DIRICHLET PROCESS HMM MIXTURE MODELS WITH APPLICATION TO MUSIC ANALYSIS

Yuting Qi, John William Paisley and Lawrence Carin

Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708-0291

ABSTRACT

A hidden Markov mixture model is developed using a Dirichlet process (DP) prior, to represent the statistics of sequential data for which a single hidden Markov model (HMM) may not be sufficient. The DP prior has an intrinsic clustering property that encourages parameter sharing, naturally revealing the proper number of mixture components. The evaluation of posterior distributions for all model parameters is achieved via a variational Bayes formulation. We focus on exploring music similarities as an important application, highlighting the effectiveness of the HMM mixture model. Experimental results are presented from classical music clips.

Index Terms— Dirichlet Process, HMM mixture, Music, Variational Bayes.

1. INTRODUCTION

Music recognition, including music classification, retrieval, browsing, and recommendation systems, has been of significant recent interest. Correspondingly, ideas from statistical machine learning have attracted growing interest in the music-analysis community. For example, Gaussian mixture models have been used to represent the distribution of the MFCCs over all frames of an individual song [1][4]. However, no dynamic behavior of music is taken into account in these works. Since "the brain dynamically links a multitude of short events which cannot always be separated" [2], temporal cues are critical and contain information that should not be ignored. Therefore, music is treated as time-series data and hidden Markov models (HMMs), which can accurately represent the statistics of sequential data [8], have been introduced to model the overall music in [2][9] and more recently for music genre classification [10][12].

Building a single HMM for a song performs well when the music's "movement pattern" is relatively simple and thus the structure is of modest complexity (*e.g.*, the number of states is few). However, most real music is a complicated signal, which may have more than one "movement pattern" across the entire piece. Therefore an HMM mixture model is proposed in this paper to describe multiple "movement patterns" in music, with each pattern characterized by a single mixture component (an HMM). The work reported here develops an HMM mixture model in a Bayesian setting using a nonparametric Dirichlet process (DP) as a common prior distribution on the parameters of the individual HMMs. It has been proven that DP is rich enough to model parameters of individual components with arbitrarily high complexity, and flexible enough to fit them well without any assumptions about the functional form of the prior distribution [6][13]. Importantly, the number of mixture components need not be set *a priori* in the DP HMM mixture model. A variational Bayes [5] approach is considered to perform DP-based mixture modeling for efficient computation. In this paper we focus on HMM mixture models based on discrete observations; our method is applicable to any sequential discrete data set containing multiple underlying patterns.

The remainder of the paper is organized as follows. Section 2 provides an introduction to the Dirichlet process and its application to HMM mixture models. A variational Bayes inference method is developed in Section 3. Section 4 describes the music application as well as experimental results. Section 5 concludes the work.

2. DP-BASED HIDDEN MARKOV MIXTURE MODEL

2.1. Hidden Markov Mixture Model

The hidden Markov mixture model with K^* mixture components may be written as

$$p(\mathbf{x}|a_1,\cdots,a_{K^*},\Theta_1,\cdots,\Theta_{K^*}) = \sum_{k=1}^{K^*} a_k p(\mathbf{x}|\Theta_k), \quad (1)$$

where $\mathbf{x} = \{x_t\}_{t=1,T}$ is a sequence of observations, $p(\mathbf{x}|\Theta_k)$ represents the k^{th} HMM component with associated parameters Θ_k, and a_k represents the mixing weight for the k^{th} HMM, with $\sum_{k=1}^{K^*} a_k = 1$.

