Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Topic model for graph mining based on hierarchical Dirichlet process

Haibin Zhang, Shang Huating & Xianyi Wu

To cite this article: Haibin Zhang, Shang Huating & Xianyi Wu (2020) Topic model for graph mining based on hierarchical Dirichlet process, Statistical Theory and Related Fields, 4:1, 66-77, DOI: 10.1080/24754269.2019.1593098

To link to this article: https://doi.org/10.1080/24754269.2019.1593098

Published online: 27 Mar 2019.

STATISTICAL THEORY AND RELATED FIELDS
2020, VOL. 4, NO. 1, 66–77
https://doi.org/10.1080/24754269.2019.1593098

Taylor & Francis
Taylor & Francis Group

Topic model for graph mining based on hierarchical Dirichlet process

Haibin Zhang, Shang Huating and Xianyi Wu

East China Normal University, Shanghai, People's Republic of China

ABSTRACT

In this paper, a nonparametric Bayesian graph topic model (GTM) based on hierarchical Dirichlet process (HDP) is proposed. The HDP makes the number of topics selected flexibly, which breaks the limitation that the number of topics need to be given in advance. Moreover, the GTM releases the assumption of 'bag of words' and considers the graph structure of the text. The combination of HDP and GTM takes advantage of both which is named as HDP–GTM. The variational inference algorithm is used for the posterior inference and the convergence of the algorithm is analysed. We apply the proposed model in text categorisation, comparing to three related topic models, latent Dirichlet allocation (LDA), GTM and HDP.

ARTICLE HISTORY
Received 21 October 2018
Revised 5 March 2019
Accepted 7 March 2019

KEYWORDS
Graph topic model; hierarchical Dirichlet process; variational inference; text classification

1. Introduction

We are entering the era of big data. It becomes more difficult for people to find valuable information from the explosion of document archives. New techniques or tools need to be used to deal with automatically organising, searching and understanding large collections. Topic modelling provides a convenient way to analyse big unstructured text (Miner et al., 2012). A topic model is a kind of a probabilistic generative model that has been used widely in the field of computer science text mining and information retrieval in recent years (Gupta & Lehal, 2009).

The origin of the probabilistic topic model is latent semantic indexing (LSI) (Deerwester, 1990); it has been the basis for the development of the topic model. Based on LSI, probabilistic latent semantic analysis (PLSA) (Hofmann, 1999) was proposed by Hofmann and is probabilistic topic modelling. Latent Dirichlet allocation (LDA) proposed by Blei, Ng, and Jordan (2003) is a method using probabilistic generative models. The basic assumption of LDA is 'bag of word' which means the order of the words can be neglected. Recently, there are many methods using probabilistic topic models that are based on LDA via combination with particular tasks and relaxing the assumption of LDA (Blei, 2012). In many text analysis settings, a document is composed of the words which serve as nodes, and the relations between words serve as edges (Valle, 2011). The relations may be co-occurrence relations, association relations or other semantic relations. Then, the text is expressed as graph structure data. The research of graph structure data belongs to the field of graph mining. The idea of graph mining to the topic model can improve the accuracy of classification or clustering of text graph structure data compared with text mining which only considers nodes but ignores the relationship between them. In addition, many practical tasks can all benefit from this kind of graph mining such as products, services and website retrieval. Although topic models have proved to be very successful in discovering latent topics, the standard topic models cannot be directly applied to graph-structured data because of the 'bag-of-word' assumption. Xuan, Lu, Zhang, and Luo (2015) proposed a method using topic model for graph mining (GTM) which assumes that there is an edge between two nodes in a graph, these two nodes tend to talk similar content.

In LDA, the topics are fixed for the whole corpus, and the number of topics is assumed to be known in advance. However, it is usually hard to make such a decision without a deep knowledge of the data set. Recent advances in Bayesian nonparametric modelling, specifically hierarchical Dirichlet process (HDP) (Gershman & Blei, 2012; Teh, Jordan, Beal, & Blei, 2006), has lead to 'infinite' topic models. The number of topics does not need to be specified and is determined by collection during the posterior inference and furthermore, new documents can exhibit previously unseen topics. This paper proposes a method using graph topic model based on hierarchical Dirichlet process (HDP-GTM). HDP-GTM realises the flexible selection of topic numbers by the property of HDP, breaking through the limitation that classical topic models in advance. At the same time, this method breaks through the limitation of the assumption 'bag of word' and takes the graph structure of the text into account which can make full use of data information, thus the text classification accuracy can be significantly improved.

CONTACT Xianyi Wu xywu@stat.ecnu.edu.cn

Exactly computing posterior distributions for the HDP-GTM is very intractable. We propose a variational inference algorithm for the HDP-GTM. Variational inference (VI) is a method from machine learning for approximating probability densities (Blei, Kucukelbir, & McAuliffe, 2017; Jordan, Ghahramani, Jaakkola, & Saul, 1999; Wainwright & Jordan, 2008). Variational inference is widely used to approximate posterior densities for Bayesian models, an alternative strategy to Markov chain Monte Carlo (MCMC) sampling. Compared to MCMC, variational inference tends to be faster and easier to scale to large datasets. It has been applied to problems such as large-scale document analysis, computational neuroscience and computer vision. But variational inference has been studied less rigorously than MCMC, and its statistical properties are less well understood. In this paper, we apply a variational inference algorithm for calculating the posterior distribution and investigate its convergence property.

The structure of this paper is as follows. We provide a brief introduction of the related models LDA, HDP and GTM in Section 2. Then, in Section 3, we propose HDP–GTM based on HDP and GTM. The posterior inference is derived by the variational inference procedure and the convergence of variational inference algorithm is verified in Section 4. Finally, in Section 5, experiments are conducted to compare the performance of the HDP–GTM with LDA, HDP and GTM on the Reuter dataset and the 20-newsgroup dataset. Section 6 concludes the paper with some concluding remarks.

2. Related work

This section will provide more details about LDA, GTM, and HDP respectively.

2.1. Latent Dirichlet allocation

First, we review the underlying statistical assumptions of the LDA. The LDA is a method using three-level hierarchical Bayesian model which includes three levels of documents, topics and words. LDA adds a priori distribution of document-topic level and topic-word level. LDA assumes that the probability of document-topic is $p(z \mid d)$ extracted from Dirichlet distribution $\mathrm{Dir}(\alpha)$ and the probability of topic-word is $p(w \mid z)$ extracted multinomial distribution respectively. The Dirichlet distribution and multinomial distribution are conjugate prior distributions, which can simplify the calculation of the posterior distribution. The graphical representation of LDA is shown in Figure 1 and each document is assumed drawn from the following generative process:

(1) Draw the topic proportion $\theta_d \overset{\text{i.i.d.}}{\sim} \mathrm{Dir}(\alpha)$ for each document;

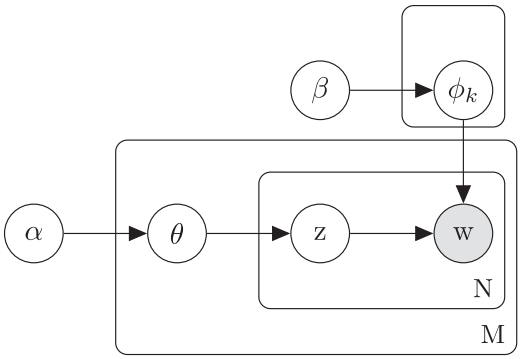

Figure 1. Graphical model representation of LDA.

(2) Draw $\phi_k \overset{\text{i.i.d.}}{\sim} \mathrm{Dir}(\beta)$ for each topic;
(3) For each word n in the document d:
   (a) Draw topic assignment $z_{dn} \overset{\text{i.i.d.}}{\sim} \mathrm{Multi}(\theta_d)$;
   (b) Draw $w_{dn} \mid \{\phi\}_{k=1}^{\infty}, z_{dn} \sim \mathrm{Multi}(\phi_{z_{dn}})$ for each word.

As the figure makes clear, there are three levels to the LDA representation. The parameters α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document. It is difficult to compute the posterior distribution directly, two commonly used methods are variational EM algorithm which aims at looking to the approximate Bayesian estimate of $p(w \mid \alpha, \beta)$ (Blei et al. 2003) and the other method is Gibbs sampling (Griffiths & Steyvers, 2004).

