Correlated Topic Modelling via Householder Flow

Luyang Liu*
Department of Computer Science and Technology, Beijing Institute of technology
Beijing, China
lly_aegis@foxmail.com

Heyan Huang
Department of Computer Science and Technology, Beijing Institute of technology
Beijing, China
hhy63@bit.edu.cn

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

ABSTRACT

Topic models can be one of the prevalent unsupervised learning approaches in natural language processing. Recent works on neural variational inference offer a powerful framework combining neural networks and interpretable probability models. However, one fundamental assumption is that topics in the latent space are independent to each other, which is actually not the case in the reality. In this paper, we propose the Correlated Householder Topic Model (CHTM) to capture the correlations among topics, and model them via Householder flow. The experiments show that ,by incorporating topic correlation, CHTM outperforms baseline methods on two standard datasets.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**;

KEYWORDS

Neural Variational Inference, Variational Auto-Encoder, Correlated topic model

ACM Reference Format:
Luyang Liu, Heyan Huang, and Yang Gao. 2018. Correlated Topic Modelling via Householder Flow. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Topic models, such as Latent Dirichlet Allocation (LDA) [4], have proven to be successful unsupervised methods for mining topics among document collections. In LDA, each document is assumed to be a mixture of topics and the topic proportion of each document is given from a Dirichlet distribution. The components of Dirichlet distribution on topic proportions are nearly independent, thus leading to the strong and unrealistic limitation in topic modelling. To reform this drawback, Correlated Topic Model (CTM) [3] replaces the multinomial topic mixture distribution with a logistic normal distribution to capture the correlation between topics. Likewise, some other correlated topic models, such as Gaussian Process Topic Models[1], also model topic as an Gaussian distribution to capture the correlation between topics. However, due to that non-conjugate prior of topic, the inference of these models tend to be complicated and tricky.

To address this issue, the neural variational topic models, such as Neural Variational Document Model (NVDM) [10], are proposed. Based on Neural Variational Inference (NVI) [7, 8, 11, 12], in neural variational topic model, neural networks is applied for model inference. The generative model usually consists of interpretable procedures for document generation. To reach the computation efficiency, the topic distribution of neural variational topic model is usually an isotropic Gaussian, which, yet, leads to the problem that the topics are independent and the correlations of topics are ignored.

To remedy this problem, the isotropic Gaussian needs to be transformed into a full covariance one. In NVI, recent efforts to do this usually refer to the flow-based methods such as Normalizing flow [14] and Householder flow [17]. These flow-based methods implement several functions to establish the correlation among different dimensions of latent variables. Householder flow, as one of the unitary flows, is an efficient approach among those flow based methods. It only relies on several linear transformations which can facilitate computation.

In this paper, to solve the aforementioned problem, we present Correlated Householder Topic Model (CHTM) which can establish the topic correlation modelling via Householder flow. To efficiently estimating the parameters of CHTM, we apply Free-energy based lower bound [14] to help the training. Notably, Our work is first approach to introduce the topic correlation modelling in neural variational topic models. In summary, the main contributions of this paper include:

(1) Householder flow is introduced to model the topic correlation in neural variational topic models.
(2) To efficiently estimate the objective function, we apply Free-energy based lower bound in estimating the objective function of CHTM.
(3) Our proposed CHTM achieves better performance on two standard datasets than those of baseline methods.

The paper organizes as follows: Section 2 gives brief introduction of neural variational topic models. Section 3 describes our proposed CHTM in detail. Section 4 describes the inference of proposed CHTM. Section 5 introduces related works. Section 6 introduces the experimental setting, evaluation metrics and baseline methods. The conclusion is given in Section 7.

*Corresponding author.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.475/123_4

2 NEURAL VARIATIONAL TOPIC MODELS

In this section, we give a brief introduction to neural variational topic models. Recent works, such as NVDM, can be interpreted as a Variational Auto Encoders: A encoder for mapping documents to latent distributions and a decoder to generate documents from given latent distributions. Thus, we first give a description of VAEs framework in topic modelling. As a typical neural variational topic model, NVDM then is discussed in detail.

