The Dynamic Hierarchical Dirichlet Process

Lu Ren, Lawrence Carin {LR, LCARIN}@EE.DUKE.EDU
Department of Electrical and Computer Engineering

David B. Dunson DUNSON@STAT.DUKE.EDU
Department of Statistical Science
Duke University, Durham, NC 27708

Abstract

The dynamic hierarchical Dirichlet process (dHDP) is developed to model the time-evolving statistical properties of sequential data sets. The data collected at any time point are represented via a mixture associated with an appropriate underlying model, in the framework of HDP. The statistical properties of data collected at consecutive time points are linked via a random parameter that controls their probabilistic similarity. The sharing mechanisms of the time-evolving data are derived, and a relatively simple Markov Chain Monte Carlo sampler is developed. Experimental results are presented to demonstrate the model.

1. Introduction

The Dirichlet process (DP) mixture model (Escobar & West, 1995) has been widely used to perform density estimation and clustering, by generalizing finite mixture models to (in principle) infinite mixtures. In order to "share statistical strength" across different groups of data, the hierarchical Dirichlet process (HDP) (Teh et al., 2005) has been proposed to model the dependence among groups through sharing the same set of discrete parameters ("atoms"), and the mixture weights associated with different atoms are varied as a function of the data group. In the HDP, it is assumed that the data groups are exchangeable. However, in many real applications, such as seasonal market analysis and gene investigation for disease, data are measured in a sequential manner, and there is information in this temporal character that should ideally be exploited; this violates the aforementioned assumption of exchangeability.

Developing models for time-evolving data has recently been the focus of significant interest, and researchers have proposed various solutions directed toward specific applications. An early example is the order-based dependent DP (Griffin & Steel, 2006), in which the model is time-reversible but is not Markovian, and it requires one to specify how the mixture weights change through time. Another related work is the time-varying Dirichlet process mixture model (Caron et al., 2007) based on a modified Polya urn scheme (Blackwell & MacQueen, 1973), implemented by changing the number and locations of clusters over time. This method is easy to understand intuitively but has computational challenges for large data sets. To examine the temporal dynamics of scientific topics, latent Dirichlet allocation (Blei et al., 2003) (Griffiths & Steyvers, 2004) has been used as a generative model for analysis of documents. In order to explicitly model the dynamics of the underlying topics, Blei (Blei & Lafferty, 2006) proposed a dynamic topic model, in which the parameter at the previous time $t-1$ is the expectation for the distribution of the parameter at the next time t, and the correlation of the samples at adjacent times is controlled through adjusting the variance of the conditional distribution. Unfortunately, the non-conjugate form of the conditional distribution requires approximations in the model inference.

Recently Dunson (Dunson, 2006) proposed a Bayesian dynamic model to learn the latent trait distribution through a mixture of DPs, in which the latent variable density changes dynamically in location and shape across levels of predictors. This dynamic structure is considered in this paper to extend HDP to incorporate time dependence, and has the following features: (i) two data samples drawn at proximate times have a higher probability of sharing the same underlying

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

model parameters (atoms) than parameters drawn at disparate times; and (*ii*) there is a possibility that temporally distant data samples may also share model parameters, thereby accounting for possible distant repetition in the data.

2. Dynamic HDP

2.1. Background

A Dirichlet process is a measure on a measure G and is parameterized as $G \sim DP(\alpha_0, G_0)$, in which G_0 is a base measure and α_0 is a positive "precision" parameter. To provide an explicit form for a G drawn from $DP(\alpha_0, G_0)$, Sethuraman (Sethuraman, 1994) developed a stick-breaking construction:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k^*}, \quad \pi_k = \tilde{\pi}_k \prod_{i=1}^{k-1}(1-\tilde{\pi}_i) \tag{1}$$

where $\{\boldsymbol{\theta}_k^*\}_{k=1}^{\infty}$ represent a set of atoms drawn i.i.d. from G_0 and $\{\pi_k\}_{k=1}^{\infty}$ represent a set of weights, with the constraint $\sum_{k=1}^{\infty}\pi_k = 1$; each $\tilde{\pi}_k$ is drawn i.i.d. from $Be(1, \alpha_0)$. According to the construction in (1), a draw G from a $DP(\alpha_0, G_0)$ is discrete with probability one. Based on this important property, Teh (Teh et al., 2005) proposed a hierarchical Dirichlet process (HDP) to link the group-specific Dirichlet processes, learning the models jointly across multiple data sets.

Assume we have J groups of data and the j^{th} data set (group) is denoted as $\{x_{j,i}\}_{i=1,\ldots,N_j}$. For each of these data sets, $x_{j,i}$ is drawn from the model $x_{j,i} \overset{ind}{\sim} F(\boldsymbol{\theta}_{j,i})$ with parameters $\boldsymbol{\theta}_{j,i} \overset{iid}{\sim} G_j$, and the parameters $\{\boldsymbol{\theta}_{j,i}\}_{i=1,\ldots,N_j}$ are likely to assume the atoms $\boldsymbol{\theta}_k^*$ for which the associated sticks $\pi_{j,k}$ are large, as a consequence of the form of G_j given by (1); for the J data sets, different group-specific G_j are drawn from $DP(\alpha_{j0}, G_0)$, in which G_0 is drawn from another DP. The generative model for HDP is represented as:

$$\begin{aligned} x_{j,i} &\overset{ind}{\sim} F(\boldsymbol{\theta}_{j,i}) \\ \boldsymbol{\theta}_{j,i} &\overset{iid}{\sim} G_j \\ G_j &\overset{ind}{\sim} DP(\alpha_{j0}, G_0) \\ G_0 &\sim DP(\gamma, H) \end{aligned} \tag{2}$$

where $j = 1, \ldots, J$ and $i = 1, \ldots, N_j$.

