Nonparametric Spherical Topic Modeling with Word Embeddings

Nematollah Kayhan Batmanghelich* CSAIL, MIT kayhan@mit.edu

Ardavan Saeedi* CSAIL, MIT ardavans@mit.edu

Karthik R. Narasimhan CSAIL, MIT karthikn@mit.edu

Samuel J. Gershman Department of Psychology Harvard University gershman@fas.harvard.edu

Abstract

Traditional topic models do not account for semantic regularities in language. Recent distributional representations of words exhibit semantic consistency over directional metrics such as cosine similarity. However, neither categorical nor Gaussian observational distributions used in existing topic models are appropriate to leverage such correlations. In this paper, we propose to use the von Mises-Fisher distribution to model the density of words over a unit sphere. Such a representation is well-suited for directional data. We use a Hierarchical Dirichlet Process for our base topic model and propose an efficient inference algorithm based on Stochastic Variational Inference. This model enables us to naturally exploit the semantic structures of word embeddings while flexibly discovering the number of topics. Experiments demonstrate that our method outperforms competitive approaches in terms of topic coherence on two different text corpora while offering efficient inference.[1]

1 Introduction

Prior work on topic modeling has mostly involved the use of categorical likelihoods (Blei et al., 2003; Blei and Lafferty, 2006; Rosen-Zvi et al., 2004). Applications of topic models in the textual domain treat words as discrete observations, ignoring the semantics of the language. Recent developments in distributional representations of words (Mikolov et al., 2013; Pennington et al., 2014) have succeeded in capturing certain semantic regularities, but have not been explored extensively in the context of topic modeling. In this paper, we propose a probabilistic topic model with a novel observational distribution that integrates well with directional similarity metrics.

One way to employ semantic similarity is to use the Euclidean distance between word vectors, which reduces to a Gaussian observational distribution for topic modeling (Das et al., 2015). The *cosine distance* between word embeddings is another popular choice and has been shown to be a good measure of semantic relatedness (Mikolov et al., 2013; Pennington et al., 2014). The von Mises-Fisher (vMF) distribution is well-suited to model such directional data (Dhillon and Sra, 2003; Banerjee et al., 2005) but has not been previously applied to topic models.

In this work, we use vMF as the observational distribution. Each word can be viewed as a point on a unit sphere with topics being canonical directions. More specifically, we use a Hierarchical Dirichlet Process (HDP) (Teh et al., 2006), a Bayesian nonparametric variant of Latent Dirichlet Allocation (LDA), to automatically infer the number of topics. We implement an efficient inference scheme based on Stochastic Variational Inference (SVI) (Hoffman et al., 2013).

We perform experiments on two different English text corpora: 20 NEWSGROUPS and NIPS and compare against two baselines - HDP and Gaussian LDA. Our model, spherical HDP (sHDP), outperforms all three systems on the measure of *topic coherence*. For instance, sHDP obtains gains over Gaussian LDA of 97.5% on the NIPS dataset and 65.5% on the 20 NEWSGROUPS dataset. Qualitative inspection reveals consistent topics produced by sHDP. We also empirically demonstrate that employing SVI leads to efficient

*Authors contributed equally and listed alphabetically.

[1]Code is available at https://github.com/Ardavans/sHDP.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 537–542, Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

topic inference.

2 Related Work

Topic modeling and word embeddings Das et al. (2015) proposed a topic model which uses a Gaussian distribution over word embeddings. By performing inference over the vector representations of the words, their model is encouraged to group words that are semantically similar, leading to more coherent topics. In contrast, we propose to utilize von Mises-Fisher (vMF) distributions which rely on the cosine similarity between the word vectors instead of euclidean distance.

vMF in topic models The vMF distribution has been used to model directional data by placing points on a unit sphere (Dhillon and Sra, 2003). Reisinger et al. (2010) propose an admixture model that uses vMF to model documents represented as vector of normalized word frequencies. This does not account for word level semantic similarities. Unlike their method, we use vMF over word embeddings. In addition, our model is nonparametric.

Nonparametric topic models HDP and its variants have been successfully applied to topic modeling (Paisley et al., 2015; Blei, 2012; He et al., 2013); however, all these models assume a categorical likelihood in which the words are encoded as one-hot representation.

