Exploring Bayesian Models for Multi-level Clustering of Hierarchically Grouped Sequential Data

Adway Mitra
Indian Institute of Science
Bangalore, India
adway@csa.iisc.ernet.in

Abstract

A wide range of Bayesian models have been proposed for data that is divided hierarchically into groups. These models aim to cluster the data at different levels of grouping, by assigning a mixture component to each datapoint, and a mixture distribution to each group. Multi-level clustering is facilitated by the sharing of these components and distributions by the groups. In this paper, we introduce the concept of Degree of Sharing (DoS) for the mixture components and distributions, with an aim to analyze and classify various existing models. Next we introduce a generalized hierarchical Bayesian model, of which the existing models can be shown to be special cases. Unlike most of these models, our model takes into account the sequential nature of the data, and various other temporal structures at different levels while assigning mixture components and distributions. We show one specialization of this model aimed at hierarchical segmentation of news transcripts, and present a Gibbs Sampling based inference algorithm for it. We also show experimentally that the proposed model outperforms existing models for the same task.

1 Introduction

In many applications we come across hierarchically grouped data. For example in a text corpus, data is grouped into documents, paragraphs and sentences. Such data can be clustered at multiple levels, based on the notion of topics. A large number of hierarchical Bayesian models have been proposed for such data, many of whom are quite similar to each other in various aspects. However, to the best of our knowledge, there has not been much research aimed at placing these models in perspective, and making a comparative study of them, except empirical comparisons. This is what we attempt in this paper. The main aspect of these models which we compare is how they share the mixture components and distributions across the groups at different levels.

The contributions of this paper are as follows: 1) We introduce a novel classification of Hierarchical Bayesian models for grouped data, based on Degree of Sharing of mixture components and distributions 2) We introduce a generalized

\[ D^1_1 = \forall i, D^1_2 = 1, D^3_2 = 2, D^2_3 = 3, D^6_4 = 4 \text{ etc, } D^1_i = 1 \text{ for } i = 1 \ldots 4, D^1_i = 2 \text{ for } i = 5 \ldots 8 \text{. Also, } D^3(1) = 1, D^3(3) = 2 \text{ etc \}

Hierarchical Bayesian model and show many existing ones to be special cases of it, and 3) We show how it can be adapted for news transcript segmentation, for which we give an inference algorithm and demonstrate experimental results.

2 Notations

Consider \( N \) datapoints \( Y_1, Y_2, \ldots, Y_N \), of any type (eg. integers, real-valued vectors) based on the application. Each of these are associated with group membership variables (positive integers), which specify the grouping of the datapoints. If there are \( L \) levels of grouping, each datapoint \( Y_i \) is associated with observed variables \( \{D^1_i, D^2_i, \ldots, D^L_i\} \). For example, a text corpus consists of a set of documents, each of which consists of word-tokens. We can consider the word-tokens as data-points \( \{Y_i\} \), which are tagged with their document memberships using \( \{D^2\} \), where \( \{D^1\} \) are the token indices, to capture the sequential ordering. This is the standard setting used in most topic models for text documents. In addition, it is possible to consider a 3-level grouping with sentences within documents. Then each word-token \( Y_i \) is associated with a sentence membership variable \( D^3_i \) and a document membership variable \( D^1_i \). In this paper, we will overload \( D^l \) \((l > 1)\) to indicate the higher-level group-memberships of lower-level-groups. For example if \( g \) is the index of a level-2 group, then \( D^3(g) \) is the level-3 group that covers all the datapoints under group \( g \), i.e. \( D^3(g) = D^3_1 \) where \( D^3_1 = g \). Please see Fig.1 for illustration.

Most topic models consider documents or sentences to be bags of words, and do not consider the sequential nature of
Figure 2: Grouped data clustered at 2 levels (l = 1, 2). Colours indicate the clustering, like \( Z^1 \), \( Z^2(2) \), etc. Different colours used at the two levels. Note that \( Z^1 \neq Z^2 \), but \( Z^2_3 = Z^2_6 \) and the datapoints are not grouped at all. There are \( K \) mixture components \( \{ \phi \} \) which are Gaussian distributions, i.e. \( \phi_k = \mathcal{N}(\mu_k, \Sigma_k) \). In general, the mixture components need not be Gaussian. The mixture distribution \( \theta^1 \) is a \( K \)-dimensional multinomial. Each datapoint is assigned to a mixture component \( Z^1_i \), which defines a clustering of the datapoints. This assignment is IID as \( Z^1_i \sim \theta^1 \) and sequential structure of the datapoints is not considered.

