The Theoretical Breakthrough of Self-Supervised Learning: Variational Autoencoders and Its Application in Big Data Analysis

Yanqiu Ji¹, Zhiyi Lu¹*

¹School of Science, Tianjin University of Commerce, Tianjin, 300134, China
²Corresponding author's e-mail: lxylzy@tjcu.edu.cn

Abstract. In the study of big data statistics, a large number of unknown latent variables bring great difficulties to modeling. Variational autoencoders (VAE) can overcome the shortcomings of traditional variational methods such as low efficiency and poor generality, and provide an efficient and extensible framework for variational posterior inference and approximate maximum likelihood learning based on gradient. On the basis of reviewing the development history of variational autoencoders and taking the deep latent variable model (DLVM) as an example, this paper introduces the basic principle of variational autoencoders and analyzes its application under the background of big data. The problems in theory and application of variational autoencoders are presented, as well as the topics to be further studied. The combination of variational autoencoders and other statistical modeling methods may become a new idea for big data research.

1. Introduction

In order to realize the intelligent recognition and analysis of image, audio, video and other data, researchers need to annotate the acquired data. At present, the data annotation is mainly completed by human. However, in the context of big data, this method relying on human to complete data annotation is costly and inevitably introduces the subjective prior bias of the annotator. At the same time, when modeling big data, there are often many latent variables that cannot be observed. Latent variables are often difficult to express and calculate, which brings great inconvenience to modeling and optimization. As a self-supervised learning algorithm, variational autoencoders (VAE) method based on deep latent variable model (DLVM) is a good way to overcome the above problems.

Autoencoder (AE) is a self-supervised neural network model proposed by Rumelhart et al. in the 1980s for learning the deep representation of input data [1]. It is a combination of two coupled but independent parameterized models [2]: encoder (or recognition model) and decoder (or generation model). The encoder provides the posterior distribution of its latent random variables to the decoder, which receives the information provided by the encoder and generates the inverse of the encoder according to Bayesian rules, from which it learns meaningful data representations, including possible category labels. The decoder uses the input data as a supervised signal to learn, and this kind of algorithm is called self-supervised learning.

Once proposed, autoencoders have become popular algorithms in the field of machine learning. However, although autoencoders are simple and effective, they can only reconstruct the original input data and cannot generate new samples. In 2014, Kingma et al. proposed a new deep generation model, namely variational autoencoders, which is mainly used for data generation [3]. The main idea of VAE
is still to want to learn the latent variable representation of the input data, but differs from AE in that it adds a condition at the coding stage: it wants the latent variables $z$ to conform to the preset prior distribution $p(z)$. So, if you can learn the original data distribution $p(x|z)$, based on the joint probability distribution $p(x, z) = p(x|z)p(z)$ sampling, it provides an effective method for the establishment of continuous deep latent variable model and the directed probability model which is more complicated to calculate the parameters of posterior distribution.

VAE method has been one of the hot popular methods in the field of big data and deep learning since it was proposed. In recent years, there have also emerged many theoretical and applied innovations. In 2017, Shang et al. proposed a cyclic channel variational autoencoder [4]. Tolstikhin et al. proposed Wasserstein autoencoder; In 2018, Huang et al. proposed an introspective variational autoencoder [5]. In 2019, Zheng et al. proposed Fisher information autoencoder [6]; In 2020, Vincent et al. proposed deep probabilistic variational autoencoder [7]. In addition, in the context of big data, VAE method has been widely used in medicine, astronomy, natural language processing, image synthesis and reconstruction (see Chapter 3 of this article).

VAE method has changed the traditional perception of unsupervised learning and developed a self-supervised learning method. It is also a useful supplement to the application of variational inference (VI) theory. In recent years, with the continuous deepening of the application of big data and deep learning research, VAE method improvement and innovation continue to emerge, the application range is also more and more extensive. On the basis of reviewing the development process of VAE method, this paper introduces the basic principle of VAE method applied to deep latent variable model, and elaborates some new theoretical breakthroughs and innovations of VAE method under the background of big data, and introduces its application in different fields. Then the problems in theory and application of variational autoencoders and the topics for further research are proposed. The combination of the variational autoencoders method and other statistical modeling methods can become a new idea for big data research.

