ISSN 2007-9737

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

Kana Ozaki, Ichiro Kobayashie

Ochanomizu University,
Japan

{ozaki.kana, 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) [2], 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 Gaussain LDA(G-LDA) [3] 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. [1] 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) [5] 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. [7] 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 emgedding space.

They reported that LCTM is well suited for extracting topics from short texts with diverse vocaburaly such as tweets. Xun et al. [15] 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. [12] 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. [14] 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. [17] 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 [10] 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. [16] 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 [8] 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. [4] 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, 5]. 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 [2] 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 dth document, respectively. The generative process is as follows:

1. for $k = 1$ to K
   a) Choose topic $\boldsymbol{\beta}_k \sim \mathrm{Dir}(\boldsymbol{\eta})$
2. for each document d in corpus D
   a) Draw topic distribution $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$
   b) For each word index n from 1 to N_d
      a) Draw a topic $z_n \sim \mathrm{Categorical}(\boldsymbol{\theta}_{\mathrm{d}})$
      b) Draw a word $w_n \sim \mathrm{Categorical}(\boldsymbol{\beta}_{\mathrm{z_n}})$.

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

2.2 Gaussian LDA (G-LDA)

Hu et al. [6] 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. [3] presented an approach for accounting for semantic regularities in language, which integrates the model proposed by Hu et al. [6] with word embeddings. They use word2vec [9], 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-dimentional 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:

1. for $k = 1$ to K
   a) Draw topic covariance $\mathbf{\Sigma}_k \sim \mathcal{W}^{-1}(\mathbf{\Psi}, \mu)$
   b) Draw topic mean $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{\beta}\mathbf{\Sigma}_k)$
2. for each document d in corpus D
   a) Draw topic distribution $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$
   b) for each word index n from 1 to N_d
      a) Draw a topic $z_n \sim \mathrm{Categorical}(\boldsymbol{\theta}_\mathrm{d})$
      b) Draw $\boldsymbol{v}_{d,n} \sim \mathcal{N}(\boldsymbol{\mu}_{z_n}, \mathbf{\Sigma}_{z_n})$.

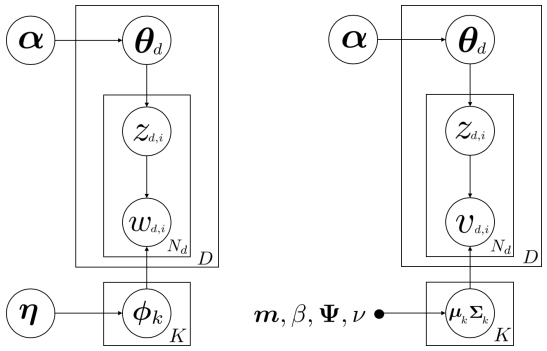

Fig. 1. Grphical representations of LDA (left) and Gaussian LDA (right).

Although $\boldsymbol{\theta}_d$ represents topic distributions of dth document as the traditional LDA does, $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ represent the mean and the covariance of the multivariate Gaussian distribution, respectively. Besides, $\boldsymbol{v}_{d,n}$ 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) [5], 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 $q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \mathbf{\Sigma})$, which is indexed by a set of free parameters $\boldsymbol{\theta}, \boldsymbol{\mu}$ and $\mathbf{\Sigma}$.

These parameters are optimized to maximize the Evidence Lower BOund (ELBO) , a lower bound

on the logarithm of the marginal probability of the observations $\log p(\boldsymbol{v})$:

$$\begin{aligned}\log p(\boldsymbol{v} \mid \boldsymbol{\alpha},\ \boldsymbol{\zeta}) \geq L(\boldsymbol{v},\ \boldsymbol{\phi},\ \boldsymbol{\gamma},\ \boldsymbol{\zeta})\\ \triangleq \mathbb{E}_q[\log p(\boldsymbol{v},\ \boldsymbol{z},\ \boldsymbol{\theta}\ ,\ \boldsymbol{\mu},\ \boldsymbol{\Sigma} \mid \boldsymbol{\alpha},\ \boldsymbol{\zeta})] -\\ \mathbb{E}_q[\log q(\boldsymbol{z},\ \boldsymbol{\theta},\ \boldsymbol{\mu},\ \boldsymbol{\Sigma})].\end{aligned} \tag{1}$$

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

$$q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = q(\boldsymbol{z})q(\boldsymbol{\theta})q(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2}$$

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

$$q(z_{di} = k) = \phi_{dw_{di}k}, \tag{3}$$

$$q(\theta_d) = \mathrm{Dir}(\theta_\mathrm{d}|\gamma_\mathrm{d}), \tag{4}$$

$$q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = NIW(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{m}_k, \beta_k, \boldsymbol{\Psi}_k, \nu_k), \tag{5}$$

$$\gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk}, \tag{6}$$

$$\phi_{dwk} \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log N(\boldsymbol{v}_{dw}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]\}. \tag{7}$$

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 $q(\boldsymbol{z}_d)$ and $q(\boldsymbol{\theta}_d)$ are optimized at each iteration.

