Infinite Mixtures of Trees

Sergey Kirshner SERGEY@CS.UALBERTA.CA
AICML, Department of Computing Science, University of Alberta, Edmonton, Canada T6G 2E8

Padhraic Smyth SMYTH@ICS.UCI.EDU
Department of Computer Science, University of California, Irvine CA 92697-3425 USA

Abstract

Finite mixtures of tree-structured distributions have been shown to be efficient and effective in modeling multivariate distributions. Using Dirichlet processes, we extend this approach to allow countably many tree-structured mixture components. The resulting Bayesian framework allows us to deal with the problem of selecting the number of mixture components by computing the posterior distribution over the number of components and integrating out the components by Bayesian model averaging. We apply the proposed framework to identify the number and the properties of predominant precipitation patterns in historical archives of climate data.

1. Introduction

Graphical models play a central role in learning of probability distributions from multivariate categorical data. Tree-structured models hold a special place among graphical models since fundamental operations such as learning and inference are much more efficient for tree-structured models than for graphs with loops. Learning of both the parameters and the edges for a tree-structured distribution from a complete multivariate data set can be done *optimally*[1] in time proportional to the size of the data set and the square of the dimensionality of the data vectors (Chow & Liu, 1968). On the other hand, if loops are allowed in a graph, learning of the optimal structure becomes NP-hard (Chickering, 1996; Srebro, 2003), and even learning of an approximate structure is computationally challenging as the estimation of the structure and the parameters cannot be performed simultaneously. Even for the cases when the tree structures are not sufficiently expressive to capture the complex interactions between the variables, one can still benefit from the computational efficiency of trees by approximating the joint distribution with a *mixture* of trees (MT) (Grim, 1984; Meilă & Jordan, 2000). While computationally not as efficient as individual Chow-Liu trees (and potentially not even optimal within the relevant class of models), mixtures can provide useful approximations at a fraction of the computational cost of learning general Markov or Bayesian networks.

Tree-structured distributions over categorical variables have another useful property: they can be described in a fully Bayesian framework with a conjugate prior (Meilă & Jaakkola, 2006). This conjugate prior defines a distribution over all possible tree structures and the parameters given tree structures. This is achieved by providing hyperparameters for each pair of variables and decomposing the posterior distribution over the structure and parameters into a product of distributions defined only on the edges of the corresponding tree. What is also remarkable is that the normalization constant for the prior distribution can be computed in closed form without prohibitive computational cost. In contrast, for general Bayesian and Markov networks no such general prior is known, and more primitive priors are sometimes used instead (Cooper & Herskovits, 1992; Lam & Bacchus, 1994). Mixtures of trees can be presented in a fully Bayesian framework with flexible priors for trees allowing one to estimate the uncertainty about the structure and parameters and possibly even integrating out (albeit numerically, not analytically) this uncertainty.

In this paper, we propose going a step beyond the standard mixtures of trees. In learning of finite mixtures, the number of mixture components is usually

[1] Within the class of tree-structured distributions.

Appearing in *Proceedings of the 24^{th} International Conference on Machine Learning*, Corvallis, OR, 2007.

set beforehand. While there are many approaches to selection of the number of components, such as scoring (e.g., BIC, DIC) or cross-validation, when the true distribution generating data is not itself a finite mixture model, these model selection methods may not yield a clear choice in terms of the number of components. Also, due to the limited amount of available data, only a small number of components may have manifested themselves in the data, i.e., the more data that is available the more components may be appropriate to model the observed data. It is possible to sidestep these issues with the help of a Dirichlet process mixture model (DPMM). The DPMM is a generalization of a finite mixture model allowing (in the limit) countably many mixture components by replacing a multinomial distribution over the mixture components with a Dirichlet process. DPMMs allow the number of mixture components to grow with the amount of data available, something not easily doable with finite mixture models. In addition, DPMMs can be formulated within a fully Bayesian framework allowing for direct estimation of the posterior probability of the number of mixture components.

Our main contribution in this paper is the extension of finite mixtures of Chow-Liu trees to Dirichlet process mixtures of trees (DPMT) in a Bayesian framework. The framework of the new model can be used to decide how many mixture components to use if one wants to build a tree-based finite mixture model, to sample from the posterior over the parameters, or to perform model averaging over both the parameters and different numbers of mixture components (potentially reaping benefits from removing both sources of uncertainty). We demonstrate the effectiveness of the model both on simulated data and on historical data of daily rainfall occurrence for networks of rain stations.

The paper is structured as follows: we recap Chow-Liu trees and a finite mixture of Chow-Liu trees (Section 2), then describe a fully Bayesian framework for Chow-Liu trees and finite mixtures of trees (Section 3) briefly discussing DPMMs and introduce infinite mixtures of trees (Section 3.3). We then analyze the use of the model on both synthetic and real-application data (Section 4). Finally, we summarize our contributions and mention several possible directions. (Section 5).

2. Trees and Mixtures of Trees

In this section, we describe the basic framework for learning with finite mixtures of trees. This framework will be extended further in the following sections.

