j|S) = \sum_{\substack{\mathbf{w}: i \to j \\ \forall k \in \mathbf{w}, k \notin S}} \phi_G(\mathbf{w}). \tag{25}$$

Moreover, we have

$$\sup_{\substack{i \in V \\ S \subset V \setminus i}} \Sigma(i,i|S) \leq (1-\alpha)^{-1} = O(1). \tag{26}$$

Proof: We have, from Rue and Held (2005, Thm. 2.5),

$$\Sigma(i,j|S) = J^{-1}_{-S,-S;G}(i,j),$$

where $\mathbf{J}_{-S,-S;G}$ denotes the submatrix of potential matrix $\mathbf{J}_G$ by deleting nodes in S. Since submatrix of a walk-summable matrix is walk-summable, we have (25) by appealing to the walk-sum expression for conditional covariances.

For (26), let $\|\mathbf{A}\|_\infty$ denote the maximum absolute value of entries in matrix $\mathbf{A}$. Using monotonicity of spectral norm and the fact that $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\|$, we have

$$\sup_{\substack{i \in V \\ S \subset V, i \notin V}} \Sigma(i,i|S) \leq \|\mathbf{J}^{-1}_{-S,-S;G}\| = (1-\|\mathbf{R}_{-S,-S;G}\|)^{-1} \leq (1-\|\overline{\mathbf{R}}_{-S,-S;G}\|)^{-1} \leq (1-\|\overline{\mathbf{R}}_G\|)^{-1} = O(1).$$

□

Thus, the conditional covariance in (25) consists of walks in the original graph G, not passing through nodes in S.

28. When two matrices $\mathbf{A}$ and $\mathbf{B}$ are such that $|A(i,j)| \geq |B(i,j)|$ for all i, j, we have $\|\mathbf{A}\| \geq \|\mathbf{B}\|$.

Appendix B. Graphs with Local-Separation Property

We now provide bounds on conditional covariance for walk-summable matrices.

B.1 Conditional Covariance between Non-Neighbors: Normalized Case

We now provide bounds on the conditional covariance for Gaussian graphical models Markov on a graph $G \sim \mathcal{G}(p;\eta,\gamma)$ satisfying the local-separation property (η,γ), as per Definition 2.

Lemma 12 (Conditional Covariance Between Non-neighbors) *For a walk-summable Gaussian graphical model, the conditional covariance between non-neighbors i and j, conditioned on S_γ, the γ-local separator between i and j, satisfies*

$$\max_{j \notin \mathcal{N}(i)} \Sigma(i;j|S_\gamma) = O(\|\overline{\mathbf{R}}_G\|^\gamma).$$

Proof: In this proof, we abbreviate S_γ by S for notational convenience. The conditional covariance is given by the Schur complement, that is, for any subset A such that $A \cap S = \emptyset$,

$$\Sigma(A|S) = \Sigma(A,A) - \Sigma(A,S)\Sigma(S,S)^{-1}\Sigma(S,A). \tag{27}$$

We use the notation $\mathbf{\Sigma}_G(A,A)$ to denote the submatrix of the covariance matrix $\mathbf{\Sigma}_G$, when the underlying graph is G. As in Lemma 10, we may decompose $\mathbf{\Sigma}_G$ as follows:

$$\mathbf{\Sigma}_G = \mathbf{\Sigma}_{H_\gamma} + \mathbf{E}_\gamma,$$

where H_γ is the subgraph spanned by γ-hop neighborhood $B_\gamma(i)$, and $\mathbf{E}_\gamma$ is the error matrix. Let $\mathbf{F}_\gamma$ be the matrix such that

$$\mathbf{\Sigma}_G(S,S)^{-1} = \mathbf{\Sigma}_{H_\gamma}(S,S)^{-1} + \mathbf{F}_\gamma.$$

We have $\Sigma_{H_\gamma}(i,j|S) = 0$, where $\Sigma_{H_\gamma}(i,j|S)$ denotes the conditional covariance by considering the model given by the subgraph H_γ. This is due to the Markov property since i and j are separated by S in the subgraph H_γ.

Thus using (27), the conditional covariance on graph G can be bounded as

$$\Sigma_G(i,j|S) = O(\max(\|\mathbf{E}_\gamma\|, \|\mathbf{F}_\gamma\|)).$$

By Lemma 10, we have $\|\mathbf{E}_\gamma\| = O(\|\overline{\mathbf{R}}_G\|^\gamma)$. Using Woodbury matrix-inversion identity, we also have $\|\mathbf{F}_\gamma\| = O(\|\overline{\mathbf{R}}_G\|^\gamma)$. □

B.2 Extension to General Precision Matrices: Unnormalized Case

We now extend the above analysis to general precision matrices $\mathbf{J}$ where the diagonal elements are not assumed to be identity. Denote the precision matrix as

$$\mathbf{J} = \mathbf{D} - \mathbf{E},$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{E}$ has zero diagonal elements. We thus have that

$$\mathbf{J}_{\text{norm}} := \mathbf{D}^{-0.5}\mathbf{J}\mathbf{D}^{-0.5} = \mathbf{I} - \mathbf{R},$$

where $\mathbf{R}$ is the partial correlation matrix. This also implies that

$$\mathbf{J} = \mathbf{D}^{0.5}\mathbf{J}_{\text{norm}}\mathbf{D}^{0.5}.$$

Thus, we have that

$$\mathbf{\Sigma} = \mathbf{D}^{-0.5}\mathbf{\Sigma}_{\text{norm}}\mathbf{D}^{-0.5}, \tag{28}$$

where $\mathbf{\Sigma}_{\text{norm}} := \mathbf{J}_{\text{norm}}^{-1}$ is the covariance matrix corresponding to the normalized model. When the model is walk-summable, that is, $\|\overline{\mathbf{R}}\| \leq \alpha < 1$, we have that $\mathbf{\Sigma}_{\text{norm}} = \sum_{k\geq 0}\mathbf{R}^k$.

We now use the results derived in the previous sections involving the normalized model (Lemma 10 and Lemma 12) to obtain bounds for general precision matrices.

Lemma 13 (Covariance Bounds for General Models) *For any walk-summable Gaussian graphical model* ($\alpha := \|\overline{\mathbf{R}}_G\| < 1$)*, we have the following results:*

1. Covariance Bounds: *The covariance entries upon limiting to a subgraph $H_{\gamma,ij}$ for any $i, j \in V$ satisfies*

$$\max_{i,j} |\Sigma_G(i,j) - \Sigma_{H_{\gamma,ij}}(i,j)| \leq \frac{\alpha^\gamma}{D_{\min}}\frac{2\alpha}{1-\alpha} = O\left(\frac{\alpha^\gamma}{D_{\min}}\right), \tag{29}$$

   where $D_{\min} := \min_i D(i,i) = \min_i J(i,i)$.

2. Conditional Covariance between Non-neighbors: *The conditional covariance between non-neighbors i and j, conditioned on S_γ, the γ-local separator between i and j, satisfies*

$$\max_{j\notin\mathcal{N}(i)} \Sigma(i;j|S_\gamma) = O\left(\frac{\alpha^\gamma}{D_{\min}}\right), \tag{30}$$

   where $D_{\min} := \min_i D(i,i) = \min_i J(i,i)$.

