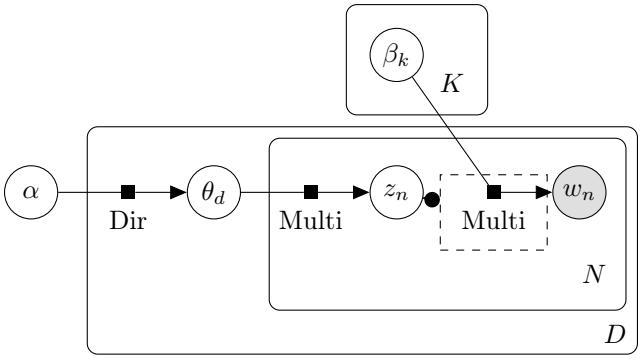


FIGURE 1

PGM REPORT
Latent Dirichlet Allocation

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

The paper we are working on is **Latent Dirichlet Allocation** [1].

1 Model

Latent Dirichlet Allocation is a probabilistic model for corpora of text documents or other collections of discrete data. We will use the language of text in this review.

It includes two level of latent variables : the so-called topics $\mathbf{z}$ at word-level (drawn one for each word) and the topic mixtures θ at document-level (drawn once for each document). Observed variables are the words w. The parameters of the model are of two kinds : α that governs the distributions of the topic mixtures θ_d (d indexes the documents) through a Dirichlet distribution and β that governs the distributions of words w_n conditioned on a topic z_n (n indexes the words) through a categorical distribution. For a fixed vocabulary size V (range of values that words w can take) and a fixed number of topics k (range of values that topics z can take) α is a k-vector and β is a $k \times V$ matrix. Words and topics are discrete variables taking respectively V and K values and as such are classically encoding as one-hot vectors.

A document is assumed to be a "bag of words" that is to say that the order of the words in the document does not matter. In a statistical point of view, words $\mathbf{w} = \{w_1, \ldots, w_N\}$ being considered as random variables, it amounts to stating infine exchangeability of the collection of words $\mathbf{w}$. Actually the topics $\mathbf{z}$ are similarly exchangeable with a document. De Finetti representation theorem therefore concludes that the distribution $p(\mathbf{w}, \mathbf{z})$ is a mixture over a latent parameter θ, that is that we draw a parameter θ and then all the variables are drawn independently conditioned on this parameter. The joint distribution

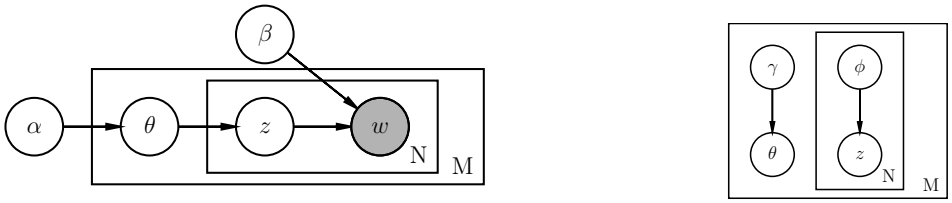
over topics and words of one document of N words reads :

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

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 2

In Figure 2a 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.

2 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 2a 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 2a. The authors propose the representation given in Figure 2b. 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 :

$$\log p(\mathbf{w}|\alpha, \beta) \geq \mathbb{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - \mathbb{E}_q[\log q(\theta, \mathbf{z}|\gamma, \phi)] \tag{1}$$

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 :

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

3 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 (1) can be summed for each document, leading to the following M-step for the document $\boldsymbol{w_d}$:

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

4 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

5 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 3 – 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

5.1 Advantages

1. LDA is an effective tool for modeling

5.2 Limitations

1. Must know the number of topics in advance
2. Dirichlet topic distribution cannot capture correlations among topics
3.

Références

[1] D. M. Blei, A. Y. Ng et M. I. Jordan, "Latent dirichlet allocation", *Journal of machine Learning research*, t. 3, n° Jan, p. 993–1022, 2003.