3 Model

In this section, we describe the generative process for documents. Rather than one-hot representation of words, we employ normalized word embeddings (Mikolov et al., 2013) to capture semantic meanings of associated words. Word n from document d is represented by a normalized M-dimensional vector x_{dn} and the similarity between words is quantified by the cosine of angle between the corresponding word vectors.

Our model is based on the Hierarchical Dirichlet Process (HDP). The model assumes a collection of "topics" that are shared across documents in the corpus. The topics are represented by the topic centers $\mu_k \in \mathbb{R}^M$. Since word vectors are normalized, the μ_k can be viewed as a direction on unit sphere. Von Mises−Fisher (vMF) is a distribution that is commonly used to model directional data. The likelihood of the topic k for word x_{dn} is:

$$f(x_{dn}; \mu_k; \kappa_k) = \exp\left(\kappa_k \mu_k^T x_{dn}\right) C_M(\kappa_k)$$

where κ_k is the concentration of the topic k, the $C_M(\kappa_k) := \kappa_k^{M/2-1} / \left((2\pi)^{M/2} I_{M/2-1}(\kappa_k)\right)$ is the normalization constant, and $I_\nu(\cdot)$ is the modified Bessel function of the first kind at order ν. Interestingly, the log-likelihood of the vMF is proportional to $\mu_k^T x_{dn}$ (up to a constant), which is equal to the cosine distance between two vectors. This distance metric is also used in Mikolov et al. (2013) to measure semantic proximity.

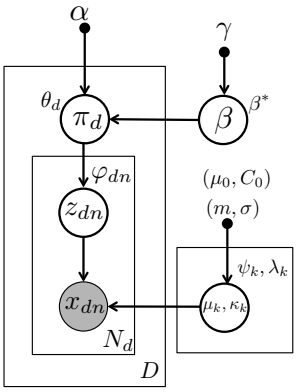

Figure 1: Graphical representation of our spherical HDP (sHDP) model. The symbol next to each random variable denotes the parameter of its variational distribution. We assume D documents in the corpus, each document contains N_d words and there are countably infinite topics represented by (μ_k, κ_k).

When sampling a new document, a subset of topics determine the distribution over words. We let z_{dn} denote the topic selected for the word n of document d. Hence, z_{dn} is drawn from a categorical distribution: $z_{dn} \sim \text{Mult}(\pi_d)$, where π_d is the proportion of topics for document d. We draw π_d from a Dirichlet Process which enables us to estimate the the number of topics from the data. The generative process for the generation of new document is as follows:

$$\begin{aligned}
&\beta \sim \text{GEM}(\gamma) && \pi_d \sim \text{DP}(\alpha, \beta) \\
&\kappa_k \sim \text{log-Normal}(m, \sigma^2) && \mu_k \sim \text{vMF}(\mu_0, C_0) \\
&z_{dn} \sim \text{Mult}(\pi_d) && x_{dn} \sim \text{vMF}(\mu_k, \kappa_k)
\end{aligned}$$

where $\text{GEM}(\gamma)$ is the stick-breaking distribution with concentration parameter γ, $\text{DP}(\alpha, \beta)$ is a Dirichlet process with concentration parameter α and stick proportions β (Teh et al., 2012). We use

log-normal and vMF as hyper-prior distributions for the concentrations (κ_k) and centers of the topics (μ_k) respectively. Figure 1 provides a graphical illustration of the model.

Stochastic variational inference In the rest of the paper, we use bold symbols to denote the variables of the same kind (*e.g.,* $\boldsymbol{x}_d = \{x_{dn}\}_n$, $\boldsymbol{z} := \{z_{dn}\}_{d,n}$). We employ stochastic variational mean-field inference (SVI) (Hoffman et al., 2013) to estimate the posterior distributions of the latent variables. SVI enables us to sequentially process batches of documents which makes it appropriate in large-scale settings.

To approximate the posterior distribution of the latent variables, the mean-field approach finds the optimal parameters of the fully factorizable q (*i.e.,* $q(\mathbf{z}, \beta, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa}) := q(\mathbf{z})q(\beta)q(\boldsymbol{\pi})q(\boldsymbol{\mu})q(\boldsymbol{\kappa})$) by maximizing the Evidence Lower Bound (ELBO),

$$\mathcal{L}(q) = \mathbb{E}_q\left[\log p(\boldsymbol{X}, \mathbf{z}, \beta, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa})\right] - \mathbb{E}_q\left[\log q\right]$$

where $\mathbb{E}_q[\cdot]$ is expectation with respect to q, $p(\boldsymbol{X}, \mathbf{z}, \beta, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa})$ is the joint likelihood of the model specified by the HDP model.