2.1 Variational Auto Encoder in Topic modelling

In terms of VAEs in topic modelling, the input is supposed to a bag-of-words document vector $\mathbf{x}$. For each document in collection $\mathcal{D} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}\}$, VAEs tries to maximize the marginal likelihood of observed data:

$$\max \log p(\mathcal{D}) = \max \sum_{i=1}^{n} \log p(\mathbf{x_i}) \tag{1}$$

The task, however, is intractable due to the fact that the generative model is parameterized by a neural network. To avoid arduous work of variational inference, the inference network is introduced to inference the model parameters [5]:

$$\log p(\mathbf{x}) \geq E_{q_\phi(\mathbf{h}|\mathbf{x})}[\log p_\psi(\mathbf{x}|\mathbf{h})] - KL(q_\phi(\mathbf{h}|\mathbf{x} \parallel p(\mathbf{h})) \tag{2}$$

where $\mathbf{h}$, in neural variational topic models, is usually the latent topic representation or other topic related factors. $q_\phi(\mathbf{h}|\mathbf{x})$ is the encoder network with parameter ϕ. $p_\psi(x|h)$ is the decoder with parameter ψ. $p(\mathbf{h})$ is the prior distribution given by $\mathcal{N}(0, \mathbf{I})$. The optimization of Eq.(2) can be efficiently done via reparameterization [7, 15] of latent variable $\mathbf{z}$.

During training of neural variational topic models, the inference network focuses on mapping observed document vectors into latent topic distributions. The generative model then tries to ensure the encoded latent topics can generate the given document representations. In neural variational topic models, actually, the inference network serves as an estimator in approximating topic posteriors with observed samples.

2.2 Neural Variational Document Model

Neural Variational Document Model (NVDM) [10] is a simple but powerful case of neural variational topic models. NVDM implements a multilayer perceptron (MLP) inference network to map the bag-of-words document vectors into continuous latent distributions. The generative model then takes samples from latent topic distribution as input of multinomial softmax network to regenerate the document representations.

Specifically, the general structure of NVDM is:

(1) The inference network maps a document vector $\mathbf{d}$ into a continuous latent normal distribution $\mathcal{N}(\mu_0, \Sigma_0)$ via a MLP neural network. The parameters of latent distribution are parametrized by two linear layers:

$$\begin{aligned} &\mathbf{d} \xrightarrow{MLP} \mathbf{h} \backsim \mathcal{N}(\mu_0, \Sigma_0) \\ &\mathbf{u} = g(MLP(\mathbf{d})) \\ &\mu_0 = \mathbf{l_1}(\mathbf{u}) \\ &\Sigma_0 = diag(\exp 2 \cdot \mathbf{l_2}(\mathbf{u})) \end{aligned} \tag{3}$$

where $\mathbf{l_1}(\cdot)$ and $\mathbf{l_2}(\cdot)$ are linear transformation layers. $g(\cdot)$ is non-linearity operation.

(2) For the generative model, it first takes some samples from latent normal distribution using the reparameterisation method[7, 15]. Multinomial softmax[10] then generates document vectors with the given samples. The process of multinomial softmax denotes in (4).

$$\begin{aligned} &p_\psi(\mathbf{x_i}|\mathbf{h}) = \frac{exp\{-F(\mathbf{w_i}; \mathbf{h}, \psi)\}}{\sum_{j=1}^{|V|} exp\{-F(\mathbf{w_i}; \mathbf{h}, \psi)\}} \\ &F(\mathbf{w_i}; \mathbf{h}, \psi) = -\mathbf{h^T R w_i} - \mathbf{b_{w_i}} \end{aligned} \tag{4}$$

where $\mathbf{R}$ is the topic-word matrix. $\mathbf{h}$ is a sample from latent topic distribution $\mathcal{N}(\mu_0, \Sigma_0)$. $\mathbf{w_i}$ is corresponding one-hot word index vector in vocabulary.

