Modeling the diversity and log-normality of data

Khoat Than*and Tu Bao Ho‡

July 4, 2013

Abstract

We investigate two important properties of real data: diversity and log-normality. Log-normality accounts for the fact that data follow the lognormal distribution, whereas diversity measures variations of the attributes in the data. To our knowledge, these two inherent properties have not been paid much attention from the machine learning community, especially from the topic modeling community. In this article, we fill in this gap in the framework of topic modeling. We first investigate whether or not these two properties can be captured by the most well-known Latent Dirichlet Allocation model (LDA), and find that LDA behaves inconsistently with respect to diversity. Particularly, it favors data of low diversity, but works badly on data of high diversity. Then, we argue that these two inherent properties can be captured well by endowing the topic-word distributions in LDA with the lognormal distribution. This treatment leads to a new model, named Dirichlet-lognormal topic model (DLN). Using the lognormal distribution complicates the learning and inference of DLN, compared with those of LDA. Hence, we used variational method, in which model learning and inference are reduced to solving convex optimization problems. Extensive experiments strongly suggest that (1) the predictive power of DLN is consistent with respect to diversity, and that (2) DLN works consistently better than LDA for datasets whose diversity is large, and for datasets which contain many log-normally distributed attributes. Justifications for these results require insights into the used statistical distributions and will be discussed in the article.

1 Introduction

Topic modeling is increasingly emerging in machine learning and data mining. More and more successful applications of topic modeling have been reported, e.g., topic discovery [12], [7], information retrieval [33], analyzing social networks

*Corresponding author: Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan. Mobile: +818042557532. E-mail: khoat@jaist.ac.jp.

†Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan. E-mail: bao@jaist.ac.jp

‡University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam.

[21], [34], [27], and trend detection [6]. Although text is often the main target, many topic models are general enough to be used in other applications with non-textual data, e.g., image retrieval [30], [8], and Bio-informatics [16].

Topic models often consider a given corpus to be composed of latent topics, each of which turns out to be a distribution over words. A document in that corpus is a mixture of these topics. These in some models imply that the order of the documents in a corpus does not play an important role. Further, the order of the words in a specific document is often discarded.

One of the most influential models having the above-mentioned assumptions is the *Latent Dirichlet Allocation* model (LDA) [7]. LDA assumes that each latent topic is a sample drawn from a Dirichlet distribution, and that the topic proportions in each document are samples drawn from a Dirichlet distribution as well. This interpretation of topic-word distributions has been utilized in many other models, such as the *Correlated Topic Model* (CTM) [6], the *Independent Factor Topic Model* (IFTM) [20], DCMLDA [11], Labeled LDA [21], and fLDA [1].

1.1 Forgotten characteristics of data

Geologists have shown that the concentration of elements in the Earth's crust distributes very skewed and fits the lognormal distribution well. The latent periods of many infectious diseases also follow lognormal distributions. Moreover, the occurrences of many real events have been shown to be log-normally distributed, see [15] and [13] for more information. In linguistics, the number of words per sentence, and the lengths of all words used in common telephone conversations, fit lognormal distributions. Recently, the number of different words per document in many collections has been observed to very likely follow the lognormal distribution as well [10]. These observations suggest that log-normality is present in many data types.

Another inherent property of data is the "diversity" of features (or attributes). Loosely speaking, diversity of a feature in a dataset is essentially the number of different values of that feature observed in the records of that dataset. For a text corpus, high diversity of a word means a high number of different frequencies observed in the corpus.[1] The high diversity of a word in a corpus reveals that the word may play an important role in that corpus. The diversity of a word varies significantly among different corpora with respect to the importance of that word. Nonetheless, to the best of our knowledge, this phenomenon has not been investigated previously in the machine learning literature.

In the topic modeling literature, log-normality and diversity have not been under consideration up to now. We will see that despite the inherent importance of the diversity of data, existing topic models are still far from appropriately

[1] For example, the word "learning" has 71 different frequencies observed in the NIPS corpus [4]. This fact suggests that "learning" appears in many (1153) documents of the corpus, and that many documents contain this word with very high frequencies, e.g. more than 50 occurrences. Hence, this word would be important in the topics of NIPS.

capturing it. Indeed, in our investigations, the most popular LDA behaved inconsistently with respect to diversity. Higher diversity did not necessarily assure a consistently better performance or a consistently worse performance. Beside, LDA tends to favor data of low diversity. This phenomenon may be reasonably explained by the use of the Dirichlet distribution to generate topics. Such a distribution often generates samples of low diversity, see Section 5 for detailed discussions. Hence the use of the Dirichlet distribution implicitly sets a severe setback on LDA in modeling data with high diversity.

1.2 Our contributions

In this article, we address those issues by using the lognormal distribution. A rationale for our approach is that such distribution often allows its samples to have high variations, and hence is able to capture well the diversity of data. For topic models, we posit that the topics of a corpus are samples drawn from the lognormal distribution. Such an assumption has two aspects: one is to capture the lognormal properties of data, the other is to better model the diversity of data. Also, this treatment leads to a new topic model, named *Dirichlet-Lognormal topic model* (DLN).

By extensive experiments, we found that the use of the lognormal distribution really helps DLN to capture the log-normality and diversity of real data. The greater the diversity of the data, the better prediction by DLN; the more log-normally distributed the data is, the better the performance of DLN. Further, DLN worked consistently with respect to diversity of data. For these reasons, the new model overcomes the above-mentioned drawbacks of LDA. Summarizing, our contributions are as follows:

- We introduce and carefully investigate an inherent property of data, named "diversity". Diversity conveys many important characteristics of real data. In addition, we extensively investigate the existence of log-normality in real datasets.
- We investigate the behaviors of LDA, and find that LDA behaves inconsistently with respect to diversity. These investigations highlight the fact that "diversity" is not captured well by existing topic models, and should be paid more attention.
- We propose a new variant of LDA, called DLN. The new model can capture well the diversity and log-normality of data. It behaves much more consistently than LDA does. This shows the benefits of the use of the lognormal distribution in topic models.

Roadmap of the article: After discussing some related work in the next section, some notations and definitions will be introduced. Some characteristics of real datasets will be investigated in Section 4. By those investigations, we will see the necessity of more attention to diversity and log-normality of data.

Insights into the lognormal and Dirichlet distributions will be discussed in Section 5. Also we will see the rationales of using the lognormal distribution to cope with diversity and log-normality. Section 6 is dedicated to presenting the DLN model. Our experimental results and comparisons will be described in Section 7. Further discussions are in Section 8. The last section presents some conclusions.

2 Related work

In the topic modeling literature, many models assume a given corpus to be composed of some hidden topics. Each document in that corpus is a mixture of those topics. The first generative model of this type is known as *Probabilistic Latent Semantic Analysis* (pLSA) proposed by Hofmann [12]. Assuming pLSA models a given corpus by K topics, then the probability of a word w appearing in document d is

$$P(w|d) = \sum_z P(w|z)P(z|d), \tag{1}$$

where $P(w|z)$ is the probability that the word w appears in the topic $z \in \{1, ..., K\}$, and $P(z|d)$ is the probability that the topic z participates in the document d. However, pLSA regards the topic proportions, $P(z|d)$, to be generated from some discrete and document-specific distributions.

Unlike pLSA, the topic proportions in each document are assumed to be samples drawn from Dirichlet distributions in LDA [7]. Such assumption is strongly supported by the de Finetti theorem on exchangeable random variables [2]. Amazingly, LDA has been reported to be successful in many applications.

Many subsequent topic models have been introduced since then that differ from LDA in endowing distributions on topic proportions. For instance, CTM and IFTM treat the topic proportions as random variables which follow logistic distributions; *Hierarchical Dirichlet Process* (HDP) considers these vectors as samples drawn from a Dirichlet process [25]. Few models differ from LDA in view of topic-word distributions, i.e., $P(\boldsymbol{w}|z)$. Some candidates in this line are *Dirichlet Forest* (DF) [3], *Markov Topic Model* (MTM) [32], and *Continuous Dynamic Topic Model* (cDTM) [31].

Unlike those approaches, we endow the topic-word distributions with the lognormal distribution. Such treatment aims to tackle diversity and log-normality of real datasets. Unlike the Dirichlet distribution used by other models, the lognormal distribution seems to allow high variation of its samples, and thus can capture well high diversity data. Hence it is a good candidate to help us cope with diversity and log-normality.

3 Definitions

The following notations will be used throughout the article.

