Hierarchical Dirichlet Scaling Process

Dongwoo Kim DW.KIM@KAIST.AC.KR
KAIST, Daejeon, Korea

Alice Oh ALICE.OH@KAIST.EDU
KAIST, Daejeon, Korea

Abstract

We present the *hierarchical Dirichlet scaling process* (HDSP), a Bayesian nonparametric mixed membership model for multi-labeled data. We construct the HDSP based on the gamma representation of the hierarchical Dirichlet process (HDP) which allows scaling the mixture components. With such construction, HDSP allocates a latent location to each label and mixture component in a space, and uses the distance between them to guide membership probabilities. We develop a variational Bayes algorithm for the approximate posterior inference of the HDSP. Through experiments on synthetic datasets as well as datasets of newswire, medical journal articles, and Wikipedia, we show that the HDSP results in better predictive performance than HDP, labeled LDA and partially labeled LDA.

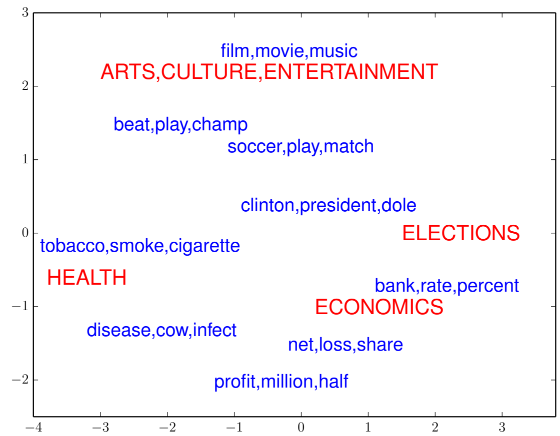


Figure 1. Locations of observed labels (capital letters in red) and latent topics (small letters in blue) inferred by HDSP from the Reuters corpus. HDSP uses the distances between labels and topics to scale the topic proportions such that the topics closer to the observed labels in a document are given higher probabilities.

1. Introduction

The Hierarchical Dirichlet process (HDP) is an important nonparametric Bayesian prior for mixed membership models, and the HDP topic model is useful for a wide variety of tasks involving unstructured text (Teh et al., 2006). To extend the HDP topic model, there has been active research in dependent random probability measures as priors for modeling the underlying association between the latent semantic structure and explanatory variables, such as time stamps and spatial coordinates (Ahmed & Xing, 2010; Ren et al., 2011).

A large body of this research is rooted in the dependent Dirichlet process (DP) (MacEachern, 1999) where the probabilistic random measure is defined as a function of some covariate. Most dependent DP approaches rely on the generalization of Sethuraman's stick breaking representation of DP (Sethuraman, 1991), incorporating the time difference between two or more data points or the spatial difference among observed data into the predictor dependent stick breaking process (Duan et al., 2007; Dunson & Park, 2008). Some of these priors can be integrated into the hierarchical construction of DP (Srebro & Roweis, 2005), resulting in topic models where temporally- or spatially-proximate data are more likely to be clustered.

Many datasets, however, come with labels, or categorical side information, which cannot be modeled with these existing dependent DP approaches. Labels, like temporal and spatial information, are correlated with the latent semantics of the documents, but they cannot be used to directly define the distance between two documents. This is because labels are categorical, so there is no simple way to measure distances between labels. Moreover, labels and documents do not have a one-to-one correspondence, as there may be zero, one, or more labels per document.

We develop the hierarchical Dirichlet scaling process

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

(HDPS) which models the latent locations of topics and labels in a space and uses the distances between them to guide the topic proportions. Figure 1 visualizes how the HDSP discovers the latent locations of the topics and the labels from the Reuters articles with news categories as labels. In this example, an article under the news category "ECONOMICS" would be endowed high probabilities for topics *<net, loss, share>* and *<bank, rate, percent>*, and low probabilities for topics *<beat, play, champ>* and *<film, movie, music>*.

In the next section, we describe the gamma process construction of the HDP and how the scale parameter is used to develop the HDSP. In section 3, we derive a variational inference for the latent variables by directly placing a prior over the distances between the latent locations. In section 4, we verify our approach on a synthetic dataset and demonstrate the improved predictive power of our model on multi- and partially-labeled corpora.

Related Work Previously proposed topic models for labeled documents take an approach quite distinct from the dependent DP literature. Labeled LDA (L-LDA) allocates one dimension of the topic simplex per label and generates words from only the topics that correspond to the labels in each document (Ramage et al., 2009). An extension of this model, partially labeled LDA (PLDA), adds more flexibility by allocating a pre-defined number of topics per label and including a background label to handle documents with no labels (Ramage et al., 2011). The Dirichlet process with mixed random measures (DP-MRM) is a nonparametric topic model which generates an unbounded number of topics per label but still excludes topics from labels that are not observed in the document (Kim et al., 2012).

2. Hierarchical Dirichlet Scaling Process

In this section, we describe the hierarchical Dirichlet scaling process (HDSP) for multi-labeled data. First we review the HDP and the gamma process construction of the second level DP. We then present the HDSP where the second level DP incorporates the latent locations for the mixture components and the labels.

