A Layered Dirichlet Process for Hierarchical Segmentation of Sequential Grouped Data

Adway Mitra[1], Ranganath B.N.[1], and Indrajit Bhattacharya[2]

[1] CSA Department, Indian Institute of Science, Bangalore, India
{adway,ranganath}@csa.iisc.ernet.in
[2] IBM India Research Lab, Bangalore, India
indrajitb@gmail.com

Abstract. We address the problem of hierarchical segmentation of sequential grouped data, such as a collection of textual documents, and propose a Bayesian nonparametric approach for this problem. Existing Bayesian nonparametric models such as the sticky HDP-HMM are suitable only for single-layer segmentation. We propose the Layered Dirichlet Process (LaDP), where each layer has a countable set of Dirichlet Processes, draws from which define a distribution over the countable set of Dirichlet Processes at the next layer. Each data item gets assigned to a distribution (index) from each layer of the hierarchy, leading to hierarchical segmentation of the sequence. The complexity of inference depends upon the exchangeability assumptions for the measures at different layers. We propose a new notion of exchangeability called Block Exchangeability, which lies between Markov Exchangeability (used in HDP-HMM) and Complete Group Exchangeability (used in HDP), and allows for faster inference than Markov Exchangeability. Using experiments on a news transcript dataset and a product review dataset, we show that LaDP generalizes better than existing non-parametric models for sequential data, and by simultaneously segmenting at multiple levels, outperforms existing models in terms of single-layer segmentation. We also show empirically that using Block Exchangeability greatly speeds up inference and allows trading off accuracy for execution time.

1 Introduction

We address the problem of hierarchical segmentation of sequential grouped data. For example, consider transcripts of news broadcast on television or radio. Here, each transcript represents a group of data points, which are the words in this case. The words in each transcript or group form a sequence that needs to be segmented. The segmentation needs to be at two layers — news categories, such as politics, sports, etc, and individual stories within a category. There are benefits to segmenting transcripts simultaneously, instead of individually. Stories are typically shared across transcripts, and transition patterns between stories (more important stories often come earlier) and categories (e.g. sports rarely comes before other categories) may also be shared across transcripts. Also, there are benefits to simultaneous segmentation into stories and categories. Inferring

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part II, LNAI 8189, pp. 465–482, 2013.

a story strongly suggests a category, while inferring a category increases the posterior probability for certain stories. Finally, while the number of categories may often be known or guessed, this is not true for the number of stories. In this paper, we propose a Bayesian approach for this problem, which is both hierarchical and non-parametric. For the news example, each story can be modeled as a distribution over words, and each category as a distribution over stories. The same stories and categories are shared between all news transcripts. Being non-parametric, the model does not require the number of stories to be specified.

The Dirichlet Process[1] is a measure over measures and is useful as a prior in Bayesian nonparametric mixture models, where the number of mixture components is not specified a-priori, and is allowed to grow with the number of data points. The Hierarchical Dirichlet Process(HDP)[5] hierarchically extends DP for grouped data, such as words partitioned into documents, so that mixture components are shared between groups. The DP is a completely exchangeable model (probability of the data is invariant to permutations in the sequence), while the HDP is completely exchangeable within each group (Group Exchangeable). As a result, these are not suitable as statistical models for segmentation. HDP variants such as the HDP-HMM [5] and sticky HDP-HMM [6], which satisfy Markov exchangeability, are more suitable for segmentation. However, these perform segmentation at a single layer and not at several layers simultaneously.

We propose the Layered Dirichlet Process, where each layer has a countable set of DP-distributed measures over integers. The integers at each layer serve as indices for the measures at the next layer. Each data item filters down this layered structure, where a measure at each layer assigns it to a measure at the next layer. Such assignments for each data item in the sequence results in a hierarchical segmentation of the sequence. The assignment of a measure to each data item at each layer depends on the exchangeability assumption at that layer. For Complete Group Exchangeability, it depends only on its assignment at the previous layer. For other partial Group Exchangeabilities, it additionally depends on the assignments of other data items at that layer. We perform inference for LaDP using collapsed Gibbs sampling. Since the assignments are coupled across layers, inference is naturally complex. We propose a new notion of exchangeability called Block Exchangeability. We show that this relaxes Complete Exchangeability to capture sequential patterns, but is stricter than Markov Exchangeability with significantly lower inference complexity.

Using experiments on multiple real datasets, we show that by modeling grouping at multiple layers simultaneously, LaDP is able to generalize better that state-of-the-art non-parametric models. We also show that simultaneous segmentation at multiple layers improves segmentation accuracy over single layer segmentation. Additionally, using Block Exchangeability leads to significantly faster inference compared to Markov Exchangeability, while incurring negligible increase in segmentation error and perplexity in some cases, and actually improving performance in some others. Interestingly, Block Exchangeability has the novel ability of trading off efficiency for accuracy.

2 Background and Related Work

In this section we briefly review some existing nonparametric models used as priors for infinite mixture models, and existing notions of exchangeability.

DP Mixture Model and Complete Exchangeability: A random measure G on Θ is said to be distributed according to a Dirichlet Process (DP) [1] ($G \sim DP(\alpha, H)$) with base distribution H and concentration parameter α if, for every finite partition $\{\Theta_1, \Theta_2, \ldots, \Theta_k\}$ of Θ, $(G(\Theta_1), G(\Theta_2), \ldots, G(\Theta_k)) \sim Dir(\alpha H(\Theta_1), \alpha H(\Theta_2), \ldots, \alpha H(\Theta_k))$. The stick-breaking representation shows the discreteness of draws G from a DP:

$$\theta_k \sim H;\ \beta_k = \hat{\beta}_k \prod_{i=1}^{k-1}(1-\hat{\beta}_i);\ \hat{\beta}_i \sim Beta(1,\alpha);\ G \triangleq \sum_k \beta_k \delta_{\theta_k}$$

We write $\beta_k \sim GEM(\alpha)$. Given n independent draws $\{\theta_i\}_{i=1}^n$ from G as above, the predictive distribution of the next draw, on integrating out G, is given by $p(\theta_{n+1}|\theta_1 \ldots \theta_n) \propto \sum_{k=1}^K n_k \delta_{\phi_k} + \alpha H$, where $\{\phi_k\}_{k=1}^K$ be the K unique values taken by $\{\theta_i\}_{i=1}^n$ with corresponding counts $\{n_k\}_{k=1}^K$. This shows the clustering nature of the DP. Using the DP as a prior results in an 'infinite mixture model' for data $\{w_i\}_{i=1}^n$ with the following generative process:

$$G \sim DP(\alpha, H); \theta_i \overset{iid}{\sim} G;\ w_i \overset{iid}{\sim} F(\theta_i), i = 1 \ldots n$$

where F is a measure defined over Θ. This is called the DP mixture model [1]. This can alternatively be represented using the stick-breaking construction and integer latent variables z_i as follows:

$$\beta \sim GEM(\alpha);\ \theta_k \sim H,\ k = 1 \ldots \infty\ ;\ z_i \sim \beta;\ w_i \sim F(\theta_{z_i}),\ i = 1 \ldots n$$

An important notion for hierarchical Bayesian modeling is that of exchangeability [11,2]. Given any assignment $\{\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_n\}$ to a sequence of random variables $\{z_n\} \in \mathcal{S}$, where $\mathcal{S}$ is a space of sequences, exchangeability (under joint distribution P on $\mathcal{S}$) defines which permutations $\{\bar{z}_{\pi(1)}, \bar{z}_{\pi(2)}, \ldots, \bar{z}_{\pi(N)}\}$ of the assignment have the same probability (under P). In general, any notion of exchangeability E is defined using a statistic, which we call Exchangeability Statistic $S_E(z)$. A model, defining a joint distribution P, is said to satisfy exchangeability E if $S_E(z_1) = S_E(z_2)$ implies $P(z_1) = P(z_2)$, for all $z_1, z_2 \in \mathcal{S}$.

