Dirichlet Enhanced Latent Semantic Analysis

1 The Goal

In latent semantic analysis (LSA), we aim at modelling a large corpus of *high-dimensional discrete data* from *probabilistic* perspective.

The Assumption: one data point can be modelled by *latent factors*, which account for the co-occurrence of items within the data.

We are also interested in the *clustering* structure of the data, which may benefit from the latent factors of the items.

For example:

- In document modelling, the data are document-word pairs.
  Latent factors: topics for words
  Data clustering: categories of documents
- In collaborative filtering, the data are user ratings (for, e.g., movies).
  Latent factors: categories or structures of movies
  Data clustering: user interest groups

We wish to build a probabilistic model that

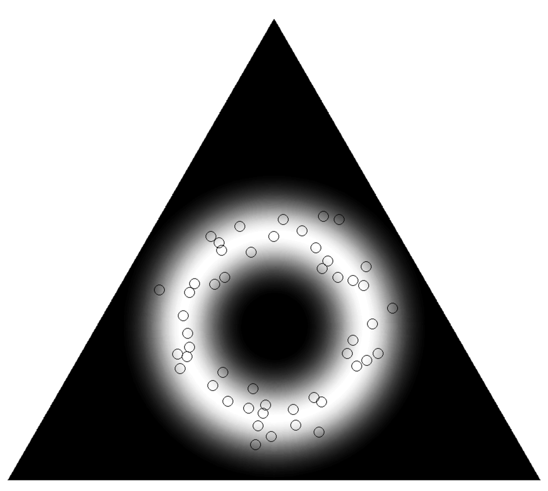
- is *flexible enough* to learn arbitrary probabilistic dependencies
- will not *overfit* the training data and generalize poorly
- facilitates *model selection*, e.g., choosing the number of clusters

We choose document modelling as the working example in this poster, and call the latent factors for words as topics.

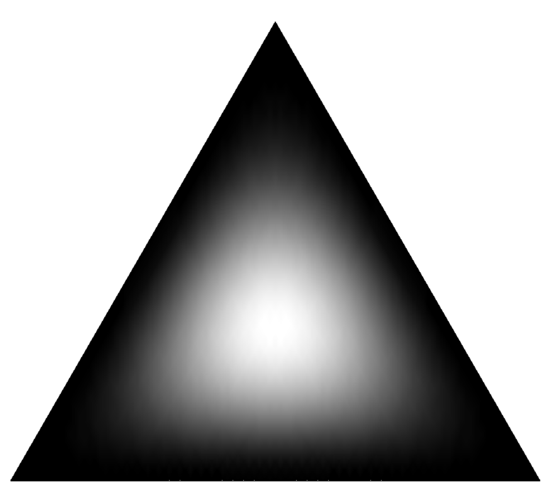
2 Previous Models

Take different assumptions and have corresponding limitations:

- **Mixture of Unigrams**
  Assign a discrete latent model to words
  - Assumption: Words are i.i.d. sampled after choosing a topic
  - Limitation: One document only belongs to one word topic
- **Probabilistic Latent Semantic Indexing (PLSI)** [Hofmann, 1999]
  Assign a discrete latent model to document-word pairs
  - Assumption: Document-word pairs are i.i.d. sampled given a topic
  - Limitation: Not a well-defined model because long documents could get higher probability in the sampling process
- **Latent Dirichlet Allocation (LDA)** [Blei et al., 2003]
  Assign a discrete latent model to words and let each document maintain a random variable θ, saying its probabilities of belonging to each topic
  - Assumption: Assign a Dirichlet prior for θ
  - Limitation: A single Dirichlet distribution is not flexible enough and no clustering structure can be found for documents



The true distribution of θ in a toy problem



The learned Dirichlet distribution in LDA

3 Dirichlet Enhanced Latent Semantic Analysis

The key point of the DELSA model is to replace the single Dirichlet distribution in LDA with a *nonparametric Dirichlet process prior*, which gains:

- a flexible enough distribution to fit an arbitrary prior
- a natural discrete *clustering* structure for documents
- automatic determination of the number of clusters

Notations: We consider a corpus $\mathcal{D}$ containing D documents. Each document d is denoted by $\mathbf{w}_d = \{w_{d,1}, \ldots, w_{d,N_d}\}$ with N_d words. $w_{d,n}$ is a variable for the n-th word in $\mathbf{w}_d$ and denotes the index of the corresponding word in a vocabulary $\mathcal{V}$ of length V.