In GMM, the number of mixture components \( K \) is fixed and known. A non-parametric model with \( L = 1 \) is the Dirichlet Process Mixture Model (DP-MM), which considers infinitely many mixture components, though only a few of them are used for a finite number of datapoints. The mixture distribution \( \theta^1 \) is an infinite-dimensional multinomial, drawn from a stick-breaking distribution. The parameters of the mixture components are drawn from base distribution \( H \).

A one-level nonparametric model which does consider the sequential structure of the data is the HDP-HMM [Fox et al., 2008]. This model considers a set of \( \theta^1 \)-distributions from which one may be chosen conditioned on the previous assignments of \( Z^1 \). The \( Z^1 \)-assignment to each datapoint \( i \) is done as \( Z^1_i \sim \theta^1 \) with \( j = Z^1_{prev(i)} \), where \( prev(i) \) is the predecessor of the current datapoint in the sequential order encoded by \( \{ D^1 \} \), i.e. \( prev(i) = i' \) where \( D^1_{i} = D^1_{i'} + 1 \).

### 3.2 2-level models

Next, we move into two-level models, i.e. where \( L = 2 \). This is the standard setting for document modelling, where the word-tokens are grouped into documents (one level of grouping). The document membership of the variables are encoded by \( D^2 \). The most standard model of this kind is the Latent Dirichlet Allocation (LDA) [Blei et al., 2003] which considers \( K \) mixture components (topics) \( \{ \phi \} \), where \( K \) is fixed and known. Each mixture component \( \phi_k \) is a multinomial distribution over the vocabulary of size \( V \). Here the level-2 groups (documents) are not clustered, i.e. \( Z^2 \) is distinct for each document. Consequently, \( \theta^2 \) is not used here, and \( \{ \theta^1 \} \) are group-specific. The \( Z^1 \)-variables of the datapoints under any group \( j \) are assigned as IID draws from \( \theta^1_j \). Once again, no sequential structure is considered. Note that the mixture components \( \phi \) are shared by all groups.

\[
\phi_k \sim \text{Dir}(\beta), k \in [1, K]; \theta^1_j \sim \text{Dir}(\alpha), j \in [1, M] \\
Z^1_i \sim \theta^1_{D^2_j}, Y_i \sim \phi_{Z^1_i}
\]  

A non-parametric generalization of LDA is the Hierarchical Dirichlet Process (HDP) [Teh et al., 2006], which is also a 2-level extension of the DP-MM discussed above. Here, the number of components is not fixed or known, so the document-specific \( \{ \theta^1 \} \)-distributions are infinite-dimensional, and drawn from a Dirichlet Process/Stick-Breaking Process instead of finite-dimensional Dirichlet.

Another nonparametric 2-level model is the Nested Dirichlet Process (NDP), where the level-2 groups (documents) are clustered using \( Z^2 \), which are drawn according to a discrete distribution \( \theta^2 \). Each cluster induced by \( Z^2 \) uses its own \( \theta^1 \). However, unlike the previous models, here the mixture components themselves are specific to the clusters induced by \( Z^2 \).
4.1 DoS Concept

As already discussed, in all the hierarchical Bayesian models, the mixture components $\{\phi\}$ and the mixture components $\{\theta^1, \ldots, \theta^L\}$ are shared among the different groups. We have seen three types of sharing:

1. **Full sharing (F):** where components/distributions are shared by all the groups. For example, in HDP, MLC-HDP etc the mixture components are shared by all the level-2 groups.

2. **Group-specific sharing (G):** where components/distributions are specific to groups, and not accessible outside the groups. For example, in HDP, STM etc the distributions $\theta^1$ are specific to the top-level groups (documents).

3. **Cluster-specific sharing (C):** where the components/distributions are specific to clusters of groups, but not accessible outside the clusters. For example, in MLC-HDP each $\theta^1$-distribution is accessible to only one cluster of level-2 groups, and each $\theta^2$-distribution is accessible to only one cluster of level-3 groups. In all the models, the mixture components are specific to clusters of level-1 groups (as datapoints are clustered by the assignment of a mixture component through $Z^1$).

Based on these notions we introduce Degree-of-sharing (DoS). For any given model, we first specify how the mixture components are shared at each of the levels- Full (F), group-specific (G) or cluster-specific (C), and we call this the DoS of $\{\phi\}$. The type of sharing at the different levels are hyphen-separated. Next, regarding the distributions $\{\theta^l\}$ at each level $l$, we specify how it is shared by the levels $(l+1)$ upwards, and we call this the DoS of $\{\theta^l\}$. Also, to indicate if sequential structure is considered at the different levels, we add $S$ to the levels where it is considered. Finally, to indicate how groups are clustered at different levels, we add $N$ to the levels where there is no clustering of groups, $P$ to the levels where the number of clusters is fixed, and $NP$ to the levels where the clustering is non-parametric. Note that this indicates the dimensionality of $\{\theta^l\}$. $P$ indicates that it is finite-dimensional, $NP$ indicates it is infinite-dimensional, and $N$ indicates it is not in use.