The goal of NVDM is to maximize the marginal log-likelihood $\log p(\mathbf{x})$ of given data $\mathbf{x}$. It is equivalent to maximize the Evidence Lower Bound(ELBO) which is lower bound of marginal log-likelihood[5]:

$$\max \log p(\mathbf{x}) \geq \max ELBO \tag{5}$$

$$ELBO = \mathbb{E}_{q_\phi(\mathbf{h}|\mathbf{x})}[\log p_\psi(\mathbf{x}|\mathbf{h})] - KL[q_\phi(\mathbf{h}|\mathbf{x}) \parallel p(\mathbf{h})] \tag{6}$$

where $p(\mathbf{h})$ is a standard Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. $KL[q_\phi(\mathbf{h}|\mathbf{x}) \parallel p(\mathbf{h})]$ can be analytically computed to reduce the lower variance of the gradients.

In NVDM, each dimension of latent topic distribution is corresponding to a topic. As it is mentioned in Eq.(3), the covariance matrix of latent normal distribution is set to a diagonal one which indicates that each topic is independent to others. In practice, the diagonal covariance matrix of latent topic distribution aims to achieve computational efficiency. However, here it leads to the failure of topic correlation modelling. On the other hand, in VAEs framework, the model tries to approximate the true posterior distribution with the latent distribution which makes the diagonal covariance matrix only explainable when it coincides with the diagonal covariance matrix of true posterior distribution. To reform this drawback, a full covariance matrix is needed.

3 CORRELATED HOUSEHOLDER TOPIC MODEL

In this section, we first review the Householder flow, an efficient approach to transform isotropic Gaussian into a full covariance one. The details of CHTM is then discussed.

3.1 Householder flow

Previous work [17] introduces a unitary transformation called Household flow to capture the correlation between each dimension in latent distribution. In NVDM, the each dimension of latent distribution also is corresponding to a topic, which is reasonable to model the topic correlation with Householder flow. Therefore,

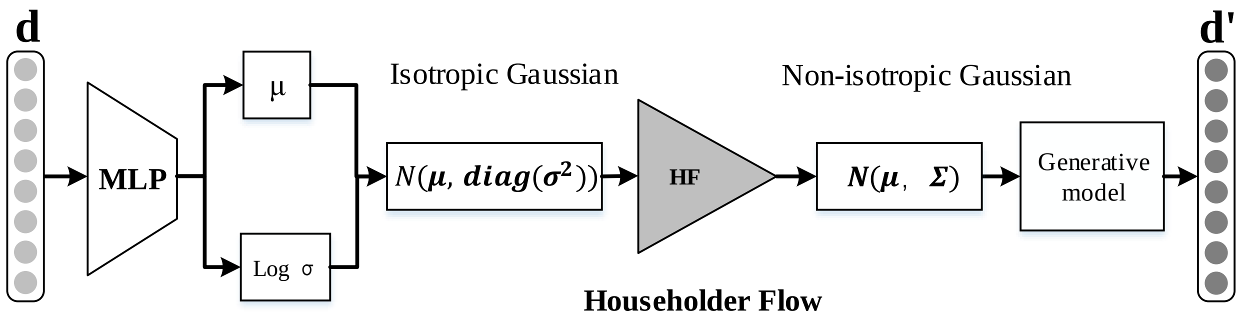

Figure 1: Schematic representation of CHTM. The triangle indicates the procedure of Householder flow.

inspired by their work, we introduce Household flow to incorporate the topic correlation in CHTM.

