ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

On the equivalence between algorithms for Non-negative Matrix Factorization and Latent Dirichlet Allocation

Thiago de Paulo Faleiros and Alneu de Andrade Lopes *

University of São Paulo - Institute of Mathematics and Computer Science
13560-970 São Carlos, SP - Brazil
Email: {thiagopf, alneu}@icmc.usp.br

Abstract. LDA (Latent Dirichlet Allocation) and NMF (Non-negative Matrix Factorization) are two popular techniques to extract topics in a textual document corpus. This paper shows that NMF with Kullback-Leibler divergence approximate the LDA model under a uniform Dirichlet prior, therefore the comparative analysis can be useful to elucidate the implementation of variational inference algorithm for LDA.

1 Introduction

The equivalence between NMF (Non-negative Matrix Factorization) and PLSI (Probabilistic Latent Semantic Indexing) have been discussed in several works [1, 2]. Ding and colleagues [3] demonstrate that both NMF and PLSI optimize the same objective function. Although LDA (Latent Dirichlet Allocation) be a full Bayesian counterpart and maximum-a-posteriori view of PLSI [4], the equivalence between NMF and LDA is not well defined. However, there are evidences that such intrinsic relations also exists [5, 6]. Here, we want to clarify these relationship by demonstrating that NMF-KL (NMF with Kullback-Leibler divergence) approximate the LDA model, and compare the multiplicative algorithm to solve NMF-KL with the variational inference algorithm for LDA.

2 NMF

The NMF method approximately factorizes a matrix of which all the elements have non-negative values into two matrices with elements having non-negative values. NMF for documents factorizes a document-term matrix $F = (F_{j,i})$, with dimension $\mathcal{D} \times \mathcal{W}$, where $\mathcal{D}$ is the number of documents, $\mathcal{W}$ is the number of words, and each entry $F_{j,i}$ is the frequency of word w_i in document d_j, into two matrices A and B such as $F \approx AB$, where A is a $\mathcal{D} \times \mathcal{K}$ matrix and B is a $\mathcal{K} \times \mathcal{W}$ matrix. The value of $\mathcal{K}$ is the number of components.

The factors matrices A and B are obtained by optimizing a cost function which can be set by using some distance measure. There are different types of cost functions [7]. Here, we are interested in NMF with KL-Divergence, defined as

$$Q_{KL-NMF} = \sum_{j,i} \left(f_{j,i} \log \frac{f_{j,i}}{(AB^T)_{j,i}} - f_{j,i} + (AB^T)_{j,i} \right). \quad (1)$$

*Thanks to FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo, Projeto 2011/23689-9) for financial support.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

The simplest technique to solve the optimization of Equation 1 is by Gradient descent method. Gradient descent based method can be implemented by the following "multiplicative update rules"

$$A_{j,k} = A_{j,k} \frac{\sum_i B_{k,i} F_{j,i}/(AB)_{j,k}}{\sum_q B_{k,q}}, \qquad B_{k,i} = B_{k,i} \frac{\sum_j A_{j,k} V_{j,i}/(AB)_{j,i}}{\sum_p A_{p,k}} \tag{2}$$

3 Variational Bayes Inference for LDA

LDA is a generative topic model for documents. The basic idea is that documents are represented as a random mixtures over latent topics, where each topic is characterized by a distribution over words [8]. In a simplified LDA formulation, the word probabilities are parametrized by a $\mathcal{K} \times \mathcal{W}$ matrix β. A topic k $(1 \leq k \leq \mathcal{K})$ is a discrete distribution over words with probability vector β_k. Each document d_j maintains a separated distribution θ_j that describes the contribution of each topic. Implementations of LDA inference algorithms typically use symmetric Dirichlet prior over $\Theta = \{\theta_1, \ldots, \theta_{\mathcal{D}}\}$, in which the concentration parameter α is fixed. A topic distribution of a document d_j and a word w_i is associate in a distribution variable $z_{j,i}$.

