Statistics and Its Interface Volume 10 (2017) 677–698

Spectral methods for learning discrete latent tree models[*]

Xiaofei Wang, Jianhua Guo[†], Lizhu Hao, and Nevin L. Zhang

We consider the problems of structure learning and parameter estimation for discrete latent tree models. For structure learning, we introduce a concept of generalized information distance between variables based on singular values of probability matrices, and use it to build a bottom-up algorithm for structure recovery. The algorithm is proved to be consistent. Moreover, a finite sample bound is given for exact structure recovery. For parameter estimation, we suggest a novel matrix decomposition algorithm for the case when every latent variable has two states. Unlike the expectation-maximization (EM) algorithm, our algorithm can avoid trapping into a local optima. Moreover, it is proved to be consistent and a finite sample bound is also given for parameter estimation.

In both structural learning and parameter estimation, empirical results were provided to support our theoretical results. In applications to real data, we analyzed the Changchun mayor hotline data, where the underlying structures were detected for Chinese words. We demonstrated that the proposed method is efficient for discovering hierarchical structures and latent information.

AMS 2000 subject classifications: Primary 62H05; secondary 62H12.

Keywords and phrases: Latent variables, Parameter estimation, Spectral distance, Structural learning.

1. INTRODUCTION

This work is motivated by mining and analyzing various topics for the Chinese text data. A typical way of modeling those potential topics is to introduce latent variables for explaining the mechanism that words are generated from topics. There are already numerous probabilistic models with latent variables from both Statistics [3] and Machine Learning [18]. This paper is concerned with tree-structure graphical probabilistic models where all the leaf nodes are observed while the internal nodes are latent. They are referred to as latent tree models in the literature. Special latent tree models such as phylogenetic trees [10] have been studied for decades. General latent tree models were first investigated by Zhang [25] (2004), where they were called hierarchical latent class models. In Machine Learning, they are used as a tool for latent structure discovery [7], density estimation [23], and multidimensional clustering [16].

Previous algorithms for learning the structures of latent tree models can be roughly divided into two groups. Algorithms in the first group [25, 26, 6] aim at searching the latent tree models that are optimal according to a scoring metric. Those algorithms are computationally expensive and do not have consistency guarantees. Algorithms in the second group focus on the phylogenetic tree reconstruction [11, 10, 8, 14, 17, 21], where the structures of models mostly confine each internal node to having the same degree except for the root. Inspired by works on the phylogenetic tree reconstruction, Choi et al. [7] (2011) propose several consistent bottom-up structure learning algorithms based on the information distance. Those algorithms require that all the variables have the same number of states. In this paper, we generalize Choi's work by allowing observed variables to have different numbers of states. A key novelty of our work is the use of a concept of generalized information distance between variables that is defined using singular values of probability matrices. Our new algorithm is proved to be consistent. Moreover, a finite sample bound is given for exact structure recovery with high probability.

For estimating the parameters of latent tree models, previous works [25] rely on the expectation-maximization (EM) algorithm. EM suffers from the possibility of being trapped in local optimum, and thus no consistency guarantees can be provided for those works. Chang [5] (1996) propose a matrix decomposition method for parameter estimation on phylogenetic trees that does not have the local optimum problem. However, it requires all variables (observed and latent) to have the same number of states, and encounters parameter unidentifiability when matrix decomposition yields equal diagonal elements. In this paper, we relax the restriction of Chang's method on the numbers of states for observed variables, and fix the parameter unidentifiability problem in the special case when all latent variables have two states. We propose a consistent algorithm, which has a finite sample size for parameter estimation with high probability.

This paper is organized as follows. In Section 2, we introduce some basic notations and assumptions used in this paper. In Section 3, we introduce the notion of general information distance based on the product of singular values,

[*] This work is supported in part by NSF of China (Grants No. 11401092, 11631003).

[†] Corresponding Author.

and provide a recursive bottom-up algorithm for learning latent tree structures. In Section 4, we discuss parameter estimation methods for discrete latent tree models, and propose a new parameter estimation algorithm. In Section 5, we report empirical results and real data applications to demonstrate the performance of our structural learning and parameter estimating algorithms. In Section 6, we give a brief summary of our work. Finally, in the Appendix, we provide the proofs for the theorems presented in our paper.

2. PRELIMINARIES

Let $G = (W, E)$ be a simple undirected graph, where W is the set of vertices and E is the set of undirected edges. The number of edges incident to a vertex x in G is called the degree of x, which is denoted by $d(x)$. A set L of distinct vertices $[x_0, x_1, \cdots, x_m]$ is referred to as a path of length m in G between x_0 and x_m if $(x_{i-1}, x_i) \in E$ for all $i = 1, \cdots, m$. Furthermore, if $(x_0, x_m) \in E$, we also refer to L as a cycle in G. If for any two vertices of W there is a path in G between them, we refer to G as a connected graph. Two disjoint vertex subsets A, B are separated by a vertex subset S in G if for any $x \in A$ and any $y \in B$ every path in G between x and y contains a vertex in S. If (W, E) is a connected simple graph with no cycles, it is referred to as a tree and denoted as T. If a vertex x of T with $d(x) = 1$, it is referred to as a leaf in T. If two leaves x and y in T are adjacent to the same vertex, we refer to $\{x, y\}$ as a sibling pair in T. The diameter of a tree T is the number of nodes on the longest path between two leaves in the tree, which we denote as $diam(T)$.

If the vertex set W of G represents a set of random variables, a graphical model on G is a family of probability distributions where A and B are conditionally independent given S when two disjoint vertex subsets A and B are separated by a vertex subset S in G. Let $T = (W, E)$ be a tree. A graphical model $\mathcal{T}$ on T is referred to as a latent tree model if the leaves of T are observed variables and the internal nodes are latent variables. Furthermore, if we limit $\mathcal{T}$ to a multinomial distribution family, we refer to it as a discrete latent tree model. We denote the set of observed variables as V and the set of latent variables as H. Thus, the vertex set W of T comprises V and H, i.e., $W = V \bigcup H$. If the path between two observed variables v_i, v_j in T contains h, then we refer to v_i, v_j as bifurcation variables of h.

If we set a root variable for T, we can obtain a directed tree $\vec{T} = (W, \vec{E})$ and the directions are fixed from the root to the leaves. The element $x \to y$ in $\vec{E}$ represents a direct edge from x to y. We refer to y as a child variable of x and denote the relation as $y \in ch(x)$. An ordered set L of distinct vertices $x_0 \to x_1 \to \cdots \to x_m$ is referred to as a directed path of length m in $\vec{T}$ from x_0 to x_m if $x_{i-1} \to x_i \in \vec{E}$ for all $i = 1, \cdots, m$. If there is a directed path in $\vec{T}$ from h to an observed variable v_i, then we refer to v_i as a directed bifurcation variable of h in $\vec{T}$.

For any random variable $z \in W$, let d_z denote the number of states of the variable z. For two random variables $x, y \in W$, the joint probability matrix between them is defined by $P_{xy} = (Pr(x = x_i, y = y_j))_{1 \le i \le d_x, 1 \le j \le d_y}$. When $x = y$, as a special case of the joint probability matrix, the marginal probability matrix P_{xx} is a diagonal matrix with diagonal element $Pr(x = x_i)$, where $i = 1, \cdots, d_x$. For two distinct random variables $x, y \in W$, the conditional probability matrix from y to x is defined by $P_{x|y} = (Pr(x = x_i | y = y_j))_{1 \le i \le d_x, 1 \le j \le d_y}$. For any variable $x \in W$, we define a probability vector $P_x = (Pr(x = x_1), \cdots, Pr(x = x_{d_x}))^T$.

For any matrix M, let $\sigma_t(M)$ denote the t-th largest singular value of M. For any vector x, its Euclidean norm is denoted by $\|x\|$ and the spectral norm of M is denoted by $\|M\|$, i.e., $\|M\| := \sup_{\|x\|=1} \|Mx\|$.

For latent tree models, two standard assumptions (see [7] and [19]) ensure that a latent tree does not include a redundant latent variable, as follows.
(A1) Each latent variable has at least three neighbors.
(A2) Any two variables connected by an edge in the tree model are neither perfectly dependent nor independent.
In the present study, the observed variables in our architecture are allowed to have different numbers of states. To obtain probably approximately correct results, we also require the following two assumptions.
(A3) Latent variables have the same number r of states.
(A4) The joint probability matrix has rank r.
To avoid a case where the generalized information distance is infinity, the assumption (A4) of the rank of the joint probability matrix is suggested, which is a generalization of the parameter identifiability conditions in latent variable models [1, 13, 17]. The following section shows that (A3) and (A4) ensure that the generalized information distance based on the product of singular values has an additive property along paths in latent trees.

3. LEARNING THE TREE STRUCTURE FOR DISCRETE LATENT TREE MODELS

3.1 Generalized information distance

We define a generalized information distance between two discrete variables x and y by

$$d_{xy} = -\log \frac{\prod_{s=1}^{r} \sigma_s(P_{xy})}{\sqrt{\det(P_{xx}) \det(P_{yy})}}, \tag{3.1}$$

where $\sigma_s(A)$ denotes the s-th largest singular value of matrix A and r is the rank of P_{xy}. This distance is an extension version of the information distance [7] and is similar as the case of continuous non-Gaussian variables [22]. For continuous non-Gaussian variables x and y, a covariance operator $\mathcal{C}_{xy}$ can be introduced to compute the expectation of the product of functions $f(x)$ and $g(y)$, using linear operations in the reproducing kernel Hilbert space $\mathcal{F}$. Formally, let $\mathcal{C}_{xy} : \mathcal{F} \to \mathcal{F}$ such that for all $f, g \in \mathcal{F}$, $E_{xy}(f(x)g(y)) = \langle f, \mathcal{C}_{xy} g\rangle$. The distance metric [22] between two continuous non-Gaussian variables x and y is defined by

$$d_{xy} = -\frac{1}{2}\log|\mathcal{C}_{xy}\mathcal{C}_{xy}^T|_* + \frac{1}{4}\log|\mathcal{C}_{xx}\mathcal{C}_{xx}^T|_* + \frac{1}{4}\log|\mathcal{C}_{yy}\mathcal{C}_{yy}^T|_*,$$

where $|\mathcal{C}|_*$ is the product of non-zero singular values of $\mathcal{C}$.

For two discrete variables x and y have the same number of states, P_{xy} is a square matrix. Furthermore, if P_{xy} has a full rank, our generalized information distance is reduced to the information distance [7]. For example, if h and g are any two latent variables in H, based on assumptions (A3) and (A4), the generalized information distance d_{hg} defined in (3.1) degenerates into the information distance $-\log\frac{|\det(P_{hg})|}{\sqrt{\det(P_{hh})\det(P_{gg})}}$. The rank of P_{hg} is r, so every row of P_{hg} has at least one nonzero element and every column also has at least one nonzero element. Furthermore, some row of P_{hg} has at least two nonzero elements if and only if some column of P_{hg} has at least two nonzero elements. While there are two or more nonzero elements in some rows of P_{hg}, we find that $|\det(P_{hg})| < \det(P_{hh})$ based on the definition of the determinant and the relation between P_{hg} and P_{hh}. Similarly, while there are two or more nonzero elements in some column of P_{hg}, we find that $|\det(P_{hg})| < \det(P_{gg})$. As discussed above, $|\det(P_{hg})| = \det(P_{hh})$ if and only if every row of P_{hg} contains only one nonzero element. $|\det(P_{hg})| = \det(P_{gg})$ if and only if every column of P_{hg} contains only one nonzero element, which occurs if and only if every row of P_{hg} has only one nonzero element. Thus, $d_{hg} = 0$ if and only if every row of P_{hg} has only one nonzero element, which means that a permutation σ exists such that $Pr(g = g_{\sigma(i)}|h = h_i) = 1$, $i = 1, \cdots, r$. Based on assumption (A2), we find that $d_{hg} > 0$ for any two distinct latent variables $h, g \in H$ in latent tree models. Similar results regarding the positive properties of information distance were reported by Lake [15], where he proposed the use of sequence distances to reconstruct evolutionary trees and illustrated some basic properties of sequence distances. It should be noted that the generalized information distance d_{xy} may be negative if x or y is an observed variable. In the following section, however, we show that the sign does not affect our bottom-up algorithm for learning latent tree structure.

Similar to the information distance [7, 15], the generalized information distance also has additivity along paths. We present the following theorem, the proof of which is given in Appendix A.1.

Theorem 3.1. *For a tree $T = (W, E)$ and a discrete latent tree model $\mathcal{T}$ on T, if $[x_0 = x, x_1, \cdots, x_m = y]$ is a path in T between two variables $x, y \in W$, then we find that:*

$$d_{xy} = \sum_{i=0}^{m-1} d_{x_i x_{i+1}}.$$

3.2 Structural learning algorithm for discrete latent tree models

In this subsection, we propose an algorithm for the structural learning of latent trees (SLLT) based on generalized information distances. This algorithm is a modification of the recursive grouping (RG) procedure proposed by Choi et al. [7]. RG is a bottom-up algorithm used to determine the relationships among leaf variables at each iteration. Similar to RG, the computational complexity of our algorithm is also $O(diam(T)n^3)$, where T is the latent tree and n is the number of observed variables.

We assume that observed variables V are all the leaves of a latent tree model. The inputs of the SLLT algorithm are generalized information distances d_{xy} for any $x, y \in V$. The output of it is the latent tree structure.

For any three variables $x, y, z \in V$, we denote the generalized information distance difference $d_{xz} - d_{yz}$ by Φ_{xyz}. At 1^0 in Step 2 and $1^0, 2^0$ in Step 3 of the SLLT algorithm, two basic local structures among variables can be found based on this difference. One is called a sibling group, the nodes of which share the same neighbor, and any two nodes of a sibling group form a sibling pair. The other is called a remaining child relation, where one node u is adjacent only to the other node w that has been found, and we refer to u as a remaining child of w. In contrast to computing the information distance difference among latent variables [7], our algorithm only use the generalized information distance difference among observed variables.

In the SLLT algorithm, the temporary set Y contains nodes, where their structural relations are checked in the current step and Y is updated after we have found all the local structures among Y. The temporary set D records all of the nodes that have been found to be sibling groups or remaining child relations. Every new latent variable is generated by a sibling group, as shown by 3^0 in Step 2 and 4^0 on Step 3, so for every latent variable there are at least two bifurcation variables in D. The temporary set $D(v)$ can be regarded as a descendant set of v and, more precisely, the adjacent relations among $D(v)\bigcup\{v\}$ can form a subtree T_v with the root v. Furthermore, it can be seen that 4^0 in Step 3 incorporates new subtrees with some smaller subtrees.

To illustrate the SLLT algorithm in more detail, we introduce a discrete time record $t = 0, 1, 2, \cdots$ as a subscript of $Y, D, D(\bullet)$ to describe the iteration process when $|Y| \geq 3$. At the beginning of each iteration step, the set Y_t is contained in $W \setminus D_t$ and $D_t(x) \bigcap D_t(y) = \emptyset$ for any two distinct nodes $x, y \in Y_t$ because of the updating mechanism of $D_t(\bullet)$. In Appendix B, we discuss the details of Steps 2 and 3, which can find the correct remaining child relations and sibling groups in $T(W \setminus D_t)$ at each iteration, and the updated set D_{t+1} is added to the nodes as the remaining children or in the sibling groups of $T(W \setminus D_t)$ after each iteration. Furthermore, $T(W \setminus D_{t+1})$ is a subtree of T and the updated set Y_{t+1} contains all the leaf variables of $T(W \setminus D_{t+1})$ after each iteration.

Algorithm 1 Structural Learning for Latent Trees (SLLT)

```
Input: Observed variables $V$ and generalized information distances $d_{xy}$ for any $x, y \in V$;
Output: A tree structure $T$;
1: $Y \leftarrow V$. $D \leftarrow \emptyset$. For any $v \in V$, $D(v) \leftarrow \emptyset$.
2: If $|Y| \geq 3$, compute $\Phi_{xyz} = d_{xz} - d_{yz}$ for any three variables $x, y, z \in Y$.
   $1^0$. For any $x, y \in Y$,
           if $\Phi_{xyz}$ is constant for any $z \in Y \setminus \{x, y\}$,
           then $\{x, y\}$ are a sibling pair in $T$.
   $2^0$. Denote maximal sibling groups by $\{\Pi_l\}_{l=1}^L$.
        $Y \leftarrow Y \setminus \bigcup_{l=1}^{L} \Pi_l$.
   $3^0$. For any $l = 1, \cdots, L$,
           add a new latent variable $h_l$ and connect $h_l$ to every node in $\Pi_l$.
           $Y \leftarrow Y \bigcup \{h_l\}$,$D(h_l) \leftarrow \bigcup_{x \in \Pi_l} \{x\}$,$D \leftarrow D \bigcup (\bigcup_{x \in \Pi_l} \{x\})$.
3: While $|Y| \geq 3$,
   $1^0$. For any $v \in Y \bigcap V$ and any $u \in Y \setminus V$,
           $M \leftarrow V \setminus (D(u) \bigcup \{v\})$ and choose bifurcation variables $i, j$ of $u$ in $D$.
           If $\Phi_{vim}$ is constant and $\Phi_{vim} \neq \Phi_{vij}$ for any $m \in M$,
             then $\{v, u\}$ is a sibling pair in $T(W \setminus D)$.
   $2^0$. For any two variables $u, w \in Y \setminus V$,
           $M \leftarrow V \setminus (D(u) \bigcup D(w))$ and choose bifurcation variables $i, j$ of $u$ and $k, l$ of $w$ in $D$.
           If $\Phi_{ikm} = \Phi_{ikl}$ and $\Phi_{kim} \neq \Phi_{kij}$ for any $m \in M$,
             then $u$ is a remaining child of $w$ in $T(W \setminus D)$.
        For any two variables $u, w \in Y \setminus V$,
           if neither $u$ nor $w$ is a remaining child,
             $M \leftarrow V \setminus (D(u) \bigcup D(w))$ and choose bifurcation variables $i, j$ of $u$ and $k, l$ of $w$ in $D$.
             If $\Phi_{ikm}$ is constant and $\Phi_{ikm} \neq \Phi_{ikl}, \Phi_{kim} \neq \Phi_{kij}$ for any $m \in M$,
                then $\{u, w\}$ is a sibling pair in $T(W \setminus D)$.
   $3^0$. Denote the remaining child relations and maximal sibling groups by $\{\Pi_l\}_{l=1}^L$. $Y \leftarrow Y \setminus \bigcup_{l=1}^{L} \Pi_l$.
   $4^0$. For any $l = 1, \cdots, L$,
           if $\Pi_l = \{u, w\}$ and $u$ is a remaining child of $w$,
             then connect $u$ and $w$.
             $Y \leftarrow Y \bigcup \{w\}$, $D(w) \leftarrow D(w) \bigcup D(u) \bigcup \{u\}$
             and $D \leftarrow D \bigcup \{u\}$.
           if $\Pi_l$ is a sibling group,
             then add a new latent variable $h_l$ and connect $h_l$ to every node in $\Pi_l$.
             $Y \leftarrow Y \bigcup \{h_l\}$, $D(h_l) \leftarrow \bigcup_{x \in \Pi_l} (D(x) \bigcup \{x\})$
             and $D \leftarrow D \bigcup (\bigcup_{x \in \Pi_l} \{x\})$.
4: If $|Y| = 2$, connect the two variables in $Y$.
5: return The structure generated by the adjacent relationship.
```

Let us consider an example to illustrate our algorithm. A latent tree T is shown in Figure 1, where $v_1, \cdots, v_{12}$ are observed variables and $h_1, \cdots, h_8$ are latent variables. According to our SLLT algorithm, the generalized information distances among observed variables $v_1, \cdots, v_{12}$ of T are used to learn the unknown latent tree structure. For any three observed variables x, y and z, all the generalized information distance differences Φ_{xyz} can be computed by $d_{xz} - d_{yz}$. When $\Phi_{v_1 v_2 v_3} = \Phi_{v_1 v_2 v_4} = \cdots = \Phi_{v_1 v_2 v_{12}}$, we find that $\{v_1, v_2\}$ is a sibling pair in T. Similarly, $\{v_1, v_3\}$, $\{v_2, v_3\}$, $\{v_4, v_5\}$, $\{v_6, v_7\}$, $\{v_8, v_9\}$, $\{v_{10}, v_{11}\}$ are sibling pairs in T. Thus, $\{v_1, v_2, v_3\}$, $\{v_4, v_5\}$, $\{v_6, v_7\}$, $\{v_8, v_9\}$, $\{v_{10}, v_{11}\}$ are five maximal sibling groups, and h_5, h_2, h_6, h_7, and h_8 are added as latent variables. $D = \{v_1, \cdots, v_{11}\}$ and $Y = \{h_2, h_5, h_6, h_7, h_8, v_{12}\}$. After dropping the variables in D from T, all of the leaf variables $\{h_5, h_6, h_7, h_8, v_{12}\}$ of $T(W \setminus D)$ are contained in Y.

