# Dirichlet Enhanced Latent Semantic Analysis

**Kai Yu**
Siemens Corporate Technology
D-81730 Munich, Germany
*Kai.Yu@siemens.com*

**Shipeng Yu**
Institute for Computer Science
University of Munich
D-80538 Munich, Germany
*spyu@dbs.informatik.uni-muenchen.de*

**Volker Tresp**
Siemens Corporate Technology
D-81730 Munich, Germany
*Volker.Tresp@siemens.com*

## Abstract

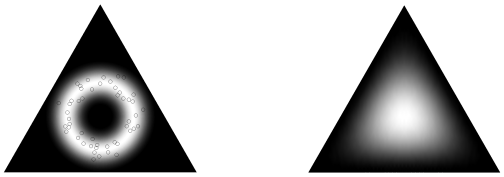
This paper describes nonparametric Bayesian treatments for analyzing records containing occurrences of items. The introduced model retains the strength of previous approaches that explore the latent factors of each record (e.g. topics of documents), and further uncovers the clustering structure of records, which reflects the statistical dependencies of the latent factors. The nonparametric model induced by a *Dirichlet process* (DP) flexibly adapts model complexity to reveal the clustering structure of the data. To avoid the problems of dealing with infinite dimensions, we further replace the DP prior by a simpler alternative, namely *Dirichlet-multinomial allocation* (DMA), which maintains the main modelling properties of the DP. Instead of relying on Markov chain Monte Carlo (MCMC) for inference, this paper applies efficient variational inference based on DMA. The proposed approach yields encouraging empirical results on both a toy problem and text data. The results show that the proposed algorithm uncovers not only the latent factors, but also the clustering structure.

## 1 Introduction

We consider the problem of modelling a large corpus of high-dimensional discrete records. Our assumption is that a record can be modelled by latent factors which account for the co-occurrence of items in a record. To ground the discussion, in the following we will identify records with documents, latent factors with (latent) topics and items with words. Probabilistic latent semantic indexing (PLSI) [7] was one of the first approaches that provided a probabilistic approach towards modelling text documents as being composed of latent topics. Latent Dirichlet allocation (LDA) [3] generalizes PLSI by treating the topic mixture parameters (i.e. a multinomial over topics) as variables drawn from a Dirichlet distribution. Its Bayesian treatment avoids overfitting and the model is generalizable to new data (the latter is problematic for PLSI). However, the parametric Dirichlet distribution can be a limitation in applications which exhibit a richer structure. As an illustration, consider Fig. 1 (a) that shows the empirical distribution of three topics. We see that the probability that all three topics are present in a document (corresponding to the center of the plot) is near zero. In contrast, a Dirichlet distribution fitted to the data (Fig. 1 (b)) would predict the highest probability density for exactly that case. The reason is the limiting expressiveness of a simple Dirichlet distribution.

This paper employs a more general nonparametric Bayesian approach to explore not only latent topics and their probabilities, but also complex dependencies between latent topics which might, for example, be expressed as a complex clustering structure. The key innovation is to replace the parametric Dirichlet prior distribution in LDA by a flexible nonparametric distribution $G(\cdot)$ that is a sample generated from a *Dirichlet process* (DP) or its finite approximation, *Dirichlet-multinomial allocation* (DMA). The Dirichlet distribution of LDA becomes the base distribution for the Dirichlet process. In this *Dirichlet enhanced* model, the posterior distribution of the topic mixture for a new document converges to a flexible mixture model in which both mixture weights and mixture parameters can be learned from the data. Thus the *a posteriori* distribution is able to represent the distribution of topics more truthfully. After convergence of the learning procedure, typically only a few components with non-negligible weights remain; thus the model is able to naturally output clusters of documents.

Nonparametric Bayesian modelling has attracted considerable attentions from the learning community



(a) (b)

Figure 1: Consider a 2-dimensional simplex representing 3 topics (recall that the probabilities have to sum to one): (a) We see the probability distribution of topics in documents which forms a ring-like distribution. Dark color indicates low density; (b) The 3-dimensional Dirichlet distribution that maximizes the likelihood of samples.

(e.g. [1, 13, 2, 15, 17]). A potential problem with this class of models is that inference typically relies on MCMC approximations, which might be prohibitively slow in dealing with the large collection of documents in our setting. Instead, we tackle the problem by a less expensive variational mean-field inference based on the DMA model. The resultant updates turn out to be quite interpretable. Finally we observed very good empirical performance of the proposed algorithm in both toy data and textual document, especially in the latter case, where meaningful clusters are discovered.

This paper is organized as follows. The next section introduces Dirichlet enhanced latent semantic analysis. In Section 3 we present inference and learning algorithms based on a variational approximation. Section 4 presents experimental results using a toy data set and two document data sets. In Section 5 we present conclusions.

## 2 Dirichlet Enhanced Latent Semantic Analysis

Following the notation in [3], we consider a corpus $\mathcal{D}$ containing $D$ documents. Each document $d$ is a sequence of $N_d$ words that is denoted by $\mathbf{w}_d = \{w_{d,1}, \ldots, w_{d,N_d}\}$, where $w_{d,n}$ is a variable for the $n$-th word in $\mathbf{w}_d$ and denotes the index of the corresponding word in a vocabulary $V$. Note that a same word may occur several times in the sequence $\mathbf{w}_d$.

