Learning Mixtures of Tree Graphical Models

Animashree Anandkumar
UC Irvine
a.anandkumar@uci.edu

Daniel Hsu
Microsoft Research New England
dahsu@microsoft.com

Furong Huang
UC Irvine
furongh@uci.edu

Sham M. Kakade
Microsoft Research New England
skakade@microsoft.com

Abstract

We consider unsupervised estimation of mixtures of discrete graphical models, where the class variable is hidden and each mixture component can have a potentially different Markov graph structure and parameters over the observed variables. We propose a novel method for estimating the mixture components with provable guarantees. Our output is a tree-mixture model which serves as a good approximation to the underlying graphical model mixture. The sample and computational requirements for our method scale as $\mathrm{poly}(p, r)$, for an r-component mixture of p-variate graphical models, for a wide class of models which includes tree mixtures and mixtures over bounded degree graphs.

Keywords: Graphical models, mixture models, spectral methods, tree approximation.

1 Introduction

The framework of graphical models allows for parsimonious representation of high-dimensional data by encoding statistical relationships among the given set of variables through a graph, known as the *Markov graph*. Recent works have shown that a wide class of graphical models can be estimated efficiently in high dimensions [1–3]. However, frequently, graphical models may not suffice to explain all the characteristics of the observed data. For instance, there may be latent or hidden variables, which can influence the observed data in myriad ways.

In this paper, we consider latent variable models, where a latent variable can alter the relationships (both structural and parametric) among the observed variables. In other words, we posit the observed data as being generated from a mixture of graphical models, where each mixture component has a potentially different Markov graph structure and parameters. The choice variable corresponding to the selection of the mixture component is hidden. Such a class of graphical model mixtures can incorporate *context-specific dependencies*, and employs multiple graph structures to model the observed data. This leads to a significantly richer class of models, compared to graphical models.

Learning graphical model mixtures is however far more challenging than learning graphical models. State-of-art theoretical guarantees are mostly limited to mixtures of product distributions, also known as *latent class models* or *naïve Bayes models*. These models are restrictive since they do not allow for dependencies to exist among the observed variables in each mixture component. Our work significantly generalizes this class and allows for general Markov dependencies among the observed variables in each mixture component.

The output of our method is a tree mixture model, which is a good approximation for the underlying graphical model mixture. The motivation behind fitting the observed data to a tree mixture is clear: inference can be performed efficiently via belief propagation in each of the mixture components.

See [4] for a detailed discussion. Thus, a tree mixture model offers a good tradeoff between using single-tree models, which are too simplistic, and general graphical model mixtures, where inference is not tractable.

1.1 Summary of Results

We propose a novel method with provable guarantees for unsupervised estimation of discrete graphical model mixtures. Our method has mainly three stages: graph structure estimation, parameter estimation, and tree approximation. The first stage involves estimation of the *union graph* structure $G_{\cup} := \cup_h G_h$, which is the union of the Markov graphs $\{G_h\}$ of the respective mixture components. Our method is based on a series of rank tests, and can be viewed as a generalization of conditional-independence tests for graphical model selection (e.g. [1, 5, 6]). We establish that our method is efficient (in terms of computational and sample complexities), when the underlying union graph has sparse vertex separators. This includes tree mixtures and mixtures with bounded degree graphs. The second stage of our algorithm involves parameter estimation of the mixture components. In general, this problem is NP-hard. We provide conditions for tractable estimation of pairwise marginals of the mixture components. Roughly, we exploit the conditional-independence relationships to convert the given model to a series of mixtures of product distributions. Parameter estimation for product distribution mixture has been well studied (e.g. [7–9]), and is based on *spectral decompositions* of the observed moments. We leverage on these techniques to obtain estimates of the pairwise marginals for each mixture component. The final stage for obtaining tree approximations involves running the standard Chow-Liu algorithm [10] on each component using the estimated pairwise marginals of the mixture components.