Given the parameter α and β, the joint distribution of a topic mixture Θ is given by

$$p(\Theta, z, w|\alpha, \beta) = \prod_{d_j \in \mathcal{D}} p(\theta_j|\alpha) \sum_{n=1}^{N_j} p(z_{j,n}|\theta_j) p(w_{i,n}^{d_j}|z_{j,n}, \beta). \tag{3}$$

where N_j is the number of tokens words in document d_j.

A wide variety of approximate inference algorithms can be considered for LDA. Here, we describe the variational inference algorithm. The main idea behind the variational method is to use a distribution with its own parameters replacing the posterior distribution $p(\theta, z, w|\alpha, \beta)$. This variational distribution for LDA is described as

$$q(\theta_j, z_j|\gamma_j, \varphi_j) = q(\theta_j|\gamma_j) \prod_{n=1}^{N} q(z_{j,n}|\theta_{j,n}), \tag{4}$$

where γ_j and φ_j are the variational parameters respectively corresponding to LDA real distributions θ_j and z_j.

The value of variational parameters are chosen by a optimization procedure that attempts to minimizing the KL-divergence between the variational distribution and the true posterior $p(\Theta, z, w|\alpha, \beta)$.

Actually, it is not possible to minimize the KL-divergence directly. However, bounding the log likelihood of a document, $p(w|\alpha, \beta)$, and using Jensen's inequality [9] it is possible to show that minimizing the KL-divergence between the variational distribution and the true posterior distribution is equivalent to maximizing the Evidence Lower Bound (ELBO) with respect to variational parameters. The ELBO is defined by the difference between the variational expectation of real posterior distribution and the variational distribution, $E_q[\log p(\theta, z, w|\alpha, \beta)] - E_q[\log q(\theta, z)]$ [8].

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

ELBO $\mathcal{L}$ can be optimized using coordinate over the variational parameters (detailed derivation in [8]):

$$\varphi_{j,i,k} \propto \beta_{k,i} \exp\left(E_q[\log(\theta_{j,k})|\gamma]\right), \quad \gamma_{j,k} \quad = \alpha + \sum_{i=1}^{\mathcal{W}} F_{j,i}\varphi_{j,i,k}, \quad \beta_{i,w_n} = \sum_{j=1}^{\mathcal{D}}\sum_{i=1}^{\mathcal{W}} \varphi_{j,i,k} \quad (5)$$

where $F_{j,i}$ is the number of words w_i in document d_j. The expectation in the multinomial update can be computed as

$$E_q[\log(\theta_{j,k})] = \psi(\gamma_{j,k}) - \psi\left(\sum_{\hat{k}=1}^{\mathcal{K}} \gamma_{j,\hat{k}}\right), \quad (6)$$

where Ψ denotes the digamma function.

4 Comparing LDA and NMF

The correspondence between NMF-KL and variational inference algorithm for LDA follows the fact that they try to minimize the divergence between word frequency, document-topic and topic-word statistics. To clarify the relationship between NMF and LDA, we will describe NMF-KL as a relaxation of variational problem. The equivalence is reached when a relaxation of functions $\log\Gamma(\cdot)$ and $\Psi(\cdot)$ are considered in the LDA derivations.

Theorem 4.1 *The objective function of NMF with KL-Divergence is a approximation of ELBO $\mathcal{L}$ of LDA with symmetric Dirichlet priors.*

Proof Initially, we expand the ELBO $\mathcal{L}$ by using factorization of LDA joint distribution p (Equation 3) and the variational distribution q (Equation 4):

