A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation

Sasha Collin - Paul Martin
Ecole Normale Superieure Paris-Saclay
Master MVA

Abstract

In this study, we present three different inference procedures for the Latent Dirichlet allocation (LDA), i.e. approximate variational bayes inference, collapsed Gibbs sampling and collapsed variational bayesian inference [3]. We then compare their performances on the dataset KOS and on a new dataset containing Reuters news.

1 Latent Dirichlet Allocation

The Latent Dirichlet Allocation is a generative model which models the documents of a corpus as a mixture of unknown topics. We assume that there are K underlying topics in a corpus composed of D documents and W unique words. In general, we remove stop words and rare words in order to use a meaningful document-word matrix. The generative process is summed up in Figure 1 and works this way:

- For each topic $k \in [\![1; K]\!]$, draw a distribution over the vocabulary $\phi_k \sim \mathcal{D}(\beta)$
- For each document $j \in [\![1, D]\!]$,
  - draw a distribution over topics $\theta_j \sim \mathcal{D}(\alpha)$
  - for $i \in [\![1, N_j]\!]$,
    * draw a topic assignment $z_{ji} \sim Multinomial(\theta_j)$
    * draw a word $w_{ji} \sim Multinomial(\phi_{z_{ji}})$

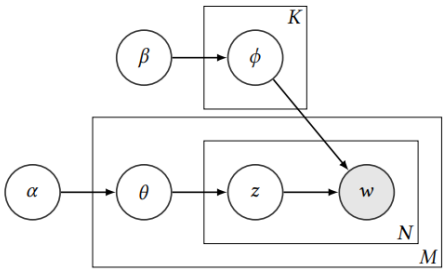

Figure 1: Graphical model for LDA [1]

2 Approximate Variational Bayes (VB) Inference

We approximate the posterior distribution $p(w|\alpha, \beta)$ by another distribution which is simpler denoted $\tilde{q}(z, \theta, \phi)$. We assume that this posterior distribution is fully factorized, that is to say :

$$\tilde{q}(z, \theta, \phi) = \prod_{ij} \tilde{q}(z_{ij}|\tilde{\gamma}_{ij}) \prod_{j} \tilde{q}(\theta_j|\tilde{\alpha}_j) \prod_{k} \tilde{q}(\phi_k|\tilde{\beta}_k)$$

With $\tilde{q}(.|\tilde{\gamma}_{ij})$ a Multinomial of parameter $\tilde{\gamma}_{ij}$, $\tilde{q}(.|\tilde{\alpha}_j)$ and $\tilde{q}(.|\tilde{\beta}_k)$ respectively two Dirichlet distribution of parameters $\tilde{\alpha}_j$ and $\tilde{\beta}_k$.

We introduce the Evidence Lower BOund of the model (ELBO) of the model :

$$\log p(w|\alpha, \beta) \geq \mathcal{L} = \mathbb{E}_{\tilde{q}}(\log p(w, z, \phi, \theta|\alpha, \beta)) - \mathbb{E}_{\tilde{q}}(\log \tilde{q}(z, \theta, \phi))$$

We can rewrite the lower bound :

$$\mathcal{L} = \sum_{j=1}^{D} \mathbb{E}_{\tilde{q}}(\log\, p(w_j|z_d, \theta_d, \phi_j)) + \mathbb{E}_{\tilde{q}}(\log\, p(z_j|\theta_j)) + \mathbb{E}_{\tilde{q}}(\log\, p(\theta_j|\alpha)) + \mathbb{E}_{\tilde{q}}(\log\, p(\phi_j|\beta))/D$$
$$- \mathbb{E}_{\tilde{q}}(\log\, \tilde{q}(z_j|\tilde{\gamma}_j)) - \mathbb{E}_{\tilde{q}}(\log\, \tilde{q}(\theta_j|\tilde{\alpha}_j)) - \mathbb{E}_{\tilde{q}}(\log\, \tilde{q}(\phi|\tilde{\beta}))/D$$

Maximizing the ELBO is equivalent to minimize the Kullback-Leibler divergence between the true posterior and the approximate posterior. We use the EM algorithm to compute the expectation and then to update the variational parameters so as to maximize the expectation. The updates are described in the algorithm 1 and we have the guarantee that the algorithm converge toward a local maxima. We can run several time the algorithm with different initialization and keep the best local maxima found. We have the following algorithm :

