




Topic change point detection using a mixed Bayesian model

Xiaoling Lu^{1,2} · Yuxuan Guo² · Jiayi Chen³ · Feifei Wang^{1,2} 

Received: 6 March 2021 / Accepted: 5 October 2021 / Published online: 17 October 2021

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2021

Abstract

Dynamic text documents, including news articles, user reviews, and blogs, are now commonly encountered in many fields. Accordingly, the topics underlying text streams also change over time. To grasp the topic changes in the increasing accumulation of text documents, there is a great need to develop automatic text analysis models to find the key changes in topics. To this end, this study proposes a topic change point detection (Topic-CD) model. Different from previous studies, we define the change point of topics from the perspective of hyperparameters associated with topic-word distributions. This allows the model to detect change points underlying the whole topic set. Under this definition, the topic modeling and change point detection are combined in a unified framework and then performed simultaneously using a Markov chain Monte Carlo algorithm. In addition, the Topic-CD model is free from setting the number of change points in advance, which makes it more convenient for practical use. We investigate the performance of the Topic-CD model numerically using synthetic data and three real datasets. The results show that the Topic-CD model can well identify the change points in topics when compared with several state-of-the-art methods.

Responsible editor: Johannes Fürnkranz.

✉ Feifei Wang
feifei.wang@ruc.edu.cn

Xiaoling Lu
xiaolinglu@ruc.edu.cn

Yuxuan Guo
yuxuanguo@ruc.edu.cn

Jiayi Chen
jane.cjy@alibaba-inc.com

¹ Center for Applied Statistics, Renmin University of China, Beijing, China

² School of Statistics, Renmin University of China, Beijing, China

³ Intelligent Marketing Platform, Alibaba Group, Beijing, China

Keywords Change point detection · Dynamic topic models · Latent Dirichlet allocation · Markov chain Monte Carlo

1 Introduction

With the rapid development of technology and exponential internet growth, there is an increasing accumulation of text documents in all fields, e.g., emails, news articles, and consumer comments. Faced with massive amounts of textual data, it is impossible for individuals to keep track of all relevant key points and detect changes in emerging trends or topics. Consequently, automatic text summarization and change point detection methods that allow users to grasp event changes in text streams quickly have received substantial attention.

To detect event changes in text streams, topic models, which can reveal the thematic structure in a large document corpus, have been widely used. However, studies identifying the *change point* during topic changes over time have only emerged in recent years. In statistics, change point detection is a well-defined problem, which identifies the moments when the probability distribution of a stochastic process or time series has changed (Bai 1997; Chib 1998). The past literature in the recent decade has seen an increasing number of works focusing on topic change point detection. Bruggermann et al. (2016) first applied the dynamic LDA model to find topic-word distributions across time. Then, a change point was defined as the moment when the word distribution of a specific topic had changed enough from the previous time stamp. Zhang et al. (2017) applied the standard LDA for topic extraction and prediction. Then they used the method of scientific evolutionary pathway to find emerging topics and death topics. Wang and Goutte (2018) adopted a similar definition of topic change points. They first applied online topic models to obtain the topic-word distributions in different times, and then applied on-line change point detection algorithms to detect the moments when significant changes happened. In summary, these studies defined topic change points from the perspective of topic-word distributions. To find the change points, they used two-step strategies by first building topic models for documents and then conducting change point detection. However, there are also limitations associated with these studies. First, they define the change point for each single topic. This definition perspective is meaningful but when the number of changing topics is small, their influence to the whole corpus may be limited. In the meanwhile, the change point for each topic may appear at different time points. As a consequence, the semantic meanings of the whole corpus would show a pattern with gradual changes. Second, the two-stage strategy might have adverse effects on topic learning, because the existence of change points inevitably affects the meaning of topics.

Another adopted definition of topic changes is constructed from the perspective of document–topic representations. A related work is the topic segmentation method (Lan et al. 2013), which applied the Pitman–Yor process to segment topic distributions represented by each single document and thus found the topic changing positions. A more recent work is the multiple latent changepoint topic model (Zhong and Schweidel 2020), which also focused on changes of document–topic distributions, but applied the Dirichlet process hidden Markov multiple change point (DPHMM) model (Ko et al.

2015) to detect the topic changes. Different from the previous works, we define topic changes as the change of topic meanings (i.e., the topic-word distributions) occurred in the whole topic set.

We develop in this work a new topic change point detection model, which we refer to as the Topic-CD model. In this model, we define the change point of topics from the perspective of hyperparameters associated with the topic-word distributions. This definition allows detection of change points in the whole topic set, as the hyperparameters can affect the meanings of all topics. Under this definition, when the hyperparameters change, the meanings of all topics would change. Only the topic meanings have changed significantly (measured by the hyperparameters), then a change point can be detected. Therefore, the Topic-CD model often detects change points as the moments that the semantic meanings of the whole corpus have changed dramatically. To model the changes in hyperparameters, we assume that they follow the DPHMM model (Ko et al. 2015). The DPHMM model avoids specifying the number of change points in hyperparameters in advance. Given the hyperparameters, the documents at each time stamp are modeled via LDA. In this way, text modeling and change point detection are combined in a unified framework and then performed simultaneously. Consequently, the Topic-CD model overcomes the disadvantages caused by separate detection. For model estimation, we propose a Markov chain Monte Carlo (MCMC) algorithm. The detection performance of the Topic-CD model is evaluated using a series of experiments on synthetic data and three real datasets.

To summarize, the contributions of this study are as follows:

- We define the change point of topics from a new perspective (i.e., the hyperparameters), which facilitates the detection of change point from the whole topic picture.
- We propose a novel Topic-CD model for topic change point detection. This model combines topic models and change point detection in a unified framework, which improves the quality of topic learning and the performance of change point detection.
- The Topic-CD model does not require the number of change points be set in advance, which makes it more flexible and convenient for practical use.
- To evaluate the performance of the proposed Topic-CD model, various experiments are conducted on both synthetic data and three real datasets.

The remainder of this paper comprises five sections. We review the related literature in Sect. 2. In Sect. 3, we introduce the Topic-CD model and its estimation algorithm. In Sect. 4, the finite sample performance of the Topic-CD model is demonstrated through various experiments on synthetic data. In Sect. 5, we apply the Topic-CD model on three real datasets. Section 6 concludes the paper with a brief discussion.

2 Related literature

Automatic text summarization and change point detection have received much attention in recent years (Guo et al. 2013; Wu et al. 2014; Hasan et al. 2017; Wang and Goutte 2018; Zhong and Schweidel 2020). Among all these methods, topic models,

which are a suite of three-level hierarchical Bayesian models, have been widely used. The most basic topic model is the latent Dirichlet allocation (LDA, Blei et al. 2003). In LDA, there are a set of latent topics underlying all documents. Each topic is represented by a specific vector of probabilities over the dictionary, and each document is assumed to be generated from a probability distribution over these topics. Because LDA is a hierarchical Bayesian model, it can be easily extended to address various problems and has wide applications in text mining, such as text classification, summarization, and sentiment analysis (Blei and McCallum 2008; Lin and He 2009; Ramage et al. 2009; Blei 2012).

There are numerous LDA extensions that focus on dynamic topic modeling for text streams (Wang and McCallum 2006; Wang et al. 2015; Pozdnoukhov and Kaiser 2011; Chae et al. 2012; Vavliakis et al. 2012; Zhou and Chen 2014). In general, the dynamic topic models can be classified roughly into three categories according to the time variable (Zhou et al. 2017). The first category is discrete time topic evolution models. Important works in this category include: the dynamic topic model (DTM, Blei and Lafferty 2006), which assumes a time evolution factor in each neighboring time slice to model topic changes; the multiscale topic tomography model (MTTM, Nallapati et al. 2007), which assumes both a data generation process and parameter generation processes over time; and some extensions such as the temporal Dirichlet process mixture model (TDPM, Ahmed and Xing 2008), the infinite dynamic topic model (iDTM, Ahmed and Xing 2010), and the dynamic topic model based on non-negative matrix factorization (Greene and Cross 2016). The second category is continuous time topic evolution models. Important works in this category include: the continuous time dynamic topic model (cDTM, Wang et al. 2015), which uses Brownian motion to model topic evolution over time; and the model of topics over time (TOT, Wang and McCallum 2006), which uses a beta distribution to jointly model time with word co-occurrence patterns to explore topic changes; and other works such as Kawamae (2011) and Dubey et al. (2013). The last category is online topic evolution models, which have attracted more attentions in the last decade. In this category, online inference algorithms are developed to model the online document streams. For example, AlSumait et al. (2008) extends the original LDA model into an online model and proposes an empirical Bayesian-based solution. Other works include Sasaki et al. (2014), He et al. (2015) and Mohamad and Bouchachia (2019).

Dynamic topic models usually focus on investigating the topic evolution patterns, which mainly characterize the *gradual changes* of topic meanings. In the recent decade, an increasing number of works pay their attentions on topic detection or topic change point detection. Different from dynamic topic models, the works related to topic change point detection often try to characterize the *sudden changes* or *big changes* of topic meanings. An early related work is Holz and Teresniak (2010), which adopts the volatility measure to find contextual shift of topics over time slices. However, this work only relies on the occurrences of words, and not applies the topic modeling techniques. In terms of identifying the *change point* of topics using topic models, there are three important works (Bruggemann et al. 2016; Zhang et al. 2017; Wang and Goutte 2018). Specifically, Bruggemann et al. (2016) uses a dynamic topic model to obtain multiple topic sequences. It defines the change point as the moment when the word distribution of a specific topic has changed over a pre-defined threshold during

any adjacent time periods. Zhang et al. (2017) first applies the LDA model to profile the hotspots of current datasets, and then predicts the possible future trends. Then, a scientific evolutionary pathway method is applied to detect the emerging topics and death topics over time slices. Wang and Goutte (2018) applies a similar definition for topic change point as Bruggermann et al. (2016). However, it focuses on the problem of real-time change point detection. Specifically, it first uses online LDA (Hoffman et al. 2010) to achieve sequences of real-time topics and then applies various online change point detection methods to identify the topic change point.

