

Spectral Methods for Learning Multivariate Latent Tree Structure

Animashree Anandkumar¹, Kamalika Chaudhuri², Daniel Hsu³, Sham M. Kakade³,
Le Song⁴, and Tong Zhang⁵

¹Department of Electrical Engineering and Computer Science, UC Irvine

²Department of Computer Science and Engineering, UC San Diego

³Microsoft Research, New England

⁴Machine Learning Department, Carnegie Mellon University

⁵Department of Statistics, Rutgers University

July 8, 2011

Abstract

This work considers the problem of learning the structure of a broad class of multivariate latent variable tree models, which include a variety of continuous and discrete models (including the widely used linear-Gaussian models, hidden Markov models, and Markov evolutionary trees). The setting is one where we only have samples from certain observed variables in the tree and our goal is to estimate the tree structure (*i.e.*, the graph of how the underlying hidden variables are connected to the observed variables). We provide the Spectral Recursive Grouping algorithm, an efficient and simple bottom-up procedure for recovering the tree structure from independent samples of the observed variables. Our finite sample size bounds for exact recovery of the tree structure elucidate certain natural dependencies on underlying statistical and structural properties of the underlying joint distribution. Furthermore, our sample complexity guarantees have no explicit dependence on the dimensionality of the observed variables, making the algorithm applicable to many high-dimensional settings. At the heart of our algorithm is a spectral quartet test for determining the relative topology of a quartet of variables, which only utilizes certain second order statistics and is based on the determinants of certain cross-covariance matrices.

1 Introduction

Graphical models are a central tool in modern machine learning applications as they provide a succinct representational methodology to model high dimensional distributions. As such, they have enjoyed much success in varied AI and machine learning applications including natural language processing, speech recognition, robotics, computer vision, and bioinformatics.

The main challenges involved in learning graphical models include estimation and inference. While the body of techniques for probabilistic inference in graphical models is rather rich [23], current methods for tackling the more challenging problems of parameter and structure estimation are less developed and understood. The problem of parameter estimation involves estimation using samples at only certain observed nodes (sampled with respect to the underlying distribution). Here, the predominant approach is the expectation maximization (EM) algorithm and, only rather recently is the understanding of this algorithm improving (*e.g.*, for mixtures of Gaussians, see [8, 4]).

E-mail: a.anandkumar@uci.edu, kamalika@cs.ucsd.edu, dahsu@microsoft.com, skakade@microsoft.com, lesong@cs.cmu.edu, tzhang@stat.rutgers.edu

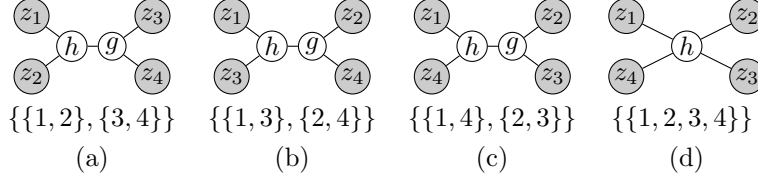


Figure 1: The four possible tree topologies over leaves $\{z_1, z_2, z_3, z_4\}$.

The problem of structure learning is to estimate the structure of the underlying graph of the graphical model. In general, this problem is NP-hard and becomes even more challenging when some variables are hidden or latent. The main approaches for structure estimation are either greedy approaches [7, 18] or more recently, based on convex relaxation [20].

This work focuses on the class of multivariate latent tree graphical models. Here, the assumption is that the underlying graph is a tree (*e.g.*, hidden Markov model, binary evolutionary tree), and only samples from a set of (multivariate) observed variables (the leaves of the tree) are available for learning the structure. Latent tree graphical models are relevant in many applications, *e.g.*, in financial modeling for discovering hidden relationships between various stocks [5], in computer vision, to learn the co-occurrences of various objects in images [6] and in phylogenetics, where the task is to estimate the evolutionary tree based on the genetic material of the surviving species [10].

Previous methods for latent-tree estimation take a rather direct approach for learning the latent tree structure through the *quartet tests* [11, 19]; roughly speaking, a quartet test seeks to discover the relative configuration of four (possibly non-adjacent) observed nodes (See Figure 1 for example configurations, which are discussed in more detail later). The quartet test and other algorithms are known as distance-based methods, since they rely on the presence of an additive metric on the unknown tree. Recently, detailed statistical and computational treatments of the special case where the hidden nodes are scalar variables have been considered [5, 9]; these works explicitly characterize the complexity of certain algorithms in terms of both natural statistical parameters (*e.g.*, how correlations between variables decay in the graph) and natural structural parameters (*e.g.*, a precise notion of *effective* depth, which could be much small than the true depth).

This work considers the more challenging multivariate case (*e.g.*, where the hidden and observed variables are random vectors). This setting is very general, allowing the nodes in the tree to even have different state spaces (*e.g.*, mixtures of Gaussians) which cannot be handled by previous works. For the case of parameter estimation, spectral methods have enjoyed much recent success, for both parameter estimation in time series and in tree models [12, 21, 22]. For structure estimation, thus far, spectral based approaches are used primarily through convex-relaxation approaches (*e.g.*, [3]), although some techniques from the phylogenetic reconstruction literature have proved relevant [14].

This work develops a *spectral quartet test* and characterizes the complexity of our algorithm based on certain natural statistical and structural parameters. While the classical multivariate statistical tests often focus on determining whether variables are uncorrelated (*e.g.*, certain canonical correlation analysis based tests [2, 15]), they are not applicable to reveal hidden tree structures. Here, our first key technical contribution is a hypothesis testing methodology for the delicate problem of correctly determining the relative tree topology of a quartet based on the determinant of certain cross-covariance matrices. We propose the Spectral Recursive Grouping algorithm, a bottom-up tree construction procedure based on a modification of the recursive grouping algorithm of [5], with our spectral quartet test at its core. Our main result shows this algorithm returns the correct latent structure (with high probability) in a provably efficient manner (both in terms of computation time and sample size dependency under appropriate conditions). Importantly, the sample complexity has no explicit dependence on the dimensionality of the observed variables, making the algorithm applicable to many high-dimensional settings that cannot be handled by previous methods. Furthermore, our methods are applicable to settings with both discrete and continuous variables. While we do not directly address the question of parameter estimation, provable parameter estimation methods may

adapted from the subspace identification methods and spectral algorithms in [16, 12, 17].

2 Preliminaries

2.1 Latent variable tree models

Let \mathcal{T} be an undirected tree graphical model with leaves $\mathcal{V}_{\text{obs}} := \{x_1, x_2, \dots, x_n\}$ and internal nodes $\mathcal{V}_{\text{hid}} := \{h_1, h_2, \dots, h_m\}$. The leaves are termed the *observed variables* and the internal nodes *hidden variables*. Note that all nodes are high-dimensional random vectors, and for the rest of the paper, the term random variable generally refers to random vectors over high dimensions.

Each observed variable $x \in \mathcal{V}_{\text{obs}}$ is modeled as random vector in \mathbb{R}^d , and each hidden variable $h \in \mathcal{V}_{\text{hid}}$ as a random vector in \mathbb{R}^k . The joint distribution over all the variables $\mathcal{V}_{\mathcal{T}} := \mathcal{V}_{\text{obs}} \cup \mathcal{V}_{\text{hid}}$ is assumed satisfy conditional independence properties specified by the tree structure over the variables. Namely, if $\text{Neigh}(v)$ is the set of neighbors of v in \mathcal{T} , then v is conditionally independent of $\mathcal{V}_{\mathcal{T}} \setminus \text{Neigh}(v)$, given $\text{Neigh}(v)$.

2.2 Structural and distributional assumptions

The class of models considered are specified by the following structural and distributional assumptions.

First, assume that each variable $v \in \mathcal{V}_{\mathcal{T}}$ is, in expectation, linearly related to each of its neighbors. Let $h \in \mathcal{V}_{\text{hid}}$ be any hidden variable. For any hidden variable $g \in \text{Neigh}(h)$, assume there exists a matrix $A_{(g|h)} \in \mathbb{R}^{k \times k}$ such that

$$\mathbb{E}[g|h] = A_{(g|h)}h;$$

and for any observed variable $x \in \text{Neigh}(h)$, assume there exists a matrix $C_{(x|h)} \in \mathbb{R}^{d \times k}$ such that

$$\mathbb{E}[x|h] = C_{(x|h)}h.$$

We refer to this class of graphical models as *linear tree models*. Such models include a variety of continuous and discrete tree distributions (as well as hybrid combinations of the two) which are widely used in practice. Continuous linear tree models include linear-Gaussian models and Kalman filters. In the discrete case, suppose that the observed variables take on d values, and hidden variables take k values. Then, each variable is represented by a binary vector in $\{0, 1\}^s$, where $s = d$ for the observed variables and $s = k$ for the hidden variables (in particular, if the variable takes value i , then the corresponding vector is the i -th coordinate vector), and any conditional distribution between the variables is represented by a linear relationship. Thus, discrete linear tree models include discrete hidden Markov models [12] and Markovian evolutionary trees [14].

In addition to the linearity, the following conditions are assumed in order to recover the hidden tree structure. For any matrix M , let $\sigma_t(M)$ denote its t -th largest singular value.

Condition 1 (Rank condition). The variables in $\mathcal{V}_{\mathcal{T}} = \mathcal{V}_{\text{hid}} \cup \mathcal{V}_{\text{obs}}$ obey the following rank conditions.

