Neural Variational Correlated Topic Modeling

Luyang Liu
Department of Computer Sciences,
Beijing Institute of Technology;
Beijing Engineering Research Center
of High Volume Language
Information Processing and Cloud
Computing Applications
Beijing, China
lly_aegis@foxmail.com

Heyan Huang*
Department of Computer Sciences,
Beijing Institute of Technology;
Zhejiang Lab
Beijing, China

Yang Gao
Department of Computer Sciences,
Beijing Institute of Technology
Beijing, China
gyang@bit.edu.cn

Xiaochi Wei
Baidu Inc.
Beijing, China
weixiaochi@baidu.com

Yongfeng Zhang
Department of Computer Sciences,
Rutgers University
New Brunswick, USA
yongfeng.zhang@rutgers.edu

ABSTRACT

With the rapid development of the Internet, millions of documents, such as news and web pages, are generated everyday. Mining the topics and knowledge on them has attracted a lot of interest on both academic and industrial areas. As one of the prevalent unsupervised data mining tools, topic models are usually explored as probabilistic generative models for large collections of texts. Traditional probabilistic topic models tend to find a closed form solution of model parameters and approach the intractable posteriors via approximation methods, which usually lead to the inaccurate inference of parameters and low efficiency when it comes to a quite large volume of data. Recently, an emerging trend of neural variational inference can overcome the above issues, which offers a scalable and powerful deep generative framework for modeling latent topics via neural networks. Interestingly, a common assumption for the most neural variational topic models is that topics are independent and irrelevant to each other. However, this assumption is unreasonable in many practical scenarios. In this paper, we propose a novel Centralized Transformation Flow to capture the correlations among topics by reshaping topic distributions. Furthermore, we present the Transformation Flow Lower Bound to improve the performance of the proposed model. Extensive experiments on two standard benchmark datasets have well-validated the effectiveness of the proposed approach.

CCS CONCEPTS

• **Information systems → Document topic models**; • **Computing methodologies → Natural language processing**.

*Corresponding author.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313561

KEYWORDS

Natural language processing;topic model;neural variational inference

ACM Reference Format:
Luyang Liu, Heyan Huang, Yang Gao, Xiaochi Wei, and Yongfeng Zhang. 2019. Neural Variational Correlated Topic Modeling. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313561

1 INTRODUCTION

Nowadays, the numerous texts, such as online news and reports, are generated by various kinds of daily web services. Mining and organizing large number of topics and knowledge from them has attracted much attention [2, 16, 28, 33, 35]. As one of the most prevalent unsupervised algorithms in data mining and natural language processing, topic model has become such a successful technique to discover topics from collections. The conventional topic models, such as Latent Dirichlet Allocation (LDA) [9], offer a practical and explainable probability generative process for modeling topics. In the LDA, topic distribution of a document is drawn as a multinomial with Dirichlet prior, and document words are drawn from their topical words distribution. Then many LDA-like probability topic models are proposed. However, since generative process is getting expressive as the data grows, the inference of topic models tends to be challenging and tricky due to the complicated intractable marginal likelihood.

Recent efforts in neural variational inference (NVI) [23] offers a powerful auto-encoding framework to handle with large amount of data, and it can also facilitate topic modeling. The advantage of the NVI is that it replaces the arduous works on inference of probabilistic models by applying flexible and powerful neural networks through stochastic back propagation [31]. In addition, the NVI can be automatically applied to a new model with a simple declarative specification of the generative process. Based on the NVI framework, several neural variational topic models are proposed, such as Neural Variational Document Model (NVDM) [23],

Neural Variation Latent Dirichlet Allocation (NVLDA) [34], and Gaussian Softmax Model (GSM) [22]. To reduce the computation complexity, their topics are always modeled as isotropic Gaussian distributions via inference network.

The covariance matrix of the isotropic Gaussian is a diagonal matrix for topical representations. It means that topics are solely independent to each other. Yet, in reality, topics in documents are usually correlated. For instance, the topics about "hardware" are probably related with topics about "software" and " company". Therefore, it is reasonable to expect the relationships among different topics can be modeled to expand the limitations of the classical neural variational topic models in which topics are generated by the isotropic Gaussian distribution. Intuitively, the isotropic Gaussian topic distribution can be replaced by a Gaussian distribution with full covariance matrix, so that the topic correlations can be captured.

Similar challenges also exist in the field of computer vision. Several efforts have been dedicated to solving this independent problem, such as Normalizing flow [30], Inverse auto regressive flow [18], and Householder flow [37]. These flow-based methods map an isotropic Gaussian samples into full covariance one by using several invertible transformations. Among them, the most efficient one is Householder Flow since it only involves linear transformations compared with the other flow-based methods [37]. However, Householder flow cannot be directly applied to the topic models designed for NLP tasks. Because Householder flow can only transform single stochastic Gaussian sample into non-isotropic one, which is obviously insufficient in computing loss function.

To tackle with the aforementioned problems, in this paper, we first propose a Centralized Transformation Flow (CTF) to make the inference network capable of modeling covariance matrix of latent distribution. Based on this, we further propose a Neural Variational Correlated Topic Model (NVCTM) that integrates the CTF with multinomial softmax generative process. The proposed NVCTM can enhance the capability of capturing the correlations among topics with the assistance of the new CTF process. In particular, inspired by involutory property of Householder matrix and linear property of Gaussian, the proposed CTF utilizes the product of several Householder transformations to get final transformation matrix. Then the isotropic Gaussian samples can be transformed into non-isotropic samples via the linear operator. Finally, the document is reconstructed by the multinomial softmax generative model given correlated topic vectors. Besides, to effectively infer with Centralized Transformation Flow, we also present the Transformation Flow Lower Bound (TFLB) to regulate KL divergence term of the objective function. The TFLB utilizes the linear property of Gaussian distribution and gives the closed form KL divergence between encoded non-isotropic Gaussian distribution and topic prior. To evaluate the proposed methods, we conduct experiments on two standard datasets. The experimental results indicate the effectiveness of the correlated topic modeling via the CTF and the TFLB is an appropriate lower bound for the inference of the NVCTM. In summary, this paper makes following contributions:

(1) We propose the CTF for inference network, which is capable of transforming isotropic Gaussian distribution samples into Gaussian samples with full covariance matrix.
(2) We propose the NVCTM model, which is able to capture the correlations among topics via the proposed CTF process.
(3) We design the new TFLB which gives an appropriate KL divergence term of the NVCTM and is capable to effectively facilitate the training of the NVCTM.

The rest of this paper is organized as follows. Section 2 focuses on the related works. In Section 3, we briefly introduce the background knowledge of neural variational topic models and Householder flow. In Section 4, we propose our NVCTM and present its inference process. In Section 5, we introduce the dataset, options of experiments, evaluating metric, results and analysis. Finally, the conclusion of this work is made in Section 6.