Input: Document Word Matrix, number of topic K, number of iterations n_{iter}, α, β
Initialize $\beta \in \mathbb{R}^{KW}$ randomly
for $n = 1, \dots n_{iter}$ **do**
  for $j = 1, \dots D$ **do**
    while $\|\gamma_d^{new} - \gamma_d^{old}\|_1 > \epsilon$ **do**
      $\tilde{\gamma}_{ijk} \propto \exp(\Psi(\tilde{\alpha}_{jk}) + \Psi(\tilde{\beta}_{kw_{ij}}) - \Psi(\sum_w \tilde{\beta}_{kw}))$
      $\tilde{\alpha}_{jk} = \alpha + \sum_w \tilde{\gamma}_{wjk}$
    end
  end
  $\tilde{\beta}_{kw} = \beta + \sum_{j=1}^{D} \sum_{i=1}^{W} \mathbb{1}(w_{ij} = w)\tilde{\gamma}_{ijk}$
end

Algorithm 1: EM algorithm for LDA

We also have an explicit expression for the variational bound given in [2]:

$$\mathcal{L} = \sum_{j=1}^{D} \sum_{w} n_{dw} \sum_{k} \tilde{\gamma}_{dwk}(\mathbb{E}_{\tilde{q}}(\log\, \theta_{dk}) + \mathbb{E}_{\tilde{q}}(\log\, \phi_{kw}) - \log\, \tilde{\gamma}_{dwk})$$
$$- \log\, \Gamma(\sum_k \tilde{\alpha}_{dk}) + \sum_k (\alpha - \tilde{\alpha}_{dk})\mathbb{E}_{\tilde{q}}(\log\, \theta_{dk}) + \log\, \Gamma(\tilde{\alpha}_{dk})$$
$$(\sum_k -\log\, \Gamma(\sum_w \tilde{\beta}_{kw}) + \sum_w (\beta - \tilde{\beta}_{kw})\mathbb{E}_{\tilde{q}}(\log\, \beta_{kw}) + \log\, \Gamma(\tilde{\beta}_{kw}))/D$$
$$+ \log\, \Gamma(K\alpha) - K\log\, \Gamma(\alpha) + (\log\, \Gamma(W\beta) - W\log\, \Gamma(\beta))/D)$$

3 Gibbs Sampling

The aim of Gibbs Sampling is to do sampling from the complete distribution. Generally, sampling directly is untractable but when the conditional distribution can be easily computed, we take benefit of it by sampling all the latent variables conditioned with all the others. Besides, the priors over ϕ and θ are conjugate (the Dirichlet distribution is a conjugate prior for the multinomial distribution) therefore we can 'collapse' these priors by integrating them. We start as proposed by [1] from the joint distribution :

$$\begin{aligned} p(w, z|\alpha, \beta) &= p(z|\alpha)p(w|z, \beta) \\ &= \int p(z|\theta)p(\theta|\alpha)\mathrm{d}\theta \int p(w|x, \phi)p(\phi|\beta)\mathrm{d}\phi \\ &= (\prod_{j=1}^{D} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k) \prod_{k=1}^{K} \Gamma(\alpha_k + n_{j,k,.})}{\Gamma(\sum_{k=1}^{K} \alpha_k + n_{j,k,.}) \prod_{k=1}^{K} \Gamma(\alpha_k)})(\prod_{k=1}^{K} \frac{\Gamma(\sum_{w=1}^{W} \beta_w) \prod_{w=1}^{W} \Gamma(\beta_w + n_{.,k,w})}{\Gamma(\sum_{w=1}^{W} \beta_w + n_{.,k,w}) \prod_{w=1}^{W} \Gamma(\beta_w)}) \end{aligned} \tag{1}$$

Then by conditioning by all the latent variables we have :

$$p(z_{ij} = k|z^{-ij}, w, \alpha, \beta) = \frac{(\alpha + n_{jk.}^{-ij})(\beta + n_{.kw_{ij}}^{-ij})(W + n_{.k.}^{-ij})^{-1}}{\sum_{k'=1}^{K}(\alpha + n_{jk'.}^{-ij})(\beta + n_{.k'w_{ij}}^{-ij})(W + n_{.k'.}^{-ij})^{-1}}$$

From this, we can deduce that :

$$\theta_{j,k} \approx \frac{\alpha + n_{jk.}}{\sum_{k'=1}^{K} \alpha + n_{jk'.}} \quad \phi_{k,w} \approx \frac{\beta + n_{.kw}}{\sum_{w'=1}^{W} \beta + n_{.kw'}}$$