The Model

$$w_{d,n}|z_{d,n};\beta \sim \mathrm{Mult}(\beta_{z_{d,n}}) \qquad \theta_d \sim G$$
$$z_{d,n}|\theta_d \sim \mathrm{Mult}(\theta_d) \qquad G;G_0,\alpha_0 \sim \mathrm{DP}(G_0,\alpha_0)$$

- Each document $\mathbf{w}_d$ maintains a variable θ_d of *topic mixtures*
- Each word $w_{d,n}$ in document $\mathbf{w}_d$ is sampled by first choosing a topic $z_{d,n}$ given θ_d, and then sampling the word given the topic-word matrix β
- Variables θ_d are sampled from the *Dirichlet process*, with a Dirichlet distribution $G_0(\cdot|\lambda)$ as the *base distribution*, and a positive scalar α_0 as the *concentration parameter*

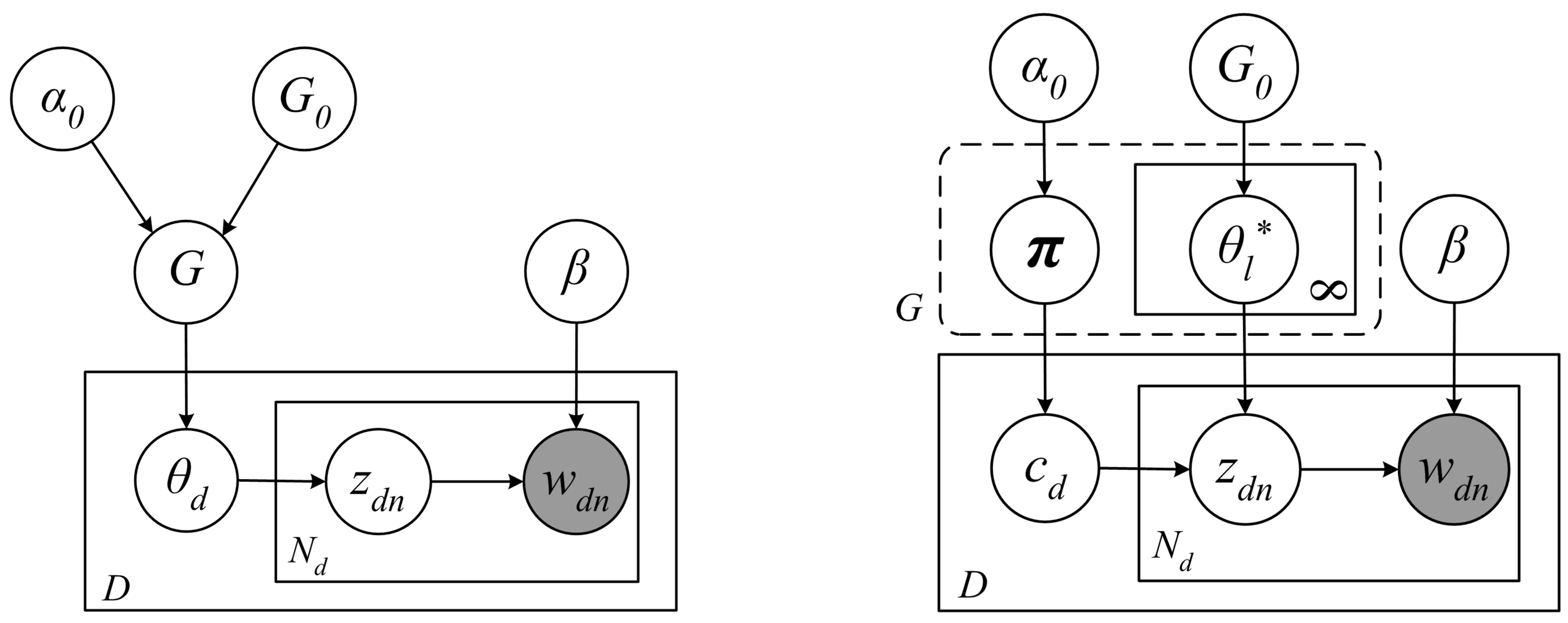


Plate models for DELSA with DP prior (left) and stick-breaking (right)

