Journal of Machine Learning Research 21 (2020) 1-35 Submitted 4/20; Revised 9/20; Published 10/20

Learning Mixed Latent Tree Models

Can Zhou ZHOUC574@NENU.EDU.CN
Xiaofei Wang[*] WANGXF341@NENU.EDU.CN
Jianhua Guo[*] JHGUO@NENU.EDU.CN
KLAS and School of Mathematics and Statistics
Northeast Normal University
Changchun, China.

Editor: Zhihua Zhang

Abstract

Latent structural learning has attracted more attention in recent years. But most related works only focuses on pure continuous or pure discrete data. In this paper, we consider mixed latent tree models for mixed data mining. We address the latent structural learning and parameter estimation for those mixed models. For structural learning, we propose a consistent bottom-up algorithm, and give a finite sample bound guarantee for the exact structural recovery. For parameter estimation, we suggest a moment estimator by exploiting matrix decomposition, and prove asymptotic normality of the estimator. Experiments on the simulated and real data support that our method is valid for mining the hierarchical structure and latent information.

Keywords: Information distance, Latent variables, Mixed latent tree, Parameter estimation, Structural learning

1. Introduction

Latent variable models are important tools for probabilistic modeling, and have been widely applied to various domains, such as speech analysis and bioinformatics. As one of classical latent variable models, the latent class model can effectively deal with clustering analysis problems (Goodman, 1974; Lazarsfeld and Henry, 1968). The latent class model has a simple assumption that all observed variables are conditionally independent given the latent class variable. But this assumption can't catch the potential complex mechanism behind observed variables. For further extension, Zhang (2004) first investigated the latent tree model, in which all leaf nodes are observed variables and all internal nodes are latent variables. The latent tree model addresses the local dependence problem with a principled manner. It can capture a hierarchical generating mechanism for observed variables, and provide more illuminating explanations on the variable groups compared to the latent class model.

The learning and application of latent tree models developed a lot over the past decades. The original work (Zhang, 2004) proposed a scoring-based algorithm for structural learning. Some further works (Chen et al., 2012; Liu et al., 2015) extended the scoring-based learning algorithm, and suggested the multidimensional clustering for the multiple partitions of data. Besides the multidimensional clustering, the model also has applications in computer vision,

[*] Corresponding authors

probabilistic inference, and hierarchical topic detection. Wang and Li (2013) proposed a more flexible and effective human pose estimation in computer vision based on latent tree models. This method can combine two parts which are unlimited to the physical connections in the human body. Furthermore, it effectively exploits the interactions between combined parts and single parts. Wang et al. (2008) applied the latent tree model to the approximate inference of the Bayesian network. They learned a latent tree model from data which is sampled from a Bayesian network, and made inference with the latent tree model instead of the original Bayesian network. It achieved good approximation accuracy at low online computational cost. Chen et al. (2016) used the latent tree models to handle hierarchical topic detection. Bottom-level variables are observed binary variables that represent whether the words appear in the document, and high-level variables are latent binary variables that give soft partitions of the documents and furthermore represent topics. The latent tree model can discover substantially better topics and topic hierarchies.

In this paper, we mainly focus on learning latent tree models from samples. The learning of models includes two aspects. One is structural learning and the other is parameter estimation. The scoring-based algorithms (Zhang, 2004; Chen et al., 2016) search the optimal structure by hill-climbing with a scoring metric such as AIC (Akaike, 1974) or BIC (Schwarz, 1978). And the distance-based algorithms (Choi et al., 2011; Wang et al., 2017) can reconstruct the latent structure with an additive information distance. They are usually faster than the scoring-based ones and have theoretical guarantees for the structural learning. Once the latent structure is given, the matrix decomposition (Chang, 1996; Wang et al., 2017) and tensor decomposition (Anandkumar et al., 2014) are efficient for the parameter estimation of discrete latent tree models. Moreover, Song et al. (2011) proposed a method based on the kernel embedding of distributions for latent tree models with continuous non-Gaussian observation.

Though above algorithms contribute a lot to learning latent tree models, they can only handle pure discrete or continuous data. Typical multivariate problems may contain both continuous and discrete variables in the population survey data, biological and biomedical data, etc. In the graphical model setting, Lee and Hastie (2014) proposed a pseudo-likelihood method for handling mixed Gaussian and multinomial data. Fan et al. (2017) assumed that the observed binary data are obtained by dichotomizing a latent continuous variable, and proposed a semi-parametric model for modelling mixed continuous and binary data. To our best knowledge, there are no existing works on learning the hierarchical tree structure with latent variables for mixed data.

In this paper, we address the latent hierarchical information mining for mixed data, and contribute three points for this mining. Firstly, we introduce a mixed latent tree model for modeling mixed data, and define an information distance between discrete and continuous variables. Secondly, we propose a bottom-up structural learning algorithm basing on this information distance. This structural learning algorithm has the probabilistic approximate consistency. Thirdly, we suggest a moment method for parameter estimation by exploiting matrix decomposition. Those moment estimators are asymptotically normal.

The rest of this paper is organized as follows. In Section 2, some notions on latent tree model are reviewed and some assumptions used in this paper are given. In Section 3, we define a new information distance and give a finite sample bound required for the exact structural recovery. In Section 4, we propose a parameter estimation method and prove its asymptotic normality for mixed latent tree models. The simulation studies and the real data analysis are conducted in Section 5.

2. Preliminaries

Let $G = (\mathbf{W}, \mathbf{E})$ be a simple graph, where $\mathbf{W}$ is the set of nodes and $\mathbf{E}$ is the set of edges. An edge between node u and node v is denoted by (u, v), and we call that u is adjacent to v. If edge $(v_{j-1}, v_j) \in \mathbf{E}$ for any $j = 1, \cdots, k$, the set of distinct nodes $[v_0, v_1, \cdots, v_k]$ is referred to as a length-k path from v_0 to v_k in G. Furthermore, a path $[v_0, v_1, \cdots, v_k]$ is referred to as a cycle in G if $v_0 = v_k$. We call G a connected graph if for any nodes $u, v \in \mathbf{W}$, there is a path $[v_0 = u, \cdots, v_k = v]$ in G. Let $\mathbf{A}, \mathbf{B}$ be two disjoint node subsets. We call that $\mathbf{A}, \mathbf{B}$ are separated by a node subset $\mathbf{S}$ if for any nodes $u \in \mathbf{A}, v \in \mathbf{B}$, every path in G from u to v contains a node in $\mathbf{S}$. A connected simple acyclic graph is called a tree and denoted as T. A pair of leaves $\{u, v\}$ is a sibling pair on T if nodes u and v in T are adjacent to a same node. The number of nodes on the longest path of a tree T is referred to as the diameter of the tree and we denote it as $diam(T)$.

Let $\mathbf{X}_W = \{X_v\}_{v \in \mathbf{W}}$ be a random vector where $\mathbf{W}$ corresponds to a set of nodes, and let $\mathbf{X}_W^{(1)}, \ldots, \mathbf{X}_W^{(n)}$ denote *i.i.d.* samples of size n. A family of probability distribution over G is referred to as a graphical model (Lauritzen, 1996), if it satisfies the conditional independence: $\mathbf{X}_\mathbf{A}, \mathbf{X}_\mathbf{B}$ are conditionally independent given $\mathbf{X}_\mathbf{S}$ when two disjoint node subsets $\mathbf{A}, \mathbf{B}$ are separated by a node subset $\mathbf{S}$ in G . Let $T = (\mathbf{W}, \mathbf{E})$ be a tree. If the leaves of T are all observed variables and the internal nodes are latent variables, the graphical model $\mathcal{T}$ on T is referred to as a latent tree model (Zhang, 2004). Furthermore, the graphical model $\mathcal{T}$ is called the mixed latent tree model, if the tree T contains both discrete and continuous variables. In this paper, we mainly discuss one mixed case that latent variables are binary, and observed variables are binary or conditional Gaussian given its adjacent variable.

We denote the set of observed nodes in $\mathbf{W}$ as $\mathbf{V}$(with $m = |\mathbf{V}|$), and the set of latent nodes in $\mathbf{W}$ as $\mathbf{H}$. Hence $\mathbf{W} = \mathbf{V} \cup \mathbf{H}$. Furthermore, we denote the set of continuous nodes in $\mathbf{V}$ as $\mathbf{V}_c$, and the set of discrete nodes in $\mathbf{V}$ as $\mathbf{V}_d$. Hence $\mathbf{V} = \mathbf{V}_c \cup \mathbf{V}_d$. We refer to u, v as bifurcation nodes of w if nodes u, v are observed and the path from u to v contains w.

If we choose a node on tree T as the root, we can obtain a directed tree $\overrightarrow{T} = (\mathbf{W}, \overrightarrow{\mathbf{E}})$ by assigning the edge direction from the root to the leaves. The element $u \to v$ in $\overrightarrow{\mathbf{E}}$ represents a directed edge from u to v. Node u is called as a parent of node v and node v is called as a child of node u. We denote all child nodes of u as $ch(u)$ and denote the parent node of v as $pa(v)$. An ordered set of distinct nodes $\mathbf{L} = [v_0, v_1, \cdots, v_k]$ is a length-k directed path from v_0 to v_k in $\overrightarrow{T}$ if the directed edge $v_{j-1} \to v_j \in \overrightarrow{\mathbf{E}}$ for all $j = 1, \cdots, k$. If there exists a directed path in $\overrightarrow{T}$ from u to an observed node v, we say v is a bifurcation node of u in $\overrightarrow{T}$.

Here we consider three assumptions to characterize the relationship between variables.

(A1) The correlation coefficient of any two variables is nonzero.
(A2) Each latent variable has three neighbors at least.
(A3) Any two variables connected by an edge on the tree are not completely dependent.

These assumptions are routinely used in the graphical model setting because they can provide a guarantee for the identifiability of the graphical tree model. If Assumption (A1) is violated, our learning model may be disconnected, thus the model is not a graphical tree model. This assumption can be relaxed when we consider a graphical forest model with several connected components. Assumptions (A2) and (A3) ensure that a latent tree does not include a redundant latent node (Choi et al., 2011). If (A2) or (A3) is violated, there may exist a redundant latent variable in the tree model. This can further cause the non-identifiability of the model.

3. Structural Learning for Mixed Latent Tree Models

In this section, we first define an information distance between binary and continuous variables. And then we propose a bottom-up structural learning algorithm basing on this information distance. Finally, we give a finite sample bound for the exact structural recovery of this learning algorithm.

3.1 Information Distance

In this subsection, we define the information distance between two variables, and prove that the distance has an additivity along paths on mixed latent trees. This additivity is an important tool for designing a structural learning algorithm of latent tree models. The learning algorithm will be suggested in the next subsection.

Let $T = (\mathbf{W}, \mathbf{E})$ be a tree and $\mathcal{T}$ be a mixed latent tree model on T. For variable X_u and X_v where $u, v \in \mathbf{W}$, we define the information distance between them:

$$d_{uv} := -\log(|\rho_{uv}|), \tag{1}$$

where $\rho_{uv} = \frac{\mathrm{Cov}(X_u,X_v)}{\sqrt{\mathrm{Var}(X_u)\cdot\mathrm{Var}(X_v)}}$ is the correlation coefficient of variables X_u and X_v.

The correlation coefficient relies on the covariance of two random variables. By the double expectation formula, the covariance may be further decomposed into a product of several quantities in term of the conditional independence on the graphical tree model. For some specific distributions, the correlation coefficient ρ_{uv} could be presented as the product of two correlation coefficients ρ_{uh} and ρ_{hv} when variables X_u and X_v are conditional independent given variable X_h. Thus the information distance (1) has the additivity on the graphical tree model for an appropriate distribution. For binary variable X_u and X_v, the form (1) is equivalent to the information distance (Chang, 1996)

$$d_{uv} := -\log\left(\frac{|\det(P_{uv})|}{\sqrt{\det(P_{uu})\det(P_{vv})}}\right),$$

where P_{uv} is the joint probability matrix $(P(X_u = a, X_v = b))_{a,b=0,1}$ of X_u and X_v. For Gaussian variables X_u and X_v, the form in (1) is also the information distance (Choi et al., 2011). In this paper, we mainly consider that variables are binary or conditional Gaussian.

According to the Assumptions (A1) and (A3), we know that the correlation coefficient $0 < |\rho_{uv}| < 1$ and further the information distance $0 < d_{uv} < +\infty$ for any $u, v \in \mathbf{W}$. The following theorem establishes that the information distance is additive along paths. The proof can be found in the Appendix.

Theorem 1 *Let $T = (\mathbf{W}, \mathbf{E})$ be a tree and $\mathcal{T}$ be a mixed latent tree model on T. If node set $[v_0 = u, v_1, \cdots, v_k = v]$ is a path from u to v on T, we obtain that:*

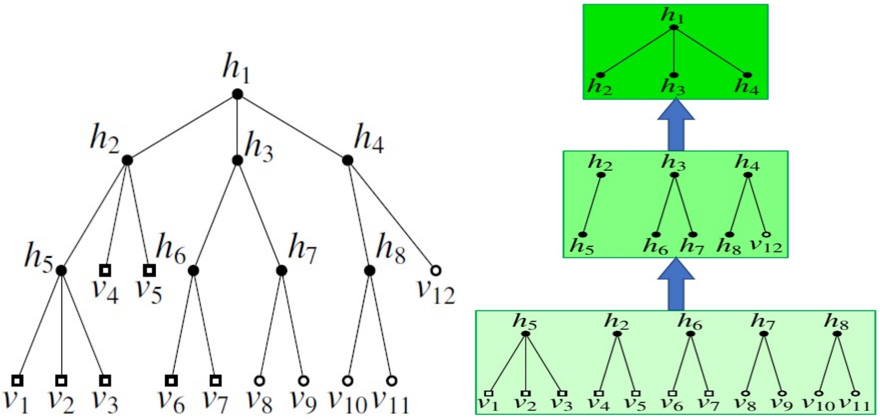
$$d_{uv} = \sum_{l=0}^{k-1} d_{v_l v_{l+1}}.$$

The proof of Theorem 1 mainly employs conditional independence in the tree model and the restriction that latent variables are binary. So the additivity of the information distance does not rely on the specific distributions of observed variables. The structural learning algorithm in the next subsection could also apply to mixed data from multinomial distributions or conditional non-Gaussian distributions.

3.2 Structural Learning Algorithm for Mixed Latent Tree Models

From Theorem 1, the information distance (1) has the additivity in the mixed latent tree model. Our structural learning method is from the bottom-up *SLLT* algorithm (Wang et al., 2017). This algorithm itself does not rely on the type of data. It only requires information distances as its inputs for recovering latent trees. This algorithm can output the mixed latent tree correctly within the time $O(\text{diam}(T)m^3)$ if the true information distances are available, where m is the number of observed variables.

In the following, we illustrate the *SLLT* algorithm in detail by using the mixed latent tree T shown in Figure 1, where $v_1, \cdots, v_{12}$ are observed variables and $h_1, \cdots, h_8$ are latent variables. We refer to $\mathbf{V} = \{v_1, \cdots, v_{12}\}$ and $\mathbf{H} = \{h_1, \cdots, h_8\}$ as the observed variable set and the latent variable set respectively, and let $\mathbf{W} = \mathbf{V} \cup \mathbf{H}$. We use node symbols $\bullet$, $\circ$, $\square$ to represent a latent variable, a discrete observed variable, a continuous observed variable respectively. The *SLLT* algorithm uses the information distances among the observed variable set $\mathbf{V}$ of T to reconstruct the unknown latent tree structure.



(a) The mixed latent tree structure (b) The process of structural learning

Figure 1: Example of structural learning.