### 2.1 The Proposed Model

We assume that each document is a mixture of $k$ latent topics and words in each document are generated by repeatedly sampling topics and words using the distributions

$$w_{d,n}|z_{d,n};\beta \sim \text{Mult}(z_{d,n},\beta) \quad (1)$$
$$z_{d,n}|\theta_d \sim \text{Mult}(\theta_d). \quad (2)$$

$w_{d,n}$ is generated given its latent topic $z_{d,n}$, which takes value $\{1,\ldots,k\}$. $\beta$ is a $k \times |V|$ multinomial parameter matrix, $\sum_j \beta_{i,j} = 1$, where $\beta_{z,w_{d,n}}$ specifies the probability of generating word $w_{d,n}$ given topic $z$. $\theta_d$ denotes the parameters of a multinomial distribution of document $d$ over topics for $\mathbf{w}_d$, satisfying $\theta_{d,i} \geq 0, \sum_{i=1}^{k} \theta_{d,i} = 1$.

In the LDA model, $\theta_d$ is generated from a $k$-dimensional Dirichlet distribution $G_0(\theta) = \text{Dir}(\theta|\lambda)$ with parameter $\lambda \in \mathbb{R}^{k\times 1}$. In our Dirichlet enhanced model, we assume that $\theta_d$ is generated from distribution $G(\theta)$, which itself is a random sample generated from a *Dirichlet process* (DP) [5]

$$G|G_0,\alpha_0 \sim \text{DP}(G_0,\alpha_0), \quad (3)$$

where nonnegative scalar $\alpha_0$ is the *precision parameter*, and $G_0(\theta)$ is the *base distribution*, which is identical to the Dirichlet distribution. It turns out that the distribution $G(\theta)$ sampled from a DP can be written as

$$G(\cdot) = \sum_{l=1}^{\infty} \pi_l \delta_{\theta_l^*}(\cdot) \quad (4)$$

where $\pi_l \geq 0$, $\sum_l^{\infty} \pi_l = 1$, $\delta_\theta(\cdot)$ are point mass distributions concentrated at $\theta$, and $\theta_l^*$ are countably infinite variables i.i.d. sampled from $G_0$ [14]. The probability weights $\pi_l$ are solely depending on $\alpha_0$ via a *stick-breaking process*, which is defined in the next subsection. The generative model summarized by Fig. 2(a) is conditioned on $(k \times |V| + k + 1)$ parameters, i.e. $\beta$, $\lambda$ and $\alpha_0$.

Finally the likelihood of the collection $\mathcal{D}$ is given by

$$\mathcal{L}_{\text{DP}}(\mathcal{D}|\alpha_0,\lambda,\beta) = \int_G \left\{ p(G;\alpha_0,\lambda) \prod_{d=1}^{D} \int_{\theta_d} \left[ p(\theta_d|G) \prod_{n=1}^{N_d} \sum_{z_{d,n}=1}^{k} p(w_{d,n}|z_{d,n};\beta) p(z_{d,n}|\theta_d) \right] d\theta_d \right\} dG. \quad (5)$$

In short, $G$ is sampled once for the whole corpus $\mathcal{D}$, $\theta_d$ is sampled once for each document $d$, and topic $z_{d,n}$ sampled once for the $n$-th word $w_{d,n}$ in $d$.

### 2.2 Stick Breaking and Dirichlet Enhancing

The representation of a sample from the DP-prior in Eq. (4) is generated in the stick breaking process in which infinite number of pairs $(\pi_l, \theta_l^*)$ are generated.

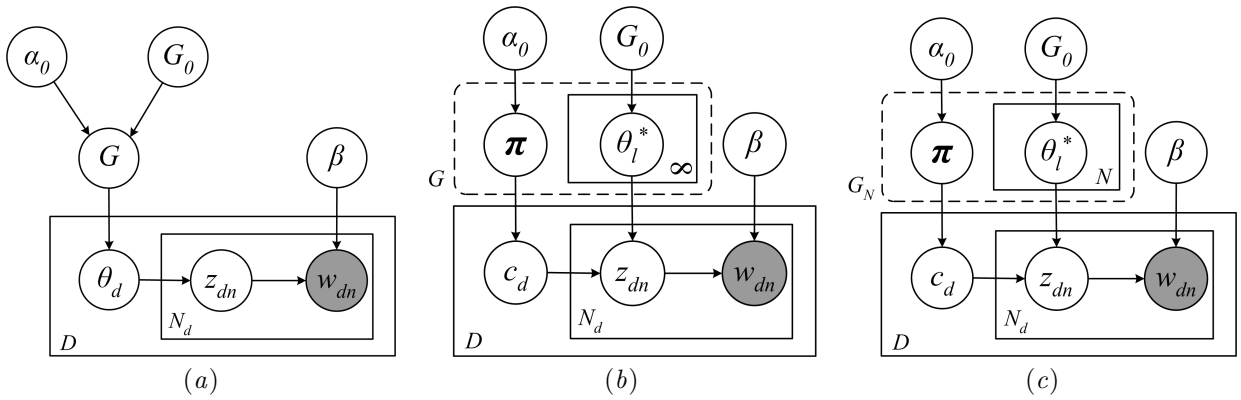



Figure 2: Plate models for latent semantic analysis. (a) Latent semantic analysis with DP prior; (b) An equivalent representation, where $c_d$ is the indicator variable saying which cluster document $d$ takes on out of the infinite clusters induced by DP; (c) Latent semantic analysis with a finite approximation of DP (see Sec. 2.3).