LDA considers the prior distribution of parameters, and the topic distribution is no longer fixed, but obtained by sampling. However, LDA is based on the 'bag of words' assumption, regardless of the order of words, and regardless of the correlation between words. In this way, some important information in the text will be omitted, resulting in information loss.

2.2. Graph topic model

The traditional topic models are based on the assumption of 'bag of words', that is, words in text are independent and exchangeable (Aldous, 1985). Exchangeability means that we don't consider the order in which the locations of words appear in the text, nor do we consider the association between words. Although topic model has been popular in the field of text mining and information retrieval, the research on topic mining of graph structure text data is insufficient. Xuan, Lu, Zhang, and Luo (2015) proposed a method using probabilistic topic model based on graph structure data which was inspired by graph structure text data.

GTM is an extension of LDA which the graph structure of text data is considered. Although the traditional

topic model is considered very successful, traditional topic models cannot be directly applied to graph structure text data because of the 'bag of word' assumption of the topic model. In the graph topic model, w_{dn} indicates a node in the document d and $e_{w_{di},w_{dj}}$ in the document d indicates the edge between the node w_{di} and the node w_{dj} corresponding to the relationship between words. Applying GTM to text data requires adding an edge parameter e in LDA model. Assuming that two nodes describe similar topics, it is considered that there is an edge between these two nodes. The edge is a parameter that describes the topic similarity of the nodes, therefore, $e_{w_{di},w_{dj}}$ can be generated by the topic distribution of node w_{di} and node w_{dj}.

The graphical representation of GTM is shown in Figure 2 and the corresponding generative process is as follows.

(1) Draw topic proportion $\theta_d \overset{\text{i.i.d.}}{\sim} \text{Dir}(\alpha)$ for each document
(2) Draw $\phi_k \overset{\text{i.i.d.}}{\sim} \text{Dir}(\beta)$ for each topic
(3) Draw topic assignment $z_{dn} \overset{\text{i.i.d.}}{\sim} \text{Multi}(\theta_d)$
(4) Draw $w_{dn}|\{\phi\}_{k=1}^{\infty}, z_{dn} \sim \text{Multi}(\phi_{z_{dn}})$ for each word node
(5) For all edges in a graph: draw $e_{w_{di},w_{dj}} \sim \text{Bernoulli}(p_{w_{di},w_{dj}})$ for each edge, where

$$p_{w_{di},w_{dj}}(e_{w_{di},w_{dj}} = 1) = \phi_{z_{di}} \cdot \phi_{z_{dj}} \tag{1}$$

$e_{w_{di},w_{dj}}$ is the edge between word w_{di} and word w_{dj} which is extracted from Bernoulli distribution, and $p_{w_{di},w_{dj}}$ is the parameter of Bernoulli distribution.

From Figure 2, $e_{w_{di},w_{dj}}$ is generated from $\{z_{di}, z_{dj}, \Phi\}$, so it means that the probability of the existence of an edge between two nodes is determined by the similarity of their topic distributions. This similarity is measured by vector inner product of $\phi_{z_{di}}$ and $\phi_{z_{dj}}$, where $\phi_{z_{di}}$ is node distribution of topic z_{di}, as shown in Equation (1). The more similar topics of two keywords are, the more likely there is an edge between these two nodes. Different from the 'bag of words' assumption of the traditional topic model, Bernoulli distribution is used to model the relationship between two keywords, and the similarity between two theme distributions of two keywords is parameterised in GTM. By considering the relationship between key words in the document, GTM is verified better than LDA through the experimental results of text classification GTM is derived from LDA, and both of these models belong to Bayesian models. Similar to LDA, the deficiency of GTM model lies in that the number of topics needs to be given in advance, it is difficult to make such a decision without a deep knowledge of the dataset. HDP is a Bayesian nonparametric method for modelling topic model, the number of topics does not need to be specified in advance and is determined by collection during posterior inference.

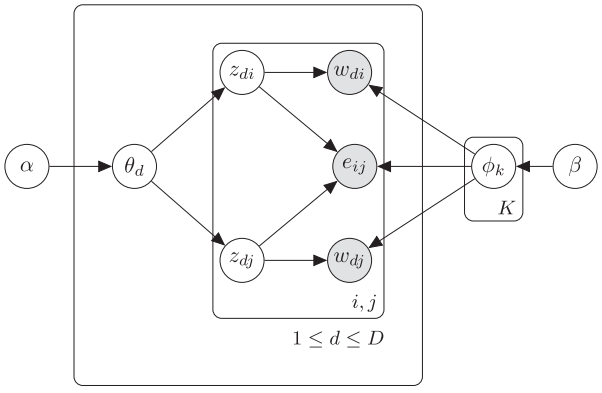

Figure 2. Graphical model representation of GTM.

2.3. Hierarchical Dirichlet process

HDP is useful in problems in which there are multiple groups of data (Teh et al., 2006; Wang et al. 2011) . The HDP is a hierarchical extension to Dirichlet process (DP) (Blackwell & MacQueen, 1973; Ferguson, 1973). The hierarchical structure provides an elegant way of sharing parameters. Mathematically,

$$\begin{aligned} G_0 &\sim \text{DP}(\gamma, H), \\ G_d &\sim \text{DP}(\alpha_0, G_0), \text{ for each document d,} \end{aligned} \tag{2}$$

where H is the base probability measure, γ and α_0 are concentration parameters. The distribution G_0 varies around H by an amount controlled by γ and the distribution G_d in group d varies around G_0 by an amount controlled by α_0.

The HDP can be seen as adding another level of smoothing on top of DP mixture models. For each d, let $\theta_{d1}, \theta_{d2}, \ldots$ be iid random variables distributed as G_d. Each θ_{dn} is a factor corresponding to a single observation w_{dn}. The likelihood is given by

$$\begin{aligned} \theta_{dn} \,|\, G_d &\sim G_d, \\ w_{dn} \,|\, \theta_{dn} &\sim \text{Multi}(\theta_{dn}). \end{aligned} \tag{3}$$

Sticking-breaking priors are rich and important class of random measures in Bayesian Nonparametric which origins in Ferguson (1973) and Sethuraman (1994) proposed infinite mixture representation of the Dirichlet process. HDP can be constructed by stick-breaking method which was shown in Teh, Jordan, Beal and Blei (2006). We describe an alternative stick-breaking construction for the HDP proposed by Wang, Paisley and Blei (2011). For the corpus-level DP draw, this representation is

$$\begin{aligned} \beta'_k &\sim \text{Beta}(1, \gamma), \\ \beta_k &= \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell), \end{aligned}$$

$$\phi_k \overset{\text{i.i.d.}}{\sim} H,$$
$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}. \tag{4}$$

Thus G_0 is discrete and has support at the atoms $\boldsymbol{\phi} = (\phi_k)_{k=1}^{\infty}$ with weights $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$. The distribution for $\boldsymbol{\beta}$ is written as $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$ which stands for Griffiths, Engen, and McCloskey (Pitman 2002). The construction of each document-level G_j by Sethuraman's stick-breaking construction is

$$\pi'_{dt} \sim \text{Beta}(1, \alpha),$$
$$\pi_{dt} = \pi'_{dt} \prod_{\ell=1}^{t-1} (1 - \pi'_{d\ell}),$$
$$\pi_{dt} \overset{\text{i.i.d.}}{\sim} G_0,$$
$$G_d = \sum_{t=1}^{\infty} \pi_{dt} \delta_{\psi_t}. \tag{5}$$

