Journal of Machine Learning Research 20 (2019) 1-30 Submitted 5/18; Published 3/19

# Learning Unfaithful $K$-separable Gaussian Graphical Models

**De Wen Soh** SOHDW@IHPC.A-STAR.EDU.SG
*Institute of High Performance Computing*
*1 Fusionopolis Way, #16-16 Connexis*
*Singapore, 138632*
**Sekhar Tatikonda** SEKHAR.TATIKONDA@YALE.EDU
*Department of Electrical Engineering*
*Yale University*
*New Haven, CT 06511, USA*

**Editor:** Alexander Ihler

## Abstract

The global Markov property for Gaussian graphical models ensures graph separation implies conditional independence. Specifically if a node set $S$ graph separates nodes $u$ and $v$ then $X_u$ is conditionally independent of $X_v$ given $X_S$. The opposite direction need not be true, that is, $X_u \perp X_v \mid X_S$ need not imply $S$ is a node separator of $u$ and $v$. When it does, the relation $X_u \perp X_v \mid X_S$ is called faithful. In this paper we provide a characterization of faithful relations and then provide an algorithm to test faithfulness based only on knowledge of other conditional relations of the form $X_i \perp X_j \mid X_S$. We study two classes of separable Gaussian graphical models, namely, weakly $K$-separable and strongly $K$-separable Gaussian graphical models. Using the above test for faithfulness, we introduce algorithms to learn the topologies of weakly $K$-separable and strongly $K$-separable Gaussian graphical models with $\Omega(K \log p)$ sample complexity. For strongly $K$-separable Gaussian graphical models, we additionally provide a method with error bounds for learning the off-diagonal precision matrix entries.

**Keywords:** Gaussian graphical model selection, separable graphs, high-dimensional statistical learning, faithful conditional independence relations, structural consistency

## 1. Introduction

Graphical models (Pearl (1988); Lauritzen (1996); Whittaker (1990); Wainwright and Jordan (2008)) are a popular and important means of representing certain conditional independence relations between random variables. In particular, the Gaussian graphical model is a popular model with applications to many areas such as object recognition and tracking (Sudderth (2006)), protein sequencing (Durbin et al. (1999)), pyschological modeling (Epskamp et al. (2018)), gene networks (Mohan et al. (2012); van Wieringen et al. (2018)), computer vision (Isard (2003)) and neuroimaging (Ryali et al. (2012); Belilovskym et al. (2016)). In a Gaussian graphical model, each variable is associated with a node in a graph, and any two nodes are connected by an undirected edge if and only if their two corresponding variables are dependent conditioned on the rest of the variables. An edge between two nodes therefore corresponds directly to the non-zero entries of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, where $\mathbf{\Sigma}$ is the covariance matrix of the multivariate Gaussian distribution in

question. With the graphical model defined in this way, the Gaussian distribution satisfies the global Markov property: for any pair of nodes $i$ and $j$, if all paths between the two pass through a set of nodes $S$, then the variables associated with $i$ and $j$ are conditionally independent given the variables associated with $S$.

The converse of the global Markov property does not always hold. When it does hold for a conditional independence relation, that relation is called faithful. If it holds for all relations in a model, that model is faithful. Faithfulness is important for the use of conditional independence relations in structural estimation of graphical models, that is, identifying the zeros of $\mathbf{\Omega}$. It can be challenging to simply invert $\mathbf{\Sigma}$. With faithfulness, to determine an edge between nodes $i$ and $j$, one could run through all possible separator sets $S$ and test for conditional independence (Liang et al. (2015); Koldanov et al. (2017)). If $S$ is small, the computation becomes more accurate. In the work of (Meinshausen and Bühlmann (2006); Ravikumar et al. (2011); Anandkumar et al. (2012); Wu et al. (2013)), different assumptions are used to bound $S$ to this end.

The main problem of faithfulness in graphical models is one of identifiability. Can we distinguish between a faithful graphical model and an unfaithful one? The idea of faithfulness was first explored for conditional independence relations that were satisfied in a family of graphs, using the notion of $\theta$-Markov perfectness (Frydenberg (1990); Kauermann (1996)). For Gaussian graphical models with a tree topology the the distribution has been shown to be faithful (Becker et al. (2005); Malouche and Rajaratnam (2009)). In directed graphical models, the class of unfaithful distributions has been studied in (Spirites et al. (1993); Meek (1995)). In more recent work, a notion of strong-faithfulness as a means of relaxing the conditions of faithfulness was defined (Uhler et al. (2013); Lin et al., and some answers were given as to whether a faithful graph representation existed for a given probability distribution (Sadeghi (2017)).

Being able to distinguish between faithful and unfaithful relations is the first step. Our main goal is to learn the structure of the graphical model. Many literature (Meinshausen and Bühlmann (2006); Ravikumar et al. (2011); Ren et al. (2015); Dalal and Rajaratnam (2017)) involve a form of sparsity to make learning of the graphical model easier. This sparsity comes in the form of node degree, so that a graph is sparse if the node degree is small. However, in studying the use of conditional independence relations, another form of sparsity may prove to be more useful. This form of sparsity is the size of the set $S$ that we are conditioning upon. We can make the set $S$ small, by limiting the number of vertex disjoint paths between any two nodes in a graph that are not neighbors. Graphs that exhibit this property are what is known as $K$-separable graphs (Cicalese and Melanič (2012)). It is natural therefore to study these kinds of graphs. We will examine graph learning for this graph class in this paper.

In our paper, we make the following contributions:

- We propose an algorithm to test the faithfulness of a conditional independence relation of the form $X_u \perp X_v \mid \boldsymbol{X}_S$, where $X_u, X_v$ and $\boldsymbol{X}_S$ are the random variables associated with the nodes $u$, $v$ and the node set $S$. This algorithm uses other conditional independence relations of the form $X_i \perp X_j \mid \boldsymbol{X}_S$, where $i, j \notin S$ to determine this. The faithfulness test does not require any assumption on the population version of covariance matrix $\mathbf{\Sigma}$ and can be applied to any Gaussian graphical model. To the best of our knowledge, this is the first algorithm that uses local information of

a matrix to test for the faithfulness of a conditional independence relation. We also provide sample complexity bounds for this algorithm.

- We propose a structure learning algorithm for weakly $K$-separable Gaussian graphical models. The quantity $K$ controls the size of $S$ that we need to condition on to learn the graph. This algorithm searches through different possible node sets $S$ and makes use of the faithfulness test to identify when a conditional independence relation implies that there is no edge between two nodes in the topology of the graphical model. There is no particular assumption on how small $K$ needs to be, except that $K$ scales according to $O(n/\log p)$ where $n$ is the number of samples and $p$ is the dimension of the multivariate Gaussian distribution. There are two main novelties of this algorithm. The first is that we make no assumption on the model of the graph other than it is weakly $K$-separable. Therefore, we do not require assumptions for the graph to be faithful or to satisfy certain irrepresentability conditions, which tend to be hard to verfiy. The second novelty is that the weakly $K$-separable condition subsumes other known degree bound (sparsity) assumptions. This means that our algorithm caters to a larger class of Gaussian graphical models.

- We propose a precision matrix learning algorithm for strongly $K$-separable Gaussian graphical models. This algorithm not only learns the structure of the graph, it learns the entries of the precision matrix (edge weights) as well. Of course, to do so we can simply invert the matrix. However, with small $K$, we can get better sample complexity in the high-dimensional setting. The algorithm is similar to the weakly $K$-separable learning algorithm in the sense that it again runs through different node sets $S$ to condition on. However, it uses the faithfulness test somewhat differently to identify separator nodes $S$. The goal is, for any pair of nodes $i$ and $j$, to find $S$ such that all paths from $i$ to $j$ that is of edge length greater than one must pass through some node in $S$. This gives us a nice expression for the precision matrix entries $\mathbf{\Omega}_{ij}$.

This paper is structured as follows: In Section 2, we discuss some preliminaries about Gaussian graphical models. In Section 3, we state our algorithm for testing the faithfulness of a conditional independence relation and the theoretical guarantees of the algorithm. In Section 4, we lay out an algorithm that learns the structure of a weakly $K$-separable Gaussian graphical model. In Section 5, we introduce our algorithm for learning the structure and precision matrix entries of a strongly $K$-separable Gaussian graphical model. In Section 6, we look at some examples where our algorithm can perform structural estimation while others that rely on sparsity cannot.

### 1.1. Related Work

The use of conditional independence relations to infer graph structure in the high dimensional setting was first introduced in Anandkumar et al. (2012). The authors in this work also use a search algorithm to search through different node sets $S$ for every node pair $i$ and $j$ so as to determine whether an edge exists between $i$ and $j$. They consider walk-summable Gaussian graphical models that satisfy a separability condition known as the local separability property. Because conditional independence relations are used to imply something about the topology of the graphical model, the problem of unfaithful conditional

independence relations needed to be dealt with. An assumption on the precision matrix (Assumption A4) restricts their models to faithful graphical models. We also note here that this assumption along with a minimum precision matrix entry assumption (Assumption A1) implies that the maximum node degree of the graph is $O(\sqrt{n/\log p})$.

Another class of papers that work on high-dimensional Gaussian graphical models structural estimation are those that make use of convex optimization on sparse precision matrices (Meinshausen and Bühlmann (2006); Ravikumar et al. (2011); Ren et al. (2015)). These papers make use of optimization procedures such as $\ell_1$-regularization to learn the structure of the graphical model. These techniques exploit the sparsity of the precision matrix of the graphical model, which takes the form of bounded node degree. The order of the maximum node degree is different for different papers, and we discuss this later when we compare our results to theirs. Besides the sparsity element, the optimization methods (Meinshausen and Bühlmann (2006); Ravikumar et al. (2011)) also require some assumptions on the precision matrix that are hard to verify in general, which are the incoherence assumptions and neighborhood stability assumptions. Part of the motivation of our work was to overcome the need for such assumptions.

## 2. Preliminaries

### 2.1. Linear Algebra

We first define some linear algebra notation. For a matrix $\boldsymbol{M}$, let $\boldsymbol{M}^T$ denote its transpose and let $|\boldsymbol{M}|$ denote its determinant. If $I$ is a subset of its row indices and $J$ a subset of its column indices, then we define the submatrix $\boldsymbol{M}_{IJ}$ as the $|I| \times |J|$ matrix with elements with both row and column indices from $I$ and $J$ respectively. If $I = J$, we use the notation $\boldsymbol{M}_I$ for convenience. In the same way, for a vector $\boldsymbol{v}$, we define $\boldsymbol{v}_I$ to be the subvector of $\boldsymbol{v}$ with indices from $I$. Let the spectral norm of $\boldsymbol{M}$ be denoted by $\|\boldsymbol{M}\|_2$, and let the trace of $\boldsymbol{M}$ be denoted by $\mathrm{tr}(\boldsymbol{M})$. For a square matrix $\boldsymbol{M}$, we denote its maximum and minimum eigenvalues by $\lambda_{\max}(\boldsymbol{M})$ and $\lambda_{\min}(\boldsymbol{M})$. In this paper, we will often refer to the $p$ by $p$ covariance matrix $\boldsymbol{\Sigma}$, so we will use the shorthand $\lambda_{\max} = \lambda_{\max}(\boldsymbol{\Sigma})$ and $\lambda_{\min} = \lambda_{\min}(\boldsymbol{\Sigma})$.

Let $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ and let $\mathcal{W} = \{1, \ldots, p\}$ be the index set of $\boldsymbol{M}$. Let $S \subset \mathcal{W}$ and let $S^c = \mathcal{W} \setminus S$. The matrix $\boldsymbol{M}$ has the block structure

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_S & \boldsymbol{M}_{SS^c} \\ \boldsymbol{M}_{S^cS} & \boldsymbol{M}_{S^c} \end{bmatrix}. \tag{1}$$

The Schur complement of $\boldsymbol{M}_S$ in $\boldsymbol{M}$ is defined by

$$\boldsymbol{M}_{S^c|S} = \boldsymbol{M}_{S^c} - \boldsymbol{M}_{S^cS}\boldsymbol{M}_S^{-1}\boldsymbol{M}_{SS^c}. \tag{2}$$

Using the Schur complement, we can write the inverse of $\boldsymbol{M}^{-1}$ in the form

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{M}_{S|S^c}^{-1} & -\boldsymbol{M}_{S|S^c}^{-1}\boldsymbol{M}_{S^cS}\boldsymbol{M}_S^{-1} \\ -\boldsymbol{M}_S^{-1}\boldsymbol{M}_{SS^c}\boldsymbol{M}_{S|S^c}^{-1} & \boldsymbol{M}_S^{-1} + \boldsymbol{M}_S^{-1}\boldsymbol{M}_{SS^c}\boldsymbol{M}_{S|S^c}^{-1}\boldsymbol{M}_{S^cS}\boldsymbol{M}_S^{-1} \end{bmatrix}. \tag{3}$$

### 2.2. Graph Theory

Let $\mathcal{G} = (\mathcal{W}, \mathcal{E})$ be an undirected graph, where $\mathcal{W} = \{1, \ldots, p\}$ is the set of nodes and $\mathcal{E}$ is the set of edges, namely, a subset of the set of all unordered pairs $\{(u, v) \mid u, v \in$

$\mathcal{W}\}$. In our paper, we are dealing with graphs that have no self-loops and no multiple edges between the same pair of nodes. For $I \subseteq \mathcal{W}$, we denote the induced subgraph on nodes $I$ by $\mathcal{G}_I$. For two distinct nodes $u$ and $v$, a path of length $t$ from $u$ to $v$ is a series $\{(u, w_1), (w_1, w_2), \ldots, (w_{t-2}, w_{t-1}), (w_{t-1}, v)\}$ of edges in $\mathcal{E}$, where $w_1, \ldots, w_{k-1} \in \mathcal{W}$.

