Submitted to the Annals of Statistics
arXiv: stat.ML/1203.0697

LEARNING HIGH-DIMENSIONAL MIXTURES OF GRAPHICAL MODELS

By Animashree Anandkumar[*,¶], Daniel Hsu[‖], Furong Huang[†,¶], and Sham Kakade[‖]

Univ. of California Irvine[¶], *Microsoft Research New England* [‖]

We consider unsupervised estimation of mixtures of discrete graphical models, where the class variable corresponding to the mixture components is hidden and each mixture component over the observed variables can have a potentially different Markov graph structure and parameters. We propose a novel moment-based approach for estimating the mixture components, and our output is a tree-mixture model which serves as a good approximation to the underlying graphical model mixture. Our method is efficient when the union graph, which is the union of the Markov graphs of the mixture components with sparse vertex separators between any pair of observed variables. This includes tree mixtures and mixtures of bounded degree graphs. For such models, we prove that our method correctly recovers the union graph structure and the tree structures corresponding to maximum-likelihood tree approximations of the mixture components. The sample and computational complexities of our method scale as $\text{poly}(p, r)$, for an r-component mixture of p-variate graphical models. Our approach offers a powerful alternative to heuristics such as expectation maximization (EM) for learning mixture models.

1. Introduction. Mixture models are applied widely and can account for changes in observed data based on hidden influences [35]. A mixture model can be thought of as selecting the distribution of the manifest variables from a fixed set, depending on the realization of a so-called choice variable, which is latent or hidden (in an unsupervised setting). In classical mixture modeling, typically, there are two goals: model-based clustering, where learning the parameters of the mixture components is the main goal, and density estimation, where the mixture components themselves are not of much interest, but the goal is to estimate the overall mixture model accurately and employ it for prediction.

We mainly focus on the former goal in this paper, viz., that we are interested in recovering the mixture components efficiently. This arises in a variety of applications, e.g. in the biological domain, it is now widely accepted that cancers arise due to various interactions among the genes (termed as cancer pathways), and different types of cancers arise due to different forms of gene interactions [49]. Learning these pathways is critical to developing effective cancer therapies. Thus, we can model different gene pathways as the mixture components that lead to different types of cancer (the choice variable), and it is important to estimate the mixture components here. Similarly, in text analysis, an important application is to infer the contexts of the documents at hand, and this is done by studying the co-occurrences of the words [11]. We can model the occurrence of various words under each context as a mixture component, and again, the goal is to infer the mixture components. Similarly, in the social domain, an important problem is community detection [33], i.e. inferring the hidden communities of people by studying their interactions. Here, the interactions among

[*] Supported in part by the NSF Career award CCF-1254106, NSF Award CCF-1219234, AFOSR Award FA9550-10-1-0310, and ARO Award W911NF-12-1-0404.

[†] Supported by the NSF Career award CCF-1254106, AFOSR Award FA9550-10-1-0310, and NSF Award 1028394.

AMS 2000 subject classifications: 62H12

Keywords and phrases: Graphical models, Mixture models, Spectral methods, Tree approximation

different communities can be modeled as mixture components, and learning these interactions is an important goal.

In the above examples, it is important to efficiently model the interactions among the variables in each mixture component. We employ the popular framework of (undirected) probabilistic graphical models to model each mixture component, i.e. we consider mixtures of graphical models. Graphical models offer a graph-based framework for representing multivariate distributions, where qualitative relationships among the variables are represented via a graph structure, while quantitative relationships are represented via values assigned to different node groups on the graph [31]. These models allow for parsimonious representation of high-dimensional data, while retaining the computational advantage of performing inference via belief propagation and its variants. Mixtures of graphical models can incorporate *context-specific dependencies*, where the structural (and parametric) relationships among the observed variables can change depending on the hidden choice variable, and this is especially relevant in the applications described above.

1.1. *Summary of Results.* We propose a novel moment-based approach to learning mixtures of discrete graphical models. It combines the techniques used in graphical model selection, based on conditional independence tests, and the moment-based spectral decomposition methods, employed for estimating the mixture components. We establish that the proposed method succeeds in recovering the underlying components under some natural and transparent conditions, and the model class includes tree mixtures and mixtures over bounded degree graphs. Moreover, the computational and sample complexities of our method scale as low order polynomials in the number of nodes and the number of mixture components. To the best of our knowledge, our work is the first to provide provable guarantees for learning non-trivial mixtures of graphical models (which are not mixtures of product distributions).

The current practice for learning mixtures of graphical models (and other mixture models) is based on local-search heuristics such as expectation maximization (EM). However, EM scales poorly in the number of dimensions, suffers from convergence issues, and lacks theoretical guarantees. Our proposed method offers a powerful and a fast alternative to EM, and our experiments demonstrate that our proposed method has superior performance in recovering the mixture components. On the other hand, EM is superior in density estimation in our experiments, which is not surprising since it aims to optimize the overall likelihood. In our experience, combining the two techniques, i.e. initializing EM with the output of our method, allows us to achieve the best of both the worlds: accurate estimation of the mixture components as well as good density estimation. The intuition behind this improvement is the fact that our moment-based spectral estimator can be improved locally by running EM, and this agrees with the classical result that taking a single step of Newton-Ralphson on the likelihood function starting from a moment-based estimate can lead to asymptotically efficient estimation [34].

1.2. *Overview of the Method and Techniques.* Our method proceeds in three main stages: (union) graph structure estimation, estimation of mixture components, and tree approximation.

In the first stage, our algorithm estimates the union graph structure, corresponding to the union of the Markov graphs of the mixture components. We propose a rank criterion for classifying a node pair as neighbors or non-neighbors in the union graph of the mixture model, and can be viewed as a generalization of conditional-independence tests for graphical model selection [6, 13, 46]. Our method is efficient when the union graph has sparse separators between any node pair, which holds for tree mixtures and mixtures of bounded degree graphs. The sample complexity of our algorithm is logarithmic in the number of nodes. Thus, our method learns the union graph structure of a graphical model mixture with similar guarantees as graphical model selection (i.e., when there is a single graphical model).

We also extend our analysis of union graph estimation to a larger family of models, where the union graph has sparse local separators [6], which is a weaker criterion, but the model is in the regime of correlation decay [6]. This family includes locally tree-like graphs (including sparse random graphs), and augmented graphs (e.g. small-world graphs where there is a local and a global graph).

In the second stage, we use the union graph estimate $\widehat{G}_\cup$ to learn the pairwise marginals of the mixture components. Since the choice variable is hidden, this involves decomposition of the observed statistics into component models in an unsupervised manner. We leverage on the spectral decomposition method developed for learning mixtures of product distributions [2, 15, 40]. In a mixture of product distributions, the observed variables are conditionally independent given the hidden class variable. We adapt this method to our setting as follows: we consider different triplets over the observed nodes and condition on suitable separator sets (in the union graph estimate $\widehat{G}_\cup$) to obtain a series of mixtures of product distributions. Thus, we obtain estimates for pairwise marginals of each mixture component (and in principle, higher order moments) under some natural non-degeneracy conditions.

In the final stage, we find the best tree approximation to the estimated component marginals via the standard Chow-Liu algorithm [18]. The Chow-Liu algorithm produces a max-weight spanning tree using the estimated pairwise mutual information as edge weights. In our view, a tree-mixture approximation offers good tradeoff between data fitting and inferential complexity of the model. Tree mixtures are attractive since inference reduces to belief propagation on the component trees [37]. Tree mixtures thus present a middle ground between tree graphical models, which are too simplistic, and general graphical model mixtures, where inference is not tractable, and our goal is to efficiently fit the observed data to a tree mixture model.

We establish that our overall algorithm recovers the correct tree structure corresponding to maximum-likelihood tree approximation of each mixture component for a wide class of models. In the special case, when the underlying distribution is a tree mixture, this implies that we can correctly recover tree structures (and parameters) corresponding to all the mixture components. Our proof techniques involve establishing the correctness of our algorithm (under exact statistics). The sample analysis involves careful use of spectral perturbation bounds to guarantee success in finding the mixture components.

1.3. *Related Work.*

Our work lies at the intersection of learning mixture models and graphical model selection. We outline related works in both these areas.

Overview of Mixture Models: . Mixture models have been extensively studied [35] and are employed in a variety of applications. More recently, the focus has been on learning mixture models in high dimensions. There are a number of recent works dealing with estimation of high-dimensional Gaussian mixtures, starting from the work of Dasgupta [20] for learning well-separated components, and most recently by [10, 39], in a long line of works. These works provide guarantees on recovery under various separation constraints between the mixture components and/or have computational and sample complexities growing exponentially in the number of mixture components r. In contrast, the so-called spectral methods have both computational and sample complexities scaling only polynomially in the number of components, and do not impose stringent separation constraints, and we outline below.

Spectral Methods for Mixtures of Product Distributions: . The classical mixture model over product distributions consists of multivariate distributions with a single latent variable H and the observed variables are conditionally independent under each state of the latent variable [32]. Hierarchical latent class (HLC) models [16, 51, 52] generalize these models by allowing for multiple latent variables. Spectral methods were first employed for learning discrete (hierarchical) mixtures of product distributions [15, 27, 40] and have been recently extended for learning general multiview mixtures [2]. The method is based on triplet and pairwise statistics of observed variables and we build on these methods in our work. Note that our setting is *not* a mixture of product distributions, and thus, these methods are not directly applicable.

Graphical Model Selection: . Graphical model selection is a well studied problem starting from the seminal work of Chow and Liu [18] for finding the best tree approximation of a graphical model. They established that maximum likelihood estimation reduces to a maximum weight spanning tree problem where the edge weights are given by empirical mutual information. However, the problem becomes more challenging when either some of the nodes are hidden (i.e., latent tree models) or we are interested in estimating loopy graphs. Learning the structure of latent tree models has been studied extensively, mainly in the context of phylogenetics [22]. Efficient algorithms with provable performance guarantees are available, e.g. [5, 17, 21, 23]. Works on high-dimensional loopy graphical model selection are more recent. The approaches can be classified into mainly two groups: non-convex local approaches [4, 6, 13, 28, 42] and those based on convex optimization [14, 38, 43, 44]. There is also some recent work on learning conditional models, e.g. [26]. However, these works are not directly applicable for learning mixtures of graphical models.

Mixtures of Graphical Models: . Works on learning mixtures of graphical models (other than mixtures of product distributions) are fewer, and mostly focus on tree mixtures. The works of Meila and Jordan [37] and Kumar and Koller [30] consider EM-based approaches for learning tree mixtures, Thiesson *et al.* [48] extend the approach to learn mixtures of graphical models on directed acyclic graphs (DAG), termed as Bayesian multinets by Geiger and Heckerman [25], using the Cheeseman-Stutz asymptotic approximation and Armstrong *et al.* [8] consider a Bayesian approach by assigning a prior to decomposable graphs. However, these approaches do not have any theoretical guarantees.

Theoretically, the problem of separating a mixture of graphical models is challenging and ill-posed in general. For instance, the works in [1, 36] discuss identifiability issues which arise in the special case of tree mixtures. Recently, Mossel and Roch [41] consider structure learning of latent tree mixtures and provide conditions under which they become identifiable and can be successfully recovered. Note that this model can be thought of as a hierarchical mixture of product distributions, where the hierarchy changes according to the realization of the choice variable. Our setting differs substantially from this work. Mossel and Roch [41] require that the component latent trees of the mixture be very different, in order for the quartet tests to distinguish them (roughly), and establish that a uniform selection of trees will ensure this condition. On the other hand, we impose no such restriction and allow for graphs of different components to be same/different (although our algorithm is efficient when the overlap between the component graphs is more). Moreover, we allow for loopy graphs while Mossel and Roch [41] restrict to learning latent tree mixtures. However, Mossel and Roch [41] do allow for latent variables on the tree, while we assume that all variables to be observed (except for the latent choice variable). Mossel and Roch [41] consider only structure learning, while we consider both structure and parameter estimations. Mossel and Roch [41] limit to finite number of mixtures $r = O(1)$, while we allow for r to scale with the number of variables p. As such, the two methods operate in significantly different settings.

2. System Model.

2.1. *Graphical Models.* We first introduce the concept of a graphical model and then discuss mixture models. A graphical model is a family of multivariate distributions Markov on a given undirected graph [31]. In a discrete graphical model, each node in the graph $v \in V$ is associated with a random variable Y_v taking value in a finite set $\mathcal{Y}$ and let $d := |\mathcal{Y}|$ denote the cardinality of the set. The set of edges[1] $E \subset \binom{V}{2}$ captures the set of conditional-independence relationships among the random variables. We say that a vector of random variables $\mathbf{Y} := (Y_1, \ldots, Y_p)$ with a joint probability mass function (pmf) P is Markov on the graph G if the *local Markov property*

$$P(y_v|\mathbf{y}_{\mathcal{N}(i)}) = P(y_v|\mathbf{y}_{V \setminus v}) \tag{1}$$

holds for all nodes $v \in V$, where $\mathcal{N}(v)$ denotes the open neighborhood of v (i.e., not including v). More generally, we say that P satisfies the *global Markov property* for all disjoint sets $A, B \subset V$

$$P(\mathbf{y}_A, \mathbf{y}_B|\mathbf{y}_{\mathcal{S}(A,B;G)}) = P(\mathbf{y}_A|\mathbf{y}_{\mathcal{S}(A,B;G)})P(\mathbf{y}_B|\mathbf{y}_{\mathcal{S}(A,B;G)}), \quad \forall A, B \subset V : \mathcal{N}[A] \cap \mathcal{N}[B] = \emptyset. \tag{2}$$

where the set $\mathcal{S}(A, B; G)$ is a *node separator*[2]between A and B, and $\mathcal{N}[A]$ denotes the closed neighborhood of A (i.e., including A). The global and local Markov properties are equivalent under the *positivity condition*, given by $P(\mathbf{y}) > 0$, for all $\mathbf{y} \in \mathcal{Y}^p$ [31]. Henceforth, we say that a graphical model satisfies Markov property with respect to a graph, if it satisfies the global Markov property.

The Hammersley-Clifford theorem [12] states that under the positivity condition, a distribution P satisfies the Markov property according to a graph G iff. it factorizes according to the cliques of G,

$$P(\mathbf{y}) = \frac{1}{Z} \exp\left(\sum_{c \in \mathcal{C}} \Psi_c(\mathbf{y}_c)\right), \tag{3}$$

where $\mathcal{C}$ is the set of cliques of G and $\mathbf{y}_c$ is the set of random variables on clique c. The quantity Z is known as the *partition function* and serves to normalize the probability distribution. The functions Ψ_c are known as *potential* functions. We will assume positivity of the graphical models under consideration, but otherwise allow for general potentials (including higher order potentials).

2.2. *Mixtures of Graphical Models.* In this paper, we consider mixtures of discrete graphical models. Let H denote the discrete hidden choice variable corresponding to the selection of a different components of the mixture, taking values in $[r] := \{1, \ldots, r\}$ and let $\mathbf{Y}$ denote the observed variables of the mixture. Denote $\boldsymbol{\pi}_H := [P(H = h)]_h^\top$ as the probability vector of the mixing weights and G_h as the Markov graph of the distribution $P(\mathbf{y}|H = h)$.

Our goal is to learn the mixture of graphical models, given n i.i.d. samples $\mathbf{y}^n = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top$ drawn from a p-variate joint distribution $P(\mathbf{y})$ of the mixture model, where each variable is a d-dimensional discrete variable. The component Markov graphs $\{G_h\}_h$ corresponding to models $\{P(\mathbf{y}|H = h)\}_h$ are assumed to be unknown. Moreover, the variable H is latent and thus, we do not a priori know the mixture component from which a sample is drawn. This implies that we cannot directly apply the previous methods designed for graphical model selection. A major challenge is thus being able to decompose the observed statistics into the mixture components.

[1] We use notations E and G interchangeably to denote the set of edges.

[2] A set $\mathcal{S}(A, B; G) \subset V$ is a separator of sets A and B if the removal of nodes in $\mathcal{S}(A, B; G)$ separates A and B into distinct components.

We now propose a method for learning the mixture components given n i.i.d. samples $\mathbf{y}^n$ drawn from a graphical mixture model $P(\mathbf{y})$. Our method proceeds in three main stages. First, we estimate the graph $G_\cup := \cup_{h=1}^r G_h$, which is the union of the Markov graphs of the mixture. This is accomplished via a series of rank tests. Note that in the special case when $G_h \equiv G_\cup$, this gives the graph estimates of the component models. We then use the graph estimate $\widehat{G}_\cup$ to obtain the pairwise marginals of the respective mixture components via a spectral decomposition method. Finally, we use the Chow-Liu algorithm to obtain tree approximations $\{T_h\}_h$ of the individual mixture components[3].

3. Estimation of the Union of Component Graphs.

Notation:. Our learning method will be based on the estimates of probability matrices. For any two nodes $u, v \in V$ and any set $S \subset V \setminus \{u, v\}$, denote the joint probability matrix

$$M_{u,v,\{S;k\}} := [P(Y_u = i, Y_v = j, \mathbf{Y}_S = k)]_{i,j}, \quad k \in \mathcal{Y}^{|S|}. \tag{4}$$

Let $\widehat{M}^n_{u,v,\{S;k\}}$ denote the corresponding estimated matrices using samples $\mathbf{y}^n$

$$\widehat{M}^n_{u,v,\{S;k\}} := [\widehat{P}^n(Y_u = i, Y_v = j, \mathbf{Y}_S = k)]_{i,j}, \tag{5}$$

where $\widehat{P}^n$ denotes the empirical probability distribution, computed using n samples. We consider sets S satisfying $|S| \leq \eta$, where η depends on the graph family under consideration. Thus, our method is based on $(\eta + 2)^{\text{th}}$ order statistics of the observed variables.

Intuitions: . We provide some intuitions and properties of the union graph $G_\cup = \cup_{h=1}^r G_h$, where G_h is the Markov graph corresponding to component $H = h$. Note that $G_\cup$ is different from the Markov graph corresponding to the marginalized model $P(\mathbf{y})$ (with latent choice variable H marginalized out). Yet, $G_\cup$ represents some natural Markov properties with respect to the observed statistics. We first establish the simple result that the union graph $G_\cup$ satisfies Markov property in each mixture component. Recall that $\mathcal{S}(u, v; G)$ denotes a vertex separator between nodes u and v in G, i.e., its removal disconnects u and v in G.

FACT 1 (Markov Property of $G_\cup$). *For any two nodes $u, v \in V$ such that $(u, v) \notin G_\cup$,*

$$Y_u \perp\!\!\!\perp Y_v | \mathbf{Y}_S, H, \quad S := \bigcup_{h=1}^r \mathcal{S}(u, v; G_h) \subseteq \mathcal{S}(u, v; G_\cup). \tag{6}$$

Proof: The set $S := \cup_{h=1}^r \mathcal{S}(u, v; G_h)$ is also a vertex separator for nodes u and v in each component graph G_h. This is because removal of S disconnects u and v in each G_h. Thus, we have Markov property in each component: $Y_u \perp\!\!\!\perp Y_v | \mathbf{Y}_S, \{H = h\}$, for $h \in [r]$, and the above result follows. Note that $\cup_h^r \mathcal{S}(u, v; G_h) \subseteq \mathcal{S}(u, v; G_\cup)$ since a separation in the union graph implies separation in its components. □

Fact 1 implies that the conditional independence relationships of each mixture component are satisfied on the union graph $G_\cup$ conditioned on the latent factor H. The above result can be exploited to obtain union graph estimate as follows: two nodes u, v are not neighbors in $G_\cup$ if a separator set S can be found which results in conditional independence, as in (6). The main

[3] Our method can also be adapted to estimating the component Markov graphs $\{G_h\}_h$ and we outline it and other extensions in Appendix A.1.

challenge is indeed that the variable H is not observed and thus, conditional independence cannot be directly inferred via observed statistics. However, the effect of H on the observed statistics can be quantified as follows:

LEMMA 1 (Rank Property). *Given an r-component mixture of graphical models with $G_\cup = \cup_{h=1}^r G_h$, for any $u, v \in V$ such that $(u,v) \notin G_\cup$ and $S := \cup_{h=1}^r \mathcal{S}(u,v;G_h)$, the probability matrix $M_{u,v,\{S;k\}} := [P[Y_u = i, Y_v = j, \mathbf{Y}_S = k]]_{i,j}$ has rank at most r for any $k \in \mathcal{Y}^{|S|}$.*

Proof: From Fact 1, $G_\cup$ satisfies Markov property conditioned on the latent factor H,

$$Y_u \perp\!\!\!\perp Y_v | \mathbf{Y}_S, H, \quad \forall\, (u,v) \notin G_\cup. \tag{7}$$

This implies that

$$M_{u,v,\{S;k\}} = M_{u|H,\{S;k\}} \operatorname{Diag}(\boldsymbol{\pi}_{H|\{S;k\}}) M^\top_{v|H,\{S;k\}} P(\mathbf{Y}_S = k), \tag{8}$$

where $M_{u|H,\{S;k\}} := [P[Y_u = i|H = j, \mathbf{Y}_S = k]]_{i,j}$ and similarly $M_{v|H,\{S;k\}}$ is defined. $\operatorname{Diag}(\boldsymbol{\pi}_{H|S;k})$ is the diagonal matrix with entries $\boldsymbol{\pi}_{H|\{S;k\}} := [P(H = i|\mathbf{Y}_S = k)]_i$. Thus, we have $\operatorname{Rank}(M_{u,v,\{S;k\}})$ is at most r. □

Thus, the effect of marginalizing the choice variable H is seen in the rank of the observed probability matrices $M_{u,v,\{S;k\}}$. Thus, when u and v are non-neighbors in $G_\cup$, a separator set S can be found such that the rank of $M_{u,v,\{S;k\}}$ is at most r. In order to use this result as a criterion for inferring neighbors in $G_\cup$, we require that the rank of $M_{u,v,\{S;k\}}$ for any neighbors $(u,v) \in G_\cup$ be strictly larger than r. This requires the dimension of each node variable $d > r$. We discuss in detail the set of sufficient conditions for correctly recovering $G_\cup$ in Section 3.1.1.

