MONDRIAN HIDDEN MARKOV MODEL FOR MUSIC SIGNAL PROCESSING

Masahiro Nakano, Yasunori Ohishi, Hirokazu Kameoka, Ryo Mukai, Kunio Kashino

NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan
nakano.masahiro@lab.ntt.co.jp

ABSTRACT

This paper discusses a new extension of hidden Markov models that can capture clusters embedded in transitions between the hidden states. In our model, the state-transition matrices are viewed as representations of relational data reflecting a network structure between the hidden states. We specifically present a nonparametric Bayesian approach to the proposed state-space model whose network structure is represented by a Mondrian Process-based relational model. We show an application of the proposed model to music signal analysis through some experimental results.

Index Terms— Bayesian nonparametrics, hidden Markov model, Mondrian process

1. INTRODUCTION

Hidden Markov models (HMMs) have been a major success story in time series modeling. In particulars, the recent nonparametric Bayesian approaches constitute one of the most successful directions. [1] described the infinite hidden Markov model (IHMM) and [2] further formalized the hierarchical Dirichlet process hidden Markov model (HDP-HMM), that develops an HMM with a (potentially) infinite state space. However, after ten years of work on Bayesian nonparametric HMMs [3, 4], a number of significant improvements are still required. This paper addresses an important issue, namely whether we can capture "clusters" in transitions between the hidden states.

When we apply HMMs (which share hidden states) to multiple time series, a network structure between the hidden states is considered to be embedded in the state-transition matrices, since they represent relational data. Actually, many sequences (whose state-transition matrices are analyzed by HMMs) often have square and rectangular clusters. Chord progressions in music are a telling example. In the context of HMMs, for the observed audio signals, each hidden state is expected to correspond to an individual $chord$, such as C♯, Dm7 and Edim. That is, the sequences of hidden states are regarded as chord progressions. Popular songs are usually in a particular key. For example, the chords used within the key "C major" are generally drawn from the "C major scale", and are C, Dm, Em, F, G, Am and Bm-5. As a result, transitions between these chords tend to be preferred. On the other hand, when we want to add some "colors" to a piece, various chords can be used, including $borrowed\ chords$, $altered\ chords$, $secondary\ dominants$. A brief example is the progression: "F→G→{Em, E7, G♯dim}→Am" (often adopted by $Eurobeat$ music), i.e., E7 or G♯ are sometimes used in place of Em. Thus, "{E7, G♯dim}→Am" may form a rectangular cluster on the state-transition matrices. In fact, the concept of $chord\ substitution$ in music theory has been a useful technique for assigning one chord in place of another. Therefore, we believe that a technique for automatically finding the clusters in chord transitions is essential when building machines that can understand music.

To capture such cluster structures, we have to deal with the unknown numbers of hidden states and the unknown numbers of clusters in the transitions. This paper addresses this issue using a Bayesian nonparametric fusion of HMMs and relational models. Bayesian nonparametric approaches have been developed in both time-series modeling and relational data analysis.

2. RELATED WORK AND PRELIMINARIES

2.1. HMM and hierarchical Dirichlet processes

Briefly, an HMM consists of a hidden state sequence $\boldsymbol{z} = (z_1, \ldots, z_T)$ and a corresponding observation sequence $\boldsymbol{y} = (y_1, \ldots, y_T)$. Each state z_t denotes the index of the state at time t. Transitions between states are governed by Markov dynamics parameterized by the transition probabilities $\boldsymbol{\pi}$, where $\pi_{i,j} = p(z_t = j \mid z_i = i)$ and $z_t \sim \boldsymbol{\pi}_{z_{t-1}}$. The current state z_t indexes the parameter θ_{z_t} which parametrizes the observation likelihood for that state: $y_t \mid z_t \sim f(\theta_{z_t})$.