Assume we are given a set of d-dimensional complete discrete-valued vector observations $\mathcal{D} = \left\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N\right\}$. We assume that vectors $\boldsymbol{x}^n$ ($n = 1, \ldots, N$) were sampled i.i.d. from some unknown joint distribution over d random variables. We will denote the set of these d variables by $\mathcal{V}$ ($|\mathcal{V}| = d$) and refer to individual variables as X_v for $v \in \mathcal{V}$. Let $\mathcal{X}_v$ be the range of X_v, and let $r_v = |\mathcal{X}_v|$ be the number of values that X_v takes on. Let $\boldsymbol{X} = (X_v)_{v\in\mathcal{V}}$ with range $\mathcal{X} = \bigotimes_{v\in\mathcal{V}} \mathcal{X}_v$ be the vector of all d variables X_v defined over $|\mathcal{X}| = \prod_{v\in\mathcal{V}} r_v$ possible values of the Cartesian product of ranges of X_v. To simplify notation, an assignment of a random variable will be denoted by a lower case letter, e.g., $\boldsymbol{X} = \boldsymbol{x}$ will be denoted simply by $\boldsymbol{x}$.

In the maximum likelihood (ML) framework, we want to find a distribution $T(\boldsymbol{X})$ maximizing the log-likelihood of the data or alternatively minimize Kullback-Leibler divergence (relative entropy) $KL(P \,\|\, T)$ where $P(\boldsymbol{X})$ is an empirical distribution as observed from $\mathcal{D}$, i.e., $P(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} \delta_{\boldsymbol{x}^n}(\boldsymbol{x})$ with $\delta_{\boldsymbol{x}'}(\boldsymbol{x}) = 1$ iff $\boldsymbol{x} = \boldsymbol{x}'$.

First, we concentrate on finding T from a family of distributions where the dependence structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be represented as a tree. T has two related sets of parameters: the set of edges $\mathcal{E}$ making up a tree, and the set of parameters $\boldsymbol{\theta}_E = \{\boldsymbol{\theta}_{uv}\}_{\{u,v\}\in\mathcal{E}}$ defining multinomial distributions for pairs of variables corresponding to the edges in the tree. Let $\theta_{uv}(i,j) = T(X_u = i, X_v = j)$ and $\theta_v(i) = T(X_v = i)$[2]. Note that for any $v \in \mathcal{V}$ such that $\{u,v\} \in \mathcal{E}$ and any $i \in \mathcal{X}_u$, $\sum_{j\in\mathcal{X}_v} \theta_{uv}(i,j) = \theta_u(i)$. Defining $\boldsymbol{\theta}_T = \{\mathcal{E}, \boldsymbol{\theta}_E\}$, T can be formulated as

$$T(\boldsymbol{x}|\boldsymbol{\theta}_T) = \left(\prod_{v\in\mathcal{V}} \theta_v(x_v)\right)\left(\prod_{\{u,v\}\in\mathcal{E}} \frac{\theta_{uv}(x_u, x_v)}{\theta_u(x_u)\,\theta_v(x_v)}\right). \tag{1}$$

Chow and Liu (1968) showed that the relative entropy is maximized when $\mathcal{E}$ is constructed by solving a maximum spanning tree problem (Cormen et al., 1990) on a graph with nodes $\mathcal{V}$ and pairwise weights equal to corresponding mutual information values $MI_P(X_u, X_v)$:

$$MI_P(X_u, X_v) = \sum_{X_u}\sum_{X_v} P(x_u, x_v) \log \frac{P(x_u, x_v)}{P(x_u)\,P(x_v)}.$$

The solution for $\boldsymbol{\theta}_E$ is just the set of corresponding bivariate marginals of P:

$$\begin{array}{lrcl} \forall \{u,v\} \in \mathcal{E} & \theta_{uv}(i,j) & = & P(X_u = i, X_v = j)\,; \\ \forall v \in \mathcal{V} & \theta_v(i) & = & P(X_v = i)\,. \end{array}$$

[2] This notation is adopted from Meilă and Jaakkola (2006).

The Chow-Liu algorithm has $\mathcal{O}\left(Nd^2r_{max}^2\right)$ time complexity where $r_{max} = \max_{v\in\mathcal{V}}\{r_v\}$ is the largest number of possible values over all variables. The distribution T is parametrized using only $\sum_{v\in\mathcal{V}} r_v - d + \sum_{u,v\in\mathcal{E}_T}(r_u-1)(r_v-1)$ free parameters[3] ($\mathcal{O}\left(dr_{max}^2\right)$) compared to $\prod_{v\in\mathcal{V}} r_v - 1$ free parameters ($\mathcal{O}\left(r_{max}^d\right)$) needed to specify a full probability distribution on $\mathcal{X}$.

2.1. Finite Mixture of Trees

Not all joint distributions can be approximated well by Chow-Liu trees. The expressive power of the model in the previous section can be improved by using a convex combination (a mixture) of tree-structured components (Meilă & Jordan, 2000). The K-component mixture of trees $Q(\boldsymbol{X})$ is defined as

$$Q(\boldsymbol{x}) = \sum_{i=1}^{K} \pi_i T^i(\boldsymbol{x})$$

where each T^i is of the form of Equation (1), $\pi_i \geq 0$, and $\sum_{i=1}^{K}\pi_i = 1$. While one can require that all K components share the dependence tree structure (Meilă & Jordan, 2000), we do not enforce this restriction on structures in this paper. The set of parameters $\boldsymbol{\Theta}_M$ for Q consists of a vector of mixture component probabilities $\boldsymbol{\pi}$ (parameters for a multinomial distribution), and K sets of parameters $\boldsymbol{\theta}_T^i = \left\{\mathcal{E}^i, \boldsymbol{\theta}^i\right\}$ for tree components. The measure G over the parameters used to generate $\mathcal{D}$ can be written as

$$G = \sum_{i=1}^{K} \pi_i \delta_{\boldsymbol{\theta}_T^i}$$

where δ is Dirac delta function. The generative model for $\mathcal{D}$ can then be described as:

$$\begin{aligned} n = 1,\ldots,N: \ \boldsymbol{\theta}_n &\sim G \\ n = 1,\ldots,N: \ \boldsymbol{x}^n &\sim T(\cdot|\boldsymbol{\theta}_n). \end{aligned}$$