Under this hierarchical structure, not only can different observations $x_{j,i}$ and $x_{j,i'}$ in the same group share the same parameters $\boldsymbol{\theta}^*$ based on the stick weights represented by G_j, but also the observations across different groups might share parameters as a consequence of the discrete form of G_0 (all G_j are composed of the same set of atoms $\{\boldsymbol{\theta}_k^*\}_{k=1}^{\infty}$). The clusters in each group j, assumed by the set $\{\boldsymbol{\theta}_{j,i}\}_{i=1,\ldots,N_j}$, are inferred via the posterior density function on the parameters, with the likelihood function selecting the set of discrete parameters $\{\boldsymbol{\theta}_k^*\}_{k=1}^{\infty}$ most consistent with the data $\{x_{j,i}\}_{i=1,\ldots,N_j}$. Meanwhile, clusters (and, hence, associated cluster parameters $\{\boldsymbol{\theta}_k^*\}_{k=1}^{\infty}$) are shared across multiple data sets, as appropriate.

Although the HDP introduces a dependency between the J groups, the data sets are assumed exchangeable. However, in many applications, the data may be collected sequentially, and one may have a prior belief that sharing of data is more probable when the data sets are collected at similar points in time. The purpose of this paper is to extend the HDP to account for such temporal information.

Before proceeding, it will prove useful to consider an alternative form of the HDP model, as derived in (Teh et al., 2005). Specifically, each draw G_j may be expressed as:

$$\begin{aligned} G_j &= \sum_{k=1}^{\infty} \pi_{j,k} \delta_{\boldsymbol{\theta}_k^*} \\ \boldsymbol{\pi}_j &\overset{ind}{\sim} DP(\alpha_{0j}, \boldsymbol{\beta}) \\ \boldsymbol{\beta} &\sim Stick(\gamma) \\ \boldsymbol{\theta}_k^* &\overset{iid}{\sim} H \end{aligned} \tag{3}$$

where $Stick(\gamma)$ stochastically generates an infinite set of sticks $\{\beta_1, \beta_2, \ldots\}$, based on a stick-breaking process of the form in (1), here with parameter γ, satisfying the constraint $\sum_{i=1}^{\infty}\beta_i = 1$.

2.2. Bayesian Dynamic Structure

Similar to HDP, we again consider J data sets but now using an explicit assumption that the data sets are collected sequentially, with $\{x_{1,i}\}_{i=1,\ldots,N_1}$ collected first, $\{x_{2,i}\}_{i=1,\ldots,N_2}$ collected second, and with $\{x_{J,i}\}_{i=1,\ldots,N_J}$ collected last. Since our assumption is that a time evolution exists between adjacent data groups, the distribution G_{j-1}, from which $\{\boldsymbol{\theta}_{j-1,i}\}_{i=1,\ldots,N_{j-1}}$ are drawn, is likely related to G_j, from which $\{\boldsymbol{\theta}_{j,i}\}_{i=1,\ldots,N_j}$ are drawn.

To specify explicitly the dependence between G_{j-1} and G_j, Dunson (Dunson, 2006) proposed a Bayesian dynamic mixture DP (DMDP), in which G_j shares features with G_{j-1} but some innovation may also occur. The DMDP has the drawback that mixture components can only be added over time, so that one ends up with more components at later times as an artifact of the model.

In the dHDP, we have

$$G_j = (1-\tilde{w}_{j-1})G_{j-1} + \tilde{w}_{j-1}H_{j-1} \tag{4}$$

where $G_1 \sim DP(\alpha_{01}, G_0)$, H_{j-1} is called an innovation distribution drawn from $DP(\alpha_{0j}, G_0)$, and $\tilde{w}_{j-1} \sim Be(a_{w(j-1)}, b_{w(j-1)})$. In this way, G_j is modified from G_{j-1} by introducing a new innovation distribution H_{j-1}, and the random variable $\tilde{w}_{j-1}$ controls the probability of innovation (*i.e.*, it defines the mixture weights). As a result, the relevant atoms adjust with time, and it is probable that proximate data will share the same atoms, but with the potential for transient innovation.

Additionally, we assume that $G_0 \sim DP(\gamma, H)$ as in the HDP to enforce that G_0 is discrete, which manifests another important aspect of the dynamic HDP: the same atoms are used for *all* G_j, but with different time-evolving weights. Consequently, the model encourages sharing between temporally proximate data, but it is also possible to share between data sets widely separated in time.

