Spectral Methods for Learning Multivariate Latent Tree Structure

Animashree Anandkumar
UC Irvine
a.anandkumar@uci.edu

Kamalika Chaudhuri
UC San Diego
kamalika@cs.ucsd.edu

Daniel Hsu
Microsoft Research
dahsu@microsoft.com

Sham M. Kakade
Microsoft Research &
University of Pennsylvania
skakade@microsoft.com

Le Song
Carnegie Mellon University
lesong@cs.cmu.edu

Tong Zhang
Rutgers University
tzhang@stat.rutgers.edu

Abstract

This work considers the problem of learning the structure of multivariate linear tree models, which include a variety of directed tree graphical models with continuous, discrete, and mixed latent variables such as linear-Gaussian models, hidden Markov models, Gaussian mixture 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 each other and to the observed variables). We propose 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 reveal 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 from second-order statistics.

1 Introduction

Graphical models are a central tool in modern machine learning applications, as they provide a natural methodology for succinctly representing high-dimensional distributions. As such, they have enjoyed much success in various AI and machine learning applications such as natural language processing, speech recognition, robotics, computer vision, and bioinformatics.

The main statistical challenges associated with graphical models include estimation and inference. While the body of techniques for probabilistic inference in graphical models is rather rich [1], current methods for tackling the more challenging problems of parameter and structure estimation are less developed and understood, especially in the presence of latent (hidden) variables. The problem of parameter estimation involves determining the model parameters from samples of certain observed variables. Here, the predominant approach is the expectation maximization (EM) algorithm, and only rather recently is the understanding of this algorithm improving [2, 3]. The problem of structure learning is to estimate the underlying graph of the graphical model. In general, structure learning is NP-hard and becomes even more challenging when some variables are unobserved [4]. The main approaches for structure estimation are either greedy or local search approaches [5, 6] or, more recently, based on convex relaxation [7].

1

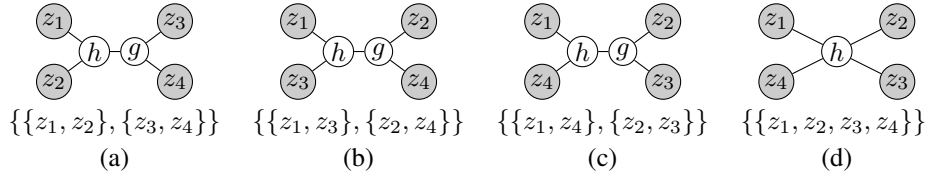



Figure 1: The four possible (undirected) tree topologies over leaves $\{z_1, z_2, z_3, z_4\}$.

This work focuses on learning the structure of multivariate latent tree graphical models. Here, the underlying graph is a directed 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, ranging from computer vision, where one may learn object/scene structure from the co-occurrences of objects to aid image understanding [8]; to phylogenetics, where the central task is to reconstruct the tree of life from the genetic material of surviving species [9].

Generally speaking, methods for learning latent tree structure exploit structural properties afforded by the tree that are revealed through certain statistical tests over every choice of four variables in the tree. These *quartet tests*, which have origins in structural equation modeling [10, 11], are hypothesis tests of the relative configuration of four (possibly non-adjacent) nodes/variables in the tree (see Figure 1); they are also related to the *four point condition* associated with a corresponding additive tree metric induced by the distribution [12]. Some early methods for learning tree structure are based on the use of *exact* correlation statistics or distance measurements (*e.g.*, [13, 14]). Unfortunately, these methods ignore the crucial aspect of estimation error, which ultimately governs their sample complexity. Indeed, this (lack of) robustness to estimation error has been quantified for various algorithms (notably, for the popular Neighbor Joining algorithm [15, 16]), and therefore serves as a basis for comparing different methods. Subsequent work in the area of mathematical phylogenetics has focused on the sample complexity of evolutionary tree reconstruction [17, 15, 18, 19]. The basic model there corresponds to a directed tree over discrete random variables, and much of the recent effort deals exclusively in the regime for a certain model parameter (the Kesten-Stigum regime [20]) that allows for a sample complexity that is polylogarithmic in the number of leaves, as opposed to polynomial [18, 19]. Finally, recent work in machine learning has developed structure learning methods for latent tree graphical models that extend beyond the discrete distributions of evolutionary trees [21], thereby widening their applicability to other problem domains.

This work extends beyond previous studies, which have focused on latent tree models with either discrete or scalar Gaussian variables, by directly addressing the multivariate setting where hidden and observed nodes may be random vectors rather than scalars. The generality of our techniques allows us to handle a much wider class of distributions than before, both in terms of the conditional independence properties imposed by the models (*i.e.*, the random vector associated with a node need not follow a distribution that corresponds to a tree model), as well as other characteristics of the node distributions (*e.g.*, some nodes in the tree could have discrete state spaces and others continuous, as in a Gaussian mixture model).

We propose the *Spectral Recursive Grouping* algorithm for learning multivariate latent tree structure. The algorithm has at its core a multivariate *spectral quartet test*, which extends the classical quartet tests for scalar variables by applying spectral techniques from multivariate statistics (specifically canonical correlation analysis [22, 23]). Spectral methods have enjoyed recent success in the context of parameter estimation [24, 25, 26, 27]; our work shows that they are also useful for structure learning. We use the spectral quartet test in a simple modification of the recursive grouping algorithm of [21] to perform the tree reconstruction. The algorithm is essentially a robust method for reasoning about the results of quartet tests (viewed simply as hypothesis tests); the tests either confirm or reject hypotheses about the relative topology over quartets of variables. By carefully choosing which tests to consider and properly interpreting their results, the algorithm is able to recover the correct latent tree structure (with high probability) in a provably efficient manner, in terms of both computational and sample complexity. The recursive grouping procedure is similar to the *short quartet method* from phylogenetics [15], which also guarantees efficient reconstruction in the context of evolutionary trees. However, our method and analysis applies to considerably more general high-dimensional settings; for instance, our sample complexity bound is given in terms of natural correlation con-

2

ditions that generalize the more restrictive *effective depth* conditions of previous works [15, 21]. Finally, we note that while we do not directly address the question of parameter estimation, provable parameter estimation methods may derived using the spectral techniques from [24, 25].

2 Preliminaries

2.1 Latent variable tree models

Let $\mathbb{T}$ be a connected, directed tree graphical model with leaves $\mathcal{V}_{\mathrm{obs}} := \{x_1, x_2, \ldots, x_n\}$ and internal nodes $\mathcal{V}_{\mathrm{hid}} := \{h_1, h_2, \ldots, h_m\}$ such that every node has at most one parent. The leaves are termed the *observed variables* and the internal nodes *hidden variables*. Note that all nodes in this work generally correspond to multivariate random vectors; we will abuse terminology and still refer to these random vectors as random variables. For any $h \in \mathcal{V}_{\mathrm{hid}}$, let $\mathrm{Children}_{\mathbb{T}}(h) \subseteq \mathcal{V}_{\mathbb{T}}$ denote the children of h in $\mathbb{T}$.

Each observed variable $x \in \mathcal{V}_{\mathrm{obs}}$ is modeled as random vector in $\mathbb{R}^d$, and each hidden variable $h \in \mathcal{V}_{\mathrm{hid}}$ as a random vector in $\mathbb{R}^k$. The joint distribution over all the variables $\mathcal{V}_{\mathbb{T}} := \mathcal{V}_{\mathrm{obs}} \cup \mathcal{V}_{\mathrm{hid}}$ is assumed satisfy conditional independence properties specified by the tree structure over the variables. Specifically, for any disjoint subsets $V_1, V_2, V_3 \subseteq \mathcal{V}_{\mathbb{T}}$ such that V_3 separates V_1 from V_2 in $\mathbb{T}$, the variables in V_1 are conditionally independent of those in V_2 given V_3.

2.2 Structural and distributional assumptions

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

Condition 1 (Linear conditional means). Fix any hidden variable $h \in \mathcal{V}_{\mathrm{hid}}$. For each hidden child $g \in \mathrm{Children}_{\mathbb{T}}(h) \cap \mathcal{V}_{\mathrm{hid}}$, 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 each observed child $x \in \mathrm{Children}_{\mathbb{T}}(h) \cap \mathcal{V}_{\mathrm{obs}}$, 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 the class of tree graphical models satisfying Condition 1 as *linear tree models*. Such models include a variety of continuous and discrete tree distributions (as well as hybrid combinations of the two, such as Gaussian mixture models) 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 [25] and Markovian evolutionary trees [24].

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

1. For all $h \in \mathcal{V}_{\mathrm{hid}}$, $\mathbb{E}[hh^{\top}]$ has rank k (*i.e.*, $\sigma_k(\mathbb{E}[hh^{\top}]) > 0$).

2. For all $h \in \mathcal{V}_{\mathrm{hid}}$ and hidden child $g \in \mathrm{Children}_{\mathbb{T}}(h) \cap \mathcal{V}_{\mathrm{hid}}$, $A_{(g|h)}$ has rank k.

3. For all $h \in \mathcal{V}_{\mathrm{hid}}$ and observed child $x \in \mathrm{Children}_{\mathbb{T}}(h) \cap \mathcal{V}_{\mathrm{obs}}$, $C_{(x|h)}$ has rank k.

The rank condition is a generalization of parameter identifiability conditions in latent variable models [28, 24, 25] which rules out various (provably) hard instances in discrete variable settings [24].

3

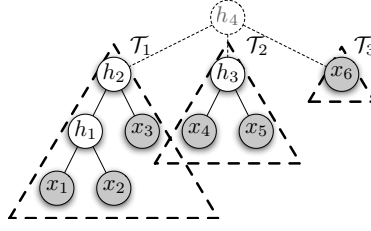



Figure 2: Set of trees $\mathcal{F}_{h_4} = \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$ obtained if h_4 is removed.

Condition 3 (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 $h, g \in \mathcal{V}_{\mathrm{hid}}$,

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

The requirement for each hidden node to have three neighbors is natural; otherwise, the hidden node can be eliminated. The quantity $\rho_{\max}$ is a natural multivariate generalization of correlation. First, note that $\rho_{\max} \leq 1$, and that 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 to a single node. Clearly, this condition is necessary for the goal of identifying the correct tree structure, and it is satisfied as soon as h and g have limited correlation in just a single direction. Previous works [13, 29] show that an analogous condition ensures identifiability for *general* latent tree models (and in fact, the conditions are identical in the Gaussian case). Condition 3 is therefore a generalization of this condition suitable for the multivariate setting.

Our learning guarantees also require a correlation condition that generalize the explicit depth conditions considered in the phylogenetics literature [15, 24]. To state this condition, first define $\mathcal{F}_h$ to be the set of subtrees of that remain after a hidden variable $h \in \mathcal{V}_{\mathrm{hid}}$ is removed from $\mathbb{T}$ (see Figure 2). Also, for any subtree $\mathcal{T}'$ of $\mathbb{T}$, let $\mathcal{V}_{\mathrm{obs}}[\mathcal{T}'] \subseteq \mathcal{V}_{\mathrm{obs}}$ be the observed variables in $\mathcal{T}'$.

Condition 4 (Correlation condition)**.** There exists $\gamma_{\min} > 0$ such that for all hidden variables $h \in \mathcal{V}_{\mathrm{hid}}$ and all triples of subtrees $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\} \subseteq \mathcal{F}_h$ in the forest obtained if h is removed from $\mathbb{T}$,

$$\max_{x_1 \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_1], x_2 \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_2], x_3 \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_3]} \min_{\{i,j\} \subset \{1,2,3\}} \sigma_k(\mathbb{E}[x_i x_j^\top]) \geq \gamma_{\min}.$$

The quantity $\gamma_{\min}$ is related to the *effective depth* of $\mathbb{T}$, which is the maximum graph distance between a hidden variable and its closest observed variable [15, 21]. 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 effective depth is 1, even though the true depth, or diameter, is $m + 1$). If the matrices giving the (conditionally) linear relationship between neighboring variables in $\mathbb{T}$ are all well-conditioned, then $\gamma_{\min}$ is at worst exponentially small in the effective depth, and therefore at worst polynomially small in the number of variables.

Finally, also define

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

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.*, $\|A\| := \sigma_1(A) = \sup\{\|Ax\| : \|x\| = 1\}$.

4

Algorithm 1 SpectralQuartetTest on observed variables $\{z_1, z_2, z_3, z_4\}$.

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

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

 1: **if** there exists a partition of $\{z_1, z_2, z_3, z_4\} = \{z_i, z_j\} \cup \{z_{i'}, z_{j'}\}$ such that

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

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

 2: **else** return $\perp$.

---

3 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 outputs $\perp$ to indicate abstinence. If the test returns $\perp$, then no guarantees are provided on 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 (SpectralQuartetTest). The notation $[a]_+$ denotes $\max\{0, a\}$ and $[t]$ (for an integer t) denotes the set $\{1, 2, \ldots, t\}$.