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. Note that if the underlying model is a tree mixture, we correctly recover the tree structures of the mixture components. The sample and computational complexities of our method scale as $\mathrm{poly}(p, r)$, for an r-component mixture of p-variate graphical models, when the union graph has sparse vertex separators between any node pair. This includes tree mixtures and mixtures with bounded degree graphs. To the best of our knowledge, this is the first work to provide provable learning guarantees for graphical model mixtures. Our algorithm is also efficient for practical implementation and some preliminary experiments suggest an advantage over EM with respect to running times and accuracy of structure estimation of the mixture components. Thus, our approach for learning graphical model mixtures has both theoretical and practical implications.

1.2 Related Work

Graphical Model Selection: Graphical model selection is a well studied problem starting from the seminal work of Chow and Liu [10] for finding the maximum-likelihood tree approximation of a graphical model. Works on high-dimensional loopy graphical model selection are more recent. They can be classified into mainly two groups: non-convex local approaches [1, 2, 6] and those based on convex optimization [3, 11]. However, these works are not directly applicable for learning mixtures of graphical models. Moreover, our proposed method also provides a new approach for graphical model selection, in the special case when there is only one mixture component.

Learning Mixture Models: Mixture models have been extensively studied, and there are a number of recent works on learning high-dimensional mixtures, e.g. [12, 13]. 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. Spectral methods are applicable for parameter estimation in mixtures of discrete product distributions [7] and more generally for latent trees [8] and general linear multiview mixtures [9]. We leverage on these techniques for parameter estimation in models beyond product distribution mixtures.

2 Graphical Models and their Mixtures

A graphical model is a family of multivariate distributions Markov on a given undirected graph [14]. 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 $\mathcal{Y}$. Let $d := |\mathcal{Y}|$ denote the cardinality of the set and $p := |V|$ denote the number of variables. 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 P satisfies the *global Markov property* for all disjoint sets $A, B \subset V$

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

where the set $\mathcal{S}(A, B; G)$ is a *node separator*[1]between A and B, and $\mathcal{N}[A]$ denotes the closed neighborhood of A (i.e., including A).

Mixtures of discrete graphical models is considered. Let H denote the discrete hidden choice variable corresponding to selection of a different mixture components, taking values in $[r] := \{1, \ldots, r\}$ and let $\mathbf{Y}$ denote the observed random vector. 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 $P(\mathbf{y}|H = h)$ of each mixture component. Given n i.i.d. samples $\mathbf{y}^n = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top$ from $P(\mathbf{y})$, our goal is to find a tree approximation for each mixture component $\{P(\mathbf{y}|H = h)\}_h$. We do not assume any knowledge of the mixing weights $\boldsymbol{\pi}_H$ or Markov graphs $\{G_h\}_h$ or parameters of the mixture components $\{P(\mathbf{y}|H = h)\}_h$. Moreover, since the variable H is latent, we do not a priori know the mixture component from which a sample is drawn. Thus, a major challenge is in decomposition of the observed statistics into the component models, and we tackle this in three main stages. First, we estimate the union graph $G_\cup := \cup_{h=1}^r G_h$, which is the union of the Markov graphs of the components. We then use this graph estimate $\widehat{G}_\cup$ to obtain the pairwise marginals of the respective mixture components $\{P(\mathbf{y}|H = h)\}_h$. Finally, Chow-Liu algorithm provides tree approximations $\{T_h\}_h$ of each mixture components.

3 Estimation of the Union of Component Graphs

We propose a novel method for learning graphical model mixtures by first estimating the union graph $G_\cup = \cup_{h=1}^r G_h$, which is the union of the graphs of the components. In the special case when $G_h \equiv G_\cup$, this gives the graph estimate of the components. However, the union graph $G_\cup$ appears to have no direct relationship with the marginalized model $P(\mathbf{y})$. We first provide intuitions on how $G_\cup$ relates to the observed statistics.

Intuitions: We first establish the simple result that the union graph $G_\cup$ satisfies Markov property in each mixture component. Recall that $\mathcal{S}(u, v; G_\cup)$ denotes a vertex separator between nodes u and v in $G_\cup$.