We assume a set $\mathbf{X} = \{\mathbf{x}_n\}_{n=1,N}$ of N sequences of data. Each data sequence $\mathbf{x}_n$ is assumed to be drawn from an associated HMM with parameters $\Theta_n = \{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n\}$, *i.e.*, $\mathbf{x}_n \sim \mathcal{H}(\Theta_n)$, where $\mathcal{H}(\Theta)$ represents the HMM. The set of associated parameters $\{\Theta_n\}_{n=1,N}$ are drawn *i.i.d* from a shared prior G, *i.e.*, $\Theta_n|G \overset{i.i.d}{\sim} G$. The distribution G is itself drawn from a distribution, in particular a Dirichlet process. The prior G encourages the clustering of the parameters $\{\Theta_n\}_{n=1,N}$ and each such cluster corresponds to an HMM mixture component in (1). The algorithm automatically determines an appropriate number of mixture components, balancing the DP-generated desire to cluster with the likelihood's desire to choose parameters that match the data $\mathbf{X}$ well. This balance between the likelihood and the DP prior is manifested in the posterior density function for parameters $\{\Theta_n\}_{n=1,N}$.

2.2. Dirichlet Process

The Dirichlet process, denoted as $DP(\alpha, G_0)$, is a random measure on measures and is parameterized by a positive scaling parameter α, often termed the "innovation parameter", and a base distribution G_0. Assume we have N random variables $\{\Theta_n\}_{n=1,N}$ distributed according to G, and G itself is a random measure drawn from a Dirichlet process,

$$\begin{aligned} \Theta_n|G &\sim G, \quad n = 1, \cdots, N, \\ G &\sim DP(\alpha, G_0), \end{aligned}$$

where G_0 is the expectation of G, $E[G] = G_0$. Define $\Theta^{-n} = \{\Theta_1, \cdots, \Theta_{n-1}, \Theta_{n+1}, \cdots, \Theta_N\}$ and let $\{\Theta_k^*\}_{k=1,K^*}$ be the distinct values taken by $\{\Theta_n\}_{n=1,N}$ and let n_k^{-n} be the number of values in Θ^{-n} that equal Θ_k^*. Integrating out G, the conditional distribution of Θ_n given Θ^{-n} follows a Pólya urn scheme and has the following form [13]

$$\Theta_n|\Theta^{-n}, \alpha, G_0 \sim \frac{1}{\alpha + N - 1}(\alpha G_0 + \sum_{k=1}^{K^*} n_k^{-n} \delta_{\Theta_k^*}). \quad (2)$$

where δ_{Θ_i} denotes the distribution concentrated at point Θ_i.

Equation (2) shows that when considering Θ_n given all other observations Θ^{-n}, this new sample is either drawn from base distribution G_0 with probability $\frac{\alpha}{\alpha+N-1}$, or is selected from the existing draws Θ_k^* according to a multinomial allocation, with probabilities proportional to existing groups sizes n_k^{-n}. Sethuraman [11] provides an explicit characterization of G in terms of a stick-breaking construction,

$$G = \sum_{k=1}^{\infty} p_k \delta_{\Theta_k^*}, \quad (3)$$

with

$$p_k = v_k \prod_{i=1}^{k-1}(1 - v_i), \quad (4)$$

where $v_k|\alpha \sim Beta(1, \alpha)$ and $\Theta_k^*|G_0 \sim G_0$. This representation shows the support of G consists of an infinite set of atoms located at Θ_k^*, drawn independently from G_0. The mixing weights p_k for atom Θ_k^* are given by successively breaking a unit length "stick" into an infinite number of pieces [11], with $0 \leq p_k \leq 1$ and $\sum_{k=1}^{\infty} p_k = 1$.

2.3. HMM mixture models with DP prior

Given the observed data $\mathbf{X} = \{\mathbf{x}_n\}_{n=1,N}$, each $\mathbf{x}_n$ is assumed to be drawn from its own HMM $\mathcal{H}(\Theta_n)$ parameterized by Θ_n with the underlying state sequence $\mathbf{s}_n$. The common prior G on all Θ_n is given as (3). Since G is discrete, different Θ_n may share the same value, Θ_k^*, and take the value of Θ_k^* with probability p_k. Introducing an indicator variable $\mathbf{c} = \{c_n\}_{n=1,N}$ and letting $c_n = k$ indicate that Θ_n takes the value of Θ_k^*, the hidden Markov mixture model with DP prior can be expressed as

$$\begin{aligned} \mathbf{x}_n|c_n, \{\Theta_k^*\}_{k=1}^{\infty} &\sim \mathcal{H}(\Theta_{c_n}^*), \\ c_n|\mathbf{p} &\sim Mult(\mathbf{p}), \\ v_k|\alpha &\sim Beta(1, \alpha), \\ \Theta_k^*|G_0 &\sim G_0, \end{aligned} \quad (5)$$

where $\mathbf{p} = \{p_k\}_{k=1,\infty}$ is given by (4) and $Mult(\mathbf{p})$ is the multinomial distribution with parameter $\mathbf{p}$.

