Journal of Machine Learning Research 26 (2025) 1-53 Submitted 11/23; Revised 12/24; Published 4/25

Fine-Grained Change Point Detection for Topic Modeling with Pitman-Yor Process

Feifei Wang FEIFEI.WANG@RUC.EDU.CN
Center for Applied Statistics
School of Statistics
Renmin University of China
Beijing, 100872, China

Zimeng Zhao ANNZHAO@RUC.EDU.CN
School of Statistics
Renmin University of China
Beijing, 100872, China

Ruimin Ye RUIMYE@TENCENT.COM
Game Security Department
Tencent Games
Shenzhen, 518057, China

Xiaoge Gu 2021201521GXG@RUC.EDU.CN
School of Statistics
Renmin University of China
Beijing, 100872, China

Xiaoling Lu XIAOLINGLU@RUC.EDU.CN
Center for Applied Statistics
School of Statistics
Renmin University of China
Beijing, 100872, China

Editor: Mohammad Emtiyaz Khan

Abstract

Identifying change points in dynamic text data is crucial for understanding the evolving nature of topics across various sources, such as news articles, scientific papers, and social media posts. While topic modeling has become a widely used technique for this purpose, capturing fine-grained shifts in individual topics over time remains a significant challenge. Traditional approaches typically use a two-stage process, separating topic modeling and change point detection. However, this separation can lead to information loss and inconsistency in capturing subtle changes in topic evolution. To address this issue, we propose TOPIC-PYP, a change point detection model specifically designed for fine-grained topic-level analysis, i.e., detecting change points for each individual topic. By leveraging the Pitman-Yor process, TOPIC-PYP effectively captures the dynamic evolution of topic meanings over time. Unlike traditional methods, TOPIC-PYP integrates topic modeling and change point detection into a unified framework, facilitating a more comprehensive understanding of the relationship between topic evolution and change points. Experimental evaluations on both synthetic and real-world datasets demonstrate the effectiveness of TOPIC-PYP in accurately detecting change points and generating high-quality topics.

Keywords: Change Point Detection, Dynamic Texts, Fine-Grained, Pitman-Yor Process, Topic Models

1. Introduction

With the rapid advancement of technology and the exponential growth of the Internet, text documents have become increasingly abundant across various fields, including news articles, user posts, consumer reviews, and so on. However, due to the overwhelming volume of textual data, it has become virtually impossible for individuals to effectively monitor and identify key points or track emerging trends and topics. As a result, there has been significant interest in developing automatic text summarization and change point detection methods that enable users to quickly comprehend shifts in events within text streams.

In the field of statistics, *change point detection* refers to a well-defined problem that involves identifying the specific moments when there is a significant change in the probability distribution of a stochastic process or time series (Bai, 1997; Chib, 1998). The goal is to accurately pinpoint the exact points in time when the underlying statistical properties of the process undergo a noticeable shift. However, detecting change points in text streams can indeed be challenging due to the unstructured nature of text documents. One plausible approach is to first model the text documents into several topics via the methods of topic models (Blei et al., 2003; Griffiths and Steyvers, 2004). Then change points can be defined from the perspective of topic changes. By analyzing the shifts in the topic-related distributions over time, it becomes possible to identify moments when there are significant changes in the themes or subject matter being discussed in the text stream.

The topic-assisted approach for text change point detection has gained attention as a promising method in recent years (Bruggermann et al., 2016; Zhang et al., 2017; Wang and Goutte, 2018; Zhong and Schweidel, 2020; Lu et al., 2022). In this regard, existing literature can be roughly divided into two streams. The first stream usually adopts a two-stage strategy to find the change points among topics. Specifically, the first stage is to build topic models for text streams while the second stage is to detect the change points using the obtained topic-word representations. In this way, the change points are defined from the perspective of topic semantics, which are represented by the topic-word distributions. In addition, by using the two-stage methods, one can detect the change points for each single topic, which we refer to as "fine-grained" topic change point detection. Typical studies in this stream include Bruggermann et al. (2016); Zhang et al. (2017); Wang and Goutte (2018). However, this two-stage strategy poses some potential issues. Firstly, it may lead to information loss, as the topic model learned in the first stage might not fully account for the influence of change points. Secondly, due to the separate nature of the two stages, the model may lack consistency in capturing subtle changes in the evolution of topics. This inconsistency could result in missing crucial topic variations when actual change points occur. Given the potential shortcomings of the two-stage methods, the second stream aims to develop a unified framework combining topic modeling and change point detection. A representative work is Lu et al. (2022), which introduces a Topic-CD model to detect obvious changes experienced by all topics, not individual topics. Other unified methods include Lan et al. (2013) and Zhong and Schweidel (2020). However, these two works define change

points from the perspective of document-topic representations. In other words, they mainly focus on detecting the changes in the discussed proportions of topics.

In this work, we aim to find the fine-grained changes for topics, i.e., detecting the change points among the evolution of each individual topic, rather than detecting the change points collectively across all topics. However, different from the commonly used two-stage strategy, we develop a unified automatic model for topic change point detection. This integrated approach facilitates a more comprehensive consideration of the relationship between topic evolution and change points, enhancing the model's sensitivity to the dynamic nature of text data. To this end, we adopt the Pitman-Yor process (PYP, Pitman and Yor, 1995). PYP is an extension of the Dirichlet process, which is commonly used in Bayesian modeling. It has two key properties. The first one is the *nonparametric nature*. PYP dynamically adjusts the number of clusters or categories as more data is observed, making it suitable for data with an unknown or growing number of components. The second one is the *power-law behavior*. It captures heavy-tailed distributions where a few categories (e.g., frequent words) dominate, while many categories (e.g., rare words) occur infrequently but remain important. These properties make PYP particularly popular for modeling topic-word distributions in topic models (Sato and Nakagawa, 2010; Lindsey et al., 2012; Lim et al., 2016; Guo et al., 2024). The nonparametric nature of PYP allows it to dynamically adapt to the increasing number of unique words over time. The power-law property of PYP enables it to model the heavy-tailed nature of word frequency distributions effectively, i.e., low-frequency words often play a crucial role in distinguishing between topics. Therefore, we identify the change points for each topic by modeling its topic-word distribution using PYP.

