Tree-Based Inference for Dirichlet Process Mixtures

Yang Xu
Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, USA

Katherine A. Heller
Department of Engineering
University of Cambridge
Cambridge, UK

Zoubin Ghahramani
Department of Engineering
University of Cambridge
Cambridge, UK

Abstract

The Dirichlet process mixture (DPM) is a widely used model for clustering and for general nonparametric Bayesian density estimation. Unfortunately, like in many statistical models, exact inference in a DPM is intractable, and approximate methods are needed to perform efficient inference. While most attention in the literature has been placed on Markov chain Monte Carlo (MCMC) [1, 2, 3], variational Bayesian (VB) [4] and collapsed variational methods [5], [6] recently introduced a novel class of approximation for DPMs based on Bayesian hierarchical clustering (BHC). These tree-based combinatorial approximations efficiently sum over exponentially many ways of partitioning the data and offer a novel lower bound on the marginal likelihood of the DPM [6]. In this paper we make the following contributions: (1) We show empirically that the BHC lower bounds are substantially tighter than the bounds given by VB [4] and by collapsed variational methods [5] on synthetic and real datasets. (2) We also show that BHC offers a more accurate predictive performance on these datasets. (3) We further improve the tree-based lower bounds with an algorithm that efficiently sums contributions from alternative trees. (4) We present a fast approximate method for BHC. Our results suggest that our combinatorial approximate inference methods and lower bounds may be useful not only in DPMs but in other models as well.

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

1 Introduction

Nonparametric Bayesian methods have become extremely popular due to their ability to flexibly model data. Whereas traditional parametric methods restrict the form of a model by fixing the number of parameters, nonparametric models allow the model complexity to grow with the number of data points and provide wide support for data from a large family of distributions. The Bayesian approach to nonparametric data modeling avoids overfitting by placing a prior on the model parameters and considering the posterior distribution in the limit as the number of parameters becomes infinite. One such prior over infinitely many parameters is provided by the Dirichlet process (DP) [7].

The DP defines a nonparametric distribution over distributions, and can therefore be used to define flexible priors over unknown distributions. A property of the distributions drawn is that they are discrete, so draws from these distributions in turn yield repeated values. Values drawn from a DP exhibit a "rich get richer" property, where values which have been commonly observed in past draws are more likely to be observed in future draws. This generally leads to the number of observed values being much smaller than the number of draws. Thus DPs can be utilized as a prior for clustering data if these drawn values are interpreted as cluster memberships for each data point. From this construction we can get the Dirichlet process mixture model (DPM) [8], which will be reviewed in detail in Section 2.

The DPM can be seen as an infinite mixture model in that instead of using a fixed number of mixture components (or clusters), it allows a countably infinite number of mixture components to model the data. Moreover, the DPM does not require model selection to determine the number of components in the mixture model, it automatically infers this from the data by allowing data points to be assigned to new clusters and not restricting membership to existing mixture compo-

nents. One drawback of the DPM is that it is generally intractable since it considers exponentially many $O(n^n)$ ways of partitioning n data points into clusters. Rasmussen (2000) and Escobar and West (1995) provide a detailed analysis of DPMs with Gaussian components and an MCMC algorithm for sampling from partitionings of the data [1, 3]. Blei et al. (2005) describe a variational Bayesian (VB) approach which optimizes a lower bound on the marginal likelihood of a DPM and they compare it thoroughly with standard MCMC methods (e.g. Gibbs sampler) showing a significant decrease in running time [4]. Kurihara et al. (2007) also introduce a class of collapsed variational approximate methods for DPMs by using a truncated stick-breaking construction and a finite mixture model with a symmetric Dirichlet prior [5].

Heller and Ghahramani (2005) developed Bayesian hierarchical clustering (BHC) as a new hierarchical clustering method and approximate inference algorithm for DPMs and proved that BHC yields a new combinatorial lower bound on the marginal likelihood of a DPM [6]. In this paper, we empirically compare BHC to VB and the collapsed variational methods on small synthetic datasets where the exact marginal likelihood of a DPM can be computed. We then compare these on three real-world datasets. We also compare the predictive performance of BHC to the other algorithms. Furthermore, we develop a new algorithm which constructs alternative tree structures to BHC and show that this method tightens the BHC lower bound on the marginal likelihood of a DPM. Finally, we present a fast approximate method for BHC based on a Bayes K-means algorithm and show that it gives significant speedups in the runtime.

The paper is organized as follows. Section 2 briefly reviews the DP mixture models. Section 3 reviews the BHC algorithm. Section 4 derives the alternative tree algorithm. Section 5 introduces a fast approximate method for BHC. Sections 6 and 7 discuss the empirical results and present conclusions.