Generally, it is not easy to determine the model complexity, i.e., the number of states K. To address this problem, we can typically employ the hierarchical Dirichlet process, which can be used to develop an HMM with a potentially infinite state space. A Dirichlet process (DP) provides a discrete random probability measure. It is very useful for constructing an (countably) infinite number of states. Given a positive concentration parameter γ and a base measure H on a

parameter space Θ, the measure drawn from a Dirichlet process, $G_0 \sim \mathrm{DP}(\gamma, H)$ can be expressed as follows: $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}$, $\theta_k \sim H$, where $\delta_\cdot$ denotes the Dirac measures. By using a well-known method, the weights $\boldsymbol{\beta}$ can be sampled via a stick-breaking construction [5]: $\beta_k = \beta'_k \prod_{l=1}^{k-1}(1 - \beta'_l)$, $\beta'_k \sim \mathrm{Beta}(1, \gamma)$. This is often expressed as $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$. Intuitively, the weight drawn from DP can be regarded as the mixture proportion of an infinite-dimensional multinomial distribution. We might thus use DP priors over the transition probabilities $\boldsymbol{\pi}_i$ $(i = 1, 2, \dots)$ to obtain an infinite HMM. However, there may be no coupling across transitions out of different states if the transition probabilities are given independent priors. To avoid such a problem, we can use the following hierarchical Dirichlet process. A hierarchical Dirichlet process (HDP) is a set of Dirichlet processes coupled through a shared base measure that is itself drawn from a DP: $G_i \sim DP(\alpha, G_0)$ $(i = 1, 2, \dots)$. Each random measure can be expressed as follows: $G_i = \sum_{k=1}^{\infty} \pi_{i,k} \delta_{\theta_k}$ $(i = 1, 2, \dots)$. Identifying $\pi_{i,j}$ and θ_k as describing the transition probabilities from state i to state j and the emission parameters, we can define the HDP-HMM as

$$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma), \quad \boldsymbol{\pi}_i \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}),$$
$$z_t \mid z_{t-1} \sim \boldsymbol{\pi}_{z_{t-1}}, \ \theta_k \sim H, \ y_t \mid z_t, \theta \sim f(\theta_{z_t}). \quad (1)$$

A number of extensions to the HDP-HMM have been proposed. We briefly review the two extensions that are most closely related to our work. [6] described a sticky HDP-HMM, which captures smoothly varying dynamics. When modeling systems with state persistence, we would like to be able to express slow dynamics. They proposed sampling transition distributions $\boldsymbol{\pi}_i$ as follows: $\boldsymbol{\pi}_i \sim \mathrm{DP}(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \kappa\delta_i}{\alpha + \kappa})$. An amount $\kappa > 0$ exhibits a bias for self-transitions.

[7] proposed a block diagonal infinite HMM (BD-IHMM), which biases the transition matrix so that it is block diagonal. When modeling sequences consisting of sub-process, there is a higher tendency of transitions occurring between states with the same sub-process, i.e. a transition matrix will exhibit nearly block diagonal. They introduced $\boldsymbol{\rho} \sim \mathrm{GEM}(\zeta)$ to partition the countably infinite number of states into blocks. Each state k belongs to the block $b_k \sim \boldsymbol{\rho}$. Then, they propose sampling transition distributions $\boldsymbol{\pi}_i$ as follows: $\boldsymbol{\pi}_i \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}^*_i)$,

$$\beta^*_{i,j} = \frac{1}{1+\xi} \beta_j (\xi^*_i)^{\delta(b_i = b_j)}, \ \ \xi^*_i = 1 + \frac{\xi}{\sum_k \beta_k \delta(b_i = b_k)},$$

where a nonnegative parameter ξ control the amount of prior bias for within-block transitions.

2.2. Relational models and the Mondrian Process

Stochastic block models have been developed to model relational data, which are observations of relationships between sets of entities. While we only focus on binary relations here, the ideas can be straightforwardly extended to other likelihood models. Suppose we are given D relational data as arrays of random binary variables $R_{d,i,j}$, where i and j index entities $x_i \in X$ and $y_j \in Y$. The goal of the stochastic block models is to assign each pair (i, j) to a cluster.

[8] presented a Bayesian nonparametric approach to relational models called the infinite relational model (IRM), which is expressed as the following generative model. Both sets $x_1, x_2, \dots$ and $y_1, y_2, \dots$ are first clustered using a Chinese restaurant process prior. Then, each relation $R_{d,i,j}$ is generated as follows: $R_{d,i,j} \sim \mathrm{Bernoulli}(\phi_{B_{i,j}})$, where $S_{i,j}$ shows the cluster to which the pair (x_i and y_j) belong, and ϕ_B denotes a cluster-specific Beta variable. The partitions obtained by the IRM are not usually parsimonious, since such partitions look like regular grids. When one area of the array requires a split, it causes the other parts to be divided, even if the data suggest there is no such structure. This issue motivated the proposal of the Mondrian process.