2 RELATED WORKS

2.1 Correlated Topic Models

Latent Dirichlet Allocation [9] is one of the most popular approaches to text analysis community. In LDA, the multinomial-based topic distribution cannot model correlations between topics. To solve this issue, a CTM [7] replaces the component-independent Dirichlet topic prior with the full covariance Gaussian to capture the correlations. Similarly, many approaches are continually proposed. The Gaussian Process Topic Model (GPTM) [1] captures the correlations among document collection and adds known similarities among documents via Gaussian process. The above CTM and GPTM models capture the topic correlations by reconstructing topic distributions. Another kind of approaches, such as Gaussian-LDA [11] and Correlated Gaussian Topic Model (CGTM) [38], leverage external resources to model the topic correlations, such as computing relations by embedding-based representations [24, 25]. Notably, for most of the correlation-based topic models, Gaussian distribution is often adopted. It means that it is flexible to model dependent distributions and easy to capture correlations among topics. Traditional topic models utilize directed probability graph to describe their generative processes. Their training methods often adopt sampling methods [3, 27] and variational inference [6, 15]. These training methods requires closed form solution of deviations for updating model parameters to approximate intractable posterior. However, as the expressiveness and structure of generative processes grows, the deviation of parameters tends to be tough and complicated, which also hinders the model's efficiency when it comes to a large scale.

2.2 Neural Variational Topic Models

In recent years, the advances in stochastic variational inference and neural networks brings about the idea of using neural networks for parameter inference. Neural variational inference (NVI) approach [23] makes variational auto- encoders (VAEs) [19] a powerful deep generative framework for topic modeling. In VAEs, the flexible neural network is served as a estimator of the target distribution, which eliminates those arduous and complicated mathematical deviations. Neural variational document model (NVDM) [23] is a typical neural variational topic model with VAEs-like architecture. The NVDM consists of two parts: an inference network to parameterize latent topic distributions and a multinomial softmax generative model to reconstruct the document based on the topic vectors from latent

topic distribution. Similarly, neural variational latent Dirichlet allocation (NVLDA) [34] implements an isotropic Gaussian topic distribution under the Laplace approximation [12]. Gaussian Softmax Model (GSM) [22] normalizes the topic representation of NVDM to get topic proportion of each document. These neural variational approaches share a same drawback: topic distribution is assumed to be an isotropic Gaussian, which makes them incapable of modeling topic correlations.

2.3 Flow-based Methods for VAEs

Recent efforts have been conducted to improve latent variable distribution for neural variational inference. For instance, the flow-based methods such as Normalizing flow [30], which implements a series of invertible functions to transform the original latent variables. Householder flow [37] further utilizes a series of Householder transformation to model the orthogonal matrix and transforms an isotropic Gaussian sample into a non-isotropic one. Additionally, several more complicated flow-based methods, i.e., Hamiltonian Variational Inference (HVI) [32], and linear Inverse Autoregressive Flow (IAF) [18], are also proposed. Among these flow-based methods, the Householder flow is the most efficient one that only involves several linear transformations.

In this paper, our model overcomes the drawbacks of above works and consider both neural variational network and correlations among topics. It is noteworthy that our work introduces neural variational correlated topic model by bringing a new concept of *centralized transformation flow* and enhances the flexibility and interpretations of topic models for text analysis.

3 PRELIMINARIES

In this section, we firstly introduce a most classical framework of neural variational topic mdoel, Neural Variational Document Model (NVDM). Then, we briefly introduce Householder flow. The notations and symbols frequently used in this paper are displayed in Table 1.

Table 1: Notations in our model.

Name	Description
$\mathbf{x}$	Bag-of-word document vector
$\mathbf{h}^{(0)}$	Latent isotropic Gaussian topic vector
k	Length of CTF
$\mathbf{h}^{(k)}$	Latent non-isotropic topic vector after k length of CTF
$\mathbf{H}_i$	$i-th$ Householder matrix in Householder flow
π	Output vector from MLP in Inference network
μ	Mean vector of isotropic Gaussian topic distribution
Σ	Diagonal covariance matrix of isotropic Gaussian topic distribution
$\mathbf{U}$	Transformation matrix of CTF
V	Length of vocabulary
θ	Parameter set of inference network
γ	Parameter set of generative model

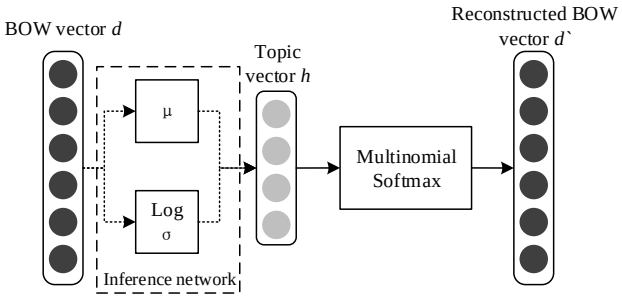


Figure 1: Schematic representation of NVDM.

3.1 Neural Variational Document Model

In traditional probability topic models, the inference of models often requires tricky and complicated math deviations for intractable posterior when the expressiveness of models' generative process grows. To fulfill efficient inference and learning in probabilistic topic models, neural variational inference (NVI) is introduced. In NVI, the inference of model parameters relies on neural networks to approximate the intractable posterior and only limited deviations of parameters are needed.

Neural Variational Document Model (NVDM) [23] is a typical NVI topic model. The schematic representation of NVDM is displayed in figure 1. Each input document is encoded into an isotropic Gaussian distribution via inference network. Then topic vectors are drawn from the Gaussian distributions and passed into a multinomial softmax generative model to reconstruct the input document. Specifically, let $\mathbf{d}$ indicate the bag-of-words document vector, which is the input of inference network; π denotes the output vector of multilayer perceptron (MLP) in inference network. The isotropic Gaussian topic distribution $\mathcal{N}(\mu, \Sigma)$ is parameterized by μ and σ. $\mathbf{l}_1(\cdot)$ and $\mathbf{l}_2(\cdot)$ are two linear neural networks.

$$\begin{aligned} \pi &= MLP(\mathbf{d}) \\ \mu &= \mathbf{l}_1(\pi) \\ \log \sigma &= \mathbf{l}_2(\pi) \\ \Sigma &= diag(\exp 2{\cdot}\mathbf{l}_2(\pi)) \end{aligned} \tag{1}$$

For M documents in corpus and N words in each document, the document generative process of NVDM then can be described as follows:

(1) Topic vector $\mathbf{h} \sim N(\mu, \Sigma)$.
(2) For each word in document $w_n \sim p(w_n|\mathbf{h})$. The definition of $p(w_n|\mathbf{h})$ is given by Equation 2.

$$p(w_n|\mathbf{h}) = \frac{exp\{-\mathbf{h}Wd_n - \mathbf{b}_{d_n}\}}{\sum_{j=1}^{|V|}\{\mathbf{h}Wd_j + \mathbf{b}_{d_i}\}} \tag{2}$$

The training objective of NVDM is to maximize the lower bound of marginal likelihood given in Equation 3.

$$L_{NVDM} = E_q[\log p(\mathbf{d}|\mathbf{h})] - KL[q(\mathbf{h}|\mathbf{d}) \parallel p(\mathbf{h})] \tag{3}$$

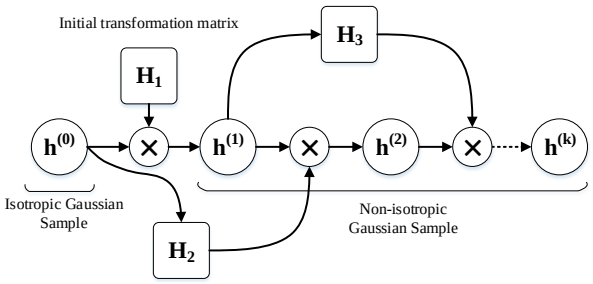


Figure 2: Schematic representation of Householder flow.