Providing now more model details, the discrete base distribution drawn from $DP(\gamma, H)$ may be expressed as:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\boldsymbol{\theta}_k^*} \tag{5}$$

where $\{\boldsymbol{\theta}_k^*\}_{k=1,2,\ldots,\infty}$ are the global parameter components (atoms), drawn independently from the base distribution H and $\{\beta_k\}_{k=1,2,\ldots,\infty}$ are drawn from a stick-breaking process $\boldsymbol{\beta} \sim Stick(\gamma)$, defined as:

$$\beta_k = \tilde{\beta}_k \prod_{l<k}(1-\tilde{\beta}_l) \qquad \tilde{\beta}_k \overset{iid}{\sim} Be(1,\gamma) \tag{6}$$

We also have J groups of data. G_j represents the prior for the mixture distribution associated with the global components in group j, H_{j-1} represents the associated prior for the innovation mixture distribution, and this yields the explicit priors used in (4):

$$\begin{gathered} G_1 = \sum_{k=1}^{\infty} \pi_{1,k} \delta_{\boldsymbol{\theta}_k^*}, H_1 = \sum_{k=1}^{\infty} \pi_{2,k} \delta_{\boldsymbol{\theta}_k^*}, \ldots, \\ H_{J-1} = \sum_{k=1}^{\infty} \pi_{J,k} \delta_{\boldsymbol{\theta}_k^*} \end{gathered} \tag{7}$$

where, analogous to the discussion at the end of Section 2.1, the different weights $\boldsymbol{\pi}_j$ are independent given $\boldsymbol{\beta}$ since $G_1, H_1, \ldots, H_{J-1}$ are independent given G_0; the relationship between $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}$ is proven (Teh et al., 2005) to be

$$\boldsymbol{\pi}_j | \alpha_{0j}, \boldsymbol{\beta} \sim DP(\alpha_{0j}, \boldsymbol{\beta}) \tag{8}$$

To further develop the dynamic relationship from G_1 to G_J, we extend the mixture structure in (4) from group to group:

$$\begin{aligned} G_j &= (1-\tilde{w}_{j-1}) G_{j-1} + \tilde{w}_{j-1} H_{j-1} \\ &= \prod_{l=1}^{j-1} (1-\tilde{w}_l) G_1 + \sum_{l=1}^{j-1} \{ \prod_{m=l+1}^{j-1} (1-\tilde{w}_m) \} \tilde{w}_l H_l \\ &= w_{j1} G_1 + w_{j2} H_1 + \ldots + w_{jj} H_{j-1} \end{aligned} \tag{9}$$

where $w_{jl} = \tilde{w}_{l-1} \prod_{m=l}^{j-1} (1-\tilde{w}_m)$, for $l = 1, 2, \ldots, j$, with $\tilde{w}_0 = 1$. It can be easily verified that $\sum_{l=1}^{j} w_{jl} = 1$ for each $\mathbf{w}_j$, which is the prior probability that the data in group j will be drawn from the mixture distribution: $G_1, H_1, \ldots, H_{j-1}$. If all $\tilde{w}_j = 0$, all of the groups share the same mixture distribution G_1 and the model reduces to a Dirichlet mixture model, and if all $\tilde{w}_j = 1$ the model reduces to the HDP. Therefore, the dynamic HDP is more general than both DP and HDP, with each a special case. A visual representation of the model is depicted in Figure 1.

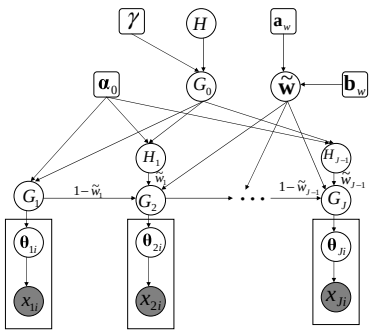


Figure 1. General graphical model for the dynamic HDP.

According to (9), the observation $x_{j,i}$ will choose a mixture distribution from $\boldsymbol{\pi}_{1:j}$ based on $Mult(\mathbf{w}_j)$ to be drawn from the global parameter components $\{\boldsymbol{\theta}_k^*\}_{k=1}^{\infty}$. We let $r_{j,i}$ be a variable to indicate which mixture distribution is taken from $\boldsymbol{\pi}_{1:j}$ to draw the observation $x_{j,i}$; $z_{j,i}$ is a parameter component indicator variable. An alternative form of the dHDP model is represented as:

$$\begin{array}{cc} \boldsymbol{\theta}_k^* | H \sim H, & \boldsymbol{\beta} | \gamma \sim Stick(\gamma) \\ \tilde{w}_j | a_{wj}, b_{wj} \sim Be(\tilde{w}_j | a_{wj}, b_{wj}), & r_{j,i} | \tilde{\mathbf{w}} \sim \mathbf{w}_j \\ \boldsymbol{\pi}_j | \alpha_{0j}, \boldsymbol{\beta} \sim DP(\alpha_{0j}, \boldsymbol{\beta}), & z_{j,i} | \boldsymbol{\pi}_{1:j}, r_{j,i} \sim \boldsymbol{\pi}_{r_{j,i}} \\ x_{j,i} | z_{j,i}, (\boldsymbol{\theta}_k^*)_{k=1}^{\infty} \sim F(\boldsymbol{\theta}_{z_{j,i}}^*), & \end{array} \tag{10}$$

and a graphical representation is shown in Figure 2, in which we add a gamma prior for γ and for the components of the vector $\boldsymbol{\alpha}_0$: $Pr(\gamma) = Ga(\gamma; \gamma_{01}, \gamma_{02})$ and $Pr(\boldsymbol{\alpha}_0) = \prod_{j=1}^{J} Ga(\alpha_{0j}; c_0, d_0)$. The form of the parametric model $F(\cdot)$ may be varied depending on the application.