Tractable Graph Families: . Another obstacle in using Lemma 1 to estimate graph $G_\cup$ is computational: the search for separators S for any node pair $u, v \in V$ is exponential in $|V| := p$ if no further constraints are imposed. Define $s(G_1, \ldots, G_r)$ to be the worst-case bound for the model under consideration:

$$\bigcup_{h=1}^{r} |\mathcal{S}(u,v;G_h)| \leq s(G_1,\ldots,G_r), \quad \forall\, (u,v) \notin G_\cup, G_\cup := G_1 \cup \ldots G_r. \tag{9}$$

Note that $\cup_{h=1}^r \mathcal{S}(u,v;G_h) \subseteq \mathcal{S}(u,v;G_\cup)$ since a separation on the union graph implies separation in its components. This implies that

$$s(G_1,\ldots,G_r) \leq s(G_\cup), \tag{10}$$

and equality holds when $G_1 = \ldots = G_r$. Similarly, we also have the bound

$$s(G_1,\ldots,G_r) \leq \sum_{h=1}^{r} s(G_h). \tag{11}$$

In light of the above bounds, we list a few graph families where $s(G_1, \ldots, G_r)$ or its bound $s(G_\cup)$ is small:

1. If $G_\cup$ is trivial (i.e., no edges) then $s(G_\cup) = 0$, we have a mixture of product distributions.
2. When $G_\cup$ is a tree, i.e., we have a mixture model Markov on the same tree, then $s(G_\cup) = 1$, since there is a unique path between any two nodes on a tree.

Algorithm 1 $\widehat{G}^n_\cup = \mathsf{RankTest}(\mathbf{y}^n; \xi_{n,p}, \eta, r)$ for estimating $G_\cup := \cup_{h=1}^r G_h$ of an r-component mixture using $\mathbf{y}^n$ samples, where η is the bound on size of vertex separators between any node pair: $\max_{u,v} \cup_{h=1}^r |\mathcal{S}(u,v;G_h)| \leq \eta$, and $\xi_{n,p}$ is a threshold on the singular values.

Rank$(A;\xi)$ denotes the effective rank of matrix A, i.e., number of singular values more than ξ. $\widehat{M}^n_{u,v,\{S;k\}} := [\widehat{P}^n(Y_u = i, Y_v = j, \mathbf{Y}_S = k)]_{i,j}$ is the empirical estimate computed using n i.i.d. samples $\mathbf{y}^n$. Initialize $\widehat{G}^n_\cup = (V, \emptyset)$. For each $u, v \in V$, estimate $\widehat{M}^n_{u,v,\{S;k\}}$ from $\mathbf{y}^n$ for some configuration $k \in \mathcal{Y}^{|S|}$, if

$$\min_{\substack{S \subset V \setminus \{u,v\} \\ |S| \leq \eta}} \text{Rank}(\widehat{M}^n_{u,v,\{S;k\}}; \xi_{n,p}) > r, \tag{12}$$

then add (u,v) to $\widehat{G}^n_\cup$.

3. For a general graph $G_\cup$ with *treewidth* tw$(G_\cup)$ and maximum degree $\Delta(G_\cup)$, we have that $s(G_\cup) \leq \min(\Delta(G_\cup), \text{tw}(G_\cup))$.
4. For an arbitrary r-component tree mixture, $G_\cup = \cup_h T_h$ where each component is a tree, we have $s(T_1, \ldots, T_r) \leq r$ since $s(T_i) = 1$ and we use (11).
5. For an arbitrary mixture of bounded degree graphs, we have $s(G_1, \ldots, G_r) \leq \sum_{h \in [r]} \Delta_h$, where Δ_h is the maximum degree in G_h using (11).

Rank Test: . We propose $\mathsf{RankTest}(\mathbf{y}^n; \xi_{n,p}, \eta, r)$ in Algorithm 1 for structure estimation of $G_\cup := \cup_{h=1}^r G_h$, the union Markov graph of an r-component mixture. The method is based on a search for potential separators S between any two given nodes $u, v \in V$, based on the effective rank[4]of $\widehat{M}^n_{u,v,\{S;k\}}$: if the effective rank is r or less, then u and v are declared as non-neighbors (and set S as their separator). If no such sets are found, they are declared as neighbors. Thus, the method involves searching for separators for each node pair $u, v \in V$, by considering all sets $S \subset V \setminus \{u, v\}$ satisfying $|S| \leq \eta$. From Lemma 1, it is clear that the rank test for structure estimation succeeds if we set $\eta \geq s(G_1, \ldots, G_r)$. The computational complexity of this procedure is $O(p^{\eta+2} d^3)$, where d is the dimension of each node variable Y_i, for $i \in V$ and p is the number of nodes. This is because the number of rank tests performed is $O(p^{\eta+2})$ over all node pairs and conditioning sets; each rank tests has $O(d^3)$ complexity since it involves performing singular value decomposition (SVD) of a $d \times d$ matrix.

From the previous observations, for a wide family of models, $\mathsf{RankTest}(\mathbf{y}^n; \xi_{n,p}, \eta, r)$ requires only a small separator bound η for success, and includes tree mixtures and mixtures over bounded degree graphs. In Section B, we relax the requirement of exact separation to that of local separation. A larger class of graphs satisfy the local separation property including mixtures of locally tree-like graphs.

3.1. *Results for the Rank Test.*

3.1.1. *Conditions for the Success of Rank Tests.* The following assumptions are made for the $\mathsf{RankTest}$ proposed in Algorithm 1 to succeed under the PAC formulation.

(A1) **Number of Mixture Components:** The number of components r of the mixture model and dimension d of each node variable satisfy

$$d > r. \tag{13}$$

[4] The effective rank is given by the number of singular values above a given threshold ξ.

The mixing weights of the latent factor H are assumed to be strictly positive

$$\pi_H(h) := P(H = h) > 0, \quad \forall\ h \in [r].$$

(A2) **Constraints on Graph Structure:** Recall that $s(G_1, \ldots, G_r)$ to be the worst-case bound on the union of separators in the component graphs $G_1, \ldots, G_r$ in (9) and we assume that $s(G_1, \ldots, G_r) = O(1)$. We choose parameter η in rank test as $\eta \geq s(G_1, \ldots, G_r)$.
In Section B, we relax the strict separation constraint to a local separation constraint in the regime of correlation decay, where η refers to the bound on the size of local separators between any two non-neighbor nodes in the union graph.

(A3) **Rank Condition:** We assume that the matrix $M_{u,v,\{S;k\}}$ in (4) has rank strictly greater than r when the nodes u and v are neighbors in graph $G_\cup = \cup_{h=1}^r G_h$ and the set satisfies $|S| \leq \eta$. Let $\rho_{\min}$ denote

$$\rho_{\min} := \min_{\substack{(u,v)\in G_\cup, |S|\leq\eta \\ S\subset V\setminus\{u,v\}}} \max_{k\in\mathcal{Y}^{|S|}} \sigma_{r+1}\left(M_{u,v,\{S;k\}}\right) > 0, \tag{14}$$

where $\sigma_{r+1}(\cdot)$ denotes the $(r+1)^{\text{th}}$ singular value, when the singular values are arranged in the descending order $\sigma_1(\cdot) \geq \sigma_2(\cdot) \geq \ldots \sigma_d(\cdot)$.

(A4) **Choice of threshold ξ:** For RankTest in Algorithm 1, the threshold ξ is chosen as

$$\xi := \frac{\rho_{\min}}{2}.$$

(A5) **Number of Samples:** Given $\delta \in (0, 1)$, the number of samples n satisfies

$$n > n_{\text{Rank}}(\delta; p) := \max\left(\frac{1}{t^2}\left(2\log p + \log\delta^{-1} + \log 2\right), \left(\frac{2}{\rho_{\min} - t}\right)^2\right), \tag{15}$$

for some $t \in (0, \rho_{\min})$ (e.g. $t = \rho_{\min}/2$,) where p is the number of nodes and $\rho_{\min}$ is given by (14).

Assumption (A1) relates the number of components to the dimension of the sample space of the variables. Note that we allow for the number of components r to grow with the number of nodes p, as long as the cardinality of the sample space of each variable d is also large enough. In principle, this assumption can be removed by considering grouping the nodes together and performing rank tests on the groups. Assumption (A2) imposes constraints on the graph structure $G_\cup$, formed by the union of the component graphs. The bound $s(G_1, \ldots, G_r)$ on the separator sets in the component graphs is a crucial parameter and the complexity of learning (both sample and computational) depends on it. We relax the assumption of separator bound to a criterion of local separation in Section B. Assumption (A3) is required for the success of rank tests to distinguish neighbors and non-neighbors in graph $G_\cup$. It rules out the presence of spurious low rank matrices between neighboring nodes in $G_\cup$ (for instance, when the nodes are marginally independent or when the distribution is degenerate). Assumption (A4) provides a natural threshold on the singular values in the rank test. In Section B, we modify the threshold to also account for distortion due to approximate vertex separation, in contrast to the setting of exact separation considered in this section. (A5) provides the finite sample complexity bound.

3.1.2. *Result on Rank Tests.* We now provide the result on the success of recovering the graph $G_\cup := \cup_{h=1}^r G_h$.

THEOREM 1 (Success of Rank Tests). *The* $\mathsf{RankTest}(\mathbf{y}^n; \xi, \eta, r)$ *outputs the correct graph* $G_\cup := \cup_{h=1}^r G_h$*, which is the union of the component Markov graphs, under the assumptions (A1)–(A5) with probability at least* $1-\delta$.

Proof: The proof is given in Appendix C. □

A special case of the above result is graphical model selection, where there is a single graphical model ($r=1$) and we are interested in estimating its graph structure.

COROLLARY 1 (Application to Graphical Model Selection). *The* $\mathsf{RankTest}(\mathbf{y}^n; \xi, \eta, 1)$ *outputs the correct Markov graph G, given n i.i.d. samples* $\mathbf{y}^n$*, under the assumptions*[5] *(A2)–(A5) with probability at least* $1-\delta$.

Remarks: Thus, the rank test is also applicable for graphical model selection. Previous works (see Section 1.3) have proposed tests based on conditional independence, using either conditional mutual information or conditional variation distances, see [6, 13]. The rank test above is thus an alternative test for conditional independence. In addition, it extends naturally to estimation of union graph structure of mixture components.

4. Parameter Estimation of Mixture Components.

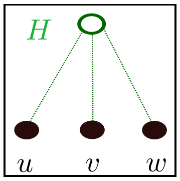
The rank test proposed in the previous section is a tractable procedure for estimating the graph $G_\cup := \cup_{h=1}^r G_h$, which is the union of the component graphs of a mixture of graphical models. However, except in the special case when $G_h \equiv G_\cup$, the knowledge of $\widehat{G}_\cup^n$ is not very useful by itself, since we do not know the nature of the different components of the mixture. In this section, we propose the use of spectral decomposition tests to find the various mixture components.

4.0.3. *Spectral Decomposition for Mixture of Product Distributions.* The spectral decomposition methods, first proposed by Chang [15], and later generalized by Mossel and Roch [40] and Hsu, Kakade and Zhang [27], and recently by Anandkumar, Hsu and Kakade [2], are applicable for mixtures of product distributions. We illustrate the method below via a simple example.

Consider the simple case of three observed variables $\{Y_u, Y_v, Y_w\}$, where a latent factor H separates them, i.e., the observed variables are conditionally independent given H

$$Y_u \perp\!\!\!\perp Y_v \perp\!\!\!\perp Y_w | H.$$

This implies that the Markov graphs $\{G_h\}_{h\in[r]}$ of the component models $\{P(Y_u, Y_v, Y_w|H=h)\}_{h\in[r]}$ are trivial (i.e., have no edges) and thus forms a special case of our setting.



We now give an overview of the spectral decomposition method. It proceeds by considering pairwise and triplet statistics of Y_u, Y_v, Y_w. Denote $M_{u|H} := [P(Y_u = i|H=j)]_{i,j}$, and similarly for $M_{v|H}, M_{w|H}$ and assume that they are full rank. Denote the probability matrices $M_{u,v} := [P(Y_u = i, Y_v = j)]_{i,j}$ and $M_{u,v,\{w;k\}} := [P(Y_u = i, Y_v = j, Y_w = k)]_{i,j}$. The parameters (i.e., matrices $M_{u|H}, M_{v|H}, M_{w|H}$) can be estimated as:

[5] When $r=1$, there is no latent factor, and the assumption $d > r$ in (A1) is trivially satisfied for all discrete random variables.

Lemma 2 (Mixture of Product Distributions). *For the latent variable model* $Y_u \perp\!\!\!\perp Y_v \perp\!\!\!\perp Y_w | H$, *when the conditional probability matrices* $M_{u|H}, M_{v|H}, M_{w|H}$ *have rank* d, *let* $\boldsymbol{\lambda}^{(k)} = [\lambda_1^{(k)}, \ldots, \lambda_d^{(k)}]^\top$ *be the column vector with the* d *eigenvalues given by*

$$(16) \qquad \boldsymbol{\lambda}^{(k)} := \mathsf{Eigenvalues}\left(M_{u,v,\{w;k\}} M_{u,v}^{-1}\right), \quad k \in \mathcal{Y}.$$

Let $\Lambda := [\boldsymbol{\lambda}^{(1)} | \boldsymbol{\lambda}^{(2)} | \ldots | \boldsymbol{\lambda}^{(d)}]$ *be the matrix where the* k^{th} *column corresponds to* $\boldsymbol{\lambda}^{(k)}$ *from above. We have that*

$$(17) \qquad M_{w|H} := [P(Y_w = i | H = j)]_{i,j} = \Lambda^\top.$$

Proof: A more general result is proven in Appendix D.1. □

Thus, we have a procedure for recovering the conditional probabilities of the observed variables conditioned on the latent factor. Using these parameters, we can also recover the mixing weights $\boldsymbol{\pi}_H := [P(H = i)]_i^\top$ using the relationship

$$M_{u,v} = M_{u|H} \operatorname{Diag}(\boldsymbol{\pi}_H) M_{v|H}^\top,$$

where $\operatorname{Diag}(\boldsymbol{\pi}_H)$ is the diagonal matrix with $\boldsymbol{\pi}_H$ as the diagonal elements.

Thus, if we have a general product distribution mixture over nodes in V, we can learn the parameters by performing the above spectral decomposition over different triplets $\{u, v, w\}$. However, an obstacle remains: spectral decomposition over different triplets $\{u, v, w\}$ results in different permutations of the labels of the hidden variable H. To overcome this, note that any two triplets (u, v, w) and (u, v', w') share the same set of eigenvectors in (16) when the "left" node u is the same. Thus, if we consider a fixed node $u_* \in V$ as the "left" node and use a fixed matrix to diagonalize (16) for all triplets, we obtain a consistent ordering of the hidden labels over all triplet decompositions. Thus, we can learn a general product distribution mixture using only third-order statistics.

4.0.4. *Spectral Decomposition for Learning Graphical Model Mixtures.* We now adapt the above method for learning more general graphical model mixtures. We first make a simple observation on how to obtain mixtures of product distributions by considering separators on the union graph $G_\cup$. For any three nodes $u, v, w \in V$, which are not neighbors on $G_\cup$, let S_{uvw} denote a *multiway* vertex separator, i.e., the removal of nodes in S_{uvw} disconnects u, v and w in $G_\cup$. On lines of Fact 1,

$$(18) \qquad Y_u \perp\!\!\!\perp Y_v \perp\!\!\!\perp Y_w | \mathbf{Y}_{S_{uvw}}, H, \quad \forall u, v, w : (u, v), (v, w), (w, u) \notin G_\cup.$$

Thus, by fixing the configuration of nodes in S_{uvw}, we obtain a product distribution mixture over $\{u, v, w\}$. If the previously proposed rank test is successful in estimating $G_\cup$, then we possess correct knowledge of the separators S_{uvw}. In this case, we can obtain estimates $\{P(Y_w | \mathbf{Y}_{S_{uvw}} = k, H = h)\}_h$ by fixing the nodes in S_{uvw} to k and using the spectral decomposition described in Lemma 2, and the procedure can be repeated over different triplets $\{u, v, w\}$. See Fig.1.

An obstacle remains, viz., the permutation of hidden labels over different triplet decompositions $\{u, v, w\}$. In case of product distribution mixture, as discussed previously, this is resolved by fixing the "left" node in the triplet to some $u_* \in V$ and using the same matrix for diagonalization over different triplets. However, an additional complication arises when we consider graphical model mixtures, where conditioning over separators is required. We require that the permutation of the hidden labels be unchanged upon conditioning over different values of variables in the separator set $S_{u_* vw}$. This holds when the separator set $S_{u_* vw}$ has no effect on node u_*, i.e., we require that

$$(19) \qquad \exists u_* \in V, s.t. \quad Y_{u_*} \perp\!\!\!\perp \mathbf{Y}_{V \setminus u_*} | H,$$

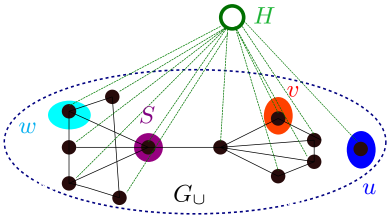


Fig 1. *By conditioning on the separator set S on the union graph $G_\cup$, we have a mixture of product distribution with respect to nodes $\{u, v, w\}$, i.e., $Y_u \perp\!\!\!\perp Y_v \perp\!\!\!\perp Y_w | \mathbf{Y}_S, H$.*

which implies that u_* is isolated from all other nodes in graph $G_\cup$.

Condition (19) is required to hold for identifiability if we only operate on statistics over different triplets (along with their separator sets). In other words, if we resort to operations over only low order statistics, we require additional conditions such as (19) for identifiability. However, our setting is a significant generalization over the mixtures of product distributions, where (19) is required to hold for all nodes.

Finally, since our goal is to estimate pairwise marginals of the mixture components, in place of node w in the triplet $\{u, v, w\}$ in Lemma 2, we need to consider a node pair $a, b \in V$. The general algorithm allows the variables in the triplet to have different dimensions, see [2] for details. Thus, we obtain estimates of the pairwise marginals of the mixture components. The computational complexity of the procedure scales as $O(p^2 d^{s(G_\cup)+6} r)$, where p is the number of nodes, d is the cardinality of each node variable and $s(G_\cup)$ is the bound on separator sets on the union graph $G_\cup$, see (9). For details on implementation of the spectral method, see Appendix A.

4.1. *Results for Spectral Decomposition.*

4.1.1. *Assumptions.* In addition to the assumptions (A1)–(A5) in Section 3.1.1, we impose the following constraints to guarantee the success of estimating the various mixture components.

(A6) **Full Rank Views of the Latent Factor:** For each node pair $a, b \in V$, and any subset $S \subset V \setminus \{a, b\}$ with $|S| \leq 2s(G_\cup)$ and $k \in \mathcal{Y}^{|S|}$, the probability matrix $M_{(a,b)|H,\{S;k\}} := [P(\mathbf{Y}_{a,b} = i | H = j, \mathbf{Y}_S = k)]_{i,j} \in \mathbb{R}^{d^2 \times r}$ has rank r.