Given a sequence $z \in \mathcal{S}$, define $S_C(z) = \{n_i\}_{i=1}^K$ as the vector of counts of the K unique values occurring in it, where n_i is the count of the i^{th} unique value. Using $S_C(z)$ as the exchangeability statistic leads to the definition of Complete Exchangeability (CE), under which all permutations are equiprobable.

De Finetti's Theorem [3] shows that if an infinite sequence of random variables z is infinitely exchangeable (meaning that every finite subset is completely exchangeable) under a joint distribution $P(z)$, then the joint distribution can be equivalently represented as a Bayesian hierarchy: $P(z) = \int_\theta P(\theta) \prod_i P(z_i|\theta) d\theta$. It can be shown that a sequence drawn from a DP mixture model, using a similar hierarchical generation process, satisfies Complete Exchangeability.

HDP Mixture Model and Group Exchangeability: Now consider grouped data of the form $\{w_i, g_i\}_{i=1}^n$, where $g_i \in \{1, G\}$ indicates the group to which w_i belongs. The Hierarchical Dirichlet Process (HDP) [5] allows sharing of mixture components $\{\phi_k\}$ across groups using two levels of DPs:

$$\phi_k \sim H,\ k = 1 \ldots \infty; \beta \sim GEM(\gamma),\ \pi_j \sim DP(\alpha, \beta),\ j = 1 \ldots G$$
$$z_i \sim \pi_{g_i};\ w_i \sim F(\phi_{z_i}),\ i = 1 \ldots n \quad (1)$$

This generative procedure for the data is called the HDP mixture model. We have modified the representation to make the group variable explicit, which we can build upon for our work. Note that the HDP can also be represented directly using measures instead of indices. The HDP mixture model can be shown to satisfy a notion of partial exchangeability called Group Exchangeability. For grouped data of the form $\{z_i, g_i\}_{i=1}^n$, where the z_i and g_i variables take K and G unique values respectively, define $S_G(z, g) = \{\{n_{j,k}\}_{k=1}^K\}_{j=1}^G$, where $n_{j,k} = \sum_{i=1}^n \delta(z_i, k)\delta(g_i, j)$. Group Exchangeability (GE) is characterized by the exchangeability statistic $S_G(z, g)$. For GE models, all intra-group permutations are equiprobable, but probability changes with exchange of values across groups.

Other Group Exchangeable Nonparametric Models: For grouped data $\{w_i, g_i\}_{i=1}^n$, the Nested Dirichlet Process (NDP) [7] proposes the following generative model with two layers of latent variables (z^2, z^1) for each data item:

$$\phi_{k,l} \sim H,\ k, l = 1 \ldots \infty; \beta_k^1 \sim GEM(\beta),\ k = 1 \ldots \infty; \beta^2 \sim GEM(\alpha);$$
$$z_g^2 \sim \beta^2,\ g = 1 \ldots G; z_i^1 \sim \beta_{z_{g_i}^2}^1; w_i \sim \phi_{z_{g_i}^2, z_i^1},\ i = 1 \ldots n$$

Unlike the HDP, only some groups share mixture components. Additionally, unlike the HDP they also share distributions over these components.

The MLC-HDP [9] models data of the form $\{w_i, g_i^1, g_i^2, g_i^3\}_{i=1}^n$, which is grouped at 3 different levels, and proposes the following generative process:

$$\phi_k \sim H,\ k = 1 \ldots \infty; \beta^3 \sim GEM(\gamma^3),\ \beta^2 \sim GEM(\gamma^2),\ \beta^1 \sim GEM(\gamma^1);$$
$$\pi^3 \sim DP(\alpha^3, \beta^3),\ \pi_k^2 \sim DP(\alpha^2, \beta^2),\ \pi_l^1 \sim DP(\alpha^1, \beta^1),\ k, l = 1 \ldots \infty;$$
$$z_a^3 \sim \pi^3\ \forall a;\ z_{ab}^2 \sim \pi_{z_a^3}^2 \forall a \forall b;\ z_{abc}^1 \sim \pi_{z_{ab}^2}^1 \forall a \forall b \forall c;\ w_i \sim \phi_{z_{g_i^3, g_i^2, g_i^1}^1},\ i = 1 \ldots n$$

Here the mixture components can be shared by all groups, and two groups can have identical distributions over these components with non-zero probability.

Segmentation, HDP-HMM and Markov Exchangeability: Now we come to the segmentation problem for a sequence $\{w_i, z_i\}$ where the the variables w_i are observed while $z_i \in \{1, 2 \ldots\}$ are latent, with distribution $P(w, z) = P(z)P(w|z)$. Given any assignment to the $\{z_i\}$ variables, segments are defined as maximal sub-sequences (s, e) such that $z_e = z_s = z_i$ for $s \leq i \leq e$. Since $\{z_i\}$ variables are random, a natural definition for the segmentation problem is to first perform inference to find the optimal assignment to $\{z_i\}$ according to the posterior distribution $P(z|w)$, and then identifying segments for this assignment. Instances

of this problem include segmentation according to topics for textual documents, and according to speaker in conversational audio. Naturally, distinguishing between different permutations is critical for segmentation of grouped (un-grouped) data, and GE (CE) assumptions for $P(z)$ are not appropriate, since all permutations are equiprobable. Therefore, HDP (DP) mixture models are not suitable for such segmentation tasks. These call for more discerning models that satisfy other notions of exchangeability that distinguish between different segmentations of $\{w_i, z_i\}$ represented by different assignments to $\{z_i\}$.

To model (ungrouped) data $\{w_i\}$ with such properties, the HDP-HMM [6] considers the mixture components z_i as states of an HMM with infinite state-space. This is done by identifying the groups as well as the mixture components in the HDP with the HMM states. Now $\pi_j \sim DP(\alpha, \beta)$ is considered as transition distribution for the j^{th} state, and is used to generate the next state:

$$\pi_j \sim DP(\alpha, \beta),\ j = 1 \ldots \infty; z_i \sim \pi_{z_{i-1}};\ w_i \sim F(\theta_{z_i}),\ i = 1 \ldots n \qquad (2)$$

A special case of this is the Sticky HDP-HMM (sHDP-HMM) [6], which increases the probability of self-transition as $\pi_j \sim DP(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_j}{\alpha+\kappa})$, to enforce sequential continuity of mixture components which occur naturally in speech (where a mixture component represents a speaker) and text (where a mixture component represents a topic). Though originally developed for single sequences, the HDP-HMM and sHDP-HMM models can also be extended for grouped data.