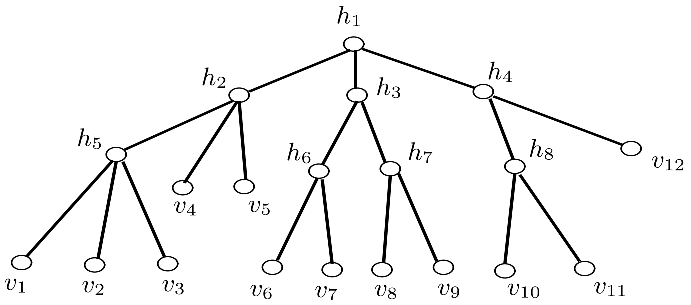


Figure 1. Example of the SLLT algorithm.

The observed variables v_{10}, v_{11} are bifurcation variables of the latent variable h_8. When $\Phi_{v_{12} v_{10} v_m}$ is constant and $\Phi_{v_{12} v_{10} v_m} \neq \Phi_{v_{12} v_{10} v_{11}}$ for $m = 1, \cdots, 9$, we find that $\{h_8, v_{12}\}$ is a sibling pair in $T(W \setminus D)$. The observed variables v_1, v_2 are bifurcation variables of latent variable h_5 and the observed variables v_4, v_5 are bifurcation variables of latent variable h_2. When $\Phi_{v_1 v_4 v_m} = \Phi_{v_1 v_4 v_5}$ and $\Phi_{v_4 v_1 v_m} \neq \Phi_{v_4 v_1 v_2}$ for $m = 6, \cdots, 12$, the latent variable h_5 is a remaining child of latent variable h_2. The observed variables v_6, v_7 are bifurcation variables of latent variable h_6 and the observed variables v_8, v_9 are bifurcation variables of latent variable h_7. When $\Phi_{v_6 v_8 v_m}$ is constant and $\Phi_{v_6 v_8 v_m} \neq \Phi_{v_6 v_8 v_9}$, $\Phi_{v_8 v_6 v_m} \neq \Phi_{v_8 v_6 v_7}$ for $m = 1, \cdots, 5, 10, 11, 12$, we find that $\{h_6, h_7\}$ is a sibling pair in $T(W \setminus D)$. Thus, $\{h_8, v_{12}\}$ and $\{h_6, h_7\}$ are two maximal sibling groups in $T(W \setminus D)$, and h_4, h_3 are added as latent variables. $D = \{v_1, \cdots, v_{12}, h_5, \cdots, h_8\}$ and $Y = \{h_2, h_3, h_4\}$. After dropping the variables in D from T, all of the leaf variables $\{h_2, h_3, h_4\}$ of $T(W \setminus D)$ are contained in Y. Finally, $\{h_2, h_3, h_4\}$ forms a sibling group based on similar checks.

The following theorem illustrates the correctness and computational complexity of our SLLT algorithm, and its proof is provided in Appendix A.2.

Theorem 3.2. *If the joint probability matrixes P_{xy}, $x, y \in V$, are available, then SLLT outputs the true tree T correctly within the time $O(diam(T)n^3)$.*

To apply our SLLT algorithm to data, we need to use the empirical estimation $\hat{P}_{xy}$ of the joint probability matrix

P_{xy} for any two observed variables $x, y \in V$. We compute the empirical estimation matrix $\hat{P}_{xy} = (\frac{1}{N}\sum_{k=1}^{N} \mathbb{I}(x^{(k)} = x_i, y^{(k)} = y_j))_{1 \le i \le d_x, 1 \le j \le d_y}$, where N is the sample size. Furthermore, the empirical estimation $\hat{d}_{xy}$ of the generalized information distance d_{xy} can be obtained from $\hat{P}_{xy}$ and the empirical estimation $\hat{\Phi}_{xyz}$ of the generalized information distance difference Φ_{xyz} can be computed by $\hat{\Phi}_{xyz} = \hat{d}_{xz} - \hat{d}_{yz}$ for $x, y, z \in V$.

In the SLLT algorithm, to determine the relations between two variables, we only need to check whether Φ_{xyz} is equal to some constant or not. However, in the sample-based SLLT algorithm, $\hat{\Phi}_{xyz}$ and $\hat{\Phi}_{xyw}$ may not be exactly equal even if $\Phi_{xyz} = \Phi_{xyw}$. The difference $|\hat{\Phi}_{xyz} - \hat{\Phi}_{xyw}|$ tends to $|\Phi_{xyz} - \Phi_{xyw}|$ when $N \to \infty$, so we use this difference to determine the equality of Φ_{xyz} and Φ_{xyw} in the sample-based version of the algorithm. Thus, we introduce a prescribed threshold $\epsilon > 0$ such that $|\hat{\Phi}_{xyz} - \hat{\Phi}_{xyw}| < \epsilon$ if and only if $\Phi_{xyz} = \Phi_{xyw}$ when the sample size N is sufficiently large. In fact, we define a lower bound notation $\rho_{min} = \min\{|\Phi_{xyz} - \Phi_{xyw}| : \Phi_{xyz} \neq \Phi_{xyw}, x, y, z, w \in V\}$, and choose a threshold value $\epsilon < \min\{\frac{1}{2}\rho_{min}, 1\}$. Furthermore, if $|(\hat{\Phi}_{xyz} - \hat{\Phi}_{xyw}) - (\Phi_{xyz} - \Phi_{xyw})| < \epsilon$ when N is sufficiently large, then we find that $|(\hat{\Phi}_{xyz} - \hat{\Phi}_{xyw})| < \epsilon$ if and only if $\Phi_{xyz} = \Phi_{xyw}$. Therefore, if the event $\{|(\hat{\Phi}_{xyz} - \hat{\Phi}_{xyw}) - (\Phi_{xyz} - \Phi_{xyw})| < \epsilon \text{ for any } x, y, z, w \in V\}$ occurs with a high probability when the sample size is sufficiently large, we can learn the true latent tree structure with a high probability from the sample-based SLLT algorithm based on the correction of Theorem 3.2.

The following theorem shows the structural consistency of the sample-based SLLT algorithm. Moreover, it illustrates the relation between the sample size and the intrinsic parameters of models when learning the true latent tree. The following notations are used in the following theorem. We denote σ_{min} as the minimum of the non-zero singular values of all the joint probability matrices for the observed variables, i.e., $\sigma_{min} = \min_{x,y \in V} \sigma_{rank(P_{xy})}(P_{xy})$. We denote d_{max} as the maximum number of states of the observed variables, i.e., $d_{max} = \max_{x \in V} d_x$.

Theorem 3.3. *Let $\eta \in (0,1)$. Assume that the SLLT algorithm is provided with N independent samples from the distribution over the observed variables set V. If the sample size N is sufficiently large such that*

$$\frac{\sqrt{t_0} + 1}{\sqrt{N}} < \frac{\epsilon \sigma_{min}}{16 d_{max}}, \tag{3.2}$$

where $t_0 = -\log \frac{\eta}{n^2 d_{max}}$, then with a probability of at least $1 - \eta$, the SLLT algorithm returns the true latent tree.

The proof of this Theorem 3.3 is provided in Appendix A.3, where we show that the event $\{|(\hat{\Phi}_{xyz} - \hat{\Phi}_{xyw}) - (\Phi_{xyz} - \Phi_{xyw})| < \epsilon$ for any $x, y, z, w \in V\}$ occurs with a high probability if the sample size is sufficiently large. From (3.2), we know that if N is sufficiently large such that $\frac{16 d_{max}(\sqrt{t_0}+1)}{\sigma_{min}\sqrt{N}} < \min\{\frac{1}{2}\rho_{min}, 1\}$, there exists a suitable threshold value $\epsilon < \min\{\frac{1}{2}\rho_{min}, 1\}$.

The SLLT algorithm based on the generalized information distance is a modification of the recursive grouping (RG) algorithm [7]. The major differences between the SLLT and the RG lie in three points. First, the SLLT adopts the generalized information distance which is an extension version of the information distance used in the RG. Second, the computation in SLLT relies on distances between observed variables while the RG needs to use distances on latent variables. Third, the RG utilizes one extra tuning parameter to control the learning of latent tree. Inspired by the RG algorithm, we introduce two tuning parameters for controlling the empirical distance, and obtain a modified version of SLLT with a better structural learning ability. Since a longer distance estimate is less accurate for a given number of samples, not all estimated distances can be used for structural learning reliably [7]. So we only consider possible sibling pairs for nodes x, y whose estimated distances $\hat{d}_{xy}, \hat{d}_{xz}, \hat{d}_{yz}$ are controlled by two thresholds τ_1, τ_2. Specifically, for each pair of nodes x, y such that $\hat{d}_{xy} < \tau_1$, $\hat{\Phi}_{xyz}$ is computed for node z in $\mathcal{K}_{xy} = \{z \in V \setminus \{x, y\} | \max\{\hat{d}_{xz}, \hat{d}_{yz}\} < \tau_2\}$. So we can obtain a modified algorithm SLLT2 with two thresholds τ_1 and τ_2 for structural learning. The threshold τ_1 can control the relationship of nodes in sibling groups. A small τ_1 makes variables in a sibling group close to each other. The threshold τ_2 can control the judgement of the sibling pair relationship. A large τ_2 introduces more nodes into $\mathcal{K}_{xy}$. It increases the possibility to find that $\hat{\Phi}_{xyz}$ is not a constant. So this case tends to obtain a tree with small branches. A possible way for trimming nodes from the tree is to reduce the value of τ_2.

To our best knowledge, there is no widely accepted framework of tuning parameters selection for latent tree models. Tuning parameters τ_1 and τ_2 can be interpreted as an upper bound of distances among sibling pairs and an upper bound of reliable distances among observed variables respectively. In this paper, τ_1 and τ_2 are chosen such that $\tau_1 < \tau_2$ and τ_1, τ_2 are less than the mean plus two standard deviations of the distances computed in our experiments. Specifically, we empirically set $\tau_1 = 3$ and $\tau_2 = 5$.

4. ESTIMATING THE PARAMETERS FOR DISCRETE LATENT TREE MODELS

The methods described in this section were inspired by Chang's matrix decomposition technique [5] for discrete Markov models of evolution. A similar technique [17] is also discussed for phylogenies and hidden Markov models. Another method [2] based on the randomized spectral decomposition can obtain the parameter estimates for mixture models, but this method does not lead directly to practical algorithms because of the amplification of errors due to the reliance of a high order polynomial factor on the rank parameter.

In general, we do not know the state number for latent variables. Thus, the discrete latent tree model with only two-state latent variables is a simple and efficient choice that meets Occam's Razor principle, although the BIC could give other suggestions based on the penalized likelihood. Moreover, based on models with two-state latent variables, our matrix decomposition method does not require the control of eigenvalue separation using the additional randomized methods employed by [2], so our method has a more compact sample size bound shown in the following Theorem 4.1.

In this section, the observed variables are allowed to have different numbers of states, but we will start with an illuminating case where the observed variables are two-state.

4.1 Computing conditional probability matrices by spectral decomposition

In this subsection, we assume that all the variables in W are two-state, i.e., $d_x = 2$ for any $x \in W$. According to the basic assumption (A1), for each latent variable h, there are at least three neighbors of h. Thus, there are at least three leaf variables a, b, c in the tree such that a, b, c are conditionally independent given h. As considered in [5], we focus on the conditional probabilities $Pr(a = a_i, b = b_j | c = c_k)$ for $i, j, k = 1, 2$, which are well defined when $Pr(c = c_k)$ is positive. According to the conditional independence relationship in latent tree models,

$$\begin{aligned}&Pr(a = a_i, b = b_j | c = c_k)\\&= \sum_l Pr(h = h_l, a = a_i, b = b_j | c = c_k),\end{aligned}$$

where $Pr(h = h_l, a = a_i, b = b_j | c = c_k) = Pr(b = b_j | h = h_l) Pr(a = a_i | h = h_l) Pr(h = h_l | c = c_k)$. We select i as 1 and denote the matrix $(Pr(a = a_1, b = b_j | c = c_k))_{1 \le j \le 2, 1 \le k \le 2}$ as $P_{b|c}^{a=a_1}$, thus $P_{b|c}^{a=a_1}$ has the following decomposition:

$$P_{b|h} \begin{pmatrix} Pr(a = a_1 | h = h_1) & 0 \\ 0 & Pr(a = a_1 | h = h_2) \end{pmatrix} P_{h|c}.$$

Furthermore, it can be decomposed into:

$$P_{b|h} \begin{pmatrix} Pr(a = a_1 | h = h_1) & 0 \\ 0 & Pr(a = a_1 | h = h_2) \end{pmatrix} P_{b|h}^{-1} P_{b|c}.$$

Then, we obtain the spectral decomposition form:

$$\begin{aligned}&P_{b|c}^{a=a_1} P_{b|c}^{-1}\\&= P_{b|h} \begin{pmatrix} Pr(a = a_1 | h = h_1) & 0 \\ 0 & Pr(a = a_1 | h = h_2) \end{pmatrix} P_{b|h}^{-1}.\end{aligned}$$

We denote the matrix $(Pr(a = a_1, b = b_j, c = c_k))_{1 \le j \le 2, 1 \le k \le 2}$ as $P_{bc}^{a=a_1}$. Since $P_{bc}^{a=a_1} = P_{b|c}^{a=a_1} P_{cc}$ and $P_{bc} = P_{b|c} P_{cc}$, we have:

$$\begin{aligned}&P_{bc}^{a=a_1} P_{bc}^{-1}\\&= P_{b|h} \begin{pmatrix} Pr(a = a_1 | h = h_1) & 0 \\ 0 & Pr(a = a_1 | h = h_2) \end{pmatrix} P_{b|h}^{-1}.\end{aligned} \tag{4.1}$$

The left side of this equation can be estimated by using the observed variables. By spectral decomposition, the parameters of the latent variables can be obtained from the right side of this equation. From assumption (A4), we find that $Pr(a = a_1 | h = h_1) \neq Pr(a = a_1 | h = h_2)$, although we do not know the actual states of h_1 and h_2. From equation (4.1), the conditional probabilities $Pr(a = a_1 | h = h_1)$ and $Pr(a = a_1 | h = h_2)$ are the two eigenvalues of $P_{bc}^{a=a_1} P_{bc}^{-1}$, which we denote as λ_1 and λ_2, respectively, and we assume that $\lambda_1 > \lambda_2$ without any loss of generalization. Furthermore, the vector $x_k = (Pr(b = b_1 | h = h_k), Pr(b = b_2 | h = h_k))^T$ is the eigenvector of $P_{bc}^{a=a_1} P_{bc}^{-1}$ that is associated with the eigenvalue λ_k for $k = 1, 2$. Thus, from the spectral decomposition of $P_{bc}^{a=a_1} P_{bc}^{-1}$, we can obtain $P_{b|h}$ by using the restriction relation that $Pr(b = b_1 | h = h_k) + Pr(b = b_2 | h = h_k) = 1$ for $k = 1, 2$.

To apply this spectral decomposition method to data, we need to estimate the two matrices $P_{bc}^{a=a_1}$ and P_{bc}^{-1} from the observed data. We use the frequency matrix $\hat{P}_{bc}^{a=a_1} = (\frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(a^{(k)} = a_1, b^{(k)} = b_j, c^{(k)} = c_l))_{1 \le j \le 2, 1 \le l \le 2}$ to estimate the conditional probability matrix $P_{bc}^{a=a_1}$, where N is the sample size. Similarly, the conditional probability matrix P_{bc} is estimated by using the frequency matrix $\hat{P}_{bc} = (\frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(b^{(k)} = b_j, c^{(k)} = c_l))_{1 \le j \le 2, 1 \le l \le 2}$, where N is the sample size. As discussed above, based on the spectral decomposition of $\hat{P}_{bc}^{a=a_1} \hat{P}_{bc}^{-1}$, we can obtain the estimation $\hat{P}_{b|h}$ for $P_{b|h}$.

The following Theorem 4.1 shows that the eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2$ with $\hat{\lambda}_1 \ge \hat{\lambda}_2$ and the corresponding restricted eigenvectors $\hat{x}_1, \hat{x}_2$ of $\hat{P}_{bc}^{a=a_1} \hat{P}_{bc}^{-1}$ are consistent with the true eigenvalues λ_1, λ_2 and the corresponding restricted eigenvectors x_1, x_2 of $P_{bc}^{a=a_1} P_{bc}^{-1}$. Thus, the estimation $\hat{P}_{b|h}$ from this spectral decomposition method is consistent with the true conditional probability matrix $P_{b|h}$. The proof of Theorem 4.1 is shown in Appendix B.1. For any invertible matrix A, $\kappa(A)$ denotes the condition number of A, i.e., $\kappa(A) = \|A\| \|A^{-1}\|$. Based on our assumptions, both P_{bc} and $P_{a|h}$ are invertible. For any $\eta \in (0, 1)$, we define the events $E = \bigcap_{i=1,2} \{|\hat{\lambda}_i - \lambda_i| \le 3\kappa(P_{b|h}) \|P_{bc}^{-1}\|^2 \frac{1+\sqrt{t_0}}{\sqrt{N}}, \|\hat{x}_i - x_i\| \le \frac{18}{|\lambda_1 - \lambda_2|}(1 + \kappa(P_{b|h})) \|P_{bc}^{-1}\|^2 \frac{1+\sqrt{t_0}}{\sqrt{N}}\}$, where $t_0 = -\log \frac{1}{2}\eta$.