The variational distributions for $\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}$ have the following parametric forms,

$$\begin{aligned} q(\mathbf{z}) &= \text{Mult}(\mathbf{z}|\boldsymbol{\varphi}) \\ q(\boldsymbol{\pi}) &= \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\theta}) \\ q(\boldsymbol{\mu}) &= \text{vMF}(\boldsymbol{\mu}|\boldsymbol{\psi}, \boldsymbol{\lambda}), \end{aligned}$$

where Dir denotes the Dirichlet distribution and $\boldsymbol{\varphi}, \boldsymbol{\theta}, \boldsymbol{\psi}$ and $\boldsymbol{\lambda}$ are the parameters we need to optimize the ELBO. Similar to (Bryant and Sudderth, 2012), we view β as a parameter; hence, $q(\beta) = \delta_{\beta^*}(\beta)$. The prior distribution κ does not follow a conjugate distribution; hence, its posterior does not have a closed-form. Since κ is only one dimensional variable, we use importance sampling to approximate its posterior. For a batch size of one (*i.e.,* processing one document at time), the update equations for the parameters are:

$$\begin{aligned} \varphi_{dwk} &\propto \exp\{\mathbb{E}_q[\log \text{vMF}(x_{dw}|\psi_k, \lambda_k)] \\ &\qquad + \mathbb{E}_q[\log \pi_{dk}]\} \\ \theta_{dk} &\leftarrow (1-\rho)\theta_{dk} + \rho(\alpha\beta_k + D\sum_{n=1}^{W}\omega_{wj}\varphi_{dwk}) \\ t &\leftarrow (1-\rho)t + \rho s(\boldsymbol{x}_d, \varphi_{dk}) \\ \psi &\leftarrow t/\|t\|_2, \quad \lambda \leftarrow \|t\|_2 \end{aligned}$$

where D, ω_{wj}, W, ρ are the total number of documents, number of word w in document j, the total number of words in the dictionary, and the step size, respectively. t is a natural parameter for vMF and $s(\boldsymbol{x}_d, \varphi_{dk})$ is a function computing the sufficient statistics of vMF distribution of the topic k. We use numerical gradient ascent to optimize for β^*. For exact forms of $\mathbb{E}_q \log[\text{vMF}(x_{dw}|\psi_k, \lambda_k)]$ and $\mathbb{E}_q[\log \pi_{dk}]$, see Appendix.

4 Experiments

Setup We perform experiments on two different text corpora: 11266 documents from 20 NEWSGROUPS[2] and 1566 documents from the NIPS corpus[3]. We utilize 50-dimensional word embeddings trained on text from Wikipedia using *word2vec*[4]. The vectors are normalized to have unit ℓ^2-norm, which has been shown to provide superior performance (Levy et al., 2015)).

We evaluate our model using the measure of topic coherence (Newman et al., 2010), which has been shown to effectively correlate with human judgement (Lau et al., 2014). For this, we compute the Pointwise Mutual Information (PMI) using a reference corpus of 300k documents from Wikipedia. The PMI is calculated using co-occurence statistics over pairs of words (u_i, u_j) in 20-word sliding windows:

$$\text{PMI}(u_i, u_j) = \log \frac{p(u_i, u_j)}{p(u_i) \cdot p(u_j)}$$

Additionally, we also use the metric of normalized PMI (NPMI) to evaluate the models in a similar fashion:

$$\text{NPMI}(u_i, u_j) = \frac{\log \frac{p(u_i, u_j)}{p(u_i) \cdot p(u_j)}}{-\log p(u_i, u_j)}$$

We compare our model with two baselines: HDP and the Gaussian LDA model. We ran G-LDA with various number of topics (k).