2.1. The gamma process construction of HDP

In the HDP[1], there are two levels of the DP where the measure drawn from the upper level DP is the base distribution of the lower level DP. The hierarchical representation of the process is

$$G_0 \sim \mathrm{DP}(\alpha H), \qquad G_m \sim \mathrm{DP}(\beta G_0), \tag{1}$$

[1] In this paper, we limit our discussions of the HDP to the two level construction of the DP and refer to it simply as the HDP.

where H is a base distribution, α, and β are concentration parameters, and index m represents multiple draws from the second level DP. For the mixed membership model, x_{mn}, observation n in group m, can be drawn from

$$\theta_{mn} \sim G_m, \qquad \mathrm{x}_{mn} \sim f(\theta_{mn}), \tag{2}$$

where $f(\cdot)$ is a data distribution parameterized by θ. In the context of topic models, the base distribution H is usually a Dirichlet distribution over the vocabulary, so the atoms of the first level random measure G_0 are an infinite set of topics drawn from H. The second level random measure G_m is distributed based on the first level random measure G_0, so the second level shares the same set of topics, the atoms of the first level random measure.

The constructive definition of the DP can be represented as a stick breaking process (Sethuraman, 1991), and in the HDP inference algorithm based on stick breaking, the first level DP is given by the following conditional distributions:

$$V_k \sim \mathrm{Beta}(1, \alpha) \qquad p_k = V_k \prod_{j=1}^{j<k} (1 - V_j)$$
$$\phi_k \sim H \qquad G_0 = \sum_{k=1}^{\infty} p_k \delta_{\phi_k}, \tag{3}$$

where V_k defines a corpus level topic distribution for topic ϕ_k. The second level random measures are conditionally distributed on the first level discrete random measure G_0:

$$\pi_{ml} \sim \mathrm{Beta}(1, \beta) \qquad p_{ml} = \pi_{ml} \prod_{j=1}^{j<l} (1 - \pi_{mj})$$
$$\theta_{ml} \sim G_0 \qquad G_m = \sum_{l=1}^{\infty} p_{ml} \delta_{\theta_{ml}}, \tag{4}$$

where the second level atom θ_{ml} corresponds to one of the first level atoms ϕ_k. This stick breaking construction is the most widely used method for the hierarchical construction (Wang et al., 2011; Teh et al., 2006).

An alternative construction of the HDP is based on the normalized gamma process (Paisley et al., 2012). While the first level construction remains the same, the gamma process changes the second level construction from Eq. 4 to

$$\pi_{mk} \sim \mathrm{Gamma}(\beta p_k, 1)$$
$$G_m = \sum_{k=1}^{\infty} \frac{\pi_{mk}}{\sum_{j=1}^{\infty} \pi_{mj}} \delta \phi_k, \tag{5}$$

where $\mathrm{Gamma}(x; a, b) = b^a x^{(a-1)} e^{-bx} / \Gamma(a)$. Unlike the stick breaking construction, the atom of the π_{mk} of the gamma process is the same as the atom of the kth stick of the first level. Therefore, during inference, the model does not need to keep track of which second level atoms correspond to which first level atoms. Furthermore, by placing a proper random variable on the rate parameter of the gamma distribution, the model can infer the correlations among the topics (Paisley et al., 2012) through the Gaussian process (Rasmussen & Williams, 2005).

2.2. Hierarchical Dirichlet Scaling Process

In the hierarchical Dirichlet scaling process (HDSP), we start with the gamma process construction of the HDP with a proper prior for the rate parameter to guide the topic proportions based on the labels of the document. In the model, each topic and label has a latent location, and the topic proportion of a document is proportional to the distances between the topics and the labels. With the assumption that the locations of topics and labels are drawn from a distribution over the space, the first level DP of HDSP is drawn from the product of two base distributions,

$$G_0 \sim \text{DP}(\alpha H \otimes L), \qquad G_0 = \sum_{k=1}^{\infty} p_k \delta_{\{\phi_k, l_k\}} \tag{6}$$

where H is a distribution over the topic parameter θ_k, L is a distribution over the latent locations of topic k and label j, and p_k is a stick length for topic k, $p_k = V_k \prod_{k'=1}^{k'<k}(1 - V_{k'})$. There are J observable labels, and for each observable label j, a latent location l_j is drawn from the distribution L. Through the first level DP, the model draws an infinite number of topics ϕ_k as well as their corresponding locations l_k.

In the second level DP, the gamma process is used to incorporate the distances between the location of topics and observed labels. Let r_{mj} be an indicator variable, if the label j is observed in document m then r_{mj} is 1 otherwise 0. First, as in the HDP, draw a random measure $G'_m \sim \text{DP}(\beta G_0)$ for each document. Second, scale the topics based on the product of the inverse distances between topics and the observed labels of the document

$$G_m(\{\phi_k, l_k\}) \propto G'_m(\{\phi_k, l_k\}) \prod_{j=1}^{J} d(l_j, l_k)^{-r_{mj}}, \tag{7}$$

where $d(l_j, l_k)$ is a distance measure of the location of label l_j and the location of topic l_k. Through this process, topics that are closely located to the observed labels would have larger proportions in the final random measure G_m.

