PGM Report
Latent Dirichlet Allocation

Quentin LEROY, Vincent MATTHYS, Bastien PONCHON
{qleroy,vmatthys,bponchon}@ens-paris-saclay.fr

1 Model

Latent Dirichlet Allocation is a generative probabilistic model, presentend in [1], for corpora of text documents or other collections of discrete data.

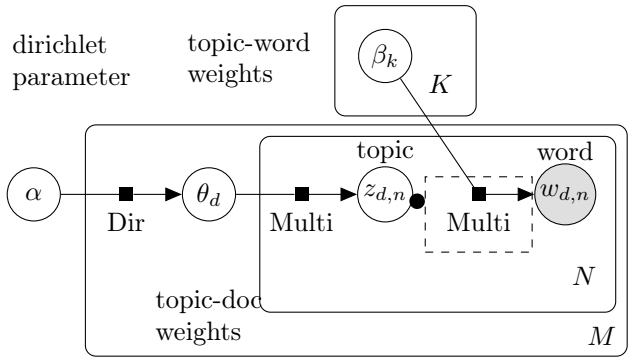


Figure 1: Graphical model representation of LDA

Using the language of text collections, we define:

- $\mathcal{D}$ corpus of M documents of N words
- K number of topics
- V length of vocabulary in corpus $\mathcal{D}$
- $w_{d,n} \in \{0,1\}^V$ one-hot encoded word n in document d
- $\alpha \in]0,1]^K$ prior on the per-document topic distributions
- $\theta_d \in \mathbb{R}^K$ probability of topics in document d: topic mixture at document level.
- $z_{d,n} \in \{0,1\}^K$ one-hot encoded topic of $w_{d,n}$
- $\beta \in [0,1]^{KV}$ with $\beta_{z_{d,n},w_{d,n}}$ probability of word $w_{d,n}$ given the topic $z_{d,n}$

So that we can express the following conditional probabilities:

$$p(\theta_d \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_{d,i}^{\alpha_i - 1}$$

$$p(z_{d,n} \mid \theta_d) = \theta_{d,z_{d,n}} = \prod_{k=1}^K \theta_{d,k}^{z_{d,n}^k} = \prod_{k=1}^K \prod_{j=1}^V \theta_{d,k}^{z_{d,n}^k w_{d,n}^j}$$

$$p(w_{d,n} \mid z_{n,d}, \beta) = \beta_{z_{d,n},w_{d,n}} = \prod_{k=1}^K \prod_{j=1}^V \beta_{k,j}^{z_{d,n}^k w_{d,n}^j}$$

The joint distribution according to the graphical model in figure 1 reads:

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta) = \prod_{d=1}^M p(\theta_d \mid \alpha) \prod_{n=1}^N p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid z_{n,d}, \beta) \tag{1}$$

$$= \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{d=1}^M \prod_{i=1}^K \theta_{d,i}^{\alpha_i - 1} \prod_{n=1}^N \prod_{k=1}^K \prod_{j=1}^V (\theta_{d,k} \beta_{k,j})^{z_{d,n}^k w_{d,n}^j} \tag{2}$$

2 Inference

Using the equation (2), the marginal distribution over words $\boldsymbol{w}$ can be written as:

$$p(\boldsymbol{w} \mid \alpha, \beta) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \sum_{\boldsymbol{z}} \prod_{d=1}^{M} \int_{\theta_d} \left(\prod_{i=1}^{K} \theta_{d,i}^{\alpha_i - 1} \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{j=1}^{V} (\theta_{d,k} \beta_{k,j})^{z_{d,n}^k w_{d,n}^j} \right) d\theta_d \tag{3}$$

which is not tractable due to the sum and integral over multidimensional parameters. The exact inference problem is not solvable, and there are different ways to do approximate inference.