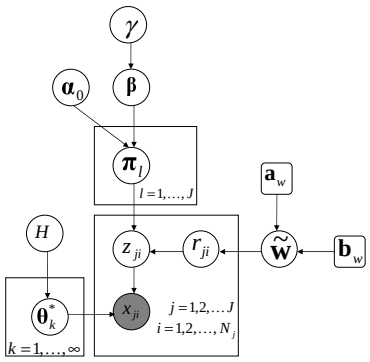


Figure 2. Graphical representation of the dHDP from a stick-breaking view.

2.3. Sharing Properties

To see the mixture structure in a discrete partition space $\mathcal{A} = (A_1, \ldots, A_K)$, we consider

$$\begin{aligned} &G_j(A_1,\ldots,A_K)|G_{j-1},\tilde{w}_{j-1} \sim \\ &(1-\tilde{w}_{j-1})G_{j-1}(A_1,\ldots,A_K)+\tilde{w}_{j-1}H_{j-1}(A_1,\ldots,A_K) \\ &\triangleq G_{j-1}(A_1,\ldots,A_K)+\triangle_j(A_1,\ldots,A_K) \end{aligned} \tag{11}$$

where $\triangle_j(A_1,\ldots,A_K) = \tilde{w}_{j-1}\{H_{j-1}(A_1,\ldots,A_K) - G_{j-1}(A_1,\ldots,A_K)\}$ is the random deviation from G_{j-1} to G_j.

Theorem 1. Given any discrete partition $\mathcal{A}$, we have:

$$\begin{aligned} &E\{\triangle_j(\mathcal{A})|G_{j-1},\tilde{w}_{j-1},H,\gamma,\alpha_{0j}\} \\ =&\tilde{w}_{j-1}\{H(\mathcal{A})-G_{j-1}(\mathcal{A})\} \end{aligned} \tag{12}$$

$$\begin{aligned} &V\{\triangle_j(\mathcal{A})|G_{j-1},\tilde{w}_{j-1},H,\gamma,\alpha_{0j}\} \\ =&\tilde{w}_{j-1}^2\frac{(1+\gamma+\alpha_{0j})H(\mathcal{A})(1-H(\mathcal{A}))}{(1+\alpha_{0j})(1+\gamma)} \end{aligned} \tag{13}$$

According to Theorem 1, given the previous mixture distribution G_{j-1}, the expectation of the deviation from G_{j-1} to G_j is controlled by $\tilde{w}_{j-1}$. Meanwhile, the variance of the deviation is both related with $\tilde{w}_{j-1}$ and the precision parameters γ, α_{0j}. To consider limiting cases, we observe the following:

- if $\tilde{w}_{j-1} \to 0$, $G_j = G_{j-1}$;
- if $G_{j-1} \to H$, $E(G_j(\mathcal{A})|G_{j-1},\tilde{w}_{j-1},H,\gamma,\alpha_{0j}) = G_{j-1}(\mathcal{A})$;
- if $\gamma \to \infty$ and $\alpha_{0j} \to \infty$, $V(\triangle_j(\mathcal{A})|G_{j-1},\tilde{w}_{j-1},H,\gamma,\alpha_{0j}) \to 0$.

These limiting cases yield insights on the underlying dependence between adjacent groups.

Theorem 2. The correlation coefficient of the distributions between two adjacent groups G_{j-1} and G_j for $j = 2,\ldots,J$ is

$$\begin{aligned} &Corr(G_{j-1},G_j) \\ =&\frac{E\{G_j(\mathcal{A})G_{j-1}(\mathcal{A})\}-E\{G_j(\mathcal{A})\}E\{G_{j-1}(\mathcal{A})\}}{[V\{G_j(\mathcal{A})\}V\{G_{j-1}(\mathcal{A})\}]^{1/2}} \\ =&\frac{\sum_{l=1}^{j-1}\frac{w_{jl}w_{j-1,l}}{1+\alpha_{0l}}\cdot\frac{\alpha_{0l}+\gamma+1}{\gamma+1}}{[\sum_{l=1}^{j}\frac{w_{jl}^2}{1+\alpha_{0l}}\cdot\frac{\alpha_{0l}+\gamma+1}{\gamma+1}]^{1/2}[\sum_{l=1}^{j-1}\frac{w_{j-1,l}^2}{1+\alpha_{0l}}\cdot\frac{\alpha_{0l}+\gamma+1}{\gamma+1}]^{1/2}} \end{aligned} \tag{14}$$

To compare the similarity of two data groups, the correlation coefficient defined in Theorem 2 can be calculated from the posterior expectation of $\mathbf{w}$, $\boldsymbol{\alpha}_0$ and γ as a local similarity measure.