The constructive definition of HDSP is similar to the HDP, but the difference comes from the location variables and the distance terms. The first level stick breaking for HDSP is

$$V_k \sim \text{Beta}(1, \alpha) \qquad p_k = V_k \prod_{j=1}^{j<k}(1 - V_j)$$
$$\phi_k \sim H, \quad l_k \sim L \qquad G_0 = \sum_{k=1}^{\infty} p_k \delta_{\{\phi_k, l_k\}}. \tag{8}$$

Also, for each observable label, l_j is drawn i.i.d from L.

The second level gamma process for HDSP is

$$\pi_{mk} \sim \text{Gamma}(\beta p_k, \prod_{j=1}^{J} d(l_j, l_k)^{r_{mj}})$$
$$G_m = \sum_{k=1}^{\infty} \frac{\pi_{mk}}{\sum_{j=1}^{\infty} \pi_{mj}} \delta_{\phi_k}. \tag{9}$$

The distance terms are directly incorporated into the second parameter of the gamma distribution since the scaled gamma random variable $y = kx \sim \text{Gamma}(a, 1)$ is equal to $y \sim \text{Gamma}(a, k^{-1})$. For the mixed membership model, nth observation in mth group is drawn as follows:

$$\phi_k \sim G_m, \quad \text{x}_{mn} \sim f(\phi_k). \tag{10}$$

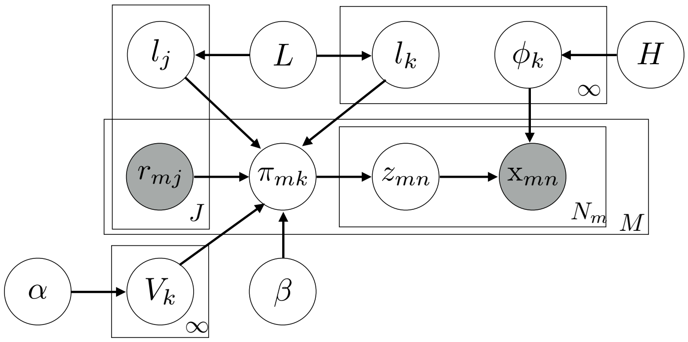
For topic modeling, G_m and x_{mn} correspond to document m and word n in document m, respectively.

3. Variational Inference for HDSP

The posterior inference for Bayesian nonparametric models is important because it is intractable to compute the posterior over an infinite dimensional space. Approximation algorithms, such as marginalized MCMC (Escobar & West, 1995; Teh et al., 2006) and variational inference (Blei & Jordan, 2006; Teh et al., 2008), have been developed for the Bayesian nonparametric mixture models. We develop a mean field variational inference (Jordan et al., 1999; Wainwright & Jordan, 2008) algorithm for approximate posterior inference of the HDSP topic model. The objective of variational inference is to minimize the KL divergence between a distribution over the hidden variables and the true posterior, which is equivalent to maximizing the lower bound of the marginal log likelihood of observed data.

For simple and efficient inference, we devise a simple model from the same perspective. Since we are interested in the distances between topics and labels, we directly place a prior over the distance between a topic and label $d(l_j, l_k)$. Let w_{jk} be the inverse distance between the latent location of label j and topic k, i.e. $w_{jk} = d(l_j, l_k)^{-1}$, we approximate w_{jk} by placing an inverse-Gamma prior over w_{jk}:

$$w_{jk} \sim \text{invGamma}(a^w, b^w) \tag{11}$$



Notations

α, β	: concentration parameters
H, L	: base distribution for topics and locations
l_k, l_j	: location of topics and labels
ϕ_k	: *k* th topic
V_k, π_{mk}	: first and second level topic proportion
z_{mn}	: topic of *n* th word in *m* th document
r_{mj}	: *j* th label of *m* th document {0, 1}
x_{mn}	: observed *n* th word in *m* th document

Figure 2. Graphical model of the hierarchical Dirichlet scaling process.

As a consequence, the second level gamma process can be rewritten from Equation 9 to

$$\pi_{mk} \sim \text{Gamma}(\beta p_k, \prod_{j=1}^{J} w_{jk}^{-r_{mj}}). \tag{12}$$

Excluding the locations from the model, it becomes unnecessary to place a proper distribution L over the space, in order to place a proper distance measure and to infer the latent locations of labels and topics. Hence, we can reduce the model complexity and derive a relatively simple and efficient inference algorithm, excluding the need for more complex metric learning problems (Xing et al., 2002) or the kernel based dependent DP construction (Dunson & Park, 2008; Ren et al., 2011). Now, we derive a variational Bayes inference using this approximation approach.