Although the above studies have achieved good performance in finding changes in topics, they still have some limitations. First, they identify the change point for each single topic, whereas in practice, we are also interested in detecting changes from the perspective of the whole topic set. Second, all these works apply two-step strategies, which conduct topic learning and topic change point detection separately. Given that the topic change point is an important characteristic of documents, combining the two modeling processes can help better summarize document meanings and detect change points. Recently, a multiple latent changepoint topic model is proposed to capture changes in social media content (Zhong and Schweidel 2020). Similar with the Topic-CD model, it also combines the standard LDA model with the DPHMM process. However, this work assumes the DPHMM prior on α , while the Topic-CD model assumes the DPHMM prior on β . In other words, Zhong and Schweidel (2020) defines change point from the perspective of topic distributions represented by documents, while our work defines change point from the perspective of topic meanings. Another typical work is the topic segmentation method using a structured topic model (Lan et al. 2013). Similar with Zhong and Schweidel (2020), the topic segmentation method tries to segment the topic distributions represented by each document (i.e., the θ_d) and thus finds the topic changing positions. On the contrary, the Topic-CD model focuses on the change of topic meanings represented by all documents (i.e., the ϕ_k). Moreover, the topic segmentation method is designed for static documents, while the Topic-CD model is designed for dynamic documents and attempts to find topic changes across time.

3 The topic-CD model

3.1 Model description

Assume that there are D_t documents in the t -th moment with $1 \leq t \leq T$. Accordingly, the total number of documents is $D = \sum_{t=1}^T D_t$. Further assume that there are K topics underlying all D documents. Then, for the d -th document in the t -th moment ($1 \leq d \leq D_t$), assume it has a vector of topic probabilities $\theta_{td} = (\theta_{td,1}, \dots, \theta_{td,K})^\top$ over K topics. Although the number of topics remains unchanged in different moments, the meanings of the topics may change over time. The changes in topic meanings would be directly reflected by the topic-word probability distributions. Theoretically, in the t -th moment, assume that the k -th topic has a vector of word probability distribution $\phi_{tk} = (\phi_{tk,1}, \dots, \phi_{tk,V})^\top$ over the whole dictionary with size V . The past literature often defines change point on each single topic (Bruggermann et al. 2016; Wang and

Goutte 2018). Specifically, for the k -th topic, they assume a change point occurs at the moment when the topic-word distribution ϕ_{t+1k} has changed dramatically when compared with ϕ_{tk} .

However in this work, we attempt to investigate the change point from the perspective of all topics. To this end, we define the change point on the hyperparameters, which control the topic-word distribution $\{\phi_{tk}\}$ s. Specifically, at the t -th moment, assume that ϕ_{tk} follows a homogeneous Dirichlet distribution with hyperparameter β_t . During the time series $\{\beta_t : 1 \leq t \leq T\}$, assume that there are Q change points, which occur at moments τ_1, \dots, τ_Q . Here, the occurrence of a change point means the value of β_t has changed at this moment, that is, $\beta_{\tau_q} \neq \beta_{\tau_q-1}$ for $1 \leq q \leq Q$. It is notable that, Q change points can split T moments into $P = Q + 1$ partitions. To further represent the specific partition each moment belongs to, a state variable $s_t \in \{1, \dots, P\}$ is introduced in the t -th moment with $1 \leq t \leq T$. Under the assumption of change points, we can simplify the notations for topic-word distributions and their corresponding hyperparameters. Specifically, assume that $\beta = (\beta_1, \dots, \beta_P)$ are the hyperparameters in each partition. In the i -th partition with $1 \leq i \leq P$, assume that the topic meanings remain the same within moments in this partition. Therefore, the k -th topic has a vector of word probabilities $\phi_{ik} = (\phi_{ik,1}, \dots, \phi_{ik,v})^\top$ in the i -th partition, and ϕ_{ik} follows a homogeneous Dirichlet distribution with hyperparameter β_i . With the word-topic distribution $\{\phi_{ik}\}$ s having been specified, the generation process for documents in the i -th partition is similar to that in the LDA.

Next, we discuss the settings for the state variables $\{s_t\}$ s. It is notable that the values of $s_t (t = 1, \dots, T)$ can determine both the number and location of change points. This is because, if $s_t \neq s_{t+1}$, then a change point is observed at moment $t + 1$. Therefore, the task of change point detection is to detect the state variables accurately. To model the state variables, we assume that s_1, \dots, s_T follow the DPHMM (Ko et al. 2015). The DPHMM assumes that all state variables follow a Markov process. That is, s_t depends only on the state at the previous moment (i.e., s_{t-1}) and has nothing to do with other states in the past. In DPHMM, the conditional distribution of s_{t+1} given all previous state variables is specified as

$$p(s_{t+1} = j | s_t = i, s_1, \dots, s_{t-1}) = \begin{cases} \frac{n_{ii} + \lambda}{n_{ii} + \zeta + \lambda}, & j = i \\ \frac{\zeta}{n_{ii} + \zeta + \lambda}, & j \text{ is a new partition} \end{cases} \tag{1}$$

where $n_{ii} = \sum_{t'=1}^{t-2} \delta(s_{t'}, i)\delta(s_{t'+1}, i)$, and $\delta(a, b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}$. Here, n_{ii} represents the number of self-transition times, and λ and ζ are both hyperparameters, where λ controls the prior tendency to stay in a partition, and ζ controls the tendency to explore new partitions. For convenience, we assume $\eta = (\lambda, \zeta)$. It is also notable that, by introducing a Dirichlet process, DPHMM can determine the number of change points simultaneously in the parameter estimation process without the need for specifying in advance. In other words, after the state variables are generated, the number of

partitions (i.e., P) equals to the unique values of the state variables. Then the number of change points is also decided.

Finally, we present the generative process of the Topic-CD model.

1. Generate all state variables s_1, \dots, s_T from the DPHMM(η). Given $\{s_t\}$ s, the number of change points Q and their corresponding locations τ_1, \dots, τ_Q are determined.
2. For the i -th ($i = 1, \dots, Q + 1$) partition, the generative process of documents is given below:
 - (a) Generate β_i from a uniform distribution $\beta_i \sim U(b_0, b_1)$;
 - (b) Generate topic probabilities ϕ_{ik} ($k = 1, \dots, K$) independently from a homogeneous Dirichlet distribution: $\phi_{ik} \sim Dir(\beta_i, \dots, \beta_i)$.
 - (c) Each document d at moment t ($t = \tau_{i-1} + 1, \dots, \tau_i$) in this partition is independently generated as follows:
 - (i) Generate θ_{td} from a Dirichlet distribution with hyperparameter $(\alpha_1, \dots, \alpha_K)$: $\theta_{td} \sim Dir(\alpha_1, \dots, \alpha_K)$.
 - (ii) Each word n in document d ($n = 1, \dots, N_{td}$) is independently generated as follows:
 - (A) Choose a topic z_{tdn} from a multinomial distribution with probabilities given by θ_{td} : $z_{tdn} \sim Multi(\theta_{td})$.
 - (B) Choose a word w_{tdn} from a multinomial distribution with probabilities given by $\phi_{i, z_{tdn}}$: $w_{tdn} \sim Multi(\phi_{i, z_{tdn}})$.

Different from the standard LDA model, we do not use symmetric Dirichlet distributions for θ_{td} , but use asymmetric Dirichlet distributions with hyperparameters $\alpha = (\alpha_1, \dots, \alpha_K)^\top$. The usage of vector α can significantly improve the performance of LDA. It is notable that, there are also various works investigating how to estimate the hyperparameters α , such as Ishwaran and James (2001) and Teh et al. (2006). Therefore in real applications, the hyperparameters α can be estimated together with the Topic-CD model. To further illustrate the Topic-CD model, we present the graphical model of the generative process of documents with one single topic change point as an example in Fig. 1. Given the generative process of the Topic-CD model, we can derive the full posterior distribution of all variables. Then, the model can be estimated using the MCMC method. We describe the details of model estimation in the next section.

3.2 Model estimation

We apply the MCMC method for model estimation. Let $\mathbf{S} = (s_1, \dots, s_T)^\top$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)^\top$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\top$, $\mathbf{z} = \{z_{tdn} : 1 \leq n \leq N_{td}, 1 \leq d \leq D_t, 1 \leq t \leq T\}$, $\mathbf{w} = \{w_{tdn} : 1 \leq n \leq N_{td}, 1 \leq d \leq D_t, 1 \leq t \leq T\}$, $\boldsymbol{\Theta} = \{\theta_{td} : 1 \leq d \leq D_t, 1 \leq t \leq T\}$, and $\boldsymbol{\Phi} = \{\phi_{ik} : 1 \leq i \leq P, 1 \leq k \leq K\}$. For simplicity purpose, we first discuss model estimation with fixed η . Then we talk about how to estimate the hyperparameters η in practice. Specifically, with fixed η , the full posterior distribution

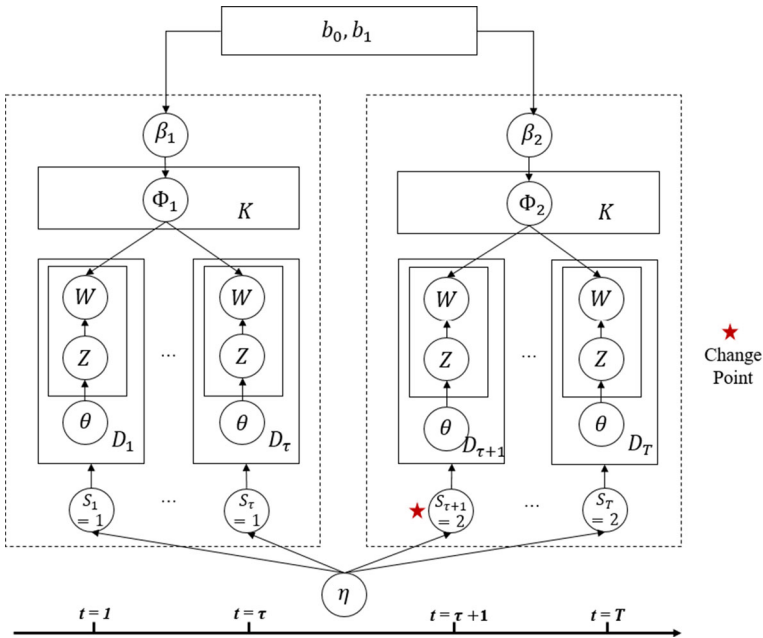


Fig. 1 Generative process for the Topic-CD model with one single topic change point appearing at $\tau + 1$. For simplicity, we omit the hyperparameters α for θ

of $(S, \beta, \Theta, \Phi, z)$ can be derived given the generative process of the Topic-CD model as:

$$\begin{aligned}
 & f(S, \beta, \Theta, \Phi, z | w, \alpha, b_0, b_1, \eta) \\
 & \propto f(S | \eta) f(\beta | S, b_0, b_1) f(\Theta | \alpha) f(\Phi | \beta) f(w, z | \Theta, \Phi).
 \end{aligned}
 \tag{2}$$