While the maximum likelihood estimation of $\boldsymbol{\Theta}_M$ cannot be performed in closed form, they can be estimated using the Expectation-Maximization (EM) (Dempster et al., 1977) algorithm that treats the mixture components for all examples in the training set as hidden variables. (See Meilă and Jordan (2000) for details.)

3. Bayesian Framework for Mixtures of Chow-Liu Trees

Instead of computing point estimates of parameters (e.g., via ML or maximum aposteriori (MAP)), it may be desirable to use a posterior distribution over the parameters given the data or to integrate over (or to average) the parameters in order to compute conditional probability distributions for new data points. This approach is especially useful when the posterior has multiple modes or if the posterior probability surface is flat near its modes. While the posterior distribution can rarely be described analytically, one can often efficiently obtain samples from it, and use these samples to approximate the true posterior.

We can place the mixture of trees from Section 2.1 into a Bayesian framework by introducing priors over the set of parameters $\boldsymbol{\pi}$ and over sets of parameters $\boldsymbol{\theta}_T^i$. Let $p(\boldsymbol{\pi}|\alpha) = D\left(\frac{\alpha}{K},\ldots,\frac{\alpha}{K}\right)$ be a symmetric Dirichlet prior for the parameters of the mixing component multinomial distribution (the Dirichlet is a conjugate prior for multinomial distributions), and let $p\left(\boldsymbol{\theta}_T^i|\boldsymbol{\Theta}_\theta\right)$ be the prior for the parameters for the Chow-Liu tree components (assuming all components have the same prior distribution). The generative model for $\mathcal{D}$ is then (as shown in Figure 1)

$$\begin{aligned} i = 1,\ldots,K: \ \boldsymbol{\theta}_T^i &\sim G_0 = p(\cdot|\boldsymbol{\Theta}_\theta) \\ \boldsymbol{\pi} &\sim D\left(\frac{\alpha}{K},\ldots,\frac{\alpha}{K}\right) \\ G &= \sum_{i=1}^{K}\pi_i\delta_{\boldsymbol{\theta}_T^i} \\ n = 1,\ldots,N: \ \boldsymbol{\theta}_n &\sim G \\ n = 1,\ldots,N: \ \boldsymbol{x}^n &\sim T(\cdot|\boldsymbol{\theta}_n). \end{aligned}$$

Then the posterior distribution over the parameters of the mixture of trees can be written as

$$\begin{aligned} p(\boldsymbol{\Theta}_M|\mathcal{D},\alpha,\boldsymbol{\Theta}_\theta) &= \frac{p(\boldsymbol{\Theta}_M,\mathcal{D}|\alpha,\boldsymbol{\Theta}_\theta)}{p(\mathcal{D}|\alpha,\boldsymbol{\Theta}_\theta)} \\ &= \frac{p(\boldsymbol{\pi}|\alpha)\prod_{i=1}^{K}p\left(\boldsymbol{\theta}_T^i|\boldsymbol{\Theta}_\theta\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{K}\pi_iT^i(\boldsymbol{x}^n)\right)}{\int p(\boldsymbol{\pi}|\alpha)\prod_{i=1}^{K}P\left(\boldsymbol{\theta}_T^i|\boldsymbol{\Theta}_\theta\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{K}\pi_iT^i(\boldsymbol{x}^n)\right)\mathrm{d}\boldsymbol{\Theta}_M}. \end{aligned}$$

While the numerator can be computed tractably by direct evaluation, the integral in the denominator cannot be evaluated analytically and numerical integration is challenging due to high dimensionality of the variable of integration. Similarly, we are also unable to com-

[3] Not including parameters needed to indicate the selected edges.

pute analytically the probability of unseen data

$$P\left(\boldsymbol{x}|\mathcal{D},\alpha,\boldsymbol{\Theta}_\theta\right) = \int p\left(\boldsymbol{\Theta}_M|\mathcal{D},\alpha,\boldsymbol{\Theta}_\theta\right) P\left(\boldsymbol{x}|\boldsymbol{\Theta}_M\right) \mathrm{d}\boldsymbol{\Theta}_M.$$