Assuming $\mathbf{A}, \mathbf{B}$ and $\boldsymbol{\pi}$ are independent of each other, the base distribution G_0 is represented as $G_0 = p(\mathbf{A})p(\mathbf{B})p(\boldsymbol{\pi})$. For computational convenience (use of appropriate conjugate priors), we have the following prior distributions

$$P(\mathbf{A}|\mathbf{u}^A) = \prod_{i=1}^{I} \text{Dir}(\{a_{i1}, \cdots, a_{iI}\}; \mathbf{u}^A) \quad (6)$$

$$p(\mathbf{B}|\mathbf{u}^B) = \prod_{i=1}^{I} \text{Dir}(\{b_{i1}, \cdots, b_{iM}\}; \mathbf{u}^B) \quad (7)$$

$$p(\boldsymbol{\pi}|\mathbf{u}^\pi) = \text{Dir}(\{\pi_1, \cdots, \pi_I\}; \mathbf{u}^\pi), \quad (8)$$

where $\mathbf{u}^A = \{u_i^A\}_{i=1,I}$, $\mathbf{u}^B = \{u_m^B\}_{m=1,M}$, and $\mathbf{u}^\pi = \{u_i^\pi\}_{i=1,I}$ are parameters of the Dirichlet distribution. To learn α from the data, we place a prior distribution on it,

$$p(\alpha) = Ga(\alpha; \gamma_{01}, \gamma_{02}), \quad (9)$$

where $Ga(\alpha; \gamma_{01}, \gamma_{02})$ is the Gamma distribution with selected parameters γ_{01} and γ_{02}.

3. VARIATIONAL INFERENCE

Considering computational complexity in the infinite stick-breaking model, in practice we select an appropriate truncation level K (*i.e.*, finite sticks) that leads to a model virtually indistinguishable from the infinite DP model [5]. Since $\{\Theta_n\}_{n=1,N}$ may only take a subset of values from $\{\Theta_k^*\}_{k=1,K}$, the utilized number of mixture components K^* may be less than K (and the clustering properties of DP almost always yield less than K mixture components, unless α is very large) [7]. From Bayes' rule, we have

$$p(\Phi|\mathbf{X}, \Psi) = \frac{p(\mathbf{X}|\Phi)p(\Phi|\Psi)}{\int p(\mathbf{X}|\Phi)p(\Phi|\Psi)d\Phi}, \quad (10)$$

where $\Phi = \{\mathbf{A}^*, \mathbf{B}^*, \boldsymbol{\pi}^*, \mathbf{v}, \alpha, \mathbf{S}, \mathbf{c}\}$ are hidden variables of interest and $\Psi = \{\mathbf{u}^A, \mathbf{u}^B, \mathbf{u}^\pi, \gamma_{01}, \gamma_{02}\}$ are fixed parameters. The integration in the denominator of (10), the *marginal likelihood*, is generally intractable analytically. Variational methods are thus introduced to seek a distribution $q(\Phi)$ to approximate the true posterior distribution $p(\Phi|\mathbf{X}, \Psi)$. Consider the log marginal likelihood

$$\log p(\mathbf{X}|\Psi) = \mathcal{L}(q(\Phi)) + \mathcal{D}_{KL}(q(\Phi)||p(\Phi|\mathbf{X}, \Psi)), \quad (11)$$

where

$$\mathcal{L}(q(\Phi)) = \int q(\Phi) \log \frac{p(\mathbf{X}|\Phi)p(\Phi|\Psi)}{q(\Phi)} d\Phi \leq \log p(\mathbf{X}|\Psi), \quad (12)$$

and $\mathcal{D}_{KL}(q||p)$ is the KL divergence between q and p. The approximation of $p(\Phi|\mathbf{X}, \Psi)$ using $q(\Phi)$ can be achieved by minimizing $\mathcal{D}_{KL}(q(\Phi)||p(\Phi|\mathbf{X}, \Psi))$, which is equivalent to maximization of $\mathcal{L}(q(\Phi))$.