Given that the Dirichlet priors are conjugate to the multinomial distributions, we first integrate out Φ and Θ from the full posterior distribution. Specifically,

$$\begin{aligned}
 f(w, z | \Theta, \Phi) &= f(z | \Theta) f(w | z, \Phi) \\
 &= \left\{ \int f(z | \Theta) f(\Theta | \alpha) d\Theta \right\} \left\{ \int f(w | z, \Phi) f(\Phi | \beta) d\Phi \right\} \\
 &= \left\{ \prod_{t=1}^T \prod_{d=1}^{D_t} \int f(z_{td} | \theta_{td}) f(\theta_{td} | \alpha) d(\theta_{td}) \right\} \\
 & \quad \left\{ \prod_{i=1}^P \prod_{k=1}^K \int f(w | \phi_{ik}) f(\phi_{ik} | \beta_i) d(\phi_{ik}) \right\} \\
 &= \left\{ \prod_{t=1}^T \prod_{d=1}^{D_t} \frac{\Delta(\vec{n}_{td}^{(1)} + \vec{\alpha})}{\Delta(\vec{\alpha})} \right\} \left\{ \prod_{i=1}^P \prod_{k=1}^K \frac{\Delta(\vec{n}_{ik}^{(2)} + \vec{\beta}_i)}{\Delta(\vec{\beta}_i)} \right\}.
 \end{aligned}
 \tag{3}$$

Here, $\vec{n}_{td}^{(1)} = (n_{td1}^{(1)}, \dots, n_{tdK}^{(1)})^\top$, where $n_{tdk}^{(1)}$ denotes the number of words associated with topic k in document d at moment t ; $\vec{n}_{ik}^{(2)} = (n_{ik1}^{(2)}, \dots, n_{ikV}^{(2)})^\top$, where $n_{ikv}^{(2)}$ denotes the number of occurrences of word v with topic k in partition i . Here, $\vec{\alpha} = \alpha = (\alpha_1, \dots, \alpha_K)^\top$ is the vector of hyperparameters in the Dirichlet distributions for $\{\theta_{td}\}$ s, and $\vec{\beta}_i = (\beta_i, \dots, \beta_i)^\top$ is the vector of hyperparameters in the symmetric Dirichlet distributions for $\{\phi_{ik}\}$ s. $\Delta(\vec{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)}$, $\Delta(\vec{n}_{td}^{(1)} + \vec{\alpha})$, $\Delta(\vec{\beta}_i)$, and $\Delta(\vec{n}_{ik}^{(2)} + \vec{\beta}_i)$ are defined similarly. Given (3), the full posterior distribution is reduced as $f(\mathbf{S}, \boldsymbol{\beta}, \mathbf{z} | \mathbf{w}, \boldsymbol{\alpha}, b_0, b_1, \eta)$. Then, the collapsed MCMC method is applied for model estimation.

Below, we derive the full conditional distributions and updating strategies for \mathbf{S} , $\boldsymbol{\beta}$, and \mathbf{z} in detail.

(1) Updating \mathbf{S} .

The full conditional distribution of $s_t (t = 1, \dots, T)$ can be derived as follows:

$$f(s_t | \mathbf{S}_{t-1}, \mathbf{S}_{t+1}, \boldsymbol{\beta}, \mathbf{z}, \mathbf{w}, \eta) \propto f(s_t | s_{t-1}, \mathbf{S}_{t-2}, \eta) f(s_{t+1} | s_t, \mathbf{S}_{t+2}, \eta) f(\mathbf{w}_t | \mathbf{z}_t, \boldsymbol{\beta}_{s_t}), \tag{4}$$

where $\mathbf{S}_{t-1} = (s_1, \dots, s_{t-1})^\top$, $\mathbf{S}_{t+1} = (s_{t+1}, \dots, s_n)^\top$, and \mathbf{w}_t and \mathbf{z}_t denote all words and topic indicators at moment t . Because we assume that all state variables \mathbf{S} follow the DPHMM model, we obtain a change point only when $s_{t-1} \neq s_t + 1$. That is, we do not allow continuous change. Therefore, when $s_{t-1} = i$ and $s_{t+1} = i + 1$, the value of s_t could be either i or $i + 1$. According to the conditional distribution of state variables in (1) derived from the DPHMM model, we can obtain that when $s_t = i$, $f(s_t | s_{t-1}, \mathbf{S}_{t-2}, \eta) = \frac{n_{ii} + \lambda}{n_{ii} + \zeta + \lambda}$ and $f(s_{t+1} | s_t, \mathbf{S}_{t+2}, \eta) = \frac{\zeta}{n_{ii} + 1 + \zeta + \lambda}$; and when $s_t = i + 1$, $f(s_t | s_{t-1}, \mathbf{S}_{t-2}, \eta) = \frac{\zeta}{n_{ii} + \zeta + \lambda}$ and $f(s_{t+1} | s_t, \mathbf{S}_{t+2}, \eta) = \frac{n_{i+1,i+1} + \lambda}{n_{i+1,i+1} + \zeta + \lambda}$.

Next, we derive the full conditional distributions for the first state s_1 and last state s_T , which are given in (5) and (6), respectively.

$$f(s_1 | s_2, \dots, s_T, \eta) = \begin{cases} c \cdot \frac{\lambda}{\zeta + \lambda} \cdot \frac{\zeta}{\zeta + \lambda} \cdot f(\mathbf{w}_1 | \mathbf{z}_1, \boldsymbol{\beta}_{s_1}), & s_2 \text{ changes} \\ c \cdot \frac{\zeta}{\zeta + \lambda} \cdot \frac{n_{s_2 s_2} + \lambda}{n_{s_2 s_2} + \zeta + \lambda} \cdot f(\mathbf{w}_1 | \mathbf{z}_1, \boldsymbol{\beta}_{s_2}), & s_2 = s_1 \end{cases} \tag{5}$$

$$f(s_T | s_{T-1}, \dots, s_1, \eta) = \begin{cases} c \cdot \frac{n_{s_{n-1} s_{n-1}} + \lambda}{n_{s_{n-1} s_{n-1}} + \zeta + \lambda} \cdot f(\mathbf{w}_T | \mathbf{z}_T, \boldsymbol{\beta}_{s_{T-1}}), & s_T = s_{T-1} \\ c \cdot \frac{\zeta}{n_{s_{n-1} s_{n-1}} + \zeta + \lambda} \cdot f(\mathbf{w}_T | \mathbf{z}_T, \boldsymbol{\beta}_{s_T}), & s_T \text{ changes} \end{cases} \tag{6}$$

where $n_{s_2 s_2} = \sum_{t'=2}^{T-1} \delta(s_{t'}, s_2) \delta(s_{t'+1}, s_2)$, $n_{s_{n-1} s_{n-1}} = \sum_{t'=1}^{T-1} \delta(s_{t'}, s_{T-1}) \times \delta(s_{t'+1}, s_{T-1})$, c_1 and c_2 are some normalization constants. It is notable that

after all state variables \mathbf{S} are updated, the number of change points as well as their corresponding locations are determined.

(2) Updating β .

The full conditional distribution of $\beta_i (i = 1, \dots, P)$ is derived as follows:

$$\begin{aligned}
 f(\beta_i | \mathbf{w}, \mathbf{S}, \mathbf{z}, \beta_{-i}) &\propto f(\mathbf{w}_i | \mathbf{z}_i, \beta_i) f(\beta_i | b_0, b_1) \\
 &\propto f(\mathbf{w}_i | \mathbf{z}_i, \beta_i) \propto \prod_{k=1}^K \frac{\Delta(\vec{n}_{ik}^{(2)} + \vec{\beta}_i)}{\Delta(\vec{\beta}_i)}, \tag{7}
 \end{aligned}$$

where β_{-i} is β without β_i . Because it is difficult to sample from the full conditional distribution of β_i , we use the Metropolis-Hastings algorithm. The proposal distribution for β_i is a uniform distribution $U(\beta_i - \epsilon, \beta_i + \epsilon)$, where ϵ is a tuning parameter to help achieve a reasonable acceptance rate.

(3) Updating z .

The full conditional distribution of z_{tdn} is derived as follows:

$$\begin{aligned}
 f(z_{tdn} = k | \mathbf{S}, \mathbf{w}, \beta, \mathbf{z}_{-tdn}) &= \frac{f(\mathbf{S}, \mathbf{w}, \beta, \mathbf{z})}{f(\mathbf{S}, \mathbf{w}, \beta, \mathbf{z}_{-tdn})} \\
 &= \frac{\binom{n_{tdk, -tdn}^{(1)} + \alpha_k}{\sum_k n_{tdk, -tdn}^{(1)} + \alpha_k}}{\sum_v \frac{\binom{n_{s_t kv}^{(2)} - 1}{\sum_v (n_{s_t kv}^{(2)} - 1)} + \beta_{s_t}}{\sum_v (n_{s_t kv}^{(2)} - 1)} + \beta_{s_t}}, \tag{8}
 \end{aligned}$$

where \mathbf{z}_{-tdn} is \mathbf{z} with z_{tdn} being omitted, $n_{tdk, -tdn}^{(1)}$ denotes the number of words (excluding w_{tdn}) belonging to topic k in document d at moment t , v denotes the order of word w_{tdn} in the vocabulary, and $n_{s_t kv}^{(2)}$ denotes the number of occurrences of word v belonging to topic k in partition s_t . The full conditional distribution of z_{tdn} can be explained intuitively, where the term on the left represents the probability of sampling topic k in document d at moment t whereas the term on the right represents the probability of sampling word v for topic k in partition s_t .

Given the full conditional distributions above, we design a Gibbs sampling algorithm with an embedded Metropolis-Hastings step for model estimation, in which the variables \mathbf{S} , β , and \mathbf{z} are updated sequentially.