Instead we can use Markov-Chain Monte Carlo (MCMC) techniques to sample from the posterior of the unseen variables and parameters. We will do so using a Gibbs sampler by fixing some of the parameters and variables and sampling from the posterior of the rest of the unobserved parameters. Let $\boldsymbol{s} = \left(s^1,\ldots,s^N\right)$ be the set of indicators where s^n refers to the mixture component that generated $\boldsymbol{x}^n$, and let $\boldsymbol{s}^{-n}$ refer to the indicators of all mixture components except for s^n. For the Bayesian finite mixture of trees, the sampling can be performed in the following sequence:

$$\begin{aligned} n=1,\ldots,N:\ s^n|\boldsymbol{\pi},\boldsymbol{\theta}_T^1,\ldots,\boldsymbol{\theta}_T^K &\sim \frac{\pi_i T^i\left(\boldsymbol{x}^n\right)}{\sum_{i=1}^K \pi_i T^i\left(\boldsymbol{x}^n\right)} \\ i=1,\ldots,K:\ \boldsymbol{\theta}_T^i|\boldsymbol{s} &\sim p\left(\cdot|\boldsymbol{s},\boldsymbol{\Theta}_\theta,\mathcal{D}\right) \\ \boldsymbol{\pi}|\boldsymbol{s} \sim D\left(\frac{\alpha}{K}+\#s_1,\ldots,\frac{\alpha}{K}+\#s_K\right) \end{aligned}$$

where

$$p\left(\boldsymbol{\theta}_T^i|\boldsymbol{s},\boldsymbol{\Theta}_\theta,\mathcal{D}\right) \propto p\left(\boldsymbol{\theta}_T^i|\boldsymbol{\Theta}_\theta\right) \prod_{n:s^n=i} T^i\left(\boldsymbol{x}^n\right) \qquad (2)$$

and $\#\boldsymbol{s}_i$ is the number of entries of $\boldsymbol{s}$ equal to i.

Finite mixture models can be extended to allow countably many mixture components by replacing a multinomial distribution (with a conjugate Dirichlet prior) with a Dirichlet process, a distribution over distributions (Ferguson, 1973; Antoniak, 1974). The resulting Dirichlet process mixture model (DPMM) can be viewed as a convex combination of distributions sampled from some base measure G_0 (in our case, $p\left(\cdot|\boldsymbol{\Theta}_\theta\right)$) and mixing coefficients from $D\left(\frac{\alpha}{K},\ldots,\frac{\alpha}{K}\right)$ as $K\to\infty$. This view is possible because distribution $G=\sum_{i=1}^\infty \pi_i\delta_{\boldsymbol{\phi}_i}$ with $\boldsymbol{\phi}_i\sim G_0$ is well-defined. One possible generative model for $\mathcal{D}$ under a DPMM is due to Blackwell and MacQueen (1973):

$$\begin{aligned} n=1,\ldots,N:\ \boldsymbol{\theta}_T^n|\boldsymbol{\theta}_T^1,\ldots,\boldsymbol{\theta}_T^{n-1} &\sim \frac{\sum_{j=1}^{n-1}\delta_{\boldsymbol{\theta}_T^j}+\alpha G_0}{n-1+\alpha} \\ n=1,\ldots,N:\ \boldsymbol{x}^n &\sim T\left(\cdot|\boldsymbol{\theta}_n\right). \end{aligned}$$

We still need to define a prior distribution for $\boldsymbol{\theta}_T^i$ so that Equation (2) has a form that we can sample from. Recently, Meilă and Jaakkola (2006) proposed a conjugate prior for tree-structured distributions. This prior is described in the next subsection; we will use it to obtain a closed form posterior distribution for (2).

3.1. Conjugate Prior for Tree-Structured Distributions

The prior consists of two components, $P\left(\mathcal{E}|\boldsymbol{\beta}\right)$ and $p\left(\boldsymbol{\theta}_E|\mathcal{E},\boldsymbol{\Psi}\right)$. The first probability distribution is over all possible spanning tree structures. The second distribution is over parameters of the bivariate distributions on the edges given the structure of the tree. $\boldsymbol{\beta}$ is a symmetric matrix of non-negative weights for each pair of distinct variables and zeros on the diagonal ($\beta_{uv}=\beta_{vu}\geq 0\ \forall u,v\in\mathcal{V},\ u\neq v$ and $\beta_{vv}=0\ \forall v\in\mathcal{V}$). The probability over all spanning tree structures is defined as

$$P\left(\mathcal{E}|\boldsymbol{\beta}\right) = \frac{1}{Z}\prod_{\{u,v\}\in\mathcal{E}}\beta_{uv} \qquad (3)$$

where the normalization constant is

$$Z = \sum_{\mathcal{E}}\prod_{\{u,v\}\in\mathcal{E}}\beta_{uv}. \qquad (4)$$

Remarkably, Z can be computed in closed form (in $\mathcal{O}\left(d^3\right)$ time) even though there are d^{d-2} trees to sum over.

For the conditional distribution of the parameters $\boldsymbol{\theta}$ given the edges $\mathcal{E}$, Meilă and Jaakkola (2006) propose using a product of Dirichlet priors, one for each edge:

$$\begin{aligned} p\left(\boldsymbol{\theta}|\boldsymbol{\Psi}\right) &= \prod_{v\in\mathcal{V}} p\left(\boldsymbol{\theta}_v|\boldsymbol{\Psi}\right)\prod_{\{u,v\}\in\mathcal{E}}\frac{P\left(\boldsymbol{\theta}_{uv}|\boldsymbol{\Psi}\right)}{P\left(\boldsymbol{\theta}_u|\boldsymbol{\Psi}\right)p\left(\boldsymbol{\theta}_v|\boldsymbol{\Psi}\right)} \\ &= \prod_{v\in\mathcal{V}} D\left(\boldsymbol{N}'_v\right)\prod_{\{u,v\}\in\mathcal{E}}\frac{D\left(\boldsymbol{N}'_{uv}\right)}{D\left(\boldsymbol{N}'_u\right)D\left(\boldsymbol{N}'_v\right)} \end{aligned} \qquad (5)$$

where $\boldsymbol{N}'_{uv} = \left\{N'_{uv}\left(i,j\right)>0: i\in\mathcal{X}_u, j\in\mathcal{X}_v\right\}$ and $\boldsymbol{N}'_v=\left\{N'_v\left(i\right)>0: v\in\mathcal{X}_v\right\}$, and $\forall v\in\mathcal{V}\ \forall j\in\mathcal{X}_v\ \forall u\neq v\in\mathcal{V}$

$$\sum_{i\in\mathcal{X}_u} N'_v\left(j\right) = \sum_{i\in\mathcal{X}_u} N'_{uv}\left(u,v\right).$$