$$\begin{aligned}
\mathcal{L} \triangleq\ & E_q[\log p(\Theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\Theta, \mathbf{z})] \\
=\ & E_q[\log p(\Theta|\alpha)] + E_q[\log p(\mathbf{z}|\Theta)] + E_q[\log p(\mathbf{w}|\mathbf{z}, \beta)] - E_q[\log q(\Theta)] - E_q[\log q(\mathbf{z})] \\
& \textit{(Expanding each of the five terms in the bound)} \\
=\ & \prod_{d_j \in \mathcal{D}} \left\{ \left[\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi\left(\gamma_{j,k}\right) - \Psi\left(\sum_{l=1}^{K}\gamma_{j,l}\right)\right)\right] \right. \\
& + \left[\sum_{n=1}^{N_j}\sum_{k=1}^{K}\varphi_{j,n,k}\left(\Psi(\gamma_{j,k}) - \Psi\left(\sum_{l=1}^{K}\gamma_{j,k}\right)\right)\right] \\
& + \left[\sum_{n=1}^{N_j}\sum_{k=1}^{K}\sum_{i=1}^{V}\varphi_{n,i} w_{i,n}^{d_j}\log\beta_{k,i}\right] \\
& + \left[-\log\Gamma\left(\sum_{k=1}^{K}\gamma_{j,k}\right) + \sum_{l=1}^{K}\log\Gamma\left(\gamma_{j,l}\right) - \sum_{k=1}^{K}(\gamma_{j,k} - 1)\left(\Psi(\gamma_{j,k}) - \Psi\left(\sum_{l=1}^{K}\gamma_{j,l}\right)\right)\right] \\
& \left. + \left[-\sum_{n=1}^{N_j}\sum_{k=1}^{K}\varphi_{j,n,k}\log\varphi_{j,n,k}\right]\right\} \quad (7)
\end{aligned}$$

Now, we will approximate the Equation 7 by replacing the Gamma function $\Gamma(\cdot)$ and digamma function $\Psi(\cdot)$. The Gamma function is defined by $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$, for $x > 0$. In general, $\Gamma(x+1) = x\Gamma(x)$, and for

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

integer arguments, $\Gamma(x+1) = x!$. For practical purposes, we can consider the Stirlings approximation of function $\Gamma(\cdot)$:

$$\log\Gamma(x) = \log x! = \sum_{i=1}^{n} \log i \approx \int_{i=1}^{x} \log(i)di \approx x\log x - x. \tag{8}$$

The digamma function is defined by $\Psi(x) = \frac{d}{dx}\log\Gamma(x)$, and can be approximated by [10]

$$\Psi(n) \approx \log n - c, \tag{9}$$

where c is a constant.

We can relate γ_j distribution with the vector A_j associated to document d_j. In LDA setting, we treat β as a factor matrix B. Then, considering LDA with the fixed symmetric Dirichlet hyperparameters α, we can rewrite the ELBO using the correspondent approximation of functions gamma, Equation 8, and digamma, Equation 9:

$$\begin{aligned}
\mathcal{L} \approx \prod_{j=1}^{\mathcal{D}} \Bigg\{ & \left[\sum_{k=1}^{\mathcal{K}} (\alpha_k - 1)\left(\log\frac{A_{j,k}}{\sum_{l=1}^{\mathcal{K}} A_{j,l}}\right)\right] \\
& + \left[\sum_{i=1}^{\mathcal{W}}\sum_{k=1}^{\mathcal{K}} f_{j,i}\varphi_{j,i,k}\left(\log\frac{A_{j,k}}{\sum_{l=1}^{\mathcal{K}} A_{j,l}}\right)\right] \\
& + \left[\sum_{i=1}^{\mathcal{W}}\sum_{k=1}^{\mathcal{K}} F_{j,i}\varphi_{j,i,k}\left(\log\frac{B_{i,k}}{\sum_{p=1}^{\mathcal{W}} B_{p,k}}\right)\right] \\
& + \left[\sum_{k=1}^{\mathcal{K}}\left(A_{j,k}\left(\log A_{j,k} - 1\right) - (A_{j,k}-1)\left(\log\frac{A_{j,k}}{\sum_{l=1}^{\mathcal{K}} A_{j,l}}\right)\right)\right] \\
& + \left[\sum_{i=1}^{\mathcal{W}}\sum_{k=1}^{\mathcal{K}} -F_{j,i}\varphi_{j,i,k}\log C_{e_{j,i},k}\right]\Bigg\} \\
= & \sum_{j}^{\mathcal{D}}\sum_{i}^{\mathcal{W}}\sum_{k=1}^{\mathcal{K}}\Bigg(F_{j,i}\varphi_{j,i,k}\log\frac{\frac{A_{j,k}}{\sum_{l=1}^{\mathcal{K}} A_{j,l}}\frac{B_{i,k}}{\sum_{p=1}^{\mathcal{W}} B_{p,k}}}{\varphi_{j,i,k}} \\
& + (\alpha_k - A_{j,k})\left(\log\frac{A_{j,k}}{\sum_{l=1}^{\mathcal{K}} A_{j,l}}\right) - A_{j,k}(\log A_{j,k} - 1)\Bigg)
\end{aligned} \tag{10}$$