2.1 Mean-field variational inference

Similarly to the derivation of the EM algorithm, finding a lower-bound on the log-likelihood $\log p(\boldsymbol{w} \mid \alpha, \beta)$ involves using Jensen's inequality and a distribution q over latent variables $\boldsymbol{\theta}$ and $\boldsymbol{z}$. It leads to the evidence lower bound $\mathcal{L}$, called ELBO:

$$\log p(\boldsymbol{w} \mid \alpha, \beta) \geq \mathbb{E}_q[\log p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta)] - \mathbb{E}_q[\log q(\boldsymbol{\theta}, \boldsymbol{z})] = \mathcal{L}(q; \alpha, \beta) \tag{4}$$

In the EM-algorithm, this ELBO is exactly equal to the log-likelihood for a distribution $q(\boldsymbol{\theta}, \boldsymbol{z}) = p(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{w}, \alpha, \beta) = p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta)/p(\boldsymbol{w} \mid \alpha, \beta)$. Nevertheless the denominator is not tractable, as seen in equation (3) thus the expectation under $q(\boldsymbol{\theta}, \boldsymbol{z})$ is not computable. Therefore a variational distribution has to be used to maximize the lower-bound through an optimization problem.

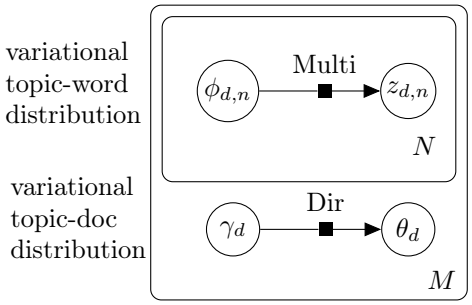


Figure 2: Graph of considered mean-field family

Specifying the form of $q(\boldsymbol{\theta}, \boldsymbol{z})$ using the mean-field family with variational parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, factorizing according to the graph in figure 2:

$$\begin{aligned} q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\phi}, \boldsymbol{\gamma}) &= q(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) q(\boldsymbol{z} \mid \boldsymbol{\phi}) \\ &= \prod_{d=1}^{M} q(\theta_d \mid \gamma_d) \prod_{n=1}^{N} q(z_{d,n} \mid \phi_{d,n}) \\ &= \prod_{d=1}^{M} q(\theta_d \mid \gamma_d) \prod_{n=1}^{N} q(z_{d,n} \mid \phi_{d,n}) \end{aligned} \tag{5}$$

The ELBO connects the variational distribution $q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\phi}, \boldsymbol{\gamma})$ to the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta)$ through the KL-divergence as shown in equation (6):

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \alpha, \beta) = -\mathcal{D}(q(\boldsymbol{\theta}, \boldsymbol{z} \mid \gamma, \phi) \mid\mid p(\boldsymbol{\theta}, \boldsymbol{z} \mid \alpha, \beta)) + \log p(\boldsymbol{w} \mid \alpha, \beta) \tag{6}$$

which ensures that the mean-field variational distribution used will get closed to the untractable posterior distribution once solved the variational E-step optimization problem. Given fixed $\boldsymbol{w}$ the variational parameters are solution to:

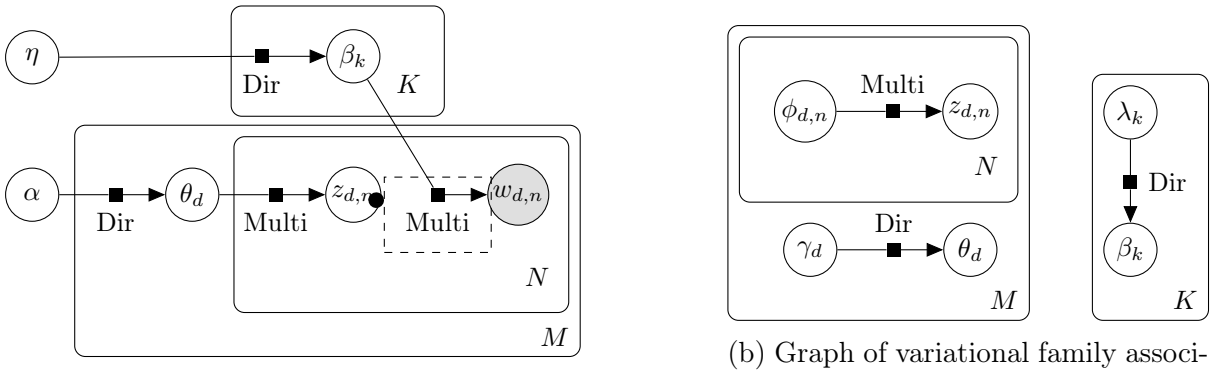
$$(\boldsymbol{\gamma}, \boldsymbol{\phi}) = \arg\min_{(\boldsymbol{\gamma},\boldsymbol{\phi})} \mathcal{D}(q(\boldsymbol{\theta}, \boldsymbol{z} \mid \gamma, \phi) \mid\mid p(\boldsymbol{\theta}, \boldsymbol{z} \mid \alpha, \beta)) \qquad \text{(variational E-step)}$$