In the above MCMC estimation, we assume the hyperparameters η is fixed, which should be pre-defined in practice. As pointed by Ko et al. (2015), the value of η (i.e., λ and ζ) should influence the number of change points. Therefore, in real applications, the values of λ and ζ should be estimated appropriately. To address this issue, we follow Ko et al. (2015) to estimate λ and ζ . Specifically, first assume vague Gamma priors for λ and ζ . Then, the DPHMM prior reduces to the generalized Dirichlet distribution and the posterior distribution can be constructed. Finally, by solving the posterior distribution using the Newton–Raphson method, we can get the maximum-a-posteriori (MAP) estimates for λ and ζ . Except for the MAP method, one can also apply the Metropolis-Hastings sampler for estimation.

4 Experiments on synthetic data

4.1 Experimental setup

To demonstrate the finite sample performance of the Topic-CD model, we present a variety of experiments on synthetic data. To model the topic changes underlying dynamic documents, the time span T should not be too short. In real applications, it is also difficult to observe a long time span. Therefore in all experiments, we assume the total number of moments $T = (60, 100)$ to account for different scenarios of time span. As for the size of vocabulary, we set $V = 1000$ for illustration. A larger V can also be considered in the synthetic experiments, but would increase the computation cost. In the t -th moment ($t = 1, \dots, T$), the number of documents D_t is fixed as 100. For the d -th document in the t -th moment, the number of words N_{td} is initially generated from a Poisson distribution with parameter $\kappa_w = 100$ and then increased by 100 to prevent the inclusion of documents without the required number of words.

Regarding the number of change points, we consider three scenarios. In the first, we assume that the total number of moments is $T = 60$. Then we assume that there is only one change point (i.e., $Q = 1$), which exists at moment $t = 30$. The hyperparameters β in the two partitions separated by this change point are set as $\beta = (0.1, 1)^\top$. In the second scenario, the total number of moments is also assumed as $T = 60$. Then, we assume that the number of change points is $Q = 2$, whose positions are at moments $t = 20$ and $t = 40$. The hyperparameters β in the three partitions separated by two change points are set as $\beta = (0.1, 1, 10)^\top$. In the last scenario, assume that the total number of moments is $T = 100$. Further assume that there are $Q = 3$ change points, which are at moments $t = 25, 50, 75$, respectively. The corresponding hyperparameters in the four partitions are set as $\beta = (0.1, 1, 10, 50)^\top$.

In each of the three described scenarios, we consider two settings for the number of topics, i.e., $K = (10, 20)$, which is consistent with the three real applications in Sect. 5. As a result, there are a total of $3 \times 2 = 6$ experimental setups. In each experimental setup, with the positions of change points and β fixed, we can then generate the topic probabilities ϕ_{ik} for each topic under each partition from Dirichlet distributions. Next, we generate each specific document. For experimental setups with $K = 10$, we assume the hyperparameters α_k with $1 \leq k \leq K$ generated from the uniform distribution $U(0.05, 0.15)$; while under the experimental setups with $K = 20$, we fix $\alpha = (0.1, \dots, 0.1)^\top$. Then, for each document d at moment t , we generate θ_{td} from the Dirichlet distribution with hyperparameter α . Finally, each document can be generated according to the generative process shown in Sect. 3.1.

4.2 Evaluation metrics and comparison methods

For a reliable evaluation, we repeat the experiment $B = 100$ times. To assess the performance of the Topic-CD method, we compare it with the two-stage methods in topic change point detection. Specifically, in the two-stage methods, the dynamic topic model (Blei and Lafferty 2006) is first applied on the documents to obtain the multiple topic probability sequences. Then, two distance measurements are applied to

the topic-word distributions to evaluate the topic changes for each adjacent time period. They are the cosine similarity (Bruggemann et al. 2016) and the Jensen-Shannon divergence (Lau et al. 2012; Wang and Goutte 2018). Finally, to detect the change point based on the distances of each topic in adjacent time periods, three methods are applied: (1) setting the threshold, that is, if the distance of topic-word distributions is larger than the threshold, then the corresponding time moment is regarded as a change point (Bruggemann et al. 2016); (2) the dynamic programming method for offline change point detection (Truong et al. 2020); and (3) the binary segmentation method for offline change point detection (Truong et al. 2020). Together with the 2 distance measurements plus the 3 change point detection methods, there are a total of $2 \times 3 = 6$ two-stage methods for comparison. It is notable that, in each two-stage method, the number of change points is regarded as an input parameter, which should be defined in advance. To meet this requirement, we assume that the true Q is already known for these two-stage methods. To focus on the change point detection performance, we assume the true number of topics K and the true hyperparameters α are already known in all the methods. To estimate the Topic-CD model, we assume $\lambda = \zeta = 1$ for simplicity purpose. It is notable that, the hyperparameters λ and ζ can also be estimated using the MAP method, as which we do in real data analysis.

We first investigate the performance of Topic-CD in detecting the number of change points (i.e., Q). Let $\hat{Q}^{(b)}$ denote the estimate for Q in the b -th replication ($b = 1, \dots, B$) and $\hat{\tau}_1^{(b)}, \dots, \hat{\tau}_{\hat{Q}^{(b)}}^{(b)}$ denote the corresponding estimated locations for all change points. To evaluate the detection performance of our Topic-CD method, we calculate the average number of detected change points (AvgNum) and the correctly detected percentage of the number of change points (PerNum),

$$\text{AvgNum} = \frac{1}{B} \sum_{b=1}^B \hat{Q}^{(b)}, \quad \text{PerNum} = \frac{1}{B} \sum_{b=1}^B I(\hat{Q}^{(b)} = Q^{(b)}).$$

Then, we focus on the location of the detected topic change points. To evaluate the accuracy of the locations, we calculate the precision and recall of the detected change points. Denote the location set of detected change points as $\hat{\Omega}^{(b)} = \{\hat{\tau}_1^{(b)}, \dots, \hat{\tau}_{\hat{Q}^{(b)}}^{(b)}\}$. Accordingly, the location set of true change points is $\Omega^{(b)} = \{\tau_1^{(b)}, \dots, \tau_{Q^{(b)}}^{(b)}\}$. Define $M^{(b)} = \sum_{q=1}^Q I(\hat{\tau}_q^{(b)} \in (\tau_q - h, \tau_q + h))$ as the number of correctly detected locations, where $I(\cdot)$ is the indicator function and h is the bandwidth. Define $|\cdot|$ as the function counting the number of items in a set. Then, the precision and recall of detected change points are defined as follows,

$$\text{Precision} = \frac{1}{B} \sum_{b=1}^B (|M^{(b)}|/|\hat{\Omega}^{(b)}|), \quad \text{Recall} = \frac{1}{B} \sum_{b=1}^B (|M^{(b)}|/\Omega^{(b)}).$$

Furthermore, we use perplexity to evaluate the topic modeling performance. The measure of perplexity is first proposed by Blei et al. (2003). It is a commonly used measure to evaluate the predictive ability of a model. The lower the perplexity, the better the

model. Define $p(k|d)$ as the document-topic distribution and $p(w|k)$ as the topic-word distribution. Then the likelihood of document d is $p(w) = \sum_{k=1}^K p(k|d)p(w|k)$. Further define N_d is the number of words in document d . The perplexity is defined as follow

$$\text{perplexity} = \exp \left(- \frac{\sum_{d=1}^D \sum_{w=1}^V \log(p(w))}{\sum_{d=1}^D N_d} \right).$$

Finally, inspired by Lan et al. (2013), we apply two measures P_k and WindowDiff to evaluate the change point detection performance. The measure P_k is introduced by Beeferman et al. (1999). It first chooses a bandwidth $h = T/2$, and then applies a moving window with bandwidth h on the whole time span. For each fixed window, check whether the partition statuses of two ends of the window have been correctly detected. Then the score P_k measures the incorrect proportions when moving the window along the time span. The measure WindowDiff is an extension of P_k . It also applies a moving window on the time span. Then in each fixed window, it calculates the number of time points having incorrectly estimated partitions during the window. The detailed description of P_k and WindowDiff can be found in Pevzner and Hearst (2002). The lower the two measures, the better the detection performance.

4.3 Experimental results

We first report the experimental results for detection of the number of change points under different scenarios. It is notable that the number of change points should be pre-defined for two-stage methods. Therefore, we only report the detection results of the Topic-CD method. In the first scenario with a single change point, the average number of detected change points are 1.00 and 1.02 under $K = 10$ and $K = 20$, respectively. In addition, in nearly all replications, the change point has been correctly detected because the PerNum are 100% and 98% under $K = 10$ and $K = 20$, respectively. These results suggest that, the Topic-CD model is effective under the scenario of a single change point. In the second scenario with two change points, the AvgNum are 2.30 and 1.90 under $K = 10$ and $K = 20$, respectively. The corresponding PerNum results are 74% and 82%. Consequently, the Topic-CD model also achieves good performance in this scenario. However, in the third scenario with three change points, the detection performance of Topic-CD worsens because the situation is more complicated. Specifically, the AvgNum under $K = 10$ and $K = 20$ are 2.64 and 2.32, both of which are smaller than the true number of change points. As a result, the corresponding values of PerNum are 64% and 50%.