By combining the DoS of $\{\phi\}, \{\theta^1\}, \ldots, \theta^L$ in that order, we have the DoS-classification of the model. The DoS of the different variables are semicolon-separated. The number of components in any of these models is $(L + 1)$, so that the DoS-classification of any model will have $(L + 1)$ semicolon-separated parts. Also, the first part (corresponding to $\phi$) will consist of $L$ hyphen-separated letters, and the number of these letters will keep decreasing by one for each of the following parts (corresponding to $\theta^1, \theta^2$ etc), but followed by the letters specifying dimensionality and sequence structure.

4.2 Classification of Models

Let us illustrate the concept of DoS-classification with a case study of all the models discussed in Section 3.

**Level-1** parametric models like GMM have mixture-components $\{\phi\}$ specific to clusters of datapoints, so that its DoS is $C$. But the mixture distribution $\theta^1$ is fully shared by...
all the datapoints, so that its DoS is $F$. The number of clusters formed at level 1 (i.e. the dimensionality of $\theta^1$) is fixed (i.e. $P$) and sequential structure is not considered. Hence the DoS-classification of GMM is $C; F \rightarrow P$. In case of DP-MM, $\theta$ is infinite-dimensional (NP), i.e. DoS-classification is $C; F \rightarrow N P$.

In case of HDP-HMM, the mixture-components $\{\phi\}$ are again specific to clusters of datapoints, so that its DoS is $C$. Here, the $\{\theta^1\}$ are non-parametric (NP), and the sequential structure is also considered. So, the DoS-classification of HDP-HMM is $C; F \rightarrow N P \rightarrow S$. Note that here $\{\theta^1\}$ is a collection of distributions, from which one is chosen for each data-point $i$, depending on the assignment to $prev(i)$.

**Level-2 models:** In HDP or LDA, the mixture-components are shared cluster-specific in level-1 and fully at level-2, so the DoS for $\phi$ is $C - F$. The $\{\theta^1\}$ are specific to level-2 groups, so the DoS for $\theta^1$ is $G$. Sequential structure is not considered at any level. In case of LDA the number of clusters of datapoints (level-1) is fixed ($P$), and for HDP it is $NP$. The level-2 groups are not clustered ($N$) in either model. So we say the DoS-classification of LDA is $C = F - G - P; N$, and for HDP it is $C = F - G - N P; N$. In case of NDP, the $\{\phi\}$ are cluster-specific at both levels, so its DoS is $C - C - G$. $\theta^1$ is specific to clusters of level-2 groups ($C$), and it is non-parametric (NP). The $\theta^2$ are also non-parametric (NP). So, the DoS-classification is $C - C; C - N P; N P$.

**Level-3 models:** In MLC-HDP, the mixture-components $\{\phi\}$ are cluster-specific in level-1, but fully at both levels 2 and 3, so that its DoS is $C = F - P$. The $\{\theta^1\}$ are specific to clusters of level-2 groups but fully shared by level-3 groups, and they are nonparametric, so the notation is $C = F - N P$. The $\{\theta^2\}$ are specific to clusters of level-3 groups and nonparametric, and finally $\theta^3$ is nonparametric. So the DoS-classification of MLC-HDP is $C = F - P; C = F - N P; C - N P; N P$.

For Topic-segmentation model, the topics $\{\phi\}$ are shared by all sentences and documents, i.e. its DoS is $C = F - F$. The $\{\theta^1\}$ are specific to clusters of sentences inside individual documents, i.e. the DoS is $C = G$, and they are of fixed dimension ($P$). The $\{\theta^2\}$ used to cluster sentences is document-specific ($G$). The number of clusters (segments) of sentences to be formed is not fixed, and sequential structure is also taken into account, so the notation is $G = N P - S$. Finally, the documents themselves are not clustered ($N$), and the DoS-classification of TSM is $C = F - F; C = G - P; G - N P - S; N$.

Finally we come to Subtle Topic Model (STM), where the topics are shared by all sentences and documents, i.e. the DoS is $C = F - F$. The $\theta^1$ are shared by all sentences in a document but are specific to documents, and they are nonparametric, i.e. the notation is $F - G - N P$. The sentences and documents are not clustered, so the DoS-classification of STM is $C = F - F; F = G - N P; N; N$.

### 5 Generalized Bayesian Model for Grouped Sequential Data

Having discussed the DoS-classification of various existing models, it is clear that despite over a decade of research on topic models, there are several DoS-classifications for which there are no existing models. But instead of trying to point out those classifications individually and propose models following them, we now propose a generalized Bayesian Model for grouped sequential data. We will show that by specific settings of this model, it is possible to recover all the previously discussed models (or their close variants). Other models, not explored so far, can also be obtained from it.