1. For each hidden variable $h \in \mathcal{V}_{\text{hid}}$, $\mathbb{E}[hh^\top]$ has rank k (i.e., $\sigma_k(\mathbb{E}[hh^\top]) > 0$).
2. For each observed variable $x \in \mathcal{V}_{\text{obs}}$ with neighbor $h \in \mathcal{V}_{\text{hid}}$, $C_{(x|h)}$ has rank k .

The rank condition is a generalization of parameter identifiability conditions in latent variable models [1, 14, 12] which rules out various (provably) hard instances in discrete variable settings [14].

Condition 2 (Non-redundancy condition). Each hidden variable has at least three neighbors. Furthermore, there exists $\rho_{\max}^2 > 0$ such that for each pair of distinct hidden variables $h, g \in \mathcal{V}_{\text{hid}}$,

$$\frac{\det(\mathbb{E}[hg^\top])^2}{\det(\mathbb{E}[hh^\top])\det(\mathbb{E}[gg^\top])} \leq \rho_{\max}^2 < 1.$$

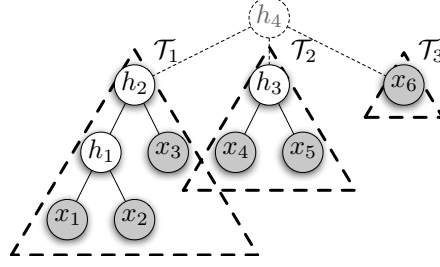


Figure 2: Set of trees $\mathcal{F}_{h_4} = \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$ obtained if h_4 is removed.

The requirement for each hidden node to have three neighbors is natural; otherwise, the hidden node can be eliminated. This condition is also analogous to the conditions presented in the foundational work in [19] (In fact, in the multivariate Gaussian case, this condition is identical to that in [19]. In general, the above condition is stronger in that it only considers second moments.) The quantity ρ_{\max} is a natural multivariate generalization of correlation. First, note $\rho_{\max} \leq 1$. If $\rho_{\max} = 1$ is achieved with some h and g , then h and g are completely correlated, implying the existence of a deterministic map between hidden nodes h and g ; hence simply merging the two nodes into a single node h (or g) resolves this issue. Therefore the non-redundancy condition simply means that any two hidden nodes h and g cannot be further reduced into one node. Clearly, this condition is necessary for the goal of identifying the correct tree structure. Previous works [18] show that a similar condition ensures identifiability for *general* latent tree models. The non-degeneracy condition is a generalization of this condition suitable for the multivariate setting.

Our learning guarantees also utilize a correlation condition that is a natural generalization of depth conditions considered in the phylogenetics literature [11, 14]. To state the condition, first define \mathcal{F}_h to be the set of trees of that remain after a hidden variable $h \in \mathcal{V}_{\text{hid}}$ is removed from \mathcal{T} (see Figure 2). Also, for any subtree \mathcal{T}' of \mathcal{T} , let $\mathcal{V}_{\text{obs}}[\mathcal{T}'] \subseteq \mathcal{V}_{\text{obs}}$ be the observed variables in \mathcal{T}' .

Condition 3 (Correlation condition). There exists $\gamma_{\min} > 0$ such that for all hidden variables $h \in \mathcal{V}_{\text{hid}}$ and all pairs of trees $\{\mathcal{T}_1, \mathcal{T}_2\} \subseteq \mathcal{F}_h$ in the forest obtained if h is removed from \mathcal{T} ,

$$\max_{x_1 \in \mathcal{V}_{\text{obs}}[\mathcal{T}_1], x_2 \in \mathcal{V}_{\text{obs}}[\mathcal{T}_2]} \sigma_k(\mathbb{E}[x_1 x_2^\top]) \geq \gamma_{\min}.$$

The quantity γ_{\min} is related to the *effective depth* of \mathcal{T} , which is the maximum graph distance between a hidden variable and its closest observed variable [11, 5]. The effective depth is at most logarithmic in the number of variables (as achieved by a complete binary tree), though it can also be a constant if every hidden variable is close to an observed variable (*e.g.*, in a hidden Markov model the true depth, or diameter, is $m + 1$ while the effective depth is 1). If all of the matrices $A_{(g|h)}$ and $C_{(x|h)}$ relating neighboring variables in \mathcal{T} are all well-conditioned, then γ_{\min} is at worst exponentially small in the effective depth, and therefore at worst polynomially small in the number of variables.

Finally, also define

$$\gamma_{\max} := \max_{\{x_1, x_2\} \subseteq \mathcal{V}_{\text{obs}}} \{\sigma_1(\mathbb{E}[x_1 x_2^\top])\} = \max_{\{x_1, x_2\} \subseteq \mathcal{V}_{\text{obs}}} \{\|\mathbb{E}[x_1 x_2^\top]\|\}$$

to be the largest spectral norm of any second-moment matrix between observed variables. Note $\gamma_{\max} \leq 1$ in the discrete case, and, in the continuous case, $\gamma_{\max} \leq 1$ if each observed random vector is in isotropic position.

In this work, the Euclidean norm of a vector x is denoted by $\|x\|$, and the (induced) spectral norm of a matrix A is denoted by $\|A\|$, *i.e.*, $\|A\| := \sup\{\|Ax\| : \|x\| = 1\}$.

3 Structural consistency

Our main learning guarantee is given in the following theorem.

Theorem 1. *Let $\eta \in (0, 1)$. Assume the linear tree model, given by a tree \mathcal{T} over variables $\mathcal{V}_{\mathcal{T}} = \mathcal{V}_{\text{obs}} \cup \mathcal{V}_{\text{hid}}$, satisfies Conditions 1, 2, and 3. Suppose the Spectral Recursive Grouping algorithm (Algorithm 2) is provided N independent samples from the distribution over \mathcal{V}_{obs} , and uses parameters given by*

$$\Delta_{x_i, x_j} := \sqrt{\frac{2B_{x_i, x_j}^2 t_{x_i, x_j}}{N}} + \frac{M_{x_i} M_{x_j} t_{x_i, x_j}}{3N}, \quad \theta := \left(\frac{8k + \gamma_{\min}/\gamma_{\max}}{8k + 2\gamma_{\min}/\gamma_{\max}} \right) \cdot \gamma_{\min} \quad (1)$$

where

$$\begin{aligned} B_{x_i, x_j} &:= \sqrt{\max\{\|\mathbb{E}[\|x_i\|^2 x_j x_j^\top]\|, \|\mathbb{E}[\|x_j\|^2 x_i x_i^\top]\|\}}, & M_{x_i} &\geq \|x_i\| \quad \text{almost surely,} \\ \bar{d}_{x_i, x_j} &:= \frac{\mathbb{E}[\|x_i\|^2 \|x_j\|^2] - \text{tr}(\mathbb{E}[x_i x_j^\top] \mathbb{E}[x_j x_i^\top])}{\max\{\|\mathbb{E}[\|x_j\|^2 x_i x_i^\top]\|, \|\mathbb{E}[\|x_i\|^2 x_j x_j^\top]\|\}}, & t_{x_i, x_j} &:= 4 \ln(4 \bar{d}_{x_i, x_j} n / \eta). \end{aligned}$$

Let $B := \max_{x_i, x_j \in \mathcal{V}_{\text{obs}}} \{B_{x_i, x_j}\}$, $M := \max_{x_i \in \mathcal{V}_{\text{obs}}} \{M_{x_i}\}$, $t := \max_{x_i, x_j \in \mathcal{V}_{\text{obs}}} \{t_{x_i, x_j}\}$. If

$$N > \frac{200 \cdot k^2 \cdot B^2 \cdot t}{\left(\frac{\gamma_{\min}^2}{\gamma_{\max}} \cdot (1 - \rho_{\max})\right)^2} + \frac{7 \cdot k \cdot M^2 \cdot t}{\frac{\gamma_{\min}^2}{\gamma_{\max}} \cdot (1 - \rho_{\max})},$$

then with probability at least $1 - \eta$, the Spectral Recursive Grouping algorithm returns $\hat{\mathcal{T}} = \mathcal{T}$.

Consistency is implied by the above theorem with an appropriate scaling of δ with n . The theorem reveals that the sample complexity of the algorithm depends solely on intrinsic spectral properties of the distribution. Note that there is no explicit dependence on the dimensions of the observable variables, which makes the result applicable to high-dimensional settings.

4 Spectral quartet tests

This section describes the core of our learning algorithm, a spectral quartet test that determines topology of the subtree induced by four observed variables $\{z_1, z_2, z_3, z_4\}$. There are four possibilities for the induced subtree, as shown in Figure 1. Our quartet test either returns the correct induced subtree among possibilities in Figure 1(a)–(c); or it abstains and outputs \perp . If the test returns \perp , then no guarantees are provided about the induced subtree topology. If it does return a subtree, then the output is guaranteed to be the correct induced subtree (with high probability).

The quartet test proposed is described in Algorithm 1. The notation $[a]_+$ denotes $\max\{0, a\}$ and $[t]$ (for an integer t) denotes the set $\{1, 2, \dots, t\}$.

The quartet test is defined with respect to four observed variables $\mathcal{Z} := \{z_i : i \in [4]\}$. For each pair of variables z_i and z_j , it takes an input an empirical estimate $\hat{\Sigma}_{ij}$ of the second-moment matrix $\Sigma_{ij} = \mathbb{E}[z_i z_j^\top]$, and confidence bound parameters $\Delta_{i,j}$ which are functions of N , the number of samples used to compute the $\hat{\Sigma}_{ij}$'s, a high-probability confidence parameter δ , and of properties of the distribution of z_i and z_j . In practice, one uses a single threshold Δ for all pairs, which is tuned by the algorithm. Our theoretical analysis also applies to this case. The output of the test is either \perp or a *pairing* of the variables $\{\{i, j\}, \{i', j'\}\}$. For example, if the output is the pairing is $\{\{1, 2\}, \{3, 4\}\}$, then Figure 1(a) is the output topology.

