Topic Modeling For NIPS Words

CAI Bibi
Department of Mathematics
HKUST
bcaiaa@connect.ust.hk

QIU Zhenyu
Department of Mathematics
HKUST
zqiuai@connect.ust.hk

WANG Zhiwei
Department of Mathematics
HKUST
zhiwei.wang@connect.ust.hk

Abstract

Scientific problems: Can we figure out the main topics of the papers in the NIPS words dataset by using machine learning methods? More precisely, can we know the topic proportion of each paper and give an explicit representation of each topic? Based on the above analysis, can we know the relationship between each topic or the relationship between each paper by using some clustering and dimension reduction methods? Moreover, can we see how these topics have changed in popularity over time?

In this article, we perform statistical topic modeling to analyze the NIPS words dataset. Specifically, we use the latent Dirichlet allocation (LDA), which introduces the latent variable *topic* to model the generative process of each word in the document. In the LDA model, each topic is characterized by a distribution of words. With the word distribution provided by LDA, we can easily figure out the topics based on the weights of each word inner each topic. In the context of text modeling, LDA can output an explicit and meaningful representation of a document, which can be viewed as topic proportion. We further use clustering and dimension reduction methods, such as K-means, MDS, and tSNE, to analyze this latent representation to gain more insights into this dataset.

1 Introduction

The Conference and Workshop on Neural Information Processing Systems (NIPS) is a machine learning and computational neuroscience conference. We chose the dataset NIPS words, which collects the distribution of words in the full text of the NIPS conference papers published from 1987 to 2015, to do some statistical analysis. The dataset is in the form of a 11463×5812 matrix of word counts, containing 11463 words and 5811 NIPS conference papers (the first column contains the list of words). Each column contains the number of times each word appears in the corresponding document. The names of the columns give information about each document and its timestamp in the following format: **Xyear-paperID**.

Specifically, we use the latent Dirichlet allocation (LDA), which introduces the latent variable *topic* to help model the generative process of each word in the document. In the LDA model, each topic is characterized by a distribution of words. Under the probabilistic framework, the goal is to maximize the marginal likelihood, which is intractable in this case. Therefore, LDA exploits the variational EM algorithm, which is an approximation-based approach to performing approximate Bayesian inference, and enjoys a fast computation speed.

LDA can provide the word distribution as the representation of the latent topics, and the topic proportion as the representation of each paper. With the topic representation and topic proportion, we can figure out the meaning of each topic and analyze the papers' topics. In this article, we further use clustering and dimension reduction methods, such as K-means, MDS, and tSNE, to analyze the output of the LDA to gain more insights into this dataset.

Submitted to CSIC5011 Topological and Geometric Data Reduction and Visualization, 2023 Spring, HKUST.

2 Latent Dirichlet Allocation

In this part, we will briefly introduce some notations and recall the model of Latent Dirichlet Allocation (LDA) [1].

Formally, we define the following terms:

- A *word* is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \ldots, V\}$. We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the vth word in the vocabulary is represented by a V-vector w such that $w^v = 1$ and $w^u = 0$ for $u \neq v$.
- A *document* is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, where w_n is the nth word in the sequence.
- A *corpus* is a collection of M documents denoted by $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

LDA is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus D:

1. Choose $N \sim Possion(\xi)$.
2. Choose $\theta \sim Dir(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim$ Multinomial(θ).
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality k of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed to be known and fixed. Second, the word probabilities are parameterized by a $k \times V$ matrix β where $\beta_{ij} = p(w^j = 1|z^i = 1)$, which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that N is independent of all the other data generating variables (θ and z). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.

A k-dimensional Dirichlet random variable θ can take values in the $(k-1)$-simplex (a k-vector θ lies in the $(k-1)$-simplex if $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$), and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \tag{1}$$

where the parameter α is a k-vector with components $\alpha_i > 0$, where $\Gamma(x)$ is the Gamma function.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is given by

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta), \tag{2}$$

where $p(z_n|\theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. Integrating over θ ans summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \Big(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \Big) \, d\theta. \tag{3}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$P(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\Big(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\Big)\, d\theta_d. \tag{4}$$

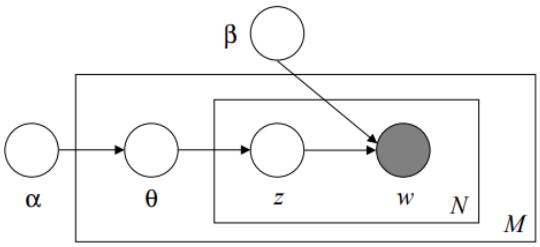

Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

The LDA model is represented as a probabilistic graphical model in Figure 1. As the figure makes clear, there are three levels to the LDA representation. The parameters α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables $z_{dn}$and w_{dn} are word-level variables and are sampled once for each word in each document.