Notation	Meaning
$\mathcal{C}$	a corpus consisting of M documents
$\mathcal{V}$	the vocabulary of the corpus
$\boldsymbol{w}_i$	the ith document of the corpus
w_{dn}	the nth word in the dth document
w_j	the jth term in the vocabulary $\mathcal{V}$, represented by a unit vector
w_j^i	the ith component of the word vector w_j; $w_j^i = 0, \forall i \neq j, w_j^j = 1$
V	the size of the vocabulary
K	the number of topics
N_d	the length of the dth document
$\boldsymbol{\beta}_k$	the kth topic-word distribution
$\boldsymbol{\theta}_d$	the topic proportion of the dth document
z_{dn}	the topic index of the nth word in the dth document
$\|S\|$	the cardinal of the set S
$Dir(\cdot)$	the Dirichlet distribution
$LN(\cdot)$	the lognormal distribution
$Mult(\cdot)$	the multinomial distribution

Each dataset $\mathcal{D} = \{d_1, d_2, ..., d_D\}$ is a set of D records, composed from a set of features, $\mathcal{A} = \{A_1, A_2, ..., A_V\}$; each record $d_i = (d_{i1}, ..., d_{iV})$ is a tuple of which d_{ij} is a specific value of the feature A_j.

3.1 Diversity

Diversity is the main focus of this article. Here we define it formally in order to avoid confusion with the other possible meanings of this word.

Definition 1 (Observed value set)**.** *Let $\mathcal{D} = \{d_1, d_2, ..., d_D\}$ be a dataset, composed from a set $\mathcal{A}$ of features. The observed value set of a feature $A \in \mathcal{A}$, denoted $OV_{\mathcal{D}}(A)$, is the set of all values of A observed in $\mathcal{D}$.*

Note that the observed value set of a feature is very different from the domain that covers all possible values of that feature.

Definition 2 (Diversity of feature)**.** *Let $\mathcal{D}$ be a dataset, and be composed from a set $\mathcal{A}$ of features. The diversity of the feature A in the data set $\mathcal{D}$ is*

$$Div_{\mathcal{D}}(A) = \frac{|OV_{\mathcal{D}}(A)|}{|\mathcal{D}|}$$

Clearly, diversity of a feature defined above is the normalized version of the number of different values of that feature in the data set. This concept is introduced in order to compare different datasets.

The diversity of a dataset is defined via averaging the diversities of the features of that dataset. This number will provide us an idea about how variation a given dataset is.

Definition 3 (Diversity of dataset)**.** *Let $\mathcal{D}$ be a dataset, composed from a set $\mathcal{A}$ of features. The diversity of the dataset $\mathcal{D}$ is*

$$Div_{\mathcal{D}} = average\{Div_{\mathcal{D}}(A) : A \in \mathcal{A}\}$$

Note that the concept of diversity defined here is completely different from the concept of variance. Variance often relates to the variation of a random variable from the true statistical mean of that variable whereas diversity provides the extent of variation in general of a variable. Furthermore, diversity only accounts for a given dataset, whereas variance does not. The diversity of the same feature may vary considerably among different datasets.

By means of averaging over all features, the diversity of a dataset surfers from outliers. In other words, the diversity of a dataset may be overly dominated by very few features, which have very high diversities. In this case, the diversity is not a good measure of the variation of the considered dataset. Overcoming this situation will be our future work.

We will often deal with textual datasets in this article. Hence, for the aim of clarity, we adapt the above definitions for text and discuss some important observations regarding such a data type.

If the dataset $\mathcal{D}$ is a text corpus, then the observed value set is defined in terms of frequency. We remark that in this article each document is represented by a sparse vector of frequencies, each component of which is the number of occurrences of a word occurred in that document.

Definition 4 (Observed frequency set)**.** *Let $\mathcal{C} = \{d_1, d_2, ..., d_M\}$ be a text corpus of size M, composed from a vocabulary $\mathcal{V}$ of V words. The observed frequency set of the word $w \in \mathcal{V}$, denoted $OV_{\mathcal{C}}(w)$, is the set of all frequencies of w observed in the documents of $\mathcal{C}$.*

$$OV_{\mathcal{C}}(w) = \{freq(w) : \exists d_i \text{ that contains exactly } freq(w) \text{ occurrences of } w\}$$

In this definition, there is no information about how many documents have a certain $freq(w) \in OV_{\mathcal{C}}(w)$. Moreover, if a word w appears in many documents with the same frequency, the frequency will be counted only once. The observed frequency set tells much about the behavior and stability of a word in a corpus. If $|OV_{\mathcal{C}}(w)|$ is large, w must appear in many documents of $\mathcal{C}$. Moreover, many documents must have high frequency of w. For example, if $|OV_{\mathcal{C}}(w)| = 30$, w must occur in at least 30 documents, many of which contain at least 20 occurrences of w.

Definition 5 (Diversity of word)**.** *Let $\mathcal{C}$ be a corpus, composed from a vocabulary $\mathcal{V}$. The diversity of the word $w \in \mathcal{V}$ in the corpus is*

$$Div_{\mathcal{C}}(w) = \frac{|OV_{\mathcal{C}}(w)|}{|\mathcal{C}|}$$

Definition 6 (Diversity of corpus)**.** *Let $\mathcal{C}$ be a corpus, composed from a vocabulary $\mathcal{V}$. The diversity of the corpus is*

$$Div_{\mathcal{C}} = average\{Div_{\mathcal{C}}(w) : w \in \mathcal{V}\}$$

It is easy to see that if a corpus has high diversity, a large number of its words would have a high number of different frequencies, and thus have high variations in the corpus. These facts imply that such kind of corpora seem to be hard to deal with. Moreover, provided that the sizes are equal, a corpus with higher diversity has higher variation, and hence may be more difficult to model than a corpus with lower diversity. Indeed, we will see this phenomenon in the later analyses.

3.2 Topic models

Loosely speaking, a topic is a set of semantically related words [14]. For examples, {*computer, information, software, memory, database*} is a topic about "computer"; {*jazz, instrument, music, clarinet*} may refer to "instruments for Jazz"; and {*caesar, pompay, roman, rome, carthage, crassus*} may refer to a battle in history.

Formally, we define a topic to be a distribution over a fixed vocabulary. Let $\mathcal{V}$ be the vocabulary of V terms, a topic $\boldsymbol{\beta}_k = (\beta_{k1}, ..., \beta_{kV})$ satisfies $\sum_{i=1}^{V} \beta_{ki} = 1$ and $\beta_{ki} \geq 0$ for any i. Each component β_{ki} shows the probability that term i contributes to topic k. A *topic model* is a statistical model of those topics. A corpus is often assumed to be composed of K topics, for some K.

Each document is often assumed to be a mixture of the topics. In other words, in a typical topic model, a document is assumed to be composed from some topics with different proportions. Hence each document will have another representation, says $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ where θ_k shows the probability that topic k appears in that document. $\boldsymbol{\theta}$ is often called *topic proportion.*

The goal of topic modeling is to automatically discover the topics from a collection of documents [5]. In reality, we can only observe the documents, while the *topic structure* including topics and topic proportions is *hidden*. The central problem for topic modeling is to use the observed documents to infer the topic structure.

Topic models provide a way to do dimension reduction if setting $K < V$. Learning a topic model implies we are learning a topical space, in which each document has a latent representation $\boldsymbol{\theta}$. Therefore, $\boldsymbol{\theta}$ can be used for many tasks including text classification, spam filtering, and information retrieval [12], [7], [26].

3.3 Dirichlet and lognormal distributions

In this article, we will often mention lognormal and Dirichlet distributions. Hence we include here their mathematical definitions. The lognormal distribution of a random variable $\boldsymbol{x} = (x_1, ..., x_n)^T$, with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, has the following density function

$$LN(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\boldsymbol{\Sigma}|} x_1 ... x_n} \exp\{-\frac{1}{2}(\log \boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{x} - \boldsymbol{\mu})\}.$$

Table 1: Datasets for experiments.

Data set	AP	NIPS	KOS	SPAM	Comm-Crime
Number of documents	2246	1500	3430	4601	1994
Vocabulary size	10473	12419	6906	58	100
Document length	194.05	1288.24	136.36		
#unique words per doc	134.48	497.54	102.96		

Similarly, the density function of the Dirichlet distribution is

$$Dir(\boldsymbol{x}; \alpha_1, \ldots, \alpha_n) = \frac{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n x_i^{\alpha_i - 1},$$

where $\sum_{i=1}^n x_i = 1, x_i > 0$. The constraint means that the Dirichlet distribution is in fact in $(n-1)$-dimensional space.

4 Diversity and Log-normality of real data

We first describe our initial investigations on 5 real datasets from the UCI Machine Learning Repository [4] and Blei's webpage.[2] Some information on these datasets is reported in Table 1, in which the last two rows have been averaged. In fact, the Communities and Crime dataset (Comm-Crime for short) is not a usual text corpus. This data set contains 1994 records each of which is the information of a US city. There are 123 attributes, some of which are missing for some cities [22]. In our experiments, we removed the attributes from all records if they are missing in some records. Also, we removed the first 5 non-predictive attributes, and the remainings consist of only 100 real attributes including crime.