For the HMM mixture model proposed we assume

$$q(\Phi) = q(\alpha)q(\mathbf{v})\left\{\prod_{k=1}^{K}[q(\mathbf{A}_k^*)q(\mathbf{B}_k^*)q(\boldsymbol{\pi}_k^*)]\right\}\cdot \cdot\left\{\prod_{n=1}^{N}\prod_{c_n=1}^{K}[q(c_n)q(\mathbf{s}_{nc_n})]\right\}, \quad (13)$$

where $q(\mathbf{A}_k^*)$, $q(\mathbf{B}_k^*)$, $q(\boldsymbol{\pi}_k^*)$ have the same form as in (6)-(8) respectively but different parameters, $q(\mathbf{v}) = \prod_{k=1}^{K-1} q(v_k)$ with $q(v_k) = Beta(v_k; \beta_{1k}, \beta_{2k})$, and $q(\alpha) = Ga(\alpha; \gamma_1, \gamma_2)$. Once we learn the parameters of these variational distributions from the data, we obtain the approximation of $p(\Phi|\mathbf{X}, \Psi)$ by $q(\Phi)$. The joint distribution of Φ and observations $\mathbf{X}$ are given as

$$p(\mathbf{X}, \Phi|\Psi) = p(\alpha)p(\mathbf{v}|\alpha)\prod_{k=1}^{K}[p(\mathbf{A}_k^*)p(\mathbf{B}_k^*)p(\boldsymbol{\pi}_k^*)]\cdot \cdot\prod_{n=1}^{N}\prod_{c_n=1}^{K}[p(c_n|\mathbf{v})p(\mathbf{x}_n, \mathbf{s}_{nc_n}|\mathbf{A}^*, \mathbf{B}^*, \boldsymbol{\pi}^*, c_n)], \quad (14)$$

where priors $p(\mathbf{A}_k^*)$, $p(\mathbf{B}_k^*)$, $p(\boldsymbol{\pi}_k^*)$, and $p(\alpha)$ are given in (6)-(9) respectively, and $p(\mathbf{v}|\alpha) = \prod_{k=1}^{K-1} p(v_k|\alpha)$ with $p(v_k|\alpha) = Beta(v_k; 1, \alpha)$.

The term $\mathcal{L}(q)$ can be obtained by substituting (13) and (14) into (12). The optimization of the lower bound $\mathcal{L}(q)$ is realized by taking functional derivatives with respect to each of the $q(\cdot)$ distributions [3]. The update equations for the variational posteriors can be found in [7] and are omitted here for brevity.

The local maximum of the lower bound $\mathcal{L}(q)$ is achieved by iteratively updating the parameters of $q(\cdot)$ according to the update equations. We terminate the algorithm when the change in $\mathcal{L}(q)$ is negligibly small. Assuming that the states and the model parameters are independent and the model can be evaluated at the mean (or mode) of the variational posterior as suggested in [3], the prediction for a new observation sequence $\mathbf{y}$ can be easily obtained.

4. MUSIC EXPERIMENTS

The music clips are sampled at 22 kHz and we divide each clip into 25 ms non-overlapping frames. A 10-dimensional MFCC feature vector is extracted for each frame and then quantized into discrete symbols with LBG algorithm. For our experiments, we use a sequence of 1 second. This transforms the music into a collection of sequences, with each sequence assumed to originate from an HMM. All data and [7] can be found at http:// www.ee.duke.edu/∼jwp4/HMMMix.