$\theta_l^*$ is sampled independently from $G_0$ and $\pi_l$ is defined as

$$\pi_1 = B_1, \quad \pi_l = B_l \prod_{j=1}^{l-1}(1 - B_j),$$

where $B_l$ are i.i.d. sampled from Beta distribution $\text{Beta}(1, \alpha_0)$. Thus, with a small $\alpha_0$, the first "sticks" $\pi_l$ will be large with little left for the remaining sticks. Conversely, if $\alpha_0$ is large, the first sticks $\pi_l$ and all subsequent sticks will be small and the $\pi_l$ will be more evenly distributed. In conclusion, the base distribution determines the locations of the point masses and $\alpha_0$ determines the distribution of probability weights. The distribution is nonzero at an infinite number of discrete points. If $\alpha_0$ is selected to be small the amplitudes of only a small number of discrete points will be significant. Note, that both locations and weights are not fixed but take on new values each time a new sample of $G$ is generated. Since $\mathbb{E}(G) = G_0$, initially, the prior corresponds to the prior used in LDA. With many documents in the training data set, locations $\theta_l^*$ which agree with the data will obtain a large weight. If a small $\alpha_0$ is chosen, parameters will form clusters whereas if a large $\alpha_0$, many representative parameters will result. Thus Dirichlet enhancement serves two purposes: it increases the flexibility in representing the posterior distribution of mixing weights and encourages a clustered solution leading to insights into the document corpus.

The DP prior offers two advantages against usual document clustering methods. First, there is no need to specify the number of clusters. The finally resulting clustering structure is constrained by the DP prior, but also adapted to the empirical observations. Second, the number of clusters is not fixed. Although the parameter $\alpha_0$ is a control parameter to tune the tendency for forming clusters, the DP prior allows the creation of new clusters if the current model cannot explain upcoming data very well, which is particularly suitable for our setting where dictionary is fixed while documents can be growing.

By applying the stick breaking representation, our model obtains the equivalent representation in Fig. 2(b). An infinite number of $\theta_l^*$ are generated from the base distribution and the new indicator variable $c_d$ indicates which $\theta_l^*$ is assigned to document $d$. If more than one document is assigned to the same $\theta_l^*$, clustering occurs. $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_\infty\}$ is a vector of probability weights generated from the stick breaking process.

### 2.3 Dirichlet-Multinomial Allocation (DMA)

Since infinite number of pairs $(\pi_l, \theta_l^*)$ are generated in the stick breaking process, it is usually very difficult to deal with the unknown distribution $G$. For inference there exist Markov chain Monte Carlo (MCMC) methods like Gibbs samplers which directly sample $\theta_d$ using Pólya urn scheme and avoid the difficulty of sampling the infinite-dimensional $G$ [4]; in practice, the sampling procedure is very slow and thus impractical for high dimensional data like text. In Bayesian statistics, the *Dirichlet-multinomial allocation* $\text{DP}_N$ in [6] has often been applied as a finite approximation to DP (see [6, 9]), which takes on the form

$$G_N = \sum_{l=1}^{N} \pi_l \delta_{\theta_l^*},$$

where $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_N\}$ is an $N$-vector of probability weights sampled once from a Dirichlet prior $\text{Dir}(\alpha_0/N, \ldots, \alpha_0/N)$, and $\theta_l^*$, $l = 1, \ldots, N$, are i.i.d. sampled from the base distribution $G_0$. It has been shown that the limiting case of $\text{DP}_N$ is DP [6, 9, 12], and more importantly $\text{DP}_N$ demonstrates similar stick breaking properties and leads to a similar clustering effect [6]. If $N$ is sufficiently large with

respect to our sample size $D$, $\mathrm{DP}_N$ gives a good approximation to DP.

