Sequential Bayesian Nonparametric Multimodal Topic Models for Video Data Analysis*

SUMMARY  Topic modeling as a well-known method is widely applied for not only text data mining but also multimedia data analysis such as video data analysis. However, existing models cannot adequately handle time dependency and multimodal data modeling for video data that generally contain image information and speech information. In this paper, we therefore propose a novel topic model, sequential symmetric correspondence hierarchical Dirichlet processes (Seq-Sym-cHDP) extended from sequential conditionally independent hierarchical Dirichlet processes (Seq-CI-HDP) and sequential correspondence hierarchical Dirichlet processes (Seq-cHDP), to improve the multimodal data modeling mechanism via controlling the pivot assignments with a latent variable. An inference scheme for Seq-Sym-cHDP based on a posterior representation sampler is also developed in this work. We finally demonstrate that our model outperforms other baseline models via experiments.

key words: bayesian nonparametric methods, multimedia data mining, hierarchical Dirichlet processes, topic models

1. Introduction

Bayesian topic modeling is well-known as a high-performance method for text data mining, such as Latent Dirichlet Allocation (LDA) [2] and hierarchical Dirichlet processes (HDP) [3]. According to their theory, topic models are based on a hypothesis that text words in each document are generated from a mixture distribution of latent topics, where each latent topic is represented as a word distribution. Recently, some researchers become interested in multimedia data analysis (such as video data analysis) combined with topic modeling methods [4]–[7]. For video data, they usually have more complex data components (multimodal data enriched with image information and speech information) than general text data. To be specific, the image information represents a sequence of video frames, while the speech information represents speech transcript. From the perspective of topic models, the image information within each video frame can be characterized with a number of visual words, which are drawn from each video frame by using some image feature extraction techniques. However, a tricky problem for video data analysis is how to deal with the visual words and speech words simultaneously within one unified topic model. Besides, the time dependency issue among sequential video frames should also be taken into account when modeling a video with a topic model.

For these reasons, we previously proposed a sequential correspondence HDP (Seq-cHDP) model [7] extended from HDP for learning video data. Seq-cHDP provides a unified video data modeling framework incorporating two important mechanisms in accordance with the features of video data: one is a time-dependency mechanism that models sequential video frames with a time dependent first-order Markovian assumption among Dirichlet processes, and the other is the correspondence mechanism that models the multimodal data by establishing a unidirectional dependency among them. However, we realize that such correspondence mechanism within Seq-cHDP also has two drawbacks: one is that the pivot data type (modality) is required to be manually specified beforehand, and the other is that the pivot is always fixed throughout the whole generative process. In this paper, we focus on exploring a way to further improve the multimodal data modeling scheme in Bayesian nonparametric topic models. Based on the framework of sequential HDP (Seq-HDP) [7], we firstly propose a novel topic model called sequential conditionally independent HDP (Seq-CI-HDP) inspired by [8] to show the performance when representing the correspondence mechanism in a loose way. Then, Seq-cHDP is reviewed to illustrate another correspondence mechanism with a unidirectional dependency approach. Combining the features of these two correspondence mechanisms, we finally develop a more flexible model called sequential symmetric correspondence HDP (Seq-Sym-cHDP), which can spontaneously select the pivot data type by using a latent variable during the generative process of video multimodal data. We infer Seq-Sym-cHDP based on a posterior representation sampler [9]. Finally, we demonstrate that our model outperforms other baselines by experiments.

2. Related Work

As the two most classic and fundamental topic models, latent Dirichlet allocation (LDA) [2] presents a good theory on the Bayesian topic model for text mining, where each text document is represented as a mixture of topics, and each topic is represented as a word distribution. Then, HDP [3] was proposed as an alternative non-parametric topic model that models multiple text documents as multiple infinite Dirichlet processes connected by sharing the same mixture

Copyright © 2018 The Institute of Electronics, Information and Communication Engineers
of components.

Based on LDA and HDP, researchers recently attempted to apply topic models for modeling the video data. Wang et al. [4] extended LDA and HDP to model activities and interactions within videos. However, they did not consider the time-dependency issue among sequential clips. Hospedales et al. [5] provides a sophisticated framework for the unsupervised learning of scene characteristics, dynamically screening and identifying irregular spatiotemporal patterns with a proposed Markov clustering topic model (MCTM). However, MCTM extended from LDA is restricted to manually specify the number of topics beforehand. Kuettel et al. [6] proposed a cascaded topic model called dependent Dirichlet processes and hidden Markov models (DDP-HMM) for jointly modeling the spatio-temporal dependencies of moving agents in complex dynamic scenes (DDP-HMM). Then on the third layer, G^f denotes the local measure corresponding to the j-th frame within the f-th video. In fact, these local measures are not independent, since video frames belonging to the same video are sequential and time related. Hence, we establish a transfer link based on a Markovian assumption between two adjacent local measures G^f_{j−1} and G^f_j in Seq-HDP with a transfer weight w^f_j. We describe the generative process of Seq-HDP as follows.