Even though the configuration in Figure 1(d) is a possibility, the spectral quartet test never returns $\{\{1, 2, 3, 4\}\}$, as there is no correct pairing of \mathcal{Z} . The topology $\{\{1, 2, 3, 4\}\}$ can be viewed as a degenerate case of $\{\{1, 2\}, \{3, 4\}\}$ where the hidden variables h and g are deterministically identical, and Condition 2 fails to hold with respect to h and g .

Algorithm 1 Spectral quartet test for four observed variables $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$.

Input: For each $\{i, j\} \subset \{1, 2, 3, 4\}$, an empirical estimate $\hat{\Sigma}_{ij}$ of the second-moment matrix $\mathbb{E}[z_i z_j^\top]$ and a confidence parameter $\Delta_{i,j} > 0$.

Output: Either a pairing $\{\{i, j\}, \{i', j'\}\}$, or \perp .

- 1: For each $\{i, j\} \subset \{1, 2, 3, 4\}$, compute $\{\sigma_s(\hat{\Sigma}_{ij}) : s \in [k]\}$, the k largest singular values of $\hat{\Sigma}_{ij}$.
- 2: If there exists a partition of $\{1, 2, 3, 4\} = \{i, j\} \cup \{i', j'\}$ such that

$$\prod_{s=1}^k (\sigma_s(\hat{\Sigma}_{i'j'}) + \Delta_{i',j'}) (\sigma_s(\hat{\Sigma}_{ij}) + \Delta_{i,j}) < \prod_{s=1}^k [\sigma_s(\hat{\Sigma}_{ij}) - \Delta_{i,j}]_+ [\sigma_s(\hat{\Sigma}_{i'j'}) - \Delta_{i',j'}]_+$$

then return the pairing $\{\{i, j\}, \{i', j'\}\}$.

- 3: Otherwise, return \perp .
-

4.1 Properties of the spectral quartet test

With exact second moments: The spectral quartet test is motivated by the following lemma, which shows the relationship between the singular values of second-moment matrices of the z_i 's and the induced topology among them in the latent tree.

Lemma 1 (Perfect quartet test). *Suppose that the observed variables $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$ have the true induced tree topology shown in Figure 1(a), and the tree model satisfies Condition 1. If $\Sigma_{ij} := \mathbb{E}[z_i z_j^\top]$, $\Sigma_{hh} := \mathbb{E}[hh^\top]$, $\Sigma_{gg} := \mathbb{E}[gg^\top]$, $\Sigma_{hg} := \mathbb{E}[hg^\top]$, and $\sigma_1(\Sigma_{ij}), \dots, \sigma_k(\Sigma_{ij})$ are the k largest singular values of Σ_{ij} , then*

$$\frac{\prod_{s=1}^k \sigma_s(\Sigma_{13}) \sigma_s(\Sigma_{24})}{\prod_{s=1}^k \sigma_s(\Sigma_{12}) \sigma_s(\Sigma_{34})} = \frac{\prod_{s=1}^k \sigma_s(\Sigma_{14}) \sigma_s(\Sigma_{23})}{\prod_{s=1}^k \sigma_s(\Sigma_{12}) \sigma_s(\Sigma_{34})} = \frac{\det(\Sigma_{hg})^2}{\det(\Sigma_{hh}) \det(\Sigma_{gg})} \leq 1 \quad (2)$$

$$\prod_{s=1}^k \sigma_s(\Sigma_{13}) \sigma_s(\Sigma_{24}) = \prod_{s=1}^k \sigma_s(\Sigma_{14}) \sigma_s(\Sigma_{23}).$$

This lemma shows that with the true second-moment matrices (as opposed to the empirical) then the inequality in (2) can be utilized to deduce the correct topology.

Reliability: The next lemma shows that even if the singular values of Σ_{ij} are not known exactly, then with valid confidence intervals (that contain these singular values) a robust test can be constructed which is reliable in the following sense: if it does not output \perp , then the output topology is indeed the correct topology.

Lemma 2 (Reliability). *Consider the setup of Lemma 1, and suppose that Figure 1(a) is the correct topology. If for all $\{i, j\}$ and for all $s \in [k]$,*

$$\sigma_s(\hat{\Sigma}_{ij}) - \Delta_{i,j} \leq \sigma_s(\Sigma_{ij}) \leq \sigma_s(\hat{\Sigma}_{ij}) + \Delta_{i,j},$$

and if Algorithm 1 returns a pairing $\{\{i, j\}, \{i', j'\}\}$, then $\{\{i, j\}, \{i', j'\}\} = \{\{1, 2\}, \{3, 4\}\}$.

In other words, the spectral quartet test never returns an incorrect pairing as long as the singular values of Σ_{ij} lie in an interval of length $2\Delta_{i,j}$ around the singular values of $\hat{\Sigma}_{ij}$. The lemma below shows how to set the $\Delta_{i,j}$ s as a function of N , δ and properties of the distributions of z_i and z_j so that this holds with probability $1 - \delta$.

Lemma 3 (Confidence intervals). *Let $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$ be four random vectors. Let $\|z_i\| \leq M_i$ almost surely, and let $\delta \in (0, 1/6)$. If each empirical second-moment matrix $\hat{\Sigma}_{ij}$ is computed using N iid copies of*

z_i and z_j , and if

$$\bar{d}_{i,j} := \frac{\mathbb{E}[\|z_i\|^2 \|z_j\|^2] - \text{tr}(\Sigma_{ij} \Sigma_{ij}^\top)}{\max\{\|\mathbb{E}[\|z_j\|^2 z_i z_i^\top]\|, \|\mathbb{E}[\|z_i\|^2 z_j z_j^\top]\|\}}, \quad t_{i,j} := 1.55 \ln(24\bar{d}_{i,j}/\delta),$$

$$\Delta_{i,j} \geq \sqrt{\frac{2 \max\{\|\mathbb{E}[\|z_j\|^2 z_i z_i^\top]\|, \|\mathbb{E}[\|z_i\|^2 z_j z_j^\top]\|\}}{N}} + \frac{M_i M_j t_{i,j}}{3N},$$

then with probability $1 - \delta$, for all $\{i, j\}$ and all $s \in [k]$,

$$\sigma_s(\hat{\Sigma}_{ij}) - \Delta_{i,j} \leq \sigma_s(\Sigma_{ij}) \leq \sigma_s(\hat{\Sigma}_{ij}) + \Delta_{i,j}.$$

Conditions for returning a correct pairing: The conditions under which Algorithm 1 returns an induced topology (as opposed to \perp) are now provided. For simplicity, only the case where $\mathbb{E}[hg^\top]$ is rank k is considered; the other case is analyzed in the appendix.

An important parameter in this analysis is the level of non-redundancy between the hidden variables h and g . Let $\Sigma_{hh} := \mathbb{E}[hh^\top]$, $\Sigma_{gg} := \mathbb{E}[gg^\top]$, $\Sigma_{hg} := \mathbb{E}[hg^\top]$, and

$$\rho^2 := \det(\Sigma_{hg})^2 / [\det(\Sigma_{hh}) \det(\Sigma_{gg})]. \quad (3)$$

If Figure 1(a) is the correct induced topology among $\{z_1, z_2, z_3, z_4\}$, then the smaller ρ is, the sharper is the distinction between $\det(\Sigma_{12}) \det(\Sigma_{34})$ and either of $\det(\Sigma_{13}) \det(\Sigma_{24})$ and $\det(\Sigma_{14}) \det(\Sigma_{23})$. Thus, Δ_{ij} s need to be smaller as ρ increases; as the Δ_{ij} s depend inversely on the number of samples, this can be done by using more samples. Lemma 4 quantifies how small the Δ_{ij} s need to be for the quartet test to return a correct pairing.

Lemma 4 (Correct pairing, $\text{rank}(\Sigma_{hg}) = k$). *Suppose that (i) the observed variables $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$ have the true induced tree topology shown in Figure 1(a), (ii) the tree model satisfies Condition 1 and $\rho < 1$ (where ρ is defined in (3)), (iii) Σ_{hg} has rank k , and (iv) the confidence bounds in (2) hold for all $\{i, j\}$ and all $s \in [k]$. If*

$$\Delta_{i,j} < \frac{1}{8k} \cdot \min\left\{1, \frac{1}{\rho} - 1\right\} \cdot \min_{\{i,j\}} \{\sigma_k(\Sigma_{ij})\}$$

for each $\{i, j\}$, then Algorithm 1 returns the correct pairing $\{\{1, 2\}, \{3, 4\}\}$.

5 The Spectral Recursive Grouping algorithm

The Spectral Recursive Grouping algorithm is presented in Figure 2, for learning the multivariate latent tree structure based on the spectral properties discussed in the previous section. It is based on a modification of the recursive grouping (RG) procedure proposed in [5]. RG builds the tree in a bottom-up fashion, where the initial working set of variables are the observed variables. In each iteration, the algorithm determines which pairs of variables in the working set are siblings; a new hidden variable is then added to the tree with the siblings as its children. The siblings are then removed from the working set, and the hidden variable is added. The process repeats until the entire tree is constructed.

