Neural Mixed Counting Models for Dispersed Topic Discovery

Jiemin Wu[1,*], **Yanghui Rao**[1,†], **Zusheng Zhang**[1],
Haoran Xie[2], **Qing Li**[3], **Fu Lee Wang**[4], **Ziye Chen**[1]
[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
[2]Department of Computing and Decision Sciences, Lingnan University, Hong Kong
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[4]School of Science and Technology, The Open University of Hong Kong, Hong Kong
wujm29@mail2.sysu.edu.cn, raoyangh@mail.sysu.edu.cn,
hrxie2@gmail.com, csqli@comp.polyu.edu.hk,
pwang@ouhk.edu.hk, {zhangzsh3, chenzy35}@mail2.sysu.edu.cn

Abstract

Mixed counting models that use the negative binomial distribution as the prior can well model over-dispersed and hierarchically dependent random variables; thus they have attracted much attention in mining dispersed document topics. However, the existing parameter inference method like Monte Carlo sampling is quite time-consuming. In this paper, we propose two efficient neural mixed counting models, i.e., the Negative Binomial-Neural Topic Model (NB-NTM) and the Gamma Negative Binomial-Neural Topic Model (GNB-NTM) for dispersed topic discovery. Neural variational inference algorithms are developed to infer model parameters by using the reparameterization of Gamma distribution and the Gaussian approximation of Poisson distribution. Experiments on real-world datasets indicate that our models outperform state-of-the-art baseline models in terms of perplexity and topic coherence. The results also validate that both NB-NTM and GNB-NTM can produce explainable intermediate variables by generating dispersed proportions of document topics.

1 Introduction

Mixture modeling is an essential topic in statistics and machine learning areas, owing to generating the random probability measure of data samples belonging to multiple clusters. In unsupervised learning tasks such as topic discovery, mixture modeling has gained increasing attention from researchers (Wang et al., 2011; Zhou and Carin, 2012, 2015; Zhou, 2018; Zhao et al., 2019). Specifically, mixture modeling over document words devotes to assign these words to different topics via random probability measures. Hierarchical Dirichlet Process (HDP) (Teh et al., 2004) is one of the representative methods in mixture modeling, which can characterize the two-level dependency of random probability measures. Although we can use Monte Carlo sampling or variational inference to estimate the parameters in HDP, it requires the help of indirect construction of random variables such as the Chinese Restaurant Franchise (Teh et al., 2004) or the Stick-Breaking construction (Wang et al., 2011) due to the lack of conjugation between the two-tier Dirichlet processes. This makes the inference of HDP mostly complicated (Zhou et al., 2016).

The mixed counting models represented by the Negative Binomial (NB) process (Titsias, 2007) and the Gamma Negative Binomial (GNB) process (Zhou and Carin, 2012) have solved this problem to a certain extent, in which, the normalized GNB process has been proven to be equivalent to HDP (Zhou and Carin, 2012). Because both NB and GNB processes satisfy the properties of completely random measures (Charles Kingman, 1967), the generative process of random probability measures among various mixed components is independent and becomes straightforward. Moreover, they naturally introduce non-negative constraints and have been proven as able to model over-dispersed data. In the case of mining latent topics of documents, the over-dispersed property indicates that the variance is larger than the mean for document-topic distributions. When compared to the NB process, the GNB process has an extra feature of describing more flexible stochastic phenomena with hierarchical dependencies. Despite the above advantages, with the increase of data size and observable information, the aforementioned parameter inference method like Monte Carlo sampling or variational inference has gradually become an important factor limiting the usage scenarios of mixed counting models (Miao et al., 2016). The reason is that Monte Carlo sampling has a high computational

[*] The first two authors contributed equally to this work which was finished when Jiemin Wu was an undergraduate student of his final year.
[†] The corresponding author.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6159–6169

cost, and variational inference becomes intractable when applied to models with complex variable dependencies (Acharya et al., 2015).

Neural variational inference (NVI) is a flexible and fast parameter inference framework based on neural networks (Mnih and Gregor, 2014). It can be regarded as a generalization of variational autoencoder applicable to natural language processing tasks. Based on NVI, several neural topic models had been proposed and achieved encouraging performance in document modeling (Miao et al., 2016; Srivastava and Sutton, 2017; Miao et al., 2017). These models used the neural network to learn the distribution relationship between input documents and latent topics due to its excellent function fitting ability and scalability. Particularly, the neural network parameters can be trained by back-propagation through the reparameterization of a continuous distribution (Naesseth et al., 2017) or using variance reduction techniques for a discrete distribution (Mnih and Gregor, 2014). However, the hidden variables in the above neural topic models lack good interpretability, and it is also impossible to model over-dispersed and hierarchically dependent document sets for these methods.

In this paper, we propose two novel neural mixed counting models dubbed the Negative Binomial-Neural Topic Model (NB-NTM) and the Gamma Negative Binomial-Neural Topic Model (GNB-NTM) based on NB and GNB processes, respectively. The general motivation is to combine the advantages of NVI and mixed counting models. On the one hand, NVI-based models are fast and easy to estimate but hard to interpret. On the other hand, document modeling via mixed counting models is easy to interpret but difficult to infer. In our NB-NTM and GNB-NTM, we develop NVI algorithms to infer parameters by using the reparameterization of Gamma distribution and the Gaussian approximation of Poisson distribution. Extensive experiments on real-world datasets validate the effectiveness of our proposed models in perplexity, topic coherence, and dispersed topic learning. Furthermore, the proposed models can describe the hierarchical dependence of random probability measures and introduce non-negative constraints, which renders the intermediate variables generated by our methods to have good interpretability.

The remainder of this article is organized as follows. In Section 2, we summarize the related studies on topic discovery. In Section 3, we introduce the definitions and properties of background methods. The proposed models are described in Section 4, the experimental evaluations are shown in Section 5, and we draw the conclusions in Section 6.