Stick-breaking and Dirichlet Enhancing

The unknown distribution G in DP has a *stick-breaking* representation:

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

- θ_l^* are countably infinite variables i.i.d. sampled from G_0
- $\delta_\theta(\cdot)$ are point mass distributions concentrated at θ
- $\pi_l \geq 0$, $\sum_l^\infty \pi_l = 1$ are sampled by *stick-breaking process*:
  $$\pi_1 = B_1, \qquad \pi_l = B_l \prod_{j=1}^{l-1}(1-B_j), \quad l > 1$$
  where B_l are i.i.d. sampled from Beta distribution $\mathrm{Beta}(1,\alpha_0)$.

Parameters of the model (total number $k + 2 + k \times (V-1)$):

- We fix the number of word topics to be k
- $G_0(\theta) \sim \mathrm{Dir}(\theta|\lambda)$ is the base distribution, which tells *how the distinct θ's are sampled*. It reflects our prior knowledge of the *cluster centers*
- α_0 is the concentration parameter, which *controls the flexibility of generating new clusters*. Larger α_0 results more clusters.
- β is a $k \times V$ matrix. $\beta(i,n) = p(w_n|z_i)$ tells the probability of generating word w_n from topic z_i. Each row of β sums to 1

An equivalent graphical model for DELSA with stick-breaking is shown.

Kai Yu[1], Shipeng Yu[2], Volker Tresp[1]

[1] Dept. of Neural Computation (CT IC4)
Siemens Corporate Technology, Munich, Germany

[2] Institute for Computer Science
University of Munich, Germany

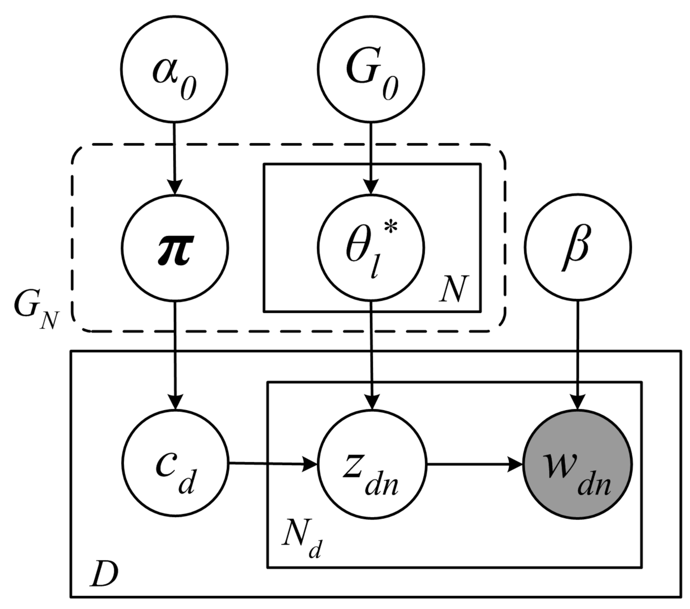
Paper and poster online at `http://www.dbs.ifi.lmu.de/~spyu/`

Dirichlet-Multinomial Allocation

Great difficulty emerges when we turn to inference, since we have to deal with the unknown distribution G with infinite number of pairs (π_l, θ_l^*). Markov chain Monte Carlo methods can be applied based on Pólya urn scheme, but could be very slow for high dimensional data like text.

Therefore we turn to Dirichlet-multinomial allocation (DMA), a finite approximation to DP denoted as DP_N:

$$G_N(\cdot) = \sum_{l=1}^{N} \pi_l \delta_{\theta_l^*}(\cdot)$$



where

- N is a large positive integer
- $\{\pi_1, \dots, \pi_N\} \sim \mathrm{Dir}(\frac{\alpha_0}{N}, \dots, \frac{\alpha_0}{N})$
- θ_l^* are i.i.d. sampled from G_0
- with $N \to +\infty$, $\mathrm{DP}_N \to \mathrm{DP}$

Now the model also has a very intuitive explanation from the perspective of *finite mixture modelling*. By setting N to be very large, the model can automatically discover *a small number of clusters*, leaving others empty.