Notice that each document-level atom ψ_{dt} maps to a corpus-level atom ϕ_k in G_0 according to the distribution defined by G_0. The distribution for $\boldsymbol{\pi}_d$ is also written as $\boldsymbol{\pi}_d \sim \text{GEM}(\alpha)$. Let $\boldsymbol{c}_d = (c_{dt})_{t=1}^{\infty}$ be a series of indicator variables which are drawn i.i.d.,

$$c_{dt} \overset{\text{i.i.d.}}{\sim} \text{Mult}(\boldsymbol{\beta}), \tag{6}$$

where $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$. Then let

$$\psi_{dt} = \phi_{c_{dt}}. \tag{7}$$

Thus we do not need to explicitly represent the document atoms $\boldsymbol{\psi}_j$. The property that multiple document-level atoms ψ_{dt} can map to the same corpus-level atom ϕ_k in this representation which is similar in spirit to the Chinese restaurant franchise (CRF) (Teh et al., 2006), where each restaurant can have multiple tables serving the same dish ϕ_k. In the CRF representation, a hierarchical Chinese restaurant process allocates dishes to tables. Here, we use a series of random indicator variables $\mathbf{c}_d$ to represent this structure. Given the representation in Sethuraman's stick-breaking construction, the generative process for the observed words in the dth document is as follows:

$$z_{dn} \overset{\text{i.i.d.}}{\sim} \text{Mult}(\boldsymbol{\pi}_d),$$
$$\theta_{dn} = \psi_{dz_{dn}} = \phi_{c_{dz_{dn}}},$$
$$w_{dn} \overset{\text{i.i.d.}}{\sim} \text{Mult}(\theta_{dn}). \tag{8}$$

The indicator z_{dn} selects topic parameter ψ_{dt}, which maps to one topic ϕ_k through the indicators c_{dt}. The graphical representation of GTM is shown in Figure 3 and HDP topic model is described below:

(1) Draw $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$ and $\boldsymbol{\pi}_d \sim \text{GEM}(\alpha)$

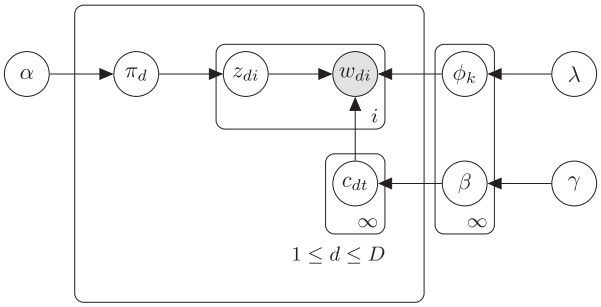

Figure 3. Graphical model representation of the HDP.

(2) Draw $\phi_k \overset{\text{i.i.d.}}{\sim} \text{Dir}(\lambda_k)$
(3) For each word w_{dn} in the document d where $n \in \{1, 2, \ldots, N_d\}$ and $t \in \{1, 2 \cdots, \}$,
  (a) Draw $c_{dt} \overset{\text{i.i.d.}}{\sim} \text{Mult}(\boldsymbol{\beta})$ and $z_{dn} \overset{\text{i.i.d.}}{\sim} \text{Mult}(\boldsymbol{\pi}_d)$;
  (b) Draw $w_{dn} \mid \{\phi_k\}_{k=1}^{\infty}, z_{dn}, c_{dt} \sim \text{Mult}(\phi_{c_{dz_{dn}}})$.

In an HDP, the number of topics does not need to be specified in advance and is determined by collection during posterior inference and furthermore, new documents can exhibit previously unseen topics. However, HDP cannot take the relationship of the words into consideration. The new model which combines GTM with HDP is required.

3. HDP–GTM

This section will introduce the proposed method HDP–GTM. HDP–GTM is an extension of GTM which includes the co-occurrence relationship between words. Besides, HDP–GTM considers the HDP, so that the number of topics is no longer fixed but determined flexibly according to the document itself. HDP–GTM suppose that the topic distribution of the text is extracted from an HDP rather than a Dirichlet distribution whose dimension is fixed. The graph model representation of HDP–GTM is presented in Figure 4 and the generative HDP–GTM is described below:

(1) Draw $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$ and $\boldsymbol{\pi}_d \sim \text{GEM}(\alpha)$
(2) Draw $\phi_k \overset{\text{i.i.d.}}{\sim} \text{Dir}(\lambda)$
(3) For each word w_{dn} in the document d where $n \in \{1, 2, \ldots, N_d\}$ and $t \in \{1, 2 \cdots\}$,
  (a) Draw $c_{dt} \overset{\text{i.i.d.}}{\sim} \text{Mult}(\boldsymbol{\beta})$ and $z_{dn} \overset{\text{i.i.d.}}{\sim} \text{Mult}(\boldsymbol{\pi}_d)$;
  (b) Draw $w_{dn} \mid \{\phi_k\}_{k=1}^{\infty}, z_{dn}, c_{dt} \sim \text{Mult}(\phi_{c_{dz_{dn}}})$;
  (c) $e_{w_{di}, w_{dj}}$ is the edge between word w_{di} and word w_{dj} drawn from the Bernoullii distribution with parameter $\phi_{c_{dz_{di}}} \cdot \phi_{c_{dz_{dj}}}$.

Then, we have

$$p(e_{w_{di}, w_{dj}} = 1) = \phi_{c_{dz_{di}}} \cdot \phi_{c_{dz_{dj}}}.$$

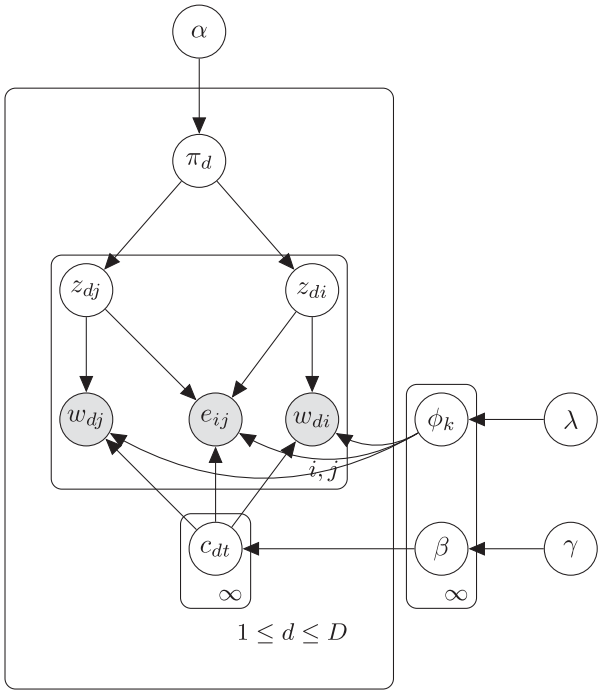

Figure 4. Graphical model representation of HDP–GTM.

The principle of edge generation in HDP–GTM is consistent with GTM, assuming that there is association between words with similar topic distribution, and the similarity of topic distribution is determined by the inner product of the topic distribution of corresponding words.

In the HDP–GTM, the joint likelihood of the whole corpus is

$$
\begin{aligned}
&P(\mathbf{W}, \mathbf{E}, \mathbf{Z}, \boldsymbol{\Theta} \mid \alpha, \gamma, \lambda) \\
&= \left\{ \prod_{d=1}^{D} \prod_{k=1}^{\infty} \prod_{t=1}^{\infty} P(\boldsymbol{\pi}_d \mid \alpha) P(c_{dt} \mid \boldsymbol{\beta}) P(\beta'_k \mid \gamma) P(\phi'_k \mid \lambda) \right. \\
&\quad \times \left[\prod_{n=1}^{N_d} P(w_{dn} \mid z_{dn}, \{\phi_k\}_{k=1}^{\infty}, c_{dt}) P(z_{dn} \mid \boldsymbol{\pi}_d) \right] \\
&\quad \left. \times \prod_{i,j} [P(e_{w_{di}, w_{dj}})] \right\}, \qquad (9)
\end{aligned}
$$

where $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{C})$, $\mathbf{W} = \{w_{dn}\}$, $\mathbf{E} = \{e_{w_{di}, w_{dj}}\}$, $\mathbf{Z} = \{z_{dn}\}$, $\boldsymbol{\pi} = \{\pi'_d\}$, $\boldsymbol{\beta} = \{\beta'_k\}$ and $\boldsymbol{c} = \{c_{dt}\}$.