Given a new corpus $\mathcal{C}$, we can estimate the log-likelihood of this corpus with :

$$\mathcal{L}(\mathcal{C}) = \sum_{j=1}^{D} \sum_{w=1}^{W} \log(\sum_{k=1}^{K} \theta_{jk}\phi_{kw_{ij}})$$

In order to have a more robust estimate, we can average this results using several samples of the Markov chain generated by the Gibbs' algorithm. This algorithm is pretty easy to implement and not too costly computationally although it's quite difficult to know when the chain has converged. In order to reduce the complexity, we keep the estimate N_{wkj}, N_{jk} and N_{kw} that we update afer each sampling so as to avoid full computation of these high dimension matrices.

4 Collapsed Variational Bayesian (CVB) Inference for LDA

Contrary to classical VB inference (see section 1), Collapsed Variational Bayesian Inference models the dependence of the parameters θ and ϕ on the latent variables z in an exact manner. However, as in VB inference, latent variables z are still assumed to be mutually independent. Thus, we approximate the posterior distribution as:

$$\hat{q}(z, \theta, \phi) = \hat{q}(\theta, \phi|z)\Pi_{ij}\hat{q}(z_{ij}|\hat{\gamma}_{ij}) \tag{2}$$

where $\hat{\gamma}(z_{ij}|\hat{\gamma}_{ij})$ is again a multinomial with parameter $\hat{\gamma_{ij}}$. The variational free energy can be rewritten as follows:

$$\begin{aligned}
\hat{\mathcal{F}}(\hat{q}(z, \theta, \phi)) &= \hat{\mathcal{F}}(\hat{q}(z)\hat{q}(\theta, \phi|z)) = E_{\hat{q}(z)\hat{q}(\theta,\phi|z)}[-\log p(w, z, \theta, \phi|\alpha, \beta)] - \mathcal{H}(\hat{q}(z)\hat{q}(\theta, \phi|z)) \\
&= E_{\hat{q}(z)}[E_{\hat{q}(\theta,\phi|z)}[-\log p(w, z, \theta, \phi|\alpha, \beta)] - \mathcal{H}(\hat{q}(\theta, \phi|z))] - \mathcal{H}(\hat{q}(\theta, \phi|z)) \\
&= E_{\hat{q}(z)}[E_{\hat{q}(\theta,\phi|z)}[-\log p(\theta, \phi|w, z, \alpha, \beta) - \log p(w, z|\alpha, \beta)] - \mathcal{H}(\hat{q}(\theta, \phi|z))] - \mathcal{H}(\hat{q}(\theta, \phi|z))
\end{aligned}$$

First minimizing with respect to $\hat{q}(\theta, \phi|z)$, the minimum is achieved at the true posterior $\hat{q}(\theta, \phi|z) = p(\theta, \phi|x, z, \alpha, \beta)$, as there is no constraint on the posterior form. Thus the free energy becomes:

$$\begin{aligned}
\hat{\mathcal{F}}(\hat{q}(z)) &= E_{\hat{q}(z)}[E_{p(\theta,\phi|x,z,\alpha,\beta)}[-\log p(\theta, \phi|w, z, \alpha, \beta) - \log p(w, z|\alpha, \beta)] - \mathcal{H}(p(\theta, \phi|x, z, \alpha, \beta))] - \mathcal{H}(\hat{q}(\theta, \phi|z)) \\
&= E_{\hat{q}(z)}[\mathcal{H}(p(\theta, \phi|x, z, \alpha, \beta)) - \log p(w, z|\alpha, \beta) - \mathcal{H}(p(\theta, \phi|x, z, \alpha, \beta))] - \mathcal{H}(\hat{q}(\theta, \phi|z)) \\
&= E_{\hat{q}(z)}[-\log p(w, z|\alpha, \beta)] - \mathcal{H}(\hat{q}(\theta, \phi|z))
\end{aligned} \tag{3}$$

Thus, we have marginalized out θ and ϕ. As the assumptions made over the variational posterior for CVB are weaker than those done for VB, we have:

$$\hat{\mathcal{F}}(\hat{q}(z)) \leq \tilde{\mathcal{F}}(\tilde{q}(z)) \tag{4}$$

which shows that CVB better approximates the posterior distribution. Minimizing now 3 with respect to $\hat{\gamma}_{ijk} = \hat{q}(z_{ij} = k)$, we obtain the following updates:

$$\hat{\gamma}_{ijk} \propto \exp(E_{\hat{q}(z^{-ij})}[p(x, z^{-ij}, z_{ij} = k|\alpha, \beta]) \tag{5}$$