Given this resulting variational distribution, the expectation in equation (12) can be computed and the parameters α and β can be estimated through the classical corresponding M-step:

$$(\alpha, \beta) = \arg\max_{(\alpha,\beta)} \mathbb{E}_q \left[\log p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \beta)\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi})\right] \qquad \text{(M-step)}$$

The variational E-step has to be updated as the model parameters α and β are updated.

Due to the sparsity of large corpora, if a document in the test corpus does contain a word which doesn't appear in train corpus, the maximum likelihood estimates of the multinomial parameters would assign 0 probability to such a word. To avoid this, a smoothing is realized by adding a Dirichlet prior for each $\theta_k \in \mathbb{R}^V$, leading to the new grahical model 3.



(a) Graph of smoothed model

(b) Graph of variational family associated with smoothed model

Figure 3: Smooth LDA model

2.2 Gibbs sampling

Using Bayesian rule and Markov's blanckets:

$$\begin{aligned} p(z_{d,n} \mid z_{-(d,n)}, \boldsymbol{w}, \alpha, \eta) &= p(z_{d,n} \mid z_{-(d,n)}, \boldsymbol{w_{-(d,n)}}, \alpha, \eta) p(w_{d,n} \mid z_{d,n}, z_{-(d,n)}, \boldsymbol{w_{-(d,n)}}, \alpha, \eta) \\ &= p(z_{d,n} \mid z_{d,-n}, \alpha) p(w_{d,n} \mid z_{d,n}, z_{-(d,n)}, \boldsymbol{w_{-(d,n)}}, \eta) \end{aligned} \tag{7}$$

Given $z_{d,n} \mid \theta_d \sim Mult(\theta_d)$, and given θ_d has a Dirichlet prior α, via conjugacy, the posterior of θ_d is also a Dirichlet distribution. After T observations, noting N_d^k the number of times topic k was assigned in document d, $\theta_d \mid \alpha, \{z_{d,t}\}_{t=1}^T \sim Dir(\alpha_1 + N_d^1, \ldots, \alpha_K + N_d^K)$, and the predictive probability of a new observation $z_{d,T+1}$ is given by the posterior mean of the Dirichlet distribution:

$$p(z^k_{d,T+1} = 1 \mid \alpha, \{z_{d,t}\}^T_{t=1}) = \mathbb{E}(\theta^k_d \mid \alpha, \{z_{d,t}\}^T_{t=1}) = \frac{\alpha_k + N^k_d}{\sum^K_{j=1}\left(\alpha_j + N^j_d\right)} \tag{8}$$

Given $z^k_{d,n} = 1$, $w_{d,n} \mid z^k_{d,n} = 1, \beta_k \sim Mult(\beta_k)$, and the same reasoning applies. The resulting conditional probabilities for $z_{d,n}$ and $w_{d,n}$ are the following:

$$p(z^k_{d,n} = 1 \mid z_{d,-n}, \alpha) = \frac{\alpha_k + N^k_{d,-n}}{\sum^K_{j=1}\left(\alpha_j + N^j_{d,-n}\right)} \tag{9}$$

$$p(w_{d,n} \mid z^k_{d,n} = 1, z_{-(d,n)}, \boldsymbol{w_{-(d,n)}}, \eta) = \frac{\lambda_{w_{d,n}} + C^{w_{d,n}}_{k,-(d,n)}}{\sum^V_{v=1}\left(\lambda_v + C^v_{k,-(d,n)}\right)} \tag{10}$$

where $N^k_{d,-n}$ is the number of times topic k was assigned in document d excluding word $w_{d,n}$, and $C^{w_{d,n}}_{k,-(d,n)}$ is the number of time the word $w_{d,n}$ of vocabulary V is assigned to topic k in all documents, excluding current word $w_{d,n}$.