Proof: Using (28) and Lemma 10, we have (29). Similarly, it can be shown that for any $S \subset V \setminus \{i,j\}$, $i, j \in V$,

$$\Sigma(i,j|S) = \mathbf{D}^{-0.5}\Sigma_{\text{norm}}(i,j|S)\mathbf{D}^{-0.5},$$

where $\Sigma_{\text{norm}}(i,j|S)$ is the conditional covariance corresponding to the model with normalized precision matrix. From Lemma 12, we have (30). □

B.3 Conditional Covariance between Neighbors: General Case

We provide a lower bound on conditional covariance among the neighbors for the graphs under consideration. Recall that $J_{\min}$ denotes the minimum edge potentials. Let

$$K(i,j) := \|\mathbf{J}(V\setminus\{i,j\},\{i,j\})\|^2,$$

where $\mathbf{J}(V\setminus\{i,j\},\{i,j\})$ is a sub-matrix of the potential matrix $\mathbf{J}$.

Lemma 14 (Conditional Covariance Between Neighbors) *For an α-walk summable Gaussian graphical model satisfying*

$$D_{\min}(1-\alpha)\min_{(i,j)\in G_p}\frac{J(i,j)}{K(i,j)} > 1+\delta, \tag{31}$$

for some $\delta > 0$ (not depending on p), where $D_{\min} := \min_i J(i,i)$, we have

$$|\Sigma_G(i,j|S)| = \Omega(J_{\min}),$$

for any $(i,j) \in G$ such that $j \in \mathcal{N}(i)$ and any subset $S \subset V$ with $i,j \notin S$.

Proof: First note that for attractive models,

$$\begin{aligned}\Sigma_G(i,j|S) &\overset{(a)}{\geq} \Sigma_{G_1}(i,j|S) \\ &\overset{(b)}{=} \frac{-J(i,j)}{J(i,i)J(j,j)-J(i,j)^2} = \Omega(J_{\min}),\end{aligned}$$

where G_1 is the graph consisting only of edge (i,j). Inequality (a) arises from the fact that in attractive models, the weights of all the walks are positive, and thus, the weight of walks on G_1 form a lower bound for those on G (recall that the covariances are given by the sum-weight of walks on the graphs). Equality (b) is by direct matrix inversion of the model on G_1.

For general models, we need further analysis. Let $A = \{i,j\}$ and $B = V \setminus \{S \cup A\}$, for some $S \subset V \setminus A$. Let $\mathbf{\Sigma}(A,A)$ denote the covariance matrix on set A, and let $\widetilde{\mathbf{J}}(A,A) := \mathbf{\Sigma}(A,A)^{-1}$ denote the corresponding marginal potential matrix. We have for all $S \subset V \setminus A$

$$\widetilde{\mathbf{J}}(A,A) = \mathbf{J}(A,A) - \mathbf{J}(A,B)\mathbf{J}(B,B)^{-1}\mathbf{J}(B,A).$$

Recall that $\|\mathbf{A}\|_\infty$ denotes the maximum absolute value of entries in matrix $\mathbf{A}$.

$$\begin{aligned}\|\mathbf{J}(A,B)\mathbf{J}(B,B)^{-1}\mathbf{J}(B,A)\|_\infty &\overset{(a)}{\leq} \|\mathbf{J}(A,B)\mathbf{J}(B,B)^{-1}\mathbf{J}(B,A)\| \\ &\overset{(b)}{\leq} \|\mathbf{J}(A,B)\|^2 \|\mathbf{J}(B,B)^{-1}\| \\ &= \frac{\|\mathbf{J}(A,B)\|^2}{\lambda_{\min}(\mathbf{J}(B,B))}, \\ &\overset{(c)}{\leq} \frac{K(i,j)^2}{D_{\min}(1-\alpha)}\end{aligned}$$

where inequality (a) arises from the fact that the ℓ_∞ norm is bounded by the spectral norm, (b) arises from sub-multiplicative property of norms and (c) arises from walk-summability property. Inequality (b) is from the bound on edge potentials and α-walk summability of the model and since $K(i,j) \geq \|\mathbf{J}(A,B)\|$. Assuming (31), we have

$$|\widetilde{J}(i,j)| > J_{\min} - \frac{\|\mathbf{J}(A,B)\|^2}{D_{\min}(1-\alpha)} = \Omega(J_{\min}).$$

Since

$$\Sigma_G(i,j|S) = \frac{-\widetilde{J}(i,j)}{\widetilde{J}(i,i)\widetilde{J}(j,j)-\widetilde{J}(i,j)^2},$$

we have the result. □

B.4 Analysis of Loopy Belief Propagation

Proof of Proposition 9: From Lemma 10 in Section A.3, for any α-walk-summable Gaussian graphical model, we have, for all nodes $i \in V$ conditioned on the event $\Gamma(i;\gamma,G)$,

$$|\Sigma_G(i,i) - \Sigma_{\mathrm{LBP}}(i,i)| = O(\|\overline{\mathbf{R}}_G\|^{\gamma}).$$

This is because conditioned on $\Gamma(i;\gamma,G)$, it is shown that the series expansions based on walk-sums corresponding to the variances $\Sigma_{H_{\gamma;ij}}(i,i)$ and $\Sigma_{\mathrm{LBP}}(i,i)$ are identical up to length γ walks, and the effect of walks beyond length γ can be bounded as above. Moreover, for a sequence of α-walk-summable, we have $\Sigma(i,i) \le M$ for all $i \in V$, for some constant M and similarly $\Sigma_{\mathrm{LBP}}(i,j) \le M'$ for some constant M' since it is obtained by the set of self-avoiding walks in G. We thus have

$$\mathbb{E}\left[|\Sigma_G(i,i) - \Sigma_{\mathrm{LBP}}(i,i)|\right] \le \left[O(\|\overline{\mathbf{R}}_G\|^{\gamma}) + P[\Gamma^c(i;\gamma)]\right] = o(1),$$

where $\mathbb{E}$ is over the expectation of ensemble $\mathcal{G}(p)$. By Markov's inequality,[29] we have the result. $\square$

Appendix C. Sample-based Analysis

We now extend our analysis to the setting where we have access to samples instead of exact statistics.

C.1 Concentration of Empirical Quantities

For our sample complexity analysis, we recap the concentration result by Ravikumar et al. (2011, Lemma 1) for sub-Gaussian matrices and specialize it to Gaussian matrices.

Lemma 15 (Concentration of Empirical Covariances) *For any p-dimensional Gaussian random vector $\mathbf{X} = [X_1,\ldots,X_p]$, the empirical covariance obtained from n samples satisfies*

$$P\left[|\widehat{\Sigma}(i,j) - \Sigma(i,j)| > \varepsilon\right] \le 4\exp\left[-\frac{n\varepsilon^2}{3200M^2}\right], \tag{32}$$

for all $\varepsilon \in (0, 40M)$ and $M := \max_i \Sigma(i,i)$.