4.1. Music Similarity Measure

Music similarity is computed based on the distance between the respective HMM mixture models. Let $\mathcal{M}_g$ be the learned HMM mixture model for music g, and $\mathcal{M}_h$ for music h. We draw a sample set S_g from $\mathcal{M}_g$ and S_h from $\mathcal{M}_h$. The distance between any two HMM mixture models is defined as

$$D(\mathcal{M}_g, \mathcal{M}_h) = \frac{1}{2}\left[L(\mathcal{M}_g|\mathcal{M}_h) + L(\mathcal{M}_h|\mathcal{M}_g)\right], \quad (15)$$

where $L(\mathcal{M}_a|\mathcal{M}_b) = \log p(S_b|\mathcal{M}_a) - \log p(S_b|\mathcal{M}_b)$ is a measure of how well model $\mathcal{M}_a$ matches observations generated by model $\mathcal{M}_b$, relative to how well $\mathcal{M}_b$ matches the observations generated by itself. The similarity $Sim(g, h)$ of the music g and h is defined by a kernel function as

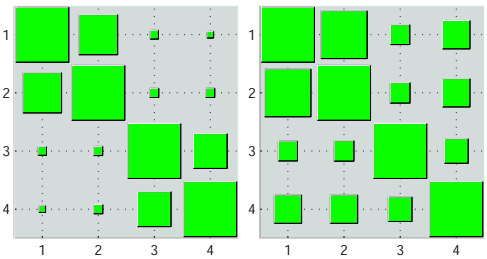
$$Sim(g, h) = \exp(-\frac{|D(\mathcal{M}_g, \mathcal{M}_h)|^2}{\sigma^2}), \quad (16)$$

where σ is a fixed parameter; we notice that σ will not change the order of similarities.

4.2. Results

We explore music similarity within the classical genre with two experiments. For comparison, we also model each piece of music as a DP Gaussian mixture model (DP GMM) [5][13], where the 10-dimensional MFCC feature vector of a frame corresponds to one data point in the feature space.

For our first experiment, we choose four 3-minute violin concerto clips from two different composers. Clips 1 and 2 are from Bach and are considered similar, clips 3 and clip 4 are from Stravinsky are also considered similar. The two pairs are considered different from each other. All four music clips are played using the same instruments, but their styles vary, indicating a high overlap in feature space, but significantly different movement. We built an HMM mixture model for each with truncation level, set to $K = 50$ and number of states to $I = 8$. The truncation level of the DP GMM was set to 50 as well. Fig. 1 shows the computed similarity between each clip for both HMM mixture and GMM modeling using a Hinton diagram, in which the size of a block is proportional to the value of the corresponding matrix elements. HMM mixture modeling produces results that fit with our intuition. However, our GMM results do not catch the connection between clips 3 and 4, and, proportionally, do not contrast clips 1 and 2 from 3 and 4 as well. The improved similarity recognition can be attributed to the temporal consideration given by



1: Bach-Violin Concerto BWV 1041 Mvt I 3: Stravinsky- Violin Concerto Mvt I
2: Bach-Violin Concerto BWV 1042 Mvt I 4: Stravinsky- Violin Concerto Mvt IV

Fig. 1. Hinton diagram for the similarity matrix for 4 violin clips. (a) by DP HMM mixture models; (b) by DP GMMs.

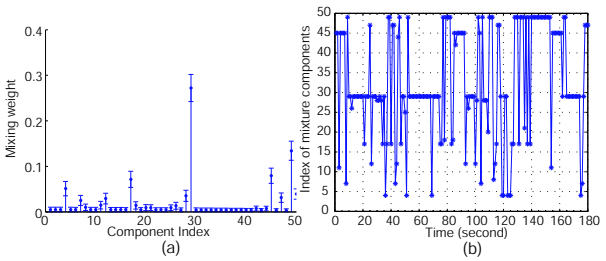


Fig. 2. Clip 4 (a) Mixing weights. (b) Memberships.

the HMM mixture model while the feature spaces are highly overlapped.

Fig. 2(a) shows the mixing weights of the DP HMM mixture models for clip 4 as an example. Although the number of significant weights is initially high, the algorithm automatically reduces this number by suppressing the superfluous components to that necessary to model each clip: the expected mixing weights for these unused HMMs are near zero with high confidence, indicated by the small variance of the mixing weights. The posterior membership (which is $\arg\max_k q(c_n = k)$) for clip 4 is displayed in Fig. 2(b), where those parts having similar styles should be drawn from the same HMM. The fact that the first 20 seconds of this clip are repeated during the last 20 can be seen in their similar membership patterns.