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

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

Proof: The separator set in $G_\cup$, denoted by $S := \mathcal{S}(u, v; G_\cup)$, is also a vertex separator for u and v in each of the component graphs 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 each $h \in [r]$, and the above result follows. □

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 (1). The main 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 $S := \mathcal{S}(u, v; G_\cup)$, 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 $k \in \mathcal{Y}^{|S|}$.*

[1] 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.

The proof is given in [15]. Thus, the effect of marginalizing the choice variable H is seen in the rank of the observed probability matrices $M_{u,v,\{S;k\}}$. 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.

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. We consider graph families where a vertex separator can be found for any $(u,v) \notin G_\cup$ with size at most η. Under our framework, the hardness of learning a union graph is parameterized by η. Similar observations have been made before for graphical model selection [1].

There are many natural families where η is small:

1. If $G_\cup$ is trivial (i.e., no edges) then $\eta = 0$, 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 $\eta = 1$, since there is a unique path between any two nodes on a tree.
3. For an arbitrary r-component tree mixture, $G_\cup = \cup_h T_h$ where each component is a tree distribution, we have that $\eta \leq r$ (since for any node pair, there is a unique path in each of the r trees $\{T_h\}$, and separating the node pair in each T_h also separates them on $G_\cup$).
4. For an arbitrary mixture of bounded degree graphs, we have $\eta \leq \sum_{h\in[r]} \Delta_h$, where Δ_h is the maximum degree in G_h, i.e., the Markov graph corresponding to component $\{H = h\}$.

In general, η depends on the respective bounds η_h for the component graphs G_h, as well as the extent of their overlap. In the worst case, η can be as high as $\sum_{h\in[r]} \eta_h$, while in the special case when $G_h \equiv G_\cup$, the bound remains the same $\eta_h \equiv \eta$. Note that for a general graph $G_\cup$ with *treewidth* $\mathrm{tw}(G_\cup)$ and maximum degree $\Delta(G_\cup)$, we have that $\eta \leq \min(\Delta(G_\cup), \mathrm{tw}(G_\cup))$.

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 in $G_\cup$ and $\xi_{n,p}$ is a threshold on the singular values.

$\mathrm{Rank}(A;\xi)$ 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

$$\min_{\substack{S\subset V\setminus\{u,v\} \\ |S|\leq\eta}} \mathrm{Rank}(\widehat{M}^n_{u,v,\{S;k\}}; \xi_{n,p}) > r, \tag{2}$$

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

Rank Test: Based on the above observations, we propose a rank test to estimate $G_\cup := \cup_{h\in[r]} G_h$, the union graph in Algorithm 1. 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 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 $|S| \leq \eta$. 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.

3.1 Analysis of the Rank Test

We now provide guarantees for the success of rank tests in estimating $G_\cup$. As noted before, we require that the number of components r and the dimension d of each node variable satisfy $d > r$. Moreover, we assume bounds on the size of separator sets, $\eta = O(1)$. This includes tree mixtures and mixtures over bounded degree graphs. In addition, the following parameters determine the success of the rank tests.

(A1) **Rank condition for neighbors:** Let $M_{u,v,\{S;k\}} := [P(Y_u = i, Y_v = j, \mathbf{Y}_S = k)]_{i,j}$ and

$$\rho_{\min} := \min_{\substack{(u,v)\in G_\cup, |S|\leq\eta \\ S\subset V\setminus\{u,v\}}} \max_{k\in\mathcal{Y}^{|S|}} \sigma_{r+1}\left(M_{u,v,\{S;k\}}\right) > 0, \tag{3}$$

where $\sigma_{r+1}(\cdot)$ denotes the $(r+1)^{\text{th}}$ singular value, when the singular values are arranged in the descending order $\sigma_1(\cdot) \geq \sigma_2(\cdot) \geq \ldots \sigma_d(\cdot)$. This ensures that the probability matrices for neighbors $(u, v) \in G_\cup$ have (effective) rank of at least $r + 1$, and thus, the rank test can correctly distinguish neighbors from non-neighbors. 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).

(A2) **Choice of threshold ξ:** The threshold ξ on singular values is chosen as $\xi := \frac{\rho_{\min}}{2}$.

