

A Study on Stochastic Variational Inference for Topic Modeling with Word Embeddings

Kana Ozaki¹ and Ichiro Kobayashi²

¹ Ochanomizu University, Japan ozaki.kana@is.ocha.ac.jp

² Ochanomizu University, Japan koba@is.ocha.ac.jp

Abstract. Probabilistic topic models based on Latent Dirichlet Allocation (LDA) is widely used to extract latent topics from document collections. In recent years, a number of extended topic models have been proposed, especially Gaussian LDA (G-LDA) has attracted a lot of attention. G-LDA integrates topic modeling with word embeddings by replacing discrete topic distributions over words with multivariate Gaussian distributions on the word embedding space. This can reflect semantic information into topics. In this paper, we use G-LDA for our base topic model and apply Stochastic Variational Inference (SVI), an efficient inference algorithm, to estimate topics. Through experiments, we could extract the topics with high coherence in practical time.

Keywords: Topic Model · Latent Dirichlet Allocation · Word Embeddings · Stochastic Variational Inference.

1 Introduction

Probabilistic topic models such as Latent Dirichlet Allocation (LDA)[1], are widely used to uncover hidden topics within text corpus. In LDA, each document may be viewed as a mixture of latent topics where each topic is a distribution over words. With statistical inference algorithms, LDA reveals latent topics using document-level word co-occurrence. In recent years, a number of extended topic models have been proposed, especially Gaussian LDA(G-LDA)[2] that integrates LDA with word embeddings has gained much attention. G-LDA uses Gaussian distribution as the topic distribution over words. Furthermore, Batmanghelich et al.[3] proposed spherical Hierarchical Dirichlet Process (sHDP) which use the von Mises-Fisher distribution as the topic distribution to model the density of words over unit sphere. They used the Hierarchical Dirichlet Process (HDP) for their base topic model and apply Stochastic Variational Inference (SVI) [14] for efficient inference. They showed that sHDP is able to exploit the semantic structures of word embeddings and flexibly discovers the number of topics. Hu et al.[4] proposed Latent Concept Topic Model (LCTM) which introduces latent concepts to G-LDA. LCTM models each topic as a distribution over latent concepts, where each concept is a localized Gaussian distribution in word embedding space. They reported that LCTM is well suited for extracting topics from short texts with diverse vocabulary such as tweets. Xun et al.[5] proposed

a correlated topic model using word embeddings. Their model enables us to exploit the additional word-level correlation information in word embeddings and directly models topic correlation in the continuous word embedding space. Nguyen et al. [6] proposed Latent Feature LDA (LF-LDA) which integrates word embeddings into LDA by replacing the topic-word Dirichlet multinomial component with a mixture of a Dirichlet multinomial component and a word embedding component. They compared the performance of LF-LDA to vanilla LDA on topic coherence, document clustering and document classification evaluations and showed that LF-LDA improves both topic-to-word mapping and document-topic assignments compared to vanilla LDA, especially on datasets with few or short documents. Kumar et. al. [7] presented an unsupervised topic model for short texts that performs soft clustering over word embedding space. They modeled the low-dimensional semantic vector space represented by word embeddings using Gaussian mixture models (GMMs) whose components capture the notion of latent topics. Their proposed framework outperforms vanilla LDA on short texts through both subjective and objective evaluation, and showed its usefulness in learning topics and classifying short texts on Twitter data for several foreign languages. Zhao et. al. [8] proposed a focused topic model where how a topic focuses on words is informed by word embeddings. Their models are able to discover more informed and focused topics with more representative words, leading to better modelling accuracy and topic quality. Moody [9] proposed a model, called lda2vec, which learns dense word vectors jointly with Dirichlet distributed latent document-level mixtures of topic vectors. His method is simple to incorporate into existing automatic differentiation frameworks and allows for unsupervised document representations geared for use by scientists while simultaneously learning word vectors and the linear relationships between them. Yao et al.[10] proposed Knowledge Graph Embedding LDA (KGE-LDA), which combines topic model and knowledge graph embeddings. KGE-LDA models document level word co-occurrence with knowledge encoded by entity vectors learned from external knowledge graphs and can extract more coherent topics and better topic representation. In this paper, we use G-LDA as our base topic model. Compared with vanilla LDA, G-LDA produces higher Pointwise Mutual Information (PMI) in each topic because it has semantic information of words as prior knowledge. In addition, because G-LDA operates on the continuous vector space, it can handle out of vocabulary (OOV) words in held-out documents whereas the conventional LDA cannot. On the other hand, the cost for estimating the posterior probability distribution for latent topics in word embedding space is costly because of dealing with the high dimensional information of words. So, it is unpractical to use the methods which take much time to estimate the posterior probability distribution such as Gibbs sampling. To reduce the cost for estimating the posterior probability distribution, G-LDA utilizes Cholesky decomposition of covariance matrix and applies Alias Sampling[11] for that. In a similar case of dealing with high dimensional data, it is also difficult to estimate latent topics in massive documents using sampling methods. To deal with this problem, Hoffman et al. [12] developed online Variational Bayes (VB) for

