Integrating Topic Modeling with Word Embeddings by Mixtures of vMFs

Ximing Li^{1,2} **Jinjin Chi**^{1,2} **Changchun Li**^{1,2} **Jihong Ouyang**^{1,2} **Bo Fu**^{3}

1. College of Computer Science and Technology, Jilin University, China
2. Key Laboratory of Symbolic Computation and Knowledge Engineering of
Ministry of Education, China
3. College of Computer and Information Technology, Liaoning Normal University, China
`liximing86@gmail.com`

Abstract

Gaussian LDA integrates topic modeling with word embeddings by replacing discrete topic distribution over word types with multivariate Gaussian distribution on the embedding space. This can take semantic information of words into account. However, the Euclidean similarity used in Gaussian topics is not an optimal semantic measure for word embeddings. Acknowledgedly, the cosine similarity better describes the semantic relatedness between word embeddings. To employ the cosine measure and capture complex topic structure, we use von Mises-Fisher (vMF) mixture models to represent topics, and then develop a novel mix-vMF topic model (MvTM). Using public pre-trained word embeddings, we evaluate MvTM on three real-world data sets. Experimental results show that our model can discover more coherent topics than the state-of-the-art baseline models, and achieve competitive classification performance.

1 Introduction

Topic models such as latent Dirichlet allocation (LDA) (Blei et al., 2003) are hierarchical probabilistic models of document collections. They can effectively uncover the main themes of corpora by using latent topics learnt from observed collections (Blei, 2012), however, they neglect semantic information of words. In topic modeling, a "topic" is a multinomial distribution over a fixed vocabulary, i.e., a word type proportion. Because words are represented by unordered indexes, with statistical inference algorithms, related words are grouped into topics mainly by using document-level word co-occurrence information (Wang and McCallum, 2006), rather than semantics of words. That is why LDA often outputs many low-quality topics, and views in (Das et al., 2015) even suggest that any such observation of semantically coherent topics in topic models is, in some sense, accidental.

To mix with semantics of words, a recent Gaussian LDA (G-LDA) (Das et al., 2015) model integrates topic modeling with word embeddings, which can effectively capture lexico-semantic regularities in language from a large unlabeled corpus (Mikolov et al., 2013). This hot technique transforms words into vectors (i.e., word vector). To model documents of word vectors, G-LDA replaces the discrete topic distributions over word types with multivariate Gaussian distributions on the word embedding space. Because words with similar semantic properties are closer to each other in the embedding space, semantic information of words can be taken into consideration by using Gaussian distributions to describe semantic centrality location of topics.

An issue of G-LDA is that the word weights in Gaussian topics are measured by the Euclidean similarity between word embeddings. However, the Euclidean similarity is not an optimal semantic measure, since most of word embedding algorithms use exponentiated cosine similarity as the link function (Li et al., 2016a). The cosine similarity may be a better choice to describe the semantic relatedness between word embeddings. Following this idea, in this paper we use von Mises-Fisher (vMF) distributions on the embedding space to represent topics, replacing Gaussian topics in G-LDA. The vMF distribution defines a probability density over vectors on a unit sphere, parameterized by mean μ and concentration

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 151–160, Osaka, Japan, December 11-17 2016.

parameter κ. Its density function for $x \in \mathcal{R}^M$, $\|x\| = 1$, $\|\mu\| = 1$, $\kappa \geq 0$ is given by:

$$p\left(x|\mu, \kappa\right) = c_p\left(\kappa\right) \exp\left(\kappa \mu^T x\right) \tag{1}$$

where $c_p\left(\kappa\right)$ is the normalization constant. Note that vMF concerns the cosine similarity defined by $\mu^T x$. It is a better way to represent topics of word embeddings.

Another issue we face is that topics often contain many words that are far away from each other in the embedding space. That is, the true distributions of topics often form two or more dominant clumps. However, a simple vMF distribution is unable to capture such structure. For example, the topic $\langle software, user, net, feedback, grade\rangle$ contains some "dissimilar" words, such as *net* and *grade*[1]. In this case, a simple vMF topic distribution can not simultaneously place high probabilities on these "dissimilar" words.