- **Step 1:** The overall measure G manipulating all the components within Seq-HDP is sampled from a Dirichlet process as G ∼ DP(ξ, H), where ξ is a concentration parameter for generating G.
- **Step 2:** The global measure G^f_0 for each video is drawn from G^f_0 ∼ DP(γ^f, G), where f is the index of the video, and γ^f is a concentration parameter for generating G^f_0.
- **Step 3:** Due to the time-dependency mechanism, both the upper global measure G^f_0 and previous local measure G^f_{j−1} contribute to the generation of G^f_j. It is expressed as follows:

\[
G^f_j \sim DP(\alpha^f_0, w^f_{j−1}G^f_{j−1} + (1−w^f_{j−1})G^f_0)
\]

where α^f_0 is a concentration parameter for generating G^f_j, and w^f_{j−1} is the transfer weight.
- **Step 4:** The parameters of the component densities θ^f_{ijkl} is drawn from G^f_j, then the data sample x^f_{ijkl} is extracted from a distribution parameterized with θ^f_{ijkl}.

\[
θ^f_{ijkl} \sim G^f_j, \quad x^f_{ijkl} \sim F(\chi|θ^f_{ijkl})
\]

Note that i is the index of the data sample, and F(·|·) denotes a conditional distribution.

3. Sequential Hierarchical Dirichlet Processes

Video consists of a mass of sequential video frames. From the perspective of HDP, we model these frames within a video as multiple Dirichlet processes. However, these video frames different from the text documents are time dependent. Here, we introduce a three-layer Seq-HDP [7] model to show the time dependencies among neighboring frames within each video.

3.1 Generative Process

We depict a graphical model representation of Seq-HDP shown in Fig. 1. In this three-layer HDP model, a measure G located in the top of the model structure represents an overall measure of Seq-HDP, which is generated from a base measure H. Similarly, we define G^f_j as the global measure reflecting the f-th video on the second layer of Seq-HDP. Then on the third layer, G^f_j denotes the local measure corresponding to the j-th frame within the f-th video. In fact, these local measures are not independent, since video frames belonging to the same video are sequential and time related. Hence, we establish a transfer link based on a Markovian assumption between two adjacent local measures G^f_{j−1} and G^f_j in Seq-HDP with a transfer weight w^f_j. We describe the generative process of Seq-HDP as follows.

- **Step 1:** The overall measure G manipulating all the components within Seq-HDP is sampled from a Dirichlet process as G ∼ DP(ξ, H), where ξ is a concentration parameter for generating G.
- **Step 2:** The global measure G^f_0 for each video is drawn from G^f_0 ∼ DP(γ^f, G), where f is the index of the video, and γ^f is a concentration parameter for generating G^f_0.
- **Step 3:** Due to the time-dependency mechanism, both the upper global measure G^f_0 and previous local measure G^f_{j−1} contribute to the generation of G^f_j. It is expressed as follows:

\[
G^f_j \sim DP(\alpha^f_0, w^f_{j−1}G^f_{j−1} + (1−w^f_{j−1})G^f_0)
\]

where α^f_0 is a concentration parameter for generating G^f_j, and w^f_{j−1} is the transfer weight.
- **Step 4:** The parameters of the component densities θ^f_{ijkl} is drawn from G^f_j, then the data sample x^f_{ijkl} is extracted from a distribution parameterized with θ^f_{ijkl}.

\[
θ^f_{ijkl} \sim G^f_j, \quad x^f_{ijkl} \sim F(\chi|θ^f_{ijkl})
\]

Note that i is the index of the data sample, and F(·|·) denotes a conditional distribution.

3.2 Stick-Breaking Construction

Measures extracted from a Dirichlet process can be expressed by using independent sequences of i.i.d. random variables. This method is known as a stick-breaking construction (SBC) [3], which uses infinite latent components for formulating the model.

The SBC for Seq-HDP is shown in Fig. 2, where we respectively use component weights ν, β^f, and π^f to construct corresponding measures in Seq-HDP, the process is as below:

\[
G = \sum_{k=1}^{∞} ν_k δ_{θ_k}, \quad G^f_0 = \sum_{k=1}^{∞} β^f_k δ_{θ_k}, \quad G^f_j = \sum_{k=1}^{∞} π^f_k δ_{θ_k}
\]  

where each component weight variable consists of an infinite number of corresponding component weights, i.e., ν = {ν_1, ν_2, ..., ν_k, ...}, β^f = {β^f_1, β^f_2, ..., β^f_k, ...}, and π^f = {π^f_1, π^f_2, ..., π^f_k, ...}. δ_{θ_k} is a probability measure concentrated at component θ_k.
Then, these component weights can be successively drawn with:

\[
\nu \sim GEM(\xi), \beta^l \sim DP(\gamma^l, \nu), \pi^f_j \sim DP(\alpha_0^f, \pi^f_j) \tag{4}
\]

where \(GEM()\) represents a GEM process [3]. \(DP()\) stands for a Dirichlet process. Note that the generation of \(\pi^f_j\) is influenced by both \(\beta^l\) and \(\pi^f_{j-1}\) due to the time dependency. Hence, \(\pi^f_j\) is expressed as

\[
\pi^f_j = w^f_{j-1} \pi^f_{j-1} + (1 - w^f_{j-1}) \beta^f.
\]

Additionally, \(z_{fji}\) drawn from \(\pi^f_k\) is the index of the component, which generates the sample \(x_{fji}\) along with component \(\phi_k\).