(A7) **Existence of an Isolated Node:** There exists a node $u_* \in V$ which is isolated from all other nodes in $G_\cup = \cup_{h=1}^r G_h$, i.e.

$$Y_{u_*} \perp\!\!\!\perp \mathbf{Y}_{V \setminus u_*} | H. \tag{20}$$

(A8) **Spectral Bounds and Random Rotation Matrix:** Refer to various spectral bounds used to obtain $K(\delta; p, d, r)$ in Appendix D.3, where $\delta \in (0, 1)$ is fixed. Further assume that the rotation matrix $Z \in \mathbb{R}^{r \times r}$ in $\mathsf{FindMixtureComponents}$ is chosen uniformly over the Stiefel manifold $\{Q \in \mathbb{R}^{r \times r} : Q^\top Q = I\}$.

(A9) **Number of Samples:** For fixed $\delta, \epsilon \in (0, 1)$, the number of samples satisfies

$$n > n_{\text{spect}}(\delta, \epsilon; p, d, r) := \frac{4K^2(\delta; p, d, r)}{\epsilon^2}, \tag{21}$$

where $K(\delta; p, d, r)$ is defined in (64).

Assumption (A6) is a natural condition required for the success of spectral decomposition, and is imposed in [40], [27] and [2]. It is also known that learning singular models, i.e., those which do not satisfy (A6), is at least as hard as learning parity with noise, which is conjectured to be computationally hard [40]. The condition in (A7) is indeed an additional constraint on graph $G_\cup$, but is required to ensure alignment of hidden labels over spectral decompositions of different groups of variables, as discussed before[6] Condition (A8) assumes various spectral bounds and (A9) characterizes the sample complexity.

4.1.2. *Guarantees for Learning Mixture Components.* We now provide the result on the success of recovering the tree approximation T_h of each mixture component $P(\mathbf{y}|H=h)$. Let $\|\cdot\|_2$ on a vector denote the ℓ_2 norm.

THEOREM 2 (Guarantees for FindMixtureComponents). *Under the assumptions (A1)–(A9), the procedure in Algorithm 3 outputs $\widehat{P}^{\text{spect}}(Y_a, Y_b|H=h)$, for each $a, b \in V$, such that for all $h \in [r]$, there exists a permutation $\tau(h) \in [r]$ with*

$$\|\widehat{P}^{\text{spect}}(Y_a, Y_b|H=h) - P(Y_a, Y_b|H=\tau(h))\|_2 \le \epsilon, \tag{22}$$

with probability at least $1 - 4\delta$.

Proof: The proof is given in Appendix D. □

Remarks: . Recall that p denotes the number of variables, r denotes the number of mixture components, d denotes the dimension of each node variable and $s(G_\cup)$ denotes the bound on separator sets between any node pair in the union graph. The quantity $K(\delta; p, d, r)$ in (64) in Appendix D.3 is $O\left(p^{2s(G_\cup)+2} d^{2s(G_\cup)} r^5 \delta^{-1} \operatorname{poly}\log(p, d, r, \delta^{-1})\right)$. Thus, we require the number of samples scaling in (21) as $n = \Omega\left(p^{4s(G_\cup)+4} d^{4s(G_\cup)} r^{10} \delta^{-2} \epsilon^{-2} \operatorname{poly}\log(p, d, r, \delta^{-1})\right)$. Since we operate in the regime where $s(G_\cup) = O(1)$ is a small constant, this implies that we have a polynomial sample complexity in p, d, r. Note that the special case of $s(G_\cup) = 0$ corresponds to the case of mixture of product distributions, and it has the best sample complexity.

4.1.3. *Analysis of Tree Approximation.* We now consider the final stage of our approach, viz., learning tree approximations using the estimates of the pairwise marginals of the mixture components from the spectral decomposition method. We now impose a standard condition of non-degeneracy on each mixture component to guarantee the existence of a unique tree structure corresponding to the maximum-likelihood tree approximation to the mixture component.

(A10) **Separation of Mutual Information:** Let T_h denote the Chow-Liu tree corresponding to the model $P(\mathbf{y}|H=h)$ when exact statistics are input and let

$$\vartheta := \min_{h \in [r]} \min_{(a,b) \notin T_h} \min_{(u,v) \in \text{Path}(a,b;T_h)} \left(I(Y_u, Y_v|H=h) - I(Y_a, Y_b|H=h)\right), \tag{23}$$

where $\text{Path}(a, b; T_h)$ denotes the edges along the path connecting a and b in T_h.

(A11) **Number of Samples:** For ϵ^{tree} defined in (75), the number of samples is now required to satisfy

$$n > n_{\text{spect}}(\delta, \epsilon^{\text{tree}}; p, d, r), \tag{24}$$

where n_{spect} is given by (21).

[6] (A7) can be relaxed as follows: if graph $G_\cup$ has at least three connected components C_1, C_2, C_3, then we can choose a reference node in each of the components and estimate the marginals in the other components. We can then align these different estimates and obtain all the marginals.

The condition in (A10) assumes a separation between mutual information along edges and non-edges of the Chow-Liu tree T_h of each component model $P(\mathbf{y}|H = h)$. The quantity ϑ represents the minimum separation between the mutual information along an edge and any set of non-edges which can replace the edge in T_h. Note that $\vartheta \geq 0$ due to the max-weight spanning tree property of T_h (under exact statistics). Intuitively ϑ denotes the "bottleneck" where errors are most likely to occur in tree structure estimation. Similar observations were made by Tan, Anandkumar and Willsky [47] for error exponent analysis of Chow-Liu algorithm. The sample complexity for correctly estimating T_h using samples is based on ϑ_h and given in (A11). This ensures that the mutual information quantities are estimated within the separation bound ϑ.

Theorem 3 (Tree Approximations of Mixture Components). *Under (A1)–(A11), the Chow-Liu algorithm outputs the correct tree structures corresponding to maximum-likelihood tree approximations of the mixture components $\{P(\mathbf{y}|H = h)\}$ with probability at least $1 - 4\delta$, when the estimates of pairwise marginals $\{\widehat{P}^{\text{spect}}(Y_a, Y_b|H = h)\}$ from spectral decomposition method are input.*

Proof: See Section D.5. □

Remarks: . Thus our approach succeeds in recovering the correct tree structures corresponding to ML-tree approximations of mixture components with computational and sample complexities scaling polynomially in the number of variables p, number of components r and the dimension of each variable d.

Note that if the underlying model is a tree mixture, we recover the tree structures of the mixture components. For this special case, we can give a slightly better guarantee by estimating Chow-Liu trees which are subgraphs of the union graph estimate $\widehat{G}_\cup$, and this is discussed in Appendix D.4. The improved bound $K^{\text{tree}}(\delta; p, d, r)$ is

$$K^{\text{tree}}(\delta; p, d, r) = O\left(p^2 (d\Delta)^{2s(G_\cup)} r^5 \delta^{-1} \operatorname{poly}\log(p, d, r, \delta^{-1})\right), \tag{25}$$

where Δ is the maximum degree in $G_\cup$.

5. Triplet Tensor Decomposition.

In the spectral decomposition approach in Procedure 4, we utilize second and third order moments for learning mixture components and we provide learning guarantees for this approach. In this approach, we reduce the third-order tensor to a matrix by projecting along a random vector. This dimensionality reduction allows us to use popular linear algebraic techniques such as SVD to undertake simultaneous diagonalization. However, at the same time, this reduction from a tensor to a matrix generally results in a big loss in information and poor learning accuracy in practice. To resolve this, we appeal to higher order moment decomposition techniques, and specifically, using third order moments, to estimate the conditional probabilities via tensor decomposition. This allows us to fully utilize the information present in the third order moment tensor, and we observe a big improvement in estimation performance both on synthetic and real datasets via this tensor approach; see section 6 for experimental details. We call this approach the *triplet tensor decomposition*, and this is discussed in detail in [7], and we describe it below.

5.1. *Tensor Notations.*

A real third order tensor with dimensionality $[n_1, n_2, n_3]$ is a multi-dimensional array, and is denoted by $\mathcal{T} = \bigotimes_{i=1}^{3} \mathbb{R}^{n_i}$. A fiber of a tensor is a column vector obtained by fixing all but one of the dimension indices. A *slice* of a tensor is a matrix obtained by fixing all but two dimension indices.

A tensor $\mathcal{T}$ can be formulated in the Kruskal form as

$$vec(\mathcal{T}) = \sum_{j=1}^{r} \lambda_j U_1(:,j) \otimes U_2(:,j) \otimes U_3(:,j) \tag{26}$$

where $\lambda \in \mathbb{R}^r$, $U_i \in \mathbb{R}^{n_i \times r}$ and $vec(\mathcal{T})$ is the vectorized tensor. The decomposition of the above form with minimal number of terms is termed as canonical polyadic or CP decomposition. This can be formulated as an optimization problem:

$$\|\mathcal{T} - \mathcal{X}\|_F^2 = \left\| vec(\mathcal{T}) - \sum_{j=1}^{r} \lambda_j U_1(:,j) \otimes U_2(:,j) \otimes U_3(:,j) \right\|_2^2 \tag{27}$$

is minimized.

Given a tensor $\mathcal{T} = \bigotimes_{i=1}^{3} \mathbb{R}^{n_i}$, there are many ways to unfold or assemble its entries into matrices $T \in \mathbb{R}^{N_1 \times N_2}$ such that $N_1 N_2 = n_1 n_2 n_3$. A popular family of tensor unfoldings is the *mode-k unfoldings*. Let $\mathcal{T}_{(k)}$ be the mode-k unfolding of a third order tensor $\mathcal{T}$, then $\mathcal{T}_{(k)} \in \mathbb{R}^{n_k \times \frac{n_1 n_2 n_3}{n_k}}$ $\forall k \in \{1,2,3\}$ is the assembly of mode-k fibers. For the tensor of the form in (26), the mode-k unfoldings are given by

$$\begin{aligned}
\mathcal{X}_{(1)} &= \sum_{j=1}^{r} \lambda_j U_1(:,j) \otimes \left(U_3(:,j) \otimes U_2(:,j)\right)^\top = U_1 \operatorname{Diag}(\lambda_j) \left(U_3 \odot U_2\right)^\top \\
\mathcal{X}_{(2)} &= \sum_{j=1}^{r} \lambda_j U_2(:,j) \otimes \left(U_3(:,j) \otimes U_1(:,j)\right)^\top = U_2 \operatorname{Diag}(\lambda_j) \left(U_3 \odot U_1\right)^\top \\
\mathcal{X}_{(3)} &= \sum_{j=1}^{r} \lambda_j U_3(:,j) \otimes \left(U_2(:,j) \otimes U_1(:,j)\right)^\top = U_3 \operatorname{Diag}(\lambda_j) \left(U_2 \odot U_1\right)^\top,
\end{aligned}$$

Minimizing $\|\mathcal{T} - \mathcal{X}\|_F$ is equivalent to minimizing the error in each modal unfolding: $\left\|\mathcal{T}_{(k)} - \mathcal{X}_{(k)}\right\|$ $\forall k \in \{1,2,3\}$. Thus, we can apply alternating least squares (ALS) to each of the modal unfoldings: for instance, the objective function $\|\mathcal{T}_{(1)} - U_1 \operatorname{Diag}(\lambda_j) \left(U_3 \odot U_2\right)^\top\|$ is used to update $\{\lambda_1, U_1\}$ while keeping $\{\lambda_i, U_i\}$ fixed for $i = 2, 3$, and so on. This is a popular approach since it involves solving only simple least squares in each step, see the review on tensors [29] for a detailed discussion. We use $\mathsf{CP} - \mathsf{als}$ function from the tensor toolbox [9] in our experiments. Refer to Procedure 2 for the implementation details.

5.2. *Tensor Decomposition for Mixtures of Product Distributions.* We now demonstrate the relationship between learning mixtures and tensor decomposition. Specifically, we show that the third-order moments of a mixture of product distributions satisfy the tensor CP decomposition of the form in (26). For the mixture of products model, where $\{x_u \perp\!\!\!\perp x_v | H,\ \forall u, v \in V\}$, let $A^u \in \mathbb{R}^{d \times r}$ be the transition matrix $P\left(x_u | H\right)$, i.e. $A_{:,h} := P(x_u | H = h)$. Let the mixing weights be denoted by the vector $\pi \in \mathbb{R}^r$. If we set $x_u = e_j$, the basis vector in coordinate j, when u is at state j, and $H = e_h$ when H is in category h, the moments are of the following form: Since

$$\mathbb{E}(x_i)_j = P(x_i \text{ in state } j) = \sum_{h=1}^{r} P(x_i = e_j | H = e_h) P(H = e_h) = \sum_{h=1}^{r} A_{j,h} \pi_h,$$

or in other words, the first order moment satisfies $\mathbb{E}[x_u] = A^u\pi$. Similarly, the third order moment satisfies

$$\mathcal{T}_{a,b,c} = \mathbb{E}[x_u \otimes x_v \otimes x_w]_{a,b,c} = P(x_u \text{ is at state } a, x_v \text{ is at state } b, x_w \text{ is at state } c) = \sum_{h=1}^{r} \pi_h A^u_{a,h} A^v_{b,h} A^w_{c,h},$$

and the above tensor is equivalent to

$$\mathcal{T} = \sum_{h=1}^{r} \pi_h A^u_{:,h} \otimes A^v_{:,h} \otimes A^w_{:,h},$$

which is nothing but the CP decomposition of $\mathcal{T}$, of the form in (26).

5.3. *Tensor Decomposition for Mixtures of Tree Distributions.* The above tensor form for mixtures of product distributions can be easily adapted to the tree mixtures model by considering the conditional probabilities, conditioned on the relevant separator sets. We have

$$\widehat{M}^n_{u,v,w|S} = \sum_{h=1}^{r} \lambda_h \widehat{P}\left(Y_u | \mathbf{Y}_S, H = h\right) \otimes \widehat{P}\left(Y_v | \mathbf{Y}_S, H = h\right) \otimes \widehat{P}\left(Y_w | \mathbf{Y}_S, H = h\right)$$

where $\widehat{M}^n_{u,v,w|S} := \left[\widehat{P}^n\left(Y_u = i, Y_v = j, Y_w = k | \mathbf{Y}_S\right)\right]_{i,j,k}$ denotes the empirical joint probability tensor, and $\widehat{P}\left(Y_u | \mathbf{Y}_S, H = h\right)$, $\widehat{P}\left(Y_v | \mathbf{Y}_S, H = h\right)$, $\widehat{P}\left(Y_w | \mathbf{Y}_S, H = h\right)$ are the component probabilities to be estimated. Thus, our goal is to find λ, $\widehat{P}\left(Y_u | \mathbf{Y}_S, H\right)$, $\widehat{P}\left(Y_v | \mathbf{Y}_S, H\right)$ and $\widehat{P}\left(Y_w | \mathbf{Y}_S, H\right)$ so that

$$\|M^n_{u,v,w|S} - \widehat{M}^n_{u,v,w|S}\|_F^2 = \|M^n_{u,v,w|S} - \sum_{h=1}^{r} \lambda_h \widehat{P}\left(Y_u | \mathbf{Y}_S, H = h\right) \otimes \widehat{P}\left(Y_v | \mathbf{Y}_S, H = h\right) \otimes \widehat{P}\left(Y_w | \mathbf{Y}_S, H = h\right)\|_2^2$$

is minimized, and this can be solved via alternating least squares, as described previously.

5.4. *Permutation Alignment.* Since we estimate component probabilities over different triplets $\{u, v, w\}$, the components are in general permuted with respect to one another, and we need to align them. Recall that previously, we resolved this by considering u as an isolated node in the union graph, and fixing it in all the triplets, see Assumption A7 in Section 4.1. Then, the transition matrices $P(Y_u^{\{u,v,w\}} | \mathbf{Y}_S, H)$ and $P(Y_u^{\{u,v',w'\}} | \mathbf{Y}'_S, H)$ for node u, estimated using two different sets of triplets $\{u, v, w\}$ and $\{u, v', w'\}$, need to be aligned. This is done via a column permutation matrix Γ such that

$$P(Y_u^{\{u,v,w\}} | \mathbf{Y}_S, H) = P(Y_u^{\{u,v',w'\}} | \mathbf{Y}'_S, H)\Gamma.$$

Therefore,

$$P(Y_v^{\{u,v,w\}} | \mathbf{Y}_S, H) \text{ is aligned with } P(Y_{v'}^{\{u,v',w'\}} | \mathbf{Y}'_S, H)\Gamma$$

Now using Γ, the other estimated transition matrices can also be aligned:

$$P(Y_w^{\{u,v,w\}} | \mathbf{Y}_S, H) \text{ is aligned with } P(Y_{w'}^{\{u,v',w'\}} | \mathbf{Y}'_S, H)\Gamma.$$

6. Experiments. In this section experimental results are presented on both synthetic and real data. We estimate the graph using proposed algorithm and compare the performance of our method with EM [37].

Procedure 2 $[\{\widehat{P}(Y_w, \mathbf{Y}_S|H=h), \widehat{\boldsymbol{\pi}}_H(h)\}_h] \leftarrow \mathsf{TensorDecom}(u, v, w; S, \mathbf{y}^n, r)$ for finding the components of an r-component mixture from $\mathbf{y}^n$ samples at w, given witnesses u, v and separator S on graph $\widehat{G}^n$.

```
Let $\widehat{M}^n_{u,v,w,\{S;q\}} := [\widehat{P}^n(Y_u = i, Y_v = j, Y_w = k, \mathbf{Y}_S = q)]_{i,j,k}$ where $\widehat{P}^n$ is the empirical distribution computed using samples $\mathbf{y}^n$.
for $q \in \mathcal{Y}^{|S|}$ do
    Obtain $\left[\widehat{M}_{u|H,\{S;k\}}, \widehat{M}_{v|H,\{S;k\}}, \widehat{M}_{w|H,\{S;k\}}\right] \leftarrow \mathsf{CP-als}(\widehat{M}^n_{u,v,w,\{S;q\}})$.
    Obtain $\widehat{\boldsymbol{\pi}}_{H|S;q}$
end for
Output $\{\widehat{P}(Y_w, \mathbf{Y}_S|H=h), \widehat{\boldsymbol{\pi}}_H(h)\}_{h\in[r]}$.
```

6.1. *Tools.* Comprehensive results based on the normalized edit distances and log-likelihood scores between the estimated and the true graphs are delivered. Proposed algorithm and EM algorithm are implemented in MATLAB. "UGM"[7] package is used for sampling from the Model. A.Ihler's "factor" class is used when calculating empirical distributions[8]. "Tensor Toolbox" is used for tensor decomposition [9]. K-shortest path algorithm is also used[9].