The objective function, which is also known as evidence lower bound (ELBO), has two components, i.e., the negative reconstruction term and the KL divergence term. The reconstruction term is mainly determined by generative model and KL divergence term denotes the measurement between the encoded isotropic Gaussian and its prior. The objective of training is equivalent to minimize reconstruction loss as well as KL divergence.

3.2 Householder flow

Previous work [37] introduces Household flow to break independent assumption in latent distributions.

Householder flow utilizes unitary and involutary properties of Householder matrix and establishes iterative invertible linear transformations to map an isotropic Gaussian sample into a non-isotropic one.

Intuitively, to transform an isotropic Gaussian sample into a full covariance one, an orthogonal matrix is needed. Generally, for any Gaussian sample $\mathbf{h}$ $N(\mu, \Sigma)$, it can be generated from a standard Gaussian distribution via $\mathbf{h} = \mu + \Sigma^{\frac{1}{2}} \cdot \epsilon, \epsilon \sim \mathcal{N}(0, I)$ where μ, Σ are parameters of Gaussian distribution. The Σ in isotropic Gaussian is a diagonal matrix. To model an orthogonal matrix, Householder flow can be represented as the product of k Householder transformation [5, 36]. Therefore, according to [37], the Householder matrix $\mathbf{H_i}$ is calculated as

$$\mathbf{H_t} = \begin{cases} \mathbf{I} - 2\dfrac{\pi \cdot (\pi)^{\mathbf{T}}}{\| \pi \|^2}, & t = 1 \\ \mathbf{I} - 2\dfrac{\mathbf{h^{(t-1)}} \cdot (\mathbf{h^{(t-1)}})^{\mathbf{T}}}{\| \mathbf{h^{(t-1)}} \|^2}, & t \geq 2 \end{cases} \tag{4}$$

where $\mathbf{h^{(1)}}$ is a random sample from isotropic Gaussian distribution. Householder matrix, mathematically, refers to an unitary, Hermitian and involutory matrix. Therefore, the product of Householder matrices can be simplified in terms of those useful properties. The vector π is the output vector of MLP in the inference network. We can iteratively get the transformed non-isotropic Gaussian sample via the following equation.

$$\mathbf{h^{(m)}} = \begin{cases} \mathbf{h^{(1)}} & , m = 1 \\ \mathbf{H_{m-1}} & \cdot \mathbf{h^{(m-1)}}, m \geq 2 \end{cases} \tag{5}$$

The schematic representation of the iterative process of Householder flow is illustrated in figure 2. From figure, we can see that Householder Flow only maps single isotropic Gaussian sample to non-isotropic one. However, this is obviously insufficient in computing reconstruction loss mentioned at Equation 1.

4 OUR MODEL

In this section, we describe the proposed Neural Variational Correlated Topic Model (NVCTM) in details. Generally, the structure of NVCTM consists of two main parts, i.e., the inference network with Centralized Transformation Flow and the multinomial softmax generative model. Specifically, the proposed Centralized Transformation Flow in inference network first generates a distributional transformation matrix. The transformation matrix is product of several Householder matrices. The isotropic Gaussian samples are then transformed into non-isotropic ones by multiplying CTF's transformation matrix. The multinomial softmax generative model reconstructs the given documents from the topic vectors that are drawn from the non-isotropic Gaussian distribution. In the following subsections, we introduce details of the Centralized Transformation Flow and the proposed correlated topic modeling. The whole procedure of the model is illustrated in figure 3.

4.1 Centralized Transformation Flow

To transform multiple isotropic stochastic samples into non-isotropic ones, a distribution-level transformation matrix is needed. Therefore, we need to generate it.

Specifically, for each isotropic latent distribution, we choose the mean vector μ as the input vector to generate series of Householder matrices like the iterative process in Householder flow. The iteration process for Householder matrices is given by Equation 4 and Equation 6 where $\pi = MLP(\mathbf{x})$ is used for generating initial Householder matrix $\mathbf{H}_1$ and $MLP(\mathbf{x})$ is the output of MLP inference network.

$$\mathbf{v_m} = \begin{cases} \mu & , m = 1 \\ \mathbf{H_{m-1}} & \cdot \mathbf{v_{m-1}}, m \geq 2 \end{cases} \tag{6}$$

For given length of flow k, the distributional transformation matrix $\mathbf{U}$ can be computed via:

$$\mathbf{U} = \prod_{i=1}^{k} \mathbf{H_i}, where \mathbf{H}_i \in \{\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3, \cdots, \mathbf{H}_k\}. \tag{7}$$

Then we take advantage of the linear property of Gaussian. The final non-isotropic Gaussian samples can be computed via linear transformation mentioned in Equation 8.

$$\begin{aligned} \mathbf{h}^{(0)} &\sim N(\mu, \Sigma) \\ \mathbf{h}^{(k)} &= \mathbf{h}^{(0)} \cdot \mathbf{U} \\ \mathbf{h}^{(k)} &\sim N(\mathbf{U}\mu, \mathbf{U}\Sigma\mathbf{U}) \end{aligned} \tag{8}$$

The procedure of k-th Centralized Transformation flow is depicted in the upper part of figure 3.

Householder flow can only transform single isotropic Gaussian into non-isotropic one, which leads to insufficient approximating accurate the intractable posterior. While in CTF, when transformation matrix is defined, CTF is able to map multiple isotropic Gaussian samples into the non-isotropic ones with linear transformation.