To address the problem mentioned above, we further use mixtures of vMFs to describe topics, rather than a single vMF. We then develop a novel mix-vMF topic model (MvTM). Mixtures of vMFs can help us capture complex topic structure that forms more dominant clumps. In MvTM, we consider two settings with respect to the topic, i.e., disjoint setting and overlapping setting. Naturally, in disjoint settings all mixtures of vMFs use disjoint vMF bases; and in overlapping setting some mixtures of vMFs share the same vMF bases. An advantages of the overlapping setting is that it can describe topic correlation in some degree. We have conducted a number of experiments on three real-world data sets. Experimental results show that our MvTM can discover more coherent topics than the state-of-the-art baseline topic models, and achieve competitive performance on the classification task.

2 Model

In this section, we simply review LDA and G-LDA.

2.1 LDA

LDA (Blei et al., 2003) is a representative probabilistic topic model of document collections. In LDA, the main themes of corpora are described by topics, where each topic is a multinomial distribution ϕ over a fixed vocabulary (i.e., a word type proportion). Each document is a multinomial distribution θ over topics (i.e., a topic proportion). For simplification, distributions ϕ and θ are designed to be sampled from the conjugate Dirichlet priors parameterized by β and α, respectively. Suppose that *D*, *K* and *V* denote the number of documents, topics and word types. The generative process of LDA is as follows:

1. For each topic $k \in \{1, 2, \cdots, K\}$
   (a) Sample a topic $\phi_k \sim Dir\left(\beta\right)$
2. For each document $d \in \{1, 2, \cdots, D\}$
   (a) Sample a topic proportion: $\theta_d \sim Dir\left(\alpha\right)$
   (b) For each of the N_d words embeddings
      i. Sample a topic indicator $z_{dn} \sim Multinomial\left(\theta_d\right)$
      ii. Sample a word $w_{dn} \sim Multinomial\left(\phi_{z_{dn}}\right)$

Reviewing the definition above, we note that a topic in LDA is a discrete distribution over observable word types (i.e., word indexes). In this sense, LDA neglects semantic information of words and precludes new word types to be added into topics.

2.2 G-LDA

G-LDA (Das et al., 2015) integrates topic modeling with word embeddings. This model replaces the discrete topic distributions over word types with multivariate Gaussian distributions on an M-dimensional

[1] This means that the cosine similarity between word embeddings of *net* and *grade* is small.

embedding space, and concurrently replaces the Dirichlet priors with the conjugate Normal-Inverse-Wishart (NIW) priors on Gaussian topics. Because word embeddings learnt from large unlabeled corpora effectively capture semantic information of words (Bengio et al., 2003), G-LDA can handle, in some sense, words' semantics and new word types. Let $\mathcal{N}\left(\mu_k, \Sigma_k\right)$ be the Gaussian topic *k* with mean μ_k and covariance matrix Σ_k. The generative process of G-LDA is as follows:

1. For each topic $k \in \{1, 2, \cdots, K\}$
   (a) Sample a Gaussian topic $\mathcal{N}\left(\mu_k, \Sigma_k\right) \sim NIW\left(\mu_0, \kappa_0, \Psi_0, \nu_0\right)$
2. For each document $d \in \{1, 2, \cdots, D\}$
   (a) Sample a topic proportion: $\theta_d \sim Dir\left(\alpha\right)$
   (b) For each of the N_d word embeddings
      i. Sample a Gaussian topic indicator $z_{dn} \sim Multinomial\left(\theta_d\right)$
      ii. Sample a word embedding $w_{dn} \sim \mathcal{N}\left(\mu_{z_{dn}}, \Sigma_{z_{dn}}\right)$