We use a fully factorized variational distribution and perform a mean-field variational inference. There are five latent variables of interest: the corpus level stick proportion V_k, the document level stick proportion π_{mk}, the inverse distance between topic and label w_{jk}, the topic assignment for each word z_{mn}, and the word topic distribution ϕ_k. Thus the variational distribution $q(z, \pi, V, w, \phi)$ can be factorized into

$$q(z, \pi, V, w, \phi) = \prod_{k=1}^{T} \prod_{m=1}^{M} \prod_{j=1}^{J} \prod_{n=1}^{N_m} q(z_{mn}) q(\pi_{mk}) q(V_k) q(\phi_k) q(w_{jk}), \tag{13}$$

where the variational distributions are

$$\begin{aligned} q(z_{mn}) &= \text{Multinomial}(z_{mn}|\gamma_{mn}) \\ q(\pi_{mk}) &= \text{Gamma}(\pi_{mk}|a^{\pi}_{mk}, b^{\pi}_{mk}) \\ q(w_{jk}) &= \text{InvGamma}(w_{jk}|a^{w}_{jk}, b^{w}_{jk}) \\ q(\phi_k) &= \text{Dirichlet}(\phi_k|\eta_k) \\ q(V_k) &= \delta_{V_k}. \end{aligned}$$

For the corpus level stick proportion V_k, we use the delta function as a variational distribution for simplicity and tractability in inference steps (Liang et al., 2007). Infinite dimensions over the posterior is a key problem in Bayesian nonparametric models and requires an approximation method. In variational treatment, we truncate the unbounded dimensionality to T by letting $V_T = 1$. Thus the model still keeps the infinite dimensionality while allowing approximation to be carried out under the bounded variational distributions.

Using standard variational theory, we derive the lower bound of the marginal log likelihood of the observed data $\mathcal{D} = (\mathbf{x}_m, \mathbf{r}_m)_{m=1}^{M}$,

$$\begin{aligned} &\log p(\mathcal{D}|\alpha, \beta, a^w, b^w, \eta) \\ &\quad \geq \mathbb{E}_q[\log p(\mathcal{D}, z, \pi, V, w, \phi)] + H(q) = \mathcal{L}(q), \end{aligned} \tag{14}$$

where $H(q)$ is the entropy for the variational distribution. By taking the derivative of this lower bound, we derive the following coordinate ascent algorithm.

Document-level Updates: At the document level, we update the variational distribution for the topic assignment z_{mn} and the document level stick proportion π_{mk}. The update for $q(z_{mn}|\gamma_{mn})$ is

$$\gamma_{mnk} \propto \exp\left(\mathbb{E}_q[\ln \eta_{k,x_{mn}}] + \mathbb{E}_q[\ln \pi_{mk}]\right). \tag{15}$$

Updating $q(\pi_{mk}|a^{\pi}_{mk}, b^{\pi}_{mk})$ requires computing the expectation term $\mathbb{E}[\ln \sum_{k=1}^{T} \pi_{mk}]$. Following (Blei & Lafferty, 2007), we approximate the lower bound of the expectation by using the first-order Taylor expansion,

$$-\mathbb{E}_q[\ln \sum_{k=1}^{T} \pi_{mk}] \geq -\ln \xi_m - \frac{\sum_{k=1}^{T} \mathbb{E}_q[\pi_{mk}] - \xi_m}{\xi_m}, \tag{16}$$

where the update for $\xi = \sum_{k=1}^{K} \mathbb{E}_q[\pi_{mk}]$. Then, the update for π_{mk} is

$$a^{\pi}_{mk} = \beta p_k + \sum_{n=1}^{N_m} \gamma_{mnk}$$
$$b^{\pi}_{mk} = \prod_j \mathbb{E}_q[w_{jk}^{-r_{mj}}] + \frac{N_m}{\xi_m}. \quad (17)$$

Note again r_{mj} is equal to 1 when jth label is observed in mth document, otherwise 0.

Corpus-level Updates: At the corpus level, we update the variational distribution for the inverse distance w_{jk}, corpus level stick length V_k and word topic distribution η_{ki}.

The optimal form of a variational distribution can be obtained by exponentiating the variational lower bound with all expectations except the parameter of interest (Bishop & Nasrabadi, 2006). For w_{jk}, we can derive the optimal form of variational distribution as follows

$$q(w_{jk}) \sim \text{InvGamma}(a', b') \quad (18)$$
$$a' = \mathbb{E}_q[\beta p_k] \sum_m r_{mj} + a^w$$
$$b' = \sum_{m'} \prod_{j'/j} \mathbb{E}_q[w^{-1}_{j'k}] \mathbb{E}_q[\pi_{m'k}] + b^w,$$

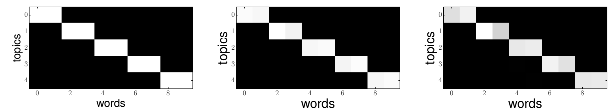
where $m' = \{m : r_{mj} = 1\}$ and $j'/j = \{j' : r_{mj'} = 1, j' \neq j\}$. See the appendix for the complete derivation. There is no closed form update for V_k, instead we use the steepest ascent algorithm to jointly optimize V_k. The gradient of V_k is

$$\frac{\partial \mathcal{L}}{\partial V_k} = -\frac{\alpha - 1}{1 - V_k} \quad (19)$$
$$- \frac{\beta p_k}{V_k} \{ \sum_{m,j} r_{mj} \mathbb{E}_q[\ln \pi_{mk}] - \mathbb{E}_q[\pi_{mk}] + \psi(\beta p_k) \}$$
$$+ \sum_{k'>k} \frac{\beta p_{k'}}{1 - V_k} \{ \sum_{m,j} r_{mj} \mathbb{E}_q[\ln \pi_{mk'}] - \mathbb{E}_q[\pi_{mk'}] + \psi(\beta p_{k'}) \},$$