Synthetic data. We generate samples from a mixture over two different randomly generated trees ($r = 2$) with mixing weights $\pi = [0.7, 0.3]$. Each mixture component is generated from the standard Potts model on $p = 60$ nodes, where the node variables are ternary ($d = 3$), and the number of samples $n \in [10^3, 10^4]$. The joint distribution of nodes in each mixture component is given by

$$P\left(X|H=h\right) \propto \exp\left[\sum_{(i,j)\in G} J_{i,j;h}(\mathbb{I}(Y_i = Y_j) - 1) + \sum_{i \in V} K_{i;h} Y_i,\right]$$

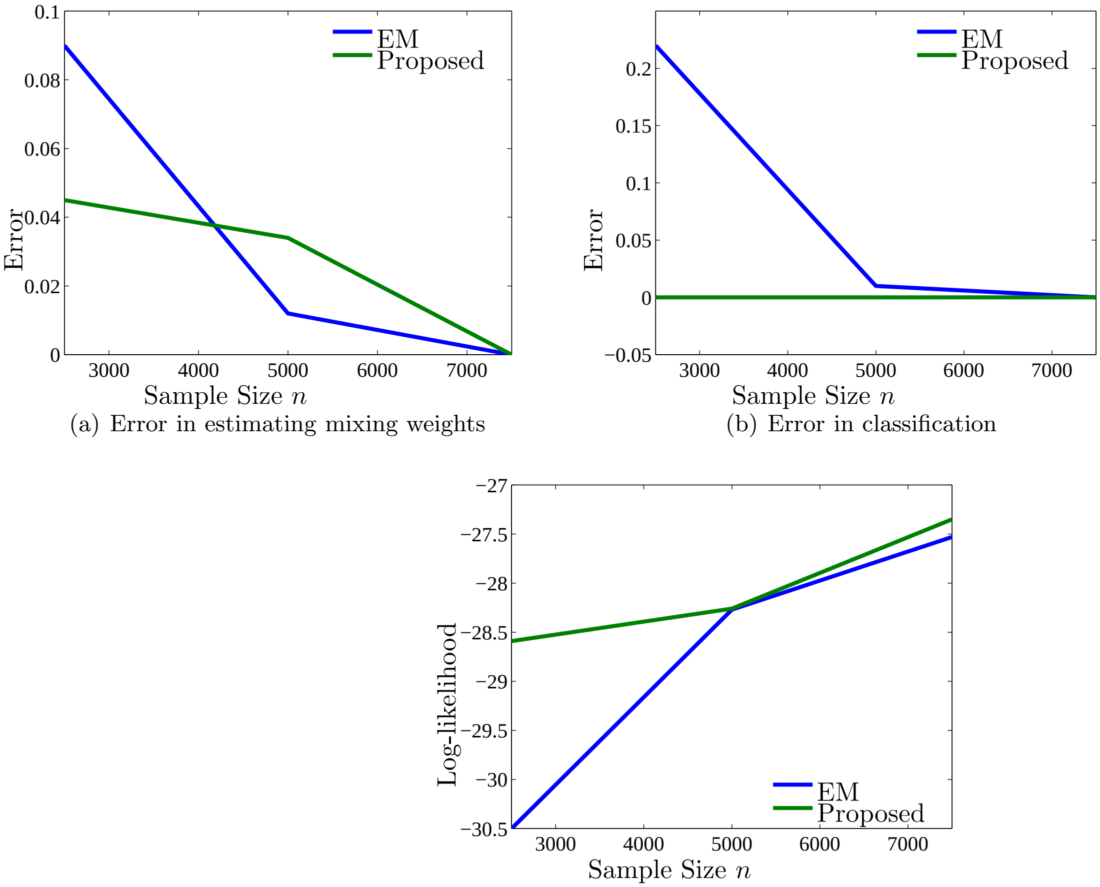
where $\mathbb{I}$ is the indicator function and $\mathbf{J}_h := \{J_{i,j;h}\}$ are the edge potentials in the model. Based on the generated graph topologies, we generate the potential matrices $\mathbf{J}_1$ and $\mathbf{J}_2$ whose sparsity pattern corresponds to those of the two tree components. By convention, we set diagonal elements $\mathbf{J}_h(i,i) = 0$ for all $i \in V$. For the first component ($H = 1$), the edge potentials $\mathbf{J}_1$ are chosen uniformly from $[5, 5.05]$, while for the second component ($H = 2$), $\mathbf{J}_2$ are chosen from $[0.5, 0.55]$. We refer to the first component as *strong* and the second as *weak* since the correlations vary widely between the two models due to the choice of parameters. The *node potentials* are all set to zero ($K_{i;h} = 0$) except at the isolated node u_* in the union graph. The samples are generated via Gibbs sampling with a burn-in period of 10^7 and a step size of 10 for selecting independent samples.

Synthetic results. The performance of the proposed method is compared with EM. We consider 10 random initializations of EM and run it to convergence. Errors in estimating the mixing weights between EM and proposed method are compared in Fig 2(a). We find that the proposed method excels EM over smaller sample size. We also evaluate the classification errors in Fig 2(b) and find that our method has superior classification performance over EM. In Fig 2(c), we plot the normalized log-likelihood values, and find that the overall likelihoods are comparable. Our method results in higher likelihoods especially over smaller sample size. This is especially encouraging since the objective of EM is to locally optimize the overall likelihood, while our method is a moment-based estimator and does not use likelihood as the optimizing criterion. In Figs 3(a) and 3(b), we

[7] UGM is at http://www.di.ens.fr/~mschmidt/Software/UGM.html

[8] http://sli.ics.uci.edu/Code/Matlab-Factor

[9] http://www.mathworks.com/matlabcentral/fileexchange/32513



(a) Error in estimating mixing weights

(b) Error in classification

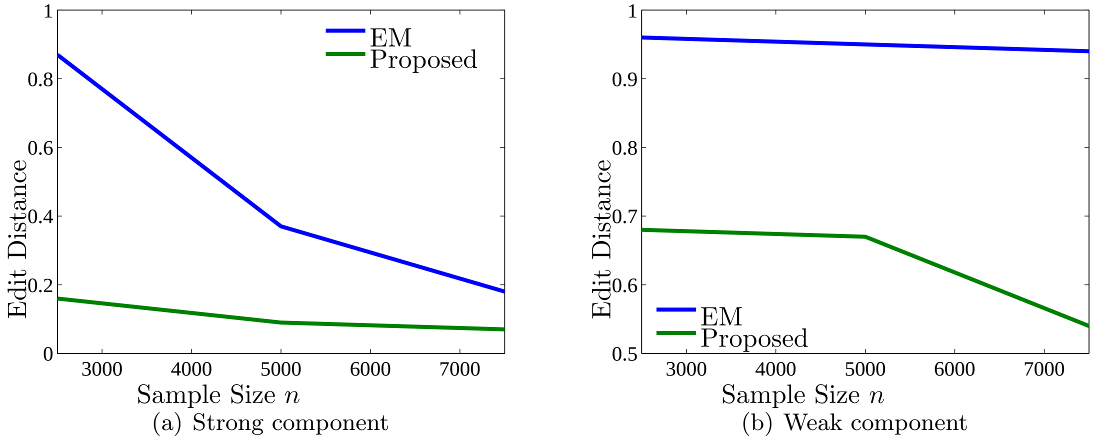
(c) Overall log-likelihood

Fig 2. *Performance of the proposed method and EM method for a tree mixture with two components.*

plot the normalized edit distances, which are evaluation measures for structures estimation on the tree components. Our algorithm has significantly superior performance with respect to the edit distances. In fact, EM never manages to recover the structure of one of the components (which turns out to be the component with weak correlations). On the other hand, our algorithm recovers both the tree structures. Intuitively, this is because EM uses the overall likelihood as criterion for tree selection in the two components. Under the above choice of parameters, the weak component has a much lower contribution to the overall likelihood, and thus, EM is unable to recover it. Thus, optimizing overall likelihood of the mixture, does not always lead to good structure estimation.

We also observe in Fig 4(a) and Fig 4(b), that our proposed method has superior performance in terms of conditional likelihood for both the components (conditional likelihood is evaluated by conditioning on the true label of the sample). Our method also has significantly faster running times than EM. For running 10 random initializations of EM algorithm to convergence, the running time of EM algorithm is approximately 10^4 secs on average while proposed algorithm takes approximately 10^2 secs when running on 8 threads with Intel(R) Core(TM)i7-3770K CPU@3.50GHz.

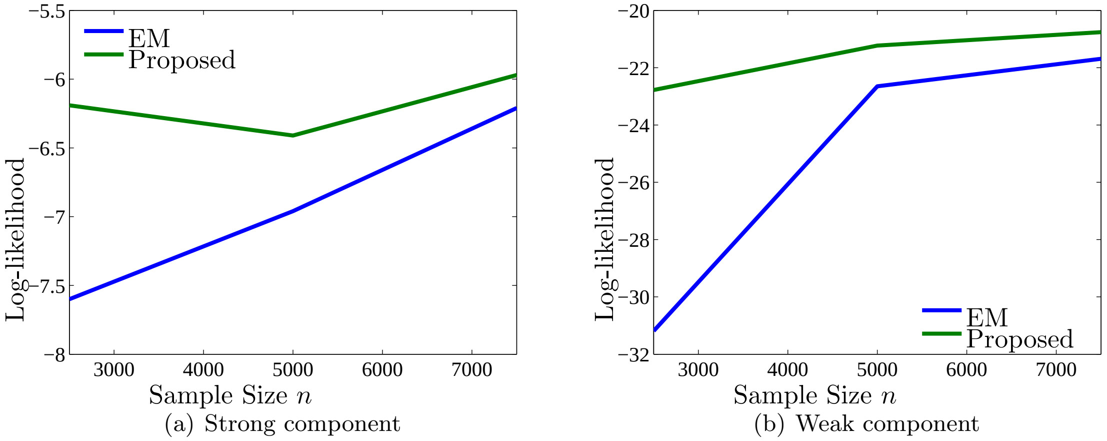
SPLICE data. DNA SPLICE-junctions is a popular data set in molecular biology [24]. Each SPLICE-junction is quantitatively represented by a sequence of DNA bases with length 60. In other words, there are 60 features (variables) for each SPLICE-junction. A total number of 3175



(a) Strong component

(b) Weak component

FIG 3. *Edit distances under the proposed method and EM method for a tree mixture with two components. Strong component refers to the component with strong correlations and vice versa.*



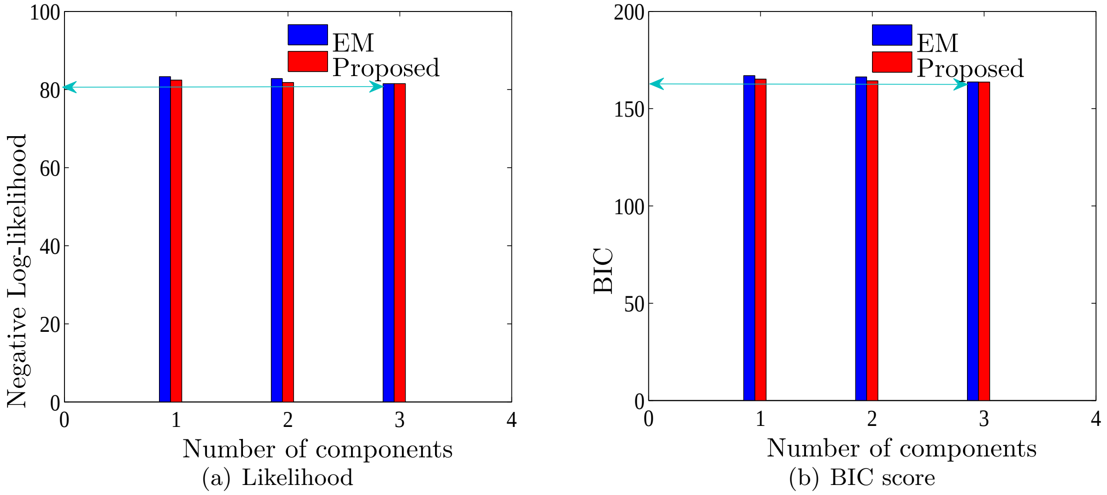
(a) Strong component

(b) Weak component

FIG 4. *Component likelihoods under the proposed method and EM method for a tree mixture with two components.*

SPLICE-junctions are included in this data set. The SPLICE-junctions can be divided into 3 types, one is EI, one is IE and the other is neither. The data set additionally provides the ground truth of which one of the three categories each SPLICE junction belongs to as the labels. We estimate the mixing weights of the three types of SPLICE-junctions and uncover a tree structure approximation under each type of SPLICE junctions. Although we have no prior information about how many SPLICE-junction types are there or what proportion each type takes, we undertake model selection and compare some model scores (explained below). We hold out 1035 samples randomly as test set, and train the model on a training set of 2000 samples. Results are averaged over 10 trials consisting of different training and test sets. The stopping threshold for EM is set to be 0.05.

SPLICE results. As an unsupervised approach for learning mixture models, both spectral and EM algorithm entail model selection. We model the sequence of DNA bases with r types of SPLICE-junction types and evaluate the penalized likelihood scores for different values of r, and select r that maximizes the score. Since the 60 observed DNA bases are categorical variables with $d = 4$ states, corresponding to the 4 DNA bases (C,A,G and T), the model space we select from spans $r \in \{1, 2, 3\}$



(a) Likelihood

(b) BIC score

FIG 5. *Overall negative likelihood and BIC score estimation of the proposed method and EM method over different r's in SPLICE data.*

(recall we require $r < d$). Normalized negative log-likelihood scores are shown in Fig 5(a). We use the normalized BIC score as the model selection criterion. Normalized BIC score is given by

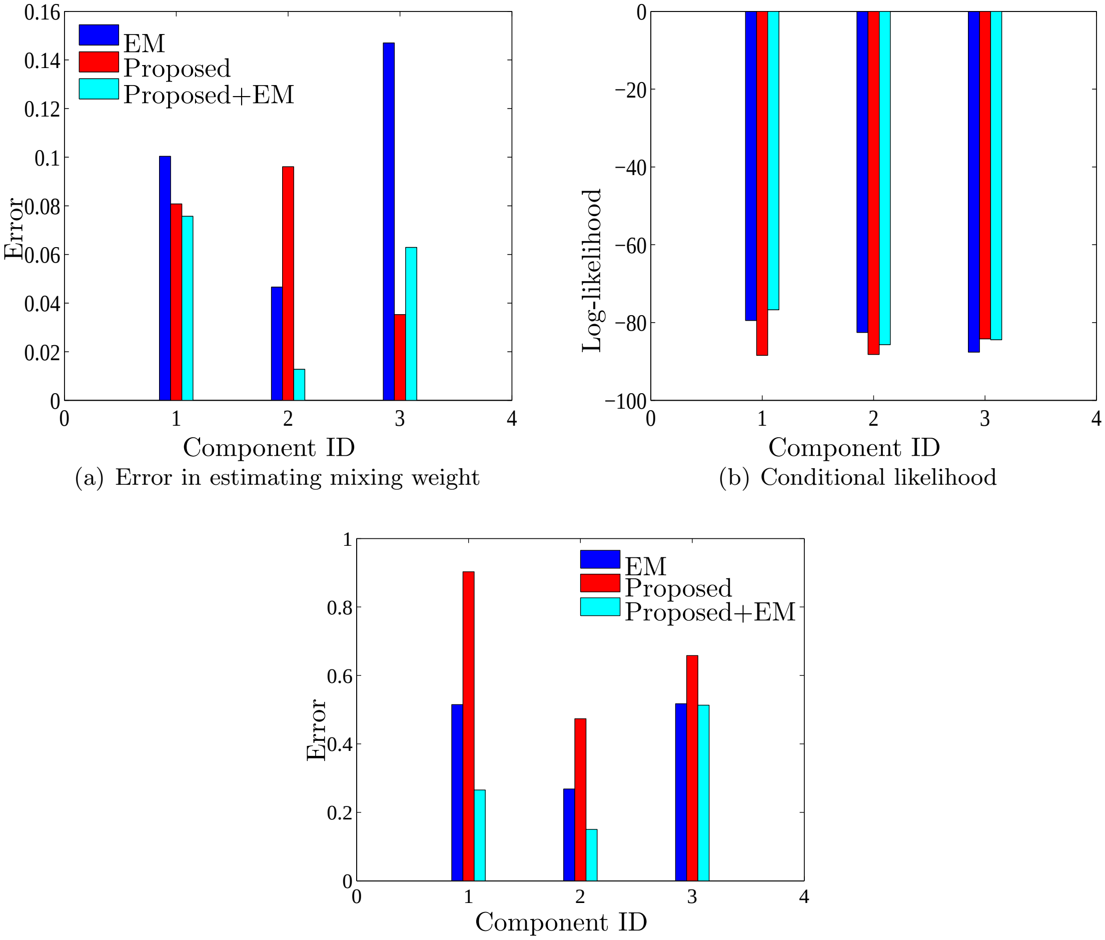
$$\text{normalized-}BIC := \frac{-2\ln(L) + k\ln(n)}{n},$$

where n is the sample size, L is the likelihood, and k is the number of model parameters. Thus, the model with the lower BIC score is the preferred one. According to Fig 5(b), $r = 3$ is selected, which agrees with the ground truth. Then, we evaluate the mixing weight error, classification error and conditional likelihood under $r = 3$. In terms of estimating the components, our method is superior to EM, as seen by the error in estimating the mixing weights (Fig 6(a)) and the component likelihoods (Fig 6(b)) under both the methods. On the other hand, EM does a better job in density estimation, as seen by the classification error in recovering the labels of each SPLICE-junction sample on the held-out dataset (Fig 6(c))[10].

This presents a scenario to exploit the advantages of both the methods: we initialize EM with the estimate from our spectral approach, termed as Proposed+EM method. We observe a big improvement, both in component parameter estimation, as well as classification error. Note that for the (pure) EM approach, we use multiple (10) restarts, and select the best value, and yet, initializing with our method yields a big improvement over multiple random initializations. This is especially relevant in high dimensions, since it is hard to find a good initialization point for EM. The intuition behind this result is that our moment-based spectral estimator can be improved locally by running EM, and this agrees with the classical result that taking a single step of Newton-Ralphson on the likelihood function starting from a moment-based estimate can lead to asymptotically efficient estimation [34].

The above experimental results confirm our theoretical analysis and suggest the advantages of our basic technique over more common approaches. Our method provides a point of tractability in the spectrum of probabilistic models, and extending beyond the class we consider here is a promising direction of future research.

[10] Although the paper [37] uses this dataset, the results they present are for supervised case, and therefore, they differ from results presented here.



(a) Error in estimating mixing weight

(b) Conditional likelihood

(c) Classification error

Fig 6. *Performance estimation of the proposed method, EM method and propose-initialized EM under $r = 3$ in SPLICE data.*

7. Conclusion. In this paper, we considered learning tree approximations of graphical model mixtures. We proposed novel methods which combined techniques used previously in graphical model selection, and in learning mixtures of product distributions. We provided provable guarantees for our method, and established that it has polynomial sample and computational complexities in the number of nodes p, number of mixture components r and cardinality of each node variable d. Our guarantees are applicable for a wide family of models. In future, we plan to investigate learning mixtures of continuous models, such as Gaussian mixture models.

Acknowledgements. The authors thank the co-editor Peter Bühlmann (ETH), associate editor and the anonymous reviewers for comments which substantially improved the paper. The authors also thank helpful discussions with Marina Meila (UW) and Alan Willsky (MIT).

APPENDIX A: IMPLEMENTATION OF SPECTRAL DECOMPOSITION METHOD

Overview of the algorithm: . We provide the procedure in Algorithm 3. The algorithm computes the pairwise statistic of each node pair $a, b \in V \setminus \{u_*\}$, where u_* is the reference node which is isolated in $\widehat{G}_\cup$, the union graph estimate obtained using Algorithm 1. The spectral decomposition is carried out on the triplet $\{u_*, c, (a,b); \{S = k\}\}$, where c is any node not in the neighborhood of a or b in graph $\widehat{G}_\cup$. Set $S \subset V \setminus \{a, b, u_*\}$ is separates a, b from c in $\widehat{G}_\cup$. See Fig.7. We fix the configuration of the separator set to $\mathbf{Y}_S = k$, for each $k \in \mathcal{Y}^{|S|}$, and consider the empirical distribution of n samples, $\widehat{P}^n(Y_{u_*}, Y_a, Y_a, Y_c, \{\mathbf{Y}_S = k\})$. Upon spectral decomposition, we obtain the mixture components $\widehat{P}^{\text{spect}}(Y_a, Y_b, \mathbf{Y}_S|H = h)$ for $h \in [r]$. We can then employ the estimated pairwise marginals to find the Chow-Liu tree approximation $\{\widehat{T}_h\}_h$ for each mixture component. This routine can also be adapted to estimate the individual Markov graphs $\{G_h\}_h$ and is described briefly in Section A.1. Also, if the underlying model is a tree mixture, we can slightly modify the algorithm and obtain better guarantees, and we outline it in Section A.1.

Procedure 3 FindMixtureComponents$(\mathbf{y}^n, \widehat{G}; r)$ for finding the tree-approximations of the components $\{P(\mathbf{y}|H = h)\}_h$ of an r-component mixture using samples $\mathbf{y}^n$ and graph $\widehat{G}$, which is an estimate of the graph $G_\cup := \cup_{h=1}^r G_h$ obtained using Algorithm 1.

```
$\widehat{M}^n_{A,B,\{C;k\}} := [P(\mathbf{Y}_A = i, \mathbf{Y}_B = j, \mathbf{Y}_C = k]_{i,j}$ denotes the empirical joint probability matrix estimated using samples $\mathbf{y}^n$, where $A \cap B \cap C = \emptyset$. Let $\mathcal{S}(A, B; G_\cup)$ be a minimal vertex separator separating $A$ and $B$ in graph $\widehat{G}_\cup$.
Choose a uniformly random orthonormal basis $\{z_1, \ldots, z_r\} \in \mathbb{R}^r$. Let $Z \in \mathbb{R}^{r\times r}$ be a matrix whose $l^{\text{th}}$ row is $\mathbf{z}_l^\top$.
Let $u_* \in V$ be isolated from all the other nodes in graph $\widehat{G}$. Otherwise declare fail.
for $a, b \in V \setminus \{u_*\}$ do
    Let $c \notin \mathcal{N}(a; \widehat{G}) \cup \mathcal{N}(b; \widehat{G})$ (if no such node is found, go to the next node pair). $S \leftarrow \mathcal{S}((a,b), c; \widehat{G})$.
    $\{\widehat{P}^{\text{spect}}(Y_a, Y_b, \mathbf{Y}_S|H = h)\}_h \leftarrow$ SpecDecom$(u_*, c, (a,b); S, \mathbf{y}^n, r, Z)$.
end for
for $h \in [r]$ do
    $\left[\widehat{T}_h, \{\widehat{P}^{\text{tree}}(Y_a, Y_b|H = h)\}_{(a,b)\in\widehat{T}_h}\right] \leftarrow$ ChowLiuTree $\left(\{\widehat{P}^{\text{spect}}(Y_a, Y_b|H = h)\}_{a,b\in V\setminus\{u_*\}}\right)$.
end for
Output $\left[\widehat{\boldsymbol{\pi}}^{\text{spect}}_H(h),\ \widehat{T}_h,\ \left\{\widehat{P}^{\text{tree}}(Y_a, Y_b|H = h) : (a,b) \in \widehat{T}_h\right\}\right]_{h\in[r]}$.
```

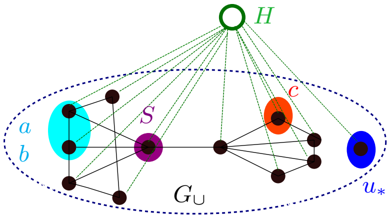


Fig 7. *By conditioning on the separator set S on the union graph $G_\cup$, we have a mixture of product distribution with respect to nodes $\{u_*, c, (a,b)\}$, i.e., $Y_{u_*} \perp\!\!\!\perp Y_c \perp\!\!\!\perp Y_{a,b}|\mathbf{Y}_S, H$.*

A.1. Discussion and Extensions.