Let $S \subseteq \mathcal{W} \setminus \{u, v\}$. We say that the node set $S$ is a node separator of $u$ and $v$ if all paths from $u$ to $v$ must pass through some node in $S$. The graph $\mathcal{G}$ is connected if for any distinct nodes $u, v \in \mathcal{W}$, there is at least one path from $u$ to $v$. Otherwise, the graph $\mathcal{G}$ is disjoint. A connected component of $\mathcal{G}$ is a subgraph of $\mathcal{G}$ that is connected. A disjoint graph can be divided into a number of connected components, where nodes from distinct connected components are not connected by a path. Therefore, $S$ is a node separator of $u$ and $v$ if and only if $\mathcal{G}_{S^c}$ is a disjoint graph with $u$ and $v$ in distinct connected components.

**Definition 1** *A graph $\mathcal{G} = (\mathcal{W}, \mathcal{E})$ is weakly $K$-separable if for every $(u, v) \notin \mathcal{E}$, there exists a node set $S \subset \mathcal{W} \setminus \{u, v\}$, $|S| \leq K$, such that $S$ separates nodes $i$ and $j$ in $\mathcal{G}$.*

**Definition 2** *A graph $\mathcal{G} = (\mathcal{W}, \mathcal{E})$ is strongly $K$-separable if for any two nodes $u$ and $v$, there exists a node set $S \subset \mathcal{W} \setminus \{u, v\}$, $|S| \leq K$, such that every path from $u$ to $v$ consisting of more than two nodes must contain a node in $S$.*

For any two nodes, let $\mathcal{G}_{-(u,v)}$ be the resulting graph with the edge $(u, v)$ deleted from $\mathcal{G}$. This means that $\mathcal{G}_{-(u,v)} = \mathcal{G}$ if and only if $(u, v) \notin \mathcal{E}$. The following proposition is an equivalent definition for a strongly $K$-separable graph.

**Proposition 3** *Let $\mathcal{G}$ be an undirected graph. $\mathcal{G}$ is strongly $K$-separable if and only if for any two nodes $u$ and $v$, there exists a node set $S \subset \mathcal{W} \setminus \{u, v\}$, $|S| \leq K$, such that $(\mathcal{G}_{-(u,v)})_{S^c}$ is a disjoint graph with nodes $u$ and $v$ in distinct connected components.*

A graph that is strongly $K$-separable graph is weakly $K$-separable as well.

**Example 1 (Tree Graphs)** *A tree graph is a weakly 1-separable graph, since any two nodes on a tree graph that are not neighbors are connected by one unique path (there are no cycles), so any node on that unique path would separate the two nodes. A tree graph is also trivially strongly 1-separable, since any two nodes connected by an edge is separated in the resultant graph where that particular edge is removed.*

**Example 2 (Degree Bounded Graphs)** *A graph $\mathcal{G}$ where the degree of each node is bounded by $K$ is a weakly $K$-separable graph. For any pair of non-neighboring nodes $u$ and $v$, the neighborhood of $u$ separates $u$ from $v$. Since this neighborhood size is bounded by $K$, so the graph must also be a weakly $K$-separable graph. This degree bounded graph is also a strongly $K$-separable graph as well, by the same logic. Let $v$ be a neighbor node of $u$, that is, they are connected by the edge $(u, v)$. Then the rest of the neighbors of $u$ must separate $u$ and $v$ in $\mathcal{G}_{-(u,v)}$. This fact, along with the fact that the graph is already weakly $K$-separable, makes $\mathcal{G}$ strongly $K$-separable as well. It is important to note here that weakly $K$-separable graphs are not degree bounded graphs. A star graph of arbitrary degree is a tree and is thus a weakly 1-separable graph. However, it is not degree bounded.*

**Example 3 (Locally Separated Graphs)** *In Anandkumar et al. (2012), the notion of local separability was introduced. Let $\mathcal{G}$ be a graph with $p$ nodes. For any two nodes $u, v$, we say that the number of hops required to reach $v$ from $u$ is the minimum number of edges in a path from $u$ to $v$. If $v$ is a neighbor of $u$, then $v$ is 1-hop from $u$. For any $i \in \mathcal{W}$, let $B_\gamma(i)$ be the set of all nodes that are $k$-hop from $i$, including the node $i$, where $k \leq \gamma$. Let $\mathcal{G}_{B_\gamma(u) \cup B_\gamma(v)}$ be the induced subgraph of $\mathcal{G}$ on the node set $B_\gamma(u) \cup B_\gamma(v)$. The graph satisfies the $(\eta, \gamma)$-local separation property if for any two non-neighboring nodes $u$ and $v$, the minimum number of nodes required to separate $u$ and $v$ in $\mathcal{G}_{B_\gamma(u) \cup B_\gamma(v)}$ is less than or equal to $\eta$. Of course, any graph satisfies the local separation property with the appropriate $\eta$ and $\gamma$ values. Let $S$ be such a local separator node set for nodes $u$ and $v$ in the induced subgraph $\mathcal{G}_{B_\gamma(u) \cup B_\gamma(v)}$. Then any path in $\mathcal{G}$ from $u$ to $v$ that does not pass through $S$ must have at least $2\gamma$ number of edges. Using the pigeonhole principle, it is easy to see that the graph is also weakly $K$-separable, where $K = \eta + \frac{p-\eta}{2\gamma}$. The graph may not be strongly $K$-separable however.*

**Example 4 (Weakly K-separable Graphs that are not Strongly K-separable)** *There are many examples of graphs that are weakly $K$-separable but not strongly $K$-separable. A simple example is the complete graph, with number of nodes $p > K + 2$. Because each node is a neighbor of every other node, it is trivially weakly $K$-separable. However, there are at least $p - 2$ vertex disjoint paths from $u$ to $v$ in $\mathcal{G}_{-(u,v)}$, namely, the paths of length 2 that go from $u$ to any other node in $\mathcal{G}$, excluding $u$ and $v$, to $v$. By Menger's theorem, at least $p - 2$ nodes are required to separate $u$ and $v$ in $\mathcal{G}_{-(u,v)}$, thus making the complete graph not strongly $K$-separable for $K < p - 2$.*

### 2.3. Gaussian Graphical Model

Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For the rest of this paper, we will only consider zero mean Gaussian distributions, that is, $\boldsymbol{\mu} = \mathbf{0}$. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ be the precision or concentration matrix of the graph. The random variable $\boldsymbol{X}$ has the distribution function

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Omega} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}. \tag{4}$$

For any set $S \subset \mathcal{W}$, we define $\boldsymbol{X}_S = \{X_i \mid i \in S\}$. We note here that $\boldsymbol{\Sigma}_{uv} = 0$ if and only if $X_u$ is independent of $X_v$, which we denote by $X_u \perp X_v$. If $X_u$ is independent of $X_v$ conditioned on some random variable $Z$, we denote this independence relation by $X_u \perp X_v \mid Z$. Note that $\boldsymbol{\Omega}_{uv} = 0$ if and only if $X_u \perp X_v \mid \boldsymbol{X}_{\mathcal{W} \setminus \{u,v\}}$.

For any set $S \subseteq \mathcal{W}$, the conditional distribution of $\boldsymbol{X}_{S^c}$ given $\boldsymbol{X}_S = \boldsymbol{x}_S$ follows a multivariate Gaussian distribution with conditional mean $\boldsymbol{\mu}_{S^c} - \boldsymbol{\Sigma}_{S^c S} \boldsymbol{\Sigma}_S^{-1} (\boldsymbol{x}_S - \boldsymbol{\mu}_S)$ and conditional covariance matrix

$$\boldsymbol{\Sigma}_{S^c|S} = \boldsymbol{\Sigma}_{S^c} - \boldsymbol{\Sigma}_{S^c S} \boldsymbol{\Sigma}_S^{-1} \boldsymbol{\Sigma}_{S S^c}. \tag{5}$$

Observe that the conditional covariance is the Schur complement of $\boldsymbol{\Sigma}_S$ in $\boldsymbol{\Sigma}$. For distinct nodes $u, v \in \mathcal{W}$ and $S \subseteq \mathcal{W} \setminus \{u, v\}$, we have

$$(\boldsymbol{\Sigma}_{S^c|S})_{uv} = \boldsymbol{\Sigma}_{uv} - \boldsymbol{\Sigma}_{uS} \boldsymbol{\Sigma}_S^{-1} \boldsymbol{\Sigma}_{Sv}. \tag{6}$$

The following property easily follows.

**Proposition 4** *$X_u \perp X_v \mid \boldsymbol{X}_S$ if and only if $(\boldsymbol{\Sigma}_{S^c|S})_{uv} = 0$ .*

To denote conditional covariance, we use the notations $\boldsymbol{\Sigma}(u, v \mid S)$ and $(\boldsymbol{\Sigma}_{S^c|S})_{uv}$ interchangeably.

The concentration graph $\mathcal{G}_{\boldsymbol{\Sigma}} = (\mathcal{W}, \mathcal{E})$ of a multivariate Gaussian distribution $\boldsymbol{X}$ is defined as follows: We have node set $\mathcal{W} = \{1, \ldots, p\}$, with random variable $X_u$ associated with node $u$, and edge set $\mathcal{E}$ where unordered pair $(u, v)$ is in $\mathcal{E}$ if and only if $\boldsymbol{\Omega}_{uv} \neq 0$. The multivariate Gaussian distribution, along with its concentration graph, is also known as a Gaussian graphical model. Any Gaussian graphical model satisfies the global Markov property, which is the following.

**Proposition 5 (Global Markov Property)** *If $S$ is a node separator of nodes $u$ and $v$ in $\mathcal{G}_{\boldsymbol{\Sigma}}$, then $X_u \perp X_v \mid \boldsymbol{X}_S$.*

The converse, however, is not necessarily true. Therefore, this motivates us to define faithfulness in a graphical model.

**Definition 6** *The conditional independence relation $X_u \perp X_v \mid \boldsymbol{X}_S$ is said to be faithful if $S$ is a node separator of $u$ and $v$ in the concentration graph $\mathcal{G}_{\boldsymbol{\Sigma}}$. Otherwise, it is unfaithful. A multivariate Gaussian distribution is faithful if all its conditional independence relations are faithful. The distribution is unfaithful if it is not faithful.*

**Example 5 (Example of an unfaithful Gaussian distribution)** *Consider the multivariate Gaussian distribution $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ with zero mean and positive definite covariance matrix*

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 2 & 1 & 2 \\ 2 & 4 & 2 & 1 \\ 1 & 2 & 7 & 1 \\ 2 & 1 & 1 & 6 \end{bmatrix}. \tag{7}$$

*By Proposition 4, we have $X_1 \perp X_3 \mid X_2$ since $\boldsymbol{\Sigma}_{13} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{23}$. However, the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ has no zero entries, so the concentration graph is a complete graph. This means that node 2 is not a node separator of nodes 1 and 3. The independence relation $X_1 \perp X_3 \mid X_2$ is thus not faithful and the distribution $\boldsymbol{X}$ is not faithful as well.*

We can think of the submatrix $\boldsymbol{\Sigma}_{S\cup\{u,v\}}$ as a local patch of the covariance matrix $\boldsymbol{\Sigma}$. When $X_u \perp X_v \mid \boldsymbol{X}_S$, nodes $u$ and $v$ are not connected by an edge in the concentration graph of the local patch $\boldsymbol{\Sigma}_{S\cup\{u,v\}}$, that is, we have $(\boldsymbol{\Sigma}^{-1}_{S\cup\{u,v\}})_{uv} = 0$. This does not imply that $u$ and $v$ are not connected in the concentration graph $\mathcal{G}_{\boldsymbol{\Sigma}}$. If $X_u \perp X_v \mid \boldsymbol{X}_S$ is faithful, then the implication follows. If $X_u \perp X_v \mid \boldsymbol{X}_S$ is unfaithful, then $u$ and $v$ may be connected in $\mathcal{G}_{\boldsymbol{\Sigma}}$ (See Figure 1).

Faithfulness is important in structural estimation, especially in high-dimensional settings. If we assume faithfulness, then finding a node set $S$ such that $X_u \perp X_v \mid \boldsymbol{X}_S$ would imply that there is no edge between $u$ and $v$ in the concentration graph. When we have access only to the sample covariance instead of the population covariance matrix, if the size of $S$ is small compared to $n$, the error of computing $X_u \perp X_v \mid \boldsymbol{X}_S$ is much less than

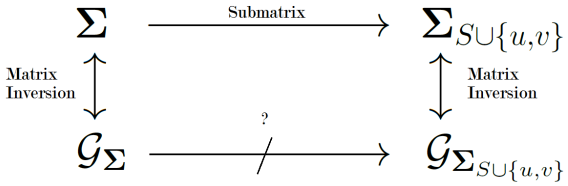


Figure 1: Even though $\mathbf{\Sigma}_{S\cup\{u,v\}}$ is a submatrix of $\mathbf{\Sigma}$, $\mathcal{G}_{\mathbf{\Sigma}_{S\cup\{u,v\}}}$ need not be a subgraph of $\mathcal{G}_{\mathbf{\Sigma}}$. Edge properties do not translate as well. That means the local patch $\mathbf{\Sigma}_{S\cup\{u,v\}}$ need not reflect the edge properties of the global graph structure of $\mathbf{\Sigma}$.

the error of inverting the entire covariance matrix. This method of searching through all possible node separator sets of a certain size is employed in Anandkumar et al. (2012); Wu et al. (2013). As mentioned before, these authors impose other restrictions on their models to overcome the problem of unfaithfulness. We do not place any restriction on the Gaussian models.