By incorporating the Pitman-Yor process, we develop TOPIC-PYP, a unified model for topic change point detection. Assume there are K topics underlying a dynamic textual corpus from $1 \leq t \leq T$. During this period, each topic k has Q_k change points. With the occurrence of each change point, the meaning of this topic, which is represented by the topic-word distribution (i.e., ϕ_k), is expected to change. To model the changing meanings of a specific topic during the period, we use the Pitman-Yor process to generate the topic-word distributions. Based on the meanings of topics, the generation of documents at each time stamp is modeled similarly with the classic latent Dirichlet allocation (LDA, Blei et al., 2003). In this way, topic modeling and change point detection are integrated into a unified framework and performed simultaneously. To estimate the TOPIC-PYP model, we employ the collapsed Gibbs sampling algorithm (Liu, 1994; Griffiths and Steyvers, 2004). Additionally, we provide theoretical guarantees regarding the posterior consistency and numerical convergence of TOPIC-PYP. The detection performance of the model is further evaluated through a series of experiments on both synthetic data and two real-world datasets. The results demonstrate that TOPIC-PYP can accurately detect change points and generate high-quality topics.

The remainder of this paper comprises five sections. Section 2 reviews the related literature. Section 3 presents the TOPIC-PYP model and then discusses its estimation algorithm and theoretical properties in detail. In Section 4, the finite sample performance of the TOPIC-PYP model is demonstrated through various experiments on synthetic data. In Section 5 and Section 6, the TOPIC-PYP model is applied to two real datasets. Section 7 presents the conclusions and a brief discussion.

2. Related Work

2.1 Probabilistic topic models for change point detection

In recent years, there has been significant interest in automatic text summarization and change point detection. One popular approach for these tasks is the utilization of topic models, which are a collection of three-level hierarchical Bayesian models. Latent Dirichlet allocation (LDA) is the most basic topic model (Blei et al., 2003), which has gained considerable attention and usage. LDA assumes there is a set of K topics underlying all documents. Each topic is characterized by a probability distribution over the dictionary of words (referred to as the topic-word distribution), and each document is characterized by a probability distribution over these latent topics (referred to as the document-topic distribution). As LDA is a static model, many researchers have extended LDA to dynamic topic modeling for text streams. Among various dynamic extensions of LDA, the most notable one is the dynamic topic model (DTM, Blei and Lafferty, 2006b), which applies the Gaussian process to model the evolution of topic-word distributions. Other important works include the topics over time model (TOT, Wang and Mccallum, 2006), the multiscale topic tomography model (MTTM, Nallapati et al., 2007), temporal Dirichlet process mixture model (TDPM, Ahmed and Xing, 2008), infinite dynamic topic model (iDTM, Ahmed and Xing, 2010), continuous time dynamic topic model (Wang et al., 2012), joint dynamic topic model (Zhu et al., 2022), and so on.

Although dynamic topic models can find topic evolution patterns, they mainly detect the *gradual changes* of topics. To detect the *sudden changes* or *obvious changes* of topics, the change point detection methods are often applied. An early work to find topic change points is Holz and Teresniak (2010), which utilized a volatility measure to identify contextual shifts in topics over time. Later on, Bruggermann et al. (2016); Zhang et al. (2017); Wang and Goutte (2018) combined topic models with change point detection methods. The basic idea of these works is to first apply topic models to extract the dynamic patterns of topic meanings; then some change point detection methods are adopted for each topic to find change points. For example, Bruggermann et al. (2016) employed a dynamic topic model to generate the dynamic sequences of each topic. Then they detected a change point when the distance of any two topic-word distributions in adjacent periods had exceeded a predefined threshold. Wang and Goutte (2018) focused on real-time change point detection. They first used the online LDA model to obtain the dynamic sequences of topic-word distributions and then applied online change point detection methods to identify the change points for each topic. Additionally, Rieger et al. (2022) developed a Rolling LDA method, leveraging rolling window techniques to detect changes in topics. In this way, it allows for the timely identification of shifts in topic-word distributions. Furthermore, Zhang and Lauw (2022) extended dynamic topic modeling to account for the document networks over time, enabling a comprehensive analysis of topic evolution and change point detection within network structures.

It is notable that, the above methods are all two-stage approaches. That is, they conduct topic modeling and topic change point detection separately in two steps. However, as topic change point is a significant feature of documents, combining topic modeling and change point detection would enhance the ability to summarize document meanings and capture important change points, leading to a more comprehensive understanding of the textual

data. In this regard, Lu et al. (2022) developed a unified model called Topic-CD for topic change point detection. They assumed the change points existed for the hyperparameter β_t, which controlled the generation of topic-word distributions. The changing pattern of β_t was further modeled by the Dirichlet process hidden Markov multiple change-point (DPHMM) process (Ko et al., 2015). However, Topic-CD aims to detect change points for the whole topic set, not for each single topic. In other words, it implicitly assumes all topics share the same change points, which seems too strong and might be not realistic in practice. Another unified work is Zhong and Schweidel (2020), which also adopted DPHMM to model change points. However, it focuses on the hyperparameter α_t, which controls the generation of document-topic representations.