4. Multimodal Data Modeling

In this section, we focus on how to handle multimodal data modeling for videos on the basis of Seq-HDP.

4.1 Conditionally Independent Model

According to the theory on conditionally independent LDA (CI-LDA) [8], it assumes that all the types of data (modalities) are conditionally independent for generating their own data samples, however, only their assigned topics are drawn from the same multinomial distribution over topics for each document. Based on this idea, we incorporate a conditionally independent mechanism into Seq-HDP, and propose a sequential conditionally independent HDP (Seq-CI-HDP) model.

A simplified graphical model representation for the SBC of Seq-CI-HDP is depicted in Fig. 3 (a), where we only show the generative process of multimodal data for convenience and omit the other variables that are the same as the ones in Seq-HDP. In other words, Seq-CI-HDP has the same generative process as Seq-HDP before generating the variable \(z\). We describe the rest of the generative process of Seq-CI-HDP as follows.

- **Step 1:** For the \(i\)th word of data type (modality) \(l\) in frame \(j\) of video \(f\), draw a topic \(z_{fji}^{(l)}\) from \(\pi^{f}_{j}\).

- **Step 2:** Conditioned on sampled topic \(z_{fji}^{(l)}\), a word \(x_{fji}^{(l)}\) is drawn from \(f(x_{fji}^{(l)}|\phi_k^{(l)}, k = z_{fji}^{(l)})\).

Note that data type \(l \in \{1, \ldots, L\}\), and \(N_{fj}^{(l)}\) are the total count for the words of data type \(l\) in frame \(j\) of video \(f\). \(i\) is the index of word, and \(f(x_{fji}^{(l)}|\phi_k^{(l)})\) denotes a word distribution conditioned on \(\phi_k^{(l)}\). In fact, all the \(\{\phi_k^{(l)}\}_{l=1}^{L}\) drawn from \(H\) share the same \(K\) topics. The sampling processes for all the topics are only related with \(\pi^{f}_{j}\), which implies that the dependencies among all the multimodal data are relatively weak.

4.2 Correspondence Model

Blei et al. [10] provides an alternative way that incorporates a correspondence mechanism into LDA to cope with annotated image data. As opposed to CI-LDA, the correspondence mechanism predefines a pivot data type, and then makes other non-pivot data types learn their latent topics based on the topic distribution over pivot data. Based on this idea, a Seq-cHDP [7] model was proposed for the same purpose.

We also show a simplified graphical model representation for the SBC of Seq-cHDP in Fig. 3 (b), and describe its generative process for multimodal data as below.

- **Step 1:** Assume that the data type \(l \in \{1\}\) holds the pivot flag all the time. Then a topic \(z_{fji}^{(1)}\) for the word \(x_{fji}^{(1)}\) is drawn from \(\pi^{f}_{j}\). Conditioned on sampled \(z_{fji}^{(1)}\), a specific word \(x_{fji}^{(1)}\) is drawn from \(f(x_{fji}^{(1)}|\phi_k^{(1)}, k = z_{fji}^{(1)})\).

- **Step 2:** For the other non-pivot data types \(l \in \{2, \ldots, L\}\), a topic \(z_{fji}^{(l)}\) for the word \(x_{fji}^{(l)}\) is drawn from \(Unif(z_{fji}^{(1)}, \ldots, z_{fji}^{(l)})\).

Conditioned on sampled \(z_{fji}^{(l)}\), a word \(x_{fji}^{(l)}\) is drawn from \(f(x_{fji}^{(l)}|\phi_k^{(l)}, k = z_{fji}^{(l)})\).

Here, \(Unif()\) denotes a uniform distribution, for the explanation of other parameters, we can refer to the similar parameters described in Seq-CI-HDP. For Seq-cHDP, the topic assignment for non-pivot data completely depends on the topic distribution over pivot data. However, such unidirectional dependency for multimodal data makes the model lack flexibility, for instance, the model may show poor performance when the data in some non-pivot data types become more dominant than the data in the pivot data type.

4.3 Symmetric Correspondence Model

In Sects. 4.1 and 4.2, we discussed Seq-CI-HDP and Seq-cHDP for modeling multimodal data. However, these two methods have some drawbacks: (1) Seq-CI-HDP handles the generative process for each data type independently conditioned on latent topics, but the dependencies among all the multimodal data are relatively weak; (2) Seq-cHDP can take
To improve these issues discussed above, we present a novel model called sequential symmetric correspondence HDP (Seq-Sym-cHDP), which can spontaneously select the pivot data type when generating a topic for a corresponding word. A similar idea also appears in our previous work [11], [12] proposed for multilingual text analysis. We show the SBC of Seq-Sym-cHDP with a simplified graphical model representation in Fig. 3(c). The generative process for multimodal data in Seq-Sym-cHDP is as below.