This translates to bounds for empirical conditional covariance.

Corollary 16 (Concentration of Empirical Conditional Covariance) *For a walk-summable p-dimensional Gaussian random vector $\mathbf{X} = [X_1,\ldots,X_p]$, we have*

$$P\left[\max_{\substack{i\neq j\\ S\subset V,|S|\le\eta}} |\widehat{\Sigma}(i,j|S) - \Sigma(i;j|S)| > \varepsilon\right] \le 4p^{\eta+2}\exp\left(-\frac{n\varepsilon^2}{K}\right), \tag{33}$$

where $K \in (0,\infty)$ is a constant which is bounded when $\|\mathbf{\Sigma}\|_\infty$ is bounded, for all $\varepsilon \in (0, 40M)$ with $M := \max_i \Sigma(i,i)$, and $n \ge \eta$.

29. By Markov's inequality, for a non-negative random variable X, we have $P[X > \delta] \le \mathbb{E}[X]/\delta$. By choosing $\delta = \omega(\mathbb{E}[X])$, we have the result.

Proof: For a given $i, j \in V$ and $S \subset V$ with $\eta \leq n$, using (27),

$$P\left[|\widehat{\Sigma}(i,j|S) - \Sigma(i;j|S)| > \varepsilon\right] \leq \mathbb{P}\left[\left(|\widehat{\Sigma}(i,j) - \Sigma(i;j)| > \varepsilon\right) \bigcup_{k \in S} \left(|\widehat{\Sigma}(i,k) - \Sigma(i;k)| > K'\varepsilon\right)\right],$$

where K' is a constant which is bounded when $\|\mathbf{\Sigma}\|_\infty$ is bounded. Using Lemma 15, we have the result. □

C.2 Proof of Theorem 4

We are now ready to prove Theorem 4. We analyze the error events for the conditional covariance threshold test CMIT. For any $(i,j) \notin G_p$, define the event

$$\mathcal{F}_1(i,j;\{\mathbf{x}^n\},G_p) := \left\{|\widehat{\Sigma}(i,j|S)| > \xi_{n,p}\right\},$$

where $\xi_{n,p}$ is the threshold in (10) and S is the γ-local separator between i and j (since the minimum in (1) is achieved by the γ-local separator). Similarly for any edge $(i,j) \in G_p$, define the event that

$$\mathcal{F}_2(i,j;\{\mathbf{x}^n\},G_p) := \left\{\exists S \subset V : |S| \leq \eta, |\widehat{\Sigma}(i,j|S)| < \xi_{n,p}\right\}.$$

The probability of error resulting from CMIT can thus be bounded by the two types of errors,

$$\mathbb{P}[\mathsf{CMIT}(\{\mathbf{x}^n\};\xi_{n,p}) \neq G_p] \leq \mathbb{P}\left[\bigcup_{(i,j)\in G_p} \mathcal{F}_2(i,j;\{\mathbf{x}^n\},G_p)\right] + \mathbb{P}\left[\bigcup_{(i,j)\notin G_p} \mathcal{F}_1(i,j;\{\mathbf{x}^n\},G_p)\right] \tag{34}$$

For the first term, applying union bound for both the terms and using the result (33) of Lemma 15,

$$\mathbb{P}\left[\bigcup_{(i,j)\in G_p} \mathcal{F}_2(i,j;\{\mathbf{x}^n\},G_p)\right] = O\left(p^{\eta+2}\exp\left[-\frac{n(C_{\min}(p)-\xi_{n,p})^2}{K^2}\right]\right) \tag{35}$$

where

$$C_{\min}(p) := \inf_{\substack{(i,j)\in G_p \\ S\subset V, i,j\notin S \\ |S|\leq\eta}} |\Sigma(i,j|S)| = \Omega\left(J_{\min}\right), \quad \forall\, p \in \mathbb{N},$$

from (37). Since $\xi_{n,p} = o(J_{\min})$, (35) is $o(1)$ when $n > L\log p/J_{\min}^2$, for sufficiently large L (depending on η and M). For the second term in (34),

$$\mathbb{P}\left[\bigcup_{(i,j)\notin G_p} \mathcal{F}_1(i,j;\{\mathbf{x}^n\},G_p)\right] = O\left(p^{\eta+2}\exp\left[-\frac{n(\xi_{n,p}-C_{\max}(p))^2}{K^2}\right]\right), \tag{36}$$

where

$$C_{\max}(p) := \max_{(i,j)\notin G_p} |\Sigma(i,j|S)| = O\left(\frac{\alpha^\gamma}{D_{\min}}\right).$$

For the choice of $\xi_{n,p}$ in (10), (36) is $o(1)$ and this completes the proof of Theorem 4. □

C.3 Conditional Mutual Information Thresholding Test

We now analyze the performance of conditional mutual information threshold test. We first note bounds on conditional mutual information.

Proposition 17 (Conditional Mutual Information) *Under the assumptions (A1)–(A5), we have that the conditional mutual information among non-neighbors, conditioned on the γ-local separation satisfies*

$$\max_{(i,j)\notin G} I(X_i;X_j|\mathbf{X}_{S_\gamma}) = O(\alpha^{2\gamma}),$$

and the conditional mutual information among the neighbors satisfy

$$\min_{\substack{(i,j)\in G \\ S\subset V\setminus\{i,j\}}} I(X_i;X_j|\mathbf{X}_S) = \Omega(J_{\min}^2). \tag{37}$$

Proof: The conditional mutual information for Gaussian variables is given by

$$I(X_i;X_j|\mathbf{X}_S) = -\frac{1}{2}\log\left[1-\rho^2(i,j|S)\right],$$

where $\rho(i,j|S)$ is the conditional correlation coefficient, given by

$$\rho(i,j|S) := \frac{\Sigma(i,j|S)}{\sqrt{\Sigma(i,i|S)\Sigma(j,j|S)}}.$$

From (26) in Proposition 11, we have $\Sigma(i,i|S) = O(1)$ and thus, the result holds. □

We now note the concentration bounds on empirical mutual information.

Lemma 18 (Concentration of Empirical Mutual Information) *For any p-dimensional Gaussian random vector $\mathbf{X} = [X_1,\ldots,X_p]$, the empirical mutual information obtained from n samples satisfies*

$$P(|\widehat{I}(X_i;X_j) - I(X_i;X_j)| > \varepsilon) \leq 24\exp\left(-\frac{nM\varepsilon^2}{204800L^2}\right), \tag{38}$$

for some constant L which is finite when $\rho_{\max} := \max_{i\neq j}|\rho(i,j)| < 1$, and all $\varepsilon < \rho_{\max}$, and for $M := \max_i \Sigma(i,i)$.