2.4. Posterior Computation

A modification of the block Gibbs sampler (Ishwaran & James, 2001) is proposed for dHDP inference. Since in practice the $\{\pi_k\}_{k=1}^{\infty}$ in (1) diminish quickly with increasing k, a truncated stick-breaking process (Ishwaran & James, 2001) is employed here, with a large truncation level K, to approximate the infinite stick breaking process. In the dHDP, the second level of DPs associated with the dynamic structure is the only part different from HDP (see Figure 2). Due to the limited space, we only give the conditional posterior distributions for $\tilde{\mathbf{w}}$, $\tilde{\boldsymbol{\pi}}$, $\mathbf{r}$ and $\mathbf{z}$.

The conditional distribution of $\tilde{w}_l$, for $l = 1,\ldots,J-1$ has the simple form:

$$(\tilde{w}_l|\cdots) \sim Be(a_w + \sum_{j=l+1}^{J} n_{j,l+1}, b_w + \sum_{j=l+1}^{J}\sum_{h=1}^{l} n_{jh}) \tag{15}$$

where $n_{jh} = \sum_{i=1}^{N_j}\delta(r_{ji} = h)$. In (15) and in the results that follow, for simplicity, the distributions $Be(a_{wj}, b_{wj})$ are set with fixed parameters $a_{wj} = a_w$ and $b_{wj} = b_w$ for all time samples.

The conditional distribution of $\tilde{\pi}_{lk}$, for $l = 1,\ldots,J$ and $k = 1,\ldots,K$, is updated under the conjugate prior: $\tilde{\pi}_{lk} \sim Be(\alpha_{0,l}\beta_k, \alpha_{0,l}(1-\sum_{m=1}^{k}\beta_m))$, which is specified in (Teh et al., 2005). Then the conditional posterior of $\tilde{\pi}_{lk}$ has the form

$$\begin{aligned} &(\tilde{\pi}_{lk}|\cdots) \sim Be(\alpha_{0l}\beta_k + \sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(r_{ji}=l,z_{ji}=k), \\ &\alpha_{0l}(1-\sum_{l=1}^{k}\beta_l) + \sum_{j=1}^{J}\sum_{i=1}^{N_j}\sum_{k'=k+1}^{K}\delta(r_{ji}=l,z_{ji}=k')) \end{aligned} \tag{16}$$

The update of the indicator variables r_{ji} and z_{ji}, for $j = 1,\ldots,J$ and $i = 1,\ldots,N_j$ are completed by generating samples from multinomial distributions with

entries as follows:

$$Pr(r_{ji} = l|\cdots) \propto \tilde{w}_{l-1} \prod_{m=l}^{j-1}(1-\tilde{w}_m) \cdot \tilde{\pi}_{lz_{ji}} \prod_{q=1}^{z_{ji}-1}(1-\tilde{\pi}_{lq}) \cdot Pr(x_{ji}|\boldsymbol{\theta}^*_{z_{ji}}) \tag{17}$$

where $l = 1, \ldots, j$. The posterior probability $Pr(r_{ji} = l)$ is normalized so that $\sum_{l=1}^{j} Pr(r_{ji} = l) = 1$.

$$Pr(z_{ji} = k|\cdots) \propto \tilde{\pi}_{r_{ji}k} \prod_{k'=1}^{k-1}(1-\tilde{\pi}_{r_{ji}k'}) \cdot Pr(x_{ji}|\boldsymbol{\theta}^*_k) \tag{18}$$

where $k = 1, \ldots, K$ and the posterior is also normalized by a constant $\sum_{k=1}^{K} Pr(z_{ji} = k)$.

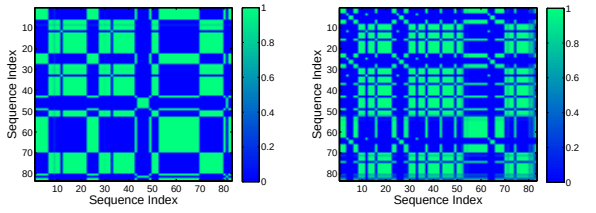
The remaining variables specified in (10) are sampled in the same ways as in HDP (Teh et al., 2005). The component parameters $\boldsymbol{\theta}^*_k$ for $k = 1, \ldots, K$ are considered for different model forms depending on the specific applications. For the results that follow, it is of interest to consider a hidden Markov model (HMM) mixture (Qi et al., 2007) and Gaussian mixture model (GMM), in which $\boldsymbol{\theta}^*_k$ respectively represent the state-transition matrix, the observation matrix, the initial-state distribution for the HMM and the mean vector and covariance matrix for GMM. For more details about sampling for such models, see (Qi et al., 2007) and (Escobar & West, 1995). The Gibbs sampling algorithm was tested carefully under different initializations and the diagnostic method in (Raftery & Lewis, 1992) is used to demonstrate rapid convergence and good mixing (for the results considered, convergence based on this method was observed for a burn-in of 200 samples, followed by a subsequent 4000 samples).

3. Experimental Results

3.1. Music Segmentation

It is of interest to segment music, to infer inter-relationships between different parts of a given piece, as well as between different pieces. Here we consider segmentation of music, where a given piece is divided into contiguous subsequences, with each subsequence modeled via a hidden Markov model (HMM). The dHDP model is useful in this application in enforcing the idea that contiguous subsequences are likely to be within the same music segment, and therefore are likely to share HMM parameters. However, when the segment changes, these changes are detected via innovation within the dHDP.