Consider the following statistic: $S_M(z) = (\{n_{ij}\}_{i=1,j=1}^{K,K}, s)$, where n_{ij} is the number of transitions from the i^{th} unique value to the j^{th} unique value in the sequence z, and $z_1 = s$. Using S_M as the exchangeability statistic leads to the definition of Markov Exchangeability (ME) [2]. Intuitively, this means that two different sequences are equiprobable under the joint distribution, if they begin with the same value and preserve the transition counts between unique values. Representation theorems, similar to De Finetti's theorem, exist for Markov Exchangeability as well [2]. It can be shown that the HDP-HMM and sticky HDP-HMM mixture models satisfy Markov Exchangeability.

3 Hierarchical Segmentation and LaDP

We now discuss hierarchical segmentation of grouped data and propose Bayesian nonparametric models for it, using existing notions of partial exchangeability.

Hierarchical Segmentation: Consider grouped data of the form $\{w_i, g_i\}$, where $g_i \in \{1 \ldots G\}$ indicates the group to which each data point w_i belongs. The data $\{w_i : g_i = g\}$ in each of the G groups forms a sequence. In the news transcript example, each group corresponds to one transcript, and the words in each transcript form a sequence. We call such data sequential grouped data.

Our task is to segment the sequential data in each of the groups at multiple layers. We define an L-layer segmentation of the data as follows. Instead of a single latent variable z_i as before, we associate L latent variables $\{z_i^l\}_{l=1}^L$, each

taking integer values, with the i^{th} data point. We call z_i^l the group for the i^{th} data point at layer l. We will assume that the grouped structure of the input data provides the grouping at the highest layer, i.e. $z_i^{L+1} = g_i$. Given any assignment to these n sets of group variables, the hierarchical segmentation at any layer l $(1 \leq l \leq L)$ is defined using $\{\{z_i^{l'}\}_{l' \geq l}\}_{i=1}^n$, which are all the group variables at layer l or higher. Two data points at positions i and j $(i < j)$ belong to the same hierarchical segment at layer l if the group variables of intermediate data points k are identical at layers l and above: $z_i^{l'} = z_k^{l'} = z_j^{l'}$, $\forall k, i \leq k \leq j$, $l \leq l' \leq L+1$. This may also be defined recursively. Two data points belong to the same segment at layer l if they belong to the same segment at layer $l+1$, and all intermediate group variables z_k^l have the same value at layer l. In the case of news transcripts, the group variables z_i^{L+1} at the highest layer indicate which transcript the i^{th} data point (word) belongs to, which is provided as input. Imagine the next layers to correspond to categories (layer 2) and stories (layer 1). Then, two words would belong to the same category segment (layer 2), if they are in the same transcript and share the same category label with all intermediate words. Similarly, two words belong to the same story segment (layer 1) if they belong to the same category segment and have the same story labels as all intermediate words. The problem is to find the L-layer hierarchical segmentation at layer 1.

Completely Exchangeable Layered Dirichlet Process: We define a joint probability distribution over the n sets of group variables $\{\{z_i^l\}_{l=1}^L\}_{i=1}^n$ and the n data points $\{w_i\}_{i=1}^n$ using a hierarchical Bayesian approach. For each layer l, $1 \leq l \leq L$, we have a countable set of measures $\{\pi_g^l\}_{g=1}^\infty$ defined over positive integers. The group variables $\{z_i^l\}_{i=1}^n$ at layer l serve as indexes for these measures. Using this countable property, the atoms of all of these measures at layer l, which are integers, correspond one-to-one with the measures at the next layer $l-1$. This gives us a hierarchy of measures, in the sense that each π_g^l forms a measure over the measures $\{\pi_{g'}^{l-1}\}_{g'=1}^\infty$ at the next layer. Finally, at the lowest layer, each $F(\phi_k)$ is a measure over the space $\mathcal{W}$ of the observations $\{w_i\}$. For discrete text data, these are multinomial distributions over the vocabulary.

Next we need to define the measures $\{\pi_g^l\}_{g=1}^\infty$ and the exchangeability properties at each layer l. In LaDP, we define each of these distributions to be DP-distributed. We begin with the simplest case, which assumes complete exchangeability at every layer. The generative process looks as follows:

$$\begin{aligned} \phi_k &\sim H,\ k = 1 \ldots \infty \\ \beta_g^l \sim GEM(\gamma^l);\ \pi_g^l &\sim DP(\alpha^l, \beta_g^l),\ g = 1 \ldots \infty,\ l = L \ldots 1 \\ z_i^l \sim \pi^l_{z_i^{l+1}},\ l = L \ldots 1, \quad w_i &\sim F(\phi_{z_i^1}),\ i = 1 \ldots n \end{aligned} \tag{3}$$

In each layer, a countable set of measures is first constructed by drawing from a DP with a distribution over integers as a base distribution. These measures as a result also have support over integers, which serve as indexes to the measures at the next lower layer, which also form a countable set. Once we have this hierarchy of measures, the group variable z_i^l for each data point at each layer l is sampled

from the measure indexed by the group z_i^{l+1} assigned at the previous (higher) layer. The measures at the lowest layer (layer 1) are sampled from a suitable base distribution H. H could be Dirichlet when each ϕ_g is a multinomial parameter. It is easy to verify that the above process satisfies Complete Exchangeability (CE). As such, we call this model the CE-LaDP.

Layered Dirichlet Process for Segmentation: Since CE models are not useful for segmentation, we next incorporate Markov Exchangeability within LaDP. The key to incorporating ME is to relax the *iid* assumption for the group variables, within a layer as in the HDP-HMM, and additionally across layers, and generate z_i^l conditioned on some of the previously sampled groups $\{z_j^{l'} : j < i,\ l' > l\}$. The HDP-HMM identifies groups with states and makes the Markovian independence assumption that $P(z_i|z_{<i}) = P(z_i|z_{i-1})$. Accordingly, it defines transition distribution π_g over next states for each group (state) g. In our case, we make the following independence assumption: $P(z_i^l|z^{>l}, z_{<i}^l) = P(z_i^l|z_i^{l+1}, z_{p(i,l)}^l)$, where $z^{>l} \equiv \{z_i^{l'} :\ l' > l\}$, $z_{<i}^l \equiv \{z_{i'}^l :\ i' < i\}$, and $p(i,l) \equiv \{j : z_j^{l+1} = z_i^{l+1}, j < i,\ z_k^{l+1} \neq z_i^{l+1}, j < k < i\}$ is the previous datapoint having the same group as i at layer $l+1$. This means that the group assignment to data point i at layer l depends on its group at the layer $l+1$ (like in CE-LaDP), and also on the group assignment at layer l of its parent datapoint $p(i,l)$. (We later overload the notation $p(i,l)$ for brevity to refer to the group value $z_{p(i,l)}^l$ as well. We accordingly define transition distribution $\pi_{g,g'}^l$ over next groups from each parent group g' at layer l, in each assigned group g in layer $(l+1)$. The generative process for layer l $(L \geq l \geq 1)$ is defined as:

$$\beta_g^l \sim GEM(\gamma^l),\ \pi_{g,g'}^l \sim DP(\alpha^l, \beta_g^l),\ g, g' = 1 \ldots \infty,$$
$$z_i^l \sim \pi_{z_i^{l+1}, p(i,l)}^l,\ i = 1 \ldots n$$

Part of the graphical model is shown in Fig. 1.