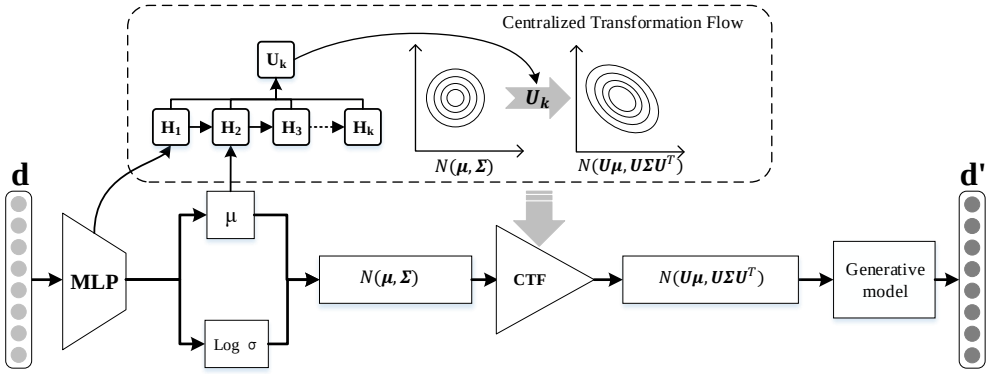


Figure 3: Schematic representation of NVCTM. The dashed rectangle indicates the procedure of CTF. The dashed arrow is short for the iterative process to generate $\mathbf{h^{(k)}}$.

4.2 Correlated Topic Modeling via Centralized Transformation Flow

The isotropic Gaussian topic distribution is unable to capture the correlation between topics. To address this issue, we propose our NVCTM which incorporates the proposed CTF to capture the correlation between topics.

In NVCTM, the proposed CTF is employed in the inference network of the NVCTM to transform the isotropic Gaussian topic distribution into a Gaussian with full covariance matrix. Then the multinomial softmax generative model generates the documents with the correlated topic vectors. The structure of NVCTM denotes in figure 3. The observed bag-of-word vector $\mathbf{d}$ is an input of MLP in inference network. The CTF then utilizes the output of MLP and the mean vector of isotropic Gaussian distribution as initial values to generate the distributional transformation matrix. The correlated topic vectors can be computed via multiplication between isotropic Gaussian topic vectors and transformation matrix in CTF. Finally, the multinomial softmax generative model takes several correlated topic vectors to reconstruct the input document vector $\mathbf{d'}$. After the employment of CTF, the topic distribution in NVCTM is a Gaussian with fully covariance matrix and the topic correlations can be captured. The objective function of NVCTM is denoted as

$$\begin{aligned} &L_{NVCTM} \\ &= E_q[\log p(\mathbf{d}|\mathbf{h^{(k)}})] - KL[q(\mathbf{h^{(k)}}|\mathbf{d})||p(\mathbf{h})] \\ &= E_{q\in N(\mathrm{U}\mu,\mathrm{U}\Sigma\mathrm{U})}[\log p(\mathbf{d}|\mathbf{h^{(k)}})] - KL[N(\mathrm{U}\mu,\mathrm{U}\Sigma\mathrm{U})||N(0,\mathbf{I})]. \end{aligned} \tag{9}$$

The negative reconstruction loss can be automatically computed via stochastic sampling multiple samples from the latent topic space. With the help of reparameterization trick, the whole objective function can be trained with stochastic gradient methods. Accordingly, the proposed CTF refines the latent topic distribution, which indicates that the KL divergence term needs to be modified. In the next subsection, we will precisely describe the deviation of new KL divergence term after applying CTF.

4.3 Inference with Centralized Transformation Flow

After employing CTF, the model inference needs to be modified accordingly. The conventional approach of estimating objective function when flow-based methods applied in VAEs usually refers to the Flow-based Free Energy lower Bound (FELB) [30]. It is calculated as

$$\begin{aligned} L_{FELB} = E_q[\log p(\mathbf{d}|\mathbf{h^{(k)}}) + \sum_{t=1}^{k} \log|det\frac{\partial f^{(t)}}{\partial h^{(t-1)}}|] \\ - KL[q(\mathbf{h^{(0)}}|\mathbf{d})||p(\mathbf{h^{(k)}})], \end{aligned} \tag{10}$$

where $\mathbf{h^{(0)}}$ indicates the isotropic Gaussian sample. In Householder flow, $\log|det\frac{\partial f^{(t)}}{\partial h^{(t-1)}}| = \log det|\mathbf{H_t}|$ equals to zero because Householder matrix is a unitary matrix. In CTF, $\log|det\frac{\partial f^{(t)}}{\partial h^{(t-1)}}| = \log det|\mathbf{U}|$ equals to zero for $\mathbf{U} = \prod_{i=1}^{k}\mathbf{H_i}$, where $\mathbf{H_i}$ is the Householder matrix. Therefore, in the NVCTM, the FELB is calculated as

$$\begin{aligned} L_{FELB} &= E_q[\log p(\mathbf{d}|\mathbf{h^{(k)}}) - KL[q(\mathbf{h^{(0)}}|\mathbf{d})||p(\mathbf{h})] \\ &= E_q[\log p(\mathbf{d}|\mathbf{h^{(k)}})] + 0.5[n - \mu^2 + |\Sigma| + \log|\Sigma|] \end{aligned}$$

where μ and Σ are in Equation 1. $p(\mathbf{h})$ is the prior, which is usually considered as $\mathcal{N}(0, \mathbf{I})$.

However, during experiments, we observes that FELB slows down the training of the model. Here, we present our Transformation Flow Lower Bound (TFLB) to improve the performance of perplexity and facilitate the training process of model. The TFLB utilizes the linear property of Gaussian and can assist get explicit parameters of the final non-isotropic Gaussian distribution. The linear property of Gaussian [26] can be described as follows : Assuming that the random variable and its distribution are given by $\mathbf{h_0} \sim \mathcal{N}(\mu, \Sigma)$, we can transform it into latent variable $\mathbf{h'} = \mathrm{U} \cdot \mathbf{h}$. The corresponding transformed latent distribution is denoted as

$$\mathbf{h'} \sim \mathcal{N}(\mathrm{U}\mu, \mathrm{U}\Sigma\mathrm{U^T}) \tag{11}$$

Then the KL divergence term in the objective function can be computed as

$$\begin{aligned} & KL[\mathcal{N}(\mu_1, \Sigma_1) \parallel \mathcal{N}(\mu_2, \Sigma_2)] \\ =& \frac{1}{2}[\log \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathrm{tr}(\Sigma_2^{-1} \cdot \Sigma_1) \\ & + (\mu_2 - \mu_1)^{\mathrm{T}} \Sigma_2^{-1} (\mu_2 - \mu_1)], \end{aligned} \tag{12}$$

where n is the number of dimensionality in multivariate Gaussian. Due to the fact of unitary property of Householder matrix, the regularization term of TFLB then can be denoted as

$$\begin{aligned} & L_{TFLB} \\ & = E_q[\log p(\mathbf{d}|\mathbf{h}^{(\mathbf{k})})] - KL[q(\mathbf{h}^{(\mathbf{k})}|\mathbf{d}) \parallel p(\mathbf{h})] \\ & = E_q[\log p(\mathbf{d}|\mathbf{h}^{(\mathbf{k})})] - \frac{1}{2}[-\log|\Sigma| - n + \\ & tr(\mathbf{U}\Sigma\mathbf{U}^{\mathrm{T}}) + (\mathbf{U}\mu)^{\mathrm{T}}\mathbf{U}\mu]. \end{aligned} \tag{13}$$

At last, the pseudo code of our method, NVCTM, is depicted in Algorithm 1.

