VARIATIONAL NONPARAMETRIC BAYESIAN HIDDEN MARKOV MODEL

Nan Ding, Zhijian Ou

Department of Electronic Engineering
Tsinghua University, Beijing, China
ssnding@gmail.com, ozj@tsinghua.edu.cn

ABSTRACT

The Hidden Markov Model (HMM) has been widely used in many applications such as speech recognition. A common challenge for applying the classical HMM is to determine the structure of the hidden state space. Based on the Dirichlet Process, a nonparametric Bayesian Hidden Markov Model is proposed, which allows an infinite number of hidden states and uses an infinite number of Gaussian components to support continuous observations. An efficient variational inference method is also proposed and applied on the model. Our experiments demonstrate that the variational Bayesian inference on the new model can discover the HMM hidden structure for both synthetic data and real-world applications.

***Index Terms*—** Nonparametric Bayesian, Hidden Markov Model, Variational Inference, Speech Recognition

1. INTRODUCTION

The Hidden Markov Model (HMM) has been widely used in many areas of pattern recognition and machine learning, such as speech recognition and gene clustering [1, 2]. The HMM includes a sequence of multinomial state variables $s_1, ..., s_T$, and a sequence of observations $o_1, ..., o_T$. Each state variable takes its value in the state space $\{1, ..., N\}$, and each observation o_t is drawn independently of the other observations conditional on s_t.

Varying the size of the state space N greatly affects the performance of HMM. Because of this reason, there are lots of works trying to find out an optimal N. Among those works, nonparametric Bayesian methods have attracted more and more attention in recent years. Some of the nonparametric Bayesian models such as the Dirichlet Process [3, 4] and the Indian Buffet Process [5] have been widely applied.

In this paper, we extend the Bayesian Hidden Markov Model [1, 6] to its nonparametric counterpart, by replacing the Dirichlet distribution by the Dirichlet process. The size of the state space of this new nonparametric Bayesian HMM model (NBHMM) is infinite, in which the ”effective” states correspond to the states with ”large” posterior probabilities. Because the exact inference of this model is intractable, we derive an variational inference method which is efficient even for large-scale problems.

The new NBHMM is different from other existing nonparametric Bayesian HMMs, which include the infinite HMM (iHMM) proposed in [7] and the hierarchical Dirichlet process HMM (HDP-HMM) proposed in [3]. First, both existing models employ sampling-based inference which is usually much slower for large-scale problems, while we apply the efficient variational inference in the NBHMM. Second, the iHMM deals only with discrete observations, while the NBHMM supports continuous observations via Gaussian mixtures. Third, note that the transition distribution in both the iHMM and the HDP-HMM is generated from a hierachical Dirichlet process. Instead, the transition distribution in the NBHMM is directly created from a stickbreaking construction, which is simpler and thus allows more efficent inference.

The rest of paper is organized as follows. Section 2 describes the new NBHMM. Section 3 introduces the variational inference for the NBHMM. The experimental results in Section 4 demonstrate the effectiveness of the NBHMM on learning the structure of the hidden state space.

