Robust Estimation of Tree Structured Gaussian Graphical Models

Ashish Katiyar [1] **Jessica Hoffmann** [1] **Constantine Caramanis** [1]

Abstract

Consider jointly Gaussian random variables whose conditional independence structure is specified by a graphical model. If we observe realizations of the variables, we can compute the covariance matrix, and it is well known that the support of the inverse covariance matrix corresponds to the edges of the graphical model. Instead, suppose we only have noisy observations. If the noise at each node is independent, we can compute the sum of the covariance matrix and an unknown diagonal. The inverse of this sum is (in general) dense. We ask: can the original independence structure be recovered? We address this question for tree structured graphical models. We prove that this problem is unidentifiable, but show that this unidentifiability is limited to a small class of candidate trees. We further present additional constraints under which the problem is identifiable. Finally, we provide an $\mathcal{O}(n^3)$ algorithm to find this equivalence class of trees.

1. Introduction

Graphical models are a way of efficiently representing the conditional independence relationships satisfied by a collection of random variables. They form the starting point for many efficient estimation and inference algorithms. Thus, learning the graphical model of a collection of random variables is a fundamental, and very well-studied problem.

For jointly Gaussian random variables, the graphical model is given by the non-zeros in the inverse of the covariance matrix, also known as the precision matrix. We ask a natural variant of this fundamental problem: suppose we observe the random variables with independent additive noise. Thus, in the infinite sample limit, rather than knowing the covariance matrix, Σ, we have access only to $M = \Sigma + D$, the sum of the covariance matrix and a diagonal matrix. In general, $(\Sigma + D)^{-1}$ does not share the sparsity structure of Σ^{-1}. In the language of probability, if two random variables X and Y are independent conditioned on Z, then we do not expect that $(X + W_1)$ and $(Y + W_2)$ are independent when conditioned on $(Z + W_3)$, even when W_1, W_2 and W_3 are independent.

We ask: when is it possible to recover the conditional independence structure (graphical model) of the underlying variables, i.e., when can we recover the sparsity pattern of Σ^{-1}? Despite the voluminous literature on Gaussian graphical models, to the best of our knowledge, there has been no answer to this question.

Contributions of this paper. We show the following:

- A negative result of unidentifiability (Theorem 1): Even for a simple Markov chain on three nodes, the problem is unidentifiable even when an arbitrarily small amount of independent noise is added. That is, there are covariance matrices that differ only on their diagonal entries, and yet whose inverses have different sparsity patterns.
- A positive result of limited unidentifiability (Theorem 2): While unidentifiable, even for large independent noise, the ambiguity is highly limited. Specifically, we show that for tree-structured graphical models, distinguishing leaves from their immediate neighbors is impossible, but the remaining structure of the graph is identifiable (see Figure 1 for an illustration).
- Identifiability with Side Information:
  - (Theorem 3) We characterize an upper bound on the noise which, if given as side information, makes the problem identifiable.
  - (Theorem 4) If there is side information that in the precision matrix, for a leaf node, the diagonal entry is greater than the absolute value of the other non-zero entry, the problem is identifiable.
  - (Theorems 5, 6) Given a lower bound on the minimum eigenvalue of the true covariance matrix as side information, we characterize the upper bound on the noise for which the problem is identifiable. We also characterize a lower bound on the noise which makes the problem unidentifiable.

[1] Department of Electrical and Computer Engineering, The University of Texas at Austin, Texas, USA. Correspondence to: Ashish Katiyar <a.katiyar@utexas.edu>, Constantine Caramanis <constantine@utexas.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

- We provide, an $\mathcal{O}(n^3)$ algorithm that identifies the equivalence class of the underlying tree (Section 5).

Related Work

Estimating Gaussian graphical models has been a very widely explored topic. Various algorithms based on the ℓ^1 penalized log likelihood maximization have been used in, e.g., (Banerjee et al., 2008; Raskutti et al., 2009; Friedman et al., 2008; Yuan & Lin, 2007; Rothman et al., 2008). A parameter free Bayesian approach was presented in (Wong et al., 2013). In (Meinshausen et al., 2006) and (Yuan, 2010), another approach was proposed which finds conditional independence relations by regression using one random variable as output and the remaining random variables as input. The output variable is conditionally independent of the input variables with regression coefficient zero.

For learning the special class of tree structured Gaussian graphical models a classical algorithm is proposed in (Chow & Liu, 1968), now known as the Chow-Liu algorithm. The authors prove that the maximum likelihood estimate of Markov tree structure is given by the maximum-weight spanning tree (MWST) where the edge weights are the empirical mutual information. If the number of samples is infinite, this algorithm provides the exact tree structure. This algorithm inherently induces some robustness against additive independent Gaussian noise. This is because the MWST estimate remains the same if the ordering of mutual information from smaller to larger remains the same. Therefore, if the noise does not alter the order of mutual information, the algorithm still correctly identifies the tree structure. However, this is not the case in general as we show in Section 4. Moreover, whether the noise has or has not altered the MWST is not checkable from the data.

In (Tan et al., 2009), an error analysis of the Chow-Liu algorithm is presented which considers the statistical error due to finite samples. There are other papers which study the class of tree structured Gaussian graphical models based on the Chow-Liu algorithm (Choi et al., 2011; Li et al., 2016; Mossel et al., 2013). None of these, however, are able to offer guarantees in the face of noise.

There has been a lot of research on the robust estimation of graphical models (Loh & Wainwright, 2011; Yang & Lozano, 2015; Wang & Gu, 2017; Kolar & Xing, 2012; Lounici, 2014; Wang & Lin, 2014; Liu et al., 2012). However, the robustness is against outliers or missing data or Gaussian noise with known covariance or bounded noise. To the best of our knowledge, there is no work that addresses the natural setting of (unknown) additive independent Gaussian noise. This is precisely the setting that we tackle in this paper. In (Zhang et al., 2017) the authors address the problem of measurement error in the directed graphical models setting. These results do not extend to the setting of undirected graphical models.