Our initial investigations studied the diversity of the above data sets. These three textual corpora, AP, NIPS, and KOS, were preprocessed to remove all function words and stopwords, which are often assumed to be meaningless to the gists of the documents. The remaining are content words. Some statistics are given in Table 2.

One can easily realize that the diversity of NIPS is significantly larger than that of AP and KOS. Among 12419 words of NIPS, 5900 words have at least 5 different frequencies; 1633 words have at least 10 different frequencies.[3] These facts show that a large number of words in NIPS vary significantly within the corpus, and hence may cause considerable difficulties for topic models.

AP and KOS are comparable in terms of diversity. Despite this fact, AP seems to have quite greater variation compared with KOS. The reason is that

[2]The AP corpus: http://www.cs.princeton.edu/∼blei/lda-c/ap.tgz

[3]The three words which have greatest number of different frequencies, $|OV|$, are "network", "model", and "learning". Each of these words appears in more than 1100 documents of NIPS. To some extent, they are believed to compose the main theme of the corpus with very high probability.

Table 2: Statistics of the 3 corpora. Although NIPS has least documents among the three corpora, all of its statistics here are much greater than those of the other two corpora.

Data set	AP	KOS	NIPS
Diversity	0.0012	0.0011	0.004
No. of words with $\|OV\| \geq 5$	1267	1511	5900
No. of words with $\|OV\| \geq 10$	99	106	1633
No. of words with $\|OV\| \geq 20$	1	4	345
Three greatest $\|OV\|$'s	{25; 19; 19}	{26; 21; 21}	{86; 80; 71}

although the number of documents in AP is nearly 10/15 of that in KOS, the number of words with $|OV| \geq 5$ in AP is approximately 12/15 of that in KOS. Furthermore, KOS and AP have nearly the same number of words with $|OV| \geq 10$. Another explanation for the larger variation of AP over KOS is is that the documents in AP are much longer on average than those of KOS, see Table 1. Longer documents would generally provide more chances for occurrences of words, and thus would probably encourage greater diversity for a corpus.

Comm-Crime and SPAM are non-textual datasets. Their diversities are 0.0458 and 0.0566, respectively. Almost all attributes have $|OV| \geq 30$, except one in each data set, and the greatest $|OV|$ in SPAM is 2161 which is far greater than that in the textual counterparts. The values of attributes are mostly real numbers, and vary considerably. This is why their diversities are much larger than those of textual corpora.

The next investigations were on how individual content words distribute in a corpus. We found that many words (attributes) of SPAM and Comm-Crime very likely follow lognormal distributions. Figure 1 shows the distributions of some representative words. To see whether or not these words are likely log-normally distributed, we fitted the data with lognormal distributions by maximum likelihood estimation. The solid thin curves in the figure are density functions of the best fitted lognormal distributions. We also fitted the data with the Beta distribution.[4] Interestingly, Beta distributions, as plotted by dashed curves, fit data very badly. By more investigations, we found that more than 85% of attributes in Comm-Crime very likely follow lognormal distributions. This amount in SPAM is 67%. For AP, NIPS and KOS, not many words seem to be log-normally distributed.

5 Insights into the Lognormals and Dirichlets

The previous section provided us an overview on the diversity and log-normality of the considered datasets. Diversity differs from dataset to dataset, and in some

[4]Note that Beta distributions are 1-dimensional Dirichlet distributions. We fitted the data with this distribution for the aim of comparison in terms of goodness-of-fit between the Dirichlet and lognormal distributions.

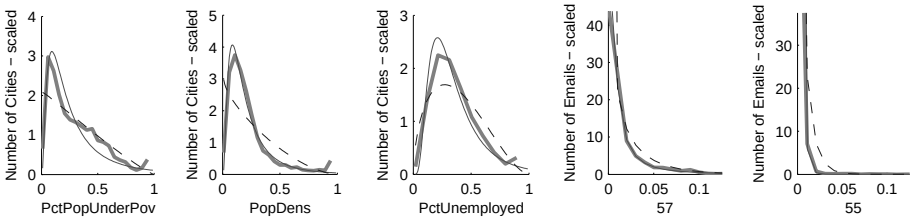

Figure 1: Distributions of some attributes in Comm-Crime and SPAM. Bold curves are the histograms of the attributes. Thin curves are the best fitted Lognormal distributions; dashed curves are the best fitted Beta distributions.

respects represents characteristics of data types. Textual data often have much less diversity than non-textual data. There are non-negligible differences in terms of diversity between text corpora. We also have seen that many datasets have many log-normally distributed properties. These facts raise an important question of how to model well diversity and log-normality of real data.

Taking individual attributes (words) into account in modeling data, one may immediately think about using the lognormal distribution to deal with the log-normality of data. This naive intuition seems to be appropriate in the context of topic modeling. As we shall see, the lognormal distribution is not only able to capture log-normality, but also able to model well diversity. Justifications for those abilities may be borrowed from the characteristics of the distribution.

Attempts to understand the lognormal and Dirichlet distributions were initiated. We began by illustrating the two distributions in 2-dimensional space. Depicted in Figure 2 are density functions with different parameter settings.

As one can easily observe, the mass of the Dirichlet distribution will shift from the center of the simplex to the corners as the values of the parameters decrease. Conversely, the mass of the lognormal distribution will shift from the origin to regions which are far from the origin as σ decreases. From more careful observations, we realized that the lognormal distribution often has long (thick) tails as σ is large, and has quickly-decreased thin tails as σ is small. Nonetheless, the reverse phenomenon is the case for the Dirichlet distribution.

The tails of a density function tell us much about that distribution. A distribution with long (thick) tails would often generate many samples which are outside of its mass. This fact suggests that the variations of individual random variables in such a multivariate distribution might be large. As a consequence, such probability distributions often generate samples of high diversity.

Unlike distributions with long tails, those with short (thin) tails considerably restrict variations of theirs samples. This implies that individual random variables in such distributions may be less free in terms of variation than those in long-tail distributions. Therefore, probability distributions with short thin tails are likely to generate samples of low diversity.

The above arguments suggest at least two implications. First, the lognormal distribution probably often generates samples of high diversity, and hence is

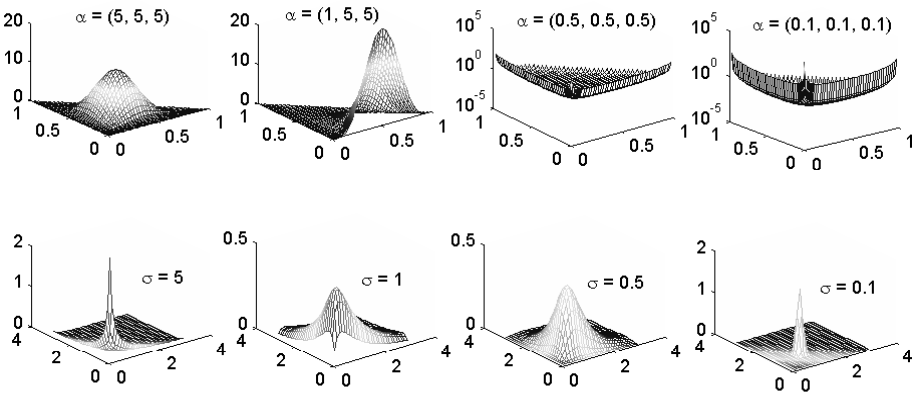

Figure 2: Illustration of two distributions in the 2-dimensional space. The top row are the Dirichlet density functions with different parameter settings. The bottom row are the Lognormal density functions with parameters set as $\boldsymbol{\mu} = 0, \boldsymbol{\Sigma} = Diag(\sigma)$.

capable of modeling high diversity data, since it often has long (thick) tails. Second, the Dirichlet distribution is appropriate to model data of low diversity like text corpora. As a result, it seems to be inferior in modeling data of high diversity, compared with the lognormal distribution.

With the aim of illustrating the above conclusions, we simulated an experiment as follows. Using tools from Matlab, we made 6 synthetic datasets from samples organized into documents. 3 datasets were constructed from samples drawn from the Beta distribution with parameters $\alpha = (0.1, 0.1)$; the others were from 1-dimensional lognormal distribution with parameters $\mu = 0, \sigma = 1$. All samples were rounded to the third decimal. Note that the Beta distribution is the 1-dimensional Dirichlet distribution. Some information of the 6 synthetic datasets is reported in Table 3. Observe that with the same settings, the lognormal distribution gave rise to datasets with significantly higher diversity than the Beta distribution. Hence, this simulation supports further our conclusions above.