Mondrian processes generate random partitions on product spaces not constrained to be regular grids. [9] and [10] presented a Markov process on hierarchical partitioning. As shown in [9] and [10], we can construct a Mondrian process based on the following famous recursive procedure. For simplicity, we focus only on a Mondrian process $\mathcal{M} \sim \mathrm{MP}(\lambda, (a, A), (b, B))$ on a 2-dimensional rectangle $(a, A) \times (b, B)$ [11]. Starting with an initial *budget* λ, the Mondrian process makes a sequence of cuts that split rectangles into sub-rectangles. Each cut incurs a random cost E. If E exceeds the budget, the process halts and returns the current partitions. Otherwise, the Mondrian process makes an axis-aligned cut uniformly at random along (a, A) and (b, B). The cost E to cut the rectangle $(a, A) \times (b, B)$ is distributed exponentially: $E \sim \mathrm{Exp}(A - a + B - b)$. After a cut is made (for example, suppose a cut split (a, A) into (a, x) and (x, A)), the cost E is subtracted from the budget λ and a new budget $\lambda' = \lambda - E$ is calculated. Both sub-rectangles are drawn independently from $\mathcal{M}_< \sim \mathrm{MP}(\lambda', (a, x) \times (b, B))$ and $\mathcal{M}_> \sim \mathrm{MP}(\lambda', (x, A) \times (b, B))$. The partition on $(a, A) \times (b, B)$ can be expressed as: $\mathcal{M}_< \cup \mathcal{M}_>$.

As shown in [9], we obtain the MP-based relational model as follows: We first generate a hierarchical partitioning of $[0, 1]^2$ from the Mondrian process, and then the column/row location of each entry is generated based on a geometrical interpretation of the $\mathrm{Uniform}[0, 1]$ random variable.

3. MONDRIAN HIDDEN MARKOV MODEL

The goal of a Mondrian HMM is not only to analyze the temporal dynamics of time series but also to capture clusters embedded in transitions between the hidden states. Suppose we are given D time series $\{Y_{d,1}, Y_{d,2}, \dots, Y_{d,T_d}\}$ $(d = 1, \dots, D)$ (Fig. 1 (left) illustrates a case of two time series). Although the lengths of the time series are not necessarily the same, subscripts are omitted for simplicity of notation, i.e.,

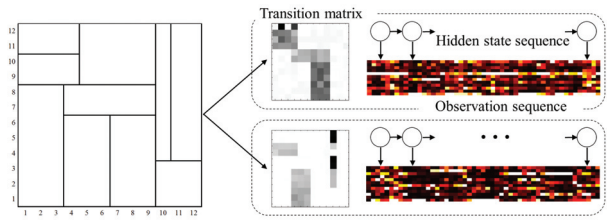

Fig. 1. Illustration of Mondrian HMM.

we express $T_1, \ldots, T_D$ as T. We assume that D time series share the hidden states and their corresponding parameters for emission distributions. A network structure between the hidden states is considered embedded in state-transition matrices. As relational models explain the relationships between sets of entities, the clusters on state-transition matrices provide an insight into the network structure of the hidden states. We thus consider that state-transition matrices should be partitioned into square and rectangular clusters.

Using a Mondrian process with the Lebesgue measure, we first sample a random partition $\mathcal{M}$ of the unit square into blocks: $\mathcal{M} \sim \mathrm{MP}(\lambda, [0,1], [0,1])$. We next sample hypothetical locations S_k (related to transitions "out") and S'_k (related to transitions "in") of the hidden state k on the unit $[0,1]$: $S_k \sim \mathrm{Uniform}[0,1]$, $S'_k \sim \mathrm{Uniform}[0,1]$. By using the partition $\mathcal{M}$, and the hypothetical locations S_k $(k = 1, 2, \ldots)$ and S'_k $(k = 1, 2, \ldots)$, we can express "clusters" in the transitions between the hidden states.