The algorithm in (Janzamin & Anandkumar, 2014) comes closest to our setting, and in fact is complementary. In that work, the goal is to recover the graph structure in the presence of corruption in those off-diagonal terms of the covariance matrix which are not conditionally independent. Specifically, the results there do not consider (and cannot address) noise in the diagonal elements. Thus, this setting considers a perfectly complementary setting, as in this work there is noise only in the diagonal elements of the covariance matrix and not in the off diagonal elements. It would be interesting to consider if these results can be merged to obtain a general result.

2. Problem Statement

Let $X = [X_1, X_2 \ldots, X_n]^T$ denote a jointly Gaussian random variable whose conditional independence structure is given by a tree. We call this the *true tree* T^*. We denote the covariance matrix of X by Σ^* and the precision matrix by Ω^*. That is, $X \sim \mathcal{N}(0, \Sigma^*)$. We denote the noise covariance matrix by D^*. This is a non-negative diagonal matrix. We denote the observed noisy covariance matrix by:

$$\Sigma^o = \Sigma^* + D^*.$$

Given Σ^o as an input, recovering Σ^* exactly is never possible. Consider, for instance, independent noise added only to a leaf node. Instead, we would like to recover the underlying tree T^*. We show that in general, recovering T^* exactly is not possible. However, we show that the ambiguity is limited. We characterize this explicitly. That is, we characterize the set of possible trees T' that correspond to a covariance matrix, Σ', and a nonnegative diagonal matrix D' such that $\Sigma^o = \Sigma' + D'$.

Notation

For any matrix Σ, $(\Sigma)^T$ represents the transpose of the matrix. Σ_{ij} denotes the element at the i, j position. $\Sigma_{:,i}$ represents the i^{th} column. $\Sigma_{-i,-j}$ represents the submatrix after deleting row i and column j from Σ. $\Sigma_{-i,j}$ represents the j^{th} column without the i^{th} element. Similarly, $\Sigma_{i,-j}$ represents the i^{th} row without the j^{th} element. We use $\det(\Sigma)$ to represent the determinant of the matrix. For a random vector $X = [X_1, X_2, \ldots, X_n]^T$, X_i denotes the i^{th} component and X_{-i} denotes the subvector after removing the i^{th} component.

3. Identifiability Result

Let the set of all the leaf nodes of T^* be $\mathcal{L}$:

$$\mathcal{L} = \{a \mid \text{node } a \text{ is a leaf node in } T^*\}.$$

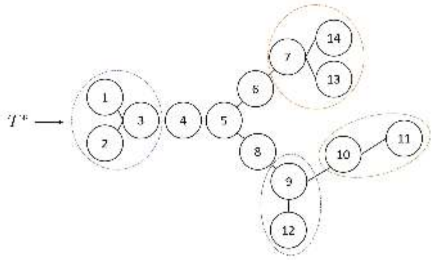


Figure 1. For this T^*, $\mathcal{T}_{T^*}$ is the set of all the trees obtained by permuting the nodes within each of the dotted regions. We prove that while T^* is unidentifiable, under our noise model, we can recover $\mathcal{T}_{T^*}$. In other words, the tree structure is recoverable up to permutation of leaves with their neighbors.

Consider all the subsets of $\mathcal{L}$ such that no two nodes in the subset share a common neighbor. Let p be the number of such subsets. Let $\mathcal{S}^q$ be the q^{th} subset. Let T^q be the tree obtained by exchanging the position of nodes in $\mathcal{S}^q$ with their neighbor node in T^*. Therefore, for every tree T^q, there is a corresponding set $\mathcal{S}^q$. We define a set of these trees as $\mathcal{T}_{T^*}$.

$$\mathcal{T}_{T^*} = \{T^q \mid q \in \{1, 2, \dots p\}\}.$$

Figure 1 gives an example of $\mathcal{T}_{T^*}$.

3.1. Identifiability Results without Side Information

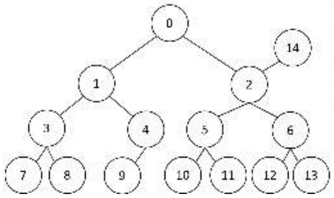
Theorem 1. *(Negative Result - Unidentifiability) Consider a covariance matrix Σ^* whose independence structure is given by the tree T^*. Suppose we are given a noisy covariance matrix $\Sigma^o = \Sigma^* + D^*$ where $D^*_{ii} > 0$ when i is a neighbor of a leaf node. For any tree $T^q \in \mathcal{T}_{T^*}$, it is always possible to decompose $\Sigma^o = \Sigma^q + D^q$ where the conditional independence for Σ^q is given by the tree T^q and D^q is a non-negative diagonal matrix.*

Proof Outline. We give an explicit construction that demonstrates that any tree $T^q \in \mathcal{T}_{T^*}$ is achievable. Consider any tree $T^q \in \mathcal{T}_{T^*}$ and its corresponding leaf subset $\mathcal{S}^q$. The required decomposition of $\Sigma^o = \Sigma^q + D^q$ is given as follows:

$$\Sigma^q_{ij} = \begin{cases} \Sigma^*_{ij} - \frac{1}{\Omega^*_{ij}} & \text{if } i = j \in \mathcal{S}^q \\ \Sigma^*_{ij} + c^i_1 & \text{if } i = j \in Neighbor(\mathcal{S}^q) \\ \Sigma^*_{ij} & \text{otherwise,} \end{cases} \tag{1}$$

where $Neighbor(\mathcal{S}^q)$ is the set of neighbor nodes of all the nodes in $\mathcal{S}^q$. Also, c^i_1 is chosen such that $0 < c^i_1 \leq D^*_{ii}$.

$$D^q_{ii} = \begin{cases} D^*_{ii} + \frac{1}{\Omega^*_{ii}} & \text{if } i \in \mathcal{S}^q \\ D^*_{ii} - c^i_1 & \text{if } i \in Neighbor(\mathcal{S}^q) \\ D^*_{ii} & \text{otherwise.} \end{cases} \tag{2}$$



Node Set	Classification	i_k, pairing (If Applicable)
{0, 1, 3, 7}	Non Star	1, {{0,1}, {3,7}}
{7, 8, 9, 10}	Non Star	1, {{7,8}, {9,10}}
{14, 2, 10, 11}	Non-Star	2, {{14,2}, {10,11}}
{7, 9, 1, 6}	Star	N.A.
{1, 14, 10, 6}	Star	N.A.