$\boldsymbol{N}'_{uv}$ and $\boldsymbol{N}'_v$ can be thought of as data pseudo-counts for pairs and singletons of variables. Together with edge weights $\boldsymbol{\beta}$ they serve as sufficient statistics for the prior distribution. It is worth noting that this prior preserves the probability mass assigned to a set of parameters under reparametrizations due to changes in edge orientation if the undirected graphical model is converted into an equivalent directed graphical model. This likelihood equivalence property becomes very useful for sampling from this prior.[4]

[4] Meilă and Jaakkola (2006) showed that, subject to certain assumptions, this factored Dirichlet prior is the only prior that preserves likelihood equivalence.

Given a complete data set $\mathcal{D}$, a posterior distribution $p\left(\mathcal{E}, \boldsymbol{\theta} | \mathcal{D}, \boldsymbol{\beta}, \boldsymbol{\Psi}\right)$ has the same functional form as its prior. The update of the parameters can be computed in $\mathcal{O}\left(Nd^2r_{max}^2\right)$ time. Conjugacy also allows analytical computation of the probability of a new vector $\boldsymbol{x}$ given a data set $\mathcal{D}$. However, this computation requires computation of the determinant of a $d \times d$ matrix, and requires therefore $\mathcal{O}\left(d^3\right)$ update time. Details can be found in Meilă and Jaakkola (2006).

3.2. Sampling for Chow-Liu Trees

The description of the prior distribution for Chow-Liu trees in Section 3.1 suggests sampling first the edges $\mathcal{E}$ of the spanning tree and then the parameters $\boldsymbol{\theta}$ given the structure $\mathcal{E}$. Even though there are d^{d-2} spanning tree structures, there are algorithms to generate samples from the distribution (3) in polynomial time. Among the deterministic algorithms, Colbourn et al. (1996) proposed an algorithm with running time proportional to the time it takes to multiple two matrices of size $d \times d$ (a less cumbersome deterministic algorithm runs in $\mathcal{O}\left(d^3\right)$). Broder (1989) suggested a randomized algorithm for sampling from Equation (3) which takes time $\mathcal{O}\left(d \ln d\right)$ based on a random walk on the graph $\mathcal{V}$ with weights $\boldsymbol{\beta}$, which is dominated by the $\mathcal{O}\left(d^2\right)$ time needed to compute the transition probabilities for this walk. We implemented Broder's algorithm that switches to a slower Metropolis-Hastings sampler (based on Broder's algorithm with modified edge weights) if the original sampler does not stop in reasonable time. To sample parameters $\boldsymbol{\theta}$ given $\mathcal{E}$, one can first convert an undirected tree with edges $\mathcal{E}$ into a directed tree $\overline{\mathcal{E}}$ by choosing an arbitrary node in $\mathcal{V}$ as the root, and then incrementally changing undirected edges into directed edges such that there is a directed path from the root to each other node in the graph. One can then sample the parameters of the directed trees and transform them into their counterpart for the undirected tree. Because of the likelihood equivalence property, the resulting sample has the same distribution as if it were sampled directly from Equation (5).

3.3. Dirichlet Process Mixture of Trees

We finally have all of the pieces in place to define Dirichlet process mixture of trees (DPMT). A directed graphical model representation for the priors and parameters is shown in Figure 1. While computation of the posterior over $\boldsymbol{s}$ and $\boldsymbol{\theta}_T$ is impossible analytically, one can sample from this posterior using MCMC since only a finite number of mixture components can be responsible for generating $\mathcal{D}$ (e.g., Neal (2000)). Having a conjugate prior $p\left(\cdot|\boldsymbol{\beta}, \boldsymbol{\Psi}\right)$ allows us to use a collapsed Gibbs sampling scheme, as described in Figure 2 and based on Algorithm 2 from Neal (2000); however, it also contains a step to resample the concentration parameter α. West (1992) showed that a conjugate prior for α is a mixture of gamma distributions; here we use a single gamma distribution $G\left(a, b\right)$ with $a, b > 0$. For complete data, $P_T\left(\boldsymbol{x}^n | \boldsymbol{\Theta}_\theta\right)$ needs only be computed once ($\mathcal{O}\left(Nd^3\right)$ total). The running time for one iteration of this sampling scheme is $\mathcal{O}\left(NKd\right)$ for all evaluations of $T^i\left(\boldsymbol{x}^n\right)$ (N evaluations of K current components, $\mathcal{O}\left(d\right)$ per evaluation) and takes time $\mathcal{O}\left(d^2\right)$ on average for every newly added component.