Under the $\mathrm{DP}_N$ model, the plate representation of our model is illustrated in Fig. 2(c). The likelihood of the whole collection $\mathcal{D}$ is

$$\mathcal{L}_{\mathrm{DP}_N}(\mathcal{D}|\alpha_0,\lambda,\beta) = \int_{\boldsymbol{\pi}}\int_{\boldsymbol{\theta}^*} \prod_{d=1}^{D}\Big[\sum_{c_d=1}^{N} p(\mathbf{w}_d|\boldsymbol{\theta}^*, c_d;\beta) \; p(c_d|\boldsymbol{\pi})\Big]\, dP(\boldsymbol{\theta}^*;G_0)\, dP(\boldsymbol{\pi};\alpha_0) \quad (6)$$

where $c_d$ is the indicator variable saying which unique value $\theta_l^*$ document $d$ takes on. The likelihood of document $d$ is therefore written as

$$p(\mathbf{w}_d|\boldsymbol{\theta}^*, c_d;\beta) = \prod_{n=1}^{N_d}\sum_{z_{d,n}=1}^{k} p(w_{d,n}|z_{d,n};\beta)p(z_{d,n}|\theta^*_{c_d}).$$

### 2.4 Connections to PLSA and LDA

From the application point of view, PLSA and LDA both aim to discover the latent dimensions of data with the emphasis on *indexing*. The proposed Dirichlet enhanced semantic analysis retains the strengths of PLSA and LDA, and further explores the clustering structure of data. The model is a generalization of LDA. If we let $\alpha_0 \to \infty$, the model becomes identical to LDA, since the sampled $G$ becomes identical to the finite Dirichlet base distribution $G_0$. This extreme case makes documents mutually independent given $G_0$, since $\theta_d$ are i.i.d. sampled from $G_0$. If $G_0$ itself is not sufficiently expressive, the model is not able to capture the dependency between documents. The Dirichlet enhancement elegantly solves this problem. With a moderate $\alpha_0$, the model allows $G$ to deviate away from $G_0$, giving modelling flexibilities to explore the richer structure of data. The exchangeability may not exist within the whole collection, but between groups of documents with respective atoms $\theta_l^*$ sampled from $G_0$. On the other hand, the increased flexibility does not lead to overfitting, because inference and learning are done in a Bayesian setting, averaging over the number of mixture components and the states of the latent variables.

## 3 Inference and Learning

In this section we consider model inference and learning based on the $\mathrm{DP}_N$ model. As seen from Fig. 2(c), the inference needs to calculate the *a posteriori* joint distribution of latent variables $p(\boldsymbol{\pi}, \boldsymbol{\theta}^*, \mathbf{c}, \mathbf{z}|\mathcal{D}, \alpha_0, \lambda, \beta)$, which requires to compute Eq. (6). This integral is however analytically infeasible. A straightforward Gibbs sampling method can be derived, but it turns out to be very slow and inapplicable to high dimensional data like text, since for each word we have to sample a latent variable $z$. Therefore in this section we suggest efficient *variational* inference.

### 3.1 Variational Inference

The idea of variational mean-field inference is to propose a joint distribution $Q(\boldsymbol{\pi}, \boldsymbol{\theta}^*, \mathbf{c}, \mathbf{z})$ conditioned on some free parameters, and then enforce $Q$ to approximate the *a posteriori* distributions of interests by minimizing the KL-divergence $D_{KL}(Q\|p(\boldsymbol{\pi}, \boldsymbol{\theta}^*, \mathbf{c}, \mathbf{z}|\mathcal{D}, \alpha_0, \lambda, \beta))$ with respect to those free parameters. We propose a variational distribution $Q$ over latent variables as the following

$$Q(\boldsymbol{\pi}, \boldsymbol{\theta}^*, \mathbf{c}, \mathbf{z}|\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\varphi}, \boldsymbol{\phi}) = Q(\boldsymbol{\pi}|\boldsymbol{\eta}) \cdot \prod_{l=1}^{N} Q(\theta_l^*|\gamma_l) \prod_{d=1}^{D} Q(c_d|\varphi_d) \prod_{d=1}^{D}\prod_{n=1}^{N_d} Q(z_{d,n}|\phi_{d,n}) \quad (7)$$

where $\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\varphi}, \boldsymbol{\phi}$ are *variational parameters*, each tailoring the variational *a posteriori* distribution to each latent variable. In particular, $\boldsymbol{\eta}$ specifies an $N$-dimensional Dirichlet distribution for $\boldsymbol{\pi}$, $\gamma_l$ specifies a $k$-dimensional Dirichlet distribution for distinct $\theta_l^*$, $\varphi_d$ specifies an $N$-dimensional multinomial for the indicator $c_d$ of document $d$, and $\phi_{d,n}$ specifies a $k$-dimensional multinomial over latent topics for word $w_{d,n}$. It turns out that the minimization of the KL-divergence is equivalent to the maximization of a lower bound of the $\ln p(\mathcal{D}|\alpha_0, \lambda, \beta)$ derived by applying Jensen's inequality [10]. Please see the Appendix for details of the derivation. The lower bound is then given as

$$\begin{aligned}\mathcal{L}_Q(\mathcal{D}) = &\sum_{d=1}^{D}\sum_{n=1}^{N_d} \mathbb{E}_Q[\ln p(w_{d,n}|z_{d,n},\beta)p(z_{d,n}|\boldsymbol{\theta}^*, c_d)] \\ &+ \mathbb{E}_Q[\ln p(\boldsymbol{\pi}|\alpha_0)] + \sum_{d=1}^{D}\mathbb{E}_Q[\ln p(c_d|\boldsymbol{\pi})] \quad (8)\\ &+ \sum_{l=1}^{N}\mathbb{E}_Q[\ln p(\theta_l^*|G_0)] - \mathbb{E}_Q[\ln Q(\boldsymbol{\pi}, \boldsymbol{\theta}^*, \mathbf{c}, \mathbf{z})].\end{aligned}$$