Proof: The result on empirical covariances can be found in Ravikumar et al. (2011, Lemma 1). The result in (38) will be shown through a sequence of transformations. First, we will bound

$P(|\widehat{\rho}(i,j)-\rho(i,j)|>\varepsilon)$. Consider,

$$
\begin{aligned}
&P(|\widehat{\rho}(i,j)-\rho(i,j)|>\varepsilon)\\
&=P\left(\left|\frac{\widehat{\Sigma}(i,j)}{(\widehat{\Sigma}(i,i)\widehat{\Sigma}(j,j))^{1/2}}-\frac{\Sigma(i,j)}{(\Sigma(i,i)\Sigma(j,j))^{1/2}}\right|>\varepsilon\right)\\
&=P\left(\left|\frac{\widehat{\Sigma}(i,j)}{\Sigma(i,j)}\left(\frac{\Sigma(i,i)}{\widehat{\Sigma}(i,i)}\frac{\Sigma(j,j)}{\widehat{\Sigma}(j,j)}\right)^{1/2}-1\right|>\frac{\varepsilon}{|\rho(i,j)|}\right)\\
&\overset{(a)}{\leq}P\left(\frac{\widehat{\Sigma}(i,j)}{\Sigma(i,j)}>\left(1+\frac{\varepsilon}{|\rho(i,j)|}\right)^{1/3}\right)+P\left(\frac{\widehat{\Sigma}(i,j)}{\Sigma(i,j)}<\left(1-\frac{\varepsilon}{|\rho(i,j)|}\right)^{1/3}\right)+\dots\\
&\quad+P\left(\frac{\Sigma(i,i)}{\widehat{\Sigma}(i,i)}>\left(1+\frac{\varepsilon}{|\rho(i,j)|}\right)^{2/3}\right)+P\left(\frac{\Sigma(i,i)}{\widehat{\Sigma}(i,i)}<\left(1-\frac{\varepsilon}{|\rho(i,j)|}\right)^{2/3}\right)+\dots\\
&\quad+P\left(\frac{\Sigma(j,j)}{\widehat{\Sigma}(j,j)}>\left(1+\frac{\varepsilon}{|\rho(i,j)|}\right)^{2/3}\right)+P\left(\frac{\Sigma(j,j)}{\widehat{\Sigma}(j,j)}<\left(1-\frac{\varepsilon}{|\rho(i,j)|}\right)^{2/3}\right)\\
&\overset{(b)}{\leq}P\left(\frac{\widehat{\Sigma}(i,j)}{\Sigma(i,j)}>1+\frac{\varepsilon}{8|\rho(i,j)|}\right)+P\left(\frac{\widehat{\Sigma}(i,j)}{\Sigma(i,j)}<1-\frac{\varepsilon}{8|\rho(i,j)|}\right)+\dots\\
&\quad+P\left(\frac{\Sigma(i,i)}{\widehat{\Sigma}(i,i)}>1+\frac{\varepsilon}{3|\rho(i,j)|}\right)+P\left(\frac{\widehat{\Sigma}(i,i)}{\Sigma(i,i)}<1-\frac{\varepsilon}{3|\rho(i,j)|}\right)+\dots\\
&\quad+P\left(\frac{\widehat{\Sigma}(j,j)}{\Sigma(j,j)}>1+\frac{\varepsilon}{3|\rho(i,j)|}\right)+P\left(\frac{\widehat{\Sigma}(j,j)}{\Sigma(j,j)}<1-\frac{\varepsilon}{3|\rho(i,j)|}\right)\\
&\overset{(c)}{\leq}24\exp\left(-\frac{nM\varepsilon^2}{204800|\rho(i,j)|^2}\right)\overset{(d)}{\leq}24\exp\left(-\frac{nM\varepsilon^2}{204800}\right)
\end{aligned}
$$

where in (a), we used the fact that $P(ABC>1+\delta)\leq P(A>(1+\delta)^{1/3}$ or $B>(1+\delta)^{1/3}$ or $C>(1+\delta)^{1/3})$ and the union bound, in (b) we used the fact that $(1+\delta)^3\leq 1+8\delta$ and $(1+\delta)^{-2/3}\leq 1-\delta/3$ for $\delta=\varepsilon/|\rho(i,j)|<1$. Finally, in (c), we used the result in (32) and in (d), we used the bounds on $\rho<1$.

Now, define the bijective function $I(|\rho|):=-1/2\log(1-\rho^2)$. Then we claim that there exists a constant $L\in(0,\infty)$, depending only on $\rho_{\max}<1$, such that

$$|I(x)-I(y)|\leq L|x-y|, \tag{39}$$

that is, the function $I:[0,\rho_{\max}]\to\mathbb{R}^+$ is $L=L(\rho_{\max})$-Lipschitz. This is because the slope of the function I is bounded in the interval $[0,\rho_{\max}]$. Thus, we have the inclusion

$$\{|\widehat{I}(X_i;X_j)-I(X_i;X_j)|>\varepsilon\}\subset\{|\widehat{\rho}(i,j)-\rho(i,j)|>\varepsilon/L\} \tag{40}$$

since if $|\widehat{I}(X_i;X_j)-I(X_i;X_j)|>\varepsilon$ it is true that $L|\widehat{\rho}(i,j)-\rho(i,j)|>\varepsilon$ from (39). We have by monotonicity of measure and (40) the desired result. □

We can now obtain the desired result on concentration of empirical conditional mutual information.

Lemma 19 (Concentration of Empirical Conditional Mutual Information) *For a walk-summable p-dimensional Gaussian random vector* $\mathbf{X} = [X_1, \ldots, X_p]$*, we have*

$$P\left[\max_{\substack{i\neq j \\ S\subset V\setminus\{i,j\},|S|\leq\eta}} |\widehat{I}(X_i;X_j|\mathbf{X}_S) - I(X_i;X_j|\mathbf{X}_S)| > \varepsilon\right] \leq 24p^{\eta+2}\exp\left(-\frac{nM\varepsilon^2}{204800L^2}\right),$$

for constants $M, L \in (0,\infty)$ *and all* $\varepsilon < \rho_{\max}$*, where* $\rho_{\max} := \max_{\substack{i\neq j \\ S\subset V\setminus\{i,j\},|S|\leq\eta}} |\rho(i,j|S)|$.

Proof: Since the model is walk-summable, we have that $\max_{i,S}\Sigma(i,i|S) = O(1)$ and thus, the constant M is bounded. Similarly, due to strict positive-definiteness we have $\rho_{\max} < 1$ even as $p \to \infty$, and thus, the constant L is also finite. The result then follows from union bound. □

The sample complexity for structural consistency of CMIT follows on lines of analysis for CMIT.

Appendix D. Necessary Conditions for Model Selection

We now provide proofs for necessary conditions for model selection.

D.1 Necessary Conditions for Exact Recovery

We provide the proof of Theorem 6 in this section. We collect four auxiliary lemmata whose proofs (together with the proof of Lemma 8) will be provided at the end of the section. For information-theoretic notation, the reader is referred to Cover and Thomas (2006).