2.2 Neural topic models for change point detection

Traditional topic models are primarily based on probabilistic frameworks. In recent years, with advancements in artificial intelligence, neural topic models (NTMs) that utilize neural network architectures have gained significant popularity due to their superior ability to enhance topic representation and adaptability. For example, the neural variational document model captures latent topic distributions through encoder-decoder frameworks (Miao et al., 2016). The adversarial-neural topic model improves topic learning through generative adversarial networks (Wang et al., 2019). To model dynamic documents, Gupta et al. (2019) developed a neural autoregressive topic model by incorporating neural language models such as RNN and LSTM. Dieng et al. (2019) extended the classic DTM by using word embeddings and developed the D-ETM model. Cvejoski et al. (2023) introduced a neural dynamic focused topic model to dynamically adjust topic focus based on context, enabling a refined understanding of topic representations in evolving text. In a similar vein, Kellert and Zaman (2022) employed the BERTopic model to track the contextual shifts of words during the COVID-19 pandemic. Miyamoto et al. (2023) proposed a dynamic structured neural topic model (DSNTM) that leverages a self-attention mechanism to capture temporal dependencies among topics, enabling the modeling of topic branching and merging process. Wu et al. (2024a) introduced a chain-free dynamic topic model (CFDTM) that uses evolution-tracking contrastive learning and word exclusion to capture topic evolution while addressing repetitive and unassociated topics. Rahimi et al. (2024) proposed the aligned neural topic model (ANTM), which uses pre-trained transformer embeddings, and an overlapping sliding window for temporal document clustering. More discussions about NTMs can be found in the survey paper of Wu et al. (2024b).

Overall, NTMs have the advantage of capturing complex, non-linear patterns and incorporating rich contextual information, especially when training on large-scale text data. However, to the best of our knowledge, there are currently no neural topic models specifically designed for the task of topic change point detection. To take advantage of dynamic neural topic models for change point detection, a two-stage approach should be also required. This gap highlights an important distinction between NTMs and our proposed TOPIC-PYP method. First, unlike existing NTMs, our model explicitly incorporates mechanisms to detect change points, combining topic modeling and change point detection in a unified framework. This approach aligns more closely with the generation process of dynamic documents with topic change points. Second, while NTMs rely on embeddings that can

sometimes lack transparency, our method retains the interpretability inherent in classical probabilistic topic models. Last, our model often requires fewer training resources compared to NTMs, the latter of which often involve complex neural architectures. However, as a probabilistic topic model, TOPIC-PYP also has some limitations compared to NTMs, especially in capturing complex relationships and handling large-scale dynamic documents. Thus how to extend TOPIC-PYP by incorporating modern deep learning techniques is an important direction in future work.

2.3 Topic models using PYP

The Pitman-Yor process (PYP) is a nonparametric Bayesian process that extends the Dirichlet process by introducing an additional "discount" parameter (Pitman and Yor, 1995). This parameter enables PYP to effectively capture heavy-tailed word frequency distributions (i.e., the power-law phenomenon) in topic models. Furthermore, its nonparametric nature allows PYP to remain flexible when encountering new vocabulary in dynamic documents. These properties make PYP particularly well-suited for capturing topic dynamics and word distribution patterns in documents. As a result, PYP has become increasingly popular in topic modeling. For example, the Pitman-Yor topic model (PYTM) combines PYP and LDA to better reflect the power-law distribution observed in natural language, in which certain words and topics can become more prevalent (Sato and Nakagawa, 2010). The phrase-discovering LDA (PDLDA) uses the hierarchical Pitman-Yor process (HPYP) to uncover interpretable topical phrases while capturing linguistic dependencies and non-compositional structures (Lindsey et al., 2012). Buntine and Mishra (2014) reinforced PYP's advantages in capturing the intricate and hierarchical structures of topics in the text, highlighting the potential of nonparametric approaches for handling complex data patterns in textual corpora. The hierarchical Pitman-Yor topic model (HPYTM) extends PYTM to a multi-level structure, allowing the number of topics to vary across documents (Lim et al., 2016). Integrating PYP with the hidden Markov topic model, Guo et al. (2024) further combined power-law behavior with temporal topic dependencies, enhancing the model's ability to represent inter-topic relationships. The above studies demonstrate the utility of PYP in topic modeling. In this work, to enhance the detection of topic change points, we also leverage the strengths of PYP to model the generation of topic-word distributions.

3. Methodology

3.1 Model description

Suppose there exists a dynamic document corpus observed over T time points. At each time point t ($1 \leq t \leq T$), we observe D_t documents. Consequently, the total number of documents is $D = \sum_{t=1}^{T} D_t$. Further assume there are a total of K topics underlying all D documents. Although the number of topics is fixed for all T moments, the topic meanings could change over time. In this work, the change point detection problem is defined for each single topic, i.e., we allow each topic to have its own change points. Therefore, we aim to conduct "fine-grained" change point detection, which is in line with the past literature (Bruggermann et al., 2016; Wang and Goutte, 2018). Below, we first describe the generation process of change points.