6 The DLN model

We have discussed in Section 5 that the Dirichlet distribution seems to be inappropriate with data of high diversity. It will be shown empirically in the next section that this distribution often causes a topic model to be inconsistent with respect to diversity. In addition, many datasets seem to have log-normally distributed properties. Therefore, it is necessary to derive new topic models that can capture well diversity and log-normality. In this section, we describe a new variant of LDA, in which the Dirichlet distribution used to generate topics is

Table 3: Synthetic datasets originated from the Beta and lognormal distributions. As shown in this table, the Beta distribution very often yielded the same samples. Hence it generated datasets with diversity which is often much less than the number of attributes. Conversely, the lognormal distribution sometimes yielded repeated samples, and thus resulted in datasets with very high diversity.

Dataset	Drawn from	#Documents	#Attributes	Diversity
1	lognormal	1000	200	193.034
2	beta	1000	200	82.552
3	lognormal	5000	200	193.019
4	beta	5000	200	82.5986
5	lognormal	5000	2000	1461.6
6	beta	5000	2000	456.6768

replaced with the lognormal distribution.

Similar with LDA, the DLN model assumes the bag-of-words representations for both documents and corpus. Let $\mathcal{C}$ be a given corpus that consists of M documents, composed from the vocabulary $\mathcal{V}$ of V words. Then the corpus is assumed to be generated by the following process:

1. For each topic $k \in \{1, ..., K\}$, choose
   $\boldsymbol{\beta}_k | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim LN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
2. For each document d in the corpus:
   (a) Choose topic proportions $\boldsymbol{\theta}_d | \alpha \sim Dir(\alpha)$
   (b) For the nth word w_{dn} in the document,
   - Choose topic index $z_{dn} | \boldsymbol{\theta}_d \sim Mult(\boldsymbol{\theta}_d)$
   - Generate the word
     $w_{dn} | \boldsymbol{\beta}, z_{dn} \sim Mult(f(\boldsymbol{\beta}_{z_{dn}}))$.

Here $f(\cdot)$ is a mapping which maps $\boldsymbol{\beta}_k$ to parameters of multinomial distributions. In DLN, the mapping is

$$f(\boldsymbol{\beta}_k) = \frac{\boldsymbol{\beta}_k}{\sum_{j=1}^{V} \beta_{kj}}.$$

The graphical representation of the model is depicted in Figure 3. We note that the distributions used to endow the topics are the main differences between DLN and LDA. Using the lognormal distribution also results in various difficulties in learning the model and inferring new documents. To overcome those difficulties, we used variational methods. For detailed description of model learning and inference, please see Section A.

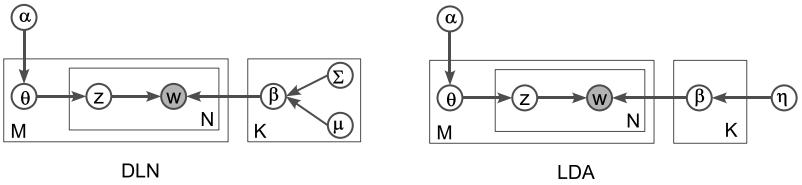

Figure 3: Graphical model representations of DLN and LDA.

7 Evaluation

This section is dedicated to presenting evaluations and comparisons for the new model. The topic model that will be used to compare with DLN is LDA. As previously mentioned, LDA is very popular and is the core of various topic models, where the topic-word distributions are endowed with the Dirichlet distribution. This view on topics is the only point in which DLN differs from LDA. Hence, any advantages of DLN over LDA can be applied to other variants of LDA. Further, any LDA-based model can be readily modified to become a DLN-based model. From these observations, it is reasonable to compare performances of DLN and LDA.

Our strategy is as follows:

- We want to see how good the predictive power of DLN is in general. Perplexity will be used as a standard measure for this task.
- Next, stability of topic models with respect to diversity will be considered. Additionally, we will also study whether LDA and DLN likely favor data of low or high diversity. See subsection 7.2.
- Finally, we want to see how well DLN can model data having log-normality and high diversity. This will be measured via classification on two non-textual datasets, Comm-Crime and SPAM. Details are in subsection 7.3.

7.1 Perplexity as a goodness-of-fit measure

We first use perplexity as a standard measure to compare LDA and DLN. Perplexity is a popular measure which evaluates the goodness-of-fit of a statistical model, and is widely used in the language modeling community. It is known to correlate closely with the precision-recall measure in information retrieval [12]. The measure is often used to compare predictive powers of different topic models as well.

Let $\mathcal{C}$ be the training data, and $\mathcal{D} = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_T\}$ be the test set. Then perplexity is calculated by

$$Perp(\mathcal{D}|\mathcal{C}) = \exp\left(-\frac{\sum_{d=1}^{T} \log P(\boldsymbol{w}_d|\mathcal{C})}{\sum_{d=1}^{T} |\boldsymbol{w}_d|}\right).$$

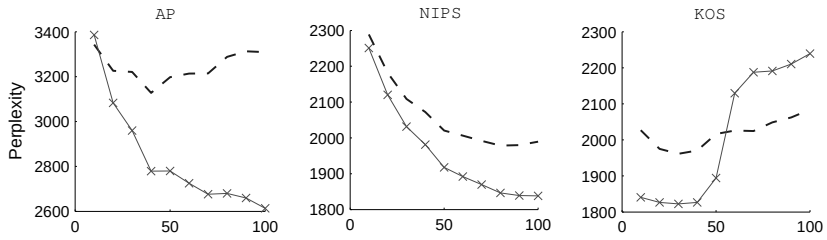

Figure 4: Perplexity as the number of topics increases. Solid curves are DLN, dashed curves are LDA. The lower is the better.

The data for this task were the 3 text corpora. The two non-textual data sets were not considered, since perplexity is implicitly defined for text. For each of the 3 text corpora, we selected randomly 90% of the data to train DLN and LDA, and the remainings were used to test their predictive powers. Both models used the same convergence settings for both learning and inference. Figure 4 shows the results as the number of topics increases. We can see clearly that DLN achieved better perplexity for AP and NIPS than LDA. However, it behaved worse than LDA on the KOS corpus.

Remember that NIPS has the greatest diversity among these 3 corpora as investigated in Section 4. That is, the variations of the words in that corpus are very high. Besides, the lognormal distribution seems to favor data of high diversity as analyzed in Section 5. The use of this distribution in DLN aims to capture the diversity of individual words better. Hence the better perplexity of DLN over LDA for the NIPS corpus is apparently justified.

The better result of DLN on NIPS also suggests more insights into the LDA model. In Section 5 we have argued that the Dirichlet distribution seems to favor data of low diversity, and seems inappropriate for high diversity data. These hypotheses are further supported by our experiments in this section.

Note that AP and KOS have nearly equal diversity. Nevertheless, the performances of both models on these corpora were quite different. DLN was much better than LDA on AP, but not on KOS. This phenomenon should be further investigated. In our opinion, some explanations for this may be borrowed from some observations in Section 4. Notice that although the number of documents of KOS is approximately 50% larger than that of AP, the number of words having at least 5 different frequencies ($|OV| \geq 5$) in KOS is only about 20% larger than that of AP. This fact suggests that the words in AP seem to have higher variations than those in KOS. Besides, $Div_{\text{AP}} > Div_{\text{KOS}}$. Combining these observations, we can conclude that AP has higher variation than KOS. This is probably the reason why DLN performed better than LDA on AP.

7.2 Stability in predictive power

Next we would like to see whether the two models can work stably with respect to diversity. The experiments described in the previous subsection are not good

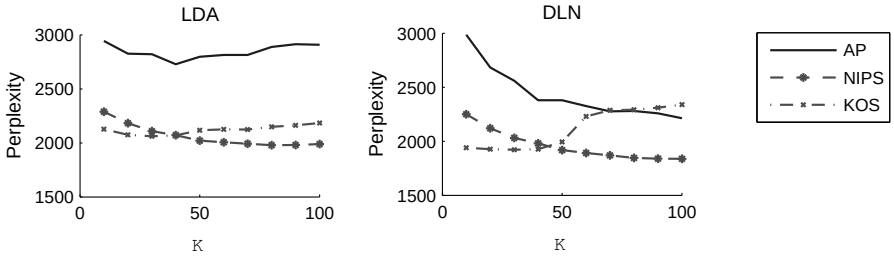

Figure 5: Sensitivity of LDA and DLN against diversity, measured by perplexity as the number of topics increases. The testing sets were of same size and same document length in these experiments. Under the knowledge of $Div_{\text{NIPS}} > Div_{\text{AP}} > Div_{\text{KOS}}$, we can see that LDA performed inconsistently with respect to diversity; DLN performed much more consistently.

enough to see this. The reason is that both topic models were tested on corpora of different numbers of documents, each with different document length. It means comparisons across various corpora by perplexity would not be fair if based on those experiments. Hence we need to conduct other experiments for this task.