Then, plugging 5 in 1, and using the fact that $\forall \eta > 0, \log \frac{\Gamma(\eta+n)}{\Gamma(\eta)} = \sum_{l=0}^{n-1} \log(\eta + l)$, we obtain:

$$\hat{\gamma}_{ijk} \propto \exp(E_{\hat{q}(z^{-ij}}[\log(\alpha + n_{jk.}^{-ij}) + \log(\beta + n_{.kw_{ij}}^{-ij} - \log(W\beta + n_{.k.}^{-ij}]) \tag{6}$$

However, computing $\hat{\gamma}_{ijk}$ using 6 is too expensive in practice, and the authors of [3] propose a Gaussian approximation which in practice works very accurately and has a low computational cost, and leads to the following simplified and easily implementable expression of $\hat{\gamma}_{ijk}$:

$$\hat{\gamma}_{ijk} \propto \left(\alpha + E_{\hat{q}}[n_{jk.}^{-ij}]\right)\left(\beta + E_{\hat{q}}[n_{.kw_{ij}}^{-ij}]\right)\left(W\beta + E_{\hat{q}}[n_{.k.}^{-ij}]\right)^{-1}$$
$$\exp\left(-\frac{Var_{\hat{q}}[n_{jk.}^{-ij}]}{2\left(\alpha + E_{\hat{q}}[n_{jk.}^{-ij}]\right)^2} - \frac{Var_{\hat{q}}[n_{.kw_{ij}}^{-ij}]}{2\left(\beta + E_{\hat{q}}[n_{.kw_{ij}}^{-ij}]\right)^2} + \frac{Var_{\hat{q}}[n_{.k.}^{-ij}]}{2\left(W\beta + E_{\hat{q}}[n_{.k.}^{-ij}]\right)^2}\right) \tag{7}$$

where

$$E_{\hat{q}}[n_{jk.}^{-ij}] = \sum_{i' \neq i} \hat{\gamma}_{i'jk} \qquad Var_{\hat{q}}[n_{jk.}^{-ij}] = \sum_{i' \neq i} \hat{\gamma}_{i'jk}(1 - \hat{\gamma}_{i'jk})$$

We have similar expressions for the expectation and variances under $\hat{q}$ of $n_{.kw_{ij}}^{-ij}$ and $n_{.k.}^{-ij}$. This new expression 7 enables to design a similar iterative algorithm than the one presented in section 1 to compute the posterior distribution of the latent variable z.

5 Experiments

We define the log-perplexity [2] of a dataset as the mean of the inverse marginal probability of each word in the considered dataset:

$$\text{log-perplexity}(n^{data}, \alpha, \phi) = -\sum_i \log p(n_i^{data}|\alpha, \phi))/(\sum_{i,w} n_{iw}^{data})$$

where h_i^{data} is the vector of word counts for the ith document. $\log p(n_i^{data}|\alpha, \phi)$ can not be computed directly, so we derive an lower bound. Noticing that:

$$\log p(w|\alpha, \phi) = E[\log p(\theta, z, w|\alpha, \phi)] - E[\log(\theta, z)] + KL(q(\theta, z\|p(\theta, z))$$

where $KL(q(\theta, z\|p(\theta, z)) \geq 0$. Thus, we obtain the following lower bound for the per-word log-probabilities:

$$\text{log-perplexity}(n^{data}, \alpha, \phi) \leq -(\sum_i \mathbb{E}_q[\log p(n_i^{data}, \theta_i, z_i|\alpha, \phi)] - \mathbb{E}_q[\log q(\theta_i, z_i)])/(\sum_{i,w} n_{iw}^{data}) \tag{8}$$

In the following experiments, we plot the evolution of this lower bound to assess the performances of each algorithm.