2 Dirichlet Process Mixture Models

We briefly review Dirichlet process mixture models by starting with a finite mixture model and taking the limit as the number of mixture components goes to infinity, as in [2, 3]. We start from a finite mixture model with C components:

$$p(\mathbf{x}^{(i)}|\phi) = \sum_{j=1}^{C} p(\mathbf{x}^{(i)}|\theta_j)p(c_i = j|\mathbf{p}) \qquad (1)$$

where $c_i \in \{1, \ldots, C\}$ is a cluster assignment for data point i, $\mathbf{p}$ are the parameters of a multinomial distribution with $p(c_i = j|\mathbf{p}) = \mathrm{p}_j$, θ_j are the parameters of the jth component, and $\phi = (\theta_1, \ldots, \theta_C, \mathbf{p})$. Let the parameters of each component have conjugate priors (e.g. Normal-Inverse-Wishart priors for Normal continuous data) $p(\theta|\beta)$ (where β are the hyperparameters of the conjugate distribution) and the multinomial parameters also have a conjugate, Dirichlet prior, $p(\mathbf{p}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/C)^C} \prod_{j=1}^{C} \mathrm{p}_j^{\alpha/C-1}$, where α is the concentration parameter. The marginal likelihood of the mixture model for a data set $\mathcal{D} = \{\mathbf{x}^{(1)} \ldots, \mathbf{x}^{(n)}\}$ is:

$$p(\mathcal{D}|\alpha, \beta) = \int \left[\prod_{i=1}^{n} p(\mathbf{x}^{(i)}|\phi)\right] p(\phi|\alpha, \beta)\, d\phi \qquad (2)$$

where $p(\phi|\alpha, \beta) = p(\mathbf{p}|\alpha) \prod_{j=1}^{C} p(\theta_j|\beta)$. This can be re-written as:

$$p(\mathcal{D}|\alpha, \beta) = \sum_{\mathbf{c}} p(\mathbf{c}|\alpha)p(\mathcal{D}|\mathbf{c}, \beta) \qquad (3)$$

where $\mathbf{c} = (c_1, \ldots, c_n)$ and $p(\mathbf{c}|\alpha) = \int p(\mathbf{c}|\mathbf{p})p(\mathbf{p}|\alpha)d\mathbf{p}$. The quantity (3) is well-defined even in the limit $C \to \infty$. The number of possible ways of partitioning n points remains finite although the number of possible settings of $\mathbf{c}$ diverges as $C \to \infty$. Let $\mathcal{V}$ denote the set of all possible partitioning of n points, we can rewrite (3) as:

$$p(\mathcal{D}|\alpha, \beta) = \sum_{v \in \mathcal{V}} p(v|\alpha)p(\mathcal{D}|v, \beta) \qquad (4)$$

Finally, it is not hard to show [6] that the marginal likelihood of a DPM as can be explicitly written as:

$$p(\mathcal{D}|\alpha, \beta) = \sum_{v \in \mathcal{V}} \frac{\alpha^{m_v} \prod_{\ell=1}^{m_v} \Gamma(n_\ell^v)}{\left[\frac{\Gamma(n+\alpha)}{\Gamma(\alpha)}\right]} \prod_{\ell=1}^{m_v} p(\mathcal{D}_\ell^v|\beta) \qquad (5)$$

where $\mathcal{V}$ is the set of all possible partitionings of $\mathcal{D}$, n is the number of data points in $\mathcal{D}$, m_v is the number of clusters in partitioning v, and n_ℓ^v is the number of points in cluster ℓ of partitioning v, and $p(\mathcal{D}_\ell^v|\beta) = \int \left[\prod_{i \in \mathcal{D}_\ell^v} p(\mathbf{x}^{(i)}|\theta_\ell)\right] p(\theta_\ell|\beta)\, d\theta_\ell$.

3 Bayesian Hierarchical Clustering and Combinatorial Lower Bounds

In this section, we review the Bayesian hierarchical clustering algorithm following [6]. BHC provides a new fast approximate inference method for Dirichlet process mixture models. Rather than summing over all possible partitions of the data using MCMC, BHC builds a binary tree (dendrogram) and provides a lower bound on the marginal likelihood of a DPM by summing over exponentially many clusterings of the data in polynomial time ($O(n^2)$).

Consider a data set $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ and tree T where $\mathcal{D}_i \subset \mathcal{D}$ is the set of data points at the leaves of the sub-tree T_i of T. BHC is similar to traditional agglomerative clustering [9] in that it is a one-pass, bottom-up agglomerative method which initializes n clusters (leaves of the hierarchy) each containing a single data point $\mathcal{D}_i = \{\mathbf{x}^{(i)}\}$. It then iteratively merges pairs of clusters to construct the hierarchy. The main difference between BHC and traditional hierarchical clustering methods is that BHC uses a statistical hypothesis test to choose which clusters to merge, instead of a distance metric.