For the first data point in any group in layer $(l+1)$, $p(i,l)$ is undefined, and z_i^l is sampled from $\beta_{z_i^{l+1}}^l$. It can be shown that this generative process satisfies ME within each group at layer l. When this process is used at all layers, we call the model ME-LaDP. As in sticky HDP-HMM, we may add more probability κ^l for self-transitions: $\pi_{g,g'}^l \sim DP(\alpha^l + \kappa^l, \frac{\alpha^l \beta_g^l + \kappa^l \delta_{g'}}{\alpha^l + \kappa^l})$, where κ^l is a *c*ontinuity parameter. This is done to encourage the same mixture component for adjacent data points. This captures the temporally smooth nature of most real-world data, and also encourages segmentations (based on group index assignments).

Layer-Specific Exchangeability: We have defined CE-LaDP as using CE at all layers, and ME-LaDP as using ME at all layers. However, each of the processes can be defined specific to a single layer, and it is possible to use layer-specific exchangeability assumptions, as demanded by particular applications. Indeed, we use such *mixed exchangeability models* in our experiments. As example, the generative process of such a model is described in the Appendix [13].

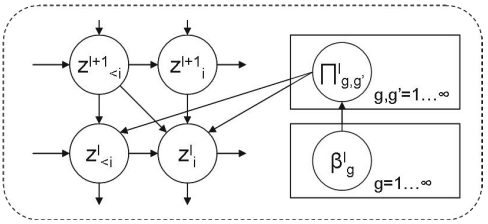

Fig. 1. Graphical model of LaDP focused on the i^{th} data point in two adjacent layers

Incorporation of Domain Knowledge: The $\{\beta_g^l\}$ variables at each layer l in Eqns. 3 and 4 are group-specific distributions over indexes (and measures) at the next layer $l-1$. These are useful for incorporating domain knowledge such as distribution over categories for specific news transcripts. For example, we can indicate that the j^{th} transcript is dominated by category index k by setting distribution β_j^L over category indexes at the appropriate layer ($L=2$) to $\sum_{c=1}^{\infty}\delta_k(c)$.

In some cases, one may also wish to bias the ϕ-distributions using domain knowledge. One option is to directly specify these ϕ_k. For weaker supervision, we may introduce an additional layer $l=0$:

$$\begin{aligned}\beta_g^0 \sim GEM(\gamma^0);\ \pi_g^0 \sim DP(\alpha^0, \beta_g^0),\ g=1\ldots\infty\\ z_i^0 \sim \pi^0_{z_i^1};\ w_i = \mathcal{W}_{z_i^0},\quad i=1\ldots n\end{aligned} \tag{4}$$

Now, on specifying some of the $\{\beta_g^0\}$ distributions, the corresponding distributions $\{\pi_g^0\}$ will be similar to these, depending on the concentration parameter α^0, and the words will be drawn from these $\{\pi_g^0\}$ distributions. Observe that we use complete (group) exchangeability at layer $l=0$.

Relation with Other Models: Observe the relation between the CE-LaDP (Eqn. 3) and the HDP mixture model (Eqn. 1). Recall that the group at the highest layer z_i^{L+1} is the input group label g_i. For $L=1$, this is exactly the HDP mixture model. However, by separating the group index in the HDP generative model, and identifying the z_i variable as the random group variable leading to the next layer, the CE-LaDP naturally extends the HDP generative process to generate layered grouping. A similar relation holds between the ME-LaDP with $L=1$ and the HDP-HMM. The MLC-HDP [9] extends HDP to 3 layers, with each data point w_i having input group indices g_i^3, g_i^2, g_i^3. When the group indices z^3 are observed (rather than sampled from π^3, as in [9]) and are identified with the indices g_i for LaDP, and additionally the input group indices $g_i^3 = i$, $g_i^2 = 1$ and $g_i^1 = 1$ are shared by data points w_i, we get back the CE-LaDP with $L=2$. Thus the LaDP framework can be used to generalize existing models to any number of layers. Secondly, the LaDP enables incorporation of domain knowledge in all layers. Among existing models, only the recently-proposed DP-MRM [10] is equipped to incorporate such domain knowledge, though only for a single layer. Finally, while all existing methods only use a single exchangeability property

(CE or ME), LaDP has the attractive property that different layers can have different exchangeability properties. In the next section, we define a new notion of exchangeability, and show how it can be incorporated in any layer of LaDP.

4 Block-Exchangeability and BE-LaDP

The models that we have introduced for segmentation satisfy ME. However, as we analyze later, there is a significant price to be paid in terms of complexity of inference as we move from CE and GE to ME. This is particularly severe for us, since we need to segment simultaneously at multiple layers. In this section, we explore an alternative notion of exchangeability, called Block Exchangeability (BE), that allows segmentation, but is less expensive than ME for inference.

Block Exchangeability. Consider the following statistic for a sequence z (with k unique values in it): $S_B(z) = (\{n_{i,i}, n_{i,-i}\}_{i=1}^{k}, e)$, where $n_{i,i}$ is the number of transitions from the i-th unique value of z to itself, $n_{i,-i}$ is the number of transitions from the i-th unique value to all other values and e is the value at the last position. Using $S_B(z)$ as the exchangeability statistic defines Block Exchangeability (BE) for a sequence z with distribution $P(z)$, or for a model that defines $P(z)$. First we observe some properties of a block exchangeable model.

Theorem 1. *If a model defining a joint distribution P is Completely Exchangeable then it is necessarily Block Exchangeable, but not the converse.*

Theorem 2. *If a model defining a joint distribution P is Block Exchangeable then it is necessarily Markov Exchangeable, but not the converse.*

BE-DP Mixture Model: Consider grouped data of the form $\{w_i, z_i, g_i\}_{i=1}^{n}$, where $g_i \in \{1 \dots G\}$ indicates the pre-assigned group corresponding to the i^{th} data point, and $z_i \in \{1 \dots \infty\}$ is the (latent) index of the mixture component corresponding to w_i. We now define a DP-based non-parametric mixture model for sequential grouped data that satisfies Block Exchangeability, as follows:

$$\begin{aligned} \phi_k \sim H,\ k = 1 \dots \infty;\ \pi_g \sim DP(\alpha, \beta_g),\ g = 1 \dots G \\ q_{gk} \sim Beta(1, \kappa);\ \hat{\pi}_{gk} = q_{gk}\pi_g + (1 - q_{gk})\delta_k,\ g = 1 \dots G,\ k = 1 \dots \infty \\ z_i \sim \hat{\pi}_{g_i, p(i)};\ w_i \sim F(\phi_{z_i}),\ i = 1 \dots n \end{aligned} \quad (5)$$

The first two lines describe the BE-DP prior, and the last line shows data generation using the mixture indices z_i. $p(i)$ is the group assignment to the data point just before i in group g_i. For the first data point in any group, $z_i \sim \pi_{g_i}$.