Theorem 4.1. *For any $\eta \in (0, 1)$, when the sample size N is sufficiently large such that*

$$\frac{18}{|\lambda_1 - \lambda_2|}(1 + \kappa(P_{b|h})) \|P_{bc}^{-1}\|^2 \frac{1 + \sqrt{t_0}}{\sqrt{N}} < 1, \tag{4.2}$$

where $t_0 = -\log \frac{1}{2}\eta$, we find that $P(E) > 1 - \eta$.

Chang [5] (1996) proposed this spectral decomposition method for the evolutionary tree model. This method requires that all of the variables have an equal number of states and it does not limit the state number of the latent variables to two. However, when the state number of latent variables is more than two, this direct decomposition method may cause the non-identifiablity of the eigenvector because the dimension of some eigenspaces may be larger than one. In the following subsection, we generalize Chang's method to latent tree models with observed variables that are allowed to have different number of states. However, the restriction of models with two-state latent variables is used to avoid non-identifiability.

4.2 Case with only two-state latent variables

In this subsection, we assume that the state number of the latent variables is two and that the state number of the observed variables may be greater than two. As mentioned in the previous subsection, for any latent variable h, there are three leaf variables $a, b, c \in V$ such that a, b, c are conditionally independent given h. We consider the case where $d_a, d_b, d_c \geq 2$ and we make some modifications of the matrix decomposition method discussed in the previous subsection.

For any state $1 \leq i \leq d_a$, we can view variable a as a two-state variable $a^{(i)}$. The first state $a_1^{(i)}$ of $a^{(i)}$ is the i-th state a_i of a, and the second state $a_2^{(i)}$ represents the other states $\{a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_{d_a}\}$ of a. Similarly, for any $1 \leq j \leq d_b, 1 \leq k \leq d_c$, variables b and c can be viewed as the two-state variables $b^{(j)}$ and $c^{(k)}$. The three two-state variables $a^{(i)}$, $b^{(j)}$, and $c^{(k)}$ are also conditionally independent given h.

If $P_{b^{(j)}c^{(k)}}$ is nonsingular, $P^{a^{(i)}=a_1^{(i)}}_{b^{(j)}c^{(k)}} P^{-1}_{b^{(j)}c^{(k)}}$ exists a spectral decomposition:

$$P_{b^{(j)}|h} \begin{pmatrix} Pr(a = a_i|h = h_1) & 0 \\ 0 & Pr(a = a_i|h = h_2) \end{pmatrix} P^{-1}_{b^{(j)}|h}.$$

Since the rank of $P_{a|h}$ is two, a minimum subscript i' exists such that $Pr(a = a_{i'}|h = h_1) \neq Pr(a = a_{i'}|h = h_2)$. We refer to $a_{i'}$ as the label state of a from h and record the values $Pr(a = a_{i'}|h = h_1)$ and $Pr(a = a_{i'}|h = h_2)$. Based on the spectral decompositions of $P^{a^{(i')}=a_1^{(i')}}_{b^{(j)}c^{(k)}} P^{-1}_{b^{(j)}c^{(k)}}$, we can obtain the conditional probability matrix $P_{b^{(j)}|h}$ because $Pr(a = a_{i'}|h = h_1) \neq Pr(a = a_{i'}|h = h_2)$. The first row $(Pr(b^{(j)} = b_1^{(j)}|h = h_1), Pr(b^{(j)} = b_1^{(j)}|h = h_2))$ of $P_{b^{(j)}|h}$ is simply the jth row $(Pr(b = b_j|h = h_1), Pr(b = b_j|h = h_2))$ of $P_{b|h}$.

If $P_{b^{(j)}c^{(k)}}$ is singular, this means that $P_{b^{(j)}|h}$ or $P_{c^{(k)}|h}$ is singular because $P_{b^{(j)}c^{(k)}} = P_{b^{(j)}|h} P_{hh} P^T_{c^{(k)}|h}$. We adjust the value of k and if $P_{b^{(j)}c^{(k)}}$ is nonsingular after the adjustment, the jth row of $P_{b|h}$ can be obtained by spectral decompositions of $P^{a^{(i')}=a_1^{(i')}}_{b^{(j)}c^{(k)}} P^{-1}_{b^{(j)}c^{(k)}}$. If $P_{b^{(j)}c^{(k)}}$ is singular irrespective of the adjustment of the value of k, $P_{b^{(j)}|h}$ must be singular because there at least one state of c exists such that $P_{c^{(k)}|h}$ is nonsingular. Thus, we find that $Pr(b = b_j|h = h_1) = Pr(b = b_j|h = h_2)$. We can adjust the value of k such that $P_{a^{(i')}c^{(k)}}$ is nonsingular. Indeed, because the rank of $P_{c|h}$ is two, we find that $Pr(c = c_k|h = h_1) \neq Pr(c = c_k|h = h_2)$ for some $1 \leq k \leq d_c$, thus $P_{a^{(i')}c^{(k)}}$ is nonsingular since $P_{a^{(i')}c^{(k)}} = P_{a^{(i')}|h} P_{hh} P^T_{c^{(k)}|h}$ and both $P_{a^{(i')}|h}$ and $P_{c^{(k)}|h}$ are nonsingular. Using spectral decomposition of

$$\begin{aligned} &P^{b^{(j)}=b_1^{(j)}}_{a^{(i')}c^{(k)}} P^{-1}_{a^{(i')}c^{(k)}} \\ &= P_{a^{(i')}|h} \begin{pmatrix} Pr(b = b_j|h = h_1) & 0 \\ 0 & Pr(b = b_j|h = h_2) \end{pmatrix} P^{-1}_{a^{(i')}|h}, \end{aligned}$$

we can obtain two equal conditional probabilities, $Pr(b = b_j|h = h_1)$ and $Pr(b = b_j|h = h_2)$, which form the jth row of $P_{b|h}$.

As discussed above, regardless of whether $P_{b^{(j)}c^{(k)}}$ is singular or not, we can obtain the jth row of $P_{b|h}$. Thus, the conditional matrix $P_{b|h}$ can be obtained step by step. Since the rank of $P_{b|h}$ is two, a minimum subscript j' exists such that $Pr(b = b_{j'}|h = h_1) \neq Pr(b = b_{j'}|h = h_2)$. We refer to $b_{j'}$ as the label state of b from h and record the values $Pr(b = b_{j'}|h = h_1)$ and $Pr(b = b_{j'}|h = h_2)$.

To compute the conditional probability matrix $P_{c|h}$, we switch $b^{(j)}$ and $c^{(k)}$, and $P_{c|h}$ can be obtained via the repeated spectral decompositions of $P^{a^{(i')}=a_1^{(i')}}_{c^{(k)}b^{(j)}} P^{-1}_{c^{(k)}b^{(j)}}$ or $P^{c^{(k)}=c_1^{(k)}}_{a^{(i')}b^{(j)}} P^{-1}_{a^{(i')}b^{(j)}}$. To compute the conditional probability matrix $P_{a|h}$, we use the label state $b_{j'}$ of b from h and the values $Pr(b = b_{j'}|h = h_1)$ and $Pr(b = b_{j'}|h = h_2)$, as well as performing the decompositions of $P^{b^{(j')}=b_1^{(j')}}_{a^{(i)}c^{(k)}} P^{-1}_{a^{(i)}c^{(k)}}$ to obtain the two eigenvectors, or the decompositions of $P^{a^{(i)}=a_1^{(i)}}_{b^{(j')}c^{(k)}} P^{-1}_{b^{(j')}c^{(k)}}$ to obtain the two eigenvalues. The conditional probability matrix $P_{a|h}$ can be obtained by performing these decompositions repeatedly.

During the process used to generate the conditional probability matrix $P_{b|h}$, we actually obtain each of its rows in each decomposition step. It should be noted that the corresponding states of latent variable h in each column of $P_{b|h}$ must be the same. To achieve this, we fix two eigenvalues $Pr(a = a_{i'}|h = h_1) \neq Pr(a = a_{i'}|h = h_2)$, and obtain the two eigenvectors that correspond to $Pr(a = a_{i'}|h = h_1)$ and $Pr(a = a_{i'}|h = h_2)$ in each decomposition step. Moreover, we record the label states $b_{j'}$ of b obtained from h and the obtained values $Pr(b = b_{j'}|h = h_1)$ and $Pr(b = b_{j'}|h = h_2)$.

To determine the conditional probability $P_{c|h}$, we decompose the matrix $P^{a^{(i')}=a_1^{(i')}}_{c^{(k)}b^{(j)}} P^{-1}_{c^{(k)}b^{(j)}}$ to obtain the two eigenvectors according to the values $Pr(a = a_{i'}|h = h_1)$ and $Pr(a = a_{i'}|h = h_2)$. To determine the conditional probability $P_{a|h}$, we decompose the matrix $P^{b^{(j')}=b_1^{(j')}}_{a^{(i)}c^{(k)}} P^{-1}_{a^{(i)}c^{(k)}}$ to obtain the two eigenvectors according to the values $Pr(b = b_{j'}|h = h_1)$ and $Pr(b = b_{j'}|h = h_2)$. As shown above, recording the label states $a_{i'}, b_{j'}$ guarantees that the states of latent variable h that correspond to the two columns of $P_{c|h}$ and $P_{a|h}$ match those of the two columns obtained for $P_{b|h}$.

4.3 Parameter estimation algorithm for discrete latent tree models

As discussed above, we only need the joint distributions of three observed variables to obtain the parameters of all the conditional probability matrices. Thus, we propose the following PELT algorithm.

Algorithm 2 Parameter Estimation for Latent Trees (PELT)

```
Input: A latent tree $T$ with a root and the joint distributions
    of three observed variables;
Output: All of conditional probability matrices on edges in $T$;
1: Construct the directed tree $\vec{T}$.
2: For every latent variable $h$,
       find all the child variables of $h$.
       For every child variable of $h$,
           find a directed bifurcation variable of child variable.
       Collect the set $C$ of all the directed bifurcation variables
       $\{a, b, \cdots\}$ of all the child variables of $h$.
       If $C$ contains exactly two variables,
           find an observed variable $c$ such that the path between
           $c$ and $h$ in $T$ does not contain any child of $h$.
       Compute $P_{a|h}, P_{b|h}, \cdots$ by spectral decomposition.
3: For every latent variable $h$ and every child variable $q$ of $h$,
       if $q$ is a latent variable,
           choose a common directed bifurcation variable $s$ of $q$
           and $h$, and compute conditional probability matrix
           $P_{q|h} = P^+_{s|q} P_{s|h}$.
4: return All the conditional probability matrices.
```

According to assumption (A1), every latent variable h has at least three neighbors. If h has exactly two child variables as described in step 2 of this algorithm, an observed variable c must exist such that the path between c and h in T does not contain any child of h. To guarantee that the column states of $P_{a|h}, P_{b|h}, \cdots$ for h in step 2 are matched, we need to record the label states of a and b from h and perform the corresponding matrix decompositions according to the label states. In step 3, since $P_{s|q}P_{q|h} = P_{s|h}$ and the rank of $P_{s|q}$ is two, $P_{q|h} = P^+_{s|q}P_{s|h}$, where $P^+_{s|q} = (P^T_{s|q}P_{s|q})^{-1}P^T_{s|q}$. Moreover, we can also obtain the probability vector P_{root} of the root using $P_{root} = P^+_{a|root}P_a$ for any observed variable a since $P_{a|root}P_{root} = P_a$.

In the PELT algorithm, we can use the frequency matrices to replace the true probability matrices and obtain the empirical version of this algorithm to estimate all the conditional probability matrices of the discrete latent tree models. If only the latent variables are restricted to being two-state, we find that for any latent variable $h \in H$ and any $v \in ch(h)$, the estimation matrix $\hat{P}_{v|h}$ obtained by using the PELT algorithm is consistent for the true conditional probability matrix $P_{v|h}$. The consistency theorem and proof are shown in Appendix B.2.

Table 1. Time costs of the EM and PELT algorithms using a sample size of 300 k

Model	EM (s)	PELT (s)
Model 1	4342.64	0.05393
Model 2	8659.58	0.11697
Model 3	3949.07	0.03561
Model 4	8528.13	0.09418

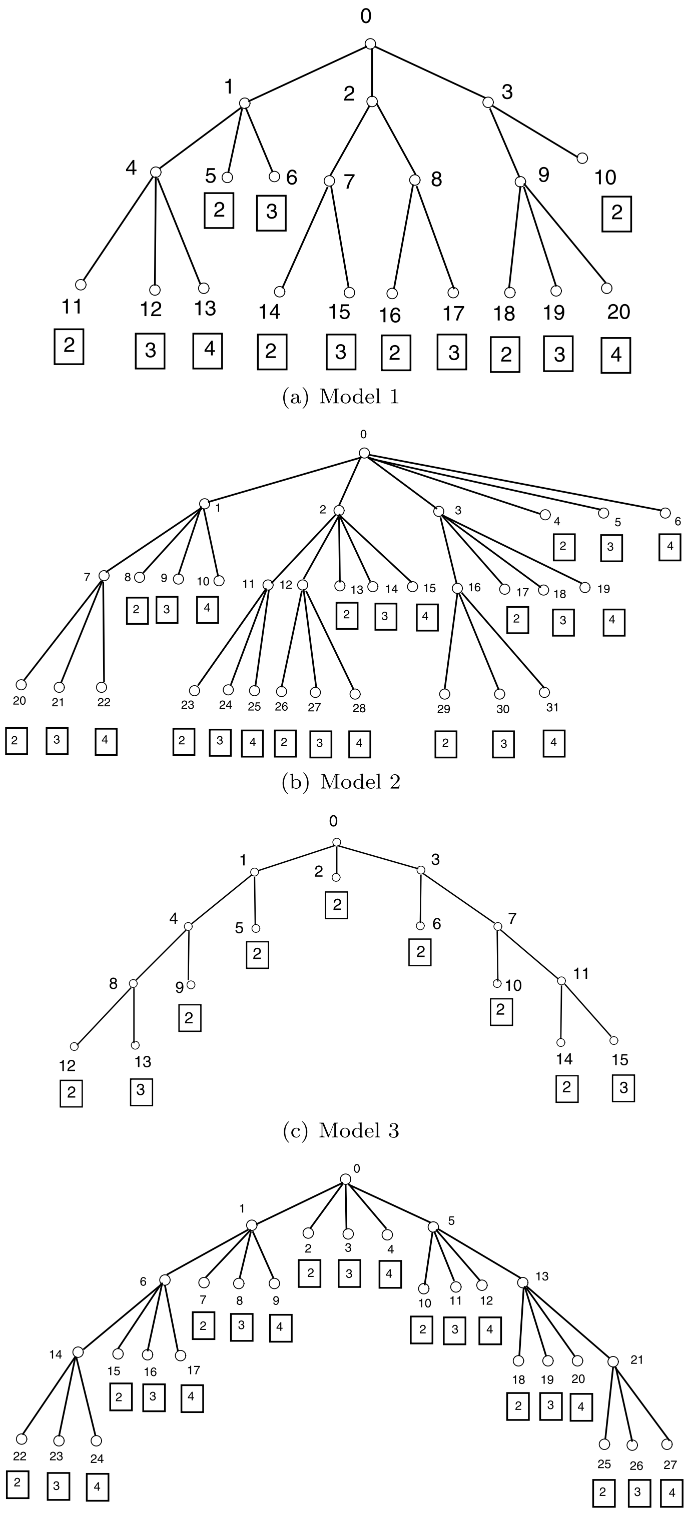
5. NUMERICAL EXPERIMENT

In this section, numerical experiments were designed for both simulated and real datum. In Section 5.1, we demonstrated the consistency of our algorithms on simulated data from four concrete latent tree structures. In Sections 5.2, we handled the Chinese text data on Public Security Bureau (PSB) from the Changchun Mayor Public Hotline. It can be seen that our algorithms provide valid hierarchical clusterings for observed variables in the datum. All of the experiments were performed using C++ on a desktop with an Intel Core i5-4590 CPU 3.3 GHz and 8 GB RAM.

5.1 Simulation study

In this subsection, we applied our SLLT and PELT algorithms to synthetic datasets generated from known latent tree models. We first presented the simulation results to demonstrate the consistency of our algorithms. And then we studied the performance of the SLLT algorithm, its modified version and Choi et al.'s RG algorithm [7] for structural learning. Finally, we compared the PELT algorithm and the EM algorithm for parameter estimation. As shown in Figures 3 and 4, the structural learning and parameter estimation errors of our algorithms decreased as the sample size increased. Figure 5 illustrates that the modified version SLLT2 performed much better than the SLLT algorithm and the RG algorithm. Figure 6 shows that the PELT algorithm partly outperformed the EM algorithm in terms of the estimation error and Table 1 shows that the execution speed of PELT was far faster than that of EM. Detailed descriptions and analyses of the simulation experiments are given in the following section.

We generated datasets from latent tree models with four different topologies, as shown in Figure 2. The structures of models 1 and 3 were similar to those of models 2 and 4, where we restricted every latent variable that was located adjacent to the three observed variables. The numbers in the boxes below the leaves are the numbers of state of observed variables. We wanted to compare the performances of the EM and PELT algorithms, so we fixed the number of states as two for the latent variables. The model parameters were generated randomly such that the determinant of the first 2 × 2 sub-matrix of every conditional probability matrix was no less than 0.3, which satisfied the full column rank requirement of the conditional probability matrices.



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure 2. Four latent tree models used in the simulation experiments.

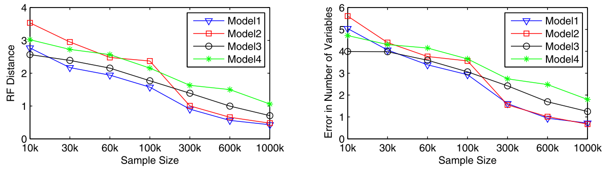


Figure 3. Performance of the SLLT algorithm using models 1, 2, 3, and 4.

As shown in Section 3.2, the basic sibling pair structure is determined using a prescribed threshold $\epsilon > 0$. In particular, if $|\hat{\Phi}_{xyz} - \hat{\Phi}_{xyw}| < \epsilon$ for all $z, w \in V \setminus \{x, y\}$, we consider that $\{x, y\}$ is a sibling pair. It is apparent that when we increase the value of this threshold, more observed variables tend to belong to a common sibling group and less become singletons. So it is much easier to obtain a latent tree structure for a larger ϵ. In the simulation for structural learning, we started the threshold ϵ from 0.1, and let it increase with a step size of 0.1 until the SLLT algorithm generates a tree structure. As shown in Section 4.2, we have two ways to estimate the conditional probability matrix $P_{b|h}$ according to whether $P_{b^{(j)}c^{(k)}}$ is singular or not. In the practical implement, a threshold ϵ_1 is set to 0.001 for judging whether $P_{b^{(j)}c^{(k)}}$ is singular or not based on the empirical estimation $\det(\hat{P}_{b^{(j)}c^{(k)}})$.