To get a full covariance matrix, a orthogonal matrix is needed to transform the diagonal matrix into a full covariance one. Generally, any full covariance matrix can be factorized into two orthogonal matrices:

$$\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathbf{T}} \tag{7}$$

where $\mathbf{D}$ is an eigenvalue matrix, $\mathbf{U}$ is an orthogonal matrix. In addition, any orthogonal matrix can be represented in following form[2, 16, 17]:

Theorem 3.1. *For any $M \times M$ matrix U, there exists a full-rank $M \times K$ matrix Y(basis) and a non-singular $K \times K$ matrix S, $K \leq M$, such that:*

$$\mathbf{U} = \mathbf{I} - \mathbf{Y}\mathbf{S}\mathbf{Y}^{\mathbf{T}} \tag{8}$$

The degree of the orthogonal matrix is given by K. Additionally, according to the [2, 16, 17], any orthogonal matrix with degree K can be expressed with the product of Householder transformations:

Theorem 3.2. *Any orthogonal matrix with basis acting on the K-dimensional subspace can be expressed as a product of exactly K Householder transformations:*

$$\mathbf{U} = \mathbf{H_K}\mathbf{H_{K-1}}\cdots\mathbf{H_1} \tag{9}$$

where $\mathbf{H_k} = \mathbf{I} - \mathrm{S_{kk}}\mathrm{Y}_{\cdot\mathbf{k}}(\mathrm{Y}_{\cdot\mathbf{k}})^{\mathbf{T}}$*, for* $k = 1, \cdots K$.

Theorem 3.2 indicates that we can model any orthogonal matrix via a series of K Householder transformations. The initial Householder matrix $\mathbf{H_1}$ is given by Eq.(10) with input of the inference network. For other $t \geq 2$, $\mathbf{H_t}$ indicates the corresponding Householder matrix with input of $\mathbf{h^{(t-1)}}$. $\mathbf{h^{(0)}}$ is the random variable sampled from original latent distribution. $\mathbf{h^{(k)}}$ is the final transformed random variable after a Householder flow at length of k.

$$\begin{aligned}\mathbf{h^{(t)}} &= \mathbf{H_t(x)}\mathbf{h^{(t-1)}} \\ &= (\mathbf{I} - 2\frac{\mathbf{v_t(x)} \cdot (\mathbf{v_t(x)})^{\mathbf{T}}}{\parallel \mathbf{v_t(x)} \parallel^2}) \cdot \mathbf{h^{(t-1)}}\end{aligned} \tag{10}$$

Mathematically, Householder matrix is an unitary, Hermitian and involutory matrix. With these properties, it can facilitate the derivation of objective function when Householder flow is applied. Moreover, the Householder flow only requires an invertible linear transformation which will facilitate the computation in training.

3.2 Incorporating Correlated Topic Modelling via Householder flow

To establish topic correlation modelling, Householder flow is applied to transform the original isotropic Gaussian topic distribution into a full covariance one. Notably, our CHTM is the first approach incorporating topic correlation modelling in neural variational topic models.

Specifically, in CHTM, the MLP inference network takes bag-of-words vector $\mathbf{d}$ as the input. Then the initial topic distribution is parameterized by two linear vector. The mean vector μ and variance vector σ is given by Eq.(3). Householder flow is then applied to transform isotropic distribution sample $h^{(0)}$ into a full covariance Gaussian sample $h^{(t)}$ by Eq.(10). Finally, the multinomial softmax layer reconstruct document $\mathbf{d'}$ according to the covariance Gaussian sample $h^{(t)}$ by Eq.(4). The structure of CHTM is display in Fig.1.

4 INFERENCE IN CHTM

In this section, we focus on detailed discussion of estimating objective function of the proposed CHTM.

Specifically, the objective function of CHTM is given by (11).