In considering each merge, two hypotheses are compared. The first hypothesis ($\mathcal{H}_1^k$) is that all the data in $\mathcal{D}_k$ were generated independently and identically from the same probabilistic model, $p(\mathbf{x}|\theta)$ with unknown parameters θ (e.g. a Gaussian with $\theta = (\mu, \Sigma)$). We compute the probability of data $\mathcal{D}_k$ under $\mathcal{H}_1^k$ by specifying some prior over the parameters of the model (if we use conjugate priors the following integral is tractable):

$$p(\mathcal{D}_k|\mathcal{H}_1^k) = \int p(\mathcal{D}_k|\theta)p(\theta|\beta)d\theta$$
$$= \int \Big[\prod_{\mathbf{x}^{(i)} \in \mathcal{D}_k} p(\mathbf{x}^{(i)}|\theta) \Big] p(\theta|\beta)d\theta \quad (6)$$

The alternative hypothesis ($\mathcal{H}_2^k$) would be that $\mathcal{D}_k$ has two or more clusters in it. Summing over the exponentially many possible ways of dividing $\mathcal{D}_k$ into two or more clusters is intractable. However, if we restrict ourselves to clusterings that partition the data in a manner that is consistent with the sub-trees T_i and T_j (see Figure 1(a) on the concept of *tree-consistent partitions*), we can efficiently sum over exponentially many alternative clusterings using recursion. The probability of the data under the alternative hypothesis is then simply $p(\mathcal{D}_k|\mathcal{H}_2^k) = p(\mathcal{D}_i|T_i)p(\mathcal{D}_j|T_j)$. The marginal probability of the data in any sub-tree T_k is computed as follows:

$$p(\mathcal{D}_k|T_k) = \pi_k p(\mathcal{D}_k|\mathcal{H}_1^k) + (1 - \pi_k)p(\mathcal{D}_i|T_i)p(\mathcal{D}_j|T_j) \quad (7)$$

where $\pi_k \stackrel{\text{def}}{=} p(\mathcal{H}_1^k)$. Note that this equation is defined recursively, there the first term considers the hypothesis that there is a single cluster in $\mathcal{D}_k$ and the second term efficiently sums over all other clusterings in $\mathcal{D}_k$ which are consistent with the tree structure. At each iteration, BHC merges the two clusters that have the highest posterior probability of the merged hypothesis $r_k \stackrel{\text{def}}{=} p(\mathcal{H}_1^k|\mathcal{D}_k)$ which is defined by the Bayes rule:

$$r_k = \frac{\pi_k p(\mathcal{D}_k|\mathcal{H}_1^k)}{p(\mathcal{D}_k|T_k)} \quad (8)$$

The quantity π_k, which can also be computed bottom up as the tree is built, is defined to be the relative prior mass in a DPM with hyperparameter α, of the partition where all data points are in one cluster, versus all the other partitions consistent with the subtrees. As shown in [6], $\pi_k = \frac{\alpha\Gamma(n_k)}{d_k}$ where $d_k = \alpha\Gamma(n_k) + d_{\text{left}_k}\ d_{\text{right}_k}$, right (left) refer to the children of internal node k, and at the leaves, $d_i = \alpha$, $\pi_i = 1$. The BHC algorithm automatically infers the number of clusters by cutting the tree at $r_k < 0.5$. The original paper [6] describes in detail the clustering performance of BHC and quantitatively compares it to traditional hierarchical clustering methods.

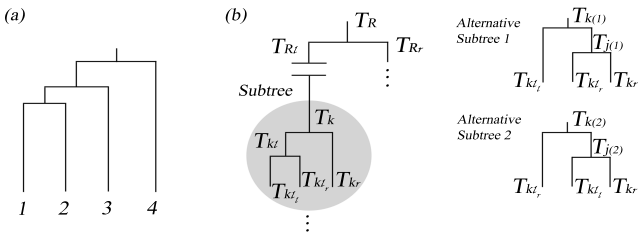

Figure 1: (a) An example tree with 4 data points. A tree-consistent clustering is given by cutting the hierarchy at any level. The clusterings (1 2 3)(4) and (1 2)(3)(4) are tree-consistent partitions of this data. The clustering (1)(2 3)(4) is not tree-consistent. (b) An example of the alternative tree algorithm. T_k is a generic subtree under internal node k in the BHC tree. $T_{k(1)}$ and $T_{k(2)}$ are two alternative subtrees obtained from relocating or swapping the branches under T_k. Thus we obtain two alternative trees to the BHC tree from relocations at a single node. Meanwhile, the marginal probability of the data under the alternative trees tightens the BHC lower bound.

Heller and Ghahramani proved that for any binary tree T_k with the data points $\mathcal{D}_k$ at its leaves, BHC gives a lower bound on the marginal likelihood of a DPM [6]:

$$\frac{d_k\Gamma(\alpha)}{\Gamma(n_k + \alpha)} p(\mathcal{D}_k|T_k) \leq p(\mathcal{D}_k|\alpha, \beta) \quad (9)$$

where the left hand side of the inequality is the BHC lower bound, $p(\mathcal{D}_k|T_k)$ is the BHC approximation to the marginal likelihood of a DPM, and $p(\mathcal{D}_k|\alpha, \beta)$ is the exact marginal likelihood of a DPM given by Equation 5. An important question, which we address in Section 6.1.1 is how well the BHC lower bound (9) compares to the class of variational lower bounds for DPMs [4, 5]. BHC also offers a predictive distribution for new test points which is discussed in detail in [10]. We will compare its predictive performance to those of the class of variational methods in Section 6.1.2.