- **Step 1:** Draw a multinomial pivot flag generator \( \sigma_{fj} \) from \( \text{Dir}(\lambda) \).
- **Step 2:** For the \( i \)th word of data type \( l \) in frame \( j \) of video \( f \), a pivot flag \( s_{fj}^{(l)} \) is drawn from \( \text{Multinomial}(\sigma_{fj}) \).
- **Step 3:** If \( s_{fj}^{(l)} = l \), draw a topic \( z_{fj}^{(l)} \) from \( \pi_{fj}^{(l)} \). Otherwise if \( s_{fj}^{(l)} = h \neq l \), draw a topic \( z_{fj}^{(l)} \) from \( \text{Unif}(z_{fj}^{(h)}, \ldots, z_{fj}^{(h)}) \).
- **Step 4:** Conditioned on sampled \( z_{fj}^{(l)} \), a word \( x_{fj}^{(l)} \) is drawn from \( f(x_{fj}^{(l)}|\theta_{k}^{(l)}), k = z_{fj}^{(l)} \).

Here, \( \text{Dir}(\lambda) \) denotes a Dirichlet distribution with a hyperparameter \( \lambda \). The pivot flag is sampled for selecting the current pivot data type, which influences the way of generating the topic for the current word. For the word \( x_{fj}^{(l)} \) with data type \( l \), if its sampled pivot flag points to the same data type (i.e., \( s_{fj}^{(l)} = l \)), the process of its topic assignment will be independent from other topic assignments and only associated with \( \pi_{fj}^{(l)} \). However, if its sampled pivot flag points to the other data type \( h \) (i.e., \( s_{fj}^{(l)} = h \neq l \)), its topic will be drawn from a uniform distribution that consists of all the topics assigned to \( H_{fj}^{(h)} \) words with data type \( h \). Different from \( N_{fj}^{(h)} \), \( H_{fj}^{(h)} \) denotes the number of words that have been already assigned with topics at the current step. Hence, there is \( H_{fj}^{(h)} \leq N_{fj}^{(h)} \).

For the other parameters, we can refer to the description in Seq-CI-HDP.

In this way, the Seq-Sym-cHDP can manage the pivot flag assignment well and balance the dependencies among all the multimodal data. Such symmetric correspondence mechanism can enhance the model with more flexibilities when encountering more complex scenario of multimodal data. For example in video data modeling, when the number of the speech words (or visual words) drastically decreases and the number of the visual words (or speech words) remains almost unchanged in some frames of the video, the speech words (or visual words) may fail to express the entire semantical information for the current frames of this video, since the number of a few speech words (or visual words) is insufficient. In this scenario, the symmetric correspondence method incorporated in Seq-Sym-cHDP can still make the model work well via balancing the dependencies between the speech words and the visual words with pivot assignments. More specifically, Seq-Sym-cHDP can let visual words (or speech words) enriched with entire information be informatively complementary to these insufficient speech words (or visual words) via assigning pivot flags to some of the visual words (or speech words).

5. **Inference Scheme**

5.1 **Posterior Representation**

In the following, we will describe the sampling method for each variable mentioned in Seq-Sym-cHDP based on a posterior representation sampler [9].

- **Sampling Component Weights**

Based on the theory of the posterior representation sampler, the overall measure \( G \) is formulated by:

\[
G = \sum_{k=1}^{K} v_{k} \delta_{b_{k}} + v_{u} G_{u}, \quad G_{u} \sim DP(\xi, H)
\] (5)
and the component weight $\nu$ is sampled by:

$$
\nu = (\nu_1, \ldots, \nu_K, \nu_a) \sim \text{Dir}(M_1, \ldots, M_K, \xi)
$$

where the original infinite vector $\nu$ is represented with a new augmentable finite vector that consists of $K$ components and the promising component $u$. Such reformulation also occurs in $\beta_f$ and $\pi_f$. Here, a new variable $M'_k$ used for sampling $\nu$ denotes the metatable count in the Chinese restaurant franchise (CRF) [3] of three-layer HDP, which is explained in [7]. $M'_k$ is the marginal form of $M'_k$ (i.e., $M'_k = \sum_f M'_k$).

Similarly, the global measure $G_0$ is formulated by:

$$
G_0 = \sum_{k=1}^{K} \beta_k' \delta_{\beta_k} + \beta_a' G_a, \quad G_a \sim DP(\xi, H)
$$

and the component weight $\beta$ is sampled by:

$$
\beta = (\beta_1', \ldots, \beta_K', \beta_u') \sim \text{Dir}(\pi_1', \ldots, \pi_K', \pi_u')
$$

where $\beta_k' = \gamma' v_k + T_{j_k}^{0-i-j}$ and $\beta_u' = \gamma' v_a$. The parameter $T_{j_k}^{0-i-j}$ is the table count associated with $\beta_k'$ in CRF, and $T_{j_k}^{0-i-j}$ is the marginal form of $T_{j_k}^{0-i-j}$ (i.e., $T_{j_k}^{0-i-j} = \sum_f T_{j_k}^{0-i-j}$).