Theorem 3. *The BE-DP prior distribution and the corresponding mixture model satisfy Block Exchangeability.*

The proofs of the theorems are available in our supplementary material [13].

We now provide an alternative representation of the BE-DP equivalent to that in Eqn. 5, but which provides a justification for its nomenclature by capturing the

structure of equi-probable permutations of any sequence. Consider the sequence $\{w_i, z_i, c_i, g_i\}_{i=1}^n$, where the variables w_i, z_i and g_i are as before, and we have added a binary variable $c_i \in \{0, 1\}$ for each data point. We change the generative process in Eqn. 5 to include the c_i variables as follows.

$$c_i \sim Ber(q_{g_i,p(i)}); \quad z_i = p(i), \text{ if } c_i = 0; \; z_i \sim \pi_{g_i}, \text{ else, } \; i = 1 \ldots n \tag{6}$$

Clearly, this version is equivalent to the generative process in Eqn. 5, in the sense that the marginal $P(\mathbf{w}, \mathbf{z}, \mathbf{g})$ obtained by summing out $\mathbf{c}$ from $P(\mathbf{w}, \mathbf{z}, \mathbf{c}, \mathbf{g})$ is identical to that obtained from Eqn. 5. But this version has the advantage that the introduction of the auxiliary variables makes inference more tractable, as we will see in Section 5.

Separately, introduction of the the c variables provides some new insights into Block Exchangeability. Observe that as long as c_i remains 0, z_i retains the value $p(i)$ of the previous mixture index in its group g_i. When c_i takes value 1, z_i takes a new random value based on a group-specific distribution over mixtures π_{g_i}, that does not depend on the previous mixture index $p(i)$. Thus c_i acts as a change-point indicator variable. The distribution of c_i, and therefore the continuity of the the current mixture component, depends on the group and also the mixture component. Thus mixture components are characterized by how long they persist, but not by what follows them in the sequence. We use the term *block* to refer to a sub-sequence $\{s, s+1, \ldots, s+m\}$ such that $p(i+1) = p(i), \forall i \in [s+1, s+m]$, $c_s = 1$ and $c_{s+k} = 0 \; \forall k \in [1, m]$. Note that this implies $z_s = z_{s+1} = \cdots = z_{s+m}$. For a block exchangeable sequence, permutations of blocks as a whole does not change probability of the sequence. Consider two different assignments x and y to $\{c_i, z_i\}_{i=1}^n$. We say that a block of x corresponds to another block of y if they are of same length, and have the same value of $\{z_i\}$ for the data points within them. Then, if there exists a bijection between the blocks of x and those of y, then they should have same probability for a BE model.

BE-LaDP: We now show that BE can be incorporated into any layer l of the LaDP instead of CE or ME. Consider the generative process in Eqn. 3 or in Eqn. 4. We modify the random variables for layer l as follows:

$$\begin{aligned} \beta_g^l \sim GEM(\gamma^l), \; \pi_g^l \sim DP(\alpha^l, \beta_g^l), \quad g = 1 \ldots \infty \\ q_{g,g'}^l \sim Beta(1, \kappa); \; \pi_{g,g'}^l = q_{g,g'}^l \pi_g^l + (1 - q_{g,g'}^l)\delta_{g'}, \; g, g' = 1 \ldots \infty \\ z_i^l \sim \pi_{z_i^{l+1}, p(i,l)}^l \quad i = 1 \ldots n \end{aligned} \tag{7}$$

where $p(i, l)$ is now the layer specific parent group. Note that κ again plays the role of a continuity parameter as for the ME-LaDP. When BE is used in every layer of LaDP as in Eqn. 7 we call the model BE-LaDP.

5 Inference Using LaDP

The inference problem in LaDP, given observations $\{w_i\}$, is to find posterior distributions over the group variables $\{z_i^l\}$ at all layers l for each data point.

As for models such as HDP, HDP-HMM and sHDP-HMM, exactly computing this posterior distribution is not tractable, and we resort to Gibbs Sampling for approximate inference as for the other models. One possibility is to perform collapsed Gibbs Sampling using only the group variables after integrating out all the parameter variables such as π_g^l and β_g^l. When the β_g^l variable takes the same value across groups in any layer l, the distribution of the variables at that layer is identical to the HDP. The predictive distribution of the z_i^l in that case is given by the CRF equations as for the HDP [5]. However, in cases where some of the β_g^l distributions are specified through domain knowledge, we integrate out only the π_g^l distributions.

Predictive Distributions: For the different LaDP models, we first derive the predictive distributions for z_i^l, the i^{th} group variable in the l^{th} layer, given the assignments to all group variables in the layers above (denoted $z^{>l}$), and the first $i-1$ group variables in layer l (denoted $z_{<i}^l$), after integrating out the $\pi_{g,g'}^l$ distributions from which they are drawn.

If the l^{th} layer uses CE (Eq. 3), the predictive distribution is given by

$$p(z_i^l = a | z_{<i}^l, z^{l+1}) \propto n^l_{z_i^{l+1},i,a} + \alpha^l \beta^l_{z_i^{l+1}}(a)$$

where $n^l_{j,i,a} = |\{t : z_t^l = a, z_t^{l+1} = j, t \in [1, i-1]\}|$. This is the number of data points before datapoint i in group j of layer $l+1$ were assigned to group a in layer l. If the l^{th} layer uses ME (Eq. 4), the predictive distribution becomes

$$p(z_i^l = a | z_{<i}^l, z^{l+1}) \propto n^l_{z_i^{l+1},i,p(i,l),a} + \kappa\delta(p(i,l),a) + \alpha^l \beta^l_{z_i^{l+1}}(a)$$

where $n^l_{j,i,b,a} = |\{t : z_t^l = a, p(t,l) = b, z_t^{l+1} = j, t \in [1, i-1]\}|$ is the number of times successive data points before datapoint i in group j of layer $l+1$ assigned to groups b and a respectively in layer l. For BE at layer l, we consider joint predictive distribution of z_i^l and the change-point indicator c_i^l at layer l. The conditional probabilities are as follows:

$$P(c_i^l = 0, z_i^l = p(i,l) | c_{<i}^l, z_{<i}^l, z^{l+1}) \propto a^l_{z_i^{l+1},i,0,p(i,l)} + \kappa$$
$$P(c_i^l = 1, z_i^l = k | c_{<i}^l, z_{<i}^l, z^{l+1}) \propto (a^l_{z_i^{l+1},i,1,p(i,l)} + 1)(v^l_{z_i^{l+1},i,1,k} + \kappa\beta^l_{z_i^{l+1}}(k))$$

where $a^l_{j,i,c,p} \equiv |\{t : c_t^l = c, p(t,l) = p, z_t^{l+1} = j, t \in [1, i-1]\}|$ is the number of times data points before datapoint i in group j of layer $l+1$ were assigned to group p in layer l, and the adjacent change-point value in group j is c; and $v^l_{j,i,c,a} = |\{t : c_t^l = c, z_t^l = a, z_t^{l+1} = j, t \in [1, i-1]\}|$ is the number of times data points before datapoint i in group j of layer $l+1$ were assigned to group a in layer l, and the change-point value at the same position is c.