Our modification of RG uses the spectral quartet tests from Section 4 rather than an explicit additive tree metric as in [5]. Note that because the test may return \perp (a null result), this algorithm uses the quartet tests to rule out possible siblings. For example, if a test returns $\{\{1, 2\}, \{3, 4\}\}$, then $\{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}$ are ruled out as possible siblings. Moreover, a test involving a hidden variable is performed by examining the results of tests among all choices of its descendants. The specific choice made in the test allows us to work with the closest two observed variables in any two neighboring nodes, which have a relatively strong correlation that is lower bounded (in the proof) by the structural quantity γ_{\min} specified in Condition 3. This means dependencies on unreliable estimates of long range correlations are avoidable. Under appropriate

Algorithm 2 Spectral Recursive Grouping.

Input: N independent samples from the distribution over $\mathcal{V}_{\text{obs}} = \{x_1, x_2, \dots, x_n\}$; thresholds $\Delta_{x,y} > 0$ for pairs $\{x, y\} \subseteq \mathcal{V}_{\text{obs}}$, and $\theta > 0$ (as given in (1)).

Output: Estimated tree structure $\hat{\mathcal{T}}$.

- 1: $\hat{\mathcal{V}}_1 := \mathcal{V}_{\text{obs}}$; $\hat{\mathcal{L}}(x) := \{x\}$ for all $x \in \mathcal{V}_{\text{obs}}$; and let $\hat{\mathcal{T}}$ have all variables in \mathcal{V}_{obs} as leaves.
 - 2: For each pair $\{x, y\} \subseteq \mathcal{V}_{\text{obs}}$, let $\hat{\Sigma}_{x,y}$ be the empirical second moment matrix for variables x and y computed using the N samples.
 - 3: **for** iteration $\ell = 1, 2, \dots$ until $|\hat{\mathcal{V}}_\ell| < 4$ **do**
 - 4: Let $\hat{\mathcal{S}}_\ell$ be the set of all pairs of variables $\{u, v\} \subseteq \hat{\mathcal{V}}_\ell$ such that $\sigma_k(\hat{\Sigma}_{x,y}) \geq \theta$ for some $x \in \hat{\mathcal{L}}(u)$ and $y \in \hat{\mathcal{L}}(v)$.
 - 5: **for** each quartet $\{v_1, v_2, v_3, v_4\}$ in $\hat{\mathcal{V}}_\ell$, and each choice of $(x_1, x_2, x_3, x_4) \in \hat{\mathcal{L}}(v_1) \times \hat{\mathcal{L}}(v_2) \times \hat{\mathcal{L}}(v_3) \times \hat{\mathcal{L}}(v_4)$ **do**
 - 6: Perform spectral quartet test on $\{x_1, x_2, x_3, x_4\}$ in $\hat{\mathcal{V}}_\ell$ using the $\hat{\Sigma}_{x_i, x_j}$ and thresholds Δ_{x_i, x_j} .
 - 7: If the test returns a result that separates some x_i and x_j (note: the result $\{\{x_i, x_{i'}\}, \{x_j, x_{j'}\}\}$ separates x_i and x_j , x_i and $x_{j'}$, etc.), and if $\{v_i, v_j\} \in \hat{\mathcal{S}}_\ell$, then mark the pair $\{v_i, v_j\}$ as *broken*.
 - 8: **end for**
 - 9: Let $\tilde{\mathcal{S}}_\ell \subseteq \hat{\mathcal{S}}_\ell$ be the set of *unbroken* pairs (i.e., not marked as *broken*) in $\hat{\mathcal{S}}_\ell$.
 - 10: **for** each connected component $C \subseteq \hat{\mathcal{V}}_\ell$ in the graph $(\hat{\mathcal{V}}_\ell, \tilde{\mathcal{S}}_\ell)$ **do**
 - 11: If $C = \{v\}$ is a singleton, then add v to $\hat{\mathcal{V}}_{\ell+1}$.
 - 12: If $C = \{v_1, v_2, \dots, v_q\}$ is not a singleton, then let h be a new hidden variable in $\hat{\mathcal{T}}$ with children $\{v_1, v_2, \dots, v_q\}$, add h to $\hat{\mathcal{V}}_{\ell+1}$, and let $\hat{\mathcal{L}}(h) := \bigcup_{i=1}^q \hat{\mathcal{L}}(v_i)$.
 - 13: **end for**
 - 14: If $\hat{\mathcal{V}}_{\ell+1} = \hat{\mathcal{V}}_\ell$ (i.e., all variables formed singleton components), then exit the for-loop.
 - 15: **end for**
 - 16: If $|\hat{\mathcal{V}}_\ell| \geq 3$, then let $\hat{\mathcal{V}}_\ell$ be the children of a new hidden node h in $\hat{\mathcal{T}}$. If $|\hat{\mathcal{V}}_\ell| = 2$, then let these two variables be adjacent in $\hat{\mathcal{T}}$. Return $\hat{\mathcal{T}}$.
-

conditions, the proof shows that the pairs that are not eliminated by any of the tests are the true sibling nodes.

For clarity, define the *correct* recursive grouping process—i.e., the process that the algorithm seeks to mimic—as follows. Consider a sequence of *levels*, which corresponds to the iterations $\ell = 1, 2, \dots$ of the algorithm. First, let $\bar{\mathcal{V}}_1 := \mathcal{V}_{\text{obs}}$ and $\bar{\mathcal{L}}(x) = \{x\}$ for all $x \in \mathcal{V}_{\text{obs}}$. Now, in level ℓ (for $\ell = 1, 2, \dots$), let $\bar{\mathcal{S}}_\ell$ be the set of *sibling pairs* among variables in $\bar{\mathcal{V}}_\ell$ (i.e., pairs $\{u, v\} \subseteq \bar{\mathcal{V}}_\ell$ such that u and v have a common neighbor). These sibling relationships define a graph $(\bar{\mathcal{V}}_\ell, \bar{\mathcal{S}}_\ell)$ whose connected components are groups of siblings with the same common neighbor, or *parent*; singleton groups are those $v \in \bar{\mathcal{V}}_\ell$ that do not have any siblings in $\bar{\mathcal{V}}_\ell$. The variables in a non-singleton group $\{v_1, v_2, \dots, v_q\}$ are neighbors of a hidden variable $h \in \mathcal{V}_{\text{hid}}$ (i.e., have h as a parent); let $\bar{\mathcal{L}}(h) := \bigcup_{i=1}^q \bar{\mathcal{L}}(v_i) \subseteq \mathcal{V}_{\text{obs}}$ be the *descendants* of h . Now, define $\bar{\mathcal{V}}_{\ell+1}$ to consist of (i) all variables in $\bar{\mathcal{V}}_\ell$ not involved in sibling pairs in $\bar{\mathcal{S}}_\ell$, and (ii) also the hidden variables that are the parents of any sibling pairs in $\bar{\mathcal{S}}_\ell$.

It is clear then that in each level ℓ , $\bar{\mathcal{V}}_\ell$ can be viewed as the current set of leaves of the tree, and that sibling groups are recursively pruned before going to the next level $\ell + 1$. The process repeats while $|\bar{\mathcal{V}}_\ell| \geq 4$. See Figure 3 for an example. The goal of the recursive grouping algorithm is to discover the groupings $\bar{\mathcal{L}}(u)$ for all $u \in \bar{\mathcal{V}}_\ell$ in each iteration ℓ , as this recursively defines the tree structure \mathcal{T} .

References

- [1] E.S. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.
- [2] M. S Bartlett. Further aspects of the theory of multiple regression. *Proc. Camb. Phil. Soc.*, 34:33–40, 1938.

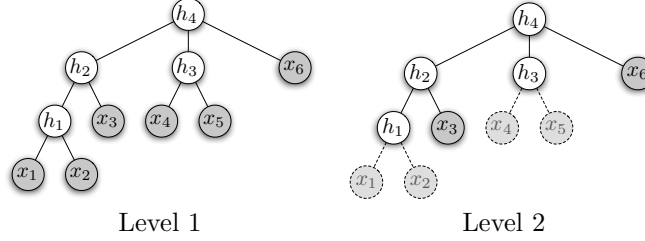


Figure 3: Example of a latent variable tree model \mathcal{T} . In the first level, $\bar{\mathcal{V}}_1 = \{x_1, x_2, \dots, x_6\}$ and $\bar{\mathcal{S}}_1 = \{\{x_1, x_2\}, \{x_4, x_5\}\}$. In the second level, $\bar{\mathcal{V}}_2 = \{h_1, x_3, h_3, x_6\}$ and $\bar{\mathcal{S}}_2 = \{\{h_1, x_3\}, \{h_3, x_6\}\}$. In the third level, $\bar{\mathcal{V}}_3 = \{h_2, h_4\}$. Also, $\bar{\mathcal{L}}(h_1) = \{x_1, x_2\}$, $\bar{\mathcal{L}}(h_3) = \{x_4, x_5\}$, $\bar{\mathcal{L}}(h_2) = \{x_1, x_2, x_3\}$, and $\bar{\mathcal{L}}(h_4) = \{x_4, x_5, x_6\}$.