2. Variational autoencoders based on deep latent variable model

2.1. Deep latent variable model and its approximation

Suppose the data set is $\mathcal{D}$, the sample data is $x$, the data set is composed of $N \geq 1$ data points, that is, $\mathcal{D} = \{x_1, x_2, ..., x_N\} = \{x_i\}_{i=1}^N$, the latent variables in the model is $z$, the model parameter is $\theta$, the model of the deep latent variables can be expressed as $p_\theta(x, z)$, which can be decomposed into $p_\theta(x, z) = p_\theta(z)p_\theta(x|z)$, and the parameters in DLVM are optimized by the neural network.

There are two problems that are difficult to solve in the calculation of DLVM. One is that the calculation of the marginal distribution $p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz$ is very complicated, so the true posterior $p_\theta(z|x) = p_\theta(x|z)p_\theta(z)/p_\theta(x)$ of the latent variables is also difficult to calculate. Second, in the context of big data, the calculation cost is too high for large data sets. VAE is an effective way to solve the above problems, VAE first constrains the prior $p_\theta(z)$ of the potential variables $z$ have some kind of distribution $p(z)$, such as the normal distribution, and then introduces an encoder $q_\phi(z|x)$, $\phi$ as the variational parameter, as an approximation of the true posterior $p_\theta(z|x)$ with higher complexity in calculation, that is:

$$q_\phi(z|x) = p_\theta(z|x)$$

The selection of encoders is flexible and more computable than a real posterior. In this way, $z$ can be sampled from $q_\phi(z|x)$, and then the sampled $z$ can be sent into the decoder network $p_\theta(x|z)$ to calculate $p_\theta(x|z)$. It is relatively easy to calculate the approximate distribution of the deep latent variable model $p_\theta(x, z)$ by multiplying it with the preset prior distribution $p(z)$, and new data close to the real data can be generated from the sampling.
2.2. Evidence lower bound

Like other variational methods, the Kullback-Leibler (KL) divergence is used to measure the degree of approximation between the encoder \( q_\phi(z|x) \) and the true posterior distribution \( p_\theta(z|x) \), and the optimization goal is the evidence lower bound (ELBO). For any encoder \( q_\phi(z|x) \) and variational parameters \( \phi \), the following equation holds:

\[
\log p_\theta(x) = \mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x, z) / p_\theta(z|x)] = \mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x, z) / q_\phi(z|x)] + \mathbb{E}_{z \sim q_\phi(z|x)} [\log q_\phi(z|x) / p_\theta(z|x)]
\]

(1)

Where, \( \mathbb{E}_{z \sim q_\phi(z|x)} [\cdot] \) represents the expectation about \( q_\phi(z|x) \). The second item in the last row of the equation (1) is the KL divergence between \( q_\phi(z|x) \) and \( p_\theta(z|x) \), which is non-negative, that is \( D_{KL}(q_\phi(z|x) || p_\theta(z|x)) \geq 0 \). The smaller the value of KL divergence, the closer \( q_\phi(z|x) \) is to the true posterior distribution \( p_\theta(z|x) \).

The first item in the last line of Equation (1) is the variational lower bound, also known as the evidence lower bound, which is expressed as:

\[
\mathcal{L}_{\theta, \phi}(x) = \mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x, z) - \log q_\phi(z|x)]
\]

(2)

Because of the non-negative nature of KL divergence, ELBO is the lower bound of the logarithmic likelihood of the data:

\[
\mathcal{L}_{\theta, \phi}(x) = \log p_\theta(x) - D_{KL}(q_\phi(z|x) || p_\theta(z|x)) \geq \log p_\theta(x)
\]

(3)

It can be seen from Equation (3) that KL divergence not only defines the "distance" between the approximate posterior and the true posterior, but also gives the "distance" between the ELBO and the marginal likelihood, so it is also called the tightness of the boundary. It can also be seen that maximizing the function ELBO about \( \theta \) and \( \phi \) minimizes the KL divergence between the encoder \( q_\phi(z|x) \) and the true posterior \( p_\theta(z|x) \), and the approximation of \( q_\phi(z|x) \) gets better while maximizing the marginal likelihood function \( p_\theta(x) \), which means that the data generated by the decoder is closer to the real data from a likelihood point of view.