The quartet test is defined with respect to four observed variables $\mathcal{Z} := \{z_1, z_2, z_3, z_4\}$. For each pair of variables z_i and z_j, it takes as input an empirical estimate $\hat{\Sigma}_{i,j}$ of the second-moment matrix $\mathbb{E}[z_i z_j^\top]$, and confidence bound parameters $\Delta_{i,j}$ which are functions of N, the number of samples used to compute the $\hat{\Sigma}_{i,j}$'s, a confidence parameter δ, and of properties of the distributions 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 $\{\{z_i, z_j\}, \{z_{i'}, z_{j'}\}\}$. For example, if the output is the pairing is $\{\{z_1, z_2\}, \{z_3, z_4\}\}$, 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 $\{\{z_1, z_2, z_3, z_4\}\}$, as there is no correct pairing of $\mathcal{Z}$. The topology $\{\{z_1, z_2, z_3, z_4\}\}$ can be viewed as a degenerate case of $\{\{z_1, z_2\}, \{z_3, z_4\}\}$ (say) where the hidden variables h and g are deterministically identical, and Condition 3 fails to hold with respect to h and g.

3.1 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. Let $\det_k(M) := \prod_{s=1}^{k} \sigma_s(M)$ denote the product of the k largest singular values of a matrix M.

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 and Condition 2. Then*

$$\frac{\det_k(\mathbb{E}[z_1 z_3^\top])\det_k(\mathbb{E}[z_2 z_4^\top])}{\det_k(\mathbb{E}[z_1 z_2^\top])\det_k(\mathbb{E}[z_3 z_4^\top])} = \frac{\det_k(\mathbb{E}[z_1 z_4^\top])\det_k(\mathbb{E}[z_2 z_3^\top])}{\det_k(\mathbb{E}[z_1 z_2^\top])\det_k(\mathbb{E}[z_3 z_4^\top])} = \frac{\det(\mathbb{E}[hg^\top])^2}{\det(\mathbb{E}[hh^\top])\det(\mathbb{E}[gg^\top])} \leq 1 \tag{1}$$

$$and \quad \det_k(\mathbb{E}[z_1 z_3^\top])\det_k(\mathbb{E}[z_2 z_4^\top]) = \det_k(\mathbb{E}[z_1 z_4^\top])\det_k(\mathbb{E}[z_2 z_3^\top]).$$

This lemma shows that given the true second-moment matrices and assuming Condition 3, the inequality in (1) becomes strict and thus can be used to deduce the correct topology: the correct pairing is $\{\{z_i, z_j\}, \{z_{i'}, z_{j'}\}\}$ if and only if

$$\det_k(\mathbb{E}[z_i z_j^\top])\det_k(\mathbb{E}[z_{i'} z_{j'}^\top]) > \det_k(\mathbb{E}[z_{i'} z_j^\top])\det_k(\mathbb{E}[z_i z_{j'}^\top]).$$

5

Reliability: The next lemma shows that even if the singular values of $\mathbb{E}[z_i z_j^\top]$ 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 $\bot$, 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 pairs $\{z_i, z_j\} \subset \mathcal{Z}$ and all $s \in [k]$, $\sigma_s(\hat{\Sigma}_{i,j}) - \Delta_{i,j} \leq \sigma_s(\mathbb{E}[z_i z_j^\top]) \leq \sigma_s(\hat{\Sigma}_{i,j}) + \Delta_{i,j}$, and if* SpectralQuartetTest *returns a pairing* $\{\{z_i, z_j\}, \{z_{i'}, z_{j'}\}\}$, *then* $\{\{z_i, z_j\}, \{z_{i'}, z_{j'}\}\} = \{\{z_1, z_2\}, \{z_3, z_4\}\}$.

In other words, the spectral quartet test never returns an incorrect pairing as long as the singular values of $\mathbb{E}[z_i z_j^\top]$ lie in an interval of length $2\Delta_{i,j}$ around the singular values of $\hat{\Sigma}_{i,j}$. 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 required event holds with probability at least $1 - \delta$. We remark that any valid confidence intervals may be used; the one described below is particularly suitable when the observed variables are high-dimensional random vectors.

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 $\hat{\Sigma}_{i,j}$ is computed using N iid copies of z_i and z_j, and if*

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

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

then with probability $1 - \delta$, for all pairs $\{z_i, z_j\} \subset \mathcal{Z}$ and all $s \in [k]$,

$$\sigma_s(\hat{\Sigma}_{i,j}) - \Delta_{i,j} \leq \sigma_s(\mathbb{E}[z_i z_j^\top]) \leq \sigma_s(\hat{\Sigma}_{i,j}) + \Delta_{i,j}. \tag{2}$$

Conditions for returning a correct pairing: The conditions under which SpectralQuartetTest returns an induced topology (as opposed to $\bot$) are now provided.

An important quantity in this analysis is the level of non-redundancy between the hidden variables h and g. Let

$$\rho^2 := \frac{\det(\mathbb{E}[hg^\top])^2}{\det(\mathbb{E}[hh^\top]) \det(\mathbb{E}[gg^\top])}. \tag{3}$$

If Figure 1(a) is the correct induced topology among $\{z_1, z_2, z_3, z_4\}$, then the smaller ρ is, the greater the gap between $\det_k(\mathbb{E}[z_1 z_2^\top])\det_k(\mathbb{E}[z_3 z_4^\top])$ and either of $\det_k(\mathbb{E}[z_1 z_3^\top])\det_k(\mathbb{E}[z_2 z_4^\top])$ and $\det_k(\mathbb{E}[z_1 z_4^\top])\det_k(\mathbb{E}[z_2 z_3^\top])$. Therefore, ρ also governs how small the $\Delta_{i,j}$ need to be for the quartet test to return a correct pairing; this is quantified in Lemma 4. Note that Condition 3 implies $\rho \leq \rho_{\max} < 1$.

Lemma 4 (Correct pairing). *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, Condition 2, and $\rho < 1$ (where ρ is defined in (3)), and (iii) the confidence bounds in (2) hold for all $\{i, j\}$ and all $s \in [k]$. If*

$$\Delta_{i,j} < \frac{1}{8k} \cdot \min\left\{1, \frac{1}{\rho} - 1\right\} \cdot \min_{\{i,j\}}\{\sigma_k(\mathbb{E}[z_i z_j^\top])\}$$

for each pair $\{i, j\}$, then SpectralQuartetTest *returns the correct pairing* $\{\{z_1, z_2\}, \{z_3, z_4\}\}$.

4 The Spectral Recursive Grouping algorithm

The Spectral Recursive Grouping algorithm, presented as Algorithm 2, uses the spectral quartet test discussed in the previous section to estimate the structure of a multivariate latent tree distribution from iid samples of the observed leaf variables.[1] The algorithm is a modification of the recursive

---

[1] To simplify notation, we assume that the estimated second-moment matrices $\hat{\Sigma}_{x,y}$ and threshold parameters $\Delta_{x,y} \geq 0$ for all pairs $\{x, y\} \subset \mathcal{V}_{\mathrm{obs}}$ are globally defined. In particular, we assume the spectral quartet tests use these quantities.

6

Algorithm 2 Spectral Recursive Grouping.

Input: Empirical second-moment matrices $\widehat{\Sigma}_{x,y}$ for all pairs $\{x, y\} \subset \mathcal{V}_{\text{obs}}$ computed from N iid samples from the distribution over $\mathcal{V}_{\text{obs}}$; threshold parameters $\Delta_{x,y}$ for all pairs $\{x, y\} \subset \mathcal{V}_{\text{obs}}$.

Output: Tree structure $\widehat{\mathbb{T}}$ or "failure".

1: **let** $\mathcal{R} := \mathcal{V}_{\text{obs}}$, and for all $x \in \mathcal{R}$, $\mathcal{T}[x] :=$ rooted single-node tree x and $\mathcal{L}[x] := \{x\}$.
2: **while** $|\mathcal{R}| > 1$ **do**
3:     **let** pair $\{u, v\} \in \{\{\tilde{u}, \tilde{v}\} \subseteq \mathcal{R} : \text{Mergeable}(\mathcal{R}, \mathcal{L}[\cdot], \tilde{u}, \tilde{v}) = \text{true}\}$ be such that $\max\{\sigma_k(\widehat{\Sigma}_{x,y}) : (x, y) \in \mathcal{L}[u] \times \mathcal{L}[v]\}$ is maximized. If no such pair exists, then halt and return "failure".
4:     **let** `result` $:= \text{Relationship}(\mathcal{R}, \mathcal{L}[\cdot], \mathcal{T}[\cdot], u, v)$.
5:     **if** `result` $=$ "siblings" **then**
6:         Create a new variable h, create subtree $\mathcal{T}[h]$ rooted at h by joining $\mathcal{T}[u]$ and $\mathcal{T}[v]$ to h with edges $\{h, u\}$ and $\{h, v\}$, and set $\mathcal{L}[h] := \mathcal{L}[u] \cup \mathcal{L}[v]$.
7:         Add h to $\mathcal{R}$, and remove u and v from $\mathcal{R}$.
8:     **else if** `result` $=$ "u is parent of v" **then**
9:         Modify subtree $\mathcal{T}[u]$ by joining $\mathcal{T}[v]$ to u with an edge $\{u, v\}$, and modify $\mathcal{L}[u] := \mathcal{L}[u] \cup \mathcal{L}[v]$.
10:        Remove v from $\mathcal{R}$.
11:     **else if** `result` $=$ "v is parent of u" **then**
12:        {Analogous to above case.}
13:     **end if**
14: **end while**
15: Return $\widehat{\mathbb{T}} := \mathcal{T}[h]$ where $\mathcal{R} = \{h\}$.

---

grouping (RG) procedure proposed in [21]. RG builds the tree in a bottom-up fashion, where the initial working set of variables are the observed variables. The variables in the working set always correspond to roots of disjoint subtrees of $\mathbb{T}$ discovered by the algorithm. (Note that because these subtrees are rooted, they naturally induce parent/child relationships, but these may differ from those implied by the edge directions in $\mathbb{T}$.) In each iteration, the algorithm determines which variables in the working set to combine. If the variables are combined as siblings, then a new hidden variable is introduced as their parent and is added to the working set, and its children are removed. If the variables are combined as neighbors (parent/child), then the child is removed from the working set. The process repeats until the entire tree is constructed.

Our modification of RG uses the spectral quartet tests from Section 3 to decide which subtree roots in the current working set to combine. Note that because the test may return $\perp$ (a null result), our algorithm uses the tests to *rule out* possible siblings or neighbors among variables in the working set—this is encapsulated in the subroutine Mergeable (Algorithm 3), which tests quartets of observed variables (leaves) in the subtrees rooted at working set variables. For any pair $\{u, v\} \subseteq \mathcal{R}$ submitted to the subroutine (along with the current working set $\mathcal{R}$ and leaf sets $\mathcal{L}[\cdot]$):

- Mergeable returns false if there is evidence (provided by a quartet test) that u and v should first be joined with different variables (u' and v', respectively) before joining with each other; and

- Mergeable returns true if no quartet test provides such evidence.

The subroutine is also used by the subroutine Relationship (Algorithm 4) which determines whether a candidate pair of variables should be merged as neighbors (parent/child) or as siblings: essentially, to check if u is a parent of v, it checks if v is a sibling of each child of u. The use of unreliable estimates of long-range correlations is avoided by only considering highly-correlated variables as candidate pairs to merge (where correlation is measured using observed variables in their corresponding subtrees as proxies). This leads to a sample-efficient algorithm for recovering the hidden tree structure.

The Spectral Recursive Grouping algorithm enjoys the following guarantee.

Theorem 1. *Let $\eta \in (0, 1)$. Assume the directed tree graphical model $\mathbb{T}$ over variables (random vectors) $\mathcal{V}_{\mathbb{T}} = \mathcal{V}_{\text{obs}} \cup \mathcal{V}_{\text{hid}}$ satisfies Conditions 1, 2, 3, and 4. Suppose the Spectral Recursive*

7

Algorithm 3 Subroutine Mergeable($\mathcal{R}, \mathcal{L}[\cdot], u, v$).

Input: Set of nodes $\mathcal{R}$; leaf sets $\mathcal{L}[v]$ for all $v \in \mathcal{R}$; distinct $u, v \in \mathcal{R}$.
Output: true or false.
  1: **if** there exists distinct $u', v' \in \mathcal{R} \setminus \{u, v\}$ and $(x, y, x', y') \in \mathcal{L}[u] \times \mathcal{L}[v] \times \mathcal{L}[u'] \times \mathcal{L}[v']$ s.t. SpectralQuartetTest($\{x, y, x', y'\}$) returns $\{\{x, x'\}, \{y, y'\}\}$ or $\{\{x, y'\}, \{x', y\}\}$ **then** return false.
  2: **else** return true.

Algorithm 4 Subroutine Relationship($\mathcal{R}, \mathcal{L}[\cdot], \mathcal{T}[\cdot], u, v$).

Input: Set of nodes $\mathcal{R}$; leaf sets $\mathcal{L}[v]$ for all $v \in \mathcal{R}$; rooted subtrees $\mathcal{T}[v]$ for all $v \in \mathcal{R}$; distinct $u, v \in \mathcal{R}$.
Output: "siblings", "u is parent of v" ("$u \to v$"), or "v is parent of u" ("$v \to u$").
  1: **if** u is a leaf **then** assert $u \not\to v$.
  2: **if** v is a leaf **then** assert $v \not\to u$.
  3: **let** $\mathcal{R}[w] := (\mathcal{R} \setminus \{w\}) \cup \{w' : w' \text{ is a child of } w \text{ in } \mathcal{T}[w]\}$ for each $w \in \{u, v\}$.
  4: **if** there exists child u_1 of u in $\mathcal{T}[u]$ s.t. Mergeable($\mathcal{R}[u], \mathcal{L}[\cdot], u_1, v$) $=$ false **then** assert "$u \not\to v$".
  5: **if** there exists child v_1 of v in $\mathcal{T}[v]$ s.t. Mergeable($\mathcal{R}[v], \mathcal{L}[\cdot], u, v_1$) $=$ false **then** assert "$v \not\to u$".
  6: **if** both "$u \not\to v$" and "$v \not\to u$" were asserted **then** return "siblings".
  7: **else if** "$u \not\to v$" was asserted **then** return "v is parent of u" ("$v \to u$").
  8: **else** return "u is parent of v" ("$u \to v$").