2. NONPARAMETRIC BAYESIAN HMM

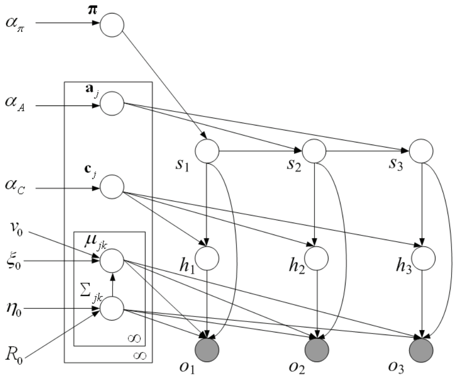


Fig. 1. Nonparametric Bayesian HMM

The graphical model of the NBHMM is shown in Fig.1. In this model, the dark nodes o_t are observations which take continuous values. A chain of mixtures of Gaussian models is considered to generate the sequence of observations. The white nodes s_t are the hidden states, h_t are the mixture components, and both of them take discrete values. In many applications, $p(s_t|s_{t-1})$ and $p(h_t|s_t)$ are regarded as the same for different t. We can represent $p(s_1)$ with $\boldsymbol{\pi} = (\pi_i)_{i=1}^N$, $p(s_t|s_{t-1})$ with $\mathbf{A} = (\mathbf{a}_j)_{j=1}^N, \mathbf{a}_j = (a_{ji})_{i=1}^N$, and $p(h_t|s_t)$ with $\mathbf{C} = (\mathbf{c}_j)_{j=1}^N, \mathbf{c}_j = (c_{jk})_{k=1}^K$. Here N denotes the size of the state space and K the size of the component space. $\boldsymbol{\pi}$ is a normalized vector, $\mathbf{A}$ is the state transition matrix and $\mathbf{C}$ is the state-to-component matrix. $\boldsymbol{\mu}$ and Σ are the parameters of the Gaussian distribution. For different s_t and h_t, μ_{s_t,h_t} and Σ_{s_t,h_t} are different.

For the Bayesian HMM, the main difference from the classical HMM is that the parameters $\boldsymbol{\pi}$, $\mathbf{A}$, $\mathbf{C}$ are not treated as unknown

This work was supported by National Natural Science Foundation of China (60402029) and China 863 program (2006AA01Z149).

values, but as random variables.

$$\begin{aligned}&p(\mathbf{s},\mathbf{h},\mathbf{o},\boldsymbol{\pi},\mathbf{A},\mathbf{C},\boldsymbol{\mu},\Sigma^{-1}) \\ &= p(\mathbf{s},\mathbf{h},\mathbf{o}|\boldsymbol{\pi},\mathbf{A},\mathbf{C},\boldsymbol{\mu},\Sigma)p(\boldsymbol{\pi})p(\mathbf{A})p(\mathbf{C})p(\boldsymbol{\mu}|\Sigma^{-1})p(\Sigma^{-1})\end{aligned} \tag{1}$$

where $\mathbf{s} = (s_t)_{t=1}^T$, $\mathbf{h} = (h_t)_{t=1}^T$, $\mathbf{o} = (o_t)_{t=1}^T$.

$$\begin{aligned}&p(\mathbf{s},\mathbf{h},\mathbf{o}|\boldsymbol{\pi},\mathbf{A},\mathbf{C},\boldsymbol{\mu},\Sigma^{-1}) \\ &= p(s_1|\boldsymbol{\pi})\prod_{t=2}^{T}p(s_t|s_{t-1},\mathbf{A})\prod_{t=1}^{T}p(h_t|s_t,\mathbf{C})p(o_t|\mu_{s_t,h_t},\Sigma_{s_t,h_t})\end{aligned} \tag{2}$$

Assuming the covariance matrix Σ is diagonal with the dimension of D, we place Gaussian-Gamma prior distribution on Gaussian parameters $\boldsymbol{\mu}$ and Σ in this paper. For each dimension $d = 1, ..., D$,

$$\begin{aligned}p(\mu_{jkd}|\Sigma_{jkd}^{-1}) &= \mathcal{N}(v_0, \xi_0^{-1}\Sigma_{jkd}) \\ p(\Sigma_{jkd}^{-1}) &= Gamma(\eta_0, R_0)\end{aligned}$$

One main problem for both the classical HMM and Bayesian HMM is the difficulty in determining the optimal size of the state space N and the component space K. The NBHMM tries to circumvent the problem by setting the number of states and components (i.e. N and K) to be infinite. In order to have an infinite-length multinomial distribution, we use the Dirichlet process [3] for the priors $p(\boldsymbol{\pi})$, $p(\mathbf{A})$, $p(\mathbf{C})$. In particular, we apply one of the commonly-used representations of the Dirichlet process called the "stickbreaking construction" [8],

$$\begin{aligned}p(\pi_i') &= Beta(1,\alpha_\pi) & \pi_i &= \pi_i'\prod_{n=1}^{i-1}(1-\pi_n') \\ p(a_{ji}') &= Beta(1,\alpha_A) & a_{ji} &= a_{ji}'\prod_{n=1}^{i-1}(1-a_{jn}') \\ p(c_{jk}') &= Beta(1,\alpha_C) & c_{jk} &= c_{jk}'\prod_{l=1}^{k-1}(1-c_{jl}')\end{aligned}$$

where $\sum_{i=1}^{\infty}\pi_i = 1$ and the same for a_{ji} and c_{jk}. The elegancy of nonparametric Bayesian method is that, although the state space is infinite, the posteriors $p(\boldsymbol{\pi}|\mathbf{o})$, $p(\mathbf{a}_j|\mathbf{o})$ and $p(\mathbf{c}_j|\mathbf{o})$ will only have "large" probabilities in a finite number of states while all others are nearly equal to zero. In fact, only the states corresponding to "large" probabilities are effective in explaining the observed data.