Inference Using Gibbs Sampling: We sample each of the z_i^l variables conditioned on all the others sequentially in each iteration until convergence. In each iteration we traverse all group variables for one data point before moving to the next data point, and for a specific data point we traverse layers top down.

The conditional distribution is given by $p(z_i^l|z_{-i}^l, z^{l-1}, z^{l+1}) \propto p(z_i^l|z_{-i}^l, z^{l+1})$ $p(z^{l-1}|z^l)$. The second term can be computed using the chain rule and the predictive distributions described above: $p(z^{l-1}|z^l) = p(z_1^{l-1}|z_1^l) \prod_{i=2}^{n} p(z_i^{l-1}|z_{<i}^{l-1}, z_i^l)$. At layer $l = 1$ this is the likelihood of the data, conditioned on the table assignments of layer 1. The form of the first term depends on the exchangeability assumption.

If layer l uses CE the i^{th} variable can be swapped with the last to get

$$p(z_i^l = a|z_{-i}^l, z^{l+1}) \propto n^l_{-i,z_i^{l+1},a} + \alpha^l \beta^l_{z_i^{l+1}}(a)$$

where $n^l_{-i,j,a} = |\{t \neq i : z_t^l = a, z_t^{l+1} = j\}|$. Swapping is possible by CE property.

If layer l uses ME with sticky transitions, we make use of the conditional distribution for the sHDP-HMM [6] to get:

$$p(z_i^l = a|z_{-i}^l, z^{l+1}) = (\alpha^l \beta_j^l(a) + s(p(i,l),a) + \kappa\delta(p(i,l),a)) \times \frac{\alpha^l \beta_j^l(c(i,l)) + s(a,c(i,l)) + \kappa\delta(c(i,l),a) + \delta(c(i,l),a)\delta(p(i,l),a)}{\alpha^l + s(a,.) + \kappa + \delta(p(i,l),a)}$$

where $j = d_i^{l+1}$, $s(a,b) = |\{t : z_t^l = a, c(t,l) = b\}|$. $p(i,l)$ is as defined before Eqn. 4, and $c(i,l)$ is defined analogously with $i+1 \leq j \leq n$ instead of $1 \leq j \leq i-1$.

Recall the $p(i,l)$ and $c(i,l)$ notations defined for the z_i^l variables. To define the conditionals for BE, we need to extend these to have equivalent notations for the c_i^l variables. So, we will use $p^z(i,l)$ and $c^z(i,l)$ for the earlier definitions, and $p^c(i,l)$ and $c^c(i,l)$ for equivalent definitions using c_i^l instead of z_i^l. Then the joint conditional $p(c_i^l, z_i^l|c_{-i}^l, z_{-i}^l, z^{l+1})$ has the following cases (omitting conditioning variables for notational brevity): For $c^c(i,l) = 0$

$$\begin{aligned}
p(c_i^l = 1, z_i^l = c^z(i,l)|.) &= 1 && \text{if } p^z(i,l) \neq c^z(i,l)\\
p(c_i^l = 0, z_i^l = p^z(i,l)|.) &\propto (\kappa + n_2^{z,c}(p^z(i,l),0)) && \text{if } p^z(i,l) = c^z(i,l)\\
p(c_i^l = 1, z_i^l = p^z(i,l)|.) &\propto \frac{(n_1^{z,c}(p^z(i,l),1) + \alpha_j^l \beta_j^l(p^z(i,l)))}{(n_1^c(1)+\delta)} \times (1 + n_2^{z,c}(p^z(i,l),1))
\end{aligned}$$

For $(c^c(i,l) = 1)$

$$\begin{aligned}
p(c_i^l = 0, z_i^l = p^z(i,l)|.) &\propto \frac{(\kappa + n_2^{z,c}(p^z(i,l),0))}{(1 + \kappa + n_2^{z,c}(p^z(i,l),.))}\\
p(c_i^l = 1, z_i^l = b|.) &\propto \frac{(n_1^{z,c}(b,1) + \alpha_j^l \beta_j^l(b))}{(n_1^c(1) + \alpha_j^l)} \frac{(1 + n_2^{z,c}(b,1))}{(1 + \kappa + n_2^{z,c}(b,.))}
\end{aligned}$$

where $n_1^c(1)$ is the count for $c_t^l = 1$ where $t \neq i$, $n_1^{z,c}(u,k)$ for $z_t^l = u$ and $c_t^l = k$, for all $t(\neq i)$ satisfying $z_t^{l+1} = z_i^{l+1}$, $n_2^{z,c}(u,k)$ for $p^z(t,l) = u$ and $c_t^l = k$, for all $t(\neq i)$ satisfying $z_t^{l+1} = z_i^{l+1}$. The equations are modified appropriately for the first and last data points of each group.

Complexity of Inference: ME vs. BE: Consider the conditional distributions at any layer l. For ME, for each data point, we need to sample from a k dimensional multinomial, where k is the current number of unique group values at layer l. This leads to a complexity of $\mathcal{O}(nk)$. In case of BE, the variable c_i^l determines whether the i^{th} data point continues with the value of $p(i, l)$. When $c_i^l = 1$, we need to sample a new value of z_i^l from a k-dimensional multinomial. Hence, the complexity of each iteration of inference in BE is $\mathcal{O}(n + bk)$, where b is the number of data-points with $c_i^l = 1$. This can be significantly less than $\mathcal{O}(nk)$, particularly for high values of k. The value of b depends on κ in Eqn. 7.

Discussion: The number of unobserved variables in the model grows linearly with the number of layers, and dependencies also become more complex, leading to slower mixing of the Gibbs sampler. Our hypothesis is that the order in which the variables are sampled in each iteration of Gibbs sampling influences the rate of convergence. Specifically, updating directly dependent variables immediately after updating any variable may lead to faster convergence. Note that there are two kinds of the dependencies in the model. There are 'vertical dependencies' between group variables for the same data point across layers, and 'horizontal dependencies' between group variables of data points and their parents within a layer. Currently, our inference algorithm orders variable updates only based on vertical dependencies — we sample all the $\{z_i^l\}_{l=1}^L$ variables corresponding to the data point i, before moving to the next data point. Future work would include investigating other possibilities.