Grouping algorithm (Algorithm 2) is provided N independent samples from the distribution over $\mathcal{V}_{\text{obs}}$, and uses parameters given by

$$\Delta_{x_i, x_j} := \sqrt{\frac{2 B_{x_i, x_j} t_{x_i, x_j}}{N}} + \frac{M_{x_i} M_{x_j} t_{x_i, x_j}}{3N} \tag{4}$$

where

$$B_{x_i, x_j} := \max\{\|\mathbb{E}[\|x_i\|^2 x_j x_j^\top]\|, \|\mathbb{E}[\|x_j\|^2 x_i x_i^\top]\|\}, \qquad M_{x_i} \geq \|x_i\| \quad \text{almost surely,}$$

$$\bar{d}_{x_i, x_j} := \frac{\mathbb{E}[\|x_i\|^2 \|x_j\|^2] - \text{tr}(\mathbb{E}[x_i x_j^\top] \mathbb{E}[x_j x_i^\top])}{\max\{\|\mathbb{E}[\|x_j\|^2 x_i x_i^\top]\|, \|\mathbb{E}[\|x_i\|^2 x_j x_j^\top]\|\}}, \qquad t_{x_i, x_j} := 4 \ln(4 \bar{d}_{x_i, x_j} n / \eta).$$

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

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

then with probability at least $1 - \eta$, the Spectral Recursive Grouping algorithm returns a tree $\widehat{\mathbb{T}}$ with the same undirected graph structure as $\mathbb{T}$.

Consistency is implied by the above theorem with an appropriate scaling of η 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.

Acknowledgements

Part of this work was completed while DH was at the Wharton School of the University of Pennsylvania and at Rutgers University. AA was supported by in part by the setup funds at UCI and the AFOSR Award FA9550-10-1-0310.

References

[1] 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.

8

[2] S. Dasgupta and L. Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.

[3] K. Chaudhuri, S. Dasgupta, and A. Vattani. Learning mixtures of Gaussians using the k-means algorithm, 2009. arXiv:0912.0086.

[4] D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

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

[6] N. Friedman, I. Nachman, and D. Peér. Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm. In *Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

[7] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

[8] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

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

[10] J. Wishart. Sampling errors in the theory of two factors. *British Journal of Psychology*, 19:180–187, 1928.

[11] K. Bollen. *Structural Equation Models with Latent Variables*. John Wiley & Sons, 1989.

[12] P. Buneman. The recovery of trees from measurements of dissimilarity. In F. R. Hodson, D. G. Kendall, and P. Tautu, editors, *Mathematics in the Archaeological and Historical Sciences*, pages 387–395. 1971.

[13] J. Pearl and M. Tarsi. Structuring causal trees. *Journal of Complexity*, 2(1):60–77, 1986.

[14] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[15] P. L. Erdös, L. A. Székely, M. A. Steel, and T. J. Warnow. A few logs suffice to build (almost) all trees: Part II. *Theoretical Computer Science*, 221:77–118, 1999.

[16] M. R. Lacey and J. T. Chang. A signal-to-noise analysis of phylogeny estimation by neighbor-joining: insufficiency of polynomial length sequences. *Mathematical Biosciences*, 199(2):188–215, 2006.

[17] P. L. Erdös, L. A. Székely, M. A. Steel, and T. J. Warnow. A few logs suffice to build (almost) all trees (I). *Random Structures and Algorithms*, 14:153–184, 1999.

[18] E. Mossel. Phase transitions in phylogeny. *Transactions of the American Mathematical Society*, 356(6):2379–2404, 2004.

[19] C. Daskalakis, E. Mossel, and S. Roch. Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel's conjecture. *Probability Theory and Related Fields*, 149(1–2):149–189, 2011.

[20] H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional galton-watson processes. *Annals of Mathematical Statistics*, 37:1463–1481, 1966.

[21] M. J. Choi, V. Tan, A. Anandkumar, and A. Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.

[22] M. S. Bartlett. Further aspects of the theory of multiple regression. *Mathematical Proceedings of the Cambridge Philosophical Society*, 34:33–40, 1938.

[23] 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.

[24] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. *Annals of Applied Probability*, 16(2):583–614, 2006.

[25] 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.

[26] S. M. Siddiqi, B. Boots, and G. J. Gordon. Reduced-rank hidden Markov models. In *Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

[27] L. Song, S. M. Siddiqi, G. J. Gordon, and A. J. Smola. Hilbert space embeddings of hidden Markov models. In *International Conference on Machine Learning*, 2010.

[28] 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.

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

[30] D. Hsu, S. M. Kakade, and T. Zhang. Dimension-free tail inequalities for sums of random matrices, 2011. arXiv:1104.1672.

9

A Sample-based confidence intervals for singular values

We show how to derive confidence bounds for the singular values of $\Sigma_{i,j} := \mathbb{E}[z_i z_j^\top]$ 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,

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

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

We state exponential tail inequalities for the spectral norm of the estimation error $\hat{\Sigma}_{i,j} - \Sigma_{i,j}$. 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*

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

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

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

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

Proof. Define the random matrix

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

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

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

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,

$$\left\| \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \right\| \leq \left\| \mathbb{E}[Z^2] \right\|$$
$$\leq \max\{\|\mathbb{E}[\|z_j\|^2 z_i z_i^\top]\|, \|\mathbb{E}[\|z_i\|^2 z_j z_j^\top]\|\}$$

and

$$\operatorname{tr}(\mathbb{E}[Z^2] - \mathbb{E}[Z]^2) = \operatorname{tr}(\mathbb{E}[\|z_j\|^2 z_i z_i^\top]) + \operatorname{tr}(\mathbb{E}[\|z_i\|^2 z_j z_j^\top]) - \operatorname{tr}(\Sigma_{i,j}\Sigma_{i,j}^\top) - \operatorname{tr}(\Sigma_{i,j}^\top \Sigma_{i,j})$$
$$= 2\Big(\mathbb{E}[\|z_i\|^2 \|z_j\|^2] - \operatorname{tr}(\Sigma_{i,j}\Sigma_{i,j}^\top) \Big).$$

Moreover,

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

By the matrix Bernstein inequality [30], for any $t > 0$,

$$\Pr\left[\left\| \hat{\Sigma}_{i,j} - \Sigma_{i,j} \right\| > \sqrt{\frac{2\left(\max\{\|\mathbb{E}[\|z_j\|^2 z_i z_i^\top]\|, \|\mathbb{E}[\|z_i\|^2 z_j z_j^\top]\|\}\right) t}{N}} + \frac{M_i M_j t}{3N} \right]$$

$$\leq 2 \cdot \frac{2\Big(\mathbb{E}[\|z_i\|^2 \|z_j\|^2] - \operatorname{tr}(\Sigma_{i,j}\Sigma_{i,j}^\top) \Big)}{\max\{\|\mathbb{E}[\|z_j\|^2 z_i z_i^\top]\|, \|\mathbb{E}[\|z_i\|^2 z_j z_j^\top]\|\}} \cdot t(e^t - t - 1)^{-1} = 4\bar{d}_{i,j}t(e^t - t - 1)^{-1}.$$

The claim follows. $\qquad\square$

10

In the case of discrete random variables (modeled as random vectors as described in Section 2), the following lemma from [25] can give a tighter exponential tail inequality.

Lemma 6 ([25]). *Let z_i and z_j be random vectors, each with support on the vertices of a probability simplex. Let $\Sigma_{i,j} := \mathbb{E}[z_i z_j^\top]$ and let $\hat{\Sigma}_{i,j}$ be the empirical average of N independent copies of $z_i z_j^\top$. Pick any $t > 0$. With probability at least $1 - e^{-t}$,*

$$\left\| \hat{\Sigma}_{i,j} - \Sigma_{i,j} \right\| \le \left\| \hat{\Sigma}_{i,j} - \Sigma_{i,j} \right\|_F \le \frac{1 + \sqrt{t}}{\sqrt{N}}$$

(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 $B_{i,j} \le \max\{M_i^2, M_j^2\}\sigma_*^2$. 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

For any hidden variable $h \in \mathcal{V}_{\mathrm{hid}}$, let $\mathrm{Descendants}_{\mathbb{T}}(h) \subseteq \mathcal{V}_{\mathbb{T}}$ be the descendants of h in $\mathbb{T}$. For any $g \in \mathrm{Descendants}_{\mathbb{T}}(h) \cap \mathcal{V}_{\mathrm{hid}}$ such that the (directed) path from h to g is $h \to g_1 \to g_2 \to \cdots \to g_q = g$, define $A_{(g|h)} \in \mathbb{R}^{k \times k}$ to be the product

$$A_{(g|h)} := A_{(g_q|g_{q-1})} \cdots A_{(g_2|g_1)} A_{(g_1|h)}.$$

Similarly, for any $x \in \mathrm{Descendants}_{\mathbb{T}}(h) \cap \mathcal{V}_{\mathrm{obs}}$ such that the (directed) path from h to x is $h \to g_1 \to g_2 \to \cdots \to g_q \to x$, define $C_{(x|h)} \in \mathbb{R}^{d \times k}$ to be the product

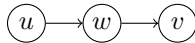
$$C_{(x|h)} := C_{(x|g_q)} A_{(g_q|g_{q-1})} \cdots A_{(g_2|g_1)} A_{(g_1|h)}.$$

B.1 $\log \det_k$ metric

Define the function $\mu \colon \mathcal{V}_{\mathbb{T}} \times \mathcal{V}_{\mathbb{T}} \to \mathbb{R}$ by

$$\mu(u,v) := \begin{cases} \log \det_k(\mathbb{E}[uu^\top]^{-1/2} \mathbb{E}[uv^\top] \mathbb{E}[vv^\top]^{-1/2}) & \text{if } u, v \in \mathcal{V}_{\mathrm{hid}} \\ \log \det_k(\mathbb{E}[uv^\top] \mathbb{E}[vv^\top]^{-1/2}) & \text{if } u \in \mathcal{V}_{\mathrm{obs}}, v \in \mathcal{V}_{\mathrm{hid}} \\ \log \det_k(\mathbb{E}[uu^\top]^{-1/2} \mathbb{E}[uv^\top]) & \text{if } u \in \mathcal{V}_{\mathrm{hid}}, v \in \mathcal{V}_{\mathrm{obs}} \\ \log \det_k(\mathbb{E}[uv^\top]) & \text{if } u, v \in \mathcal{V}_{\mathrm{obs}} \end{cases}.$$

Proposition 1 ($\log \det_k$ metric). *Assume Conditions 1 and 2 hold, and pick any $u, v \in \mathcal{V}_{\mathbb{T}}$. If $w \in \mathcal{V}_{\mathbb{T}} \setminus \{u, v\}$ is on the (undirected) path $u \rightsquigarrow v$, then $\mu(u,v) = \mu(u,w) + \mu(w,v)$.*

Proof. Suppose the induced topology over u, v, w in $\mathbb{T}$ is the following.

Assume for now that $u, v \in \mathcal{V}_{\mathrm{hid}}$. Then, using Condition 1,

$$\mathbb{E}[uv^\top] = \mathbb{E}[uw^\top] A_{(v|w)}^\top = (\mathbb{E}[uw^\top] \mathbb{E}[ww^\top]^{-1/2})(\mathbb{E}[ww^\top]^{-1/2} \mathbb{E}[wv^\top])$$

11

so, because $\operatorname{rank}(\mathbb{E}[uu^\top]^{-1/2}\mathbb{E}[uw^\top]\mathbb{E}[ww^\top]^{-1/2}) = \operatorname{rank}(\mathbb{E}[ww^\top]^{-1/2}\mathbb{E}[wv^\top]\mathbb{E}[vv^\top]^{-1/2}) = k$ by Condition 2,

$$\mu(u,v) = \log\det_k(\mathbb{E}[uu^\top]^{-1/2}\mathbb{E}[uw^\top]\mathbb{E}[ww^\top]^{-1/2}\mathbb{E}[ww^\top]^{-1/2}\mathbb{E}[wv^\top]\mathbb{E}[vv^\top]^{-1/2})$$
$$= \log\det_k(\mathbb{E}[uu^\top]^{-1/2}\mathbb{E}[uw^\top]\mathbb{E}[ww^\top]^{-1/2}) + \log\det_k(\mathbb{E}[ww^\top]^{-1/2}\mathbb{E}[wv^\top]\mathbb{E}[vv^\top]^{-1/2})$$
$$= \mu(u,w) + \mu(w,v).$$

If $u \in \mathcal{V}_{\mathrm{hid}}$ but $v \in \mathcal{V}_{\mathrm{obs}}$, then let $U_v \in \mathbb{R}^{d\times k}$ be a matrix of orthonormal left singular vectors of $C_{(v|w)}$. Then $\mathbb{E}[uv^\top] = (\mathbb{E}[uw^\top]\mathbb{E}[ww^\top]^{-1/2})(\mathbb{E}[ww^\top]^{-1/2}\mathbb{E}[wv^\top])$ as before, and

$$\det_k(\mathbb{E}[uu^\top]^{-1/2}\mathbb{E}[uv^\top]) = |\det(\mathbb{E}[uu^\top]^{-1/2}\mathbb{E}[uv^\top]U_v)|$$
$$= |\det(\mathbb{E}[uu^\top]^{-1/2})| \cdot |\det(\mathbb{E}[uv^\top]U_v)|$$
$$= \det_k(\mathbb{E}[uu^\top]^{-1/2}\mathbb{E}[uw^\top]\mathbb{E}[ww^\top]^{-1/2}) \cdot \det_k(\mathbb{E}[ww^\top]^{-1/2}\mathbb{E}[wv^\top]U_v)$$
$$= \det_k(\mathbb{E}[uu^\top]^{-1/2}\mathbb{E}[uw^\top]\mathbb{E}[ww^\top]^{-1/2}) \cdot \det_k(\mathbb{E}[ww^\top]^{-1/2}\mathbb{E}[wv^\top]),$$

so

$$\mu(u,v) = \log\det_k(\mathbb{E}[uu^\top]^{-1/2}\mathbb{E}[uw^\top]\mathbb{E}[ww^\top]^{-1/2}) + \log\det_k(\mathbb{E}[ww^\top]^{-1/2}\mathbb{E}[wv^\top]) = \mu(u,w) + \mu(w,v).$$

Suppose now that the induced toplogy over u, v, w in $\mathbb{T}$ is the following.