Binary variables $R_{d,i,j}$ show whether the transition from state i to state j is preferable. That is, whether the transition $i \to j$ will be more likely, when $R_{d,i,j} = 1$. They are controlled independently according to the clusters: $R_{d,i,j} \sim \mathrm{Bernoulli}(\phi_{d,B_{i,j}})$, $\phi_{d,B} \sim \mathrm{Beta}(a_0, a_1)$, where $B_{i,j}$ denotes the block $B \in \mathcal{M}$ such that $(S_i, S'_j) \in B$. The transition probabilities are formulated as: $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$, $\boldsymbol{\pi}_{d,i} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}^*_{d,i})$, $\xi^*_{d,i} = 1 + \frac{\xi_d}{\sum_l R_{d,i,l}\beta_l}$, $\beta^*_{d,i,j} = \frac{1}{1+\xi_d}\beta_j(\xi^*_{d,i})^{R_{d,i,j}}$. These modified weights $\boldsymbol{\beta}^*_{d,i}$ are reminiscent of the BD-IHMM [7]. $\boldsymbol{\beta}^*_{d,i}$ are scaled to favor relatively higher probabilities $\pi_{d,i,j}$ for transition from state i to state j such that $R_{d,i,j} = 1$. As a result, the Markov chains are controlled by the nearly Mondrian-textured transition matrices. Additionally, we would like to avoid rare transitions causing unreasonable relational structures in the transition between hidden states. We thus regard rare transitions as the spiky noise of sequences. We introduce binary indicators, as used in image processing [12]:

$$c_{d,t} \sim \mathrm{Bernoulli}(c'_d), \quad c'_d \sim \mathrm{Beta}(a_2, a_3). \tag{2}$$

Thus, the Markov chains are generated from the mixture of the general transition probabilities $\boldsymbol{\pi}_{d,i}$ and the special proportions $\boldsymbol{\pi}'_d$: $\boldsymbol{\pi}'_d \sim \mathrm{DP}(\alpha, \boldsymbol{\beta})$,

$$Z_{d,t} \mid Z_{d,t-1} \sim \mathbb{I}[c_{d,t} = 1]\boldsymbol{\pi}_{d,Z_{d,t-1}} + \mathbb{I}[c_{d,t} = 0]\boldsymbol{\pi}'_d.$$

Finally, emission parameters for $\boldsymbol{\theta}$ are drawn from H and the observation sequences $Y_{d,t}$ are generated from $f(\theta_{Z_{d,t}})$.

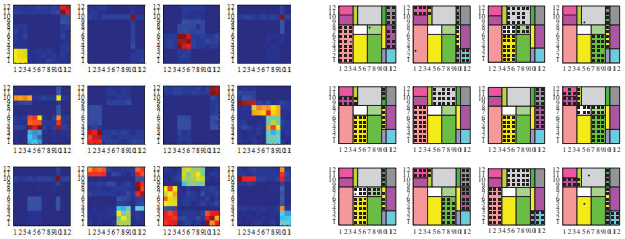

Fig. 2. Matrices of transition counts from ground truth (left), and Mondrian partitions and means of binary relations on transition matrices inferred by the Mondrian HMM (right). Larger black squares have larger $\mathbb{E}[R_{d,i,j}]$ values. Mondrian HMM can capture an active network structure between the hidden states.

4. EXPERIMENTS

"Chord progression analysis" in music is an interesting and attractive application for machine learning. The goal of the experiments was to capture clusters in the transitions between chords from music signals. Note that chord recognition tasks themselves are not our main focus although chord detection implicitly becomes part of the procedure. We are interested in not only chord recognition but also chord progression analysis from audio signals, i.e., we want to find clusters in transitions between the hidden states. We used a standard dataset consisting of 180 Beatles songs collected from 12 albums. The chord annotations are provided for these songs [13]. We employed the beat-synchronous chroma features [14] as input data. Each frame of the features represents the intensity associated with each of the 12 semitones (e.g. piano keys) of the musical octave onto which the entire spectrum is projected (i.e., all octaves are folded together). The beat-synchronous chroma features consist of one (12-dimensional) feature vector per beat, which identifies the beat segmentation times in the music audio. We excerpted major scale sections and transposed them to C major, because chords depend on the musical key. That is, chroma features were circularly shifted so that all the songs were transposed to C major. The pre-measured sequences had 64703 frames.