#### 5.1 GBM-GSD

We consider sequential data with $L$-levels of grouping, where the groups are sequential in every level (e.g. in document modelling, we will consider the sentences within each document, and the documents themselves, are sequentially arranged). We consider that clustering happens at all levels, i.e. $\{\theta^1\}, \ldots, \{\theta^L\}$ all exist. To capture the sequential nature, we will assume that at every level (say $l$), there is a collection of distributions $\{\theta^l\}$ from which one can be chosen for each group, conditioned on the previous assignments (as considered in sHDP-HMM and TSM). We also consider that all the distributions are infinite-dimensional (i.e. $NP$), i.e. neither the number of mixture components nor the number of clusters formed at each level is known in advance. We also consider that all the mixture components are accessible to all the level-2 groups, but introduce a binary random vector $B_l$ specific to each datapoint. This vector indicates which all mixture components are accessible to each datapoint. We will show that using this vector, we can make the mixture components group-specific or cluster-specific, and also capture other more intricate structures that would not be possible without it.

**Algorithm 1 Generalized Bayesian Model for Group Sequential Data (GBM-GSD)**

```plaintext
1: $\phi_k \sim H, \forall k$
2: for $g = 1 : G_1$, do
3: \hspace{1cm} $Z^g(1) \sim \theta^1 \mid Z^{g \text{prev}}(1), \ldots, Z^{g \text{prev}}(G_1)$
4: \hspace{1cm} end for
5: for $l = 1 : 2$ do
6: \hspace{1cm} for $g = 1 : G_l$, do
7: \hspace{2cm} $Z^g(l) \sim \theta^l \mid Z^{g \text{prev}}(1), \ldots, Z^{g \text{prev}}(G_l)$ where $j = Z^{l+1}(D^{l+1}(g))$
8: \hspace{1cm} end for
9: \hspace{1cm} end for
10:\hspace{1cm} for $i = 1 : N$, do
11: \hspace{2cm} $B_i = f(B_{i-1}, \ldots, Z^i_{g-1}, Z^i_g, \ldots, Z^i_L)$
12: \hspace{2cm} $Z^i_k \sim B_i \odot \theta^1 \mid Z^i_1, \ldots, Z^i_{g-1}$ where $j = Z^i(D^i_k)$
13: \hspace{2cm} $Y_i \sim \phi_k$ where $k = Z^i_k$
14: \hspace{1cm} end for
```

generative process hierarchically clusters the groups from top to bottom level. At every intermediate level $l$, it assigns $Z^i(g)$ to each group $g$ at level $l$. But for that it will have access to only those $\theta^l$-distributions, that are specific to the cluster $Z^i_g$ of group $g$ as a result of the clustering at level $(l + 1)$. If group $g$ is part of group $m = D^{l+1}(g)$ at level $(l + 1)$ then the $\theta^m$-distributions corresponding to $Z^{l+1}(m)$ must be used. Finally, at level 1, each datapoint $i$ is assigned a binary vector $B_i$ conditioned on the $B$-vectors corresponding to all previous datapoints. The distribution $\theta^1$ is convoluted with this vector $B_i$, so that a subset of the components are available to datapoint $i$. 

5.2 Recovery of Existing Models

The level-1 model for news transcripts can be recovered easily. By setting $B_i$ as a vector of all 1-s for all the datapoints, and by making $\theta^1$ conditioned only on $Z_{prev}^1$ and GEM-distributed we get back HDP-HMM. In case $\theta^1$ is also independent of the previous assignments, we can have DP-MM, and if it is finite-dimensional it will GMM provided the base distribution $H$ is Gaussian.

When $L = 2$, to recover HDP we need to define $\theta^2$ such that $Z^2_i(g) = g$ for all the groups $g$, so that groups are not clustered. Then we again set $B_i$ to be the vector of all 1-s, and make $\{\theta^1\}$ independent of all previous assignments of $Z^1_i$. The $\{\theta^1\}$ should be drawn from a GEM. If the $\{\theta^1\}$ are finite-dimensional and drawn IID from a Dirichlet, and if all the $\{\phi_k\}$ are also drawn from a Dirichlet, then we have LDA.

NDP involves nonparametric clustering of level-2 groups without sequential ordering, so generation of $Z^L_i(g)$ should be independent of previous assignments, and $\theta^L$ should be drawn from a GEM. NDP also has the special characteristic that the different level-2 clusters do not share the same mixture components. This can be managed by setting $B_i$ through an appropriate function $f$, which return a vector with 0 for those mixture components that have been assigned in other level-2 clusters, i.e. $B_{ik} = 0$ if $\exists j$ such that $Z^2_{D_i} \neq Z^2_{D_j}$ and $Z^1_i = k$.