### 2.4. Sample Covariance

Let $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \in \mathbb{R}^p$ be $n$ samples of the random variable $\boldsymbol{X}$ with distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. The scatter matrix $\boldsymbol{S}$ is defined as

$$\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}^{(i)} (\boldsymbol{x}^{(i)})^T. \tag{8}$$

The sample covariance matrix determined by these $n$ samples is defined as

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n} \boldsymbol{S}. \tag{9}$$

In determining the sample conditional covariances, we will make use of the scatter matrix $\boldsymbol{S}$ instead of $\widehat{\mathbf{\Sigma}}$. Let $u$ and $v$ be distinct elements of $\mathcal{W}$ and let $S \subseteq \mathcal{W} \setminus \{u, v\}$. The sample conditional covariance of $X_u$ and $X_v$ given $\boldsymbol{X}_S$ is denoted by

$$\widehat{\mathbf{\Sigma}}(u, v \mid S) = \frac{1}{n - |S|} \left( \boldsymbol{S}_{uv} - \boldsymbol{S}_{uS} \boldsymbol{S}_S^{-1} \boldsymbol{S}_{Sv} \right). \tag{10}$$

In our algorithms, we usually have to decide whether a conditional independence relation holds. We have to determine whether $X_u \perp X_v \mid \boldsymbol{X}_S$ or $X_u \not\perp X_v \mid \boldsymbol{X}_S$. To do so with the sample covariance matrix, we need to define a conditional independence threshold $\alpha > 0$, such that if

$$|\widehat{\mathbf{\Sigma}}(u, v \mid S)| < \alpha, \tag{11}$$

we will decide that $X_u \perp X_v \mid \boldsymbol{X}_S$. Otherwise, we decide that $X_u \not\perp X_v \mid \boldsymbol{X}_S$. In our analysis, $\alpha$ will scale depending on $p, n$ and $|S|$.

## 3. Testing Conditional Independence Relations

In this section, we will describe a novel algorithm for testing faithfulness of a conditional independence relation $X_u \perp X_v \mid \boldsymbol{X}_S$. Our main goal is to decide if a conditional independence relation $X_u \perp X_v \mid \boldsymbol{X}_S$ is faithful or not. For convenience, we will denote $\mathcal{G}_{\boldsymbol{\Sigma}}$ simply by $\mathcal{G} = (\mathcal{W}, \mathcal{E})$ for the rest of this paper. Now let us suppose that it is faithful; $S$ is a node separator for $u$ and $v$ in $\mathcal{G}$. Then we should not be able to find a path from $u$ to $v$ in the induced subgraph $\mathcal{G}_{S^c}$. The main idea therefore is to search for a path between $u$ and $v$ in $\mathcal{G}_{S^c}$. If this fails, then we know that the conditional independence relation is faithful.

By the global Markov property, for any two distinct nodes $i, j \in S^c$, if $X_i \not\perp X_j \mid \boldsymbol{X}_S$, then we know that there is a path between $i$ and $j$ in $\mathcal{G}_{S^c}$. Thus, if we find some $w \in \mathcal{W} \setminus (S \cup \{i, j\})$ such that $X_u \not\perp X_w \mid \boldsymbol{X}_S$ and $X_v \not\perp X_w \mid \boldsymbol{X}_S$, then a path exists from $u$ to $w$ and another exists from $v$ to $w$, so $u$ and $v$ are connected in $\mathcal{G}_{S^c}$. This would imply that $X_u \perp X_v \mid \boldsymbol{X}_S$ is unfaithful. However, testing for paths this way does not necessarily rule out all possible paths in $\mathcal{G}_{S^c}$. The problem is that some paths may be obscured by other unfaithful conditional independence relations. There may be some $w$ whereby $X_u \not\perp X_w \mid \boldsymbol{X}_S$ and $X_v \perp X_w \mid \boldsymbol{X}_S$, but the latter relation is unfaithful. This path from $u$ to $v$ through $w$ is thus not detected by these two independence relations.

We will show however, that if there is no path from $u$ to $v$ in $\mathcal{G}_{S^c}$, then we cannot find a series of distinct nodes $w_1, \ldots, w_t \in \mathcal{W} \setminus (S \cup \{u, v\})$ for some natural number $t > 0$ such that $X_u \not\perp X_{w_1} \mid \boldsymbol{X}_S$, $X_{w_1} \not\perp X_{w_2} \mid \boldsymbol{X}_S$, …, $X_{w_{t-1}} \not\perp X_{w_t} \mid \boldsymbol{X}_S$, $X_{w_t} \not\perp X_v \mid \boldsymbol{X}_S$. This is to be expected because of the global Markov property. What is more surprising about our result is that the converse is true. If we cannot find such nodes $w_1, \ldots, w_t$, then $u$ and $v$ are not connected by a path in $\mathcal{G}_{S^c}$. This means that if there is a path from $u$ to $v$ in $\mathcal{G}_{S^c}$, even though it may be hidden by some unfaithful conditional independence relations, ultimately there are enough conditional dependence relations to reveal that $u$ and $v$ are connected by a path in $\mathcal{G}_{S^c}$. This gives us an equivalent condition for faithfulness that is in terms of conditional independence relations.

This is the backbone of our algorithm to test for the faithfulness of a conditional independence relation. We define a new graph $\bar{\mathcal{G}} = (\bar{\mathcal{W}}, \bar{\mathcal{E}})$, where $\bar{\mathcal{W}} = S^c$, and $(i, j) \in \bar{\mathcal{E}}$ if and only if $X_i \not\perp X_j \mid \boldsymbol{X}_S$. The algorithm seeks to find a path from $u$ to $v$ in $\bar{\mathcal{G}}$. If there exists a path, then $X_u \perp X_j \mid \boldsymbol{X}_S$ is unfaithful. Otherwise, it is faithful. If we consider each test of whether two nodes are conditionally independent given $\boldsymbol{X}_S$ as one step, the running time of the algorithm is the that of the algorithm used to determine set $U$. If a breadth-first search is used, the running time is $O(|S^c|^2)$.

**Theorem 7** *Suppose $X_u \perp X_v \mid \boldsymbol{X}_S$. If $S$ is a node separator of $u$ and $v$ in the concentration graph, then Algorithm 1 will classify $X_u \perp X_v \mid \boldsymbol{X}_S$ as faithful. Otherwise, Algorithm 1 will classify $X_u \perp X_v \mid \boldsymbol{X}_S$ as unfaithful.*

**Example 6 (Testing an Unfaithful Distribution (1))** *Let us take a look again at the 4-dimensional Gaussian distribution in Example 5. Suppose we want to test if $X_1 \perp X_3 \mid X_2$ is faithful or not. From its covariance matrix, we have $\boldsymbol{\Sigma}_{14} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_{24} = 2 - 2 \cdot 1/4 = 3/2 \neq 0$, so this implies that $X_1 \not\perp X_4 \mid X_2$. Similarly, $X_3 \not\perp X_4 \mid X_2$. So there exists a path from $X_1$ to $X_3$ in $\mathcal{G}_{\{1,3,4\}}$ that passes through node 4. Indeed there is such a path since the concentration graph is complete. Therefore, the relation $X_1 \perp X_3 \mid X_2$ is unfaithful.*

**Algorithm 1:** Testing Faithfulness of Relation $X_u \perp X_v \mid \boldsymbol{X}_S$

**Input:** Covariance matrix $\boldsymbol{\Sigma}$.

1. Generate set $U$ to be the set of all nodes in $\bar{\mathcal{W}}$ that are connected to $u$ by a path in $\bar{\mathcal{G}}$, including $u$. (A breadth-first search could be used.)
2. If $v \in U$, there exists a path from $u$ to $v$ in $\bar{\mathcal{G}}$, output $X_u \perp X_v \mid \boldsymbol{X}_S$ as unfaithful.
3. If $v \notin U$, let $V = \bar{W} \setminus U$. Output $X_u \perp X_v \mid \boldsymbol{X}_S$ as faithful.

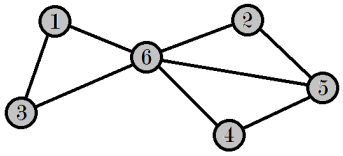


Figure 2: The concentration graph of the distribution in Example 8.

**Example 7 (Testing an Unfaithful Distribution (2))** *Consider a 6-dimensional Gaussian distribution* $\boldsymbol{X} = (X_1, \ldots, X_6)$ *that has the covariance matrix*

$$\boldsymbol{\Sigma} = \begin{bmatrix} 7 & 1 & 2 & 2 & 3 & 4 \\ 1 & 8 & 2 & 1 & 2.25 & 3 \\ 2 & 2 & 10 & 4 & 3 & 8 \\ 2 & 1 & 4 & 9 & 1 & 6 \\ 3 & 2.25 & 3 & 1 & 11 & 9 \\ 4 & 3 & 8 & 6 & 9 & 12 \end{bmatrix}. \tag{12}$$

*We want to test if the relation* $X_1 \perp X_2 \mid X_6$ *is faithful or unfaithful. Working out the necessary conditional independence relations to obtain* $\bar{\mathcal{G}}$ *with* $S = \{6\}$*, we observed that* $(1,3), (3,5), (5,4), (4,2) \in \bar{\mathcal{E}}$ *This means that* 2 *is reachable from* 1 *in* $\mathcal{G}$*, so the relation is unfaithful. In fact, the concentration graph is the complete graph* $K_6$*, and* 6 *is not a node separator of* 1 *and* 2*.*

**Example 8 (Testing a Faithful Distribution)** *We consider a 6-dimensional Gaussian distribution* $\boldsymbol{X} = (X_1, \ldots, X_6)$ *that has a covariance matrix which is similar to the distribution in Example 7,*

$$\boldsymbol{\Sigma} = \begin{bmatrix} 7 & 1 & 2 & 2 & 3 & 4 \\ 1 & 8 & 2 & 1 & 2.25 & 3 \\ 2 & 2 & 10 & 4 & 6 & 8 \\ 2 & 1 & 4 & 9 & 1 & 6 \\ 3 & 2.25 & 6 & 1 & 11 & 9 \\ 4 & 3 & 8 & 6 & 9 & 12 \end{bmatrix}. \tag{13}$$

*Observe that only* $\boldsymbol{\Sigma}_{35}$ *is changed. We again test the relation* $X_1 \perp X_2 \mid X_6$*. Running the algorithm produces a viable partition with* $U = \{1, 3\}$ *and* $V = \{2, 4, 5\}$*. This agrees with the concentration graph, as shown in Figure 2.*

The proof of Theorem 7 reflects another interesting property of the graph $\bar{\mathcal{G}}$, which is that the connected components of $\bar{\mathcal{G}}$ are exactly the same connected components of $\mathcal{G}_{S^c}$. Suppose $\bar{\mathcal{G}}$ is a disjoint graph. Then the node set $S^c$ can be partitioned into sets $S_1, \ldots, S_k$, for some $k \geq 2$, such that $\bar{\mathcal{G}}_{S_i}$ is connected for all $i = 1, \ldots, k$, and there are no edges in $\bar{\mathcal{E}}$ between nodes belonging to different partitions. This also implies that the graph $\mathcal{G}_{S^c}$ is disjoint in the same manner, in that $(\mathcal{G}_{S^c})_{S_i}$ is connected for all $i = 1, \ldots, k$, and there are no edges in $\mathcal{E}$ between nodes belonging to different partitions.

So far, we have not placed any assumption on the multivariate Gaussian distribution. Given the exact covariance matrix $\mathbf{\Sigma}$, for any general Gaussian graphical model, we can determine whether a node set $S$ is a separator of two nodes $u$ and $v$ using conditional independence relationships to test for faithfulness. If the node set $S$ is indeed a separator set of $i$ and $j$, we can then conclude that there is no edge between nodes $i$ and $j$. The next step therefore is to learn the entire structure of the graphical model.

### 3.1. Faithfulness Test Using Sample Conditional Covariances

Suppose now instead of the true covariance matrix $\mathbf{\Sigma}$, we only have access to $n$ samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \in \mathbb{R}^p$ of the $p$-dimensional multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Since the faithfulness test only considers conditional relationships between pairs of nodes, we make use of the scatter matrix $\boldsymbol{S}$ to calculate the conditional covariances, according to Section 2.4.