We then focus on the estimation performance of the locations of detected change points. We compare the Topic-CD model with six state-of-the-art methods, which have been described in Sect. 4.2. To denote these methods, we use ‘‘CS’’ and ‘‘JS’’ to represent the two distance measures, i.e., cosine similarity and Jensen-Shannon, respectively. We refer to the three offline change point detection methods, i.e. the dynamic programming, binary segmentation, and threshold method as ‘‘BS’’, ‘‘DP’’, and ‘‘T’’, respectively. Then, the six two-stage methods are referred to as the combi-

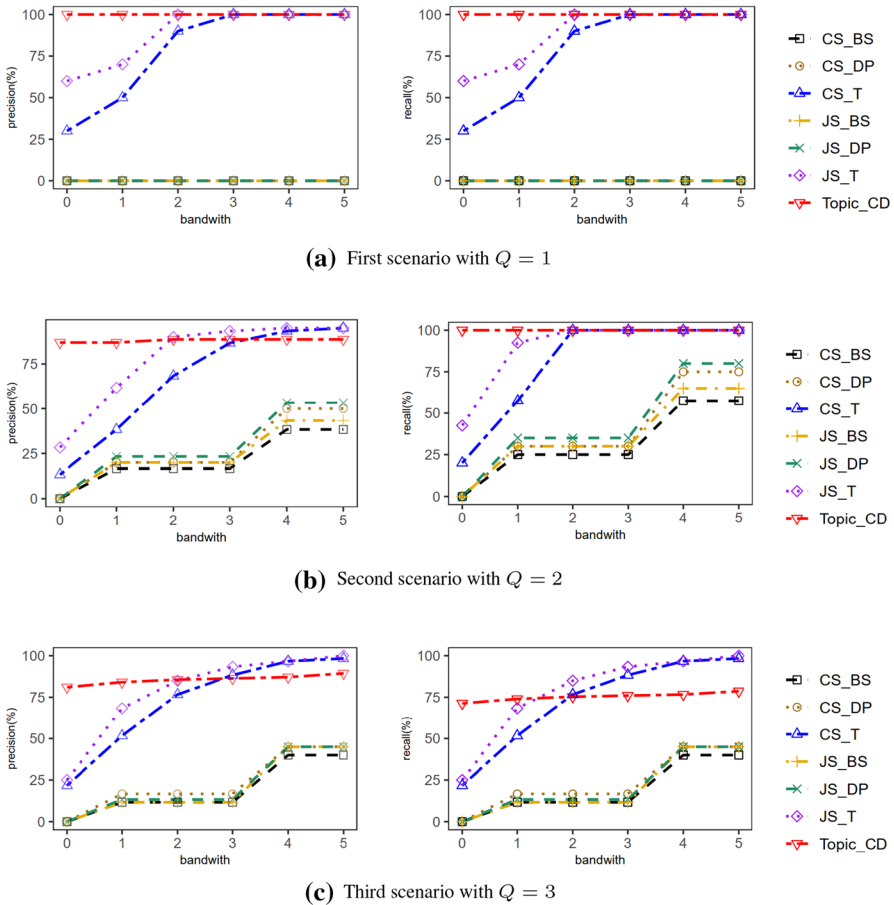


Fig. 2 The precision and recall for different methods on synthetic data with $K = 10$ topics

nation of one distance measure and one detection method. For example, the method “CS_BS” means the two-stage method using the cosine similarity measure and the dynamic programming algorithm for change point detection. To allow flexibility of the detected change points, we consider different bandwidths. Specifically, we set the bandwidth $h = 0, 1, 2, 3, 4, 5$. For $h = 0$, the precision and recall evaluate the performance of “accurately detected” change point locations. For $h > 0$, we allow the detected change points to lie in a flexible interval centered by the true location.

Figures 2 and 3 present the experimental results for the detection of change point locations in three scenarios with $K = 10$ and $K = 20$, respectively. From these results, we can draw the following conclusions. First, in different scenarios, we target on the estimation accuracy (i.e., the precision and recall) obtained by different methods under the bandwidth $h = 0$. It is clear that, among all the methods, only the Topic-CD model can achieve relatively high precision and recall. This finding implies that only our proposed Topic-CD model detects the change

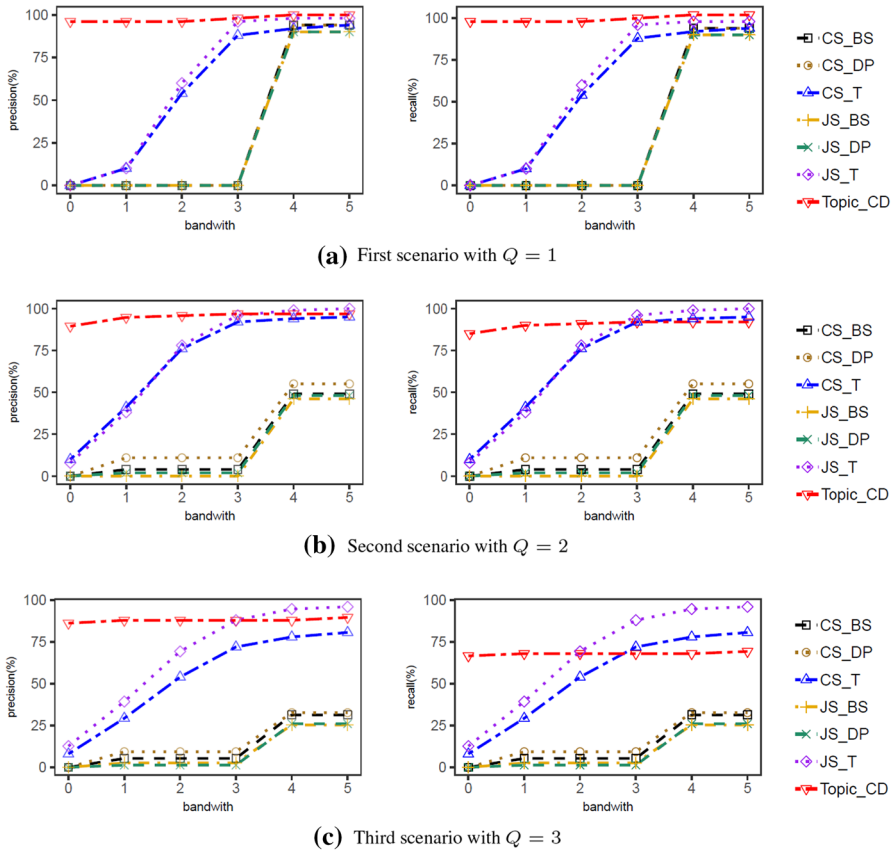


Fig. 3 The precision and recall for different methods on synthetic data with $K = 20$ topics

points precisely in most situations. Second, we focus on the influence of bandwidth on the detection performance. We find that, as the bandwidth h increases, the precision and recall of all methods improve. This is natural, because a larger $h > 0$ indicates more flexible conditions for evaluation of change point locations. Finally, we compare the detection performance of different methods with $h > 0$. As shown by Figs. 2 and 3, even with an increasing h , the Topic-CD model has obvious advantages over its competitors. Specifically, in the first scenario with one change point, both the precision and recall obtained by Topic-CD are approximately 100%. In scenarios of multiple change points, the precision and recall obtained by Topic-CD are lower, but they still behave better than most two-stage methods.

Finally, we present the experimental results measured by perplexity, P_k and WindowDiff in Table 1. The measure perplexity evaluates the topic modeling performance. In the six two-stage methods, the dynamic topic model is first built on the synthetic dataset, and then different two-stage methods are applied to detect the change points. Therefore, all two-stage methods share the same value of per-

Table 1 The model perplexity, P_k and WindowDiff of different methods in synthetic data

Measure	Topic-CD	CS_BS	CS_DP	CS_T	JS_BS	JS_DP	JS_T
$K = 10, Q = 1$							
Perplexity	558.901	954.9277					
Measure P_k	0.000	0.577	0.577	0.043	0.610	0.610	0.023
WindowDiff	0.000	0.577	0.577	0.043	0.610	0.610	0.023
$K = 10, Q = 2$							
Perplexity	819.299	1186.808					
Measure P_k	0.039	0.264	0.382	0.571	0.315	0.395	0.583
WindowDiff	0.233	0.989	1.092	1.202	1.052	1.122	1.226
$K = 10, Q = 3$							
Perplexity	784.185	1087.589					
Measure P_k	0.000	0.172	0.199	0.423	0.205	0.145	0.408
WindowDiff	0.270	0.924	0.949	1.400	0.959	0.889	1.405
$K = 20, Q = 1$							
Perplexity	656.262	717.4905					
Measure P_k	0.003	0.153	0.153	0.091	0.163	0.163	0.079
WindowDiff	0.018	0.153	0.153	0.091	0.163	0.163	0.079
$K = 20, Q = 2$							
Perplexity	858.764	876.3759					
Measure P_k	0.064	0.138	0.164	0.352	0.135	0.142	0.339
WindowDiff	0.147	0.23	0.248	0.385	0.228	0.228	0.382
$K = 20, Q = 3$							
Perplexity	833.228	900.0591					
Measure P_k	0.000	0.112	0.108	0.462	0.04	0.034	0.422
WindowDiff	0.426	0.754	0.742	1.462	0.58	0.615	1.404

plexity, which is calculated on the estimation results of the dynamic topic model. As shown in Table 1, the Topic-CD model has achieved lower perplexity than the two-stage methods in all experimental setups, which indicates better topic modeling performance in the Topic-CD model. We then focus on P_k and WindowDiff, two new measures evaluating the change point detection performance. It is obvious that, when compared with all two-stage methods, the Topic-CD model has obtained lower values of P_k and WindowDiff. As a result, the Topic-CD model also shows better change point detection performance when evaluated by P_k and WindowDiff. In summary, all the above results imply that, by combining topic models and change point detection in a unified framework, the Topic-CD model can improve the quality of topic learning and the change point detection performance.

5 Experiments on real data

5.1 Data description

We apply the Topic-CD model to three real datasets. The first dataset is Amazon cellphone reviews, which is public and can be downloaded from <http://jmcauley.ucsd.edu/data/amazon/links.html>. This dataset contains 194,000 reviews for cellphones, posted from April 2007 to July 2014. Each review contains the posting time, rating score (on a five-point scale from 1 = awful to 5 = excellent), the cellphone rated, and full textual content. After a preliminary analysis, we found that the total number of reviews in the earlier years was small. Thus, we only considered reviews in the last three years; that is, from August 2011 to July 2014. The second dataset is papers published in two statistical journals: *Journal of the Royal Statistical Society: Series B* and *Biometrics*. These two journals are typical examples in the fields of *statistics* and *biostatistics*. We crawled the information of papers published in the two journals from 2000 to 2019. The title, authors, published time, abstract, and key words were collected for each paper. We use all the abstracts as the text corpus. The third dataset is the UN General Debates data (Dieng et al. 2019), which can be downloaded from <https://www.kaggle.com/unitednations/un-general-debates>. This dataset contains statements from leaders and other senior officials in UN member states from 1970 to 2016. These statements present the perspectives of the governors on major issues in world politics.

We preprocess the three datasets with the following steps. Following the common practice in text mining, we first apply the *nltk* module in Python to remove punctuation, numbers, participles, and stop words. Subsequently, we remove low-frequency words that appear less than five times. It is notable that, large variation in document lengths should affect the performance of topic models. The lengths of the abstracts in the journal dataset and statements in the UN debate dataset are similar, but the review lengths in the Amazon dataset are quite different. Therefore, to maintain a reasonable range for review length, we remove the top 20% of the longest reviews and those with a length of less than 20 words. After preprocessing, the final Amazon dataset contains 19,247 reviews with a vocabulary size of 20,832 unique words, the final journal dataset contains 3188 abstracts with a vocabulary size of 4554 unique words, and the final UN debate dataset contains 7507 statements with a vocabulary size of 68,602 unique words.