4 Alternative Tree Algorithm

BHC is a greedy algorithm which yields a single binary tree and may not capture uncertainty associated

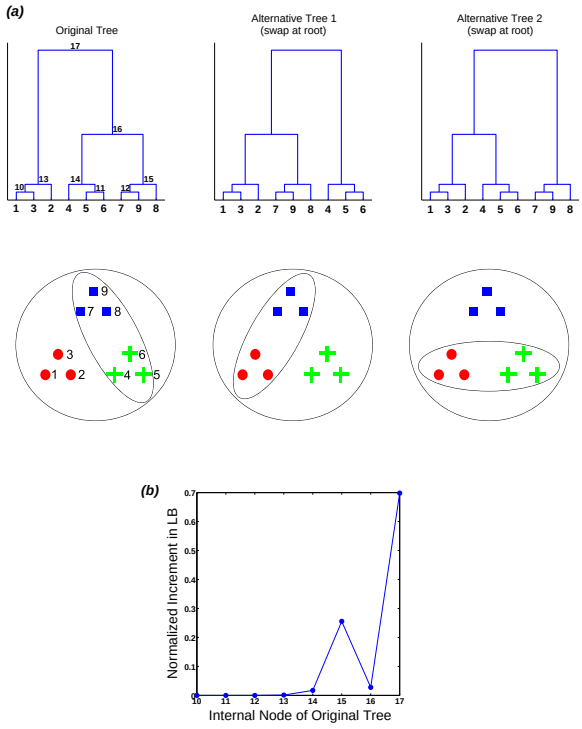

Figure 2: (a) An example with 9 data points in 3 equi-distant clusters each with 3 equi-distant data points. The top row shows 3 hierarchies: the first is from BHC and the rest are from the alternative tree algorithm with relocations at the root of tree. The bottom row shows the corresponding clusterings. (b) The distribution on the lower bound increment from relocations at each internal node. The magnitude of the increment reflects the uncertainty of the clustering. Note the values at node 13, 14, 15 and 17 indicate raised uncertainty when clustering equi-distant clusters.

with alternative clusterings of the data (Figure 2(a)). In this section, we present a new algorithm which constructs alternative tree structures to the one given by BHC. Importantly, we can use this algorithm to improve the BHC lower bound described in Section 3 and account for alternative clusterings of the data.

Our algorithm uses the original BHC tree as a starting point and alters its *subtrees* by relocating the branches (we can also think of a relocation as a swap of the branches under an internal node) as illustrated in Figure 1(b). Thus for any node that has more than 2 leaves we obtain 2 alternative trees (trees under nodes with less than or equal to 2 leaves remain unchanged). We perform these relocations at one node at a time and keep the rest of the tree unchanged, iterating this process from a user-specified node upward on the path to the root of the entire tree. By relocating at a single node for each iteration we can efficiently compute the additional marginal probability of the data under the resulting alternative subtrees. Moreover, over all iterations the additional probability mass is propagated to the root of the tree which yields an improved BHC lower bound of the DPM. While the original BHC lower bound sums over exponentially many tree-consistent partitions, the new lower bound considers alternative trees and efficiently sums over partitions which are not present in the original BHC tree. The computational complexity of our algorithm is $O(\log(n))$, and for a balanced tree with n data points at its leaves (i.e. $n-1$ internal nodes), our algorithm gives $n-2$ alternative tree structures (i.e. trees under nodes at the bottom level are not altered, and trees under nodes at one level up yield $\frac{n}{2}$ possible alternatives, and so on), which can be computed as follows:

$$\lim_{n\to\infty}(0+\frac{n}{2}+\frac{n}{4}+...+\frac{n}{2^{\log_2(n)-1}})=n-2.$$

Consider relocating branch $T_{k_{l_r}}$ from being a sibling of $T_{k_{l_l}}$ to being a sibling of T_{k_r} in a general subtree T_k (see Figure 1(b)). This gives $T_{j(1)}$ under new node $j(1)$ (note that the total number of nodes is conserved), and alternative subtree $T_{k(1)}$. Similarly, relocating $T_{k_{l_l}}$ gives $T_{j(2)}$ and alternative subtree $T_{k(2)}$. We want to compute the additional probability of the data $\mathcal{D}_k$ under these alternative trees (i.e. $p(\mathcal{D}_k|T_{k(1)})$ and $p(\mathcal{D}_k|T_{k(2)})$) which offer partitions that are not present in the original BHC tree. Since we have already represented all partitionings where the data $\mathcal{D}_k$ under T_k are in one cluster in the BHC tree , we only have to consider the non-merged hypothesis. So for $T_{k(1)}$ we have (analogously with equation (7)):

$$p(\mathcal{D}_k|T_{k(1)})=p(\mathcal{D}_{j(1)}|T_{j(1)})p(\mathcal{D}_{k_{l_l}}|T_{k_{l_l}}) \qquad (10)$$

where the quantities $p(\mathcal{D}_{j(1)}|T_{j(1)}) = p(\mathcal{D}_{j(1)}|\mathcal{H}_1^{j(1)})$ (i.e. merged hypothesis, see (6)) and $p(\mathcal{D}_{k_{l_l}}|T_{k_{l_l}})$ have already been computed in constructing the BHC tree. The priors on merging can also be obtained:

$$d_{k(1)}=d_{j(1)}d_{k_{l_l}} \qquad (11)$$

initialize: $i=k$
while $i<R\,(root)$ **do**
  $p(\mathcal{D}_{i+1}|T_{i+1})=p(\mathcal{D}_{l_{(i+1)}}|T_{l_{(i+1)}})p(\mathcal{D}_{r_{(i+1)}}|T_{r_{(i+1)}})$
  $d_{i+1}=d_{l_{(i+1)}}d_{r_{(i+1)}}$
  $i\leftarrow i+1$
end while