For any three observed variables v_i, v_j and v_k, we compute all the information distance differences $\Phi_{v_i v_j v_k} = d_{v_i v_k} - d_{v_j v_k}$. When $\Phi_{v_1 v_2 v_3} = \Phi_{v_1 v_2 v_4} = \cdots = \Phi_{v_1 v_2 v_{12}}$, then

Algorithm 1: Structural Learning for Latent Trees (*SLLT*)

```
Input  : Observed variables $\mathbf{V}$ and information distances $d_{uv}$ for any $u, v \in \mathbf{V}$;
Output: A tree structure $T$;

1. $\mathbf{A} \leftarrow \mathbf{V}$. $\mathbf{D} \leftarrow \phi$. For any $u \in \mathbf{V}$, $\mathbf{D}(u) \leftarrow \phi$;
2. If $|\mathbf{A}| \geqslant 3$, compute $\Phi_{v_1 v_2 v_3} = d_{v_1 v_3} - d_{v_2 v_3}$ for any three variables $v_1, v_2, v_3 \in \mathbf{A}$;

    1°. For any $v_1, v_2 \in \mathbf{A}$,
            if $\Phi_{v_1 v_2 z}$ is constant for any $z \in \mathbf{A}\backslash\{v_1, v_2\}$, then
            $\{v_1, v_2\}$ are a sibling pair in $T$.
    2°. Denote maximal sibling groups by $\{\Pi_l\}_{l=1}^{L}$,
            $\mathbf{A} \leftarrow \mathbf{A}\backslash \overset{L}{\underset{l=1}{\cup}} \Pi_l$.
    3°. For any $l = 1, \cdots, L$,
            add a new latent variable $h_l$ and connect $h_l$ to every node in $\Pi_l$.
            $\mathbf{A} \leftarrow \mathbf{A} \cup \{h_l\}$, $\mathbf{D}(h_l) \leftarrow \underset{u \in \Pi_l}{\cup}\{u\}$, $\mathbf{D} \leftarrow \mathbf{D} \cup (\underset{u \in \Pi_l}{\cup}\{u\})$.

3. While $|\mathbf{A}| \geqslant 3$,

    1°. For any $v \in \mathbf{A} \cap \mathbf{V}$ and $u \in \mathbf{A}\backslash\mathbf{V}$,
          $\mathbf{Z} \leftarrow \mathbf{V}\backslash(\mathbf{D}(u) \cup \{v\})$ and choose bifurcation variables $v_1, v_2$ of $u$ in $\mathbf{D}$.
            If $\Phi_{v v_1 z}$ is constant and $\Phi_{v v_1 z} \neq \Phi_{v v_1 v_2}$ for any $z \in \mathbf{Z}$, then $\{v, u\}$ is a sibling pair in $T(\mathbf{W}\backslash\mathbf{D})$.
    2°. For any two variables $u, w \in \mathbf{A}\backslash\mathbf{V}$,
          $\mathbf{Z} \leftarrow \mathbf{V}\backslash(\mathbf{D}(u) \cup \mathbf{D}(w))$ and choose bifurcation variables $v_1, v_2$ of $u$ and $v_3, v_4$ of $w$ in $\mathbf{D}$.
            If $\Phi_{v_1 v_3 z} = \Phi_{v_1 v_3 v_4}$ and $\Phi_{v_3 v_1 z} \neq \Phi_{v_3 v_1 v_2}$ for any $z \in \mathbf{Z}$, then $u$ is a remaining child of $w$ in
            $T(\mathbf{W}\backslash\mathbf{D})$.
        For any two variables $u, w \in \mathbf{A}\backslash\mathbf{V}$,
          if neither $u$ nor $w$ is a remaining child,
            $\mathbf{Z} \leftarrow \mathbf{V}\backslash(\mathbf{D}(u) \cup \mathbf{D}(w))$ and choose bifurcation variables $v_1, v_2$ of $u$ and $v_3, v_4$ of $w$ in $\mathbf{D}$.
            If $\Phi_{v_1 v_3 z}$ is constant and $\Phi_{v_1 v_3 z} \neq \Phi_{v_1 v_3 v_4}$, $\Phi_{v_3 v_1 z} \neq \Phi_{v_3 v_1 v_2}$ for any $z \in \mathbf{Z}$, then $\{u, w\}$ is a
            sibling pair in $T(\mathbf{W}\backslash\mathbf{D})$.
    3°. Denote the remaining child relations and maximal sibling groups by $\{\Pi_l\}_{l=1}^{L}$.
          $\mathbf{A} \leftarrow \mathbf{A}\backslash \overset{L}{\underset{l=1}{\cup}} \Pi_l$.
    4°. For any $l = 1, \cdots, L$,
          if $\Pi_l = \{u, w\}$ and $u$ is remaining child of $w$, then connect $u$ and $w$.
            $\mathbf{A} \leftarrow \mathbf{A} \cup \{w\}$, $\mathbf{D}(w) \leftarrow \mathbf{D}(w) \cup \mathbf{D}(u) \cup \{u\}$ and $\mathbf{D} \leftarrow \mathbf{D} \cup \{u\}$.
          if $\Pi_l$ is a sibling group, then add a new latent variable $h_l$ and connect $h_l$ to every node in $\Pi_l$.
            $\mathbf{A} \leftarrow \mathbf{A} \cup \{h_l\}$, $\mathbf{D}(h_l) \leftarrow \underset{u \in \Pi_l}{\cup}(\mathbf{D}(u) \cup \{u\})$ and $\mathbf{D} \leftarrow \mathbf{D} \cup (\underset{u \in \Pi_l}{\cup}\{u\})$.

4. If $|\mathbf{A}| = 2$, connect the two remaining variables in $\mathbf{A}$;
5. Return the structure generated;
```

$\{v_1, v_2\}$ is a sibling pair in T. Similarly, we obtain that $\{v_1, v_3\}$, $\{v_2, v_3\}$, $\{v_4, v_5\}$, $\{v_6, v_7\}$, $\{v_8, v_9\}$, $\{v_{10}, v_{11}\}$ are also 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 five latent variables h_5, h_2, h_6, h_7, h_8 are detected as their parent variables respectively. Then we have $\mathbf{D}_1 = \{v_1, \cdots, v_{11}\}$ and $\mathbf{A}_1 = \{h_2, h_5, h_6, h_7, h_8, v_{12}\}$, where the subscripts of $\mathbf{D}$ and $\mathbf{A}$ are used to indicate the iterative step. We construct a subtree $T(\mathbf{W}\backslash\mathbf{D}_1)$ by discarding all variables in $\mathbf{D}_1$ from T, and $\mathbf{A}_1$ contains all of the leaf variables $\{h_5, h_6, h_7, h_8, v_{12}\}$ of $T(\mathbf{W}\backslash\mathbf{D}_1)$.

We can recover local structures among observed variables at step 2 of the *SLLT* algorithm. At step 3, we reconstruct the structures with latent variables. Firstly, we find out the observed-latent sibling pairs in $\mathbf{A}_1$. For $v_{12}, h_8 \in \mathbf{A}_1$, we choose the observed variables v_{10}, v_{11} as bifurcation variables of h_8 in $\mathbf{D}_1$. Since $\Phi_{v_{12}v_{10}v}$ is constant and $\Phi_{v_{12}v_{10}v} \neq \Phi_{v_{12}v_{10}v_{11}}$ for $v = v_1, v_2, \cdots, v_9$, we have that $\{v_{12}, h_8\}$ is a sibling pair in $T(\mathbf{W}\backslash\mathbf{D}_1)$. Secondly, we find out the remaining-child relationship in $\mathbf{A}_1$. For $\{h_2, h_5\}$ in $\mathbf{A}_1$, we choose $\{v_4, v_5\}$ as bifurcation variables of h_2 and $\{v_1, v_2\}$ as bifurcation variables of h_5. Since $\Phi_{v_1v_4v} = \Phi_{v_1v_4v_5}$ and $\Phi_{v_4v_1v} \neq \Phi_{v_4v_1v_2}$ for $v = v_6, \cdots, v_{12}$, we find that h_5 is a remaining child variable of h_2. Thirdly, we judge the latent-latent sibling pair relationship in $\mathbf{A}_1$. For $\{h_6, h_7\}$ in $\mathbf{A}_1$, we choose $\{v_6, v_7\}$ as bifurcation variables of h_6 and $\{v_8, v_9\}$ as bifurcation variables of h_7. Since $\Phi_{v_6v_8v}$ is constant and $\Phi_{v_6v_8v} \neq \Phi_{v_6v_8v_9}$, $\Phi_{v_8v_6v} \neq \Phi_{v_8v_6v_7}$ for $v = v_1, \cdots, v_5, v_{10}, v_{11}, v_{12}$, we have that $\{h_6, h_7\}$ is a sibling pair in $T(\mathbf{W}\backslash\mathbf{D}_1)$. Thus, $\{h_6, h_7\}$ and $\{h_8, v_{12}\}$ are two maximal sibling groups, and two latent variables h_3, h_4 are added as their parent variables respectively. Then, we have $\mathbf{D}_2 = \{v_1, \cdots, v_{12}, h_5, \cdots, h_8\}$ and $\mathbf{A}_2 = \{h_2, h_3, h_4\}$. Similarly, $\mathbf{A}_2$ contains all the leaf variables $\{h_2, h_3, h_4\}$ in the subtree $T(\mathbf{W}\backslash\mathbf{D}_2)$. Finally, we obtain that $\{h_2, h_3, h_4\}$ forms a sibling group through similar steps. Then we add a latent variable h_1 as its parent variable and the algorithm ends.

3.3 Finite Sample Bound for the Structural Learning Algorithm

In this subsection, we give a finite sample bound for the exact structural recovery of this learning algorithm. To apply the *SLLT* algorithm to data, we replace the correlation coefficient ρ_{uv} with its sample version

$$\hat{\rho}_{uv} = \frac{\sum\limits_{k=1}^{n}\left(X_u^{(k)} - \bar{X}_u\right)\left(X_v^{(k)} - \bar{X}_v\right)}{\sqrt{\sum\limits_{k=1}^{n}\left(X_u^{(k)} - \bar{X}_u\right)^2}\sqrt{\sum\limits_{k=1}^{n}\left(X_v^{(k)} - \bar{X}_v\right)^2}}$$

for nodes $u, v \in \mathbf{V}$. Furthermore, we compute the sample information distances $\hat{d}_{uv} = -\log|\hat{\rho}_{uv}|$, and put them into the *SLLT* algorithm.

In the algorithm, we identify the relations between variables by checking whether the information distance difference Φ_{uvw} is equal to some constant or not. However, in the sample-based *SLLT* algorithm, $\hat{\Phi}_{uvw}$ and $\hat{\Phi}_{uvz}$ are almost impossible to be exactly equal even if the true differences are equal due to the error of estimate. Since $|\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}| \to |\Phi_{uvw} - \Phi_{uvz}|$ when $n \to \infty$, we determine the equality of Φ_{uvw} and Φ_{uvz} if $|\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}| < \varepsilon$, where ε is a prescribed positive threshold. Moreover, we define a lower bound notation $\phi_{min} := \min\{|\Phi_{uvw} - \Phi_{uvz}| : \Phi_{uvw} \neq \Phi_{uvz}, u, v, w, z \in \mathbf{V}\}$ and take a threshold $\varepsilon \leqslant \min\{\frac{1}{2}\phi_{min}, 1\}$. If the difference $|(\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}) - (\Phi_{uvw} - \Phi_{uvz})| < \varepsilon$ when the sample size n is sufficiently large, we obtain that $\Phi_{uvw} = \Phi_{uvz}$ if and only if $|\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}| < \varepsilon$. Therefore, if the event $\{|(\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}) - (\Phi_{uvw} - \Phi_{uvz})| < \varepsilon$ for any $u, v, w, z \in \mathbf{V}\}$ occurs with a high probability when the sample size n is sufficiently large, we can learn the true latent tree structure from the sample-based *SLLT* algorithm with a high probability.

The consistency of the sample-based *SLLT* algorithm is built on the tail probability inequality on the sample covariance

$$S_{uv} = \frac{1}{n}\sum_{l=1}^{n}(X_u^{(l)} - \overline{X}_u)(X_v^{(l)} - \overline{X}_v),$$

where u, v are two leaf nodes on the latent tree. We need the following notations:

$$\begin{aligned}\mu &= \max\left\{\left|\mathrm{E}\left(X_v|X_{pa(v)} = x\right)\right| : v \in \mathbf{V}_c, x = 0, 1\right\},\\ \sigma^2 &= \max\left\{\mathrm{Var}\left(X_v|X_{pa(i)} = x\right) : v \in \mathbf{V}_c, x = 0, 1\right\},\\ c &= C \cdot \max\{\sigma^4,\ \sigma^2\mu^2,\ \mu^4,\ \sigma^2,\ \mu^2,\ 1\},\end{aligned}$$

where the constant C does not depend on nodes on the tree.

Theorem 2 *The inequality*

$$P\left(\left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| > \sqrt{\frac{c(t + \log 48)}{n}}\right) \leqslant e^{-t} \tag{2}$$

holds for any two leaf nodes u, v in T and any $t > 0$.

The consistency of the *SLLT* algorithm relies on two intrinsic parameters $\phi_{min} := \min\{|\Phi_{uvw} - \Phi_{uvz}| : \Phi_{uvw} \neq \Phi_{uvz}, u, v, w, z \in \mathbf{V}\}$ and $c_{min} := \min_{u,v\in\mathbf{V}} |\mathrm{Cov}(X_u, X_v)|$ where $\mathbf{V}$ is the set of observed nodes. The following theorem shows the relationship between the sample size and the intrinsic parameters of the model when the true latent tree structure is learned.

Theorem 3 *Let $\eta \in (0, 1)$. The SLLT algorithm can return the true mixed latent tree with a probability of at least $1 - \eta$, if the sample size n is sufficiently large such that the inequality*

$$\sqrt{\frac{c(\log(48m^2) - \log\eta)}{n}} < \frac{c_{min}\min\{\frac{1}{2}\phi_{min}, 1\}}{16}, \tag{3}$$

holds.

Theorem 3 is obtained from Theorem 2, and it provides a lower bound

$$\frac{256 \cdot c(\log(48m^2) - \log\eta)}{c_{min}^2 \min\{\frac{1}{4}\phi_{min}^2, 1\}}$$

of the sample size n for recovering the true structure with a probability of at least $1 - \eta$. Intrinsic parameters ϕ_{min} and c_{min} depend on the number m of observed nodes and the true joint distributions. When the number m is large, intrinsic parameters may be small. Furthermore, a large sample size is required for recovering mixed latent tree structures with a high probability.

4. Parameter Estimation for Mixed Latent Tree Models

In this section, we consider the parameter estimation for the mixed latent tree model $\mathcal{T}$ with its tree structure T given. We choose a latent variable r as the root and further construct a directed tree $\overrightarrow{T}$ from the tree T and the root r. Model parameters (Chen et al., 2017) in the latent tree model consist of a marginal distribution for the root X_r, and all the conditional distributions for variables given their parents on the directed tree $\overrightarrow{T}$. In Subsections 4.1 and 4.2, we first assume that the expectation $\mu_u := \mathrm{E}X_u$ is zero for any continuous variable $u \in \mathbf{V}_c$. Under this assumption, we propose a moment method to estimate all the model parameters along the directed tree $\overrightarrow{T}$ by using matrix decomposition. Finally, we also discuss the parameter estimates for non-zero expectation continuous variables.

First, we introduce some notations. For a latent node $h \in \mathbf{H}$, let

$$\begin{aligned}
p_h &:= P(X_h = 1), \\
\mu_{v|X_h=x_h} &:= \mathrm{E}(X_v|X_h = x_h) \text{ for a continuous observed node } v, \\
\mu^{(2)}_{v|X_h=x_h} &:= \mathrm{E}(X_v^2|X_h = x_h) \text{ for a continuous observed node } v, \\
p_{v|X_h=x_h} &:= P(X_v = 1|X_h = x_h) \text{ for a discrete observed node } v,
\end{aligned}$$

where $x_h = 0, 1$. In particular, if node h is a parent of node v on the tree $\overrightarrow{T}$, we simplify the notations $\mu_{v|X_h=x_h}$, $\mu^{(2)}_{v|X_h=x_h}$, $p_{v|X_h=x_h}$ by $\mu_{v|x_h}$, $\mu^{(2)}_{v|x_h}$, $p_{v|x_h}$ respectively.

4.1 Parameter Representation for Mixed Latent Tree Models

Motivated by Chang (1996), we study a local structure consisting of three observed nodes $u, v, w \in \mathbf{V}$ and a latent node $h \in \mathbf{H}$. u and v are bifurcation nodes of h. v and w are also bifurcation nodes of h. Figure 2 illustrates this local structure, which implies that observed variables X_u, X_v, X_w are conditionally independent given the latent variable X_h. Unlike Chang's work (Chang, 1996) only considering discrete variables, we allow variables X_u, X_v, X_w to be continuous. So there are eight cases of three observed variables depending on whether the variable is binary or continuous. In the following, we first provide the parameter representation of models in the case that X_u, X_v, X_w are all continuous. And then we show the general representation in other cases.