3. VARIATIONAL INFERENCE ON NBHMM

The inference problem for the NBHMM model is to compute the posterior $p(\mathbf{s},\mathbf{h},\boldsymbol{\pi},\mathbf{A},\mathbf{C},\boldsymbol{\mu},\Sigma|\mathbf{o})$, which is intractable in general. However, the variational inference provides us a way to approximately compute the posterior efficiently even for large-scale problems. The basic idea of variational inference is to use a tractable distribution q to approximate the true posterior distribution p, and then to minimize the Kullback-Leibler divergence between the two distribution as measured by $KL(q|p) = \int q\log(q/p)$.

For the approximate posterior distribution q, we make two approximations. First, we assume that ($\boldsymbol{\pi}$, $\mathbf{A}$, $\mathbf{C}$, $\boldsymbol{\mu}$, Σ) and ($\mathbf{s}$, $\mathbf{h}$) are mutually independent. Second, we only compute the probabilities of the L states of the infinite large state-space. L is called the truncation level of stickbreaking, which should be sufficiently large to ensure the accuracy. Note that using the truncation level is quite different from setting a finite state-space in a statistical perspective, in that the truncation level is just an approximation of the infinite states. Similar truncation is applied to the state-dependent component distribution (i.e. each row of $\mathbf{C}$). Finally, the approximate distribution can be represented as follows,

$$\begin{aligned}&q(\mathbf{s},\mathbf{h})q(\boldsymbol{\pi}')q(\mathbf{A}')q(\mathbf{C}')q(\boldsymbol{\mu},\Sigma^{-1}) \\ &= q(s_1)\prod_{t=2}^{T}q(s_t|s_{t-1})\prod_{t=1}^{T}q(h_t|s_t) \\ &\cdot\prod_{i=1}^{L}q(\pi_i')\prod_{j=1}^{L}\prod_{i=1}^{L}q(a_{ji}')\prod_{j=1}^{L}\prod_{k=1}^{L}q(c_{jk}') \\ &\cdot\prod_{j=1}^{L}\prod_{k=1}^{L}\prod_{d=1}^{D}q(\mu_{jkd}|\Sigma_{jkd}^{-1})q(\Sigma_{jkd}^{-1})\end{aligned} \tag{3}$$

where,

$$\begin{aligned}q(\pi_i') &= Beta(\tau_{1(\pi_i')}, \tau_{2(\pi_i')}) \\ q(a_{ji}') &= Beta(\tau_{1(a_{ji}')}, \tau_{2(a_{ji}')}) \\ q(c_{jk}') &= Beta(\tau_{1(c_{jk}')}, \tau_{2(c_{jk}')}) \\ q(\mu_{jkd}|\Sigma_{jkd}^{-1}) &= \mathcal{N}(\tilde{v}_{jkd}, \tilde{\xi}_{jkd}^{-1}\Sigma_{jkd}) \\ q(\Sigma_{jkd}^{-1}) &= Gamma(\tilde{\eta}_{jkd}, \tilde{R}_{jkd})\end{aligned}$$

The parameters $\tau_{\pi'}$,$\tau_{a'}$, $\tau_{c'}$, $\tilde{v}$, $\tilde{\xi}$, $\tilde{\eta}$, and $\tilde{R}$ of the approximate distribution q is computed by minimizing $KL(q|p)$ by a coordinate descent algorithm. The resulting variational update steps are as follows:

$$\tau_{1(\pi_i')} = 1 + q(s_1 = i) \tag{4}$$

$$\tau_{2(\pi_i')} = \alpha_\pi + q(s_1 > i) \tag{5}$$

$$\tau_{1(a_{ji}')} = 1 + \sum_{t=2}^{T}q(s_{t-1} = j, s_t = i) \tag{6}$$

$$\tau_{2(a_{ji}')} = \alpha_A + \sum_{t=2}^{T}q(s_{t-1} = j, s_t > i) \tag{7}$$

$$\tau_{1(c_{jk}')} = 1 + \sum_{t=1}^{T}q(s_t = j, h_t = k) \tag{8}$$

$$\tau_{2(c_{jk}')} = \alpha_C + \sum_{t=1}^{T}q(s_t = j, h_t > k) \tag{9}$$

$$\tilde{v}_{jkd} = \big(v_0\xi_0 + \sum_{t=1}^{T}q(s_t = j, h_t = k)o_{td}\big)/\tilde{\xi}_{jk} \tag{10}$$