Figure 2. Examples of classification of 4 nodes as star shape or non star shape. If they form a non star shape, the nodes are grouped in pairs of 2.

The full proof which includes arriving at this decomposition and showing that the conditional independence structure of Σ^q is given by T^q is in Appendix A.

Theorem 2. *(Positive Result - Limit on unidentifiability) Consider any decomposition $\Sigma^o = \Sigma' + D'$ such that the conditional independence for Σ' is given by a tree T' and D' is a non-negative diagonal matrix. Then $T' \in \mathcal{T}_{T^*}$. Equations 1 and 2 provide a decomposition that results in this T'.*

Proof Outline. The proof of the theorem relies on showing that the off-diagonal terms of the covariance matrix suffice to specify the structure of the underlying tree up to the equivalence set $\mathcal{T}_{T^*}$. Our proof is constructive, and hence can be considered as a proto- or conceptual- algorithm for recovering $\mathcal{T}_{T^*}$. As any construction suffices to prove the result, we ignore questions of computational complexity. The ideas of this proof are then used and refined in order to provide an efficient algorithm in Section 5.

The main building block of this proof and of the algorithm presented in Section 5 is to categorize any set of 4 nodes as a *star-shape* or a *non-star-shape* (we define this below). Moreover, if it is a non star shape, we show that it is always possible to partition the four nodes into two pairs that each lie in separate connected components of the tree.

Definition 1.
- *Four nodes $\{i_1, i_2, i_3, i_4\}$ form a **non-star shape** if there exists a node i_k in the tree T^*[1] such that exactly two nodes among the four lie in the same connected component of $T^* \setminus i_k$.*
- *If $\{i_1, i_2, i_3, i_4\}$ do not form a non-star shape, we say they form a **star shape**.*

It is easy to see that in the event that a set of 4 nodes forms a non star, there exists a grouping such that the 2 nodes in the same connected component form the first pair and the other 2 nodes form the second pair. Figure 2 gives examples of star shape and non star shape. This categorization is done using only the off-diagonal elements of the covariance matrix, hence this property remains invariant to diagonal perturbations, that is, every set of 4 nodes falls in the same

[1] Note that nothing prevents i_k to be one of the four nodes.

category in any tree obtained from the decomposition of $\Sigma^o = \Sigma' + D'$ as $\Sigma'_{ij} = \Sigma^*_{ij} \; \forall \; i \neq j$. The proof of this theorem is split in 3 parts:

(i) Prove that it is possible to categorize any set of 4 nodes as star shape or non star shape using only off diagonal elements of the covariance matrix. Moreover, if the 4 nodes have a non star shape, we can find their grouping in two halves.

(ii) Prove that this categorization of all the possible sets of 4 nodes completely defines all the possible partitions of the original tree in 2 connected components such that the connected components have at least 2 nodes.

(iii) Prove that these partitions of a tree into connected components completely define the tree structure up to the equivalence set $\mathcal{T}_{T^*}$.

For part (i), we prove that a set of 4 nodes $\{i_1, i_2, i_3, i_4\}$ forms a non star shape such that nodes i_1 and i_2 form one pair and i_3 and i_4 form the second pair if and only if:

$$\begin{aligned} \frac{\Sigma^*_{i_1 i_3}}{\Sigma^*_{i_1 i_4}} &= \frac{\Sigma^*_{i_2 i_3}}{\Sigma^*_{i_2 i_4}}, \\ \frac{\Sigma^*_{i_2 i_1}}{\Sigma^*_{i_3 i_1}} &\neq \frac{\Sigma^*_{i_2 i_4}}{\Sigma^*_{i_3 i_4}}. \end{aligned} \tag{3}$$

We also prove that a set of 4 nodes $\{i_1, i_2, i_3, i_4\}$ forms a star if and only if:

$$\begin{aligned} \frac{\Sigma^*_{i_1 i_3}}{\Sigma^*_{i_1 i_4}} &= \frac{\Sigma^*_{i_2 i_3}}{\Sigma^*_{i_2 i_4}}, \\ \frac{\Sigma^*_{i_2 i_1}}{\Sigma^*_{i_3 i_1}} &= \frac{\Sigma^*_{i_2 i_4}}{\Sigma^*_{i_3 i_4}}. \end{aligned} \tag{4}$$

For part (ii), we first define a subtree.

Definition 2. *Let $\mathcal{A}$ denote the set of all the nodes in T^*. A **subtree** $\mathcal{B}$ of a tree T^* is a set of nodes such that $\mathcal{B}$ and $\mathcal{A} \setminus \mathcal{B}$ both form connected components in T^*. The pair of subtrees $\mathcal{B}$ and $\mathcal{A} \setminus \mathcal{B}$ are called **complementary subtrees**.*

We prove that if we start with a set of nodes $\{i_1, i_2, i_3, i_4\}$ that form a non star such that nodes i_1 and i_2 form a pair, we can get a partition of T^* into the smallest subtree containing i_1 and i_2 and the remaining tree. This is done using the function SMALLESTSUBTREE(Σ^o, $\{i_1, i_2, i_3, i_4\}$), the details of which are provided in Appendix B.2. Upon doing this for different initializations, we get all the possible partitions of the tree such that each partition has at least 2 nodes.