First, to evaluate the consistency of the performance of our SLLT and PELT algorithms, we varied the sample size among 10 k, 30 k, 60 k, 100 k, 300 k, 600 k, and 1000 k, and determined the average from 100 independent runs using different model parameters with each of the four structures shown in Figure 2.

We used two metrics to assess the performance of the consistency of the SLLT algorithm. The metric shown in the left subgraph of Figure 3 is the Robinson-Foulds metric [20], which quantifies the difference between the learned and true structures. The metric shown in the right subgraph of Figure 3 is the absolute difference between the number of learned and true latent variables. Figure 3 shows that the SLLT algorithm performed better as the sample size increased. The performance of the SLLT algorithm with model 3 was better than that with model 4. The extra leaves in model 4 compared with model 3 added the number of structure judgments in the SLLT algorithm, and also increased the possibility of making incorrect judgments.

The comparison of model 1 and model 3 in Figure 3 shows that the algorithm performed much better with model 1 than model 3 when the sample size was large, i.e., 300 k, 600 k, and 1000 k, although model 1 had more observed variables and latent variables than model 3. The tree width was 8 with model 3 and 6 with model 1, so the longer path in model 3 may have reduced the minimum non-zero singular value of the joint probability matrices. When the sample size was fixed and the minimum non-zero singular value became smaller, it was more difficult to obtain an exact estimate of the minimum non-zero singular value, which has a major effect on the generalized information distance. Thus, the

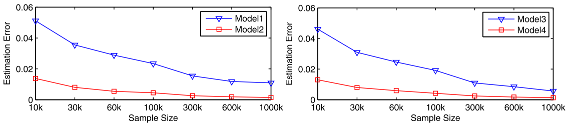


Figure 4. Performance of the PELT algorithm using models 1, 2, 3, and 4.

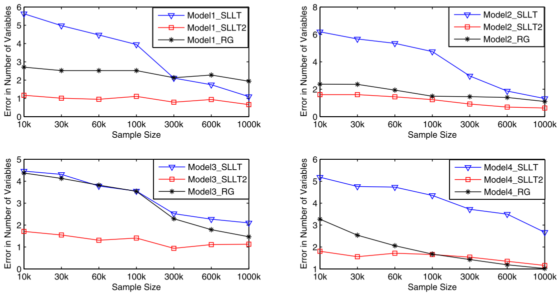


Figure 5. Performance of SLLT, SLLT2 and RG using models 1, 2, 3, and 4.

structure of model 3 was more difficult to recover than that of model 1 because the width of model 3 was greater than that of model 1.

For the PELT algorithm, we computed the average error between the estimated and true parameter based on the true structure. The left subgraph of Figure 4 shows that the PELT algorithm performed better with model 2 than model 1, although the structure of model 2 contained more observed variables than that of model 1. The increased number of observed variables made structure learning more difficult, as shown in Figure 3, although the extra leaves below the latent variables enhanced the parameter estimation accuracy. Indeed, when the latent variables bifurcated to more observed variables, we could choose b, c as the two end vertices of a path that was as short as possible, which helped move P_{bc} away from the singularity. Thus, based on matrix computation theory, we know that the estimation $\hat{P}_{bc}^{-1}$ of P_{bc}^{-1} and the further decomposition $\hat{P}_{bc}^{a=a_1}\hat{P}_{bc}^{-1}$ of $P_{bc}^{a=a_1}P_{bc}^{-1}$ would be more accurate. Similar differences in performance were also detected between models 4 and 3, which are shown in the right subgraph of Figure 4.

To compare our algorithms with Choi et al.'s RG algorithm [7], we used the four latent tree structures in Figure 2. The numbers of the state of variables were limited to two since the RG algorithm requires variables of the same cardinality. As shown in Figure 5, the performance of our SLLT2 algorithm was much better than the RG algorithm for model 1, 2 and 3. And the RG algorithm, utilizing one extra threshold to control the computation of $\hat{\Phi}_{xyz}$, outperformed the SLLT algorithm for model 2, 3 and 4.

Finally, we compared the performance of the EM algorithm and the PELT algorithm using latent tree models. We varied the sample size among 10 k, 30 k, 60 k, 100 k, and 300 k, and generated 100 random datasets with different model parameters using each of the four structures shown in Figure 2. For each dataset and every sample size, we ran the EM algorithm with five random initializations and 50 iterations, and we recorded the best estimation error obtained by using the EM algorithm. We compared the average estimation errors with the EM algorithm and the PELT algorithm for the 100 datasets and each sample size. As shown in Figure 6, the PELT algorithm outperformed the EM based on the estimation errors with models 1 and 3, and the performance of the EM algorithm was almost stable at every sample size, which appeared to be trapped by a local optimum with models 1 and 3. With models 2 and 4, the EM algorithm outperformed the PELT algorithm, but both had a high estimation accuracy with a sufficiently large sample size. The average time costs with the EM and the PELT algorithms using a sample size of 300 k are shown in Table 1. The execution speed was faster with the PELT algorithm, and the EM algorithm had huge time costs with a large sample size because the EM algorithm updates the statistics based on every sample and numerous iterations of all the samples are required.

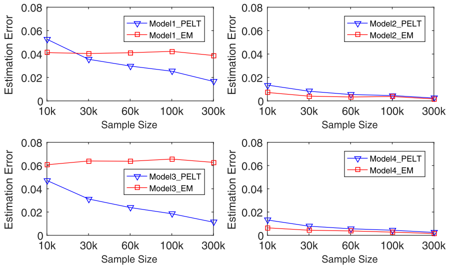


Figure 6. Estimation error using the EM and PELT algorithms with models 1, 2, 3, and 4.

5.2 Chinese text application

In this subsection, we tested our algorithms on real-world data sets. The probability distributions that govern these data sets may not satisfy the assumptions required for consistent learning of the latent tree models. However, the experiments here pointed out that our algorithms can still be useful in mining the latent hierarchical structures behind observed variables.

The Changchun mayor hotline [12] is an important project led by the local government of Changchun city, which is the capital of Jilin Province in Northeast China. This project provides local residents with the opportunity to call the mayor's office and report various public issues via the "12345" hotline. The typical issues reported include local crime, public utility, and problems with transportation. Each phone call is recorded and converted into a text message in Chinese by an operator.

We studied the Changchun mayor hotline dataset with 28,910 keywords for more than 640,000 documents. Our

Table 2. The first one hundred keywords for the PSB (The English version is shown in Table 3)

Id	Keywords	Id	Keywords	Id	Keywords	Id	Keywords
1	派出所	26	出警	51	警察	76	红绿灯
2	交警	27	饲养	52	嫌疑人	77	赌博机
3	市公安局	28	音箱	53	监控	78	犯罪
4	民警	29	每天	54	养狗	79	警员
5	犬	30	打伤	55	长春市公安局	80	一辆
6	报案	31	播放	56	禁	81	消防
7	大型	32	赌博	57	警	82	执勤
8	报警	33	此案	58	交警大队	83	声音
9	麻将馆	34	处罚	59	货车	84	建议
10	扰民	35	停放	60	派出	85	停车
11	宣传	36	疏导	61	户口	86	消防设施
12	公安部门	37	公安机关	62	单行线	87	牌照
13	音响	38	交通信号灯	63	治安	88	驾驶
14	高音喇叭	39	交警支队	64	路口	89	遛狗
15	车辆	40	罚款	65	发生	90	噪音扰民
16	狗	41	身份证	66	叫声	91	录像
17	交通	42	警号	67	麻将	92	打人
18	堵车	43	卖淫嫖娼	68	罚单	93	聚众赌博
19	养	44	交警队	69	被	94	交通堵塞
20	公安局	45	交通事故	70	告诫	95	市局
21	违章	46	三轮车	71	警力	96	左转
22	信号灯	47	喇叭	72	结案	97	殴打
23	案件	48	狗叫声	73	叫卖	98	休息
24	驾驶证	49	办案	74	严重	99	色情
25	立案	50	行驶	75	车速	100	播放音乐

method can handle this dataset in 9 hours and obtain a latent tree with 30,435 nodes, where contains 1,525 latent variables. To give a clear illustration for real data application in this paper, we built a small database consisting of 100 keywords, shown in Table 2 and 3, from 52,920 documents assigned to the Public Security Bureau (PSB). Those keywords were selected by positive correlation and large χ^2 value with PSB. We are concerned about the potential relationships between these keywords for PSB, because they can reflect various topics related to PSB.

In the practical implement, we used the SLLT2 algorithm to learn the latent tree structure as discussed in Section 5.1. This algorithm only considers possible sibling pairs for nodes x, y whose estimated distances $\hat{d}_{xy}, \hat{d}_{xz}, \hat{d}_{yz}$ are controlled by two thresholds τ_1, τ_2. Specifically, for each pair of nodes x, y such that $\hat{d}_{xy} < \tau_1$, $\hat{\Phi}_{xyz}$ is computed for node z in $\mathcal{K}_{xy} = \{z \in V \setminus \{x, y\} | \max\{\hat{d}_{xz}, \hat{d}_{yz}\} < \tau_2\}$. The threshold τ_1 can control the relationship of nodes in sibling groups. The threshold τ_2 can control the judgement of the sibling pair relationship. As discussed in Section 3.2, we empirically set $\tau_1 = 3$ and $\tau_2 = 5$.

By using the SLLT2 algorithm, we obtained a whole latent tree as shown in Figure 7. Its vertex set consists of 100 observed leaf nodes and 20 latent nodes. Sibling groups of observed nodes present various topics from the residents' complaint call. The main issues reflected by those groups include local crime, traffic problem, noise nuisance and keeping dogs. Latent node 101 of the largest degree is adjacent to seventeen leaf nodes. It reflects a common topic on local crime based on seventeen keywords, which is shown in Figure 8. Another important topic is on traffic problem (Figure 9) from latent node 102 adjacent to eleven keywords.

6. CONCLUSION

In this study, we proposed two algorithms for structure learning and parameter estimation in discrete latent tree models. We also presented provable guarantees for our algorithms and determined the relationship between the sample size and the intrinsic parameters of the models. The simulations showed that our algorithms are computationally efficient, even with large sample sizes, and the empirical results also support our theoretical results. The Chinese text application of our method illustrated that the latent tree model can capture common topics in the documents and mining latent structures behind observed variables.

APPENDIX A. PROOFS IN SECTION 3

A.1 Proof of Theorem 3.1

To prove the additivity of the generalized information distance, we need the following lemma.

Table 3. The English translation of Table 2

Id	Keywords	Id	Keywords	Id	Keywords	Id	Keywords
1	police station	26	patrol	51	police	76	traffic light
2	traffic police	27	feed	52	suspect	77	gambling machine
3	municipal public security bureau	28	loudspeaker box	53	surveillance	78	commit a crime
4	policeman	29	everyday	54	keep a dog	79	policeman
5	dog	30	wound	55	Changchun public security bureau	80	a (an)
6	report	31	broadcast	56	forbid	81	fire fighting
7	big	32	gambling	57	police	82	on duty
8	call the police	33	the case	58	traffic police brigade	83	sound
9	mahjong parlor	34	punish	59	truck	84	advice
10	disturb residents	35	place	60	dispatch	85	parking
11	propagate	36	regulate the traffic	61	registered residence	86	fire facilities
12	public security sector	37	public security organization	62	one-way street	87	license plate
13	sound box	38	traffic signals	63	public security	88	drive
14	althorn	39	traffic police detachment	64	intersection	89	walk the dog
15	vehicle	40	fine	65	happen	90	noise nuisance
16	dog	41	identity card	66	yell	91	videotape
17	traffic	42	badge number	67	mahjong	92	strike
18	traffic jam	43	prostitution	68	ticket	93	organize gambling
19	feed	44	traffic departments	69	be forced	94	traffic jam
20	public security bureau	45	traffic accident	70	warn	95	city bureau
21	rule-breaking	46	tricycle	71	police force	96	turn left
22	signal lamp	47	horn	72	close a case	97	beat
23	case	48	barking	73	cry one's wares	98	have a rest
24	driver's license	49	handle a case	74	serious	99	pornographic
25	filing	50	travel	75	speed	100	play music

Lemma A.1. *Let U be a $d \times r$ column orthogonal matrix and let C be a $d \times r$ matrix with full column rank. If the range of C is contained in the range of U, there are two $r \times r$ orthogonal matrices P and Q such that $C = UP\Lambda Q^T$, where Λ is a diagonal matrix that comprises all the singular values of C.*

Proof. Since C is a $d \times r$ matrix with full column rank, there is an $r \times r$ orthogonal matrix Q such that $Q^T C^T C Q$ is a diagonal matrix with positive diagonal elements by spectral decomposition. We denote $Q^T C^T C Q$ by Λ^2, where Λ is a diagonal matrix that comprises all the singular values of C. Since the range of C is within the range of U, there exists an $r \times r$ matrix Y such that $CQ = UY$. Thus, $Y^T Y = \Lambda^2$ and $\Lambda^{-1} Y^T Y \Lambda^{-1} = I$. We denote $Y\Lambda^{-1}$ by P, where P is an $r \times r$ orthogonal matrix and $Y = P\Lambda$. Thus, $C = UP\Lambda Q^T$. □

Now, let us prove Theorem 3.1. When we discuss the additivity of generalized information distance along paths in the tree, there are actually three basic cases, i.e., paths between latent variables, paths between observed variables, and paths between an observed variable and a latent variable. The first case is the same as the additivity information distance [7] and the second case is similar to the third case. Thus, we provide the proof of the third case, as follows.

Let v be an observed variable and let h be a latent variable. Assume that there is a path $[v, y, h]$ in T. We can see that $d_{vy} + d_{yh} = d_{vh}$. Let us consider the expression $\frac{\prod_{s=1}^{r} \sigma_s(P_{vy})}{\sqrt{\det(P_{vv})\det(P_{yy})}} \times \frac{\prod_{s=1}^{r} \sigma_s(P_{yh})}{\sqrt{\det(P_{yy})\det(P_{hh})}}$, which equals $\frac{\prod_{s=1}^{r} \sigma_s(P_{vy}) \prod_{s=1}^{r} \sigma_s(P_{yh})}{\sqrt{\det(P_{vv})\det(P_{hh})}\det(P_{yy})}$. Since $P_{vy} = P_{v|y} P_{yy}$ and (A4), the rank of $P_{v|y}$ is r. According to singular value decomposition, there are two column orthogonal matrices P and Q such that

$$P_{vy} = P \begin{pmatrix} \sigma_1(P_{vy}) & & \\ & \cdots & \\ & & \sigma_r(P_{vy}) \end{pmatrix} Q^T,$$

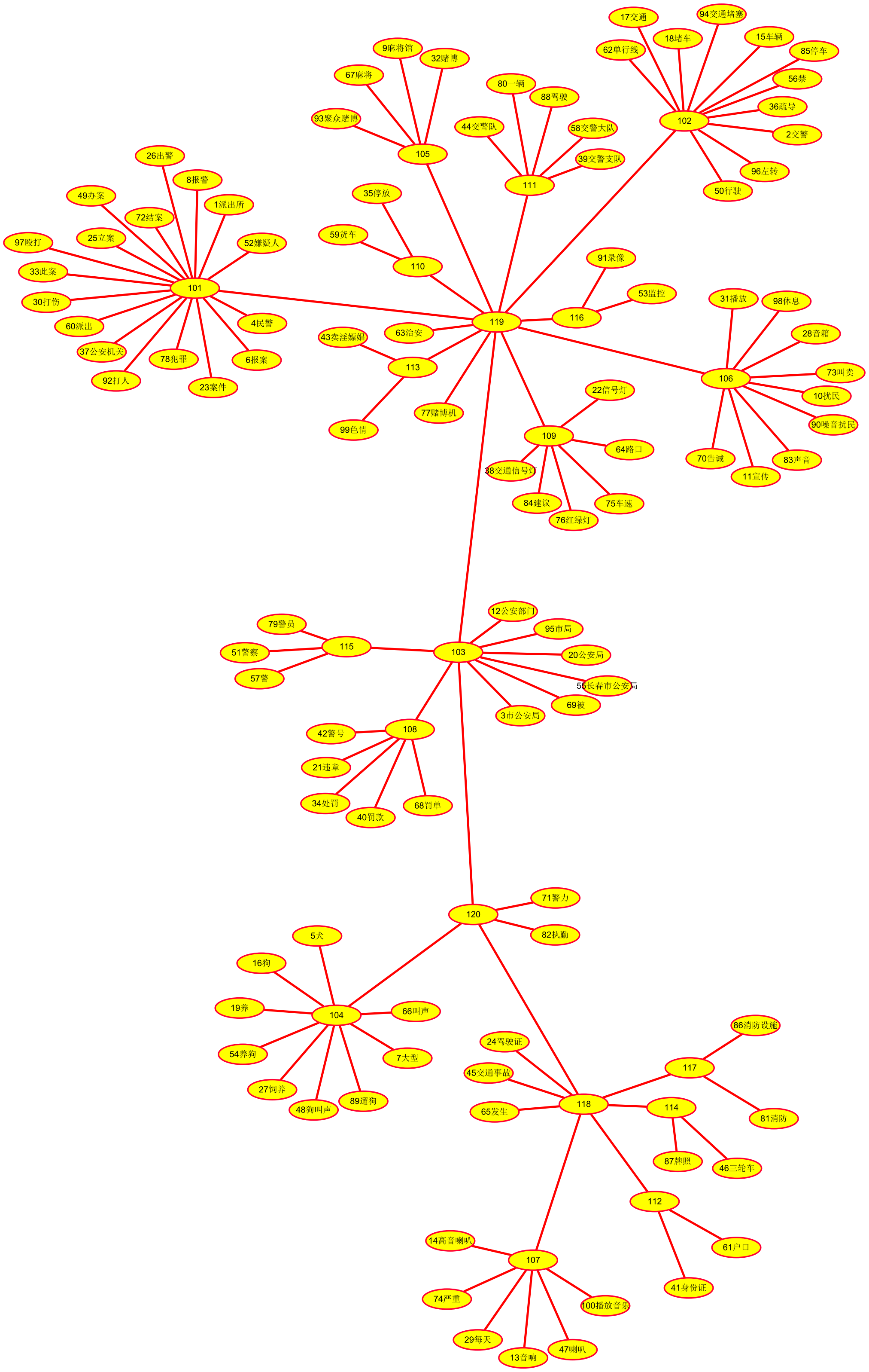


Figure 7. Latent tree structure from 100 keywords of the Public Security Bureau.

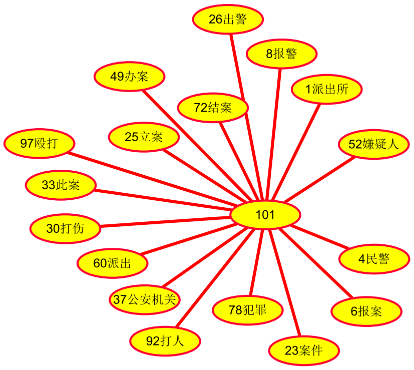


Figure 8. A common topic on crime.