Assume each topic has a topic shift probability π_k with $1 \leq k \leq K$, which describes the possibility that topic k could change over time. We assume the shift probability π_k follows a Beta distribution with hyperparameters λ_0 and λ_1. Then for the t-th moment, a shift indicator $i_{k,t}$ (valued 0 or 1) is used to characterize whether topic k has experienced a change point or not. Assume the shift indicator is generated from a Bernoulli distribution with parameter π_k. Let $i_k = (i_{k,1}, ..., i_{k,T})^\top \in \mathbb{R}^T$ denote the shift indicator vector for topic k. Given the topic shift indicators, the total number of change points for topic k is $Q_k = \sum_{t=1}^{T} i_{k,t}$, which partitions the T moments into $S_k = Q_k + 1$ segments. Further define $s_{k,t} = \sum_{j=1}^{t} i_{k,j} + 1$ as the index of segment at time point t for topic k. For easy understanding, consider a toy example illustrated in Figure 1. Assume there is a time period with $T = 7$ moments. Assume a certain topic k has the shift indicator $i_k = (0, 1, 0, 0, 1, 0, 0)^\top$, which means it has two change points occurring at $t = 2$ and $t = 5$. The two change points split the whole time period into $S_k = 3$ segments. The segment index vector is computed as $s_k = (1, 2, 2, 2, 3, 3, 3)^\top$, which indicates there are three segments $\{1\}, \{2, 3, 4\}$ and $\{5, 6, 7\}$.

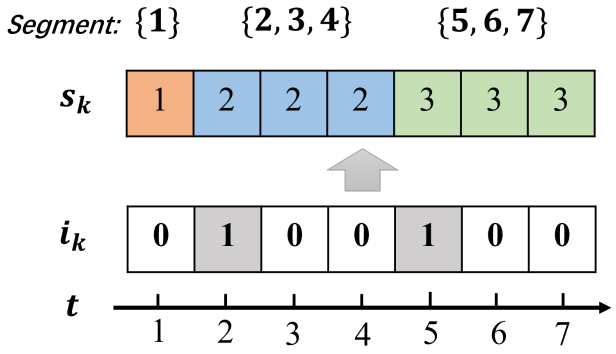


Figure 1: A topic example with two change points occurring at $t = 2$ and $t = 5$. The resulting three segments are $\{1\}, \{2, 3, 4\}$ and $\{5, 6, 7\}$.

With the locations of change points for each single topic, we then describe the generation process of topic meanings. Let $\phi_{k,t} = (\phi_{k,t,1}, ..., \phi_{k,t,V})^\top \in \mathbb{R}^V$ denote the probability distribution for topic k at the t-th moment over the whole dictionary with V words. If there exists a change point at the tth moment for topic k, then $\phi_{k,t+1}$ should be different from $\phi_{k,t}$. In other words, $\phi_{k,t}$ should remain the same within moments in the same segment but behave differently after a change point occurs. Specifically, for the sth segment with $1 \leq s \leq S_k$, define the topic-word distribution in this segment to be $\phi_{k,s} = (\phi_{k,s,1}, ..., \phi_{k,s,V})^\top \in \mathbb{R}^V$. To model the changing pattern of $\phi_{k,s}$ within different segments, we adopt the Pitman-Yor process (PYP, Pitman and Yor, 1995). Generally, there are three hyperparameters in a Pitman-Yor process: (1) the discount parameter a, (2) the concentration parameter b, and (3) a basis discrete prior distribution $H(\cdot)$. Therefore, a Pitman-Yor process is often referred to as PYP$(a, b, H(\cdot))$. The detailed discussion of PYP is given in Section 3.2.

PYP is a stochastic Bayesian process with two key features: one is the *nonparametric nature* and the other is the *power-law behavior*. These two features make PYP particularly popular for modeling the topic-word distributions in topic models (Sato and Nakagawa,

2010; Lindsey et al., 2012; Lim et al., 2016; Guo et al., 2024). Its nonparametric nature allows it to dynamically adapt to the increasing number of unique words as time goes by. Its power-law property enables it to model the heavy-tailed nature of word frequencies, i.e., low-frequency words often play a crucial role in distinguishing between topics. Additionally, by using PYP to characterize the changing patterns of topic-word distributions, we can retain a certain proportion of similarities within the same topic but also allow for significant change in different segments. This balance is mainly achieved by the power-law property of PYP, which inherently leads the model toward reusing dominant topics but still permits new words to emerge. Therefore, PYP is well suited to our task of detecting topic change points.

Last, we discuss the generation process of documents. Based on the topic-word distributions $\phi_{k,s}$ for each topic on each segment, the generation of documents is similar to that of the LDA model (Blei et al., 2003). Specifically, assume each document is a mixture of K topics, which is represented by the probability vector $\theta_{t,d} = (\theta_{t,d,1}, \ldots, \theta_{t,d,K})^\top \in \mathbb{R}^K$. Each word in the document can represent one specific topic $z_{t,d,n} \in \{1, ..., K\}$, which is generated from a multinomial distribution with parameter $\theta_{t,d}$. Under the represented topic $z_{t,d,n}$, the observed word $w_{t,d,n}$ is drawn from the corresponding topic-word distribution ϕ_{k^*,s^*}, where $k^* = z_{t,d,n}$ and s^* is the corresponding segment index $s_{k^*,t}$.

Overall, the generative process of TOPIC-PYP is presented below, which is illustrated in Figure 2. The generative process contains three stages. Stage 1 describes the process of determining the number and locations of the change points, Stage 2 employs the Pitman-Yor process to model the changing patterns of topic meanings given the identified change points, and Stage 3 is the final process of generating the documents.

1. Stage 1: *Generation of Change Points.*

   For topic k with $1 \leq k \leq K$:

   (a) Generate the topic shift probability π_k: $\pi_k \sim \text{Beta}(\lambda_0, \lambda_1)$.

   (b) For the t-th moment with $1 \leq t \leq T$:

      i. Generate the topic shift indicator: when $t = 1$, set $i_{k,t} = 0$; when $t > 1$, generate $i_{k,t} \sim \text{Bernoulli}(\pi_k)$.

      ii. Compute the index of segment: $s_{k,t} = \sum_{j=1}^{t} i_{k,j} + 1$.