For part (iii) we define equivalence clusters and edges between equivalence clusters as follows:

Definition 3. *A set containing an internal node and all the leaf nodes connected to it forms an **equivalence cluster**. We say that there is an edge between two equivalence clusters if there is an edge between any node in one equivalence cluster and any node in the other equivalence cluster.*

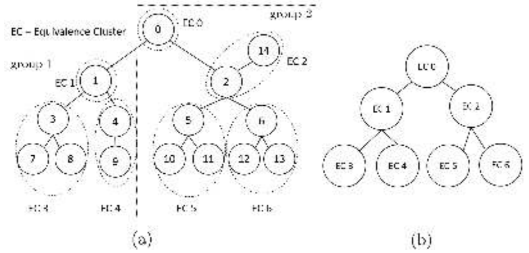


Figure 3. (a) Suppose $\{i_1, i_2, i_3, i_4\} = \{7, 9, 5, 2\}$, part (ii) partitions the nodes in group 1 and group 2. All the equivalence clusters are also shown. (b) Edges between equivalence clusters.

The subtrees obtained from part (ii) completely specify the equivalence clusters and the edges between the equivalence clusters. This gives us the set $\mathcal{T}_{T^*}$. Partitioning in part (ii) and equivalence clusters in part (iii) are illustrated in Figure 3. The detailed proof of each part is presented in Appendix B.

3.2. Identifiability Results with Side Information

Theorem 3. *(Maximum Noise Identifiability Condition) Suppose the noise is upper bounded by*

$$D^*_{aa} < \frac{1}{\Omega^*_{aa}}, \; \forall \; a \in \mathcal{L} \tag{5}$$

and suppose that this upper bound is known as side information. In this case, the decomposition of $\Sigma^o = \Sigma' + D'$ results in Σ' whose independence structure is given by T^*.

Proof. From Equation 2, for a leaf node a to exchange position with its neighbor, we need:

$$D'_{aa} \geq \frac{1}{\Omega^*_{aa}}.$$

The constraint in Equation 5 makes this solution infeasible. Hence any feasible solution cannot have a leaf node exchanged with its neighbor. □

Theorem 4. *(Leaf Diagonal Majorization Identifiability Condition) Suppose Ω^* satisfies the condition that for any leaf node a and its neighbor node b in T^*, $\Omega^*_{aa} > |\Omega^*_{ab}|$. Then for any decomposition of $\Sigma^o = \Sigma' + D'$ which satisfies the same property, the tree structure of Σ' is the same as that of Σ^*, that is, $T' = T^*$.*

Proof Outline. To prove this claim, we consider the decomposition of $\Sigma^o = \Sigma' + D'$ such that the conditional independence structure T' for Σ' has leaf node b and its neighbor node a. We show that $\Omega'_{bb} < |\Omega'_{ab}|$, that is, the leaf node b in T' violates the constraint. Hence, any decomposition of Σ^o which results in an exchange of a leaf node with its neighbor is infeasible. Hence the problem becomes identifiable.

Relabeling if necessary, assume that node n is a leaf node connected to node $n-1$ in T^*. From Equation 1, the decomposition of $\Sigma^o = \Sigma' + D'$ to obtain a tree structure T' in which node $n-1$ is a leaf node connected to node n is given by:

$$\Sigma'_{ij} = \begin{cases} \Sigma^*_{ij} - \frac{1}{\Omega^*_{ij}} & \text{if } i = j = n \\ \Sigma^*_{ij} + c_1^i & 0 < c_1^i < D^*_{n-1n-1} \text{ if } i = j = n-1 \\ \Sigma^*_{ij} & \text{otherwise.} \end{cases}$$

We derive the expression of $\Omega' = (\Sigma')^{-1}$. We denote B^1 and B^2 as follows:

$$B^1_{ij} = \begin{cases} c_1^i & 0 < c_1^i < D^*_{n-1n-1} \text{ if } i = j = n-1 \\ 0 & \text{otherwise} \end{cases},$$

$$B^2_{ij} = \begin{cases} -\frac{1}{\Omega^*_{nn}} & \text{if } i = j = n \\ 0 & \text{otherwise} \end{cases}.$$

This gives us $\Sigma' = \Sigma^* + B^1 + B^2$. The calculation of $\Omega' = (\Sigma')^{-1}$ is presented in Appendix C. At positions $(n-1, n-1)$ and $(n-1, n)$ of Ω', we get:

$$\Omega'_{n-1n-1} = \frac{1}{c_1^{n-1}},$$

$$\Omega'_{n-1n} = \frac{\Omega^*_{nn}}{c_1^{n-1}\Omega^*_{n-1n}}.$$

By the original assumption we have $\Omega^*_{nn} > |\Omega^*_{n-1n}|$, hence $\Omega'_{n-1n-1} < |\Omega'_{n-1n}|$. Therefore any exchange of leaf node with its neighbor gives an infeasible solution.