The music under consideration is the first movement "Largo - Allegro" from the Beethoven piano Sonata No. 17 (Newman, 1972). As is widely employed for analysis of such audio data, MFCC features are extracted and discretized with vector quatization (Qi et al., 2007); each of the aforementioned subsequences corresponds to a sequence of codewords (we here employ a discrete HMM). The basic form of the Bayesian representation of a discrete HMM is as discussed in (Qi et al., 2007). The piece is transformed into 4980 discrete symbols, divided into 83 subsequences of equal length (the codebook has 16 codes, and 8 states are employed for each HMM); each subsequence corresponds to 6 secs in the music. To model the time dependence between adjacent subsequences, each subsequence corresponds to one group in the dHDP HMM mixture and will choose one set of HMM parameters according to the corresponding mixture weights. In the dHDP framework, one subsequence can share the old DP mixture distributions with the previous ones or it might be drawn from an innovation DP mixture, which may be also shared by the following time series in a similar manner. To encourage that adjacent subsequences be shared, the prior for $\tilde{\mathbf{w}}$ is specified as $E(\tilde{\mathbf{w}}) < 0.5$. The product of most interest here is the segmentation of the music, with the specific HMM parameters of secondary importance.



(a) (b)

Figure 3. Similarity matrix $E(z'z)$ from HMM mixture modeling of the Sonata. (a) dHDP-HMM, (b) HDP-HMMs.

To represent the time dependence of the piece, the similarity measure $E(\mathbf{z}'\mathbf{z})$ (see $\mathbf{z}$ in Eq. (18)) is computed across each pair of subsequences, as shown in Figure 3, in which larger values represent higher probability of the two corresponding subsequences being shared during parameter inference. Based upon a discussion in (Newman, 1972), the movement alternates seeming peacefulness with sudden turmoil (1st-6th subsequences), after some time expanding into a haunting "storm" in which the peacefulness is lost (7th-21st subsequences). After the recurrence of the same pattern (22nd-42nd subsequences) and a small transition, the movement starts a long recitative section in a slow tone (53rd-69th subsequences). Then through the crescendo, previous disturbed tones come back again until the music goes to the peaceful epilogue

(after the 70th subsequence). See (Newman, 1972) for more details on the Sonata. This is deemed to be an interesting piece for study because it is well characterized in the music literature, as briefly summarized above, and because it is anticipated to have repeated segments over the length of the piece. In Figures 3(a) and (b) we compare the dHDP and HDP, respectively, the latter computed by fixing all $\tilde{\boldsymbol{w}} = 1$ in the dHDP model. The dHDP and HDP yield related results, but the former yields a smoother segmentation, in good agreement with the music theory discussed above.

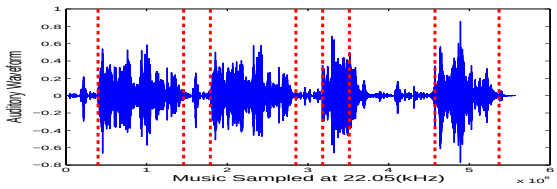


Figure 4. Segmentation results for Beethoven piano music from the dHDP HMMs (red dash lines represent segment positions and blue curves represent the auditory waveform).

Based on the results from the dHDP HMM, which effectively yields a model with smoothly time-evolving statistics, we segment the music and present the associated auditory waveform in Figure 4. By examining the waveform and the results in Figure 3, we note that the dHDP segments the music into dominant auditory phenomena, but it is less sensitive to noticeable but temporally localized events in the music, yielding a segmentation that is consistent with the music theory. By contrast, the HDP results in Figure 3(b) are evidently more sensitive to these local temporal bursts in the waveform.

3.2. Gene Expression Data

As a second example, we consider the time-evolving characteristics of gene-expression data, here for a Dengue virus study (Hibberd et al., 2006). Concerning a model for the gene-expression data at one time snapshot, Dunson (Dunson, 2006) proposed a latent response model based on a linear regression structure; we extend this model for time-evolving gene-expression data via dHDP (with comparison as well to HDP).

Assume $\mathbf{y}_{ji}$ is a feature vector with dimension p for $j = 1, \ldots, J$ and $i = 1, \ldots, N_j$ (index j corresponds to time, i represents a particular cell from which a sample is collected, and p denotes the number of genes being modeled). Each $\mathbf{y}_{ji}$ is represented as

$$\mathbf{y}_{ji} = \boldsymbol{\mu} + \boldsymbol{\lambda}\eta_{ji} + \boldsymbol{\varepsilon}_{ji} \tag{19}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ is the intercept vector and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ represents factor loadings. We define a hidden variable η_{ji} underlying the observation $\mathbf{y}_{ji}$ to be associated with the i^{th} sample at time t_j. The error term $\boldsymbol{\varepsilon}_{ji}$ is also a vector of dimension p and each coefficient $\varepsilon_{ji,d}$ is independently drawn from a Student-t distribution. To eliminate the problem of model identifiability, we incorporate the constraints that $\mu_1 = 0$ and $\lambda_1 = 1$, as (Dunson, 2006) discusses. In the present model, one cannot explicitly associate $\boldsymbol{\eta}$ exclusively with the virus; however, since these are cell data, it is anticipated that the virus represents the dominant phenomena.