Simplification for Tree Mixtures ($G_h = T_h$): . We can simplify the above method by limiting to tree approximations which are subgraphs of graph $G_\cup$. This procedure coincides with the original method when all the component Markov graphs $\{G_h\}_h$ are trees, i.e., $G_h = T_h$, $h \in [r]$. This is because in this case, the Chow-Liu tree coincides with $T_h \subset G_\cup$ (under exact statistics). This implies that we need to compute pairwise marginals *only* over the edges of $G_\cup$ using SpecDecom routine, instead of over all the node pairs, and the ChowLiuTree procedure computes a maximum weighted spanning tree over $G_\cup$, instead of the complete graph. This leads a slight improvement of

Procedure 4 $[\{\widehat{P}(Y_w, \mathbf{Y}_S|H = h), \widehat{\boldsymbol{\pi}}_H(h)\}_h] \leftarrow \mathsf{SpecDecom}(u, v, w; S, \mathbf{y}^n, r, Z)$ for finding the components of an r-component mixture from $\mathbf{y}^n$ samples at w, given witnesses u, v and separator S on graph $\widehat{G}^n$.

```
Let $\widehat{M}^n_{u,v,\{S;k\}} := [\widehat{P}^n(Y_u = i, Y_v = j, \mathbf{Y}_S = k)]_{i,j}$ where $\widehat{P}^n$ is the empirical distribution computed using samples $\mathbf{y}^n$. Similarly, let $\widehat{M}^n_{u,v,\{S;k\},\{w;l\}} := [\widehat{P}^n(Y_u = i, Y_v = j, \mathbf{Y}_S = k, Y_w = l)]_{i,j}$. For a vector $\boldsymbol{\lambda}$, let $\text{Diag}(\boldsymbol{\lambda})$ denote the corresponding diagonal matrix.
for $k \in \mathcal{Y}^{|S|}$ do
  Choose $U_u$ as the set of top $r$ left orthonormal singular vectors of $\widehat{M}^n_{u,v,\{S;k\}}$ and $V_v$ as the right singular vectors.
  Similarly for node $w$, let $U_w$ be the top $r$ left orthonormal singular vectors of $\widehat{M}^n_{w,u,\{S;k\}}$.
  for $l \in [r]$ do
    $\mathbf{m}_l \leftarrow U_w \mathbf{z}_l$, $A \leftarrow U_u^\top \widehat{M}^n_{u,v,\{S;k\}} V_v$ and $B_l \leftarrow U_u^\top \left(\sum_q m_l(q) \widehat{M}^n_{u,v,\{S;k\},\{w;q\}}\right) V_v$.
    if $A$ is invertible (Fail Otherwise) then
      $C_l \leftarrow B_l A^{-1}$. $\text{Diag}(\boldsymbol{\lambda}^{(l)}) \leftarrow R^{-1} C_l R$. {Find $R$ which diagonalizes $C_l$ for the first triplet. Use the same matrix $R$ for all other triplets.}
    end if
  end for
  Form the matrix from the above eigenvalue computations: $\Lambda = [\boldsymbol{\lambda}^{(1)}|\boldsymbol{\lambda}^{(2)}|\ldots, \boldsymbol{\lambda}^{(r)}]$
  Obtain $\widehat{M}_{w|H,\{S;k\}} \leftarrow U_w Z^{-1} \Lambda^\top$. Similarly obtain $\widehat{M}_{v|H,\{S;k\}}$.
  Obtain $\widehat{\boldsymbol{\pi}}_H$: $\widehat{M}^n_{v,w,\{S;k\}} = \widehat{M}_{v|H,\{S;k\}} \text{Diag}(\widehat{\boldsymbol{\pi}}_{H|\{S;k\}})(\widehat{M}_{w|H,\{S;k\}})^\top \widehat{P}^n(\mathbf{Y}_S = k)$.
end for
Output $\{\widehat{P}(Y_w, \mathbf{Y}_S|H = h), \widehat{\boldsymbol{\pi}}_H(h)\}_{h\in[r]}$.
```

Procedure 5 $[\widehat{T}, \{\widehat{P}^{\text{tree}}(Y_a, Y_b)\}_{(a,b)\in\widehat{T}}] \leftarrow \mathsf{ChowLiuTree}(\{\widehat{P}(Y_a, Y_b)\}_{a,b\in V\setminus\{u_*\}}$ for finding a tree approximation given the pairwise statistics.

```
for $a, b \in V \setminus \{u_*\}$ do
  Compute mutual information $\widehat{I}(Y_a; Y_b)$ using $\widehat{P}(Y_a, Y_b)$.
end for
$\widehat{T} \leftarrow \mathsf{MaxWtTree}(\{\widehat{I}(Y_a; Y_b)\})$ is max-weight spanning tree using edge weights $\{\widehat{I}(Y_a; Y_b)\}$.
for $(a, b) \in \widehat{T}$ do
  $\widehat{P}^{\text{tree}}(Y_a, Y_b) \leftarrow \widehat{P}(Y_a, Y_b)$.
end for
```

sample complexity, and we note it in the remarks after Theorem 2.

Estimation of Component Markov Graphs $\{G_h\}_h$*: .* We now note that we can also estimate the component Markov graphs $\{G_h\}$ using the spectral decomposition routines and we briefly describe it below. Roughly, we can do a suitable conditional independence test on the estimated statistics $\widehat{P}^{\text{spect}}(\mathbf{Y}_{\mathcal{N}[a;\widehat{G}_\cup]}|H = h)$ obtained from spectral decomposition, for each node neighborhood $\mathcal{N}[a; \widehat{G}_\cup]$, where $a \in V \setminus \{u_*\}$ and $\widehat{G}_\cup$ is an estimate of $G_\cup := \cup_{h\in[r]} G_h$. We can estimate these statistics by selecting a suitable set of witnesses $C := \{c_1, c_2, \ldots, \}$ such that $\mathcal{N}[a]$ can be separated from C in $\widehat{G}_\cup$. We can employ Procedure $\mathsf{SpecDecom}$ on this configuration by using a suitable separator set and then doing a threshold test on the estimated component statistics $\widehat{P}^{\text{spect}}$: if for each $(a, b) \in \widehat{G}_\cup$, the following quantity

$$\min_{k,l\in\mathcal{Y}} \|\widehat{P}^{\text{spect}}(Y_a|Y_b = k, \mathbf{Y}_{\mathcal{N}(a)\setminus b} = \mathbf{y}, H = h) - \widehat{P}^{\text{spect}}(Y_a|Y_b = l, \mathbf{Y}_{\mathcal{N}(a)\setminus b} = \mathbf{y}, H = h)\|_1,$$

is below a certain threshold, for some $\mathbf{y} \in \mathcal{Y}^{|\mathcal{N}(a)\setminus b|}$, then it is removed from $\widehat{G}_\cup$, and we obtain $\widehat{G}_h$ in this manner. A similar test was used for graphical model selection (i.e., not a mixture model) in [13]. We note that we can obtain sample complexity results for the above test, on lines of the analysis in Section 4.1 and this method is efficient when the maximum degree in $G_\cup$ is small.

APPENDIX B: EXTENSION TO GRAPHS WITH SPARSE LOCAL SEPARATORS

B.1. Graphs with Sparse Local Separators. We now extend the analysis to the setting where the graphical model mixture has the union graph $G_\cup$ with sparse local separators, which is a weaker criterion than having sparse exact separators. We now provide the definition of a local separator. For detailed discussion, refer to [6].

For $\gamma \in \mathbb{N}$, let $B_\gamma(i; G)$ denote the set of vertices within distance γ from i with respect to graph G. Let $F_{\gamma,i} := G(B_\gamma(i))$ denote the subgraph of G spanned by $B_\gamma(i; G)$, but in addition, we retain the nodes not in $B_\gamma(i)$ (and remove the corresponding edges).

Definition 1 (γ-Local Separator). *Given a graph G, a* γ-local separator $\mathcal{S}_{\rm local}(i, j; G, \gamma)$ *between i and j, for $(i, j) \notin G$, is a* minimal *vertex separator*[11] *with respect to the subgraph $F_{\gamma,i}$. In addition, the parameter γ is referred to as the* path threshold *for local separation. A graph is said to be η-locally separable, if*

$$\max_{(i,j)\notin G} |\mathcal{S}_{\rm local}(i, j; G, \gamma)| \leq \eta. \tag{28}$$

A wide family of graphs possess the above property of sparse local separation, i.e., have a small η. In addition to graphs considered in the previous section, this additionally includes the family of locally tree-like graphs (including sparse random graphs), bounded degree graphs, and augmented graphs, formed by the union of a bounded degree graph and a locally tree-like graph (e.g. small-world graphs). For detailed discussion, refer to [6].

B.2. Regime of Correlation Decay. We consider learning mixtures of graphical models Markov on graphs with sparse local separators. We assume that these models are in the regime of correlation decay, which makes learning feasible via our proposed methods.

We formally define the notion of correlation decay below[12] and incorporate it to provide learning guarantees. See [50] for details.

Let $P(Y_v|\mathbf{Y}_A; G)$ denote the conditional distribution of node v given a set $A \subset V \setminus \{v\}$ under model P with Markov graph G. For some subgraph $F \subset G$, let $P(Y_v|\mathbf{Y}_A; F)$ denote the conditional distribution on corresponding to a graphical model Markov on subgraph F instead of G, i.e., by setting the potentials of edges (and hyperedges) in $G \setminus F$ to zero. For any two sets $A_1, A_2 \subset V$, let $\mathrm{dist}(A_1, A_2) := \min_{u\in A_1, v\in A_2} \mathrm{dist}(u, v)$ denote the minimum graph distance. Let $B_l(v)$ denote the set of nodes within graph distance l from node v and $\partial B_l(v)$ denote the boundary nodes, i.e., exactly at l from node v. Let $F_l(v; G) := G(B_l(v))$ denote the induced subgraph on $B_l(v; G)$. For any vectors $\mathbf{a}, \mathbf{b}$, let $\|\mathbf{a} - \mathbf{b}\|_1 := \sqrt{\sum_i |a(i) - b(i)|}$ denote the ℓ_1 distance between them.

Definition 2 (Correlation Decay). *A graphical model P Markov on graph $G = (V, E)$ with p nodes is said to exhibit correlation decay with a non-increasing rate function $\zeta(\cdot) > 0$ if for all $l, p \in \mathbb{N}$,*

$$\|P(Y_v|\mathbf{Y}_A = \mathbf{y}_A; G) - P(Y_V|\mathbf{Y}_A = \mathbf{y}_A; F_l(i; G))\|_1 = \zeta(\mathrm{dist}(A, \partial B_l(i))), \quad \forall\, v \in V,\ A \subset V \setminus \{v\}. \tag{29}$$

[11] A minimal separator is a separator of smallest cardinality.

[12] We slightly modify the definition of correlation decay compared to the usual notion by considering models on different graphs, where one is an induced subgraph of the neighborhood of the other graph, instead of models with different boundary conditions.

Remark: For the class of Ising models (binary variables), the regime of correlation decay can be explicitly characterized, in terms of the maximum edge potential of the model. When the maximum edge potential is below a certain threshold, the model is said to be in the regime of correlation decay. The threshold that can be explicitly characterized for certain graph families. See [6] for derivations.

B.3. Rank Test Under Local Separation. We now provide sufficient conditions for the success of $\mathsf{RankTest}(\mathbf{y}^n; \xi_{n,p}, \eta, r)$ in Algorithm 1. Note that the crucial difference compared to the previous section is that η refers to the bound on local separators in contrast to the bound on exact separators. This can lead to significant reduction in computational complexity of running the rank test for many graph families, since the complexity scales as $O(p^{\eta+2}d^3)$ where p is the number of nodes and d is the cardinality of each node variable.

(B1) **Number of Mixture Components:** The number of components r of the mixture model and dimension d of each node variable satisfy

$$d > r. \tag{30}$$

The mixing weights of the latent factor H are assumed to be strictly positive

$$\pi_H(h) := P(H = h) > 0, \quad \forall\ h \in [r].$$

(B2) **Constraints on Graph Structure:** Recall that $G_\cup = \cup_{h=1}^r G_h$ denotes the union of the Markov graphs of the mixture components and we assume that $G_\cup$ is η-locally separable according to Definition 1, i.e., for the chosen path threshold $\gamma \in \mathbb{N}$, we assume that

$$|\mathcal{S}_{\text{local}}(u, v; G_\cup, \gamma)| \le \eta = O(1), \quad \forall (u, v) \notin G_\cup.$$

(B3) **Rank Condition:** We assume that the matrix $M_{u,v,\{S;k\}}$ in (4) has rank strictly greater than r when the nodes u and v are neighbors in graph $G_\cup = \cup_{h=1}^r G_h$ and the set satisfies $|S| \le \eta$. Let $\rho_{\min}$ denote

$$\rho_{\min} := \min_{\substack{(u,v)\in G_\cup, |S|\le\eta \\ S\subset V\setminus\{u,v\}}} \max_{k\in\mathcal{Y}^{|S|}} \sigma_{r+1}\left(M_{u,v,\{S;k\}}\right) > 0. \tag{31}$$

(B4) **Regime of Correlation Decay:** We assume that all the mixture components $\{P(\mathbf{y}|H = h; G_h)\}_{h\in[r]}$ are in the regime of correlation decay according to Definition 2 with rate functions $\{\zeta_h(\cdot)\}_{h\in[r]}$. Let

$$\zeta(\gamma) := 2\sqrt{d} \max_{h\in[r]} \zeta_h(\gamma). \tag{32}$$

We assume that the minimum singular value $\rho_{\min}$ in (14) and $\zeta(\gamma)$ above satisfy $\rho_{\min} > \zeta(\gamma)$.

(B5) **Choice of threshold ξ:** For $\mathsf{RankTest}$ in Algorithm 1, the threshold ξ is chosen as

$$\xi := \frac{\rho_{\min} - \zeta(\gamma)}{2} > 0,$$

where $\zeta(\gamma)$ is given by (32) and $\rho_{\min}$ is given by (14), and γ is the path threshold for local separation on graph $G_\cup$.

(B6) **Number of Samples:** Given an $\delta > 0$, the number of samples n satisfies

$$n > n_{\mathrm{LRank}}(\delta; p) := \max\left(\frac{1}{t^2}\left(2\log p + \log\delta^{-1} + \log 2\right), \left(\frac{2}{\rho_{\min} - \zeta(\gamma) - t}\right)^2\right), \tag{33}$$

where p is the number of nodes, for some $t \in (0, \rho_{\min} - \zeta(\gamma))$.

The above assumptions (B1)–(B6) are comparable to assumptions (A1)–(A5) in Section 3.1.1. The conditions on r and d in (A1) and (B1) are identical. The conditions (A2) and (B2) are comparable, with the only difference being that (A2) assumes bound on exact separators while (B2) assumes bound on local separators, which is a weaker criterion. Again, the conditions (A3) and (B3) on the rank of matrices for neighboring nodes are identical. The condition (B4) is an additional condition regarding the presence of correlation decay in the mixture components. This assumption is required for approximate conditional independence under conditioning with local separator sets in each mixture component. In addition, we require that $\zeta(\gamma) < \rho_{\min}$. In other words, the threshold γ on path lengths considered for local separation should be large enough (so that the corresponding value $\zeta(\gamma)$ is small). (B5) provides a modified threshold to account for distortion due to the use of local separators and (B6) provides the modified sample complexity.

B.3.1. *Success of Rank Tests.* We now provide the result on the success of recovering the union graph $G_\cup := \cup_{h=1}^r G_h$ for η-locally separable graphs.

THEOREM 4 (Success of Rank Tests). *The* $\mathsf{RankTest}(\mathbf{y}^n; \xi, \eta, r)$ *outputs the correct graph* $G_\cup := \cup_{h=1}^r G_h$*, which is the union of the component Markov graphs, under the assumptions (B1)–(B6) with probability at least* $1 - \delta$.

Proof: See Appendix C. □

B.4. Results for Spectral Decomposition Under Local Separation. The procedure $\mathsf{FindMixtureComponents}(\mathbf{y}^n, \widehat{G}; r)$ in Algorithm 3 can also be implemented for graphs with local separators, but with the modification that we use local separators $\mathcal{S}_{\mathrm{local}}((a,b), c; \widehat{G})$, as opposed to exact separators, between nodes a, b and c under consideration. We prove that this method succeeds in estimating the pairwise marginals of the component model under the following set of conditions. We find that there is additional distortion introduced due to the use of local separators in $\mathsf{FindMixtureComponents}$ as opposed to exact separators.

B.4.1. *Assumptions.* In addition to the assumptions (B1)–(B6), we impose the following constraints to guarantee the success of estimating the various mixture components.

(B7) **Full Rank Views of the Latent Factor:** For each node pair $a, b \in V$, and any subset $S \subset V \setminus \{a, b\}$ with $|S| \leq 2\eta$ and $k \in \mathcal{Y}^{|S|}$, the probability matrix $M_{(a,b)|H,\{S;k\}} := [P(\mathbf{Y}_{a,b} = i | H = j, \mathbf{Y}_S = k)]_{i,j} \in \mathbb{R}^{d^2 \times r}$ has rank r.

(B8) **Existence of an Isolated Node:** There exists a node $u_* \in V$ which is isolated from all other nodes in $G_\cup = \cup_{h=1}^r G_h$, i.e.

$$Y_{u_*} \perp\!\!\!\perp \mathbf{Y}_{V \setminus u_*} | H. \tag{34}$$

(B9) **Spectral Bounds and Random Rotation Matrix:** Refer to various spectral bounds used to obtain $K(\delta; p, d, r)$ in Appendix D.3, where $\delta \in (0, 1)$ is fixed. Further assume that the rotation matrix $Z \in \mathbb{R}^{r \times r}$ in $\mathsf{FindMixtureComponents}$ is chosen uniformly over the Stiefel manifold $\{Q \in \mathbb{R}^{r \times r} : Q^\top Q = I\}$.

(B10) **Number of Samples:** For fixed $\delta \in (0,1)$ and $\epsilon > \epsilon_0$, the number of samples satisfies

$$n > n_{\text{local-spect}}(\delta, \epsilon; p, d, r) := \frac{4K^2(\delta; p, d, r)}{(\epsilon - \epsilon_0)^2}, \tag{35}$$

where

$$\epsilon_0 := 2K'(\delta; p, d, r)\zeta(\gamma), \tag{36}$$

and $K'(\delta; p, d, r)$ and $K(\delta; p, d, r)$ are defined in (63) and (64), and $\zeta(\gamma)$ is given by (32).

The assumptions (B7)-(B9) are identical with (A6)-(A8). In (B10), the bound on the number of samples is slightly worse compared to (A9), depending on the correlation decay rate function $\zeta(\gamma)$. Moreover, the perturbation ϵ now has a lower bound ϵ_0 in (36), due to the use of local separators in contrast to exact vertex separators. As before, below, we impose additional conditions in order to obtain the correct Chow-Liu tree approximation T_h of each mixture component $P(\mathbf{y}|H = h)$.

(B11) **Separation of Mutual Information:** Let T_h denote the Chow-Liu tree corresponding to the model $P(\mathbf{y}|H = h)$ when exact statistics are input[13] and let

$$\vartheta := \min_{h\in[r]} \min_{(a,b)\notin T_h} \min_{(u,v)\in \text{Path}(a,b;T_h)} \left(I(Y_u, Y_v|H = h) - I(Y_a, Y_b|H = h)\right), \tag{37}$$

where $\text{Path}(a, b; T_h)$ denotes the edges along the path connecting a and b in T_h.