The likelihood of the whole collection $\mathcal{D}$ (conditional on α_0, λ, β) is

$$\mathcal{L}_{\mathrm{DP}_N}(\mathcal{D}) = \int_\pi \int_{\theta^*} \prod_d^D \left[\sum_{c_d}^N p(c_d|\pi) \prod_n^{N_d} \sum_{z_{d,n}}^k p(w_{d,n}|z_{d,n};\beta) p(z_{d,n}|\theta_{c_d}^*) \right] dP(\theta^*; G_0) dP(\pi; \alpha_0)$$

Variational Inference and Learning

To overcome the intractability of the integral, we apply *mean-field approximation* to the posterior of hidden variables with following tractable form:

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

By applying Jensen's inequality we obtain a lower bound of the likelihood and get the updates for variational parameters in variational E-step:

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

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

$$\gamma_{l,i} = \sum_{d=1}^D \sum_{n=1}^{N_d} \varphi_{d,l} \phi_{d,n,i} + \lambda_i\ , \qquad \eta_l = \sum_{d=1}^D \varphi_{d,l} + \frac{\alpha_0}{N}$$

The parameters $(\alpha_0, \lambda, \beta)$ can be updated in variational M-step by maximizing the lower bound with respect to them. β can be updated by

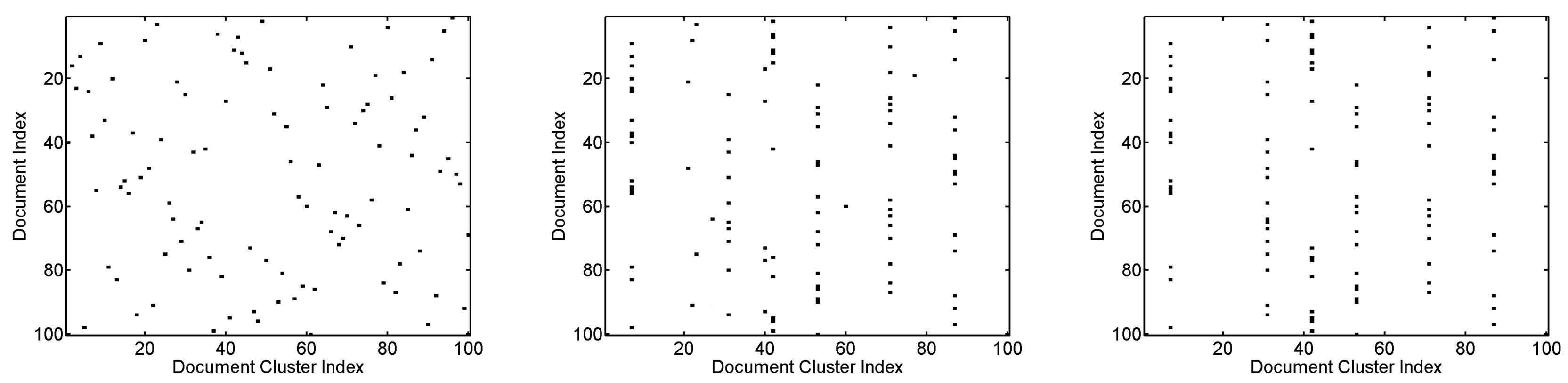
$$\beta_{i,j} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \phi_{d,n,i} \delta_j(w_{d,n})$$

α_0 and λ can be updated using Newton-Raphson method.

4 Evaluation

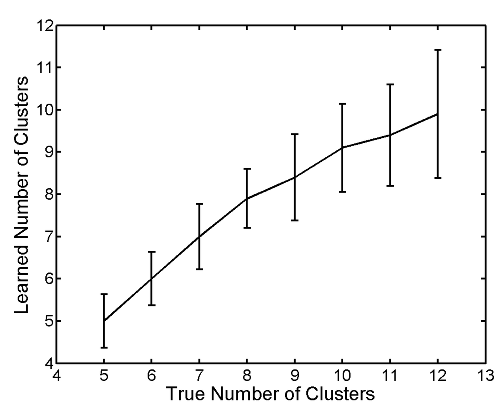
Toy Data

A dictionary of 200 words are associated with 5 latent topics. 100 documents are generated with 6 document clusters. $N = 100$ before learning.