When $L = 3$, MLC-HDP can be recovered by removing the conditioning on previous assignments in the assignment of $Z^2_i, Z^3_i$, and by setting $B_i$ to be the vector of all 1-s. The $\{\theta^1\}$ should be drawn from GEMs. For TSM, $\theta^1$ should ensure that documents are not clustered, $B_i$ should be the vector of all 1-s, and assignment of $Z^3_i$ should be independent of all previous assignments. Regarding $Z^2_i, \theta^2$ should ensure that for any sentence (level-2 group) $g$, $Z^2_i(g)$ should be either $Z^2(prev(g))$ or $Z^2(prev(g)) + 1$.

6 News Transcript Segmentation

We want to extend the generative framework for grouped sequential data (Algorithm 1) for modeling news transcripts. This data is hierarchical since there are broad news categories like politics, sports etc., under which there are individual stories or topics. In the Bayesian approach, we consider mixture components $\{\phi\}$ that correspond to these stories, and the broad categories are represented with distributions $\{\rho\}$ over these stories. As usual, each $\theta^1$-distribution is specific to a level-2 cluster (segment), and such clustering is induced by $\{\theta^2\}$, specific to the level-3 groups (the transcripts). The transcripts are not clustered. The observed datapoints $Y_t$ are word-tokens, each represented as an integer (index of the word in the vocabulary). We define $prev(i) = i - 1$ if $(i, i - 1)$ are in the same sentence, otherwise $prev(i) = -1$. Similarly, $next(i)$ is defined within sentences. Also, $prev$ and $next$ are defined for sentences. $Z^1_i$ indicates the news story (level-1) and $Z^2_i$ indicates the news category (level-2) that token $i$ is associated with. Each sentence is a level-2 group.

6.1 LaDP model for news transcripts

News transcripts were first modelled by Layered Dirichlet Process (LaDP) [Mitra et al., 2013]. Several versions of this model was proposed, with different combinations of exchangeability properties at the different layers (which included MLC-HDP). Here, the level-2 groups were not sentences, but the word-tokens themselves. In the most successful models, sequential structure was considered at both level-1 and level-2, i.e. the assignment of $Z^1_i$ and $Z^2_i$ are conditioned on $Z^1_{prev(i)}$ and $Z^2_{prev(i)}$, respectively. The clustering at both level-2 and level-1 are nonparametric. The DoS classification of LaDP is $C = F - F; C = F - NP - S; F - NP - S; N$.

6.2 Modeling Temporal Structure

News transcripts have characteristic temporal features regarding assignments of $Z^2_i$ and $Z^1_i$ for which the GBP-GSD needs to be modified appropriately. These features are discussed below. LaDP is insufficient for news transcripts, because it does not capture all of them.

Number of Level-2 clusters (segments) are fixed and known. In case of news transcripts from a particular source, it can be expected to have $K$ news categories in fixed order (say politics, national affairs, international affairs, business and sports, in that order). Segment is the task of linear clustering of words/sentences, i.e. each word/sentence $s$ can be assigned to either $Z^2_{prev(s)}$ or $Z^2_{prev(s)} + 1$. In LaDP, each datapoint $i$ is assigned a value of $Z^1_i$ and $Z^2_i$ based on the assignments of $prev(i)$, and segmentation happens based on these assignments. But this does not guarantee the formation of $K$ segments. To overcome this issue, let it be known to model that the observed data sequence has $K$ level-2 segments. Then the sequence can be partitioned into $K$ parts of sizes $N_1, N_2, \ldots, N_K$. These sizes may be modeled by a Dirichlet distribution where the parameters $\gamma_k$ signify the relative lengths/importance of the news categories.

$$\left(\frac{N_1}{\sum_{k=1}^{K} N_k}, \ldots, \frac{N_K}{\sum_{k=1}^{K} N_k}\right) \sim Dir(\gamma_1, \ldots, \gamma_K); Z^2_i = s$$

where

$$\sum_{k=1}^{s-1} N_k < j \leq \sum_{k=1}^{s} N_k$$

In the GBS-GSD, the $\theta^2$ needs to be defined as a deterministic function, conditioned on $\{N_1, \ldots, N_K\}$.

Topic Coherence has been considered in various text segmentation paper like [Eisenstein and Barzilay, 2008] [Mitra et al., 2013]. This is the property that within the same level-2 segment, successive datapoints are likely to be assigned to the same topic (mixture component). This can be easily modelled by the Markovian approach, i.e.

$$Z^1_i \sim \rho_{Z^1_{prev(i)}}(i - 1) + \rho_{Z^1_{prev(i)}}(i \mid \phi)$$

This means that the $i$-th datapoint can be assigned the $Z^1_i$-value of its predecessor $prev(i)$ with probability $\rho$, or any value with probability $(1 - \rho)$. The other available values are dictated by $B_i$, as discussed next. This is similar to the BE mixture model [Mitra et al., 2013].