Synthetic data: Our first experiment was intended to illustrate the behavior of the Mondrian HMM. We generated 12 sequences from a 12-state HMM with single-Gaussian emissions. Each state randomly chose one label from the chord annotation data, and learned means and variances from the corresponding chroma features. Transition probabilities favored active blocks at a rate of 20/21. Each sequence had 5000 steps generated from the synthetic HMMs. Fig. 2 (left) shows the matrices of transition counts in the true state sequences (ground truth). As shown in Fig. 2 (right), the Mondrian HMM can capture nearly the active network structure even though there are little-used states. To evaluate the predictive power of learned models, we compare the test-set log probabilities of the Mondrian HMM and the HDP-HMM on two types of synthetic data: (1) 100 randomly-generated 1-dimensional Poisson emission dataset, and (2) 10 randomly-generated 12-dimensional Poisson emission dataset. Each

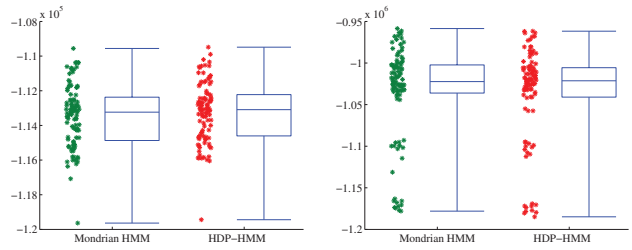

Fig. 3. **Left:** Test-set log probabilities for the 100 synthetic datasets (1-dimensional Poisson emissions) for models learned by the Mondrian HMM and the HDP-HMM. **Right** Test-set log probabilities for the 10 synthetic datasets (each data was tested 10 times) (12-dimensional Poisson) for models learned by the Mondrian HMM and the HDP-HMM. We can verify that the flexibility of the Mondrian HMM does not significantly degrade the predictive power.

Table 1. Chord recognition performance

	HDP	sticky	IRM	Mondrian
30 songs	57.2	62.8	61.6	61.1
60 songs	58.6	63.8	64.2	64.5

consists of 12 training sequences (each has 5000 steps) and 12 test sequences (each has 5000 frames) generated from 12-state HMMs. We performed the Mondrian HMM and the HDP-HMM once per each data. The result of the test-set log probabilities for 100 synthetic data is shown in Fig. 3.

Beatles chord progression analysis: Our second experiment applies the Mondrian HMM to real-world data, 180 Beatles songs. It is very challenging to identify "chords" of hidden states in a fully unsupervised manner because the model does not know the correspondence relationship between "chord names" and chroma features. We therefore employed a semi-supervised learning approach. Of the 180 pre-measured songs, we divided into test songs and training songs. The songs for training provided pairs of the chroma features and the "known" chord labels, but did not tell us anything about the state transitions. The state-transition probabilities were inferred only from the test data. Chords were reduced to 25 labels (12 major + 12 minor + no chord) by analogy with the classical chord recognition tasks. We used an infinite Dirichlet process mixture of Gaussians as emission distributions similar to [6], since we found single-Gaussian emission distributions to be less effective.

We compared the performance of the HDP-HMM, the sticky HMM, the Mondrian HMM. For the Mondrian HMM, we evaluate two variants: one is the original version; the other is the IRM for the Mondrian process (called IRM-HMM). We evaluated the chord recognition performance based on the best of the five runs. We ran each algorithm for 2000 iterations, and then the last-sampled hidden sequences were compared with the ground truth to calculate the accuracy of the chord detection. Table 1 shows results for 3-fold and 6-fold cross validations. This experimental result is strongly in favor of the sticky HMM. The reason is that, for example, *4 beat* songs typically stay in the same states three times. However, the Mondrian HMM performs similarly to the sticky HMM. We thus confirm that the flexibility of the Mondrian HMM does not degrade the performance of hidden state assignments.