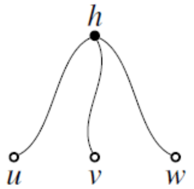


Figure 2: A local structure of three observed nodes and one latent node.

If nodes $u, v, w \in \mathbf{V}_c$, the moment $\mathrm{E}X_u^{l_u} X_v^{l_v} X_w$ can be decomposed into

$$\mathrm{E}X_u^{l_u} X_v^{l_v} X_w = \begin{pmatrix} \mu^{(l_u)}_{u|X_h=0} & \mu^{(l_u)}_{u|X_h=1} \end{pmatrix} \begin{pmatrix} (1-p_h)\,\mu_{w|X_h=0} & 0 \\ 0 & p_h\,\mu_{w|X_h=1} \end{pmatrix} \begin{pmatrix} \mu^{(l_v)}_{v|X_h=0} \\ \mu^{(l_v)}_{v|X_h=1} \end{pmatrix}$$

for exponents $l_u, l_v \in \{1,2\}$, since X_u, X_v, X_w are conditionally independent given X_h. We denote E_{uvw}, $\Gamma_{u|h}$, and $\Gamma_{v|h}$ as matrices $\begin{pmatrix} \mathrm{E}X_uX_vX_w & \mathrm{E}X_uX_v^2X_w \\ \mathrm{E}X_u^2X_vX_w & \mathrm{E}X_u^2X_v^2X_w \end{pmatrix}$, $\begin{pmatrix} \mu_{u|X_h=0} & \mu_{u|X_h=1} \\ \mu^{(2)}_{u|X_h=0} & \mu^{(2)}_{u|X_h=1} \end{pmatrix}$, and $\begin{pmatrix} \mu_{v|X_h=0} & \mu_{v|X_h=1} \\ \mu^{(2)}_{v|X_h=0} & \mu^{(2)}_{v|X_h=1} \end{pmatrix}$, respectively. We further have that

$$E_{uvw} = \Gamma_{u|h} \begin{pmatrix} (1-p_h)\,\mu_{w|X_h=0} & 0 \\ 0 & p_h\,\mu_{w|X_h=1} \end{pmatrix} \Gamma^T_{v|h}. \tag{4}$$

Denote E_{uv} as $\begin{pmatrix} \mathrm{E}X_uX_v & \mathrm{E}X_uX_v^2 \\ \mathrm{E}X_u^2X_v & \mathrm{E}X_u^2X_v^2 \end{pmatrix}$. We also have that

$$E_{uv} = \Gamma_{u|h} \cdot \begin{pmatrix} 1-p_h & 0 \\ 0 & p_h \end{pmatrix} \cdot \Gamma^T_{v|h}. \tag{5}$$

By Assumption (A1), the matrix E_{uv} is invertible since continuous variables have zero means. Thus

$$A_{uvw} := E_{uvw}E_{uv}^{-1} = \Gamma_{u|h} \cdot \begin{pmatrix} \mu_{w|X_h=0} & 0 \\ 0 & \mu_{w|X_h=1} \end{pmatrix} \cdot \Gamma^{-1}_{u|h}.$$

This eigen-decomposition of matrix A_{uvw}, determined by the joint distribution of observed variables (X_u, X_v, X_w), forms the representation of model parameters. Specifically, conditional means $\mu_{w|X_h=0}, \mu_{w|X_h=1}$ are eigenvalues of matrix A_{uvw}. Columns of conditional moment matrix $\Gamma_{u|h}$ are eigenvectors of matrix A_{uvw}. To compute $\Gamma_{u|h}$ from the eigenvector space of matrix A_{uvw}, we still need two other restrictions:

$$\mathrm{E}X_w = (1-p_h)\cdot\mu_{w|X_h=0} + p_h\cdot\mu_{w|X_h=1}, \tag{6}$$

and

$$\begin{pmatrix} \mathrm{E}X_u \\ \mathrm{E}X_u^2 \end{pmatrix} = \Gamma_{u|h} \cdot \begin{pmatrix} 1-p_h \\ p_h \end{pmatrix}. \tag{7}$$

Combining equations (4), (5), (6) and (7), we can compute the conditional expectation matrix $\Gamma_{u|h}$ and the marginal probability $p_h = P(X_h = 1)$ using the moments of observed variables. Particularly, if node h is the root r, we can actually obtain the marginal probability of the root.

In above discussion, we show that in the case that X_u, X_v, X_w are all continuous, model parameters can be obtained by solving the moment equations (4), (5), (6) and (7). In the following part, we handle the general case that allows observed variables binary. For any

$w \in \mathbf{V}$, let

$$E_{uvw} := \begin{cases} \begin{pmatrix} \mathrm{E}X_uX_vX_w & \mathrm{E}X_uX_v^2X_w \\ \mathrm{E}X_u^2X_vX_w & \mathrm{E}X_u^2X_v^2X_w \end{pmatrix}, & \text{for } u,v \in \mathbf{V}_c; \\[2ex] \begin{pmatrix} \mathrm{E}X_u(1-X_v)X_w & \mathrm{E}X_uX_vX_w \\ \mathrm{E}X_u^2(1-X_v)X_w & \mathrm{E}X_u^2X_vX_w \end{pmatrix}, & \text{for } u \in \mathbf{V}_c \text{ and } v \in \mathbf{V}_d; \\[2ex] \begin{pmatrix} \mathrm{E}(1-X_u)X_vX_w & \mathrm{E}(1-X_u)X_v^2X_w \\ \mathrm{E}X_uX_vX_w & \mathrm{E}X_uX_v^2X_w \end{pmatrix}, & \text{for } u \in \mathbf{V}_d \text{ and } v \in \mathbf{V}_c; \\[2ex] \begin{pmatrix} \mathrm{E}(1-X_u)(1-X_v)X_w & \mathrm{E}(1-X_u)X_vX_w \\ \mathrm{E}X_u(1-X_v)X_w & \mathrm{E}X_uX_vX_w \end{pmatrix}, & \text{for } u,v \in \mathbf{V}_d, \end{cases}$$

$$E_{uv} := \begin{cases} \begin{pmatrix} \mathrm{E}X_uX_v & \mathrm{E}X_uX_v^2 \\ \mathrm{E}X_u^2X_v & \mathrm{E}X_u^2X_v^2 \end{pmatrix}, & \text{for } u,v \in \mathbf{V}_c; \\[2ex] \begin{pmatrix} \mathrm{E}X_u(1-X_v) & \mathrm{E}X_uX_v \\ \mathrm{E}X_u^2(1-X_v) & \mathrm{E}X_u^2X_v \end{pmatrix}, & \text{for } u \in \mathbf{V}_c \text{ and } v \in \mathbf{V}_d; \\[2ex] \begin{pmatrix} \mathrm{E}(1-X_u)X_v & \mathrm{E}(1-X_u)X_v^2 \\ \mathrm{E}X_uX_v & \mathrm{E}X_uX_v^2 \end{pmatrix}, & \text{for } u \in \mathbf{V}_d \text{ and } v \in \mathbf{V}_c; \\[2ex] \begin{pmatrix} \mathrm{E}(1-X_u)(1-X_v) & \mathrm{E}(1-X_u)X_v \\ \mathrm{E}X_u(1-X_v) & \mathrm{E}X_uX_v \end{pmatrix}, & \text{for } u,v \in \mathbf{V}_d, \end{cases}$$

and also let A_{uvw} denote the matrix $E_{uvw}E_{uv}^{-1}$. Let

$$\Gamma_{u|h} := \begin{cases} \begin{pmatrix} \mu_{u|X_h=0} & \mu_{u|X_h=1} \\ \mu^{(2)}_{u|X_h=0} & \mu^{(2)}_{u|X_h=1} \end{pmatrix}, & u \in \mathbf{V}_c; \\ \begin{pmatrix} 1-p_{u|X_h=0} & 1-p_{u|X_h=1} \\ p_{u|X_h=0} & p_{u|X_h=1} \end{pmatrix}, & u \in \mathbf{W}\backslash(\mathbf{V}_c \cup \{r\}), \end{cases}$$

$$\Lambda_{w|h} := \begin{cases} \begin{pmatrix} \mu_{w|X_h=0} & 0 \\ 0 & \mu_{w|X_h=1} \end{pmatrix}, & w \in \mathbf{V}_c; \\ \begin{pmatrix} p_{w|X_h=0} & 0 \\ 0 & p_{w|X_h=1} \end{pmatrix}, & w \in \mathbf{V}_d. \end{cases}$$

Similar to the case that X_u, X_v, X_w are all continuous, the eigen-decomposition

$$A_{uvw} = \Gamma_{u|h} \cdot \Lambda_{w|h} \cdot \Gamma_{u|h}^{-1} \tag{8}$$

holds for any three observed nodes $u, v, w \in \mathbf{V}$. If node $u \in \mathbf{V}_c$, the matrix $\Gamma_{u|h}$ can be computed by a similar way in the case that X_u, X_v, X_w are all continuous. If node $u \in \mathbf{V}_d$, we can also obtain the matrix $\Gamma_{u|h}$ by replacing the restriction equations (6) and (7) with a natural restriction $\mathbf{1}^T\Gamma_{u|h} = (1, 1)$ (Wang et al., 2017). So for any observed node $u \in \mathbf{V}$, we can obtain the matrix $\Gamma_{u|h}$ using the eigen-decomposition. If node h is just a parent of node u on the tree $\overrightarrow{T}$, the matrix $\Gamma_{u|h}$ is exactly model parameters of conditional distribution for variable X_u given its parent variable X_h.

4.2 Parameter Estimation Algorithm for Mixed Latent Tree Models

Beyond the local structure shown in Figure 2, we consider a little more complex structure shown in Figure 3 (a) of four observed nodes and two latent nodes. Let node h_2 be the root of the tree. Nodes u, v_1 are two bifurcation nodes of h_1, and nodes v_2, w are two bifurcation nodes of h_2. By the separation in Figure 3 (b), variables X_u, X_{v_1}, X_w are conditionally independent given X_{h_1}, and variables X_u, X_{v_2}, X_w are conditionally independent given X_{h_2}. Hence parameter matrices $\Gamma_{u|h_1}$ and $\Gamma_{u|h_2}$ can be computed as discussed in Subsection 4.1. Furthermore, we get the parameter matrix

$$\Gamma_{h_1|h_2} = \Gamma_{u|h_1}^{-1} \cdot \Gamma_{u|h_2} \tag{9}$$

related to two latent variables.

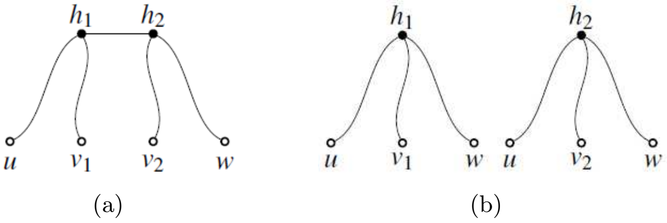


Figure 3: A local structure of four observed nodes and two latent nodes.

As discussed above, we can compute all the model parameters by using the joint distributions of three observed variables. Then we further suggest the $PEMT$ algorithm for the parameter estimation of mixed latent tree models. According to Assumption (A2), every latent variable h has three neighbors at least. If $|C| = 2$ in step 8 of the $PEMT$ algorithm, there exists an observed variable $c \in \mathbf{V}$ such that the path from h to c on T does not contain any child of h. To guarantee that the column states of all $\Gamma_{u_1|h}, \Gamma_{u_2|h}, \cdots$ are matched for h, we need to record the label states of u_1 and u_2 from h and perform the corresponding matrix decomposition in equation (8) according to the label states. Furthermore, since the parameter matrix $\Gamma_{u|h_1}$ is invertible, we can compute the parameter matrix $\Gamma_{h_1|h_2}$ related to two latent variables by the equation (9).

Our parameter estimation algorithm can reduce to the $PELT$ algorithm (Wang et al., 2017) if the data is pure binary. Specifically, the matrices

$$\begin{aligned} E_{uvw} &= \begin{pmatrix} \mathrm{E}(1-X_u)(1-X_v)X_w & \mathrm{E}(1-X_u)X_vX_w \\ \mathrm{E}X_u(1-X_v)X_w & \mathrm{E}X_uX_vX_w \end{pmatrix} \\ &= \begin{pmatrix} Pr(X_u=0, X_v=0, X_w=1) & Pr(X_u=0, X_v=1, X_w=1) \\ Pr(X_u=1, X_v=0, X_w=1) & Pr(X_u=1, X_v=1, X_w=1) \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned} E_{uv} &= \begin{pmatrix} \mathrm{E}(1-X_u)(1-X_v) & \mathrm{E}(1-X_u)X_v \\ \mathrm{E}X_u(1-X_v) & \mathrm{E}X_uX_v \end{pmatrix} \\ &= \begin{pmatrix} Pr(X_u=0, X_v=0) & Pr(X_u=0, X_v=1) \\ Pr(X_u=1, X_v=0) & Pr(X_u=1, X_v=1) \end{pmatrix}. \end{aligned}$$

Algorithm 2: Parameter Estimation for Mixed Latent Trees ($PEMT$)

```
Input : A latent tree with a root, the first order moments $\mathrm{E}X_u$ for $u \in \mathbf{V}$, the
        second order moments $\mathrm{E}X_u^2$ for $u \in \mathbf{V}_c$ and the moment matrices $E_{uvw}$
        and $E_{uv}$ for $u, v, w \in \mathbf{V}$.
Output: All conditional probability matrices on edges in $T$.
Construct a directed tree $\overrightarrow{T}$ and compute the matrices $A_{uvw}$ for $u, v, w \in \mathbf{V}$.
for $h \in \mathbf{H}$ do
    find all child variables $ch(h)$ of $h$ in $T$ ;
    for $z \in ch(h)$ do
        find a directed bifurcation variable $u$ of $z$.
    end for
    Collect the set $C$ of all the bifurcation variables $\{u_1, u_2, \cdots\}$ of all child variables
    $ch(h)$ of $h$.
    if $|C| = 2$ then
        find an observed variable $c \in \mathbf{V}$ such that the path from $h$ to $c$ on $T$ does not
        contain any child of $h$.
    end if
    Compute $\Gamma_{u_1|h}, \Gamma_{u_2|h}, \cdots$ and $p_h$ by matrix decomposition in equation (8).
end for
for $h_2 \in \mathbf{H}$ and $h_1 \in ch(h_2)$ do
    if $h_1 \in \mathbf{H}$ then
        choose a common directed bifurcation variable $u$ of $h_1$ and $h_2$, and compute
        parameter matrix $\Gamma_{h_1|h_2}$ by equation (9) ;
    end if
end for
Return the parameters $p_r$ and $\Gamma_{u|pa(u)}$, $\Gamma_{h|pa(h)}$ for $u \in \mathbf{V}, h \in \mathbf{H} \backslash \{r\}$.
```

By direct computation, the matrix $E_{uvw}E_{uv}^{-1}$ has the spectral decomposition form used in the $PELT$ algorithm. For the sample-based version of the $PELT$, the work (Wang et al., 2017) lacks of the asymptotic normality guarantee for the algorithm's output. For the $PEMT$, we provide the asymptotic normality in the following Theorem 4, which is also applied to the $PELT$ for handling pure binary data.

For obtaining a sample-based $PEMT$ algorithm, we replace the moments by their sample moments. For the basic structure in Figure 2, we take the case that $u, v, w \in \mathbf{V}_c$ as an example. Moments E_{uvw}, E_{uv} $\mathrm{E}X_u, \mathrm{E}X_u^2$, and $\mathrm{E}X_w$ are replaced by their sample moments $\hat{E}_{uvw}$, $\hat{E}_{uv}$, $\bar{X}_u$, $\overline{X_u^2}$, and $\bar{X}_w$ respectively, where

$$\hat{E}_{uvw} = \begin{pmatrix} \frac{1}{n}\sum_{l=1}^n X_u^{(l)} X_v^{(l)} X_w^{(l)} & \frac{1}{n}\sum_{l=1}^n X_u^{(l)} (X_v^{(l)})^2 X_w^{(l)} \\ \frac{1}{n}\sum_{l=1}^n (X_u^{(l)})^2 X_v^{(l)} X_w^{(l)} & \frac{1}{n}\sum_{l=1}^n (X_u^{(l)})^2 (X_v^{(l)})^2 X_w^{(l)} \end{pmatrix},$$

$$\hat{E}_{uv} = \begin{pmatrix} \frac{1}{n}\sum_{l=1}^n X_u^{(l)} X_v^{(l)} & \frac{1}{n}\sum_{l=1}^n X_u^{(l)} (X_v^{(l)})^2 \\ \frac{1}{n}\sum_{l=1}^n (X_u^{(l)})^2 X_v^{(l)} & \frac{1}{n}\sum_{l=1}^n (X_u^{(l)})^2 (X_v^{(l)})^2 \end{pmatrix},$$

$$\bar{X}_u = \frac{1}{n}\sum_{l=1}^{n} X_u^{(l)}, \quad \overline{X_u^2} = \frac{1}{n}\sum_{l=1}^{n}(X_u^{(l)})^2, \quad \bar{X}_w = \frac{1}{n}\sum_{l=1}^{n} X_w^{(l)}.$$

By further solving equations (4), (5), (6) and (7), we obtain a moment estimator $\hat{\Gamma}_{u|h}$ of the true parameter matrix $\Gamma_{u|h}$. From the property of moment estimators in van der Vaart (2000), we have the following theorem illustrating that our estimator converges to the true one in the meaning of asymptotic normality. The detailed proof is put into the Appendix.