3 Fitting the LDA Model

In order to use the LDA model, we need to estimate parameters by optimizing the marginal likelihood $P(D|\alpha,\beta)$. With the estimated parameters $\{\alpha,\beta\}$, we can do topic inference by computing the posterior distribution of the hidden variables given a document

$$p(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta) = \frac{p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)}{P(\mathbf{w}|\alpha,\beta)}. \tag{5}$$

However, this is intractable since the marginal likelihood $P(D|\alpha,\beta) = \prod_{d=1}^{M} P(\mathbf{w}_d|\alpha,\beta)$ cannot be computed by marginalizing all latent variables. To tackle this computation problem, variational inference (VI) [2, 3] is exploited here.

VI is an approximation-based approach to performing approximate Bayesian inference. Compared to the alternative method, Markov Chain Monte Carlo (MCMC) [4], which is a sampling-based approach, the advantage of VI is its scalability to large datasets. To apply variational approximation, we first define $q(\theta,\mathbf{z} \mid \gamma,\phi) = q(\theta \mid \gamma)\prod_{n=1}^{N} q\left(z_n \mid \phi_n\right)$ as an approximated distribution of posterior $p(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta)$, using mean-field factorization[2,3]. $\{\gamma,\phi\}$ are variational parameters, where γ is Dirichlet parameter and $\{\phi_1,\ldots,\phi_N\}$ are multinomial parameters. Then we obtain the evidence

lower bound (ELBO) of the logarithm of the marginal likelihood by Jensen's inequality

$$\begin{aligned}
\log p(\mathbf{w} \mid \alpha, \beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) d\theta \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \\
&\geq \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) d\theta - \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log q(\theta, \mathbf{z}) d\theta \\
&= \mathrm{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] - \mathrm{E}_q[\log q(\theta, \mathbf{z})] \\
&\equiv \mathrm{ELBO}\left(q; \alpha, \beta\right),
\end{aligned} \tag{6}$$

where the equality holds if and only if $q(\theta, \mathbf{z})$ is the exact posterior $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$. Instead of maximizing the logarithm of the marginal likelihood, we can iteratively maximize the ELBO with respect to the variational approximate posterior q and the model parameters $\{\alpha, \beta\}$

$$\left(\widehat{q}; \widehat{\alpha}, \widehat{\beta}\right) = \arg\max_{q;\alpha,\beta} \mathrm{ELBO}\left(q; \alpha, \beta\right). \tag{7}$$

Using the terminology in the EM algorithm, the maximization of ELBO with respect to q is known as the E-step, and the maximization of ELBO with respect to $\{\alpha, \beta\}$ is known as the M-step.

In the E-step, we aim to find the optimal solution for the approximate posterior with the current estimate for model parameters $\{\alpha, \beta\}$. Since q is determined by the variational parameters $\{\gamma, \phi\}$, then the optimization problem becomes

$$\left(\widehat{\gamma}, \widehat{\phi}\right) = \arg\min_{(\gamma,\phi)} \mathrm{D}(q(\theta, \mathbf{z} \mid \gamma, \phi) \| p(\theta, \mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})), \tag{8}$$

where D denotes the Kullback-Leibler (KL) divergence. By computing the derivatives of the KL divergence and setting them equal to zero, we obtain the following update equations

$$\begin{aligned}
\phi_{ni} &\propto \beta_{iw_n} \exp\left\{\mathrm{E}_q\left[\log\left(\theta_i\right) \mid \gamma\right]\right\} \\
\gamma_i &= \alpha_i + \sum_{n=1}^{N} \phi_{ni}.
\end{aligned} \tag{9}$$

With the variational approximation q updated in the E-step, we can update model parameters $\{\alpha, \beta\}$ in the M-step

$$\left(\widehat{\alpha}, \widehat{\beta}\right) = \arg\max_{\alpha,\beta} \mathrm{ELBO}\left(q; \alpha, \beta\right). \tag{10}$$

By setting the derivative of the above equation with respect to β equal to zero, we can get the updated equations

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}. \tag{11}$$

And α can be updated using an efficient Newton-Raphson method in which the Hessian is inverted in linear time.