3 MvTM

G-LDA defines Gaussian topics, which measure word weights in topics by the Euclidean similarity between word embeddings. However, the Euclidean similarity is not an optimal semantic measure of word embeddings. People often prefer the cosine similarity (Li et al., 2016a). To upgrade G-LDA, a novel mix-vMF topic model (MvTM) is proposed, where we replace the Gaussian topic in G-LDA with mixture of vMFs. In this work, we use mixture of vMFs with *C* mixture components (Banerjee et al., 2005) described by:

$$p\left(x|\pi_{1:C}, \mu_{1:C}, \kappa\right) = \sum_{c=1}^{C} \pi_c p_c\left(x|\mu_c, \kappa\right) \tag{2}$$

where $p_c\left(x|\mu_c, \kappa\right)$ is the mixture vMF component (i.e., base); π_c is the mixture weight and such that $\sum_{c=1}^{C} \pi_c = 1$. The design of MvTM has two advantages. First, the vMF distribution defines a probability density over normalized vectors on a unit sphere. Reviewing Eq.1, it can be seen that vMF concerns the cosine similarity. Second, using linear vMF mixture model can help us capture complex topic structure, which forms two or more dominant clumps.

Formally, MvTM models documents consisting of normalized word embeddings w in an *M*-dimensional space, i.e., $\|w\| = 1$ and $w \in R^M$. Suppose that there are *K* topics in total. We characterize each topic *k* as a mixture of vMFs with parameter $\Delta_k = \left\{\pi_{k|_{1:C}}, \mu_{k|1:C}, \kappa_k\right\}$. Besides the topic design, again suppose that each document is a topic proportion θ, drawn from a Dirichlet prior α. Let *D* and N_d be the number of documents and the number of words in document *d*, respectively. The generative process of MvTM is given by:

1. For each document $d \in \{1, 2, \cdots, D\}$
   (a) Sample a topic proportion: $\theta_d \sim Dir\left(\alpha\right)$
   (b) For each of the N_d word embeddings
      i. Sample a vMF mixture topic indicator $z_{dn} \sim Multinomial\left(\theta_d\right)$
      ii. Sample a word vector $w_{dn} \sim v\mathcal{MF}\left(\Delta_{z_{dn}}\right)$

In MvTM, the vMF bases of different topics can be either disjoint or overlapping. For disjoint MvTM (abbr. MvTM$_d$), the vMF bases of different topics are disjoint. In MvTM$_d$, the total number of vMF bases is $C \times K$. For overlapping MvTM (abbr. MvTM$_o$), vMF bases are allowed to be shared by different topics. An advantage is that the overlapping setting can describe topic correlation in some degree. For example, if two topics share a same vMF base and their corresponding mixture weights are close to each other, they may be semantically correlated. In previous study, we have examined several overlapping patterns, e.g., all topics share a same set of vMF bases. However, an issue is that such patterns often output many twinborn topics. In this work, we use the following overlapping scheme: suppose that there are *G* groups of K' topics. In a group, each topic consists of C' personal vMF bases, and all topics in this group share *P* public vMF bases, where $C' + P = C$. In this setting, the total number of vMF bases is $G \times (K' \times C' + P)$, and topics in a group $\mathcal{G}_g$ use a same κ_g, i.e., $\kappa_g = \kappa_k = \cdots = \kappa_{k'}$ if $k \cdots k' \in \mathcal{G}_g$. The intuition behind overlapping by topic groups is that only a small set of topics may be semantically correlated. Besides, the personal vMF base design can effectively avoid the outputs of twinborn topics.

3.1 Inference

For MvTM, the topic proportions $\{\theta_d\}_{d=1}^{d=D}$ and the topic assignments $\{z_{dn}\}_{d=1,n=1}^{d=D,n=N_d}$ are hidden variables; and the topics $\{v\mathcal{MF}(\Delta_k)\}_{k=1}^{k=K}$ are model parameters. Given an observable document collection *W* consisting of word embeddings, we wish to compute the posterior distribution over θ and *z*, and to estimate $v\mathcal{MF}(\Delta)$.