- [3] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent Variable Graphical Model Selection via Convex Optimization. *ArXiv e-prints*, August 2010.
- [4] Kamalika Chaudhuri, Sanjoy Dasgupta, and Andrea Vattani. Learning mixtures of gaussians using the k-means algorithm, 2009. arXiv:0912.0086.
- [5] M.J. Choi, V. Tan, A. Anandkumar, and A. Willsky. Learning Latent Tree Graphical Models. *accepted to Journal of Machine Learning Research, available on Arxiv*, Feb. 2011.
- [6] Myung Jin Choi, Joseph J. Lim, Antonio Torralba, and Alan S. Willsky. Exploiting hierarchical context on a large database of object categories. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [7] C. Chow and C. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Tran. on Information Theory*, 14(3):462–467, 1968.
- [8] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.
- [9] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Optimal phylogenetic reconstruction. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, STOC ’06, pages 159–168, New York, NY, USA, 2006. ACM.
- [10] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, 1999.
- [11] P. L. Erdős, L. A. Székely, M. A. Steel, and T. J. Warnow. A few logs suffice to build (almost) all trees: Part II. *Theoretical Computer Science*, 221:153–184, 1999.
- [12] D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *Twenty-Second Annual Conference on Learning Theory*, 2009.
- [13] D. Hsu, S. M. Kakade, and T. Zhang. Dimension-free tail inequalities for sums of random matrices, 2011. arXiv:1104.1672.
- [14] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. *Annals of Applied Probability*, 16(2):583–614, 2006.
- [15] R. J. Muirhead and C. M. Waternaux. Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations. *Biometrika*, 67(1):31–43, 1980.
- [16] P. V. Overschee and B. De Moor. *Subspace Identification of Linear Systems*. Kluwer Academic Publishers, 1996.
- [17] A. Parikh, L. Song, and E. P. Xing. A spectral algorithm for latent tree graphical models. In *Intl. Conf. on Machine Learning*, 2011.
- [18] J. Pearl. *Probabilistic Reasoning in Intelligent Systems—Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- [19] J. Pearl and M. Tarsi. Structuring causal trees. *Journal of Complexity*, 2(1):60–77, 1986.
- [20] P. Ravikumar, M.J. Wainwright, and J. Lafferty. High-dimensional Ising Model Selection Using l1-Regularized Logistic Regression. *Annals of Statistics*, 2008.

- [21] Sajid M. Siddiqi, Byron Boots, and Geoffrey J. Gordon. Reduced-rank hidden markov models. <http://arxiv.org/abs/0910.0902>, abs/0910.0902, 2009.
- [22] Le Song, Sajid M. Siddiqi, Geoffrey J. Gordon, and Alexander J. Smola. Hilbert space embeddings of hidden markov models. In *International Conference on Machine Learning*, pages 991–998, 2010.
- [23] 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.

A Sample-based confidence intervals for singular values

We show how to derive confidence bounds for the singular values of $\Sigma_{ij} := \mathbb{E}[z_i z_j^\top]$ for $\{i, j\} \subset \{1, 2, 3, 4\}$ from N iid copies of the random vectors $\{z_1, z_2, z_3, z_4\}$. That is, we show how to set $\Delta_{i,j}$ so that, with high probability,

$$\sigma_s(\hat{\Sigma}_{ij}) - \Delta_{ij} \leq \sigma_s(\Sigma_{ij}) \leq \sigma_s(\hat{\Sigma}_{ij}) + \Delta_{ij}$$

for all $\{i, j\}$ and all $s \in [k]$.

We state exponential tail inequalities for the spectral norm of the estimation error $\hat{\Sigma}_{ij} - \Sigma_{ij}$. The first exponential tail inequality is stated for general random vectors under Bernstein-type conditions, and the second is specific to random vectors in the discrete setting.

Lemma 5. *Let z_i and z_j be random vectors such that $\|z_i\| \leq M_i$ and $\|z_j\| \leq M_j$ almost surely, and let*

$$\bar{d}_{ij} := \frac{\mathbb{E}[\|z_i\|^2 \|z_j\|^2] - \text{tr}(\Sigma_{ij} \Sigma_{ij}^\top)}{\max\{\|\mathbb{E}[\|z_j\|^2 z_i z_i^\top]\|, \|\mathbb{E}[\|z_i\|^2 z_j z_j^\top]\|\}} \leq \max\{\dim(z_i), \dim(z_j)\}.$$

Let $\Sigma_{ij} := \mathbb{E}[z_i z_j^\top]$ and let $\hat{\Sigma}_{ij}$ be the empirical average of N independent copies of $z_i z_j^\top$. Pick any $t > 0$. With probability at least $1 - 4\bar{d}_{ij}t(e^t - t - 1)^{-1}$,

$$\|\hat{\Sigma}_{ij} - \Sigma_{ij}\| \leq \sqrt{\frac{2 \max\{\|\mathbb{E}[\|z_j\|^2 z_i z_i^\top]\|, \|\mathbb{E}[\|z_i\|^2 z_j z_j^\top]\|\} t}{N}} + \frac{M_i M_j t}{3N}.$$

Remark 1. For any $\delta \in (0, 1/6)$, we have $4\bar{d}_{ij}t(e^t - t - 1)^{-1} \leq \delta$ provided that $t \geq 1.55 \ln(4\bar{d}_{ij}/\delta)$.

Proof. Define the random matrix

$$Z := \begin{bmatrix} & z_i z_j^\top \\ z_j z_i^\top & \end{bmatrix}.$$

Let Z_1, \dots, Z_N be independent copies of Z . Then

$$\Pr \left[\|\hat{\Sigma}_{ij} - \Sigma_{ij}\| > t \right] = \Pr \left[\left\| \frac{1}{N} \sum_{\ell=1}^N Z_\ell - \mathbb{E}[Z] \right\| > t \right].$$

Note that

$$\mathbb{E}[Z^2] = \mathbb{E} \begin{bmatrix} \|z_j\|^2 z_i z_i^\top & \\ & \|z_i\|^2 z_j z_j^\top \end{bmatrix}$$

so by convexity,

$$\begin{aligned} \|\mathbb{E}[Z^2] - \mathbb{E}[Z]^2\| &\leq \|\mathbb{E}[Z^2]\| \\ &\leq \max\{\|\mathbb{E}[\|z_j\|^2 z_i z_i^\top]\|, \|\mathbb{E}[\|z_i\|^2 z_j z_j^\top]\|\} \end{aligned}$$

and

$$\begin{aligned} \text{tr}(\mathbb{E}[Z^2] - \mathbb{E}[Z]^2) &= \text{tr}(\mathbb{E}[\|z_j\|^2 z_i z_i^\top]) + \text{tr}(\mathbb{E}[\|z_i\|^2 z_j z_j^\top]) - \text{tr}(\Sigma_{ij} \Sigma_{ij}^\top) - \text{tr}(\Sigma_{ij}^\top \Sigma_{ij}) \\ &= 2 \left(\mathbb{E}[\|z_i\|^2 \|z_j\|^2] - \text{tr}(\Sigma_{ij} \Sigma_{ij}^\top) \right). \end{aligned}$$

Moreover,

$$\|Z\| \leq \|z_i\| \|z_j\| \leq M_i M_j.$$

By the matrix Bernstein inequality [13], for any $t > 0$,

$$\begin{aligned} \Pr \left[\left\| \hat{\Sigma}_{ij} - \Sigma_{ij} \right\| > \sqrt{\frac{2 \left(\max \{ \left\| \mathbb{E}[\|z_j\|^2 z_i z_i^\top] \right\|, \left\| \mathbb{E}[\|z_i\|^2 z_j z_j^\top] \right\| \right) t}{N}} + \frac{M_i M_j t}{3N} \right] \\ \leq 2 \cdot \frac{2 \left(\mathbb{E}[\|z_i\|^2 \|z_j\|^2] - \text{tr}(\Sigma_{ij} \Sigma_{ij}^\top) \right)}{\max \{ \left\| \mathbb{E}[\|z_j\|^2 z_i z_i^\top] \right\|, \left\| \mathbb{E}[\|z_i\|^2 z_j z_j^\top] \right\|}} \cdot t(e^t - t - 1)^{-1} = 4\bar{d}_{ij} t(e^t - t - 1)^{-1}. \end{aligned}$$

The claim follows. \square

In the case of discrete random variables (modeled as random vectors as described in Section 2), the following lemma from [12] can give a tighter exponential tail inequality.

Lemma 6 ([12]). *Let z_i and z_j be random vectors, each with support on the vertices of a probability simplex. Let $\Sigma_{ij} := \mathbb{E}[z_i z_j^\top]$ and let $\hat{\Sigma}_{ij}$ be the empirical average of N independent copies of $z_i z_j^\top$. Pick any $t > 0$. With probability at least $1 - e^{-t}$,*

$$\left\| \hat{\Sigma}_{ij} - \Sigma_{ij} \right\| \leq \left\| \hat{\Sigma}_{ij} - \Sigma_{ij} \right\|_F \leq \frac{1 + \sqrt{t}}{\sqrt{N}}$$

(where $\|A\|_F$ denotes the Frobenius norm of a matrix A).

For simplicity, we only work with Lemma 5, although it is easy to translate all of our results by changing the tail inequality. The proof of Lemma 3 is immediate from combining Lemma 5 and Weyl's Theorem.

Lemma 3 provides some guidelines on how to set the $\Delta_{i,j}$ as functions of N , δ , and properties of z_i and z_j . The dependence on the properties of z_i and z_j comes through the quantities M_i , M_j , $\bar{d}_{i,j}$, and

$$B_{i,j} := \max_{i,j} \{ \left\| \mathbb{E}[\|z_j\|^2 z_i z_i^\top] \right\|, \left\| \mathbb{E}[\|z_i\|^2 z_j z_j^\top] \right\| \}.$$

In practice, one may use plug-in estimates for these quantities, or use loose upper bounds based on weaker knowledge of the distribution. For instance, $\bar{d}_{i,j}$ is at most $\max\{\dim(z_i), \dim(z_j)\}$, the larger of the explicit vector dimensions of z_i and z_j . Also, if the maximum directional standard deviation σ_* of any z_i is known, then $B_{i,j} \leq \max\{M_i^2, M_j^2\} \sigma_*^2$. We note that as these are additive confidence intervals, some dependence on the properties of z_i and z_j is inevitable.