$$\tilde{\xi}_{jkd} = \xi_0 + \sum_{t=1}^{T}q(s_t = j, h_t = k) \tag{11}$$

$$\tilde{\eta}_{jkd} = \eta_0 + \sum_{t=1}^{T}q(s_t = j, h_t = k) \tag{12}$$

$$\begin{aligned}\tilde{R}_{jkd} &= R_0 + \xi_0(v_0 - \tilde{v}_{jkd})^2 \\ &\quad + \sum_{t=1}^{T}q(s_t = j, h_t = k)(o_{td} - \tilde{v}_{jkd})^2\end{aligned} \tag{13}$$

The values of $q(s_t)$, $q(s_t, s_{t-1})$, $q(s_t, h_t)$ can be computed by the forward-backward propagation algorithm similar to the classical

HMM given that,

$$\log q(s_1 = i) \tag{14}$$
$$= \left[\Psi(\tau_{1(\pi'_i)}) - \Psi(\tau_{1(\pi'_i)} + \tau_{2(\pi'_i)})\right]$$
$$+ \sum_{n=1}^{i-1} \left[\Psi(\tau_{2(\pi'_n)}) - \Psi(\tau_{1(\pi'_n)} + \tau_{2(\pi'_n)})\right] + const.$$

$$\log q(s_t = i | s_{t-1} = j) \tag{15}$$
$$= \left[\Psi(\tau_{1(a'_{ji})}) - \Psi(\tau_{1(a'_{ji})} + \tau_{2(a'_{ji})})\right]$$
$$+ \sum_{n=1}^{i-1} \left[\Psi(\tau_{2(a'_{jn})}) - \Psi(\tau_{1(a'_{jn})} + \tau_{2(a'_{jn})})\right] + const.$$

$$\log q(h_t = k | s_t = j) \tag{16}$$
$$= \left[\Psi(\tau_{1(c'_{jk})}) - \Psi(\tau_{1(c'_{jk})} + \tau_{2(c'_{jk})})\right]$$
$$+ \sum_{l=1}^{k-1} \left[\Psi(\tau_{2(c'_{jl})}) - \Psi(\tau_{1(c'_{jl})} + \tau_{2(c'_{jl})})\right] + const.$$

$$\log q(o_t | s_t = j, h_t = k) \tag{17}$$
$$= -\frac{1}{2} \sum_{d=1}^{D} \Big(\log 2\pi + \frac{1}{\tilde{\xi}_{jkd}} - \Psi(\frac{\tilde{\eta}_{jkd}}{2})$$
$$+ \log(\frac{\tilde{R}_{jkd}}{2}) + \frac{(o_{td} - \tilde{v}_{jkd})^2}{\tilde{\eta}_{jkd}^{-1} \tilde{R}_{jkd}} \Big) + const.$$

where $\Psi(\bullet)$ is the digamma function. In conclusion, the variational inference iteratively updates the parameters, which is guaranteed to converge to a local minimum of the divergence $KL(q|p)$.

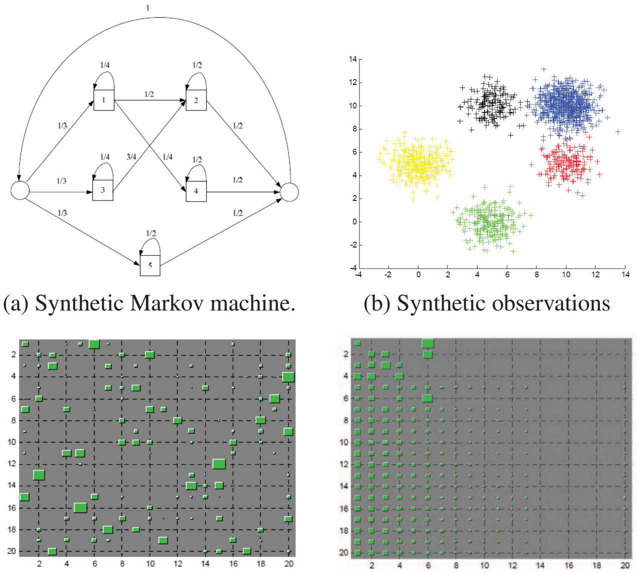
4. EXPERIMENTS

The hyperparameters of the NBHMM in the experiments are: $\alpha_\pi = 1, \alpha_A = 1, \alpha_C = 1, v_0 = 0, \xi_0 = 1, \eta_0 = 1, R_0 = 0.01$.