   (c) Compute the total number of segments: $S_k = \sum_{t=1}^{T} i_{k,t} + 1$.

2. Stage 2: *Generation of Topics.*

   For topic k with $1 \leq k \leq K$:

   (a) Generate the basis prior topic-word distribution from a homogeneous Dirichlet distribution: $h_k \sim \text{Dir}(\gamma)$.

   (b) For each specific segment $s \in \{1, \ldots, S_k\}$:

      i. Generate the topic-word distribution using a Pitman-Yor process: $\phi_{k,s} \sim \text{PYP}(a, b, h_k)$

3. Stage 3: *Generation of Documents.*

   For document d with $1 \leq d \leq D_t$ and $1 \leq t \leq T$:

(a) Generate its document-topic distribution over K topics: $\theta_{t,d} \sim \text{Dir}(\alpha)$

(b) For word $n \in \{1, \ldots, N_d\}$:

i. Generate the word topic indicator: $z_{t,d,n} \sim \text{Multinomial}(\theta_{t,d})$, and denote $z_{t,d,n}$ by k^* for easy illustration.

ii. Find the index segment for topic k^*: $s^* = s_{k^*,t}$.

iii. Generate the specific word: $w_{t,d,n} \sim \text{Multinomial}(\phi_{k^*,s^*})$.

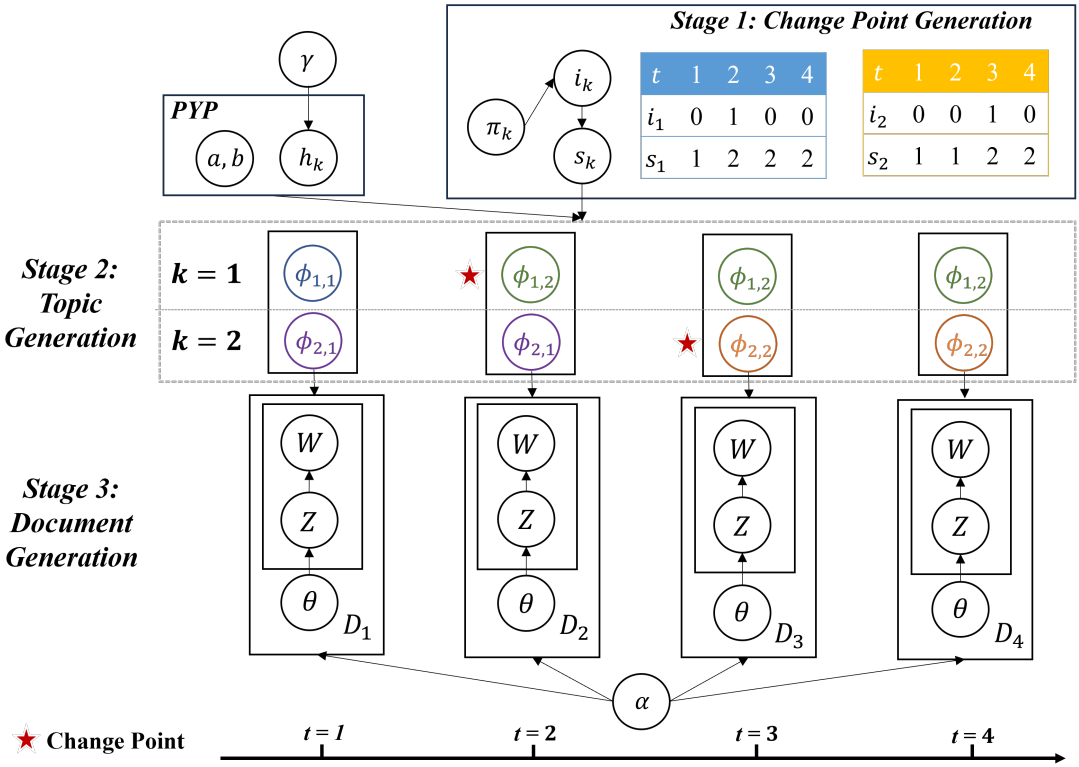


Figure 2: The generative process of TOPIC-PYP, illustrated by an example with two topics and one change point in each topic within four time periods. Specifically, the first topic has a change point at $t = 2$, while the second topic has a change point at $t = 3$. Overall, the generative process of TOPIC-PYP involves three stages: the generation of change points (Stage 1), the generation of topic meanings assisted by PYP (Stage 2), and the generation of documents (Stage 3).

We make two remarks about the above generative process. First, we adopt a homogeneous Dirichlet distribution for the basis prior topic-word distribution h_k. The homogeneous assumption has been widely adopted in the past literature; see Blei et al. (2003); Blei and Lafferty (2006a); Blei and Mcauliffe (2008); Rosen-Zvi et al. (2012); Lu et al. (2022) for examples. It actually assumes that all words have the same occurring probability under a given topic, without any inherent preference for specific words. Wallach et al. (2009) further discussed this problem for LDA and found an asymmetric prior over the topic–word

distributions provides no real benefit. Therefore, in this work, we just follow the common practice to adopt the homogeneous assumption on topic-word distributions.