B Analysis of the spectral quartet test

B.1 Proofs from Section 4

Proof of Lemma 1. For $i \in \{1, 2\}$ and $j \in \{3, 4\}$,

$$\Sigma_{ij} = \mathbb{E}[z_i z_j^\top] = \mathbb{E}[\mathbb{E}[z_i z_j^\top | h, g]] = C_{(z_i|h)} \mathbb{E}[h g^\top] C_{(z_j|g)}^\top = C_{(z_i|h)} \Sigma_{hg} C_{(z_j|g)}^\top.$$

Let $U_{ij} \in \mathbb{R}^{d \times k}$ and $V_{ij} \in \mathbb{R}^{d \times k}$ be matrices of singular vectors corresponding to the largest singular values of Σ_{ij} . If Σ_{hg} has rank k , then by Condition 1, so does Σ_{ij} . Therefore $\text{range}(U_{ij}) = \text{range}(C_{(z_i|h)})$ and $\text{range}(V_{ij}) = \text{range}(C_{(z_j|g)})$. This implies

$$\begin{aligned} \prod_{s=1}^k \sigma_s(\Sigma_{ij}) &= |\det(U_{ij}^\top \Sigma_{ij} V_{ij})| = |\det(U_{ij}^\top C_{(z_i|h)})| \cdot |\det(\Sigma_{hg})| \cdot |\det(V_{ij}^\top C_{(z_j|g)})| \\ &= \prod_{s=1}^k \sigma_s(C_{(z_i|h)}) \cdot |\det(\Sigma_{hg})| \cdot \prod_{s=1}^k \sigma_s(C_{(z_j|g)}). \end{aligned}$$

If Σ_{hg} has rank less than k , then $\det(\Sigma_{hg}) = \prod_{s=1}^k \sigma_s(\Sigma_{ij}) = 0$. By similar arguments,

$$\Sigma_{12} = C_{(z_1|h)} \Sigma_{hh} C_{(z_2|h)}^\top \quad \text{and} \quad \Sigma_{34} = C_{(z_3|g)} \Sigma_{gg} C_{(z_4|g)}^\top$$

so

$$\begin{aligned} \prod_{s=1}^k \sigma_s(\Sigma_{12}) &= \prod_{s=1}^k \sigma_s(C_{(z_1|h)}) \cdot \det(\Sigma_{hh}) \cdot \prod_{s=1}^k \sigma_s(C_{(z_2|h)}) \\ \text{and} \quad \prod_{s=1}^k \sigma_s(\Sigma_{34}) &= \prod_{s=1}^k \sigma_s(C_{(z_3|g)}) \cdot \det(\Sigma_{gg}) \cdot \prod_{s=1}^k \sigma_s(C_{(z_4|g)}). \end{aligned}$$

Finally, note that $u^\top \Sigma_{hh}^{-1/2} \Sigma_{hg} \Sigma_{gg}^{-1/2} v \leq \|u\| \|v\|$ for all vectors u and v by Cauchy-Schwarz, so

$$\frac{\det(\Sigma_{hg})^2}{\det(\Sigma_{hh}) \det(\Sigma_{gg})} = \det(\Sigma_{hh}^{-1/2} \Sigma_{hg} \Sigma_{gg}^{-1/2})^2 \leq 1.$$

The claim follows. \square

Note that if Condition 2 also holds, then Lemma 1 implies the strict inequalities

$$\max \left\{ \prod_{s=1}^k \sigma_s(\Sigma_{13}) \sigma_s(\Sigma_{24}), \prod_{s=1}^k \sigma_s(\Sigma_{14}) \sigma_s(\Sigma_{23}) \right\} < \prod_{s=1}^k \sigma_s(\Sigma_{12}) \sigma_s(\Sigma_{34}).$$

Proof of Lemma 2. Furthermore, given that (2) holds for all pairs $\{i, j\}$ and all $s \in \{1, 2, \dots, k\}$, if the spectral quartet test returns a pairing $\{\{i, j\}, \{i', j'\}\}$, it must be that

$$\begin{aligned} \prod_{s=1}^k \sigma_s(\Sigma_{ij}) \sigma_s(\Sigma_{i'j'}) &\geq \prod_{s=1}^k [\sigma_s(\hat{\Sigma}_{ij}) - \Delta_{ij}^{(N, \delta)}]_+ [\sigma_s(\hat{\Sigma}_{i'j'}) - \Delta_{i'j'}^{(N, \delta)}]_+ \\ &> \prod_{s=1}^k (\sigma_s(\hat{\Sigma}_{i'j'}) + \Delta_{i'j'}^{(N, \delta)}) (\sigma_s(\hat{\Sigma}_{ij}) + \Delta_{ij}^{(N, \delta)}) \geq \prod_{s=1}^k \sigma_s(\Sigma_{i'j}) \sigma_s(\Sigma_{ij'}) \end{aligned}$$

so

$$\prod_{s=1}^k \sigma_s(\Sigma_{ij}) \sigma_s(\Sigma_{i'j'}) > \prod_{s=1}^k \sigma_s(\Sigma_{i'j}) \sigma_s(\Sigma_{ij'}).$$

But by Lemma 1, the above inequality can only hold if $\{\{i, j\}, \{i', j'\}\} = \{\{1, 2\}, \{3, 4\}\}$. \square

Proof of Lemma 4. The assumptions in the statement of the lemma imply

$$\max\{\Delta_{12}, \Delta_{34}\} < \frac{\epsilon_0}{8k} \min\{\sigma_k(\Sigma_{12}), \sigma_k(\Sigma_{34})\}$$

where $\epsilon_0 := \min \left\{ \frac{1}{\rho} - 1, 1 \right\}$. Therefore

$$\begin{aligned} \prod_{s=1}^k [\sigma_s(\hat{\Sigma}_{12}) - \Delta_{12}]_+ [\sigma_s(\hat{\Sigma}_{34}) - \Delta_{34}]_+ &\geq \prod_{s=1}^k [\sigma_s(\Sigma_{12}) - 2\Delta_{12}]_+ [\sigma_s(\Sigma_{34}) - 2\Delta_{34}]_+ \\ &> \left(\prod_{s=1}^k \sigma_s(\Sigma_{12}) \sigma_s(\Sigma_{34}) \right) \left(1 - \frac{\epsilon_0}{4k} \right)^{2k} \\ &\geq \left(\prod_{s=1}^k \sigma_s(\Sigma_{12}) \sigma_s(\Sigma_{34}) \right) (1 - \epsilon_0/2). \end{aligned} \tag{4}$$

If Σ_{hg} has rank k , then so do Σ_{ij} for $i \in \{1, 2\}$ and $j \in \{3, 4\}$. Therefore, for $\{i', j'\} = \{1, 2, 3, 4\} \setminus \{i, j\}$,

$$\max\{\Delta_{ij}, \Delta_{i'j'}\} < \frac{\epsilon_0}{8k} \min\{\sigma_k(\Sigma_{i'j'}), \sigma_k(\Sigma_{i'j'})\}.$$

This implies

$$\begin{aligned} \prod_{s=1}^k (\sigma_s(\hat{\Sigma}_{ij}) + \Delta_{ij})(\sigma_s(\hat{\Sigma}_{i'j'}) + \Delta_{i'j'}) &\leq \prod_{s=1}^k (\sigma_s(\Sigma_{ij}) + 2\Delta_{ij})(\sigma_s(\Sigma_{i'j'}) + 2\Delta_{i'j'}) \\ &< \left(\prod_{s=1}^k \sigma_s(\Sigma_{ij}) \sigma_s(\Sigma_{i'j'}) \right) \left(1 + \frac{\epsilon_0}{4k} \right)^{2k} \\ &\leq \left(\prod_{s=1}^k \sigma_s(\Sigma_{ij}) \sigma_s(\Sigma_{i'j'}) \right) (1 + \epsilon_0). \end{aligned} \quad (5)$$

Therefore, combining (4), (5), and Lemma 1,

$$\begin{aligned} &\prod_{s=1}^k [\sigma_s(\hat{\Sigma}_{12}) - \Delta_{12}]_+ [\sigma_s(\hat{\Sigma}_{34}) - \Delta_{34}]_+ \\ &> \frac{1 - \epsilon_0/2}{1 + \epsilon_0} \cdot \frac{\det(\Sigma_{hh}) \det(\Sigma_{gg})}{\det(\Sigma_{hg})^2} \cdot \prod_{s=1}^k (\sigma_s(\hat{\Sigma}_{ij}) + \Delta_{ij})(\sigma_s(\hat{\Sigma}_{i'j'}) + \Delta_{i'j'}) \\ &\geq \frac{1}{(1 + \epsilon_0)^2} \cdot \frac{\det(\Sigma_{hh}) \det(\Sigma_{gg})}{\det(\Sigma_{hg})^2} \cdot \prod_{s=1}^k (\sigma_s(\hat{\Sigma}_{ij}) + \Delta_{ij})(\sigma_s(\hat{\Sigma}_{i'j'}) + \Delta_{i'j'}) \\ &\geq \prod_{s=1}^k (\sigma_s(\hat{\Sigma}_{ij}) + \Delta_{ij})(\sigma_s(\hat{\Sigma}_{i'j'}) + \Delta_{i'j'}), \end{aligned}$$

so the spectral quartet test will return the correct pairing $\{\{1, 2\}, \{3, 4\}\}$, proving the lemma. \square

B.2 Conditions for returning a correct pairing when $\text{rank}(\Sigma_{hg}) < k$

We consider the case in which Σ_{hg} has rank $r < k$. In this case, the widths of the confidence intervals are allowed to be wider than in the case where $\text{rank}(\Sigma_{hg}) = k$. Define

$$\sigma_{\min} := \min\left(\{\sigma_k(\Sigma_{12}), \sigma_k(\Sigma_{34})\} \cup \{\sigma_r(\Sigma_{ij}) : i \in \{1, 2\}, j \in \{3, 4\}\}\right).$$

$$\rho_1^2 = \frac{\sigma_{\min}^{2(k-r)} \cdot \max_{i,j,i',j'} \prod_{s=1}^r \sigma_s(\Sigma_{ij}) \sigma_s(\Sigma_{i'j'})}{\prod_{s=1}^k \sigma_s(\Sigma_{12}) \sigma_s(\Sigma_{34})}.$$

Instead of depending on $\min_{i,j} \{\sigma_k(\Sigma_{ij})\}$ and ρ as in the case where $\text{rank}(\Sigma_{hg}) = k$, we only depend on σ_{\min} and ρ_1 .