2 Related Work

Topic discovery aims to use the statistical information of word occurrences to obtain the abstract semantic structure embedded in a document set. From Bayesian methods represented by latent semantic analysis (LSA) (Deerwester et al., 1990), probabilistic latent semantic analysis (PLSA) (Hofmann, 1999), latent Dirichlet allocation (LDA) (Blei et al., 2003), and Hierarchical Dirichlet Process (HDP) (Teh et al., 2004), topic discovery had been widely researched in natural language processing and applied to many scenarios. For instance, the above models were extended to capture topic relevance (Blei and Lafferty, 2005) and topic evolution over time (Wang and McCallum, 2006; Blei and Lafferty, 2006). Algorithms for short text (Yan et al., 2013), tagged data (Ramage et al., 2009), and stream data (Yao et al., 2009) were also proposed. Considering the importance of prior distributions in LDA-based models, some research efforts tried to use beta and Gaussian distributions instead of the Dirichlet distribution as the prior of probabilistic graphical models (Thibaux and Jordan, 2007; Das et al., 2015). Although the Bayesian method is a natural way to represent the latent structure of a document set in topic discovery, as the structure of such a model becomes deeper and more complex, pure Bayesian inference becomes intractable due to the high dimensional integrals required (Miao et al., 2016). To address this issue, Cheng and Liu (2014) proposed a parallel Monte Carlo sampling method for HDP based on multi-threading. Unfortunately, it needs to traverse every word of all topics (i.e., threads) in the whole corpus when updating the topic-word distribution, rendering a large time cost for thread communication.

With the development of deep learning, especially the introduction of NVI, there is a new direction to discover topics based on neural networks. For example, Miao et al. (2016) assumed that word distributions in each document could be represented by hidden variables sampled from multiple Gaussian distributions, and they used the variational lower bound as the objective function of their model named NVDM. Srivastava and Sutton (2017) employed the logical Gaussian distribution to approxi-

mate the Dirichlet distribution, which improved the variational auto-encoder and LDA simultaneously. Miao et al. (2017) proposed a method named GSM to model the document-topic distribution explicitly. In their study, the topic-word distribution was introduced into the decoder. Besides the above NVI-based methods, Nalisnick and Smyth (2017) developed a stick-breaking variational auto-encoder for image generation. Nan et al. (2019) proposed a model named W-LDA in the Wasserstein auto-encoder framework. They employed the Maximum Mean Discrepancy (MMD) in W-LDA to match the proposed distribution and the prior distribution. However, the accuracy of MMD relied heavily on the number of samples for each distribution, and the kernel function in MMD had a significant influence on the performance. By leveraging word embeddings, Gupta et al. (2019) proposed a neural autoregressive topic model dubbed iDocNADE to enrich the context of short text. Experiments indicate that iDocNADE outperformed state-of-the-art generative topic models.

The recent relevant work to ours is the method proposed in (Zhao et al., 2019), which regarded the NB distribution as the prior in modeling the over-dispersed discrete data. However, the parameters of this method were still derived from the latent variables that obey the Gaussian distribution. Thus, these latent variables do not satisfy the non-negative constraint and lack good interpretability. Furthermore, the above method did not model topics explicitly, making it hard to generate document-topic and topic-word distributions.

3 Background

3.1 Negative Binomial Process

Let $X \sim \mathrm{NBP}(G_0, p)$ denote a NB process defined on the product space $\mathbb{R}_+ \times \Omega$, where G_0 is a finite continuous basic measure on a completely separable measure space Ω, and p is a scale parameter. For each Borel set $A \subset \Omega$, we use $X(A)$ to denote a count random variable describing the number of observations that reside within A. Then, $X(A)$ obeys the NB distribution $\mathrm{NB}(G_0(A), p)$. Given the k^{th} component π_k and its weight r_k on Ω, if G_0 is expressed as $G_0 = \sum_{k=1}^{\infty} r_k \delta_{\pi_k}$, where δ is the Dirac delta function, then $X \sim \mathrm{NBP}(G_0, p)$ can be expressed by $X = \sum_{k=1}^{\infty} n_k \delta_{\pi_k}$, where $n_k \sim \mathrm{NB}(r_k, p)$.

The NB distribution $m \sim \mathrm{NB}(r, p)$ has a probability density function $f_M(m) = \frac{\Gamma(r+m)}{m!\Gamma(r)}(1-p)^r p^m$, where $\Gamma(\cdot)$ denotes the gamma function. For the above probability density function, the mean and the variance are $\mu = r/(1-p)$ and $\sigma^2 = rp/(1-p)^2 = \mu + r^{-1}\mu^2$, respectively. Because the mean is smaller than the variance, i.e., the variance-to-mean ratio is greater than 1, NB distributions have shown great advantages in over-dispersed data modeling (Zhou and Carin, 2012). Moreover, since the NB distribution $m \sim \mathrm{NB}(r, p)$ can be extended to a Gamma distribution and a Poisson distribution, i.e., $m \sim \mathrm{Poisson}(\lambda)$ and $\lambda \sim \mathrm{Gamma}(r, p/(1-p))$, the NB process mentioned earlier can be extended to a Gamma-Poisson process (Zhou and Carin, 2015) as follows: $X \sim \mathrm{PP}(\Lambda)$, and $\Lambda \sim \mathrm{GaP}(G_0, (1-p)/p)$, where $\mathrm{PP}(\cdot)$ and $\mathrm{GaP}(\cdot)$ denote the Poisson process and the Gamma process, respectively. The random probability measure corresponding to each mixed component in the NB process can be directly sampled from the NB distribution without resorting to the Chinese Restaurant Franchise, the Stick-Breaking, or other construction methods, because each random measure is independent of the others, i.e., the NB process is completely random.

3.2 Gamma Negative Binomial Process

In the NB process, each Poisson process shares the same Gamma process prior with a fixed mean. Based on the NB process, the GNB process assigns another Gamma process as a prior to its mean, making it easier to model over-dispersed data (Zhou et al., 2016). Particularly, the generative process of random variables for the GNB process is as follows: $G \sim \mathrm{GaP}(G_0, \eta)$, $\Lambda_j \sim \mathrm{GaP}(G, (1-p_j)/p_j)$, and $X_j \sim \mathrm{PP}(\Lambda_j)$, where j is the subset index, η is the scale parameter of the first-level Gamma process, and the basic measure G_0 in the NB process is replaced by another random measure G. It has been shown that HDP is a normalized form of the GNB process in (Zhou and Carin, 2012). However, unlike HDP, the GNB process explicitly introduces the parameter p_j to control the dispersion degree of instantaneous measurement, making the latter model more flexible.