Then, the local measure $G_{j_k}'$ is formulated by:

$$
G_{j_k}' = \sum_{k=1}^{K} \pi_k' \delta_{\beta_k} + \pi_u' G_u, \quad G_u \sim DP(\xi, H)
$$

and the component weight $\pi_f$ is sampled by:

$$
\pi_f = (\pi_{j_1}', \ldots, \pi_{j_k}', \pi_u') \sim \text{Dir}(\pi_{j_1}', \ldots, \pi_{j_k}', \pi_u')
$$

note that the $\pi_f$ is affected by not only the $\beta_f$ but also the $\pi_{f-1-j}$. The two estimators $\hat{\pi}_{j_k}'$ and $\tilde{\pi}_{j_k}'$ are therefore derived as follows:

$$
\hat{\pi}_{j_k}' = \alpha_1' \frac{f_{j_k}'}{T_{j_k}^{0-i-j}} + \alpha_0' (1 - w_{j_k}') \delta_{\beta_k} + C_{j_k} \cdot T_{j_k}^{0-i-j}
$$

$$
\tilde{\pi}_{j_k}' = \alpha_1' \frac{f_{j_k}'}{T_{j_k}^{0-i-j}} + \alpha_0' (1 - w_{j_k}') \delta_{\beta_k} + C_{j_k} \cdot T_{j_k}^{0-i-j}
$$

where $w_{j_k}'$ is a controlling parameter ranging from 0 to 1, and $C_{j_k}$ is the count for the word with data type $l$ in frame $j$ of video $f$ possessing the pivot flag with the same data type $l$ conditioned on topic $k$, and $C_{j_k} = \sum_l C_{j_k}^{(l)}$. $T_{j_k}^{0-i-j}$ is another table count that occurs due to the time-dependency mechanism (from $\pi_{j_k}'$ to $\pi_{j_k}'^{(l)}$) in CRF.

**Sampling Table Counts of CRF**

For the CRF of a three-layer HDP [7], the metatable count $M'_k$ quantizes the generative process from $\nu_k$ to $\beta_k'$. We use $T_{j_k}'$ to represent the total table count that can be formulated with $T_{j_k}' = T_{j_k}^{0-i-j} + T_{j_k}^{j-i-j}$, where $T_{j_k}^{j-i-j}$ quantizes the generative process from $\pi_{j_k}'$ to $\beta_{j_k}'$ and $T_{j_k}^{0-i-j}$ quantizes the generative process from $\beta_{j_k}'$ to $\nu_k$. Here, we use a Bernoulli trial method to sample these counts, and its probability on getting a value of 1 is defined with $P(n|a, b) = \frac{ab}{ab+nb-1}$. Then, $T_{j_k}'$ can be sampled with:

$$
T_{j_k}' = \sum_{a=1}^{m_n} m_n, \quad m_n \sim \text{Bern}(P(n|a, b))
$$

where $\text{Bern}(\cdot)$ denotes a Bernoulli distribution, $a = a_0$, and $b = w_{j_k} - \pi_{j_k}^{(0-i-j)} + (1 - w_{j_k}) \pi_{j_k}^{(0-i-j)}$. Then, $T_{j_k}^{j-i-j}$ and $\pi_{j_k}^{0-i-j}$ can be sampled based on a binomial distribution with $T_{j_k}'$

$$
(T_{j_k}^{j-i-j}, T_{j_k}^{0-i-j}) \sim \text{Binomial}(T_{j_k}', [P_T, 1 - P_T])
$$

Similarly, we can sample $M'_k$ with:

$$
M'_k = \sum_{a=1}^{m_n} m_n, \quad m_n \sim \text{Bern}(P(n|a, b))
$$

where $T_{j_k}'$ can be obtained with $T_{j_k}' = \sum_j T_{j_k}'$, $a = \gamma_f$, and $b = \nu_k$.

**Sampling Topics and Pivot Flags for Words**

For Seq-Sym-cHDP, the topic sampling is associated with the pivot selection. We thus derive a full conditional joint likelihood for estimating both topic $z_{f,j}$ and pivot flag $\phi_{f,j}$ for each word:

$$
P(c_{f,j} = k, s_{f,j} = l | h_{f,j}, \lambda, \ldots) \propto P(x_{f,j} = k | \lambda) P(z_{f,j} = k | \phi_{f,j}) P(\phi_{f,j} = l | \lambda)$$

$$
= \sum_{j_k} C_{f,j_k} \cdot \frac{f_{j_k}'}{C_{f,j_k}} \cdot \frac{\pi_{j_k}'}{C_{f,j_k}^{(l)}} \cdot \frac{T_{j_k}^{0-i-j}}{C_{f,j_k}^{(l)}} \cdot \frac{T_{j_k}^{j-i-j}}{C_{f,j_k}^{(l)}} \cdot \frac{1}{C_{f,j_k}^{(l)}}
$$

$$
P(c_{f,j} = k, s_{f,j} = l | h_{f,j}, \lambda, \ldots) \propto P(x_{f,j} = k | \lambda) P(z_{f,j} = k | \phi_{f,j}) P(\phi_{f,j} = l | \lambda)$$

$$
= \sum_{j_k} \frac{C_{f,j_k}^{(l)}}{C_{f,j_k}^{(l)}} \cdot \frac{f_{j_k}'}{C_{f,j_k}^{(l)}} \cdot \frac{\pi_{j_k}'}{C_{f,j_k}^{(l)}} \cdot \frac{T_{j_k}^{0-i-j}}{C_{f,j_k}^{(l)}} \cdot \frac{T_{j_k}^{j-i-j}}{C_{f,j_k}^{(l)}} \cdot \frac{1}{C_{f,j_k}^{(l)}}
$$