4 Results

4.1 Latent Topics and Papers Clustering

At first, we want to use the LDA to find the latent topics in the dataset. After preprocessing the words, including stemming words and removing a standard list of stop words, we want to find the Dirichlet and conditional multinomial parameters for a 10-topic LDA model. The output (Figure 2) is a plot of topics, each represented as a bar plot using the top words based on weights in the specific topic.

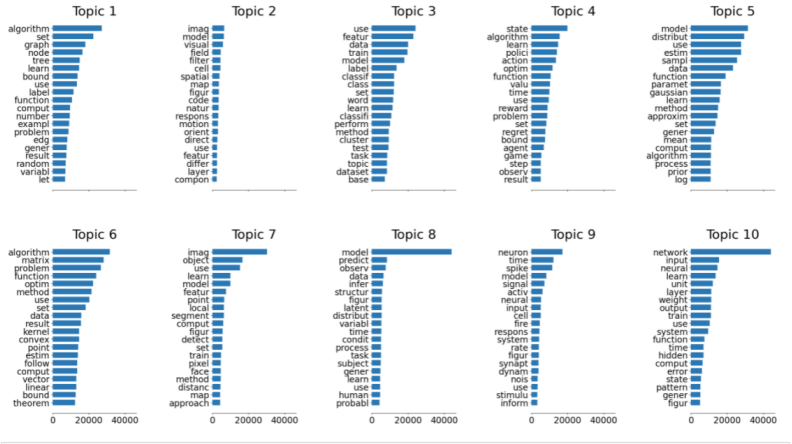

Figure 2: Top words in ten topics identified by the LDA model.

Based on the top words in each topic, we summarized and named the ten topics as follows:

- topic 1: graph theory
- topic 2: biology application
- topic 3: classification & clustering
- topic 4: reinforcement learning
- topic 5: probabilistic model & statistical machine learning
- topic 6: matrix computation & convex optimization
- topic 7: computer vision & image segmentation
- topic 8: probabilistic model & prediction model
- topic 9: neuron science
- topic 10: neural network

LDA model can also give the probability matrix of articles belonging to each topic. Based on the matrix, we used K-means method to cluster the articles. We use "calinski_harabaz_score", "silhouette_score" and "inertia_score" to figure out the best value of **k**, and the output is 10, which is the exact value of the number of topics in LDA. As we can see in Figure 3, each category grouped by K-means mainly represents a specific topic explored by LDA. Also, each category shows some relationships among topics in LDA. For example, topic 10 mainly talks about the LDA topic 3 and focuses on classification and clustering, which is a general methodology in Machine Learning. Then, the topic 10 also includes other related topics, like LDA topic 5: statistical machine learning and LDA topic 6: convex optimization, but excludes some topics with different methodologies, like LDA topic 3: reinforcement learning.

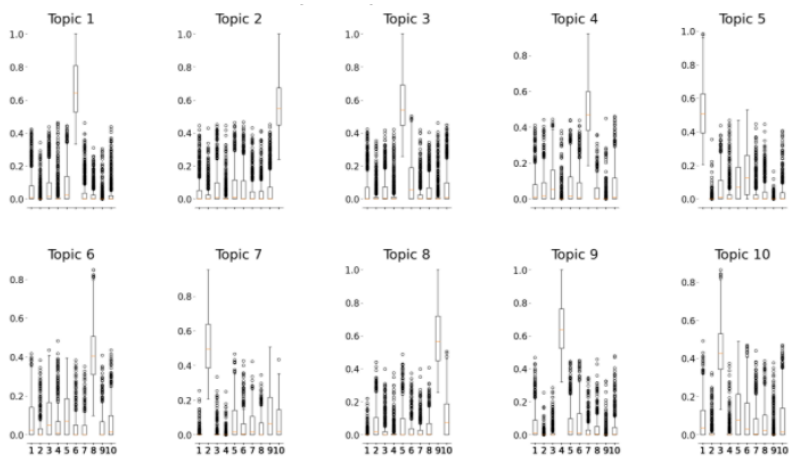

Figure 3: Papers' topics in K-means

K-means topics:

- topic 1 → LDA topic 6: matrix computation & convex optimization
- topic 2 → LDA topic 10: neural network
- topic 3 → LDA topic 5: probabilistic model & statistical machine learning
- topic 4 → LDA topic 7: computer vision & image segmentation
- topic 5 → LDA topic 1: graph theory
- topic 6 → LDA topic 8: probabilistic model & prediction model
- topic 7 → LDA topic 2: biology application
- topic 8 → LDA topic 9: neuron science
- topic 9 → LDA topic 4: reinforcement learning
- topic 10 → LDA topic 3: classification and clustering

4.2 Topics in Trend

In this subsection, we want to use **heatmap** to analyze the paper topics in trend. Based on the results of K-means, we sum up the occurrence of topics each year and perform cross-sectional normalization, then we can use **heatmap** to visualize the relationship of topics and year, see Figure 4.