Second, although the generative process contains Stage 1, Stage 2, and Stage 3 sequentially, we cannot sequentially estimate these stages. This is because, practically, the estimation of the number and positions of change points in Stage 1 should rely on the topic-word distributions (i.e., the $\phi_{k,s}$). However, the $\phi_{k,s}$s are further modeled using the Pitman-Yor process in Stage 2. In other words, $\phi_{k,s}$ serves as the input in estimating Stage 1, but the output from the estimation of Stage 2. Therefore, if we were to estimate the Stage 1 and Stage 2 sequentially and independently, the topic-word distributions would become inconsistent between the two stages. Due to this inherent dependency problem, existing two-stage estimation methods in fact invert the estimation order. That is, they first perform Stages 2&3 to estimate the topic-word distributions using some pre-defined models, and then proceed Stage 1 to detect the change points using the obtained $\phi_{k,s}$. Please see Bruggermann et al. (2016), Zhang et al. (2017), and Wang and Goutte (2018) for more discussions about the existing two-stage methods. In this work, we do not adopt a two-stage approach. Instead, we integrate these stages into a unified framework, combining topic modeling and change point detection. This integrated approach should align more closely with the natural generation process of documents.

Given the generative process of TOPIC-PYP, we can derive the posterior distribution for all variables, hyperparameters, and the data. Then by integrating out some latent variables, the proposed TOPIC-PYP can be estimated using the collapsed Gibbs sampling algorithm (Liu, 1994; Griffiths and Steyvers, 2004). We describe the details of model estimation in Section 3.3.

3.2 The Pitman-Yor process

In this work, we adopt the Pitman-Yor process (PYP) to model the generation of topics. Assume there is a Pitman-Yor process denoted by $\text{PYP}(a, b, H(\cdot))$. Here $a \in [0, 1)$ is the discount parameter, $b > -a$ is the concentration parameter, and $H(\cdot)$ is the basis word distribution of topics. The discount parameter a controls the degree of sparsity and the heavy-tail behavior in the topic-word distributions. A larger a would encourage a heavy-tailed distribution and allow more rare words to appear with high probability. The concentration parameter b influences the diversity of the topic-word distributions and the probability at which new words are introduced. A larger b would increase the diversity by encouraging the generation of new words with high probability. The basis distribution $H(\cdot)$ serves as the baseline word distribution for a given topic. For example, a uniform basis distribution $H(\cdot)$ treats all words with equal probability initially, promoting unbiased exploration of the vocabulary.

For a better understanding of the Pitman-Yor process, we use the stick-breaking process to illustrate its mechanism. Specifically, PYP can be described using the following steps.

(1) Assume there exists a stick of length 1. We sample a random value V_1 from $\text{Beta}(1 - a, b)$ and split the stick into two parts. One stick is V_1 long and the other stick is $1 - V_1$ long. Then define the first probability value $\tilde{p}_1 = V_1$.

(2) Sample another random value V_2 from Beta$(1-a, b+a)$. Multiply the remaining $1-V_1$ with V_2 to get the length $(1-V_1)V_2$ of the second stick. Then we define the second probability value $\tilde{p}_2 = (1-V_1)V_2$.

(3) Repeat the previous steps. For the ith step, sample a random value V_i from Beta$(1-a, b+i\times a)$. Then compute $\tilde{p}_i = (1-V_i)\prod_{j=1}^{i-1}(1-V_j)$.

By using the stick-breaking process, we have a series of probability values $\tilde{p}_1, \tilde{p}_2, \ldots$, which satisfies $\sum_i^{\infty} \tilde{p}_i = 1$. Next, assume a series of samples $v_1, v_2, \ldots$ are drawn sequentially from the prior distribution $H(\cdot)$. Define a discrete distribution $\phi = (\tilde{p}_1, \tilde{p}_2, \ldots)^\top$ satisfying $P(w = v_i) = \tilde{p}_i$. Then we say ϕ follows PYP$(a, b, H(\cdot))$ and w follows a multinomial distribution with parameter ϕ.

In the context of TOPIC-PYP, assume $\phi_{k,s} \sim \text{PYP}(a, b, h_k)$ for the given topic k and segment s, where h_k is a basis prior distribution. Suppose a sequence of words $w_1, w_2, \ldots, w_N$ is generated from the multinomial distribution with parameter $\phi_{k,s}$. Then we try to derive the conditional posterior probability of a new word w_{N+1} given the N observable words. As shown in Buntine and Hutter (2010) and Chen et al. (2011), by integrating out $\phi_{k,s}$, the conditional posterior probability of w_{N+1} can be derived as follows:

$$p(w_{N+1} \mid w_1, \ldots, w_N, a, b, h_k) = \frac{b+Ma}{b+N} h_k + \sum_{m\in\mathcal{M}} \frac{n_{v_m} - a}{b+N} I(w_{N+1} = v_m). \tag{1}$$

To better understand Equation (1), we explain the Pitman-Yor process from the perspective of Chinese restaurant process (CRP), which is discussed in Appendix A.1. Simply speaking, we can regard the word generation process as a customer choosing a dish in a restaurant. Assume a document represents a restaurant. Each word w_i refers to a new customer entering the restaurant, its representing topic z_i refers to the table, and the dish enjoyed by the customer refers to the selected value v_i from the vocabulary. Each time a new customer (word) enters the restaurant (document), it should make two choices. First, it should choose a table (topic). In this step, it either chooses an existing table or a new table. After sitting at one table, it should choose a dish (a specific value in vocabulary) on the table. If the customer joins an existing table, it automatically selects the dish one that table. Note that one table only has one dish. If the customer opens a new table, it selects a dish from the menu h_k.

To explain Equation (1), note that we have already observed N words $(w_1, \ldots, w_N)$. Let $\mathcal{M}$ contain the indices of N words directly selected from the prior distribution h_k (i.e., opening new tables), and M be the corresponding count. Then we should have $M \leq N$. For $m \in \mathcal{M}$, let v_m be the specific value of w_m (i.e., the dish on the table). Further denote n_{v_m} to be the number of times v_m being selected among the N words. Then Equation (1) implies that, the value of a new word w_{N+1} is drawn from the prior distribution h_k with probability $(b+Ma)/(b+N)$, or equals to one previously appeared word v_m with probability $(n_{v_m} - a)/(b+N)$. We remark that, based on the definition of PYP, we have $0 \leq a < 1$. Recall $M \leq N$. Therefore we should have $Ma < N$. Accordingly, the inequality $(b+Ma) \leq (b+N)$ holds, which guarantees the probability $(b+Ma)/(b+N)$ lying between 0 and 1.