Theorem 5. *(Minimum Eigenvalue Identifiability Condition) Suppose that a lower bound on the minimum eigenvalue $\lambda_{\min}$ of Σ^* is such that for every neighbor node b of a leaf node a in T^*, $D^*_{bb} < \lambda_{min}$. Then for any decomposition of $\Sigma^o = \Sigma' + D'$ such that the minimum eigenvalue of Σ' is at least λ_{min}, the tree structure of Σ' is the same as that of Σ^*, i.e., $T' = T^*$.*

Corollary 1. *If the smallest eigenvalue of Σ^* is larger than every element of the diagonal noise matrix D^*,* and we know that this fact holds as side information*, then T^* is identifiable.*

Proof. Relabeling if necessary, assume that node n is a leaf node and node $n-1$ is its neighbor in T^*. We again consider the decomposition of $\Sigma^o = \Sigma' + D'$ such that the conditional independence structure T' for Σ' has leaf node $n-1$ and its neighbor node n. In order to prove this theorem we first consider an intermediate matrix Σ^I:

$$\Sigma^I = \Sigma^* + B^2.$$

Σ^I has minimum eigenvalue 0 (This is proved in the Appendix A during the proof of Theorem 1). Σ' is obtained as follows:

$$\Sigma' = \Sigma^I + B^1.$$

We denote the minimum eigenvalue of Σ' by λ'_{min} and Σ^I by λ^I_{min}. Using a standard result in matrix perturbation theory for symmetric matrices (Stewart & Sun, 1990) we have:

$$\begin{aligned} \lambda'_{min} &\leq \lambda^I_{min} + c_1^{n-1} \\ &= c_1^{n-1} \\ &\leq D^*_{n-1n-1}. \end{aligned}$$

If $D^*_{n-1n-1} < \lambda_{min}$ then $\lambda'_{min} < \lambda_{min}$ making this decomposition infeasible. Hence any decomposition resulting in the exchange of a leaf node a with its neighbor b is infeasible if $D^*_{bb} < \lambda_{min}$. □

Theorem 5 gives a sufficient condition on the noise for identifiability if the minimum eigenvalue is lower bounded. Next, we present a sufficient condition for unidentifiability in the same setting.

Before the theorem statement, we define the following quantities for any pair of a leaf node a and its neighbor b in T^*:

$$\begin{aligned} e^{ab} &= 1 + \frac{\Omega^*_{aa}}{|\Omega^*_{ab}|}, \\ f^{ab} &= \frac{(\Omega^*_{aa})^2}{(\Omega^*_{ab})^2} + \frac{\Omega^*_{aa}}{|\Omega^*_{ab}|}, \\ g^{ab} &= \frac{\Omega^*_{aa}(\Omega^*_{aa}\Omega^*_{bb} - (\Omega^*_{ab})^2)}{(\Omega^*_{ab})^2} + \sum_{\substack{j=1 \\ j \neq a,b}}^{n} \frac{\Omega^*_{aa}|\Omega^*_{bj}|}{|\Omega^*_{ab}|}, \\ h^{ab} &= \max_{\substack{i=1\dots n \\ i \neq a,b}} \Big(\sum_{\substack{j=1 \\ j \neq a,b}}^{n} |\Omega^*_{ij}| + \frac{\Omega^*_{aa}|\Omega^*_{bi}|}{|\Omega^*_{ab}|} \Big). \end{aligned} \tag{6}$$

Theorem 6. *(Minimum Eigenvalue Unidentifiability Condition) Suppose that a lower bound on the minimum eigenvalue of Σ^* is λ_{min}. If for any decomposition of $\Sigma^o = \Sigma' + D'$, the same constraint holds, the problem will be unidentifiable if, for a leaf node a and its neighbor b, the noise in node b is lower bounded as follows:*

$$D^*_{bb} \geq \begin{cases} e^{ab}\lambda_{min} & \textit{if } \lambda_{min} \leq \frac{(e^{ab} - f^{ab})}{e^{ab}g^{ab}}, \\ \frac{f^{ab}}{1/\lambda_{min} - g^{ab}} & \textit{if } \frac{(e^{ab} - f^{ab})}{e^{ab}g^{ab}} < \lambda_{min} < \frac{1}{g^{ab}}, \frac{1}{h^{ab}}. \end{cases}$$

If this holds, there exists a feasible Σ' with conditional independence structure T' which has node b as a leaf node and node a as its neighbor.

Proof Outline. Suppose Σ' has node b as leaf node and node a as its neighbor and the rest of the structure is the same as

T^*. We provide a lower bound on the minimum eigenvalue of Σ' by upper bounding the maximum eigenvalue of Ω' using Gerschgorin's Theorem (Stewart & Sun, 1990). The details are provided in Appendix D.

Note that a lower bound on the noise for unidentifiability can be given only below a threshold of λ_{min}. If λ_{min} is above this threshold, we cannot draw a conclusion about identifiability using this theorem.

4. Examples and Illustrations

In this section we provide an example to illustrate the theorem statements.

Consider a Markov Chain (MC) on 4 nodes whose covariance matrix is given as follows:

$$\Sigma^* = \begin{bmatrix} 1.1508 & -0.1885 & 0.0548 & -0.0069 \\ -0.1885 & 0.2356 & -0.0686 & 0.0086 \\ 0.0548 & -0.0686 & 0.7472 & -0.0934 \\ -0.0069 & 0.0086 & -0.0934 & 0.1367 \end{bmatrix},$$

Then its precision matrix is:

$$\Omega^* = \begin{bmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 5 & 0.4 & 0 \\ 0 & 0.4 & 1.5 & 1 \\ 0 & 0 & 1 & 8 \end{bmatrix}.$$

and T^* is given in Figure 4(a). Let the noise matrix be:

$$D^* = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}.$$

We have $\Sigma^o = \Sigma^* + D^*$.