The full conditional distribution in equation (7) can be then sampled using equations (10) and (9):

$$p(z_{d,n} \mid z_{-(d,n)}, \boldsymbol{w}, \alpha, \eta) = \frac{\alpha_k + N^k_{d,-n}}{\sum^K_{j=1}\left(\alpha_j + N^j_{d,-n}\right)} \frac{\lambda_{w_{d,n}} + C^{w_{d,n}}_{k,-(d,n)}}{\sum^V_{v=1}\left(\lambda_v + C^v_{k,-(d,n)}\right)} \tag{11}$$

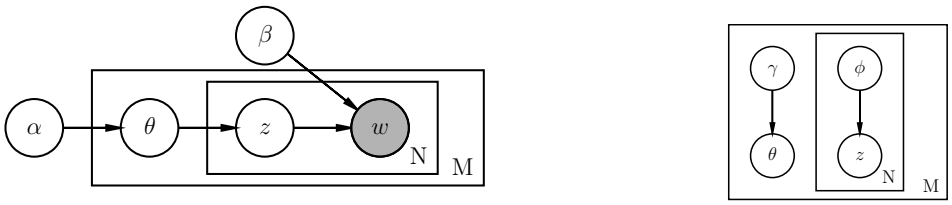
which allows to sample from the inferred posterior distribution for Bayesian inferance.

$$p(\mathbf{w}, \mathbf{z}) = \int \Big(\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n) \Big) p(\theta) d\theta$$

LDA makes use of the exchangeability of the words and topics within documents and this representation of the joint distribution over words and topics. Moreover it adds a Dirichlet prior over θ and model the conditional distributions $p(w_n|z_n)$ as multinomials parameterized by β as the graphical models shows in Figure 1 as well as the following N-word document generation procedure makes clear.

Algorithm 1 N-word document generation

$\theta \sim Dirichlet(\alpha)$
for $n = 1 : N$ **do**
  Choose topic $z_n \sim Categorical(\theta)$
  Choose word $w_n \sim Categorical(\beta_{z_n})$, that is to say $p(w_n^j = 1|z_n^i = 1, \beta) = \beta_{ij}$
end for



(a) Graphical model representation of LDA

(b) Graphical model representation of the variational distribution in Section Inference

Figure 4

In Figure 4a M is the number of documents and N the number of words. It should be again emphasized that θ and $\mathbf{z}$ are latent variables while α and β are parameters of the model that are to be learned from data. Figure **??** shows an example with $M = 2$ documents and $N = 2$ words per document with the plates unfolded.

Latent variable models usually takes the advantage of the simplicity and flexibility of introducing unobserved variable to cut modeling into subproblems. For example Gaussian Mixture Models introduce a "soft-assignment" variable that permit to group observed variables and model according to the group. In the LDA model at the word-level the topic latent variable z reflects the intuitive idea that words are drawn accoding to a fix number of different discrete processes, the number of topics. The topic mixture θ sampled once per document is another latent variable that governs the distribution over topics for one document, the topic mixture at document-level favors certain topics over other. One word w "belongs" to a topic through z and the topic z "belongs" to a topic mixture through θ.

LDA has conceptual advantages over previously applied mixture models for text data. These advantages will not be covered in this review.

3 Inference

The inferential problem tackled in the original article is that of computing the conditional distribution of the hidden variables (θ and $\mathbf{z}$):

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

However it is intractable since computing the marginal $p(\mathbf{w} \mid \alpha, \beta)$ involves integrating out the topic mixture θ and summing over the topics $\mathbf{z}$ a quantity coupling θ and β. The nodes w in Figure 4a have several parents that need to be marginalizing out (θ and β), this makes the marginal intractable. In general when we are presented with a non-tree like graphical model we have to resort to approximation techniques; sum-product algorithm for exact inference does not work.