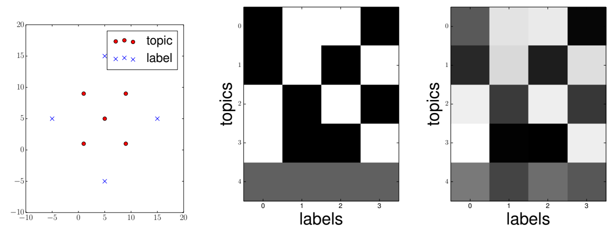
where $\psi(\cdot)$ is a digamma function. Finally, the update for the word topic distribution $q(\phi_k | \eta_k)$ is

$$\eta_{ki} = \eta + \sum_{m,n} \gamma_{mnk} \mathbf{1}(\mathrm{x}_{mn} = i), \quad (20)$$

where i is a word index, and $\mathbf{1}$ is an indicator function (Blei et al., 2003).



(a) Synthetic Topics (b) HDSP Topics (c) HDP Topics



(d) Latent Locations (e) Syn Distances (f) HDSP Distances

Figure 3. Experiments with synthetic data. (a) is the synthetic topic distribution of 5 topics over 10 terms. (b) and (c) are topic distributions inferred by the HDSP and the HDP. Both models recover the original topics. (d) shows the original locations of topics and labels. (e) shows the original distances between the locations of topics and labels. (f) shows that the HDSP recovers the distances. Note that the HDP cannot compute the distances.

The expectations under the variational distribution q are

$$\mathbb{E}_q[\pi_{mk}] = a^{\pi}_{mk} / b^{\pi}_{mk}$$
$$\mathbb{E}_q[\ln \pi_{mk}] = \psi(a^{\pi}_{mk}) - \ln b^{\pi}_{mk}$$
$$\mathbb{E}_q[w_{jk}] = b^{w}_{jk} / (a^{w}_{jk} - 1)$$
$$\mathbb{E}_q[w^{-1}_{jk}] = a^{w}_{jk} / b^{w}_{jk}$$
$$\mathbb{E}_q[\ln w_{jk}] = \ln b^{w}_{jk} - \psi(a^{w}_{jk})$$
$$\mathbb{E}_q[\ln \phi_{ki}] = \psi(\eta_{ki}) - \psi(\sum_i \eta_{ki}).$$

4. Experiments

We train the HDSP topic model with synthetic and real data to verify the model and show the advantages over existing models.

4.1. Synthetic data

There is no naturally-occurring dataset with observable locations of topics and labels, so we synthesize data based on the model assumptions to verify our model and the approximate inference. First, we check the difference between the original topics and the inferred topics via simple visualization. Then, we focus on the inferred locations and the original locations. For all experiments with synthetic data, we set the truncation level T at twice the number of topics. We terminate variational inference when the fractional change of the lower bound falls below 10^{-3}, and we average all results over 10 individual runs with different initializations.

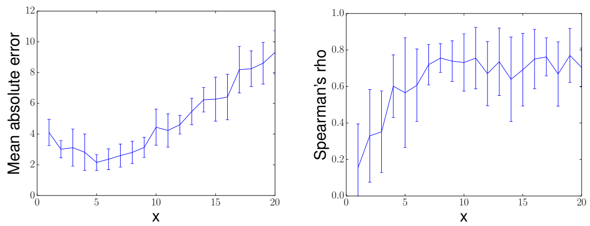


Figure 4. Spearman's correlation coefficient and mean absolute error of the synthetic data with various volume of space (x^3). As the volume of space for locations increases, the mean absolute error also increases (left). However, the model preserves the relative distances between topics and labels, shown by the high and stabilized correlation between the original ordering and the recovered ordering of label-topic pairs in terms of the distance between the two (right). This is a key characteristic of the HDSP model which scales the mixture components according to the inverse of the distance.

With the first experiment, we show that HDSP correctly recovers the underlying topics and distances between topics and labels. For the dataset, we generate 500 documents using the following steps. We define five topics over ten terms shown in Figure 3(a) and locations of five topics and four labels shown in Figure 3(d). For each document, we randomly draw N_m from the Poisson distribution and r_{mj} from the Bernoulli distribution. The average length of a document is 20, and the average number of labels per document is 2. We generate topic proportions of corpus and documents by using Equations 8 and 9. For each word in a document, we draw the topic and the word by using Equation 10. We set both α and β to 1.

Figure 3 shows the results of the HDP and the HDSP on the synthetic dataset. Figure 3(b) and Figure 3(c) are the heat maps of topics inferred from each model. We match the inferred topics to the original topics using KL divergence between the two topic distributions. There are no significant differences between the inferred topics of HDSP and HDP. In addition to the topics, HDSP infers the distances between topics and labels, which are shown in Figure 3(f).