Algorithm 1 NVCTM

Input: Document bag-of-word vector $\mathbf{d}$; document collection $\mathbf{D}$; parameters set of inference network θ; parameter set of generative model γ; length of CTF k

Output: model parameters θ and γ

Remove stop words and low-frequency words and covert document into bag-of-word vector
Initialize θ and γ
for $\mathbf{d} \in \mathbf{D}$ **do**
    Compute μ and Σ using Eq 1.
    Generate $\mathbf{H}_1, \mathbf{H}_1, \mathbf{H}_3, \cdots, \mathbf{H}_k$ via Eq 4 and Eq 6.
    Generate matrix $\mathbf{U}$ by Eq 7.
    Randomly draw 20 topic vector samples h from $N(\mu, \Sigma)$.
    Multiply $\mathbf{U}$ and $\mathbf{h}^{(0)}$ to get $\mathbf{h}^{(k)}$.
    Compute objective function with $\mathbf{h}^{(k)}$ and KL divergence term via Eq 13.
    Update θ via Adam method with previously calculated objective.
    Compute μ and Σ using Eq 1.
    Generate $\mathbf{H}_1, \mathbf{H}_1, \mathbf{H}_3, \cdots, \mathbf{H}_k$ via Eq 4 and Eq 6.
    Generate matrix $\mathbf{U}$ by Eq 7.
    Randomly draw 20 topic vector samples h from $N(\mu, \Sigma)$.
    Multiply $\mathbf{U}$ and $\mathbf{h}^{(0)}$ to get $\mathbf{h}^{(k)}$.
    Compute objective function with $\mathbf{h}^{(k)}$ and KL divergence term via Eq 13.
    Update γ via Adam method with previously calculated objective.
end for

5 EXPERIMENTS

In this section, we first introduce the experimental settings, evaluate the proposed model, then analyse our proposed model with extensive experiments.

Dataset: To evaluate our efforts, we select *20NewsGroups*[1] and *Reuters RCV1-v2*[2] for experiments. 20NewsGroups is a collection of newsgroup documents which consists of 11,314 training and 7,531 testing articles. And Reuters RCV1-v2 is a huge dataset that consists of Reuters newswire stories with 794,414 training and 10,000 testing cases. For data preprocessing, we remove stopwords and take the most frequent 2,000 words and 10,000 words as the vocabularies.

Baseline Methods: We compare the following methods to demonstrate the priority of the proposed method:

- **LDA** [9]: A classical topic model, which models the topic distribution as a multinomial distribution with Dirichlet prior. Here, we utilize the variational inference to implement the LDA [29].
- **CTM** [8]: It implements a log normal topic distribution. The variational inference of CTM implemented by the authors[3] is used for evaluation.
- **NVDM** [23]: The model implements a typical neural variational inference approach. The inference network of NVDM consists of a MLP network, and the generative model of NVDM reconstructs documents from reparameterized isotropic Gaussian distribution. We also use author's codes[4] for evaluation.
- **GSM** [22]: This model extends NVDM by normalizing the topic vector Gaussian sample and words distribution.
- **NVLDA** [34]: The model leverages the Laplace approximation for the LDA generative process. The topic distribution of NVLDA is an isotropic Gaussian.
- **NVCTM:** This is the proposed method in this paper.

5.1 Settings

For all of the NVI-based methods (i.e., NVDM, GSM and the proposed NVCTM), the topic vector is an average of 20 samples drawn from the latent topic distribution. For NVDM, GSM and NVCTM, we follow the authors' setting where the MLP in inference network has 256 hidden units and a hyperbolic tangent activation function. The dropout with probability of 0.8 is applied to the output of MLP network. The linear layers to parameterize μ and σ both have the same amount of hidden units with the number of topics. The training process is divided into two stages: Stage-1: optimizing the encoder's parameters while fixing the parameters of generative model; Stage-2: optimizing the generative model's parameters and keeping the inference network unchanged, which is also called wake-sleep algorithm [13]. The optimization method we use to train NVDM, GSM, NVLDA and NVCTM is Adam [17] with learning rate of $1e-5$. We also utilize early-stop [39] to stop training and export topics and topic vectors for further evaluations. The training methods of LDA and CTM are online variational inference [14] and variational inference [8]. The hyper parameters of the LDA and the CTM, which are initial values of topic distribution prior α, topic word distribution prior β and number of variational EM iterations, are determined via grid search [4] to find its optimal performance on corpus. For the proposed NVCTM, we also carry out grid search

[1] http://qwone.com/jason/20Newsgroups
[2] http://trec.nist.gov/data/reuters/reuters.html
[3] https://github.com/blei-lab/ctm-c
[4] https://github.com/ysmiao/nvdm

Table 2: Performance of perplexity of all the methods.

	20News			RCV1-v2		
T	25	50	100	25	50	100
LDA	1049	1020	1016	1123	1062	1043
CTM	1011	944	896	987	972	1021
NVDM	801	832	846	694	651	628
NVLDA	1105	1073	1034	1345	1321	1378
GSM	879	850	**843**	643	674	768
NVCTM	**715**	**749**	945	**482**	**491**	**528**

to decide the length of CTF k. For experimental environment, we run all baseline methods and proposed NVCTM on a workstation with an eight-cores Xenon E5-2630V4 CPU and GTX TITAN XP GPU.

5.2 Perplexity

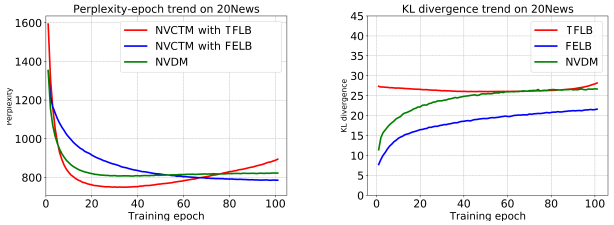
The traditional way of evaluating topic model usually refers to the perplexity computed on unseen documents. For language modeling, it refers to the inverse of geometric average per-word likelihood. The lower perplexity usually indicates the better generalization performance. The perplexity is given by:

$$Perplexity = \exp\left[-\frac{1}{D}\sum_{n}^{N_d}\frac{1}{N_d}\log p(\mathbf{X}_d)\right] \tag{14}$$

We train the baseline models and the proposed NVCTM on the training set of 20News and RCV1-v2, and compute perplexity of the corresponding models on test set. To evaluate the performance of the proposed model more comprehensively, we also conduct experiments with different number of topics, i.e., T=25, 50, 100. The performance on perplexity is displayed in Table 2. From the table, it can be observed that the NVI models (i.e., NVDM, GSM, NVCTM except for NVLDA) always have better perplexity than those of traditional models (i.e., LDA and CTM). Specifically, the proposed NVCTM approach has better performance than other baselines in most cases. For NVDM and NVCTM, which both have the same structure of generative model, we find that NVCTM significantly outperforms NVDM, which indicates that the proposed CTF approach can contribute to document modeling on perplexity. Interestingly, the perplexity performance of NVCTM is better when it comes to the smaller number of topics. This phenomenon is probably caused by the fact that the perplexity of the documents is increased as the complexity of covariance matrix of latent topics increases when the topic number grows. Besides, we also observe that for unnormalized topic vector models, such as NVDM and NVCTM, they usually have better perplexity performance than other normalized topic vector models. This effect may be due to the fact that topic vector normalization makes models difficult to optimize.