The optimum is found setting the partial derivatives with respect to each variational parameter to be zero,

which gives rise to the following updates

$$\phi_{d,n,i} \propto \beta_{i,w_{d,n}} \exp\Big\{ \sum_{l=1}^{N} \varphi_{d,l} \big[\Psi(\gamma_{l,i}) - \Psi(\sum_{j=1}^{k} \gamma_{l,j})\big] \Big\} \tag{9}$$

$$\varphi_{d,l} \propto \exp\Big\{ \sum_{i=1}^{k} \Big[ \Big(\Psi(\gamma_{l,i}) - \Psi(\sum_{j=1}^{k} \gamma_{l,j})\Big) \sum_{n=1}^{N_d} \phi_{d,n,i} \Big] + \Psi(\eta_l) - \Psi(\sum_{j=1}^{N} \eta_j) \Big\} \tag{10}$$

$$\gamma_{l,i} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \varphi_{d,l} \phi_{d,n,i} + \lambda_i \tag{11}$$

$$\eta_l = \sum_{d=1}^{D} \varphi_{d,l} + \frac{\alpha_0}{N} \tag{12}$$

where $\Psi(\cdot)$ is the digamma function, the first derivative of the log Gamma function. Some details of the derivation of these formula can be found in Appendix. We find that the updates are quite interpretable. For example, in Eq. (9) $\phi_{d,n,i}$ is the *a posteriori* probability of latent topic $i$ given one word $w_{d,n}$. It is determined both by the corresponding entry in the $\beta$ matrix that can be seen as a *likelihood* term, and by the possibility that document $d$ selects topic $i$, i.e., the *prior* term. Here the prior is itself a weighted average of different $\theta_l^*$ s to which $d$ is assigned. In Eq. (12) $\eta_l$ is the *a posteriori* weight of $\pi_l$, and turns out to be the tradeoff between empirical responses at $\theta_l^*$ and the prior specified by $\alpha_0$. Finally since the parameters are coupled, the variational inference is done by iteratively performing Eq. (9) to Eq. (12) until convergence.

### 3.2 Parameter Estimation

Following the empirical Bayesian framework, we can estimate the hyper parameters $\alpha_0, \lambda$, and $\beta$ by iteratively maximizing the lower bound $\mathcal{L}_Q$ both with respect to the variational parameters (as described by Eq. (9)-Eq. (12)) and the model parameters, holding the remaining parameters fixed. This iterative procedure is also referred to as variational EM [10]. It is easy to derive the update for $\beta$:

$$\beta_{i,j} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{d,n,i} \delta_j(w_{d,n}) \tag{13}$$

where $\delta_j(w_{d,n}) = 1$ if $w_{d,n} = j$, and 0 otherwise. For the remaining parameters, let's first write down the parts of $\mathcal{L}$ in Eq. (8) involving $\alpha_0$ and $\lambda$:

$$\mathcal{L}_{[\alpha_0]} = \ln\Gamma(\alpha_0) - N \ln\Gamma\big(\frac{\alpha_0}{N}\big) + (\frac{\alpha_0}{N} - 1) \sum_{l=1}^{N} \big[\Psi(\eta_l) - \Psi(\sum_{j=1}^{N} \eta_j)\big],$$

$$\mathcal{L}_{[\lambda]} = \sum_{l=1}^{N} \Big\{ \ln\Gamma(\sum_{i=1}^{k} \lambda_i) - \sum_{i=1}^{k} \ln\Gamma(\lambda_i) + \sum_{i=1}^{k} (\lambda_i - 1) \big[\Psi(\gamma_{l,i}) - \Psi(\sum_{j=1}^{k} \gamma_{l,j})\big] \Big\}.$$

Estimates for $\alpha_0$ and $\lambda$ are found by maximization of these objective functions using standard methods like Newton-Raphson method as suggested in [3].

## 4 Empirical Study

### 4.1 Toy Data

We first apply the model on a toy problem with $k = 5$ latent topics and a dictionary containing 200 words. The assumed probabilities of generating words from topics, i.e. the parameters $\beta$, are illustrated in Fig. 3(d), in which each colored line corresponds to a topic and assigns non-zero probabilities to a subset of words. For each run we generate data with the following steps: (1) one cluster number $M$ is chosen between 5 and 12; (2) generate $M$ document clusters, each of which is defined by a combination of topics; (3) generate each document $d$, $d = 1, \ldots, 100$, by first randomly selecting a cluster and then generating 40 words according to the corresponding topic combinations. For $\mathrm{DP}_N$ we select $N = 100$ and we aim to examine the performance for discovering the latent topics and the document clustering structure.