2.3. The gradient of ELBO

ELBO allows for joint optimization where all parameters \( \theta \) and \( \phi \) can be optimized simultaneously using the stochastic gradient descent method (SGD). And, although the ELBO of a single data point and its gradient \( \nabla_{\theta, \phi} \mathcal{L}_{\theta, \phi}(x) \) is hard to calculate, however, we can find a good unbiased estimator \( \nabla_{\theta, \phi} \mathcal{L}_{\theta, \phi}(x) \), so it can still be optimized with minibatch SGD.

The objective function of ELBO optimization is the sum (or average) of ELBO of a single data point, that is:

\[
\mathcal{L}_{\theta, \phi}(\mathcal{D}) = \sum_{x \in \mathcal{D}} \mathcal{L}_{\theta, \phi}(x)
\]

(4)

The unbiased gradient of decoder parameter \( \theta \) in ELBO is:

\[
\nabla_{\theta} \mathcal{L}_{\theta, \phi}(x) = \nabla_{\theta} \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x, z) - \log q_\phi(z|x)]
\]

\[
= \mathbb{E}_{q_\phi(z|x)} [\nabla_{\theta} (\log p_\theta(x, z) - \log q_\phi(z|x))]
\]

\[
= \nabla_{\theta} (\log p_\theta(x, z) - \log q_\phi(z|x))
\]

(5)

The last term in equation (5) is the simple Monte Carlo estimator of the second term, and \( z \) in the last two terms is a random sample of the sampled autoencoder \( q_\phi(z|x) \).

In general, the gradient of the variational parameter in ELBO is:

\[
\nabla_{\phi} \mathcal{L}_{\theta, \phi}(x) = \nabla_{\phi} \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x, z) - \log q_\phi(z|x)]
\]

\[
= \mathbb{E}_{q_\phi(z|x)} [\nabla_{\phi} (\log p_\theta(x, z) - \log q_\phi(z|x))]
\]

(6)

This is often hard to calculate. However, in the case of continuous latent variables, a
reparameterization trick (see below) can be used to obtain an unbiased estimate of \( \nabla \theta, \phi \mathcal{L}_{\theta, \phi}(x) \), which can be used to optimize ELBO using SGD.

2.4. Reparameterization trick for continuous latent variables

The above VAE model has a serious flaw in its implementation. The latent variables \( z \) is sampled from the output of the autoencoder \( q_\theta(z|x) \). When both \( q_\theta(z|x) \) and \( p(z) \) are assumed to be normal distribution, the encoder outputs the mean \( \mu \) and variance \( \sigma^2 \) of the normal distribution, and the input of the decoder is sampled from the distribution \( \mathcal{N} (\mu, \sigma^2) \). Due to the existence of sampling operation, the gradient propagation is discontinuous, so it is impossible to train the VAE network through the gradient descent algorithm.

The reparameterization trick [4] provides a solution to make the gradient continuously differentiable by sampling the potential variables \( z \) in the form of \( z = \mu + \sigma \odot \epsilon \) (\( \odot \) is the product of elements), and \( \epsilon \) is sampled from the standard normal distribution \( \mathcal{N}(0, I) \), i.e. \( p(\epsilon) \sim \mathcal{N}(0, I) \), \( \mu \) and \( \sigma \) are generated by the encoder network. Where \( \partial z/\partial \mu \) and \( \partial z/\partial \sigma \) are continuously differentiable, thus connecting the gradient propagation.

The steps of the reparameterization trick are illustrated below with an example of the most common factorized Gaussian encoder.

Step 1: The distribution of \( q_\theta(z|x) \) is set as the joint distribution of independent univariate normal distribution, so the probability density function of \( q_\phi(z_i|x) \) is \( \mathcal{N}(z_i; \mu, \sigma_i^2) \). The mean \( \mu \) and variance \( \sigma^2 \) of the normal distribution are obtained by using the encoder neural network optimization, and the distribution of \( q_\theta(z|x) \) is obtained:

\[
(q_\phi(z|x) = \prod_i q_\phi(z_i|x) = \prod_i \mathcal{N}(z_i; \mu, \sigma_i^2)
\]