Because the exact posterior distribution $p(\theta, z|W, \alpha, \Delta)$ is intractable to be computed, we must resort approximation inference algorithms. Due to the multinomial-Dirichlet design, the topic proportion θ can be analytically integrated out. We then use hybrid variational-Gibbs (HVG) (Mimno et al., 2012) to approximate a posterior over the topic assignment *z*: $p(z|W, \alpha, \Delta)$. A variational distribution of the following form is used:

$$q(z) = \prod_{d=1}^{D} q(z_d) \tag{3}$$

where $q(z_d)$ is a single distribution over the K^{N_d} possible topic configurations, rather than a product of N_d distributions. By using this variational distribution, we obtain an Evidence Lower BOund (ELBO) $\mathcal{L}$ as follows :

$$\log p(z|W, \alpha, \Delta) \geq \mathcal{L}(z_d, \Delta) \triangleq \mathrm{E}_q[\log p(W, z|\alpha, \Delta)] - \mathrm{E}_q[\log q(z)] \tag{4}$$

We then develop an expectation maximization (EM) process to optimize this ELBO, where in the E-step we maximize $\mathcal{L}$ with respect to the variational distribution $q(z)$, and in the M-step we maximize $\mathcal{L}$ with respect to the model parameter Δ, holding $q(z)$ fixed. Optimizing $q(z)$ directly is expensive because for each document *d* it needs to enumerate all K^{N_d} possible topic configurations. We therefore apply Monte-Carlo approximation to this ELBO $\mathcal{L}$ in Eq.4 by:

$$\begin{aligned}\mathcal{L}(z_d, \Delta) &\triangleq \mathrm{E}_q[\log p(W, z|\alpha, \Delta)] - \mathrm{E}_q[\log q(z)] \\ &\approx \frac{1}{B}\sum_{b=1}^{B}\left(\log p(W, z^{(b)}|\alpha, \Delta) - \log q\left(z^{(b)}\right)\right)\end{aligned} \tag{5}$$

where $\left\{z^{(b)}\right\}_{b=1}^{b=B}$ are samples drawn from $q(z)$. Because the variational distributions $q(z_d)$ are independent from each other, reviewing Eq.3, each document *d* drives a personal sampling process with respect to $q(z_d)$.

In the **E-step**, for each document *d* we use Gibbs sampling to draw *B* samples from $q(z_d)$. This sequentially samples topic assignment to each word embedding from the posterior distribution conditioned on all other variables and the data. The sampling equation is given by:

$$p(z_{dn} = k|z_d^{-n}, \alpha, \Delta) \propto (N_{dk}^{-n} + \alpha) \times v\mathcal{MF}(w_{dn}|\Delta_k) \tag{6}$$

where N_{dk} is the number of word embeddings assigned to topic *k* in document *d*; the superscript "-n" is a quantity that excludes the word embedding w_{dn}. During per-document Gibbs sampling, we iteratively run the MCMC chain a fixed number of times and save the last *B* samples.

In the **M-step**, we optimize Δ given all samples of *z* obtained in E-step. This is achieved by maximizing the following approximate ELBO $\mathcal{L}'$:

$$\mathcal{L}' \triangleq \frac{1}{B}\sum_{b=1}^{B}\left(\log p(W, z^{(b)}|\alpha, \Delta) + const\right) \tag{7}$$

For the disjoint setting, i.e., MvTM$_d$, the optimization of $\mathcal{L}'$ is equivalent to independently estimate Δ_k for each topic *k*. Due to space limit, we omit the derivation details (Gopal and Yang, 2014). Extracting all N_k word embeddings assigned to topic *k*, for each word embedding w_i we compute its weights for all *C* vMF bases by:

$$weight_{ic} = \frac{\pi_{k|c} v\mathcal{MF}(w_i|\mu_{k|c}, \kappa_k)}{\sum_{j=1}^{C} \pi_{k|j} v\mathcal{MF}(w_i|\mu_{k|j}, \kappa_k)} \tag{8}$$

and then update Δ_k by:

$$\begin{gathered} R_{k|c} = \sum_{i=1}^{N_k} weight_{ic} \times w_i, \quad r_k = \sum_{c=1}^{C} \frac{\left\| R_{k|c} \right\|}{N_k} \\ \mu_{k|c} = \frac{R_{k|c}}{\left\| R_{k|c} \right\|}, \quad \pi_{k|c} = \sum_{i=1}^{N_k} \frac{weight_{ic}}{N_k}, \quad \kappa_k = \frac{r_k M - r_k^3}{1 - r_k^2} \end{gathered} \tag{9}$$