Theorem 4 *Assume that the expectation* $\mathrm{E}X_u$ *is zero for any continuous node* $u \in \mathbf{V}_c$*. In the sample-based* $PEMT$ *algorithm, the moment estimators* $\hat{p}_r$, $\hat{\Gamma}_{u|pa(u)}$ *and* $\hat{\Gamma}_{h|pa(h)}$ *satisfy* $\sqrt{n}(\hat{p}_r - p_r)$, $\sqrt{n}(\hat{\Gamma}_{u|pa(u)} - \Gamma_{u|pa(u)})$ *and* $\sqrt{n}(\hat{\Gamma}_{h|pa(h)} - \Gamma_{h|pa(h)})$ *are asymptotically normal, where* n *is the sample size,* r *is the root node and* $u \in \mathbf{V}, h \in \mathbf{H}\backslash\{r\}$*.*

Note that Theorem 4 requires a zero-expectation assumption. For node $u \in \mathbf{V}_c$, the zero expectation of variable X_u guarantees the non-singularity of the conditional moment matrix $\Gamma_{u|h}$ and the diagonal matrix $\Lambda_{w|h}$. So it also ensures that the eigen-decomposition (8) for matrix A_{uvw} is valid. Theoretically, we can replace non-zero mean continuous variables with zero mean continuous variables. Specifically, for a continuous node $u \in \mathbf{V}_c$, let $\widetilde{X}_u = X_u - \mu_u$ where $\mu_u = \mathrm{E}X_u$. Variable $\widetilde{X}_u$ has a zero expectation. Moreover, the conditional distribution $\widetilde{X}_u|X_{pa(u)} = x \sim N\left(\widetilde{\mu}_{u|x}, \widetilde{\sigma}^2_{u|x}\right)$, where the parameters $\widetilde{\mu}_{u|x} = \mu_{u|x} - \mu_u$ and $\widetilde{\sigma}^2_{u|x} = \sigma^2_{u|x}$. So there is an one-to-one map between the parameters for X_u and those for $\widetilde{X}_u$, if the expectation μ_u is known. By replacing all the continuous variables with their centralized variables, we obtain new matrices $\widetilde{E}_{uvw}, \widetilde{E}_{uv}$ and $\widetilde{A}_{uvw}$ for any observed nodes $u, v, w \in \mathbf{V}$. By the similar eigen-decomposition in Subsection 4.1, we can obtain $\widetilde{\mu}_{u|0}, \widetilde{\mu}_{u|1}, \widetilde{\mu}^{(2)}_{u|0}, \widetilde{\mu}^{(2)}_{u|1}$ for any node $u \in \mathbf{V}_c$. We can further compute the original model parameters $\mu_{u|0}, \mu_{u|1}, \sigma^2_{u|0}, \sigma^2_{u|1}$ by equations:

$$\mu_{u|x} = \widetilde{\mu}_{u|x} + \mu_u, \ \sigma^2_{u|x} = \widetilde{\sigma}^2_{u|x} = \widetilde{\mu}^{(2)}_{u|x} - \widetilde{\mu}^2_{u|x}, \quad \text{and} \quad \mu^{(2)}_{u|x} = \sigma^2_{u|x} + \mu^2_{u|x}, \tag{10}$$

where $x = 0, 1$.

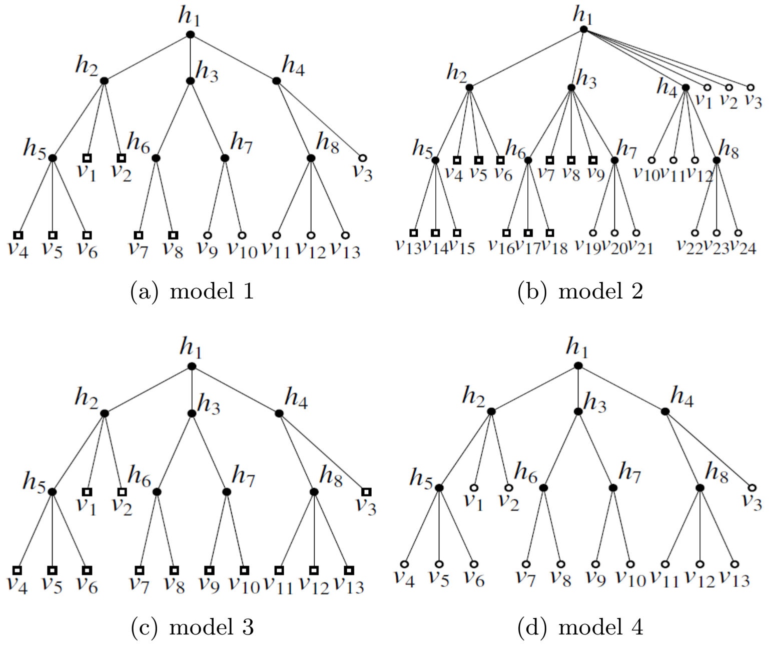
In the numerical computation, we also suggest to centralize all the continuous variables before using the $PEMT$ algorithm. For any node $u \in \mathbf{V}_c$, let $\widetilde{X}^{(l)}_u = X^{(l)}_u - \bar{X}_u$ and replace the original observation $X^{(l)}_u$ with $\widetilde{X}^{(l)}_u$ where $l = 1, \cdots, n$. We can obtain the parameter estimation $\{\hat{\widetilde{\mu}}_{u|0}, \hat{\widetilde{\mu}}_{u|1}, \hat{\widetilde{\mu}}^{(2)}_{u|0}, \hat{\widetilde{\mu}}^{(2)}_{u|1}, u \in \mathbf{V}_c\}$ by the sample-based $PEMT$ algorithm. Furthermore, the original parameters can be computed using equations in (10).

5. Numerical Experiment

In this section, we performed numerical experiments on both simulated and real data sets. In Subsection 5.1, we show the consistency of the $SLLT$ algorithm and the $PEMT$ algorithm on the simulated data, which was generated from four mixed latent tree structures. For parameter estimation, we compared the $PEMT$ algorithm with the conventional EM algorithm. For structural learning, we further performed a simulated experiment for high-dimensional data with one thousand observed variables. In Subsection 5.2, we apply our algorithm to a Forest Cover Type dataset for mining the hierarchical structure and latent information. All of the experiments were performed using R on a desktop with an Intel Core i5-3470 CPU 3.2 GHz and 16 GB RAM.

5.1 Simulation Study

We generated data sets from four mixed latent tree models shown in Figure 4. We use node symbols $\bullet$, $\circ$, $\square$ to represent a latent variable, a discrete observed variable, a continuous observed variable respectively. Models $1, 3$ and 4 have similar structures, but the ratios of the number of continuous variables to the number of observed variables set $\mathbf{V}$ are different. And the structure of model 1 is similar to that of model 2, where we restricted every latent variable to have three observed neighbors. The model parameters were generated randomly such that $|p_{u|0} - p_{u|1}| \geqslant 0.3$ for $u \in \mathbf{V}_d$ and $|\mu_{u|0} - \mu_{u|1}| \geqslant 0.5$ for $u \in \mathbf{V}_c$, which ensure that the information distances are limited.


h_1
h_2
h_3
h_4
h_5
v_1
v_2
h_6
h_7
h_8
v_3
v_4 v_5 v_6 v_7 v_8 v_9 v_{10} v_{11} v_{12} v_{13}
h_1
h_2
h_3
h_4 v_1 v_2 v_3
h_5 v_4 v_5 v_6
h_6 v_7 v_8 v_9
h_7 v_{10} v_{11} v_{12}
h_8
v_{13} v_{14} v_{15} v_{16} v_{17} v_{18} v_{19} v_{20} v_{21} v_{22} v_{23} v_{24}
h_1
h_2
h_3
h_4
h_5
v_1
v_2
h_6
h_7
h_8
v_3
v_4 v_5 v_6 v_7 v_8 v_9 v_{10} v_{11} v_{12} v_{13}
h_1
h_2
h_3
h_4
h_5
v_1
v_2
h_6
h_7
h_8
v_3
v_4 v_5 v_6 v_7 v_8 v_9 v_{10} v_{11} v_{12} v_{13}

(a) model 1

(b) model 2

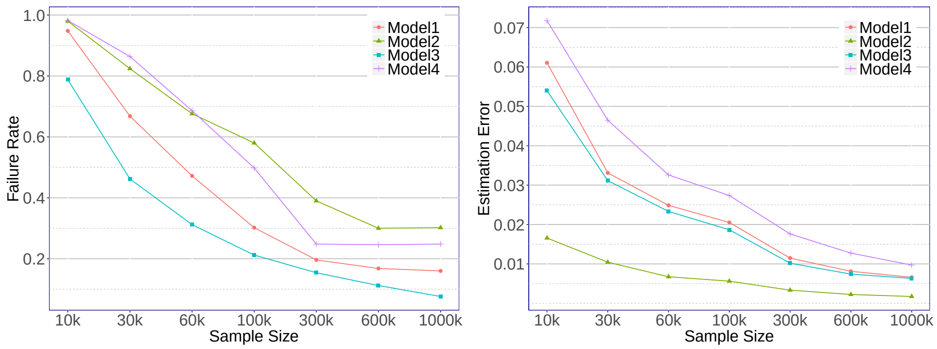
(c) model 3

(d) model 4

Figure 4: Four mixed latent tree models used in the simulation study.

As shown in Subsection 3.2, we determine the basic sibling pair by using a prescribed threshold $\varepsilon > 0$. Particularly, if the difference $|\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}| < \varepsilon$ for $w, z \in \mathbf{V}\backslash\{u, v\}$, we judge that $\{u, v\}$ is a sibling pair. According to Wang et al. (2017), 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. We only considered the possible sibling pair $\{u, v\}$ whose estimated distances $\hat{d}_{uv}, \hat{d}_{uw}, \hat{d}_{vw}$ are controlled by two thresholds τ_1, τ_2. In particular, for each pair of nodes $\{u, v\}$ satisfied $\hat{d}_{uv} < \tau_1$, the estimated difference $\hat{\Phi}_{uvw}$ was computed only for node $w \in \mathcal{K}_{uv} = \{w \in \mathbf{V}\backslash\{u, v\} \big| \max\{\hat{d}_{uw}, \hat{d}_{vw}\} < \tau_2\}$. Furthermore, if the difference $|\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}| < \varepsilon$ for any $w, z \in \mathcal{K}_{uv}$ and $\hat{d}_{uv} < \tau_1$, we consider $\{u, v\}$ to be a sibling pair. When we increase the threshold ε, it is apparent that the number of observed nodes belonging to the same sibling group tends to increase, while the number of individual nodes tends to decrease. So a larger ε makes it easier to obtain a tree structure. In the structural learning simulation, we started the value of the threshold ε from 0.1 and let it increase with the step size 0.1 until the $SLLT$ algorithm obtained a tree. We set $\tau_1 = 3$ and $\tau_2 = 5$.

To assess the consistency of the $SLLT$ algorithm and the $PEMT$ algorithm, we varied the sample size among $10k, 30k, 60k, 100k, 300k, 600k, 1000k$. For each sample size, we did 500 experiments with randomly generated parameters in four mixed latent tree models. The $SLLT$ algorithm may fail in one experiment if it does not find the real latent tree structure. The performance of the $SLLT$ was evaluated by its failure rate. For the $PEMT$ algorithm, its performance was assessed by the average estimate error in 500 experiments. Figure 5 shows that the failure rate and the estimation error decreased as the sample size increased.



(a) The $SLLT$ algorithm

(b) The $PEMT$ algorithm

Figure 5: The consistency of algorithms.

As shown in Figure 5 (a), the $SLLT$ algorithm performed much better with model 3 than model 4, which indicates that the structure of continuous variables is better to be recovered than that of discrete ones. The $SLLT$ algorithm worked better with model 1 than model 2, since the additional nodes in model 2 relative to model 1 increase the probability of misjudgment. Similarly, as shown in Figure 5 (b), the $PEMT$ algorithm had a better performance on the model 3 than the model 1. Moreover, the algorithm performed better with model 1 than model 4. As the proportion of the continuous parts increased in the observed variables, the performance of the algorithm could get better. The $PEMT$ algorithm performed best with model 2, since the additional leaves in model 2 relative to model 1 can enhance the parameter estimation accuracy (Wang et al., 2017).

We also compared the performance of the EM algorithm and the $PEMT$ algorithm using those mixed latent tree models. We varied the sample size among $10k, 30k, 60k, 100k$. Given a group of model parameters, we generated 100 random datasets for each sample size and each of the four structures in Figure 4, where node symbols $\bullet$, $\circ$, $\Box$ represent a latent variable, a discrete observed variable, a continuous observed variable respectively. We ran the EM with random initialization and 100 iterations. In Table 1, we list the average estimation accuracy and the average time costs for both algorithms. The estimation accuracy is measured by the mean absolute error (MAE) and the mean squared error (MSE). The MSEs are listed in the brackets of Table 1.

Table 1: Estimation Accuracy

Model	Sample Size	Estimation Accuracy ($\times 10^{-2}$)		Time Costs (seconds)	
		EM	$PEMT$	EM	$PEMT$
1	10k	3.38542(0.45116)	5.96665(1.79734)	245.27	1.48
	30k	3.18377(0.43035)	3.74196(0.79397)	706.07	1.50
	60k	3.12205(0.43014)	2.78707(0.439)	1380.06	1.54
	100k	3.09361(0.42536)	2.01234(0.20053)	2328.86	1.61
2	10k	0.71304(0.01021)	1.93799(0.21545)	379.61	1.75
	30k	0.41582(0.00354)	1.14539(0.03704)	1103.09	1.80
	60k	0.30062(0.00183)	0.79623(0.01761)	2187.71	1.88
	100k	0.23239(0.00114)	0.62655(0.01071)	3701.19	1.97
3	10k	1.72369(0.17865)	5.34733(1.54988)	379.64	1.46
	30k	1.49627(0.17578)	3.18196(0.43873)	1098.84	1.49
	60k	1.42979(0.1789)	2.23121(0.19593)	2193.00	1.54
	100k	1.38426(0.17731)	1.6833(0.11438)	3664.98	1.60
4	10k	6.93965(1.06076)	6.73677(1.4187)	86.61	1.47
	30k	6.81237(1.05467)	4.37126(0.67131)	225.16	1.50
	60k	6.77787(1.04613)	3.39273(0.40884)	443.07	1.54
	100k	6.77533(1.04971)	2.58861(0.24045)	742.43	1.61

As the sample size went up, the estimation accuracy of the $PEMT$ improved quickly with a little increasing time cost, while the time cost of the EM increased rapidly with a slow accuracy improvement for some models. Specifically, the EM algorithm had a slow improvement on the MSE for models 1, 3, and 4 as the sample size increased. So the EM failed to provide the maximum likelihood estimates in some experiments for models 1, 3, and 4 since the asymptotic variance of the maximum likelihood estimate depends on the reciprocal of the sample size. The $PEMT$ algorithm performed better than the EM with models 1 and 4 when the sample size is large enough. For model 2, the EM outperformed the $PEMT$, and both algorithms had a low MAE and a low MSE. The reason may lie in that compared to models 1, 3, and 4, more observed variables in model 2 can provide more information from data for the parameter estimation. If the sample size is equal to $100k$, the MAE and the MSE of the $PEMT$ algorithm were 0.62655×10^{-2} and 0.01071×10^{-2} respectively. The execution speed was much faster with the $PEMT$ 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.