$\text{C}_{f,j}$ indicates the count for the pivot flag pointing to data type $l$ over all the multimodal data in frame $j$ of video $f$, the superscript $-fjl$ for a variable denotes the terms excluding the $fjl$th term. $Z_{f,j}$ is a vector that records all the topic assignments for the words of data type $h$ in frame $j$ of video $f$, $n_{f,j}$ indicates the count for topic $k$ assigned to words of data type $h$ in frame $j$ of video $f$, and $\text{N}_{f,j}$ indicates the total count for the words in the same domain. Note that there are two kinds of conditional word likelihood functions ($f_{x_{f,j}}(x_{f,j})$ and $f_{s_{f,j}}(s_{f,j})$) conditioned on topic $k$. If the topic $k$ is previously used, $f_{x_{f,j}}(x_{f,j}, k)$ is expressed by:

$$
f_{x_{f,j}}(x_{f,j}, k) = \frac{\text{N}_{f,j}}{\sum_{j_k} \text{N}_{f,j_k}}
$$

If the topic $k$ is new, $f_{s_{f,j}}(s_{f,j})$ is expressed by:

$$
f_{s_{f,j}}(s_{f,j}) = \frac{1}{\text{N}_{f,j}}
$$
Algorithm 1: Inference Scheme with CGS

Initialization;
for \((f = 1; f \leq F; f++)\) do
  for \((j = J^f; j \geq 1; j--)\) do
    for \((l = 1; l \leq L; l++)\) do
      Sample pivot flags \(s_{lj}^f\) and topics \(z_{lj}^f\);
    end
    for \((k = 1; k \leq K; k++)\) do
      Sample \(t_{j;k}, t_{j;f}^{(f'-1)}\) and \(T_{j;k}^{f_0}\);
    end
  end
end
Sample \(\xi, \gamma^f\) and \(\alpha_0^f\);
Sample \(\nu^f\);
for \((f = 1; f \leq F; f++)\) do
  Sample \(\beta^f\);
  for \((j = 1; j \leq J^f; j++)\) do
    Sample \(\pi_j^f\);
  end
end

index of words in terms of the vocabulary in data type \(l,\) \(n_{k(\nu)}\) is the count for the word \(v_{k(\nu)}^f\) assigned to topic \(k.\) \(V(\nu)\) is the vocabulary size of data type \(l.\) \(\tau\) is a hyperparameter for a Dirichlet distribution, and \(H = \text{Dir}(\tau).\)

- **Sampling Hyperparameters**

  The hyperparameters \(\xi, \gamma^f\) and \(\alpha_0^f\) corresponding to \(\nu, \beta^f\) and \(\pi_j^f\) are sampled from a gamma prior. In addition, the hyperparameter \(\lambda\) is considered to be another controlling parameter that will be determined in the parameter tuning stage, and the hyperparameter \(\tau\) is a constant parameter in this paper.

5.2 Cascade Gibbs Sampling Implementation

Based on the posterior representation sampler described in Sect. 5.1, we devise a cascaded Gibbs sampling (CGS) method to implement the inference scheme for Seq-Sym-cHDP. Its pseudo-code is shown in Algorithm 1. In this algorithm, some hyperparameters and controlling parameters will be preset in the initialization phase. During the sampling phase, we need to notice that the total number of topics \(K\) is not always constant, and it might increase if the promising component is triggered to release a newborn topic when sampling \(Z_{lj}^f\). To handle this case, all the component weights are required to update their components to instantiate this newborn topic in the sampler [9].

6. Experiments

6.1 Experimental Setup

In this experiment, we utilize the video datasets provided by MediaEval-2011 Genre Tagging Task [13] (MEGTT), where two types of data were extracted from videos. Firstly, a sequence of key frames deemed as image data were drawn for each video. Secondly, a number of speech transcript words deemed as speech data were allocated to each key frame by an automatic speech recognition (ASR) method [13].

Then, we need to convert the image data and speech data into visual words and speech words in accordance with Seq-Sym-cHDP. SIFT descriptor [14] is used to draw visual words from each key frame, where the SIFT descriptor for every \(10 \times 10\) pixel grid is computed conditioning in which the patch size is randomly sampled between scales of 10 to 30 pixels. Then, all the extracted SIFT descriptors are clustered into \(K\) types of visual words by a \(k\)-means algorithm. In our experiment, we set \(k\) to 1000. For the speech transcript, we remove standard stop words and some rare words (which appear in less than five videos), and manage the remaining 17596 types of general words as speech words for experiment. We summarize the details of the experimental dataset in Table 1.

| Number of Video Files | 1726 |
|-----------------------|------|
| Number of Genre Labels| 26   |
| Number of Key Frames  | 40588|
| Number of Visual Words| 166803k|
| Number of Speech Words| 1313k|

1 Only test data of MEGTT is used in this experiments.

We evaluate the model by using genre classification with a five-fold nested cross-validation scheme, where all the video datasets are evenly split into five subsets. For each round in the cross validation, one subset is treated as the test set while the other four subsets are used for four-fold cross validation on parameter tuning in advance.