**Theorem 8** *Let $\boldsymbol{S}$ be used according to* (10) *to determine the sample conditional covariances in Algorithm 1 for testing the faithfulness of a conditional independence relation $X_u \perp X_v \mid \boldsymbol{X}_S$. Let $\beta = \min_{X_i \not\perp X_j \mid \boldsymbol{X}_S} |\mathbf{\Sigma}(i, j \mid S)|$, and let $\alpha = \beta/2$. Let $\Upsilon$ be the event in which:*

- *Using $\boldsymbol{S}$, Algorithm 1 correctly outputs whether or not $X_u$ is conditional independent of $X_v$ given $\boldsymbol{X}_S$.*
- *If $X_u \perp X_v \mid \boldsymbol{X}_S$, Algorithm 1 correctly outputs whether or not it is faithful using $\boldsymbol{S}$.*

*Then,*

$$\mathbb{P}(\Upsilon) \geq 1 - \epsilon. \tag{14}$$

*for $n = \Omega\left(\frac{\lambda_{\max}^2 + \beta\lambda_{\max}}{\beta^2}\left(\log(p - |S|) + \log(\epsilon^{-1})\right) + |S|\right)$.*

Some remarks:

- The algorithm to test the faithfulness of a conditional independent relation requires no assumption on the true covariance matrix $\mathbf{\Sigma}$. We only require that the distribution be multivariate Gaussian.
- The threshold $\alpha$ is used to decide if two variables are conditionally dependent or independent. If $|\widehat{\mathbf{\Sigma}}(i, j \mid S)| < \alpha$, then we decide $X_i \perp X_j \mid \boldsymbol{X}_S$, otherwise $X_i \not\perp X_j \mid \boldsymbol{X}_S$. Since we are using $\alpha$ to decide whether $\mathbf{\Sigma}(i, j \mid S)$ is zero or not, it is natural to set $\alpha = \beta/2$ where $\beta = \min_{X_i \not\perp X_j \mid \boldsymbol{X}_S} |\mathbf{\Sigma}(i, j \mid S)|$.

- Using the conditional independence test to determine faithfulness, we see that the sample complexity, if the dimension $p$ is much larger than separator size $|S|$, then the $\log(p-|S|)$ term dominates. In this case, the separator size $|S|$ behaves like a kind of sparsity measure of the graph. However, as $|S|$ increases and gets closer to $p$, we see that the term $|S|$ dominates in the sample complexity. This is to be expected because when $|S|$ tends towards $p-2$, we get closer to inverting almost the entire matrix. Of course, when $|S| = p-2$, there is no faithfulness test needed to be done since $X_u \perp X_v \mid \boldsymbol{X}_{\mathcal{W}\setminus\{i,j\}}$ implies that there is no edge between nodes $u$ and $v$. However the sample complexity needed is larger. Therefore, the faithfulness test allows us to reduce the size of $|S|$ and with that the sample complexity, even though there are more conditional relations to test between variables in $S^c$.

- The quantity $\beta = \min_{X_i \not\perp X_j \mid \boldsymbol{X}_S} |\boldsymbol{\Sigma}(i,j \mid S)|$ measures how far away the conditional covariance is from zero when it isn't zero. When this term is close to zero, more samples are required to test whether the conditional covariance is zero or non-zero for a pair of random variables. The relationship between $\beta$ and the entries of precision matrix will be discussed later when we compare our recovery results with some of the other literature. Intuitively, it makes sense that this entry plays a role in the testing for faithfulness because the foundation of the algorithm is the testing for whether a pair of random variables are conditionally independent or not.

## 4. Weakly $K$-separable Gaussian graphical models

In the process of using conditional independence relationships to infer the structure of a Gaussian graphical model, if we let the separator size $|S|$ be $p-2$, then we do not have to use any faithfulness test to determine whether any two nodes $i$ and $j$ are connected by an edge. Therefore, the true value of the faithfulness test is in the case where $|S|$ is small compared to the number of nodes $p$.

The main idea of this section is to learn the structure of a weakly $K$-separable Gaussian graphical model. For two nodes $i$ and $j$ that are not connected by an edge, there exists a node set $S$ with $|S| \leq K$, such that $S$ separates $i$ and $j$. Thus, the variable $K$ places a bound on the minimal node separator size for a node set that separates $i$ and $j$. Consequently, $K$ affects directly the sample complexity required for structural estimation. For $K$ significantly smaller than $p$, the sample complexity involved in computing each of the conditional independence relations $X_i \perp X_j \mid \boldsymbol{X}_S$ is also significantly smaller than inverting the entire covariance matrix. Using the faithfulness test described in Algorithm 1 of the previous section, Algorithm 2 is able to learn the structure of a weakly $K$-separable graph, that is, it can estimate the edge set $\mathcal{E}$.

Again, considering a computation of a conditional independence relation as one step, the running time of the algorithm is $O(p^{K+4})$. This comes from exhaustively checking through all $\binom{p-2}{K}$ possible separation sets $S$ for each of the $\binom{p}{2}$ $(i,j)$ pairs. Each time there is a conditional independence relation, we have to check for faithfulness using Algorithm 1, and the running time for that is $O(p^2)$. The novelty of the algorithm is in its ability to learn graphical models that are unfaithful.

**Algorithm 2:** Learning topology of weakly $K$-separable GGM

**Input:** Covariance matrix $\mathbf{\Sigma}$.

1. For each node pair $(i, j)$:
   - Let $F = \{S \subset \mathcal{W} \setminus \{i, j\} : |S| = K, X_i \perp X_j \mid \boldsymbol{X}_S, \text{ and it is faithful}\}$.
   - If $F \neq \phi$, output $(i, j) \notin \mathcal{E}$. If $F = \phi$, output $(i, j) \in \widehat{\mathcal{E}}$.
2. Output $\widehat{\mathcal{E}}$.

**Theorem 9** *For a weakly $K$-separable Gaussian graphical model, given the exact covariance matrix $\mathbf{\Sigma}$ as the input in Algorithm 2, the corresponding output will be the correct edge set $\mathcal{E}$.*

**Theorem 10** *Let $\mathcal{G}$ be a weakly $K$-separable graph. Let $\boldsymbol{S}$ be used according to* (10) *to determine the sample conditional covariances in Algorithm 2, instead of the true covariance matrix $\mathbf{\Sigma}$. Let $\beta \leq \min_{|S|=K, X_i \not\perp X_j | \boldsymbol{X}_S} |\mathbf{\Sigma}(i, j \mid S)|$, and let $\alpha = \beta/2$. Then,*

$$\mathbb{P}\left(\hat{\mathcal{E}} = \mathcal{E}\right) \geq 1 - \epsilon, \tag{15}$$

*for $n = \Omega(\frac{\lambda_{\max}^2 + \beta\lambda_{\max}}{\beta^2}(K \log p + \log(\epsilon^{-1})) + K)$.*

### 4.1. Comparisons to existing work

We compare this graph recovery algorithm to ones in existing work in the following aspects:

- *Faithfulness Assumptions*: In our result, we make no assumptions about the conditional independence relations having to be faithful. The purpose of Algorithm 1 is specifically to test for the faithfulness or unfaithfulness of a conditional independence relation. In Anandkumar et al. (2012), the authors make use of conditional covariances in the same way to test whether nodes are neighbors or not. Because they are using conditional covariances to infer graph structure, they naturally have to overcome the hurdle of unfaithful conditional independence relations. This is done by assumption A4 of their paper, which is an assumption on the covariance matrix $\mathbf{\Sigma}$ ensuring that pairs of nodes which are not neighbors and not separated by a set $S$ will not have small conditional covariances when conditioned on the variables $\boldsymbol{X}_S$. This assumption is strong; together with Assumption A1, it implies that the node degree of the graph is bounded by $O(\sqrt{n/\log p})$. Our algorithm is novel in the sense that we can test for faithfulness and unfaithfulness using only conditional independence relations. Thus, we do not need an assumption like that of Assumption A4, that prevents the conditional independence relations from being unfaithful.
- *Degree Bounds*: The only assumption placed on the graph structure is that it is weakly $K$-separable. According to Theorem 10, for consistent structural estimation,

we require

$$K = O\left(\frac{n}{\log p}\right). \tag{16}$$

This subsumes all graphs that have bounded node degree $K$. Degree boundedness is one of the main type of sparsity constraint in other high-dimensional Gaussian graphical model learning. In Anandkumar et al. (2012), even though the main structural assumption is the local separation property, their assumptions A1 and A4 imply that their graphs have node degree bounded by $O(\sqrt{n/\log p})$. In Ravikumar et al. (2011), the degree of each node is bounded by the same order as well. In Ren et al. (2015); Meinshausen and Bühlmann (2006), the degree bound required is $O(n/\log p)$. Thus, the weakly $K$-separable assumption subsumes all the assumptions on degree bounds, and Algorithm 2 caters to a wider class of graphical models.

- *General Gaussian Graphs*: Besides requiring the graph to be weakly $K$-separable, we place no other assumptions on the Gaussian graphical model. In Anandkumar et al. (2012), the authors require the model to be walk-summable, so that the normalized covariance matrix can be written using the Neumann series for matrix inverses. They also require assumption A4 to hold, which is a restriction on the entries of the precision matrix with respect to its corresponding row entries. In Ravikumar et al. (2011) and other works using $\ell$-1 minimization, certain incoherence conditions need to hold, and these incoherence conditions are a restriction on the precision matrix and are hard to check in general. In Meinshausen and Bühlmann (2006), they place assumptions on neighborhood stability that allows them to do support recovery. The neighborhood stability assumption is hard to verify and the precision matrices that satisfy form a subset of diagonally dominant matrices.

## 5. Strongly $K$-separable Gaussian graphical models

In this section, we impose an additional assumption on the graphical model so that we can not only learn the topology of the graph, but learn the entries of the precision matrix as well. We can think of the precision matrix as the matrix of edge weights, where $\mathbf{\Omega}_{ij}$ is the weight of edge $(i, j)$. If the edge weight is zero, it means that there is no edge between the corresponding two nodes.

The additional assumption is that the graphical model must not only be weakly $K$-separable, but it must be strongly $K$-separable as well. This means that even for edges $(i, j)$ that belong to $\mathcal{E}$, there is a node set $S \subset \mathcal{W} \setminus \{i, j\}$, $|S| \leq K$, such that the removal of edge $(i, j)$ from $\mathcal{G}$ results in a graph where $S$ separates nodes $i$ and $j$. This separation property is exactly where we can apply our faithfulness test. If we can remove the edge $(i, j)$ from the graph and use our faithfulness test to find a node separator $S$ of $i$ and $j$ in the resultant graph $\mathcal{G}_{-(i,j)}$, we can deduce the precision matrix entry $\mathbf{\Omega}_{ij}$. As shown in the following algorithm, if we can find such an $S$, the entry $\mathbf{\Omega}_{ij}$ can be calculated from the conditional covariance of $X_i$ and $X_j$, and the conditional variances of $X_i$ and of $X_j$, all of which are conditioned on $\boldsymbol{X}_S$.

The main idea behind the algorithm is to find such a node separator $S$, by appropriately "removing" the edge $(i, j)$. We cannot simply condition on $\boldsymbol{X}_S$, because, if $(i, j)$ is in $\mathcal{E}$, we

will have the condition dependence relation $X_i \not\perp X_j \mid \boldsymbol{X}_S$. We could remove the influence of the edge $(i,j)$ in the graph by conditioning on $\boldsymbol{X}_{S\cup\{i,j\}}$, however this does not ensure that $i$ and $j$ are separated by $S$ in $\mathcal{G}_{-(i,j)}$.

To overcome this problem, we condition on both $\boldsymbol{X}_{S\cup\{i\}}$ and $\boldsymbol{X}_{S\cup\{j\}}$. We use the conditional independence relations given these random variables to deduce that $S$ is a node separator of $i$ and $j$ in $\mathcal{G}_{-(i,j)}$. Running through node subsets $S \subset \mathcal{W}\setminus\{i,j\}$ of size $k$, we first condition on $\boldsymbol{X}_{S\cup\{j\}}$ to see how $S$ separates $\mathcal{G}_{S^c\setminus\{j\}}$. We then condition on $\boldsymbol{X}_{S\cup\{i\}}$ to see how $S$ separates $\mathcal{G}_{S^c\setminus\{i\}}$. Using these two pieces of information, we can infer whether $S$ separates $i$ and $j$ in $\mathcal{G}_{-(i,j)}$.

In the faithfulness test for a relation $X_u \perp X_v \mid \boldsymbol{X}_S$ in Section 3.1, we defined the graph of conditional dependences, $\bar{\mathcal{G}}$. The connectivity of $\bar{\mathcal{G}}$ reflects the connectivity of $\mathcal{G}_{S^c}$, which further implies whether $S$ separates $u$ and $v$. Here, we need to define $\bar{\mathcal{G}}$ for different node subsets $S$. For any subset $S \subset \mathcal{W}$, we denote the graph $\bar{\mathcal{G}}^{S^c} = (S^c, \bar{\mathcal{E}}^{S^c})$, where $(i,j) \in \bar{\mathcal{E}}^{S^c}$ if and only if $\boldsymbol{X}_i \not\perp X_j \mid \boldsymbol{X}_S$. For a node $h \in S^c$, let the connected node set component of $\bar{\mathcal{G}}_{S^c}$ containing $h$ be denoted by $\bar{U}_{S^c}(h)$. Therefore $\bar{U}_{S^c}(h)$ is the set of nodes in $S^c$ that are connected to $h$ by a path in $\bar{\mathcal{G}}_{S^c}$, including $h$.