Lemma 20 (Upper Bound on Differential Entropy of Mixture) *Let* $\alpha < 1$*. Suppose asymptotically almost surely each precision matrix* $\mathbf{J}_G = \mathbf{I} - \mathbf{R}_G$ *satisfies* (7)*, that is, that* $\|\overline{\mathbf{R}}_G\| \leq \alpha$ *for a.e.* $G \in \mathcal{G}(p)$*. Then, for the Gaussian model, we have*

$$h(\mathbf{X}^n) \leq \frac{pn}{2}\log_2\left(\frac{2\pi e}{1-\alpha}\right),$$

where recall that $\mathbf{X}^n|G \sim \prod_{i=1}^n f(\mathbf{x}_i|G)$*.*

For the sake of convenience, we define the random variable:

$$W = \begin{cases} 1 & G \in \mathcal{T}_\varepsilon^{(p)} \\ 0 & G \notin \mathcal{T}_\varepsilon^{(p)} \end{cases}.$$

The random variable W indicates whether $G \in \mathcal{T}_\varepsilon^{(p)}$.

Lemma 21 (Lower Bound on Conditional Differential Entropy) *Suppose that each precision matrix* $\mathbf{J}_G$ *has unit diagonal. Then,*

$$h(\mathbf{X}^n|G,W) \geq -\frac{pn}{2}\log_2(2\pi e).$$

Lemma 22 (Conditional Fano Inequality) *In the above notation, we have*

$$\frac{H(G|\mathbf{X}^n, G \in \mathcal{T}_\varepsilon^{(p)}) - 1}{\log_2(|\mathcal{T}_\varepsilon^{(p)}| - 1)} \leq P(\widehat{G}(\mathbf{X}^n) \neq G | G \in \mathcal{T}_\varepsilon^{(p)}).$$

Lemma 23 (Exponential Decay in Probability of Atypical Set) *Define the rate function* $K(c,\varepsilon) := \frac{c}{2}[(1+\varepsilon)\ln(1+\varepsilon) - \varepsilon]$. *The probability of the* ε*-atypical set decays as*

$$P((\mathcal{T}_\varepsilon^{(p)})^c) = P(G \notin \mathcal{T}_\varepsilon^{(p)}) \leq 2\exp(-pK(c,\varepsilon)) \tag{41}$$

for all $p \geq 1$.

Note the non-asymptotic nature of the bound in (41). The rate function $K(c,\varepsilon)$ satisfies $\lim_{\varepsilon \downarrow 0} K(c,\varepsilon)/\varepsilon^2 = c/4$. We prove Theorem 6 using these lemmata.

Proof: Consider the following sequence of lower bounds:

$$\begin{aligned}
\frac{pn}{2}\log_2\left(\frac{2\pi e}{1-\alpha}\right) &\overset{(a)}{\geq} h(\mathbf{X}^n) \\
&\overset{(b)}{\geq} h(\mathbf{X}^n|W) \\
&= I(\mathbf{X}^n; G|W) + h(\mathbf{X}^n|G,W) \\
&\overset{(c)}{\geq} I(\mathbf{X}^n; G|W) - \frac{pn}{2}\log_2(2\pi e) \\
&= H(G|W) - H(G|\mathbf{X}^n, W) - \frac{pn}{2}\log_2(2\pi e),
\end{aligned} \tag{42}$$

where (a) follows from Lemma 20, (b) is because conditioning does not increase differential entropy and (c) follows from Lemma 21. We will lower bound the first term in (42) and upper bound the second term in (42). Now consider the first term in (42):

$$\begin{aligned}
H(G|W) &= H(G|W=1)P(W=1) + H(G|W=0)P(W=0) \\
&\overset{(a)}{\geq} H(G|W=1)P(W=1) \\
&\overset{(b)}{\geq} H(G|G \in \mathcal{T}_\varepsilon^{(p)})(1-\varepsilon) \\
&\overset{(c)}{\geq} (1-\varepsilon)\binom{p}{2} H_b\left(\frac{c}{p}\right),
\end{aligned} \tag{43}$$

where (a) is because the entropy $H(G|W=0)$ and the probability $P(W=0)$ are both non-negative. Inequality (b) follows for all p sufficiently large from the definition of W as well as Lemma 8 part 1. Statement (c) comes from fact that

$$\begin{aligned}
H(G|G \in \mathcal{T}_\varepsilon^{(p)}) &= -\sum_{g \in \mathcal{T}_\varepsilon^{(p)}} P(g|g \in \mathcal{T}_\varepsilon^{(p)}) \log_2 P(g|g \in \mathcal{T}_\varepsilon^{(p)}) \\
&\geq -\sum_{g \in \mathcal{T}_\varepsilon^{(p)}} P(g|g \in \mathcal{T}_\varepsilon^{(p)}) \left[-\binom{p}{2} H_b\left(\frac{c}{p}\right)\right] = \binom{p}{2} H_b\left(\frac{c}{p}\right).
\end{aligned}$$

We are now done bounding the first term in the difference in (42).

Now we will bound the second term in (42). First we will derive a bound on $H(G|\mathbf{X}^n, W=1)$. Consider,

$$\begin{aligned}
P_e^{(p)} &:= P(\widehat{G}(\mathbf{X}^n) \neq G) \\
&\stackrel{(a)}{=} P(\widehat{G}(\mathbf{X}^n) \neq G | W=1)P(W=1) + P(\widehat{G}(\mathbf{X}^n) \neq G | W=0)P(W=0) \\
&\geq P(\widehat{G}(\mathbf{X}^n) \neq G | W=1)P(W=1) \\
&\stackrel{(b)}{\geq} P(\widehat{G}(\mathbf{X}^n) \neq G | G \in \mathcal{T}_\varepsilon^{(p)}) \left(\frac{1}{1+\varepsilon}\right) \\
&\stackrel{(c)}{\geq} \frac{H(G|\mathbf{X}^n, G \in \mathcal{T}_\varepsilon^{(p)}) - 1}{\log_2 |\mathcal{T}_\varepsilon^{(p)}|} \left(\frac{1}{1+\varepsilon}\right),
\end{aligned} \tag{44}$$

where (a) is by the law of total probability, (b) holds for all p sufficiently large by Lemma 8 part 1 and (c) is due to the conditional version of Fano's inequality (Lemma 22). Then, from (44), we have

$$\begin{aligned}
H(G|\mathbf{X}^n, W=1) &\leq P_e^{(p)}(1+\varepsilon)\log_2 |\mathcal{T}_\varepsilon^{(p)}| + 1 \\
&\leq P_e^{(p)}(1+\varepsilon)\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right) + 1.
\end{aligned} \tag{45}$$

Define the *rate function* $K(c,\varepsilon) := \frac{c}{2}[(1+\varepsilon)\ln(1+\varepsilon) - \varepsilon]$. Note that this function is positive whenever $c, \varepsilon > 0$. In fact it is monotonically increasing in both parameters. Now we use (45) to bound $H(G|\mathbf{X}^n, W)$:

$$\begin{aligned}
H(G|\mathbf{X}^n, W) &= H(G|\mathbf{X}^n, W=1)P(W=1) + H(G|\mathbf{X}^n, W=0)P(W=0) \\
&\stackrel{(a)}{\leq} H(G|\mathbf{X}^n, W=1) + H(G|\mathbf{X}^n, W=0)P(W=0) \\
&\stackrel{(b)}{\leq} H(G|\mathbf{X}^n, W=1) + H(G|\mathbf{X}^n, W=0)(2e^{-pK(c,\varepsilon)}) \\
&\stackrel{(c)}{\leq} H(G|\mathbf{X}^n, W=1) + p^2(2e^{-pK(c,\varepsilon)}) \\
&\stackrel{(d)}{\leq} P_e^{(p)}(1+\varepsilon)\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right) + 1 + 2p^2 e^{-pK(c,\varepsilon)},
\end{aligned} \tag{46}$$

where (a) is because we upper bounded $P(W=1)$ by unity, (b) follows by Lemma 23, (c) follows by upper bounding the conditional entropy by p^2 and (d) follows from (45).