4. Posterior inference for the HDP–GTM

In this section, we turn our attention to the procedures of posterior inference of the HDP–GTM. The key inferential problem we need to solve is the posterior distribution of latent variables given documents when we use probabilistic topic models. The posterior distribution for the HDP–GTM topic model is

$$P(\mathbf{Z}, \boldsymbol{\Theta} \mid \mathbf{W}, \mathbf{E}). \qquad (10)$$

For nonparametric models, exact inference of the model parameters is intractable in general. In order to solve this problem, we discuss variational Bayes (VB) inference methods for the HDP–GTM. VB presents an efficient approximate method for the true posterior $P(\mathbf{Z}, \boldsymbol{\Theta} \mid \mathbf{W}, \mathbf{E})$ by minimising $KL(Q \,||\, P)$. Q is called variational distribution which is a simplification of the true posterior $P(\mathbf{Z}, \boldsymbol{\Theta} \mid \mathbf{W}, \mathbf{E})$. We propose the following factorised family of variational distributions for mean-field variational inference:

$$
\begin{aligned}
Q(\mathbf{Z}, \boldsymbol{\Theta}) = &\prod_{k=1}^{K-1} q(\beta'_k \mid a_{k1}, a_{k2}) \prod_{k=1}^{K} q(\phi_k \mid \kappa_k) \\
&\times \prod_{d=1}^{D} \left[\prod_{t=1}^{\mathrm{T}} q(c_{dt} \mid \eta_{dt}) \prod_{t=1}^{T-1} q(\pi'_{dt} \mid \gamma_{dt1}, \gamma_{dt2}) \right. \\
&\left. \times \prod_{n=1}^{N} q(z_{dn} \mid \zeta_{dn}) \right], \qquad (11)
\end{aligned}
$$

where

$$
\begin{aligned}
&\beta'_k \sim \mathrm{Beta}(a_{k1}, a_{k2}), \quad \phi_k \sim \mathrm{Dir}(\kappa_k), \\
&c_{dt} \sim \mathrm{Multi}(\eta_{dt}), \ \pi'_{dt} \sim \mathrm{Beta}(\gamma_{dt1}, \gamma_{dt2}), \\
&z_{dn} \sim \mathrm{Multi}(\zeta_{dn}). \qquad (12)
\end{aligned}
$$

Let Σ be the set of the parameters in Equation (12), then

$$
\begin{aligned}
\mathrm{KL}(Q \,||\, P) &= E_Q[\log Q(\mathbf{Z}, \boldsymbol{\Theta})] - E_Q[\log P(\mathbf{Z}, \boldsymbol{\Theta} \mid \mathbf{W}, \mathbf{E}) \\
&= E_Q[\log Q(\mathbf{Z}, \boldsymbol{\Theta})] - E_Q[\log P(\mathbf{Z}, \boldsymbol{\Theta}, \mathbf{W}, \mathbf{E})] \\
&\quad + \log P(\mathbf{W}, \mathbf{E}). \qquad (13)
\end{aligned}
$$

Let

$$L(\Sigma) = E_Q[\log P(\mathbf{Z}, \boldsymbol{\Theta}, \mathbf{W}, \mathbf{E})] - E_Q[\log Q(\mathbf{Z}, \boldsymbol{\Theta})], \qquad (14)$$

where $L(\Sigma)$ is called the evidence lower bound (ELBO) which is dependent on the variational parameters Σ. The goal is to minimise the objective function KL-divergence which is equivalent to maximise the evidence lower bound function. We use truncations for inferencing the number of factors at T and K. Having specified a simplified family of probability distributions, the next step is to set up an optimisation problem that determines the values of the variational parameters. The derivation details can be found in Appendix 1. For document-level updates, at the document level we update the parameters to the per-document stick, and the parameters are

$$
\begin{aligned}
a_{k1} &= 1 + \sum_{d=1}^{D} \sum_{t=1}^{\mathrm{T}} \eta_{dtk}, \\
a_{k2} &= \gamma + \sum_{d=1}^{D} \sum_{t=1}^{\mathrm{T}} \sum_{f=k+1}^{K} \eta_{dtk},
\end{aligned}
$$

$$\eta_{dtk} = \exp\left\{\sum_{e=1}^{k-1}(\Psi(a_{e2}) - \Psi(a_{e1} + a_{e2})) + (\Psi(a_{k1}) - \Psi(a_{k1} + a_{k2})) + \sum_{n=1}^{N}\sum_{v=1}^{V} w_{dn}^{v}\zeta_{dnt}\left(\Psi(\lambda_{kv}) - \Psi\left(\sum_{l=1}^{V}\lambda_{kl}\right)\right)\right\}. \quad (15)$$

There exists no closed-form solution for ζ_{dtk}. We can derive Iterative expression of ζ_{dtk} by the Newton's method.

$$\zeta_{dtk}^{(n+1)} = \zeta_{dtk}^{(n)} - H(\zeta_{dtk}^{(n)})^{-1}\Delta_{\zeta_{dtk}^{(n)}}L(\zeta_{dtk}^{(n)}), \quad (16)$$

where $f(\zeta_{dtk}^{(n)})$ and $H(\zeta_{dtk}^{(n)})$ are provided in Appendix 1. At the corpus level, we update the parameters to top-level sticks and the topics,

$$\gamma_{dt1} = 1 + \sum_{n=1}^{N}\zeta_{dnt},$$

$$\gamma_{dt2} = \alpha_0 + \sum_{n=1}^{N}\sum_{b=t+1}^{T}\zeta_{dnt},$$

$$\kappa_{jv} = \beta_{jv} + \sum_{d=1}^{D}\sum_{n=1}^{N_d} w_{dn}^{v}\zeta_{dnt}\eta_{dtk}. \quad (17)$$

For each iterative step, define

$$\Sigma^{(n)} = M(\Sigma^{(n-1)}), \quad (18)$$

for $n = 1, \ldots,$ where M denotes a certain mapping and $\{\Sigma^{(n)}\}$ denotes the sequences of iterates that are produced. To verify the convergence of Algorithm 1, we need to combine Wang and Titterington (2006) and the convergence of the Newton's method. Therefore, by Theorem 1 of Wang and Titterington (2006), we have the following theorem, of which a proof is given in Appendix 2.

Algorithm 1: Variational inference for GTM- -HDP topic model

Input: Documents $\{\mathbf{w}_d\}_{d=1}^{D}$
Output: Variational parameters Σ
initialise: Initialise the variational parameters $\Sigma^{(0)}$ and hyperparameters $\{\alpha, \gamma, \lambda\}$
while *the ELBO not converge* **do**
  for $d \in \{1, \ldots, D\}$ **do**
    Update a_{k1},a_{k2}, η_{dtk} and ζ_{dtk} by Equations (15) and (16).
  end
  Compute γ_{dt1}, γ_{dt2} and κ_{jv} by Equation (17)
end

Theorem 4.1: *as n approaches infinity, the iterative procedure* (18) *with probability 1 converges locally to the true value* $\Sigma^{\star}$; *that is, the iterative procedure* (18) *converges to the true value* $\Theta^{\star}$ *and the starting values are sufficiently near to* $\Sigma^{\star}$.

5. Experiments

This section mainly carries out experimental design and verifies the superiority of new model HDP–GTM proposed in the previous section. The organisational structure of this section is as follows: first, the dataset used in this experiment is explained, including the preprocessing of the dataset. Then the design scheme is tested, Support Vector Machine (SVM) algorithm is adopted to classify the text, and indicators to evaluate the experimental results are given. Finally, the text classification effects of HDP–GTM, GTM, HDP and LDA are compared.

5.1. Data description

Two text libraries were used in this chapter experiment, Reuters-21578[1] and 20-newsgroup[2]. Reuters-21578 is a series of news articles published on Reuters news website in 1987. It is widely used in text classification tasks and has been manually marked by Reuters personnel. This group consists of 21,578 documents, some of which are unmarked, and some of which are marked with one or more of 672 different categories. These categories are divided into five different categories: communication, organisation, personnel, location and theme. The 20-newsgroup data set is a news text library, which is often used for text classification, text mining and information retrieval. This text library integrates about 20,000 news texts, and is evenly divided into 20 collections according to different news themes.

Before text analysis, the dataset should be preprocessed first. The main purpose of text preprocessing is to reduce the dimension of the text data by controlling vocabulary so as to analyse the subsequent process. Text preprocessing steps commonly used in text classification are tagging, normalisation, stop-word removal and word stem.