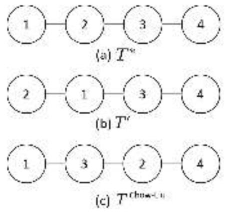
4.1. Example for Theorem 1

By Theorem 1, there exists a decomposition of $\Sigma^o = \Sigma' + D'$ such that the conditional independence structure of Σ' is given by a tree T' with node 2 as a leaf node. A possible decomposition is as follows:

$$\begin{gathered} \Sigma' = \begin{bmatrix} 0.1508 & -0.1885 & 0.0548 & -0.0069 \\ -0.1885 & 10.2356 & -0.0686 & 0.0086 \\ 0.0548 & -0.0686 & 0.7472 & -0.0934 \\ -0.0069 & 0.0086 & -0.0934 & 0.1367 \end{bmatrix}, \\ D' = \begin{bmatrix} 1.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}. \end{gathered} \tag{7}$$

The precision matrix Ω' is then:

$$\Omega' = \begin{bmatrix} 6.9687 & 0.1250 & -0.5 & 0 \\ 0.1250 & 0.1 & 0 & 0 \\ -0.5 & 0 & 1.5 & 1 \\ 0 & 0 & 1 & 8 \end{bmatrix}. \tag{8}$$



(a) T^*

(b) T'

(c) $T^{Chow-Liu}$

Figure 4. (a) T^* is a Markov Chain on 4 nodes. (b) T' is an element of $\mathcal{T}_{T^*}$, thus $\exists\Sigma', D'$ such that $\Sigma^o = \Sigma' + D'$, D' is diagonal with non-negative entries and the conditional independence structure of Σ' is given by T'. (c) Running the Chow-Liu algorithm on the Σ^o gives a tree which is not in $\mathcal{T}_{T^*}$, hence it gives an infeasible solution.

Thus, in the conditional independence structure of Σ', node 2 is a leaf node attached to node 1 as shown in Figure 4(b).

Chow-Liu. We now note that running the Chow-Liu algorithm on Σ^o gives a MC as shown in Figure 4(c). This tree does not belong to $\mathcal{T}_{T^*}$. This is an example of how the Chow-Liu algorithm can give an infeasible solution.

4.2. Example of Theorem 3

The noise matrix D^* satisfies the condition of Theorem 3:

$$D^*_{11} < \frac{1}{\Omega^*_{11}}, D^*_{44} < \frac{1}{\Omega^*_{44}}.$$

Hence by the theorem statement, with side information that $D'_{11} < 1$, the decomposition in Equation 7 is no longer feasible. Similarly a decomposition with node 3 as a leaf node is also not feasible. Hence the only feasible solutions have the same structure as T^* and the problem is identifiable.

4.3. Example of Theorem 4

Ω^* satisfies the condition of Theorem 4, that is, for leaf nodes 1 and 4:

$$\Omega^*_{11} > |\Omega^*_{12}|, \Omega^*_{44} > |\Omega^*_{34}|.$$

In the presence of side information that for any leaf node b connected to node a in T', $\Omega'_{bb} > |\Omega'_{ab}|$, the decomposition in Equation 7 becomes infeasible as $\Omega'_{22} < |\Omega'_{12}|$. Similarly, exchanging nodes 3 and 4 also results in an infeasible Σ'. Hence the problem becomes identifiable with this side information.

4.4. Example of Theorem 5.

A lower bound on the minimum eigenvalue of Σ^* is $\lambda_{min} = 0.6$. The noise in node 2 does not satisfy the condition of Theorem 5, that is:

$$D^*_{22} > \lambda_{min}.$$

Therefore, we cannot say anything about the feasibility of the decomposition when node 2 becomes a leaf node connected to node 1. However, the condition of Theorem 5 is satisfied by node 3, that is:

$$D_{33}^* < \lambda_{min}.$$

Therefore any decomposition which results in node 3 becoming a leaf node violates the minimum eigenvalue constraint (if Σ' were such that node 3 were a leaf node, the minimum eigenvalue of Σ' could at most be $0.0046 < \lambda_{min}$).

4.5. Example of Theorem 6

In order to illustrate Theorem 6, we consider leaf node 1 and its neighbor node 2. The values $e^{12}, f^{12}, g^{12}, h^{12}$ for the current example are:

$$e^{12} = 2.25, f^{12} = 2.8125, g^{12} = 7.3125, h^{12} = 9.$$

If $\lambda_{min} = 0.6$, we cannot draw a conclusion about the identifiability of the problem using Theorem 6 as $\lambda_{min} > 1/h^{12}$. If instead $\lambda_{min} = 0.1$, it satisfies $\lambda_{min} < 1/h^{12}, 1/g^{12}$. Hence we can arrive at a lower bound on the noise for unidentifiability using Theorem 5 which is given as follows:

$$D_{22}^* > 1.0465.$$

5. Algorithm

In this section we present an algorithm which takes the noisy covariance matrix Σ^o as an input and outputs $\mathcal{T}_{T^*}$. We use the classification of 4 nodes as a star shape or non star shape, the concept of subtrees, complementary subtrees and equivalence cluster (EC) that we introduced in the proof of Theorem 2.

1. We start by obtaining a subtree $\mathcal{B}$ and a node from the closest EC outside of this subtree $i_{outside_B}$. To do so:
   (a) We partition all the nodes into complementary subtrees $\mathcal{B}$ and $\mathcal{B}'$ with at least 2 nodes using only the off diagonal terms of Σ^o. This is implemented in PartitionNodes(Σ^o).
   (b) We pick any node i_B in $\mathcal{B}$.
   (c) We find the EC in $\mathcal{B}'$ that has an edge with a node in $\mathcal{B}$ in T^* by calling GetClosestEquivalenceCluster. We select one node from this EC, $i_{outside_B}$.
2. We learn all the ECs and the edges between ECs in $\mathcal{B}$ by calling LearnEdges which uses a node from the closest EC outside $\mathcal{B}$. The sets of ECs and edges are initialized as null sets. We perform the following steps:
   (a) We first call GetClosestEquivalenceCluster to obtain EC_{close}, the EC closest to $i_{outside_B}$. We add this EC to the set of ECs and select one node i_{close}.