Level-2 segments do not share mixture components, because each individual news story (topic) can come under only one news category. Also, Topics do not repeat inside a Level-2 segment. Inside a level-2 segment $s$, successive datapoints are expected to be assigned to the same mixture component due to temporal coherence. However, in news transcript, a news story will be told only once, which means that
which are the level-2 changepoints, i.e. the set of datapoints \( I \) about 300-350 sentences per transcript. In this dataset, the datapoints per sequence are too few for all following points in the segment, so that \( \phi_k \) cannot be sampled again. The generative process is as follows:

**Algorithm 2 Generative Model for News Transcripts**

1. \( H_c \sim \text{Dir}(\beta) \) \( \forall c \)
2. \( c \sim U(K), \phi_0 \sim H_c(\alpha) \) \( \forall k \)
3. \( \theta^1 \sim \text{GEM}(\alpha) \) where \( s \in [1, K] \)
4. for \( g = 1 \) to \( G^2 \) do
5. \( B_{gk} = 1 \) \( \forall k \)
6. \( \{ s^g_1, \ldots, s^g_K \} \sim \text{Dir}(\gamma) \)
7. end for
8. for \( j = 1 \) to \( G^2 \) do
9. \( Z^2_j = s \) based on \( (N_{g1}, \ldots, N_{gK}) \) where \( g = D^2_j \)
10. end for
11. for \( i = 1 : N \) do
12. if \( Z^2_{D^2(i)} \neq Z^2_{D^2(\text{pred}(i))} \) set \( p = 0 \)
13. \( Z^1_i \sim p \phi^{Z^2_{D^2(i)}} + (1 - p)(B_{g \circ \theta^1}) \) where \( s = Z^2_{D^2(i)}, g = D^2_{i} \)
14. if \( (Z^1_i \neq Z^2_{\text{pred}(i)}) \) set \( B_{gk} = 0 \) where \( k = Z^1_i \), \( g = D^2_{i} \)
15. \( Y_i \sim \text{mult}(\phi_k) \) where \( k = Z^1_i \)
16. end for

Here \( G^3 \) is the number of transcripts, and \( G^2 \) the number of sentences across all the transcripts. Clearly this model has 3 levels, and sequential structure is considered at level 2 (sentences) and at level 1 (word-tokens). Any topic \( k \) belongs to a broad category \( c(k) \) \((\in \{1, \ldots, K\}) \) uniformly at random, and corresponding to each category we have a base distribution \( H_c \) which in turn are all drawn from a common base distribution \( \text{Dir}(\beta) \). This helps to capture the fact that mixture components are specific to level-2 segments. The documents are not clustered, the sentences are clustered (segmented) with fixed number of segments, and the number of topics (word-clusters) is not fixed. The topics are shared across all transcripts, but are specific clusters of sentences, the \( \theta^1 \)-distributions are specific to level-2 segments (clusters of sentences) but shared across transcripts, the \( \theta^2 \)-distributions are transcripts-specific (parametrized by \( \{N_{g}\} \)) and \( \theta^3 \) are not used. So the DoS-classification for the generative model of news transcripts is \( C - C - F; C - F - N P - S; G - P - S; N \).

6.3 Inference Algorithm

We now discuss inference for this model. We need an inference algorithm which ensures that \( K \) segments are formed. We start with the joint distribution.

\[
p(Y, Z^1, Z^2, B, N, \Phi, \beta, \{t\}, \{H\}) \propto \prod_{i=1}^{G^3} p(N_{gi}) \prod_{k=1}^{K} p(\theta^2_k) \\
\times \prod_{s} p(H_{c(i)}|H) \prod_{s} p(\phi_{s}|H_{c(i)}) \prod_{j=2}^{G^2} p(Z^2_j|Z^2_{\text{pred}(j)}, \{N\}) \\
\times \prod_{i=1}^{N} p(Z^1_i, B_{\text{next}(i)}|Z^1_{\text{pred}(i)}, B_{i}, Z^2_i, \theta^1)p(Y_i|Z^1_i, \Phi) \\
\]