3.3 Neural Variational Inference

NVI is often used as an efficient parameter inference framework for complex and deep-seated structural models. Inspired by the variational auto-encoder, NVI assumes that the observed data $\boldsymbol{d}$ is subject to a certain probability distribution determined by a hidden variable $\boldsymbol{h}$. In contrast to

variational auto-encoders on handling the case of continuous latent variables (Kingma and Welling, 2014), NVI can deal with both discrete and continuous latent variables. Specifically, a neural network is used to infer the proposed distribution $q(\boldsymbol{h}|\boldsymbol{d})$. As stated in (Miao et al., 2017), Monte Carlo estimates of the gradient must be employed for models with discrete latent variables. In the case of $q(\boldsymbol{h}|\boldsymbol{d})$ being continuous, the hidden variable $\boldsymbol{h}$ is firstly obtained by sampling from $q(\boldsymbol{h}|\boldsymbol{d})$ through the corresponding reparameterization approach. Then, the likelihood $p(\boldsymbol{d}|\boldsymbol{h})$ is used to reproduce the observed data from hidden variables, and the objective is to minimize the Kullback-Leibler (KL) divergence of the proposed distribution and the actual posterior distribution. Finally, the variational lower bound is obtained by $\mathcal{L} = \mathbb{E}_{q(\boldsymbol{h}|\boldsymbol{d})} \log p\left(\boldsymbol{d}|\boldsymbol{h}\right) - D_{\mathrm{KL}}\left[q(\boldsymbol{h}|\boldsymbol{d})\|p(\boldsymbol{h})\right]$, where the first term is the expectation of the log-likelihood, and the second one is the KL divergence between the inferred distribution and a predefined prior. To sum up, NVI first uses a neural network to infer the proposed distribution $q(\boldsymbol{h}|\boldsymbol{d})$, and then maximizes the variational lower bound by back-propagation to fit the actual posterior distribution $p(\boldsymbol{h}|\boldsymbol{d})$. Such a framework learns the distribution of input data well, enabling it to combine with the traditional probability graphical models (e.g., LDA) and infer model parameters quickly (Srivastava and Sutton, 2017). However, how to effectively integrate the distributed dependencies in mixed counting models into the framework of variational inference is still quite a challenging problem.

4 Proposed Models

In this section, we respectively detail our NB-NTM and GNB-NTM for dispersed topic discovery.

4.1 Negative Binomial-Neural Topic Model

With a NB process prior, we propose the NB-NTM to model the counting of document words. Furthermore, a novel NVI framework is developed for parameter inference. Let $\boldsymbol{D} = \{\boldsymbol{d}^1, ..., \boldsymbol{d}^{|D|}\}$ be the input with $|D|$ documents and each document $\boldsymbol{d} \in \mathbb{R}^V$ be a bag-of-words representation, where V is the vocabulary size. Since it is impossible to draw all the countably infinite atoms of a Gamma process, we first employ the finite truncation strategy, in which, a number of topics K (i.e., the truncated level) is set manually (Nalisnick and Smyth, 2017; Zhou, 2018). Note that although K is fixed, if K is set to be large enough, not necessarily all topics would be used and hence a truncated model still preserves its nonparametric ability; whereas if K is set to be small, asymmetric priors on the topic weights are also maintained (Zhou, 2018). Then we can express the generative process of NB-NTM for document $\boldsymbol{d}$ as follows:

$$\boldsymbol{r} = f_1(\boldsymbol{d}), \quad \boldsymbol{p} = f_2(\boldsymbol{d}), \tag{1}$$

$$\boldsymbol{\lambda} \sim \mathrm{Gamma}\left(\boldsymbol{r}, \boldsymbol{p}/\left(\boldsymbol{1}-\boldsymbol{p}\right)\right), \tag{2}$$

$$\boldsymbol{n} \sim \mathrm{Poisson}\left(\boldsymbol{\lambda}\right), \tag{3}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are two multilayer perceptrons (MLPs) applying to generate the variational parameters $\boldsymbol{r}$ and $\boldsymbol{p}$. Specifically, $\boldsymbol{r}$ is the component weight of G, i.e., the topic measure at the corpus level, and $G = \sum_{k=1}^{K} r_k \delta_{\pi_k}$. $\boldsymbol{\lambda}$ represents the weights of topics at the document level, which can be used to estimate the topic measure on $\boldsymbol{d}$ by $\Lambda = \sum_{k=1}^{K} \lambda_k \delta_{\pi_k}$. In the above, λ_k denotes the k^{th} component of $\boldsymbol{\lambda}$. Finally, $\boldsymbol{n}$ is the component weight of Π that represents a Poisson process at the word level, and $\Pi = \sum_{k=1}^{K} n_k \delta_{\pi_k}$. The framework of NB-NTM is shown in Figure 1, and the parameter inference process is described as follows.

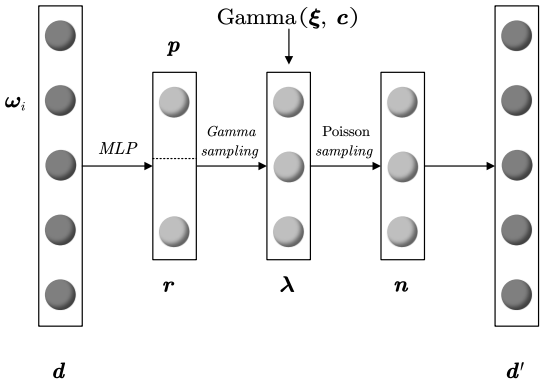


Figure 1: Framework of NB-NTM.