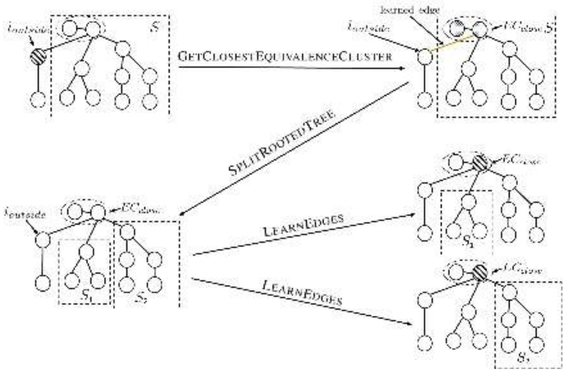


Figure 5. One recursive step of LearnEdges.

   (b) We add the edge between the EC containing $i_{outside_B}$ and EC_{close} in the edge set.
   (c) We then call SplitRootedTree to split $\mathcal{B} \setminus EC_{close}$ into the subtrees $\mathcal{B}_1, \ldots, \mathcal{B}_k$.
   (d) For any $\mathcal{B}_j$, i_{close} is a node from the closest EC. We recursively call LearnEdges on all the subtrees.
3. We repeat 1.b) - 2.d) with $\mathcal{B}'$ instead of $\mathcal{B}$.

This is illustrated in Figure 5. We next present the implementation and proof overview of all the functions.

5.1. Algorithm to partition all the nodes into two subtrees - PartitionNodes

The function PartitionNodes can be split in two parts:

(i) Find a set of 4 nodes which forms a non star. Let this set be $\{i_1, i_2, i_3, i_4\}$ such that nodes i_1 and i_2 form a pair.

(ii) Call the function SmallestSubtree to obtain complementary subtrees $\mathcal{B}$ and $\mathcal{B}'$ such that $\mathcal{B}$ is the smallest subtree containing i_1 and i_2.

For part (i), we fix two nodes and scan through all the pairs of the remaining nodes. If there exists a set of 4 nodes which forms a non star shape, this procedure finds that set. If there is no such set, T^* has a single EC.

Part (ii) is the same as discussed in the proof of Theorem 2. In Appendix E, we give the pseudo-code, proof of correctness and prove that this function is $\mathcal{O}(n^2)$

5.2. Algorithm to find the closest equivalence cluster - GetClosestEquivalenceCluster

As an input, GetClosestEquivalenceCluster takes the set of the nodes of the subtree $\mathcal{B}$, an external node $i_{outside_B}$ which belongs in $\mathcal{A} \setminus \mathcal{B}$, and the observed covariance matrix Σ^o. It outputs the EC in $\mathcal{B}$ closest to $i_{outside_B}$.

We first find a node from the closest EC. We initialize its estimate i_{close} to be the first node of $\mathcal{B}$. We notice an important fact: if $\{i_{outside_B}, i_2, i_3, i_4\}$ forms a non-star shape, nodes from the closest EC always pair with $i_{outside_B}$. Therefore, we can compare two nodes i_{close} and $i_{candidate}$: if there exists a node j in $\mathcal{B}$ such that $\{i_{outside_B}, i_{close}, i_{candidate}, j\}$ forms a non-star shape and $i_{candidate}$ is paired with $i_{outside_B}$, then i_{close} is ruled out and $i_{candidate}$ becomes the next estimate i_{close}. We use this fact to find a node i_{close} in the closest EC to $i_{outside_B}$, by scanning through all the values of $i_{candidate}$ and j.

Further, we find the remaining nodes in the EC of i_{close}. A node $i_{equivalent} \in \mathcal{B}$ is in the EC of i_{close} if $\{i_{outside_B}, i_{close}, i_{equivalent}, j\}$ forms a star shape $\forall j \in \mathcal{B} \setminus \{i_{close}\}$. In Appendix E, we give the pseudo-code, proof of correctness and prove that this function is $\mathcal{O}(n^2)$.

5.3. Algorithm to split a subtree - SPLITROOTEDTREE

As inputs, SPLITROOTEDTREE takes the subset $\mathcal{B}$, an external node $i_{outside}$, the EC to be removed EC_{close}, and the observed covariance matrix Σ^o. It outputs a list of the largest subtrees $\mathcal{B}_1, \ldots, \mathcal{B}_k$ containing all the nodes of $\mathcal{B} \setminus EC_{close}$.

Choose any $i_{close} \in EC_{close}$. To get these subtrees, we notice an important fact: i_1 and i_2 belong in the same subtree of $\mathcal{B} \setminus EC_{close}$, if and only if $\{i_{outside}, i_{close}, i_1, i_2\}$ forms a non-star shape. Therefore, we pick any node of $\mathcal{B} \setminus EC_{close}$, and use it to initialize $\mathcal{B}_1$. Then, for each new node j in $\mathcal{B}$, for each subset $\mathcal{B}_i$ containing a node $i_{\mathcal{B}_i}$ we check if $\{i_{outside}, i_{close}, j, i_{\mathcal{B}_i}\}$ forms a non-star shape. If it does, we add j to $\mathcal{B}_i$. Otherwise, we create a new subset containing only j. In Appendix E, we give the pseudo-code, proof of correctness and prove that this function is $\mathcal{O}(n^2)$

5.4. Algorithm to find equivalence clusters and edges between equivalence clusters - LEARNEDGES

We use GETCLOSESTEQUIVALENCECLUSTER and SPLITROOTEDTREE to find the ECs and the edges between the ECs.

As inputs, LEARNEDGES takes a subtree $\mathcal{B}$, an external node $i_{outside_B}$ which is a node in the EC in $\mathcal{A} \setminus \mathcal{B}$ closest to $\mathcal{B}$ and the observed covariance matrix Σ^o. The set of ECs ($equivalence\_clusters$) and the edges between ECs ($cluster\_edges$) are initialized as empty sets. This function updates these sets.