4.1. A Simple Comparison

The synthetic data is generated by a 5-state Markov machine as in Fig.2 (a). The number in the square node denotes the state-number. The circle node is introduced to simplify the plotting. This is intended as a toy example of continuous speech recognition which uses four phonetic states (no.1-4) plus a silence state (no. 5). The data contains 50 chains, and the length of each chain is 20. The observations take 2-d continuous values being synthetic samples from Gaussian distributions, as shown in Fig.2 (b). Different colors mean that the observations are generated by different hidden states. We fit both the classical HMM with the size of state-space $N = 20$ and the NBHMM with the truncation level $L = 20$. The Hinton graphs for the learned transition matrix $\mathbf{A}$ of the classical HMM and the mean of $q(\mathbf{A})$ of the NBHMM are plotted in Fig.2 (c)(d). (In the Hinton graph, a bigger blot represents a larger probability in the transition matrix.)

It is clear from Fig.2 that given the improper setup of the size of the state space, the classical HMM cannot learn the structure of the Markov machine that generates the data. In contrast, the Hinton graph of the NBHMM indicates that there are five states, corresponding to row 1,2,3,4,6 in Fig.2(d), whose posteriors are different from their priors due to the impact of the observations. It is also found that, each of the corresponding 5 rows in the $\mathbf{C}$ matrix for the NBHMM places nearly all weights on only one component. Thus, only these 5 states are effective in explaining the data. And it can be easily read from Fig.2(d), that the transitions between these 5 states correspond exactly to Fig. 2(a), discovering the true structure of the Markov machine.



(a) Synthetic Markov machine. (b) Synthetic observations

(c) Hinton graph for classical HMM (d) Hinton graph for NBHMM

Fig. 2. A simple comaprison of classical HMM and NBHMM

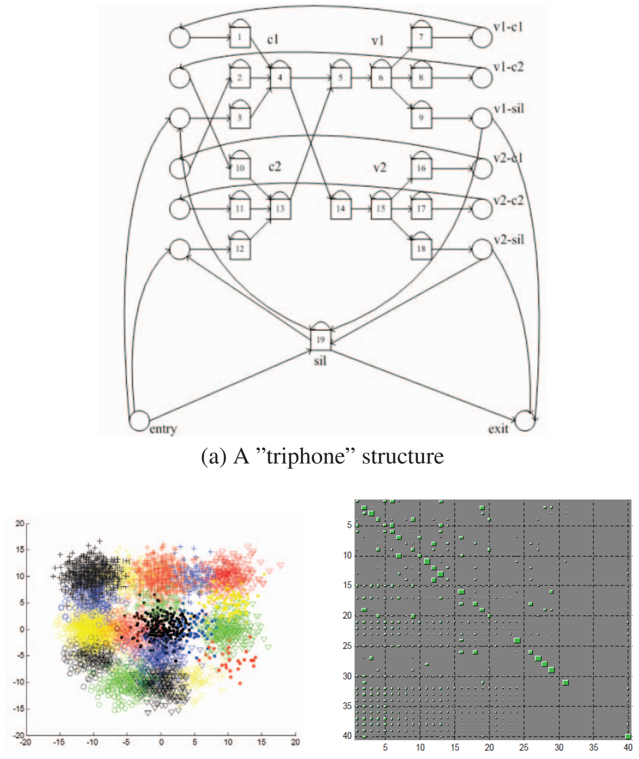
4.2. Simulated Triphone Structure

In order to illustrate the ability of the NBHMM in learning more complex structures, we simulate an important structure which is widely used in current speech recognition system - triphone structure for context-dependent acoustic modeling [9]. It is supposed that there are two consonants - c1 and c2, and two vowels - v1 and v2, each being modeled as two-states. The vocabulary consists of three words - c1v1, c1v2 and c2v1, plus a silence unit. Then the cross-word triphone structure is shown in Fig.3(a). We generate 5 chains, and the length of each chain is 1000. The observations take 2-d continuous values as shown in Fig.3(b).