LDA. Their model is handily applied to massive and streaming document collections. Their proposed method, online variational Bayes, becomes well known as “Stochastic Variational Inference” [13],[14].

Referring to their approach, in this paper, we propose a method to efficiently estimate latent topics in the high dimensional space of word embeddings by adopting SVI (Stochastic Variational Inference).

2 LDA and Gaussian LDA

2.1 Latent Dirichlet Allocation (LDA)

LDA[1] is a probabilistic generative model of document collections. In LDA, each topic has a multinomial distribution β over a fixed vocabulary and each document has a multinomial distribution θ over K topics. Distributions β and θ are designed to be sampled from the conjugate Dirichlet priors parameterized by η and α , respectively. Suppose that D and N_d denote the number of documents and words in d th document, respectively.

The generative process is as follows.

1. for $k = 1$ to K
 - (a) Choose topic $\beta_k \sim \text{Dir}(\eta)$
2. for each document d in corpus D
 - (a) Draw topic distribution $\theta_d \sim \text{Dir}(\alpha)$
 - (b) for each word index n from 1 to N_d
 - (a) Draw a topic $z_n \sim \text{Categorical}(\theta_d)$
 - (b) Draw a word $w_n \sim \text{Categorical}(\beta_{z_n})$

The graphical model for LDA is shown in the left side of Figure 1.

2.2 Gaussian LDA (G-LDA)

Hu et al.[15] proposed a new method to model the latent topic in the task of audio retrieval, in which each topic is directly characterized by Gaussian distribution over audio features. Das et al.[2] presented an approach for accounting for semantic regularities in language, which integrates the model proposed by Hu et al[15] with word embeddings. They use word2vec[16], to generate skip-gram word embeddings from unlabeled corpus. In this model, they characterize each topic k as a multivariate Gaussian distribution with mean μ_k and covariance Σ_k in an M -dimensional embedding space, and concurrently replaces the Dirichlet priors with the conjugate Normal-Inverse-Wishart (NIW) priors on Gaussian topics. Because the observations are no longer discrete values but continuous vectors, word vectors are sampled from continuous topic distributions. They reported that G-LDA produced higher PMI score than conventional LDA as the result of the experiment, which means topical coherence was improved. Because G-LDA uses continual distributions as the topic distributions over words, it can assign latent topics to OOV words without training the model again, whereas the original LDA cannot deal with those words.

The generative process is as follows.

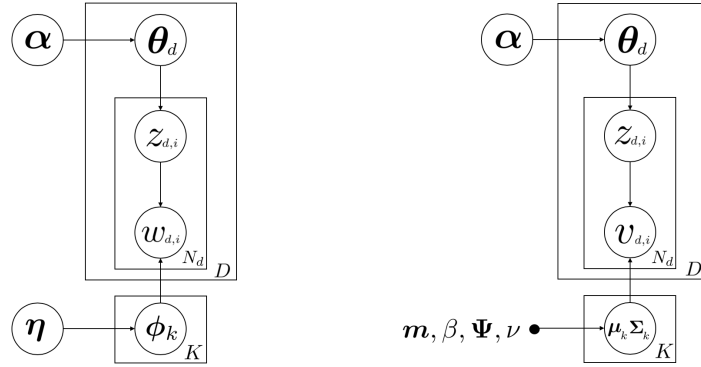


Fig. 1. Graphical representations of LDA(left) and Gaussian LDA(right).