For the logarithmic likelihood of each document $\boldsymbol{d}$, we can derive the variational lower bound by $\mathcal{L} = -D_{KL}(q(\boldsymbol{\lambda}|\boldsymbol{d})||p(\boldsymbol{\lambda})) + \mathbb{E}_{q(\boldsymbol{\lambda}|\boldsymbol{d})}\left[\sum_{i=1}^{N_{\boldsymbol{d}}} \log p(\boldsymbol{\omega}_i|\boldsymbol{\lambda})\right]$. In the above, $q(\boldsymbol{\lambda}|\boldsymbol{d})$ is the encoder's inference of posterior probability, i.e., $\mathrm{Gamma}(\boldsymbol{r}, \boldsymbol{p})$, $\boldsymbol{\omega}_i \in \mathbb{R}^V$ is the one-hot representation of the word at the i^{th} position, $N_{\boldsymbol{d}}$ is the number of words in document $\boldsymbol{d}$, and $p(\boldsymbol{\lambda})$ is the Gamma prior for $\boldsymbol{\lambda}$, i.e., $\mathrm{Gamma}(\boldsymbol{\xi}, \boldsymbol{c})$. The KL divergence between $q(\boldsymbol{\lambda}|\boldsymbol{d})$ and $p(\boldsymbol{\lambda})$, i.e., $\mathrm{Gamma}(\boldsymbol{r}, \boldsymbol{p})$ and $\mathrm{Gamma}(\boldsymbol{\xi}, \boldsymbol{c})$, is calculated by following (Mathiassen et al., 2002): $D_{KL}(q(\boldsymbol{\lambda}|\boldsymbol{d})||p(\boldsymbol{\lambda})) = \sum_{k=1}^{K}[(r_k - 1)\Psi(r_k) -$

$\log p_k - r_k - \log\Gamma(r_k) - (\boldsymbol{\xi}-1)(\Psi(r_k)+\log p_k) + \log\Gamma(\boldsymbol{\xi}) + \boldsymbol{\xi}\log \boldsymbol{c} + \frac{r_k p_k}{\boldsymbol{c}}]$, where $\Psi(\cdot)$ is the Digamma function. The conditional probability over each word $p(\boldsymbol{\omega}_i|\boldsymbol{\lambda})$ is modeled by softmax function, as follows: $p(\boldsymbol{\omega}_i|\boldsymbol{\lambda}) = \frac{exp\{\sigma(\boldsymbol{n}^T R\boldsymbol{\omega}_i + \boldsymbol{b}_i)\}}{\sum_{j=1}^{V} exp\{\sigma(\boldsymbol{n}^T R\boldsymbol{\omega}_j + \boldsymbol{b}_j)\}}$, where R and $\boldsymbol{b}$ denote the weight matrix and the bias term, respectively. We present the parameter inference process of NB-NTM in Algorithm 1, in which, the variational lower bound $\mathcal{L}$ is used to calculate gradients and model parameters are updated by Adam (Kingma and Ba, 2015).

Algorithm 1: Parameter Inference for NB-NTM

Input: Number of topics K, gamma priors $\boldsymbol{\xi}$ and $\boldsymbol{c}$, document set $\boldsymbol{D}$;
Output: Document-topic distribution $\boldsymbol{\theta}$, topic-word distribution $\boldsymbol{\phi}$.

repeat
  for *document* $\boldsymbol{d} \in \boldsymbol{D}$ **do**
    Compute gamma distribution parameters $\boldsymbol{r} = f_1(\boldsymbol{d}), \boldsymbol{p} = f_2(\boldsymbol{d})$;
    Compute the KL divergence between $\text{Gamma}(\boldsymbol{r}, \boldsymbol{p})$ and $\text{Gamma}(\boldsymbol{\xi}, \boldsymbol{c})$;
    for $k \in [1, K]$ **do**
      Sample the Poisson distribution parameter by $\lambda_k \sim \text{Gamma}(r_k, p_k/(1-p_k))$;
      Sample word numbers by $n_k \sim \text{Poisson}(\lambda_k)$;
    end
    for $\boldsymbol{\omega}_i \in \boldsymbol{d}$ **do**
      Compute log-likelihood $\log p(\boldsymbol{\omega}_i|\boldsymbol{\lambda})$;
    end
    Compute variational lower bound $\mathcal{L}$;
    Update $f_1(\cdot)$, $f_2(\cdot)$, R, and $\boldsymbol{b}$;
  end
until *convergence*;
for *document* $\boldsymbol{d} \in \boldsymbol{D}$ **do**
  Normalize $\boldsymbol{\lambda}$ to obtain $\theta_{\boldsymbol{d}}$;
end
Apply softmax to R in row to obtain $\boldsymbol{\phi}$.

4.2 Gamma Negative Binomial-Neural Topic Model

Based on the NB-NTM, we further propose the GNB-NTM by assigning another Gamma process as a prior to the NB process. As shown in Figure 2, the generative process of GNB-NTM for document $\boldsymbol{d}$ is given below:

$$\boldsymbol{\gamma} = f_1(\boldsymbol{d}), \quad \boldsymbol{\eta} = f_2(\boldsymbol{d}), \tag{4}$$

$$\boldsymbol{r} \sim \text{Gamma}(\boldsymbol{\gamma}, \boldsymbol{\eta}), \tag{5}$$

$$\boldsymbol{\lambda} \sim \text{Gamma}(\boldsymbol{r}, p/(1-p)), \tag{6}$$

$$p = f_3(\boldsymbol{d}), \quad \boldsymbol{n} \sim \text{Poisson}(\boldsymbol{\lambda}). \tag{7}$$

In the above, $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are the parameters of the first-level Gamma process, and p is the scale parameter of the second-level Gamma process. The differences between GNB-NTM and NB-NTM are three-fold. Firstly, another Gamma process G_0 is introduced over the existing Gamma process G as a prior of its shape parameter, so as to characterize the multi-level dependencies of random variables. In particular, $G_0 = \sum_{k=1}^{K} \gamma_k \delta_{\pi_k}$. Secondly, a scale parameter p is introduced for each document to describe the dispersion degree of all words in the document. Thirdly, the GNB-NTM employs $\boldsymbol{n} + \boldsymbol{r}$ as the input of the decoder by following the production rule of the observed variable in (Zhou and Carin, 2012). Using $\boldsymbol{n} + \boldsymbol{r}$ as the input also helps to incorporate the global topic information into the decoder's inference of posterior probability $q(\boldsymbol{r}|\boldsymbol{d})$. Thus, the conditional probability over each word $p(\boldsymbol{\omega}_i|\boldsymbol{r})$ is modeled as follows: $p(\boldsymbol{\omega}_i|\boldsymbol{r}) = \frac{exp\{\sigma((\boldsymbol{n}+\boldsymbol{r})^T R\boldsymbol{\omega}_i + \boldsymbol{b}_i)\}}{\sum_{j=1}^{V} exp\{\sigma((\boldsymbol{n}+\boldsymbol{r})^T R\boldsymbol{\omega}_j + \boldsymbol{b}_j)\}}$.