We designed an experiment on learning a latent tree structure with 1000 observed variables. A schematic diagram is shown in Figure 6 for the latent tree structure. We considered three cases of observed variables for this structure. The first case is that observed variables are pure discrete. The second one is that the first half of observed variables $\{v_1, \cdots, v_{500}\}$ are continuous and the others are discrete. The last one is that observed variables are pure continuous.

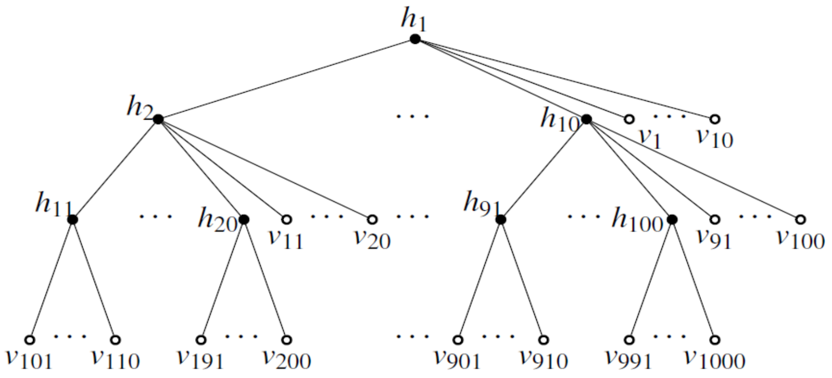
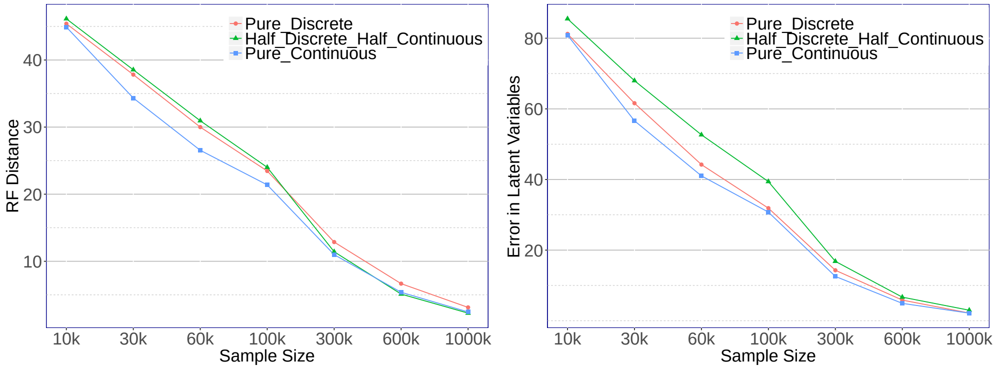


Figure 6: A schematic diagram for latent tree with pure discrete observed variables.



(a) RF Distance

(b) Error in Latent Variables

Figure 7: The consistency of the *SLLT* algorithm.

The performance of our method was assessed using the Robinson Foulds (RF) distance (Robinson and Foulds, 1981) and the error in the number of latent variables. Figure 7 illustrates the consistency performance of the *SLLT* algorithm for this high-dimensional structural recovery. In three cases, both the RF distance and the error in latent variables decreased as the sample size increased. So our method can still work well for learning high-dimensional mixed latent tree models when the sample size is large enough.

5.2 Real Data Analysis

In this part, we applied the *SLLT* algorithm to a Forest Cover Type dataset, which is available from the University of California, Irvine (UCI) machine learning data set repository. This dataset with 581012 samples includes ten continuous variables, forty soil types, and seven forest cover types in the Roosevelt National Forest of northern Colorado. Since the cover type *Lodgepole pine* accounts for more than 48% of the total samples, we studied the latent hierarchical structure related to the type *Lodgepole pine*.

Table 2: Observed variables and their Ids

Id	Variable	Id	Variable
1	Elevation	30	Haplocryolls
2	Aspect	31	Haplustalfs
3	Slope	32	Haplustolls
4	Vertical distance to nearest surface water features	33	Histic Cryaquolls
5	Horizontal distance to nearest surface water features	34	Hiwan family
6	Horizontal distance to nearest roadway	35	Legault family
7	Horizontal distance to nearest wildfire ignition points	36	Leighcan family
8	Hillshade index at 9am	37	Lithic Cryorthents
9	Hillshade index at Noon	38	Matcher family
10	Hillshade index at 3pm	39	Moran family
11	Aquic Argiudolls	40	Pachic Argiustolls
12	Argicryolls	41	Pachic Haplustolls
13	Argiustolls	42	Ratake family
14	Barrett family	43	Rock land
15	Bross family	44	Rock outcrop
16	Bullwark family	45	Rogert family
17	Catamount family	46	Rubble land
18	Cathedral family	47	Scout family
19	Cerro family	48	Supervisor family
20	Cryaquepts	49	Tolby family
21	Cryaquolls	50	Troutville family
22	Cryofluvents	51	Typic Argiustolls
23	Cryohemists	52	Typic Cryaquepts
24	Cryorthents	53	Typic Cryorthents
25	Cypher family	54	Typic Haplocryolls
26	Dystrocryepts	55	Typic Haplustolls
27	Eutrocryepts	56	Water
28	Frisco family	57	Wetmore family
29	Gateview family	58	Cover Type - Lodgepole pine

The original soil types are tagged using the US Forest Service Ecological Landtype Units (ELUs)[∗], and each ELU consists of one or more basic soil components[†]. We replaced the original types by those basic soil components, and considered the total 58 observed variables shown in Table 2 after preprocessing. Those observed variables are the elevation, the aspect, the slope, the distances (4), the hillshade indices (3), the soil components (47), and the forest cover type. The first 10 variables are continuous. The following 47 soil components are binary. The forest cover type is also binary depending on whether the *Lodgepole pine* exists or not.

Figure 8 presents the learned mixed latent tree structure for this Forest Cover Type dataset. The tree consists of 58 observed variables and nine latent variables. The diameter of the tree is seven. The node set $\{1, 4, 5, 6, 7, 9, 13, 16, 17, 18, 20, 21, 23, 25, 26, 27, 30, 31, 33, 34, 35, 36, 37, 39, 41, 42, 43, 44, 47, 49, 50, 52, 53, 55, 56, 58\}$ is the largest sibling group in the tree. From the additivity of the information distance, the cover type on *Lodgepole*

[∗] For more details about ELU, please visit the following link: `https://archive.ics.uci.edu/ml/datasets/Covertype`

[†] For more details about basic soil components, please visit the following link: `https://casoilresource.lawr.ucdavis.edu/soil_web/ssurgo.php?action=list_mapunits&areasymbol=co645`

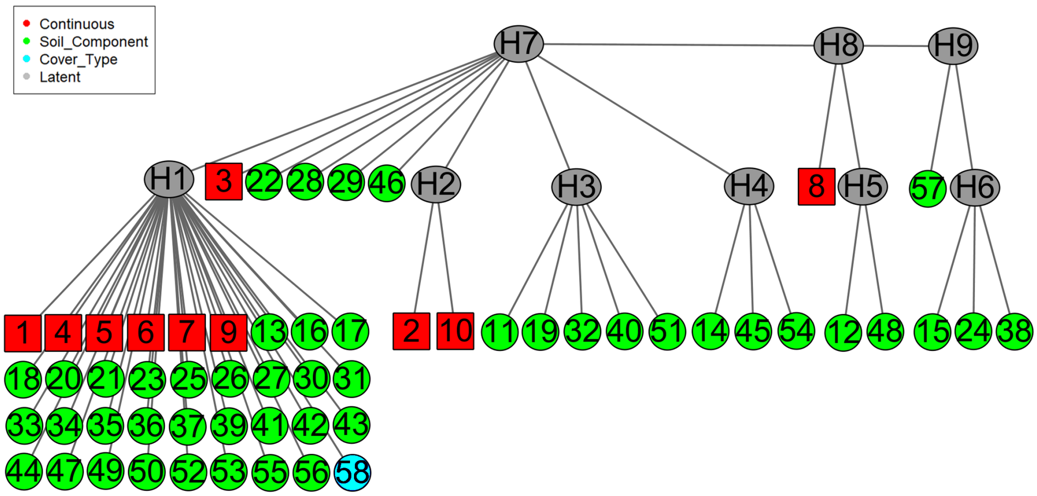


Figure 8: Mixed latent tree structure for the Forest Cover Type dataset.

Pine is close to six continuous variables and 29 soil types. Latent node H1 synthesizes all the information close to *Lodgepole pine*. Further statistical inferences on *Lodgepole pine* can be done on this largest sibling group. The Aspect (node 2) and the Hillshade index at 3pm (node 10) form a sibling pair with a latent node H2, which may be related to the solar incident angle. Soil components {11, 19, 32, 40, 51} form another sibling group. These five components belong to a common soil order Mollisols. Similarly, soil components {12,48} is from a soil suborder Cryolls. Observed variables {14,45,54} are basic components in two Ecological Landtype Units ELU3502 and ELU6731, and observed variables {15,24,38} are components in the ELU8707. So the latent nodes for soil components may reveal the potential background information on the soil taxonomy, and are related to the US Forest Ecological Landtype Units.

6. Discussion and Conclusion

To provide moderate theoretical proofs on mixed latent tree learning, we mainly consider that discrete observed variables are binary in this paper. But our algorithms also work for observed variables with more than two categories. For structural learning, we have mentioned in Subsection 3.1 that the information distance does not rely on the specific distributions of observed variables. So the structural learning algorithm can also handle general categorical observed variables. For parameter estimation, our strategy is to view categorical observed variables as binary variables. Assume that a categorical observed variable X has categories in $\{C_1, \cdots, C_K\}$. For any fixed category C_k, variable X can be viewed as a binary variable $X_{(k)}$ with two states. One is that variable X takes the category C_k and the other is that X takes a category in the remaining set $\{C_1, \cdots, C_{k-1}, C_{k+1}, \cdots, C_K\}$. Furthermore, for the fixed category C_k, we replace the original variable X by the binary variable $X_{(k)}$. This replacement can neither change the original tree structure, nor impact the conditional independence relations in the graphical tree model. So with this replacement, the Assumption

(A2) also holds that each latent variable has three neighbors at least. Similarly, we can also convert other categorical observed variables into binary variables. Thus the proposed $PEMT$ algorithm can estimate the parameter on the categorical variable X for the category C_k. The parameters on variable X for other categories can be computed in a similar way.

In summary, we propose an information distance for the mixed latent tree model and prove that the distance is additive along paths. Furthermore, we suggest a consistent bottom-up algorithm and give a finite sample bound guarantee for the exact structural recovery. For estimating model parameters, we study the moment estimator using matrix decomposition and prove that this estimator is asymptotically normal. The simulations support that our algorithms perform well when the sample size is large. In the real data application, we show that our structural learning algorithm can detect the latent hierarchical structure of observed variables.

Acknowledgments

The authors would like to thank the Action Editor and the reviewers for their constructive suggestions and comments that improved the quality of this paper. This work is supported by the National Key Research and Development Program of China (SQ2020YFA070264), the National Natural Science Foundation of China (11690012, 11631003, 11726629, 11701079), the Fundamental Research Funds for the Central Universities (2412019FZ030).

Appendix A. Proof of Theorem 1

Here we give a proof of Theorem 1.

Proof Consider a path $[v_0 = u, v_1 = w, v_2 = v]$ on the latent tree T between two nodes $u, v \in \mathbf{W}$. By the conditional independence on T and the double expectation formula, we have that

$$\mathrm{Cov}(X_u, X_v) = \mathrm{Var}(X_w)\big(\mathrm{E}(X_u|X_w = 1) - \mathrm{E}(X_u|X_w = 0)\big)\big(\mathrm{E}(X_v|X_w = 1) - \mathrm{E}(X_v|X_w = 0)\big).$$

Similarly, we have

$$\begin{aligned}\mathrm{Cov}(X_u, X_w) &= \mathrm{Var}(X_w)\big(\mathrm{E}(X_u|X_w = 1) - \mathrm{E}(X_u|X_w = 0)\big),\\ \mathrm{Cov}(X_v, X_w) &= \mathrm{Var}(X_w)\big(\mathrm{E}(X_v|X_w = 1) - \mathrm{E}(X_v|X_w = 0)\big).\end{aligned}$$

Thus

$$\mathrm{Cov}(X_u, X_v) = \frac{\mathrm{Cov}(X_u, X_w)\mathrm{Cov}(X_v, X_w)}{\mathrm{Var}(X_w)},$$

and $d_{uv} = d_{uw} + d_{wv}$. ■

Appendix B. Proof of Theorem 2

A random variable X with zero expectation is sub-exponential if there is a positive number λ such that its moment function satisfies $E(e^{tX}) \leqslant e^{\frac{\lambda^2 t^2}{2}}$ for any $|t| \leqslant \frac{1}{\lambda}$. Denote a sub-exponential distribution by $sub\,E(\lambda)$. We further have the following proposition (Boucheron et al., 2013) for controlling the tail probability of the sample mean:

Proposition 1 *If independent random variables $Z_1, \cdots, Z_n$ with mean zero and $Z_i \sim sub\,E(\lambda)$ for all i, we have that*

$$P(\overline{Z} > t) \vee P(\overline{Z} < -t) \leqslant \exp\left\{-\frac{n}{2}\left(\frac{t^2}{\lambda^2} \wedge \frac{t}{\lambda}\right)\right\}$$

for any $t > 0$, where $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

Now we give the proof of Theorem 2.

Proof Let X, Y be two observed variables in T and denote $\varepsilon = \sqrt{\frac{c(t+\log 48)}{n}}$, then the inequality (2) is equivalent to

$$P\left(\left|S_{XY} - \mathrm{Cov}(X,Y)\right| > \varepsilon\right) \leqslant 48\exp\left\{-\frac{n\varepsilon^2}{c}\right\} \tag{11}$$

for any $\varepsilon > 0$. We prove this inequality in two cases. Case I: X and Y are the same variable. Case II: X and Y are not the same variable. Furthermore, we divide Case I into Case I.i, variable X is continuous, and Case I.ii, X is binary. Similarly, we divide Case II into Case II.i, both variables X and Y are continuous, and Case II.ii, X is continuous and Y is binary, and Case II.iii, both X and Y are binary.
Case I.i:

$$\begin{aligned}
&P(|S_{XX} - \mathrm{Var}(X)| > \varepsilon)\\
&= P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}X_i + \mathrm{E}X_i - \overline{X})^2 - \mathrm{E}(X_i - \mathrm{E}X_i)^2\right| > \varepsilon\right)\\
&\leqslant P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)^2 - \mathrm{E}(X_i - \mathrm{E}X_i)^2\right| > \frac{\varepsilon}{2}\right) + P\left(\left|(\overline{X} - \mathrm{E}X_i)\right| > \sqrt{\frac{\varepsilon}{2}}\right)\\
&=: P_1 + P_2.
\end{aligned} \tag{12}$$

Then we compute the upper bounds of P_1 and P_2 in the following.