Random initialization — After 1 EM step — After 5 EM step (final)

We then vary the number of clusters from 5 to 12 and randomize the data for 20 trials. We record the detected number of clusters.



- We can correctly detect number of clusters
- The calculation is fast without overfitting
- The recovered parameter β is very good

Document Modelling

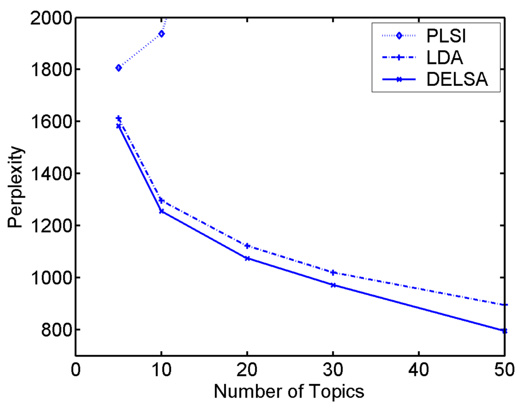
We compare DELSA with PLSI and LDA on Reuters-21578 and 20-Newsgroup in terms of *perplexity*: $\mathrm{Perp}(\mathcal{D}_t) = \exp\left(-\ln p(\mathcal{D}_t)/\sum_d |\mathbf{w}_d|\right)$.

- DELSA is consistently better than PLSI and LDA without overfitting
- Better for data set with strong clustering structure (like 20-Newsgroup)

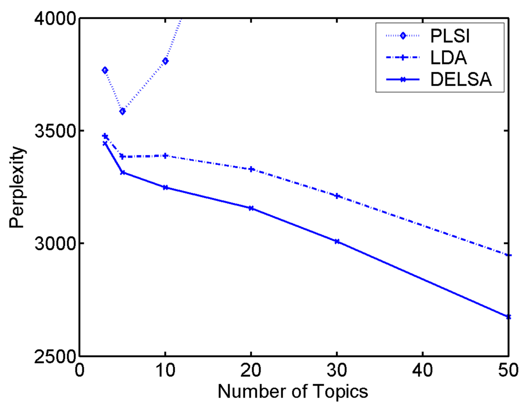
Clustering

We test DELSA on 20-Newsgroup data with 4 categories *autos*, *motorcycles*, *baseball* and *hockey*, each taking 446 documents. 6 clusters are found.

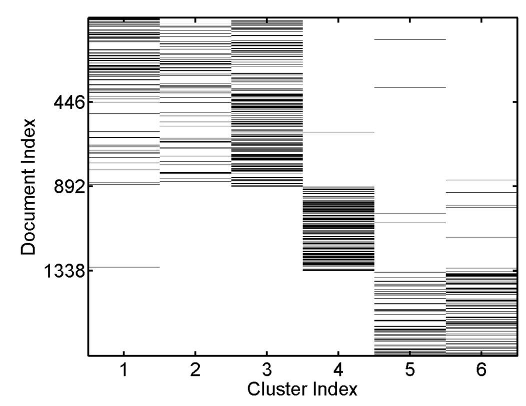
- Documents in one category show similar behavior
- Clear difference observable for different categories except related



Perplexity for Reuters



Perplexity for Newsgroup



Clustering for Newsgroup

5 Things to Keep in Mind

- Nonparametric Bayesian modelling with Dirichlet enhancement is *flexible enough to fit any prior distribution without overfitting*
- A natural discrete structure of DP results in a *clustering* structure for the data, with automatically determined number of clusters
- Variational methods for inference and learning are available for DP enhanced models, with which good performance can be obtained
- Future works include investigating other DP enhancement (e.g., [Teh et al., 2005]), and comparing different approximation methods for DP enhanced models (e.g., Blei and Jordan [2004])

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

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

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

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.