For the overlapping setting, i.e., MvTM$_o$, there are a few changes to the optimization of $\mathcal{L}'$. In each topic group $\mathcal{G}_g$, the updates of π and μ of personal vMF bases remain unchanged, whereas the mean μ of public vMF bases and κ of this group are updated by:

$$r_g = \sum_{c=1}^{C} \frac{\left\| \sum_{k \in \mathcal{G}_g} R_{k|c} \right\|}{\sum_{k \in \mathcal{G}_g} N_k}, \quad \mu_{k|p} = \frac{\sum_{k \in \mathcal{G}_g} R_{k|p}}{\left\| \sum_{k \in \mathcal{G}_g} R_{k|p} \right\|}, \quad \kappa_g = \frac{r_g M - r_g^3}{1 - r_g^2} \tag{10}$$

where $\mu_{k|p}$ is the mean of the *p*th public vMF base for topic *k* and note that $\mu_{k|p} = \mu_{k'|p}$ if $k, k' \in \mathcal{G}_g$.

For clarity, the overall EM inference algorithm for MvTM is outlined in *Algorithm 1*.

Algorithm 1 Inference for MvTM

```
**Initialize** parameters.
**For** $t = 1, 2, \cdots, Max\_iter$ **do**
    **E-step**
        **For** document *d*=1 to *D* **do**
            Gibbs sampling for *B* topic assignments $z_d^{(b)}$ using Eq.6
        **End for**
    **M-step**
        For MvTM$_d$, optimize $\Delta$ using Eq.8 and 9.
        For MvTM$_0$, optimize $\Delta$ using Eq.8, 9 and 10.
**End for**
```

3.2 Time Complexity

We first analyze the time complexities of E-step and M-step, and then present the overall time cost of MvTM.

In the **E-step**, the main time cost is the topic assignment sampling of each word embedding over *K* topics. Reviewing Eq.6, one sampling process needs to compute the probabilities of the current word embedding, i.e., $v\mathcal{MF}(w_{dn}|\Delta_k)$, in all *K* topics, which requires $O(KCM)$ time. Fortunately, the topics are fixed in the E-step, thus we only need to compute the value of $v\mathcal{MF}(w|\Delta_k)$ for each word embedding at the beginning of each EM sweep, and save them in the memory. This requires $O(VKCM)$ time, where *V* is the number of word embeddings. Consequently, the topic sampling process of MvTM is equivalent to the sampling of Gibbs sampling LDA, requiring $O(K)$ time. We present that the per-iteration time complexity of E-step is given by $O(VKCM + \zeta N_V K)$, where ζ is the iteration number in per-document Gibbs sampling and N_V is the total number of word embeddings occurred in a corpus. Recently, sparse sampling algorithms (Yao et al., 2009; Li et al., 2014) effectively accelerate the sampling

Table 1: Summarization of data sets used in our experiments. N_V is the total number of word tokens; N_V/D is the average document length; "label" denotes the number of pre-assigned classes.

Data set	V	D	N_V	N_V/D	label
NG	18,127	18,768	1,946,487	104	20
NIPS	4,805	1,740	2,097,746	1,206	–
Wiki	7,702	44,819	6,851,615	153	–

of topic models. Inspired by (Li et al., 2016b), we employ the Alias method (Walker, 1977; Marsaglia et al., 2004) to reduce the per-word sampling cost from $O(K)$ to $O(K_d)$, where K_d is the number of instantiated topics in document *d* and commonly $K_d \ll K$. The per-iteration time complexity of E-step now is $O(VKCM + \zeta N_V K_d)$.