Again, first assume that $u, v \in \mathcal{V}_{\mathrm{hid}}$. Then, by Condition 1,

$$\mathbb{E}[uv^\top] = A_{(u|w)}\mathbb{E}[ww^\top]A_{(v|w)}^\top = (\mathbb{E}[uw^\top]\mathbb{E}[ww^\top]^{-1/2})(\mathbb{E}[ww^\top]^{-1/2}\mathbb{E}[wv^\top]),$$

so $\mu(u,v) = \mu(u,w) + \mu(v,w)$ as before. The cases where one or both of u and v is in $\mathcal{V}_{\mathrm{obs}}$ follow by similar arguments as above. $\qquad\square$

B.2   Proof of Lemma 1

By Proposition 1,

$$\det_k(\mathbb{E}[z_1 z_3^\top]) \cdot \det_k(\mathbb{E}[z_2 z_4^\top]) = \exp(\mu(z_1, z_3) + \mu(z_2, z_4))$$
$$= \exp(\mu(z_1, h) + \mu(h, g) + \mu(g, z_3) + \mu(z_2, h) + \mu(h, g) + \mu(g, z_4))$$
$$= \exp(\mu(z_1, h) + \mu(h, g) + \mu(g, z_4) + \mu(z_2, h) + \mu(h, g) + \mu(g, z_3))$$
$$= \exp(\mu(z_1, z_4) + \mu(z_2, z_3))$$
$$= \det_k(\mathbb{E}[z_1 z_4^\top]) \cdot \det_k(\mathbb{E}[z_2 z_3^\top]).$$

Moreover,

$$\frac{\det_k(\mathbb{E}[z_1 z_3^\top]) \cdot \det_k(\mathbb{E}[z_2 z_4^\top])}{\det_k(\mathbb{E}[z_1 z_2^\top]) \cdot \det_k(\mathbb{E}[z_3 z_4^\top])} = \frac{\exp(\mu(z_1, z_3) + \mu(z_2, z_4))}{\exp(\mu(z_1, z_2) + \mu(z_3, z_4))}$$
$$= \frac{\exp(\mu(z_1, h) + \mu(h, g) + \mu(g, z_3) + \mu(z_2, h) + \mu(h, g) + \mu(g, z_4))}{\exp(\mu(z_1, h) + \mu(h, z_2) + \mu(z_3, g) + \mu(g, z_4))}$$
$$= \exp(2\mu(h, g))$$
$$= \det(\mathbb{E}[hh^\top]^{-1/2}\mathbb{E}[hg^\top]\mathbb{E}[gg^\top]^{-1/2})^2$$
$$= \frac{\det(\mathbb{E}[hg^\top])^2}{\det(\mathbb{E}[hh^\top]) \cdot \det(\mathbb{E}[gg^\top])}.$$

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

$$\frac{\det(\mathbb{E}[hg^\top])^2}{\det(\mathbb{E}[hh^\top]) \cdot \det(\mathbb{E}[gg^\top])} = \det(\mathbb{E}[hh^\top]^{-1/2}\mathbb{E}[hg^\top]\mathbb{E}[gg^\top]^{-1/2})^2 \leq 1$$

as required. $\qquad\square$

Note that if Condition 3 also holds, then Lemma 1 implies the strict inequalities

$$\max\left\{\det_k(\mathbb{E}[z_1 z_3^\top]) \cdot \det_k(\mathbb{E}[z_2 z_4^\top]),\ \det_k(\mathbb{E}[z_1 z_4^\top]) \cdot \det_k(\mathbb{E}[z_2 z_3^\top])\right\} < \det_k(\mathbb{E}[z_1 z_2^\top]) \cdot \det_k(\mathbb{E}[z_3 z_4^\top]).$$

12

B.3 Proof of Lemma 2

Given that (2) holds for all pairs $\{i, j\}$ and all $s \in \{1, 2, \ldots, k\}$, if the spectral quartet test returns a pairing $\{\{z_i, z_j\}, \{z_{i'}, z_{j'}\}\}$, it must be that

$$\prod_{s=1}^{k} \sigma_s(\mathbb{E}[z_i z_j^\top]) \sigma_s(\mathbb{E}[z_{i'} z_{j'}^\top]) \geq \prod_{s=1}^{k} [\sigma_s(\hat{\Sigma}_{i,j}) - \Delta_{i,j}]_+ [\sigma_s(\hat{\Sigma}_{i',j'}) - \Delta_{i',j'}]_+$$

$$> \prod_{s=1}^{k} (\sigma_s(\hat{\Sigma}_{i',j}) + \Delta_{i',j})(\sigma_s(\hat{\Sigma}_{i,j'}) + \Delta_{i,j'}) \geq \prod_{s=1}^{k} \sigma_s(\mathbb{E}[z_{i'} z_j^\top]) \sigma_s(\mathbb{E}[z_i z_{j'}^\top]).$$

Therefore

$$\det_k(\mathbb{E}[z_i z_j^\top]) \cdot \det_k(\mathbb{E}[z_{i'} z_{j'}^\top]) = \prod_{s=1}^{k} \sigma_s(\mathbb{E}[z_i z_j^\top]) \sigma_s(\mathbb{E}[z_{i'} z_{j'}^\top])$$

$$> \prod_{s=1}^{k} \sigma_s(\mathbb{E}[z_{i'} z_j^\top]) \sigma_s(\mathbb{E}[z_i z_{j'}^\top]) = \det_k(\mathbb{E}[z_{i'} z_j^\top]) \cdot \det_k(\mathbb{E}[z_i z_{j'}^\top]).$$

But by Lemma 1, the above inequality can only hold if $\{\{z_i, z_j\}, \{z_{i'}, z_{j'}\}\} = \{\{z_1, z_2\}, \{z_3, z_4\}\}$.
□

B.4 Proof of Lemma 4

Let $\Sigma_{i,j} := \mathbb{E}[z_i z_j^\top]$. The assumptions in the statement of the lemma imply

$$\max\{\Delta_{1,2}, \Delta_{3,4}\} < \frac{\epsilon_0}{8k} \min\{\sigma_k(\Sigma_{1,2}), \sigma_k(\Sigma_{3,4})\}$$

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

$$\prod_{s=1}^{k} [\sigma_s(\hat{\Sigma}_{1,2}) - \Delta_{1,2}]_+ [\sigma_s(\hat{\Sigma}_{3,4}) - \Delta_{3,4}]_+ \geq \prod_{s=1}^{k} [\sigma_s(\Sigma_{1,2}) - 2\Delta_{1,2}]_+ [\sigma_s(\Sigma_{3,4}) - 2\Delta_{3,4}]_+$$

$$> \left(\prod_{s=1}^{k} \sigma_s(\Sigma_{1,2}) \sigma_s(\Sigma_{3,4})\right) \left(1 - \frac{\epsilon_0}{4k}\right)^{2k}$$

$$\geq \left(\prod_{s=1}^{k} \sigma_s(\Sigma_{1,2}) \sigma_s(\Sigma_{3,4})\right) (1 - \epsilon_0/2). \qquad (5)$$

If $\mathbb{E}[hg^\top]$ has rank k, then so do $\Sigma_{i,j}$ for $i \in \{1, 2\}$ and $j \in \{3, 4\}$. Therefore, for $\{i', j'\} = \{1, 2, 3, 4\} \setminus \{i, j\}$,

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

This implies

$$\prod_{s=1}^{k} (\sigma_s(\hat{\Sigma}_{i,j}) + \Delta_{i,j})(\sigma_s(\hat{\Sigma}_{i',j'}) + \Delta_{i',j'}) \leq \prod_{s=1}^{k} (\sigma_s(\Sigma_{i,j}) + 2\Delta_{i,j})(\sigma_s(\Sigma_{i',j'}) + 2\Delta_{i',j'})$$

$$< \left(\prod_{s=1}^{k} \sigma_s(\Sigma_{i,j}) \sigma_s(\Sigma_{i',j'})\right) \left(1 + \frac{\epsilon_0}{4k}\right)^{2k}$$

$$\leq \left(\prod_{s=1}^{k} \sigma_s(\Sigma_{i,j}) \sigma_s(\Sigma_{i',j'})\right) (1 + \epsilon_0). \qquad (6)$$

13

Therefore, combining (5), (6), and Lemma 1,

$$\prod_{s=1}^{k}[\sigma_s(\hat{\Sigma}_{1,2}) - \Delta_{1,2}]_+[\sigma_s(\hat{\Sigma}_{3,4}) - \Delta_{3,4}]_+$$

$$> \frac{1 - \epsilon_0/2}{1 + \epsilon_0} \cdot \frac{\det(\mathbb{E}[hh^\top])\det(\mathbb{E}[gg^\top])}{\det(\mathbb{E}[hg^\top])^2} \cdot \prod_{s=1}^{k}(\sigma_s(\hat{\Sigma}_{i,j}) + \Delta_{i,j})(\sigma_s(\hat{\Sigma}_{i',j'}) + \Delta_{i',j'})$$

$$\geq \frac{1}{(1 + \epsilon_0)^2} \cdot \frac{\det(\mathbb{E}[hh^\top])\det(\mathbb{E}[gg^\top])}{\det(\mathbb{E}[hg^\top])^2} \cdot \prod_{s=1}^{k}(\sigma_s(\hat{\Sigma}_{i,j}) + \Delta_{i,j})(\sigma_s(\hat{\Sigma}_{i',j'}) + \Delta_{i',j'})$$

$$\geq \prod_{s=1}^{k}(\sigma_s(\hat{\Sigma}_{i,j}) + \Delta_{i,j})(\sigma_s(\hat{\Sigma}_{i',j'}) + \Delta_{i',j'}),$$

so the spectral quartet test will return the correct pairing $\{\{z_1, z_2\}, \{z_3, z_4\}\}$, proving the lemma.
$\square$

B.5 Conditions for returning a correct pairing when $\mathrm{rank}(\mathbb{E}[hg^\top]) < k$

The spectral quartet test is also useful in the case where $\mathbb{E}[hg^\top]$ has rank $r < k$. In this case, the widths of the confidence intervals are allowed to be wider than in the case where $\mathrm{rank}(\mathbb{E}[hg^\top]) = k$. Define

$$\sigma_{\min} := \min\Big(\{\sigma_k(\Sigma_{1,2}), \sigma_k(\Sigma_{3,4})\} \cup \{\sigma_r(\Sigma_{i,j}) : i \in \{1, 2\}, j \in \{3, 4\}\}\Big).$$

$$\rho_1^2 = \frac{\sigma_{\min}^{2(k-r)} \cdot \max_{i,j,i',j'} \prod_{s=1}^{r} \sigma_s(\Sigma_{i,j})\sigma_s(\Sigma_{i',j'})}{\prod_{s=1}^{k} \sigma_s(\Sigma_{1,2})\sigma_s(\Sigma_{3,4})}.$$

Instead of depending on $\min_{i,j}\{\sigma_k(\Sigma_{i,j})\}$ and ρ as in the case where $\mathrm{rank}(\mathbb{E}[hg^\top]) = k$, we only depend on $\sigma_{\min}$ and ρ_1.

Lemma 7 (Correct pairing, rank $r < k$). *Suppose that (i) the observed variables $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$ have the true induced (undirected) topology shown in Figure 1(a), (ii) the tree model satisfies Condition 1 and Condition 2, (iii) $\mathbb{E}[hg^\top]$ 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 $\{\{z_1, z_2\}, \{z_3, z_4\}\}$.