6.2 Parameter Tuning

In the Seq-Sym-cHDP, we predefine \(w^f\) and \(\lambda\) as controlling parameters, which need to be tuned to optimize the model. For simplicity, all the transfer weights \(w^f\) are assumed to be the same parameter \(w.\) We specify that all the hyper-parameters \(\xi, \gamma^f\), and \(\alpha_0^f\) are drawn from a gamma prior \(\text{Gamma}(1.0, 1.0),\) and set \(\tau\) to 1.0. The CGS system will run 1000 iterations in total so as to let all the variables in the model fully converge.

In experiments, we respectively sweep the controlling parameter \(w\) and \(\lambda\) in \([0.1, 0.3, 0.5, 0.7, 0.9]\) and \([0.1, 0.5, 1.0, 1.5, 2.0]\) and conduct this parameter tuning task by utilizing the validation set from training set in five-fold nested cross validation. We can gather \(25 \times 4 \times 5\) sets of experimental results in total. Then, the average parameter tuning results are shown in Fig. 4, where we can know that

\(^{1}\)An extra classifier such as SVM is required for estimating a genre label for each video based on learned topic distributions.
Seq-Sym-cHDP performs the best when we select $w = 0.5$ and $\lambda = 1.0$.

6.3 Qualitative Evaluation

Here, we focus on qualitative evaluation for Seq-Sym-cHDP. In experiments, we proceed to use the same parameter settings in Sect. 6.2, i.e., $w = 0.5$ and $\lambda = 1.0$.

- **Dynamic Pivot Assignments**

The major advantage of Seq-Sym-cHDP is that it can balance the dependencies well among all the multimodal data via managing the pivot assignments. We illustrate the performance of dynamic pivot assignments with three different examples occurring in three corresponding videos, as shown in Fig. 5. Here, we need to introduce a new parameter $\sigma'$ that indicates the proportion of pivot assignments, and define $\sigma'_{fjl} = \frac{C_{fjl}}{\sum_l C_{fjl}}$. In Fig. 5, both the variation curves of $\sigma'$ on visual words and speech words are plotted against key frames with solid lines, for comparison, the real ratios of the number of visual words and speech words among all the multimodal words are also plotted with dotted lines.

We observe from the examples that the proportion of pivot assignments for each data type is dynamically changing along with the key frame evolution, and the variation trend on $\sigma'$ follows the variation trend on the ratio of the words in the current data type, which means the model will increase (or decrease) the pivot assignments to the current data type when the ratio of the words in the current data type ascends (or descends). These results demonstrate that the symmetric correspondence mechanism can make Seq-Sym-cHDP more effective and flexible on balancing the dependencies among all the multimodal data.

- **Trend Estimation for Latent Topics**

Estimating the trend of latent topics based on learned topic distributions is another key feature for time dependent topic models. Especially for Seq-Sym-cHDP, its particular symmetric correspondence mechanism ensure that the model can learn and track the trend of latent topics with more stability.

Here, we still use the same three example video clips (which are shown in Fig. 5.), and draw their area graphs of topic distributions to illustrate the performance of topic trend estimation with Seq-Sym-cHDP, as shown in Fig. 6, where each colored stripe superimposed with words stands for each topic with assigned speech words, note that we only show some high-frequency speech words on this graph due to space limitation, the font size of the word reflects its appearance frequency (bigger font size means higher frequency) in a topic, and the width of stripes on the vertical axis depends on the topic distributions over visual words.

In Fig. 6, the topic trend within a video is predicted well with sequential topic distributions due to the time-dependency mechanism in Seq-Sym-cHDP. For instance in video No.511, the real video content for this clip is a talk show focusing on politics, therefore the model predicted that the major topic with key speech words such as military, government, political ran through all the key frames. However, when encountered more extreme scenarios such as extremely imbalanced multimodal data, Seq-Sym-cHDP still can handle such tricky task well. In video No.161, the visual words are very dominant in Frame #37 and #38, since the number of speech words is very few. Similarly in video No.984, the same phenomenon occurs in Frame #2 and #5. In these scenarios, the speech words fail to express the entire semantical information for the current key frame. Consequently, this might result in some incorrectness or a drastic fluctuation on speech or visual word topic trend estimation if using general time dependent topic mod-
els such as Seq-CI-HDP and Seq-cHDP. Figure 7 shows the results of topic trend estimation in video No.161 with Seq-CI-HDP and Seq-cHDP, respectively. In Fig. 7 (a) with Seq-CI-HDP, although the topic trend estimation on visual words looks smooth and steady, when the number of speech words drastically decreases in Frame #37 and #38, the quality of the topic estimation for these speech words becomes worse, and some speech words are assigned to some wrong topics. This is because such weak dependencies on multimodal modeling cannot fix such incorrectness on topic estimation caused by insufficient speech words, since all the modalities are conditionally independent for assigning topics to their own words in Seq-CI-HDP. In Fig. 7 (b) with Seq-cHDP, we can find the topic trend estimation on visual words shows a drastic fluctuation when the number of the pivot speech words drastically decreases in Frame #37 and #38, the effectiveness of topic trend estimation becomes unstable. This is because such unidirectional dependency on multimodal modeling makes the topic learning for non-pivot visual words follow the learned topic distribution over pivot speech words, and the topic trend estimation for visual words is misled by a few pivot speech words in this scenario. However, according to Fig. 6 (a), Seq-Sym-cHDP effectively fixes these problems caused by such imbalanced multimodal data, and improves the quality of topic trend estimation via balancing the pivot assignments between speech words and visual words with the symmetric correspondence mechanism, as we discussed in Sect. 4.3.