In hierarchical Bayesian non-parametric models, the conditional distributions of latent variables, given assignments to earlier ones are typically associated with restaurant analogies. For the LaDP, we may consider a hypothetical restaurant that has layers consisting of infinite number of tables, each layer possibly corresponding to one course in the menu. Each customer, unlike in a formal dinner, has to move from one layer to the next after each course. The restaurant has multiple entrances, corresponding to each input group, and in the first layer, each customer randomly chooses a table based on table assignments of previous customers who came in through the same entrance. After completing the i^{th} course, each customer randomly chooses a table for the next $(i + 1)^{th}$ course based on tables assigned to previous customers who shared his table in the i^{th} course. Clearly, the dependencies are more complex for ME and BE.

6 Experiments

In this section, we empirically evaluate our proposed models on two datasets for the tasks of document modeling and segmentation. We first check if learning multiple layers of grouping leads to better fit on held-out data, and also if the resultant simultaneous segmentation at multiple layers is better than single layer segmentation performed by models such as the sticky HDP-HMM. We also compare the performance using the proposed notion of BE and that using ME in terms of generalization ability, segmentation quality and execution time.

Datasets: Our first dataset is a set of semi-synthetic **News** transcripts. We crawled archived pages from 5 news websites (Yahoo! News, The Hindu, The Times of India, Deccan Herald, The Telegraph) for a 30 day period (April 1-30, 2012), where news articles for each day were arranged in sequence like news transcripts. We selected stories from 5 categories — politics, national affairs, international affairs, business and sports, to create one transcript for each day for each news source. This produced a dataset of $150 (30 \times 5)$ virtual news transcripts, consisting of 2600 individual news articles, spread over the 5 categories. From these, 60 transcripts were used for training and the rest for testing. After eliminating stop-words and rare words, we had a vocabulary of size 7204, with a total of 0.4 million tokens in the complete dataset. Our second dataset is on customer-generated laptop **Reviews** from Amazon.com. Here each document is a single review, consisting of parts discussing different product facets, like appearance, weight, screen size, image clarity, connectivity etc. The vocabulary size was 7147 and there were 1.5 million tokens in the entire dataset. We used 11510 documents for training, and 1000 for testing. In 100 of the test documents, we annotated the facet segments manually for use as gold standard segmentation for evaluation.[1]

Weak Supervision: Our models can accept weak supervision through the group-specific β_g^l base distributions at any layer l. In the news dataset we have gold-standard on the category labels. In the topmost layer L of any LaDP model, the groups correspond to news documents, each belonging to one of the 5 categories. For some of the models (as discussed later) the training documents j were provided supervision by setting β_j^L to a δ-distribution peaked at the label of the category. We do not have such unique labels for stories. Separately, we ran HDP in advance on the entire set of news articles (considering each article as a document) and manually selected 136 meaningful topics, which we used as β_g^0 (for $g = 1 \ldots 136$), which serve as base distributions for the stories. (Eqn. 4).

Evaluated Models: We evaluate models with different number of layers, different exchangeability properties at each layer, and with and without supervision at specific layers. We choose a naming convention that clearly identifies these choices. For example, the name ME_s^r-BE-CE_s-LaDP indicates that the model has 2 layers, with ME used at layer 2, BE at layer 1 and CE at layer 0 for words. The s subscripts indicate that supervision is used at layers 2 and 0. The r superscript indicates that the number of mixture components is restricted at layer 2, instead of an infinite mixture. All of our models use complete exchangeability at the layer of words, but we still include it in the name of the model, since we have the option of using supervision at that layer. In our experiments, we use 2 and 1 layer models (i.e. with $L = 2$ and $L = 1$). Note that CE-CE-LaDP is the same as HDP [5], ME-CE-LaDP as sHDP-HMM [6], and CE-CE-CE-LaDP as MLC-HDP [9].

[1] The data is available at
`http://clweb.csa.iisc.ernet.in/adway/ladp/data.tar.gz`

Performance Measures: We aim to evaluate generalization ability and segmentation of the models. To evaluate generalization ability, we measure *perplexity*(PP) [11] on test data: $exp(-\frac{\sum_d \log p(W_d)}{\sum_d N_d})$, where W_d are the words, and N_d the number of words in the d^{th} test document. A lower value of perplexity indicates better performance. For evaluating *segmentation*, we use the P_k measure [12], which is the probability that two tokens, k positions apart in the same document, are reported to be in different segments despite being in same gold-standard segment, or the other way round. Since different models perform well for different ranges of k, we report the average over three different values of k (short, medium and long), which we denote as $S2$ for layer 2 and $S1$ for layer 1. The performance of the proposed models involving ME or BE depends critically on the parameter κ (Eqns. 4 and 7). We tune these parameters for all models using a validation set of 5 transcripts to optimize performance.

Experiments on News: For the news dataset, we have gold standard segmentation at the level of categories as well as at the level of stories. We evaluate five versions of 2-layered LaDP ($L = 2$), considering layer 2 as categories and layer 1 as stories. The first four use combinations of BE and ME for layers 2 and 1, with the number of components restricted to 5 at $l = 2$. We test a version that uses CE at both layers, (MLC-HDP [9] model restricted at layer 2). For models with $L = 1$, we consider $l = 1$ to correspond to stories, leading to the models ME-CE-LaDP (sHDP-HMM) and BE-CE-LaDP, which can be evaluated for story segmentation ($S1$) and perplexity (PP). These models do not have a layer corresponding to categories. Alternatively, we could consider $l = 1$ to correspond to categories, with no layer for stories, leading to ME^r-CE-$LaDP$ and BE^r-CE-$LaDP$, and evaluate for category segmentation ($S2$). For $S2$ we use k values of 700(long), 200(medium), and 50(short), while for $S1$ we use 160(long), 50(medium) and 20(short), based on the typical lengths of category and story segments in the gold-standard. The results are shown in Table 1. We separately evaluate all these models with weak supervision at layers $l = 2$ and $l = 0$ as discussed. The results are shown in Table 2.

Table 1. Perplexity and Segmentation Error for News without supervision

Model	PP	$S2$	$S1$
CE-CE	5245	-	0.60
$CE^r - CE$	5969	0.69	-
ME-CE	3751	-	0.59
$ME^r - CE$	7204	**0.33**	-
BE-CE	3371	-	0.61
$BE^r - CE$	3975	0.69	-
CE^r-CE-CE	3656	0.68	0.61
ME^r-ME-CE	**3326**	0.45	0.53
ME^r-BE-CE	3856	0.68	0.61
BE^r-ME-CE	4475	0.49	0.42
BE^r-BE-CE	3713	0.45	**0.37**

Table 2. Perplexity and Segmentation Error for News with supervision

Model	PP	$S2$	$S1$
CE-CE_s	5309	-	0.60
$CE^r - CE_s$	6248	0.69	-
ME-CE_s	2763	-	0.59
$ME^r_s - CE_s$	7204	0.45	-
BE-CE_s	**2173**	-	0.59
$BE^r_s - CE_s$	2830	0.44	-
CE^r_s-CE-CE_s	3632	0.68	0.61
ME^r_s-ME-CE_s	2546	0.33	**0.42**
ME^r_s-BE-CE_s	2830	0.46	0.49
BE^r_s-ME-CE_s	3000	0.49	**0.42**
BE^r_s-BE-CE_s	3184	**0.28**	0.44