5.3 TFLB vs FELB

In the previous subsection, the results of perplexity demonstrate the effectiveness of CTF and TFLB in topic modeling. Previously,



(a) Perplexity-epoch trend on 20News-group. (b) KL divergence trend on 20Newsgroup.

Figure 4: The perplexity trend and KL divergence trend of TFLB and FELB on 20Newsgroup.

the original version of Householder flow utilizes the flow-based energy lower bound (FELB) to compute the KL divergence term. In this subsection, we will quantitatively evaluate our proposed TFLB and previous FELB in topic modeling. Accordingly, we also select perplexity as the main evaluation metric. The results are displayed in figure 4. From the perplexity trend illustrated in figures 4(a), we can see that NVCTM with TFLB can achieve relative lower perplexity than that of FELB. Moreover, compared with FELB, TFLB requires less epoch to reach the lowest perplexity, which indicates that TFLB can facilitate the training of the model. On KL divergence trend graph illustrated in figure 4(b), it shows that the KL divergence of TFLB is slightly larger than that of FELB, while the model with TFLB gets lower perplexity. This indicates that TFLB can reduce the reconstruction loss even if the KL divergence of the model slightly increases. The main reason might be that TFLB explicitly takes covariance matrix to compute the gradient and this enable it to accumulate more gradient on the inference network.

5.4 Topic Coherence

Besides the perplexity, another common method of evaluating topic models is topic coherence. Topic coherence evaluates each topic with given reference corpus and compute corresponding coherence score. Here, we adopt normalized point-wise mutual information (NPMI) topic coherence proposed by [20] to automatically evaluate the proposed model as well as baselines. The NPMI coherence score is shown to be close to human judgments [20], which is why we adopt this method for evaluation. The reference corpus of corresponding corpus is the training set of corpus. We extract top-10 and top-5 words of each topic and compute the NPMI score for each word set with the equation of

$$\text{NPMI(N)} = \sum_{j=2}^{N}\sum_{i=1}^{j-1}\frac{\log\frac{P(w_j,w_i)}{P(w_i)P(w_j)}}{-\log P(w_i,w_j)}. \tag{15}$$

Then we take the average value of two NPMI scores as the final topic coherence score. We use the author implemented version script[5] to automatically evaluate the baseline methods as well as NVCTM. To fully evaluate the topic coherence, we also compute the topic coherence score of the corresponding model on topic number T=25, 50, 100 . The topic coherence scores achieved by our

[5]https://github.com/jhlau/topic_interpretability

model and the baselines are depicted in Table 3. From the table, we find that NVCTM slightly outperforms the baselines in most cases. On 20News, NVCTM usually has better topic coherence scores than those of the baseline methods when the number of topic are 25 and 50. When the number of topic increases to 100, the topic coherence score of NVCTM decreases to 0.158, which is pretty close to the best performance of CTM. For RCV1-v2, NVCTM outperforms baselines when the number of topics are 25 and 50. Similarly, when the number of topics increases, the topic coherence of NVCTM, NVDM and GSM tends to decrease, which indicates that those models are effective on relative small amount of topics. For NVDM and proposed NVCTM, which both have the same structure of generative model, the result also indicates that proposed CTF can improve the topic coherence score of the multinomial softmax generative model.

5.5 Document Classification

To further evaluate the performance of the proposed method, document classification experiments are conducted on both 20News and RCV1-v2 datasets. To efficiently evaluate the proposed approaches and baselines, we randomly select 11 categories of RCV1-v2, namely RCV-115k, which consists of 114, 324 documents as training set and 1, 439 documents as test set evaluation. Similar to what is usually done in document classification, the topic vectors of the corresponding documents can be regarded as the low dimensional representation of the sparse document vectors. We then use these vectors to train an SVM classifier [10] with multinomial kernel function and evaluate the precision, recall and F-1 measure of the corresponding methods. The topic number is set to 50 equally of all the topic models for evaluation, in terms of the performance of document classification, which is depicted in Table 4. the results indicate that the proposed NVCTM method outperforms the baseline methods on 20News and RCV-115k. Compared with NVDM, which has the same structure generative model with NVCTM, NVCTM improves its document classification performance on both datasets, which shows the effectiveness of CTF in inference network. Among all approaches, the topic vectors of NVDM, NVLDA and NVCTM are assumed to be real-valued vectors, while the topic vectors of other methods are normalized vectors. This indicates that real-valued topic vectors can enhance the performance of document classification. To further demonstrate the topic vectors of documents and their labels, we export the topic vectors for 20News and RCV1-v2. We then visualize them by t-SNE algorithm. The visualizations of the topic vectors on 20News and RCV-115k are illustrated in figure 5. The dots with same color indicate they are the documents from same category. From figure, we find that the aggregation of vectors in RCV-115k is better than that of 20News. The better aggregation of vectors will facilitate the performance of classification.

5.6 Visualization of Topics

To further quantitatively investigate the quality of topics and topic correlation mined by NVCTM, we export the topic and the corresponding word distributions, and visualize the topics with their corresponding top 10 words on the 20NewsGroup dataset. To clearly make the visualization, we select 5 topics, which are considered to have correlations by NVCTM, and then utilize t-SNE [21] for visualization. Visualization of the topics and the topic word correlation graphs is depicted in figure 6. The points with different shapes and colors indicate different topics. The dashed circles denote the corresponding topic word distribution, and those circles with the same color are recognized to be correlated by NVCTM. We also manually annotate each topic with the corresponding label. Accordingly, they can be annotated as “comp.os.ms-windows”, “comp.graphics”, “comp.sys.ibm.pc.hardware”, “talk.politics.guns”, “talk.politics.mideast” from the ground truth labels of 20NewsGroup. From figure 6, we can see that the topics of “hardware”, “windows” and “graphics” are correlated in continuous vector space. Similarly, another two correlated topics are “mideast” and “guns”. This verifies the effectiveness of the proposed method in mining topics and correlations.