Thus, we apply the stochastic natural gradient descent to update the parameters $\boldsymbol{\zeta} = (\boldsymbol{m}, \beta, \boldsymbol{\Psi}, \nu)$. At dth document containing n_d words, we optimize ϕ_d and γ_d, holding $\boldsymbol{\zeta}$ fixed. Next, we calculate intermediate global parameters $\boldsymbol{\zeta}^* = (\boldsymbol{m}^*, \beta^*, \boldsymbol{\Psi}^*, \nu^*)$ as follows:

$${\beta_k}^* = \beta + D\sum_w n_{dw}\phi_{dwk}, \tag{8}$$

$$\nu_k^* = \nu + D\sum_w n_{dw}\phi_{dwk}, \tag{9}$$

$$\boldsymbol{m}_k^* = \frac{\beta\boldsymbol{m} + D\sum_w n_{dw}\phi_{dwk}\bar{\boldsymbol{v}}_k}{\beta_k^*}, \tag{10}$$

$$\boldsymbol{\Psi}_k^* = \boldsymbol{\Psi} + \boldsymbol{C}_k + \frac{\beta D\sum_w n_{dw}\phi_{dwk}}{\beta_k^*}(\bar{\boldsymbol{v}}_k - \boldsymbol{m})(\bar{\boldsymbol{v}}_k - \boldsymbol{m})^\mathrm{T}. \tag{11}$$

Here:

$$\bar{\boldsymbol{v}}_k = \frac{\sum_w n_{dw}\phi_{dwk}\boldsymbol{v}_{dw}}{\sum_w n_{dw}\phi_{dwk}}, \tag{12}$$

$$\boldsymbol{C}_k = D\sum_w n_{dw}\phi_{dwk}(\boldsymbol{v}_{dw} - \bar{\boldsymbol{v}}_k)(\boldsymbol{v}_{dw} - \bar{\boldsymbol{v}}_k)^\mathrm{T}. \tag{13}$$

D denotes the number of corpus, which means that $\boldsymbol{\zeta}$ is optimized if the entire corpus consisted of the singe document n_d repeated D times. By this operation, it becomes possible to update the parameters $\boldsymbol{\phi}$, $\boldsymbol{\gamma}$ 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:

$$\boldsymbol{\zeta} = (1 - \rho_d)\boldsymbol{\zeta} + \rho_d\boldsymbol{\zeta}^*. \tag{14}$$

The weight for $\boldsymbol{\zeta}^*$ is given by $\rho_d \triangleq (\tau_0 + d)^{-\kappa}$, where $\kappa \in (0.5, 1]$ controls the rate at which old values of $\boldsymbol{\zeta}$ 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 $\log N(\boldsymbol{v}_{dw}|\boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k)$ are:

$$\mathbb{E}_q[\log \theta_{dk}] = \Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^{K}\gamma_{di}), \tag{15}$$

$$\begin{aligned}\mathbb{E}_q[\log N(\boldsymbol{v}_{dw}|\boldsymbol{m}\boldsymbol{u}_k, \boldsymbol{\Sigma}_k)] = -\frac{1}{2}\boldsymbol{v}_{dw}^\mathrm{T}\langle\boldsymbol{\Sigma}_k^{-1}\rangle\boldsymbol{v}_{dw}\\ +\boldsymbol{v}_{dw}^\mathrm{T}\langle\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\rangle - \frac{1}{2}\langle\log|\boldsymbol{\Sigma}_k|\rangle.\end{aligned} \tag{16}$$

where, Ψ and $\langle\ \rangle$ denote the digamma function and the expectation, respectively. Algorithm 1 presents the full algorithm of SVI for Gaussian LDA.

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.