For our second experiment, we compute the similarities between ten 3-minute clips, which were chosen deliberately with the following intended clustering: 1) clip 1 is unique in style and instrumentation; 2) clips 2 and 3, 4 and 5, 6 and 7, and 9 and 10 are intended to be paired together 3) clip 8 is also unique, but is the same format (instrumentation) as clips 6 and 7. The Hinton diagrams of the corresponding similarity matrices are shown in Fig. 3. Again, our intuition is consistent in this experiment with HMM mixture modeling, but less accurate with GMM modeling. Though the GMM model does not contradict our intuition, the similarities are not as stark as in the HMM mixture, especially in the case of clip 1, which was selected to be unique.

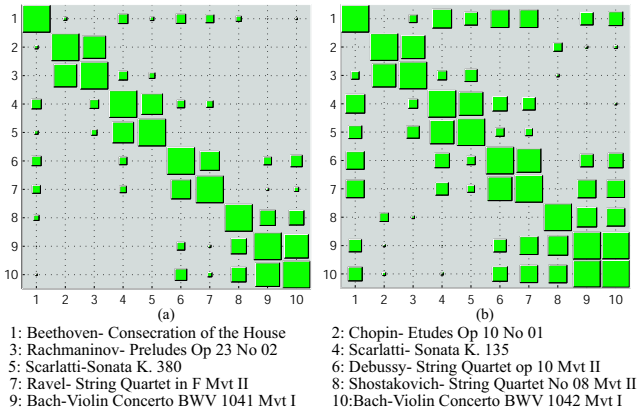


Fig. 3. Hinton diagram for the similarity matrix for 10 music clips. (a) by DP HMM mixture models; (b) by DP GMMs.

5. CONCLUSION

We have developed a discrete HMM mixture model in a Bayesian setting using DP priors, which has the advantage of avoiding the need to select the number of mixture components, through the encouragement of parameter sharing. A VB approach is employed for inference. The performance of HMM mixture modeling was demonstrated on music data sets and compared to the GMM, computing similarities between music as a measure of performance. In our experiments HMM mixture modeling outperforms the GMM.

6. REFERENCES

[1] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high's the sky?" *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, 2004.

[2] J.-J. Aucouturier and M. Sandler, "Segmentation of musical signals using hidden markov models," *in Proceedings of the 110th Convention of the Audio Engineering Society*, May 2001.

[3] M. J. Beal, "Variational algorithms for approximate bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

[4] A. Berenzweig, B. Logan, D. P. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music similarity measures," *Computer Music Journal*, vol. 28, no. 2, pp. 63–76, 2004.

[5] D. Blei and M. Jordan, "Variational methods for the dirichlet process," *ICML*, 2004.

[6] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[7] Y. Qi, J. W. Paisley, and L. Carin, "Hidden markov mixture models with dirichlet process priors," ECE Department, Duke University," Technical report, 2006.

[8] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *in Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[9] C. Raphael, "Automatic segmentation of acoustic musical signals using hidden markov models," *IEEE Trans. on PAMI*, vol. 21, no. 4, pp. 360–370, 1999.

[10] N. Scaringella and G. Zoia, "On the modeling of time information for automatic genre recognition systems in audio signals," *in Proceedings of the 6th ISMIR*, pp. 666–671, 2005.

[11] J. Sethuraman, "A constructive definition of the dirichlet prior," *Statistica Sinica*, vol. 2, pp. 639–650, 1994.

[12] X. Shao, C. Xu, and M. Kankanhalli, "Unsupervised classification of musical genre using hidden markov model," *ICME*, pp. 2023–2026, 2004.

[13] M. West, P. Muller, and M. Escobar, "Hierarchical priors and mixture models with applications in regression and density estimation," in *Aspects of Uncertainty*, P. R. Freeman and A. F. Smith, Eds. John Wiley, 1994, pp. 363–386.