Step 2: Sample the noise \( \epsilon \) from \( \mathcal{N}(0, I) \) and construct the parameterized latent variable:

\[
\epsilon \sim \mathcal{N}(0, I), \quad z = \mu + \sigma \odot \epsilon
\]

Step 3: Calculate the Jacobian from \( \epsilon \) to \( z \) and its absolute value:

\[
\frac{\partial z}{\partial \epsilon} = \text{diag}(\sigma)
\]

Since the determinant of a diagonal (or more generally, triangular) matrix is the product of its diagonal terms, the logarithm of the absolute value \( d_\phi(x, \epsilon) \) of the Jacobian determinant is:

\[
\log d_\phi(x, \epsilon) = \log |\det(\partial z/\partial \epsilon)| = \sum_i \log \sigma_i
\]

Step 4: Calculate the posterior density of \( \log q_\phi(z|x) \) [1]:

\[
\log q_\phi(z|x) = \log p(\epsilon) - \log d_\phi(x, \epsilon) = \sum_i (\log(\epsilon_i; 0, 1) - \log \sigma_i)
\]

Where \( z = g(\epsilon, \phi, x) \).

Step 5: After reparameterization, the expectation about \( q_\phi(z|x) \) is replaced with the expectation about \( p(\epsilon) \), and ELBO can be re-expressed as:

\[
\mathcal{L}_{\theta, \phi}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x, z) - \log q_\phi(z|x)] = \mathbb{E}_{p(\epsilon)}[\log p_\theta(x, z) - \log q_\phi(z|x)]
\]

Step 6: Using a single noise sample \( \epsilon \) from \( p(\epsilon) \), a simple Monte Carlo estimator \( \hat{\mathcal{L}}_{\theta, \phi}(x) \) of ELBO for a single data point is formed:

\[
\hat{\mathcal{L}}_{\theta, \phi}(x) = \log p_\theta(x, z) - \log q_\phi(z|x)
\]

After reparameterization, it is easy to differentiate \( \theta \) and \( \phi \), and the resulting gradient \( \nabla_{\theta, \phi} \hat{\mathcal{L}}_{\theta, \phi}(x) \) can be used to optimize ELBO with minibatch SGD.

2.5. Optimize ELBO with minibatch SGD

The most commonly used parameter estimation method for probabilistic models is maximum logarithmic likelihood estimation. It can be seen from Equation (4) that this is equivalent to the minimization of KL divergence and the maximization of ELBO between data and model distribution. For large data sets, due to the linear relationship between the complexity of calculation and the size of
the data set, an effective optimization method is stochastic gradient descent method, which uses randomly selected data $N_M$ of size $M \subset D$ to carry out minibatch processing. Through such minibatch processing, the maximum likelihood estimate of the decoder parameter $\theta$ in ELBO can be obtained by the following equation:

$$\frac{1}{N_D} \log p_\theta(D) \approx \frac{1}{N_M} \log p_\theta(M) = \frac{1}{N_M} \sum_{x \in M} \log p_\theta(x)$$

(7)

Where $N_D$ represents the original sample size. In this case, the noise source on the right side of $\approx$ in equation (7) is generated when minibatch of data $M$ is randomly extracted. The unbiased estimator $\log p_\theta(M)$ is differentiable, resulting in an unbiased random gradient:

$$\frac{1}{N_D} \nabla_\theta \log p_\theta(D) \approx \frac{1}{N_M} \nabla_\theta \log p_\theta(M) = \frac{1}{N_M} \sum_{x \in M} \nabla_\theta \log p_\theta(x)$$

(8)

The reparameterization trick described above can be used to derive an estimate of the unbiased gradient of $\phi$ for the encoder parameter in ELBO similar to the formula above. The objective function can be optimized by repeatedly performing a small gradient iteration step in the direction of the random gradient.

The optimization of ELBO using SGD is a double stochastic optimization process, because the noise comes from the sampling of minibatch data and the sampling of $p(\varepsilon)$. In the context of deep latent variable model, this method is called Auto-Encoding Variational Bayes algorithm (AEVB) [3]. After optimizing the parameters $\theta$ of DLVM and the variational parameters $\phi$ of the encoder through AEVB, the new data which is very close to the real data can be generated by VAE method to achieve effective self-supervised learning. AEVB algorithm can effectively combine VAE with deep latent variable model, providing an effective, extensible, gradient-based variational posteriori inference and approximate maximum likelihood learning framework, and can be widely used in big data analysis.