(B12) **Constraint on Distortion:** For function $\phi(\cdot)$ defined in (72) in Appendix D.5, and for some $\tau \in (0, 0.5\vartheta)$, let $\epsilon^{\text{tree}} := \phi^{-1}\left(\frac{0.5\vartheta - \tau}{3d}\right) > \epsilon_0$, where ϵ_0 is given by (36). The number of samples is now required to satisfy

$$n > n_{\text{local-spect}}(\delta, \epsilon^{\text{tree}}; p, d, r), \tag{38}$$

where $n_{\text{local-spect}}$ is given by (35).

Conditions (B11) and (B12) are identical to (A10) and (A11), except that the required bound ϵ^{tree} in (B12) is required to be above the lower bound ϵ_0 in (36).

B.4.2. *Guarantees for Learning Mixture Components.* We now provide the result on the success of recovering the tree approximation T_h of each mixture component $P(\mathbf{y}|H = h)$ under local separation.

THEOREM 5 (Guarantees for FindMixtureComponents). *Under the assumptions (B1)–(B10), the procedure in Algorithm 3 outputs $\widehat{P}^{\text{spect}}(Y_a, Y_b|H = h)$, for $a, b \in V \setminus \{u_*\}$, with probability at least $1 - 4\delta$, such that for all $h \in [r]$, there exists a permutation $\tau(h) \in [r]$ with*

$$\|\widehat{P}^{\text{spect}}(Y_a, Y_b|H = h) - P(Y_a, Y_b|H = \tau(h))\|_2 \le \epsilon. \tag{39}$$

Moreover, under additional assumptions (B11)-(B12), the method outputs the correct Chow-Liu tree T_h of each component $P(\mathbf{y}|H = h)$ with probability at least $1 - 4\delta$.

Remark: The sample and computational complexities are significantly improved, since it only depends on the size of local separators (while previously it depended on the size of exact separators).

[13] Assume that the Chow-Liu tree T_h is unique for each component $h \in [r]$ under exact statistics, and this holds for generic parameters.

APPENDIX C: ANALYSIS OF RANK TEST

Bounds on Empirical Probability: . We first recap the result from [27, Proposition 19], which is an application of the McDiarmid's inequality. Let $\|\cdot\|_2$ the ℓ_2 norm of a vector.

PROPOSITION 1 (Bound for Empirical Probability Estimates). *Given empirical estimates $\widehat{P}^n$ of a probability vector P using n i.i.d. samples, we have*

$$\mathbb{P}[\|\widehat{P}^n - P\|_2 > \epsilon] \leq \exp\left[-n\left(\epsilon - 1/\sqrt{n}\right)^2\right], \quad \forall \epsilon > 1/\sqrt{n}. \tag{40}$$

Remark: The bound is independent of the cardinality of the sample space.

This implies concentration bounds for $\widehat{M}_{u,v,\{S;k\}}$. Let $\|\cdot\|_2$ and $\|\cdot\|_{\mathbb{F}}$ denote the spectral norm and the Frobenius norms respectively.

LEMMA 3 (Bounds for $\widehat{M}_{u,v,\{S;k\}}$). *Given n i.i.d. samples $\mathbf{y}^n$, the empirical estimate $\widehat{M}^n_{u,v,\{S;k\}} := [\widehat{P}^n[Y_u = i, Y_v = j, \mathbf{Y}_S = k]]_{i,j}$ satisfies*

$$\mathbb{P}\left[\max_{\substack{l\in[d]\\ k\in\mathcal{Y}^{|S|}}} |\sigma_l(\widehat{M}^n_{u,v,\{S;k\}}) - \sigma_l(M_{u,v,\{S;k\}})| > \epsilon\right] \leq \exp\left[-n\left(\epsilon - 1/\sqrt{n}\right)^2\right], \quad \forall \epsilon > 1/\sqrt{n}. \tag{41}$$

Proof: Using proposition 1, we have

$$\mathbb{P}[\max_{k\in\mathcal{Y}^{|S|}} \|\widehat{P}^n(Y_u, Y_v, \mathbf{Y}_S = k) - P(Y_u, Y_v, \mathbf{Y}_S = k)\|_2 > \epsilon] \leq \exp\left[-n\left(\epsilon - 1/\sqrt{n}\right)^2\right], \quad \epsilon > 1/\sqrt{n}. \tag{42}$$

In other words,

$$\mathbb{P}[\max_{k\in\mathcal{Y}^{|S|}} \|\widehat{M}^n_{u,v,\{S;k\}} - M_{u,v,\{S;k\}}\|_{\mathbb{F}} > \epsilon] \leq \exp\left[-n\left(\epsilon - 1/\sqrt{n}\right)^2\right], \quad \epsilon > 1/\sqrt{n}. \tag{43}$$

Since $\|A\|_2 \leq \|A\|_{\mathbb{F}}$ for any matrix A and applying the Weyl's theorem, we have the result. □

From Lemma 1 and Lemma 3, it is easy to see that

$$\mathbb{P}[\widehat{G}^n_\cup \neq G_\cup] \leq 2p^2 \exp\left[-n\left(\rho_{\min}/2 - 1/\sqrt{n}\right)^2\right],$$

and we have the result. Similarly, we have Theorem 4 from Lemma 11 and Lemma 3. □

APPENDIX D: ANALYSIS OF SPECTRAL DECOMPOSITION

D.1. Analysis Under Exact Statistics. We now prove the success of FindMixtureComponents under exact statistics. We first consider three sets $A_1, A_2, A_3 \subset V$ such that $\mathcal{N}[A_i; G_\cup] \cap \mathcal{N}[A_j; G_\cup] = \emptyset$ for $i, j \in [3]$ and $G_\cup := \cup_{h\in[r]} G_h$ is the union of the Markov graphs. Let $S \subset V \setminus \cup_i A_i$ be a multiway separator set for A_1, A_2, A_3 in graph $G_\cup$. For A_i, $i \in \{1, 2, 3\}$, let $U_i \in \mathbb{R}^{d^{|A_i|}\times r}$ be a matrix such that $U_i^\top M_{A_i|H,\{S;k\}}$ is invertible, for a fixed $k \in \mathcal{Y}^{|S|}$. Then $U_1^\top M_{A_1,A_2,\{S;k\}} U_2$ is invertible, and for all $\mathbf{m} \in \mathbb{R}^{d^{|A_3|}}$, the observable operator $\widetilde{C}(\mathbf{m}) \in \mathbb{R}^{r\times r}$, given by

$$\widetilde{C}(\mathbf{m}) := \left(U_1^\top \left(\sum_q m(q) M_{A_1,A_2,\{S;k\},\{A_3;q\}}\right) U_2\right) \left(U_1^\top M_{A_1,A_2,\{S;k\}} U_2\right)^{-1}. \tag{44}$$

Note that the above operator is computed in SpecDecom procedure. We now provide a generalization of the result in [3].

LEMMA 4 (Observable Operator). *Under assumption (A6), the observable operator in* (44) *satisfies*

$$\widetilde{C}(\mathbf{m}) = \left(U_1^T M_{A_1|H,\{S;k\}}\right) \operatorname{Diag}\left(M_{A_3|H,\{S;k\}}^\top \mathbf{m}\right) \left(U_1^T M_{A_1|H,\{S;k\}}\right)^{-1}. \tag{45}$$

In particular, the r roots of the polynomial $\lambda \mapsto \det(\widetilde{C}(\mathbf{m}) - \lambda I)$ are $\{\langle \mathbf{m}, M_{A_3|H,\{S;k\}}\mathbf{e}_j\rangle : j \in [r]\}$.

Proof: We have

$$U_1^\top M_{A_1,A_2,\{S;k\}} U_2 = (U_1^\top M_{A_1|H,\{S;k\}}) \operatorname{Diag}(\boldsymbol{\pi}_{H,\{S;k\}})(M_{A_2|H,\{S;k\}}^\top U_2)$$

on lines of (8), which is invertible by the assumptions on U_1, U_2 and Assumption (A6). Similarly,

$$U_1^\top M_{A_1,A_2,\{S;k\},\{A_3;q\}} U_2 = (U_1^\top M_{A_1|H,\{S;k\}}) \operatorname{Diag}(\boldsymbol{\pi}_{H,\{S;k\},\{A_3;q\}})(M_{A_2|H,\{S;k\}}^\top U_2),$$

and we have the result. □

The above result implies that we can recover the matrix $M_{A|H,\{S;k\}}$ for any set $A \subset V$, by using a suitable reference node, a witness and a separator set. We set the isolated node u_* as the reference node (set A_1 in the above result). Since we focus on recovering the edge marginals of the mixture components, we consider each node pair $a, b \in V \setminus \{u_*\}$ (set A_3 in the above result), and any node $c \notin \mathcal{N}(a; G_\cup) \cup \mathcal{N}(b; G_\cup)$ (set A_2 in the above result), where $G_\cup := \cup_{h\in[r]} G_h$, as described in FindMixtureComponents. Thus, we are able to recover $M_{a,b|H,\{S;k\}}$ under exact statistics. Since $\mathbf{Y}_S$ are observed, we have the knowledge of $P(\mathbf{Y}_S = k)$, and can thus recover $M_{a,b|H}$ as desired. The spectral decompositions of different groups are aligned since we use the same node u_*, and since u_* is isolated in $G_\cup$, fixing the variables $\mathbf{Y}_S = k$ has no effect on the conditional distribution of Y_{u_*}, i.e., $P(Y_{u_*}|H, \mathbf{Y}_S = k) = P(Y_{u_*}|H)$. Since we recover the edge marginals $M_{a,b|H}$ correctly we can recover the correct tree approximation T_h, for $h \in [r]$.

D.2. Analysis of SpecDecom$(u, v, w; S)$. We consider success of SpecDecom$(u, v, w; S)$ for estimating the statistics of w using node $u \in V$ as the reference node (which is conditionally independent of all other nodes given H) and witness $v \in V$ and separator set S. We will use this to provide sample complexity results on FindMixtureComponents using union bounds. The proof borrows heavily from [3].

Recall that $\widehat{U}_u$ is the set of top r left orthonormal singular vectors of $\widehat{M}^n_{u,v,\{S;k\}}$ and $\widehat{V}_v$ as the right orthonormal vectors. For $l \in [r]$, let $\mathbf{m}_l = \widehat{U}_w \mathbf{z}_l$, where $\mathbf{z}_l$ is uniformly distributed in $\mathbb{S}^{r-1}$ and $\widehat{U}_w$ is the top r left singular vectors of $\widehat{M}^n_{w,u,\{S;k\}}$. By Lemma 13, we have that $U_u^\top M_{u,v,\{S;k\}} V_v$ is invertible. Recall the definition of the observable operator in (44)

$$\widetilde{C}_l := \widetilde{C}(\mathbf{m}_l) = \widehat{U}_u^\top \left(\sum_q m_l(q) M_{u,v,\{S;k\},\{w;q\}}\right) \widehat{V}_v \left(U_u^\top M_{u,v,\{S;k\}} V_v\right)^{-1}, \tag{46}$$

where exact matrices M are used. Denote $\widehat{C}_l$ when the sample versions $\widehat{M}^n$ are used

$$\widehat{C}_l := \widehat{U}_u^\top \left(\sum_q m_l(q) \widehat{M}^n_{u,v,\{S;k\},\{w;q\}}\right) \widehat{V}_v \left(U_u^\top \widehat{M}^n_{u,v,\{S;k\}} V_v\right)^{-1}, \tag{47}$$

We have the following result.

LEMMA 5 (Bounds for $\|\widehat{C}_l - \widetilde{C}_l\|_2$). *The matrices $\widetilde{C}_l$ and $\widehat{C}_l$ defined in (46) and (47) satisfy*

$$\|\widehat{C}_l - \widetilde{C}_l\|_2 \leq \frac{2\|\sum_q m_l(q)(\widehat{M}^n_{u,v,\{S;k\},\{w;q\}} - M_{u,v,\{S;k\},\{w;q\}})\|_2}{\sigma_r(M_{u,v,\{S;k\}})} + \frac{2\|\sum_q m_l(q) M_{u,v,\{S;k\},\{w;q\}}\|_2 \|\widehat{M}^n_{u,v,\{S;k\}} - M_{u,v,\{S;k\}}\|_2}{\sigma_r(M_{u,v,\{S;k\}})^2}. \tag{48}$$

Proof: Using Lemma 14 and Lemma 4. □

We now provide perturbation bounds between estimated matrix $\widehat{M}_{w|H,\{S;k\}}$ and the true matrix $M_{w|H,\{S;k\}}$. Define

$$\beta(w) := \min_{k\in\mathcal{Y}^{|S|}} \min_{i\in[r]} \min_{j\neq j'} |\langle \mathbf{z}^{(i)}, \widehat{U}_w^\top M_{w|H,\{S;k\}}(\vec{e}_j - \vec{e}_{j'})\rangle| \tag{49}$$

$$\lambda_{\max}(w) := \max_{i,j\in[r]} |\langle \mathbf{z}^{(i)}, \widehat{U}_w^\top M_{w|H,\{S;k\}}\vec{e}_j\rangle|, \tag{50}$$

where $\mathbf{z}_l$ is uniformly distributed in $\mathbb{S}^{r-1}$.

LEMMA 6 (Relating $\widehat{M}_{w|H,\{S;k\}}$ and $M_{w|H,\{S;k\}}$). *The estimated matrix $\widehat{M}_{w|H,\{S;k\}}$ using samples and the true matrix $M_{w|H,\{S;k\}}$ satisfy, for all $j \in [r]$,*

$$\|\widehat{M}_{w|H,\{S;k\}}\mathbf{e}_j - M_{w|H,\{S;k\}}\mathbf{e}_{\tau(j)}\|_2 \leq 2\|M_{w|H,\{S;k\}}\mathbf{e}_{\tau(j)}\|_2 \cdot \frac{\|\widehat{M}^n_{u,w,\{S;k\}} - M_{u,w,\{S;k\}}\|_2}{\sigma_r(M_{u,w,\{S;k\}})} + \Big(12\sqrt{r}\cdot\kappa(M_{u|H})^2 + 256r^2\cdot\kappa(M_{u|H})^4\cdot\lambda_{\max}(w)/\beta(w)\Big)\cdot\|\widehat{C}_l - \widetilde{C}_l\|_2. \tag{51}$$

Proof: Define a matrix $R := \widehat{U}_u^\top M_{u|H}\,\mathrm{Diag}(\|\widehat{U}_u^\top M_{u|H}\mathbf{e}_1\|_2, \ldots, \|\widehat{U}_u^\top M_{u|H}\mathbf{e}_r\|_2)^{-1}$. Note that R has unit norm columns and R diagonalizes $\widetilde{C}_l$, i.e.,

$$R^{-1}\widetilde{C}_l R = \mathrm{Diag}(M^\top_{w|H,\{S;k\}}\mathbf{z}_l).$$

Using the fact that for any stochastic matrix $d\times r$ matrix A, $\|A\|_2 \leq \sqrt{r}\|A\|_1 = \sqrt{r}$ and Lemma 17, we have

$$\|R^{-1}\|_2 \leq 2\kappa(\widehat{U}_u^\top M_{u|H}), \quad \kappa(R) \leq 4\kappa(M_{u|H}).$$

From above and by Lemma 16, there exist a permutation τ on $[r]$ such that, for all $j, l \in [r]$,

$$\begin{aligned}|\widehat{\lambda}^{(l)}(j) - \lambda^{(l)}(\tau(j))| &\leq \Big(3\kappa(R) + 16r^{1.5}\cdot\kappa(R)\cdot\|R^{-1}\|_2^2\cdot\lambda_{\max}(w)/\beta(w)\Big)\cdot\|\widehat{C}_l - \widetilde{C}_l\|_2\\ &\leq \Big(12\kappa(M_{u|H})^2 + 256r^{1.5}\cdot\kappa(M_{u|H})^4\cdot\lambda_{\max}(w)/\beta(w)\Big)\cdot\|\widehat{C}_l - \widetilde{C}_l\|_2,\end{aligned} \tag{52}$$

where $\beta(w)$ and $\lambda_{\max}(w)$ are given by (49) and (50). Let $\widehat{\boldsymbol{\nu}}^{(j)} := (\widehat{\lambda}^{(1)}(j), \widehat{\lambda}^{(2)}(j), \ldots, \widehat{\lambda}^{(r)}(j)) \in \mathbb{R}^r$ be the row vector corresponding to j^{th} row of $\widehat{\Lambda}$ and $\vec{\nu}^{(j)} := (\lambda^{(1)}(j), \lambda^{(2)}(j), \ldots, \lambda^{(r)}(j)) \in \mathbb{R}^r$. Observe that $\vec{\nu}^{(j)} = Z\widehat{U}^\top_{w|H,\{S;k\}} M_{w|H,\{S;k\}}\vec{e}_j$. By the orthogonality of Z, the fact $\|\vec{v}\|_2 \leq \sqrt{r}\|\vec{v}\|_\infty$

for $\vec{v} \in \mathbb{R}^r$, and the above inequality,

$$\begin{aligned}
&\|Z^{-1}\widehat{\boldsymbol{\nu}}^{(j)} - \widehat{U}_{w|H,\{S;k\}}^{\top} M_{w|H,\{S;k\}} \vec{e}_{\tau(j)}\|_2 \\
&= \|Z^{-1}(\widehat{\boldsymbol{\nu}}^{(j)} - \vec{\nu}^{(\tau(j))})\|_2 \\
&= \|\widehat{\boldsymbol{\nu}}^{(j)} - \vec{\nu}^{(\tau(j))}\|_2 \\
&\leq \sqrt{r} \cdot \|\widehat{\boldsymbol{\nu}}^{(j)} - \vec{\nu}^{(\tau(j))}\|_\infty \\
&\leq \Big(12\sqrt{r} \cdot \kappa(M_{u|H})^2 + 256 r^2 \cdot \kappa(M_{u|H})^4 \cdot \lambda_{\max}(w)/\beta(w)\Big) \cdot \|\widehat{C}_l - \widetilde{C}_l\|_2.
\end{aligned}$$

By Lemma 13 (as applied to $\widehat{M}^n_{u,w,\{S;k\}}$ and $M_{u,w,\{S;k\}}$), we have

$$\begin{aligned}
\|\widehat{M}_{w|H,\{S;k\}}\mathbf{e}_j - M_{w|H,\{S;k\}}\mathbf{e}_{\tau(j)}\|_2 &\leq \|Z^{-1}\widehat{\boldsymbol{\nu}}^{(j)} - \widehat{U}_{w|H,\{S;k\}}^{\top} M_{w|H,\{S;k\}} \vec{e}_{\tau(j)}\|_2 \\
&\quad + 2\|M_{w|H,\{S;k\}}\mathbf{e}_{\tau(j)}\|_2 \cdot \frac{\|\widehat{M}^n_{u,w,\{S;k\}} - M_{u,w,\{S;k\}}\|_2}{\sigma_r(M_{u,w,\{S;k\}})}. \qquad (53)
\end{aligned}$$

□

D.3. Analysis of FindMixtureComponents. We now provide results for Procedure FindMixtureComponents by using the previous result, where w is set to each node pair $a, b \in V \setminus \{u_*\}$. We condition on the event that $\widehat{G}_\cup = G_\cup$, where $G_\cup := \cup_{h\in[r]} G_h$ is the union of the component graph.

We now give concentration bounds for β and $\lambda_{\max}$ in (49) and (50). Define

$$\alpha_{\min} := \min_{a,b\in V\setminus\{u_*\}} \min_{\substack{k\in\mathcal{Y}^{|S|}, |S|\leq 2s(G_\cup) \\ S\subset V\setminus\{a,b,u_*\}}} \min_{i\neq i'} \|M_{(a,b)|H,\{S;k\}}(\mathbf{e}_i - \mathbf{e}_{i'})\|_2 \qquad (54)$$

$$\alpha_{\max} := \max_{a,b\in V\setminus\{u_*\}} \max_{\substack{k\in\mathcal{Y}^{|S|}, |S|\leq 2s(G_\cup) \\ S\subset V\setminus\{a,b,u_*\}}} \max_{j\in[r]} \|M_{(a,b)|H,\{S;k\}}\mathbf{e}_j\|_2, \qquad (55)$$

and let

$$\alpha := \frac{\alpha_{\max}}{\alpha_{\min}}. \qquad (56)$$

Lemma 7 (Bounds for β and $\lambda_{\max}$). *Fix $\delta \in (0,1)$, given any $a, b \in V \setminus \{u_*\}$ and any set $S \subset V \setminus \{a, b, u_*\}$ with $|S| \leq 2s(G_\cup)$, we have with probability at least $1 - \delta$,*

$$\beta(a,b) \geq \frac{\alpha_{\min} .\delta}{2\sqrt{er}\binom{r}{2} r p^2 (pd)^{2s(G_\cup)}} \qquad (57)$$

$$\lambda_{\max}(a,b) \leq \frac{\alpha_{\max}}{\sqrt{r}}\left(1 + \sqrt{2\ln(r^2p^2(pd)^{2s(G_\cup)}/\delta)}\right) \qquad (58)$$

This implies that with probability at least $1 - 2\delta$,

$$\frac{\lambda_{\max}(a,b)}{\beta(a,b)} \geq \frac{\sqrt{e}\alpha}{\delta} r^3 p^2 (pd)^{2s(G_\cup)}\left(1 + \sqrt{2\ln(r^2p^2(pd)^{2s(G_\cup)}/\delta)}\right), \qquad (59)$$

where α is given by (56).