- Tagging: based on the 'bag of words' assumptions, the text is divided into words or other meaningful parts regardless of the order between words. After being split, the whole text library is like a bag filled with words, and each word will correspond to a unique code.
- Normalisation: To convert and add or delete words in the document such as changing all uppercase parts of words into lowercase, discarding words

[1] http://www.davidlewis.com/resources/test collections/Reuters 21578/.
[2] http://www.qwone.com/Jason/20 news groups/.

with vocabulary less than 10 words in articles, discarding words with vocabulary less than 3 or more than 20 characters, deleting numbers and non-alphabetic characters, etc.
- Delete stop words: Stop words refer to words that are often encountered in the text but have nothing to do with analysis, such as prepositions, articles, conjunctions and so on.
- Words stem: Stemming is used to identify the root/stem of a word in the text.

After preprocessing, the variance threshold of a simple baseline method is adopted to select features, and the features whose variance does not meet the threshold are removed. In the traditional theme model, we get the document vector through the 'bag of words' (Aldous, 1985). Word bag is a simplified representation in which documents are represented by word bag regardless of grammar or even word order.

This paper studies the text data of graph structure at the word level and considers the co-occurrence relationship between words. Therefore, before applying the theme model HDP–GTM, we also need to calculate the correlation coefficient between words and convert the text data into graph structure data. On the basis of traditional data preprocessing, it is necessary to add one step operation at the last step to convert text data into graph structure data. First, according to the following formula:

$$f_{co} = \frac{|N_{e_i} \cap N_{e_j}|}{|D|}, \tag{19}$$

where D is the number of documents in the corpus, N_{e_i} and N_{e_j} denote the number of words e_i and e_j in the corpus D respectively. We need to calculate the frequency of common occurrence of each pair of words f_{CO}, and then compare it with the preset threshold ρ, to determine whether there are edges between words. If $f_{co}(i,j) \geq \rho$, the results of parameter derivation in the previous chapter show that the existence of edges will affect the calculation of posterior parameters.

5.2. Experiment design

The 'bag of words' representations of the documents are unable to recognise synonyms from a given term set and unable to recognise semantic relationships between terms. We apply the topic-model approach to cluster the words into a set of topics. Words assigned into the same topic are semantically related. Our main goal is to compare the performance of classification among different topic models. We also apply and compare whether the different threshold values have effects on the classification. This experimental process is divided into three stages, text preprocessing stage, topic models training stage and text classification stage(Figure 5).

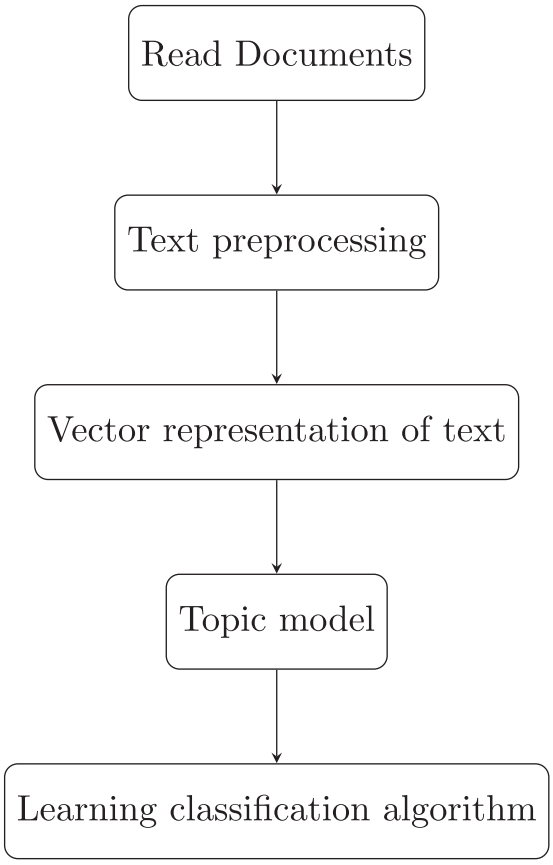

Figure 5. Graphical model representation of the process for text classification.

In the topic models training stage, all documents are used to train the topic model, and the purpose is to extract valid feature information of the text. In the text classification stage, the text set is divided into training set and test set, and support vector machine (SVM) is adopted to classify the text trained by different topic models, and the classification effect of different models under SVM is compared. Because we consider the association between words in the graph theme model, it is necessary to set the threshold value to determine whether edges exist or not in advance before experiments are carried out. When we set different thresholds ρ, the number of edges existing in the text is different. Figure 1 shows the relationship between threshold and average graph density in the Reuters dataset.

5.3. Evaluation measures

In this paper, accuracy, recall and F-score are selected as evaluation index of experimental results. Accuracy and recall are often used in the fields of information retrieval and text mining as important indicators to test model results. As far as text classification is concerned, accuracy is used to measure the number of correctly classified texts in the extracted texts accounting for the number of samples. The recall is the proportion of correctly classified samples to all samples in the sample. In Table 1, TP indicates the actual number of samples in

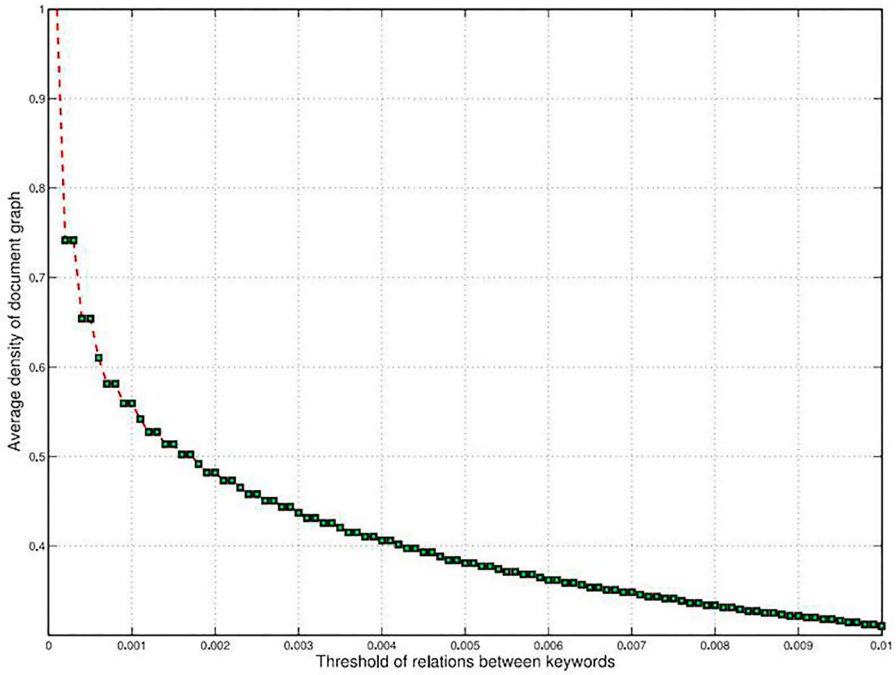

Figure 6. The relation between threshold value of word relation and average graph density.

class a which are finally assigned to class a, FP indicates the actual number of samples in class b which are predicted as the class a, FN indicates the actual number of samples in class a which are predicted as class b and TN indicates the actual number of samples in class b which are predicted as the class b. The calculation formula of the precision rate is $P = TP/(TP + FP)$. The calculation formula of the recall rate is $R = TP/(TP + FN)$. In the actual classification, we hope that the higher the value of P is, the better it is, and the higher the value of R is, but sometimes when p and r values conflict, we can refer to the value of F at this time. F is a harmonic average of accuracy and recall $F = 2PR/(P + R)$. The larger F is, the better the classification effect.

5.4. Result analysis

The subject categories belong to which in Reuters data set and 20-newsgroup data set are known. At the beginning of the experiment, first, the topic distribution of each document is deduced by HDP–GTM, HDP, GTM and LDA models, and then SVM algorithm is applied to classify the topic of each document. Finally, the classification effect of different models is compared by calculating the classification accuracy and recall rate. We set different correlation coefficient threshold ρ and use SVM classification method to compare the text topic classification effects of HDP–GTM and other three models HDP, GTM and LDA.