Figure 3: Message propagation from node k of subtree T_k to root R of the entire tree.

where $d_{j(1)} = \alpha\Gamma(n_{j(1)})$, and $d_{k_{l_l}}$ remains unchanged as computed in BHC. This gives $\pi_{k(1)} = 0$ and $\pi_{j(1)} = 1$. The computation of $p(\mathcal{D}_k|T_{k(2)})$ and $d_{k(2)}$ is analogous. We can now propagate the quantities $p(\mathcal{D}_k|T_{k(1)})$, $d_{k(1)}$, $p(\mathcal{D}_k|T_{k(2)})$ and $d_{k(2)}$ (treating them as messages) from node k to the root of tree R (see Figure 3). Note that by relocating in two ways under an internal node we obtain 2 alternative trees and 2 sets of messages (i.e. $\{p(\mathcal{D}_{root}|T_{root(1)}), d_{root(1)}\}$ and $\{p(\mathcal{D}_{root}|T_{root(2)}), d_{root(2)}\}$) at the root of tree. The idea is that we can add the additional marginal probability of the data under these alternative trees to the marginal probability under the BHC tree and hence increase the total marginal likelihood and improve the BHC lower bound as discussed in Section 3.

We can now compute the improvement in the BHC lower bound that results from the alternative trees. Specifically, if we denote $T_{k(i)}$ as one alternative tree to T_k, then its corresponding contribution to the increment of the original BHC lower bound $\delta_{LB(i)}$ can be defined as follows (using equation (9)):

$$\delta_{LB(i)} = \frac{d_{k(i)}\Gamma(\alpha)}{\Gamma(n_k+\alpha)} p(\mathcal{D}_k|T_{k(i)}) \tag{12}$$

Theorem 1 *For any binary tree T_k with the data points $\mathcal{D}_k$ at its leaves, the following is a new lower bound on the marginal likelihood of a DPM:*

$$\frac{d_k\Gamma(\alpha)}{\Gamma(n_k+\alpha)} p(\mathcal{D}_k|T_k) + \sum_{i\in\mathcal{N}} \delta_{LB(i)} \leq p(\mathcal{D}_k|\alpha,\beta)$$

where the first term on the left hand side of the inequality is the original BHC lower bound as in (9), and the second term is the sum of increments in lower bound from set $\mathcal{N}$ of alternative trees to T_k.

Proof The proof follows from the fact that (1) we sum over all partitions that are consistent with the set of alternative trees and (2) each alternative tree is unique in that it results from a single change in the original BHC tree, hence there is no accumulated change or double counting of partitionings of the data.

The general alternative tree algorithm is summarized in Figure 4. Our algorithm is valid for any starting node l. Note that although there can be many ways of generating alternative tree structures, our algorithm has the advantages that (1) by confining to local changes in the tree we are able to efficiently compute and propagate the probabilities in a manner that is consistent with the original BHC algorithm (2) given (1) we can effectively improve the BHC lower bound on the marginal likelihood of a DPM with very few additional computations (3) we further accommodate uncertainty in the clusterings of BHC which have already been shown to be of high quality [6].

input: starting node l, outputs from BHC: tree, $p(\mathcal{D}_k|T_k)$, d_k & $p(\mathcal{D}_k|\mathcal{H}_1^k)$ for $k = l, ..., R(root)$
initialize: $i = l$
while $i < R$ **do**
  if number of leaves under node $i > 2$ **then**
    Change subtree under i (in two ways)
    Compute $p(\mathcal{D}_i|T_i)$ & d_i for 2 alternative subtrees
  else
    Retain $p(\mathcal{D}_i|T_i)$ & d_i from original BHC
  end if
  Propagate message $\{p(\mathcal{D}_i|T_i),d_i\}$ from i to R
  Store $p(\mathcal{D}_R|T_R)$ and d_R for 2 subtrees under i
  $i \leftarrow i+1$
end while
output: alternative trees to original BHC tree and a new lower bound