5.7 Visualization of Topical Correlations

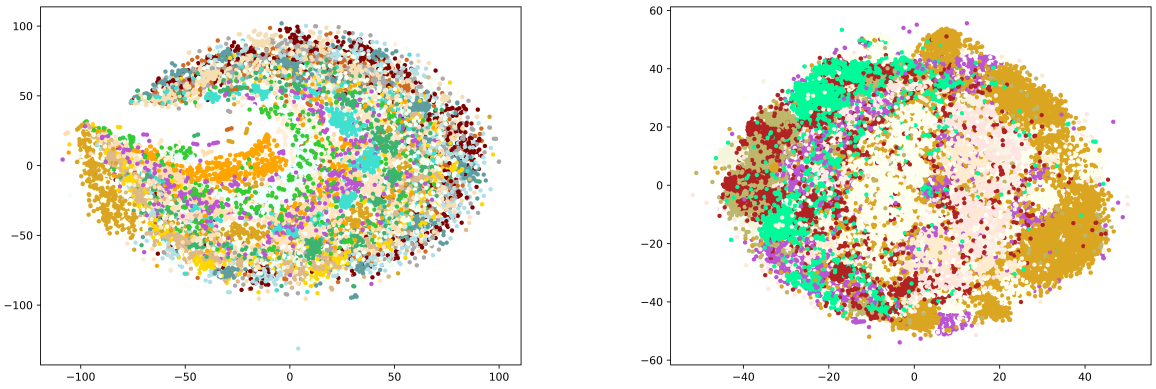
To further demonstrate effectiveness of proposed model, we export and visualize the covariance matrix of latent topic distribution on 20News and RCV1-v2. Unlike traditional mean-field variational inference or mean-field neural variational inference methods, the proposed NVCTM is capable of modeling the full covariance matrix of latent topic distributions. The main advantage of our model is that, with help of proposed CTF, NVCTM is capable of modeling the full covariance matrix rather than the diagonal one. Mathematically, the diagonal elements of covariance matrix are usually the variance values for corresponding dimension and non-diagonal elements are correlation of different two dimension. Therefore, to reach clarified visualization, we get the final correlation matrix by zeroing all the diagonal elements of covariance matrix and normalizing the matrix with rest values. The numbers of topics in NVCTM are 50 for 20News and 100 for RCV1-v2. The visualization of correlation matrix is illustrated in figure 7. From figure 7(a), we can obviously notice that the correlation matrix is non-diagonal. Besides, the most values of correlation matrix are nearly zero. This phenomenon is mainly due to the latent topic prior $N(0, \mathbf{I})$, where $\mathbf{I}$ is a diagonal unit matrix. On the other hand, each document in 20News is corresponding to a specific class of 20News, which also demonstrates the sparsity in topic correlation. The optimization of KL divergence term between CTF refined distributions and prior will make them similar to each other, which indicates that prior can also act as the sparse regularization on correlation matrix. Besides, we also notice that there are 3 topics with the higher correlation with other topics (i.e., topic No.3, 8 and 12. These 3 topics are visualized with light color which represents high correlations with other topics.). We then export these topical word distributions and find that the words are some common words, i.e, “who”, “are”, “etc”. The main reason of this effect is that these common words often exists in many documents, which makes model believe that they are correlated. On the other hand, we can still find several points with relatively large correlation coefficients. For instance, the correlation coefficients of 37 and 49, 40 and 49 are around $0.6 \sim 0.8$ which indicates the correlations exist among those topics.For figure 7(b), we can see that the number of non-zero values are larger than that of 20News. This probably results from the fact that each documents of RCV1-v2 dataset has multiple labels, which also indicates that topics in the

Table 3: Topic coherence on 20News and RCV1-v2 datasets.

	20News			RCV1-v2		
The numbers of topics	25	50	100	25	50	100
LDA	0.112	0.140	0.151	0.112	0.131	**0.143**
CTM	0.149	0.154	0.161	0.094	0.072	0.076
NVDM	0.163	0.165	0.140	0.125	0.118	0.137
NVLDA	0.167	0.145	0.110	0.121	0.136	0.107
GSM	0.141	0.132	0.111	0.101	0.076	0.062
NVCTM	**0.180**	**0.176**	0.158	**0.146**	**0.139**	0.113

Table 4: Performance of document classification on datasets of 20News and RCV-115k. The topic number of all methods is set to 50**.**

	20News			RCV-115k		
	Precision	Recall	F-1 measure	Precision	Recall	F-1 measure
LDA	0.421	0.501	0.458	0.446	0.447	0.446
CTM	0.503	0.431	0.464	0.484	0.502	0.493
NVDM	0.539	0.527	0.533	0.781	**0.798**	0.789
NVLDA	0.451	0.471	0.461	0.501	0.503	0.502
GSM	0.347	0.346	0.347	0.432	0.419	0.425
NVCTM	**0.577**	**0.564**	**0.570**	**0.818**	0.786	**0.802**



(a) The visualization of topic vectors on 20News. The length of CTF k is 5 and the number of topics is 50.

(b) The visualization of topic vectors RCV-115k.The length of CTF k is 3 and the number of topics is 100.

Figure 5: t-SNE Visualization of averaged estimated topic vector $\mathbf{h}^{(k)}$ of each document. The points with same color indicates that they have same class label in corresponding dataset.

document are likely to be correlated. Therefore, the topic correlation matrix in RCV1-v2 are reasonable to be a non-sparse matrix. To sum up, the above visualizations of correlation matrix further indicate the effectiveness of proposed approach.

6 CONCLUSION

In this paper, we first propose the Neural Variational Correlated Topic Model (NVCTM) model, which incorporates the Centralized Transformation Flow (CTF) to capture the topic correlations. CTF can enable the model to capture the correlations among topics as well as to sufficiently compute the marginal likelihood. We then