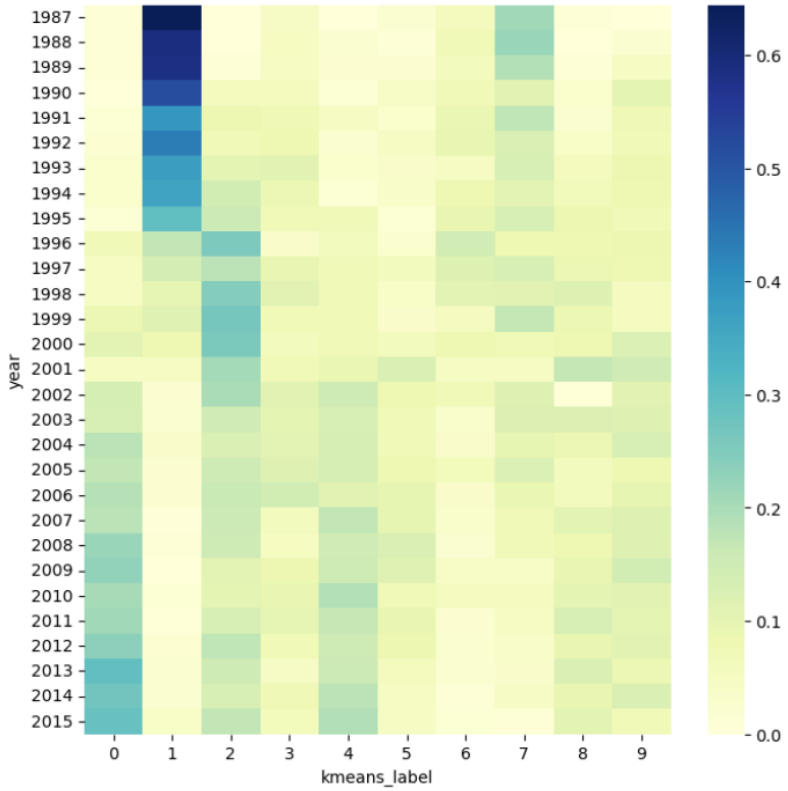

Figure 4: Heatmap for papers' trends

K-means topics

- topic 1 $\rightarrow$ LDA topic 6: matrix computation & convex optimization
- topic 2 $\rightarrow$ LDA topic 10: neural network
- topic 3 $\rightarrow$ LDA topic 5: probabilistic model & statistical machine learning
- topic 4 $\rightarrow$ LDA topic 7: computer vision & image segmentation
- topic 5 $\rightarrow$ LDA topic 1: graph theory
- topic 6 $\rightarrow$ LDA topic 8: probabilistic model & prediction model
- topic 7 $\rightarrow$ LDA topic 2: biology application
- topic 8 $\rightarrow$ LDA topic 9: neuron science
- topic 9 $\rightarrow$ LDA topic 4: reinforcement learning
- topic 10 $\rightarrow$ LDA topic 3: classification and clustering

As we can see in Figure 4, the topic "neural network" was extremely hot from 1987 to 1995, since more than 40% papers in NIPS were talking about it at that time. Then, from 1996 to 2002, the "probabilistic model & statistical machine learning" became the most popular topic and still remained active till 2015. The topics "matrix computation & convex optimization" and "graph theory" have become more and more popular in recent years. In addition, the topics of "reinforcement learning" and "classification and clustering" keep moderately active in these years.

4.3 Visualization with MDS and tSNE

In this subsection, we use MDS and tSNE to reduce the dimensions of these ten LDA topics to the major two dimensions. Then, we visualize papers in the 2-dimensional subspace and the K-means labels of the papers.