In Fig. 3(a)-(c) we illustrate the process of clustering documents over EM iterations with a run containing 6 document clusters. In Fig. 3(a), we show the initial random assignment $\varphi_{d,l}$ of each document $d$ to a cluster $l$. After one EM step documents begin to accumulate to a reduced number of clusters (Fig. 3(b)), and converge to exactly 6 clusters after 5 steps (Fig. 3(c)). The learned word distribution of topics $\beta$ is shown in Fig. 3(e) and is very similar to the true distribution.

By varying $M$, the true number of document clusters, we examine if our model can find the correct $M$. To determine the number of clusters, we run the variational inference and obtain for each document a weight vector $\varphi_{d,l}$ of clusters. Then each document takes the cluster with largest weight as its assignment, and we calculate the cluster number as the number of non-empty clusters. For each setting of $M$ from 5 to 12, we randomize the data for 20 trials and obtain the curve in Fig. 3(f)

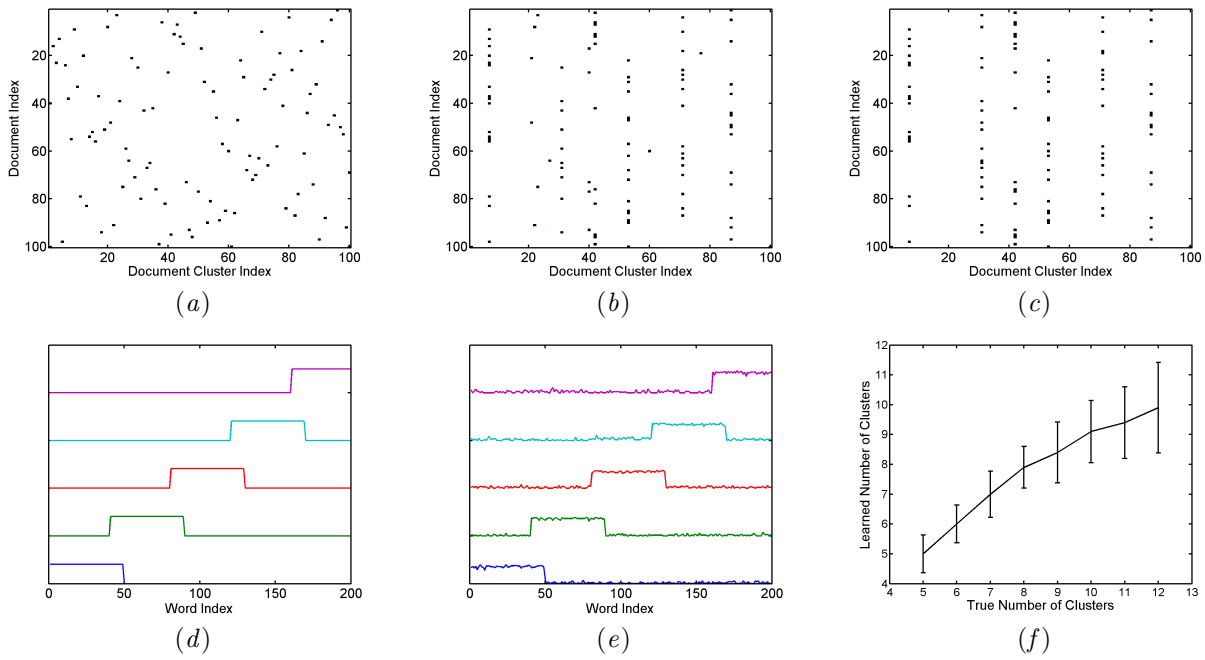



Figure 3: Experimental results for the toy problem. (a)-(c) show the document-cluster assignments $\varphi_{d,l}$ over the variational inference for a run with 6 document clusters: (a) Initial random assignments; (b) Assignments after one iteration; (c) Assignments after five iterations (final). The multinomial parameter matrix $\beta$ of true values and estimated values are given in (d) and (e), respectively. Each line gives the probabilities of generating the 200 words, with wave mountains for high probabilities. (f) shows the learned number of clusters with respect to the true number with mean and error bar.

which shows the average performance and the variance. In 37% of the runs we get perfect results, and in another 43% runs the learned values only deviate from the truth by one. However, we also find that the model tends to get slightly fewer than $M$ clusters when $M$ is large. The reason might be that, only 100 documents are not sufficient for learning a large number $M$ of clusters.

### 4.2 Document Modelling

We compare the proposed model with PLSI and LDA on two text data sets. The first one is a subset of the Reuters-21578 data set which contains 3000 documents and 20334 words. The second one is taken from the 20-newsgroup data set and has 2000 documents with 8014 words. The comparison metric is *perplexity*, conventionally used in language modelling. For a test document set, it is formally defined as

$$\text{Perplexity}(\mathcal{D}_{\text{test}}) = \exp(-\ln p(\mathcal{D}_{\text{test}})/\sum_d |\mathbf{w}_d|).$$

We follow the formula in [3] to calculate the perplexity for PLSI. In our algorithm $N$ is set to be the number of training documents. Fig. 4(a) and (b) show the comparison results with different number $k$ of latent topics. Our model outperforms LDA and PLSI in all the runs, which indicates that the flexibility introduced by DP enhancement does not produce overfitting and results in a better generalization performance.