Results Table 2 details the topic coherence averaged over all topics produced by each model. We observe that our sHDP model outperforms G-LDA by 0.08 points on 20 NEWSGROUPS and by 0.17 points in terms of PMI on the NIPS dataset. The NPMI scores also show a similar trend with sHDP obtaining the best scores on both datasets. We can also see that the individual topics inferred

[2] http://qwone.com/~jason/20Newsgroups/

[3] http://www.cs.nyu.edu/~roweis/data.html

[4] https://code.google.com/p/word2vec/

Gaussian LDA							
vector	shows	network	hidden	performance	net	figure	size
image	feature	learning	term	work	references	shown	average
gaussian	show	model	rule	press	introduction	neurons	present
equation	motion	neural	word	tion	statistical	point	family
generalization	action	input	means	ing	related	large	versus
images	spike	data	words	eq	comparison	neuron	spread
gradient	series	function	approximate	performed	source	small	median
theory	final	time	derived	em	statistics	fig	physiology
dimensional	robot	set	describe	vol	free	cells	children
1.16	0.4	0.35	0.29	0.25	0.25	0.21	0.2
Spherical HDP							
neural	function	analysis	press	pattern	problem	noise	algorithm
layer	linear	theory	cambridge	fig	process	gradient	error
neurons	functions	computational	journal	temporal	method	propagation	parameters
neuron	vector	statistical	vol	shape	optimal	signals	computation
activation	random	field	eds	smooth	solution	frequency	algorithms
brain	probability	simulations	trans	surface	complexity	feedback	compute
cells	parameter	simulation	springer	horizontal	estimation	electrical	binary
cell	dimensional	nonlinear	volume	vertical	prediction	filter	mapping
synaptic	equation	dynamics	review	posterior	solve	detection	optimization
1.87	1.73	1.51	1.44	1.41	1.19	1.12	1.03

Table 1: Examples of top words for the most coherent topics (column-wise) inferred on the NIPS dataset by Gaussian LDA (k=40) and Spherical HDP. The last row for each model is the topic coherence (PMI) computed using Wikipedia documents as reference.

Model	**Topic Coherence**			
	20 NEWS		NIPS	
	pmi	npmi	pmi	npmi
HDP	0.037	0.014	0.270	0.062
G-LDA (k=10)	-0.061	-0.006	0.214	0.055
G-LDA (k=20)	-0.017	0.001	0.215	0.052
G-LDA (k=40)	0.052	0.015	0.248	0.057
G-LDA (k=60)	0.082	0.021	0.137	0.034
sHDP	**0.162**	**0.046**	**0.442**	**0.102**

Table 2: Average topic coherence for various baselines (HDP, Gaussian LDA (G-LDA)) and sHDP. k=number of topics. Best scores are shown in bold.

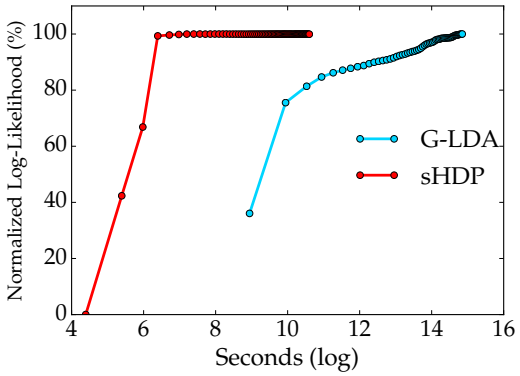

Figure 2: Normalized log-likelihood (in percentage) over a training set of size 1566 documents from the NIPS corpus. Since the log-likelihood values are not comparable for the Gaussian LDA and the sHDP, we normalize them to demonstrate the convergence speed of the two inference schemes for these models.

by sHDP make sense qualitatively and have higher coherence scores than G-LDA (Table 1). This supports our hypothesis that using the vMF likelihood helps in producing more coherent topics. sHDP produces 16 topics for the 20 NEWSGROUPS and 92 topics on the NIPS dataset.

Figure 2 shows a plot of normalized log-likelihood against the runtime of sHDP and G-LDA.[5] We calculate the normalized value of log-likelihood by subtracting the minimum value from it and dividing it by the difference of maximum and minimum values. We can see that sHDP converges faster than G-LDA, requiring only around five iterations while G-LDA takes longer to converge.

5 Conclusion

Classical topic models do not account for semantic regularities in language. Recently, distributional

[5]Our sHDP implementation is in Python and the G-LDA code is in Java.

representations of words have emerged that exhibit semantic consistency over directional metrics like cosine similarity. Neither categorical nor Gaussian observational distributions used in existing topic models are appropriate to leverage such correlations. In this work, we demonstrate the use of the von Mises-Fisher distribution to model words as points over a unit sphere. We use HDP as the base topic model and propose an efficient algorithm based on Stochastic Variational Inference. Our model naturally exploits the semantic structures of word embeddings while flexibly inferring the number of topics. We show that our method outperforms three competitive approaches in terms of topic coherence on two different datasets.