Figure 4: Alternative Tree Algorithm

5 Fast Bayes K-Means BHC

BHC is a greedy algorithm with $O(n^2)$ complexity. Heller and Ghahramani [11] describe two fast algorithms for BHC which reduce its complexity to $O(n)$ and $O(n \log n)$. Here we develop a new approximate method called Bayes K-means BHC. We first introduce Bayes K-means (similar to ME algorithm in [12]). Our algorithm is like traditional K-means, but instead of using Euclidean distance, we use marginal likelihood as the criterion for cluster assignment. Moreover, our algorithm uses DPM as the generative model and greedily chooses a clustering configuration that maximizes the marginal likelihood by marginalizing over the parameters at each iteration (using Dirichlet-Multinomial conjugacy and DPM priors, we can compute marginal likelihood from Equation 3). We start by randomly sampling k points (e.g. $k = \sqrt{n}$) and assigning them to their own clusters. We then iteratively assign each of the remaining points to one of the existing clusters or create their own new cluster by considering the marginal likelihood over different clusterings, e.g. suppose we want to assign $(k+1)th$ point, we would compute the marginal likelihood for k clustering configurations resulted from tentatively assigning the point to each of k clusters in turn, and for the case that the point is assigned to a new cluster. We then choose the clustering that gives the highest marginal likelihood. Once all points have been assigned, we further merge the clusters by greedily optimizing the global marginal likelihood. Note that our algorithm automatically infers the number of clusters. Bayes K-means BHC then follows two steps. First, it partitions the data into a number of clusters via Bayes K-means.

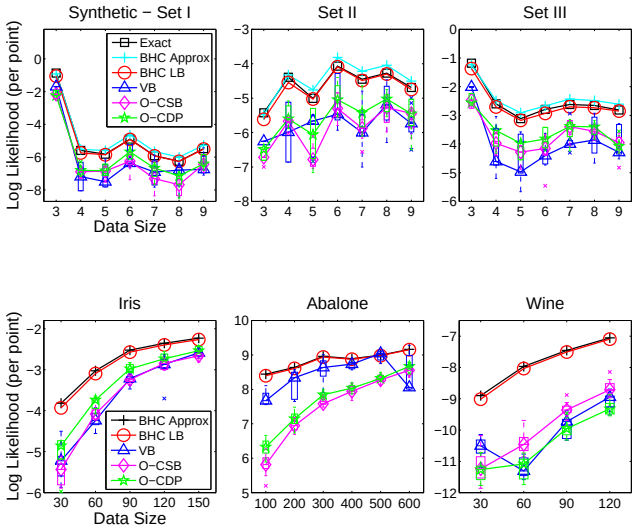

Figure 5: Comparison of marginal likelihood approximate among BHC, VB and collapsed variational methods ('O-CSB' and 'O-CDP' are collapsed variational methods with truncated stick-breaking and a finite mixture model with symmetric Dirichlet prior). The top row are 3 synthetic Gaussian datasets: Set I has 2 separate mixtures; Set II has 2 close mixtures; Set III has 1 mixture. The bottom row corresponds to 3 real datasets: iris, abalone and wine data.

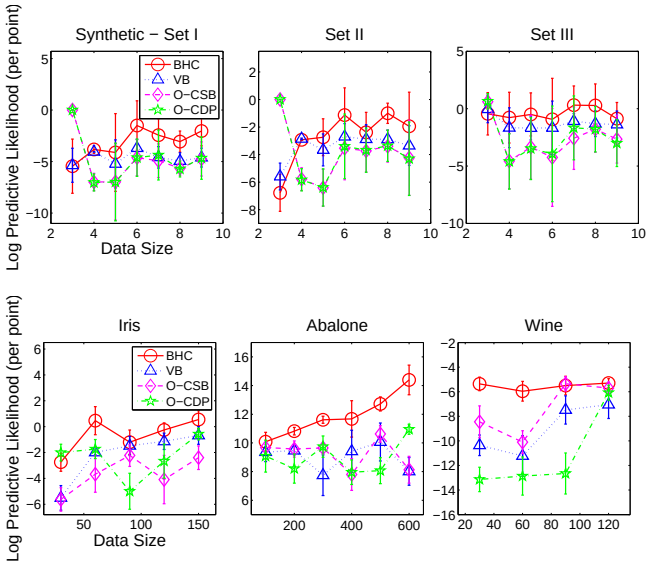

Figure 6: Comparison of predictive performance among BHC, VB and collapsed variational methods on the same synthetic and real datasets as laid out in Figure 5.

Second, it runs BHC on these clusters and constructs a hierarchical tree. The complexity of Bayes K-means BHC is approximately $O(nk + k^2)$ $(k < n)$. We show empirically in Section 6.3 that it gives large speedups and good approximation to BHC.

6 Results

6.1 Empirical Comparison

6.1.1 Comparison on Marginal Likelihood

We compared the marginal likelihood approximate for DPMs from BHC to those from variational Bayesian approximations (VB) [4] and two collapsed variational approximations [5], i.e. the variational method with truncated stick-breaking construction, and one with a finite mixture model using a symmetric Dirichlet prior. We used optimal cluster label reordering for the collapsed methods as this was shown to produce tighter lower bounds than those without [5]. We first compared these algorithms on 3 small synthetic datasets where the exact marginal likelihood of a DPM could be computed. We then compared on 3 large real-world datasets. Due to space limitations, we do not review VB and the collapsed variational methods here. Detailed description of these and comparisons of VB to MCMC (e.g. Gibbs sampling) can be found in [4, 5].