Substituting (43) and (46) back into (42) yields

$$\begin{aligned}
\frac{pn}{2}\log_2\left[2\pi e\left(\frac{1}{1-\alpha}+1\right)\right] &\geq (1-\varepsilon)\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right) - P_e^{(p)}(1+\varepsilon)\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right) - 1 - 2p^2 e^{-pK(c,\varepsilon)} \\
&= \binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right)\left[(1-\varepsilon) - P_e^{(p)}(1+\varepsilon)\right] - \Theta(p^2 e^{-pK(c,\varepsilon)}),
\end{aligned}$$

which implies that

$$n \geq \frac{2}{p\log_2\left[2\pi e\left(\frac{1}{1-\alpha}+1\right)\right]}\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right)\left[(1-\varepsilon) - P_e^{(p)}(1+\varepsilon)\right] - \Theta(pe^{-pK(c,\varepsilon)}).$$

Note that $\Theta(pe^{-pK(c,\varepsilon)}) \to 0$ as $p \to \infty$ since the rate function $K(c,\varepsilon)$ is positive. If we impose that $P_e^{(p)} \to 0$ as $p \to \infty$, then n has to satisfy (14) by the arbitrariness of $\varepsilon > 0$. This completes the proof of Theorem 6. □

D.2 Proof of Lemma 8

Proof: Part 1 follows directed from the law of large numbers. Part 2 follows from the fact that the Binomial pmf is maximized at its mean. Hence, for $G \in \mathcal{T}_\varepsilon^{(p)}$, we have

$$P(G) \leq \left(\frac{c}{p}\right)^{cp/2} \left(1-\frac{c}{p}\right)^{\binom{p}{2}-cp/2}.$$

We arrive at the upper bound after some rudimentary algebra. The lower bound can be proved by observing that for $G \in \mathcal{T}_\varepsilon^{(p)}$, we have

$$\begin{aligned} P(G) &\geq \left(\frac{c}{p}\right)^{cp(1+\varepsilon)/2} \left(1-\frac{c}{p}\right)^{\binom{p}{2}-cp(1+\varepsilon)/2} \\ &= \exp_2\left[\binom{p}{2}(\frac{c}{p}\log_2\frac{c}{p})(1+\varepsilon) + [1-c(1+\varepsilon)/p]\log_2(1-\frac{c}{p})\right] \\ &\geq \exp_2\left[\binom{p}{2}(\frac{c}{p}\log_2\frac{c}{p})(1+\varepsilon) + (1+\varepsilon)(1-\frac{c}{p})\log_2(1-\frac{c}{p})\right]. \end{aligned}$$

The result in Part 2 follows immediately by appealing to the symmetry of the binomial pmf about its mean. Part 3 follows by the following chain of inequalities:

$$\begin{aligned} 1 = \sum_{G\in\mathfrak{G}_n} P(G) &\geq \sum_{G\in\mathcal{T}_\varepsilon^{(p)}} P(G) \geq \sum_{G\in\mathcal{T}_\varepsilon^{(p)}} \exp_2\left[-\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}(1+\varepsilon)\right)\right] \\ &= |\mathcal{T}_\varepsilon^{(p)}| \exp_2\left[-\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right)(1+\varepsilon)\right]. \end{aligned}$$

This completes the proof of the upper bound on $|\mathcal{T}_\varepsilon^{(p)}|$. The lower bound follows by noting that for sufficiently large n, $P(\mathcal{T}_\varepsilon^{(p)}) \geq 1-\varepsilon$ (by Lemma 8 Part 1). Thus,

$$1-\varepsilon \leq \sum_{G\in\mathcal{T}_\varepsilon^{(p)}} P(G) \leq \sum_{G\in\mathcal{T}_\varepsilon^{(p)}} \exp_2\left[-\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right)\right] = |\mathcal{T}_\varepsilon^{(p)}| \exp_2\left[-\binom{p}{2} H_{\mathrm{b}}\left(\frac{c}{p}\right)\right].$$

This completes the proof. □

D.3 Proof of Lemma 20

Proof: Note that the distribution of $\mathbf{X}$ (with G marginalized out) is a Gaussian mixture model given by $\sum_{G\in\mathfrak{G}_p} P(G)\mathcal{N}(\mathbf{0}, \mathbf{J}_G^{-1})$. As such the covariance matrix of $\mathbf{X}$ is given by

$$\mathbf{\Sigma}_{\mathbf{X}} = \sum_{G\in\mathfrak{G}_p} P(G)\mathbf{J}_G^{-1}. \tag{47}$$

This is not immediately obvious but it is due to the zero-mean nature of each Gaussian probability density function $\mathcal{N}(\mathbf{0}, \mathbf{J}_G^{-1})$. Using (47), we have the following chain of inequalities:

$$
\begin{aligned}
h(\mathbf{X}^n) &\leq nh(\mathbf{X}) \\
&\overset{(a)}{\leq} \frac{n}{2} \log_2 \left((2\pi e)^p \det(\mathbf{\Sigma_X}) \right) \\
&= \frac{n}{2} \left[p \log_2(2\pi e) + \log_2 \det(\mathbf{\Sigma_X}) \right] \\
&\overset{(b)}{\leq} \frac{n}{2} \left[p \log_2(2\pi e) + p \log_2 \lambda_{\max}(\mathbf{\Sigma_X}) \right] \\
&= \frac{n}{2} \left[p \log_2(2\pi e) + p \log_2 \lambda_{\max} \left(\sum_{G \in \mathfrak{G}_p} P(G) \mathbf{J}_G^{-1} \right) \right] \\
&\overset{(c)}{\leq} \frac{n}{2} \left[p \log_2(2\pi e) + p \log_2 \left(\sum_{G \in \mathfrak{G}_p} P(G) \lambda_{\max} \left(\mathbf{J}_G^{-1} \right) \right) \right] \\
&= \frac{n}{2} \left[p \log_2(2\pi e) + p \log_2 \left(\sum_{G \in \mathfrak{G}_p} P(G) \frac{1}{\lambda_{\min}(\mathbf{J}_G)} \right) \right] \\
&\overset{(d)}{\leq} \frac{n}{2} \left[p \log_2(2\pi e) + p \log_2 \left(\sum_{G \in \mathfrak{G}_p} P(G) \frac{1}{1-\alpha} \right) \right] \\
&= \frac{pn}{2} \log_2 \left(\frac{2\pi e}{1-\alpha} \right),
\end{aligned}
$$

where (a) uses the maximum entropy principle (Cover and Thomas, 2006, Chapter 13), that is, that the Gaussian maximizes entropy subject to an average power constraint (b) uses the fact that the determinant of $\mathbf{\Sigma_X}$ is upper bounded by $\lambda_{\max}(\mathbf{\Sigma_X})^n$, (c) uses the convexity of $\lambda_{\max}(\cdot)$ (it equals to the operator norm $\|\cdot\|_2$ over the set of symmetric matrices, (d) uses the fact that $\alpha \geq \|\overline{\mathbf{R}}_G\|_2 \geq \|\mathbf{R}_G\|_2 = \|\mathbf{I} - \mathbf{J}_G\|_2 = \lambda_{\max}(\mathbf{I} - \mathbf{J}_G) = 1 - \lambda_{\min}(\mathbf{J}_G)$ a.a.s. This completes the proof. □