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{h^{(t)}}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{h^{(t)}})] - KL\left[q_\phi(\mathbf{h^{(t)}}|\mathbf{x})\|p(\mathbf{h})\right] \tag{11}$$

where $p(\mathbf{h})$ is a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. However, the topic distribution refined by Householder flow involve a implicitly non diagonal matrix, which is unable to carry out mean-field variational inference used in NVDM. Previously, [14, 17] offers an efficient approach for estimating (11), namely Flow-based Free Energy Bound. The Flow-based Free Energy Bound denotes in Eq.(12). Finally, the parameters of generative model and inference network can be updated by awake-sleep algorithm .

$$L_{FFEB} = E_q[\log p(\mathbf{d}|\mathbf{h^{(k)}}) + \sum_{t=1}^{k} \log|det\frac{\partial f^{(t)}}{\partial h^{(t-1)}}|] - KL[q(\mathbf{h^{(0)}}|\mathbf{d})\|p(\mathbf{h^{(k)}})] \tag{12}$$

From Eq.(12), Flow-based Free Energy Bound has a term involving the inverse transformation of flow function. For proposed CHTM, it can be simplified due to the unitary property of Householder matrix: $|\mathbf{H}| = 1$ so that $\log|\frac{\partial f^{(t)}}{\partial h^{(t-1)}}| = log|\mathbf{H}| = 0$. This property suggests that we actually don't have to get explicit parameters of final topic distribution while the neural variational inference can work properly. The objective function of CHTM can then be written as Eq.(13).

$$L_{FFEB} = E_q[\log p(\mathbf{d}|\mathbf{h^{(k)}})] + 0.5[n - \mu^2 + |\Sigma| + \log|\Sigma|] \quad (13)$$

5 RELATED WORKS

CTM [3] is a typical topic model incorporating topic correlation modelling. The topic mixture distribution in CTM is a logistic normal distribution with full covariance matrix. The following works such as Gaussian Process Topic Models [1] also relies on log normal distribution with full covariance matrix to model the topic correlations.

In neural variational topic models, NVDM [10] is a typical VAEs-based topic model. It implements a multilayer perceptron inference network for inference and a multinomial softmax generative model. The multinomial softmax can be regarded as the multinomial distribution under the restriction of softmax simplex[10]. Similarly, Gaussian Softmax Model [9] normalized the topic vector and generative process of NVDM. It replaces the multinomial softmax generative with a normalized layer to enhance the interpretability of NVDM. The structure of inference network remains unchanged. To reach computation efficiency, the topic distributions of NVDM and GSM are isotropic Gaussian distributions.

To transform isotropic Gaussian distributions into a full covariance one, normalizing flow[14] is the typical way of solution. Normalizing flow implements several invertible functions to map the original distribution variable into a full covariance one. What's more, Flow-based Free Energy Bound[14] is proposed to simplify the deviation of evidence lower bound. Householder flow is another solution. Householder flow involves a series Householder transformation and matrix multiplication, which can be efficiently computed.

6 EXPERIMENTS

6.1 Dataset and Setup

To evaluate our efforts, we choose *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 test articles. Reuters RCV1-v2 is a large dataset which consists of Reuters newswire stories with 794,414 training and 10,000 test cases. For data preprocessing, we follow the samilar procedure and setup in [10]. The vocabulary sizes of experiments conducted on these dataset are 2,000 and 10,000.

To make direct comparison with the prior works, we choose following methods as baselines:

(1) Latent Dirichlet Allocation: The widely used topic model in community. Here, we utilize the online variational inference implement of LDA in Gensim Gensim[13].
(2) Correlated Topic Model: CTM replaces the component independent Dirichlet distribution with a log normal Gaussian distribution to capture the correlation between topics. The author implemented version of CTM [3] is choosed.
(3) Neural Variational Document Model: An typical neural variational topic model. It implements a MLP network as inference network and a multinomial softmax generative model to model the document construction process.
(4) Gaussian Softmax Model: GSM normalized the topic vector of NVDM and implement a normalized multinomial softmax generative model to represent the topic word distribution.
(5) CHTM: The proposed method.

Table 1: Perplexity on corresponding dataset. The number of topic is 50.