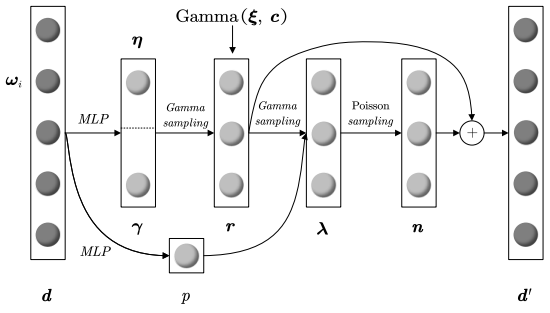


Figure 2: Framework of GNB-NTM, where both $\boldsymbol{r}$ and $\boldsymbol{n}$ are used as input in the decoder.

Similar to NB-NTM, the variational lower bound is derived by: $\mathcal{L} = \mathbb{E}_{q(\boldsymbol{r}|\boldsymbol{d})}\left[\sum_{i=1}^{N_{\boldsymbol{d}}} \log p(\boldsymbol{\omega}_i|\boldsymbol{r})\right] - D_{KL}(q(\boldsymbol{r}|\boldsymbol{d})||p(\boldsymbol{r}))$, where $p(\boldsymbol{r})$ is the Gamma prior for $\boldsymbol{r}$, i.e., $\text{Gamma}(\boldsymbol{\xi}, \boldsymbol{c})$. The parameter inference for GNB-NTM is presented in Algorithm 2. We use the variational lower bound to calculate gradients and apply Adam to update parameters of GNB-NTM, which are the same as NB-NTM.

4.3 Reparameterization Approach

The Gamma and Poisson sampling operation cannot be differentiated, making it intractable to up-

Algorithm 2: Parameter Inference for GNB-NTM

Input: Number of topics K, gamma priors $\boldsymbol{\xi}$ and $\boldsymbol{c}$, document set $\boldsymbol{D}$;
Output: Document-topic distribution $\boldsymbol{\theta}$, topic-word distribution $\boldsymbol{\phi}$.

repeat
  for *document* $\boldsymbol{d} \in \boldsymbol{D}$ **do**
    Compute the 1^{st} gamma distribution parameters $\boldsymbol{\gamma} = f_1(\boldsymbol{d}), \boldsymbol{\eta} = f_2(\boldsymbol{d})$;
    Compute the KL divergence between $\text{Gamma}(\boldsymbol{\gamma}, \boldsymbol{\eta})$ and $\text{Gamma}(\boldsymbol{\xi}, \boldsymbol{c})$;
    Compute the 2^{nd} gamma distribution parameter $p = f_3(\boldsymbol{d})$;
    for $k \in [1, K]$ **do**
      Sample the 2^{nd} gamma distribution parameter by $r_k \sim \text{Gamma}(\gamma_k, \eta_k)$;
      Sample the Poisson distribution parameter by $\lambda_k \sim \text{Gamma}(r_k, p/(1-p))$;
      Sample word numbers by $n_k \sim \text{Poisson}(\lambda_k)$;
    end
    for $\boldsymbol{\omega}_i \in \boldsymbol{d}$ **do**
      Compute log-likelihood $\log p(\boldsymbol{\omega}_i | \boldsymbol{r})$;
    end
    Compute variational lower bound $\mathcal{L}$;
    Update $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$, R, and $\boldsymbol{b}$;
  end
until *convergence*;
for *document* $\boldsymbol{d} \in \boldsymbol{D}$ **do**
  Normalize $\boldsymbol{\lambda}$ to obtain $\theta_{\boldsymbol{d}}$;
end
Apply softmax to R in row to obtain $\boldsymbol{\phi}$.

date model parameters through back-propagation. Here, we describe the reparameterization approach for smoothing gradients. For the Gamma distribution $x \sim \text{Gamma}(\alpha, \beta)$ with $\alpha > 1$, the reparameterization can be obtained by the reject-sampling method (Naesseth et al., 2017), i.e., $x = \frac{1}{\beta}\left(\alpha - \frac{1}{3}\right)\left(1 + \frac{\varepsilon}{\sqrt{9\alpha - 3}}\right)^3, \epsilon \sim \mathcal{N}(0, 1)$. Besides, the shape augmentation method (Naesseth et al., 2017) is applied to convert $\alpha \leq 1$ to $\alpha > 1$ to increase the accept rate of each rejection sampler. For the Poisson distribution which is discrete, we use the Gaussian distribution as an approximation (Rezende et al., 2014; Kingma and Welling, 2014). Based on the central limit theorem, $\mathcal{N}(\mu = \lambda, \sigma^2 = \lambda)$ can approximate $\text{Poisson}(\lambda)$. Thus, we sample from the Poisson distribution directly to avoid the issue of discretization and use the Gaussian distribution as an approximation when calculating the Poisson distribution's gradient. Particularly, the reparameterization of a Gaussian distribution $x \sim \mathcal{N}(\mu, \sigma^2)$ is $x = \mu + \epsilon \cdot \sigma, \epsilon \sim \mathcal{N}(0, 1)$.

5 Empirical Results

5.1 Datasets

We employ the following three datasets to evaluate the effectiveness of our models: Reuters[1], 20News, and MXM song lyrics (Miao et al., 2017). The Reuters dataset contains 7,758 training documents and 3,005 testing documents. The 20News corpus consists of 18,773 news articles under 20 categories. These news articles are divided into 11,268 training documents and 7,505 testing documents. The 20 categories include sports, electronics, automotive, and so forth, and the number of documents under each category is almost the same. MXM is the official lyrics collection of the Million Song Dataset, which contains 210,519 training documents and 27,143 testing documents, respectively. By following (Miao et al., 2017), we use the originally provided vocabulary with 5,000 words for MXM, while for Reuters and 20News, we use stemming, stop words filtering, and the 2,000 most frequently occurred words as vocabularies. The statistics of these datasets are presented in Table 1.