D.4 Proof of Lemma 21

Proof: Firstly, we lower bound $h(\mathbf{X}^n|G,W=1)$ as follows:

$$\begin{aligned}
h(\mathbf{X}^n|G) &= \sum_{g\in\mathfrak{G}_p} P(g)h(\mathbf{X}^n|G=g) \\
&\overset{(a)}{=} n\sum_{g\in\mathfrak{G}_p} P(g)h(\mathbf{X}|G=g) \\
&\overset{(b)}{=} \frac{n}{2}\sum_{g\in\mathfrak{G}_p} P(g)\log_2[(2\pi e)^p \det(\mathbf{J}_g^{-1})] \\
&= -\frac{n}{2}\sum_{g\in\mathfrak{G}_p} P(g)\log_2[(2\pi e)^p \det(\mathbf{J}_g)] \\
&\overset{(c)}{\geq} -\frac{n}{2}\sum_{g\in\mathfrak{G}_p} P(g)\log_2[(2\pi e)^p] \\
&\geq -\frac{pn}{2}\log_2(2\pi e),
\end{aligned}$$

where (a) is because the samples in $\mathbf{X}^n$ are conditionally independent given $G=g$, (b) is by the Gaussian assumption, (c) is by Hadamard's inequality

$$\det(\mathbf{J}_g) \leq \prod_{i=1}^{p}[\mathbf{J}_g]_{ii} = 1$$

and the assumption that each diagonal element of each precision matrix $\mathbf{J}_g = \mathbf{I} - \mathbf{R}_g$ is equal to 1 a.a.s. This proves the claim. □

D.5 Proof of Lemma 22

Proof: Define the "error" random variable

$$E = \begin{cases} 1 & \widehat{G}(\mathbf{X}^n) \neq G \\ 0 & \widehat{G}(\mathbf{X}^n) = G \end{cases}.$$

Now consider

$$\begin{aligned}
H(E,G|\mathbf{X}^n,W=1) &= H(E|\mathbf{X}^n,W=1) + H(G|E,\mathbf{X}^n,W=1) &(48)\\
&= H(G|\mathbf{X}^n,W=1) + H(E|G,\mathbf{X}^n,W=1). &(49)
\end{aligned}$$

The first term in (48) can be bounded above by 1 since the alphabet of the random variable E is of size 2. Since $H(G|E=0,\mathbf{X}^n,W=1)=0$, the second term in (48) can be bounded from above as

$$\begin{aligned}
H(G|E,\mathbf{X}^n,W=1) &= H(G|E=0,\mathbf{X}^n,W=1)P(E=0|W=1) \\
&\quad + H(G|E=1,\mathbf{X}^n,W=1)P(E=1|W=1) \\
&\leq P(\widehat{G}(\mathbf{X}^n)\neq G|G\in\mathcal{T}_\varepsilon^{(p)})\log_2(|\mathcal{T}_\varepsilon^{(p)}|-1).
\end{aligned}$$

The second term in (49) is 0. Hence, we have the desired conclusion. □

D.6 Proof of Lemma 23

Proof: The proof uses standard Chernoff bounding techniques but the scaling in p is somewhat different from the usual Chernoff (Cramér) upper bound. For simplicity, we will use $M := \binom{p}{2}$. Let $Y_i, i = 1,\ldots,M$ be independent Bernoulli random variables such that $P(Y_i = 1) = c/p$. Then the probability in question can be bounded as

$$\begin{aligned} P(G \notin \mathcal{T}_{\varepsilon}^{(p)}) &= P\left(\left|\frac{1}{cp}\sum_{i=1}^{M} Y_i - \frac{1}{2}\right| > \frac{\varepsilon}{2}\right) \\ &\overset{(a)}{\leq} 2P\left(\frac{1}{cp}\sum_{i=1}^{M} Y_i > \frac{1+\varepsilon}{2}\right) \\ &\overset{(b)}{\leq} 2\mathbb{E}\left[\exp\left(t\sum_{i=1}^{M} Y_i - pt\frac{c}{2}(1+\varepsilon)\right)\right] \qquad (50) \\ &= 2\exp\left(-pt\frac{c}{2}(1+\varepsilon)\right)\prod_{i=1}^{M}\mathbb{E}[\exp(tY_i)], \qquad (51) \end{aligned}$$

where (a) follows from the union bound, (b) follows from an application of Markov's inequality with $t \geq 0$ in (50). Now, the moment generating function of a Bernoulli random variable with probability of success q is $qe^t + (1-q)$. Using this fact, we can further upper bound (51) as follows:

$$\begin{aligned} P(G \notin \mathcal{T}_{\varepsilon}^{(p)}) &= 2\exp\left(-pt\frac{c}{2}(1+\varepsilon) + M\ln(\frac{c}{p}e^t + (1-\frac{c}{p})\right) \\ &\overset{(a)}{\leq} 2\exp\left(-pt\frac{c}{2}(1+\varepsilon) + \frac{p(p-1)}{2}\frac{c}{p}(e^t-1)\right) \\ &\leq 2\exp\left(-p\left[t\frac{c}{2}(1+\varepsilon) - \frac{c}{2}(e^t-1)\right]\right), \qquad (52) \end{aligned}$$

where in (a), we used the fact that $\ln(1+z) \leq z$. Now, we differentiate the exponent in square brackets with respect to $t \geq 0$ to find the tightest bound. We observe that the optimal parameter is $t^* = \ln(1+\varepsilon)$. Substituting this back into (52) completes the proof. □

D.7 Necessary Conditions for Recovery with Distortion

We now provide the proof for Corollary 7.

The proof of Corollary 7 follows from the following generalization of the conditional Fano's inequality presented in Lemma 22. This is a modified version of an analogous theorem in Kim et al. (2008).

Lemma 24 (Conditional Fano's Inequality (Generalization)) *In the above notation, we have*

$$\frac{H(G|\mathbf{X}^n, G \in \mathcal{T}_{\varepsilon}^{(p)}) - 1 - \log_2 L}{\log_2(|\mathcal{T}_{\varepsilon}^{(p)}| - 1)} \leq P(d(G, \widehat{G}(\mathbf{X}^n)) > D | G \in \mathcal{T}_{\varepsilon}^{(p)}) \qquad (53)$$

where $L = \binom{p}{2} H_b(\beta)$ *and* β *is defined in* (16).