Perplexity was used again for this investigation. To arrive at fair comparisons and conclusions, we need to measure perplexity on corpora of the same size and same document length. In order to have such corpora, we did as follows. We used 3 text corpora as above. For each corpus, 90% were randomly chosen for training, and the remaining were used for testing. In each testing set, each document was randomly cut off to remain only 100 occurrences of words in total. This means the resulting documents for testing were of the same length across testing sets. Additionally, we randomly removed some documents to remain only 100 documents in each testing set. Finally, we have 3 testing sets which are equal in size and document length.

After learning both topic models, the testing sets were inferred to measure their predictive powers. The results are summarized in Figure 5. As known in Section 4, the diversity of NIPS is greater than those of AP and KOS. However, LDA performed inconsistently in terms of perplexity on these corpora as the number of topics increased. Higher diversity led to neither consistently better nor consistently worse perplexity. This fact suggests that LDA cannot capture well the diversity of data.

In comparison with LDA, DLN worked more consistently on these corpora. It achieved the best perplexity on NIPS, which has the largest diversity among 3 corpora. The gap in perplexity between NIPS and the others is quite large. This implies that DLN may capture well data of high diversity. However, since the perplexity for AP was worse than that for KOS while $Div_{\text{AP}} = 0.0012 > Div_{\text{KOS}} = 0.0011$, we do not know clearly whether DLN can cope well with data of low diversity or not. Answers for this question require more sophisticated investigations.

Another observation from the results depicted in Figure 5 is that LDA seems to work well on data of low diversity, because its perplexity on KOS was consistently better than on other corpora. A reasonable explanation for this behavior is the use of the Dirichlet distribution to generate topics. Indeed, such distribution favors low diversity, as analyzed in Section 5. Nonetheless, it is still unclear to conclude that LDA really works well on data of low diversity, because its perplexity for KOS was much better than that for AP while $Div_{\text{AP}} \simeq Div_{\text{KOS}}$.

7.3 Document classification

Our next experiments were to measure how well the two models work, via classification tasks, when data have high diversity and log-normality. As is well-known, topic models are basically high-level descriptions of data. In other words, the most interesting characteristics of data are expected to be captured in topic models. Hence topic models provide new representations of data. This interpretation implicitly allows us to apply them to many other applications, such as classification [26], [7].

The datasets for these tasks are SPAM and Comm-Crime. We used micro precision [23] as a measure for comparison. Loosely speaking, precision can be interpreted as the extent of our confidence in assigning labels to documents. It is believed, at least in the text categorization community, that this measure is more reliable than the accuracy measure for classification [23]. Thus it is reasonable to use it for our tasks in this section.

SPAM is straightforward to understand, and is very suitable for the classification task. The main objective is to predict whether a given document is spam or not. Thus, we keep the spam attribute unchanged, and multiply all values of other attributes in all records by 10000 to make sure that the obtained values are integers. Resulting records are regarded as documents in which each value of an attribute is the frequency of the associated word.

The nature of Comm-Crime is indirectly related to classification. The goal of Comm-Crime is to predict how many violent crimes will occur per 100K population. In this corpus, all cities have these values that can be used to train or test a learning algorithm. Since predicting an exact number of violent crimes is unrealistic, we predicted the interval in which the number of violent crimes of a city most probably falls.[5]

Since all crime values in the original data were normalized to be in [0,1], two issues arise when performing classification on this dataset. First, how many intervals are appropriate? Second, how to represent crime values, each belonging to exactly one interval, as class labels. The first issue is easier to deal with in practice than the latter. In our experiments, we first tried 10 intervals, and then 15 intervals. For the second issue, we did as follows: each attribute was

[5]Be aware that this dataset is also suitable to be used in regression, since the data were previously normalized to be in [0, 1]. However, this section is devoted to comparing topic models in terms of how well they can capture diversity and log-normality of data. SPAM and Comm-Crime are good datasets for these tasks, because they both have high diversity and many likely log-normally distributed attributes.

Table 4: Average precision in crime prediction.

#intervals	SVM	DLN+SVM	LDA+SVM
10	0.56	**0.61**	0.58
15	0.43	**0.48**	0.46

associated with a word except crime. The values of the attributes were scaled by the same number to make sure that all are integers, and then were regarded as frequencies of the associated words. For the crime attribute, we associated each interval with each class label. Each record then corresponds to a document, where the crime value is associated with a class label.

We considered performances on Comm-Crime of 3 approaches: SVM, DLN+SVM, LDA+SVM. Here we used multi-class SVM implemented in the package by Joachims.[6] It was trained and tested on the original dataset to ensure fair comparisons. DLN+SVM (and LDA+SVM) worked in the same way as in previous works [7], i.e., we first modeled the data by DLN (LDA) to find latent representations of the documents in terms of topic proportions vectors, and then used them as feature vectors for SVM. Note that different kernels can be used for SVM, DLN+SVM, LDA+SVM, which could lead to different results [24]. Nonetheless, our main aims are to compare performance of topic models. Hence, using the linear kernel for three methods seems sufficient for our aims. For each classification method, the regularization constant C was searched from $\{1, 10, 100, 1000\}$ to find the best one. We further used 5-fold cross-validation and reported the averaged results.

For topic models, the number of topics should be chosen appropriately. In [29], Wallach et al. empirically showed that LDA may work better as the number of topics increases. Nevertheless, the subsections 7.1 and 7.2 have indicated that large values of K did not lead to consistently better perplexity for LDA. Moreover, the two models did not behave so badly at $K = 50$. Hence we chose 50 topics for both topic models in our experiments. The results are presented in Table 4.

Among the 3 approaches, DLN+SVM consistently performed best. These results suggest that DLN worked better than LDA did. We remark that Comm-Crime has very high diversity and seems to have plenty of log-normality. Hence the better performance of DLN over LDA suggests that the new model can capture well log-normality of data, and can work well on data of high diversity.

One can realize that the precisions obtained from these approaches were quite low. In our opinion, this may be due to the inherent nature of that data. To provide evidence for our belief, we conducted separately regression on the original Comm-Crime dataset with two other well-known methods, Bagging and Linear Regression implemented in Weka.[7] Experiments with these methods used default parameters and used 5-fold cross-validation. Mean absolute errors from these experiments varied from 0.0891 to 0.0975. Note that all values of

[6] Available from http://svmlight.joachims.org/svm_multiclass.html

[7] Version 3.7.2 at http://www.cs.waikato.ac.nz/∼ml/weka/

Table 5: Average precision in spam filtering.

SVM	DLN+SVM	LDA+SVM
0.81	**0.95**	0.92

the attributes in the dataset had been normalized to be in [0, 1]. Therefore the resulting errors are problematic. After scaling and transforming the regression results to classification, the consequent precisions vary from 0.3458 to 0.4112. This variation suggests that Comm-Crime seems to be difficult for current learning methods.

The above experiments on Comm-Crime provide some supporting evidence for the good performance of DLN. We next conducted experiments for classification on SPAM. We used the same settings as above, 50 topics for topic models and 5-fold cross-validation. The results are described in Table 5. One can easily observe the consistently better performance of our new model over LDA, working in combination with SVM. Note that precisions for SPAM are much greater than those for Comm-Crime. The reasons are that SPAM is inherently for binary classification, which is often easier than multi-class counterparts, and that the training set for SPAM is much bigger than that for Comm-Crime which enables better learning.

8 Discussion

In summary, we now have strong evidence from the empirical results and analyses for the following conclusions. First, DLN can get benefits from data that have many likely log-normally distributed properties. It seems to capture well log-normality of data. Second, DLN is more suitable than LDA on data of high diversity, since consistently better performances have been observed. Third, topic models are able to model well data that are non-textual, since the combinations of topic models with SVM often got better results than SVM did alone in our experiments.

LDA and DLN have been compared in various evaluations. The performance of DLN was consistent with the diversity of data, whereas LDA was inconsistent. Furthermore, DLN performed consistently better than LDA on data that have high diversity and many likely log-normally distributed properties. Note that in our experiments, the considered datasets have different diversities. This treatment aimed to ensure that each conclusion will be strongly supported. In addition, the lognormal distribution is likely to favor data of high diversity as demonstrated in Section 5. Hence, the use of the lognormal distribution in our model really helps the model to capture diversity and log-normality of real data.