### 4.3 Clustering

In our last experiment we demonstrate that our approach is suitable to find relevant document clusters. We select four categories, *autos, motorcycles, baseball* and *hockey* from the 20-newsgroups data set with 446 documents in each topic. Fig. 4(c) illustrates one clustering result, in which we set topic number $k = 5$ and found 6 document clusters. In the figure the documents are indexed according to their true category labels, so we can clearly see that the result is quite meaningful. Documents from one category show similar membership to the learned clusters, and different categories can be distinguished very easily. The first two categories are not clearly separated because they are both talking about vehicles and share many terms, while the rest of the categories, baseball and hockey, are ideally detected.

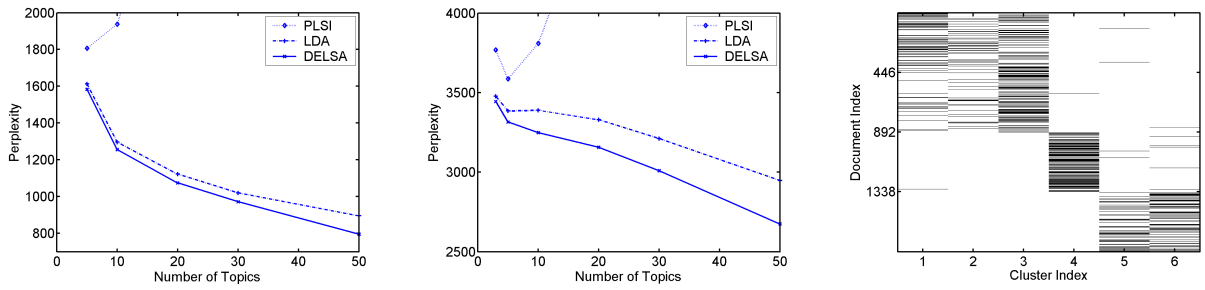



Figure 4: (a) and (b): Perplexity results on Reuters-21578 and 20-newsgroups for DELSA, PLSI and LDA; (c): Clustering result on 20-newsgroups dataset.

## 5 Conclusions and Future Work

This paper proposes a Dirichlet enhanced latent semantic analysis model for analyzing co-occurrence data like text, which retains the strength of previous approaches to find latent topics, and further introduces additional modelling flexibilities to uncover the clustering structure of data. For inference and learning, we adopt a variational mean-field approximation based on a finite alternative of DP. Experiments are performed on a toy data set and two text data sets. The experiments show that our model can discover both the latent semantics and meaningful clustering structures.

In addition to our approach, alternative methods for approximate inference in DP have been proposed using expectation propagation (EP) [11] or variational methods [16, 2]. Our approach is most similar to the work of Blei and Jordan [2] who applied mean-field approximation for the inference in DP based on a truncated DP (TDP). Their approach was formulated in context of general exponential-family mixture models [2]. Conceptually, $\mathrm{DP}_N$ appears to be simpler than TDP in the sense that the *a posteriori* of $G$ is a symmetric Dirichlet while TDP ends up with a generalized Dirichlet (see [8]). In another sense, TDP seems to be a tighter approximation to DP. Future work will include a comparison of the various DP approximations.

## Acknowledgements

The authors thank the anonymous reviewers for their valuable comments. Shipeng Yu gratefully acknowledges the support through a Siemens scholarship.

## References

[1] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems (NIPS) 14*, 2002.

[2] D. M. Blei and M. I. Jordan. Variational methods for the Dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[3] D. M. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430), June 1995.

[5] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

[6] P. J. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. unpublished paper, 2000.

[7] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM SIGIR Conference*, pages 50–57, Berkeley, California, August 1999.

[8] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[9] H. Ishwaran and M. Zarepour. Exact and approximate sum-representations for the Dirichlet process. *Can. J. Statist*, 30:269–283, 2002.

[10] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[11] T. Minka and Z. Ghahramani. Expectation propagation for infinite mixtures. In *NIPS'03 Workshop on Nonparametric Bayesian Methods and Infinite Models*, 2003.

[12] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal*

*of Computational and Graphical Statistics*, 9:249–265, 2000.

[13] C. E. Rasmussen and Z. Ghahramani. Infinite mixtures of gaussian process experts. In *Advances in Neural Information Processing Systems 14*, 2002.

[14] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Technical Report 653, Department of Statistics, University of California, Berkeley, 2004.

[16] V. Tresp and K. Yu. An introduction to nonparametric hierarchical bayesian modelling with a focus on multi-agent learning. In *Proceedings of the Hamilton Summer School on Switching and Learning in Feedback Systems*. Lecture Notes in Computing Science, 2004.

[17] K. Yu, V. Tresp, and S. Yu. A nonparametric hierarchical Bayesian framework for information filtering. In *Proceedings of 27th Annual International ACM SIGIR Conference*, 2004.