(A3) **Number of Samples:** Given $\delta \in (0, 1)$, the number of samples n satisfies

$$n > n_{\text{Rank}}(\delta; p) := \max\left(\frac{1}{t^2}\left(2\log p + \log\delta^{-1} + \log 2\right), \left(\frac{2}{\rho_{\min} - t}\right)^2\right), \tag{4}$$

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

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

Theorem 1 (Success of Rank Tests) *The* $\mathsf{RankTest}(\mathbf{y}^n; \xi, \eta, r)$ *recovers the correct graph* $G_\cup$*, which is the union of the component Markov graphs, under (A1)–(A3) with probability at least* $1 - \delta$.

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) *Given* n *i.i.d. samples* $\mathbf{y}^n$*, the* $\mathsf{RankTest}(\mathbf{y}^n; \xi, \eta, 1)$ *is structurally consistent under (A1)–(A3) with probability at least* $1 - \delta$.

Remarks: Thus, the rank test is also applicable for graphical model selection. Previous works (see Section 1.2) have proposed tests based on conditional independence, using either conditional mutual information or conditional variation distances, see [1, 6]. The rank test above is thus an alternative test for conditional independence in graphical models, resulting in graph structure estimation. In addition, it extends naturally to estimation of union graph structure of mixture components. Our above result establishes that our method is also efficient in high dimensions, since it only requires logarithmic samples for structural consistency $(n = \Omega(\log p))$.

4 Parameter Estimation of Mixture Components

Having obtained an estimate of the union graph $G_\cup$, we now describe a procedure for estimating parameters of the mixture components $\{P(\mathbf{y}|H = h)\}$. Our method is based on spectral decomposition, proposed previously for mixtures of product distributions [7–9]. We recap it briefly below and then describe how it can be adapted to the more general setting of graphical model mixtures.

Recap of Spectral Decomposition in Mixtures of Product Distributions: Consider the case where $V = \{u, v, w\}$, and $Y_u \perp\!\!\!\perp Y_v \perp\!\!\!\perp Y_w | H$. For simplicity assume that $d = r$, i.e., the hidden and observed variables have the same dimension. This assumption will be removed subsequently. 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:

Lemma 2 (Mixture of Product Distributions) *Given the above model, let* $\boldsymbol{\lambda}^{(k)} = [\lambda_1^{(k)}, \ldots, \lambda_d^{(k)}]^\top$ *be the column vector with the* d *eigenvalues given by*

$$\boldsymbol{\lambda}^{(k)} := \mathsf{Eigenvalues}\left(M_{u,v,\{w;k\}} M_{u,v}^{-1}\right), \quad k \in \mathcal{Y}. \tag{5}$$

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

$$M_{w|H} := [P(Y_w = i|H = j)]_{i,j} = \Lambda^\top. \tag{6}$$

For the proof of the above result and for the general algorithm (when $d \geq r$), see [9]. 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 (5) 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 (5) for all triplets, we obtain a consistent ordering of the hidden labels over all triplet decompositions.

Parameter Estimation in Graphical Model Mixtures: We now adapt the above procedure for estimating components of a general graphical model mixture. 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,

$$Y_u \perp\!\!\!\perp Y_v \perp\!\!\!\perp Y_w | \mathbf{Y}_{S_{uvw}}, H, \quad \forall u, v, w : (u, v), (v, w), (w, u) \notin G_\cup. \tag{7}$$

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} and using the spectral decomposition described in Lemma 2, and the procedure can be repeated over different triplets $\{u, v, w\}$.

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

$$\exists u_* \in V, s.t. \quad Y_{u_*} \perp\!\!\!\perp \mathbf{Y}_{V \setminus u_*} | H, \tag{8}$$

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

Condition (8) is required 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 (8) for identifiability. However, our setting is a significant generalization over the mixtures of product distributions, where (8) 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 [9] for details. Thus, we obtain estimates of the pairwise marginals of the mixture components. For details on implementation, refer to [15].