Although the new model has many distinguishing characteristics for real applications, it suffers from some limitations. First, due to the complex nature of the lognormal distribution, learning the model from real data is complicated and time-consuming. Second, the memory for practical implementation is large, $O(K.V.V + M.V + K.M)$, since we have to store K different lognormal distri-

butions corresponding to K topics. Hence it is suitable with corpora of average vocabularies, and datasets with average numbers of attributes.

Some concerns may arise when applying DLN in real applications: what characteristics of data ensure the good performance of DLN? Which data types are suitable for DLN? The followings are some of our observations.

- For non-textual datasets, DLN is very suitable if diversity is high. Our experiments suggest that the higher diversity the data have, the better DLN can perform. Note that diversity is basically proportional to the number of different values of attributes observed in a dataset. Hence, by intuition, if there are many attributes that vary significantly in a dataset, then the diversity of that dataset would be probably high, and thus DLN would be suitable.

- Log-normality of data is much more difficult to see than diversity.[8] Nonetheless, if once we know that a given dataset has log-normally distributed properties, DLN would probably work better on it than LDA.

- For text corpora, the diversity of a corpus is essentially proportional to the number of different frequencies of words observed in the corpus. Hence if a corpus has words that vary significantly, DLN would probably work better than LDA. The reason is that DLN favors data of high diversity.

- A corpus whose documents are often long will allow high variations of individual words. This implies that such a corpus is very likely to have high diversity. Therefore, DLN would probably work better than LDA, as observed in the previous section. Corpora with short documents seem to be suitable for LDA.

- A corpus that is made from different sources with different domains would very likely have high diversity. As we can see, each domain may result in a certain common length for its documents, and thus the average document length would vary significantly among domains. For instance, scientific papers in NIPS and news in AP differ very much in length; conversations in blogs are often shorter than scientific papers. For such mixed corpora, DLN seems to work well, but LDA is less favorable.

The concept of "diversity" in this work is limited to a fixed dataset. Therefore, it is an open problem to extend the concept to the cases that our data is dynamic or streams. When the data is dynamic, it is very likely that behaviors of features often will be complex. Another limitation of the concept is that data are assumed to be free of noises and outliers. When noises or outliers appear in a dataset, the diversity of features will be probably high. This could cause

[8] In principle, checking the presence of log-normality in a dataset is possible. Indeed, checking the log-normality property is equivalent to checking the normality property. This is because if a variable x follows the normal distribution, then $y = e^x$ will follow the log-normal distribution [13], [15]. Hence, checking the log-normality property of a dataset $\mathcal{D}$ can be reduced to checking the normality property of the logarithm version of $\mathcal{D}$.

the modeling more difficult. In our work, we found that the lognormal distribution can model well high diversity of data. Therefore, in the cases of noises or outliers, it seem better to employ this distribution to develop robust models. Nevertheless, this conjecture is left open for future research.

9 Conclusion

In this article, we studied a fundamental property of real data, phrased as "diversity", which has not been paid enough attention from the machine learning community. Loosely speaking, diversity measures average variations of attributes within a dataset. We showed that diversity varies significantly among different data types. Textual corpora often have much less diversity than non-textual datasets. Even within text, diversity varies significantly among different types of text collections.

We empirically showed that diversity of real data non-negligibly affects performance of topic models. In particular, the well-known LDA model [7] worked inconsistently with respect to diversity. In addition, LDA seems not to model well data of high diversity. This fact raises an important question of how to model well the diversity of real corpora.

To deal with the inherent diversity property, we proposed a new variant of LDA, called DLN, in which topics are samples drawn from the lognormal distribution. In spite of being a simple variant, DLN was demonstrated to model well the diversity of data. It worked consistently and seemingly proportionally as diversity varies. On the other hand, the use of the lognormal distribution also allows the new model to capture lognormal properties of many real datasets [15], [10].

Finally, we remark that our approach here can be readily applied to various topic models since LDA is their core. In particular, the Dirichlet distribution used to generate topics can be replaced with the lognormal distribution to cope with diversity of data.

Acknowledgments

We would like to thank the reviewers for many helpful comments. K. Than was supported by MEXT, and T.B. Ho was partially supported by Vietnam's National Foundation for Science and Technology Development (NAFOSTED Project No. 102.99.35.09). This work was partially supported by JSPS Kakenhi Grant Number 23300105.

A Variational method for learning and posterior inference

There are many learning approaches to a given model. Nonetheless, the lognormal distribution used in DLN is not conjugate with the multinomial distribution. So learning the parameters of the model is much more complicated than that of LDA. We use variational methods [28] for our model.

The main idea behind variational methods is to use simpler variational distributions to approximate the original distributions. Those variational distributions should be tractable to learn their parameters, but still give good approximations.

Let $\mathcal{C}$ be a given corpus of M documents, say $\mathcal{C} = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_M\}$. $\mathcal{V}$ is the vocabulary of the corpus and has V words. The jth word of the vocabulary is represented as the jth unit vector of the V-dimensional space $\mathcal{R}^V$. More specifically, if w_j is the jth word in the vocabulary $\mathcal{V}$ and w_j^i is the ith component of w_j, then $w_j^i = 0$ for all $i \neq j$, and $w_j^j = 1$. These notations are similar to those in [7] for ease of comparison.

The starting point of our derivation for learning and inference is the joint distribution of latent variables for each document d, $P(\boldsymbol{z}_d, \boldsymbol{\theta}_d, \boldsymbol{\beta}|\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. This distribution is so complex that it is intractable to deal with. We will approximate it by the following variational distribution:

$$\begin{aligned} Q(\boldsymbol{z}_d, \boldsymbol{\theta}_d, \boldsymbol{\beta}|\boldsymbol{\phi}_d, \boldsymbol{\gamma}_d, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) &= Q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) Q(\boldsymbol{z}_d|\boldsymbol{\phi}_d) \prod_{k=1}^{K} Q(\boldsymbol{\beta}_k|\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k) \\ &= Q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \prod_{n=1}^{N_d} Q(z_{dn}|\boldsymbol{\phi}_{dn}) \prod_{k=1}^{K} \prod_{j=1}^{V} Q(\beta_{kj}|\widehat{\mu}_{kj}, \widehat{\sigma}_{kj}^2) \end{aligned}$$

Where $\widehat{\boldsymbol{\Sigma}}_k = diag(\widehat{\sigma}_{k1}^2, ..., \widehat{\sigma}_{kV}^2), \widehat{\boldsymbol{\mu}}_k = (\widehat{\mu}_{k1}, ..., \widehat{\mu}_{kV})^T \in \mathcal{R}^V$. The variational distribution of discrete variable z_{dn} is specified by the K-dimensional parameter $\boldsymbol{\phi}_{dn}$. Likewise, the variational distribution of continuous variable $\boldsymbol{\theta}_d$ is specified by the K-dimensional parameter $\boldsymbol{\gamma}_d$. The topic-word distributions are approximated by much simpler variational distributions $Q(\boldsymbol{\beta}_k|\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k)$ which are decomposable into 1-dimensional lognormals.

We now consider the log likelihood of the corpus $\mathcal{C}$ given the model $\{\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$.

$$\begin{aligned} \log P(\mathcal{C}|\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{d=1}^{M} \log P(\boldsymbol{w}_d|\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \sum_{d=1}^{M} \log \int d\boldsymbol{\theta}_d \int d\boldsymbol{\beta} \sum_{z_d} P(\boldsymbol{w}_d, \boldsymbol{z}_d, \boldsymbol{\theta}_d, \boldsymbol{\beta}|\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \sum_{d=1}^{M} \log \int d\boldsymbol{\theta}_d \int d\boldsymbol{\beta} \sum_{z_d} P(\boldsymbol{w}_d, \Xi|\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{Q(\Xi|\Lambda)}{Q(\Xi|\Lambda)}. \end{aligned}$$

Where we have denoted $\Xi = \{\boldsymbol{z}_d, \boldsymbol{\theta}_d, \boldsymbol{\beta}\}$, $\Lambda = \{\boldsymbol{\phi}_d, \boldsymbol{\gamma}_d, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\}$. By Jensen's inequality [28] we have

$$\begin{aligned}\log P(\mathcal{C}|\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &\geq \sum_{d=1}^{M} \int d\boldsymbol{\theta}_d \int d\boldsymbol{\beta} \sum_{z_d} Q(\Xi|\Lambda) \log \frac{P(\boldsymbol{w}_d, \Xi|\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{Q(\Xi|\Lambda)} \\ &\geq \sum_{d=1}^{M} \left[\boldsymbol{E}_Q \log P(\boldsymbol{w}_d, \Xi|\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}) - \boldsymbol{E}_Q \log Q(\Xi|\Lambda)\right]. \qquad (2)\end{aligned}$$

The task of the variational EM algorithm is to optimize the equation (2), i.e., to maximize the lower bound of the log likelihood. The algorithm alternates E-step and M-step until convergence. In the E-step, the algorithm tries to maximize the lower bound w.r.t variational parameters. Then for fixed values of variational parameters, the M-step maximizes the lower bound w.r.t model parameters. In summary, the EM algorithm for the DLN model is as follows.