1. for $k = 1$ to K
 - (a) Draw topic covariance $\Sigma_k \sim \mathcal{W}^{-1}(\Psi, \mu)$
 - (b) Draw topic mean $\mu_k \sim \mathcal{N}(\mu, \frac{1}{\beta} \Sigma_k)$
2. for each document d in corpus D
 - (a) Draw topic distribution $\theta_d \sim \text{Dir}(\alpha)$
 - (b) for each word index n from 1 to N_d
 - (a) Draw a topic $z_n \sim \text{Categorical}(\theta_d)$
 - (b) Draw $v_{d,n} \sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n})$

Although θ_d represents topic distributions of d th document as the traditional LDA does, μ_k and Σ_k represent the mean and the covariance of the multivariate Gaussian distribution, respectively. Besides, $v_{d,n}$ represents word vector.

The graphical model for G-LDA is shown in the right side of Figure 1.

3 Posterior Inference with SVI

Sampling method such as Gibbs sampler is widely used to perform approximate inference in topic modeling. Although Gibbs sampler has an advantage for easy implementation, it takes much time to estimate a posterior distribution. Hence, we employ an efficient inference algorithm based on VB, i.e., Stochastic Variational Inference (SVI)[14], to estimate the posterior probability distributions of the latent variables. SVI is an efficient algorithm for large datasets because it can sequentially process batches of documents. With VB inference, the true posterior probability distribution is approximated by a simpler distribution $q(z, \theta, \mu, \Sigma)$, which is indexed by a set of free parameters θ, μ and Σ . These parameters are optimized to maximize the Evidence Lower Bound (ELBO), a lower bound on the logarithm of the marginal probability of the observations $\log p(v)$.

$$\log p(v|\alpha, \zeta) \geq L(v, \phi, \gamma, \zeta) \triangleq \mathbb{E}_q[\log p(v, z, \theta, \mu, \Sigma|\alpha, \zeta)] - \mathbb{E}_q[\log q(z, \theta, \mu, \Sigma)]. \quad (1)$$

Based on the assumption that variables are independent in the mean-field family, approximate distribution q is fully factorized as follows.

$$q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = q(\mathbf{z})q(\boldsymbol{\theta})q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (2)$$

Let ϕ be the parameter for the latent variables z , γ be the parameter for the distribution θ over topics and $\boldsymbol{\zeta} = (\mathbf{m}, \beta, \boldsymbol{\Psi}, \nu)$ be the parameter of the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the topic distribution over word types. Factorized distributions of q are:

$$\begin{aligned} q(z_{di} = k) &= \phi_{dwi k}; & q(\theta_d) &= \text{Dir}(\theta_d | \gamma_d) \\ q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= NIW(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | \mathbf{m}_k, \beta_k, \boldsymbol{\Psi}_k, \nu_k) \end{aligned} \quad (3)$$

$$\begin{aligned} \phi_{dwk} &\propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log N(\mathbf{v}_{dw} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]\} \\ \gamma_{dk} &= \alpha + \sum_w n_{dw} \phi_{dwk}. \end{aligned} \quad (4)$$

SVI needs not analyze the whole data set before improving the global variational parameters and can apply new data which is constantly arriving, while VB requires a full pass through the entire corpus at each iteration. $q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the object for sequential learning, while $q(\mathbf{z}_d)$ and $q(\boldsymbol{\theta}_d)$ are optimized at each iteration. Thus, we apply the stochastic natural gradient descent to update the parameters $\boldsymbol{\zeta} = (\mathbf{m}, \beta, \boldsymbol{\Psi}, \nu)$. At d th document containing n_d words, we optimize ϕ_d and γ_d , holding $\boldsymbol{\zeta}$ fixed. Next, we calculate intermediate global parameters $\boldsymbol{\zeta}^* = (\mathbf{m}^*, \beta^*, \boldsymbol{\Psi}^*, \nu^*)$ as follows.