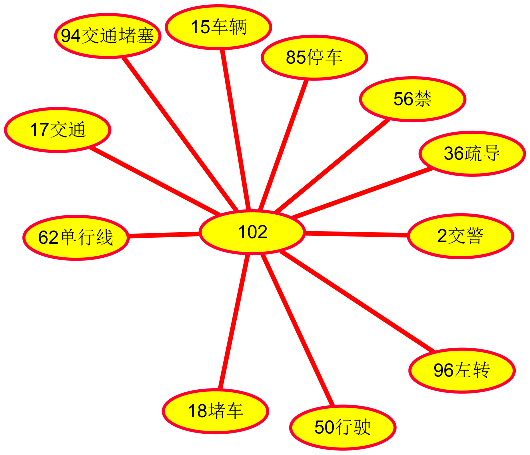


Figure 9. A common topic on traffic problem.

where P is a $d_v \times r$ matrix and Q is a $r \times r$ matrix. So

$$P_{vy} = P \begin{pmatrix} \sigma_1(P_{vy}) & & \\ & \cdots & \\ & & \sigma_r(P_{vy}) \end{pmatrix} Q^T = P_{v|y} P_{yy},$$

and

$$\begin{pmatrix} \sigma_1(P_{vy}) & & \\ & \cdots & \\ & & \sigma_r(P_{vy}) \end{pmatrix} = P^T P_{vy} Q = P^T P_{v|y} P_{yy} Q.$$

It can be verified that the range of $P_{v|y}$ is within the range of P. From Lemma A.1, we find that $\prod_{s=1}^{r} \sigma_s(P_{vy}) = \prod_{s=1}^{r} \sigma_s(P_{v|y}) \det(P_{yy})$. Thus,

$$\frac{\prod\limits_{s=1}^{r} \sigma_s(P_{vy}) \prod\limits_{s=1}^{r} \sigma_s(P_{yh})}{\sqrt{\det(P_{vv}) \det(P_{hh}) \det(P_{yy})}} = \frac{\prod\limits_{s=1}^{r} \sigma_s(P_{v|y}) \prod\limits_{s=1}^{r} \sigma_s(P_{yh})}{\sqrt{\det(P_{vv}) \det(P_{hh})}}.$$

Based on the Markov property of latent tree models, we find that $P_{vh} = P_{v|y} P_{yh}$. As discussed above, we also find that $\prod_{s=1}^{r} \sigma_s(P_{vh}) = \prod_{s=1}^{r} \sigma_s(P_{v|y}) \prod_{s=1}^{r} \sigma_s(P_{yh})$ by Lemma A.1. Thus, we find that:

$$\frac{\prod\limits_{s=1}^{r} \sigma_s(P_{vy})}{\sqrt{\det(P_{vv}) \det(P_{yy})}} \frac{\prod\limits_{s=1}^{r} \sigma_s(P_{yh})}{\sqrt{\det(P_{yy}) \det(P_{hh})}}$$
$$= \frac{\prod\limits_{s=1}^{r} \sigma_s(P_{vh})}{\sqrt{\det(P_{vv}) \det(P_{hh})}}.$$

Therefore, $d_{vy} + d_{yh} = d_{vh}$. If there is a much longer path between v and h in T, the proof of the third case is similar. Hence, we have completed the proof of Theorem 3.1.

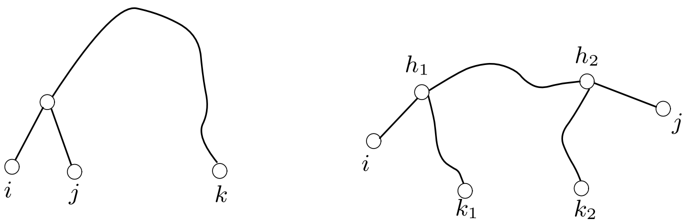


Figure 10. Structure relations between two observed variables i, j.

A.2 Proof of Theorem 3.2

To prove the correctness of the SLLT algorithm, we need the following Lemmas A.2, A.3, and A.4 to describe the characteristics of the generalized information distance difference with various local structure relations. The corresponding actual cases are shown in Figures 10, 11, and 12, where a straight line represents an edge and a curve denotes a path. The proofs of these lemmas can be obtained directly by Theorem 3.1, so we omit them.

Lemma A.2. *Given two observed variables i, j, $\{i, j\}$ is a sibling pair if and only if Φ_{ijk} is constant for any observed variable $k \neq i, j$.*

From Lemma A.2, we know that Step 2 of the SLLT algorithm can find sibling groups among the observed variables. When we consider the relationship between an observed variable v and a latent variable u, we only need to judge whether there is a sibling pair relationship. Indeed, every latent variable is generated by a sibling group in the SLLT algorithm so if v is a remaining child node of u, when the sibling pair $\{x, y\}$ generates u, $\{v, x\}$ is also a sibling pair.

Lemma A.3. *For an observed variable v and a latent variable u, assume that i, j are two bifurcation variables of u. If the relationship between v and u is shown in Fig. 11 (a), then:*
(1) Φ_{vim} is constant and $\Phi_{vim} \neq \Phi_{vij}$ for any observed variable m, as shown in Fig. 11 (a).
If the relationship between v and u is shown in Fig. 11 (b),

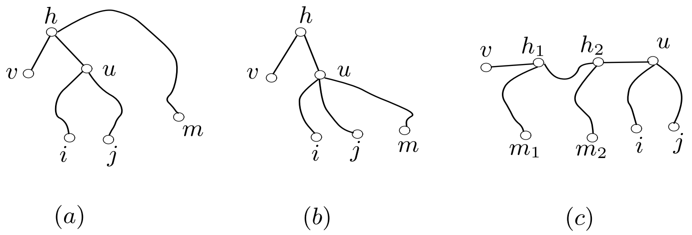


Figure 11. Structure relations between observed variable v and latent variable u.

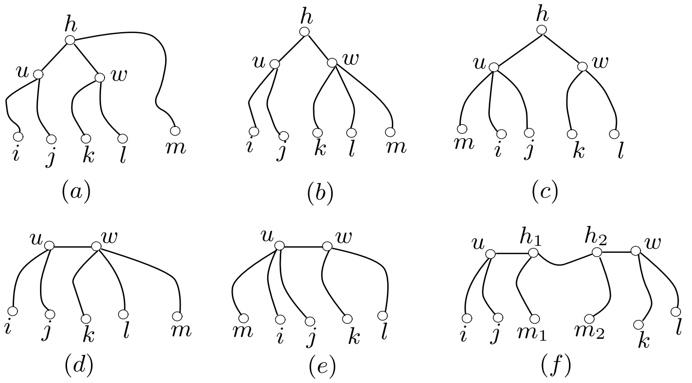


Figure 12. Structure relations between two latent variables u, w.

then:
(2) $\Phi_{vim} = \Phi_{vij}$ for any observed variable m, as shown in Fig. 11 (b).
If the relation between v and u is shown in Fig. 11 (c), then:
(3) $\Phi_{vim_1} \neq \Phi_{vim_2}$ for any observed variables m_1, m_2, as shown in Fig. 11 (c).

In 1^0 of Step 3, for any $v \in Y \bigcap V$ and $u \in Y \setminus V$, select the bifurcation variables i, j of u in D. We want to judge the relation between v and u in $T(W \setminus D)$. We assume that $T(W \setminus D)$ is a subtree of T and Y contains all the leaf nodes in $T(W \setminus D)$. Since $T(W \setminus D)$ is a subtree of T, there is a path of length 2, or more than 2, between u and v in $T(W \setminus D)$.

If Φ_{vim} is constant and $\Phi_{vim} \neq \Phi_{vij}$ for any $m \in V \setminus (D(u) \bigcup \{v\})$, then any path between u and v in $T(W \setminus D)$ has a length of 2. Otherwise, if there is a path of length more than 2 in $T(W \setminus D)$ between v and u, this case is similar to that shown in Fig. 11 (c) because every latent variable has at least three neighbors. Thus, two observed variables $m_1, m_2 \in V \setminus (D(u) \bigcup \{v\})$ exist such that $\Phi_{vim_1} \neq \Phi_{vim_2}$, which is a contradiction. Therefore, any path between u and v in $T(W \setminus D)$ has a length of 2. Furthermore, u is a leaf node of $T(W \setminus D)$. Otherwise, a path in $T(W \setminus D)$ passes through u between a leaf node w of $T(W \setminus D)$ and v. We know that $w \in Y$ because Y contains all of the leaf nodes in $T(W \setminus D)$. Since $w, u \in Y$, $D(w) \bigcap D(u) = \emptyset$. We can select an observed variable $m' \in D(w) \bigcup \{w\}$. Hence, $m' \in V \setminus (D(u) \bigcup \{v\})$, and this case is similar to that shown in Fig. 11 (b). Thus, $\Phi_{vim'} = \Phi_{vij}$, which is a contradiction. Therefore, u is also a leaf node of $T(W \setminus D)$ and $\{v, u\}$ is a sibling pair in $T(W \setminus D)$.

Lemma A.4. *For two latent variable u, w, assume that i, j are two bifurcation variables of u and k, l are two bifurcation variables of w.*
If the relation between v and u is shown in Fig. 12 (a), then:
(1) Φ_{ikm} is constant and $\Phi_{ikm} \neq \Phi_{ikl}, \Phi_{kim} \neq \Phi_{kij}$ for any observed variable m as shown in Fig. 12 (a).
If the relation between v and u is shown in Fig. 12 (b), then:
(2) $\Phi_{ikm} = \Phi_{ikl}$ and $\Phi_{kim} \neq \Phi_{kij}$ for any observed variable m in Fig. 12 (b).
If the relation between v and u is shown in Fig. 12 (c), then:
(3) $\Phi_{ikm} \neq \Phi_{ikl}$ and $\Phi_{kim} = \Phi_{kij}$ for any observed variable m in Fig. 12 (c).
If the relation between v and u is shown in Fig. 12 (d), then:
(4) $\Phi_{ikm} = \Phi_{ikl}$ and $\Phi_{kim} \neq \Phi_{kij}$ for any observed variable m in Fig. 12 (d).
If the relation between v and u is shown in Fig. 12 (e), then:
(5) $\Phi_{ikm} \neq \Phi_{ikl}$ and $\Phi_{kim} = \Phi_{kij}$ for any observed variable m in Fig. 12 (e).
If the relation between v and u is shown in Fig. 12 (f), then:
(6) $\Phi_{ikm_1} \neq \Phi_{ikm_2}$ for any observed variables m_1, m_2 in Fig. 12 (f).

In 2^0 of Step 3, for any $u, w \in Y \setminus V$, select the bifurcation variables i, j of u and k, l of w in D. We need to judge the relationship between u and w in $T(W \setminus D)$. We assume that $T(W \setminus D)$ is a subtree of T and that Y contains all the leaf nodes in $T(W \setminus D)$. Since $T(W \setminus D)$ is a subtree of T, there is a path of length 1, or more than 1, between u and w in $T(W \setminus D)$.

If Φ_{ikm} is constant and $\Phi_{ikm} \neq \Phi_{ikl}, \Phi_{kim} \neq \Phi_{kij}$ for any $m \in V \setminus (D(u) \bigcup D(w))$, then any path between u and w in $T(W \setminus D)$ has a length of no more than 2. Otherwise, if there is a path of a length more than 2 in $T(W \setminus D)$ between u and w, this case is similar to that shown in Fig. 12 (f) because every latent variable has at least three neighbors. Thus, there are two observed variables $m_1, m_2 \in V \setminus (D(u) \bigcup D(w))$ such that $\Phi_{ikm_1} \neq \Phi_{ikm_2}$. This is a contradiction. Therefore, any path between u and w in $T(W \setminus D)$ has a length of no more than 2. Furthermore, w is a leaf node of $T(W \setminus D)$. Otherwise, a path in $T(W \setminus D)$ passes through w between a leaf node w_1 of $T(W \setminus D)$ and u. In addition, we know that $w_1 \in Y$ since Y contains all of the leaf nodes in $T(W \setminus D)$. Since $w_1, u, w \in Y$, $D(w_1) \bigcap D(u) = \emptyset$ and $D(w_1) \bigcap D(w) = \emptyset$. We can select an observed variable $m' \in D(w_1) \bigcup \{w_1\}$. Therefore, $m' \in V \setminus (D(u) \bigcup D(w))$ and this case is similar to that shown in Fig. 12 (b) or (d). Thus, $\Phi_{ikm'} = \Phi_{ikl}$, which is a contradiction. Similarly, u is also a leaf node of $T(W \setminus D)$. If u is adjacent to w, then u or w is not a leaf node in $T(W \setminus D)$ when $|Y| \geqslant 3$, which is a contradiction. Therefore, $\{u, w\}$ is a sibling pair in $T(W \setminus D)$.

If $\Phi_{ikm} = \Phi_{ikl}$ and $\Phi_{kim} \neq \Phi_{kij}$ for any $m \in V \setminus (D(u) \bigcup D(w))$, then any path between u and w in $T(W \setminus D)$

has a length of no more than 2. Otherwise, if there is a path of a length more than 2 in $T(W \setminus D)$ between u and w, this case is similar to that shown in Fig. 12 (f) because every latent variable has at least three neighbors. Thus, there are two observed variables $m_1, m_2 \in V \setminus (D(u) \bigcup D(w))$ such that $\Phi_{ikm_1} \neq \Phi_{ikm_2}$, which is a contradiction. Therefore, any path between u and w in $T(W \setminus D)$ has a length of no more than 2. Furthermore, u is a leaf node of $T(W \setminus D)$. Otherwise, a path in $T(W \setminus D)$ passes through u between a leaf node u_1 of $T(W \setminus D)$ and w. We know that $u_1 \in Y$ since Y contains all of the leaf nodes in $T(W \setminus D)$. Thus, $u_1, u, w \in Y$, $D(u_1) \bigcap D(u) = \emptyset$ and $D(u_1) \bigcap D(w) = \emptyset$. We can select an observed variable $m' \in D(u_1) \bigcup \{u_1\}$. Hence, $m' \in V \setminus (D(u) \bigcup D(w))$ and this case is similar to that shown in Fig. 12 (c) or (e). Thus, $\Phi_{ikm'} \neq \Phi_{ikl}$ and $\Phi_{kim'} = \Phi_{kij}$, which is a contradiction. Therefore, u is a leaf node of $T(W \setminus D)$. If there is a path with a length of 2 in $T(W \setminus D)$ between v and u, this case is similar to that shown in Fig. 12 (a) because every latent variable has at least three neighbors. Thus, an observed variable $m' \in V \setminus (D(u) \bigcup D(w))$ exists such that $\Phi_{ikm'} \neq \Phi_{ikl}$, which is a contradiction. Therefore, u is a remaining child of w in $T(W \setminus D)$.

To prove the SLLT algorithm, we use the discrete time record $t = 0, 1, 2, \cdots$ to describe the iteration process when $|Y| \geq 3$.

When $t = 0$, $D_0 = \emptyset$ and Y_0 is the set V of all the leaves of $T(W)$. From Lemma A.2, Step 2 can find all the sibling groups in T correctly. The updated set D_1 that comprises all the maximal sibling groups in T is a leaf subset of T. Hence, $T(W \setminus D_1)$ is a subtree of T and all the leaves of $T(W \setminus D_1)$ are contained in Y_1. According to Lemma A.2, the structure relation among D_1 is true and the adjacent relation between D_1 and $W \setminus D_1$ is also true.

When $t = 1$, $T(W \setminus D_1)$ is a subtree of T and all the leaves of $T(W \setminus D_1)$ are contained in Y_1. According to Lemmas A.3 and A.4, Step 3 can correctly find all the sibling groups and remaining child relations in $T(W \setminus D_1)$. The difference subset $D_2 \setminus D_1$ is a leaf subset of $T(W \setminus D_1)$. Hence, $T(W \setminus D_2)$ is a subtree of T and all of the leaves of $T(W \setminus D_2)$ are contained in Y_2. According to Lemmas A.3 and A.4, the structure relation among $D_2 \setminus D_1$ is true and the adjacent relation between $D_2 \setminus D_1$ and $W \setminus D_2$ is also true. The structure relation among D_1 is true and the adjacent relation between D_1 and $W \setminus D_1$ is also true, so the adjacent relation between D_2 and $W \setminus D_2$ is true and the structure relation among D_2 is true.

When $t = s$, $T(W \setminus D_s)$ is a subtree of T and all of the leaves of $T(W \setminus D_s)$ are contained in Y_s. According to Lemmas A.3 and A.4, Step 3 can correctly find all the sibling groups and remaining child relations in $T(W \setminus D_s)$. The difference subset $D_{s+1} \setminus D_s$ is a leaf subset of $T(W \setminus D_s)$. Hence, $T(W \setminus D_{s+1})$ is a subtree of T and all of the leaves of $T(W \setminus D_{s+1})$ are contained in Y_{s+1}. According to Lemmas A.3 and A.4, the structure relation among $D_{s+1} \setminus D_s$ is true. The adjacent relation between $D_{s+1} \setminus D_s$ and $W \setminus D_{s+1}$ is also true. The structure relation among D_s is true and the adjacent relation between D_s and $W \setminus D_s$ is also true, thus the adjacent relation between D_{s+1} and $M \setminus D_{s+1}$ is true, and the structure relation among D_{s+1} is true.

As discussed above, our SLLT algorithm can learn the true latent tree structure by recursive reconstruction. The computational complexity is bounded by $O(diam(T)n^3)$, as described by [7]. Hence, we have completed the proof of Theorem 3.2.

A.3 Proof of Theorem 3.3

In this section, we give the proof of Theorem 3.3. First, we introduce the following proposition A.1 from [13], which gives the tail bound inequality of the difference between the empirical estimate and the true probability.

For three observed variables $x, y, z \in V$, let P_{xyz} denote the third-order joint probability tensor $(Pr(x = x_i, y = y_j, z = z_k))_{1 \leq i \leq d_x, 1 \leq j \leq d_y, 1 \leq k \leq d_z}$, and let $\hat{P}_{xyz}$ denote its empirical estimation $(\frac{1}{N}\sum_{l=1}^{N} \mathbb{I}(x^{(l)} = x_i, y^{(l)} = y_j, z^{(l)} = z_k))_{1 \leq i \leq d_x, 1 \leq j \leq d_y, 1 \leq k \leq d_z}$, where N is the sample size.

Proposition A.1. *For any three observed variables $x, y, z \in V$, we find that for any $t > 0$, $Pr(\|\hat{P}_{xy} - P_{xy}\|_F > \frac{1+\sqrt{t}}{\sqrt{N}}) \leq e^{-t}$ and $Pr(\|\hat{P}_{xyz} - P_{xyz}\|_F > \frac{1+\sqrt{t}}{\sqrt{N}}) \leq e^{-t}$, where $\| \bullet \|_F$ is the Frobenius norm.*

The following lemma describes the tail bound inequality on the maximum singular value of $\hat{P}_{xy} - P_{xy}$.