## Appendix

To simplify the notation, we denote $\mathbf{\Xi}$ for all the latent variables $\{\boldsymbol{\pi}, \boldsymbol{\theta}^*, \mathbf{c}, \mathbf{z}\}$. With the variational form Eq. (7), we apply Jensen's inequality to the likelihood Eq. (6) and obtain

$$\begin{aligned}
&\ln p(\mathcal{D}|\alpha_0, \lambda, \beta) \\
&\quad = \ln \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\theta}^*} \sum_{\mathbf{c}} \sum_{\mathbf{z}} p(\mathcal{D}, \mathbf{\Xi}|\alpha_0, \lambda, \beta) d\boldsymbol{\theta}^* d\boldsymbol{\pi} \\
&\quad = \ln \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\theta}^*} \sum_{\mathbf{c}} \sum_{\mathbf{z}} \frac{Q(\mathbf{\Xi}) p(\mathcal{D}, \mathbf{\Xi}|\alpha_0, \lambda, \beta)}{Q(\mathbf{\Xi})} d\boldsymbol{\theta}^* d\boldsymbol{\pi} \\
&\quad \geq \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\theta}^*} \sum_{\mathbf{c}} \sum_{\mathbf{z}} Q(\mathbf{\Xi}) \ln p(\mathcal{D}, \mathbf{\Xi}|\alpha_0, \lambda, \beta) d\boldsymbol{\theta}^* d\boldsymbol{\pi} \\
&\qquad - \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\theta}^*} \sum_{\mathbf{c}} \sum_{\mathbf{z}} Q(\mathbf{\Xi}) \ln Q(\mathbf{\Xi}) d\boldsymbol{\theta}^* d\boldsymbol{\pi} \\
&\quad = \mathbb{E}_Q[\ln p(\mathcal{D}, \mathbf{\Xi}|\alpha_0, \lambda, \beta)] - \mathbb{E}_Q[\ln Q(\mathbf{\Xi})],
\end{aligned}$$

which results in Eq. (8).

To write out each term in Eq. (8) explicitly, we have, for the first term,

$$\sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_Q[\ln p(w_{d,n}|z_{d,n}, \beta)] = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{d,n,i} \beta_{i,\nu},$$

where $\nu$ is the index of word $w_{d,n}$.

The other terms can be derived as follows:

$$\begin{aligned}
&\sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_Q[\ln p(z_{d,n}|\boldsymbol{\theta}^*, c_d)] = \\
&\qquad \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{l=1}^{N} \psi_{d,l} \phi_{d,n,i} \big[\Psi(\gamma_{l,i}) - \Psi(\sum_{j=1}^{k} \gamma_{l,j})\big], \\
&\mathbb{E}_Q[\ln p(\boldsymbol{\pi}|\alpha_0)] = \ln \Gamma(\alpha_0) - N \ln \Gamma(\frac{\alpha_0}{N}) \\
&\qquad + \Big(\frac{\alpha_0}{N} - 1\Big) \sum_{l=1}^{N} \big[\Psi(\eta_l) - \Psi(\sum_{j=1}^{N} \eta_j)\big], \\
&\sum_{d=1}^{D} \mathbb{E}_Q[\ln p(c_d|\boldsymbol{\pi})] = \sum_{d=1}^{D} \sum_{l=1}^{N} \psi_{d,l} \big[\Psi(\eta_l) - \Psi(\sum_{j=1}^{N} \eta_j)\big], \\
&\sum_{l=1}^{N} \mathbb{E}_Q[\ln p(\theta_l^*|G_0)] = \sum_{l=1}^{N} \Big\{ \ln \Gamma(\sum_{i=1}^{k} \lambda_i) - \sum_{i=1}^{k} \ln \Gamma(\lambda_i) \\
&\qquad + \sum_{i=1}^{k} (\lambda_i - 1) \big[\Psi(\gamma_{l,i}) - \Psi(\sum_{j=1}^{k} \gamma_{l,j})\big] \Big\},
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_Q[\ln Q(\boldsymbol{\pi}, \boldsymbol{\theta}^*, \mathbf{c}, \mathbf{z})] = \ln \Gamma(\sum_{l=1}^{N} \eta_l) - \sum_{l=1}^{N} \ln \Gamma(\eta_l) \\
&\quad + \sum_{l=1}^{N} (\eta_l - 1) \big[\Psi(\eta_l) - \Psi(\sum_{j=1}^{N} \eta_j)\big] \\
&\quad + \sum_{l=1}^{N} \Big\{ \ln \Gamma(\sum_{i=1}^{k} \gamma_{l,i}) - \sum_{i=1}^{k} \ln \Gamma(\gamma_{l,i}) \\
&\qquad + \sum_{i=1}^{k} (\gamma_{l,i} - 1) \big[\Psi(\gamma_{l,i}) - \Psi(\sum_{j=1}^{k} \gamma_{l,j})\big] \Big\} \\
&\quad + \sum_{d=1}^{D} \sum_{l=1}^{N} \psi_{d,l} \ln \psi_{d,l} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{d,n,i} \ln \phi_{d,n,i}.
\end{aligned}$$

Differentiating the lower bound with respect to different latent variables gives the variational E-step in Eq. (9) to Eq. (12). M-step can also be obtained by considering the lower bound with respect to $\beta$, $\lambda$ and $\alpha_0$.