$$\begin{aligned} \beta_k^* &= \beta + D \sum_w n_{dw} \phi_{dwk}; & \nu_k^* &= \nu + D \sum_w n_{dw} \phi_{dwk}, \\ \mathbf{m}_k^* &= \frac{\beta \mathbf{m} + D \sum_w n_{dw} \phi_{dwk} \bar{\mathbf{v}}_k}{\beta_k^*}, \\ \boldsymbol{\Psi}_k^* &= \boldsymbol{\Psi} + \mathbf{C}_k + \frac{\beta D \sum_w n_{dw} \phi_{dwk}}{\beta_k^*} (\bar{\mathbf{v}}_k - \mathbf{m})(\bar{\mathbf{v}}_k - \mathbf{m})^\top. \end{aligned} \quad (5)$$

Here,

$$\begin{aligned} \bar{\mathbf{v}}_k &= \frac{\sum_w n_{dw} \phi_{dwk} \mathbf{v}_{dw}}{\sum_w n_{dw} \phi_{dwk}}, \\ \mathbf{C}_k &= D \sum_w n_{dw} \phi_{dwk} (\mathbf{v}_{dw} - \bar{\mathbf{v}}_k)(\mathbf{v}_{dw} - \bar{\mathbf{v}}_k)^\top. \end{aligned} \quad (6)$$

D denotes the number of corpus, which means that $\boldsymbol{\zeta}$ is optimized if the entire corpus consisted of the single document n_d repeated D times. By this operation, it becomes possible to update the parameters ϕ , γ and $\boldsymbol{\zeta}$ at each iteration without whole documents, so that it can analyze massive document collections, including those arriving in a stream. We then update $\boldsymbol{\zeta}$ using a weighted average of its previous value and the estimated $\boldsymbol{\zeta}^*$. The update is

$$\boldsymbol{\zeta} = (1 - \rho_d) \boldsymbol{\zeta} + \rho_d \boldsymbol{\zeta}^*. \quad (7)$$

The weight for ζ^* is given by $\rho_d \triangleq (\tau_0 + d)^{-\kappa}$, where $\kappa \in (0.5, 1]$ controls the rate at which old values of ζ are forgotten and $\tau_0 \geq 0$ slows down in the early iterations of the algorithm.

The expectations under q of $\log \theta_{dk}$ and $\log N(\mathbf{v}_{dw} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are

$$\begin{aligned} \mathbb{E}_q[\log \theta_{dk}] &= \Psi(\gamma_{dk}) - \Psi\left(\sum_{i=1}^K \gamma_{di}\right), \\ \mathbb{E}_q[\log N(\mathbf{v}_{dw} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] &= -\frac{1}{2} \mathbf{v}_{dw}^T \langle \boldsymbol{\Sigma}_k^{-1} \rangle \mathbf{v}_{dw} \\ &\quad + \mathbf{v}_{dw}^T \langle \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \rangle - \frac{1}{2} \langle \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \rangle - \frac{1}{2} \langle \log |\boldsymbol{\Sigma}_k| \rangle, \end{aligned} \quad (8)$$

where, Ψ and $\langle \cdot \rangle$ denote the digamma function and the expectation, respectively.

Algorithm 1 presents the full algorithm of SVI for Gaussian LDA.

Algorithm 1 SVI for Gaussian LDA

Define $\rho_d \triangleq (\tau_0 + d)^{-\kappa}$
Initialize $\mathbf{m}, \beta, \boldsymbol{\Psi}, \nu$ randomly.
for $d = 0$ to ∞ **do**
 Estep:
 Initialize $\gamma_{dk} = 1$ (The constant 1 is arbitrary.)
 repeat
 Set $\phi_{dwk} \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log N(\mathbf{v}_{dw} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]\}$
 Set $\gamma_{dk} = \alpha + \sum_w n_{dw} \phi_{dwk}$
 until $\frac{1}{K} \sum_k |\text{change in } \gamma_{dk}| < 0.00001$
 Mstep:
 Compute ζ_k^* with Eq.(5)
 Set $\zeta = (1 - \rho_d)\zeta + \rho_d \zeta^*$
end for

4 Experiments

We construct a model which integrates SVI to word vector topic model following Algorithm 1 and conduct the experiment of topic extraction. In this paper, we evaluate whether our model is able to find coherent and meaningful topics compared with the conventional LDA.