- ***E-step****: maximize the lower bound in (2) w.r.t* $\boldsymbol{\phi}, \boldsymbol{\gamma}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$.
- ***M-step****: maximize the lower bound in (2) w.r.t* $\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.
- *Iterate these two steps until convergence.*

Note that DLN differs from LDA only in topic-word distributions. Thus $\boldsymbol{\phi}, \boldsymbol{\gamma}$, and α can be learnt as in [7], with a slightly different formula for $\boldsymbol{\phi}$.

$$\phi_{dni} \propto \left[\widehat{\mu}_{i\nu} - \log \sum_{t=1}^{V} \exp(\widehat{\mu}_{it} + \frac{1}{2}\widehat{\sigma}_{it}^2)\right] \exp\left(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{K} \gamma_{dj})\right) \qquad (3)$$

To complete the description of the learning algorithm for DLN, we next deal with the remaining variational parameters and model parameters. For the aim of clarity, we begin with the lower bound in (2).

$$\begin{aligned}\boldsymbol{E}_Q \log P(\boldsymbol{w}_d, \Xi|\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \boldsymbol{E}_Q \log P(\boldsymbol{w}_d|\boldsymbol{z}_d, \boldsymbol{\beta}) + \boldsymbol{E}_Q \log P(\boldsymbol{z}_d|\boldsymbol{\theta}_d) \\ &\quad + \boldsymbol{E}_Q \log P(\boldsymbol{\theta}_d|\alpha) + \boldsymbol{E}_Q \log P(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\end{aligned}$$

$$\begin{aligned}\boldsymbol{E}_Q \log Q(\Xi|\boldsymbol{\phi}_d, \boldsymbol{\gamma}_d, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) &= \boldsymbol{E}_Q \log Q(\boldsymbol{z}_d|\boldsymbol{\phi}_d) + \boldsymbol{E}_Q \log Q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \\ &\quad + \sum_{i=1}^{K} \boldsymbol{E}_Q \log Q(\boldsymbol{\beta}_i|\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)\end{aligned}$$

Thus the log likelihood now is

$$\log P(\mathcal{C}|\alpha,\boldsymbol{\mu},\boldsymbol{\Sigma}) \geq \sum_{d=1}^{M} \boldsymbol{E}_Q \log P(\boldsymbol{w}_d|\boldsymbol{z}_d,\boldsymbol{\beta}) - \sum_{d=1}^{M} \left[KL\left(Q(\boldsymbol{z}_d|\boldsymbol{\phi}_d)||P(\boldsymbol{z}_d|\boldsymbol{\theta}_d)\right) - KL\left(Q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)||P(\boldsymbol{\theta}_d|\alpha)\right)\right] - \sum_{d=1}^{M}\sum_{i=1}^{K} KL\left(Q(\boldsymbol{\beta}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i)||P(\boldsymbol{\beta}_i|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right) \tag{4}$$

Where $KL(\cdot||\cdot)$ is the Kullback-Leibler divergence of two distributions. Since $Q(\boldsymbol{z}_d|\boldsymbol{\phi}_d)$ and $P(\boldsymbol{z}_d|\boldsymbol{\theta}_d)$ are multinomial distributions, according to [18], we have

$$KL\left(Q(\boldsymbol{z}_d|\boldsymbol{\phi}_d)||P(\boldsymbol{z}_d|\boldsymbol{\theta}_d)\right) = \sum_{n=1}^{N_d}\sum_{i=1}^{K} \phi_{dni}\log\phi_{dni} - \sum_{n=1}^{N_d}\sum_{i=1}^{K}\phi_{dni}\left[\Psi(\gamma_{di}) - \Psi(\sum_{t=1}^{K}\gamma_{dt})\right] \tag{5}$$

Where $\Psi(\cdot)$ is the digamma function. Note that the first term is the expectation of $\log Q(\boldsymbol{z}_d|\boldsymbol{\phi}_d)$, and the second one is the expectation of $\log P(\boldsymbol{z}_d|\boldsymbol{\theta}_d)$ for which we have used the expectation of the sufficient statistics $\boldsymbol{E}_Q[\log\theta_{di}|\gamma_d] = \Psi(\gamma_{di}) - \Psi(\sum_{t=1}^{K}\gamma_{dt})$ for the Dirichlet distribution [7].

Similarly, for Dirichlet distributions as implicitly shown in [7],

$$\begin{aligned} &KL\left(Q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)||P(\boldsymbol{\theta}_d|\alpha)\right) = \\ &-\log\Gamma(\sum_{i=1}^{K}\alpha_i) + \sum_{i=1}^{K}\log\Gamma(\alpha_i) - \sum_{i=1}^{K}(\alpha_i-1)\left(\Psi(\gamma_{di}) - \Psi(\sum_{t=1}^{K}\gamma_{dt})\right) \\ &+\log\Gamma(\sum_{j=1}^{K}\gamma_{dj}) - \sum_{i=1}^{K}\log\Gamma(\gamma_{di}) + \sum_{i=1}^{K}(\gamma_{di}-1)\left(\Psi(\gamma_{di}) - \Psi(\sum_{t=1}^{K}\gamma_{dt})\right) \end{aligned} \tag{6}$$

By a simple transformation, we can easily show that the KL divergence of two lognormal distributions, $Q(\boldsymbol{\beta}|\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}})$ and $P(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, is equal to that of other normal distributions, $Q^*(\boldsymbol{\beta}|\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}})$ and $P^*(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. Hence using the KL divergence of two Normals as in [19], we obtain the divergence of two lognormal distributions.

$$\begin{aligned} &KL\left(Q(\boldsymbol{\beta}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i)||P(\boldsymbol{\beta}_i|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right) \\ &= \frac{1}{2}\log|\widehat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{\Sigma}_i| + \frac{1}{2}Tr\left((\widehat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{\Sigma}_i)^{-1}\right) - \frac{V}{2} + \frac{1}{2}(\widehat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\widehat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i) \end{aligned} \tag{7}$$

Where $Tr(A)$ is the trace of the matrix A.

The remaining term in (4) is the expectation of the log likelihood of the document $\boldsymbol{w}_d$. To find more detailed representations, we observe that, since $\boldsymbol{\beta}_i$

is a log-normally random variable,

$$\begin{aligned}
\boldsymbol{E}_Q \log \beta_{ij} &= \widehat{\mu}_{ij}, j \in \{1, ..., V\} \\
\boldsymbol{E}_Q \log \sum_{t=1}^{V} \beta_{it} &= \log \exp \left(\boldsymbol{E}_Q \log \sum_{t=1}^{V} \beta_{it} \right) && (8) \\
&\leq \log \boldsymbol{E}_Q \sum_{t=1}^{V} \beta_{it} && (9) \\
&\leq \log \sum_{t=1}^{V} \exp(\widehat{\mu}_{it} + \widehat{\sigma}_{it}^2/2) && (10)
\end{aligned}$$

Note that the inequality (9) has been derived from (8) using Jensen's inequality. The last inequality (10) is simply another form of (9), replacing the expectations of individual variables by their detailed formulas [13].

From those observations, we have

$$\begin{aligned}
&\boldsymbol{E}_Q \log P(\boldsymbol{w}_d | \boldsymbol{z}_d, \boldsymbol{\beta}) \\
&= \sum_{n=1}^{N_d} \boldsymbol{E}_Q \log P(w_{dn} | z_{dn}, \boldsymbol{\beta}) && (11) \\
&= \sum_{n=1}^{N_d} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{dni} w_{dn}^j \boldsymbol{E}_Q \left[\log \beta_{ij} - \log \sum_{t=1}^{V} \beta_{it} \right] && (12) \\
&\geq \sum_{n=1}^{N_d} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{dni} w_{dn}^j \left[\widehat{\mu}_{ij} - \log \sum_{t=1}^{V} \exp(\widehat{\mu}_{it} + \widehat{\sigma}_{it}^2/2) \right] && (13)
\end{aligned}$$

There is a little strange in the right-hand side of (12) resulting from (11). The reason is that in DLN each topic $\boldsymbol{\beta}_i$ has to be transformed by the mapping $f(\cdot)$ into parameters of the multinomial distribution. Hence the derived formula is more complicated than that of LDA.

A lower bound of the log likelihood of the corpus $\mathcal{C}$ is finally derived from combining (4), (5), (6), (7), and (13). We next have to incorporate this lower bound into the variational EM algorithm for DLN by describing how to maximize the lower bound with respect to the parameters.