For any node $i \in \mathcal{W}$, we denote the set

$$\Gamma_{(i,j)} = \{S \subset \mathcal{W}\setminus\{i,j\} : |S| \leq K\}. \tag{17}$$

of all possible node subsets $S$ of size $K$ in $\mathcal{W}\setminus\{i,j\}$. We define a subset of this set, which is

$$\Gamma_{i|j} = \{S \in \Gamma_{(i,j)} : \exists h \in S^c\setminus\{i,j\} \text{ s.t. } \boldsymbol{\Sigma}(i,h \mid S\cup\{j\}) = 0 \text{ and is faithful}\}. \tag{18}$$

This quantity encompasses the different sets $S$ such that $\mathcal{G}_{S\cup\{j\}}$ is a disjoint graph. However, this set does not subsume all possible $S$ that separate $i$ and $j$ in $\mathcal{G}_{-(i,j)}$. To include all such possible node sets $S$, we specify a subset of $\Gamma_{i|j}$, namely,

$$\Psi_{i|j} = \{S \in \Gamma_{(i,j)} : |S| \leq K, \boldsymbol{\Sigma}(i,h \mid S\cup\{j\}) = 0, \ \forall h \in S^c\setminus\{i,j\}\}. \tag{19}$$

These quantities allow us to bring definition to $\Lambda_1$, $\Lambda_2$ and $\Lambda_3$, which are given in Algorithm 3. Basically, all $S$ in $\Lambda_1 \cup \Lambda_2 \cup \Lambda_3$ are node sets that separate $i$ and $j$ in $\mathcal{G}_{-(i,j)}$. The set $\Lambda_1 \cup \Lambda_2 \cup \Lambda_3$ also has to be non-empty, which we prove in the appendix.

We only need to find one element of the set $\Lambda_1 \cup \Lambda_2 \cup \Lambda_3$. It is easy to test if $S$ is an element of the set $\Psi_{i|j}$ or $\Psi_{j|i}$. Finding an element of $\Psi_{i|j}$ or $\Psi_{j|i}$ is a special case of the test for faithfulness. To do so, in the case of $\Psi_{i|j}$, run through the subsets $S$ of size $K$, and determine if $\boldsymbol{\Sigma}(i,h \mid S\cup\{j\}) = 0$ for all $h \in S^c\setminus\{i,j\}$.

To test if $S$ belongs to $\Gamma_{i|j}$, we have to run through the $K+1$ node set of $S\cup\{h\}$, where $h \in S^c\setminus\{i,j\}$. For each node set $S\cup\{h\}$, we employ Algorithm 1, the faithfulness test, to test whether $\boldsymbol{\Sigma}(i,h \mid S\cup\{j\})$ is zero and is a faithful conditional independence relation. If we can find such a node set $S\cup\{h\}$, then $S$ belongs to $\Gamma_{i|j}$. If there is no $h \in S^c\setminus\{i,j\}$ whereby $\boldsymbol{\Sigma}(i,h \mid S\cup\{j\})$ is zero and faithful, then $S$ belongs to $\Gamma_{(i,j)}\setminus\Gamma_{i|j}$.

In this way, using Algorithm 1, we can determine whether $S$ belongs to the set $\Lambda_1 \cup \Lambda_2 \cup \Lambda_3$. Let each computation of a conditional independence relation be one step. For each $(i,j)$ pair, there are $\binom{p-2}{K}$ possible sets $S$. For each $S$, there are $p-2-K$ possible nodes $h \in S^c\setminus\{i,j\}$ to pick from, with each $h$ possibly require a faithfulness test, of which the running time is $O(p^2)$. Therefore, Algorithm 3 has a running time of $O(p^{K+5})$.

**Algorithm 3:** Learning the precision matrix of strongly $K$-separable GGM

**Input:** Covariance matrix $\mathbf{\Sigma}$

1. For each pair $(i,j)$:
   - Let $\Lambda_1 = \left\{S \in \Gamma_{i|j} \cap \Gamma_{j|i} : \bar{U}_{S^c\setminus\{i\}}(j) \subseteq \left(S^c \setminus \bar{U}_{S^c\setminus\{j\}}(i)\right)\right\}$.
   - Let $\Lambda_2 = \Psi_{j|i}$.
   - Let $\Lambda_3 = \Psi_{i|j}$.
   - Choose a set $S^* \in \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$.
   - Let $\widehat{\mathbf{\Omega}}_{ij} = -\mathbf{\Sigma}(i,j \mid S^*)/\left[\mathbf{\Sigma}(i,i \mid S^*)\mathbf{\Sigma}(j,j \mid S^*) - \mathbf{\Sigma}(i,j \mid S^*)^2\right]$.
2. Output $\widehat{\mathbf{\Omega}}$.

**Theorem 11** *For a strongly $K$-separable Gaussian graphical model, given the exact covariance matrix $\mathbf{\Sigma}$ as the input in Algorithm 3, the corresponding output will be the correct precision matrix $\mathbf{\Omega}$.*

For each node pair $i,j$, we define

$$\Lambda_0(i,j) = \{S \in \Gamma_{(i,j)}, S \text{ separates } i \text{ and } j \text{ in } \mathcal{G}_{-(i,j)}\}. \tag{20}$$

**Theorem 12** *Let $K \leq p-2$. Let $\mathcal{G}$ be a strongly $K$-separable graph, and let $\boldsymbol{S}$ be used according to (10) to determine the sample conditional covariances in Algorithm 3, instead of the true covariance matrix $\mathbf{\Sigma}$. Let $L$ be a constant such that $0 < L < 1$. Let*

$$C_1 = \frac{L\lambda_{\min}^2 + 2\lambda_{\max}(1+2\lambda_{\max})}{(1-L)L\lambda_{\min}^4}. \tag{21}$$

*Then, for*

$$\epsilon \in \left(0, C_1 \cdot \min\left\{1, \frac{L\lambda_{\min}^2}{2(1+2\lambda_{\max})}\right\}\right], \tag{22}$$

*we have*

$$\mathbb{P}\left(\max_{\substack{i,j\in\mathcal{W}, i\neq j,\\ S\in\Lambda_0(i,j)}} |\widehat{\mathbf{\Omega}}_{ij} - \mathbf{\Omega}_{ij}| \geq \epsilon\right) \leq 4p^{K+2}\exp\left\{-\frac{(n-K)\epsilon^2}{6C_1^2\lambda_{\max}^2 + 4C_1\epsilon\lambda_{\max}}\right\}. \tag{23}$$

Again, when we use the scatter matrix $\boldsymbol{S}$ to determine $\widehat{\Omega}_{ij}$, we get

$$\widehat{\mathbf{\Omega}}_{ij} = -\frac{\widehat{\mathbf{\Sigma}}(i,j \mid S^*)}{\widehat{\mathbf{\Sigma}}(i,i \mid S^*)\widehat{\mathbf{\Sigma}}(j,j \mid S^*) - \widehat{\mathbf{\Sigma}}(i,j \mid S^*)^2}, \tag{24}$$

where the sample conditional covariance are derived from $\boldsymbol{S}$ using (10).

Some remarks:

- Similar to the weakly $K$-separable case, we make no faithfulness assumptions about the model and we also do not make other assumptions besides the graph being strongly $K$-separable.

- Structural Estimation of a strongly $K$-separable graph using samples depends on accurately identifying the conditional independence relations. Therefore, the sample complexity required to estimate the structure of a strongly $K$-separable graph is the same as that of a weakly $K$-separable graph.

- The bounded degree assumption of other work is still subsumed under the strongly $K$-separable assumption. A $K$ bounded degree graph is a strongly $K$-separable graph. As such, this algorithm learns a broader variety of Gaussian graphical models.

## 6. Discussion

So far, in both the weakly and strongly $K$-separable cases, the graphs with bounded degree can be learned as well. In this section, we will examine some graphs without bounded degree that our recovery algorithms can learn.

*The complete graph:* The complete graph on $p$ nodes, as shown before, is a weakly $K$-separable graph, for any $K \geq 1$. In particular, it is weakly 1-separable graph. However, the node degree is the maximum possible, which is $p-1$. In this dense graph, our algorithm for weakly $K$-separable graphs can be used to learn the structure of the complete graph. We can learn the structure of the complete graph in $O(p^3)$, since every conditional independence in a weakly 1-separable graph is faithful. Edge estimation methods that require degree bounds will fail in this case. In fact, a weakly $K$-separable graph can contain many cliques of arbitrary size (). The presence of a large clique in a graph would render these methods inaccurate. For example, consider the $p$ node graph with $p-1$ of its nodes forming a clique and the last node is connected to only one of the other $p-1$ nodes. This graph is weakly 1-separable as well. Therefore, our algorithm is able to deal with graphs that have arbitrarily large cliques in them. These graphs do not have bounded degree. However, the complete graph and graphs with large cliques are not strongly $K$-separable.

*The star graph:* Consider the star graph on $p$ nodes. One node has degree $p-1$ and every other node has degree 1. This graph is weakly 1-separable and it is strongly 1-separable as well. Using Algorithm 3, we can learn the precision matrix entries of the star graph with low sample complexity. This algorithm has a running time of $O(p^4)$ in order to recover the structure of the graph, since all conditional independence relations in strongly $K$-separable graph are faithful. The precision matrix entries can then be derived according to 3. The ensemble of star graphs, however, clearly do not fall under the bounded degree regime. The star graph, in its essence, describes other graphs that are strongly $K$-separable, but not degree bounded. For example, as long as there is a node that is connected to a large number of degree 1 nodes, the graph is no longer degree bounded. However, in this case, the graph is still strongly $K$-separable, and we can use our algorithm to retrieve the graph structure and the entries of the precision matrix.

## 7. Conclusion

We have presented an equivalence condition for faithfulness in Gaussian graphical models and an algorithm to test whether a conditional independence relation is faithful or not. Gaussian distributions are special because its conditional independence relations depend on its covariance matrix, whose inverse, the precision matrix, provides us with a graph structure. The question of faithfulness in other Markov random fields, like Ising models, is an area of study that has much to be explored. The same questions can be asked, such as when unfaithful conditional independence relations occur, and whether they can be identified. Being able to test faithfulness allows us to learn a wider class of Gaussian graphical models, such as the weakly and strong $K$-separable graphs. In the future, we plan to extend some of these results to other Markov random fields.

## Acknowledgments

This work was partially supported by the National Science Foundation under Grant CNS-0963989 and Grant CCF-1217023.

## Appendix A. Correctness of Algorithms based on exact covariance matrix $\mathbf{\Sigma}$

### A.1 Proof of Theorem 7

**Proof** Suppose Algorithm 1 decides that $X_u \perp X_v \mid \boldsymbol{X}_S$ is unfaithful. It does so by finding a series of distinct nodes $w_1, \ldots, w_t \in S^c \setminus \{u, v\}$ for some natural number $t > 0$ such that $X_u \not\perp X_{w_1} \mid \boldsymbol{X}_S$, $X_{w_1} \not\perp X_{w_2} \mid \boldsymbol{X}_S$, ..., $X_{w_{t-1}} \not\perp X_{w_t} \mid \boldsymbol{X}_S$, $X_{w_t} \not\perp X_v \mid \boldsymbol{X}_S$. By the global Markov property, this means that $u$ is connected to $w_1$ by a path in $\mathcal{G}_{S^c}$, $w_i$ is connected to $w_{i+1}$ a path in $\mathcal{G}_{S^c}$ for $i = 1, \ldots, t-1$, and $w_t$ is connected to $v$ by a path in $\mathcal{G}_{S^c}$. This implies $u$ is connected to $v$ by a path in $\mathcal{G}_{S^c}$, so $X_u \perp X_v \mid \boldsymbol{X}_S$ is unfaithful and Algorithm 1 has correctly deduced that it is so.

Now suppose Algorithm 1 decides that $X_u \perp X_v \mid \boldsymbol{X}_S$ is faithful. That means that there is no path from $u$ to $v$ in $\bar{\mathcal{G}}$. Thus, $\bar{\mathcal{G}}$ is a disjoint graph with $u$ and $v$ in separate distinct components. The graph $\bar{\mathcal{G}}_U$ is the connected component that contains $u$. By the way we defined $\bar{\mathcal{G}}$, it follows that $X_i \perp X_j \mid \boldsymbol{X}_S$ for all $i \in U$ and $j \in V$. Equivalently, by Proposition 4,

$$(\mathbf{\Sigma}_{S^c|S})_{ij} = 0, \quad \forall i \in U,\, j \in V. \tag{25}$$

The matrix $\mathbf{\Sigma}_{S^c|S}$ therefore has a block diagonal structure, with

$$(\mathbf{\Sigma}_{S^c|S})_{UV} = (\mathbf{\Sigma}_{S^c|S})_{VU}^T = \mathbf{0}. \tag{26}$$

From (5) and (3), it follows that

$$(\mathbf{\Sigma}_{S^c|S})^{-1} = (\mathbf{\Sigma}_{S^c} - \mathbf{\Sigma}_{S^cS}\mathbf{\Sigma}_S^{-1}\mathbf{\Sigma}_{SS^c})^{-1} = \mathbf{\Omega}_{S^c}. \tag{27}$$

Since the inverse of a block diagonal matrix is also block diagonal, it follows that

$$(\mathbf{\Omega}_{S^c})_{UV} = \mathbf{\Omega}_{UV} = \mathbf{0}. \tag{28}$$

As the non-zero entries of $\mathbf{\Omega}_{S^c}$ reflect the edges between the nodes in $\mathcal{G}_{S^c}$, the last equation implies that for any $i \in U$ and $j \in V$, the edge $(i, j)$ is not in the edge set $\mathcal{E}$. This means $u$ is not connected to $v$ by a path in $\mathcal{G}_{S^c}$, which further implies that $S$ is a separator of $u$ and $v$ in $\mathcal{G}$. Thus, Algorithm 1 has correctly deduced that $X_u \perp X_v \mid \boldsymbol{X}_S$ is a faithful conditional independence relation. ■

### A.2 Proof of Theorem 9

**Proof** If $F \neq \phi$, $F$ is non-empty so there exists an $S$ such that $X_i \perp X_j \mid \boldsymbol{X}_S$ is faithful. Therefore, $S$ separates $i$ and $j$ in $\mathcal{G}$ and $(i, j) \notin \mathcal{E}$. If $F = \phi$, then for any $S \subseteq \mathcal{W}$, $|S| \leq K$, we have either $X_i \not\perp X_j \mid \boldsymbol{X}_S$ or $X_i \perp X_j \mid \boldsymbol{X}_S$ but it is unfaithful. In both cases, $S$ does not separate $i$ and $j$ in $\mathcal{G}$, for any $S \subseteq \mathcal{W}$, $|S| \leq K$. By the assumption on the graphical model, $(i, j)$ must be in $\mathcal{E}$. This shows that Algorithm 2 will correctly output the edges of $\mathcal{G}$. ■

### A.3 Conditional covariance in terms of the precision matrix

We establish some properties of the covariance matrix in terms of the precision matrix. In most of the work in this paper, we are trying to learn properties of the precision matrix, such as the support or the entries of the matrix, using conditional independence relations. These conditional independence relations are reflected by entries of the covariance matrix. Here, in this section, we further describe some of the relationship between the covariance matrix $\mathbf{\Sigma}$ and the precision matrix $\mathbf{\Omega}$.