5.2 Change point detection in the Amazon dataset

We apply the Topic-CD model to detect change points among common topics in all cellphone reviews. Because the user-generated Amazon reviews describe the true sentiments of consumers concerning the cellphones and services, the change points detected in the review stream reflect the changes in consumer preferences from 2011 to 2014. These findings are crucial for cellphone manufacturers and Amazon to design better products and services through understanding consumer preferences.

To apply the Topic-CD model, we organize the Amazon dataset in a monthly format, which results in 36 months in total. The number of documents in each month is shown

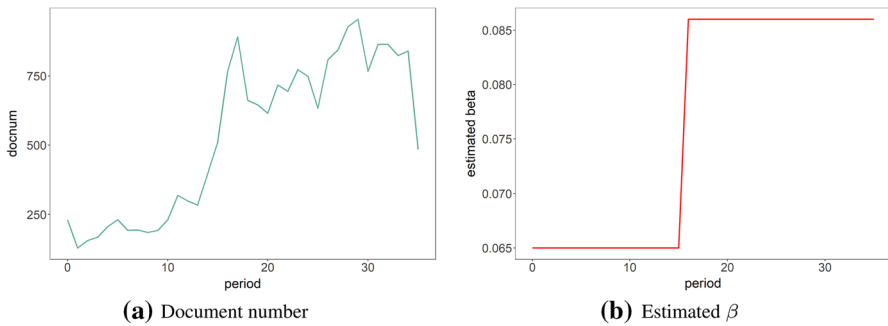


Fig. 4 The left panel is the trend of the number of documents in each month for the Amazon dataset from August 2011 to December 2012. The right panel is the estimated values of β in each month are also reported. As shown, there exists one change point at $t = 16$ (December 2012)

in Fig. 4a, which presents an obvious increasing trend from August 2011 to December 2012. To estimate the Topic-CD model, we set the number of topics as $K = 10$, the hyperparameters $\alpha = (0.1, \dots, 0.1)$, $b_0 = 0.01$ and $b_1 = 0.1$. The hyperparameters λ and ζ are estimated using the MAP method described in Sect. 3.2. After model estimation, the Topic-CD model has detected one change point at moment $t = 16$ (i.e., December 2012), which is approximately the moment with the greatest number of documents. The estimated values of β in each month are shown in Fig. 4b. As shown, before the change point, the estimated β is 0.042, while after the change point, the estimated β has increased to 0.054. The larger β value after the change point implies that there are more diversified topics. For an intuitive understanding, the number of documents and vocabulary size after the change point are relatively larger than those before the change point. In other words, the reviews become more abundant after the change point. Therefore, the associated topics become more diversified.

We then investigate the use of words before and after the change point. In total, there are 3897 newly appearing words after the detected change point, which we refer to as the *appearing vocabulary*. The left panel in Fig. 5 shows the top fifteen words with highest frequencies in the appearing vocabulary. As shown, the high-frequency words include “powerpack,” which represents power bank, and “oxa,” “bolse,” “tylt,” and “maxboost,” which are manufacturers of power banks and cell phone accessories. Therefore, after the change point, power bank become the new field for cell phone accessories. On the contrary, there are 334 words that no longer appear after the change point, which we refer to as the *disappearing vocabulary*. The right panel in Fig. 5 shows the top fifteen words with highest frequencies in the disappearing vocabulary. We find that most high-frequency words, such as “liveaction,” “easygo,” and “dvp” represent mobile accessories that become redundant.

To illustrate the changes in word frequencies more clearly, we take six words (i.e., “Samsung,” “Galaxy,” “Power,” “Droid,” “Evo” and “Motorola”) as examples and present the trend of the count of their appearance in the documents of each month. Figure 6 shows the corresponding results. It is obvious that, the words “Samsung,” “Galaxy,” and “Power” have increasing appearance count trends, even after the moment of the topic change point. This is because, the phone brand *Samsung* with its star prod-

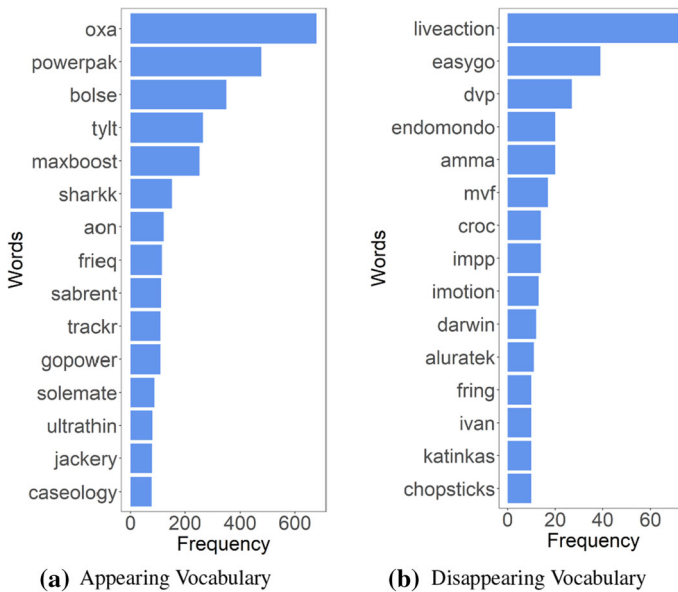


Fig. 5 The top fifteen words with highest frequencies in the appearing vocabulary and disappearing vocabulary after the change point in the Amazon review dataset

uct *Galaxy* become more famous after the moment of change point. In addition, with the increasing phone usage, the power of phones has been placed more attentions. Typical example reviews after the topic change point include: “This charger woks wonderfully on the Samsung Galaxy S4”, “They both have Samsung Galaxy phones” and “battery of my Samsung Galaxy S2”. On the contrary, the words “Droid”, “Evo,” and “Motorola” have decreasing appearance count trends after the moment of the topic change point. Take “Motorola” as an example. Its decreasing appearance counts indicate the decline of this phone brand. Typical review examples include: “The Motorola T505 doesn’t come with an AC charger”, and “Motorola only includes a car charger”.

Finally, we investigate the topics extracted before and after the change point. These include *brands*, *earphones*, *battery*, *USB adapters*, *phone cases*, and *automotive devices*. However, as stated above, the topics after the change point are more diverse. Therefore, some new topics appear after the change point, such as *tablet* and *power bank*. In addition, even for topics that have the same meaning before and after the change point, the associated high-frequency words under the topics have changed.

Table 2 presents some example topics extracted before and after the change point. As shown in Table 2, Topics 1 and 2 before and after the change point are all related to phone brands. However, the high-frequency words before the change point include “blackberry,” “motorola,” and “virgin,” whereas those after the change point include other cellphone names, such as “samsung” and “galaxy.” For Topic 3, the meaning changed from *handset* to *power bank* given the high-frequency words under each topic.

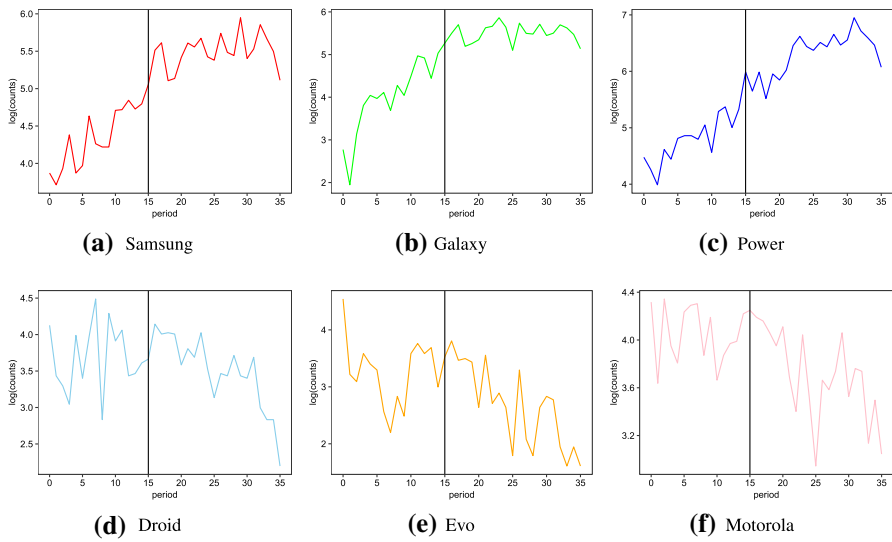


Fig. 6 The appearance count trend (in logarithm) of six example words in each month in the Amazon dataset. The dark vertical line indicates the topic change point moment (i.e., $t = 15$)

Table 2 Comparison of some example topics extracted before and after the change point for the Amazon dataset

Period	Topic	Words with high probabilities
Aug. 2011–Oct. 2012	1	Blackberry , like, bold, screen, device, stylus, keyboard, camera, rim...
	2	Virgin , motorola , device, great, price, mobile, optimus, service...
	3	Handset , iphone, device, base, cell, like, design, hold, hand, button...
Nov. 2012–Jul. 2014	1	Galaxy , samsung , screen, battery, android, camera, card, apps, storage...
	2	Samsung , quality, galaxy , price, black, good, design, box, packaging, cover, packaging...
	3	Power , battery, charge, mah, bank , usb, capacity, led, charger, cable, pack, small, external,...