4.1 Experimental Setting

We perform experiments on two different text corpora: 18846 documents from 20Newsgroups¹ and 1740 documents from the NIPS². We utilize 50-dimensional word embeddings trained on text from Wikipedia using word2vec and run out the model with various number of topics ($K = 20 \sim 60$). The document distribution over topics θ is designed to be sampled from the conjugate Dirichlet prior parameterized by $\alpha = 1/K$. In equation (7), we set parameters $\tau_0 \in \{1, 4, 16, 64, 256, 1024\}$ and $\kappa \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. and set batch size according to the number of documents: $S \in \{4, 16, 64, 256, 1024\}$ for 20Newsgroups-dataset, $S \in \{4, 10, 16\}$ for NIPS-dataset. Our model implementation is in Python³.

In the experiments, we used the conventional LDA as a baseline model. The hyper parameters α and η in Dirichlet distribution are $1/K$ and 0.01, respectively.

4.2 Evaluation

We use PMI score to evaluate the quality of topics learnt by our models as well as it is used to evaluate the ability of G-LDA [2]. Newman et al [17] showed that PMI has relatively good agreement with human scoring. We use a reference corpus of documents from Wikipedia and use co-occurrence statistics over pairs of words (w_i, w_j) in the same document. The PMI score of topic k is computed by:

$$PMI(k) = \frac{1}{\binom{N}{2}} \sum_{j=2}^N \sum_{i=1}^{j-1} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}. \quad (9)$$

We use the average of the score of top 10 words of each topic. A higher PMI score implies a more coherent topic as it means the topic words usually co-occur in the same document.

4.3 Result

The experimental results of PMI on 20Newsgroups and NIPS-datasets are shown in Figure 2. We plot the average of the PMI scores for the top 10 words in each topic, the result of 20Newsgroups with the parameters of $S = 16$, $\kappa = 1.0$, and $\tau_0 = 1024$, and NIPS with those of $S = 4$, $\kappa = 1.0$, and $\tau_0 = 1024$. It is clearly seen that our model outperforms the conventional LDA in terms of PMI score.

Some examples of top topic words are listed in Table 1 and Table 2. The parameter settings is the same as above and we present the top 10 topics in descending order. In the last line of the tables, we present the PMI score for 10 topics for both our model and the traditional LDA. We see that the topics of our model seems more coherent than the baseline model. In addition, our model

¹ <http://qwone.com/~jason/20Newsgroups/>

² <https://cs.nyu.edu/~roweis/data.html>

³ https://github.com/KanaOzaki/SVI_GLDA

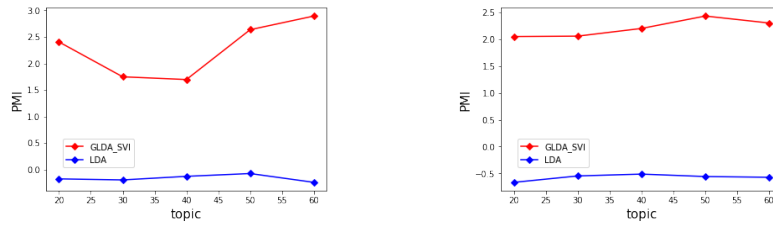


Fig. 2. PMI performance of the top 10 words on 20Newsgroups(left) and NIPS(right).

is able to capture several intuitive topics in the corpus such as natural science, mythology and cargo in Table1, mathematics and car in Table2. In particular, our model discovered the collection of human names, which was not captured by traditional LDA.