5.4.1. Reuter dataset

The Reuters data set is preprocessed, and finally 8025 texts are left. In the process of SVM text classification, 5770 of these texts are used as training sets, and the rest 2255 documents are used as test sets. In this experiment, four thresholds ρ are selected to compare the precision, recall and F_1 value of topic classification of Reuters dataset by three models under different thresholds.

Table 1. Evaluation measures.

	True classification	False classification
Predicted true classification	TP (true positive)	FP (false positive)
Predicted false classification	FN (false negative)	TN (true negative)

From Table 2, the classification accuracy of HDP–GTM is generally higher than that of other three models. The selection of the average density threshold ρ of the graph basically has no influence on the classification effect of LDA and HDP, because LDA and HDP models are based on the 'bag of words' assumption, and the correlation between words is not considered, so the classification accuracy of LDA and HDP is not different in different graph densities. GTM and HDP–GTM are relatively sensitive to average plot density, but it is not that the larger the plot density, the better. As shown in Figure 7, when $\rho = 0.0042$, HDP–GTM has the best classification effect, and the figure density at this time is 0.4017.

5.4.2. 20-newsgroup dataset

The 20-news group data set has 18,846 documents left after data preprocessing. In the process of SVM text classification, 11,314 documents are used as test sets, and the rest 7532 documents are used as test sets. In order to compare the topic classification effects of the three models for the text library, the thresholds in this section are consistent with those in the previous

Table 2. Classification effects of the Reuters.

Reuters dataset								
	$\rho = 0.01$				$\rho = 0.0042$			
	HDP–GTM	HDP	GTM	LDA	HDP–GTM	HDP	GTM	LDA
P	**0.742**	0.708	0.705	0.582	**0.763**	0.714	0.714	0.588
R	**0.904**	0.878	0.880	0.580	**0.920**	0.876	0.884	0.600
F	**0.818**	0.784	0.783	0.566	**0.834**	0.794	0.790	0.562
	$\rho = 0.0017$				$\rho = 0.0006$			
	HDP–GTM	HDP	GTM	LDA	HDP–GTM	HDP	GTM	LDA
P	**0.725**	0.707	0.690	0.580	**0.737**	0.716	0.710	0.581
R	**0.888**	0.881	0.885	0.582	**0.900**	0.873	0.876	0.583
F	**0.798**	0.781	0.775	0.564	**0.811**	0.786	0.784	0.568

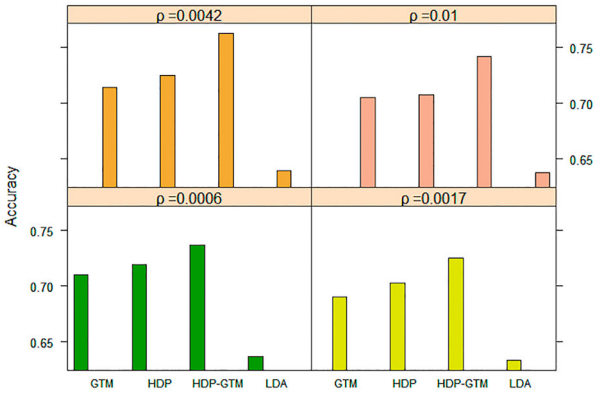

Figure 7. Classification effect of Reuters dataset under different ρ.

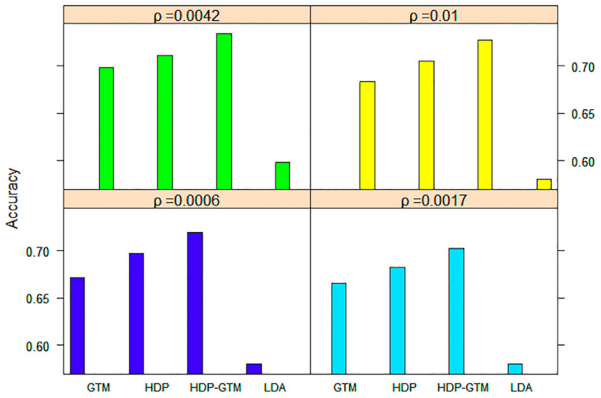

Figure 8. Classification effect of 20-newsgroup dataset under different ρ.

section. Table 3 shows the classification effect of three different models under different thresholds. Overall, the classification effect of HDP–GTM model is better than that of other three models.

Comparing the classification results of Table 3, we can find that the classification accuracy of LDA and HDP model for two datasets is not different which indicates that LDA and HDP modes are relatively stable for topic classification. At the same time, the sensitivity of the two datasets to the threshold is different. As shown in Figure 8, HDP–GTM performs best when $\rho = 0.0006$.

5.5. Selection of the threshold

For graph structure data, the selection of threshold value is related to the number of edges in the graph, and the range of threshold value is $\rho \in (0, 1)$. Consider the extreme situation: when ρ approaches zero, it means that the density of the graph approaches 1, that is to say, this graph is a connected graph, that is, there are edges between each pair of node pairs. When $\rho = 0$, it means that two nodes appear simultaneously in every document, which is almost impossible in practice. The edges in the graph are almost zero, and HDP–GTM at this time is equivalent to the HDP model and the GTM at this time is equivalent to the LDA model.

Through the analysis of the above two data sets, it is found that the accuracy of GTM and HDP–GTM model for the topic classification of Reuters data set is higher than that of 20-newsgroup data set, as shown in Figure 9. This difference is mainly related to the data

Table 3. Classification effects of the 20-newsgroup.

20-newsgroup dataset								
	$\rho = 0.01$				$\rho = 0.0042$			
	HDP–GTM	HDP	GTM	LDA	HDP–GTM	HDP	GTM	LDA
P	**0.727**	0.702	0.683	0.581	**0.702**	0.698	0.665	0.580
R	**0.823**	0.813	0.848	0.734	**0.815**	0.802	0.798	0.723
F_1	**0.772**	0.755	0.757	0.684	**0.754**	0.748	0.725	0.681
	$\rho = 0.0017$				$\rho = 0.0006$			
	HDP–GTM	HDP	GTM	LDA	HDP–GTM	HDP	GTM	LDA
P	**0.734**	0.695	0.698	0.598	**0.719**	0.697	0.671	0.580
R	**0.873**	0.811	0.858	0.735	**0.835**	0.809	0.816	0.776
F_1	**0.754**	0.751	0.725	0.681	**0.796**	0.751	0.770	0.697

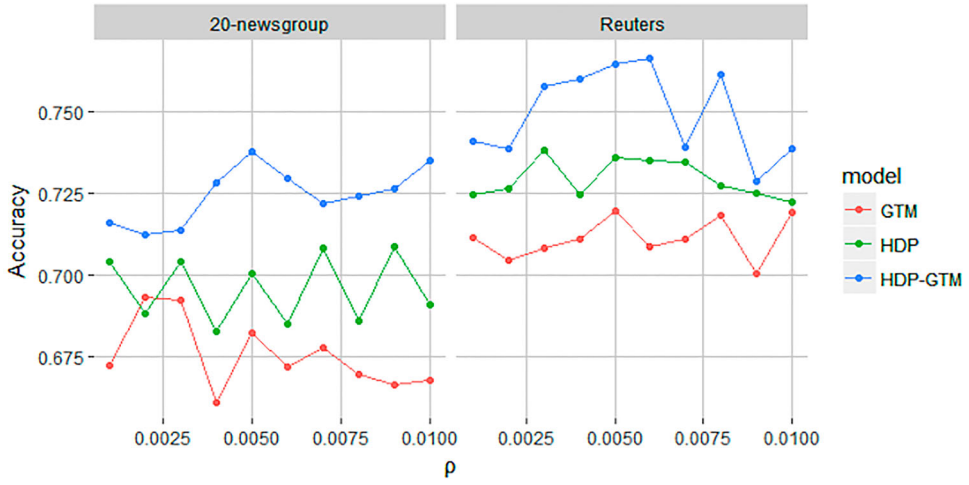

Figure 9. Classification effect of two datasets under HDP–GTM model.

structure of the database itself. Figure 9 shows the classification effect of GTM model and HDP–GTM model on two datasets.