We wish to approximate the intractable $p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)$. Two kinds of approximate inference techniques can be undertaken to solve this problem: stochastic techniques involving sampling such as MCMC, and variational inference which is deterministic. Variational inference also comes with several flavors. In general it consists of approaching the target distribution with an adujstable lower-bound taken from a restricted family of distributions, for example distributions that factorizes (mean field).

In our case the family considered is derived from simplifications of the original graphical model of Figure 4a. The authors propose the representation given in Figure 4b. The variational parameters are ϕ and γ and the variational distributions associated with this graphical model indexed by these parameters read:

$$q(\theta, \mathbf{z}|\phi, \gamma) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n)$$

γ is the parameter for the Dirichlet distribution followed by θ while ϕ is the parameter for the multinomial distribution followed by z_n. Similarly to the derivation of the EM algorithm, finding a lower-bound on $p(\mathbf{w}|\alpha, \beta)$ involves using Jensen's inequality on the expression $\log(p(\mathbf{w}|\alpha, \beta))$ and just like we have seen in the course the margin of the inequality :

$$\log p(\mathbf{w}|\alpha, \beta) \geq \mathbb{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - \mathbb{E}_q[\log q(\theta, \mathbf{z}|\gamma, \phi)] \tag{12}$$

is the Kullback-Leibler divergence between the variational distribution $q(\theta, \mathbf{z}|\phi, \gamma)$ and the target distribution $p(\theta, \mathbf{z}|\mathbf{w}|\alpha, \beta)$: $\mathcal{D}(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}, \alpha, \beta))$. Thus, maximizing the lower-bound is equivalent to minimizing the KL divergence so that the variational distribution adjusts to the target p as desired. The authors show how to maximize the right-side of the inequality by optimizing with respect to ϕ then with respect to γ and it follows two update equations of ϕ and γ that are coupled so are solved with an iterative fixed point algorithm. It leads to the following E-step:

$$(\gamma_d, \phi_d) = \arg\max_{(\boldsymbol{\gamma}, \boldsymbol{\phi})} \mathbb{E}_q \left[\log p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}_d \mid \boldsymbol{\alpha}, \boldsymbol{\beta})\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\theta}, \boldsymbol{z} \mid \gamma, \phi)\right] \qquad \text{(E-step)}$$

4 Parameters estimation

Considering the entire set of documents $\mathbb{D} = (\boldsymbol{w_1}, \ldots, \boldsymbol{w_M})$, with the hypothesis of exchangeability of documents, the likelihood (12) can be summed for each document, leading to the following M-step for the document $\boldsymbol{w_d}$:

$$(\boldsymbol{\alpha}, \boldsymbol{\beta})^{(t+1)} = \arg\min_{(\boldsymbol{\alpha},\boldsymbol{\beta})} \sum_{d=1}^{M} \left(\mathbb{E}_q \left[\log p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}_d \mid \boldsymbol{\alpha}, \boldsymbol{\beta})\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\theta}, \boldsymbol{z} \mid \gamma, \phi)\right]\right) \quad \text{(M-step)}$$

5 Generalization

Note that in our review, we assumed that all documents of the corpus had the same length N. This assumption clearly not always holds in real data sets. However the author proposes

6 Conclusion

This article The authors applied the LDA model for several tasks such as document modeling and document classification. At the time it outperformed .The literature on topic modeling that followed this article is abundant. Other approximation techniques have proven to be successful, for example Gibbs sampling.

Note that in our review, we have

topic 0	topic 1	topic 2	topic 3	topic 4
bush	city	soviet	percent	court
president	people	government	million	police
campaign	new	people	year	federal
house	years	police	billion	case
congress	two	military	new	state

Figure 5: A LDA model is learned on 5 topics with 2246 documents from the Associated Press corpus using a c-implementation (https://github.com/blei-lab/lda-c). The figure shows

6.1 Advantages

1. LDA is an effective tool for modeling

6.2 Limitations

1. Must know the number of topics in advance
2. Dirichlet topic distribution cannot capture correlations among topics

3.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation", *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.