Lemma A.5. *For any observed variables $x, y \in V$, we find that for any $t > 0$, $Pr(\sigma_1(\hat{P}_{xy} - P_{xy}) > \frac{1+\sqrt{t}}{\sqrt{N}}) \leq e^{-t}$.*

Proof. Since $\sigma_1(\hat{P}_{xy} - P_{xy})$ is no more than $\|\hat{P}_{xy} - P_{xy}\|_F$, this lemma is obtained from Proposition A.1. □

To prove Theorem 3.3, we only need to show that if the sample size is sufficiently large, the probability of the event $\{|(\hat{\Phi}_{xyz} - \hat{\Phi}_{xyw}) - (\Phi_{xyz} - \Phi_{xyw})| < \epsilon$ for any $x, y, z, w \in V\}$ is sufficiently large. According to the definition of Φ, we only need to show that when the sample size is sufficiently large, the probability of the event $\{|\hat{d}_{xy} - d_{xy}| < \frac{1}{4}\epsilon$ for any $x, y \in V\}$ is sufficiently large.

In the following, we show how to make $|\hat{d}_{xy} - d_{xy}| < \frac{1}{4}\epsilon$. For any variables $x, y \in V$, we consider the generalized information distance d_{xy}. Since $d_{xy} = -\sum_{s=1}^{r} \log \sigma_s(P_{xy}) + \frac{1}{2}\sum_{s=1}^{d_x} \log \sigma_s(P_{xx}) + \frac{1}{2}\sum_{s=1}^{d_y} \log \sigma_s(P_{yy})$, hence $|\hat{d}_{xy} - d_{xy}| < \sum_{s=1}^{r} |\log \sigma_s(\hat{P}_{xy}) - \log \sigma_s(P_{xy})| + \frac{1}{2}\sum_{s=1}^{d_x} |\log \sigma_s(\hat{P}_{xx}) - \log \sigma_s(P_{xx})| + \frac{1}{2}\sum_{s=1}^{d_y} |\log \sigma_s(\hat{P}_{yy}) - \log \sigma_s(P_{yy})|$.

We denote σ_{min} as $\min_{x,y \in V} \sigma_{rank(P_{xy})}(P_{xy})$ (this allows $x = y$). If $\Delta > 0$ exists such that $\Delta < \frac{1}{2}\sigma_{min}$ and $|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| \leq \Delta$ for any $1 \leq s \leq rank(P_{xy})$ and any $x, y \in V$, we have $\sigma_s(\hat{P}_{xy}) > \sigma_s(P_{xy}) - |\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| > \frac{1}{2}\sigma_{min}$. Since $\sigma_s(P_{xy}), \sigma_s(\hat{P}_{xy}) > \frac{1}{2}\sigma_{min}$, then $|\log \sigma_s(\hat{P}_{xy}) - \log \sigma_s(P_{xy})| < \frac{2}{\sigma_{min}}|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})|$. Furthermore, we denote d_{max} as $\max_{x\in V} d_x$. We have $|\hat{d}_{xy} - d_{xy}| < \frac{4d_{max}\Delta}{\sigma_{min}}$. Since $\epsilon \leq 1$, we can see that $\Delta < \frac{\sigma_{min}\epsilon}{16d_{max}}$ implies $\Delta < \frac{1}{2}\sigma_{min}$. Thus, if a suitable Δ exists such that $\Delta < \frac{\sigma_{min}\epsilon}{16d_{max}}$ and $|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| \leq \Delta$ for $1 \leq s \leq rank(P_{xy})$, then $|\hat{d}_{xy} - d_{xy}| < \frac{1}{4}\epsilon$.

We show how to select Δ such that $\Delta < \frac{\sigma_{min}\epsilon}{16d_{max}}$ and $Pr(\bigcap_{x,y,s}(|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| \leq \Delta))$ is sufficiently large. If we can shows this successfully, we complete the proof of Theorem 3.3. According to Lemma A.5, for any observed variables $x, y \in V$, we find that for any $t > 0$,

$$Pr\left(\sigma_1(\hat{P}_{xy} - P_{xy}) \leq \frac{\sqrt{t}+1}{\sqrt{N}}\right) > 1 - e^{-t}.$$

Based on standard matrix analysis theory, we know that $|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| < \sigma_1(\hat{P}_{xy} - P_{xy})$ for any $x, y \in V$ and any $1 \leq s \leq rank(P_{xy})$. Thus, $Pr(|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| \leq \frac{\sqrt{t}+1}{\sqrt{N}}) > 1 - e^{-t}$ for any $x, y \in V$ and any $1 \leq s \leq rank(P_{xy})$. Therefore, for any $t > 0$, we have:

$$\begin{aligned} & Pr\left(\bigcap_{x,y,s}(|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| \leq \frac{\sqrt{t}+1}{\sqrt{N}})\right) \\ = \quad & 1 - Pr\left(\bigcup_{x,y,s}(|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| > \frac{\sqrt{t}+1}{\sqrt{N}})\right) \\ \geq \quad & 1 - \sum_{x,y,s} Pr\left(|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| > \frac{\sqrt{t}+1}{\sqrt{N}}\right) \\ \geq \quad & 1 - n^2 d_{max} e^{-t} \triangleq 1 - \eta, \end{aligned}$$

where $\eta \in (0, 1)$. Hence, $\eta = n^2 d_{max} e^{-t}$ and we denote t_0 as $-\log \frac{\eta}{n^2 d_{max}}$. Thus, $Pr(\bigcap_{x,y,s}(|\sigma_s(\hat{P}_{xy}) - \sigma_s(P_{xy})| \leq \frac{\sqrt{t_0}+1}{\sqrt{N}})) \geq 1 - \eta$. Therefore, we can select $\Delta = \frac{\sqrt{t_0}+1}{\sqrt{N}}$. For any η, if the sample size N is sufficiently large such that $\Delta < \frac{\sigma_{min}\epsilon}{16d_{max}}$, then our SLLT algorithm learns the true latent tree structure with a probability of at least $1 - \eta$.

APPENDIX B. PROOFS IN SECTION 4

B.1 Proof of Theorem 4.1

In this section, we give the proof of Theorem 4.1. For a square matrix $A \in \mathbb{R}^{k\times k}$, if A is an invertible matrix, we denote the condition number $\|A\|\|A^{-1}\|$ as $\kappa(A)$. For two square matrices $A, B \in \mathbb{R}^{k\times k}$, we denote the class $\{\lambda_1, \cdots, \lambda_k\}$ of all the eigenvalues of A as $e(A)$, and we denote the class $\{\mu_1, \cdots, \mu_k\}$ of all the eigenvalues of B as $e(B)$. We define the optimal matching distance between $e(A)$ and $e(B)$ as $d(e(A), e(B)) = \min_\tau \max_{1\leq j\leq k} |\lambda_j - \mu_{\tau(j)}|$, where τ is a permutation. From this definition, if $d(e(A), e(B)) \leq t$, a permutation τ exists such that $|\lambda_j - \mu_{\tau(j)}| \leq t$ for $j = 1, \cdots, k$. We define $\mathbb{1}_k$ as a k-dimensional vector where each coordinate is 1. If the context does not require the distinction of subscript k, we also use the symbol $\mathbb{1}$.

First, we introduce a lemma regarding the spectral variation of diagonalizable matrices.

Lemma B.1. *Assume that $A \in \mathbb{R}^{k\times k}$ is a diagonalizable matrix and that $A = SDS^{-1}$, where D is a diagonal matrix and S is an invertible matrix. We define $r_A = \min_{i\neq j}\{|D_{ii} - D_{jj}| : D_{ii} \neq D_{jj}\}$. If $B \in \mathbb{R}^{k\times k}$ is a matrix such that $\kappa(S)\|A - B\| < \frac{1}{2}r_A$, then $d(e(A), e(B)) \leq \kappa(S)\|A - B\|$.*

Proof. When combined with the Bauer-Fike Theorem, this lemma can be verified by a similar proof as Theorem VI.5.1 in [4] based on continuity arguments. □

Next, we present a lemma regarding the perturbation of eigenvalues and eigenvectors.

Lemma B.2. *Assume that $A \in \mathbb{R}^{k\times k}$ is a diagonalizable matrix with k distinct real eigenvalues $\lambda_1, \cdots, \lambda_k \in \mathbb{R}$. The corresponding eigenvectors $x_1, \cdots, x_k$ satisfy $\mathbb{1}^T x_i = 1, 1 \leq i \leq k$, and every element in the matrix $S = (x_1, \cdots, x_k)$ is nonnegative. Let $\hat{A} \in \mathbb{R}^{k\times k}$ be a matrix. Define $\varepsilon_A = \|\hat{A} - A\|$, $r_A = \min_{i\neq j}\{|\lambda_i - \lambda_j| : \lambda_i \neq \lambda_j\}$, $\alpha_A = \max_i\{\|\begin{pmatrix} A - \lambda_i I \\ \mathbb{1}^T \end{pmatrix}^+\|\}$. If $\kappa(S)\varepsilon_A < \frac{1}{2}r_A$, we reach the following conclusions.*
1. $\hat{A}$ has k distinct real eigenvalues $\hat{\lambda}_1, \cdots, \hat{\lambda}_k \in \mathbb{R}$ such that for any $1 \leq i \leq k$, $|\hat{\lambda}_i - \lambda_i| \leq \kappa(S)\varepsilon_A$.
2. Assume that the corresponding eigenvectors of $\hat{A}$ are $\hat{x}_1, \cdots, \hat{x}_k$, which satisfy $\mathbb{1}^T\hat{x}_i = 1$ for any $1 \leq i \leq k$. If $\alpha_A(1 + \kappa(S))\varepsilon_A < 1$, then $\|\hat{x}_i - x_i\| \leq \frac{\alpha_A(1+\kappa(S))\varepsilon_A}{1-\alpha_A(1+\kappa(S))\varepsilon_A}$ for any $1 \leq i \leq k$.

Proof. First, we give the proof of conclusion 1. From Lemma B.1, since A is diagonalizable, for any eigenvalue λ_i of A, there is an eigenvalue $\hat{\lambda}$ of $\hat{A}$ such that $|\hat{\lambda} - \lambda_i| < \kappa(S)\varepsilon_A < \frac{1}{2}r_A$. Thus, these eigenvalues of $\hat{A}$ are distinct and we can order them as $\hat{\lambda}_1, \cdots, \hat{\lambda}_k$ such that $|\hat{\lambda}_i - \lambda_i| < \kappa(S)\varepsilon_A < \frac{1}{2}r_A$. Since $\hat{A}$ is a real matrix, if $\hat{A}$ has a complex eigenvalue λ, then $\hat{A}x = \lambda x$ and $\hat{A}\overline{x} = \overline{\hat{A}x} = \overline{\lambda x}$. Therefore, $\overline{\lambda}$ is also an eigenvalue of $\hat{A}$. Since all the eigenvalues of A are real, there is an eigenvalue λ_i of A such that $|\lambda_i - \lambda| = |\lambda_i - \overline{\lambda}| < \frac{1}{2}r_A$, which is a contradiction. Thus all of the eigenvalues of $\hat{A}$ are real.

Next, we provide the proof of conclusion 2. Since x_i is an eigenvector of A that satisfies $\mathbb{1}^T x_i = 1$, x_i is the solution of the joint equations

$$\begin{pmatrix} A - \lambda_i I \\ \mathbb{1}^T \end{pmatrix} x_i = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Similarly, $\hat{x}_i$ is the solution of the joint equations

$$\begin{pmatrix} \hat{A} - \hat{\lambda}_i I \\ \mathbb{1}^T \end{pmatrix} \hat{x}_i = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Denote $\begin{pmatrix} A - \lambda_i I \\ \mathbb{1}^T \end{pmatrix}$ as B_i and $\begin{pmatrix} \hat{A} - \hat{\lambda}_i I \\ \mathbb{1}^T \end{pmatrix}$ as $\hat{B}_i$. It can be verified that both B_i and $\hat{B}_i$ have full column rank.

From Theorem 5.1 in [24], we can obtain the perturbation bound of x_i as $\|\hat{x}_i - x_i\| \leq \frac{\|B_i^+\|\|\hat{B}_i - B_i\|}{1-\|B_i^+\|\|\hat{B}_i - B_i\|}\|x_i\| \leq \frac{\|B_i^+\|(1+\kappa(S))\varepsilon_A}{1-\|B_i^+\|(1+\kappa(S))\varepsilon_A}\|x_i\|$ when $\|B_i^+\|(1+\kappa(S))\varepsilon_A < 1$. Since every element of x_i is nonnegative and $\mathbb{1}^T x_i = 1$, we know that $\|x_i\| \leq 1$. Thus, if $\alpha_A(1+\kappa(S))\varepsilon_A < 1$, $\|\hat{x}_i - x_i\| \leq \frac{\alpha_A(1+\kappa(S))\varepsilon_A}{1-\alpha_A(1+\kappa(S))\varepsilon_A}$ for any $1 \leq i \leq k$. □

In the following, we assume that all of the variables are two-state. We focus mainly on the 2×2 matrix $A = P_{bc}^{a=a_1} P_{bc}^{-1}$ when P_{bc} is invertible. As discussed in Section 4.1, A has a decomposition $P_{b|h}\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}P_{b|h}^{-1}$, where $\lambda_1 = Pr(a = a_1|h = h_1)$ and $\lambda_2 = Pr(a = a_1|h = h_2)$. Let x_1, x_2 be the two eigenvectors of A that correspond to λ_1 and λ_2, which satisfy $\mathbb{1}^T x_i = 1$ for $i = 1, 2$. Thus, $x_1 = (Pr(b = b_1|h = h_1), Pr(b = b_2|h = h_1))^T$ and $x_2 = (Pr(b = b_1|h = h_2), Pr(b = b_2|h = h_2))^T$. We present the following lemma to set an upper bound on α_A in Lemma B.2.

Lemma B.3. *Assume that the 2×2 matrix A is defined as above. If $\lambda_1 \neq \lambda_2$, we have $\left\|\begin{pmatrix} A - \lambda_i I \\ \mathbb{1}^T \end{pmatrix}^+\right\| \leq \frac{3}{|\lambda_1 - \lambda_2|}$ for $i = 1, 2$.*

Proof. For $i = 1, 2$, we denote $\begin{pmatrix} A - \lambda_i I \\ \mathbb{1}^T \end{pmatrix}$ as B_i. Since B_i has full column rank, we have $\|B_i^+\| = \frac{1}{\sigma_2(B_i)} = \frac{1}{(\lambda_2(B_i^T B_i))^{\frac{1}{2}}}$. In the following, we prove that $\lambda_2(B_1^T B_1) \geq \frac{(\lambda_1-\lambda_2)^2}{9}$. The case of $\lambda_2(B_2^T B_2)$ is similar.

We denote the matrix $A - \lambda_1 I$ as C and denote the matrix $C^T C$ as D. Since A has a decomposition $P_{b|h}\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}P_{b|h}^{-1}$, the matrix C has the form of

$$\frac{1}{p_1 - p_2}\begin{pmatrix} p_1 & p_2 \\ 1-p_1 & 1-p_2 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & \lambda_2 - \lambda_1 \end{pmatrix}\begin{pmatrix} 1-p_2 & -p_2 \\ -(1-p_1) & p_1 \end{pmatrix},$$

where $p_1 = Pr(b = b_1|h = h_1)$ and $p_2 = Pr(b = b_1|h = h_2)$. It can be verified that the two eigenvalues $\lambda_1(D) = \frac{(\lambda_1-\lambda_2)^2(p_1^2+(1-p_1)^2)(p_2^2+(1-p_2)^2)}{(p_1-p_2)^2}$ and $\lambda_2(D) = 0$. The eigenvector $\xi_2(D) = (\frac{p_1}{\sqrt{p_1^2+(1-p_1)^2}}, \frac{1-p_1}{\sqrt{p_1^2+(1-p_1)^2}})^T$, which corresponds to $\lambda_2(D)$.

We define $y = \xi_2(D)^T\mathbb{1}$, thus $y = \frac{1}{\sqrt{p_1^2+(1-p_1)^2}}$. From Theorem 3.3 in [9], which provides the lower bound of the smallest eigenvalue of the rank one updates, we obtain $\lambda_2(B_1^T B_1) \geq y^2\frac{\lambda_1(D)}{\lambda_1(D)+\nu^2}$, where $\nu = y + \sqrt{2-y^2}$. Furthermore, we have

$$\begin{aligned} y^2\frac{\lambda_1(D)}{\lambda_1(D)+\nu^2} &\geq \frac{(p_2^2+(1-p_2)^2)}{(p_1^2+(1-p_1)^2)(p_2^2+(1-p_2)^2)+4\frac{(p_1-p_2)^2}{(\lambda_1-\lambda_2)^2}} \\ &\geq \frac{(p_2^2+(1-p_2)^2)}{(p_2^2+(1-p_2)^2)+\frac{4}{(\lambda_1-\lambda_2)^2}} \\ &\geq \frac{(\lambda_1-\lambda_2)^2}{(\lambda_1-\lambda_2)^2+\frac{4}{(p_2^2+(1-p_2)^2)}} \geq \frac{(\lambda_1-\lambda_2)^2}{9}, \end{aligned}$$

where the inequalities follow from $\frac{1}{2} \leq p_1^2+(1-p_1)^2 \leq 1$, $\frac{1}{2} \leq p_2^2+(1-p_2)^2 \leq 1$, $(p_1-p_2)^2 \leq 1$ and $(\lambda_1-\lambda_2)^2 \leq 1$. Thus, we obtain $\left\|\begin{pmatrix} A - \lambda_i I \\ \mathbb{1}^T \end{pmatrix}^+\right\| \leq \frac{3}{|\lambda_1-\lambda_2|}$ for $i = 1, 2$. □

If the empirical estimate $\hat{P}_{bc}$ of P_{bc} is invertible, we can define the empirical estimation $\hat{A} = \hat{P}_{bc}^{a=a_1}\hat{P}_{bc}^{-1}$ of A, and denote the two eigenvalues of $\hat{A}$ as $\hat{\lambda}_1, \hat{\lambda}_2$. Furthermore, if $\hat{\lambda}_1 \neq \hat{\lambda}_2$, $\hat{A}$ has two eigenvectors $\hat{x}_1$ and $\hat{x}_2$ that correspond to $\hat{\lambda}_1$ and $\hat{\lambda}_2$, which also satisfy $\mathbb{1}^T\hat{x}_i = 1$ for $i = 1, 2$.

Corollary B.1. *We define $\varepsilon_A = \|\hat{A} - A\|$. If P_{bc} and $\hat{P}_{bc}$ are invertible, the two following conclusions can be made:*
1. If $\lambda_1 = \lambda_2$, then $|\hat{\lambda}_i - \lambda_i| \leq \kappa(P_{b|h})\varepsilon_A$ for $i = 1, 2$.
2. If $\lambda_1 \neq \lambda_2$ and $\frac{6}{|\lambda_1-\lambda_2|}(1+\kappa(P_{b|h}))\varepsilon_A < 1$, then:
$\hat{A}$ has two distinct real eigenvalues $\hat{\lambda}_1$ and $\hat{\lambda}_2$, such that $|\hat{\lambda}_1 - \hat{\lambda}_2| > \frac{1}{2}|\lambda_1 - \lambda_2|$ and $|\hat{\lambda}_i - \lambda_i| < \kappa(P_{b|h})\varepsilon_A$, $\|\hat{x}_i - x_i\| \leq \frac{6}{|\lambda_1-\lambda_2|}(1+\kappa(P_{b|h}))\varepsilon_A$ for $i = 1, 2$.