Dataset	Reuters	20News	MXM
Train.Size	7,758	11,268	210,519
Test.Size	3,005	7,505	27,143
Label number	90	20	-
Vocabulary size	2,000	2,000	5,000

Table 1: Statistics of the datasets.

5.2 Experimental Setup

The following models are adopted as baselines: HDP (Teh et al., 2004), NVDM (Miao et al., 2016), NVLDA and ProdLDA (Srivastava and Sutton, 2017), GSM (Miao et al., 2017), and iDocNADE (Gupta et al., 2019). Among these baselines, HDP is a classical mixture modeling method followed

[1] https://www.nltk.org/book/ch02.html

the equivalence with the normalized GNB process (Zhou and Carin, 2012). In HDP, the model parameters are estimated by Monte Carlo sampling. NVDM, NVLDA, ProdLDA, and GSM are all neural topic models based on NVI. Considering that word embeddings have shown to capture both the semantic and syntactic relatedness in words and demonstrated impressive performance in natural language processing tasks, we also present the result of a neural autoregressive topic model that leverages word embeddings (i.e., iDocNADE). Particularly, the publicly available codes of HDP[2], NVDM[3], NVLDA and ProdLDA[4], and iDocNADE[5] are directly used. As an extended model of NVDM, the baseline of GSM is implemented by us based on the code of NVDM. To ensure fair comparisons on various NVI-based methods, unless explicitly specified, we set the number of topics to 50, the hidden dimension of MLP to 256, and use one sample for NVI by following (Miao et al., 2017). For the batch size, the learning rate, and other model parameters, grid search is carried out on the training set to determine their optimal values and achieve the held-out performance.

To evaluate the quality of topics generated by different models, we use perplexity and topic coherence as evaluation criteria. The perplexity of each model on a testing set $\widetilde{\boldsymbol{D}}$ is: $\text{perplexity}\,(\widetilde{\boldsymbol{D}}) = \exp\left(-\frac{1}{|\widetilde{\boldsymbol{D}}|}\sum_{\widetilde{\boldsymbol{d}}}\frac{1}{N_{\widetilde{\boldsymbol{d}}}}\log p(\widetilde{\boldsymbol{d}})\right)$, where $\log p(\widetilde{\boldsymbol{d}})$ represents the log-likelihood of the model on document $\widetilde{\boldsymbol{d}}$, and $N_{\widetilde{\boldsymbol{d}}}$ is the number of words in $\widetilde{\boldsymbol{d}}$. The lower the perplexity is, the more likely for a model to generate $\widetilde{\boldsymbol{D}}$. Therefore, if a model obtains a lower perplexity than others in the testing set, it can be considered as the better one. For all NVI-based topic models, the variational lower bound, which is proven to be the upper bound of perplexity (Mnih and Gregor, 2014), is used to calculate the perplexity by following (Miao et al., 2016, 2017). When calculating the topic coherence, we use the normalised pointwise mutual information (NPMI) which measures the relationship between word w_i and other $T-1$ top words (Lau et al., 2014) as follows: $\text{NPMI}\,(w_i) = \sum_{j=1}^{T-1}[\log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}/-\log P\,(w_i,w_j)]$. The higher the value of topic coherence, the more explainable the topic is.

[2] https://github.com/soberqian/TopicModel4J

[3] https://github.com/ysmiao/nvdm

[4] https://github.com/akashgit/autoencoding_vi_for_topic_models

[5] https://github.com/pgcool/iDocNADEe

5.3 Performance Comparison

Table 2 shows the perplexity and topic coherence of different models on the test datasets. We can observe that NB-NTM outperforms most baselines, and GNB-NTM performs the best in all cases. The results validate that the NB distribution can model over-dispersed documents well. Furthermore, the latent semantics of these corpora may be hierarchically dependent. In other words, the topics at the corpus level and those of each document are not independent but correlated with one another.

Model	Perplexity			Topic coherence		
	Reuters	20News	MXM	Reuters	20News	MXM
HDP	302.3	730.7	319.5	0.305	0.223	0.356
NVDM	224.9	855.0	252.3	0.133	0.138	0.109
NVLDA	578.6	1252.2	668.7	0.253	0.240	0.216
prodLDA	648.1	1267.2	852.9	0.332	0.329	0.313
GSM	266.2	963.5	330.5	0.192	0.211	0.177
iDocNDAE	202.8	844.6	294.7	0.130	0.151	0.222
NB-NTM	181.0	740.6	247.5	0.341	0.343	0.340
GNB-NTM	**146.7**	**602.8**	**216.8**	**0.377**	**0.375**	**0.427**

Table 2: Perplexity and topic coherence results, where the latter is an average of three coherence scores by calculating 5, 10, and 15 top words for each topic.

In terms of the model efficiency, neural topic models can be trained much faster than HDP on a large corpus by GPU acceleration. Take the large-scaled MXM dataset as an example, the training time of both NB-NTM and GNB-NTM is around one hour using a *GeForce GTX 960* GPU, while HDP needs more than three hours to converge using an *AMD R5 3600* CPU. Under the same environment, the training time of all NVI-based topic models is close. In general, NVLDA, prodLDA, and NVDM run slightly faster than NB-NTM because the Gaussian reparameterization approach is simpler than the Gamma one. GSM and GNB-NTM are slightly slower than others because the former introduces more parameters to model the topic-word distribution, while the latter introduces more sampling operations.

As an illustration, we also qualitatively evaluate the semantic information learned by different models on the 20News training set. The baselines of HDP, NVLDA, and prodLDA, which achieve competitive topic coherence scores, are selected for comparison. Table 3 presents 5 of the most representative topics with the corresponding top 10 words, from which we can observe that although all these models can identify the chosen topics reasonably, our NB-NTM and GNB-NTM perform better than the other baselines in most cases.