4.1 Analysis and Guarantees

In addition to (A1)–(A3) in Section 3.1 to guarantee correct recovery of $G_\cup$ and the conditions discussed above, the success of parameter estimation depends on the following quantities:

(A4) **Non-degeneracy:** 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.

(A5) **Spectral Bounds and Number of Samples:** Refer to various spectral bounds used to obtain $K(\delta; p, d, r)$ in **(??)** in [15], where $\delta \in (0, 1)$ is fixed. Given any fixed $\epsilon \in (0, 1)$, assume that the number of samples satisfies

$$n > n_{\text{spect}}(\delta, \epsilon; p, d, r) := \frac{4K^2(\delta; p, d, r)}{\epsilon^2}. \tag{9}$$

Note that (A4) is a natural condition required for success of spectral decomposition and has been previously imposed for learning product distribution mixtures [7–9]. Moreover, when (A4) does not hold, i.e., when the matrices are not full rank, parameter estimation is computationally at least as hard as learning parity with noise, which is conjectured to be computationally hard [8]. Condition (A5) is required for learning product distribution mixtures [9], and we inherit it here.

We now provide guarantees for estimation of pairwise marginals of the mixture components. Let $\|\cdot\|_2$ on a vector denote the ℓ_2 norm.

Theorem 2 (Parameter Estimation of Mixture Components) *Under the assumptions (A1)–(A5), the spectral decomposition method 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*

$$\|\widehat{P}^{\text{spect}}(Y_a, Y_b|H = h) - P(Y_a, Y_b|H = \tau(h))\|_2 \leq \epsilon, \tag{10}$$

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

Remark: Recall that p denotes the number of variables, r is the number of mixture components, d is the dimension of each node variable and η is the bound on separator sets between any node pair in the union graph. We establish that $K(\delta; p, d, r)$ is $O\left(p^{2\eta+2} d^{2\eta} r^5 \delta^{-1} \operatorname{poly}\log(p, d, r, \delta^{-1})\right)$ in [15]. Thus, we require the number of samples in (9) scaling as $n = \Omega\left(p^{4\eta+4} d^{4\eta} r^{10} \delta^{-2} \epsilon^{-2} \operatorname{poly}\log(p, d, r, \delta^{-1})\right)$. Since we consider models where $\eta = O(1)$ is a small constant, this implies that we have a polynomial sample complexity in p, d, r.

Tree Approximation of Mixture Components: The final step involves using the estimated pairwise marginals of each component $\{\widehat{P}^{\text{spect}}(Y_a, Y_b|H = h)\}$ to obtain tree approximation of the component via Chow-Liu algorithm [10]. 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.

(A6) **Separation of Mutual Information:** Let T_h denote the maximum-likelihood tree approximation corresponding to the model $P(\mathbf{y}|H = h)$ when exact statistics are input and let

$$\vartheta := \min_{h \in [r]} \min_{(a,b) \notin T_h} \min_{(u,v) \in \text{Path}(a,b;T_h)} \left(I(Y_u, Y_v|H = h) - I(Y_a, Y_b|H = h)\right), \tag{11}$$

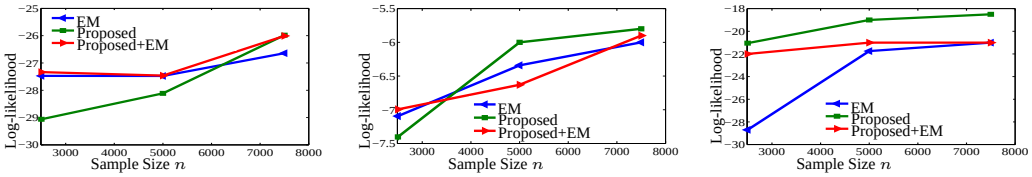
where $\text{Path}(a, b; T_h)$ denotes the edges along the path connecting a and b in T_h. Intuitively ϑ denotes the "bottleneck" where errors are most likely to occur in tree structure estimation. See [16] for a detailed discussion.