5.3 Change point detection in the journal dataset

Papers published in top journals often discuss cutting-edge topics in statistics. Therefore, detecting the change points in the stream of papers can reflect the development and changing trend in the discipline of statistics. To this end, we apply the Topic-CD model to the journal dataset. Specifically, we organize the journal dataset in a yearly format, which leads to 20 years in total. The number of documents in each year is shown in Fig. 7a, which presents a relatively stable pattern. To apply the Topic-CD model, we set the number of topics as $K = 15$. Other settings of hyperparameters

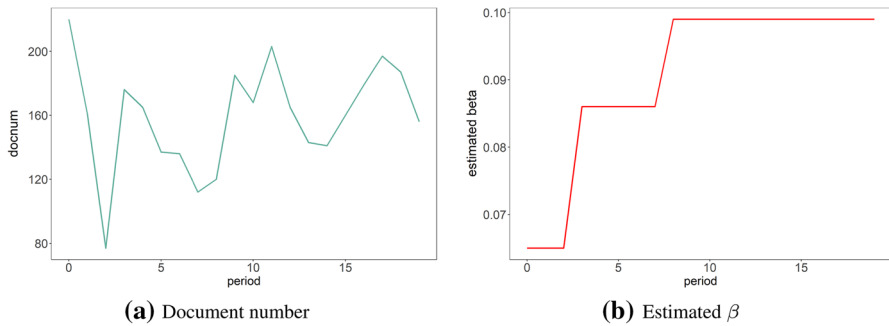


Fig. 7 The left panel is the trend of the number of documents in each year for the Journal dataset from 2000 to 2019. The right panel is the estimated values of β in each year are also reported. As shown, there exists two change points at $t = 3$ (year 2003) and $t = 8$ (year 2008)

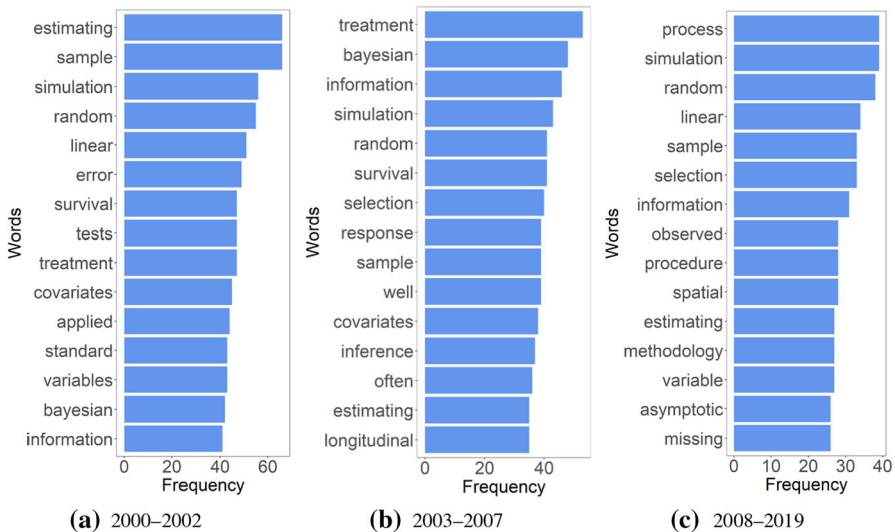


Fig. 8 The top fifteen words with highest frequencies in the 2000–2002, 2003–2007, and 2008–2019 periods for the journal dataset

are similar with the Amazon dataset. As a result, the Topic-CD model detected two change points at $t = 3$ (year 2003) and $t = 8$ (year 2008). The estimated β in each year is also shown in Fig. 7b. Specifically, the estimated β s before and after each detected change point are 0.065, 0.086, and 0.099. As we mentioned before, a larger β value implies more diversified topics. Therefore, the topics discussed in the statistical papers become more diversified as time goes by.

To illustrate the content changes before and after each change point, we first explore the high-frequency words in the three partitions split by the two change points. Figure 8 presents the top fifteen words with the highest frequencies in the periods 2000–2002, 2003–2007, and 2008–2019. As one can see, there are some words shared by all three time periods, such as “estimating,” “sample,” “simulation,” “treatment,” and

Table 3 Comparison of the first four topics extracted in the three time periods split by the two estimated change points at 2003 and 2008 for the journal dataset

Period	Topic	Words with high probabilities
2000–2002	1	Survival , inference, edition, ROC, sample , editors, statistics, population,...
	2	Test , procedure, conditional, group, estimating, asymptotic, censoring, simulation,...
	3	Process, generalized , test, linear , responses, count, multivariate ,...
	4	Sample , size, local, population, sampling , design , first, bias,...
2003–2007	1	Bayesian , variable, size, treatment , sample, outcome , response, expression,...
	2	Population, inference, treatment , series, disease, cancer, individual, species,...
	3	Tests, subjects, exact, design , latent, treatment , expression, response ,...
	4	Work, tests, prior , genes, Bayesian , measures, treatment , estimating,...
2008–2019	1	Treatment , longitudinal , outcome, random, exposure, patient , simulation, first,...
	2	Selection , variable , information, simulation, prior , gene, Bayesian , sample,...
	3	Spatial , Bayesian , random, selection , well, important, applied, methodology,...
	4	High , sparse , dimensional , real, algorithm, linear, properties,...

“Bayesian.” These words reflect the themes that have always been valued by statistical researchers. By contrast, we also find big changes in the top words in the three periods. For example, the 2000–2002 period also discusses “tests,” “error,” and “linear models,” which are basic issues in statistics. In the 2003–2007 period, researchers begin to discuss “variable selection” and “longitudinal data”. In the last period (2008–2019), the top words include “spatial,” “methodology,” “asymptotic,” and “missing,” which indicate new themes of concern in the last decade.

Finally, we investigate the content differences before and after each change point from the perspective of topics. For illustration purpose, we focus on the first four topics estimated by the Topic-CD model. Then we present the top words with the highest probabilities under each topic in the three partitions split by the two estimated change points. The corresponding results are shown in Table 3. By summarizing the meanings of the top words under each topic, we can characterize the meaning of the topic. As shown, the meanings of the first four topics in the three periods are quite different. Specifically, the topics during the 2000–2002 period discuss survival analysis, generalized linear models, sampling methods, and experimental design. These topics are classical problems in statistics and have been well studied in the early development of statistics. During the 2003–2007 period, the first four topics all discuss Bayesian analysis, causal inference, and developing methods for medicine and epidemiology. As for the last period (2008–2019), the topics become more diverse. Except for the topic of causal inference, more topics such as variable selection, spatial analysis, and high-dimensional analysis have appeared, which are still the focus of current statis-

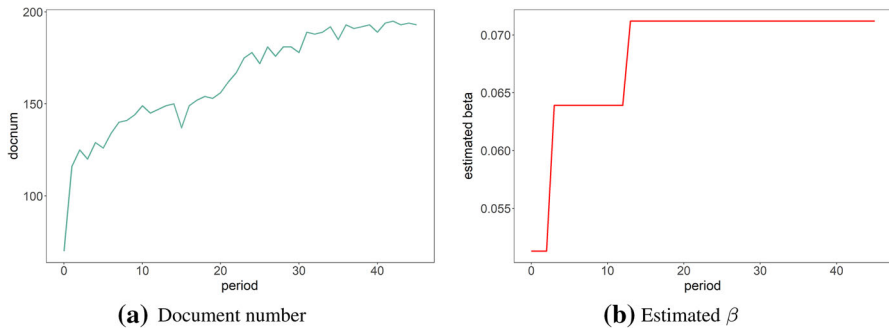


Fig. 9 The left panel is the trend of the number of documents in each year for the UN debate dataset from 1970 to 2016. The right panel is the estimated values of β in each year are also reported. As shown, there exists two change points at $t = 3$ (year 1973) and $t = 13$ (year 1983)

tical research. All the above findings verify that the Topic-CD model displays good detection performance.

5.4 Change point detection in the UN debate dataset

The representatives of UN member states gather at the annual sessions of the United Nations General Assembly. The centerpiece of each session is the General Debate. The statements of representatives are akin to the annual legislative state-of-the-union addresses in domestic politics. Detecting the change point in the stream of statements can reflect the changes of worldwide political concerns and focuses. To investigate the changes of statements, we apply the Topic-CD model to the UN debate dataset. Specifically, we organize this dataset in a yearly format, which leads to 46 years in total. The number of documents in each year is shown in Fig. 9a, which presents an obvious increasing trend from 1970 to 2016. We apply the same experimental settings as those used in the journal dataset and then estimate the Topic-CD model. As a result, the Topic-CD model detected two change points at $t = 3$ (year 1973) and $t = 13$ (year 1983). The estimated β in each year is also shown in Fig. 9b. Specifically, the estimated β s before and after each detected change point are 0.051, 0.064, and 0.071. The increase of estimated β indicates that, the topics discussed in the UN debate become more diversified as time goes by.

First, to illustrate the content changes, we explore the high-frequency words in the three partitions split by the two change points. Figure 10 presents the top fifteen words with the highest frequencies in the periods 1970–1972, 1972–1982, and 1983–2016, respectively. It is notable that, some words such as “people”, “development”, “economic” and “security” appear in all three time periods. These words reflect the common goals and pursuits of human beings. By contrast, there also exist word changes in the three periods. For example, the word “war” has high frequency in 1970–1972. The representatives also discuss a lot about “rights” and “principles”. In 1972–1982, “cooperation” and “relations” become important topics. As time goes by, with the enhanced national cooperations, common development become the goal of all

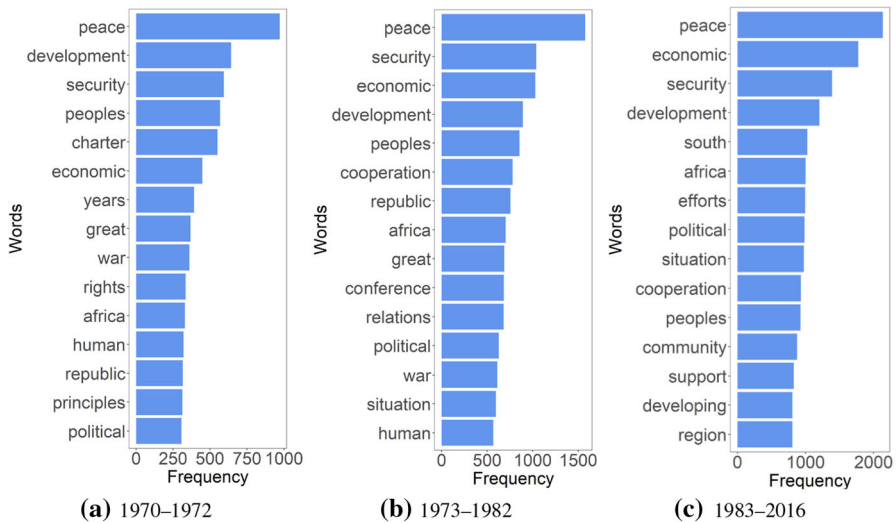


Fig. 10 The top fifteen words with highest frequencies in the 1970–1972, 1973–1982, and 1983–2016 periods in UN debate dataset

countries. Then the words “efforts”, “cooperation” and “community” have appeared in the period 1983–2016.