The Pitman-Yor process is a suitable choice to characterize the changing pattern of topic-word distributions due to its nonparametric and power-law features. Except for PYP, there

exist other stochastic Bayesian processes suitable for modeling the topic-word distributions. One example is the Polya tree process (Ferguson, 1974; Lavine, 1992). By constructing hierarchical structures, this process captures complex dependencies in data and provides a flexible way to model intricate word distributions. Another example is the hierarchical Dirichlet process (HDP, Teh et al., 2006), which is an extension of the Dirichlet process. HDP allows for shared topics across grouped data, which can effectively model both global and local topic structures. Known for modeling overlapping features, the Indian buffet process (IBP, Ghahramani and Griffiths, 2005) is particularly useful in discovering shared and unique topics across documents and thus serves as another suitable choice.

3.3 Model estimation

To estimate TOPIC-PYP, we develop a collapsed Gibbs sampling method (Liu, 1994; Griffiths and Steyvers, 2004). We first define some necessary notations. Let $\Pi = \{\pi_1, \ldots, \pi_K\}$ be the collection of topic shift probabilities and $\mathcal{I} = \{i_{k,t}, 1 \le k \le K, 1 \le t \le T\}$ be the collection of topic shift indicators. Let $\Phi = \{\phi_{k,s}, 1 \le k \le K, 1 \le s \le S_k\}$ be the collection of topic-word probabilities, and $\mathcal{H} = \{h_1, \ldots, h_K\}$ be the collection of prior topic-word probabilities. Further define $\Theta = \{\theta_{t,d}, 1 \le t \le T, 1 \le d \le D_t\}$ to be the collection of document-topic probabilities, $\mathcal{Z} = \{z_{t,d,n}, 1 \le t \le T, 1 \le d \le D_t, 1 \le n \le N_{t,d}\}$ to be the collection of topic indicators, and $\mathcal{W} = \{w_{t,d,n}, 1 \le t \le T, 1 \le d \le D_t, 1 \le n \le N_{t,d}\}$ to be the collection of all words. To facilitate the estimation procedure, we introduce a new dummy variable $\delta_{t,d,n}$. Specifically, define $\delta_{t,d,n} = 1$ when the word $w_{t,d,n}$ is the first customer in a new table representing topic $z_{t,d,n}$; and $\delta_{t,d,n} = 0$ otherwise. Then let $\Delta = \{\delta_{t,d,n}, 1 \le t \le T, 1 \le d \le D_t, 1 \le n \le N_{t,d}\}$. In the TOPIC-PYP model, the hyperparameters include λ_0, λ_1 for the Beta prior of π_k, a, b for the Pitman-Yor process, and γ, α for the Dirichlet priors of the topic-word and document-topic distributions. Define $\Xi = \{a, b, \gamma, \alpha, \lambda_0, \lambda_1\}$ to be the collection of hyperparameters.

Based on the above notations, the joint posterior distribution of all variables $\{\Pi, \mathcal{H}, \mathcal{I}, \Phi, \Theta, \mathcal{Z}, \Delta\}$ given the hyperparameters Ξ and observed words $\mathcal{W}$ can be derived as follows,

$$\begin{aligned} &f(\Pi, \mathcal{I}, \mathcal{H}, \Phi, \Theta, \mathcal{Z}, \Delta | \mathcal{W}, \Xi) \\ \propto &f(\Pi \mid \lambda_0, \lambda_1) f(\mathcal{I} \mid \Pi) f(\mathcal{H} \mid \gamma) f(\Phi \mid a, b, \mathcal{H}) f(\Theta \mid \alpha) f(\mathcal{Z} \mid \Theta) f(\mathcal{W}, \Delta \mid \Phi, \mathcal{Z}, \mathcal{I}). \end{aligned}$$

Based on Theorem 17 in Buntine and Hutter (2010), we can integrate out Π, Θ, Φ and $\mathcal{H}$ by the conjugate structure. This yields the joint posterior distribution over $\{\mathcal{I}, \Delta, \mathcal{Z}\}$ given the hyperparameters and data, which is denoted by $f(\mathcal{I}, \Delta, \mathcal{Z} | \mathcal{W}, \Xi)$. The derivation details are present in Appendix A.2. As a result, we just need to estimate $\{\mathcal{I}, \Delta, \mathcal{Z}\}$. To this end, we apply the collapsed Gibbs sampling method, which employs an iterative two-step algorithm: (1) estimate $\{\Delta, \mathcal{Z}\}$ given $\mathcal{I}$, and (2) estimate $\mathcal{I}$ given $\{\Delta, \mathcal{Z}\}$. See Appendix A.3 and Appendix A.4 for the detailed derivations of the posterior distributions $f(\Delta, \mathcal{Z} | \mathcal{I}, \mathcal{W}, \Xi)$ and $f(\mathcal{I} | \Delta, \mathcal{Z}, \mathcal{W}, \Xi)$, respectively.