With the second experiment, we show that the distances that HDSP infers preserve the relative distances between labels and topics in the dataset. Recall that HDSP does not directly infer the latent locations but instead infers the distances between labels and topics, which are then used to scale the topic proportions.

For this experiment, we generate 1,000 documents with randomly drawn 10 topics from Dirichlet(0.1) with the vocabulary size of 20. The locations of topics and labels are drawn from Uniform$(0, x)$ varying the x value from 1 to 20 for each experiment. We set the dimensionality of locations to 3, thus the volume of space is x^3. We compute the

Table 1. Datasets used for the experiments in 4.2. As the last two columns show, we experiment on datasets with a varied number of unique labels, as well as the average number of labels per document, including the Wikipedia corpus with many documents that are unlabeled.

	docs	vocab	labels	labels/doc
Wikipedia	25,547	7,702	1093	0.6
RCV	23,149	9,911	117	3.2
OHSUMED	7,505	7,056	52	5.2

mean absolute error (MAE) and the spearman's rank correlation coefficient (rho) between the original distances and the inferred distances. The spearman's rho is designed to measure the ranking correlation of two lists.

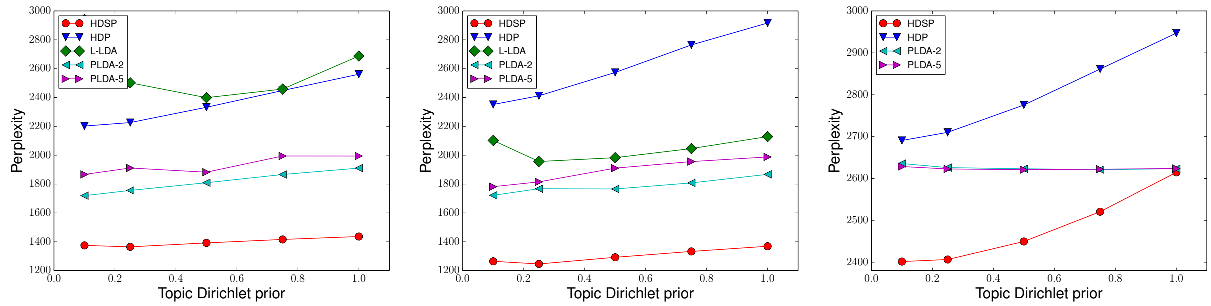
Figure 4 shows the results. The MAE increases as the volume of the space increases. However, spearman's rho stabilizes, indicating that the relative distances are preserved even when the MAE increases. Since there are an infinite number of configurations of distances that generate the same expectation $E[p(\pi_m|\beta p, w_j)]$ given π_m and βp, preserving the relative distances verifies our model's capability of capturing the underlying structure of topics and labels.

4.2. Real data

We evaluate the performance of HDSP and compare it with the HDP, labeled LDA (L-LDA) and partially labeled LDA (PLDA). The L-LDA defines a one-to-one correspondence between latent topics and labels. We use two multi-labeled corpora, RCV [2], newswire from Reuter's, and OHSUMED [3], a subset of the Medline journal articles, and one partially labeled corpus, Wikipedia.

Experimental Settings: For the HDP and HDSP, we initialize the word-topic distribution with three iterations of LDA for fast convergence to the posterior while preventing the posterior from falling into a local mode of LDA, then reorder these topics by the size of the posterior word count. For all experiments, we set the truncation level T to 200. We terminate variational inference when the fractional change of the lower bound falls below 10^{-3}, and we optimize all hyper parameters during inference except η. For the L-LDA and PLDA, we implement the collapsed Gibbs sampling algorithm. For each model, we run 5,000 iterations, the first 3,000 as burn-in and then using the samples thereafter with gaps of 100 iterations. For PLDA, we set the number of topics for each label to two and five (PLDA-2, PLDA-5). We try five different values for the topic Dirichlet parameter η: $\eta = 0.1, 0.25, 0.5, 0.75, 1.0$. Finally all results are averaged over 20 runs with different random initialization. We do not report the standard errors because

[2] http://trec.nist.gov/data/reuters/reuters.html
[3] http://ir.ohsu.edu/ohsumed/ohsumed.html



(a) OHSUMED (b) RCV (c) Wikipedia

Figure 5. Perplexity of held-out documents. For HDSP, L-LDA and PLDA, the perplexity is measured given documents and observed labels. For HDP, the model only uses the words of the documents. The HDSP which, instead of excluding topics from unobserved labels, scales all topics according to distances to observed labels, shows the best heldout perplexity.

they are small enough to ignore.

Evaluation Metric: The goal of our model is to construct the dependent random probability measure given labels. Therefore, our interest is to see the increments of predictive performance when the label information is given.