### 6.4 Quantitative Evaluation

We finally evaluate the performance of the models with genre classification. Three topic models Seq-CI-HDP, Seq-cHDP and symmetric correspondence HDP (Sym-cHDP) are adopted as baseline models. Note that Sym-cHDP is a nonsequential version of Seq-Sym-cHDP. For Seq-cHDP, we conduct two experiments by predefining the pivot to visual data and speech data respectively. Before commencing the experiments, we also performed the parameter tuning tasks for other baseline models, the results showed that all the baseline models worked most effectively when setting the controlling parameters with $w = 0.5$ and $\lambda = 1.0$ as well.

Table 2 shows the results of video genre classification for all the baseline models and Seq-Sym-cHDP, where Seq-
Sym-cHDP outperforms the other baselines with both the results of Micro-F1 and Macro-F1. This demonstrates that the symmetric correspondence mechanism for multimodal data can make the model work more effectively in the case of video categorization. In addition, Sym-cHDP shows poor performance with more instability, which means the time-dependency mechanism in Seq-Sym-cHDP plays a very important role in video data modeling as well.

7. Conclusions

This paper focused on a discussion about using different multimodal data modeling methods on the basis of sequential Bayesian nonparametric topic modeling framework to analyze video data. We firstly introduced two different topic models Seq-CI-HDP and Seq-cHDP to handle the video multimodal data in two different ways. However, these methods have drawbacks on the flexibility of modeling multimodal data. To tackle these problems, we proposed a Seq-Sym-cHDP model, whose main contribution is that it makes the process of multimodal data modeling more flexible and effective via controlling the pivot assignments in a symmetric correspondence manner. We also derived an inference method for Seq-Sym-cHDP based on posterior representation sampler. We evaluated these models by experiments, and finally demonstrated that Seq-Sym-cHDP achieved the best performance compared with other baselines. Future work will focus on applying some more flexible time-dependency mechanisms such as hidden Markov models (HMM) [3] to our model.

Acknowledgments

This work was supported in part by the Grant-in-Aid for Scientific Research (#15H02703) from JSPS, Japan.

References

[1] J. Xue and K. Eguchi, “Sequential symmetric correspondence hierarchical Dirichlet processes for video data analysis,” NIPS Workshop on Practical Bayesian Nonparametrics, pp.229–233, 2016.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet allocation,” the Journal of machine Learning research, vol.3, pp.993–1022, 2003.

[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] X. Wang, X. Ma, and E. Grimson, “Unsupervised activity perception by hierarchical Bayesian models,” IEEE Conference on Computer Vision and Pattern Recognition, pp.1–8, IEEE, 2007.

[5] T. Hospedales, S. Gong, and T. Xiang, “A Markov clustering topic model for mining behaviour in video,” IEEE 12th International Conference on Computer Vision, 2009, pp.1165–1172, IEEE, 2009.

[6] D. Kuettel, M.D. Breitenstein, L. Van Gool, and V. Ferrari, “What’s going on? discovering spatio-temporal dependencies in dynamic scenes,” IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp.1951–1958, IEEE, 2010.

[7] J. Xue and K. Eguchi, “Sequential correspondence hierarchical Dirichlet processes for video data analysis,” Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp.229–233, ACM, 2016.

[8] E. Erosheva, S. Fienberg, and J. Laﬀerty, “Mixed-membership models of scientific publications,” Proceedings of the National Academy of Sciences, vol.101, no.suppl 1, pp.5220–5227, 2004.

[9] Y.W. Teh and M.I. Jordan, “Hierarchical Bayesian nonparametric models with applications,” Bayesian nonparametrics, vol.1, pp.158–207, 2010.

[10] D.M. Blei and M.I. Jordan, “Modeling annotated data,” Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp.127–134, ACM, 2003.

[11] K. Fukumasu, K. Eguchi, and E.P. Xing, “Symmetric correspondence topic models for multilingual text analysis,” Advances in Neural Information Processing Systems, pp.1286–1294, 2012.

[12] K. Eguchi, K. Fukumasu, T. Omoto, and E.P. Xing, “Symmetric correspondence topic models for multimodal data analysis,” ArXiv, 2018.

[13] M. Larson, M. Eskevitch, R. Ordelman, C. Kofler, S. Schmiedeke, and G.J. Jones, “Overview of mediaeval 2011 rich speech task and genre tagging task,” CEUR Workshop Proceedings, 2011.

[14] D.G. Lowe, “Distinctive image features from scale-invariant keypoints,” International journal of computer vision, vol.60, no.2, pp.91–110, 2004.

Jianfei Xue is currently pursuing a Ph.D degree at the Graduate School of System Informatics, Kobe University, Japan.

Koji Eguchi is an Associate Professor at the Graduate School of System Informatics, Kobe University, Japan. His research interests include information retrieval, statistical machine learning, and data mining.