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

Proof. 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

$$\prod_{s=1}^{k}[\sigma_s(\hat{\Sigma}_{1,2}) - \Delta_{1,2}]_+[\sigma_s(\hat{\Sigma}_{3,4}) - \Delta_{3,4}]_+ > \left(\prod_{s=1}^{k}\sigma_s(\Sigma_{1,2})\sigma_s(\Sigma_{3,4})\right)(1 - \epsilon_1/2)$$

14

as in the proof of Lemma 4. Moreover,

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

$$
< \left(\prod_{s=1}^{r}\sigma_s(\Sigma_{i,j})\sigma_s(\Sigma_{i',j'})\right)\cdot(1+\epsilon_1)\cdot\left(\frac{\epsilon_1\sigma_{\min}}{8k}\right)^{2(k-r)}
$$

$$
\leq \left(\prod_{s=1}^{k}\sigma_s(\Sigma_{1,2})\sigma_s(\Sigma_{3,4})\right)\cdot\frac{\rho_1^2}{(\sigma_{\min})^{2(k-r)}}\cdot(1+\epsilon_1)\cdot\left(\frac{\epsilon_1\sigma_{\min}}{8k}\right)^{2(k-r)}
$$

$$
= \left(\prod_{s=1}^{k}\sigma_s(\Sigma_{1,2})\sigma_s(\Sigma_{3,4})\right)\cdot\rho_1^2\cdot(1+\epsilon_1)\cdot\left(\frac{\epsilon_1}{8k}\right)^{2(k-r)}
$$

$$
< \left(\prod_{s=1}^{k}[\sigma_s(\hat{\Sigma}_{1,2}) - \Delta_{1,2}]_+[\sigma_s(\hat{\Sigma}_{3,4}) - \Delta_{3,4}]_+\right)\cdot\rho_1^2\cdot\frac{1+\epsilon_1}{1-\epsilon_1/2}\cdot\left(\frac{\epsilon_1}{8k}\right)^{2(k-r)}
$$

$$
\leq \left(\prod_{s=1}^{k}[\sigma_s(\hat{\Sigma}_{1,2}) - \Delta_{1,2}]_+[\sigma_s(\hat{\Sigma}_{3,4}) - \Delta_{3,4}]_+\right)\cdot\rho_1^2\cdot(1+\epsilon_1)^2\cdot\left(\frac{\epsilon_1}{8k}\right)^{2(k-r)}
$$

$$
\leq \prod_{s=1}^{k}[\sigma_s(\hat{\Sigma}_{1,2}) - \Delta_{1,2}]_+[\sigma_s(\hat{\Sigma}_{3,4}) - \Delta_{3,4}]_+.
$$

Therefore the spectral quartet test will return the correct pairing $\{\{z_1, z_2\}, \{z_3, z_4\}\}$; the lemma follows. $\qquad\square$

C Analysis of Spectral Recursive Grouping

C.1 Overview

Here is an outline of the argument for Theorem 1.

1. First, we condition on a $1 - \eta$ probability event over the iid samples from the distribution over $\mathcal{V}_{\mathrm{obs}}$ in which the empirical second-moment matrices are sufficiently close to the true second-moment matrices in by spectral norm (Equation 8). This is required to reason deterministically about the behavior of the algorithm.

2. Next, we characterize the pairs $\{u, v\} \subseteq \mathcal{R}$ (where $\mathcal{R}$ are the roots of subtrees maintained by the algorithm) that cause the Mergeable subroutine to return true. (Lemma 11), as well as those that cause it to return false (Lemma 12).

3. We use the above characterizations to show that the main while-loop of the algorithm maintains loop invariants such that when the loop finally terminates, the entire tree structure will have been completely discovered (Lemma 13). This is achieved by showing each iteration of the while-loop

   (a) selects a "Mergeable" pair $\{u, v\} \subseteq \mathcal{R}$ that satisfies certain properties (Claim 2 and Claim 3) such that, if they are properly combined (as siblings or parent/child), the required loop invariants will be perserved; and

   (b) uses the Relationship subroutine to correctly determine whether the chosen pair $\{u, v\}$ should be combined as siblings or parent/child (Claim 4).

C.2 Proof of Theorem 1

Recall the definitions of $A_{(g|h)} \in \mathbb{R}^{k \times k}$ and $C_{(x|h)} \in \mathbb{R}^{d \times k}$ for descendants $g \in \mathrm{Descendants}_{\mathbb{T}}(h) \cap \mathcal{V}_{\mathrm{hid}}$ and $x \in \mathrm{Descendants}_{\mathbb{T}}(h) \cap C_{(x|h)}$ in $\mathbb{T}$, as given in Appendix B.

15

Let us define

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

$$\theta := \frac{\gamma_{\min}}{1 + \varepsilon}, \qquad\qquad \varsigma := \frac{\gamma_{\min}}{\gamma_{\max}} \cdot (1 - \varepsilon) \cdot \theta.$$

The sample size requirement ensures that

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

This implies conditions on the thresholds Δ_{x_i, x_j} in Lemma 4 for the spectral quartet test on $\{x_1, x_2, x_3, x_4\}$ to return a correct pairing, provided that

$$\min\{\sigma_k(\Sigma_{x_i, x_j}) : \{i, j\} \subset \{1, 2, 3, 4\}\} \ge \varsigma. \tag{7}$$

The probabilistic event we need is that in which the confidence bounds from Lemma 5 hold for each pair of observed variables. The event

$$\forall \{x_i, x_j\} \subseteq \mathcal{V}_{\mathrm{obs}} \bullet \|\widehat{\Sigma}_{x_i, x_j} - \Sigma_{x_i, x_j}\| \le \Delta_{x_i, x_j}, \tag{8}$$

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

The following is an immediate consequence of Weyl's Theorem and conditioning on the above event.

Lemma 8. *Fix any pair $\{x, y\} \subseteq \mathcal{V}_{\mathrm{obs}}$. If $\sigma_k(\Sigma_{x,y}) \ge (1+\varepsilon)\theta$, then $\sigma_k(\widehat{\Sigma}_{x,y}) \ge \theta$. If $\sigma_k(\widehat{\Sigma}_{x,y}) \ge \theta$, then $\sigma_k(\Sigma_{x,y}) \ge (1-\varepsilon)\theta$.*

Before continuing, we need some definitions and notation. First, we refer to the variables in $\mathcal{V}_{\mathbb{T}}$ interchangeably as both nodes and variables. Next, we generally ignore the direction of edges in $\mathbb{T}$, except when it becomes crucial (namely, in Lemma 10). For a node r in $\mathbb{T}$, we say that a subtree $\mathcal{T}[r]$ of $\mathbb{T}$ (ignoring edge directions) is *rooted at r* if $\mathcal{T}[r]$ contains r, and for every node u in $\mathcal{T}[r]$ and any node v not in $\mathcal{T}[r]$, the (undirected) path from u to v in $\mathbb{T}$ passes through r. Note that a rooted subtree naturally imply parent/child relationships between its constituent nodes, and it is in this sense we use the terms "parent", "child", "sibling", etc. throughout the analysis, rather than in the sense given by the edge directions in $\mathbb{T}$ (the exception is in Lemma 10). A collection $\mathcal{C}$ of disjoint rooted subtrees of $\mathbb{T}$ naturally gives rise to a *super-tree $\mathcal{ST}[\mathcal{C}]$* by starting with $\mathbb{T}$ and then collapsing each $\mathcal{T}[r] \in \mathcal{C}$ into a single node. Note that each node in $\mathcal{ST}[\mathcal{C}]$ is either associated with a subtree in $\mathcal{C}$, or is a node in $\mathbb{T}$ that doesn't appear in any subtree in $\mathcal{C}$. We say a subtree $\mathcal{T} \in \mathcal{C}$ is a *leaf component relative to $\mathcal{C}$* if it is a leaf in this super-tree $\mathcal{ST}[\mathcal{C}]$. Finally, define $\mathcal{V}_{\mathrm{hid}}[\mathcal{C}] := \{h \in \mathcal{V}_{\mathrm{hid}} : h$ does not appear in any subtree in $\mathcal{C}\}$.

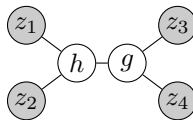
The following lemma is a simple fact about the super-tree given properties on the subtrees (which will be maintained by the algorithm).

Lemma 9 (Super-tree property). *Let $\mathcal{R} \subseteq \mathcal{V}_{\mathbb{T}}$. Let $\mathcal{C} := \{\mathcal{T}[u] : u \in \mathcal{R}\}$ be a collection of disjoint rooted subtrees, with u being the root of $\mathcal{T}[u]$, such that their leaf sets $\{\mathcal{L}[u] : u \in \mathcal{R}\}$ partition $\mathcal{V}_{\mathrm{obs}}$. Then the nodes of the super-tree $\mathcal{ST}[\mathcal{C}]$ are $\mathcal{C} \cup \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$, and the leaves of $\mathcal{ST}[\mathcal{C}]$ are all in $\mathcal{C}$.*

Proof. This follows because each leaf in $\mathbb{T}$ appears in the leaf set of some $\mathcal{T}[u]$. □

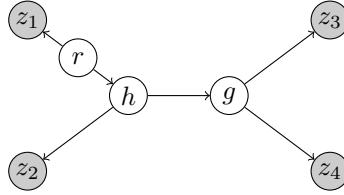
The next lemma relates the correlation between two observed variables in a quartet (on opposite sides of the bottleneck) to the correlations of the other pairs crossing the bottleneck.

Lemma 10 (Correlation transfer). *Consider the following induced (undirected) topology over $\{z_1, z_2, z_3, z_4\} \subseteq \mathcal{V}_{\mathrm{obs}}$.*

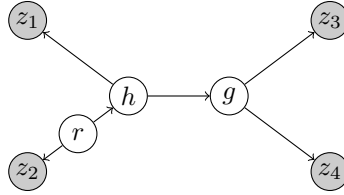
16

Then

$$\sigma_k(\mathbb{E}[z_1 z_4^\top]) \geq \frac{\sigma_k(\mathbb{E}[z_1 z_3^\top])\sigma_k(\mathbb{E}[z_2 z_4^\top])}{\sigma_1(\mathbb{E}[z_2 z_3^\top])}.$$

Proof. In this proof, the edge directions and the notion of ancestor are determined according to the edge directions in $\mathbb{T}$. Let r be the least common ancestor of $\{z_1, z_2, z_3, z_4\}$ in $\mathbb{T}$. There are effectively three possible cases to consider, depending on the location of r relative to the z_i, h, and g; we may exploit the fact that $\sigma_k(\mathbb{E}[z_1 z_4^\top]) = \sigma_k(\mathbb{E}[z_4 z_1^\top])$ to cover the remaining cases.

1. Suppose r appears between h and z_1.

By Condition 2, we can choose matrices $U_1, U_2, U_3, U_4 \in \mathbb{R}^{d \times k}$ such that the columns of U_1 are an orthonormal basis of $\mathrm{range}(C_{(z_1|r)})$, the columns of U_2 are an orthonormal basis of $\mathrm{range}(C_{(z_2|h)})$, the columns of U_3 are an orthonormal basis of $\mathrm{range}(C_{(z_3|g)})$, and the columns of U_4 are an orthonormal basis of $\mathrm{range}(C_{(z_4|g)})$. We have

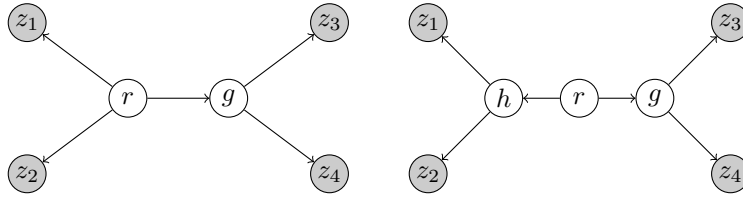
$$
\begin{aligned}
&U_1^\top \mathbb{E}[z_1 z_4^\top] U_4 \\
&= U_1^\top C_{(z_1|r)} \mathbb{E}[rr^\top] A_{(h|r)}^\top C_{(z_4|h)}^\top U_4 \\
&= (U_1^\top C_{(z_1|r)} \mathbb{E}[rr^\top]) A_{(h|r)}^\top (C_{(z_3|h)}^\top U_3)(C_{(z_3|h)}^\top U_3)^{-1} \\
&\quad (U_2^\top C_{(z_2|h)} \mathbb{E}[hh^\top])^{-1}(U_2^\top C_{(z_2|h)} \mathbb{E}[hh^\top])(C_{(z_4|h)}^\top U_4) \\
&= (U_1^\top C_{(z_1|r)} \mathbb{E}[rr^\top] A_{(h|r)}^\top C_{(z_3|h)}^\top U_3)(U_2^\top C_{(z_2|h)} \mathbb{E}[hh^\top] C_{(z_3|h)}^\top U_3)^{-1} \\
&\quad (U_2^\top C_{(z_2|h)} \mathbb{E}[hh^\top] C_{(z_4|h)}^\top U_4) \\
&= (U_1^\top \mathbb{E}[z_1 z_3^\top] U_3)(U_2^\top \mathbb{E}[z_2 z_3^\top] U_3)^{-1}(U_2^\top \mathbb{E}[z_2 z_4^\top] U_4).
\end{aligned}
$$

2. Suppose r appears between h and z_2.

By Condition 2, we can choose matrices $U_1, U_2, U_3, U_4 \in \mathbb{R}^{d \times k}$ such that the columns of U_1 are an orthonormal basis of $\mathrm{range}(C_{(z_1|h)})$, the columns of U_2 are an orthonormal basis of $\mathrm{range}(C_{(z_2|r)})$, the columns of U_3 are an orthonormal basis of $\mathrm{range}(C_{(z_3|g)})$, and the