Considering that the vectors A_j and B_i are normalized such that $\sum_{k=1}^{\mathcal{K}} A_{j,k} = 1$ and $\sum_{p=1}^{\mathcal{W}} B_{i,p} = 1$, and defining $\mathcal{R}(A_{j,k}, \alpha_k) = (\alpha_k - A_{j,k})(\log A_{j,k}) - A_{j,k}(\log A_{j,k} - 1)$, we can rewrite Equation 10 and describe the following maximization problem

$$\begin{aligned}
\max\mathcal{L} \approx & \max\sum_{j}^{\mathcal{D}}\sum_{i}^{\mathcal{W}}\sum_{k=1}^{\mathcal{K}}\left(F_{j,i}\varphi_{j,i,k}\log\frac{A_{j,k}B_{i,k}}{\varphi_{j,i,k}} + \mathcal{R}(A_{j,k},\alpha_k)\right) \\
\approx & \min\sum_{j}^{\mathcal{D}}\sum_{i}^{\mathcal{W}}\sum_{k=1}^{\mathcal{K}}\left(F_{j,i}\varphi_{j,i,k}\log\frac{\varphi_{j,i,k}}{A_{j,k}B_{i,k}} - \mathcal{R}(A_{j,k},\alpha_k)\right)
\end{aligned} \tag{11}$$

since $\sum_{i=1}^{n} a_i \log\frac{a_i}{b_i} \leq \sum_{i=1}^{n} a_i \log\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$, for any a_i and b_i non-negative, and by adding a constant value $\sum_{j,i} F_{j,i}\log F_{j,i}$,

$$\max\mathcal{L} \leq \min\sum_{j}^{\mathcal{D}}\sum_{i}^{\mathcal{W}}\left(F_{j,i}\sum_{k=1}^{\mathcal{K}}\varphi_{j,i,k}\log\frac{\sum_{k=1}^{\mathcal{K}}\varphi_{j,i,k}}{\sum_{k=1}^{\mathcal{K}} A_{j,k}B_{i,k}} - \sum_{k=1}^{\mathcal{K}}\mathcal{R}(A_{j,k},\alpha_k)\right)$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

$$\approx \min \sum_{j}^{\mathcal{D}} \sum_{i}^{\mathcal{W}} \left(F_{j,i} \log \frac{F_{j,i}}{\sum_{k=1}^{K} A_{j,k} B_{i,k}} - \sum_{k=1}^{\mathcal{K}} \mathcal{R}(A_{j,k}, \alpha_k) \right). \tag{12}$$

The last term in Equation 12 is equivalent the NMF with KL-Divergence (Equation 1) minus the term $\mathcal{R}(A_{j,k}, \alpha_k)$. The term $\mathcal{R}(A_{j,k}, \alpha_k)$ play a important role in LDA performance, it correspond to the priors influence and induce sparsity over the document-topic distribution. When it is added to NMF, it can be considered as a regularization term restricting the values of vector A_j. Then, we can conclude that maximizing the ELBO of LDA with symmetric Dirichlet prior is proportional to minimize the NMF with KL-Divergence objective function disregarding the regularization term.