Let $i, j$ be two elements of the index set $\mathcal{W} = \{1, \ldots, p\}$ of the square matrix $\mathbf{\Sigma}$ and let $Q = \{i, j\}$. Let $S$ be a subset of $\mathcal{W} \setminus Q$. Let $S^c = \mathcal{W} \setminus S$ and let $T = S^c \setminus Q$. Consider the matrix $\mathbf{\Omega}_{S^c}$. Computing the Schur complement of $\mathbf{\Omega}_{S^c \setminus Q}$ with respect to $\mathbf{\Omega}_{S^c}$ and using (3), we get

$$[(\mathbf{\Omega}_{S^c}^{-1})_Q]^{-1} = \mathbf{\Omega}_Q - \mathbf{\Omega}_{QT}\mathbf{\Omega}_T^{-1}\mathbf{\Omega}_{TQ}. \tag{29}$$

Using (27), we get

$$[(\mathbf{\Sigma}_{S^c|S})_Q]^{-1} = \mathbf{\Omega}_Q - \mathbf{\Omega}_{QT}\mathbf{\Omega}_T^{-1}\mathbf{\Omega}_{TQ}. \tag{30}$$

Now $(\mathbf{\Sigma}_{S^c|S})_Q$ has the form

$$(\mathbf{\Sigma}_{S^c|S})_Q = \begin{bmatrix} \mathbf{\Sigma}(i, i \mid S) & \mathbf{\Sigma}(i, j \mid S) \\ \mathbf{\Sigma}(i, j \mid S) & \mathbf{\Sigma}(j, j \mid S) \end{bmatrix}. \tag{31}$$

Therefore, comparing off-diagonal entries in (30), we get

$$\frac{-\mathbf{\Sigma}(i, j \mid S)}{\mathbf{\Sigma}(i, i \mid S)\mathbf{\Sigma}(j, j \mid S) - \mathbf{\Sigma}(i, j \mid S)^2} = \mathbf{\Omega}_{ij} - \mathbf{\Omega}_{iT}\mathbf{\Omega}_T^{-1}\mathbf{\Omega}_{Tj}. \tag{32}$$

We will make use of this last equation to learn the entries of the precision matrix for a strongly $K$-separable graph.

### A.4 Proof of Theorem 11

**Proof** If $S^*$ separates $i$ and $j$ in the edge-truncated graph $\mathcal{G}_{-(i,j)}$, this means that there exists node sets $T_i, T_j \subset \mathcal{W} \setminus S^*$, such that

- $T_i \cup T_j = \mathcal{W} \setminus S^*$ and $T_i \cap T_j = \phi$;
- $i \in T_i$, $j \in T_j$;
- $\mathbf{\Omega}_{h_1 h_2} = 0$, for all $h_1 \in T_i, h_2 \in T_j$, $(h_1, h_2) \neq (i, j)$.

This implies that

$$\begin{aligned} \mathbf{\Omega}_{iT}\mathbf{\Omega}_{T}^{-1}\mathbf{\Omega}_{Tj} &= \begin{bmatrix}\mathbf{\Omega}_{iT_i} & \mathbf{0}\end{bmatrix} \begin{bmatrix}\mathbf{\Omega}_{T_i} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{T_j}\end{bmatrix}^{-1} \begin{bmatrix}\mathbf{0} & \mathbf{\Omega}_{T_j j}\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{\Omega}_{iT_i} & \mathbf{0}\end{bmatrix} \begin{bmatrix}\mathbf{\Omega}_{T_i}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{T_j}^{-1}\end{bmatrix} \begin{bmatrix}\mathbf{0} & \mathbf{\Omega}_{T_j j}\end{bmatrix} \\ &= 0. \end{aligned} \tag{33}$$

This reduces (32) to

$$\mathbf{\Omega}_{ij} = \frac{-\mathbf{\Sigma}(i, j \mid S)}{\mathbf{\Sigma}(i, i \mid S)\mathbf{\Sigma}(j, j \mid S) - \mathbf{\Sigma}(i, j \mid S)^2}, \tag{34}$$

which is the correct output that we want from the algorithm.

To complete the proof, we need to show that the set $\Lambda_1 \cup \Lambda_2 \cup \Lambda_3$ is non-empty for any pair of nodes $i$ and $j$, and that any element of $\Lambda_1 \cup \Lambda_2 \cup \Lambda_3$ separates $i$ and $j$ in $\mathcal{G}_{-(i,j)}$. Let

$$\Lambda_0 = \{S \in \Gamma_{(i,j)} : S \text{ separates } i \text{ and } j \text{ in } \mathcal{G}_{-(i,j)}\}. \tag{35}$$

If $\Lambda_0 = \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$, by the definition of a strongly $K$-separable graph, there exists a set $S \in \Gamma_{i,j}$ such that $S$ separates $i$ and $j$ in $\mathcal{G}_{-(i,j)}$. This means that $\Lambda_0$ is non-empty, and by association, $\Lambda_1 \cup \Lambda_2 \cup \Lambda_3$ is non-empty as well. It also follows that any element of $\Lambda_1 \cup \Lambda_2 \cup \Lambda_3$ separates $i$ and $j$ in $\mathcal{G}_{-(i,j)}$.

Therefore, it remains to show that $\Lambda_0 = \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$. Suppose $S \in \Lambda_0$, that is, $S \in \Gamma_{i,j}$ that separates $i$ and $j$ in $\mathcal{G}_{-(i,j)}$. If $i$ is an isolated node in $(\mathcal{G}_{-(i,j)})_{S^c}$, then $S \in \Lambda_3$. If $j$ is an isolated node in $(\mathcal{G}_{-(i,j)})_{S^c}$, then $S \in \Lambda_2$. Suppose $i$ and $j$ are both not isolated nodes. Let $U_i \in S^c$ be the set of nodes connected to $i$ by some path in $(\mathcal{G}_{-(i,j)})_{S^c}$ including $i$, and let $U_j \in S^c$ be the set of nodes connected to $j$ by some path in $(\mathcal{G}_{-(i,j)})_{S^c}$ including $j$. Since $i$ and $j$ are not isolated nodes, both $U_i$ and $U_j$ contains at least two elements. Let $h \in U_i$ such that $h \neq i$. Then $\mathbf{\Sigma}(j, h \mid S \cup \{i\}) = 0$ and is faithful since $S \cup \{j\}$ separates $h$ and $j$ in $\mathcal{G}$. This implies that $S \in \Gamma_{j|i}$. Similarly, $S \in \Gamma_{i|j}$. Observe also that

$$\bar{U}_{S^c \setminus \{i\}}(j) \subseteq U_j \subseteq (S^c \setminus U_i) \subseteq \left(S^c \setminus \bar{U}_{S^c \setminus \{j\}}(i)\right). \tag{36}$$

The first relation follows from the fact that $\bar{U}_{S^c \setminus \{i\}}(j)$ contains only the nodes from $U_j$ that are connected to $j$ by some path in $(\mathcal{G}_{-(i,j)})_{S^c}$ that does not pass through node $i$. The last relation also follows from a similar argument. The middle relation follows because $S$ separates $i$ and $j$ in $\mathcal{G}_{-(i,j)}$. This implies that $S \in \Lambda_1$. Therefore $\Lambda_0 \subseteq \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$.

Now, let $S \in \Lambda_1$. For the sake of contradiction, suppose that there is a path in $(\mathcal{G}_{-(i,j)})_{S^c}$ from $i$ to $j$. This implies that there is a node $h$ such that

- $i$ is connect by a path to $h$ in $(\mathcal{G}_{-(i,j)})_{S^c}$ that does not pass through $j$;
- $j$ is connect by a path to $h$ in $(\mathcal{G}_{-(i,j)})_{S^c}$ that does not pass through $i$.

The first property is equivalent to there being a path from $i$ to $h$ in $(\mathcal{G}_{-(i,j)})_{S^c\setminus\{j\}}$, which is if and only if $h \in \bar{U}_{S^c\setminus\{j\}}(i)$. Similarly, $h \in \bar{U}_{S^c\setminus\{i\}}(j)$. This implies that $h \notin (S^c \setminus \bar{U}_{S^c\setminus\{j\}}(i))$, which means that $\bar{U}_{S^c\setminus\{i\}}(j) \not\subseteq (S^c \setminus \bar{U}_{S^c\setminus\{j\}}(i))$, which contradicts $S \in \Lambda_1$. Therefore, there is no path in $(\mathcal{G}_{-(i,j)})_{S^c}$ from $i$ to $j$, which implies that $S$ is a node separator of $i$ and $j$ in $\mathcal{G}_{-(i,j)}$, that is, $S \in \Lambda_0$.

Next, let $S \in \Lambda_2$. This means that $S \in \Psi_{j|i}$. Suppose that $j$ is connected by a path to $i$ in $(\mathcal{G}_{-(i,j)})_{S^c}$. Then there exists a node $h$ on this path such that $j$ is connected to $h$ by a path in $(\mathcal{G}_{-(i,j)})_{S^c\setminus i}$ that does not pass through $i$. Therefore, $\mathbf{\Sigma}(i, h \mid S \cup \{i\})$ cannot be zero and faithful at the same time. This contradicts $S \in \Psi_{j|i}$, by definition. Therefore, $j$ is not connected by a path to $i$ in $(\mathcal{G}_{-(i,j)})_{S^c}$, which also means that $S \in \Lambda_0$. Thus $\Lambda_2 \subseteq \Lambda_0$. Similarly, by symmetry $\Lambda_3 \subseteq \Lambda_0$. Therefore, $\Lambda_0 = \Lambda_1 cup \Lambda_2 \cup \Lambda_3$.

■

## Appendix B. Sample Analysis

### B.1 Wishart Distribution

Let $\boldsymbol{X} = (X_1, \ldots, X_p) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Let $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ be $n$ independent samples of $\boldsymbol{X}$. The random scatter matrix $\boldsymbol{S}$ follows a Wishart distribution, which depends on parameters $\mathbf{\Sigma}$, $p$ and $n$. We denote the Wishart distribution by $W(\mathbf{\Sigma}, p, n)$. For convenience, we denote

$$A_{n,\mathbf{\Sigma}} = \frac{1}{n}\|\mathbf{\Sigma}\|_2 \left(\mathrm{tr}(\mathbf{\Sigma}) + \|\mathbf{\Sigma}\|_2\right), \tag{37}$$

$$B_{n,\mathbf{\Sigma}} = \frac{2}{n}\mathrm{tr}(\mathbf{\Sigma}). \tag{38}$$

We will make use of the following proposition from Zhu (2012) in our sample analysis.

**Proposition 13** *Let $\boldsymbol{S} \sim W(\mathbf{\Sigma}, p, n)$. The following inequality*

$$\mathbb{P}\left(\left\|\frac{1}{n}\boldsymbol{S} - \mathbf{\Sigma}\right\|_2 \geq \epsilon\right) \leq 2p \exp\left\{-\frac{\epsilon^2}{2A_{n,\mathbf{\Sigma}} + 2\epsilon B_{n,\mathbf{\Sigma}}}\right\}. \tag{39}$$

*holds for all $\epsilon \geq 0$.*

We can simplify the above proposition to the following form.