In the **M-step**, the time cost of MvTM$_d$ and that of MvTM$_o$ are almost the same. We only present the time complexity of MvTM$_d$. Reviewing the M-step, we see that the most expensive updates include Eq.8, the first and the fourth equations in Eq.9. They require $O(VCM)$, $O(VCM)$ and $O(VC)$. Thus we present that the (per-iteration) time complexity of M-step is $O(VCM)$.

Overall, we see that in each EM sweep the E-step dominates the run-time, giving an approximate total per-iteration time complexity $O(VKCM + \zeta N_V K_d)$. Clearly, MvTM is much efficient than Gibbs sampling G-LDA (Das et al., 2015), because G-LDA needs to repeatedly compute the determinant and inverse of the covariance matrix in Gaussian topics. For each word occurring, this spends $O(M^2)$ time, even using Cholesky decomposition.

4 Experiment

In this section, we evaluate MvTM qualitatively and quantitatively.

4.1 Experimental Setting

Data set Three data sets were used in our experiments, including Newsgroup (NG), NIPS and Wikipedia (Wiki). The NG data set is a collection of newsgroup documents, consisting of 20 classes. We will use NG to examine the classification performance of MvTM in Section 4.3. The NIPS data set is a collection of papers in the NIPS conference. The processed versions of these two data sets were downloaded from the open source of G-LDA[2]. For the Wiki data set, we downloaded a number of documents from online English Wikipedia, and processed these documents using a standard vocabulary[3]. The statistics of the three data sets are listed in Table 1.

Baseline model: In the experiments, we used two baseline models, including LDA[4] and G-LDA[2]. For both baseline models, we use their open source codes publicly available on the net. A pre-trained 50-dimensional word embeddings[5] were used. Especially for MvTM, we normalized the word embeddings.

4.2 Evaluation on Topics

We use the PMI score (Newman et al., 2010) to evaluate the quality of topics learnt by topic models. This metric is based on the pointwise mutual information of a power-law reference corpus. For a topic *k*, given *T* most probable words the PMI score is computed by:

$$PMI\left(k\right) = \frac{1}{T\left(T-1\right)} \sum_{1 \le i \le j \le T} \log \frac{p\left(w_i, w_j\right)}{p\left(w_i\right) p\left(w_j\right)} \tag{11}$$

where $p\left(w_i\right)$ and $p\left(w_i, w_j\right)$ are the probabilities of occurring word w_i and co-occurring word pattern (w_i, w_j) estimated by the reference corpus, respectively. In the experiments, we use the Palmetto[6] tool

[2] https://github.com/rajarshd/Gaussian_LDA

[3] http://www.cs.princeton.edu/∼mdhoffma/

[4] http://gibbslda.sourceforge.net/

[5] GloVe word embeddings available at http://nlp.stanford.edu/projects/glove/

[6] http://aksw.org/Projects/Palmetto.html

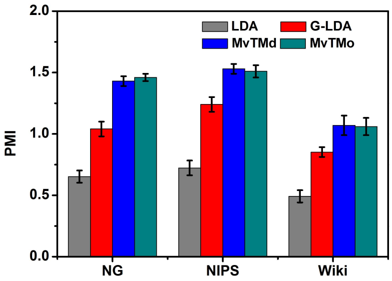


Figure 1: PMI performance of 15 top words on NG, NIPS and Wiki.

Table 2: Random selected examples of top words learnt by baseline models and our MvTM on NG.