■

5 Comparing the updates equations

In practice, LDA and NMF use iterative algorithms to reach a feasible solution. Theoretically, these updates are based on distinct methods and different mathematical foundation. However, as their objective functions, we can also indicate similarities between their update equations. Then, in this section we will compare the updates of NMF-KL, Equations 2, and the LDA with variational Inference, Equations 5.

In update rule for LDA, the exponential operation over a digamma function $\Psi(x)$ approximate a linear function when $x > 0.5$ [10]. Therefore, it is possible to approximate the value of φ only with linear operation

$$\varphi_{j,i,k} \approx \beta_{k,i} \times \frac{\gamma_{j,k}}{\sum_{k^*=1}^{\mathcal{K}} \gamma_{j,k^*}}. \tag{13}$$

Thus, the value $\varphi_{j,i}$ approximate the Hadamard product of normalized vectors γ_j and β_k. The resulting factor matrix A is closely related to document-topic distribution γ, and the resulting factor matrix B is closely related to topic-word distribution β. Thus, considering these relationships, we can approximate the update of variational parameter φ as

$$\varphi_{j,i,k} \propto \left(\frac{A_{j,k} B_{k,i}}{\sum_{k^*=1}^{K} A_{j,k^*} B_{k^*,i}} \right) \tag{14}$$

Without loss of generality, we can consider a row-wise normalization in NMF B factor matrix, such that $\sum_i B_{k,i} = 1$. Then, using Equation 14, we can rewrite update of factor $A_{j,k}$ in Equation 2, as

$$A_{j,k} = \sum_{i=1}^{\mathcal{W}} F_{j,k} \varphi_{j,i,k}. \tag{15}$$

Note that the updating equation of factor A_j, Equation 15, is similar to updating equation of parameter γ_j without the parameter α, Equation 5.

The update equation of factor $B_{k,i}$ can be rewritten considering the φ approximation, Equation 14, and the last value of $A_{j,k}$ obtained in Equation 15

$$B_{k,i} = \frac{1}{\sum_j A_{j,k}} \frac{\sum_j F_{j,k} A_{j,k} B_{k,i}}{(AB)_{j,k}}$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

$$= \frac{\sum_j F_{j,k}\varphi_{j,i,k}}{\sum_j \sum_i F_{j,k}\varphi_{j,k,i}}. \quad (16)$$

By Equation 16, we can note that the value of $B_{k,i}$ is obtained by the statistics φ for a specific word w_i and topic k for every document d_j, and normalized by every word w_i in the vocabulary. It corresponds to the topic-word distribution for a topic k, represented by distribution β_k in LDA.

6 Conclusion

In this paper, we study the relationships between NMF (with KL-Divergence objective) and LDA (with variational inference algorithm). In particular, we show that a) NMF-KL in fact is a special case of LDA where we assume uniform Dirichlet prior; and b) The NMF-KL "multiplicative updates roles" can be approximated to the updates established by variational inference algorithm for LDA.

References

[1] W. Buntine. Variational extensions to em and multinomial pca. In *In ECML 2002*, pages 23–34. Springer-Verlag, 2002.

[2] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 601–602, New York, NY, USA, 2005. ACM.

[3] Chris Ding, Tao Li, and Wei Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.*, 52(8):3913–3927, April 2008.

[4] M. Girolami and A. Kabán. On an equivalence between plsi and lda. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 433–434, New York, NY, USA, 2003. ACM.

[5] Zeng J., Z. Liu, and X. Cao. Memory-efficient topic modeling. *CoRR*, abs/1206.1147, 2012.

[6] S. J. Gershman and D. M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, February 2012.

[7] D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562, The MIT Press, 55 Hayward Street Cambridge, MA 02142-1493 USA, April 2001. MIT Press.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, March 2003.

[9] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999.

[10] Isa Muqattash and Mohammed Yahdi. Infinite family of approximations of the digamma function. *Mathematical and Computer Modelling*, 43(11 - 12):1329 – 1336, 2006.