Again, the Hinton graph resulting from the variational inference over the NBHMM with $L = 40$ discovers the nearly-correct structure. There are 22 ”effective” states, slightly more than the real 19 states, which is acceptable considering this difficult structure and the noise on the observations. Further, it can be read from Fig.3(c) that the transitions between these 22 states correspond closely to Fig.3(a). And each of the corresponding 22 rows in the $\mathbf{C}$ matrix for the NBHMM again places nearly all weights on one component.

4.3. Impact on Speech Recognition

Finally, we apply the NBHMM in the task of Chinese isolated (toned) syllable recognition. There are a total of 1254 syllables in Chinese. The database consists of 50 males, with each person speaking all 1254 syllables exactly once. We leave one person’s data for recognition and use the remaining 49 persons’ data for training. This procedure is repeated for every person, and the averaged recognition rate over 50 persons is reported here. In the front-end, the speech was parameterized into 14 MFCCs along with normalized log-energy, and their first and second order differentials.

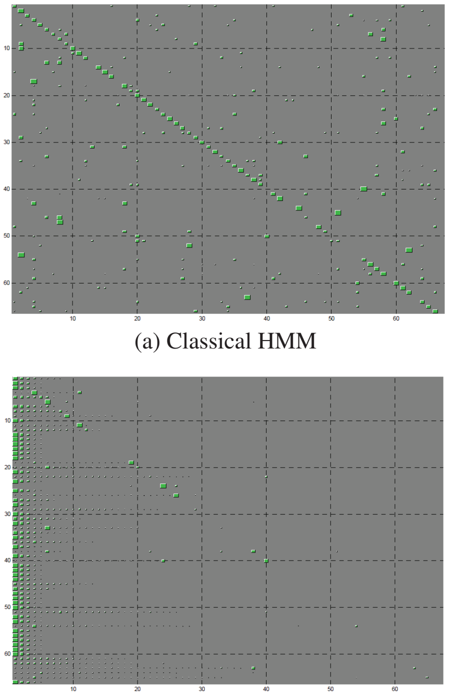


(a) A "triphone" structure

(b) Synthetic observations

(c) Hinton graph for NBHMM

Fig. 3. NBHMM for a simulated "triphone" structure



(a) Classical HMM

(b) NBHMM

Fig. 4. Hinton graph for Chinese syllable "Shi4"

If we use the whole-syllable classical HMM, some arbitrary size of the hidden state space has to be prefixed for each syllable. And, it has been found that the size of the state space has significant impact on the recognition rate. In our system with each state having 2 diagonal Gaussians, the recognition rate of the 6-state classical HMM for each syllable is 73.4%, while increasing the state-space to 16-state for each syllable gives a recognition rate of 80.1%. If we use the NBHMM model for each syllable, the variational inference automatically converges to using about 14-18 "effective" states for all the syllables, and the recognition rate is 78.9%. This resulting size of the state space coincides with the peaky recognition performance region of using the classical HMM.

We illustrate the resulting Hinton graphs of a Chinese syllable ("shi4") for the classical HMM (with $N = 66$) and the NBHMM (with $L = 66$) in Fig.4. As in the previous experiments, the classical HMM uses too many hidden states (being overfitted), while the NBHMM converges to use only 16 "effective" states. Besides, each of the corresponding rows in the $\mathbf{C}$ matrix for the NBHMM places nearly all weights on one or two components.

5. CONCLUSION

In this paper, we proposes a novel Nonparametric Bayesian HMM. The NBHMM assumes the state space is infinitely large and circumvents the difficulty of prefixing the size of state space. We also derive an efficient variational inference for this new model in the case of continuous observations. The experiments have demonstated its ability of structure discovery for both synthetic data and real-world speech recognition application.

6. REFERENCES

[1] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 365–381, 2004.

[2] M. J. Beal and P. Krishnamurthy, "Clustering gene expression time course data with countably infinte hidden Markov models," in *Uncertainty in Artificial Intelligence*, 2006.

[3] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[4] Y. W. Teh, K. Kurihara, and M. Welling, "Collapsed variational inference for HDP," in *Advances in Neural Information Processing Systems*, 2008, vol. 20.

[5] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," Tech. Rep., University College London, 2005.

[6] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Tech. Rep., University College London, 2003.

[7] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," in *Advances in Neural Information Processing Systems*, 2002.

[8] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[9] K. F. Lee, H. W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1990.