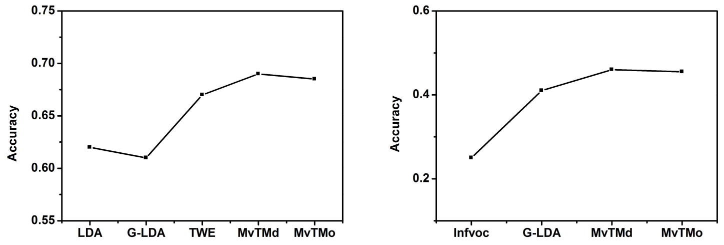
LDA				G-LDA			
president	car	treating	space	government	car	disease	space
government	cars	writes	nasa	administration	university	food	nasa
fbi	engine	medical	gov	support	ohio	treatment	spacecraft
mr	good	cancer	orbit	state	cars	doctor	earth
clinton	oil	doctor	writes	military	carolina	medical	orbit
koresh	mr	doesn	don	leaders	virginia	eat	level
children	speed	treatment	moon	groups	harvard	patients	mars
people	drive	brain	mission	public	speed	cancer	put
batf	ford	patients	solar	policy	michigan	drink	asked
administration	article	drug	water	forces	missouri	course	shuttle
MvTM$_d$				MvTM$_o$			
country	car	disease	earth	country	wheel	patients	space
western	cars	treatment	orbit	government	door	treatments	earth
arab	driver	medical	mars	state	gear	therapy	orbit
muslim	bike	patients	light	president	car	treatment	mars
territory	drivers	infection	space	public	pulling	diabetes	spacecraft
government	truck	drugs	orbiting	policy	inside	diseases	light
war	vehicle	diseases	jupiter	leaders	wheels	hiv	surface
occupation	driving	brain	solar	administration	front	treating	orbiting
eastern	vehicles	tests	orbiter	war	stuck	disease	solar
occupied	wheel	treating	spacecraft	people	rolled	vaccine	orbiter

to compute PMI scores of the top 15 words.

We train baseline models and our MvTM with 50 topics, and evaluate the average PMI score of all topics. For MvTM$_d$, the number of vMF bases is set to 2, i.e., $C = 2$. For MvTM$_o$, topics are organized into ten groups, where each group consists of five topics; and the numbers of personal vMF bases and public vMF bases are set to 2 and 3, respectively[7].

The experimental PMI results on three data sets are shown in Figure 1. It is clearly seen that MvTM performs better than LDA and G-LDA. This implies that MvTM outputs more coherent topics. Some examples of top topic words are listed in Table 2. Overall, we see that the topics of MvTM seem more coherent than those of baseline models. The topics of LDA contain some noise words, e,g., "mr" and "don"; and G-LDA contains some less relevant words, e.g., the second topic of G-LDA is incoherent. In contrast, the topics of MvTM are more precise and clean. Besides, for MvTM$_o$ we measure topic correlation by computing the cosine between vMF weights of topics in the same group. Some topic pairs with high cosine similarity scores, such as $\langle patients, treatments, therapy, treatment, diabetes\rangle$ and $\langle blood, skin, heart, stomach, breathing\rangle$, seem semantically correlated.

[7] In previous experiments, we found that using mixtures of vMFs with 2 bases is able to better represent topics.



(a)

(b)

Figure 2: Classification performance on NG: (a) original test documents and (b) test documents with new words.

4.3 Evaluation on Classification

We compare the classification performance of MvTM with baseline topic models across NG. Two new baselines are used, i.e., topical word embedding (TWE) (Liu et al., 2015) and infvoc (Zhai and Boyd-Graber, 2013). For all models, we learn the topic proportions (K=50) as features of documents, and then use the SVM classifier implemented by LibSVM[8].

The results of original test documents are shown in Figure 2(a). Clearly, MvTM achieves better performance than LDA, G-LDA and TWE. MvTM can handle absent words in training data. To examine this ability, we compare MvTM with G-LDA and infvoc[9], where the two also can handle unseen words. We replace a number of words in test documents with synonyms by using WordNet as in (Das et al., 2015). The classification results are shown in Figure 2(b). It can be seen that MvTM outperforms G-LDA and infvoc. The results imply that MvTM works well even future documents containing new words. This may be insignificant in practice.

5 Related Work

Some early works have attempted to combine topic modeling with embeddings. (Hu et al., 2012) proposed a model to describe indexing representations for audio retrieval, which is similar with G-LDA. Another work (Wan et al., 2012) jointly estimates parameters of a topic model and a neural network to represent topics of images.