17

columns of U_4 are an orthonormal basis of $\mathrm{range}(C_{(z_4|g)})$. We have

$$
U_1^\top \mathbb{E}[z_1 z_4^\top] U_4
$$
$$
= U_1^\top C_{(z_1|h)} \mathbb{E}[hh^\top] A_{(h|r)}^{-\top} C_{(z_4|r)}^\top U_4
$$
$$
= (U_1^\top C_{(z_1|h)} \mathbb{E}[hh^\top])(C_{(z_3|h)}^\top U_3)(C_{(z_3|h)}^\top U_3)^{-1} A_{(h|r)}^{-\top}
$$
$$
\quad (U_2^\top C_{(z_2|r)} \mathbb{E}[rr^\top])^{-1}(U_2^\top C_{(z_2|r)} \mathbb{E}[rr^\top])(C_{(z_4|r)}^\top U_4)
$$
$$
= (U_1^\top C_{(z_1|h)} \mathbb{E}[hh^\top] C_{(z_3|h)}^\top U_3)(U_2^\top C_{(z_2|r)} \mathbb{E}[rr^\top] A_{(h|r)}^\top C_{(z_3|h)}^\top U_3)^{-1}
$$
$$
\quad (U_2^\top C_{(z_2|r)} \mathbb{E}[rr^\top] C_{(z_4|r)}^\top U_4)
$$
$$
= (U_1^\top \mathbb{E}[z_1 z_3^\top] U_3)(U_2^\top \mathbb{E}[z_2 z_3^\top] U_3)^{-1}(U_2^\top \mathbb{E}[z_2 z_4^\top] U_4).
$$

3. Suppose either $r = h$, or r is between h and g.

In either case, by Condition 2, we can choose matrices $U_1, U_2, U_3, U_4 \in \mathbb{R}^{d \times k}$ such that the columns of U_1 are an orthonormal basis of $\mathrm{range}(C_{(z_1|h)})$, the columns of U_2 are an orthonormal basis of $\mathrm{range}(C_{(z_2|h)})$, the columns of U_3 are an orthonormal basis of $\mathrm{range}(C_{(z_3|g)})$, and the columns of U_4 are an orthonormal basis of $\mathrm{range}(C_{(z_4|g)})$. We have

$$
U_1^\top \mathbb{E}[z_1 z_4^\top] U_4
$$
$$
= U_1^\top C_{(z_1|r)} \mathbb{E}[rr^\top] C_{(z_4|r)}^\top U_4
$$
$$
= (U_1^\top C_{(z_1|r)} \mathbb{E}[rr^\top])(C_{(z_3|r)}^\top U_3)(C_{(z_3|r)}^\top U_3)^{-1}
$$
$$
\quad (U_2^\top C_{(z_2|r)} \mathbb{E}[rr^\top])^{-1}(U_2^\top C_{(z_2|r)} \mathbb{E}[rr^\top])(C_{(z_4|r)}^\top U_4)
$$
$$
= (U_1^\top C_{(z_1|r)} \mathbb{E}[rr^\top] C_{(z_3|r)}^\top U_3)(U_2^\top C_{(z_2|r)} \mathbb{E}[rr^\top] C_{(z_3|r)}^\top U_3)^{-1}
$$
$$
\quad (U_2^\top C_{(z_2|r)} \mathbb{E}[rr^\top] C_{(z_4|r)}^\top U_4)
$$
$$
= (U_1^\top \mathbb{E}[z_1 z_3^\top] U_3)(U_2^\top \mathbb{E}[z_2 z_3^\top] U_3)^{-1}(U_2^\top \mathbb{E}[z_2 z_4^\top] U_4).
$$

Therefore, in all cases,

$$
\sigma_k(\mathbb{E}[z_1 z_4^\top]) \geq \frac{\sigma_k(\mathbb{E}[z_1 z_3^\top]) \cdot \sigma_k(\mathbb{E}[z_2 z_4^\top])}{\sigma_1(\mathbb{E}[z_2 z_3^\top])}. \qquad \square
$$

The next two lemmas (Lemmas 11 and 12) show a dichotomy in the cases that cause the subroutine Mergeable return either true or false.

Lemma 11 (Mergeable pairs). *Let $\mathcal{R} \subseteq \mathcal{V}_\mathbb{T}$. Let $\mathcal{C} := \{\mathcal{T}[r] : r \in \mathcal{R}\}$ be a collection of disjoint rooted subtrees, with r being the root of $\mathcal{T}[r]$, such that their leaf sets $\{\mathcal{L}[r] : r \in \mathcal{R}\}$ partition $\mathcal{V}_{\mathrm{obs}}$. Further, suppose the pair $\{u, v\} \subseteq \mathcal{R}$ are such that one of the following conditions hold.*

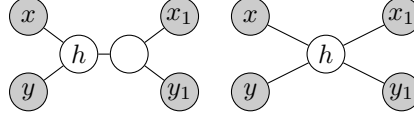
1. *$\{u, v\}$ share a common neighbor in $\mathbb{T}$, and both of $\mathcal{T}[u]$ and $\mathcal{T}[v]$ are leaf components relative to $\mathcal{C}$.*

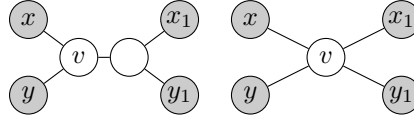
2. *$\{u, v\}$ are neighbors in $\mathbb{T}$, and at least one of $\mathcal{T}[u]$ and $\mathcal{T}[v]$ is a leaf component relative to $\mathcal{C}$.*

Then for all pairs $\{u_1, v_1\} \subseteq \mathcal{R} \setminus \{u, v\}$ and all $(x, y, x_1, y_1) \in \mathcal{L}[u] \times \mathcal{L}[v] \times \mathcal{L}[u_1] \times \mathcal{L}[v_1]$, $\mathsf{SpectralQuartetTest}(\{x, y, x_1, y_1\})$ returns $\{\{x, y\}, \{x_1, y_1\}\}$ or $\bot$. This implies that $\mathsf{Mergeable}(\mathcal{R}, \mathcal{L}[\cdot], u, v)$ returns true.

18

Remark 2. Note that if $|\mathcal{R}| < 4$, then $\mathsf{Mergeable}(\mathcal{R}, \mathcal{L}[\cdot], u, v)$ returns true for all pairs $\{u, v\} \subseteq \mathcal{R}$.

Proof. Suppose the first condition holds, and let h be the common neighbor. Since $\mathcal{T}[u]$ is a leaf component relative to $\mathcal{C}$, the (undirected) path from any node u' in $\mathcal{T}[u]$ to another node w not in $\mathcal{T}[u]$ must pass through h. Similarly, the (undirected) path from any node v' in $\mathcal{T}[v]$ to another node w not in $\mathcal{T}[v]$ must pass through h. Therefore, each choice of $\{u_1, v_1\} \subseteq \mathcal{R} \setminus \{u, v\}$ and $(x, y, x_1, y_1) \in \mathcal{L}[u] \times \mathcal{L}[v] \times \mathcal{L}[u_1] \times \mathcal{L}[v_1]$ induces one of the following topologies,

upon which, by Lemma 2, the quartet test returns either $\{\{x, y\}, \{x_1, y_1\}\}$ or $\bot$.

Now instead suppose the second condition holds. Without loss of generality, assume $\mathcal{T}[u]$ is a leaf component relative to $\mathcal{C}$, which then implies that the (undirected) path from any node u' in $\mathcal{T}[u]$ to another node w not in $\mathcal{T}[u]$ must pass through v. Moreover, since $\mathcal{T}[v]$ is rooted at v, the (undirected) path from any node v' in $\mathcal{T}[v]$ to another node w not in $\mathcal{T}[v]$ must pass through v. If $\mathcal{T}[v]$ is also a leaf component, then it must be that $\mathcal{R} = \{u, v\}$, in which case $\mathcal{R} \setminus \{u, v\} = \emptyset$. If $\mathcal{T}[v]$ is not a leaf component, then each choice of $\{u_1, v_1\} \subseteq \mathcal{R} \setminus \{u, v\}$ and $(x, y, x_1, y_1) \in \mathcal{L}[u] \times \mathcal{L}[v] \times \mathcal{L}[u_1] \times \mathcal{L}[v_1]$ induces one of the following topologies,

upon which, by Lemma 2, the quartet test returns either $\{\{x, y\}, \{x_1, y_1\}\}$ or $\bot$. $\qquad\square$

Lemma 12 (Un-mergeable pairs). *Let $\mathcal{R} \subseteq \mathcal{V}_{\mathbb{T}}$. Let $\mathcal{C} := \{\mathcal{T}[r] : r \in \mathcal{R}\}$ be a collection of disjoint rooted subtrees, with r being the root of $\mathcal{T}[r]$, such that their leaf sets $\{\mathcal{L}[r] : r \in \mathcal{R}\}$ partition $\mathcal{V}_{\mathrm{obs}}$. Further, suppose the pair $\{u, v\} \subseteq \mathcal{R}$ are such that all of the following conditions hold.*

1. *There exists $(x, y) \in \mathcal{L}[u] \times \mathcal{L}[v]$ such that $\sigma_k(\widehat{\Sigma}_{x,y}) \geq \theta$.*

2. *$\{u, v\}$ do not share a common neighbor in $\mathbb{T}$, or at least one of $\mathcal{T}[u]$ and $\mathcal{T}[v]$ is not a leaf component relative to $\mathcal{C}$.*

3. *$\{u, v\}$ are not neighbors in $\mathbb{T}$, or neither $\mathcal{T}[u]$ nor $\mathcal{T}[v]$ is a leaf component relative to $\mathcal{C}$.*

Then there exists a pair $\{u_1, v_1\} \subseteq \mathcal{R} \setminus \{u, v\}$ and $(x_1, y_1) \in \mathcal{L}[u_1] \times \mathcal{L}[v_1]$ such that $\mathsf{SpectralQuartetTest}(\{x, y, x_1, y_1\})$ returns $\{\{x, x_1\}, \{y, y_1\}\}$. This implies that $\mathsf{Mergeable}(\mathcal{R}, \mathcal{L}[\cdot], u, v)$ returns false.

Proof. First, take $(x, y) \in \mathcal{L}[u] \times \mathcal{L}[v]$ such that $\sigma_k(\widehat{\Sigma}_{x,y}) \geq \theta$. By Lemma 8, $\sigma_k(\Sigma_{x,y}) \geq (1-\varepsilon)\theta$. Lemma 9 implies that the nodes of $\mathcal{ST}[\mathcal{C}]$ are $\mathcal{C} \cup \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$, and that each leaf in $\mathcal{ST}[\mathcal{C}]$ is a subtree $\mathcal{T}[u] \in \mathcal{C}$. The second and third conditions of the lemma on $\{u, v\}$ imply that at least one of the following cases holds.

(i) Neither $\mathcal{T}[u]$ nor $\mathcal{T}[v]$ is a leaf component relative to $\mathcal{C}$.

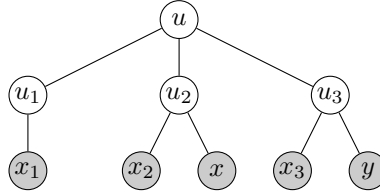
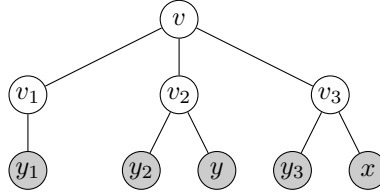
(ii) u and v are not neighbors and do not share a common neighbor.

(iii) u and v are not neighbors, and one of $\mathcal{T}[u]$ and $\mathcal{T}[v]$ is not a leaf component relative to $\mathcal{C}$.

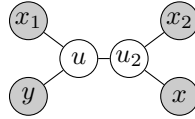
Suppose (i) holds. Then each of $\mathcal{T}[u]$ and $\mathcal{T}[v]$ have degree ≥ 2 in $\mathcal{ST}[\mathcal{C}]$. Note that neither u nor v are leaves in $\mathbb{T}$. Moreover, there exists $\{u_1, v_1\} \subseteq (\mathcal{R} \setminus \{u, v\}) \cup \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$ such that u_1 is adjacent to u

19

in $\mathbb{T}$, v_1 is adjacent to v in $\mathbb{T}$, and the (undirected) path from u_1 to v_1 in $\mathbb{T}$ intersects the (undirected) path from u to v in $\mathbb{T}$.