We can collapse the variables \( \{H\}, \{\Phi\} \), and \( \{\theta^1\} \), and perform Gibbs Sampling. The key feature of this likelihood function is the presence of the \( \{N_{g}\} \) variables. To handle these, we introduce auxiliary variables \( I_{g1}, \ldots, I_{g,K-1} \) which are the level-2 changepoints, i.e. the set of datapoints \( \{i\} \) at which \( Z^2_i \neq Z^2_{\text{pred}(i)} \). Also note that \( \{Z^2\}, \{I\} \) and \( \{N\} \) are deterministically related. We introduce the \( I \) variables to simplify the sampling. We initialize the \( Z^2 \) variables by sampling a level-2 segmentation of the datapoints into \( K \) segments. The \( B \) and \( Z^2 \) variables are sampled accordingly. In each iteration of Gibbs Sampling, we consider the state-space of \( I_{gs} \) as \( I_{gs} \in \{I_{g,s-1}, \ldots, I_{g,s+1}\} \), i.e. the level-2 potential changepoints lying in between \( I_{g,s-1} \) and \( I_{g,s+1} \). The process is described in Algorithm 3. Here, \( B_{gs} = \{B_{\text{set}}\} \) where \( \text{set} = \{i : D^2_i = g, Z^2_i = s\} \), i.e. the set of datapoints in transcript \( g \) in segment \( s \). (similarly \( Z^2_{gs}, Z^1_{gs}, Y_{gs} \)) The major part in the Gibbs sampling is to sample the values \((\{B\}, \{Z^1\})\) for any segment \( s \), conditioned on the remaining \( B \) and \( Z^2 \) variables. This can be done using the Chinese Restaurant Process (CRP), where any component \( k \) may be sampled for \( Z^1_i \) (where datapoint \( i \) is within segment \( s \)) proportional to the number of times it has been sampled, provided \( B_{\text{pred}(i),k} = 1 \). The procedure is detailed in Algorithm 3, which is called Global Inference as it considers the overall structure of the transcript (as described in Sec6.2).

**Algorithm 3 Global Inference Algorithm by Blocked Gibbs Sampling (GI-BGS)**

1. for transcript \( g = 1 \) to \( G^2 \) do
2. Initialize \( I_g \) with \((K - 1)\) points by sampling from \( \text{Dir}(\gamma) \); 3. Set \( \{Z^2\} \) according to \( I \); 4. Initialize \((B), \{Z^1\} \) variables; 5. end for
6. Estimate components \( \hat{\phi} \leftarrow E(\phi|B, Z^1, Y) \); 7. while Not Converged do 8. for transcript \( g = 1 \) to \( G^2 \) do
9. for segment \( s = 1 : K \) do
10. \( t_{gs} \leftarrow \{\text{succ}(I_{g,s-1}), \ldots, \text{pred}(I_{g,s+1})\} \)
11. Update \( Z^2 \) according to \( I \)
12. \( p(Y_{gs}|B_{gs}, Z^1_{gs}, Z^2_{gs}, \hat{\phi}) \);
13. \( p(B|gs, \{Z^1\}_{gs}, \{B\}_{-gs}, \{Z^1\}_{-gs}, \{Z^2\}_s, Y, \hat{\phi}) \);
14. end for
15. Update components \( \hat{\phi} \leftarrow E(\phi|B, Z^1, Z^2, Y) \);
16. end while

7 Experiment on News Transcript Segmentation

For news transcript segmentation, we used the news transcripts used by [Mitra et al., 2013] for hierarchical segmentation. Here, each transcript has 4-5 news categories- politics, national affairs, international affairs, business, sports- in that fixed order. Overall, each transcript is about 5000 tokens long (after removal of stopwords and infrequent words), and has about 40 news stories, spread over the 5 categories. The task is to segment the transcripts into level two at two levels. At level 1, each segment should correspond to a single story, while at level 2, each segment should correspond to a news category. The endpoints of the sentences are assumed to be known (these can be figured out based on pause durations in speech-to-text conversion), and are used to define level-2 groups. There are about 300-350 sentences per transcript.

In this dataset, the datapoints per sequence are too few in number to learn the level-1 mixture components (topics).
Moreover, as already explained, each story occurs only once in a transcript, thus reducing learnability. Hence, we considered 60 randomly chosen transcripts, and using initial segmentations of each sequence by the level-1 changepoints, 136 topics were learnt using HDP. These topics form our initial estimate of Φ, using which we performed inference on individual sequences. The inference provides us with the $Z^2$ and $Z^1$ variables, based on which we can infer the segmentation at the two levels. We have gold standard segmentation available at both layers, and so we compute the segmentation errors ($S_1, S_2$) at both layers. $S_1$ and $S_2$ can be computed by taking the average $P_k$-measure [Mitra et al., 2013] for three different values of $k$, namely the maximum, minimum and average lengths of gold-standard segments (level-1 segments for $S_1$ and level-2 segments for $S_2$).