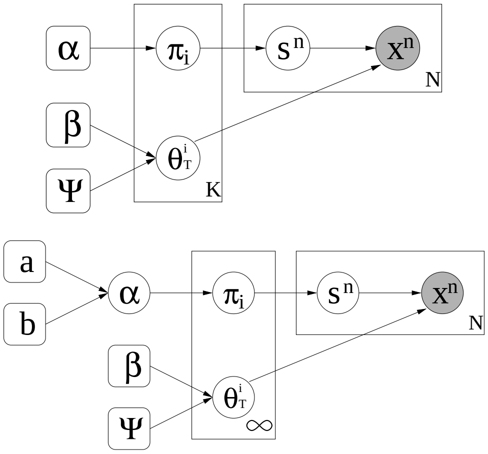

Figure 1. A graphical representation of a Bayesian finite mixture of trees (top) and a Dirichlet process mixture of trees (bottom).

4. Experiments

The experiments were performed with two tasks in mind. The first is estimation of the number of components for a finite mixture of trees (i.e., model selection). The second task is prediction via Bayesian model averaging using samples from the estimated posterior distribution over the parameters. Neither task can be solved analytically and, thus, we use approximations based on sampling. For the experiments, we used flat priors for trees, i.e., $\beta_{uv} = 1$ and $N'_{uv}\left(i, j\right) = 1$ for $u \neq v$.

4.1. Simulated Data

First, we use the DPMT model to learn a distribution from simulated data generated from a finite mixture of trees. We randomly generated 10 sets of model parameters of a 6-state MT model over $d = 30$ binary variables. From each of these models, we simulated 30-dimensional data vectors with training sets of size $N = 100, 200, 500, 1000, 2000$ and 5000. For each set of model parameters, we also generated a test set with

```
Algorithm SampleDPMT$(\boldsymbol{s}, \boldsymbol{\Theta}_T, \alpha, \mathcal{D})$
Inputs: Data set $\mathcal{D}$ and current MCMC values of
$\boldsymbol{s} = \left(s^1, \ldots, s^n\right)$, $\boldsymbol{\Theta}_T = \left\{\boldsymbol{\theta}_T^{s^n} \,:\, n = 1, \ldots, N\right\}$, $\alpha$

  • $n = 1, \ldots, N$: Update $s^n$
      – $\mathcal{S} = \{s^m \,:\, m = 1, \ldots, n-1, n+1, \ldots, N\}$
      – $w_{new} = \frac{\alpha}{N-1+\alpha}$;  $i \in \mathcal{S}:\ w_i = \frac{\# s_i^{-n}}{N-1+\alpha}$
      – $Z = \sum_{i \in \mathcal{S}} w_i T_i\left(\boldsymbol{x}^n\right) + w_{new} P_T\left(\boldsymbol{x}^n | \boldsymbol{\Theta}_\theta\right)$
      – $s^n \sim \left\{ \begin{array}{ccc} \frac{w_i}{Z} T^i\left(\boldsymbol{x}^n\right) & : & i \in \mathcal{S} \\ \frac{w_{new}}{Z} P_T\left(\boldsymbol{x}^n | \boldsymbol{\Theta}_\theta\right) & : & i = new \end{array} \right.$
      – If $s^n = new$ draw a new value for $s^n$ and
        parameters $\boldsymbol{\theta}_T^{s^n} \sim G_0$
  • $j \in \left\{s^1, \ldots, s^N\right\}$: Update $\boldsymbol{\theta}_T^j$
      – $\boldsymbol{\theta}_T^j \sim p\left(\cdot | \boldsymbol{s}, \left\{\boldsymbol{x}^n : s^n = j\right\}, \boldsymbol{\Theta}_\theta\right)$
  • Update $\alpha$ (West, 1992)

Output: New values MCMC values: $\boldsymbol{s}$, $\boldsymbol{\Theta}_T$, $\alpha$
```

Figure 2. Collapsed Gibbs sampling for the infinite mixture of trees model.

$N = 10000$. We then learned back the parameters for the 6-state MT with EM (MAP) with 100 random restarts. For each training set, we also computed 500 samples from the posterior $p\left(\boldsymbol{s}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N | \mathcal{D}\right)$ of the DPMT. Samples were collected every 20 iterations of the MCMC sampler after 1000 iterations for burn-in.

Figure 3 compares the log-likelihood of the test set under the DPMT (with a prior of $G(0.1, 1)$ for α), the learned 6-state MT, and the true model. The log-likelihood for each Bayesian model was obtained by averaging over test set log-likelihoods with parameters sampled from the posterior distribution. The results in Figure 3 show that the DPMT significantly outperforms a point estimate model. Furthermore, it performs virtually identically to a Bayesian finite mixture that knows the true number of components (and this true number is not known to the DPMT). The DPMT achieves its predictive gain over the MAP estimate model by averaging over sets of parameters.

4.2. Modeling Rainfall Occurrence Patterns

We use our approach to study predominant patterns of rainfall occurrence. Precipitation modeling is an important problem in hydrology, and accurate generative models for rainfall serve as useful tools in water management and crop modeling. Of particular interest is the problem of modeling simultaneous daily rainfall for a group of locations, usually in the same geographic region.

We consider two data sets: data from the Ceará region of Brazil (consisting of 24 90-day seasons for $d = 10$ stations, $N = 2160$), and data from the Queensland region in northeastern Australia (40 197-day seasons for $d = 11$ stations, $N = 7880$). For each set, we fit a DPMT with different priors for α $((a, b) = (1, 1), (0.5, 1), (0.2, 1), (0.1, 1), (100, 100), (50, 100), (20, 100), (10, 100))$. We then collected 1000 samples from the posterior $p\left(\boldsymbol{s}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N | \mathcal{D}\right)$ (one every 20 iterations of the MCMC sampler after 1000 iterations of burn-in).