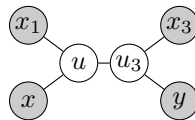
Since u is not a leaf, it has at least three neighbors by assumption, and thus there exist three sub-trees $\{\mathcal{T}_{u,1}, \mathcal{T}_{u,2}, \mathcal{T}_{u,3}\} \subseteq \mathcal{F}_u$ such that u_1 is the root of $\mathcal{T}_{u,1}$, $x \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u,2}]$ and $y \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u,3}]$. Moreover, by Condition 4, there exist $x_1 \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u,1}]$, $x_2 \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u,2}]$, and $x_3 \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u,3}]$ such that $\sigma_k(\mathbb{E}[x_i x_j^\top]) \geq \gamma_{\min}$ for all $\{i,j\} \subset \{1,2,3\}$. Note that it is possible to have $x_2 = x$ and $x_3 = y$. Let u_2 denote the node in $\mathcal{T}_{u,2}$ at which the (undirected) paths $x \rightsquigarrow u$ and $x_2 \rightsquigarrow u$ intersect (if $x_2 = x$, then let u_2 be the root of $\mathcal{T}_{u,2}$); similarly, let u_3 denote the node in $\mathcal{T}_{u,2}$ at which the (undirected) paths $y \rightsquigarrow u$ and $x_3 \rightsquigarrow u$ intersect (if $x_3 = y$, then let u_3 be the root of $\mathcal{T}_{u,3}$). The induced (undirected) topology over these nodes is shown below.

A completely analogous argument can be applied relative to v instead of u, giving the following.

Claim 1. *The following lower bounds hold.*

$$\min \left\{ \sigma_k(\Sigma_{x_1,x}),\ \sigma_k(\Sigma_{x_1,y}),\ \sigma_k(\Sigma_{y_1,y}),\ \sigma_k(\Sigma_{y_1,x}) \right\} \geq \frac{\gamma_{\min} \cdot (1-\varepsilon)\theta}{\gamma_{\max}} = \varsigma. \tag{9}$$

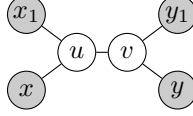
Proof. We just show the inequalities for $\sigma_k(\mathbb{E}[x_1 x^\top])$ and $\sigma_k(\mathbb{E}[x_1 y^\top])$; the other two are analogous. If $x_2 = x$, then $\sigma_k(\mathbb{E}[x_1 x^\top]) = \sigma_k(\mathbb{E}[x_1 x_2^\top]) \geq \gamma_{\min} \geq \varsigma$. If $x_2 \neq x$, then we have the following induced (undirected) topology.

Therefore, by Lemma 10,

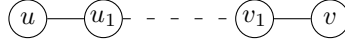
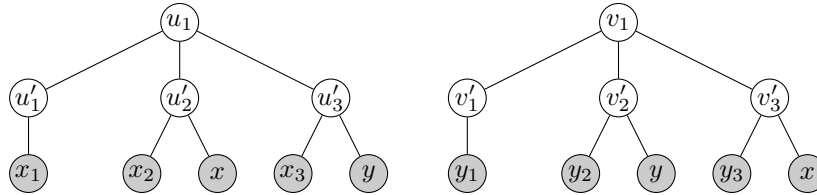
$$\sigma_k(\mathbb{E}[x_1 x^\top]) \geq \frac{\sigma_k(\mathbb{E}[x_1 x_2^\top]) \cdot \sigma_k(\mathbb{E}[yx^\top])}{\sigma_1(\mathbb{E}[yx_2^\top])} \geq \frac{\gamma_{\min} \cdot (1-\varepsilon)\theta}{\gamma_{\max}} = \varsigma.$$

This gives the first claimed inequality; now we show the second. If $x_3 = y$, then $\sigma_k(\mathbb{E}[x_1 y^\top]) = \sigma_k(\mathbb{E}[x_1 x_3^\top]) \geq \gamma_{\min} \geq \varsigma$. If $x_3 \neq y$, then we have the following induced (undirected) topology.

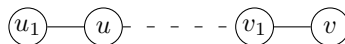
20

Again, by Lemma 10,

$$\sigma_k(\mathbb{E}[x_1 y^\top]) \geq \frac{\sigma_k(\mathbb{E}[x_1 x_3^\top]) \cdot \sigma_k(\mathbb{E}[xy^\top])}{\sigma_1(\mathbb{E}[xx_3^\top])} \geq \frac{\gamma_{\min} \cdot (1-\varepsilon)\theta}{\gamma_{\max}} = \varsigma. \qquad \square$$

Claim 1, Lemma 4, and the sample size requirement of Theorem 1 (as per (7)) imply that the spectral quartet test on $\{x, x_1, y, y_1\}$ returns the correct pairing. Since the induced (undirected) topology is

the correct pairing is $\{\{x, x_1\}, \{y, y_1\}\}$. Because the leaf sets $\{\mathcal{L}[r] : r \in \mathcal{R}\}$ partition $\mathcal{V}_{\mathrm{obs}}$, and because $x_1 \notin \mathcal{L}[u]$ and $y_1 \notin \mathcal{L}[v]$, there exists $\{u', v'\} \subseteq \mathcal{R} \setminus \{u, v\}$ such that $x_1 \in \mathcal{L}[u']$ and $y_1 \in \mathcal{L}[v']$. This proves the lemma in this case.

Now instead suppose (ii) holds. Since $\mathbb{T}$ is connected, and $\mathcal{T}[u]$ and $\mathcal{T}[v]$ are respectively rooted at u and v, there must exist a pair $\{u_1, v_1\} \subset (\mathcal{R} \setminus \{u, v\}) \cup \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$ such that neither u_1 nor v_1 are leaves in $\mathbb{T}$, u_1 is adjacent to u in $\mathbb{T}$, v_1 is adjacent to v in $\mathbb{T}$, and the (undirected) path from u to v in $\mathbb{T}$ passes through the path from u_1 to v_1.

An argument analogous to that in case (i) applies to prove the lemma in this case; we provide a brief sketch below. Because u_1 is not a leaf, there exists three subtrees $\{\mathcal{T}_{u_1,1}, \mathcal{T}_{u_1,2}, \mathcal{T}_{u_1,3}\} \subseteq \mathcal{F}_{u_1}$ such that u is the root of $\mathcal{T}_{u_1,2}$ (so $x \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u_1,2}]$) and $y \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u_1,3}]$. Moreover, there exist $x_1 \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u_1,1}]$, $x_2 \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u_1,2}]$, and $x_3 \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u_1,3}]$ such that $\sigma_k(\mathbb{E}[x_i x_j^\top]) \geq \gamma_{\min}$ for all $\{i, j\} \subset \{1, 2, 3\}$ (it is possible to have $x_2 = x$ and $x_3 = y$). Let u_1' denote the root of $\mathcal{T}_{u_1,1}$, u_2' denote the node in $\mathcal{T}_{u_1,2}$ at which the (undirected) paths $x \rightsquigarrow u_1$ and $x_2 \rightsquigarrow u_1$ intersect (if $x_2 = x$, then let $u_2' = u$, which is the root of $\mathcal{T}_{u_1,2}$), and u_3 denote the node in $\mathcal{T}_{u_1,2}$ at which the (undirected) paths $y \rightsquigarrow u_1$ and $x_3 \rightsquigarrow u_1$ intersect (if $x_3 = y$, then let u_3 be the root of $\mathcal{T}_{u_1,3}$). An analogous argument applies relative to v_1 instead of u_1; the induced (undirected) topologies are given below.

Using the arguments in Claim 1, it can be shown that the inequalities in (9) hold in this case, so by Lemma 4, the quartet test on $\{x, x_1, y, y_1\}$ returns $\{\{x, x_1\}, \{y, y_1\}\}$. Because the leaf sets $\{\mathcal{L}[r] : r \in \mathcal{R}\}$ partition $\mathcal{V}_{\mathrm{obs}}$, and because $x_1 \notin \mathcal{L}[u] = \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u_1,2}]$ and $y_1 \notin \mathcal{L}[v] = \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{v_1,2}]$, there exists $\{u', v'\} \subseteq \mathcal{R} \setminus \{u, v\}$ such that $x_1 \in \mathcal{L}[u']$ and $y_1 \in \mathcal{L}[v']$. This proves the lemma in this case.

Finally, suppose (iii) holds. Without loss of generality, assume $\mathcal{T}[u]$ is not a leaf component relative to $\mathcal{C}$. Since $\mathbb{T}$ is connected, and $\mathcal{T}[u]$ and $\mathcal{T}[v]$ are respectively rooted at u and v, there must exist $v_1 \in (\mathcal{R} \setminus \{u, v\}) \cup \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$ such that v_1 is not a leaf in $\mathbb{T}$, v_1 is adjacent to v in $\mathbb{T}$, and the (undirected) path from u to v in $\mathbb{T}$ passes through v_1. Moreover, since $\mathcal{T}[u]$ is not a leaf component relative to $\mathcal{C}$, it has degree ≥ 2 in $\mathcal{ST}[\mathcal{C}]$. Note that u is not a leaf in $\mathbb{T}$, and moreover, there exists $u_1 \in (\mathcal{R} \setminus \{u, v\}) \cup \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$ such that u_1 is adjacent to u in $\mathbb{T}$, and u_1 is not on the (undirected) path from u to v.

21

Again, an argument analogous to that in case (i) applies now to prove the lemma in this case. $\qquad\square$

Finally, we give a lemma which analyzes the while-loop of Algorithm 2 and consequently implies Theorem 1.

Lemma 13 (Loop invariants). *The following invariants concerning the state of the objects* $(\mathcal{R}, \mathcal{T}[\cdot], \mathcal{L}[\cdot])$ *hold before the while-loop in Algorithm 2, and after each iteration of the while-loop.*

1. $\mathcal{R} \subseteq \mathcal{V}_{\mathbb{T}}$, *and for each* $u \in \mathcal{R}$, $\mathcal{T}[u]$ *is a subtree of* $\mathbb{T}$ *rooted at* u. *Moreover, the rooted subtree* $\mathcal{T}[v]$ *is already defined by Algorithm 2 for every node* v *appearing in* $\mathcal{T}[u]$ *for some* $u \in \mathcal{R}$. *Finally, for each* $u \in \mathcal{R}$, *the subtree* $\mathcal{T}[u]$ *is formed by joining the subtrees* $\mathcal{T}[v]$ *corresponding to children* v *of* u *in* $\mathcal{T}[u]$ *via edges* $\{u, v\}$.

2. *The subtrees in* $\mathcal{C} := \{\mathcal{T}[u] : u \in \mathcal{R}\}$ *are disjoint, and the leaf sets* $\{\mathcal{L}[u] : u \in \mathcal{R}\}$ *partition* $\mathcal{V}_{\mathrm{obs}}$.

Moreover, no iteration of the while-loop terminates in failure.

Before proving Lemma 13, we show how it implies Theorem 1. Initially, $|\mathcal{R}| = n$, and each iteration of the while-loop decreases the cardinality of $\mathcal{R}$ by one, so there are a total of $n - 1$ iterations of the while-loop. By Lemma 13, the final iteration results in a set $\mathcal{R} = \{h\}$ such that $\widehat{\mathbb{T}} = \mathcal{T}[h]$ is a subtree of $\mathbb{T}$ rooted at h, and $\mathcal{L}[h] = \mathcal{V}_{\mathrm{obs}}$. This implies that $\widehat{\mathbb{T}}$ has the same (undirected) structure as $\mathbb{T}$, as required. This completes the proof of Theorem 1.

Proof of Lemma 13. The loop invariants clearly hold before the while-loop with the initial settings of $\mathcal{R} = \mathcal{V}_{\mathrm{obs}}$, $\mathcal{T}[x] = $ rooted single-node tree x, and $\mathcal{L}[x] = \{x\}$ for all $x \in \mathcal{R}$. So assume as the inductive hypothesis that the loop invariants hold at the start of a particular iteration (in which $|\mathcal{R}| > 1$). It remains to prove that the iteration does not terminate in failure, and that the loop invariants hold at the end of the iteration. Let $\mathcal{R}$, $\mathcal{T}[\cdot]$, and $\mathcal{L}[\cdot]$ be in their state at the beginning of the iteration.

Because the second loop invariant holds, Lemma 9 implies that the nodes of $\mathcal{ST}[\mathcal{C}]$ are $\mathcal{C} \cup \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$, and that each leaf in $\mathcal{ST}[\mathcal{C}]$ is a subtree $\mathcal{T}[u] \in \mathcal{C}$ (so we may refer to the leaves of $\mathcal{ST}[\mathcal{C}]$ as leaf components).

Claim 2. *If* $|\mathcal{R}| > 1$, *then there exists a pair* $\{u, v\} \subseteq \mathcal{R}$ *such that the following hold.*

1. *Either* u *and* v *are neighbors in* $\mathbb{T}$, *and at least one of* $\mathcal{T}[u]$ *or* $\mathcal{T}[v]$ *is a leaf component relative to* $\mathcal{C}$; *or* u *and* v *share a common neighbor in* $\mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$, *and both* $\mathcal{T}[u]$ *and* $\mathcal{T}[v]$ *are leaf components relative to* $\mathcal{C}$.

2. Mergeable$(\mathcal{R}, \mathcal{L}[\cdot], u, v) = true$.

3. $\max\{\sigma_k(\widehat{\Sigma}_{x,y}) : (x, y) \in \mathcal{L}[u] \times \mathcal{L}[v]\} \geq \theta$.