We can look upon segmentation as a retrieval problem, and define the Precision and Recall of level-2 segments (PR2 and RC2), and also for level-1 segments (PR1 and RC1). Let $i$ and $j$ be the starting and ending points of an inferred segment $s$, i.e. $Z^2_i = Z^2_{next(i)} = \ldots = Z^2_j = s$, but $Z^2_{prev(i)} \neq s$ and $Z^2_{next(j)} \neq s$. Then, if there exists $(i_0, j_0, s_0)$ such that $(i_0, j_0)$ defines a gold-standard segment $s_0$ satisfying $|i - i_0| < k$ and $|j - j_0| < k$, then inferred segment $(i, j, s)$ is aligned to gold standard segment $(i_0, j_0, s_0)$. Precision, recall of a segmentation are defined as

\[
\text{Precision} = \frac{\#\text{inferred segments aligned to a gold-standard segment}}{\#\text{inferred segments}} \\
\text{Recall} = \frac{\#\text{gold-standard segments aligned to an inferred segment}}{\#\text{gold-standard segments}}
\]

For level-2, the alignment threshold is set to 500, and at level-1 it is set to 10. As a baseline, we use sticky HDP-HMM [Fox et al., 2008] and LaDP [Mitra et al., 2013] at level-1, once again using the 136 HDP topics. For level-2, LaDP is the only baseline. We use the BE-BE-CE version, since that is the most successful according to [Mitra et al., 2013].

From the 60 news transcripts from which we learnt the 136 topics through HDP, we selected 3 (Trans1, Trans2, Trans3) to report the segmentation. Also, we selected another 3 (Trans91, Trans92, Trans93) from outside the learning set, for which we used the same initial values of Φ. The results are reported in Table 1. It is clear that in terms of all the measures we considered, GI-BGS outperformed both competitors at level-1. At level-2 also, GI-BGS is competitive on the three measures on all the transcripts except Trans1.

8 Conclusions

We carried out a study of various Bayesian models for hierarchically grouped data with emphasis on how they share mixture components and distributions among the groups. We also introduced the notion of Degree-of-Sharing (DoS) as a nomenclature for such models. We described a Generalized Bayesian model for this type of data, and showed how various existing models can be recovered from it. Next we used it to develop a new model for news transcripts, which has several peculiar temporal structures, and also provided an inference algorithm for hierarchical unsupervised segmentation of such transcripts. We showed that this model can outperform the existing LaDP model for this task. The DoS concept opens up possibilities to explore models with DoS-classifications that have not yet been considered, and the GBM-GSD can be used to capture complex temporal structures in the data.

References

[Mitra et al., 2013] A Mitra, BN Ranganath, and I Bhattacharya. A layered dirichlet process for hierarchical segmentation of sequential grouped data. In European Conference on Machine Learning and Knowledge Discovery in Databases, 2013.

[Mitra et al., 2013] A Mitra, BN Ranganath, and I Bhattacharya. A layered dirichlet process for hierarchical segmentation of sequential grouped data. In European Conference on Machine Learning and Knowledge Discovery in Databases, 2013.

[Teh et al., 2006] Y W Teh, M I Jordan, M J Beal, and D M Blei. Hierarchical dirichlet processes. Journal of the American Statistical Association, 101(476), 2006.

[Wulsem et al., 2012] D Wulsin, S Jensen, and B Litt. Hierarchical dirichlet process model with multiple levels of clustering for human eeg seizure modeling. Proceedings of the 23rd International Conference on Machine learning, 2012.

| Data | GI-BGS | LaDP | sHDP-HMM |
|------|--------|------|---------|
| PR1  | RC1 1 | S1   | PR1 1 | RC1 1 | S1   |
| Trans1 | 0.38 | 0.46 | 0.06 | 0.33 | 0.46 | 0.07 | 0.20 | 0.40 | 0.08 |
| Trans2 | 0.33 | 0.37 | 0.10 | 0.27 | 0.34 | 0.11 | 0.18 | 0.34 | 0.12 |
| Trans3 | 0.26 | 0.41 | 0.09 | 0.25 | 0.41 | 0.08 | 0.13 | 0.32 | 0.11 |
| Trans91 | 0.15 | 0.28 | 0.16 | 0.13 | 0.28 | 0.13 | 0.06 | 0.21 | 0.16 |
| Trans92 | 0.14 | 0.25 | 0.14 | 0.10 | 0.20 | 0.14 | 0.08 | 0.23 | 0.14 |
| Trans93 | 0.22 | 0.22 | 0.09 | 0.17 | 0.08 | 0.11 | 0.12 | 0.03 | 0.11 |

Table 1: Above: Comparison of news transcript segmentation at level-1 by sticky HDP-HMM, LaDP and GI-BGS. Below: News transcript segmentation at level-2 by LaDP and GI-BGS. Lower value of $S_1, S_2$ indicate better segmentation.