**Lemma 14** *Let $\boldsymbol{S} \sim W(\mathbf{\Sigma}, p, n)$. Then,*

$$\mathbb{P}\left(\left\|\frac{1}{n}\boldsymbol{S} - \mathbf{\Sigma}\right\|_2 \geq \epsilon\right) \leq 2p \exp\left\{-\frac{n\epsilon^2}{(2+C)(p+1)\|\mathbf{\Sigma}\|_2^2}\right\}. \tag{40}$$

*for all $\epsilon \in \left[0, \frac{C\|\mathbf{\Sigma}\|_2}{4}\left(1 + \frac{\|\mathbf{\Sigma}\|_2}{tr(\mathbf{\Sigma})}\right)\right]$, where $C > 0$ is a constant.*

**Proof**
For $\epsilon \in \left[0, \frac{C\|\mathbf{\Sigma}\|_2}{4}\left(1+\frac{\|\mathbf{\Sigma}\|_2}{\mathrm{tr}(\mathbf{\Sigma})}\right)\right]$, we have

$$2\epsilon B_{n,\mathbf{\Sigma}} \leq 2B_{n,\mathbf{\Sigma}} \cdot \frac{C\|\mathbf{\Sigma}\|_2}{4}\left(1+\frac{\|\mathbf{\Sigma}\|_2}{\mathrm{tr}(\mathbf{\Sigma})}\right) = CA_{n,\mathbf{\Sigma}}. \tag{41}$$

Also, since

$$\mathrm{tr}(\mathbf{\Sigma}) \leq p\|\mathbf{\Sigma}\|_2, \tag{42}$$

we have

$$A_{n,\mathbf{\Sigma}} \leq \frac{1}{n}(p+1)\|\mathbf{\Sigma}\|_2^2. \tag{43}$$

Therefore, applying both (41) and (43), we get

$$\begin{aligned} \exp\left\{-\frac{\epsilon^2}{2A_{n,\mathbf{\Sigma}}+2\epsilon B_{n,\mathbf{\Sigma}}}\right\} &\leq \exp\left\{-\frac{\epsilon^2}{(C+2)A_{n,\mathbf{\Sigma}}}\right\} \\ &\leq \exp\left\{-\frac{n\epsilon^2}{(2+C)(p+1)\|\mathbf{\Sigma}\|_2^2}\right\}, \end{aligned} \tag{44}$$

as required.

■

This inequality provides a bound for the spectral norm of the sample covariance matrix with respect to the actual covariance matrix. We however want an entry-wise bound for the sample conditional covariance matrix, conditioned on $\boldsymbol{X}_S$, where $S \subseteq \{1,\ldots,p\} = \mathcal{W}$. To do so, we make use of the following proposition from Eaton (2007), which gives us the distribution of the sample conditional covariance matrix. Let $S^c = \mathcal{W} \setminus S$. We define $\widehat{\mathbf{\Sigma}}^{|S}$ to be the $p-|S|$ by $p-|S|$ matrix, where

$$\widehat{\mathbf{\Sigma}}^{|S}_{ij} = \widehat{\mathbf{\Sigma}}(i,j \mid S), \tag{45}$$

with $\widehat{\mathbf{\Sigma}}(i,j \mid S)$ defined by (10).

**Proposition 15** *(Eaton (2007)) The conditonal covariance matrix $\boldsymbol{S}_{S^c|S}$ follows a Wishart distribution with parameters $\mathbf{\Sigma}_{S^c|S}$, $p-|S|$ and $n-|S|$, that is, $\boldsymbol{S}_{S^c|S} \sim W(\mathbf{\Sigma}_{S^c|S}, p-|S|, n-|S|)$.*

Using this fact, we can now provide an element-wise bound for the sample conditional covariance.

**Lemma 16** *For any $i,j \in \mathcal{W} \setminus S$, the sample conditional covariance satisfies*

$$\mathbb{P}\left(\left|\widehat{\mathbf{\Sigma}}(i,j \mid S) - \mathbf{\Sigma}(i,j \mid S)\right| \geq \epsilon\right) \leq 4\exp\left\{-\frac{(n-|S|)\epsilon^2}{6\lambda_{\max}^2 + 4\epsilon\lambda_{\max}}\right\}, \tag{46}$$

*for all $\epsilon \geq 0$.*

**Proof** The submatrix $\boldsymbol{S}_{S\cup\{i,j\}}$ of the scatter matrix $\boldsymbol{S}$ follows a Wishart distribution with parameters $\boldsymbol{\Sigma}_{S\cup\{i,j\}}, |S|+2$ and $n$, that is, $\boldsymbol{S}_{S\cup\{i,j\}} \sim W(\boldsymbol{\Sigma}_{S\cup\{i,j\}}, |S|+2, n)$. Let $Q = \{i, j\}$, and let

$$\boldsymbol{\Sigma}_{Q|S} = \boldsymbol{\Sigma}_Q - \boldsymbol{\Sigma}_{QS}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{SQ}. \tag{47}$$

By Proposition 15, we have $\boldsymbol{S}_{Q|S} \sim W(\boldsymbol{\Sigma}_{Q|S}, 2, n-|S|)$, where

$$\boldsymbol{S}_{Q|S} = \boldsymbol{S}_Q - \boldsymbol{S}_{QS}\boldsymbol{S}_S^{-1}\boldsymbol{S}_{SQ}. \tag{48}$$

Applying Proposition **??** to $\boldsymbol{S}_{Q|S}$, we get

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{n-|S|}\boldsymbol{S}_{Q|S} - \boldsymbol{\Sigma}_{Q|S}\right\|_2 \geq \epsilon\right)\\ &\leq 4\exp\left\{-\frac{(n-|S|)\epsilon^2}{2\|\boldsymbol{\Sigma}_{Q|S}\|_2\left(\mathrm{tr}(\boldsymbol{\Sigma}_{Q|S}) + \|\boldsymbol{\Sigma}_{Q|S}\|_2\right) + 4\epsilon(\mathrm{tr}(\boldsymbol{\Sigma}_{Q|S}))}\right\}.\end{aligned} \tag{49}$$

for all $\epsilon \geq 0$. Using the eigenvalue interlacing properties for the Schur complement and submatrices Smith (1992), we have

$$\lambda_{\min}^2 \leq \lambda_{\min}^2(\boldsymbol{\Sigma}_{Q\cup S}) \leq \|\boldsymbol{\Sigma}_{Q|S}\|_2^2 \leq \|\boldsymbol{\Sigma}_{Q\cup S}\|_2^2 \leq \|\boldsymbol{\Sigma}\|_2^2 = \lambda_{\max}. \tag{50}$$

Also, we have

$$\mathrm{tr}(\boldsymbol{\Sigma}_{Q|S}) \leq 2\lambda_{\max}(\boldsymbol{\Sigma}_{Q|S}) \leq 2\lambda_{\max}. \tag{51}$$

This give us the probabilistic bound

$$\mathbb{P}\left(\left\|\frac{1}{n-|S|}\boldsymbol{S}_{Q|S} - \boldsymbol{\Sigma}_{Q|S}\right\|_2 \geq \epsilon\right) \leq 4\exp\left\{-\frac{(n-|S|)\epsilon^2}{6\lambda_{\max}^2 + 4\epsilon\lambda_{\max}}\right\}. \tag{52}$$

for all $\epsilon \geq 0$.

For any matrix, the maximum of the absolute value of its entries is bounded above by the spectral norm. Since $\widehat{\boldsymbol{\Sigma}}(i, j \mid S) = \frac{1}{n-|S|}(\boldsymbol{S}_{Q|S})_{ij}$, we have

$$\left|\widehat{\boldsymbol{\Sigma}}(i, j \mid S) - \boldsymbol{\Sigma}(i, j \mid S)\right| \leq \left\|\frac{1}{n-|S|}\boldsymbol{S}_{Q|S} - \boldsymbol{\Sigma}_{Q|S}\right\|_2. \tag{53}$$

This gives us,

$$\mathbb{P}\left(\left|\widehat{\boldsymbol{\Sigma}}(i, j \mid S) - \boldsymbol{\Sigma}(i, j \mid S)\right| \geq \epsilon\right) \leq \mathbb{P}\left(\left\|\frac{1}{n-|S|}\boldsymbol{S}_{Q|S} - \boldsymbol{\Sigma}_{Q|S}\right\|_2 \geq \epsilon\right), \tag{54}$$

which completes the proof. ■

Using Lemma 16, we prove the following two useful corollaries.

**Corollary 17** *Let $S \subset \mathcal{W}$, with $|S| \leq p-2$. Then*

$$\mathbb{P}\left(\max_{i,j\in S^c}\left|\widehat{\boldsymbol{\Sigma}}(i, j \mid S) - \boldsymbol{\Sigma}(i, j \mid S)\right| \geq \epsilon\right) \leq 4(p-|S|)^2\exp\left\{-\frac{(n-|S|)\epsilon^2}{6\lambda_{\max}^2 + 4\epsilon\lambda_{\max}}\right\}, \tag{55}$$

*for all $\epsilon \geq 0$.*

**Proof** Let $A_{i,j}$ be the event $\left\{\left|\widehat{\mathbf{\Sigma}}(i,j \mid S) - \mathbf{\Sigma}(i,j \mid S)\right| \geq \epsilon\right\}$. Then,

$$\mathbb{P}\left(\max_{i,j\in S^c}\left|\widehat{\mathbf{\Sigma}}(i,j \mid S) - \mathbf{\Sigma}(i,j \mid S)\right| \geq \epsilon\right) = \mathbb{P}\left(\bigcup_{i,j\in S^c} A_{i,j}\right). \tag{56}$$

Applying the union bound, we get

$$\mathbb{P}\left(\bigcup_{i,j\in S^c} A_{i,j}\right) \leq \sum_{i,j\in S^c} P\left(A_{i,j}\right) \leq 4(p-|S|)^2 \exp\left\{-\frac{(n-|S|)\epsilon^2}{6\lambda_{\max}^2 + 4\epsilon\lambda_{\max}}\right\}, \tag{57}$$

since there are $\binom{p-|S|}{2}$ possible choices of $\{i,j\}$. ■

**Corollary 18** *Let $K < p-2$. The following inequality*

$$\mathbb{P}\left(\max_{\substack{S\subset\mathcal{W},|S|=K,\\ i,j\in S^c}}\left|\widehat{\mathbf{\Sigma}}(i,j \mid S) - \mathbf{\Sigma}(i,j \mid S)\right| \geq \epsilon\right) \leq 4p^{K+2}\exp\left\{-\frac{(n-K)\epsilon^2}{6\lambda_{\max}^2 + 4\epsilon\lambda_{\max}}\right\}, \tag{58}$$

*holds for all $\epsilon \geq 0$.*

**Proof** The proof follows that of Corollary 17, except that the union bound now runs over all different choices of $i, j$, and $S$, with $|S| = K$. There are altogether $\binom{p}{2}\binom{p-2}{K}$ such choices, which gives us the resulting inequality. ■

### B.2 Proof of Theorem 8

**Proof** Instead of the event $\Upsilon$, we define a subset of event $\Upsilon$ that is more useful. Let $\xi$ be the event that, using $\boldsymbol{S}$, Algorithm 1 correctly identifies, for all $i, j \in S^c$, whether $X_i$ is conditionally independent of $X_j$ or not given $\boldsymbol{X}_S$. If each of these $\binom{p-|S|}{2}$ pairs of conditional independence relations are identified correctly, Algorithm 1 will be able to correctly identify whether or not $X_u \perp X_v \mid \boldsymbol{X}_S$ is faithful. Thus, if $\xi$ occurs, $\Upsilon$ occurs as well.

There are two types of events that occurs in the complement event space $\xi^c$ of $\xi$. The first event or error, is when $X_i \perp X_j \mid \boldsymbol{X}_S$, but Algorithm 1 outputs this relation conditionally dependent. We name this event $\xi_{ij}^{(1)}$. Thus, the event $\xi_{ij}^{(1)}$ occurs when $\mathbf{\Sigma}(i,j \mid S) = 0$ but $|\widehat{\mathbf{\Sigma}}(i,j \mid S)| \geq \alpha$, where $\widehat{\mathbf{\Sigma}}(i,j \mid S)$ is defined according to (10). The second type of error occurs when $X_i \not\perp X_j \mid \boldsymbol{X}_S$ but Algorithm 1 outputs this relation conditionally independent. Let this event be $\xi_{ij}^{(2)}$. Event $xi_{ij}^{(2)}$ occurs when $\mathbf{\Sigma}(i,j \mid S) \neq 0$, but $|\widehat{\mathbf{\Sigma}}(i,j \mid S)| \leq \alpha$.

As a result, we have

$$\mathbb{P}\left(\xi^c\right) = \mathbb{P}\left(\bigcup_{i,j\in S^c, X_i\perp X_j|\boldsymbol{X}_S} \xi_{ij}^{(1)}\right) + \mathbb{P}\left(\bigcup_{i,j\in S^c, X_i\not\perp X_j|\boldsymbol{X}_S} \xi_{ij}^{(2)}\right). \tag{59}$$

We bound the first term of the expression on the right hand side of the equation. When $\xi_{ij}^{(1)}$ occurs for some $i$ and $j$, we immediately have

$$\left|\widehat{\mathbf{\Sigma}}(i,j \mid S) - \mathbf{\Sigma}(i,j \mid S)\right| = \left|\widehat{\mathbf{\Sigma}}(i,j \mid S)\right| \geq \alpha. \tag{60}$$

This gives us the upper bound on the probability

$$\mathbb{P}\left(\bigcup_{i,j\in S^c, X_i \perp X_j | \boldsymbol{X}_S} \xi_{ij}^{(1)}\right) \leq \mathbb{P}\left(\max_{i,j\in S^c} \left|\widehat{\mathbf{\Sigma}}(i,j \mid S) - \mathbf{\Sigma}(i,j \mid S)\right| \geq \alpha\right). \tag{61}$$

Applying Corollary 17, we have

$$\mathbb{P}\left(\bigcup_{i,j\in S^c, X_i \perp X_j | \boldsymbol{X}_S} \xi_{ij}^{(1)}\right) \leq 4(p-|S|)^2 \exp\left\{-\frac{(n-|S|)\alpha^2}{6\lambda_{\max}^2 + 4\alpha\lambda_{\max}}\right\}, \tag{62}$$

Since $\alpha = \beta/2$, for

$$n \geq \frac{24\lambda_{\max}^2 + 8\beta\lambda_{\max}}{\beta^2}\left(\log 8 + 2\log(p-|S|) + \log\frac{1}{\epsilon}\right) + |S|, \tag{63}$$

we have

$$\mathbb{P}\left(\bigcup_{i,j\in S^c, X_i \perp X_j | \boldsymbol{X}_S} \xi_{ij}^{(1)}\right) \leq \frac{\epsilon}{2}. \tag{64}$$

The second term is bounded similarly. Since $|\mathbf{\Sigma}(i,j \mid S)| \geq \beta$ when $\mathbf{\Sigma}(i,j \mid S) \neq 0$, we have

$$\mathbb{P}\left(\bigcup_{i,j\in S^c, X_i \not\perp X_j | \boldsymbol{X}_S} \xi_{ij}^{(2)}\right) \leq \mathbb{P}\left(\max_{i,j\in S^c} \left|\widehat{\mathbf{\Sigma}}(i,j \mid S) - \mathbf{\Sigma}(i,j \mid S)\right| \geq \beta - \alpha\right). \tag{65}$$

Since $\beta - \alpha = \beta/2$, we again have

$$\mathbb{P}\left(\bigcup_{i,j\in S^c, X_i \not\perp X_j | \boldsymbol{X}_S} \xi_{ij}^{(2)}\right) \leq \frac{\epsilon}{2}, \tag{66}$$

for $n$ satisfying (63). Substituting equations (64) and (66) into (59), we get

$$\mathbb{P}(\Upsilon) \geq \mathbb{P}(\xi) \geq 1 - \epsilon, \tag{67}$$

as required.