Let variable U be a parent of X on the tree. Then we have $X|_{U=u} \sim N(\mu_{X|u}, \sigma^2_{X|u})$ for any $u = 0, 1$. Since $X_i = \sum_{u=0}^{1} X_i I(U_i = u)$ and $\mathrm{E}X_i = \sum_{u=0}^{1} P(U_i = u)\mu_{i|u}$, let

$$A_{i,u} := X_i I(U_i = u) - P(U_i = u)\mu_{i|u}$$

for $u = 0, 1$, hence $X_i - \mathrm{E}X_i = A_{i,0} + A_{i,1}$. For any $u = 0, 1$, we obtain that

$$\begin{array}{rcccc}
A_{i,u} & = & X_i I(U_i = u) - \mu_{X|u} I(U_i = u) & + & \mu_{X|u} I(U_i = u) - P(U_i = u)\mu_{X|u}\\
& =: & C_{i,u,1} & + & C_{i,u,2}.
\end{array}$$

Therefore

$$\begin{aligned}
(X_i - \mathrm{E}X_i)^2 &= \left(\sum_{u=0}^{1}\sum_{k=1}^{2} C_{i,u,k}\right)^2 \\
&= C_{i,0,1}^2 + C_{i,1,1}^2 + 2C_{i,0,1}C_{i,0,2} + 2C_{i,1,1}C_{i,1,2} + 0 + 2C_{i,0,1}C_{i,1,2} + \\
&\quad 2C_{i,1,1}C_{i,0,2} + \left(C_{i,0,2}^2 + C_{i,1,2}^2\right) + 2C_{i,0,2}C_{i,1,2}, \\
\mathrm{E}(X_i - \mathrm{E}X_i)^2 &= \mathrm{E}C_{i,0,1}^2 + \mathrm{E}C_{i,1,1}^2 + 0 + 0 + 0 + 0 + \\
&\quad 0 + \mathrm{E}\left(2C_{i,0,2}^2 + 2C_{i,1,2}^2\right) + \mathrm{E}\left(2C_{i,0,2}C_{i,1,2}\right).
\end{aligned}$$

Then

$$P_1 \leqslant P_{1.1} + P_{1.2} + P_{1.3} + P_{1.4} + P_{1.5} + P_{1.6} + P_{1.7} + P_{1.8}, \tag{13}$$
$$P_2 \leqslant P_{2.1} + P_{2.2} + P_{2.3} + P_{2.4}, \tag{14}$$

where

$$\begin{aligned}
P_{1.1} &:= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,0,1}^2 - \mathrm{E}C_{i,0,1}^2\right| > \frac{\varepsilon}{16}\right),\ P_{1.2} := P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,1,1}^2 - \mathrm{E}C_{i,1,1}^2\right| > \frac{\varepsilon}{16}\right), \\
P_{1.3} &:= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,0,1}C_{i,0,2}\right| > \frac{\varepsilon}{32}\right),\ P_{1.4} := P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,1,1}C_{i,1,2}\right| > \frac{\varepsilon}{32}\right), \\
P_{1.5} &:= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,0,1}C_{i,1,2}\right| > \frac{\varepsilon}{32}\right),\ P_{1.6} := P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,0,2}C_{i,1,1}\right| > \frac{\varepsilon}{32}\right), \\
P_{1.7} &:= P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(C_{i,0,2}^2 + C_{i,1,2}^2) - \mathrm{E}(C_{i,0,2}^2 + C_{i,1,2}^2)\right| > \frac{\varepsilon}{16}\right), \\
P_{1.8} &:= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,0,2}C_{i,1,2} - \mathrm{E}C_{i,0,2}C_{i,1,2}\right| > \frac{\varepsilon}{32}\right),
\end{aligned}$$

$$\begin{aligned}
P_{2.1} &:= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,0,1}\right| > \frac{1}{4}\sqrt{\frac{\varepsilon}{2}}\right),\ P_{2.2} := P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,1,1}\right| > \frac{1}{4}\sqrt{\frac{\varepsilon}{2}}\right), \\
P_{2.3} &:= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,0,2}\right| > \frac{1}{4}\sqrt{\frac{\varepsilon}{2}}\right),\ P_{2.4} := P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,1,2}\right| > \frac{1}{4}\sqrt{\frac{\varepsilon}{2}}\right).
\end{aligned}$$

We only show the computation of the upper bound of $P_{1.1}, P_{1.3}, P_{1.7}$, and the others can be obtained similarly.

Firstly, we compute the upper bound of $P_{1.3}$. Since

$$P_{1.3} \leqslant P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \mu_{X|0}}{\sigma_{X|0}} \cdot I(U_i = 0)\right| > \frac{\varepsilon}{32\mu_m\sigma_m}\right),$$

hence we denote the variable $\frac{X_i-\mu_{X|0}}{\sigma_{X|0}} \cdot I(U_i=0)$ as Z_i for any $i=1,\dots,n$. Then

$$\begin{cases} \mathrm{E}Z_i = 0 \\ \mathrm{E}e^{tZ_i} = (1-P(U_i=0))\cdot 1 + P(U_i=0)\cdot e^{\frac{t^2}{2}} \leqslant e^{\frac{t^2}{2}} \end{cases},$$

therefore $Z_1,\dots,Z_n \overset{i.i.d.}{\sim} sub\,\mathrm{E}(1)$. By Proposition 1, a constant C exists such that

$$P_{1.3} \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\mu_m^2\sigma_m^2}\right\}. \tag{15}$$

Similarly, we obtain that

$$\begin{aligned} P_{1.4}, P_{1.5}, P_{1.6} &\leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\mu_m^2\sigma_m^2}\right\}, \\ P_{2.1}, P_{2.2} &\leqslant 2\exp\left\{-\frac{n\varepsilon}{C\sigma_m^2}\right\} \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\sigma_m^2}\right\}. \end{aligned} \tag{16}$$

Secondly, we compute the bound of $P_{1.7}$. Let

$$Z_i := C_{i,0,2}^2 + C_{i,1,2}^2 = \sum_{u=0}^{1}(I(U_i=u)-P(U_i=u))^2\mu_{X|u}^2.$$

Then

$$Z_i = \begin{cases} P^2(U_i=1)\left(\mu_{X|0}^2+\mu_{X|1}^2\right), & \text{for } U_i=0 \\ P^2(U_i=0)\left(\mu_{X|0}^2+\mu_{X|1}^2\right), & \text{for } U_i=1 \end{cases},$$

and $Z_i \in [a_i,b_i]$, where $b_i - a_i = |P(U_i=1)-P(U_i=0)|\left(\mu_{X|0}^2+\mu_{X|1}^2\right) \leqslant 2\mu_m^2$. By Hoeffding inequality, a constant C exists such that

$$P_{1.7} = P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathrm{E}Z_i\right| > \frac{\varepsilon}{16}\right) \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\mu_m^4}\right\}. \tag{17}$$

Similarly, we obtain that

$$\begin{aligned} P_{1.8} &\leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\mu_m^4}\right\}, \\ P_{2.3}, P_{2.4} &\leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\mu_m^2}\right\}. \end{aligned} \tag{18}$$

Thirdly, we compute the bound of $P_{1.1}$. Since,

$$\begin{aligned} P_{1.1} &= P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu_{X|0}\right)^2 I(U_i=0) - P(U_i=0)\sigma_{X|0}^2\right| > \frac{\varepsilon}{16}\right) \\ &\leqslant P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{X_i-\mu_{X|0}}{\sigma_{X|0}}\right)^2 - 1\right] I(U_i=0)\right| > \frac{\varepsilon}{32\sigma_{X|0}^2}\right) + \\ &\quad P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(U_i=0)-P(U_i=0)\right| > \frac{\varepsilon}{32\sigma_{X|0}^2}\right), \end{aligned}$$

hence let $Z_i = \left[\left(\frac{X_i - \mu_{X|0}}{\sigma_{X|0}}\right)^2 - 1\right] I(U_i = 0)$. Then we have $\mathrm{E}Z_i = 0$ and

$$\begin{aligned}
\mathrm{E}e^{tZ_i} &= (1 - P(U_i = 0)) + P(U_i = 0) \cdot \mathrm{E}e^{tZ_i|U_i=0} \\
&= (1 - P(U_i = 0)) + P(U_i = 0) \cdot e^{-t} \cdot (1 - 2t)^{-\frac{1}{2}} \\
&\leqslant e^{\frac{1}{2} \cdot 3^2 t^2}, && (\forall |t| < \tfrac{1}{3})
\end{aligned}$$

where the last inequality follows from the inequality $\frac{9}{2}t^2 + t + \frac{1}{2}\log(1 - 2t) \geqslant 0$ for any $|t| < \frac{1}{3}$. Therefore $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} sub\,\mathrm{E}(3)$. By Proposition 1 and the Hoeffding inequality, a constant C exists such that

$$\begin{aligned}
P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{X_i - \mu_{X|0}}{\sigma_{X|0}}\right)^2 - 1\right] I(U_i = 0)\right| > \frac{\varepsilon}{32\sigma^2_{X|0}}\right) \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\sigma_m^4}\right\}, \\
P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(U_i = 0) - P(U_i = 0)\right| > \frac{\varepsilon}{32\sigma^2_{X|0}}\right) \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\sigma_m^4}\right\}.
\end{aligned}$$

Furthermore,

$$P_{1.1}, P_{1.2} \leqslant 4\exp\left\{-\frac{n\varepsilon^2}{C\sigma_m^4}\right\}. \tag{19}$$

By (12), (13), (14), (15), (16), (17), (18) and (19), we obtain that

$$P(|S_{XX} - \mathrm{Var}(X)| > \varepsilon) \leqslant 28\exp\left\{-\frac{n\varepsilon^2}{c}\right\} \leqslant 48\exp\left\{-\frac{n\varepsilon^2}{c}\right\},$$

where $c := C \cdot \max\{\sigma_m^4,\ \sigma_m^2\mu_m^2,\ \mu_m^4,\ \sigma_m^2,\ \mu_m^2, 1\}$.

Case I.ii: by the Hoeffding inequality, we obtain that

$$\begin{aligned}
&P\left(|S_{XX} - \mathrm{Var}(X)| > \varepsilon\right) \\
&= P\left(\left|\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) - \left(\mathrm{E}X_i^2 - (\mathrm{E}X_i)^2\right)\right| > \varepsilon\right) \\
&\leqslant P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}X_i\right| > \frac{\varepsilon}{2}\right) + P\left(\left|\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 - (\mathrm{E}X_i)^2\right| > \frac{\varepsilon}{2}\right) \\
&\leqslant P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}X_i\right| > \frac{\varepsilon}{2}\right) + P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}X_i\right| > \frac{\varepsilon}{4}\right) \\
&\leqslant 4\exp\left\{-\frac{n\varepsilon^2}{8}\right\} \leqslant 48\exp\left\{-\frac{n\varepsilon^2}{c}\right\}.
\end{aligned} \tag{20}$$

Case II.i: we divide it into case II.i.a, X, Y are not sibling pair, and case II.i.b, X, Y are sibling pair.
Case II.i.a:

$$\begin{aligned}
&P(|S_{XY} - \mathrm{Cov}(X,Y)| > \varepsilon) \\
\leqslant\, & P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}X)(Y_i - \mathrm{E}Y) - \mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)\right| > \frac{\varepsilon}{2}\right) + \\
& P\left(\left|(\overline{X} - \mathrm{E}X)(\overline{Y} - \mathrm{E}Y)\right| > \frac{\varepsilon}{2}\right) \\
=:\, & P_1 + P_2.
\end{aligned} \tag{21}$$

Let U and V be parent of X and Y respectively. Then $X|_{U=u} \sim N(\mu_{X|u}, \sigma^2_{X|u})$, $Y|_{V=v} \sim N(\mu_{Y|v}, \sigma^2_{Y|v})$, where $u, v \in \{0,1\}$. Since $X_i = \sum_{u=0}^{1} X_i I(U_i = u)$, $\mathrm{E}X_i = \sum_{u=0}^{1} P(U_i = u)\mu_{X|u}$, and the similar to Y_i and $\mathrm{E}Y_i$, hence let

$$\begin{aligned}
A_{i,u} &= X_i I(U_i = u) - P(U_i = u)\mu_{X|u}, \\
B_{i,v} &= Y_i I(V_i = v) - P(V_i = v)\mu_{Y|v}.
\end{aligned}$$

Furthermore, let

$$\begin{array}{rccccc}
A_{i,u} & = & X_i I(U_i = u) - \mu_{X|u} I(U_i = u) & + & \mu_{X|u} I(U_i = u) - P(U_i = u)\mu_{X|u} \\
 & =: & C_{i,u,1} & + & C_{i,u,2}, \\
B_{i,v} & = & Y_i I(V_i = v) - \mu_{Y|v} I(V_i = v) & + & \mu_{Y|v} I(V_i = v) - P(V_i = v)\mu_{Y|v} \\
 & =: & D_{i,v,1} & + & D_{i,v,2}.
\end{array}$$

Then we have

$$\begin{aligned}
P_1 &\leqslant P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{1}\sum_{v=0}^{1}\sum_{k=1}^{2}\sum_{l=1}^{2}\left(C_{i,u,k}D_{i,v,l} - \mathrm{E}C_{i,u,k}D_{i,v,l}\right)\right| > \frac{\varepsilon}{2}\right) \\
&\leqslant \sum_{u=0}^{1}\sum_{v=0}^{1}\sum_{k=1}^{2}\sum_{l=1}^{2} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,u,k}D_{i,v,l} - \mathrm{E}C_{i,u,k}D_{i,v,l}\right| > \frac{\varepsilon}{32}\right), \qquad (22) \\
P_2 &\leqslant P\left(\left|\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{1}\sum_{k=1}^{2} C_{i,u,k}\right)\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{v=0}^{1}\sum_{l=1}^{2} D_{j,v,l}\right)\right| > \frac{\varepsilon}{2}\right) \\
&\leqslant \sum_{u=0}^{1}\sum_{k=1}^{2} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,u,k}\right| > \frac{1}{4}\sqrt{\frac{\varepsilon}{2}}\right) + \sum_{v=0}^{1}\sum_{l=1}^{2} P\left(\left|\frac{1}{n}\sum_{j=1}^{n} D_{j,v,l}\right| > \frac{1}{4}\sqrt{\frac{\varepsilon}{2}}\right). \qquad (23)
\end{aligned}$$

Thus, we only need to find the upper bounds of these equations:

$$\begin{aligned}
P_{1,u,v,k,l} &:= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,u,k}D_{i,v,l} - \mathrm{E}C_{i,u,k}D_{i,v,l}\right| > \frac{\varepsilon}{32}\right), \quad \forall u, v \in \{0,1\}, \forall k, l \in \{1,2\}, \\
P_{2,u,k} &:= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} C_{i,u,k}\right| > \frac{1}{4}\sqrt{\frac{\varepsilon}{2}}\right), \quad \forall u \in \{0,1\}, \forall k \in \{1,2\}.
\end{aligned}$$

From the arguments in case I.i, $\forall u, v \in \{0, 1\}$, we have

$$\begin{aligned} & P_{1,u,v,1,1} \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\sigma_m^4}\right\},\ P_{1,u,v,1,2}, P_{1,u,v,2,1} \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\mu_m^2\sigma_m^2}\right\}, \\ & P_{1,u,v,2,2} \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\mu_m^4}\right\},\ P_{2,u,1} \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\sigma_m^2}\right\},\ P_{2,u,2} \leqslant 2\exp\left\{-\frac{n\varepsilon^2}{C\mu_m^2}\right\}. \end{aligned} \tag{24}$$

By (21), (22), (23) and (24), we have

$$P(|S_{XY} - \mathrm{Cov}(X,Y)| > \varepsilon) \leqslant 48\exp\left\{-\frac{n\varepsilon^2}{c}\right\}. \tag{25}$$

Case II.i.b: since X, Y are sibling pair, hence let U be the parent of X and Y. Replace V_i in case II.i.b with U_i, then we can obtain the desired result (11).
Case II.ii: Let U be the parent of X. Then $X|_{U=u} \sim N(\mu_{X|u}, \sigma^2_{X|u})$, $Y|_{U=u} \sim b(1, p_{Y|u=u})$, where $u \in \{0, 1\}$. From similar arguments, we have

$$P(|S_{XY} - \mathrm{Cov}(X,Y)| > \varepsilon) \leqslant 48\exp\left\{-\frac{n\varepsilon^2}{c}\right\}. \tag{26}$$

Case II.iii: Let U be the parent of X, then $X|_{U=u} \sim b(1, p_{X|u})$, $Y|_{U=u} \sim b(1, p_{Y|U=u})$, where $u = 0, 1$. From similar arguments, we have

$$P(|S_{XY} - \mathrm{Cov}(X,Y)| > \varepsilon) \leqslant 24\exp\left\{-\frac{n\varepsilon^2}{C}\right\} \leqslant 48\exp\left\{-\frac{n\varepsilon^2}{c}\right\}. \tag{27}$$

In summary, for any $X, Y \in \mathbf{V}$ and $t > 0$, we have

$$P\left(\left|S_{XY} - \mathrm{Cov}(X,Y)\right| > \sqrt{\frac{c(t + \log 48)}{n}}\right) \leqslant e^{-t}, \tag{28}$$

where $c = C \cdot \max\{\sigma_m^4, \sigma_m^2\mu_m^2, \mu_m^4, \sigma_m^2, \mu_m^2, 1\}$. ■

Appendix C. Proof of Theorem 3

In this section, we give the proof of Theorem 3. For an threshold $\epsilon \leqslant \min\{\frac{1}{2}\phi_{min}, 1\}$, we only need to show that the probability of the event

$$\left\{\left|\left(\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}\right) - (\Phi_{uvw} - \Phi_{uvz})\right| < \epsilon, \forall u, v, w, z \in \mathbf{V}\right\}$$

is sufficiently large if the sample size is sufficiently large. Since

$$\begin{aligned} & \left|\left(\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}\right) - (\Phi_{uvw} - \Phi_{uvz})\right| \\ \leqslant & \left|\hat{d}_{uw} - d_{uw}\right| + \left|\hat{d}_{vw} - d_{vw}\right| + \left|\hat{d}_{uz} - d_{uz}\right| + \left|\hat{d}_{vz} - d_{vz}\right| \end{aligned}$$

for any $u, v, w, z \in \mathbf{V}$, we have

$$\begin{aligned} & P\left(\left|\left(\hat{\Phi}_{uvw} - \hat{\Phi}_{uvz}\right) - (\Phi_{uvw} - \Phi_{uvz})\right| < \epsilon, \forall u, v, w, z \in \mathbf{V}\right) \\ \geqslant\, & P\left(\left|\hat{d}_{uv} - d_{uv}\right| < \frac{1}{4}\epsilon, \forall u, v \in \mathbf{V}\right). \end{aligned}$$

Thus we only need to show that the probability of the event $\left\{\left|\hat{d}_{uv} - d_{uv}\right| < \frac{1}{4}\epsilon, \forall u, v \in \mathbf{V}\right\}$ is sufficiently large if the sample size is sufficiently large.