Table 1. Top 10 words of some topics from our model and multinomial LDA on 20Newsgroups for $K = 40$ and PMI score.

| Gaussian LDA topics | | | | | | | | | |
|------------------------|---------------|--------------|--------------|-------------|-----------|------------|--------------|-------------|--------------|
| cie | geophysics | manning | authenticity | beasts | ton | acts | disasters | provoke | normals |
| informatik | astrophysics | neely | veracity | creatures | tons | exercising | disaster | provocation | histograms |
| nos | physics | carney | credence | demons | gallon | coercion | hazards | futile | gaussian |
| gn | meteorology | brady | assertions | monsters | mv | act | catastrophic | suppress | linear |
| nr | astronomy | wilkins | inaccuracies | eleves | cargo | enforcing | devastation | resorting | symmetric |
| sta | geophysical | brett | particulars | spirits | cruiser | collective | dangers | threatening | histogram |
| vy | geology | seaver | texttual | unicorns | pound | proscribed | pollution | aggression | vectors |
| gl | astrophysical | reggie | merits | denizens | pounds | regulating | destruction | urge | inverse |
| cs | chemistry | ryan | substantiate | magical | corvettes | initiating | impacts | inflict | graphs |
| ger | microbiology | wade | refute | gods | guns | involving | destructive | expose | variables |
| 6.6429 | 6.2844 | 5.3070 | 5.0646 | 4.3270 | 3.6760 | 3.1671 | 2.8486 | 2.7723 | 2.7408 |
| Multinomial LDA topics | | | | | | | | | |
| drive | ax | subject | data | south | la | supreme | writes | key | government |
| disease | max | lines | doctors | book | goal | bell | article | code | law |
| hard | a86 | server | teams | lds | game | at&t | organization | package | gun |
| scsi | 0d | organization | block | published | cal | zoology | senate | window | clinton |
| drives | 1t | spacecraft | system | adl | period | subject | subject | data | congress |
| disk | giz | spencer | spave | armenian | bd | covenant | dod | information | clipper |
| subject | 3t | program | output | books | roy | suggesting | lines | anonymous | key |
| daughter | cx | space | pool | documents | 55.0 | lines | income | ftp | clayton |
| unit | bh | software | resources | isubject | its | off | deficit | program | federal |
| organization | kt | graphic | bits | information | season | origins | year | source | constitution |
| 2.3514 | 2.2500 | 1.3700 | 1.1216 | 1.0528 | 0.8338 | 0.7092 | 0.4531 | 0.4501 | 0.4355 |

Table 2. Top 10 words of some topics from our model and multinomial LDA on NIPS for $K = 40$ and PMI score.

| Gaussian LDA topics | | | | | | | | | |
|------------------------|------------|------------|---------------|----------|-----------------|----------|------------|----------------|---------------|
| topological | ginzburg | mitsubishi | negation | m.s | generalize | vx | gcs | behaviors | describes |
| projective | goldmann | vw | predicate | ms | analytically | xf | dcs | behaviours | describing |
| subspaces | jelinek | gm | disjunction | m/s | generalizations | vf | tss | behavior | interprets |
| symplectic | kolmogorov | motors | predicates | bd | intuitively | vz | rbp | behaviour | discusses |
| homotopy | markov | flat | propositional | tat | generalizing | r4 | sdh | biases | illustrates |
| topology | pinks | dyna | priori | dd | computable | xr | modulators | arousal | relates |
| euclidian | christof | integra | reflexive | stm | theretic | rx | signalling | behavioural | identifies |
| integrable | koenig | combi | duality | bs | solvable | tlx | mbs | behaviorally | characterizes |
| subspace | engel | gt | categorical | lond | generalization | t5 | analysers | attentional | demonstrates |
| affine | lippmann | suzuki | imperfect | bm | observable | spec | bss | predisposition | observes |
| 11.3463 | 9.4716 | 6.8211 | 6.7072 | 4.8832 | 4.4388 | 4.2489 | 3.6088 | 3.4125 | 3.3666 |
| Multinomial LDA topics | | | | | | | | | |
| model | network | learning | network | neural | network | network | learning | function | network |
| figure | model | network | algorithm | networks | networks | neural | neural | network | model |
| neural | input | figure | neural | model | neural | function | figure | model | input |
| learning | learning | data | learning | input | input | input | network | neural | learning |
| input | neural | units | training | data | learning | learning | data | training | data |
| network | networks | model | input | learning | data | model | input | learning | system |
| output | output | input | output | function | training | networks | training | set | training |
| number | function | set | networks | figure | output | figure | function | algorithm | neural |
| function | data | neural | set | units | number | output | model | data | function |
| data | figure | output | function | output | set | training | output | figure | output |
| 0.4945 | 0.4302 | 0.3506 | 0.3232 | 0.2280 | 0.1759 | -0.0473 | -0.1412 | -0.1784 | -0.2415 |