Proof. Suppose there are no pairs $\{u, v\} \subseteq \mathcal{C}$ such that u and v are neighbors in $\mathbb{T}$ and at least one of $\mathcal{T}[u]$ and $\mathcal{T}[v]$ is a leaf component relative to $\mathcal{C}$. Then each leaf component must be adjacent to some $h \in \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$ in $\mathcal{ST}[\mathcal{C}]$. Consider the tree $\mathcal{ST}'$ obtained from $\mathcal{ST}[\mathcal{C}]$ by removing all the leaf components in $\mathcal{ST}[\mathcal{C}]$. The leaves of $\mathcal{ST}'$ must be among the $h \in \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$ that were adjacent to the leaf components in $\mathcal{ST}[\mathcal{C}]$. Fix such a leaf h in $\mathcal{ST}'$, and observe that it has degree one in $\mathcal{ST}'$. By assumption, no node in $\mathbb{T}$ has degree two, so h must have been connected to at least two leaf components in $\mathcal{ST}[\mathcal{C}]$, say $\mathcal{T}[u]$ and $\mathcal{T}[v]$. The node h is therefore a common neighbor of u and v. This proves the existence of a pair $\{u, v\} \subseteq \mathcal{R}$ satisfying the first required property.

Fix the pair $\{u, v\}$ specified above. By Lemma 11, Mergeable$(\mathcal{R}, \mathcal{L}[\cdot], u, v)$ returns true, so $\{u, v\}$ satisfies the second required property.

To show the final required property, we consider two cases. Suppose first that u and v are neighbors, and that $\mathcal{T}[u]$ is a leaf component relative to $\mathcal{C}$. Note that u and v cannot both be leaves in $\mathbb{T}$. If v is not a leaf, then there exists subtrees $\mathcal{T}_{v,1}$ and $\mathcal{T}_{v,2}$ in $\mathcal{F}_v$ such that $\mathcal{T}_{v,1} = \mathcal{T}[u]$ (because $\mathcal{T}[u]$ is a leaf component) and $\mathcal{T}_{v,2} = \mathcal{T}[v']$ for some child v' of v in $\mathcal{T}[v]$ (by the first loop invariant). By

22

Condition 4, there exists $x \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{v,1}] = \mathcal{L}[u]$ and $y \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{v,2}] \subseteq \mathcal{L}[v]$ such that $\sigma_k(\Sigma_{x,y}) \geq \gamma_{\min} = (1+\varepsilon)\theta$; by Lemma 8, $\sigma_k(\widehat{\Sigma}_{x,y}) \geq \theta$. If v is a leaf but u is not, then there exists subtrees $\mathcal{T}_{u,1}$ and $\mathcal{T}_{u,2}$ in $\mathcal{F}_u$ such that $\mathcal{T}_{u,1} = v$ and $\mathcal{T}_{u,2} = \mathcal{T}[u']$ for some child u' of u in $\mathcal{T}[u]$ (by the first loop invariant). So by Condition 4, there $y \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}_{u,2}] \subseteq \mathcal{L}[u]$ such that $\sigma_k(\Sigma_{v,y}) \geq \gamma_{\min} = (1+\varepsilon)\theta$; by Lemma 8, $\sigma_k(\widehat{\Sigma}_{v,y}) \geq \theta$. Now instead suppose that u and v share a common neighbor h, and that both $\mathcal{T}[u]$ and $\mathcal{T}[v]$ are leaf components relative to $\mathcal{C}$. This latter fact implies that $\{\mathcal{T}[u], \mathcal{T}[v]\} \subset \mathcal{F}_h$, so Condition 4 implies that there exists $x \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}[u]] = \mathcal{L}[u]$ and $y \in \mathcal{V}_{\mathrm{obs}}[\mathcal{T}[v]] = \mathcal{L}[v]$ such that $\sigma_k(\Sigma_{x,y}) \geq \gamma_{\min} = (1+\varepsilon)\theta$. By Lemma 8, $\sigma_k(\widehat{\Sigma}_{x,y}) \geq \theta$. $\qquad\square$

Claim 3. *Consider any pair $\{u, v\} \subseteq \mathcal{R}$ such that $\max\{\sigma_k(\widehat{\Sigma}_{x,y}) : (x, y) \in \mathcal{L}[u] \times \mathcal{L}[v]\} \geq \theta$. If the first property from Claim 2 fails to hold for $\{u, v\}$, then* Mergeable$(\mathcal{R}, \mathcal{L}[\cdot], u, v) = false.$

Proof. This follows immediately from Lemma 12. $\qquad\square$

Taken together, Claims 2 and 3 imply that the pair $\{u, v\} \subseteq \mathcal{R}$ selected by the first step in the while-loop indeed exists (so the iteration does not terminate in failure) and satisfies the properties in Claim 2.

Now we consider the second step of the while-loop, which is the call to the subroutine Relationship.

Claim 4. *Suppose a pair $\{u, v\}$ satisfies the properties in Claim 2. Then* Relationship$(\mathcal{R}, \mathcal{L}[\cdot], \mathcal{T}[\cdot], u, v)$ *returns the correct relationship for u and v. Specifically:*

1. *If u and v share a common neighbor in $\mathbb{T}$ (and both are leaf components relative to $\mathcal{C}$), then "siblings" is returned.*

2. *If u and v are neighbors in $\mathbb{T}$ and $\mathcal{T}[v]$ is a leaf component relative to $\mathcal{C}$ but $\mathcal{T}[u]$ is not, then "u is parent of v" is returned.*

3. *If u and v are neighbors in $\mathbb{T}$ and $\mathcal{T}[u]$ is a leaf component relative to $\mathcal{C}$ but $\mathcal{T}[v]$ is not, then "v is parent of u" is returned.*

4. *If u and v are neighbors in $\mathbb{T}$ and both $\mathcal{T}[u]$ and $\mathcal{T}[v]$ are leaf components relative to $\mathcal{C}$, and u is a leaf in $\mathbb{T}$ but v is not, then "v is parent of u" is returned.*

5. *If u and v are neighbors in $\mathbb{T}$ and both $\mathcal{T}[u]$ and $\mathcal{T}[v]$ are leaf components relative to $\mathcal{C}$, and v is a leaf in $\mathbb{T}$ but u is not, then "u is parent of v" is returned.*

6. *If u and v are neighbors in $\mathbb{T}$ and both $\mathcal{T}[u]$ and $\mathcal{T}[v]$ are leaf components relative to $\mathcal{C}$, and neither u nor v are leaves in $\mathbb{T}$, then "u is parent of v" is returned.*

Proof. Fix the pair $(x, y) \in \mathcal{L}[u] \times \mathcal{L}[v]$ guaranteed by the third property of Claim 2 such that $\sigma_k(\widehat{\Sigma}_{x,y}) \geq \theta$. Now we consider the possible relationships between u and v.

Suppose u and v share a common neighbor $h \in \mathcal{V}_{\mathrm{hid}}[\mathcal{C}]$ in $\mathbb{T}$, and that both $\mathcal{T}[u]$ and $\mathcal{T}[v]$ are leaf components relative to $\mathcal{C}$. We need to show that the subroutine Relationship asserts both "$u \not\to v$" and "$v \not\to u$". To show that "$u \not\to v$" is asserted, we assume u is not a leaf (otherwise "$u \not\to v$" is immediately asserted and we're done), let $\{u_1, \dots, u_q\}$ be the children of u in $\mathcal{T}[u]$, and take $\mathcal{R}[u]$ as defined in Relationship. By the first loop invariant, the subtrees in $\mathcal{C}[u]$ are disjoint, and the leaf sets $\{\mathcal{L}[r] : r \in \mathcal{R}[u]\}$ partition $\mathcal{V}_{\mathrm{obs}}$. In particular, $x \in \mathcal{L}[u_i]$ for some $i \in \{1, \dots, q\}$. Since u_i and v are not neighbors, and do not share a common neighbor. Therefore, by Lemma 12, Mergeable$(\mathcal{R}[u], \mathcal{L}[\cdot], u_i, v) = $ false, so "$u \not\to v$" is asserted. A similar argument implies that "$v \not\to u$" is asserted. Since both "$u \not\to v$" and "$v \not\to u$" are asserted, the subroutine returns "siblings".

Now instead suppose u and v are neighbors. First, suppose $\mathcal{T}[u]$ is a leaf component relative to $\mathcal{C}$. We claim that if v is not a leaf, then "$v \not\to u$" is not asserted. Let $\{v_1, \dots, v_q\}$ be the children of v in $\mathcal{T}[v]$, and take $\mathcal{R}[v] = \{u, v_1, \dots, v_q\}$ as defined in Relationship. By the first loop invariant, the subtrees in $\mathcal{C}[v]$ are disjoint, and the leaf sets $\{\mathcal{L}[r] : r \in \mathcal{R}[v]\}$ partition $\mathcal{V}_{\mathrm{obs}}$. By Lemma 14, $\mathcal{T}[u]$ and $\mathcal{T}[v_i]$ are leaf components relative to $\mathcal{C}[v]$ for each $i \in \{1, \dots, q\}$. For each $i \in \{1, \dots, q\}$, $\{u, v_i\}$ share v as a common neighbor, and $\mathcal{T}[u]$ and $\mathcal{T}[v_i]$ are both leaf components relative to $\mathcal{C}[v]$.

23

Therefore by Lemma 11, $\text{Mergeable}(\mathcal{R}[v], \mathcal{L}[\cdot], u, v_i) = \text{true}$ for all $i \in \{1, \ldots, q\}$, so "$v \nrightarrow u$" is not asserted.

Suppose $\mathcal{T}[u]$ is a leaf component relative to $\mathcal{C}$ but $\mathcal{T}[v]$ is not. By Lemma 9, v is not a leaf in $\mathbb{T}$, so as argued above, "$v \nrightarrow u$" is not asserted. It remains to show that "$u \nrightarrow v$" is asserted. Assume u is not a leaf (or else $u \nrightarrow v$ is immediately asserted and we're done), let $\{u_1, \ldots, u_q\}$ be the children of u in $\mathcal{T}[u]$, and take $\mathcal{R}[u]$ as defined in Relationship. By the first loop invariant, the subtrees in $\mathcal{C}[u]$ are disjoint, and the leaf sets $\{\mathcal{L}[r] : r \in \mathcal{R}[u]\}$ partition $\mathcal{V}_{\text{obs}}$. In particular, $x \in \mathcal{L}[u_i]$ for some $i \in \{1, \ldots, q\}$. By Lemma 14, $\mathcal{T}[v]$ is not a leaf component relative to $\mathcal{C}[u]$. Moreover, u_i and v are not neighbors. Therefore by Lemma 12, $\text{Mergeable}(\mathcal{R}[u], \mathcal{L}[\cdot], u_i, v) = \text{false}$, so "$u \nrightarrow v$" is asserted. Since "$v \nrightarrow u$" is not asserted but "$u \nrightarrow v$" is asserted, the subroutine returns "$v \rightarrow u$". An analogous argument shows that if $\mathcal{T}[v]$ is a leaf component relative to $\mathcal{C}$ but $\mathcal{T}[u]$ is not, then the subroutine returns "$u \rightarrow v$".

Now suppose both $\mathcal{T}[u]$ and $\mathcal{T}[v]$ are leaf components relative to $\mathcal{C}$. By assumption, leaves in $\mathbb{T}$ are only adjacent to non-leaves, so it cannot be that both u and v are leaves. Therefore at least one of u and v is not a leaf in $\mathbb{T}$. Without loss of generality, say v is not a leaf in $\mathbb{T}$. Then as argued above, "$v \nrightarrow u$" is not asserted. If u is a leaf, then "$u \nrightarrow v$" is asserted, so the subroutine returns "$v \rightarrow u$". If u is not a leaf, then by symmetry, "$u \nrightarrow v$" is not asserted. Therefore the subroutine returns "$u \rightarrow v$". $\qquad\square$

Claim 4 implies that the remaining steps in the while-loop after the call to Relationship preserve the two loop invariants, simply by construction. $\qquad\square$

There is one last lemma used in the proof of Lemma 13.

Lemma 14 (Leaf components). *Suppose the invariants in Lemma 13 are satisfied. Then for each $u \in \mathcal{R}$ such that u is not a leaf in $\mathbb{T}$, the leaf components relative to the collection*

$$\mathcal{C}[u] := (\mathcal{C} \setminus \{\mathcal{T}[u]\}) \cup \{\mathcal{T}[v] : v \text{ is a child of } u \text{ in } \mathcal{T}[u]\}$$

are

$$\{\mathcal{T}[r] : r \neq u \;\wedge\; \mathcal{T}[r] \text{ is a leaf component relative to } \mathcal{C}\} \cup \{\mathcal{T}[r] : r \text{ is a child of } u \text{ in } \mathcal{T}[u]\}.$$

Proof. Pick any $u \in \mathcal{R}$ such that u is not a leaf in $\mathbb{T}$. Let $\{v_1, \ldots, v_q\}$ be the children of u in $\mathcal{T}[u]$. By the first loop invariant, each v_i is the root of a subtree $\mathcal{T}[v_i]$. This implies that the subtrees $\{\mathcal{T}[v_1], \ldots, \mathcal{T}[v_q]\}$ are disjoint and $\{\mathcal{L}[v_1], \ldots, \mathcal{L}[v_q]\}$ partition $\mathcal{L}[u]$. Therefore $\mathcal{ST}[\mathcal{C}[u]]$ is the same as $\mathcal{ST}[\mathcal{C}]$ except with the following changes.

1. $\mathcal{T}[u]$ is replaced with u.

2. For each i, $\mathcal{T}[v_i]$ is added with the edge $\{u, v_i\}$.

This means that each $\mathcal{T}[v_i]$ has degree one in $\mathcal{ST}[\mathcal{C}[u]]$ and therefore is a leaf component relative to $\mathcal{C}[u]$. $\qquad\square$

24