Lemma 7 (Correct pairing, $\text{rank}(\Sigma_{hg}) < k$). *Suppose that (i) the observed variables $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$ have the true induced tree topology shown in Figure 1(a), (ii) the tree model satisfies Condition 1, (iii) Σ_{hg} has rank $r < k$, and (iv) the confidence bounds in (2) hold for all $\{i, j\}$ and all $s \in [k]$. If*

$$\Delta_{i,j} < \frac{1}{8k} \cdot \min \left\{ 1, 8k \left(\frac{1}{2\rho_1} \right)^{\frac{1}{k-r}} \right\} \cdot \sigma_{\min}$$

for each $\{i, j\}$, then Algorithm 1 returns the correct pairing $\{\{1, 2\}, \{3, 4\}\}$.

Note that the allowed width increases (to a point) as the rank r decreases.

Proof of Lemma 7. The assumptions in the statement of the lemma imply

$$\max\{\Delta_{i,j} : \{i,j\} \subset [4]\} < \frac{\epsilon_1 \sigma_{\min}}{8k}$$

where

$$\epsilon_1 := \min \left\{ 8k \cdot \left(\frac{1}{2\rho_1} \right)^{\frac{1}{k-r}}, 1 \right\}.$$

We have

$$\prod_{s=1}^k [\sigma_s(\hat{\Sigma}_{12}) - \Delta_{12}]_+ [\sigma_s(\hat{\Sigma}_{34}) - \Delta_{34}]_+ > \left(\prod_{s=1}^k \sigma_s(\Sigma_{12}) \sigma_s(\Sigma_{34}) \right) (1 - \epsilon_1/2)$$

as in the proof of Lemma 4. Moreover,

$$\begin{aligned} & \prod_{s=1}^k (\sigma_s(\hat{\Sigma}_{ij}) + \Delta_{ij}) (\sigma_s(\hat{\Sigma}_{i'j'}) + \Delta_{i'j'}) \\ & < \left(\prod_{s=1}^r \sigma_s(\Sigma_{ij}) \sigma_s(\Sigma_{i'j'}) \right) \cdot (1 + \epsilon_1) \cdot \left(\frac{\epsilon_1 \sigma_{\min}}{8k} \right)^{2(k-r)} \\ & \leq \left(\prod_{s=1}^k \sigma_s(\Sigma_{12}) \sigma_s(\Sigma_{34}) \right) \cdot \frac{\rho_1^2}{(\sigma_{\min})^{2(k-r)}} \cdot (1 + \epsilon_1) \cdot \left(\frac{\epsilon_1 \sigma_{\min}}{8k} \right)^{2(k-r)} \\ & = \left(\prod_{s=1}^k \sigma_s(\Sigma_{12}) \sigma_s(\Sigma_{34}) \right) \cdot \rho_1^2 \cdot (1 + \epsilon_1) \cdot \left(\frac{\epsilon_1}{8k} \right)^{2(k-r)} \\ & < \left(\prod_{s=1}^k [\sigma_s(\hat{\Sigma}_{12}) - \Delta_{12}]_+ [\sigma_s(\hat{\Sigma}_{34}) - \Delta_{34}]_+ \right) \cdot \rho_1^2 \cdot \frac{1 + \epsilon_1}{1 - \epsilon_1/2} \cdot \left(\frac{\epsilon_1}{8k} \right)^{2(k-r)} \\ & \leq \left(\prod_{s=1}^k [\sigma_s(\hat{\Sigma}_{12}) - \Delta_{12}]_+ [\sigma_s(\hat{\Sigma}_{34}) - \Delta_{34}]_+ \right) \cdot \rho_1^2 \cdot (1 + \epsilon_1)^2 \cdot \left(\frac{\epsilon_1}{8k} \right)^{2(k-r)} \\ & \leq \prod_{s=1}^k [\sigma_s(\hat{\Sigma}_{12}) - \Delta_{12}]_+ [\sigma_s(\hat{\Sigma}_{34}) - \Delta_{34}]_+. \end{aligned}$$

Therefore the spectral quartet test will return the correct pairing $\{\{1, 2\}, \{3, 4\}\}$; the lemma follows. \square

C Proof of Theorem 1

Let us define

$$\epsilon_{\min} := \min \left\{ \frac{1}{\rho_{\max}} - 1, 1 \right\}, \quad \varepsilon := \frac{\gamma_{\min}/\gamma_{\max}}{8k + \gamma_{\min}/\gamma_{\max}}, \quad \varsigma := \frac{\gamma_{\min}}{\gamma_{\max}} \cdot \frac{1 - \varepsilon}{1 + \varepsilon} \cdot \gamma_{\min}.$$

The sample size requirement ensures that

$$\Delta_{x_i, x_j} < \frac{\epsilon_{\min} \cdot \varsigma}{8k} \leq \varepsilon \theta.$$

The probabilistic event we need is that in which the confidence bounds in (6) hold for each pair of observed variables. The event

$$\forall \{x_i, x_j\} \subseteq \mathcal{V}_{\text{obs}} \bullet \|\hat{\Sigma}_{x_i, x_j} - \Sigma_{x_i, x_j}\| \leq \Delta_{x_i, x_j}, \quad (6)$$

occurs with probability at least $1 - \eta$ by Lemma 5 and a union bound. We henceforth condition on the above event.

Our primary goal is to show the following claim.

Claim 1. *If $\widehat{\mathcal{V}}_\ell = \bar{\mathcal{V}}_\ell$ and $\widehat{\mathcal{L}}(u) = \bar{\mathcal{L}}(u)$ for all $u \in \bar{\mathcal{V}}_\ell$, then $\bar{\mathcal{S}}_\ell \subseteq \widetilde{\mathcal{S}}_\ell$ and $\widehat{\mathcal{S}}_\ell = \bar{\mathcal{S}}_\ell$.*

Note that if both the precondition of Claim 1 and its second conclusion hold, *i.e.*, if

1. $\widehat{\mathcal{V}}_\ell = \bar{\mathcal{V}}_\ell$ (the algorithm has the correct set of working variables in level ℓ);
2. $\widehat{\mathcal{L}}(u) = \bar{\mathcal{L}}(u) \forall u \in \bar{\mathcal{V}}_\ell$ (the algorithm has the correct groupings of observed variables in level ℓ); and
3. $\widehat{\mathcal{S}}_\ell = \bar{\mathcal{S}}_\ell$ (the algorithm correctly determines the sibling pairs in level ℓ);

then we have $\widehat{\mathcal{V}}_{\ell+1} = \bar{\mathcal{V}}_{\ell+1}$ and $\widehat{\mathcal{L}}(u) = \bar{\mathcal{L}}(u)$ for all $u \in \bar{\mathcal{V}}_{\ell+1}$ by the construction in the algorithm. Since $\widehat{\mathcal{L}}(x) = \bar{\mathcal{L}}(x)$ for all $x \in \bar{\mathcal{V}}_1 = \bar{\mathcal{V}}_1$ trivially, the theorem follows from Claim 1 and induction.

Henceforth assume $\widehat{\mathcal{V}}_\ell = \bar{\mathcal{V}}_\ell$ and $\widehat{\mathcal{L}}(u) = \bar{\mathcal{L}}(u)$ for all $u \in \bar{\mathcal{V}}_\ell$. We now show the first conclusion in Claim 1.

Claim 2. $\bar{\mathcal{S}}_\ell \subseteq \widetilde{\mathcal{S}}_\ell$.