Recently, (Liu et al., 2015) proposed a straightforward TWE model. This model separately trains a topic model and word embeddings on the same corpus, and then uses the average of embeddings assigned to the same topic as the topic embedding. A limitation of TWE is that it lacks statistical foundations. Another modification latent feature topic modeling (LFTM) (Nguyen et al., 2015) extends LDA and Dirichlet multinomial mixture by incorporating word embeddings as latent features. However, LFTM may be infeasible for large-scale data sets, since it, i.e., the code provided by its authors, is time-consuming. A most recent nonparametric model (Batmanghelich et al., 2016) also uses vMF to describe the topic over word embeddings, where a topic is represented by a single vMF on the embedding space. By contrast, it may be less effective to capture complex topic structure.

6 Conclusion and Discussion

In this paper, we investigate how to improve topic modeling with word embeddings. A previous art G-LDA defines Gaussian topics over word embeddings, however, the word weights of topics are measured by the Euclidean similarity. To address this problem and further capture complex topic structure, we use mixtures of vMFs to model topics, and then propose a novel MvTM algorithm. The vMF bases of topics in MvTM can be either disjoint or overlapping, leading to two versions of MvTM. The overlapping MvTM can describe topic correlation in some degree. In empirical evaluations, we use the per-trained GloVe word embeddings, and then compare MvTM with LDA and G-LDA on three real-world data

[8] https://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[9] For fair comparison, we train infvoc by a batch optimization procedure.

sets. The experimental results indicate that compared to the state-of-the-art baseline models MvTM can discover more coherent topics measured by PMI, and achieve competitive classification performance. In the future, we are interested in supervised versions of MvTM, directly applying to basic document tasks such as sentiment analysis.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) under the Grant No. 61133011, and 61103091. We thank the reviewers for their useful comments.

References

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382.

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. *arXiv:1604.00126v1*.

Yoshua Bengio, Rjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

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

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Annual Meeting of the Association for Computational Linguistics*, pages 795–804.

Siddharth Gopal and Yiming Yang. 2014. Von mises-fisher clustering models. In *International Conference on Machine Learning*, pages 154–162.

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.

Aaron Q. Li, Amr Ahmed, Sujith Ravi, and Alexander J. Smola. 2014. Reducing the sampling complexity of topic models. In *International Conference on Knowledge Discovery and Data Mining*.

Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. 2016a. Generative topic embedding: a continuous representation of documents. In *Annual Meeting of the Association for Computational Linguistics*.

Ximing Li, Jihong Ouyang, and Xiaotang Zhou. 2016b. Sparse hybrid variational-gibbs algorithm for latent Dirichlet allocation. In *SIAM International Conference on Data Mining*.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and MaosongSun. 2015. Topical word embeddings. In *Association for the Advancement of Artificial Intelligence*, pages 2418–2424.

George Marsaglia, Wai Wan Tsang, and Jingbo Wang. 2004. Fast generation of discrete random variables. *Journal of Statistical Software*, 11:1–8.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

David Mimno, Matthew D. Hoffman, and David M. Blei. 2012. Sparse stochastic inference for latent Dirichlet allocation. In *International Conference on Machine Learning*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwi. 2010. Automatic evaluation of topic coherence. In *Annual Conference of the North American Chapter of the ACL*, pages 100–108.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Alastair J. Walker. 1977. An efficient method for generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software*, 3(3):253–256.

Li Wan, Leo Zhu, and Rob Fergus. 2012. A hybrid neural network-latent topic model. In *International Conference on Artificial Intelligence and Statistics*, pages 1287–1294.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *International Conference on Knowledge Discovery and Data Mining*, pages 424–433.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *International Conference on Knowledge Discovery and Data Mining*.

Ke Zhai and Jordan L. Boyd-Graber. 2013. Online latent Dirichlet allocation with infinite vocabulary. In *International Conference on Machine Learning*, pages 561–569.