The predictive probability given label information for held-out documents are approximated by the conditional marginal,

$$p(\mathbf{x}'|\mathbf{r}', \mathcal{D}_{\text{train}}) = \tag{21}$$
$$\int_q \prod_{n=1}^{N} \sum_{k=1}^{T} p(x'_n|\phi_k)p(z'_n = k|\pi')p(\pi'|V, \mathbf{r}')dq(V, w, \phi),$$

where $\mathcal{D}_{\text{train}} = \{\mathbf{x}_{\text{train}}, \mathbf{r}_{\text{train}}\}$ is the training data, $\mathbf{x}'$ is the vector of N words of a held-out document, $\mathbf{r}'$ are the labels of the held-out document, z'_n is the latent topic of word n, and π'_k is the kth topic proportion of the held-out document. Since the integral is intractable, we approximate the probability

$$p(\mathbf{x}'|\mathbf{r}', \mathcal{D}_{\text{train}}) \approx \prod_{n=1}^{N} \sum_{k=1}^{T} \tilde{\pi}_k \tilde{\phi}_{k,x'_n}, \tag{22}$$

where $\tilde{\phi}_k$ and $\tilde{\pi}_k$ are the variational expectation for ϕ_k and π_k given label $\mathbf{r}'$. This approximated likelihood is then used to compute the perplexity of the held-out document

$$\text{perplexity} = \exp\left\{\frac{-\ln p(\mathbf{x}'|\mathbf{r}', \mathcal{D}_{\text{train}})}{N}\right\}. \tag{23}$$

Lower perplexity indicates better performance. We also take the same approach to compute the perplexity for L-LDA, PLDA and HDP, but HDP does not use the labels of held-out documents. To measure the predictive performance, we leave 20% of the documents for testing and use the remaining 80% to train the models.

Multi-labeled Data: We use two multi-labeled corpora, the RCV Reuters news data and the OHSUMED medical journal data. Both are multi-labeled, and every document has at least one label. The average number of labels per article is 3.2 for RCV and 5.2 for OHSUMED. Table 1 contains the details of the datasets.

The HDSP outperforms all comparison models, HDP, L-LDA and PLDA, in terms of perplexity as shown in Figure 5. For the OHSUMED data, the performance of L-LDA is worse than the HDP even though L-LDA is trained with label information. PLDA, which relaxes the assumption of L-LDA by adding an additional latent label and allowing multiple topics per label, outperforms the HDP and the L-LDA. But it excludes the topics of unobserved labels from modeling the document, and it performs worse than the HDSP. Note that the HDSP also relies on the observed labels to strongly guide the topics, but it still allows all topics to be used, even ones that are not closely located to the observed labels.

Figure 6 shows the expected topic distributions given different sets of labels. As we discussed in Section 2, the scaling effect yields more sharpened distribution given a set of labels. When multiple labels are given, the model expects high probabilities for the topics that are closely located to all given labels. The Appendix provides more examples from the other corpora as well as posterior topic count analysis.

To visualize the latent locations, we embed the inferred topics and the given labels into the two dimensional space by using multidimensional scaling (MDS) on the inferred distances (Kruskal, 1964). In Figure 1, we choose and display a few representative topics and labels.

Partially-labeled Data: We also test our model with partially labeled data which have not been explicitly covered in topic modeling. Many real-world data fall into this cate-

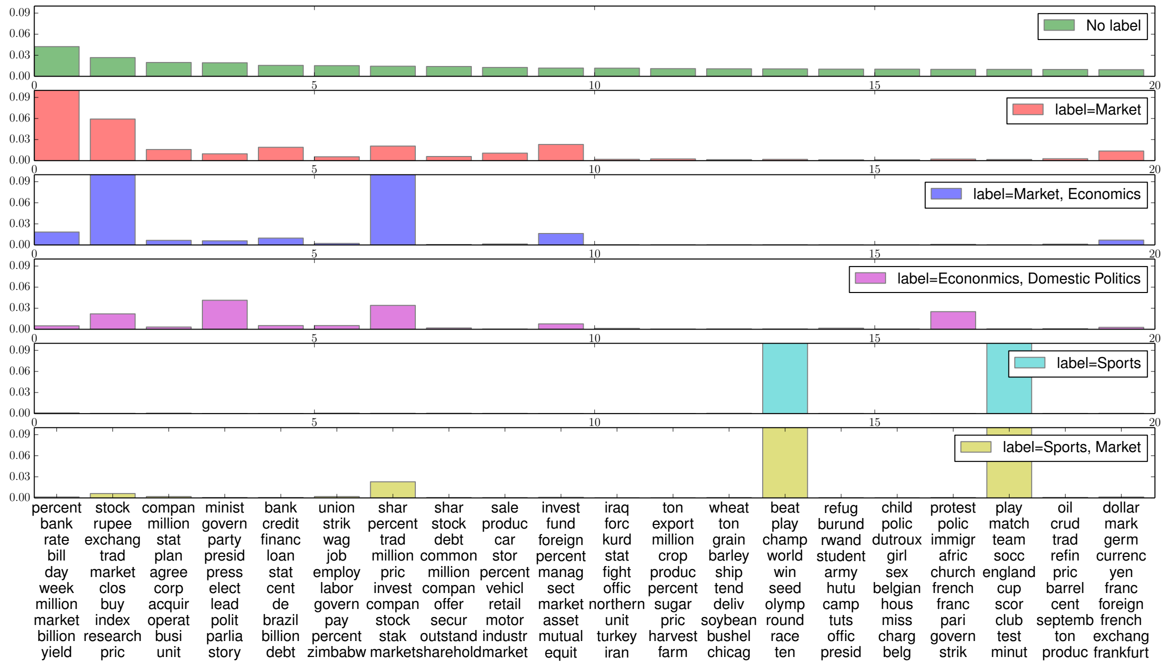


Figure 6. Expected topic distributions given labels from RCV. Topics are sorted by their posterior word counts, and the top 20 topics are displayed with the top 10 words (stemmed). From top to bottom, we compute an expected topic distribution given a set of labels.