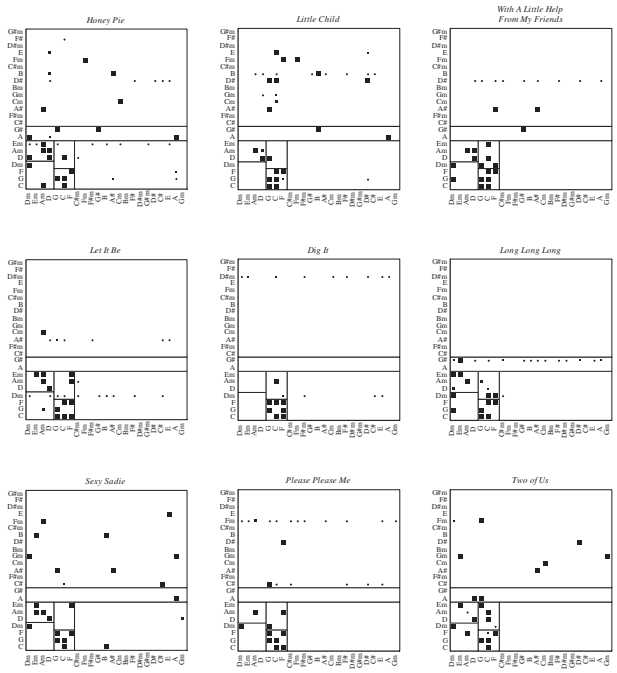

Fig. 4. Learned Mondrian partitions and $\mathbb{E}[R_{d,i,j}]$ of 9 songs. Four blocks in the bottom left show typical transitions, nearly *Diatonic chords* {C, Dm, Em, F, G, A, Bm} generated from C major scale. What is very interesting for us is that these blocks include "D". The Beatles often used "D" as *secondary dominants*, while the transition "Dm"→G is generally an extremely typical pattern (called *two-five*). Moreover, the bottom-left block captures the set {C, F, G}, which is called *three chords*. As some songs are actually built around *three chords*, these chords have strong connections. The model could find such a structure from the songs.

Finally, Fig. 4 shows the $\mathbb{E}[R_{d,i,j}]$ values of 9 songs (randomly chosen from 30 test songs) are illustrated. "A♯m" (no assignments) and "no chord" are omitted for visibility. We used a similar procedure to that used for the synthetic data to evaluate $\mathbb{E}[R_{d,i,j}]$. Larger black squares shows the "active" state transitions.

5. CONCLUSION

This paper deals with a new extension of HMMs, which can capture clusters in transitions between the hidden states. Our method is based on the nonparametric Bayesian fusion of the HDP-HMM and a Mondrian process-based relational model. While we confirmed that the Mondrian HMM sees the possibility of capturing clusters in transitions between the hidden states through experiments on both synthetic and real-world data, we anticipate many directions of improvement. In the future, we are interested in embedding relational models in the n-gram models.

6. REFERENCES

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

[2] Y. W. Teh, M. I. Jordan, M. Beal, and D. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, pp. 1566–1581, 2006.

[3] D. Wingate, N. D. Goodman, D. M. Roy, and J. B. Tenenbaum, "The infinite latent events model," in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2009.

[4] Finale Doshi-Velez and Nicholas Roy David Wingate, Joshua Tenenbaum, "Infinite dynamic Bayesian networks," in *Proceedings of the International Conference on Machine Learning*, 2011.

[5] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, 1994.

[6] E. Fox, E. Sudderth, M. I. Jordan, and A. Willsky, "An hdp-hmm for systems with state persistence," in *Proceedings of the International Conference on Machine Learning*, 2008.

[7] T. Stepleton, Z. Ghahramani, G. Gordon, and T. S. Lee, "The block diagonal infinite hidden markov model," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2009.

[8] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems ofconcepts with an infinite relational model," in *Proceedings of the National Conference onArtificial Intelligence*, 2006.

[9] D. Roy and Y. W. Teh, "The Mondrian process," in *Advances in Neural Information Processing Systems*, 2009.

[10] D. M. Roy, *Computability, inference and modeling in probabilistic programming*, Ph.D. thesis, Massachusetts Institute of Technology, 2011.

[11] P. Wang, K. B. Laskey, C. Domeniconi, and M. I. Jordan, "Nonparametric bayesian co-clustering ensembles," in *Proceedings of the International Conference on Data Mining*, 2011.

[12] L. Ren, Y. Wang, D. Dunson, and L. Carin, "The kernel beta process," in *Advances in Neural Information Processing Systems*, 2011.

[13] C. Harte, M. Sandler, S. Abdallah, and E. Gómez, "Symbolicrepresentation of musical chords: A proposed syntax for textannotations," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2005.

[14] D. Ellis and G. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2007.