Acknowledgments

Thanks to Rajarshi Das for helping with the Gaussian LDA experiments and Matthew Johnson for his help with the HDP code.

References

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von mises-fisher distributions. In *Journal of Machine Learning Research*, pages 1345–1382.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Michael Bryant and Erik B Sudderth. 2012. Truly nonparametric online variational inference for hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems*, pages 2699–2707.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*.

Inderjit S Dhillon and Suvrit Sra. 2003. Modeling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin. URL ftp://ftp. cs. utexas. edu/pub/techreports/tr03-06. ps. gz.

Siddarth Gopal and Yiming Yang. 2014. Von mises-fisher clustering models.

Yulan He, Chenghua Lin, Wei Gao, and Kam-Fai Wong. 2013. Dynamic joint sentiment-topic model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):6.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Matthew Johnson and Alan Willsky. 2014. Stochastic variational inference for bayesian time series models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1854–1862.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

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.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

John Paisley, Chingyue Wang, David M Blei, and Michael I Jordan. 2015. Nested hierarchical dirichlet processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):256–270.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J Mooney. 2010. Spherical topic models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 903–910.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2012. Hierarchical dirichlet processes. *Journal of the american statistical association*.

Appendinx

Mean field update equations

In this section, we provide the mean field update equations. The SVI update equations can be derived from the mean field update (Hoffman et al., 2013).

The following term is computed for the update equations:

$$\mathbb{E}_q[\log \text{vMF}(x_{dn}|\mu_k, \kappa_k)] = \mathbb{E}_q[\log C_M(\kappa_k)] + \mathbb{E}_q[\kappa_k] x_{dn}^T \mathbb{E}_q[\mu_k]$$

where $C_M(\cdot)$ is explained in Section 3. The difficulty here lies in computing $\mathbb{E}_q[\kappa_k]$ and $\mathbb{E}_q[C_M(\kappa_k)]$. However, κ is a scalar value. Hence, to compute $\mathbb{E}_q[\kappa_k]$, we divide a reasonable interval of κ_k into grids and compute the weight for each grid point as suggested by Gopal and Yang (2014):

$$p(\kappa_k|\cdots) \propto \exp\left(n_k \log C_M(\kappa_k) + \kappa_k \left(\sum_{d=1}^{D}\sum_{n=1}^{N_d} [\varphi_{dn}]_k \langle x_{dn}, \mathbb{E}_q[\mu_k]\rangle\right)\right) \times \text{logNormal}(\kappa_k|m, \sigma^2)$$

where $n_k = \sum_{d=1}^{D}\sum_{d=1}^{N_d} [\varphi_{dn}]_k$ and $[a]_k$ denotes the k'th element of vector a. After computing the normalized weights, we can compute $\mathbb{E}_q[\kappa_k]$ or expectation of any other function of κ_k (*e.g.,* $\mathbb{E}_q[C_M(\kappa_k)]$). The rest of the terms can be computed as follows:

$$\mathbb{E}_q[\mu_k] = \mathbb{E}_q\left[\frac{I_{M/2}(\kappa_k)}{I_{M/2-1}(\kappa_k)}\right]\psi_k,$$

$$\psi_k = \mathbb{E}_q[\kappa_k]\left(\sum_{d=1}^{D}\sum_{n=1}^{N_d}[\varphi_{dn}]_k x_{dn}\right) + C_0\mu_0$$

$$\psi_k \leftarrow \frac{\psi_k}{\|\psi_k\|_2},$$

$$[\mathbb{E}_q[\log(\pi_d)]]_k = \Psi([\theta_d]_k) - \Psi\left(\sum_k [\theta_d]_k\right),$$

$$[\varphi_{dn}]_k \propto \exp\left(\mathbb{E}_q[\log \text{vMF}(x_{dn}|\mu_k, \kappa_k)] + \mathbb{E}_q[\log([\pi_d]_k)]\right),$$

$$[\theta_d]_k = \alpha + \sum_{n=1}^{N_d}[\varphi_{dn}]_k$$

$\Psi(\cdot)$ is the digamma function.

To find β^*, similar to Johnson and Willsky (2014), we use the gradient expression of ELBO with respect to β and take a truncated gradient step on β ensuring $\beta^* \geq 0$.