This is done in the following steps:

1. Use GETCLOSESTEQUIVALENCECLUSTER to get the equivalence cluster EC_{close} in $\mathcal{B}$ closest to $i_{outside_B}$. Add EC_{close} to $equivalence\_clusters$ and the edge between EC_{close} and the EC containing $i_{outside_B}$ in $cluster\_edges$.
2. Use SPLITROOTEDTREE to split $\mathcal{B} \setminus EC_{close}$ into subtrees $\mathcal{B}_1, \ldots, \mathcal{B}_k$.
3. For each of these subtrees $\mathcal{B}_j$, $i_{close} \in EC_{close}$ is a node from the closest EC in $\mathcal{A} \setminus \mathcal{B}_j$. Recursively call LEARNEDGES with $\mathcal{B}_j$, i_{close} and Σ^o as inputs.

5.5. Complete Algorithm - LEARNCLUSTERTREE

Finally, we describe LEARNCLUSTERTREE, the complete algorithm which learns all the ECs of a tree T^* and the edges between them from the observed covariance matrix Σ^o.

As input, it takes the observed covariance matrix Σ^o. It populates the ECs, $equivalence\_clusters$, and the edges between ECs, $cluster\_edges$. LEARNCLUSTERTREE performs the following steps:

1. Partition all the nodes in complementary subtrees $\mathcal{B}$ and $\mathcal{B}'$ using the PARTITIONNODES function.
2. Using the GETCLOSESTEQUIVALENCECLUSTER function, it finds a node $i_{outside_B}$ from the closest EC (respectively $i_{outside_{B'}}$) to $\mathcal{B}$ in $\mathcal{B}'$ (respectively to $\mathcal{B}'$ in $\mathcal{B}$).
3. It finally learns all the ECs and the edges between ECs by recursively calling LEARNEDGES$(i_{outside_B}, \mathcal{B}, \Sigma^o)$ followed by LEARNEDGES$(i_{outside_{B'}}, \mathcal{B}', \Sigma^o)$.

In Appendix E, we give the pseudo-code, proof of correctness and prove that this function is $\mathcal{O}(n^3)$, hence the algorithm is $\mathcal{O}(n^3)$.

6. Finite Sample Setting

If we have finite number of noisy samples, the algorithm can be modified to use the sample covariance matrix, $\hat{\Sigma}$, by replacing the conditions in Equations 18 and 26 by $|\hat{\Sigma}_{i_1 i_3}\hat{\Sigma}_{i_2 i_4} - \hat{\Sigma}_{i_1 i_4}\hat{\Sigma}_{i_2 i_3}| < \epsilon$ and $|\hat{\Sigma}_{i_1 i_3}\hat{\Sigma}_{i_2 i_4} - \hat{\Sigma}_{i_1 i_4}\hat{\Sigma}_{i_2 i_3}| > \epsilon$. The resulting sample complexity is polynomial in ϵ and logarithmic in the number of nodes. However, due to the exponential decay of the correlations with the diameter of the tree, the parameter ϵ required could be exponentially small in the diameter. Even for (near) balanced trees, the sample complexity becomes polynomial in the number of nodes, and in some settings (e.g., a chain), this would result in an exponential sample complexity. We believe it is possible to address this and recover logarithmic sample complexity in the number of nodes, by using an algorithm that only learns local structure (within a constant diameter such that the covariance is bounded away from zero), and then stitching the results together to form the complete tree. This is left for future work.

Acknowledgements

The authors would like to thank Po-Ling Loh and the anonymous reviewers for their constructive inputs. This research was partially supported by NSF EECS-1609279, NSF CCF-1302435 and NSF CNS-1704778.

References

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

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

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

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Janzamin, M. and Anandkumar, A. High-dimensional covariance decomposition into sparse markov and independence models. *The Journal of Machine Learning Research*, 15(1):1549–1591, 2014.

Kolar, M. and Xing, E. P. Estimating sparse precision matrices from data with missing values. In *Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK*, 2012.

Li, B., Wei, S., Wang, Y., and Yuan, J. Chernoff information of bottleneck gaussian trees. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pp. 970–974. IEEE, 2016.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Loh, P.-L. and Wainwright, M. J. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pp. 2726–2734, 2011.

Lounici, K. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

Meinshausen, N., Bühlmann, P., et al. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

Mossel, E., Roch, S., and Sly, A. Robust estimation of latent tree graphical models: Inferring hidden states with inexact parameters. *IEEE transactions on information theory*, 59(7):4357–4373, 2013.

Raskutti, G., Yu, B., Wainwright, M. J., and Ravikumar, P. K. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle. In *Advances in Neural Information Processing Systems*, pp. 1329–1336, 2009.

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

Stewart, G. and Sun, J.-G. *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press, 1990. ISBN 9780126702309.

Tan, V. Y., Anandkumar, A., and Willsky, A. S. Learning gaussian tree models: Analysis of error exponents and extremal structures. *arXiv preprint arXiv:0909.5216*, 2009.

Wang, J.-K. and Lin, S.-d. Robust inverse covariance estimation under noisy measurements. In *International Conference on Machine Learning*, pp. 928–936, 2014.

Wang, L. and Gu, Q. Robust gaussian graphical model estimation with arbitrary corruption. In *International Conference on Machine Learning*, pp. 3617–3626, 2017.

Wong, E., Awate, S., and Fletcher, P. T. Adaptive sparsity in gaussian graphical models. In *International Conference on Machine Learning*, pp. 311–319, 2013.

Yang, E. and Lozano, A. C. Robust gaussian graphical modeling with the trimmed graphical lasso. In *Advances in Neural Information Processing Systems*, pp. 2602–2610, 2015.

Yuan, M. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286, 2010.

Yuan, M. and Lin, Y. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Zhang, K., Gong, M., Ramsey, J., Batmanghelich, K., Spirtes, P., and Glymour, C. Causal discovery in the presence of measurement error: Identifiability conditions. *arXiv preprint arXiv:1706.03768*, 2017.