We have access to expressions of thousands of genes from each sample (cell) for multiple consecutive times $t_1, t_2, \ldots, t_J$. For each time t_j, there are N_j samples measured from different cells (Hibberd et al., 2006). Although these samples may have different observations in gene expressions at the same time, due to individual diversity, the hidden variable $\boldsymbol{\eta}$ (see (19)) underlying the observations may have similar characteristics. Based on this consideration, the $\boldsymbol{\eta}$ underlying the observations in one group corresponding to one time are assumed to be drawn from a Gaussian mixture model. They may also share the same mixture distribution for proximate time points, under the assumption of the dHDP model.

The Dengue gene expression data (Hibberd et al., 2006) are divided into six groups of samples measured at different times and the number of samples in each group are 10, 12, 12, 10, 12, 9 (the specific time points associated with these data are respectively 3, 6, 12, 24, 48 and 72 hours); each sample has 19,143 genes. To deal with such high-dimensional data, the Fisher score (Duda & Hart, 1973) is used to preliminarily select $p = 5000$ genes as being the most relevant (variable across time and cell), and then we use the dHDP mixture model discussed above to analyze the time evolution existing in these gene samples.

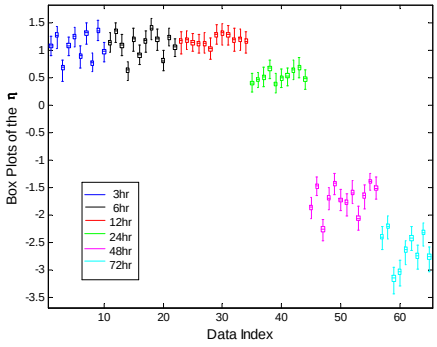


Figure 5. Median values and associated uncertainty based on posterior distributions of the hidden variables $\boldsymbol{\eta}$.

Based on the samples collected from the Gibbs sampling after burn-in, the posterior distributions (includ-

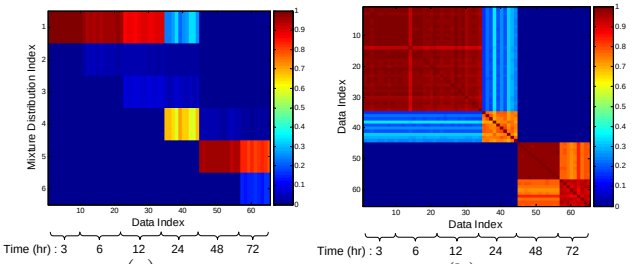


Figure 6. The dHDP GMM modeling for the gene expression data. (a) The posterior distribution of $\mathbf{r}$. (b) The similarity matrix $E[z'z]$.

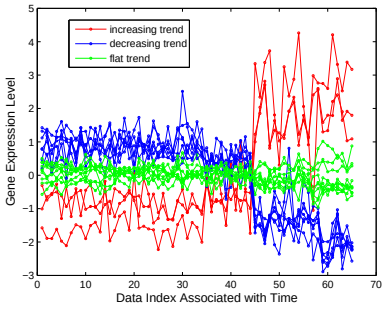


Figure 7. The first ten inferred important genes (color red and blue) and the relatively unrelated genes (color green).

ing the minimum, median, maximum, 25th and 75th percentiles of the values) for all components of $\boldsymbol{\eta}$ underlying these samples at different times are shown in Figure 5. Time points 3hr, 6hr and 12hr appear to share a similar pattern, but the $\boldsymbol{\eta}_{t=12}$ seem to have smaller diversity among different samples. From 24hrs, $\boldsymbol{\eta}$ drops slightly to a new pattern and they drop significantly again at 48hr. The posterior of indicator $\mathbf{r}$ is plotted in Figure 6(a) to show the mixture-distribution sharing relationship across different groups. Figure 6(b) shows the similarity measure $E(\mathbf{z}'\mathbf{z})$ across every pair of samples; here z_{ji} is the indicator variable for the η_{ji} associated with time t_j (see Eq. (18)).

Consider the factor loadings vector $\boldsymbol{\lambda}$, which has components linked to the p genes under consideration. The larger the value of $|\lambda_d|$, the more influence the pattern contained in $\boldsymbol{\eta}$ has on the corresponding gene at the d^{th} dimension. Therefore, according to the posterior mean of $|\lambda_d|$ for all d from the Gibbs sampling we rank the genes based on their importance.

In Figure 7 we plot the expression levels over time for the 10 most important and 10 least important genes. The red and blue curves show two different time patterns and their values have either an increasing or a decreasing trend with time, depending on whether the associated λ is positive or negative. The green curves represent the genes with no apparent relation to the virus (as determined by the analysis) due to the lack of a systematic trend over time.

As discussed in Section 2.2, if all $\tilde{w}_j$ are set to one for $j = 1, \ldots, J-1$, the dHDP reduces to HDP and all the temporal groups are conditionally exchangeable. It is of interest to compare the dHDP with HDP both in the sharing mechanism and parameter estimation. In practice, acquisition of the gene-expression data is expensive, and it is desirable to reduce the number of samples required. To consider this issue, we reduced the samples size to four at each time point, and plot the data similarity matrix $E[\boldsymbol{z}'\boldsymbol{z}]$ for HDP and dHDP respectively in Figures 8(a) and (b).