Similarly, we have bounds on $\|\widehat{M}^n_{u_*,a,b,\{S;k\}} - M_{u_*,a,b,\{S;k\}}\|_2$ using Lemma 3 and union bound.

Proposition 2 ($\|\widehat{M}^n_{u_*,a,b,\{S;k\}} - M_{u_*,a,b,\{S;k\}}\|_2$). *With probability at least $1-\delta$, we have, for all $a, b \in V \setminus \{u_*\}$, $S \subset V \setminus \{a, b, u_*\}$, $|S| \leq 2s(G_\cup)$,*

$$\|\widehat{M}^n_{u_*,a,b,\{S;k\}} - M_{u_*,a,b,\{S;k\}}\|_2 \leq \frac{1}{\sqrt{n}} \left(1 + \sqrt{\log\left(\frac{p^{2s(G_\cup)+2} d^{2s(G_\cup)}}{\delta}\right)} \right). \tag{60}$$

Define $\rho'_{1,\min}$, $\rho'_{2,\min}$ and $\rho'_{\max}$ as

$$\rho'_{1,\min} := \min_{\substack{S \subset V \setminus \{u_*, v\} \\ |S| \leq 2s(G_\cup), k \in \mathcal{Y}^{|S|}}} \ \min_{v \in V \setminus \{u_*\}} \sigma_r\left(M_{u_*,v,\{S;k\}}\right), \tag{61}$$

$$\rho'_{2,\min} := \min_{\substack{S \subset V \setminus \{u_*, a, b\} \\ |S| \leq 2s(G_\cup), k \in \mathcal{Y}^{|S|}}} \ \min_{a,b \in V \setminus \{u_*\}} \sigma_r\left(M_{u_*,a,b,\{S;k\}}\right). \tag{62}$$

Using the above defined constants, define

$$\begin{aligned} K'(\delta; p, d, r) :=& 1024 \cdot \kappa(M_{u|H})^4 \cdot \frac{\sqrt{e}\alpha}{\delta \rho'_{1,\min}} r^5 p^2 (pd)^{2s(G_\cup)} \left(1 + \sqrt{2 \ln(r^2 p^2 (pd)^{2s(G_\cup)}/\delta)} \right) \\ &+ 48 \frac{\sqrt{r}}{\rho'_{1,\min}} \cdot \kappa(M_{u|H})^2 + \frac{2\alpha_{\max}}{\rho'_{2,\min}}, \end{aligned} \tag{63}$$

and

$$K(\delta; p, d, r) := K'(\delta; p, d, r) \left(1 + \sqrt{\log\left(\frac{p^{2s(G_\cup)+2} d^{2s(G_\cup)}}{\delta}\right)} \right). \tag{64}$$

We can now provide the final bound on distortion of estimated statistics using all the previous results.

Lemma 8 (Bounds for $\|\widehat{M}_{a,b|H,\{S;k\}}\mathbf{e}_j - M_{a,b|H,\{S;k\}}\mathbf{e}_{\tau(j)}\|_2$). *For any $a, b \in V \setminus \{u_*\}$, $k \in \mathcal{Y}^{|S|}$, $j \in [r]$, there exists a permutation $\tau(j) \in [r]$ such that, conditioned on event that $\widehat{G}_\cup = G_\cup$, with probability at least $1 - 3\delta$,*

$$\|\widehat{M}_{a,b|H,\{S;k\}}\mathbf{e}_j - M_{a,b|H,\{S;k\}}\mathbf{e}_{\tau(j)}\|_2 \leq \frac{K(\delta; p, d, r)}{\sqrt{n}}. \tag{65}$$

This implies

$$\|\widehat{M}_{a,b|H}\mathbf{e}_j - M_{a,b|H}\mathbf{e}_{\tau(j)}\|_2 \leq \frac{K(\delta; p, d, r)}{\sqrt{n}} + \frac{K(\delta; p, d, r)}{K'(\delta; p, d, r)\sqrt{n}} \leq \frac{2K(\delta; p, d, r)}{\sqrt{n}}. \tag{66}$$

Results on Random Rotation Matrix: . We also require the following result from [3]. The standard inner product between vectors $\vec{u}$ and $\vec{v}$ is denoted by $\langle \vec{u}, \vec{v} \rangle = \vec{u}^\top \vec{v}$. Let $\sigma_i(A)$ denote the i^{th} largest singular value of a matrix A. Let $\mathbb{S}^{m-1} := \{\vec{u} \in \mathbb{R}^m : \|\vec{u}\|_2 = 1\}$ denote the unit sphere in $\mathbb{R}^m$. Let $\vec{e}_i \in \mathbb{R}^d$ denote the i^{th} coordinate vector where the i^{th} entry is 1, and the rest are zero.

Lemma 9. *Fix any $\delta \in (0, 1)$ and matrix $A \in \mathbb{R}^{m \times n}$ (with $m \leq n$). Let $\vec{\theta} \in \mathbb{R}^m$ be a random vector distributed uniformly over $\mathbb{S}^{m-1}$.*

1. $\Pr\left[\min_{i \neq j} |\langle \vec{\theta}, A(\vec{e}_i - \vec{e}_j) \rangle| > \frac{\sqrt{2}\sigma_m(A) \cdot \delta}{\sqrt{em}\binom{n}{2}} \right] \geq 1 - \delta.$
2. $\Pr\left[\forall i \in [m],\ |\langle \vec{\theta}, A\vec{e}_i \rangle| \leq \frac{\|A\vec{e}_i\|_2}{\sqrt{m}} \left(1 + \sqrt{2\ln(m/\delta)} \right) \right] \geq 1 - \delta.$

D.4. Improved Results for Tree Mixtures. We now consider a simplified version of the procedure FindMixtureComponents by limiting to estimation of pairwise marginals only on the edges of $\widehat{G}_\cup$, where $\widehat{G}_\cup$ is the estimate of $G_\cup := \cup_{h\in[r]} G_h$, which is the union of the component graph, as well as constructing the Chow-Liu trees $\widehat{T}_h$ as subgraphs of $\widehat{G}_\cup$. Thus, instead of considering each node pair $a, b \in V \setminus \{u_*\}$, we only need to choose $(a,b) \in \widehat{G}_\cup$. Moreover, instead of considering $S \subset V \setminus \{a, b, u_*\}$, we can follow the convention of choosing $S \subset \mathcal{N}(a; \widehat{G}_\cup) \cup \mathcal{N}(b; \widehat{G}_\cup)$, and this changes the definition of $\alpha_{\min}$, $\alpha_{\max}$, $\rho'_{1,\min}$, $\rho'_{2,\min}$ and so on. For all $(a,b) \in G_\cup$, let

$$\Delta_2 := \max_{(a,b)\in G_\cup} |\mathcal{N}(a; G_\cup) \cup \mathcal{N}(b; G_\cup)|. \tag{67}$$

We have improved bounds for β and $\lambda_{\max}$ defined in (49) and (50), when Δ_2 is small.

Lemma 10 (Improved Bounds for β and $\lambda_{\max}$). *Fix $\delta \in (0,1)$, when $|S| \le 2s(G_\cup)$ and $S \subset \mathcal{N}(a; G_\cup) \cup \mathcal{N}(b; G_\cup)$, with probability at least $1-\delta$,*

$$\beta(w) \ge \frac{\sqrt{2}\alpha_{\min}\delta}{\sqrt{er}\binom{r}{2} r p^2 d^{2s(G_\cup)} \Delta_2^{2s(G_\cup)}} \tag{68}$$

$$\lambda_{\max}(w) \le \frac{\alpha_{\max}}{\sqrt{r}}\left(1 + \sqrt{2\ln(r^2 p^2 d^{2s(G_\cup)} \Delta_2^{2s(G_\cup)}/\delta)}\right) \tag{69}$$

We can substitute the above result to obtain a better bound $K^{\text{tree}}(\delta; p, d, r)$ for learning tree mixtures.

D.5. Analysis of Tree Approximations: Proof of Theorem 3. We now relate the perturbation of probability vector to perturbation of the corresponding mutual information [19]. Recall that for discrete random variables X, Y, the mutual information $I(X;Y)$ is related to their entropies $H(X,Y)$, $H(X)$ and $H(Y)$ as

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{70}$$

and the entropy is defined as

$$H(X) := -\sum_{x\in\mathcal{X}} P(X=x)\log P(X=x), \tag{71}$$

where $\mathcal{X}$ is the sample space of X. We recall the following result from [45]. Define function $\phi(x)$ for $x \in \mathbb{R}^+$ as

$$\phi(x) = \begin{cases} 0, & x = 0, & (72a) \\ -x\log x, & x \in (0, 1/e), & (72b) \\ 1/e, & \text{o.w.} & (72c) \end{cases}$$

Proposition 3. *For any $a, b \in [0,1]$,*

$$|a\log a - b\log b| \le \phi(|a-b|), \tag{73}$$

for $\phi(\cdot)$ defined in (72).

We can thus prove bounds on the estimated mutual information $\widehat{I}^{\text{spect}}(\cdot)$ using statistics $\widehat{P}^{\text{spect}}(\cdot)$ obtained from spectral decomposition.

Proposition 4 (Bounding $|\widehat{I}^{\text{spect}}(\cdot) - I(\cdot)|$). *Under the event that* $\|\widehat{P}^{\text{spect}}(Y_a, Y_a|H = h) - P(Y_a, Y_a|H = h)\|_2 \leq \epsilon$, *we have that*

(74) $$|\widehat{I}^{\text{spect}}(Y_a; Y_a|H = h) - I(Y_a; Y_a|H = h)| \leq 3d\phi(\epsilon).$$

For success of Chow-Liu algorithm, it is easy to see that the algorithm finds the correct tree when the estimated mutual information quantities are within half the minimum separation ϑ defined in (23). This is because the only wrong edges in the estimated tree $\widehat{T}_h$ are those that replace a certain edge in the original tree T_h, without violating the tree constraint. Similar ideas have been used by Tan, Anandkumar and Willsky [47] for deriving error exponent bounds for the Chow-Liu algorithm. Define

(75) $$\epsilon^{\text{tree}} := \phi^{-1}\left(\frac{0.5\vartheta - \tau}{3d}\right).$$

Thus, using the above result and assumption (A11) implies that we can estimate the mutual information to required accuracy to obtain the correct tree approximations.

APPENDIX E: ANALYSIS UNDER LOCAL SEPARATION CRITERION

E.1. Rank Tests Under Approximate Separation. We now extend the results of the previous section when approximate separators are employed in contrast to exact vertex separators. Let $S := \mathcal{S}_{\text{local}}(u, v; G, \gamma)$ denote a local vertex separator between any non-neighboring nodes u and v in graph G under threshold γ. We note the following result on the probability matrix $M_{u,v,\{S;k\}}$ defined in (4).

Lemma 11 (Rank Upon Approximate Separation). *Given a r-mixture of graphical models with $G = \cup_{k=1}^r G_k$, for any nodes $u, v \in V$ such that $\mathcal{N}[u] \cap \mathcal{N}[v] = \emptyset$ and $S := \mathcal{S}_{\text{local}}(u, v; G, \gamma)$ be any separator of u and v on G, the probability matrix $M_{u,v,\{S;k\}} := [P[Y_u = i, Y_v = j, \mathbf{Y}_S = k]]_{i,j}$ has effective rank at most r for any $k \in \mathcal{Y}^{|S|}$*

(76) $$\text{Rank}\left(M_{u,v,\{S;k\}}\,; \zeta(\gamma)\right) \leq r, \quad \forall\, k \in \mathcal{Y}^{|S|}, (u, v) \notin G,$$

where $\zeta(\gamma) := 2\sqrt{d}\max_{h\in[r]} \zeta_h(\gamma)$, and $\zeta_h(\cdot)$ is the correlation decay rate function in (29) corresponding to the model $P(\mathbf{y}|H = h)$ and γ is the path threshold for local vertex separators.

Notation: For convenience, for any node $v \in V$, let $P(Y_v|H = h) := P(Y_v|H = h; G_h)$ denote the original component model Markov on graph G_h, and let $P(Y_v)$ denote the corresponding marginal distribution of Y_v in the mixture. Let $\check{P}^\gamma(Y_v|H = h) := P(Y_v|H = h; F_{\gamma,h})$ denote the component model Markov on the induced subgraph $F_{\gamma,h} := G_h(B_\gamma(v))$, where $B_\gamma(v; G_h)$ is the γ-neighborhood of node v in G_h. In other words, we limit the model parameters up to γ neighborhood and remove rest of the edges to obtain $\check{P}^\gamma(Y_v|H = h)$.

Proof: We first claim that

(77) $$\|M_{u|v,\{S;k\}} - M_{u|H,\{S;k\}} M_{H|v,\{S;k\}}\|_2 \leq \zeta(\gamma).$$

Note the relationship between the joint and the conditional probability matrices:

(78) $$M_{u,v,\{S;k\}} = M_{u|v,\{S;k\}}\,\text{Diag}(\boldsymbol{\pi}_{v,\{S;k\}}),$$

where $\boldsymbol{\pi}_{v,\{S;k\}} := [P(Y_v = i, \mathbf{Y}_S = k)]_i^\top$ is the probability vector and $\mathrm{Diag}(\cdot)$ is the diagonal matrix with the corresponding probability vector as the diagonal elements. Assuming (77) holds and applying (78), we have that

$$\begin{aligned}&\|M_{u,v,\{S;k\}} - M_{u|H,\{S;k\}} M_{H|v,\{S;k\}} \mathrm{Diag}(\boldsymbol{\pi}_{v,\{S;k\}})\|_2 \\ &\leq \|\mathrm{Diag}(\boldsymbol{\pi}_{v,\{S;k\}})\|_2 \zeta(\gamma) \leq \zeta(\gamma),\end{aligned} \tag{79}$$

since $\|\mathrm{Diag}(\boldsymbol{\pi}_{v,\{S;k\};G})\|_2 \leq \|\mathrm{Diag}(\boldsymbol{\pi}_{v,\{S;k\};G})\|_{\mathbb{F}} = \|\boldsymbol{\pi}_{v,\{S;k\};G}\|_2 \leq 1$ for a probability vector. From Weyl's theorem, assuming that (79) holds, we have

$$\mathrm{Rank}\left(M_{u,v,\{S;k\}}\,;\,\zeta(\gamma)\right) \leq \min(r,d) = r,$$

since we assume $r < d$ (assumption (B1) in Section B.3). Note that $\mathrm{Rank}(A;\xi)$ denotes the effective rank, i.e., the number of singular values of A which are greater than $\xi \geq 0$.

We now prove the claim in (77). Since $G = \cup_{h=1}^r G_h$, we have that the resulting set $S := \mathcal{S}_{\mathrm{local}}(u,v;G,\gamma)$ is also a local separator on each of the component subgraphs $\{G_h\}_{h\in[r]}$ of G, for all sets $A, B \subset V$ such that $\mathcal{N}[u;G] \cap \mathcal{N}[v;G] = \emptyset$. Thus, we have that for all $k \in \mathcal{Y}^{|S|}$, $y_v \in \mathcal{Y}$, $h \in [r]$,

$$\breve{P}^\gamma(Y_u | Y_v = y_v, \mathbf{Y}_S = k, H = h) = \breve{P}^\gamma(Y_u | \mathbf{Y}_S = k, H = h). \tag{80}$$

The statement in (80) is due to the fact that the nodes u and v are exactly separated by set S in the subgraph $F_{\gamma,h}(u)$.

By assumption (B4) on correlation decay we have that

$$\|P(Y_u|Y_v = y_v, \mathbf{Y}_S = k, H = h) - \breve{P}^\gamma(Y_u|Y_v = y_v, \mathbf{Y}_S = k, H = h)\|_1 \leq \zeta_h(\gamma),$$

for all $y_v \in \mathcal{Y}$, $k \in \mathcal{Y}^{|S|}$ and $h \in [r]$. Similarly, we also have

$$\|P(Y_u|\mathbf{Y}_S = k, H = h) - \breve{P}^\gamma(Y_u|\mathbf{Y}_S = k, H = h)\|_1 \leq \zeta_h(\gamma),$$

which implies that

$$\|P(Y_u|Y_v = y_v, \mathbf{Y}_S = k, H = h) - P(Y_u|\mathbf{Y}_S = k, H = h)\|_1 \leq 2\zeta_h(\gamma),$$

for all $y_v \in \mathcal{Y}$, $k \in \mathcal{Y}^{|S|}$ and $h \in [r]$, and thus,

$$\|M_{u|v,\{S;k\}} - M_{u|H,\{S;k\}} M_{H|v,\{S;k\}}\|_1 \leq 2 \max_{h\in[r]} \zeta_h(\gamma), \tag{81}$$

where $\|A\|_1$ of a matrix is the maximum column-wise absolute sum. Since $\|A\|_2 \leq \sqrt{d}\|A\|_1$, (77) follows. □

E.2. Spectral Decomposition Under Local Separation. We now extend the above analysis of spectral decomposition when a local separator is used instead of approximate separators. For simplicity consider nodes $u_*, a, b, c \in V$ (the same results can also be proven for larger sets), where u_* is an isolated node in $G_\cup$, $a, b \in V \setminus \{u_*\}$, $c \notin \mathcal{N}[a;G_\cup] \cup \mathcal{N}[b;G_\cup]$ and let $S := \mathcal{S}_{\mathrm{local}}((a,b),c;G_\cup)$ be a local separator in $G_\cup$ separating a,b from c. Since we have

$$Y_{u_*} \perp\!\!\!\perp \mathbf{Y}_{V\setminus\{u_*\}} | H,$$

the following decomposition holds

$$M_{u*,c,\{S;k\}} = M_{u*|H} \operatorname{Diag}(\boldsymbol{\pi}_{H,\{S;k\}}) M_{c|H,\{S;k\}}^\top .$$

However, the matrix $M_{u*,c,\{S;k\},\{(a,b);q\}}$ no longer has a similar decomposition. Instead define

$$\widetilde{M}_{u*,c,\{S;k\},\{(a,b);q\}} := M_{u*|H} \operatorname{Diag}(\boldsymbol{\pi}_{H,\{S;k\},\{(a,b);q\}}) M_{c|H,\{S;k\}}^\top . \tag{82}$$

Define the observable operator, on lines of (44), based on $\widetilde{M}$ above rather than the actual probability matrix M, as

$$\widetilde{\widetilde{C}}(\mathbf{m}) := \left(U_1^\top \left(\sum_q m(q) \widetilde{M}_{u*,c,\{S;k\},\{(a,b);q\}} \right) U_2 \right) \left(U_1^\top M_{u*,c,\{S;k\}} U_2 \right)^{-1}, \tag{83}$$

where U_1 is a matrix such that $U_1^\top M_{u*|H}$ is invertible and U_2 is such that $U_2^\top M_{v|H,\{S;k\}}$ is invertible. On lines of Lemma 4, we have that

$$\widetilde{\widetilde{C}}(\mathbf{m}) = \left(U_1^T M_{u*|H} \right) \operatorname{Diag}\left(M_{(a,b)|H,\{S;k\}}^\top \mathbf{m} \right) \left(U_1^T M_{u*|H} \right)^{-1}. \tag{84}$$

Thus, the r roots of the polynomial $\lambda \mapsto \det(\widetilde{\widetilde{C}}(\mathbf{m}) - \lambda I)$ are $\{\langle \mathbf{m}, M_{(a,b)|H,\{S;k\}} \mathbf{e}_j \rangle : j \in [r]\}$. We now have show that M and $\widetilde{M}$ are close under correlation decay.