Topic	HDP	NVLDA	prodLDA	NB-NTM	GNB-NTM
Religion	god•	heaven•	shall	belief•	athos•
	who	christ•	worship	religion•	beliefs•
	people	interpretation	christians•	athos•	moral•
	atheism•	scripture•	religious•	moral•	truth•
	believe•	christian•	belief•	scripture•	church•
	religion•	church•	bible•	jesus•	christ•
	does	truth•	atheists•	church•	christian•
	atheists•	lord•	heaven•	christian•	jesus•
	his	believe•	acts	christianity•	christianity•
	evidence	christianity•	religions•	god•	belief•
Encryption	key•	keys•	encryption•	rsa•	cryptography•
	unit	brad	court	agencies•	crypto•
	keyboard	chip	clipper	encrypted•	encrypted•
	keys•	crypto•	semi	cryptography•	security•
	cable	phone	escrow•	security•	keys•
	lock•	encryption•	encrypted•	scheme•	nsa•
	fit	cryptography•	drugs	government	secure•
	cross	agencies•	gun	escrow•	key•
	back	agency•	criminal	nsa•	government
	women	secure•	criminals	secure•	agencies•
Sport	game•	cup•	players•	player•	hockey•
	fighting•	toronto	team•	win•	sport•
	four	played•	teams•	play•	game•
	games•	patrick	winning•	boston	games•
	almost	wings	ice	detroit	baseball•
	level	players•	him	cup•	fans•
	police	rangers	nhl•	playoffs•	season•
	co	leafs	season•	players•	teams•
	kill	teams•	hockey•	season•	wings
	effective	baseball•	leafs	games•	leafs
Space	its	shuttle•	commercial	lunar•	mission•
	earth•	jpl•	cryptography	his	algorithm
	organizations	development	mission•	toronto	nasa•
	first	physics	image	orbit•	chip
	high	orbit•	lunar•	years•	orbit•
	mission•	rocket•	processing	mission•	development
	shell•	cost	established	dc	solar•
	their	energy	rocket•	year•	space•
	such	earth•	remote	national	technology
	program	space•	soviet	space•	satellite•
Hardware	card•	cable•	mouse•	floppy•	cache•
	bit	floppy•	floppy•	bus•	vga•
	mac•	dx	card•	ram•	display•
	mb	controller•	lib	printer•	printer•
	memory•	mb	simms	memory•	interface•
	mhz	pin•	button•	card•	pc•
	ram•	shipping	printer•	controller•	dx
	monitor•	brand	meg	motherboard•	processor•
	speed	drive	ram•	ide	motherboard•
	bus•	motherboard•	motherboard•	monitor•	ram•

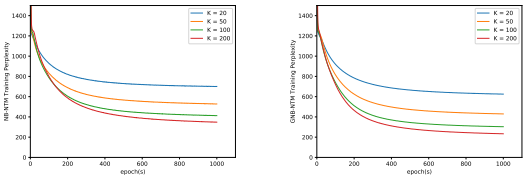
Table 3: Top 10 words of 5 topics learned by different models on 20News, where • means the word is related to the corresponding topic by checking manually.

5.4 Impact of the Number of Topics

In this part, we test the impact of the number of topics on the performance of our models. Figure 3 shows the convergence process of NB-NTM and GNB-NTM on the 20News training set with K = 20, 50, 100, 200 in terms of the perplexity. We can observe that as K increases, the perplexity values of both models decrease under each epoch. This is because the NVI framework is essentially an encoder-decoder, and the increase of the topic number enables the models to encode and reconstruct documents better. We also notice that with the continuous growth of K, the improvement of perplexity is getting lower. Table 4 presents the results of our models on the 20News testing set under the above conditions, in which a similar trend can be observed as aforementioned.

5.5 Evaluation on Learning Dispersed Topics

Compared to the existing neural topic models, another feature of our models is that the generated



(a) NB-NTM (b) GNB-NTM

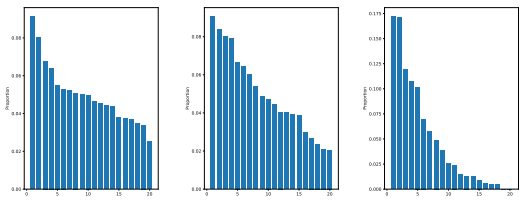
Figure 3: The convergence behavior of our models with different numbers of topics on the 20News training set.

Number of topics	Perplexity		Topic coherence	
	NB-NTM	GNB-NTM	NB-NTM	GNB-NTM
20	800.8	717.3	0.307	0.351
50	740.6	602.8	0.343	0.375
100	654.3	501.2	0.331	0.360
200	572.4	424.4	0.330	0.351

Table 4: Perplexity and topic coherence of our models on the 20News testing set with different topic numbers.

topics are dispersed, and thus, the intermediate variables can be more explainable. To validate the effectiveness of our models on learning dispersed topics, we first count the total number of words under each manually labeled category (i.e., topic) as the topic-word number distribution shown in Figure 4 (a). Then we run our NB-NTM and GNB-NTM on the entire 20News testing set to get the corresponding values of $\boldsymbol{r}$. After normalization, the proportion of different topics obtained by NB-NTM and GNB-NTM at the corpus level is presented in Figure 4 (b) and Figure 4 (c), respectively. For the convenience of the result presentation, we set the number of topics to 20 for both models. Note that the 20 topics do not need to correspond to the 20 categories, because we here focus on testing whether the topic proportions generated by our two models are in accordance with their model structures/characteristics. From these results, we can observe that the proportion of topics obtained by NB-NTM is close to the topic-word number distribution. On the other hand, GNB-NTM obtains more dispersed proportions of topics than NB-NTM. These results suggest that GNB-NTM tends to allocate less but more important topics to the corpus, i.e., the topics generated by GNB-NTM are more discriminative. Since the document-topic distribution is not directly modeled and the Gaussian distribution samples are not non-negative, the previous neural methods except GSM cannot obtain explainable intermediate variables. For the baseline of GSM, Miao et al. (2017) had demonstrated that the topics with higher probabilities were evenly

distributed on the same 20News dataset, which indicates that our models outperform GSM on learning dispersed document topics.



(a) (b) (c)

Figure 4: Qualitative analysis on the model interpretability at the corpus level, where (a) is the topic-word number distribution generated by the label information, (b) is the proportion of 20 topics obtained by NB-NTM, and (c) is the proportion of 20 topics obtained by GNB-NTM. All results are generated from the 20News testing set.