Algorithm comparison on KOS dataset

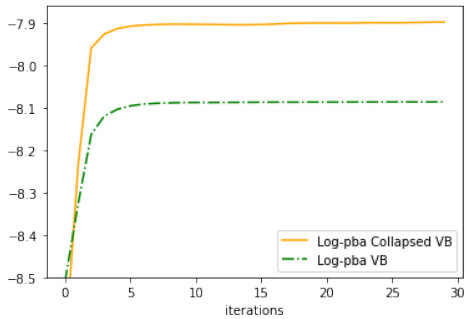

Figure 2: Per-word log-probabilities as function of number of iterations

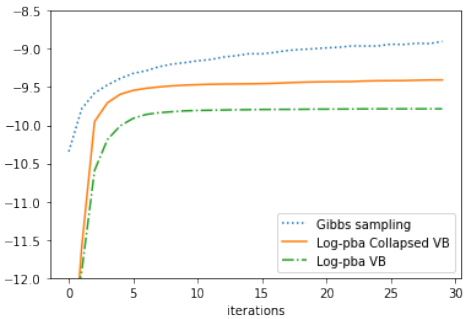

Figure 3: Test set per-word log-probabilities as function of number of iterations

We use a subset of KOS dataset containing 1000 documents with and average size of 136 words and 6909 different words. As proposed in [3], we do a train/test splitting by assigning 10% of the words to the test set. We fit 8 topic ($K = 8$) and use $\alpha = 0.1$ and $\beta = 0.1$. We observe in Figure 2 that both standard and collapsed variational Bayes converge in few iterations (≈ 10), but that the collapsed version converges toward a better local maxima than the standard algorithm. It's reasonable since the collapsed version perfectly models the dependence of the parameters on latent variables and only assumes that latent variables are collectively independent. We do not add collapsed Gibbs to this benchmark since there is no training phase, we only sample a Markov chain. On the test set (Figure 3), we notice that the performances are lower as the training does not take into account this dataset. Gibbs sampling performs better on this dataset, which is not surprising as it is supposed to be exact with enough samplings, and we could still add more iterations in order to improve the final log probabilities.

Results on Reuters Dataset

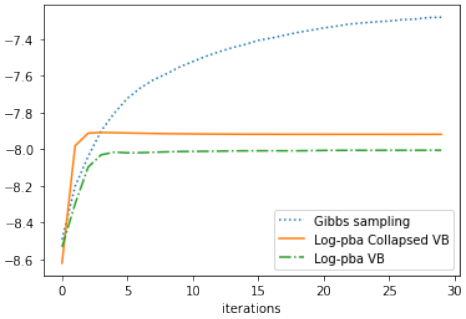

Figure 4: per word log probabilities as functions as functions of number of iterations

Topic 1	Topic 2	Topic 3
year	10th	hao
world	turks	thich
told	ottoman	thien
first	armenians	buddhism
last	slavs	hanoi
mother	thodoridis	monks
people	athos	temple
years	venizelos	dent
pope	byzantine	vietnam
church	salonika	buddhist

Figure 5: 10 most frequent word for three topics obtained with Collapsed variational Bayes

We then use a dataset containing Reuters news containing 395 articles with an average length of 212 words and 4258 different words. We use $K = 10$ and only plot the results on the test set (see Figure 4). We obtain very similar results to the ones shown previously: the collapsed VB performs better than the classical VB algorithm, and as in [3], the CVB converges faster than collapsed Gibbs sampling, but Gibbs sampling achieves a better solution in the end. In the table 5, we display the 10 first words with the highest probability to belong to a topic class (3 among 10), the words give a good idea of the topic and are very coherent as we have the lexical field of "Christian civilization" for topic 1, of "Oriental people" for topic 2 and "Vietnam" for topic 3. Thus these models prove to be very efficient when it comes to infer unknown topics and may be used for document classification after labelling these ones.

6 Conclusion

Overall, we have shown that thanks to a better modeling of the dependence of parameters on latent variables, collapsed variational Bayesian inference algorithm for LDA performs better than the classical variational bayes algorithm, which relies on a stronger assumption (independence of the parameters).

The algorithm proposed is computationally reasonable and allows an easy parameters estimation even though it's more expensive than Gibbs sampling. On top on that, the results are on par with Gibbs sampling with a restricted number of iterations as illustrated by the train and test plots on two datasets.

References

[1] Chase Geigle. Inference methods for latent dirichlet allocation. *University of Illinois at Urbana-Champaign, USA*, pages 1–29, 2016.

[2] Matthew Hoffman, David Blei, and Francis Bach. Online learning for latent dirichlet allocation. volume 23, pages 856–864, 11 2010.

[3] Yee W Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. Technical report, CALIFORNIA UNIV IRVINE SCHOOL OF INFORMATION AND COMPUTER SCIENCE, 2007.