We will only provide a proof sketch of Lemma 24 since it is similar to Lemma 22. *Proof:* The key to establishing (53) is to upper bound the cardinality of the set $\{G \in \mathfrak{G}_p : d(G,G') \leq D\}$, which is isomorphic to $\{E \in \mathfrak{E}_p : |E \triangle E'| \leq D\}$, where $\mathfrak{E}_p$ is the set of all edge sets (with p nodes). For this purpose, we order the node pairs in a labelled undirected graph lexicographically. Now, we map each edge set E into a length-$\binom{p}{2}$ bit-string $s(E) \in \{0,1\}^{\binom{p}{2}}$. The characters in the string $s(E)$ indicate whether or not an edge is present between two node pairs. Define $d_H(s,s')$ to be the Hamming distance between strings s and s'. Then, note that

$$|E \triangle E'| = d_H(s(E), s(E')) = d_H(s(E) \oplus s(E'), 0) \tag{54}$$

where $\oplus$ denotes addition in $\mathbb{F}_2$ and 0 denotes the all zeros string. The relation in (54) means that the cardinality of the set $\{E \in \mathfrak{E}_n : |E \triangle E'| \leq D\}$ is equal to the number of strings of Hamming weight less than or equal to D. With this realization, it is easy to see that

$$|\{s \in \{0,1\}^{\binom{p}{2}} : d_H(s,0) \leq D\}| = \sum_{k=1}^{D} \binom{\binom{p}{2}}{k} \leq 2^{\binom{p}{2} H_b(D/\binom{p}{2})} = 2^L.$$

By using the same steps as in the proof of Lemma 24 (or Fano's inequality for list decoding), we arrive at the desired conclusion. □

References

P. Abbeel, D. Koller, and A.Y. Ng. Learning factor graphs in polynomial time and sample complexity. *The Journal of Machine Learning Research*, 7:1743–1788, 2006.

A. Anandkumar, A. Hassidim, and J. Kelner. Topology discovery of sparse random graphs with few participants. *Accepted to J. of Random Structures and Algorithms*, Jan. 2012a.

A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky. High-dimensional structure learning of Ising models: Local separation criterion. *Accepted to Annals of Statistics*, Jan. 2012b.

P.J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36 (6):2577–2604, 2008.

A. Bogdanov, E. Mossel, and S. Vadhan. The complexity of distinguishing Markov random fields. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 331–342, 2008.

B. Bollobás. *Random graphs*. Academic Press, 1985.

G. Bresler, E. Mossel, and A. Sly. Reconstruction of Markov random fields from samples: Some observations and algorithms. In *Intl. Workshop APPROX Approximation, Randomization and Combinatorial Optimization*, pages 343–356. Springer, 2008.

J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning bayesian networks from data: an information-theory based approach. *Artificial Intelligence*, 137(1-2):43–90, 2002.

M.J. Choi, V.Y.F. Tan, A. Anandkumar, and A. Willsky. Learning latent tree graphical models. *J. of Machine Learning Research*, 12:1771–1812, May 2011.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Tran. on Information Theory*, 14(3):462–467, 1968.

F.R.K. Chung. *Spectral Graph Theory*. Amer Mathematical Society, 1997.

F.R.K. Chung and L. Lu. *Complex Graphs and Network*. Amer. Mathematical Society, 2006.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM. J. Matrix Anal. & Appl.*, 30(56), 2008.

S. Dommers, C. Giardinà, and R. van der Hofstad. Ising models on power-law random graphs. *Journal of Statistical Physics*, pages 1–23, 2010.

I. Dumitriu and S. Pal. Sparse regular random graphs: Spectral density and eigenvectors. *Arxiv preprint arXiv:0910.5306*, 2009.

R. Foygel and M. Drton. Extended bayesian information criteria for gaussian graphical models. In *Proc. of NIPS*, 2010.

A. Gamburd, S. Hoory, M. Shahshahani, A. Shalev, and B. Virag. On the girth of random cayley graphs. *Random Structures & Algorithms*, 35(1):100–117, 2009.

J.Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1), 2006.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. of Machine Learning Research*, 8:613–636, 2007.

D. Karger and N. Srebro. Learning Markov networks: maximum bounded tree-width graphs. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pages 392–401, 2001.

Y.-H. Kim, A. Sutivong, and T. M. Cover. State amplification. *IEEE Transactions on Information Theory*, 54(5):1850 – 1859, May 2008.

M. Krivelevich and B. Sudakov. The largest eigenvalue of sparse random graphs. *Combinatorics, Probability and Computing*, 12(01):61–72, 2003.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254, 2009.

S. L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.

H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. *Journal of Machine Learning Research (JMLR)*, 10:2295–2328, 2009.

H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman. Forest density estimation. *J. of Machine Learning Research*, 12:907–951, 2011.

L. Lovász, V. Neumann-Lara, and M. Plummer. Mengerian theorems for paths of bounded length. *Periodica Mathematica Hungarica*, 9(4):269–276, 1978.

D.M. Malioutov, J.K. Johnson, and A.S. Willsky. Walk-sums and belief propagation in Gaussian graphical models. *J. of Machine Learning Research*, 7:2031–2064, 2006.

B.D. McKay, N.C. Wormald, and B. Wysocka. Short cycles in random regular graphs. *The Electronic Journal of Combinatorics*, 11(R66):1, 2004.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai. Greedy learning of Markov network structure . In *Proc. of Allerton Conf. on Communication, Control and Computing*, Monticello, USA, Sept. 2010.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems—Networks of Plausible Inference*. Morgan Kaufmann, 1988.

P. Ravikumar, M.J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, (4): 935–980, 2011.

A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall, London, 2005.

N.P. Santhanam and M.J. Wainwright. Information-theoretic Limits of high-dimensional model selection. In *International Symposium on Information Theory*, Toronto, Canada, July 2008.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

P. Spirtes and C. Meek. Learning bayesian networks with discrete variables from data. In *Proc. of Intl. Conf. on Knowledge Discovery and Data Mining*, pages 294–299, 1995.

V.Y.F. Tan, A. Anandkumar, and A. Willsky. Learning Gaussian tree models: analysis of error exponents and extremal structures. *IEEE Tran. on Signal Processing*, 58(5):2701–2714, May 2010.

V.Y.F. Tan, A. Anandkumar, and A. Willsky. A large-deviation analysis for the maximum likelihood learning of tree structures. *IEEE Tran. on Information Theory*, 57(3):1714–1735, March 2011a.

V.Y.F. Tan, A. Anandkumar, and A. Willsky. Learning Markov forest models: Analysis of error rates. *J. of Machine Learning Research*, 12:1617–1653, May 2011b.

W. Wang, M.J. Wainwright, and K. Ramchandran. Information-theoretic bounds on model selection for Gaussian Markov random fields. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, Austin, Tx, June 2010.

D.J. Watts and S.H. Strogatz. Collective dynamics of small-worldnetworks. *Nature*, 393(6684): 440–442, 1998.

X. Xie and Z. Geng. A recursive method for structural learning of directed acyclic graphs. *J. of Machine Learning Research*, 9:459–483, 2008.