Proof. The first conclusion can be obtained directly from the Bauer-Fike Theorem. The second conclusion can also be obtained from Lemmas B.2 and B.3 because $\kappa(P_{b|h})\varepsilon_A < \frac{1}{4}|\lambda_1 - \lambda_2|$ when $\frac{6}{|\lambda_1-\lambda_2|}(1+\kappa(P_{b|h}))\varepsilon_A < 1$. □

Lemma B.4. *For any $t > 0$, if P_{bc} is invertible and $\frac{1+\sqrt{t}}{\sqrt{N}} < \frac{1}{2\|P_{bc}^{-1}\|}$, we have $Pr(\|\hat{P}_{bc}^{a=a_1}\hat{P}_{bc}^{-1} - P_{bc}^{a=a_1}P_{bc}^{-1}\| \leq 3\|P_{bc}^{-1}\|^2\frac{1+\sqrt{t}}{\sqrt{N}}) > 1 - 2e^{-t}$.*

Proof. When $\|P_{bc}^{-1}\|\|\hat{P}_{bc} - P_{bc}\| \leq \frac{1}{2}$, $\hat{P}_{bc}$ is invertible and $\|\hat{P}_{bc}^{-1} - P_{bc}^{-1}\| \leq \frac{\|P_{bc}^{-1}\|^2\|\hat{P}_{bc}-P_{bc}\|}{1-\|P_{bc}^{-1}\|\|\hat{P}_{bc}-P_{bc}\|} \leq 2\|P_{bc}^{-1}\|^2\|\hat{P}_{bc} - P_{bc}\|$. Next, we consider the difference between $\hat{P}_{bc}^{a=a_1}\hat{P}_{bc}^{-1}$ and $P_{bc}^{a=a_1}P_{bc}^{-1}$.

$$\begin{aligned} &\|\hat{P}_{bc}^{a=a_1}\hat{P}_{bc}^{-1} - P_{bc}^{a=a_1}P_{bc}^{-1}\| \\ &\quad\leq \|\hat{P}_{bc}^{a=a_1} - P_{bc}^{a=a_1}\|\|P_{bc}^{-1}\| + \|\hat{P}_{bc}^{-1} - P_{bc}^{-1}\|\|\hat{P}_{bc}^{a=a_1}\| \\ &\quad\leq \|P_{bc}^{-1}\|\|\hat{P}_{bc}^{a=a_1} - P_{bc}^{a=a_1}\| + \|\hat{P}_{bc}^{-1} - P_{bc}^{-1}\| \\ &\quad\leq \|P_{bc}^{-1}\|\|\hat{P}_{bc}^{a=a_1} - P_{bc}^{a=a_1}\| + 2\|P_{bc}^{-1}\|^2\|\hat{P}_{bc} - P_{bc}\| \\ &\quad\leq \|P_{bc}^{-1}\|^2(\|\hat{P}_{bc}^{a=a_1} - P_{bc}^{a=a_1}\| + 2\|\hat{P}_{bc} - P_{bc}\|). \end{aligned}$$

The second inequality holds since $\|\hat{P}_{bc}^{a=a_1}\| \leq 1$ and the final inequality holds since $\|P_{bc}^{-1}\| \geq 1$. Furthermore, since $Pr(\|\hat{P}_{bc} - P_{bc}\| \leq \frac{1+\sqrt{t}}{\sqrt{N}}, \|\hat{P}_{bc}^{a=a_1} - P_{bc}^{a=a_1}\| \leq \frac{1+\sqrt{t}}{\sqrt{N}}) > 1 - 2e^{-t}$ for any $t > 0$ from Proposition A.1, we have $Pr(\|\hat{P}_{bc}^{a=a_1}\hat{P}_{bc}^{-1} - P_{bc}^{a=a_1}P_{bc}^{-1}\| \leq 3\|P_{bc}^{-1}\|^2\frac{1+\sqrt{t}}{\sqrt{N}}) > 1 - 2e^{-t}$ when the sample size is sufficiently large such that $\frac{1+\sqrt{t}}{\sqrt{N}} \leq \frac{1}{2\|P_{bc}^{-1}\|}$. □

Next, we give the proof of Theorem 4.1. In the case where all the variables are two-state, P_{bc} is invertible and $\lambda_1 \neq \lambda_2$ from assumption (A4). Define the event $E = \{|\hat{\lambda}_i - \lambda_i| \leq 3\kappa(P_{b|h})\|P_{bc}^{-1}\|^2 \frac{1+\sqrt{t}}{\sqrt{N}}, \|\hat{x}_i - x_i\| \leq \frac{18}{|\lambda_1-\lambda_2|}(1 + \kappa(P_{b|h}))\|P_{bc}^{-1}\|^2 \frac{1+\sqrt{t}}{\sqrt{N}} : i = 1, 2\}$. Since $|\lambda_1 - \lambda_2| \leq 1$, $\kappa(P_{b|h}) \geq 1$ and $\|P_{bc}^{-1}\| \geq 1$, we have $\frac{1+\sqrt{t}}{\sqrt{N}} < \frac{1}{2\|P_{bc}^{-1}\|}$ when $\frac{18}{|\lambda_1-\lambda_2|}(1 + \kappa(P_{b|h}))\|P_{bc}^{-1}\|^2 \frac{1+\sqrt{t}}{\sqrt{N}} < 1$. Thus, $Pr(\|\hat{P}_{bc}^{a=a_1} \hat{P}_{bc}^{-1} - P_{bc}^{a=a_1} P_{bc}^{-1}\| \leq 3\|P_{bc}^{-1}\|^2 \frac{1+\sqrt{t}}{\sqrt{N}}) > 1 - 2e^{-t}$ from Lemma B.4. From the second conclusion of Corollary B.1, the event $\{\|\hat{P}_{bc}^{a=a_1} \hat{P}_{bc}^{-1} - P_{bc}^{a=a_1} P_{bc}^{-1}\| \leq 3\|P_{bc}^{-1}\|^2 \frac{1+\sqrt{t}}{\sqrt{N}}\}$ is contained in the event E when $\frac{18}{|\lambda_1-\lambda_2|}(1+ \kappa(P_{b|h}))\|P_{bc}^{-1}\|^2 \frac{1+\sqrt{t}}{\sqrt{N}} < 1$. So $P(E) > 1 - 2e^{-t}$. For any $\eta \in (0, 1)$, select $t_0 = -\log \frac{1}{2}\eta$, and thus we obtain the proof of Theorem 4.1.

B.2 Consistency of the PELT algorithm

In this section, we prove the consistency of our PELT algorithm in the case where the latent variables are two-state and the state number of the observed variables may be greater than two. We introduce two lower bounds as the intrinsic parameters: $\phi_* = \min\{\det P_{hh} : h \in H\}$ and $\theta_* = \min\{|Pr(a = a_i|h = h_1) - Pr(a = a_i|h = h_2)| : Pr(a = a_i|h = h_1) \neq Pr(a = a_i|h = h_2), 1 \leq i \leq d_a, a \in V, h \in H\}$.

For any variable $a \in V$ and any state $1 \leq i \leq d_a$, we can view variable a as a two-state variable $a^{(i)}$. The first state $a_1^{(i)}$ of $a^{(i)}$ is the ith state a_i of a, and the second state $a_2^{(i)}$ represents the other states $\{a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_{d_a}\}$ of a. Similarly, for any $b, c \in V$ and any $1 \leq j \leq d_b, 1 \leq k \leq d_c$, b and c can be viewed as the two-state variables $b^{(j)}$ and $c^{(k)}$.

The following Lemma B.5 gives the bound of singular values, the condition numbers, and the norm of matrix inversion using the intrinsic parameters ϕ_* and θ_*.

Lemma B.5. *For the observed variables $a, b \in V$, the states $1 \leq i \leq d_a, 1 \leq j \leq d_b$, and the latent variables $q, h \in H$, if $P_{a^{(i)}|h}$ and $P_{a^{(i)}b^{(j)}}$ is invertible, we have the following inequalities:*
1. $\frac{\theta_*}{\sqrt{2}} \leq \sigma_2(P_{a^{(i)}|h}) \leq \sigma_1(P_{a^{(i)}|h}) \leq \sqrt{2}$,
2. $\kappa(P_{a^{(i)}|h}) \leq \frac{2}{\theta_*}$,
3. $\frac{\theta_*}{\sqrt{2}} \leq \sigma_2(P_{q|h}) \leq \sigma_1(P_{q|h}) \leq \sqrt{2}$,
4. $\theta_*^2\phi_* \leq \sigma_2(P_{a^{(i)}b^{(j)}}) \leq \sigma_1(P_{a^{(i)}b^{(j)}}) \leq 1$,
5. $\|P_{a^{(i)}b^{(j)}}^{-1}\| \leq \frac{1}{\theta_*^2\phi_*}$,
6. $\frac{\theta_*}{\sqrt{2d_{max}}} \leq \sigma_2(P_{a|h}) \leq \sigma_1(P_{a|h}) \leq \sqrt{2}$.

Proof. The sum of each column of $P_{a^{(i)}|h}$ is one and every element in $P_{a^{(i)}|h}$ is nonnegative. Thus, $\|P_{a^{(i)}|h}\|_F \leq \sqrt{2}$. It follows that $\sigma_1(P_{a^{(i)}|h}) = \|P_{a^{(i)}|h}\| \leq \|P_{a^{(i)}|h}\|_F \leq \sqrt{2}$. Since $\sigma_1(P_{a^{(i)}|h})\sigma_2(P_{a^{(i)}|h}) = |\det(P_{a^{(i)}|h})| = |Pr(a = a_i|h = h_1) - Pr(a = a_i|h = h_2)| \geq \theta_*$, we have $\frac{\theta_*}{\sqrt{2}} \leq \sigma_2(P_{a^{(i)}|h}) \leq \sigma_1(P_{a^{(i)}|h}) \leq \sqrt{2}$. Furthermore, $\kappa(P_{a^{(i)}|h}) = \frac{\sigma_1(P_{a^{(i)}|h})}{\sigma_2(P_{a^{(i)}|h})} \leq \frac{2}{\theta_*}$.

An observed variable c exists in the latent tree such that c and h are conditionally independent given q. Some state $1 \leq k \leq d_c$ exists such that $P_{c^{(k)}|q}$ is not singular. Furthermore, since $P_{c^{(k)}|q}P_{q|h} = P_{c^{(k)}|h}$, it follows that $\sigma_1(P_{q|h})\sigma_2(P_{q|h})|\det(P_{c^{(k)}|q})| \geq \theta_*$. From $|\det(P_{c^{(k)}|q})| \leq 1$, we have $\sigma_1(P_{q|h})\sigma_2(P_{q|h}) \geq \theta_*$. Since $\sigma_1(P_{q|h}) = \|P_{q|h}\| \leq \|P_{q|h}\|_F \leq \sqrt{2}$, so $\frac{\theta_*}{\sqrt{2}} \leq \sigma_2(P_{q|h}) \leq \sigma_1(P_{q|h}) \leq \sqrt{2}$.

A latent variable s exists such that the observed variables a and b are conditionally independent given s, so we have $|\det(P_{a^{(i)}b^{(j)}})| = |\det(P_{a^{(i)}|s})|$ $|\det(P_{ss})||\det(P_{b^{(j)}|s})| \geq \theta_*^2\phi_*$. Since $\sigma_1(P_{a^{(i)}b^{(j)}}) = \|P_{a^{(i)}b^{(j)}}\| \leq \|P_{a^{(i)}b^{(j)}}\|_F \leq 1$, it follows that $\sigma_2(P_{a^{(i)}b^{(j)}}) \geq \theta_*^2\phi_*$. Thus, $\|P_{a^{(i)}b^{(j)}}^{-1}\| = \frac{1}{\sigma_2(P_{a^{(i)}b^{(j)}})} \leq \frac{1}{\theta_*^2\phi_*}$.

From the relation between $a^{(i)}$ and a, it follows that $P_{a^{(i)}|h} = BP_{a|h}$, where every element of the first row of matrix B is zero, except the ith position is one, and every element of the second row of B is one, except the ith position is zero. It follows that $\sigma_1(P_{a^{(i)}|h})\sigma_2(P_{a^{(i)}|h}) \leq \sqrt{d_{max}}\sigma_1(P_{a|h})\sigma_2(P_{a|h})$ from (III.19) in [4]. Since $\sigma_1(P_{a^{(i)}|h})\sigma_2(P_{a^{(i)}|h}) = |\det(P_{a^{(i)}|h})| \geq \theta_*$, we have $\frac{\theta_*}{\sqrt{2d_{max}}} \leq \sigma_2(P_{a|h}) \leq \sigma_1(P_{a|h}) \leq \sqrt{2}$ from $\sigma_1(P_{a|h}) \leq \|P_{a|h}\|_F \leq \sqrt{2}$. □

Based on the discussion in Section 4.2, our method for estimating parameters can be performed correctly if we can correctly judge whether $P_{b^{(j)}c^{(k)}}$ is singular or not and whether the two eigenvalues are equal or not. Furthermore, we must correctly match the latent variable states with the recorded label states. All of these requirements are considered in the following.

We introduce a threshold ϵ_1 to judge whether $P_{b^{(j)}c^{(k)}}$ is singular or not based on the empirical estimation $\hat{P}_{b^{(j)}c^{(k)}}$. If $P_{b^{(j)}c^{(k)}}$ is invertible, we have $|\det(P_{b^{(j)}c^{(k)}})| = |\det(P_{b^{(j)}|h})||\det P_{hh}||\det(P_{c^{(k)}|h})| \geq \theta_*^2\phi_*$. If $|\det(\hat{P}_{b^{(j)}c^{(k)}}) - \det(P_{b^{(j)}c^{(k)}})| \leq \epsilon_1 < \frac{1}{2}\theta_*^2\phi_*$ when the sample size is sufficiently large, then $|\det(\hat{P}_{b^{(j)}c^{(k)}})| \leq \epsilon_1$ if and only if $\det(P_{b^{(j)}c^{(k)}}) = 0$. Thus, if the event $\bigcap_{b,c,j,k}\{|\det(\hat{P}_{b^{(j)}c^{(k)}}) - \det(P_{b^{(j)}c^{(k)}})| \leq \epsilon_1\}$, where $b, c \in V, b \neq c, 1 \leq j \leq d_b$ and $1 \leq k \leq d_c$, can occur with a high probability when the sample size is sufficiently large, we can correctly judge whether all of $\det(P_{b^{(j)}c^{(k)}})$ are zero or not with a high probability. Indeed, the following lemma shows the lower bound of the probability that we can correctly judge whether all of $\det(P_{b^{(j)}c^{(k)}})$ are zero or not.

Lemma B.6. *For any $t > 0$, $Pr(\bigcap_{b,c\in V,b\neq c,1\leq j\leq d_b,1\leq k\leq d_c} \{|\det(\hat{P}_{b^{(j)}c^{(k)}}) - \det(P_{b^{(j)}c^{(k)}})| \leq \frac{2(1+\sqrt{t})}{\sqrt{N}}\}) > 1 - n^2d_{max}^2e^{-t}$, where n is the number of observed variables.*

Proof. It can be verified that $|\det(\hat{P}_{b^{(j)}c^{(k)}}) - \det(P_{b^{(j)}c^{(k)}})| \leq 2\|\hat{P}_{b^{(j)}c^{(k)}} - P_{b^{(j)}c^{(k)}}\|_F$. Thus, from Proposition A.1, it follows that $Pr(|\det(\hat{P}_{b^{(j)}c^{(k)}}) - \det(P_{b^{(j)}c^{(k)}})| \leq \frac{2(1+\sqrt{t})}{\sqrt{N}}) > 1 - e^{-t}$ for any two distinct observable variables $b, c \in V$ and any $1 \leq j \leq d_b, 1 \leq k \leq d_c$. Therefore, for any $t > 0$, we have

$$Pr\left(\bigcap_{b,c,j,k}\{|\det(\hat{P}_{b^{(j)}c^{(k)}})-\det(P_{b^{(j)}c^{(k)}})|\leq\frac{2(1+\sqrt{t})}{\sqrt{N}}\}\right)$$
$$>1-\sum_{b,c,j,k}Pr\Bigg(|\det(\hat{P}_{b^{(j)}c^{(k)}})-\det(P_{b^{(j)}c^{(k)}})|$$
$$>\frac{2(1+\sqrt{t})}{\sqrt{N}}\Bigg)>1-n^2d_{max}^2e^{-t},$$

where n is the number of observed variables and $b,c\in V, b\neq c, 1\leq j\leq d_b, 1\leq k\leq d_c$. □

As discussed above, if we set the threshold $\epsilon_1<\frac{1}{2}\theta_*^2\phi_*$, when the sample size N is sufficiently large such that $\frac{2(1+\sqrt{t})}{\sqrt{N}}<\epsilon_1$, we can correctly judge whether all of $\det(P_{b^{(j)}c^{(k)}})$ are zero or not with a probability of at least $1-n^2d_{max}^2e^{-t}$ for any $t>0$.

Now, we assume that we can correctly judge whether all of $\det(P_{b^{(j)}c^{(k)}})$ are zero or not. If the observed variables a,b,c are conditionally independent given a latent variable h, some $1\leq j\leq d_b$ and $1\leq k\leq d_c$ exist such that $P_{b^{(j)}c^{(k)}}$ is not singular. Thus, we can decompose the matrix $P_{b^{(j)}c^{(k)}}^{a^{(i)}=a_1^{(i)}}P_{b^{(j)}c^{(k)}}^{-1}$ to obtain the two eigenvalues, $\lambda_1=Pr(a=a_i|h=h_1)$ and $\lambda_2=Pr(a=a_i|h=h_2)$. We also view the two eigenvalues as two functions, $\lambda_1(a^{(i)},b^{(j)},c^{(k)},h)$ and $\lambda_2(a^{(i)},b^{(j)},c^{(k)},h)$. Since $P_{a|h}$ has full column rank, some $1\leq i\leq d_a$ exists such that $\lambda_1(a^{(i)},b^{(j)},c^{(k)},h)\neq\lambda_2(a^{(i)},b^{(j)},c^{(k)},h)$. Furthermore, we can obtain the two eigenvectors $x_1=(Pr(b^{(j)}=b_1^{(j)}|h=h_1),Pr(b^{(j)}=b_2^{(j)}|h=h_1))^T$ and $x_2=(Pr(b^{(j)}=b_1^{(j)}|h=h_2),Pr(b^{(j)}=b_2^{(j)}|h=h_2))^T$ that correspond to λ_1 and λ_2, respectively, from the restriction $\mathbb{1}^Tx_1=\mathbb{1}^Tx_2=1$. Thus, the first element x_{11} of x_1 is $Pr(b^{(j)}=b_1^{(j)}|h=h_1)$, which is simply $Pr(b=b_j|h=h_1)$, and the first element x_{21} of x_2 is $Pr(b^{(j)}=b_1^{(j)}|h=h_2)$, which is simply $Pr(b=b_j|h=h_2)$. Similarly, we view these two eigenvectors as two vector functions, $x_1(a^{(i)},b^{(j)},c^{(k)},h)$ and $x_2(a^{(i)},b^{(j)},c^{(k)},h)$.