6. Discussion

This paper proposed a new method HDP–GTM for probabilistic topic modelling. HDP–GTM takes advantage of the HDP and the GTM. We applied a variational inference algorithm for calculating the posterior distribution and investigated its convergence property. In the experimental analysis, we reported applications in text categorisation of the Reuters dataset and the 20 newsgroup data set by the HDP–GTM, compared to HDP, GTM and LDA. We found that the HDP–GTM is better than HDP, GTM and LDA. Future researches can be considered from the following aspects:

(a) The graph structure in this paper considers the co-occurrence relationship at the word level, and it is not widely used in terms of the research scope. In addition to co-occurrence relations, future research can consider other related relations, such as proximity relations, semantic relations (Griffiths et al., 2005).
(b) Besides considering the graph structure at the word level, we can also try other graph structures at other levels, such as the graph structure at the topic level or the graph structure at the document level.
(c) This paper considers the HDP for the prior distribution of topics, besides, other nonparametric Bayesian processes can be also considered, such as the Pitman–Yor process (Sato & Nakagawa, 2010).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by NSFC under grant No. 71371074 and the 111 Project under No. B14019.

Notes on contributors

Haibin Zhang is currently pursuing the doctoral degrees from East China normal University, Shanghai, China. His current research interests include text mining, machine learning, and deep learning.

Huating Shang received the master's degree from East China normal University. She is a statistical analyst in Jiangsu Hengrui Medicine. Her current research interests include text mining and clinical trial.

Xianyi Wu received the PH.D degree from East China normal University. He is a professor with school of statistics, East China normal University. His current research interests include mathematical statistics actuarial, stochastic scheduling and machine learning.

References

Aldous, D. J. (1985). *Exchangeability and related topics*. Berlin: Springer.
Blackwell, D., & MacQueen, J. (1973). Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, *1*, 353–355.
Blei, David M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.
Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. doi:10.1080/01621459.2017.1285773
Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.
Deerwester, S. (1990). Indexing by latent semantic analysis. *Journal of the Association for Information Science and Technology*, *41*(6), 391–407.
Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209–230.
Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl 1), 5228–5235.
Griffiths, T. L., Steyvers, M., Blei, D. M., Tenenbaum, J. B. (2005). *Integrating topics and syntax*. International Conference on Neural Information Processing Systems.
Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76.
Hofmann, T. (1999). *Probabilistic latent semantic analysis*. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Stockholm, Sweden.
Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, *37*(2), 183–233.
Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Waltham: Academic Press.
Pitman, J. (2002). Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, *11*(5), 501–514.
Sato, I, Nakagawa, H. (2010). *Topic models with power-law using Pitman-Yor process*. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington.
Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*(2), 639–650.
Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.
Valle, K. (2011). *Graph-based representations for textual case-based reasoning* (Master thesis). Norwegian University of Science and Technology.
Wainwright, M. J, Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference. Foundations and Trends* (Vol. 1, pp. 1–305).
Wang, C., Paisley, J. W., & Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. *Journal of Machine Learning Research*, *15*, 752–760.
Wang, B., & Titterington, D. M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, *1*(1), 625–650.
Xuan, J., Lu, J., Zhang, G., Luo, X. (2015). Topic model for graph mining. *IEEE Transactions on Cybernetics*, *45*(12), 2792–2803.

Appendices

Appendix 1. Derivation of posterior inference for GTM–HDP

First, we expand evidence lower bound function $L(\Sigma)$,

$$\begin{aligned}
L(\Sigma) &= E_Q[\log P(\mathbf{Z}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{C}, \mathbf{W}, \mathbf{E} \mid \alpha, \gamma, \lambda)] \\
&\quad - E_Q[\log Q(\boldsymbol{\beta}', \boldsymbol{\phi}', \boldsymbol{\pi}', \boldsymbol{C}', \boldsymbol{Z}')] \\
&= \sum_{d=1}^{D}[E_Q \log P(\boldsymbol{\pi}_d \mid \alpha) + \sum_{t=1}^{\infty} E_Q \log P(c_{dt} \mid \boldsymbol{\beta}) \\
&\quad + \sum_{n=1}^{N_d}\sum_{t=1}^{\infty}\sum_{k=1}^{\infty} E_Q \log P(w_{dn} \mid z_{dn}, \phi_k, c_{dt}) \\
&\quad + \sum_{n=1}^{N_d} E_Q \log P(z_{dn} \mid \boldsymbol{\pi}_d) + \sum_{i,j} E_Q \log P(e_{w_{di}, w_{dj}})] \\
&\quad + \sum_{k=1}^{\infty} E_Q \log P(\beta_k \mid \gamma) + \sum_{k=1}^{\infty} E_Q \log P(\phi_k \mid \lambda_k) \\
&\quad - \sum_{d=1}^{D}\Bigg[\sum_{t=1}^{\infty} q(\pi_{dt} \mid \gamma_{dt}^1, \gamma_{dt}^2) - \sum_{t=1}^{\infty} q(c_{dt} \mid \zeta_{dt}) \\
&\quad - \sum_{n=1}^{N_d} E_Q \log P(z_{dn} \mid \boldsymbol{\phi}_d)\Bigg] \\
&\quad + \sum_{k=1}^{\infty} E_Q \log q(\beta_k \mid a_k^1, a_k^2) + \sum_{k=1}^{\infty} E_Q \log q(\phi_k \mid \lambda_k).
\end{aligned}$$

We rewrite the first term using indicator random variables:

$$\begin{aligned}
&E_{q(\boldsymbol{\pi}_\mathbf{d})} \log P(\pi_{dt} \mid \alpha) \\
&\quad = \sum_{s=1}^{t-1} E_{q(\pi_\mathbf{d})} \log(1 - \pi'_{ds}) + E_{q(\pi_\mathbf{d})} \log(\pi'_{dt}),
\end{aligned}$$

where

$$\begin{aligned}
E_{q(\pi_\mathbf{d})}[\log \pi'_{dt}] &= \Psi(\gamma_{dt1}) - \Psi(\gamma dt1 + \gamma_{dt2}), \\
E_{q(\pi_\mathbf{d})}[\log(1 - \pi'_{dt})] &= \Psi(\gamma_{dt2}) - \Psi(\gamma_{dt1} + \gamma_{dt2}).
\end{aligned}$$

The digamma function, denoted by Ψ, arises from the derivative of the log normalisation factor in the beta distribution. Then,

$$\begin{aligned}
&E_{q(\pi_\mathbf{d})} \log P(\pi_{dt} \mid \alpha) \\
&\quad = \begin{cases} \sum_{s=1}^{t-1} E_{q(\pi_\mathbf{d})} \log(1 - \pi'_{ds}) + E_{q(\pi_\mathbf{d})} \log(\pi'_{dt}) & \text{if } t < T, \\ \sum_{s=1}^{T-1} E_{q(\pi_\mathbf{d})} \log(1 - \pi'_{ds}) & \text{if } t \geq T. \end{cases}
\end{aligned}$$

$$E_Q \log P(z_{dn} \mid \pi_d) = \sum_{t=1}^{\infty} q(z_{dn} = t) E_Q[\log P(\pi_{dt})].$$

Recall that $q(z_{dn} = t) = 0$ for $t \geq T$. Consequently, we can truncate this summation at $t = T$:

$$E_{Q(\mathbf{Z}, \boldsymbol{\Theta})} \log P(z_{dn}) = \sum_{t=1}^{T-1} q(z_{dn} = t) E_Q[\log P(\pi_{dt})],$$

where $q(z_{dn} = t) = \tilde{\pi}_{dt}$. Similarly,

$$\begin{aligned}
&E_{q(\beta_\mathbf{k})} \log P(\beta_k \mid \gamma) \\
&\quad = \begin{cases} \sum_{\ell=1}^{k-1} E_{q(\beta_\mathbf{k})} \log(1 - \beta'_\ell) + E_{q(\beta_\mathbf{l})} \log(\beta'_\ell) & \text{if } k < K, \\ \sum_{k=1}^{K-1} E_{q(\beta_\mathbf{k})} \log(1 - \beta'_\ell) & \text{if } k \geq K, \end{cases}
\end{aligned}$$

where

$$\begin{aligned}
E_q[\log \beta'_k] &= \Psi(a_{k1}) - \Psi(a_{k1} + a_{k2}), \\
E_q[\log(1 - \beta'_k)] &= \Psi(a_{k2}) - \Psi(a_{k1} + a_{k2}), \\
E_Q \log P(c_{dt} \mid \boldsymbol{\beta}) &= \sum_{k=1}^{K-1} q(c_{dt} = k) E_q[\log \beta_k],
\end{aligned}$$

where $q(c_{dt} = k) = \eta_{dtk}$.