■

### B.3 Proof of Theorem 10

**Proof** The proof follows closely the proof of Theorem 8. In the proof of Theorem 8, the separator set $S$ is fixed, and the success of the algorithm is based on correctly deciding whether each of the conditional relation between $X_i$ and $X_j$ given $\boldsymbol{X}_S$ is dependent or independent for all $i, j \in S^c$. In learning the structure of a weakly $K$-separable graph however, we have to run through various separator sets $S$ and apply our faithfulness algorithm on it. This means that if we correctly decide whether $X_i$ and $X_j$ given $\boldsymbol{X}_S$ is conditionally dependent or independent for all possible $i, j$ and $S$, that is sufficient for us to correctly determine $\mathcal{E}$.

Instead of applying Corollary 17, we apply Corollary 18, which gives us

$$\mathbb{P}\left(\hat{\mathcal{E}} \neq \mathcal{E}\right) \leq \epsilon, \tag{68}$$

for

$$n \geq \frac{24\lambda_{\max}^2 + 8\beta\lambda_{\max}}{\beta^2}\left(\log 8 + (K+2)\log p + \log\frac{1}{\epsilon}\right) + K. \tag{69}$$

■

### B.4 Proof of Theorem 12

**Lemma 19** *Let $i$ and $j$ be separate nodes in $\mathcal{W}$ and let $S$ separate $i$ and $j$ in the graph $\mathcal{G}_{-(i,j)}$. Let $L < 1$ be a positive constant, and let*

$$\epsilon \in \left(0, \frac{L\lambda_{\min}^2 + 2\lambda_{\max}(1+2\lambda_{\max})}{(1-L)L\lambda_{\min}^4} \cdot \min\left\{1, \frac{L\lambda_{\min}^2}{2(1+2\lambda_{\max})}\right\}\right]. \tag{70}$$

*Also, define*

$$\delta = \frac{(1-L)L\lambda_{\min}^4}{L\lambda_{\min}^2 + 2\lambda_{\max}(1+2\lambda_{\max})} \cdot \epsilon. \tag{71}$$

*If we have*

$$\left|\widehat{\boldsymbol{\Sigma}}(i,j \mid S) - \boldsymbol{\Sigma}(i,j \mid S)\right|, \left|\widehat{\boldsymbol{\Sigma}}(i,i \mid S) - \boldsymbol{\Sigma}(i,i \mid S)\right|, \left|\widehat{\boldsymbol{\Sigma}}(j,j \mid S) - \boldsymbol{\Sigma}(j,j \mid S)\right| \leq \delta, \tag{72}$$

*then*

$$\left|\widehat{\boldsymbol{\Omega}}_{ij} - \boldsymbol{\Omega}_{ij}\right| \leq \epsilon. \tag{73}$$

**Proof** For convenience, we denote $\boldsymbol{\Sigma}(i,j \mid S), \boldsymbol{\Sigma}(i,i \mid S)$, and $\boldsymbol{\Sigma}(j,j \mid S)$ by $a, b$, and $c$. We will also denote $\widehat{\boldsymbol{\Sigma}}(i,j \mid S), \widehat{\boldsymbol{\Sigma}}(i,i \mid S)$, and $\widehat{\boldsymbol{\Sigma}}(j,j \mid S)$ by $\hat{a}, \hat{b}$, and $\hat{c}$. Let $D$ be the determinant $bc - a^2$ and let $\hat{D} = \hat{b}\hat{c} - \hat{a}^2$. Rewriting (72), we have

$$\max\left\{|\hat{a} - a|, |\hat{b} - b|, |\hat{c} - c|\right\} \leq \delta. \tag{74}$$

We first bound $|\hat{D} - D|$. We have

$$\begin{aligned} |\hat{D} - D| &= |(\hat{b}\hat{c} - \hat{a}^2) - (bc - a^2)| \\ &\leq |\hat{b}\hat{c} - bc| + |\hat{a}^2 - a^2| \end{aligned} \tag{75}$$

The first term is bounded by

$$\begin{aligned} |\hat{b}\hat{c} - bc| &= |(\hat{b} - b)(\hat{c} - c) + c(\hat{b} - b) + b(\hat{c} - c)| \\ &\leq \delta^2 + |c|\delta + |b|\delta \\ &\leq \delta(1 + 2\lambda_{\max}), \end{aligned} \tag{76}$$

where in the last inequality, we make use of the condition that $\delta \leq 1$, which is derived from (70) and (71), and the bound on the matrix entries

$$\max\{|a|, |b|, |c|\} \leq \|\mathbf{\Sigma}_{Q|S}\|_2 \leq \|\mathbf{\Sigma}\|_2 = \lambda_{\max}. \tag{77}$$

We can then bound the second term $|\hat{a} - a|$ similarly, which gives us

$$|\hat{D} - D| \leq 2\delta(1 + 2\lambda_{\max}). \tag{78}$$

From (70), we have

$$\delta \leq \frac{L\lambda_{\min^2}}{2(1 + 2\lambda_{\max})}, \tag{79}$$

which give us

$$|\hat{D}| \geq |D| - L\lambda_{\min}^2 \geq (1 - L)|D|, \tag{80}$$

where the last inequality follows from the fact that the two eigenvalues of $\mathbf{\Sigma}_{Q|S}$ are both bounded below by $\lambda_{\min}$ by the eigenvalue interlacing property.

We are now ready to establish the upper bound for $|\widehat{\mathbf{\Omega}}_{ij} - \mathbf{\Omega}_{ij}|$. We have

$$\begin{aligned} |\widehat{\mathbf{\Omega}}_{ij} - \mathbf{\Omega}_{ij}| &= \left|\frac{\hat{a}}{\hat{D}} - \frac{a}{D}\right| \\ &= \frac{|\hat{a}D - a\hat{D}|}{|\hat{D}||D|} \\ &\leq \frac{|D||\hat{a} - a| + |a||\hat{D} - D|}{|\hat{D}||D|} \\ &\leq \frac{1}{(1 - L)|D|}\left(\delta + \frac{|a|}{L|D|} \cdot 2\delta(1 + 2\lambda_{max})\right) \\ &\leq \frac{1}{(1 - L)\lambda_{\min}^2}\left(\delta + \frac{\lambda_{\max}}{L\lambda_{\min}^2} \cdot 2\delta(1 + 2\lambda_{\max})\right). \end{aligned} \tag{81}$$

Substituting $\delta$ from (70), we get

$$|\widehat{\mathbf{\Omega}}_{ij} - \mathbf{\Omega}_{ij}| \leq \epsilon, \tag{82}$$

which is the result of this lemma. ■

We are now ready to prove Theorem 12.

**Proof** We again define $\delta$ according to (71), which can be rewritten as

$$\delta = \frac{\epsilon}{C_1}. \tag{83}$$

Applying Lemma 19, we have

$$\mathbb{P}\left(\max_{\substack{i,j\in\mathcal{W},i\neq j,\\ S\in\Lambda_0(i,j)}} |\widehat{\mathbf{\Omega}}_{ij} - \mathbf{\Omega}_{ij}| \geq \epsilon\right) \leq \mathbb{P}\left(\max_{\substack{S\subset\mathcal{W},|S|=K,\\ i,j\in S^c}} \left|\widehat{\mathbf{\Sigma}}(i,j \mid S) - \mathbf{\Sigma}(i,j \mid S)\right| \geq \delta\right), \tag{84}$$

for

$$\epsilon \in \left(0, C_1 \cdot \min\left\{1, \frac{L\lambda_{\min}^2}{2(1+2\lambda_{\max})}\right\}\right]. \tag{85}$$

Using the result of Corollary 18, we have

$$\mathbb{P}\left(\max_{\substack{S\subset\mathcal{W},|S|=K,\\ i,j\in S^c}} \left|\widehat{\mathbf{\Sigma}}(i,j \mid S) - \mathbf{\Sigma}(i,j \mid S)\right| \geq \delta\right) \leq 4p^{K+2}\exp\left\{-\frac{(n-K)\delta^2}{6\lambda_{\max}^2 + 4\delta\lambda_{\max}}\right\}. \tag{86}$$

Combining the domains for $\epsilon$, we get the result.

■

## References


A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky. High-dimensional gaussian graphical model selection: walk-summability and local separation criterion. *J. Machine Learning Research*, 13:2293–2337, Aug 2012.

A. Becker, D. Geiger, and C. Meek. Perfect tree-like markovian distributions. *Probability and Mathematical Statistics*, 25(2):231–239, 2005.

E. Belilovskym, G. Varoquaux, and M. B. Blaschko. Testing for differences in gaussian graphical models: Applications to brain connectivity. In *Advances in Neural Information Processing Systems*, Dec 2016.

F. Cicalese and M. Melanič. Graphs of separability at most 2. *Discrete Applied Mathematics*, 160 (6):685–696, April 2012.

O. Dalal and B. Rajaratnam. Sparse gaussian graphical model estimation via alternating minimization. *Biometrika*, 104(2):379–395, 2017.

R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999.

M. L. Eaton. *Multivariate Statistics: A Vector Space Approach*. Wiley, 2007.

S. Epskamp, L. J. Waldorp, R. Mõttus, and D. Borsboom. The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, pages 1–28, 2018.

M. Frydenberg. Marginalisation and collapsibility in graphical interaction models. *Annals of Statistics*, 18:790–805, 1990.

M. Isard. Pampas: real-valued graphical models for computer vision. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun 2003.

G. Kauermann. On a dualization of graphical gaussian models. *Scandinavian Journal of Statistics*, 23(1):105–116, 1996.

P. Koldanov, A. Koldanov, V. Kalyagin, and P. Pardalos. Uniformly most powerful unbiased test for conditional independence in gaussian graphical model. *Statistics and Probability Letters*, 122: 90–95, 2017.

S. L. Lauritzen. *Graphical models*. Oxford University Press, New York, 1996.

F. Liang, Q. Song, and P. Qiu. An equivalent measure of partial correlation coefficients for high-dimensional gaussian graphical models. *Journal of the American Statistical Association*, 110: 1248–1265, 2015.

S. Lin, C. Uhler, B. Sturmfels, and P. Bühlmann. Hypersurfaces and their singularities in partial correlation testing. *Preprint*.

D. Malouche and B. Rajaratnam. Gaussian covariance faithful markov trees. *Technical report, Department of Statistics, Stanford University*, 2009.

C. Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the eleventh international conference on uncertainty in artificial intelligence*, 1995.

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

K. Mohan, M. J. Chung, S. Han, D. Witten, S. Lee, and M. Fazel. Structured learning of gaussian graphical models. In *Advances in Neural Information Processing Systems*, Dec 2012.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High dimensional covariance estimation by minimizing $\ell$-1 penalized log-determinant divergence. *Electronic Journal in Statistics*, 4:935–980, 2011.

Z. Ren, T. Sun, C. Zhang, and H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical model. *Annals of Statistics*, 43(3):991–1026, 2015.

S. Ryali, T. Chen, K. Supekar, and V. Menon. Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. *Neuroimage*, 59 (4):3852–3861, February 2012.

K. Sadeghi. Faithfulness of probability distributions and graphs. *Journal of Machine Learning Research*, 18(148):1–29, 2017.

L. Smith. Some interlacing properties of schur complement of a hermitian matrix. *Linear Algebra Appl.*, 177:137–144, 1992.

P. Spirites, C. Glymore, and R. Scheines. *Causation, prediction and search*. Springer Verlag, New York, 1993.

E. B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science, 2006.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of faithfulness assumption in causal inference. *Annals of Statistics*, 41:436–463, 2013.

W. N. van Wieringen, C. F. W. Peeters, R. X. de Menezes, and M. A. van de Wiel. Testing for pathway (in)activation by using gaussian graphical models. *Journal of Royal Statistical Society*, 2018.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, 1990.

R. Wu, R. Srikant, and J. Ni. Learning loosely connected markov random fields. *Stochastic Systems*, 3, 2013.

S. Zhu. A short note on the tail bound of wishart distribution. *ArXiv e-prints*, arXiv:1212.5860, 2012.