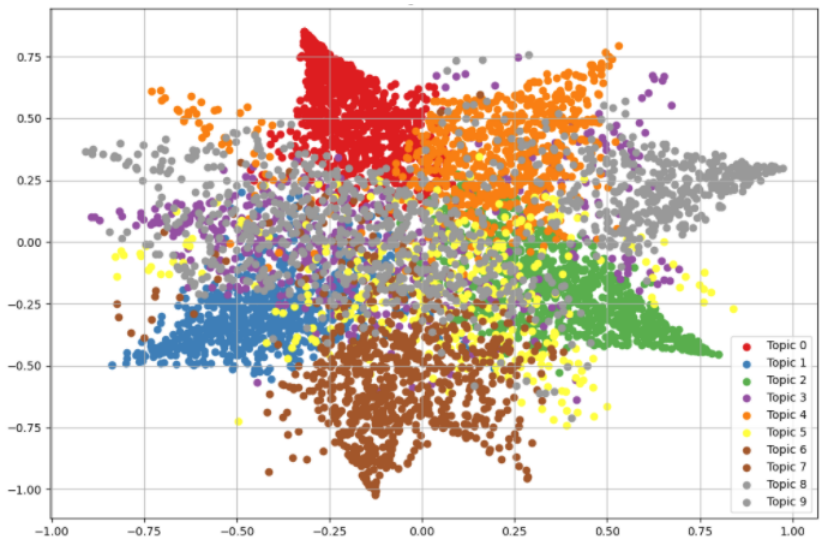

Figure 5: Visualization using reduced data with MDS

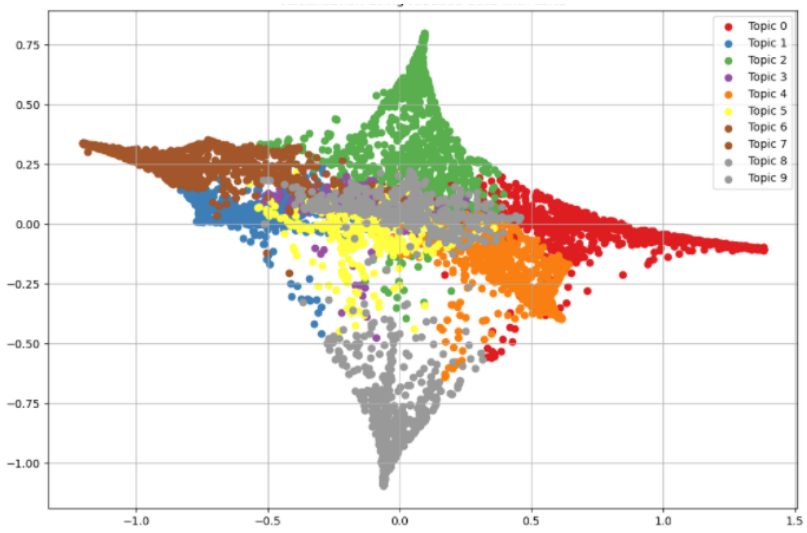

Figure 6: Visualization using reduced data with MDS

From the visualization with MDS dimension reduction, it seems that we can visualize the K-means result in 2 dimensions. But there are still overlap points, especially for Topic 6("probabilistic model & prediction model") and Topic 10("classification and clustering"). Maybe these topics are highly correlated with other topics. At the same time, other topics like topic 10("reinforcement learning") and topic 2("neural network") are clustered and located in specific concentrated areas respectively. The visualization with manifold dimension reduction methods, like tSNE, is better, since the topics may only have some locally similarities, which can be captured by tSNE. From the results, with tSNE dimension reduction, the papers in different topics are well clustered and located separately. Also, the topic 10("classification and clustering") is located at the center of these topics in the plot. Actually, the topic can be connected to all other topics and be the center in the research area of NIPS in our understanding of Machine Learning.

Contribution

All the group members discussed and shared ideas throughout the whole project. WANG Zhiwei raised the scientific problems, proposed the model and method used, and designed the research plan. QIU Zhenyu wrote the code, did the experiments, compared different methods, and visualized and analyzed the results. CAI Bibi (Abstract, Introduction, Model, Results), QIU Zhenyu (Results), and WANG Zhiwei (Abstract, Introduction, Method) wrote the report together.

References

[1] Blei, David M, Ng, Andrew Y & Jordan, Michael I (2003) Latent Dirichlet allocation. *Journal of machine Learning research* **3**: 993–1022.

[2] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[3] Blei, D. M., A. Kucukelbir & J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112 (518)**, 859–877.

[4] Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods.* Department of Computer Science, University of Toronto, Toronto, ON, Canada.