5 Conclusions and Future Work

Traditional topic models do not account for semantic regularities in language such as contextual relation of words as expressed in word embedding space. Therefore, G-LDA integrates the conventional topic model with word embeddings. However, dealing with high dimensional data such as word vectors in embedding space requires costly computation. So, G-LDA employs faster sampling using Cholesky decomposition of covariance matrix and Alias Sampling. On the other hand, Stochastic Variational Inference is much faster inference method than Markov chain Monte Carlo (MCMC) sampler such as Gibbs sampling and can deal with enormous dataset. Hence, we draw attention to SVI with expectation that SVI is also effective to handle high dimensional data. In this paper, we have proposed to apply efficient inference algorithm based on SVI to the topic model with word embeddings. As a qualitative analysis, we have verified the coherence in the extracted latent topics through the experiments and confirmed that our model is able to extract meaningful topics as G-LDA is.

In the future work, we will observe perplexity convergence to evaluate the inference speed and the soundness of our model.

References

1. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. (2003) Latent dirichlet allocation, *J. Mach. Learn. Res.*, 3:993-1022, March.
2. Rajarshi Das, Manzil Zaheer, and Chris Dyer. (2015) Gaussian LDA for topic models with word embeddings. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics.
3. Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. (2016) Nonparametric spherical topic modeling with word embeddings. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.
4. Weihua Hu and Jun'ichi Tsujii. (2016) A Latent Concept Topic Model for Robust Topic Inference Using Word Embeddings. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.
5. Guangxu Xun, Yaliang Li, Weyne Xin Zhao, Jing Gao and Aidong Zhang. (2017) A Correlated Topic Model Using Word Embeddings. In Proceedings of the 26th International Joint Conferences on Artificial Intelligence.
6. Dat Quoc Nguyen, Richard Billingsley, Lan Du, Mark Johnson. (2015) Improving Topic Models with Latent Feature Word Representations, *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 299-313, Action Editor: Kristina Toutanova.
7. Vivek Kumar Rangarajan Sridhar. (2015) Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words. Proceedings of NAACL-HLT 2015, pages 1922-1930, Denver, Colorado.
8. He Zhao, Lan Du, Wray Buntine. (2017) A Word Embeddings Informed Focused Topic Model, *Proceedings of Machine Learning Research* 77:423-438.
9. Christopher Moody. (2016) Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. <https://arxiv.org/abs/1605.02019>
10. Liang Yao, Yin Zhang, Baogang Wei, Zhe Jin, Rui Zhang, Yangyang Zhang and Qinfei Chen. (2017) Incorporating Knowledge Graph Embeddings into Topic Modeling. In Proceedings of the 31st AAAI Conference on Artificial Intelligence.
11. Aaron Q. Li, Amr Ahmed, Sujith Ravi and Alexander J. Smola. (2014) Reducing the Sampling Complexity of Topic Models. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 891-900
12. Matthew D. Hoffman, David M. Blei, Francis Bach. (2013) Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*.
13. John Paisley, David M. Blei, Michael I. Jordan. (2012) Variational Bayesian Inference with Stochastic Search, Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK.
14. Matthew D. Hoffman, David M. Blei, Chong Wang, John Paisley. (2013) Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303-1347.
15. Pengfei Hu, Wenju Liu, Wei Jiang, and Zhanlei Yang. (2012) Latent topic model based on Gaussian-LDA for audio retrieval. In *Pattern Recognition*, volume 321 of CCIS, pages 556-563. Springer.
16. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. (2013) Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111-3119.
17. David Newman, Sarvnaz Karimi, and Lawrence Cavdon. (2009) External evaluation of topic models. pages 11- 18, December.