Proof. Take any pair $\{u, v\} \subseteq \widehat{\mathcal{V}}_\ell = \bar{\mathcal{V}}_\ell$, and let h be their parent (so $u, v \in \bar{\mathcal{L}}(h)$). Suppose $u, v \in \widehat{\mathcal{V}}_\ell$ are siblings, *i.e.*, $\{u, v\} \in \bar{\mathcal{S}}_\ell$. Then there are subtrees \mathcal{T}_u and \mathcal{T}_v in \mathcal{F}_h with $\widehat{\mathcal{L}}(u) = \mathcal{V}_{\text{obs}}[\mathcal{T}_u]$ and $\widehat{\mathcal{L}}(v) = \mathcal{V}_{\text{obs}}[\mathcal{T}_v]$. Therefore, by Condition 3, there exists observed variables $x \in \bar{\mathcal{L}}(u)$ and $y \in \bar{\mathcal{L}}(v)$ such that $\sigma_k(\Sigma_{x,y}) \geq \gamma_{\min} \geq (1 + \varepsilon)\theta$. By Weyl's Theorem, the confidence bounds in (6), and the fact that $\Delta_{x,y} \leq \varepsilon\theta$,

$$\sigma_k(\widehat{\Sigma}_{x,y}) \geq \sigma_k(\Sigma_{x,y}) - \|\widehat{\Sigma}_{x,y} - \Sigma_{x,y}\| \geq (1 + \varepsilon)\theta - \Delta_{x,y} \geq \theta.$$

This implies that $\{u, v\} \in \widetilde{\mathcal{S}}_\ell$. Therefore, $\bar{\mathcal{S}}_\ell \subseteq \widetilde{\mathcal{S}}_\ell$. □

It remains to show that $\widehat{\mathcal{S}}_\ell = \bar{\mathcal{S}}_\ell$. First, by Lemma 2, no sibling pair $\{u, v\} \in \bar{\mathcal{S}}_\ell$ will be broken by a spectral quartet test. Now take any pair $\{u, v\} \subseteq \widehat{\mathcal{V}}_\ell = \bar{\mathcal{V}}_\ell$ and suppose instead that $u, v \in \widehat{\mathcal{V}}_\ell$ are not siblings but $\{u, v\} \in \widetilde{\mathcal{S}}_\ell$. That is, $\{u, v\} \in \widetilde{\mathcal{S}}_\ell \setminus \bar{\mathcal{S}}_\ell$. We need to show that the algorithm performs a spectral quartet test whose result will cause $\{u, v\}$ to be broken.

Since $\{u, v\} \in \widetilde{\mathcal{S}}_\ell$, we have that there exists $x \in \bar{\mathcal{L}}(u)$ and $y \in \bar{\mathcal{L}}(v)$ such that $\sigma_k(\widehat{\Sigma}_{x,y}) \geq \theta$. By the same arguments from the proof of Claim 2,

$$\sigma_k(\Sigma_{x,y}) \geq \sigma_k(\widehat{\Sigma}_{x,y}) - \|\widehat{\Sigma}_{x,y} - \Sigma_{x,y}\| \geq \theta - \Delta_{x,y} \geq (1 - \varepsilon)\theta.$$

Let h be the parent of u (so $u \in \bar{\mathcal{L}}(h)$), and g be the parent of v . Since h has at least three neighbors, there exist subtrees $\mathcal{T}_{h,1}$ and $\mathcal{T}_{h,2}$ in \mathcal{F}_h not containing g . Furthermore, by the Condition 3, there exist $x_1 \in \mathcal{V}_{\text{obs}}[\mathcal{T}_{h,1}]$ and $x_2 \in \mathcal{V}_{\text{obs}}[\mathcal{T}_{h,2}]$ such that $\sigma_k(\mathbb{E}[x_1 x_2^\top]) \geq \gamma_{\min}$. Without loss of generality, assume $x \in \mathcal{V}_{\text{obs}}[\mathcal{T}_{h,2}]$. Similarly, there are subtrees $\mathcal{T}_{g,1}$ and $\mathcal{T}_{g,2}$ in \mathcal{F}_g not containing h , and $y_1 \in \mathcal{V}_{\text{obs}}[\mathcal{T}_{g,1}]$ and $y_2 \in \mathcal{V}_{\text{obs}}[\mathcal{T}_{g,2}]$ such that $\sigma_k(\mathbb{E}[y_1 y_2^\top]) \geq \gamma_{\min}$. Again, without loss of generality, assume $y \in \mathcal{V}_{\text{obs}}[\mathcal{T}_{g,2}]$.

We now claim that the result of testing the quartet $\{u, v, u_1, v_1\} \subseteq \widehat{\mathcal{V}}_\ell$, where $x_1 \in \widehat{\mathcal{L}}(u_1)$ and $y_1 \in \widehat{\mathcal{L}}(v_1)$, will break the pair $\{u, v\}$. To prove this, we argue that the sample size N is large enough for the spectral quartet test to return the correct pairing $\{\{x, x_1\}, \{y, y_1\}\}$.

Claim 3. *The following lower bounds hold:*

$$\min \{ \sigma_k(\Sigma_{x_1, x}), \sigma_k(\Sigma_{x_1, y}), \sigma_k(\Sigma_{y_1, y}), \sigma_k(\Sigma_{y_1, x}) \} \geq \frac{\gamma_{\min} \cdot (1 - \varepsilon)\theta}{\gamma_{\max}}.$$

Proof. We just show the inequalities for $\sigma_k(\Sigma_{x_1,x})$ and $\sigma_k(\Sigma_{x_1,y})$; the other two are analogous. Since $\sigma_k(\mathbb{E}[xy^\top]) = \sigma_k(\Sigma_{x,y}) > 0$, we have that $\mathbb{E}[xy^\top]$ has rank k . Let the columns of U_x be an orthonormal basis of $\text{range}(C_{(x|h)})$, the columns of U_y be an orthonormal basis of $\text{range}(C_{(y|h)})$, the columns of U_1 be an orthonormal basis of $\text{range}(C_{(x_1|h)})$, and the columns of U_2 be an orthonormal basis of $\text{range}(C_{(x_2|h)})$. We have

$$\begin{aligned}
U_1^\top \mathbb{E}[x_1 x^\top] U_x &= U_1^\top C_{(x_1|h)} \Sigma_{h,h} C_{(x|h)}^\top U_x \\
&= (U_1^\top C_{(x_1|h)} \Sigma_{h,h}^{1/2}) (U_y^\top C_{(y|h)} \Sigma_{h,h}^{1/2})^{-1} (U_y^\top C_{(y|h)} \Sigma_{h,h}^{1/2}) (\Sigma_{h,h}^{1/2} C_{(x|h)}^\top U_x) \\
&= (U_1^\top C_{(x_1|h)} \Sigma_{h,h}^{1/2}) (U_y^\top C_{(y|h)} \Sigma_{h,h}^{1/2})^{-1} (U_y^\top \mathbb{E}[yx^\top] U_x) \\
&= (U_1^\top C_{(x_1|h)} \Sigma_{h,h}^{1/2}) (\Sigma_{h,h}^{1/2} C_{(x_2|h)}^\top U_2) (\Sigma_{h,h}^{1/2} C_{(x_2|h)}^\top U_2)^{-1} \\
&\quad \cdot (U_y^\top C_{(y|h)} \Sigma_{h,h}^{1/2})^{-1} (U_y^\top \mathbb{E}[yx^\top] U_x) \\
&= (U_1^\top \mathbb{E}[x_1 x_2^\top] U_2) (U_y^\top \mathbb{E}[y x_2^\top] U_2)^{-1} (U_y^\top \mathbb{E}[y x^\top] U_x).
\end{aligned}$$

Therefore

$$\sigma_k(\mathbb{E}[x_1 x^\top]) \geq \frac{\sigma_k(\mathbb{E}[x_1 x_2^\top]) \cdot \sigma_k(\mathbb{E}[y x^\top])}{\sigma_1(\mathbb{E}[y x_2^\top])} \geq \frac{\gamma_{\min} \cdot (1 - \varepsilon) \theta}{\gamma_{\max}}.$$

This gives the first claimed inequality. For the second,

$$\begin{aligned}
U_1^\top \mathbb{E}[x_1 y^\top] U_y &= U_1^\top C_{(x_1|h)} \Sigma_{h,h} C_{(y|h)}^\top U_y \\
&= (U_1^\top C_{(x_1|h)} \Sigma_{h,h}^{1/2}) (U_x^\top C_{(x|h)} \Sigma_{h,h}^{1/2})^{-1} (U_x^\top C_{(x|h)} \Sigma_{h,h}^{1/2}) (\Sigma_{h,h}^{1/2} C_{(y|h)}^\top U_y) \\
&= (U_1^\top C_{(x_1|h)} \Sigma_{h,h}^{1/2}) (U_x^\top C_{(x|h)} \Sigma_{h,h}^{1/2})^{-1} (U_x^\top \mathbb{E}[xy^\top] U_y) \\
&= (U_1^\top C_{(x_1|h)} \Sigma_{h,h}^{1/2}) (\Sigma_{h,h}^{1/2} C_{(x_2|h)}^\top U_2) (\Sigma_{h,h}^{1/2} C_{(x_2|h)}^\top U_2)^{-1} \\
&\quad \cdot (U_x^\top C_{(x|h)} \Sigma_{h,h}^{1/2})^{-1} (U_x^\top \mathbb{E}[xy^\top] U_y) \\
&= (U_1^\top \mathbb{E}[x_1 x_2^\top] U_2) (U_x^\top \mathbb{E}[x x_2^\top] U_2)^{-1} (U_x^\top \mathbb{E}[xy^\top] U_y).
\end{aligned}$$

Therefore

$$\sigma_k(\mathbb{E}[x_1 y^\top]) \geq \frac{\sigma_k(\mathbb{E}[x_1 x_2^\top]) \cdot \sigma_k(\mathbb{E}[xy^\top])}{\sigma_1(\mathbb{E}[x x_2^\top])} \geq \frac{\gamma_{\min} \cdot (1 - \varepsilon) \theta}{\gamma_{\max}}. \quad \square$$

Therefore, we have $\min\{\sigma_k(\Sigma_{x_1,x}), \sigma_k(\Sigma_{x_1,y}), \sigma_k(\Sigma_{y_1,x}), \sigma_k(\Sigma_{y_1,y})\} \geq \varsigma$. By Lemma 4, the thresholds are small enough so that the spectral quartet test on $\{x, x_1, y, y_1\}$ returns the correct pairing $\{\{x, x_1\}, \{y, y_1\}\}$, which thus breaks the non-sibling pair $\{u, v\}$. This proves that $\hat{S}_\ell = \bar{S}_\ell$, which therefore completes the proof of Claim 1 and also the proof of Theorem 1. \square