We define two sets $F_1=\{\omega=(a^{(i)},b^{(j)},c^{(k)},h)$: Three ordered distinct variables $a,b,c\in V$ are conditionally independent given $h\in H$. $1\leq i\leq d_a$, $1\leq j\leq d_b$ and $1\leq k\leq d_c$. $P_{b^{(j)}c^{(k)}}$ is invertible and $Pr(a=a_i|h=h_1)=Pr(a=a_i|h=h_2)\}$ and $F_2=\{\omega=(a^{(i)},b^{(j)},c^{(k)},h)$: Three ordered distinct variables $a,b,c\in V$ are conditionally independent given $h\in H$. $1\leq i\leq d_a$, $1\leq j\leq d_b$ and $1\leq k\leq d_c$. $P_{b^{(j)}c^{(k)}}$ is invertible and $Pr(a=a_i|h=h_1)\neq Pr(a=a_i|h=h_2)\}$. For any $\omega=(a^{(i)},b^{(j)},c^{(k)},h)\in F_1\bigcup F_2$, we denote $P_{b^{(j)}c^{(k)}}^{a^{(i)}=a_1^{(i)}}P_{b^{(j)}c^{(k)}}^{-1}$ as $A(\omega)$. Furthermore, if $\hat{P}_{b^{(j)}c^{(k)}}$ is invertible, we denote $\hat{P}_{b^{(j)}c^{(k)}}^{a^{(i)}=a_1^{(i)}}\hat{P}_{b^{(j)}c^{(k)}}^{-1}$ as $\hat{A}(\omega)$. Let $\lambda_1(\omega)$ and $\lambda_2(\omega)$ be the two eigenvalues of $A(\omega)$, and let $\hat{\lambda}_1(\omega)$ and $\hat{\lambda}_2(\omega)$ be the two eigenvalues of $\hat{A}(\omega)$. For any $\omega\in F_2$, denote the two eigenvectors of $A(\omega)$ that correspond to $\lambda_1(\omega)$ and $\lambda_2(\omega)$ as $x_1(\omega)$ and $x_2(\omega)$ with $\mathbb{1}^Tx_1(\omega)=\mathbb{1}^Tx_2(\omega)=1$. If $\hat{\lambda}_1(\omega)\neq\hat{\lambda}_2(\omega)$, denote the two eigenvectors of $\hat{A}(\omega)$ that correspond to $\hat{\lambda}_1(\omega)$ and $\hat{\lambda}_2(\omega)$ as $\hat{x}_1(\omega)$ and $\hat{x}_2(\omega)$ with $\mathbb{1}^T\hat{x}_1(\omega)=\mathbb{1}^T\hat{x}_2(\omega)=1$.

For any $\omega_1\in F_1$, we define the event $E_1(\omega_1)=\bigcap_{i=1,2}\{|\hat{\lambda}_i(\omega_1)-\lambda_i(\omega_1)|\leq\frac{6(1+\sqrt{t})}{\theta_*^5\phi_*^2\sqrt{N}}\}$. For any $\omega_2\in F_2$, we define the event $E_2(\omega_2)=\bigcap_{i=1,2}\{|\hat{\lambda}_i(\omega_2)-\lambda_i(\omega_2)|\leq\frac{6(1+\sqrt{t})}{\theta_*^5\phi_*^2\sqrt{N}},|\hat{\lambda}_1(\omega_2)-\hat{\lambda}_2(\omega_2)|>\frac{1}{2}\theta_*,|\hat{x}_{i1}(\omega_2)-x_{i1}(\omega_2)|\leq\frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}}\}$. Furthermore, The following lemma presents the lower bound of the probability of the event $\bigcap_{\omega_1\in F_1,\omega_2\in F_2}(E_1(\omega_1)\bigcup E_2(\omega_2))$:

Lemma B.7. *For any $t>0$, if $\frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}}<1$, we have*

$$Pr\left(\bigcap_{\omega_1\in F_1,\omega_2\in F_2}(E_1(\omega_1)\bigcup E_2(\omega_2))\right)\geq1-2n^3md_{max}^3e^{-t},$$

where n is the number of observed variables and m is the number of latent variables.

Proof. From the definition of θ_* and ϕ_*, it follows that $0<\theta_*\leq1$ and $0<\phi_*\leq1$. If the sample size N is sufficiently large such that $\frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}}<1$, for any $(a^{(i)},b^{(j)},c^{(k)},h)\in F_1\bigcup F_2$, it holds that $\frac{1+\sqrt{t}}{\sqrt{N}}<\frac{1}{2\|P_{b^{(j)}c^{(k)}}^{-1}\|}$ from Lemma B.5. Furthermore, $Pr(\|\hat{P}_{b^{(j)}c^{(k)}}^{a^{(i)}=a_1^{(i)}}\hat{P}_{b^{(j)}c^{(k)}}^{-1}-P_{b^{(j)}c^{(k)}}^{a^{(i)}=a_1^{(i)}}P_{b^{(j)}c^{(k)}}^{-1}\|\leq3\|P_{b^{(j)}c^{(k)}}^{-1}\|^2\frac{1+\sqrt{t}}{\sqrt{N}})>1-2e^{-t}$ from Lemma B.4. Moreover, for any $\omega=(a^{(i)},b^{(j)},c^{(k)},h)\in F_2$, we have $\frac{18}{|\lambda_1(\omega)-\lambda_2(\omega)|}(1+\kappa(P_{b^{(j)}|h}))\|P_{b^{(j)}c^{(k)}}^{-1}\|^2\frac{1+\sqrt{t}}{\sqrt{N}}<1$ from Lemma B.5.

As discussed above, when the sample size N is sufficiently large such that $\frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}}<1$, we have $Pr(E_1(\omega_1)\bigcup E_2(\omega_2))>1-2e^{-t}$ for any $\omega_1\in F_1$ and any $\omega_2\in F_2$ from Corollary B.1 and Lemma B.4. Hence, we have:

$$Pr\left(\bigcap_{\omega_1\in F_1,\omega_2\in F_2}(E_1(\omega_1)\bigcup E_2(\omega_2))\right)$$
$$\geq1-\sum_{\omega_1\in F_1,\omega_2\in F_2}Pr\left((E_1(\omega_1)\bigcup E_2(\omega_2))^c\right)$$
$$\geq1-2e^{-t}|F_1\bigcup F_2|$$
$$\geq1-2n^3md_{max}^3e^{-t},$$

where n is the number of observed variables and m is the number of latent variables. □

If $|\hat{\lambda}_1(\omega)-\hat{\lambda}_2(\omega)-\lambda_1(\omega)+\lambda_2(\omega)|<\epsilon_2<\frac{1}{2}\theta_*$, then $|\hat{\lambda}_1(\omega)-\hat{\lambda}_2(\omega)|<\epsilon_2$ if and only if $\lambda_1(\omega)=\lambda_2(\omega)$. If the event $\{|\hat{\lambda}_1(\omega)-\hat{\lambda}_2(\omega)-\lambda_1(\omega)+\lambda_2(\omega)|<\epsilon_2$ for any $\omega\in F_1\bigcup F_2\}$ occurs with a high probability when the sample size is sufficiently large, we can correctly judge whether all the $\lambda_1(\omega) - \lambda_2(\omega)$ for $\omega \in F_1 \bigcup F_2$ are zero or not with a high probability.

As discussed above, if we can correctly judge whether all the $\det(P_{b^{(j)}c^{(k)}})$ are zero or not, by setting the threshold $\epsilon_2 < \frac{1}{2}\theta_*$, when the sample size is sufficiently large such that $\frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}} < \epsilon_2$ (which implies that $\frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}} < 1$ and $|\hat{\lambda}_1(\omega) - \hat{\lambda}_2(\omega) - \lambda_1(\omega) + \lambda_2(\omega)| \le \frac{12(1+\sqrt{t})}{\theta_*^5\phi_*^2\sqrt{N}} < \epsilon_2$ for $\omega \in F_1 \bigcup F_2$), the event that we can correctly judge whether all the $\lambda_1(\omega) - \lambda_2(\omega)$ for $\omega \in F_1 \bigcup F_2$ are zero or not occurs with a probability of at least $1 - 2n^3 m d_{max}^3 e^{-t}$. At the same time, we can guarantee that for any $\omega \in F_2$, $|\hat{\lambda}_1(\omega) - \hat{\lambda}_2(\omega)| > \frac{1}{2}\theta_*$ and $|\hat{\lambda}_i(\omega) - \lambda_i(\omega)| \le \frac{1}{4}\theta_*$ for $i = 1, 2$, thus we can correctly match the states of the latent variables using the label states. Furthermore, if $Pr(b = b_j | h = h_1) = Pr(b = b_j | h = h_2)$, we only decompose $P^{b^{(j)}=b_1^{(j)}}_{a^{(i')}c^{(k)}} P^{-1}_{a^{(i')}c^{(k)}}$ once to obtain the two equal eigenvalues, as discussed in Section 4.2. This means that for any $(a^{(i)}, b^{(j)}, c^{(k)}, h) \in F_1 \bigcup F_2$, the decomposition of $P^{b^{(j)}=b_1^{(j)}}_{a^{(i)}c^{(k)}} P^{-1}_{a^{(i)}c^{(k)}}$ will be performed once at most in the PELT algorithm, and any case on $(a^{(i)}, b^{(j)}, c^{(k)}, h)$ used in the PELT algorithm is contained in $F_1 \bigcup F_2$.

Denote the event that our PELT algorithm works as G_1 and the event $\bigcap_{s,h,i,j}\{|\hat{Pr}(s = s_i | h = h_j) - Pr(s = s_i | h = h_j)| \le \frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}}\}$ as G_2, where the estimation $\hat{Pr}(s = s_i | h = h_j)$ is obtained from the PELT algorithm and $s \in V, h \in H, 1 \le i \le d_a, j = 1, 2$. Thus we have the following lemma:

Lemma B.8. *For any $t > 0$, if we set the thresholds $\epsilon_1 < \frac{1}{2}\theta_*^2\phi_*$ and $\epsilon_2 < \frac{1}{2}\theta_*$, when $\frac{2(1+\sqrt{t})}{\sqrt{N}} < \epsilon_1$ and $\frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}} < \epsilon_2$, then $Pr(G_1 \bigcap G_2) \ge 1 - 3n^3 m d_{max}^3 e^{-t}$, where n is the number of observed variables and m is the number of latent variables.*

Proof. From Lemmas B.6 and B.7, we have equalities that $Pr(G_1 \bigcap G_2) \ge (1 - 2n^3 m d_{max}^3 e^{-t})(1 - n^2 d_{max}^2 e^{-t}) \ge 1 - 3n^3 m d_{max}^3 e^{-t}$, where n is the number of observed variables and m is the number of latent variables. □

From Lemma B.8, for any $s \in V$ and $h \in H$, we have $\|\hat{P}_{s|h} - P_{s|h}\| \le \sqrt{2d_{max}}\frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}}$ and $\hat{P}_{s|h}$ also has full column rank. Since $P_{s|q}P_{q|h} = P_{s|h}$ and $\hat{P}_{s|q}\hat{P}_{q|h} = \hat{P}_{s|h}$ in the PELT algorithm, then from Theorem 5.1 in [24], we have

$$\begin{aligned}
\|\hat{P}_{q|h} - P_{q|h}\| &\le \frac{\|P^+_{s|q}\|(\sqrt{2}\|\hat{P}_{s|q} - P_{s|q}\| + \|\hat{P}_{s|h} - P_{s|h}\|)}{1 - \|P^+_{s|q}\|\|\hat{P}_{s|q} - P_{s|q}\|} \\
&\le \frac{\sqrt{2d_{max}}(\sqrt{2}\|\hat{P}_{s|q} - P_{s|q}\| + \|\hat{P}_{s|h} - P_{s|h}\|)}{\theta_* - \sqrt{2d_{max}}\|\hat{P}_{s|q} - P_{s|q}\|} \\
&\le \frac{2\sqrt{2d_{max}}}{\theta_*}\frac{\frac{5}{2}\sqrt{2d_{max}}54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}} \\
&\le \frac{540d_{max}(1+\sqrt{t})}{\theta_*^7\phi_*^2\sqrt{N}},
\end{aligned}$$

where the second inequality holds from Lemma B.5 when $\sqrt{2d_{max}}\|\hat{P}_{s|q} - P_{s|q}\| \le \frac{1}{2}\theta_*$. Since $\frac{108d_{max}(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}} < \epsilon_2 < \frac{1}{2}\theta_*$ implies that $\frac{54(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}} < \epsilon_2$ and $\sqrt{2d_{max}}\|\hat{P}_{s|q} - P_{s|q}\| \le \frac{108d_{max}(1+\sqrt{t})}{\theta_*^6\phi_*^2\sqrt{N}} < \frac{1}{2}\theta_*$, we obtain the following theorem which shows that our PELT algorithm can obtain consistent estimates of all the conditional probability matrices.

Theorem B.1. *For any $\eta \in (0, 1)$, if we set the thresholds $\epsilon_1 < \frac{1}{2}\theta_*^2\phi_*$ and $\epsilon_2 < \frac{1}{2}\theta_*$, when the sample size N is sufficiently large such that*

$$\text{(B.1)} \qquad \frac{2(1+\sqrt{t_0})}{\sqrt{N}} < \epsilon_1, \ \frac{108d_{max}(1+\sqrt{t_0})}{\theta_*^6\phi_*^2\sqrt{N}} < \epsilon_2,$$

where $t_0 = -\log\frac{\eta}{3n^3 m d_{max}^3}$, then we have $Pr(\bigcap_{h \in H, v \in ch(h)}\{\|\hat{P}_{v|h} - P_{v|h}\| \le \frac{540d_{max}(1+\sqrt{t_0})}{\theta_*^7\phi_*^2\sqrt{N}}\}) \ge 1 - \eta$, where n is the number of observed variables and m is the number of latent variables.

Received 22 April 2016

REFERENCES

[1] Allman, E. S., Matias, C. and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* **37** 3099–3132. MR2549554

[2] Anandkumar, A., Hsu, D. and Kakade, S. M. (2012). A method of moments for mixture models and hidden Markov models. *Twenty-Fifth Annual Conference on Learning Theory*.

[3] Bartholomew, D. J., Knott, M. and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd ed., Wiley. MR2849614

[4] Bhatia, R. (1997). *Matrix Analysis*, Springer-Verlag. MR1477662

[5] Chang, J. T. (1996). Full reconstruction of Markov models on evolutionary tree: Identifiability and consistency. *Mathematical Biosciences* **137** 51–73. MR1410044

[6] Chen, T., Zhang, N. L., Liu, T. F., Poon, K. M. and Wang, Y. (2011). Model-based multidimensional clustering of categorical data. *Artificial Intelligence* **176** 2246–2269. MR2896724

[7] Choi, M. J., Tan, V. Y. F., Anandkumar, A. and Willsky, A. S. (2011) Learning latent tree graphical models. *Journal of Machine Learning Research* **12** 1771–1812. MR2813153

[8] Daskalakis, C., Mossel, E. and Roch, S. (2006). Optimal phylogenetic reconstruction. *In STOC'06: Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing* 159–168. MR2277141

[9] Dickson, K. I., Kelley, C. T., Ipsen, I. C. F. and Kevrekidis, I. G. (2007). Condition estimates for pseudo-arclength continuation. *SIAM J. Numer. Anal.* **45** 678–694.MR2285854

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

[11] Erdös, P. L., Székely, L. A., Steel, M. A. and Warnow, T. J. (1999). A few logs suffice to build (almost) all trees: Part ii. *Theoretical Computer Science* **221** 153–184.

[12] Guan, G. Y., Guo, J. H. and WANG, H. S. (2014). Varying naïve bayes models with applications to classification of chinese text documents. *Journal of Business & Economic Statistics* **32** 445–56. MR3238597

[13] Hsu, D., Kakade, S. M. and Zhang, T. (2012). A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences* **78** 1460–1480. MR2926144

[14] Jiang, T., Kearney, P. E. and Li, M. (2001). A polynomial-time approximation scheme for inferring evolutionary trees from quartet topologies and its application. *SIAM J. Comput.* **30** 194–261. MR1856563

[15] Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proceedings of the National Academy of Science* **91** 1455–1459.

[16] Liu, T. F., Zhang, N. L., Chen, P. X., Liu, A. H., Poon, L. K. M. and Wang, Y. (2015). Greedy learning of latent tree models for multidimensional clustering. *Machine Learning* **98** 301–330. MR3296685

[17] Mossel, E. and Roch, S. (2006) Learning singular phylogenies and hidden Markov models. *The Annals of Applied Probability* **16** 583–614. MR2244426

[18] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*, The MIT Press.

[19] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Network and Plausible Inference*, Morgan Kaufmann, 1988. MR0965765

[20] Robinson, D. F. and Foulds, L. R. (1981). Comparision of phylogenetic trees. *Mathematical Biosciences* **53** 131–147. MR0613619

[21] Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4** 406–425.

[22] Song, L., Parikh, A. P. and Xing, E. P. (2011). Kernel embeddings of latent tree graphical models. *NIPS*.

[23] Wang, Y., Zhang, N. L. and Chen, T. (2008). Latent tree models and approximate inference in Bayesian networks. *Journal of Artificial Intelligence Research* **32** 879–900. MR2487524

[24] Wedin, P. (1973). Perturbation theory for pseudo-inverses. *BIT* **13** 217–232. MR0336982

[25] N. L. Zhang (2004). Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research* **5** 697–723. MR2247997

[26] N. L. Zhang and T. Kočka (2004). Efficient learning of hierarchical latent class models. *In ICTAI*.

Xiaofei Wang
Key Laboratory for Applied Statistics of MOE
School of Mathematics and Statistics
Northeast Normal University
Changchun 130024, Jilin Province
China
E-mail address: wangxf341@nenu.edu.cn

Jianhua Guo
Key Laboratory for Applied Statistics of MOE
School of Mathematics and Statistics
Northeast Normal University
Changchun 130024, Jilin Province
China
E-mail address: jhguo@nenu.edu.cn

Lizhu Hao
Key Laboratory for Applied Statistics of MOE
School of Mathematics and Statistics
Northeast Normal University
Changchun 130024, Jilin Province
China
E-mail address: haolizhu986@nenu.edu.cn

Nevin L. Zhang
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Clear Water Bay Road, Kowloon
Hong Kong
E-mail address: lzhang@cse.ust.hk