From the Tables 1 and 2, we first observe that supervision significantly improves performance of the models in terms of PP, and also often in terms of $S2$ or $S1$. Also, low perplexity and high segmentation errors for CE^r-CE-CE and CE^r_s-CE-CE_s confirm that capturing the sequential nature of the data is essential. More importantly, we can see that, in general, joint segmentation at two layers improves performance over independent segmentation at each layer. Only ME^r-CE in the unsupervised case for category segmentation performs better than two-layer models. Two-layer models are also better in general in terms of perplexity. Only BE-CE_s in the supervised case achieves better perplexity than two-layer models. Secondly, comparing BE and ME models, we see that the best perplexity is achieved by BE-CE_s, while BE^r-BE-CE has the best $S1$ among unsupervised models. Among supervised models, BE^r_s-BE-CE_s has the best $S2$, while BE^r_s-ME-CE_s (jointly) has the best $S1$. This improved performance using BE can be attributed to the fact that its Exchangeability Statistic S_B is simpler than that of ME (S_M), and so it has to learn fewer parameters.

Table 3. Perplexity and Segmentation Error for Reviews

Model	PP	$S2$	$S1$
CE-CE	703	-	0.49
ME-CE	399	-	0.46
BE-CE	258	-	0.4
CE-CE-CE	1786	0.50	0.50
ME-ME-CE	1549	0.49	**0.38**
ME-BE-CE	1136	0.41	0.44
BE-ME-CE	1058	0.46	0.43
BE-BE-CE	477	0.43	0.46
ME-CE-CE	742	**0.37**	0.50
BE-CE-CE	**184**	0.41	0.50
CE-ME-CE	1913	0.48	0.48
CE-BE-CE	1787	0.49	0.42

Table 4. BE-ME comparison results

	Review			News		
Model	IT	$S1$	PP	IT	$S1$	PP
ME	53	0.46	399	4	0.59	2763
BE4	5.8	0.43	245	1.0	0.59	2173
BE3	5.2	0.40	258	0.7	0.56	2996
BE2	3.1	0.39	431	0.2	0.32	8650
BE1	0.6	0.48	614	0.1	0.39	13839

Experiments on Reviews: Recall that documents in this dataset only require a single layer of segmentation. However, it is still meaningful to use 2-layer models, and then use either the first or the second layer segmentation. The corresponding measures of segmentation error are $S1$ and $S2$. We also evaluate single-layer models, with $S1$ being the segmentation error. For both $S1$ and $S2$, we consider k values 4 (short), 8 (medium) and 16 (long), again based on the typical lengths of segments (by facet) in the gold-standard. Since we did not have product facet labels for this dataset, we did not provide any supervision. As before, we used CE at layer 0 (words). Since segmentation is needed at only one layer in this case, we considered all combinations of ME, BE and CE at layers 2 and 1, leading to 3 1-layer models, and 9 2-layer models.

The results are shown in Table 3. We note that the baselines models HDP (CE-CE), sHDP-HMM (ME-CE) and MLC-HDP (CE-CE-CE) are outperformed

on all measures. More interestingly, though the data has segment information only at one layer, the best performance in terms of both PP and $S1$ is obtained by 2-layer models. Finally, BE performs well in terms of PP. Though BE does not outperform ME in terms of segmentation accuracy for this experiment, its usefulness becomes apparent in our next experiment.

Execution Time: Finally, we evaluate the effect of the continuity parameter κ (Eqn. 7) on the execution time and accuracy of Block Exchangeability. We compare the single-layer models BE-CE-LaDP and ME-CE-LaDP (sHMM-HDP) on News and Reviews in terms of inference time (per-iteration) (IT) (measured in seconds) during training, segmentation error $S1$ and perplexity PP using different values of κ. We consider 4 representative parameter settings for BE, denoted by BE1, BE2, BE3 and BE4 with κ values of 100000, 5000, 1000, 10 for news and 4000, 1500, 500, 300 for reviews. The results are shown in Table 4. To begin with, BE4 matches ME in segmentation and does better on perplexity while taking significantly less time (53 secs vs 5.8 secs for Reviews, 4 secs vs 1 sec for News). On decreasing κ, inference time reduces further with gradual degradation of perplexity, while average segmentation error decreases much below that of ME (for BE2) and then increases again. This happens because segmentation error for short k decreases monotonically with increase in block length i.e, decrease in κ, while that for long k increases monotonically. This demonstrates that using block exchangeability it is possible to trade off inference time for segmentation and modeling accuracy, unlike any existing exchangeability notion.

7 Conclusion

In this paper, we have addressed the problem of hierarchical segmentation of a collection of sequences, and proposed a Bayesian nonparametric model named the Layered Dirichlet Process, where data points filter down a layered structure of Dirichlet Processes, and get assigned to a group at every layer, depending on the exchangeability properties at that layer, leading to a hierarchical segmentation. We propose a new notion of exchangeability, that allows for more efficient inference compared to Markov exchangeability while enabling segmentation unlike complete exchangeability. We have demonstrated experimentally that using the proposed models joint segmentation at multiple layers is better than independent single-layer segmentation, and we are additionally able to trade off execution time for modeling and segmentation accuracy unlike any existing model.

References

1. Ferguson, T.: Bayesian analysis of some nonparametric problems. Annals of Statistics 1(2), 209–230 (1973)
2. Diaconis, P., Freedman, D.: De Finetti's generalizations of exchangeability. Studies in Inductive Logic and Probability 2, 233–249 (1980)
3. de Finetti, B.: Theory of probability, vol. 1-2 (1975)

4. Sethuraman, J.: A constructive definition of Dirichlet priors. Stat. Sinica 4, 639–650 (1994)
5. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.: Hierarchical Dirichlet Processes. Journal of American Statistics Association 101(476) (2006)
6. Fox, E.B., Sudderth, E.B., Jordan, M.I., Willsky, A.S.: An HDP-HMM for Systems with State Persistence. In: Intl. Conf. on Machine Learning, pp. 312–319 (2008)
7. Rodriguez, A., Dunson, D.B., Gelfand, A.E.: The nested Dirichlet process. Journal of the American Statistical Association 103(483), 1131–1154 (2008)
8. Blei, D.M., Griffiths, T.L., Jordan, M., Tanenbaum, J.B.: The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. Journal of the ACM 57(2) (2010)
9. Wulsin, D., Jensen, S., Litt, B.: A Hierarchical Dirichlet Process Model with Multiple Levels of Clustering for Human EEG Seizure Modeling. In: Intl. Conf. on Machine Learning (2012)
10. Kim, D., Kim, S., Oh, A.: Dirichlet Process with Mixed Random Measures: A Nonparametric Topic Model for Labeled Data. In: Intl. Conf. on Machine Learning (2012)
11. Blei, D.M., Ng, A.Y., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
12. Beeferman, D., Berger, A., Lafferty, J.: Statistical models for text segmentation. Machine Learning 34(1-3) (1999)
13. `http://clweb.csa.iisc.ernet.in/adway/ladp/appendix.pdf`