We also study the dispersion of intermediate variables (i.e., topics) at the document level. By randomly select a document as an example, we get the normalized document topic weight $\boldsymbol{\lambda}$ from NB-NTM and GNB-NTM to explore whether the topic distributions of the document generated by our models are reasonable. As shown in Figure 5, the document is about a standard computer, and the most related topics with large topic distributions are all related to computers, which validates the practical meaning of intermediate variables of both NB-NTM and GNB-NTM at the document level. From the keywords in the most related topics, we further observe that GNB-NTM can identify more computer-related words than NB-NTM. When compared to the whole semantic space as shown in Figure 4, both NB-NTM and GNB-NTM generate more dispersed proportions of topics at the document level. This phenomenon is consistent with the over-dispersed feature (i.e., the variance is larger than the mean) of documents.

6 Conclusion

In this paper, we present two neural mixed counting models named NB-NTM and GNB-NTM. Different from the current time consuming Bayesian methods, our models apply to large-scale datasets through the efficient back-propagation algorithm and GPU acceleration. When compared to the existing neural topic models, both NB-NTM and GNB-NTM can well model the random variables with

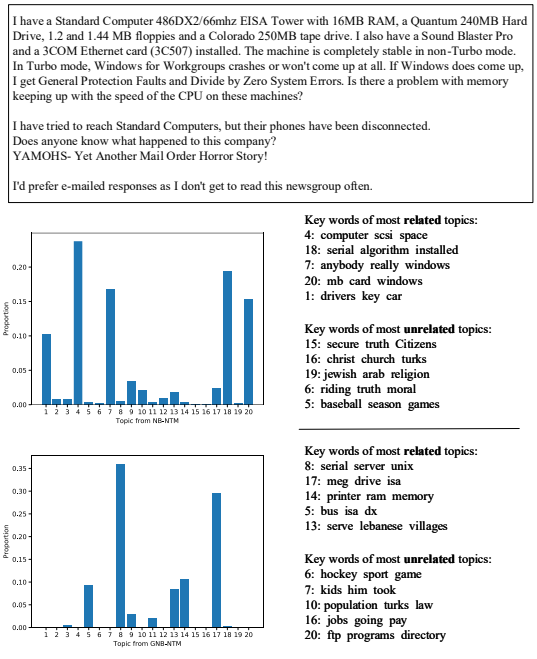


Figure 5: The proportion and key words of 20 topics obtained by our models on a document instance.

over-dispersed and hierarchically dependent characteristics. Extensive experiments on real-world datasets validate the effectiveness of our models in terms of perplexity, topic coherence, and producing explainable intermediate variables by generating dispersed proportions of document topics. The results also indicate that NB distribution families can characterize text data aptly, which is essentially due to their conformity with the over-dispersed and sparse properties of natural language.

Acknowledgment

We are grateful to the reviewers for their constructive comments and suggestions on this study. This work has been supported by the National Natural Science Foundation of China (61972426), Guangdong Basic and Applied Basic Research Foundation (2020A1515010536), HKIBS Research Seed Fund 2019/20 (190-009), the Research Seed Fund (102367), and LEO Dr David P. Chan Institute of Data Science of Lingnan University, Hong Kong. This work has also been supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E01/19), Hong Kong Research Grants Council through a General Research Fund (project no. PolyU 1121417), and by the Hong Kong Polytechnic University through a start-up fund (project no. 980V).

References

Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. 2015. Nonparametric bayesian factor analysis for dynamic count matrices. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*.

David M. Blei and John D. Lafferty. 2005. Correlated topic models. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pages 147–154.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120.

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

John Frank Charles Kingman. 1967. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

Dehua Cheng and Yan Liu. 2014. Parallel gibbs sampling for hierarchical dirichlet processes via gamma processes equivalence. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 562–571.

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 of the Asian Federation of Natural Language Processing*, pages 795–804.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Pankaj Gupta, Yatin Chaudhary, Florian Buettner, and Hinrich Schütze. 2019. Document informed neural autoregressive topic models with distributional prior. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 6505–6512.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 289–296.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

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*, pages 530–539.

John Reidar Mathiassen, Amund Skavhaug, and Ketil Bø. 2002. Texture similarity measure using kullback-leibler divergence between gamma distributions. In *Proceedings of the 7th European Conference on Computer Vision*, pages 133–147.

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*, pages 2410–2419.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 1727–1736.

Andriy Mnih and Karol Gregor. 2014. Neural variational inference and learning in belief networks. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1791–1799.

Christian A. Naesseth, Francisco J. R. Ruiz, Scott W. Linderman, and David M. Blei. 2017. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 489–498.

Eric T. Nalisnick and Padhraic Smyth. 2017. Stick-breaking variational autoencoders. In *Proceedings of the 5th International Conference on Learning Representations*.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 6345–6381.

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*, pages 248–256.

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*, pages 1278–1286.

Akash Srivastava and Charles A. Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of the 5th International Conference on Learning Representations*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, pages 1385–1392.

Romain Thibaux and Michael I. Jordan. 2007. Hierarchical beta processes and the indian buffet process. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pages 564–571.

Michalis K. Titsias. 2007. The infinite gamma-poisson feature model. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems*, pages 1513–1520.

Chong Wang, John W. Paisley, and David M. Blei. 2011. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 752–760.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd International World Wide Web Conference*, pages 1445–1456.

Limin Yao, David M. Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 937–946.

He Zhao, Piyush Rai, Lan Du, Wray L. Buntine, and Mingyuan Zhou. 2019. Variational autoencoders for sparse and overdispersed discrete data. *arXiv preprint arXiv:1905.00616*, abs/1905.00616.

Mingyuan Zhou. 2018. Nonparametric bayesian negative binomial factor analysis. *Bayesian Analysis*, 13(4):1061–1089.

Mingyuan Zhou and Lawrence Carin. 2012. Augment-and-conquer negative binomial processes. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, pages 2555–2563.

Mingyuan Zhou and Lawrence Carin. 2015. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320.

Mingyuan Zhou, Oscar Hernan Madrid Padilla, and James G Scott. 2016. Priors for random count matrices derived from a family of negative binomial processes. *Journal of the American Statistical Association*, 111(515):1144–1156.