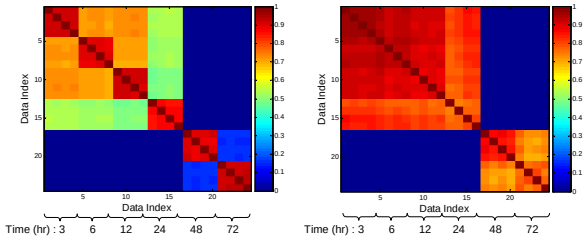


Figure 8. Similarity matrix $E[z'z]$ with four samples for each temporal group. (a) HDP, (b) dHDP.

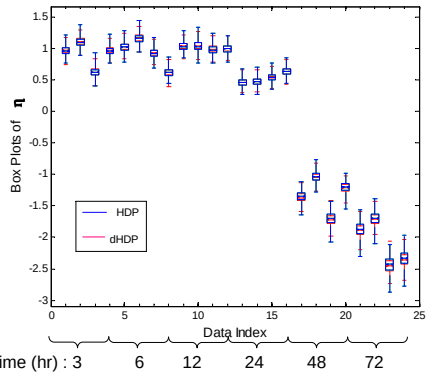
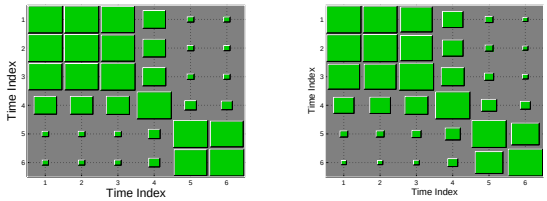


Figure 9. Comparison of dHDP and HDP with box plots of the hidden variables $\boldsymbol{\eta}$ as the sample size is reduced to four for each temporal group (the standard deviation based on dHDP is 12.1% reduced on average relative to HDP; the means are very similar).

Compared with HDP, dHDP has more sharing between the related groups (as expected from model construction), and despite the reduced data samples the dHDP yields an inter-relationship between the different times that is consistent with that in Figure 6(b) which employs all of the available data. In Figure 9 we compare dHDP and HDP estimation of $\boldsymbol{\eta}$ based on four samples per time point. These results show that dHDP has a smaller estimation uncertainty for most $\boldsymbol{\eta}$ relative to HDP, which is attributed to proper temporal

sharing explicitly imposed by dHDP. As the sample size is increased, the differences between dHDP and HDP diminish.

Finally, correlation coefficients between two groups are calculated from the samples drawn from the Gibbs sampler, according to (14) and plotted as a matrix in Figure 10; this representation is an additional benefit of the dynamic structure explicitly imposed within dHDP (of potential biological interest). The size of each small block at the i^{th} row and j^{th} column is proportional to the value of the correlation coefficient associated with group i and group j. We note based on Figure 10 that such inference appears to be accurate (or at least consistent) even with diminished sample size.



(a) (b)

Figure 10. Similarity matrix between data at different time points based on the correlation coefficients (14), as computed from the dHDP posterior. (a) using all available data, (b) using four samples for each temporal group.

4. Conclusions

The proposed dynamic hierarchical Dirichlet process (dHDP) extends the HDP (Teh et al., 2005), imposing a dynamic time dependence so that the initial mixture model and the subsequent time-dependent mixtures share the same set of components (atoms). The experiments indicate that the dHDP is an effective model for analysis of time-evolving data. Concerning future research, more efficient inference methods will be considered, such as collapsed sampling (Welling et al., 2007) and variational Bayesian inference (Blei & Jordan, 2004).

References

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via polya urn schemes. *Ann. Statist.*, *1*, 353–355.

Blei, D. M., & Jordan, M. I. (2004). Variational methods for the dirichlet process. *Proceedings of the International Conference on Machine Learning.*

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the International Conference on Machine Learning.*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Caron, F., Davy, M., & Doucet, A. (2007). Generalized poly urn for time-varying dirichlet process mixtures. *Proceedings of the International Conference on Uncertainty in Artificial Intelligence(UAI).*

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis.* Wiley.

Dunson, D. B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, *7*, 551–568.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*, 577–588.

Griffin, J. E., & Steel, M. F. J. (2006). Order-based dependent dirichlet processes. *Journal of the American Statistical Association*, *101*, 179–194.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proc Natl Acad Sci U S A*, *101, Suppl 1*, 5228–5235.

Hibberd, M. L., Vasudevan, S. G., Ling, L., & George, J. (2006). *Time course expression data of human cell lines infected with dengue virus serotype2 ngc* (Technical Report). Genome Institute of Singapore.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*, 161–173.

Newman, A. S. (1972). *Sonata in the classic era (a history of the sonata idea).* W. W. Norton.

Qi, Y., Paisley, J. W., & Carin, L. (2007). Music analysis using hidden markov mixture models. *IEEE Transactions on Signal Processing*, *55*, 5209–5224.

Raftery, A. E., & Lewis, S. (1992). How many iterations in the gibbs sampler? *Bayesian Stat.*, *4*, 763–773.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, *2*, 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). *Hierarchical dirichlet processes* (Technical Report). Dept. of Computer Science, National University of Singapore.

Welling, M., Porteous, I., & Bart, E. (2007). Infinite state bayes-nets for structured domains. *Proceedings of the International Conference on Neural Information Processing Systems.*