$$E_Q \log P(\phi_{\boldsymbol{k}} \mid \lambda_{\boldsymbol{k}}) = \sum_{v=1}^{V} (\lambda_{kv} - 1)\left(\Psi(\kappa_{jv}) - \Psi\left(\sum_{k=1}^{V} \kappa_{jk}\right)\right) + \log\left(\Gamma\left(\sum_{v=1}^{V} \lambda_{kv}\right)\right) - \sum_{v=1}^{V} \log(\Gamma(\lambda_{kv})),$$

$$E_Q \log P(w_{dn} \mid z_{dn}, \phi_k, c_{dt}) = \sum_{k=1}^{\mathrm{T}} \sum_{v=1}^{V} q(z_{dn} = t) w_{dn}^{v} \eta_{dtk} E_{Q(\mathbf{Z},\Theta)}[\log(\phi_{jv})].$$

To simplify our notation, let $w_{dn}^{v} = 1$, iff w_{dn} is the vth word in the vocabulary.

$$q(z_{dn} = t) = \zeta_{dtk},$$

$$E_Q[\log \phi_{kv}] = \Psi(\lambda_{kv}) - \Psi\left(\sum_{k=1}^{V} \lambda_{kv}\right),$$

$$\begin{aligned}\sum_{(i,j)} E_q\left[\log p(e_{w_{di},w_{dj}})\right] &= \sum_{(i,j)} E_q\left[\log(\phi_{c_{dz_{di}}} \cdot \phi_{c_{dz_{dj}}})\right] \\ &\approx \sum_{(i,j)} E_q\left[\varsigma^{-1} \sum_{k}^{K} \phi_{di,k}.\phi_{dj,k} + \log\varsigma - 1\right] \\ &= \sum_{(i,j)} \left(\varsigma^{-1} \lambda_{ki}^{\phi_{dik}}.\lambda_{kj}^{\phi_{dik}} + \log\varsigma - 1\right).\end{aligned}$$

$$\varsigma = N_{(i,j)} \sum_{k}^{K} (\lambda_{k,n_i}^{\varphi_{d,n_i,k}} \cdot \lambda_{k,n_j}^{\varphi_{d,n_j,k}}).$$

It can be seen from the forming process of edge that the probability of edge existence follows a binomial distribution, which is obtained by the inner product of the word distribution of the corresponding node subject. The above second step uses the Taylor's expansion of the logarithmic function at the point ς.

We first maximise equation with respect to η_{dtk}. Observe that this is a constrained maximisation since $\sum_{k=1}^{K} \eta_{dtk} = 1$. We form the Lagrangian by isolating the terms which contain η_{dtk} and adding the appropriate Lagrange multipliers.

$$\begin{aligned}L_{[\eta_{dtk}]} &= \eta_{dtk} E_{q(\beta)} \log(\beta'_{\ell}) + \eta_{dtk} E_{q(\beta)} \log(1 - \beta'_{\ell}) \\ &\quad - \eta_{dtk} \log(\eta_{dtk}) + \lambda\left(\sum_{i=1}^{\mathrm{T}} \eta_{dtk} - 1\right) \\ &\quad + + \sum_{n=1}^{N} \sum_{v=1}^{V} w_{dn}^{v} \eta_{dtk} \zeta_{dnt} (\Psi(\lambda_{kv}) - \Psi\left(\sum_{l=1}^{V} \lambda_{kl}\right).\end{aligned}$$

Taking derivatives with respect to η_{dtk} and setting this derivative to zero yields the maximising value of the variational parameter η_{dtk}, we obtain

$$\begin{aligned}\eta_{dtk} = \exp\Bigg\{&\sum_{e=1}^{k-1} (\Psi(a_{e2}) - \Psi(a_{e1} + a_{e2})) + (\Psi(a_{k1}) \\ &- \Psi(a_{k1} + a_{k2})) \\ &+ \sum_{n=1}^{N} \sum_{v=1}^{V} w_{dn}^{v} \zeta_{dnt} \left(\Psi(\lambda_{kv}) - \Psi\left(\sum_{l=1}^{V} \lambda_{kl}\right)\right)\Bigg\}.\end{aligned}$$

Similarly,

$$a_{k1} = 1 + \sum_{d=1}^{D} \sum_{t=1}^{\mathrm{T}} \eta_{dtk}, \quad a_{k2} = \gamma + \sum_{d=1}^{D} \sum_{t=1}^{\mathrm{T}} \sum_{f=k+1}^{K} \eta_{dtf},$$

$$\gamma_{dt1} = 1 + \sum_{n=1}^{N} \zeta_{dnt}, \quad \gamma_{dt2} = \alpha_0 + \sum_{n=1}^{N} \sum_{b=t+1}^{\mathrm{T}} \zeta_{dnb},$$

$$\kappa_{jv} = \beta_{jv} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} w_{dn}^{v} \zeta_{dnt} \eta_{dtk}.$$

For the parameter ζ_{dnt}, when finding the maximum value of $L(\zeta_{dnt})$ under the constraint of $\sum_{t}^{\mathrm{T}} \zeta_{dnt} = 1$), then the partial derivative of ζ_{dnt}:

$$\begin{aligned}\Delta_{\zeta_{dtk}} L(\zeta_{dtk}) = &\left(\Psi(\kappa_{k,n}) - \Psi\left(\sum_{n=1}^{N} \kappa_{k,n}\right)\right) \\ &+ \sum_{h=1}^{k-1} (\Psi(\gamma_{dh2}) - \Psi(\gamma_{dh1} + \gamma_{dh2})) \\ &+ \Psi(\gamma_{dk1}) - \Psi(\gamma_{dk1} + \gamma_{dk2}) - \log\zeta_{d,n,k} - 1 \\ &+ \sum_{n_j \in Ne(n_i)} (\varsigma^{-1} \cdot \log\kappa_{k,n} \cdot \kappa_{k,n}^{\zeta_{d,n,k}}) \\ &+ \sum_{k=1}^{N} \sum_{v=1}^{V} w_{dn}^{v} \eta_{dtk} \left(\Psi(\lambda_{kv}) - \Psi\left(\sum_{l=1}^{V} \lambda_{kl}\right)\right).\end{aligned}$$

Then the second derivative of ζ_{dtk} is $H(\zeta_{dtk}) = \partial L / \partial \zeta_{dtk} \partial \zeta_{dtk}$, by Newton's method,

$$\zeta_{dtk}^{(n+1)} = \zeta_{dtk}^{(n)} - H(\zeta_{dtk}^{(n)})^{-1} \Delta_{\zeta_{dtk}^{(n)}} f(\zeta_{dtk}^{(n)}).$$

Appendix 2. Proof of Theorem 4.1

We derived Σ into two parts ζ and the remaining part of Σ expressed by Γ, and note that $\Gamma^{\star}$ and $\zeta^{\star}$ are the true value of Γ and ζ respectively. Then

$$\|\Sigma - \Sigma^{\star}\| = \left\|\begin{pmatrix}\Gamma \\ \zeta\end{pmatrix} - \begin{pmatrix}\Gamma^{\star} \\ \zeta^{\star}\end{pmatrix}\right\| = \left\|\begin{pmatrix}\Gamma - \Gamma^{\star} \\ \zeta - \zeta^{\star}\end{pmatrix}\right\|,$$

the convergence of $\|\Gamma - \Gamma^{\star}\|$ is a special case of Theorem 1 in Wang and Titterington (2006), then the iterative procedure converges to the true value $\Gamma^{\star}$. The convergence of $\|\zeta - \zeta^{\star}\|$ can be proved by the Newton's method. With probability 1 as n approaches infinity, the iterative procedure (18) converges locally to the true value $\Sigma^{\star}$.