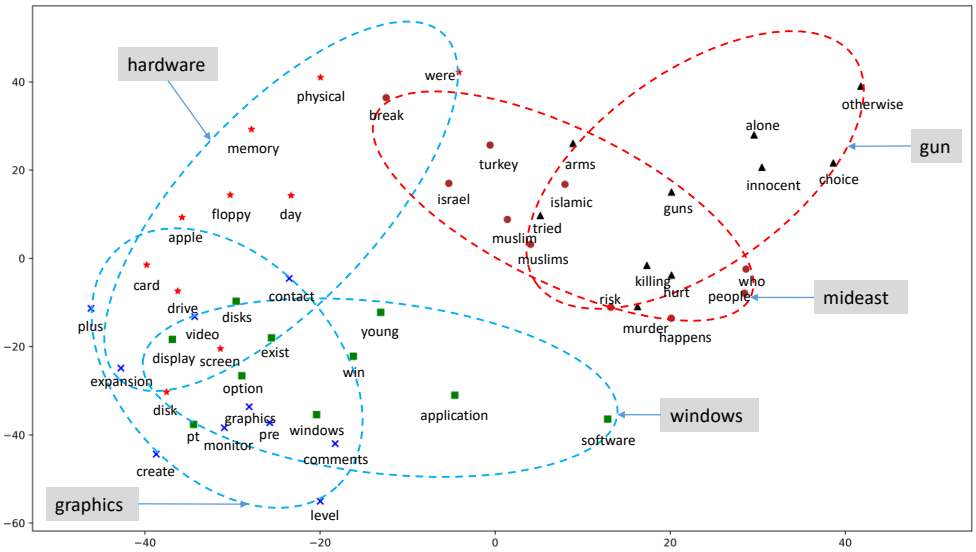
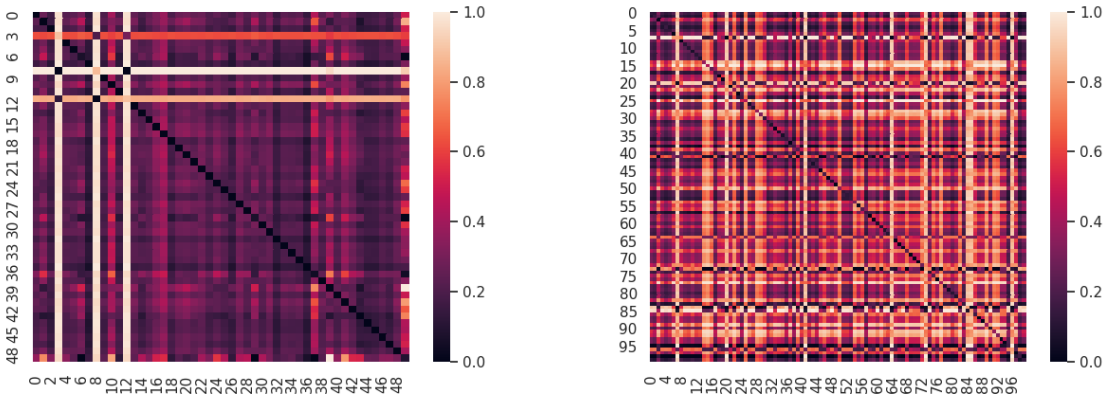


Figure 6: The t-SNE visualization of topics and topic word correlation.



(a) The visualization of Correlation matrix on 20news.

(b) The visualization of Correlation matrix on RCV1.

Figure 7: Visualization of correlation on covariance matrix. The diagonal elements are zeroed and non-diagonal elements are normalized. The deeper color indicates the smaller value.

present the Transformation Flow Lower Bound (TFLB) to regulate the objective function. It leverages the linear property of Gaussian distribution, which can regulate the model optimization and facilitates the training process of NVCTM. In order to quantitatively verify our contributions, we conduct experiments in terms of perplexity, topic coherence, and document classification tasks. The experimental results show that the proposed NVCTM approach is effective to capture topic correlations and improve the performance of topic modeling. Moreover, the visualization of the topics and their correlations qualitatively verified the effectiveness of NVCTM.

7 ACKNOWLEDGEMENT

We would like to thank Yuxiang Zhou and Yishu Miao for their insightful comments and suggestions. We also very appreciate the comments from anonymous reviewers which help further improve our work. This work is supported by National Key R&D Plan(No.2016QY03D0602), "Key technologies, system and application of Cyberspace Big Search", Major project of Zhejiang Lab (No.2019DH0ZX01), National Natural Science Foundation of China (Grant No.61602036).

REFERENCES

[1] Amrudin Agovic and Arindam Banerjee. 2012. Gaussian Process Topic Models. *CoRR* abs/1203.3462 (2012). http://arxiv.org/abs/1203.3462

[2] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. 2008. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 3–12.

[3] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. 2003. An introduction to MCMC for machine learning. *Machine learning* 50, 1-2 (2003), 5–43.

[4] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13, Feb (2012), 281–305.

[5] Christian H Bischof and Xiaobai Sun. 1994. On orthogonal block elimination. *Preprint MCS-P450-0794, Mathematics and Computer Science Division, Argonne National Laboratory* (1994).

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* 112, 518 (2017), 859–877.

[7] David M. Blei and John D. Lafferty. 2005. Correlated Topic Models. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*. 147–154. http://papers.nips.cc/paper/2906-correlated-topic-models

[8] David M. Blei and John D. Lafferty. 2005. Correlated Topic Models. (2005), 147–154. http://papers.nips.cc/paper/2906-correlated-topic-models

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani (Eds.). MIT Press, 601–608. http://papers.nips.cc/paper/2070-latent-dirichlet-allocation

[10] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[11] 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 and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 795–804.

[12] Philipp Hennig, David Stern, Ralf Herbrich, and Thore Graepel. 2012. Kernel topic models. In *Artificial Intelligence and Statistics*. 511–519.

[13] Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. 1995. The" wake-sleep" algorithm for unsupervised neural networks. *Science* 268, 5214 (1995), 1158–1161.

[14] Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*. 856–864.

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

[16] Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*. ACM, 80–88.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Diederik P. Kingma, Tim Salimans, and Max Welling. 2016. Improving Variational Inference with Inverse Autoregressive Flow. *CoRR* abs/1606.04934 (2016).

[19] Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2013). http://arxiv.org/abs/1312.6114

[20] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 530–539.

[21] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[22] Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering Discrete Latent Topics with Neural Variational Inference. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, 2410–2419. http://proceedings.mlr.press/v70/miao17a.html

[23] Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural Variational Inference for Text Processing. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016 (JMLR Workshop and Conference Proceedings)*, Maria-Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. JMLR.org, 1727–1736. http://jmlr.org/proceedings/papers/v48/miao16.html

[24] 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*. 3111–3119.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[26] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. 2008. The matrix cookbook. *Technical University of Denmark* 7, 15 (2008), 510.

[27] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 569–577.

[28] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 248–256.

[29] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[30] Danilo Jimenez Rezende and Shakir Mohamed. 2015. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings)*, Francis R. Bach and David M. Blei (Eds.), Vol. 37. JMLR.org, 1530–1538. http://jmlr.org/proceedings/papers/v37/rezende15.html

[31] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014 (JMLR Workshop and Conference Proceedings)*, Vol. 32. JMLR.org, 1278–1286. http://jmlr.org/proceedings/papers/v32/rezende14.html

[32] Tim Salimans, Diederik Kingma, and Max Welling. 2015. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*. 1218–1226.

[33] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1105–1114.

[34] Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference For Topic Models. *arXiv preprint arXiv:1703.01488* (2017).

[35] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 306–315.

[36] Xiaobai Sun and Christian Bischof. 1995. A basis-kernel representation of orthogonal matrices. *SIAM journal on matrix analysis and applications* 16, 4 (1995), 1184–1196.

[37] Jakub M. Tomczak and Max Welling. 2016. Improving Variational Auto-Encoders using Householder Flow. *CoRR* abs/1611.09630 (2016). http://arxiv.org/abs/1611.09630

[38] Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A Correlated Topic Model Using Word Embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 4207–4213. https://doi.org/10.24963/ijcai.2017/588

[39] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation* 26, 2 (2007), 289–315.