Finally, we discuss the content differences before and after each change point from the perspective of topics. For illustration purpose, we focus on three topics estimated by the Topic-CD model. We also present the top words with the highest probabilities in Table 4. We can see that the meanings of the three topics in the three periods are quite different. Specifically, the topics during the 1970–1972 period discuss war, aggression and resolution. These topics reflect that the world is not stable during this period. During the 1973–1982 period, the topics all discuss economic development. All countries are concerned about economic development and resources in this period. As for the last period in 1983–2016, the topics become more diverse, such as climate change, country cooperation and human right. With the development of economy, more and more problems need to be paid attentions and solved by the whole world. The above findings verify that the Topic-CD model has good applications in real practice.

6 Conclusion

In this study, we focus on the identification of topic change points during dynamic text streams. We define the problem of topic change point detection from the new perspective of hyperparameters β , which facilitates the detection of change points from the whole topic picture. Then, we propose a Topic-CD model to address the problem of topic change point detection. The strengths of Topic-CD model are reflected in two aspects. First, the Topic-CD model combines topic models and change point detection in a unified framework. Specifically, the LDA model is applied to extract topics underlying the dynamic text documents; whereas the LDA hyperparameters are

Table 4 Comparison of the three topics extracted in the three time periods split by the two estimated change points at 1972 and 1982 for the UN debate dataset

Period	Topic	Words with high probabilities
1970–1972	1	Peoples, republic, Soviet, struggle, Vietnam, aggression , security, independence ,...
	2	Peace, republic, Arab, security, aggression, war , charter, forces ,...
	3	Israel , peace, security, resolution , council, Arab, east, agreement...
1973–1982	1	Human, rights , political, peace, social, American, economic , Latin,...
	2	Economic, developing, development , south, conference, system, developed, resources ,...
	3	Economic , Africa, development , delegation, peace, African, developing, community,...
1983–2016	1	Development, climate, change, global , small, developing, pacific, island,...
	2	Security, human, cooperation , peace, council, development, rights, efforts,...
	3	Political, economic, peace, human, democracy , social, development, today,...

modeled by the DPHMM model to identify topic diversity changes. The LDA model and the DPHMM model are combined together and then estimated simultaneously. By using this way, the Topic-CD model can improve the quality of topic learning as well as the change point detection performance, when compared with previous two-stage methods. Second, the Topic-CD model does not require the number of change points to be set in advance, which makes it more convenient for practical use. For model estimation, we propose an MCMC algorithm. The finite sample performance of the Topic-CD model is numerically investigated using both synthetic data and three real datasets. Compared with the models in the past literature, the Topic-CD model is more suitable for integral analysis of all topics. It can be considered as the first step to test the existence of topic change points underlying a large collection of dynamic documents. Given the existence of topic change points discovered by the Topic-CD model, the researchers can further apply other change point detection methods to find the change point for each single topic or each single document.

However, the Topic-CD model also has some limitations, which can inspire further improvements in the future. First, in the Topic-CD model, the number of topics needs to be pre-specified and fixed. However, for dynamic text documents, the number of topics may also change over time. Therefore, a hierarchical Dirichlet process model can be further considered and combined with the Topic-CD model, to make the number of topics more flexible. Second, we define the topic change point from the perspective of the hyperparameter β , which controls topic meanings and topic diversity. In fact, the pattern of topic representations in each document can also change over time. Therefore, more hyperparameters (e.g., α) can also be modeled by Bayesian change point methods to help detect more accurate topic change points. Third, the Topic-CD model assumes that each topic has a probability distribution over the entire vocabulary. In fact, the used vocabulary for dynamic documents can be changed. Therefore, to make

the dynamic topics more focused on the time-specific vocabulary, a sparsity extension of Topic-CD model can be considered in future studies. Finally, the Topic-CD model uses the basic LDA model as its model foundation. In fact, more variants of topic models can be applied to handle more complex situations. For example, combing the topic segmentation method (Lan et al. 2013) with the Topic-CD model can help find topic change points for each single document.

Acknowledgements We thank Yichao Feng (now working in Jingdong company) and Yandi Zhu (now studying in Peking University) for their supports in data exploration. This work is also supported by National Natural Science Foundation of China (Nos. 72171229, 72001205, 11971504), fund for building world-class universities (disciplines) of Renmin University of China, Foundation from Ministry of Education of China (20JZD023), Ministry of Education Focus on Humanities and Social Science Research Base (Major Research Plan 17JJD910001).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

- Ahmed A, Xing EP (2008) Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: with applications to evolutionary clustering. In: Proceedings of the SIAM international conference on data mining. pp 219–230
- Ahmed A, Xing EP (2010) Timeline: a dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In: Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence. pp 20–29
- AlSumait L, Barabá D, Domeniconi C (2008) On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the 8th IEEE international conference on data mining. pp 3–12
- Bai J (1997) Estimation of a change point in multiple regression models. *Rev Econ Statist* 79(4):551–563
- Beeferman D, Berger A, Lafferty J (1999) Statistical models for text segmentation. *Mach Learn* 34(1–3):177–210
- Blei DM (2012) Probabilistic topic models. *Commun ACM* 55(4):77–84
- Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the twenty-third international conference on machine learning. pp 113–120
- Blei D, Mcauliffe JD (2008) Supervised topic models. *Adv Neural Inf Process Syst* 3:327–332
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Bruggermann D, Hermey Y, Orth C, Schneider D, Selzer S, Spanakis G (2016) Storyline detection and tracking using dynamic latent Dirichlet allocation. In: Proceedings of the 2nd workshop on computing news storylines (CNS 2016). pp 9–19
- Chae J, Thom D, Bosch H, Yun J, Maciejewski R, Ebert DS, Ertl T (2012) Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In: IEEE conference on visual analytics science and technology. pp 143–152
- Chib S (1998) Estimation and comparison of multiple change-point models. *J Econom* 86(2):221–241
- Dieng AB, Ruiz F, Blei DM (2019) The dynamic embedded topic model. [arXiv:1907.05545](https://arxiv.org/abs/1907.05545)
- Dubey A, Hefny A, Williamson S, Xing EP (2013) A nonparametric mixture model for topic modeling over time. In: Proceedings of the SIAM international conference on data mining. pp 530–538
- Greene D, Cross JP (2016) Exploring the political agenda of the European parliament using a dynamic topic modeling approach. *Polit Anal* 25(1):77–94
- Guo X, Xiang Y, Chen Q, Huang Z, Hao Y (2013) LDA-based online topic detection using tensor factorization. *J Inf Sci* 39(4):459–469
- Hasan M, Orgun MA, Schwitter R (2017) A survey on real-time event detection from the Twitter data stream. *J Inf Sci* 44(4):443–463

- He J, Chen X, Du M, Jiang H (2015) Topic evolution analysis based on improved online LDA model. *J Cent South Univ (Sci Technol)* 46(2):547–553
- Hoffman MD, Blei DM, Bach FR (2010) Online learning for latent Dirichlet allocation. In: International conference on neural information processing systems. pp 1–9
- Holz F, Teresniak S (2010) Towards automatic detection and tracking of topic change. In: The 11th international conference on computational linguistics and intelligent text processing. pp 327–339
- Ishwaran H, James LF (2001) Gibbs sampling methods for stick-breaking priors. *J Am Stat Assoc* 96(453):161–173
- Kawamae N (2011) Trend analysis model: trend consists of temporal words, topics, and timestamps. In: Proceedings of the fourth ACM international conference on web search and data mining. pp 317–326
- Ko SIM, Chong TTL, Ghosh P (2015) Dirichlet process hidden Markov multiple change-point model. *Bayesian Anal* 10(2):275–296
- Lan D, Buntine W, Johnson M (2013) Topic segmentation with a structured topic model. In: Proceedings of annual conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT). pp 190–200
- Lau J, Collier N, Baldwin T (2012) On-line trend analysis with topic models: Twitter trends detection topic model online. In: Proceedings of 24th international conference on computational linguistics. pp 1519–1534
- Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on information and knowledge management. pp 375–384
- Mohamad S, Bouchachia A (2019) Online Gaussian LDA for unsupervised pattern mining from utility usage data. [arXiv:1910.11599](https://arxiv.org/abs/1910.11599)
- Nallapati RM, Dittmore S, Lafferty JD, Ung K (2007) Multiscale topic tomography. In: International conference on knowledge discovery and data mining. pp 520–529
- Pevzner L, Hearst M (2002) A critique and improvement of an evaluation metric for text segmentation. *Comput Linguist* 28:1–19
- Pozdnoukhov A, Kaiser C (2011) Space-time dynamics of topics in streaming text. In: ACM Sigspatial international workshop on location-based social networks. pp 1–8
- Ramage D, Hall D, Nallapati R, Manning CD (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing. pp 248–256
- Sasaki K, Yoshikawa T, Furuhashi T (2014) Online topic model for Twitter considering dynamics of user interests and topic trends. In: Proceedings of the conference on empirical methods in natural language processing. pp 1977–1985
- Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. *J Am Stat Assoc* 101(476):1566–1581
- Truong C, Oudre L, Vayatis N (2020) Selective review of offline change point detection methods. *Signal Process* 167:107299
- Vavliakis KN, Tzima FA, Mitkas PA (2012) Event detection via LDA for the mediaeval 2012 SED task. In: Proceedings of the MediaEval 2012 workshop
- Wang Y, Goutte C (2018) Real-time change point detection using on-line topic models. In: Proceedings of the 27th international conference on computational linguistics. pp 2505–2515
- Wang X, McCallum A (2006) Topics over time: a non-Markov continuous-time model of topical trends. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. pp 424–433
- Wang C, Blei D, Heckerman D (2015) Continuous time dynamic topic models. [arXiv:1206.3298](https://arxiv.org/abs/1206.3298)
- Wu Q, Zhang C, Hong Q, Chen L (2014) Topic evolution based on LDA and HMM and its application in stem cell research. *J Inf Sci* 40(5):611–620
- Zhang Y, Chen H, Lu J, Zhang G (2017) Detecting and predicting the topic change of knowledge-based systems: a topic-based bibliometric analysis from 1991 to 2016. *Knowl Based Syst* 133:255–268
- Zhong N, Schweidel DA (2020) Capturing changes in social media content: a multiple latent changepoint topic model. *Mark Sci* 39(4):827–846
- Zhou X, Chen L (2014) Event detection over Twitter social media streams. *VLDB J* 23(3):381–400
- Zhou H, Yu H, Hu R (2017) Topic evolution based on the probabilistic topic model: a review. *Front Comput Sci* 11(5):786–802

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.