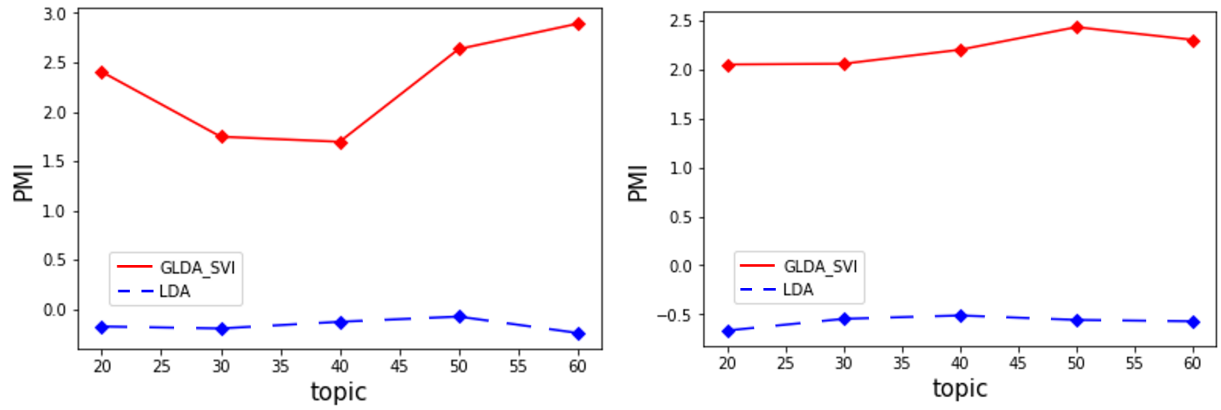

Fig. 2. PMI performance of the top 10 words on 20Newsgroups (left) and NIPS (right)

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	authenticicity	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	goverment
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

4.1 Experimental Setting

We perform experiments on two different text corpora: 18846 documents from 20Newsgroups[1] and 1740 documents from the NIPS[2]. 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 $\kappa \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ then we set the batch size according to the number of documents: $S \in \{4, 16, 64, 256, 1024\}$ for

[1] http://qwone.com/~jason/20Newsgroups/
[2] https://cs.nyu.edu/~roweis/data.html

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	mds	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	funtion	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

Algorithm 1 SVI for Gaussian LDA

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

20Newsgroups-dataset, $S \in \{4, 10, 16\}$ for NIPS -dataset. Our model implementation is in Python[3]. 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.

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

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 [3]. Newman et al. [11] showed that PMI has relatively good agreement with human scoring.

We use a reference corpus of documents from Wikipedia and use co-occurence statistics over pairs of words (w_i, w_j) in the same document. The PMI score of topic k is computed by:

$$PMI(k) = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}. \qquad (17)$$

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 is able to capture several intuitive topics in the corpus such as natural science, mythology and cargo in Table 1, mathematics and car in Table 2. In particular, our model discovered the collection of human names, which was not captured by traditional LDA.

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. **Batmanghelich, K., Saeedi, A., Narasimhan, K., Gershman, S. (2016).** Nonparametric spherical topic modeling with word embeddings.
2. **Blei, D. M., Ng, A. Y., Jordan, M. I. (2003).** Latent Dirichlet allocation. J. Mach. Learn. Res., Vol. 3, pp. 993–1022.
3. **Das, R., Zaheer, M., Dyer, C. (2015).** Gaussian LDA for topic models with word embeddings. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, pp. 795–804.
4. **Hoffman, M., Bach, F., Blei, D. (2010).** Online learning for latent Dirichlet allocation. Advances in Neural Information Processing Systems, volume 23, Curran Associates, Inc.
5. **Hoffman, M., Blei, D. M., Wang, C., Paisley, J. (2012).** Stochastic variational inference.
6. **Hu, P., Liu, W. J., Jiang, W., Yang, Z. (2012).** Latent topic model based on Gaussian-LDA for audio retrieval. pp. 556–563.
7. **Hu, W., Tsujii, J. (2016).** A latent concept topic model for robust topic inference using word embeddings. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, pp. 380–386.
8. **Li, A. Q., Ahmed, A., Ravi, S., Smola, A. J. (2014).** Reducing the sampling complexity of topic models. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, pp. 891–900.
9. **Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013).** Distributed representations of words and phrases and their compositionality.
10. **Moody, C. E. (2016).** Mixing Dirichlet topic models and word embeddings to make lda2vec.

11. **Newman, D., Karimi, S., Cavedon, L. (2011).** External evaluation of topic models. ADCS 2009 - Proceedings of the Fourteenth Australasian Document Computing Symposium.
12. **Nguyen, D. Q., Billingsley, R., Du, L., Johnson, M. (2018).** Improving topic models with latent feature word representations.
13. **Paisley, J., Blei, D., Jordan, M. (2012).** Variational Bayesian inference with stochastic search.
14. **Rangarajan Sridhar, V. K. (2015).** Unsupervised topic modeling for short texts using distributed representations of words. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, Association for Computational Linguistics, pp. 192–200.
15. **Xun, G., Li, Y., Zhao, W. X., Gao, J., Zhang, A. (2017).** A correlated topic model using word embeddings. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 4207–4213.
16. **Yao, L., Zhang, Y., Wei, B., Jin, Z., Zhang, R., Zhang, Y., Chen, Q. (2017).** Incorporating knowledge graph embeddings into topic modeling. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, No. 1.
17. **Zhao, H., Du, L., Buntine, W. (2017).** A word embeddings informed focused topic model. Proceedings of the Ninth Asian Conference on Machine Learning, volume 77, PMLR, pp. 423–438.

Article received on 15/02/2018; accepted on 11/01/2020 .
Corresponding author is Kana Ozaki.