Table 2. The most closely located category-topic pairs from Wikipedia. The categories are more specific and narrow than those of RCV.

Labels	Top words							
rivers of romania	river	water	area	bay	north	creek	valley	south
cities in iran	district	population	province	village	county	rural	area	town
italian painters	van	dutch	painter	italian	born	netherlands	portuguese	portrait
2009 albums	released	song	music	band	single	track	records	songs
public high schools	school	high	students	schools	college	education	state	girls

gory where some of the data are labeled, others are incompletely labeled, and the rest are unlabeled. For this experiment, we randomly sampled Wikipedia articles and use the categories as labels. Because most of the categories only appear once or twice in the dataset, we remove categories that appear in fewer than five articles, and our dataset contains 25,547 articles, 1,093 labels, with 0.6 labels per article on average. The label information is sparse even after removing the infrequent labels.

Figure 5(c) shows the predictive perplexity with Wikipedia. We compare the result with PLDA and HDP. We exclude L-LDA which cannot be trained on a dataset containing documents with no labels. HDSP outperforms both HDP and PLDA. We list the top ten most closely located label-topic pairs in Table 2.

5. Conclusion

We have presented the hierarchical Dirichlet scaling process (HDSP), a Bayesian nonparametric prior for a mixed membership model where categorical side information plays an important role. We discussed how the HDSP models the latent locations of the mixture components and the observed labels, and boosts the membership probability of a mixture component based on the product of the inverse distances to the labels. We showed that the application of HDSP to topic modeling correctly recovers the topics and label-topic distances of synthetic data. Furthermore, we showed the improved predictive performance of the HDSP topic model compared to the HDP, labeled LDA and partially labeled LDA. Future work on this research will explore kernel functions instead of simple products and applications of the HDSP topic model on various text mining problems.

6. Acknowledgement

This work was supported by the IT R&D program of MSIP/KEIT. [10041313, UX-oriented Mobile SW Platform]

References

Ahmed, Amr and Xing, Eric P. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 20–29, 2010.

Bishop, Christopher M and Nasrabadi, Nasser M. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

Blei, D, Ng, A, and Jordan, M. Latent dirichlet allocation. *The Journal of Machine Learning Research*, pp. 993–1022, Jan 2003.

Blei, David M and Lafferty, John D. A correlated topic model of science. *The Annals of Applied Statistics*, pp. 17–35, 2007.

Blei, D.M. and Jordan, M.I. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

Duan, Jason A, Guindani, Michele, and Gelfand, Alan E. Generalized spatial dirichlet process models. *Biometrika*, 94(4):809–825, 2007.

Dunson, David B and Park, Ju-Hyun. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

Escobar, M.D. and West, M. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, pp. 577–588, 1995.

Jordan, Michael I., Ghahramani, Zoubin, Jaakkola, Tommi S., and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, November 1999. ISSN 0885-6125.

Kim, Dongwoo, Kim, Suin, and Oh, Alice. Dirichlet process with mixed random measures: a nonparametric topic model for labeled data. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

Kruskal, Joseph B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Liang, Percy, Petrov, Slav, Jordan, Michael I, and Klein, Dan. The infinite pcfg using hierarchical dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 688–697, 2007.

MacEachern, S.N. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55, 1999.

Paisley, John, Wang, Chong, and Blei, David M. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(4):997–1034, 2012.

Ramage, D., Hall, D., Nallapati, R., and Manning, C.D. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 248–256, 2009.

Ramage, Daniel, Manning, Christopher D., and Dumais, Susan. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 457–465, New York, NY, USA, 2011.

Rasmussen, Carl Edward and Williams, Christopher K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Ren, Lu, Du, Lan, Carin, Lawrence, and Dunson, David. Logistic stick-breaking process. *The Journal of Machine Learning Research*, 12:203–239, 2011.

Sethuraman, J. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1991.

Srebro, Nathan and Roweis, Sam. Time-varying topic models using dependent dirichlet processes. *UTML, TR# 2005*, 3, 2005.

Teh, Y., Jordan, M., Beal, M., and Blei, D. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, Jan 2006.

Teh, Y., Kurihara, K., and Welling, M. Collapsed variational inference for HDP. *Advances in Neural Information Processsing Systems (NIPS)*, 20, 2008.

Wainwright, Martin J and Jordan, Michael I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Wang, Chong, Paisley, John W, and Blei, David M. Online variational inference for the hierarchical dirichlet process. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 752–760, 2011.

Xing, Eric P, Jordan, Michael I, Russell, Stuart, and Ng, Andrew. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems (NIPS)*, pp. 505–512, 2002.