We generated 3 two-dimensional Gaussian datasets: Set I has two separate Gaussian mixture components which are far apart from each other without any overlapping points; Set II has two closely-neighboring mixture components which also have no overlaps; Set III has only one mixture component[1]. We chose very small datasets (3 to 9 points) so that we were able to compare the marginal likelihoods of BHC and of the class of variational approximate methods to the exact marginal likelihood of a DPM. We varied the data size from 3 to 9 and used fixed data points. Figure 5 shows the results. We found that in all 3 data sets BHC was closer to the exact likelihood than all of the variational methods (truncation level $K = 10$). Moreover, BHC gave a deterministic lower bound whereas the estimates of the variational methods varied at each run due to local optima resulting from random initializations. We verified empirically that the BHC lower bound was never greater than the exact likelihood, and also noted that the BHC approximation (i.e. $P(D|T_{root})$) is also very close to the exact likelihood and BHC lower bound.

Although computing the exact marginal likelihood of a DPM on large datasets would be infeasible, we com-

[1]Set I: $\mu_1 = (2, 2), \sigma_1^2 = 0.5$ and $\mu_2 = (8, 8), \sigma_2^2 = 0.5$; Set II: $\mu_1 = (5, 5), \sigma_1^2 = 0.5$ and $\mu_2 = (7, 5), \sigma_2^2 = 0.5$; Set III: $\sigma^2 = 0.5$, where μ and σ^2 denote mean and variance of a 2-D Gaussian distribution.

pared the BHC lower bound to the variational methods over 3 real-world real-valued datasets. The datasets we used were the iris (150 examples, 3 classes, 4 attributes), abalone (600 examples, 25 classes, selected 5 continuous attributes) and wine (120 examples, 4 classes, selected 6 continuous attributes) data from the UCI repository. We again varied the data size with fixed data points in these experiments and set the truncation level of variational methods $K = 20, 40$ and 20 respectively for the 3 datasets. Figure 5 shows the results. We found that the BHC lower bounds were higher (and hence tighter) than the mean lower bounds of all variational methods in the three examples (in fact, BHC was often better than the best runs of the variational methods).

6.1.2 Comparison on Predictive Performance

We compared the predictive performance among BHC, VB and the class of collapsed variational approximations using the same synthetic and real datasets as described in Section 6.1.1. For the synthetic datasets, we held out 1 point as the test data and repeatedly permuted this for each data size. For the real datasets, we randomly chose a small amount (about 10%) at each data size as the holdout and trained on the rest. Figure 6 shows the results. We found that in all datasets BHC generally gave better predictive accuracy than the other algorithms.

6.2 Evaluating the New Alternative Tree BHC Lower Bound

We applied the alternative tree algorithm over similar synthetic and real-world datasets as in Section 6.1 and three more real-world binary datasets, and compared the new lower bound to the original BHC lower bound.

Figure 7 (top row) shows the results on the synthetic datasets. We set the starting node for relocations as the lowest internal node and varied the data size from 3 to 9 with randomly sampled points at each size, computing the exact likelihoods for these small data sets. We found that the new lower bound was closer to the exact likelihood than the original lower bound. We also verified the computation of the new lower bound by noting that the differences between the exact likelihood and new lower bound were zero for 3 data points. For 3 points, the alternative tree algorithm sums over all partitions, so it computes the exact likelihood.

We also applied our algorithm on 3 real-world real-valued (middle row in Figure 7) and 3 more binary datasets: (1) digits (0-9, 300 examples, 64 attributes); (2) CMU 20 newsgroups (300 examples, 500 attributes); (3) spambase (200 examples, 2 classes, 57 attributes) (bottom row in Figure 7). In these examples we set the starting node for relocations close to the root of the tree because we found empirically that swaps at high levels had much higher gains in marginal likelihood than those at lower levels. We varied the data size and selected random samples at each size. We note that in all these datasets there was a general improvement on the lower bound. We also note that the degree of improvement can be data dependent. While log differences of 0.4 to 0.8 on the binary data may seem modest, they indicate that the partitions in the alternative tree capture 50% to 120% as much posterior mass as the original BHC tree.

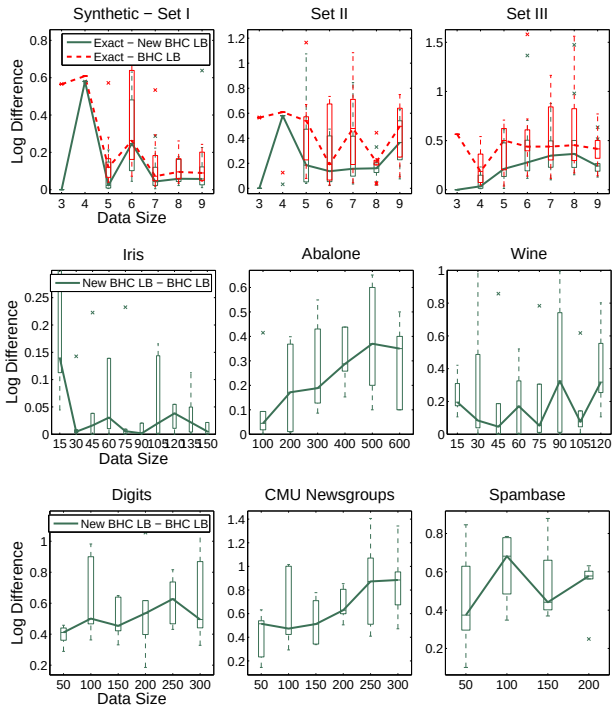

Figure 7: Comparison of new and original BHC lower bounds. The top row are 3 synthetic Gaussian datasets: Set I has 2 separate mixtures; Set II has 2 close mixtures; Set III has 1 mixture. The middle row are 3 real datasets: iris, abalone and wine data. The bottom row corresponds to 3 further real binary datasets: digits, CMU newsgroups and spambase data.