Proposition 5 (Regime of Correlation Decay). *For all $k \in \mathcal{Y}^{|S|}$ and $q \in \mathcal{Y}^2$, we have*

$$\| \widetilde{M}_{u*,c,\{S;k\},\{(a,b);q\}} - M_{u*,c,\{S;k\},\{(a,b);q\}} \|_2 \le \zeta(\gamma), \tag{85}$$

where $\zeta(\gamma)$ is given by (32).

Proof: On lines of obtaining (81) in the proof of Lemma 11, it is easy to see that

$$\|P(Y_c|\mathbf{Y}_S = k, \mathbf{Y}_{a,b} = q) - \sum_{h \in [r]} P(Y_c|\mathbf{Y}_S = k, H = h) P(H = h|\mathbf{Y}_S = k, \mathbf{Y}_{a,b} = q)\|_1 \le 2 \max_{h \in [r]} \zeta_h(\gamma).$$

This implies that for all $y \in \mathcal{Y}$,

$$\begin{aligned} \| \sum_{h \in [r]} & P(Y_{u_*} = y|H = h) P(H = h, \mathbf{Y}_S = k, \mathbf{Y}_{a,b} = q) P(Y_c|\mathbf{Y}_S = k, H = h) \\ & - P(Y_c, Y_{u_*} = y, \mathbf{Y}_S = k, \mathbf{Y}_{a,b} = q)\|_1 \le 2 \max_{h \in [r]} \zeta_h(\gamma). \end{aligned} \tag{86}$$

This is the same as

$$\| \widetilde{M}_{u*,c,\{S;k\},\{(a,b);q\}} - M_{u*,c,\{S;k\},\{(a,b);q\}} \|_\infty \le 2 \max_{h \in [r]} \zeta_h(\gamma), \tag{87}$$

where $\|A\|_\infty$ is the maximum absolute row sum and $\|A\|_2 \le \sqrt{d} \|A\|_\infty$ for a $d \times d$ matrix, and thus, we have the result. □

E.3. Spectral Bounds under Local Separation. The result follows on similar lines as Section D.3, except that the distortion between the sample version of the observable operator $\widehat{C}(\mathbf{m})$ and the desired version $\widetilde{\widetilde{C}}(\mathbf{m})$ changes. This leads to a slightly different bound

LEMMA 12 (Bounds for $\|\widehat{M}_{a,b|H,\{S;k\}}\mathbf{e}_j - M_{a,b|H,\{S;k\}}\mathbf{e}_{\tau(j)}\|_2$). *For any $a, b \in V \setminus \{u_*\}$, $k \in \mathcal{Y}^{|S|}$, $j \in [r]$, there exists a permutation $\tau(j) \in [r]$ such that, conditioned on event that $\widehat{G}_\cup = G_\cup$, with probability at least $1 - 3\delta$,*

$$\|\widehat{M}_{a,b|H,\{S;k\}}\mathbf{e}_j - M_{a,b|H,\{S;k\}}\mathbf{e}_{\tau(j)}\|_2 \leq \frac{K(\delta; p, d, r)}{\sqrt{n}} + K'(\delta; p, d, r)\zeta(\gamma), \tag{88}$$

where K' and K are given by (63) and (64), and $\zeta(\gamma)$ is given by (32). This implies

$$\|\widehat{M}_{a,b|H}\mathbf{e}_j - M_{a,b|H}\mathbf{e}_{\tau(j)}\|_2 \leq \frac{2K(\delta; p, d, r)}{\sqrt{n}} + 2K'(\delta; p, d, r)\zeta(\gamma). \tag{89}$$

APPENDIX F: MATRIX PERTURBATION ANALYSIS

We borrow the following results on matrix perturbation bounds from [3]. We denote the p-norm of a vector $\vec{v}$ by $\|\vec{v}\|_p$, and the corresponding induced norm of a matrix A by $\|A\|_p := \sup_{\vec{v} \neq \vec{0}} \|A\vec{v}\|_p / \|\vec{v}\|_p$. The Frobenius norm of a matrix A is denoted by $\|A\|_{\mathbb{F}}$. For a matrix $A \in \mathbb{R}^{m \times n}$, let $\kappa(A) := \sigma_1(A)/\sigma_{\min(m,n)}(A)$ (thus $\kappa(A) = \|A\|_2 \cdot \|A^{-1}\|_2$ if A is invertible).

LEMMA 13. *Let $X \in \mathbb{R}^{m \times n}$ be a matrix of rank k. Let $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$ be matrices with orthonormal columns such that $\operatorname{range}(U)$ and $\operatorname{range}(V)$ are spanned by, respectively, the left and right singular vectors of X corresponding to its k largest singular values. Similarly define $\widehat{U} \in \mathbb{R}^{m \times k}$ and $\widehat{V} \in \mathbb{R}^{n \times k}$ relative to a matrix $\widehat{X} \in \mathbb{R}^{m \times n}$. Define $\epsilon_X := \|\widehat{X} - X\|_2$, $\varepsilon_0 := \frac{\epsilon_X}{\sigma_k(X)}$, and $\varepsilon_1 := \frac{\varepsilon_0}{1 - \varepsilon_0}$. Assume $\varepsilon_0 < \frac{1}{2}$. Then*

1. *$\varepsilon_1 < 1$;*
2. *$\sigma_k(\widehat{X}) = \sigma_k(\widehat{U}^\top \widehat{X} \widehat{V}) \geq (1 - \varepsilon_0) \cdot \sigma_k(X) > 0$;*
3. *$\sigma_k(\widehat{U}^\top U) \geq \sqrt{1 - \varepsilon_1^2}$;*
4. *$\sigma_k(\widehat{V}^\top V) \geq \sqrt{1 - \varepsilon_1^2}$;*
5. *$\sigma_k(\widehat{U}^\top X \widehat{V}) \geq (1 - \varepsilon_1^2) \cdot \sigma_k(X)$;*
6. *for any $\widehat{\alpha} \in \mathbb{R}^k$ and $\vec{v} \in \operatorname{range}(U)$, $\|\widehat{U}\widehat{\alpha} - \vec{v}\|_2^2 \leq \|\widehat{\alpha} - \widehat{U}^\top \vec{v}\|_2^2 + \|\vec{v}\|_2^2 \cdot \varepsilon_1^2$.*

LEMMA 14. *Consider the setting and definitions from Lemma 13, and let $Y \in \mathbb{R}^{m \times n}$ and $\widehat{Y} \in \mathbb{R}^{m \times n}$ be given. Define $\varepsilon_2 := \frac{\varepsilon_0}{(1 - \varepsilon_1^2) \cdot (1 - \varepsilon_0 - \varepsilon_1^2)}$ and $\epsilon_Y := \|\widehat{Y} - Y\|_2$. Assume $\varepsilon_0 < \frac{1}{1 + \sqrt{2}}$. Then*

1. *$\widehat{U}^\top \widehat{X} \widehat{V}$ and $\widehat{U}^\top X \widehat{V}$ are both invertible, and $\|(\widehat{U}^\top \widehat{X} \widehat{V})^{-1} - (\widehat{U}^\top X \widehat{V})^{-1}\|_2 \leq \frac{\varepsilon_2}{\sigma_k(X)}$;*
2. *$\|(\widehat{U}^\top \widehat{Y} \widehat{V})(\widehat{U}^\top \widehat{X} \widehat{V})^{-1} - (\widehat{U}^\top Y \widehat{V})(\widehat{U}^\top X \widehat{V})^{-1}\|_2 \leq \frac{\epsilon_Y}{(1 - \varepsilon_0) \cdot \sigma_k(X)} + \frac{\|Y\|_2 \cdot \varepsilon_2}{\sigma_k(X)}$.*

LEMMA 15. *Let $A \in \mathbb{R}^{k \times k}$ be a diagonalizable matrix with k distinct real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{R}$ corresponding to the (right) eigenvectors $\vec{\xi}_1, \vec{\xi}_2, \ldots, \vec{\xi}_k \in \mathbb{R}^k$ all normalized to have $\|\vec{\xi}_i\|_2 = 1$. Let $R \in \mathbb{R}^{k \times k}$ be the matrix whose i^{th} column is $\vec{\xi}_i$. Let $\widehat{A} \in \mathbb{R}^{k \times k}$ be a matrix. Define $\epsilon_A := \|\widehat{A} - A\|_2$, $\gamma_A := \min_{i \neq j} |\lambda_i - \lambda_j|$, and $\varepsilon_3 := \frac{\kappa(R) \cdot \epsilon_A}{\gamma_A}$. Assume $\varepsilon_3 < \frac{1}{2}$. Then there exists a permutation τ on $[k]$ such that the following holds:*

1. *$\widehat{A}$ has k distinct real eigenvalues $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_k \in \mathbb{R}$, and $|\widehat{\lambda}_{\tau(i)} - \lambda_i| \leq \varepsilon_3 \cdot \gamma_A$ for all $i \in [k]$;*

2. *$\widehat{A}$ has corresponding (right) eigenvectors $\widehat{\xi}_1, \widehat{\xi}_2, \dots, \widehat{\xi}_k \in \mathbb{R}^k$, normalized to have $\|\widehat{\xi}_i\|_2 = 1$, which satisfy $\|\widehat{\xi}_{\tau(i)} - \vec{\xi}_i\|_2 \leq 4(k-1) \cdot \|R^{-1}\|_2 \cdot \varepsilon_3$ for all $i \in [k]$;*
3. *the matrix $\widehat{R} \in \mathbb{R}^{k \times k}$ whose i^{th} column is $\widehat{\xi}_{\tau(i)}$ satisfies $\|\widehat{R} - R\|_2 \leq \|\widehat{R} - R\|_{\mathbb{F}} \leq 4k^{1/2}(k-1) \cdot \|R^{-1}\|_2 \cdot \varepsilon_3$.*

Lemma 16. *Let $A_1, A_2, \dots, A_k \in \mathbb{R}^{k \times k}$ be diagonalizable matrices that are diagonalized by the same matrix invertible $R \in \mathbb{R}^{k \times k}$ with unit length columns $\|R\vec{e}_j\|_2 = 1$, such that each A_i has k distinct real eigenvalues:*

$$R^{-1} A_i R = \mathrm{Diag}(\lambda_{i,1}, \lambda_{i,2}, \dots, \lambda_{i,k}).$$

Let $\widehat{A}_1, \widehat{A}_2, \dots, \widehat{A}_k \in \mathbb{R}^{k \times k}$ be given. Define $\epsilon_A := \max_i \|\widehat{A}_i - A_i\|_2$, $\gamma_A := \min_i \min_{j \neq j'} |\lambda_{i,j} - \lambda_{i,j'}|$, $\lambda_{\max} := \max_{i,j} |\lambda_{i,j}|$, $\varepsilon_3 := \frac{\kappa(R) \cdot \epsilon_A}{\gamma_A}$, and $\varepsilon_4 := 4k^{1.5} \cdot \|R^{-1}\|_2^2 \cdot \varepsilon_3$. Assume $\varepsilon_3 < \frac{1}{2}$ and $\varepsilon_4 < 1$. Then there exists a permutation τ on $[k]$ such that the following holds.

1. *The matrix $\widehat{A}_1$ has k distinct real eigenvalues $\widehat{\lambda}_{1,1}, \widehat{\lambda}_{1,2}, \dots, \widehat{\lambda}_{1,k} \in \mathbb{R}$, and $|\widehat{\lambda}_{1,j} - \lambda_{1,\tau(j)}| \leq \varepsilon_3 \cdot \gamma_A$ for all $j \in [k]$.*
2. *There exists a matrix $\widehat{R} \in \mathbb{R}^{k \times k}$ whose j^{th} column is a right eigenvector corresponding to $\widehat{\lambda}_{1,j}$, scaled so $\|\widehat{R}\vec{e}_j\|_2 = 1$ for all $j \in [k]$, such that $\|\widehat{R} - R_\tau\|_2 \leq \frac{\varepsilon_4}{\|R^{-1}\|_2}$, where R_τ is the matrix obtained by permuting the columns of R with τ.*
3. *The matrix $\widehat{R}$ is invertible and its inverse satisfies $\|\widehat{R}^{-1} - R_\tau^{-1}\|_2 \leq \|R^{-1}\|_2 \cdot \frac{\varepsilon_4}{1-\varepsilon_4}$;*
4. *For all $i \in \{2, 3, \dots, k\}$ and all $j \in [k]$, the $(j,j)^{th}$ element of $\widehat{R}^{-1}\widehat{A}_i\widehat{R}$, denoted by $\widehat{\lambda}_{i,j} := \vec{e}_j^{\top} \widehat{R}^{-1} \widehat{A}_i \widehat{R} \vec{e}_j$, satisfies*

$$\begin{aligned} |\widehat{\lambda}_{i,j} - \lambda_{i,\tau(j)}| \leq & \left(1 + \frac{\varepsilon_4}{1-\varepsilon_4}\right) \cdot \left(1 + \frac{\varepsilon_4}{\sqrt{k} \cdot \kappa(R)}\right) \cdot \varepsilon_3 \cdot \gamma_A \\ & + \kappa(R) \cdot \left(\frac{1}{1-\varepsilon_4} + \frac{1}{\sqrt{k} \cdot \kappa(R)} + \frac{1}{\sqrt{k}} \cdot \frac{\varepsilon_4}{1-\varepsilon_4}\right) \cdot \varepsilon_4 \cdot \lambda_{\max}. \end{aligned}$$

If $\varepsilon_4 \leq \frac{1}{2}$, then $|\widehat{\lambda}_{i,j} - \lambda_{i,\tau(j)}| \leq 3\varepsilon_3 \cdot \gamma_A + 4\kappa(R) \cdot \varepsilon_4 \cdot \lambda_{\max}$.

Lemma 17. *Let $V \in \mathbb{R}^{k \times k}$ be an invertible matrix, and let $R \in \mathbb{R}^{k \times k}$ be the matrix whose j^{th} column is $V\vec{e}_j / \|V\vec{e}_j\|_2$. Then $\|R\|_2 \leq \kappa(V)$, $\|R^{-1}\|_2 \leq \kappa(V)$, and $\kappa(R) \leq \kappa(V)^2$.*

References.

[1] Allman, E. S., Rhodes, J. A. and Sullivant, S. (2012). When Do Phylogenetic Mixture Models Mimic Other Phylogenetic Models? *Systematic Biology.*

[2] Anandkumar, A., Hsu, D. and Kakade, S. M. (2012a). A Method of Moments for Mixture Models and Hidden Markov Models. In *Proc. of Conf. on Learning Theory.*

[3] Anandkumar, A., Hsu, D. and Kakade, S. M. (2012b). A Method of Moments for Mixture Models and Hidden Markov Models. *Preprint.*

[4] Anandkumar, A. and Valluvan, R. (2012). Learning Loopy Graphical Models with Latent Variables: Efficient Methods and Guarantees. *Preprint. Available on ArXiv:1203.3887.*

[5] Anandkumar, A., Chaudhuri, K., Hsu, D., Kakade, S. M., Song, L. and Zhang, T. (2011). Spectral Methods for Learning Multivariate Latent Tree Structure. *Preprint, ArXiv 1107.1283.*

[6] Anandkumar, A., Tan, V. Y. F., Huang, F. and Willsky, A. S. (2012a). High-Dimensional Structure Learning of Ising Models: Local Separation Criterion. *Accepted to Annals of Statistics.*

[7] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M. and Telgarsky, M. (2012b). Tensor Decompositions for Learning Latent Variable Models. *Preprint.*

[8] Armstrong, H., Carter, C. K., Wong, K. F. and Kohn, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing* **19** 303–316.

[9] Bader, B. W., Kolda, T. G. et al. (2012). MATLAB Tensor Toolbox Version 2.5. Available online.

[10] Belkin, M. and Sinha, K. (2010). Polynomial learning of distribution families. In *IEEE Annual Symposium on Foundations of Computer Science* 103–112.

[11] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* **55** 77–84.

[12] Brémaud, P. (1999). *Markov Chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer.

[13] Bresler, G., Mossel, E. and Sly, A. (2008). Reconstruction of Markov Random Fields from Samples: Some Observations and Algorithms. In *Intl. workshop APPROX Approximation, Randomization and Combinatorial Optimization* 343–356. Springer.

[14] Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2010). Latent Variable Graphical Model Selection via Convex Optimization. *Preprint. Available on ArXiv.*

[15] Chang, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences* **137** 51–73.

[16] Chen, T., Zhang, N. L. and Wang, Y. (2008). Efficient model evaluation in the search based approach to latent structure discovery. In *4th European Workshop on Probabilistic Graphical Models*.

[17] Choi, M. J., Tan, V. Y. F., Anandkumar, A. and Willsky, A. (2011). Learning Latent Tree Graphical Models. *J. of Machine Learning Research* **12** 1771–1812.

[18] Chow, C. and Liu, C. (1968). Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Tran. on Information Theory* **14** 462–467.

[19] Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. John Wiley & Sons, Inc.

[20] Dasgupta, S. (1999). Learning mixtures of Gaussians. In *Foundations of Computer Science, IEEE Annual Symposium on*.

[21] Daskalakis, C., Mossel, E. and Roch, S. (2006). Optimal phylogenetic reconstruction. In *STOC '06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing* 159–168.

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

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

[24] Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository.

[25] Geiger, D. and Heckerman, D. (1996). Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence* **82** 45–74.

[26] Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1.

[27] Hsu, D., Kakade, S. M. and Zhang, T. (2009). A spectral algorithm for learning hidden markov models. In *Proc. of COLT*.

[28] Jalali, A., Johnson, C. and Ravikumar, P. (2011). On Learning Discrete Graphical Models Using Greedy Methods. In *Proc. of NIPS*.

[29] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51** 455–500.

[30] Kumar, M. P. and Koller, D. (2009). Learning a small mixture of trees. In *Proc. of NIPS*.

[31] Lauritzen, S. L. (1996). *Graphical models: Clarendon Press*. Clarendon Press.

[32] Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

[33] Lazarsfeld, P. F., Merton, R. K. et al. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society* **18** 18–66.

[34] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer.

[35] Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*. JSTOR.

[36] Matsen, F. A. and Steel, M. (2007). Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology* **56** 767–775.

[37] Meila, M. and Jordan, M. I. (2001). Learning with mixtures of trees. *J. of Machine Learning Research* **1** 1–48.

[38] Meinshausen, N. and Bühlmann, P. (2006). High Dimensional Graphs and Variable Selection With the Lasso. *Annals of Statistics* **34** 1436–1462.

[39] Moitra, A. and Valiant, G. (2010). Settling the Polynomial Learnability of Mixtures of Gaussians. In *IEEE Annual Symposium on Foundations of Computer Science*.

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

[41] Mossel, E. and Roch, S. (2011). Phylogenetic Mixtures: Concentration of Measure in the Large-Tree Limit. *Arxiv preprint arXiv:1108.3112*.

[42] Netrapalli, P., Banerjee, S., Sanghavi, S. and Shakkottai, S. (2010). Greedy Learning of Markov Network Structure . In *Proc. of Allerton Conf. on Communication, Control and Computing*.

[43] Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2008). High-dimensional Ising Model Selection Using l1-Regularized Logistic Regression. *Annals of Statistics*.

[44] Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980.

[45] Shamir, O., Sabato, S. and Tishby, N. (2008). Learning and Generalization with the Information Bottleneck. In *Algorithmic Learning Theory. Lecture Notes in Computer Science* **5254** 92-107. Springer.

[46] Spirtes, P. and Meek, C. (1995). Learning Bayesian networks with discrete variables from data. In *Proc. of Intl. Conf. on Knowledge Discovery and Data Mining* 294–299.

[47] Tan, V. Y. F., Anandkumar, A. and Willsky, A. (2011). A Large-Deviation Analysis for the Maximum Likelihood Learning of Tree Structures. *IEEE Tran. on Information Theory* **57** 1714-1735.

[48] Thiesson, B., Meek, C., Chickering, D. and Heckerman, D. (1999). Computationally efficient methods for selecting among mixtures of graphical models. *Bayesian Statistics* **6** 569–576.

[49] Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature medicine* **10** 789–799.

[50] Weitz, D. (2006). Counting independent sets up to the tree threshold. In *Proc. of ACM symp. on Theory of computing* 140–149.

[51] Zhang, N. L. (2004). Hierarchical Latent Class Models for Cluster Analysis. *Journal of Machine Learning Research* **5** 697–723.

[52] Zhang, N. L. and Kočka, T. (2004). Efficient Learning of Hierarchical Latent Class Models. In *ICTAI*.

Electrical Engineering & Computer Science Dept.[¶]
4408 Engineering Hall, Irvine, CA, USA 92697.
E-mail: a.anandkumar@uci.edu; furongh@uci.edu

Microsoft Research New England[‖]
1 Memorial Drive #1,
Cambridge, MA 02142.
E-mail: dahsu@microsoft.com; skakade@microsoft.com