3. The application of VAE in the context of big data

Under the background of big data, VAE has been widely used in many fields thanks to the convenience of statistical software and the significant increase in the speed of calculation. In particular, applying VAE method to approximate the posterior distribution of deep latent variable model can overcome the calculation bottleneck faced by DLVM in application, so it has a very broad application prospect. At present, VAE method has been successfully applied in medicine, natural language processing, astronomy, image processing and many other fields.

In medicine, Kadurin et al. used VAE as a molecular descriptor generator to generate new molecules with potential specific anticancer properties [7]. Subsequently, Blaschke et al. used VAE to generate new structures with predictive activity for dopamine receptor type II [8]. In 2019, Sean et al. expanded VAE as a chemical decoder and applied it to small molecule recognition, drug discovery and design, chemical forensics and other aspects [12]. In 2020, Raphael et al. used VAE to learn high resolution embedding of data sets, direct generation of three dimensional coordinates of immunoglobulin [9], and so on.

In terms of natural language processing, Tran and Nguyen proposed an antagonistic training algorithm combining VAE to learn the potential semantics of utterance pairs in 2018 [10]. Du proposed a variational autoregressive encoder to generate more coherent and diverse answers [11]. Yang et al. proposed a new language stegography method based on VAE in 2020, which greatly improved the imperceptibility of stegographic statements [12].

In the aspect of image synthesis, Shang et al. proposed the circulating channel variational autoencoder, which improved the defect of VAE generated images being vague and unclear [4]. Subsequently, Chen et al. combined VAE with generative adversation network in 2019 to optimize the problems such as mode collapse and training instability in face image generation [13]. In 2020, Wu et al. extended the condition VAE to a Cross version named cross-VAE to encode and separate the facial features [14].

In astronomy, in 2019, Iwasaki et al. realized an unsupervised learning method combining VAE and Gaussian mixture model and successfully applied deep neural network to astronomy [15]. In 2020,
Michael et al. proposed a convolutional VAE model to visually check the malfunctions of astronomical telescopes [16].

4. Conclusions and discussions
Since Kingma and Welling proposed the variational autoencoders in 2014, the method of the variational autoencoder has gradually matured, and many variant methods have been developed, and its application scope has become wider and wider. The VAE overcomes the disadvantages of the traditional VI method which is inefficient and not suitable for the model which is composed of neural network, and provides an effective and extensible framework for gradient based variational posterior inference and approximate maximum likelihood learning. The appeal of variational autoencoders is not only that they can process existing data, but also that they can be combined with other models or algorithms to generate new data. VAE change the traditional cognition of approximate posterior distribution and are a useful supplement to classical modeling theory.

At the end of this paper, we propose some topics that need further study in theory and application of variational autoencoders.

(1) The improvement of the measurement of the approximate degree between distributions and the optimization objective function: At present, among the various models and algorithms based on VI, the index to measure the approximation degree between two distributions is usually KL divergence, and the optimization objective function is usually ELBO. Recently, some scholars have proposed some new methods, such as the important weighted autoencoders, which improves the calculation method of ELBO and gets a tighter lower bound. It is a research direction that needs to be innovated to explore a better measurement index and optimization objective of the degree of similarity between distributions.

(2) Improvement of network structure and optimization methods: At present, the network structure used in VAE is relatively simple, and the reparameterization trick for calculating the ELBO gradient is usually completed under the normal assumption. At present, with the vigorous development of deep learning methods, it is an important direction for future VAE research to introduce more complex and flexible network structure into VAE to achieve better learning results. At the same time, if more distribution assumptions and uncertainty of distribution are introduced into the model, and then existing methods are improved and innovated in the framework of Bayesian nonparametric statistics, it is also a research field with great potential value in both theoretical and practical applications.