In the following, we show how to make $\left|\hat{d}_{uv} - d_{uv}\right| < \frac{1}{4}\epsilon$. For any $u, v \in \mathbf{V}$, we consider the information distance d_{uv}. Since $d_{uv} = -\log|\mathrm{Cov}(X_u, X_v)| + \frac{1}{2}\log(\mathrm{Var}(X_u)) + \frac{1}{2}\log(\mathrm{Var}(X_v))$, hence

$$\begin{aligned} \left|\hat{d}_{uv} - d_{uv}\right| < & \left|\log|S_{uv}| - \log|\mathrm{Cov}(X_u, X_v)|\right| + \frac{1}{2}\left|\log(S_{uu}) - \log(\mathrm{Var}(X_u))\right| + \\ & \frac{1}{2}\left|\log(S_{vv}) - \log(\mathrm{Var}(X_v))\right|. \end{aligned}$$

We denote c_{min} as $\min_{u,v\in\mathbf{V}}\{|\mathrm{Cov}(X_u, X_v)|\}$ (this allows $u = v$). If $\Delta > 0$ exists such that $\Delta < \frac{1}{2}c_{min}$ and for any $u, v \in \mathbf{V}$, $\left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| \leqslant \Delta$, we obtain

$$\begin{aligned} |S_{uv}| & \geqslant |\mathrm{Cov}(X_u, X_v)| - \left||S_{uv}| - |\mathrm{Cov}(X_u, X_v)|\right| \\ & \geqslant |\mathrm{Cov}(X_u, X_v)| - \left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| \geqslant \frac{1}{2}c_{min}. \end{aligned}$$

Since $|\mathrm{Cov}(X_u, X_v)|, |S_{uv}| > \frac{1}{2}c_{min}$, then $\left|\log|S_{uv}| - \log|\mathrm{Cov}(X_u, X_v)|\right| < \frac{2}{c_{min}}\Delta$. Furthermore, we have $\left|\hat{d}_{uv} - d_{uv}\right| < \frac{4\Delta}{c_{min}}$. Since $\epsilon \leqslant 1$, we obtain that $\Delta < \frac{c_{min}\epsilon}{16}$ implies $\Delta < \frac{1}{2}c_{min}$. Thus if an appropriate Δ exists such that $\Delta < \frac{c_{min}\epsilon}{16}$ and $\left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| \leqslant \Delta$ for any $u, v \in \mathbf{V}$, then $\left|\hat{d}_{uv} - d_{uv}\right| < \frac{1}{4}\epsilon$.

Next, we show how to select Δ such that $P\left(\bigcap_{u,v}\left\{\left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| \leqslant \Delta\right\}\right)$ is sufficiently large and $\Delta < \frac{c_{min}\epsilon}{16}$. If we show these successfully, the proof of Theorem 3 is completed. According to Theorem 2, for any $u, v \in \mathbf{V}$ and any $t > 0$, we have

$$P\left(\left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| > \sqrt{\frac{c(t + \log 48)}{n}}\right) \leqslant e^{-t}.$$

Thus, for any $t > 0$, we have

$$\begin{aligned} & P\left(\bigcap_{u,v}\left\{\left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| \leqslant \sqrt{\frac{c(t + \log 48)}{n}}\right\}\right) \\ = & 1 - P\left(\bigcup_{u,v}\left\{\left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| > \sqrt{\frac{c(t + \log 48)}{n}}\right\}\right) \\ \geqslant & 1 - \sum_{u,v} P\left(\left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| > \sqrt{\frac{c(t + \log 48)}{n}}\right) \\ \geqslant & 1 - m^2 \cdot e^{-t} =: 1 - \eta, \end{aligned}$$

where $\eta \in (0,1)$. Therefore, $\eta = m^2 \cdot e^{-t}$ and let t be $-\log \frac{\eta}{m^2}$. Hence, we obtain

$$P\left(\min_{u,v}\left\{\left|S_{uv} - \mathrm{Cov}(X_u, X_v)\right| \leqslant \sqrt{\frac{c(\log(48m^2) - \log\eta)}{n}}\right\}\right) \geqslant 1 - \eta.$$

Thus we select $\Delta = \sqrt{\frac{c(\log(48m^2)-\log\eta)}{n}}$. For any η, if the sample size n is large enough such that $\Delta < \frac{c_{min}\epsilon}{16}$, the algorithm returns the true latent tree structure with a probability of at least $1-\eta$. Furthermore, if $\frac{16\Delta}{c_{min}} < \min\{\frac{1}{2}\phi_{min}, 1\}$, there exists an appropriate threshold $\epsilon \leqslant \min\{\frac{1}{2}\phi_{min}, 1\}$. So we get the conclusion.

Appendix D. Proof of Theorem 4

Take the case that $u, v, w \in \mathbf{V}_c$ in Figure 2 as an example, and denote θ_{uvw} as the parameter vector

$$(p_h, \mu_{u|X_h=0}, \mu_{u|X_h=1}, \mu^{(2)}_{u|X_h=0}, \mu^{(2)}_{u|X_h=1}, \mu_{v|X_h=0}, \mu_{v|X_h=1}, \mu^{(2)}_{v|X_h=0}, \mu^{(2)}_{v|X_h=1},$$
$$\mu_{w|X_h=0}, \mu_{w|X_h=1})^T,$$

and denote $f = (f_l)_{l=1}^{11}$ as a function of random vector $\mathbf{X}_{uvw} := (X_u, X_v, X_w)$:

$$f_1(x_{uvw}) = x_w,\ f_2(x_{uvw}) = x_u,\ f_3(x_{uvw}) = x_u^2,\ f_4(x_{uvw}) = x_u x_v,\ f_5(x_{uvw}) = x_u x_v^2,$$
$$f_6(x_{uvw}) = x_u^2 x_v,\ f_7(x_{uvw}) = x_u^2 x_v^2,\ f_8(x_{uvw}) = x_u x_v x_w,\ f_9(x_{uvw}) = x_u x_v^2 x_w,$$
$$f_{10}(x_{uvw}) = x_u^2 x_v x_w,\ f_{11}(x_{uvw}) = x_u^2 x_v^2 x_w.$$

It is obvious that $\mathrm{E}\mathbf{X}_w = \mathrm{E}f_1(\mathbf{X}_{uvw}), \mathrm{E}\mathbf{X}_u = \mathrm{E}f_2(\mathbf{X}_{uvw}), \mathrm{E}\mathbf{X}_u^2 = \mathrm{E}f_3(\mathbf{X}_{uvw})$,

$$E_{uv} = \begin{pmatrix} \mathrm{E}f_4(\mathbf{X}_{uvw}) & \mathrm{E}f_5(\mathbf{X}_{uvw}) \\ \mathrm{E}f_6(\mathbf{X}_{uvw}) & \mathrm{E}f_7(\mathbf{X}_{uvw}) \end{pmatrix} \text{ and } E_{uvw} = \begin{pmatrix} \mathrm{E}f_8(\mathbf{X}_{uvw}) & \mathrm{E}f_9(\mathbf{X}_{uvw}) \\ \mathrm{E}f_{10}(\mathbf{X}_{uvw}) & \mathrm{E}f_{11}(\mathbf{X}_{uvw}) \end{pmatrix}.$$

Furthermore, we denote $e = (e_l)_{l=1}^{11}$ as a function of the parameter vector θ_{uvw}:

$$\begin{aligned}
e_1(\theta_{uvw}) &= (1-p_h)\cdot\mu_{w|X_h=0} + p_h\cdot\mu_{w|X_h=1},\\
e_2(\theta_{uvw}) &= (1-p_h)\cdot\mu_{u|X_h=0} + p_h\cdot\mu_{u|X_h=1},\\
e_3(\theta_{uvw}) &= (1-p_h)\cdot\mu^{(2)}_{u|X_h=0} + p_h\cdot\mu^{(2)}_{u|X_h=1},\\
e_4(\theta_{uvw}) &= (1-p_h)\cdot\mu_{u|X_h=0}\cdot\mu_{v|X_h=0} + p_h\cdot\mu_{u|X_h=1}\cdot\mu_{v|X_h=1},\\
e_5(\theta_{uvw}) &= (1-p_h)\cdot\mu_{u|X_h=0}\cdot\mu^{(2)}_{v|X_h=0} + p_h\cdot\mu_{u|X_h=1}\cdot\mu^{(2)}_{v|X_h=1},\\
e_6(\theta_{uvw}) &= (1-p_h)\cdot\mu^{(2)}_{u|X_h=0}\cdot\mu_{v|X_h=0} + p_h\cdot\mu^{(2)}_{u|X_h=1}\cdot\mu_{v|X_h=1},\\
e_7(\theta_{uvw}) &= (1-p_h)\cdot\mu^{(2)}_{u|X_h=0}\cdot\mu^{(2)}_{v|X_h=0} + p_h\cdot\mu^{(2)}_{u|X_h=1}\cdot\mu^{(2)}_{v|X_h=1},\\
e_8(\theta_{uvw}) &= (1-p_h)\cdot\mu_{u|X_h=0}\cdot\mu_{v|X_h=0}\cdot\mu_{w|X_h=0} + p_h\cdot\mu_{u|X_h=1}\cdot\mu_{v|X_h=1}\cdot\mu_{w|X_h=1},\\
e_9(\theta_{uvw}) &= (1-p_h)\cdot\mu_{u|X_h=0}\cdot\mu^{(2)}_{v|X_h=0}\cdot\mu_{w|X_h=0} + p_h\cdot\mu_{u|X_h=1}\cdot\mu^{(2)}_{v|X_h=1}\cdot\mu_{w|X_h=1},\\
e_{10}(\theta_{uvw}) &= (1-p_h)\cdot\mu^{(2)}_{u|X_h=0}\cdot\mu_{v|X_h=0}\cdot\mu_{w|X_h=0} + p_h\cdot\mu^{(2)}_{u|X_h=1}\cdot\mu_{v|X_h=1}\cdot\mu_{w|X_h=1},\\
e_{11}(\theta_{uvw}) &= (1-p_h)\cdot\mu^{(2)}_{u|X_h=0}\cdot\mu^{(2)}_{v|X_h=0}\cdot\mu_{w|X_h=0} + p_h\cdot\mu^{(2)}_{u|X_h=1}\cdot\mu^{(2)}_{v|X_h=1}\cdot\mu_{w|X_h=1}.
\end{aligned}$$

Then our eigen-decomposition methods are equivalent to solve the equations

$$e(\theta_{uvw}) = \mathrm{E}_{\theta_{uvw}} f(\mathbf{X}_{uvw}). \tag{29}$$

By using the sample moment $\frac{1}{n}\sum_{l=1}^{n} f(\mathbf{X}_{uvw}^{(l)})$ to replace the right side of the equations (29), we obtain the moment estimation equations

$$e(\theta_{uvw}) = \frac{1}{n}\sum_{l=1}^{n} f(\mathbf{X}_{uvw}^{(l)}). \tag{30}$$

The solution of the equations (30) is the moment estimation $\hat{\theta}_{uvw}$ for θ_{uvw}.

Before proving Theorem 4, we need the following proposition which states the asymptotic normality of the moment estimation (van der Vaart, 2000).

Proposition 2 *Suppose that $e(\theta) = P_\theta f$ is one-to-one on an open set $\Theta \subset \mathbb{R}^k$ and continuously differentiable at θ_0 with nonsingular derivative e'_{θ_0}. Moreover, assume that $P_{\theta_0}\|f\|^2 < \infty$. Then moment estimators $\hat{\theta}_n$ exist with probability tending to one and satisfy*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{\theta_0}{\rightsquigarrow} N\left(0, e'^{-1}_{\theta_0} P_{\theta_0} f f^T (e'^{-1}_{\theta_0})^T\right).$$

Now we give the proof of Theorem 4.

Proof We prove this theorem on the two cases: case I, the asymptotic normality of estimators for observed variables, and case II, the asymptotic normality of estimators for latent variables.

Case I: as discussed above, we know that $\Theta_{uvw} = \{\theta_{uvw} : \mu_{w|X_h=0} > \mu_{w|X_h=1}\}$ is an open set.

For any $u, v \in \mathbf{V}, w \in \mathbf{V}_c$, since $\mu_{w|X_h=0} > \mu_{w|X_h=1}$, we obtain the unique parameters $\mu_{w|X_h=0}$, $\mu_{w|X_h=1}$ by the first step of eigen-decomposition. Then, we obtain the unique parameters $p_h, \Gamma_{u|h}$ by the second and third step of eigen-decomposition. Furthermore, according to the equation

$$E_{uv} = \Gamma_{u|h} \begin{pmatrix} 1-p_h & 0 \\ 0 & p_h \end{pmatrix} \Gamma_{v|h}^T,$$

and the matrix E_{uv} and the parameters $p_h, \Gamma_{u|h}$ are all unique, we obtain the unique parameters $\Gamma_{v|h}$. Thus $e(\theta_{uvw})$ is an one-to-one on Θ_{uvw} for any $u, v \in \mathbf{V}, w \in \mathbf{V}_c$. Similarly, $e(\theta_{uvw})$ is an one-to-one on Θ_{uvw} for $u, v \in \mathbf{V}, w \in \mathbf{V}_d$.

And $e(\theta_{uvw})$ is continuously differentiable at θ_0. According to the proposition 2, we only need to prove that the Jacobi matrix of $e(\theta_{uvw})$ at θ_0 is nonsingular.

For any $u, v, w \in \mathbf{V}_c$, we have

$$\begin{aligned}
&\det\left(\frac{\partial e(\theta_{uvw})}{\partial \theta_{uvw}^T}\right) \\
&= (1-p_1)^5 p_1^5 (\mu_{u|X_h=0}\mu^{(2)}_{u|X_h=1} - \mu_{u|X_h=1}\mu^{(2)}_{u|X_h=0})^3 (\mu_{v|X_h=0}\mu^{(2)}_{v|X_h=1} - \mu_{v|X_h=1}\mu^{(2)}_{v|X_h=0})^2 \\
&\quad \cdot (\mu_{w|X_h=1} - \mu_{w|X_h=0})^3 \\
&= (1-p_h)^2 p_h^2 \det{}^3(\Gamma_{u|h}) \det{}^2(\Gamma_{v|h}) \mathrm{Cov}^3(X_w, X_h).
\end{aligned}$$

Thus, for any $w \in \mathbf{V}$,

$$\det\left(\frac{\partial e(\theta_{uvw})}{\partial \theta_{uvw}^T}\right) = \begin{cases} (1-p_h)^2 p_h^2 \det^3(\Gamma_{u|h}) \det^2(\Gamma_{v|h}) \mathrm{Cov}^3(X_w, X_h), & u, v \in \mathbf{V}_c; \\ (1-p_h) p_h \det^2(\Gamma_{u|h}) \det^2(\Gamma_{v|h}) \mathrm{Cov}^3(X_w, X_h), & u \in \mathbf{V}_c, v \in \mathbf{V}_d; \\ (1-p_h) p_h \det^2(\Gamma_{u|h}) \det^2(\Gamma_{v|h}) \mathrm{Cov}^3(X_w, X_h), & u \in \mathbf{V}_d, v \in \mathbf{V}_c; \\ (1-p_h) p_h \det^2(\Gamma_{u|h}) \det^2(\Gamma_{v|h}) \mathrm{Cov}^2(X_w, X_h), & u, v \in \mathbf{V}_d. \end{cases}$$

Since Assumption (A1) holds and continuous variables have mean zero, we have

$$\det\left(\frac{\partial e(\theta_{uvw})}{\partial \theta_{uvw}^T}\right) \neq 0, \quad \forall u, v, w \in \mathbf{V}.$$

In the following, we prove $\mathrm{E}\|f\|^2 < +\infty$. We only need to find an upper bound $M < +\infty$ such that $\mathrm{E}f_l^2(\mathbf{X}_{uvw}) \leqslant M$ for any l.