From the histograms of the number of active mixture components obtained from 1000 samples for the Brazil data (Figure 4), it appears that depending on the parameters of the prior for α, there could be between 3 and 12 components active. However, a closer look at the traces of $\boldsymbol{s}$ paints a different picture. The top three components (in terms of the proportion of examples assigned to them) explain either all of the data or most of it (Figure 5 is representative of runs for all values of (a, b)).There are only 3 significant components, and the presence of the rest of the components may be compensating for the fact that the data was likely not generated by a finite mixture distribution. These results are consistent with evaluation of K by a cross-validation (Figure 7, top). The cross-validated log-likelihood plot also suggests that the DPMT has a better predictive accuracy than a K-state MT, again possibly because the data was not generated by a finite mixture model.

A similar story can be told about the other data sets. The DPMT posterior distribution over the number of mixture components is consistent with cross-validated log-likelihood selection for Queensland (Figures 6 (top,center) and 7, bottom). However, unlike the Brazil data, the traces for this set do not indicate a clear number of mixture components (also consistent with the cross-validated log-likelihood plots in Figure 7, bottom). However, some of the components are intermittent as the number of vectors with the same value of index s often drops to 0 (i.e., the corresponding component vanishes during MCMC). Note that by dropping these intermittent components from the counts, we get a much clearer picture of the likely number of true components in the data (e.g., see Figure 6, bottom). Also, as with the Brazil data set, the DPMT has better predictive accuracy than a finite state MT (Figure 7).

5. Conclusion

Finite mixtures of tree-structured distributions are a broadly useful class of graphical models due to their tractability. In this paper we proposed and developed the extension of such models to the infinite mixture case. We described a general non-parametric Bayesian framework for this problem. On simulated data the model was able to outperform the finite mixture model in terms of prediction on unseen data, even when the finite mixture model was using the true number of components. We also illustrated the ability of the model to extract useful information from large-scale real-world rainfall data sets. In conclusion, the non-parametric Bayesian approach to mixtures of trees allows it to achieve systematically better prediction performance compared to finite mixtures, by averaging over uncertainty in the tree mixture parameters and the number of components.

For the rainfall data in particular it would be useful to extend the DPMT approach to time series data, for example via hierarchical DPs (Teh et al., 2006). While in this paper we assumed complete data vectors, the Bayesian framework presented here could be extended to handle missing data.

Acknowledgments

This work has been supported by the Alberta Ingenuity Fund through the Alberta Ingenuity Centre for Machine Learning (SK), by the National Science Foundation under grants SCI-0225642, IIS-0431085, and ITR-0331707 (PS), and by the US Department of Energy under grant DE-FG03-01ER63260 (PS).

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, *2*, 1152–1174.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions and Pólya urn schemes. *The Annals of Statistics*, *1*, 353–355.

Broder, A. (1989). Generating random spanning trees. In *Proceedings of 30th annual symposium on foundations of computer science*, 442–447.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In D. Fisher and H.-J. Lenz (Eds.), *Learning from data: AI and statistics V*, 121–130. New York: Springer-Verlag.

Chow, C. K., & Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, *IT-14*, 462–467.

Colbourn, C. J., Myrvold, W. J., & Neufeld, E. (1996). Two algorithms for unranking arboresences. *Journal of Algorithms*, *20*, 268–281.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*, 309–347.

Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (1990). *Introduction to algorithms*. The MIT Electrical Engineering and Computer Science Series. MIT Press/McGraw Hill.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, *39*, 1–38.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209–230.

Grim, J. (1984). On structural approximating multivariate discrete probability-distributions. *Kybernetika*, *20*, 1–17.

Lam, W., & Bacchus, F. (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence*, *10*, 269–293.

Meilă, M., & Jaakkola, T. (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, *16*, 77–92.

Meilă, M., & Jordan, M. I. (2000). Learning with mixtures of trees. *Journal of Machine Learning Research*, *1*, 1–48.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Staitistics*, *9*, 249–265.

Srebro, N. (2003). Maximum likelihood bounded treewidth Markov networks. *Artificial Intelligence*, *143*, 123–138.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of American Statistical Association*, *101*, 1566–1581.

West, M. (1992). *Hyperparameter estimation in Dirichlet process mixture models* (Technical Report). Institute of Statistics and Decision Sciences, Duke University.

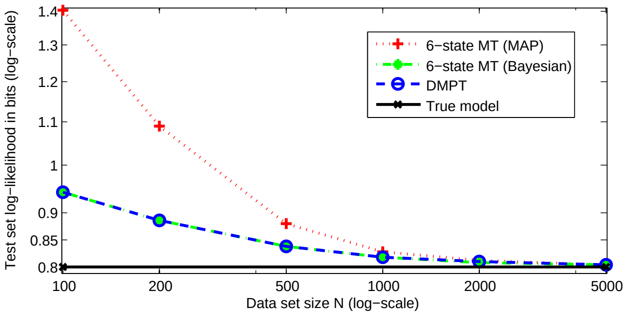

Figure 3. Test set averaged log-likelihood of the MAP estimate of the distribution with the correct parametric form (6-state mixture of trees, red pluses, dotted), Bayesian finite mixture for the correct parametric form (estimated, green stars, dash-dotted), DPMT (estimated, blue circles, dashed), and true distribution (black Xs, solid).