Finally, we ran our algorithm on a synthetic Gaussian dataset (Figure 2(a)) and explored the alternative clusterings that it offers. We set the starting node as the lowest internal node. We note that the increment in lower bound based on relocations at each internal node reflects the degree of uncertainty in that particular clustering. Moreover, the distribution of increments at different internal nodes allows us to evaluate which relocations yield effective alternative clusterings and identify good alternative tree structures (Figure 2(b)).

6.3 Comparing Bayes K-Means BHC and BHC

We ran Bayes K-means BHC and original BHC on the iris, abalone and wine datasets as in Section 6.1. We generated random samples at varied data sizes for repeated runs in each set. The inferred number of clusters from Bayes K-means for the three sets at their largest data size are on average 4, 31 and 5 respectively (real number of classes are 3, 25 and 4). Figure 8 compares the likelihood and runtime of the two algorithms. We note that the approximate method offers significant speedups (e.g. for 600-point abalone set, the speedup is about 30-fold). Furthermore, the approximate on marginal likelihood from Bayes K-means BHC to that from BHC (and to BHC lower bound) is also close. As we have shown in Section 6.1 that both BHC approximate and lower bound give accurate estimates to the exact marginal likelihood of a DPM, our result suggests that K-Means BHC can potentially be used to give very fast approximation for DPMs.

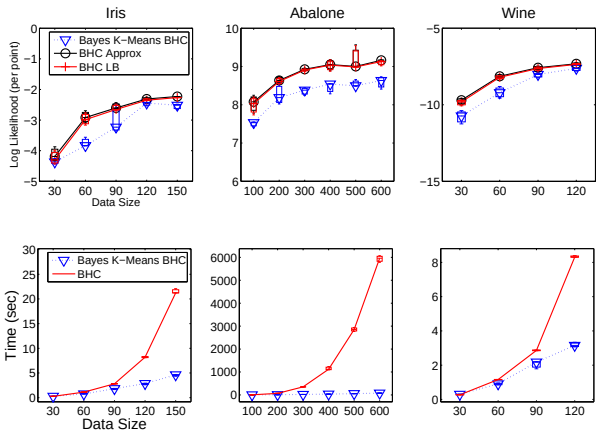

Figure 8: Comparison of Bayes K-Means BHC and orginal BHC. The columns are 3 real datasets: iris, abalone and wine data. The top row compares the likelihood of data given trees, and BHC lower bound. The bottom row compares the runtime for each dataset.

7 Conclusions

We empirically compared BHC to variational Bayes and the class of collapsed variational methods for DPMs and showed that BHC gives significantly better approximations to a DPM in most instances. Moreover, BHC offers a deterministic combinatorial lower bound to the marginal likelihood of a DPM, while the variational approximations vary at each run. We also showed that BHC gave a better predictive performance than the other algorithms. We developed a novel alternative tree algorithm which gives alternative structures to the BHC tree and allows for a tightening of the BHC lower bound to a DPM. Our algorithm identifies sensible alternative trees by examining their contribution to the marginal likelihood. Finally, we presented an approximate method Bayes K-Means BHC. We showed that it offers fast and good approximate to BHC. In the future, we would like to explore whether the combinatorial lower bounds which we have shown here to work well for DPMs, can be the basis of approximate inference algorithms for other models.

Acknowledgements

We thank Kenichi Kurihara for his code and discussion on the variational approximate methods for DPMs.

References

[1] M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of American Statistical Association*, 90:577–588, 1995.

[2] R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[3] C. E. Rasmussen. The infinite Gaussian mixture model. In *Neural Information Processing Systems 12*. MIT Press, 2000.

[4] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144, 2005.

[5] K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational Dirichlet process mixture models. In *The Twentieth International Joint Conference on Artificial Intelligence*, 2007.

[6] K. A. Heller and Z. Ghahramani. Bayesian hierarchical clustering. In *Twenty-second International Conference on Machine Learning*, 2005.

[7] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973.

[8] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 1974.

[9] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley, 1973.

[10] K.A. Heller. Efficient Bayesian methods for clustering. *PhD Thesis, University College London*, pages 35–36, 2007.

[11] K. A. Heller and Z. Ghahramani. Randomized algorithms for fast Bayesian hierarchical clustering. In *Statistics and Optimization of Clustering Workshop*, Windsor, UK, 2005.

[12] K. Kurihara and M. Welling. Bayesian K-means as a "maximization-expectation" algorithm. *Neural Computation*, 2008.