For any $u \in \mathbf{V}_c$ and $h \in \mathbf{H}$, the conditional moment $\mathrm{E}(X_u^l | X_h = x_h)$ with $x_h = 0, 1$ and $l = 2, 4$ satisfies

$$\mathrm{E}(X_u^l | X_h = x_h) = \sum_{x_{pa(u)}=0}^{1} P\left(X_{pa(u)} = x_{pa(u)} | X_h = x_h\right) \mathrm{E}\left(X_u^l | X_{pa(u)} = x_{pa(u)}\right).$$

Since $X_u | X_{pa(u)} = x_{pa(u)} \sim N(\mu_{u|x_{pa(i)}}, \sigma^2_{u|x_{pa(u)}})$ for $x_{pa(u)} = 0, 1$, we have

$$\begin{aligned} \mathrm{E}(X_u^2 | X_h = x_h) &= \sum_{x_{pa(u)}=0}^{1} P\left(X_{pa(u)} = x_{pa(u)} | X_h = x_h\right) \left(\sigma^2_{u|x_{pa(u)}} + \mu^2_{u|x_{pa(u)}}\right) \\ &\leqslant \sigma_m^2 + \mu_m^2, \\ \mathrm{E}(X_u^4 | X_h = x_h) &= \sum_{x_{pa(u)}=0}^{1} P\left(X_{pa(u)} = x_{pa(u)} | X_h = x_h\right) \cdot \Big(3\sigma^4_{u|x_{pa(u)}} + \\ &\quad 6\sigma^2_{u|x_{pa(u)}} \mu^2_{u|x_{pa(u)}} + \mu^4_{u|x_{pa(u)}}\Big) \\ &\leqslant 3\sigma_m^4 + 6\sigma_m^2 \mu_m^2 + \mu_m^4, \end{aligned}$$

where $\mu_m = \max\{|\mu_{u|X_h=0}|, |\mu_{u|X_h=1}|, \forall u \in \mathbf{V}_c\}$ and $\sigma_m^2 = \max\{\sigma^2_{u|X_h=0}, \sigma^2_{u|X_h=1}, \forall u \in \mathbf{V}_c\}$.

For any $u \in \mathbf{V}_d$ and $h \in \mathbf{H}$, the conditional moment $\mathrm{E}(X_u^l | X_h = x_h)$ with $x_h = 0, 1$ and $l = 2, 4$ satisfies

$$\mathrm{E}(X_u^l | X_h = x_h) = P(X_u = 1 | X_h = x_h) \leqslant 1.$$

Let $C = max\{1, \sigma_m^2 + \mu_m^2, 3\sigma_m^4 + 6\sigma_m^2\mu_m^2 + \mu_m^4\} < +\infty$. Thus, for any f_l, we have

$$\begin{aligned} \mathrm{E}f_l^2(\mathbf{X}) &= \sum_{x_h=0}^{1} P(X_h = x_h) \mathrm{E}(X_u^{l_u} | X_h = x_h) \mathrm{E}(X_v^{l_v} | X_h = x_h) \mathrm{E}(X_w^{l_w} | X_h = x_h) \\ &\leqslant C^3 < +\infty, \end{aligned}$$

where $l_u, l_v = 0, 2, 4$ and $l_w = 2, 4$. So, we have $\mathrm{E}\|f\|^2 < +\infty$.

From the proposition 2, we have

$$\sqrt{n}(\hat{\theta}_{uvw} - \theta_0) \xrightarrow{L} N(0, e'^{-1}(\theta_0)\mathrm{E}f(\mathbf{X}_{uvw})f^T(\mathbf{X}_{uvw})(e'^{-1}(\theta_0))^T).$$

Thus, $\hat{p}_h$, $\hat{\mu}_{u|X_h=0}, \hat{\mu}_{u|X_h=1}, \hat{\mu}^{(2)}_{u|X_h=0}, \hat{\mu}^{(2)}_{u|X_h=1}$ for $u \in \mathbf{V}_c$ (or $\hat{p}_{u|X_h=0}, \hat{p}_{u|X_h=1}$ for $u \in \mathbf{V}_d$) are all asymptotically normal.

Case II: firstly, we consider the estimating equations for the parameter of the latent variables.

As shown in Figure 3, for any observed variables $u, v_1, v_2, w \in \mathbf{V}$ and any latent variables $h_1, h_2 \in \mathbf{H}$, let $\theta_{uv_1v_2w}$ denote all the model parameters which appear in the decomposition method. As shown in Figure 3 (b), the parameter $\theta_{uv_1v_2w}$ consists of θ_{uv_1w} and θ_{uv_2w}. Since we estimate the parameters θ_{uv_1w} and θ_{uv_2w} separately when we estimate the parameter $\theta_{uv_1v_2w}$, we treat them all as free parameters. Thus, all the parameters in $\theta_{uv_1v_2w}$ are free.

Take the case where $u, v_1, v_2, w \in \mathbf{V}_c$ as an example,

$$\begin{aligned}\theta_{uv_1v_2w} &= (\theta^T_{uv_1w}, \theta^T_{uv_2w})^T\\ &= (p_{h_1}, \mu_{u|X_{h_1}=0}, \mu_{u|X_{h_1}=1}, \mu^{(2)}_{u|X_{h_1}=0}, \mu^{(2)}_{u|X_{h_1}=1}, \mu_{v_1|X_{h_1}=0}, \mu_{v_1|X_{h_1}=1}, \mu^{(2)}_{v_1|X_{h_1}=0},\\ &\quad \mu^{(2)}_{v_1|X_{h_1}=1}, \mu_{w|X_{h_1}=0}, \mu_{w|X_{h_1}=1}, p_{h_2}, \mu_{u|X_{h_2}=0}, \mu_{u|X_{h_2}=1}, \mu^{(2)}_{u|X_{h_2}=0}, \mu^{(2)}_{u|X_{h_2}=1},\\ &\quad \mu_{v_2|X_{h_2}=0}, \mu_{v_2|X_{h_2}=1}, \mu^{(2)}_{v_2|X_{h_2}=0}, \mu^{(2)}_{v_2|X_{h_2}=1}, \mu_{w|X_{h_2}=0}, \mu_{w|X_{h_2}=1})^T,\end{aligned}$$

where $\mu_{w|X_{h_1}=0} > \mu_{w|X_{h_1}=1}$ and $\mu_{w|X_{h_2}=0} > \mu_{w|X_{h_2}=1}$.

Furthermore, the estimating equations of the parameter $\theta_{uv_1v_2w}$ consist of the estimating equations of the parameters θ_{uv_1w} and θ_{uv_2w} (similar to equation (30)). Thus, we can obtain estimating equations of the parameter $\theta_{uv_1v_2w}$:

$$\begin{aligned}\frac{1}{n}\sum_{l=1}^{n} F(X^{(l)}_{uv_1v_2w}) &:= \begin{pmatrix} \frac{1}{n}\sum_{l=1}^{n} f\left(X^{(l)}_{uv_1w}\right) \\ \frac{1}{n}\sum_{l=1}^{n} f\left(X^{(l)}_{uv_2w}\right) \end{pmatrix}\\ &= \begin{pmatrix} \mathrm{E}f\left(X_{uv_1w}\right) \\ \mathrm{E}f\left(X_{uv_2w}\right) \end{pmatrix} = \begin{pmatrix} e(\theta_{uv_1w}) \\ e(\theta_{uv_2w}) \end{pmatrix} =: g(\theta_{uv_1v_2w}),\end{aligned}$$

where $f(\cdot), e(\cdot)$ are defined in the Subsection 4.2. The solution of the above equations is the moment estimation $\hat{\theta}_{uv_1v_2w}$ obtained from the $PEMT$ algorithm for $\theta_{uv_1v_2w}$. By the multiple center limit theorem, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{l=1}^{n} F(X^{(l)}_{uv_1v_2w}) - \mathrm{E}(F(X_{uv_1v_2w}))\right) \xrightarrow{L} N(0, \mathrm{Cov}(F(X_{uv_1v_2w}))).$$

As discussed above, we know that $\Theta_{uv_1v_2w} = \{\theta_{uv_1v_2w} : \mu_{w|X_{h_1}=0} > \mu_{w|X_{h_1}=1}, \mu_{w|X_{h_2}=0} > \mu_{w|X_{h_2}=1}\}$ is an open set.

For any $u, v_1, v_2, w \in \mathbf{V}$, similar to Case I, we have

$$\det\left(\frac{\partial g(\theta_{uv_1v_2w})}{\partial \theta_{uv_1v_2w}}\right) = \det\left(\frac{\partial e(\theta_{uv_1w})}{\partial \theta_{uv_1w}}\right) \cdot \det\left(\frac{\partial e(\theta_{uv_2w})}{\partial \theta_{uv_2w}}\right) \neq 0.$$

Similar to Case I, we have $\mathrm{E}\|f\|^2 < \infty$.

Thus, from proposition D.1, we have

$$\sqrt{n}(\hat{\theta}_{uv_1v_2w} - \theta_0) \xrightarrow{L} N(0, g'^{-1}(\theta_0)\mathrm{E}f(\mathbf{X}_{uv_1v_2w})f^T(\mathbf{X}_{uv_1v_2w})(g'^{-1}(\theta_0))^T).$$

Let $\theta_{u|h_1h_2}$ denote

$$(\mu_{u|X_{h_1}=0}, \mu_{u|X_{h_1}=1}, \mu^{(2)}_{u|X_{h_1}=0}, \mu^{(2)}_{u|X_{h_1}=1}, \mu_{u|X_{h_2}=0}, \mu_{u|X_{h_2}=1}, \mu^{(2)}_{u|X_{h_2}=0}, \mu^{(2)}_{u|X_{h_2}=1})^T$$

for $u \in \mathbf{V}_c$ and

$$(p_{u|X_{h_1}=0}, p_{u|X_{h_1}=1}, p_{u|X_{h_2}=0}, p_{u|X_{h_1}=1})^T$$

for $u \in \mathbf{V}_d$.

Then, we have

$$\sqrt{n}(\hat{\theta}_{u|h_1h_2} - \theta_{u|h_1h_2}) \xrightarrow{L} N(0, \Sigma),$$

where Σ is corresponding submatrix of the matrix $g'^{-1}(\theta_{uv_1v_2w}) \cdot \mathrm{E}f(\mathbf{X}_{uv_1v_2w})f^T(\mathbf{X}_{uv_1v_2w}) \cdot (g'^{-1}(\theta_{uv_1v_2w}))^T$.

In the following, we consider the asymptotic normality of the estimation of parameters $p_{h_1|X_{h_2}=0}$, $p_{h_1|X_{h_2}=1}$. Take the case that $u \in \mathbf{V}_c$ as an example. Since

$$\begin{aligned}\Gamma_{h_1|h_2} &= \Gamma^{-1}_{u|h_1} \cdot \Gamma_{u|h_2} \\ &= \frac{1}{\det(\Gamma_{u|h_1})} \begin{pmatrix} \mu^{(2)}_{u|X_{h_1}=1} & -\mu_{u|X_{h_1}=1} \\ -\mu^{(2)}_{u|X_{h_1}=0} & \mu_{u|X_{h_1}=0} \end{pmatrix} \begin{pmatrix} \mu_{u|X_{h_2}=0} & \mu_{u|X_{h_2}=1} \\ \mu^{(2)}_{u|X_{h_2}=0} & \mu^{(2)}_{u|X_{h_2}=1} \end{pmatrix},\end{aligned}$$

we have

$$\begin{pmatrix} p_{h_1|X_{h_2}=0} \\ p_{h_1|X_{h_2}=1} \end{pmatrix} = \begin{pmatrix} \frac{\mu_{u|X_{h_1}=0} \cdot \mu^{(2)}_{u|X_{h_2}=0} - \mu^{(2)}_{u|X_{h_1}=0} \cdot \mu_{u|X_{h_2}=0}}{\det(\Gamma_{u|h_1})} \\ \frac{\mu_{u|X_{h_1}=0} \cdot \mu^{(2)}_{u|X_{h_2}=1} - \mu^{(2)}_{u|X_{h_1}=0} \cdot \mu_{u|X_{h_2}=1}}{\det(\Gamma_{u|h_1})} \end{pmatrix} =: \varphi(\theta_{u|h_1h_2}).$$

Thus, $\varphi(\hat{\theta}_{u|h_1h_2})$ is the parameter estimations $(\hat{p}_{h_1|X_{h_2}=0}, \hat{p}_{h_1|X_{h_2}=1})^T$ obtained from sample *PEMT* algorithm. According to Assumption (A1) and $\mu_u = 0$, we have $\det(\Gamma_{u|h_1}) \neq 0$ and $\mu_{u|X_h=x} \neq 0$ for $h = h_1, h_2$ and $x = 0, 1$. Furthermore, the function $\varphi(\cdot)$ is Continuously differentiable. Thus, by the multiple delta theorem, we have

$$\sqrt{n}\left(\begin{pmatrix} \hat{p}_{h_1|X_{h_2}=0} \\ \hat{p}_{h_1|X_{h_2}=1} \end{pmatrix} - \begin{pmatrix} p_{h_1|X_{h_2}=0} \\ p_{h_1|X_{h_2}=1} \end{pmatrix}\right) \xrightarrow{L} N\left(0, \left(\frac{\partial\varphi(\theta_{u|h_1h_2})}{\partial\theta^T_{u|h_1h_2}}\right) \Sigma \left(\frac{\partial\varphi(\theta_{u|h_1h_2})}{\partial\theta^T_{u|h_1h_2}}\right)^T\right).$$

Then, $\hat{p}_{h_1|X_{h_2}=0}, \hat{p}_{h_1|X_{h_2}=1}$ are all asymptotically normal. ■

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

A. Anandkumar, R. Ge, D. Hsu, S. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities.* Oxford University Press, 2013.

J. T. Chang. Full reconstruction of markov models on evolutionary trees: identifiability and consistency. *Mathematical biosciences*, 137(1):51–73, 1996.

P. X. Chen, N. L. Zhang, K. M. Poon, and Z. R. Chen. Progressive em for latent tree models and hierarchical topic detection. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1498–1504, 2016.

P. X. Chen, N. L. Zhang, T. F. Liu, L. K. M. Poon, Z. R. Chen, and F. Khawar. Latent tree models for hierarchical topic detection. *Artificial Intelligence*, 250:105–124, 2017.

T. Chen, N. L. Zhang, K. M. Poon T. F. Liu, and Y. Wang. Model-based multidimensional clustering of categorical data. *Artificial Intelligence*, 176:2246–2269, 2012.

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

J. Fan, H. Liu, and Y. Ning. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B.*, 79:405–421, 2017.

L. A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231, 1974.

S. L. Lauritzen. *Graphical Models.* Oxford Clarendon Press, 1996.

P. F. Lazarsfeld and N. W. Henry. *Latent structure analysis.* Boston: Houghton Mifflin, 1968.

J. D. Lee and T. J. Hastie. Learning the structure of mixed graphical models. *Journal of Machine Learning Research*, 24:230–253, 2014.

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

D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

L. Song, A. P. Parikh, and E. P. Xing. Kernel embeddings of latent tree graphical models. *Advances in Neural Information Processing Systems*, 24:2708–2716, 2011.

A. W. van der Vaart. *Asymptotic Statistics.* Cambridge Univ. Press, 2000.

F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 596–603, 2013.

X. F. Wang, J. H. Guo, L. Z. Hao, and N. L. Zhang. Spectral methods for learning discrete latent tree models. *Statistics and Its Interface*, 10:677–698, 2017.

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

N. L. Zhang. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5:697–723, 2004.