Variational parameters:

First, we would like to maximize the lower bound by variational parameters, $\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$. Note that the term containing $\widehat{\boldsymbol{\mu}}_i$ for each $i \in \{1, ..., K\}$ is

$$\begin{aligned}
\mathcal{L}[\widehat{\boldsymbol{\mu}}_i] = &- \frac{M}{2} (\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) \\
&+ \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{j=1}^{V} \phi_{dni} w_{dn}^j \left[\widehat{\mu}_{ij} - \log \sum_{t=1}^{V} \exp(\widehat{\mu}_{it} + \widehat{\sigma}_{it}^2/2) \right].
\end{aligned}$$

Since log-sum-exp functions are convex in their variables [9], $\mathcal{L}[\widehat{\boldsymbol{\mu}}_i]$ is a concave function in $\widehat{\boldsymbol{\mu}}_i$. Therefore, we can use convex optimization methods to

maximize $\mathcal{L}[\widehat{\boldsymbol{\mu}}_i]$. In particular, we use LBFGS [17] to find the maximum of $\mathcal{L}[\widehat{\boldsymbol{\mu}}_i]$ with the following partial derivatives

$$\frac{\partial \mathcal{L}}{\partial \widehat{\mu}_{ij}} = -M\boldsymbol{\Sigma}_{ij}^{-1}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) + \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j} - \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}\frac{\exp(\widehat{\mu}_{ij}+\widehat{\sigma}_{ij}^2/2)}{\sum_{t=1}^{V}\exp(\widehat{\mu}_{it}+\widehat{\sigma}_{it}^2/2)}$$

Where $\boldsymbol{\Sigma}_{ij}^{-1}$ is the jth row of $\boldsymbol{\Sigma}_i^{-1}$.

The term in the lower bound of (4) that contains $\widehat{\boldsymbol{\Sigma}}_i$ for each i is

$$\mathcal{L}[\widehat{\boldsymbol{\Sigma}}_i] \quad = \quad \frac{M}{2}\log|\widehat{\boldsymbol{\Sigma}}_i| - \frac{M}{2}Tr(\boldsymbol{\Sigma}_i^{-1}\widehat{\boldsymbol{\Sigma}}_i) - \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}\log\sum_{t=1}^{V}\exp(\widehat{\mu}_{it}+\widehat{\sigma}_{it}^2/2)$$

We use LBFGS-B [35] to find its maximum subject to the constraints $\widehat{\sigma}_{ij}^2 > 0, \forall j \in \{1,...,V\}$, with the following derivatives

$$\frac{\partial \mathcal{L}}{\partial \widehat{\sigma}_{ij}^2} = \frac{M}{2\widehat{\sigma}_{ij}^2} - \frac{M}{2}\sigma_{ij}^{-2} - \frac{1}{2}\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}\frac{\exp(\widehat{\mu}_{ij}+\widehat{\sigma}_{ij}^2/2)}{\sum_{t=1}^{V}\exp(\widehat{\mu}_{it}+\widehat{\sigma}_{it}^2/2)}$$

Where σ_{ij}^{-2} is the jth element on the diagonal of $\boldsymbol{\Sigma}_i^{-1}$.

Model parameters:

We now want to maximize the lower bound of (4) with respect to the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, for the M-step of the variational EM algorithm. The term containing $\boldsymbol{\mu}_i$ for each i is

$$\mathcal{L}[\boldsymbol{\mu}_i] = -\frac{M}{2}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)$$

The maximum of this function is reached at

$$\boldsymbol{\mu}_i = \widehat{\boldsymbol{\mu}}_i \tag{14}$$

The term containing $\boldsymbol{\Sigma}_i^{-1}$ that is to be maximized is

$$\begin{aligned}\mathcal{L}[\boldsymbol{\Sigma}_i^{-1}] \quad = \quad & \frac{M}{2}\log|\boldsymbol{\Sigma}_i^{-1}| - \frac{M}{2}Tr(\boldsymbol{\Sigma}_i^{-1}\widehat{\boldsymbol{\Sigma}}_i) \\ & -\frac{M}{2}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\end{aligned}$$

And its derivative is

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_i^{-1}} = \frac{M}{2}\boldsymbol{\Sigma}_i - \frac{M}{2}\widehat{\boldsymbol{\Sigma}}_i - \frac{M}{2}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T$$

Setting this to 0, we can find the maximum point:

$$\boldsymbol{\Sigma}_i \quad = \quad \widehat{\boldsymbol{\Sigma}}_i + (\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T \tag{15}$$

We have derived how to maximize the lower bound of the log likelihood of the corpus $\mathcal{C}$ in (2) with respect to the variational parameters and model parameters. The variational EM algorithm now proceeds by maximizing the lower bound w.r.t $\boldsymbol{\phi}, \boldsymbol{\gamma}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$ under the fixed values of the model parameters, and then by maximizing w.r.t $\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ under the fixed values of variational parameters. Iterate these two steps until convergence. In our experiments, the convergence criterion is that the relative change of the log likelihood was no more than 10^{-4}.

For inferences on each new document, we can use the same iterative procedure as described in [7] using the formula (3) for $\boldsymbol{\phi}$. The convergence threshold for the inferences of each document was 10^{-6}.

References

[1] Deepak Agarwal and Bee-Chung Chen. fLDA: matrix factorization through latent dirichlet allocation. In *The third ACM International Conference on Web Search and Data Mining*, pages 91–100. ACM, 2010.

[2] David Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. Springer Berlin / Heidelberg, 1985.

[3] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *The 26th International Conference on Machine Learning (ICML)*, 2009.

[4] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[5] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[6] David M. Blei and John Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[8] D.M. Blei and M.I. Jordan. Modeling annotated data. In *The 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134. ACM, 2003.

[9] Mung Chiang. Geometric programming for communication systems. *Foundations and Trends in Communications and Information Theory*, 2(1-2): 1–153, 2005.

[10] Chris Ding. A probabilistic model for latent semantic indexing. *Journal of the American Society for Information Science and Technology*, 56(6): 597–608, 2005.

[11] Gabriel Doyle and Charles Elkan. Accounting for burstiness in topic models. In *The 26th International Conference on Machine Learning (ICML)*, 2009.

[12] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[13] Christian Kleiber and Samuel Kotz. *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley-Interscience, 2003.

[14] Thomas Landauer and Susan Dumais. A solution to platos problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[15] Ackhard Limpert, Werner A. Stahel, and Markus Abbt. Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352, may 2001.

[16] B. Liu, L. Liu, A. Tsykin, G.J. Goodall, J.E. Green, M. Zhu, C.H. Kim, and J. Li. Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, 26(24):3105, 2010.

[17] Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

[18] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *CoRR*, abs/0911.4863, 2009.

[19] Frank Nielsen and Richard Nock. Clustering multivariate normal distributions. In *Emerging Trends in Visual Computing*, number 5416 in LNCS, pages 164–174. Springer-Berlin / Heidelberg, 2009.

[20] D. Putthividhya, H. T. Attias, and S. Nagarajan. Independent factor topic models. In *The 26th International Conference on Machine Learning (ICML)*, 2009.

[21] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*, 2010.

[22] Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

[23] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[24] Francis E.H Tay and Lijuan Cao. Application of support vector machines in financial time series forecasting. *Omega*, 29(4): 309 – 317, 2001. doi: 10.1016/S0305-0483(01)00026-3. URL `http://www.sciencedirect.com/science/article/pii/S0305048301000263`.

[25] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[26] Khoat Than, Tu Bao Ho, Duy Khuong Nguyen, and Ngoc Khanh Pham. Supervised dimension reduction with topic models. In *ACML*, volume 25 of *Journal of Machine Learning Research: W&CP*, pages 395–410, 2012.

[27] Flora S. Tsai. A tag-topic model for blog mining. *Expert Systems with Applications*, 38(5):5330–5335, 2011.

[28] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[29] Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking lda: why priors matter. In *Neural Information Processing Systems (NIPS)*, 2009.

[30] K.W. Wan, A.H. Tan, J.H. Lim, and L.T. Chia. A non-parametric visual-sense model of images-extending the cluster hypothesis beyond text. *Multimedia Tools and Applications*, pages 1–26, 2010.

[31] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. In *The 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

[32] Chong Wang, Bo Thiesson, Christopher Meek, and David M. Blei. Markov topic models. In *Neural Information Processing Systems (NIPS)*, 2009.

[33] X. Wei and W.B. Croft. LDA-based document models for ad-hoc retrieval. In *The 29th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185. ACM, 2006.

[34] J. Weng, E.P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *The third ACM International Conference on Web Search and Data Mining*, pages 261–270. ACM, 2010.

[35] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, 1997. ISSN 0098-3500. doi: http://doi.acm.org/10.1145/279232.279236.