(A7) **Number of Samples:** Given ϵ^{tree} defined in [15], we require

$$n > n_{\text{spect}}(\delta, \epsilon^{\text{tree}}; p, d, r), \tag{12}$$

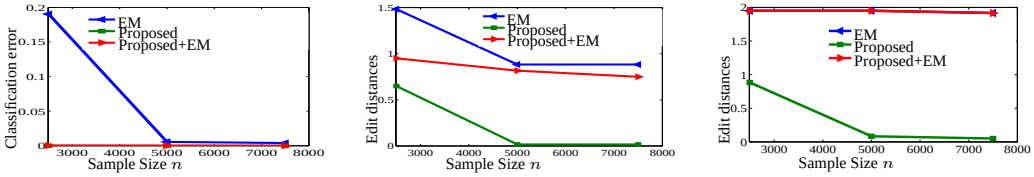
where n_{spect} is given by (9). Intuitively, ϵ^{tree} provides the bound on distortion of the estimated pairwise marginals of the mixture components, required for correct estimation of tree approximations, and depends on ϑ in (11).

Theorem 3 (Tree Approximations of Mixture Components) *Under (A1)–(A7), 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.*



(a) Overall likelihood of the mixture (b) Conditional likelihood of strong component (c) Conditional likelihood of weak component

Figure 1: Performance of the proposed method, EM and EM initialized with the proposed method output on a tree mixture with two components.



(a) Classification error (b) Strong component edit distance (c) Weak component edit distance

Figure 2: Classification error and normalized edit distances of the proposed method, EM and EM initialized with the proposed method output on the tree mixture.

5 Experiments

Experimental results are presented on synthetic data. We estimate the graph using proposed algorithm and compare the performance of our method with EM [4]. Comprehensive results based on the normalized edit distances and log-likelihood scores between the estimated and the true graphs are presented. We generate samples from a mixture over two different trees ($r = 2$) with mixing weights $\pi = [0.7, 0.3]$ using Gibbs sampling. 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 $n \in [2.5 \times 10^3, 10^4]$. 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. 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 performance of the proposed method is compared with EM. We consider 10 random initializations of EM and run it to convergence. We also evaluated EM by utilizing proposed result as the initial point (referred to as Proposed+EM in the figures). We observe in Fig 1a that the overall likelihood under our method is comparable with EM. Intuitively this is because EM attempts to maximize the overall likelihood. However, our algorithm has significantly superior performance with respect to the edit distance which is the error in estimating the tree structure in the two components, as seen in Fig 2. In fact, EM never manages to recover the structure of the weak components(i.e., the component with weak correlations). Intuitively, this is because EM uses the overall likelihood as criterion for tree selection. 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. We also observe in Fig 1b and Fig 1c, that our proposed method has superior performance in terms of conditional likelihood for both the components. Classification error is evaluated in Fig 2a. We could get smaller classification errors than EM method.

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.

Acknowledgements: The first author is supported in part by the NSF Award CCF-1219234, AFOSR Award FA9550-10-1-0310, ARO Award W911NF-12-1-0404, and setup funds at UCI. The third author is supported by the NSF Award 1028394 and AFOSR Award FA9550-10-1-0310.

References

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

[2] A. Jalali, C. Johnson, and P. Ravikumar. On learning discrete graphical models using greedy methods. In *Proc. of NIPS*, 2011.

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

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

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

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

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

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

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

[10] C. Chow and C. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Tran. on Information Theory*, 14(3):462–467, 1968.

[11] N. Meinshausen and P. Bühlmann. High Dimensional Graphs and Variable Selection With the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

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

[13] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of gaussians. In *IEEE Annual Symposium on Foundations of Computer Science*, 2010.

[14] S.L. Lauritzen. *Graphical models: Clarendon Press*. Clarendon Press, 1996.

[15] A. Anandkumar, D. Hsu, and S.M. Kakade. Learning High-Dimensional Mixtures of Graphical Models. *Preprint. Available on ArXiv:1203.0697*, Feb. 2012.

[16] V.Y.F. Tan, A. Anandkumar, and A. Willsky. A Large-Deviation Analysis for the Maximum Likelihood Learning of Tree Structures. *IEEE Tran. on Information Theory*, 57(3):1714–1735, March 2011.