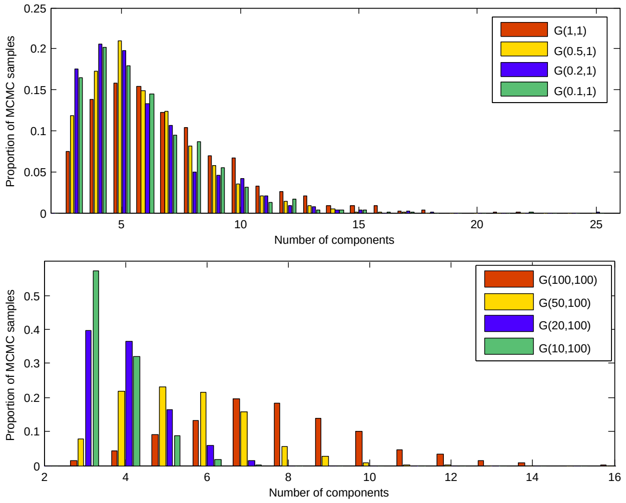

Figure 4. Histogram plots of the fraction of samples that contain k components, over 1000 MCMC samples for the Brazil data. (Top: all clusters under wide priors for α; bottom: all clusters under a concentrated prior for α)

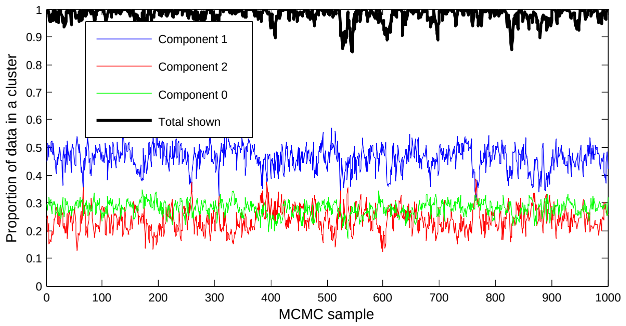

Figure 5. The proportion of data vectors in the Brazil data set that are assigned to each of the top 3 components in the model, and the combined proportion. (α with prior $G(0.1, 1)$).

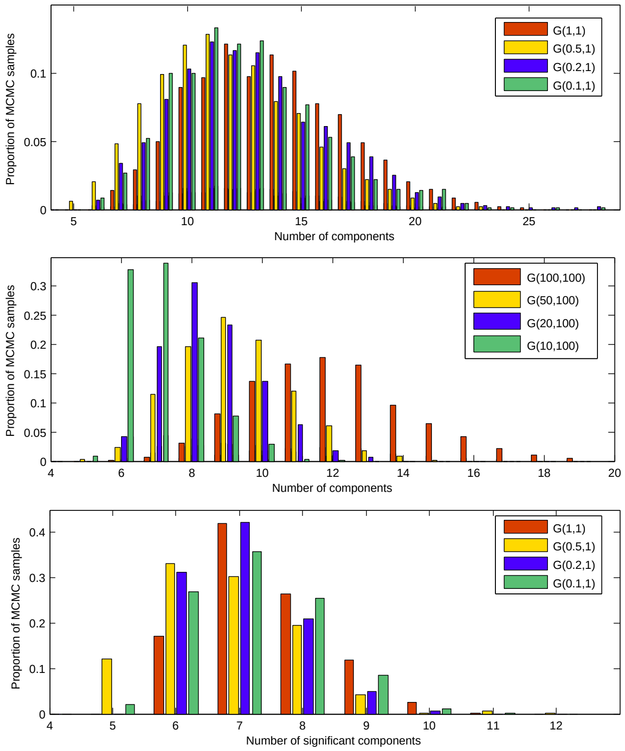

Figure 6. Histogram plots of the fraction of samples that contain k components, over 1000 MCMC samples for the Queensland data. (Top: all clusters under wide priors for α; center: all clusters under a concentrated prior for α; bottom: significant clusters under wide priors for α)

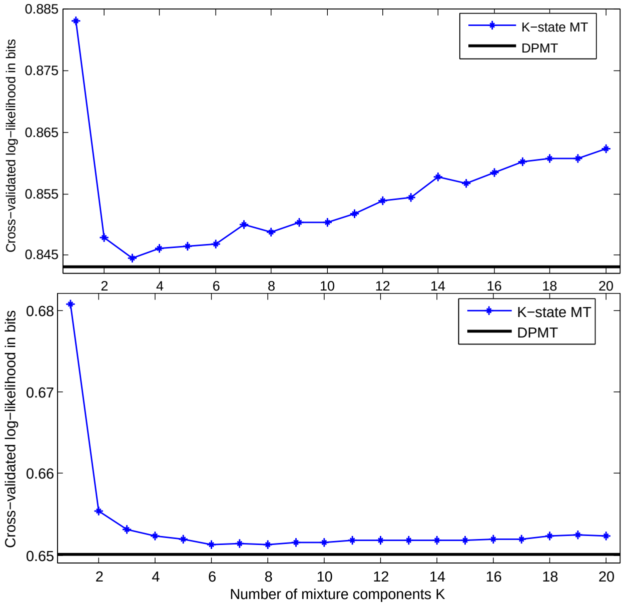

Figure 7. Out-of-sample (cross-validated, leave-one-season out) log-likelihood of K-state mixture of trees (blue stars) vs DPMT cross-validated estimate (thick black line) for Brazil (top) and Queensland (bottom) data.