(3) Explore the statistical characteristics of VAE: Although variational inference has achieved significant application effects in many fields, as Blei et al. said, compared with the simulation method of distribution approximation, namely MCMC (Markov Chain Monte Carlo) method, the variational method is not rigorous in theory and lacks in-depth research on its statistical characteristics [17]. So it is still a black box algorithm. It is a promising research direction to regard variational inference as the estimation process of parameters, and to explore the characteristics of VAE from a statistical point of view, so as to obtain more explanatory inference results.

Acknowledgments
This paper was supported by the National Social Science Foundation of China: "Research on the Actuarial Statistical Model of Catastrophe Insurance and Its Application (16ZDA052)".

References
[1] Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). Learning representations by back-propagating errors. J. nature, 323(6088): 533-536.
[2] Kingma, D. P., Welling, M. (2019). An introduction to variational autoencoders. J. Foundations and Trends in Machine Learning, 12(4): 307-392.
[3] Kingma D P, Welling M. (2014). Auto-Encoding Variational Bayes. In: International Conference on Learning Representations. Banff.
[4] Shang, W., Sohn, K., Tian, Y. (2018). Channel-recurrent autoencoding for image modeling. In: 2018 IEEE Winter Conference on Applications of Computer Vision. Nevada. pp. 1195-1204.
[5] Huang H., He R., Sun Z., Tan, T. (2018). Introvae: Introspective variational autoencoders for photographic image synthesis. In: Advances in neural information processing systems. Salt Lake City. pp. 52-63.

[6] Zheng, H., Yao, J., Zhang, Y., Tsang, I. W., Wang, J. (2019). Understanding vae in fisher-shannon plane. In: Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii. pp. 33(01):5917-5924.

[7] Kadurin, A., Nikolenko, S., Khрабrov, K., Aliper, A., Zhavoronkov, A. (2017). druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. J. Molecular pharmaceutics, 14(9): 3098-3104.

[8] Blaschke, T., Olivcrona, M., Engkvist, O., Bajorath, J., Chen, H. (2018). Application of generative autoencoder in de novo molecular design. J. Molecular informatics, 37(1-2): 1700123.

[9] Colby, S. M., Nuñez, J. R., Hodas, N. O., Corley, C. D., Renslow, R. R. (2019). Deep learning to generate in silico chemical property libraries and candidate molecules for small molecule identification in complex samples. J. Analytical chemistry, 92(2): 1720-1729.

[10] Tran, V. K., Nguyen, L. M. (2018). Adversarial domain adaptation for variational neural language generation in dialogue systems. In: International Conference on Computational Linguistics. La Rochelle. pp. 13.

[11] Du, J., Li, W., He, Y., Xu, R., Bing, L., Wang, X. (2018). Variational autoregressive decoder for neural response generation. In: Conference on Empirical Methods in Natural Language Processing. Brussels. pp. 3154-3163.

[12] Yang, Z. L., Zhang, S. Y., Hu, Y. T., Hu, Z. W., Huang, Y. F. (2020). VAE-Stega: Linguistic steganography based on variational auto-encoder. J. IEEE Transactions on Information Forensics and Security, 16: 880-895.

[13] Chen, H. Y., Lu, C. J. (2019). Nested variance estimating vae/gan for face generation. In: International Joint Conference on Neural Networks. Budapest. pp. 1-8.

[14] Wu, H., Jia, J., Xie, L., Qi, G., Shi, Y., Tian, Q. (2020). Cross-VAE: Towards Disentangling Expression from Identity For Human Faces. In: International Conference on Acoustics, Speech and Signal Processing. Toronto. pp. 4087-4091.

[15] Iwasaki, H., Ichinohe, Y., Uchiyama, Y., Yamaguchi, H. (2019). A New Implementation of Deep Neural Network for Spatio-Spectral Analysis in X-Ray Astronomy. Astronomical Data Analysis Software and Systems XXVII. J. Astronomical Society of the Pacific, 2019, 523: 79

[16] Mesarcik, M., Boonstra, A. J., Meijer, C., Jansen, W., Rangelova, E., van Nieuwpoort, R. V. (2020). Deep learning assisted data inspection for radio astronomy. J. Monthly Notices of the Royal Astronomical Society, 496(2): 1517-1529.

[17] Blei, D. M., Kucukelbir, A., McAuliffe, J. D. (2017). Variational inference: A review for statisticians. J. Journal of the American statistical Association, 112(518): 859-877.