Model	20News	RCV1-v2
LDA	1066.0309	1134.0293
CTM	944.1920	1074.4023
NVDM	832.4007	635.2824
GSM	849.0213	717.8721
CHTM	**780.7094**	**541.4351**

For LDA and CTM, the grid search is applied to find the optimal hyper parameters. For NVDM [4], GSM and CHTM, the inference network consists of a MLP with 2 layers and 500 hidden units. To fairly compare with NVDM, CHTM uses the same inference network options as NVDM. Accordingly, all baseline methods and CHTM are trained with 50 topics. For CHTM, the length of Householder flow is 1. During the training of NVDM, GSM and CHTM, we take one sample from latent distributions to compute the document vector and estimate the lower bound of document perplexity. Adam [6] and hold-out validation are applied during training. Like what is usually done in training VAEs, we alternately optimize the generative model and inference network by fixing parameters of one while updating the parameters of another.

6.2 Result

Many evaluation metrics have been applied to measure the quality of topic modelling. The typical metric in evaluating topic modelling is *perplexity*. Perplexity, in language modelling, always refers to the inverse of geometric mean per-word likelihood. The lower perplexity on test data usually indicates the better generalization performance. The perplexity in topic modelling is given by Eq.(14).

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

For CTM, the per word perplexity low bound mentioned in [3] is selected for evaluation. For neural variational methods, due to the fact that $\log p(\mathbf{X}_d)$ is intractable, we follow [11] using the variational evidence low bound (which is the upper bound of the perplexity) for evaluation.

The result is demonstrated at Table 1. Generally, the neural variational topic models have better performance than those of traditional models. Specifically, the proposed CHTM outperform other baseline methods on the both datasets. GSM and NVDM have similar performance on 20News. On RCV1-v2, the NVDM get better performance than that of GSM. On 20News dataset, CHTM

[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

achieves approximate 7% lower perplexity than NVDM. On RCV1-v2, the proposed CHTM also gets lower perplexity than that of NVDM. The generative models of NVDM and CHTM are same multinomial softmax. It indicates that the correlated topic modelling by incorporating Householder flow can improve the performance of neural variational topic models.

7 CONCLUSION

In this paper, we present CHTM: a neural variational topic model which can model topic correlation. Notably, CHTM is first approach to introduce topic correlation modelling in neural variation topic models. In CHTM, the topic correlations is established by Householder flow. The refined topic distributions can contribute the topic modelling by significantly reducing the perplexity of collections. The result of experiments shows that, compared with NVDM and GSM, CHTM can remarkably improve the performance of topic modelling. Moreover, in CHTM, the topic correlation modelling with Householder flow only involves few modifications in inference network, which indicates that the models whose latent distributions are isotropic Gaussian are suitable for Householder flow.

8 ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China (Grant No.2016QY03D0602) and the 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] Christian H Bischof and Xiaobai Sun. 1994. On orthogonal block elimination. *Preprint MCS-P450-0794, Mathematics and Computer Science Division, Argonne National Laboratory* (1994).

[3] David M. Blei and John D. Lafferty. 2005. Correlated Topic Models. In *Advances in NIPS 18 [Neural Information Processing Systems, NIPS,05, December 5-8, 2005, Vancouver, British Columbia, Canada]*. 147–154. http://papers.nips.cc/paper/2906-correlated-topic-models

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent Dirichlet Allocation. In *Advances in NIPS 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS,01, 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

[5] Carl Doersch. 2016. Tutorial on Variational Autoencoders. *CoRR* abs/1606.05908 (2016). http://arxiv.org/abs/1606.05908

[6] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). http://arxiv.org/abs/1412.6980

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

[8] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. 2017. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research* 18 (2017), 14:1–14:45. http://jmlr.org/papers/v18/papers/v18/16-107.html

[9] 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

[10] 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

[11] Andriy Mnih and Karol Gregor. 2014. Neural Variational Inference and Learning in Belief Networks. 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, 1791–1799. http://jmlr.org/proceedings/papers/v32/mnih14.html

[12] Rajesh Ranganath, Sean Gerrish, and David M. Blei. 2014. Black Box Variational Inference. *CoRR* abs/1401.0118 (2014). arXiv:1401.0118 http://arxiv.org/abs/1401.0118

[13] 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.

[14] 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

[15] 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

[16] 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.

[17] 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