To begin the algorithm, we need first set the initial value of $\mathcal{I}$. In this work, we generate the initial value of $\mathcal{I}$ following the generative process of TOPIC-PYP in Figure 2. Specifically, under the given hyperparameters λ_0 and λ_1, the topic shift probability π_k with $1 \le k \le K$ can be generated from Beta(λ_0, λ_1). Then a topic shift indicator $i_{k,t}$ for topic k at time t can be generated from Bernoulli(π_k). All the generated $i_{k,t}$s construct the initial

value $\mathcal{I}$. Subsequently, in each iteration, the first step is to estimate $\{\Delta, \mathcal{Z}\}$ given $\mathcal{I}$, and the second step is to estimate $\mathcal{I}$ given $\{\Delta, \mathcal{Z}\}$. By updating $\{\Delta, \mathcal{Z}\}$ and $\mathcal{I}$ iteratively using the Gibbs sampling method, we could obtain a series of posterior samples of $\{\Delta, \mathcal{Z}, \mathcal{I}\}$, based on which the model estimation can be conducted.

Assume we obtain a total of R posterior samples for $\{\Delta, \mathcal{Z}, \mathcal{I}\}$ after convergence, which are denoted by $\{\Delta^{(r)}, \mathcal{Z}^{(r)}, \mathcal{I}^{(r)}\}$ with $1 \leq r \leq R$. In this work, we mainly focus on the estimation of two types of parameters. The first one is $\mathcal{I}$, which represents the locations of change points for each topic. Using the R posterior samples, we can compute the maximum a posteriori (MAP) estimator for $\mathcal{I}$ as $\widehat{\mathcal{I}} = \{\widehat{i}_{k,t}, 1 \leq k \leq K, 1 \leq t \leq T\}$, where $\widehat{i}_{k,t} = \text{mode}\{i_{k,t}^{(1)}, ..., i_{k,t}^{(R)}\}$. Accordingly, the number of change points for each topic can be estimated as $\widehat{Q}_k = \sum_{t=1}^{T} \widehat{i}_{k,t}$. The second parameter of research interest is Φ, which represents the topic-word distributions. Based on Φ, we can summarize the meaning of each topic at each segment. Recall $\phi_{k,s} = (\phi_{k,s,1}, ..., \phi_{k,s,V})^\top \in \mathbb{R}^V$ denotes the word distribution for topic k at the s-th segment over the whole dictionary with V words. The estimation of $\phi_{k,s}$ should rely on the posterior samples of $\{\Delta, \mathcal{Z}\}$. First, similar to the estimation of $\mathcal{I}$, we compute the MAP estimators of $\{\Delta, \mathcal{Z}\}$ using the mode of the corresponding R posterior samples, which are denoted by $\{\widehat{\Delta}, \widehat{\mathcal{Z}}\}$. Then compute $n_{k,s,v}$ and $\tau_{k,s,v}$ using $\{\widehat{\Delta}, \widehat{\mathcal{Z}}\}$; see Step 3 in Appendix A.2 for their detailed definitions. Last, compute the estimator $\widehat{\phi}_{k,s} = (\widehat{\phi}_{k,s,1}, ..., \widehat{\phi}_{k,s,V})^\top$, where $\widehat{\phi}_{k,s,v} = (n_{k,s,v} - a\tau_{k,s,v} + bh_{k,v})/(\sum_v n_{k,s,v} - a\sum_v \tau_{k,s,v} + b)$. Here $h_k = (h_{k,1}, ..., h_{k,V})^\top$ is the basis prior distribution.

Last, we provide some theoretical guarantees for the numerical convergence of the estimation algorithm. That is, whether the posterior samples obtained by the Gibbs sampling method can converge to the true posterior distribution. Note that, the iterative method to estimate $\{\Delta, \mathcal{Z}, \mathcal{I}\}$ is indeed Gibbs sampling. It is a well-established Markov chain Monte Carlo (MCMC) technique and its convergence property has been rigorously proven under certain conditions; see Roberts and Rosenthal (2004); Gelman et al. (2013) for details. The key theoretical result of Gibbs sampling is that, as long as the conditional distributions are regular (i.e., they have sufficient support) and the Markov chain is ergodic, Gibbs sampling will converge to the true posterior distribution. In TOPIC-PYP, the posterior distribution $f(\Delta, \mathcal{Z}|\mathcal{I}, \mathcal{W}, \Xi)$ is multinomial, while the posterior distribution $f(\mathcal{I}|\Delta, \mathcal{Z}, \mathcal{W}, \Xi)$ is Bernoulli; see Appendices A.3 and A.4 for details. Multinomial distribution is well known for its broad support, encompassing all possible outcomes where the counts of categories sum to a fixed total. The probability mass function ensures that the posterior distribution $f(\Delta, \mathcal{Z}|\mathcal{I}, \mathcal{W}, \Xi)$ covers the entire support of $\{\Delta, \mathcal{Z}\}$, provided $\mathcal{I}$ and others are fixed. This ensures that the regularity condition, specifically the sufficient support, is satisfied. As for the Bernoulli distribution, it has a support of $\{0, 1\}$, and its probability mass function assigns positive probabilities to both possible outcomes as long as the success probability lies between zero and one. This property naturally holds for the posterior distribution $f(\mathcal{I}|\Delta, \mathcal{Z}, \mathcal{W}, \Xi)$, ensuring the regularity condition is met. By alternating between these two distributions during Gibbs sampling, the Markov chain transitions between states with sufficient support and retains irreducibility and aperiodicity. These conditions collectively ensure the ergodicity of the Markov chain and its convergence to the true posterior distribution. Therefore, the numerical convergence property of TOPIC-PYP should be supported. To empirically demonstrate this property, we validate the posterior convergence of TOPIC-

PYP in simulation experiments on syngenetic datasets. Please see Appendix B.1 for the detailed results.

3.4 Posterior consistency

In this section, we study the posterior consistency of parameters in TOPIC-PYP. As we mentioned before, in this work, we mainly focus on two types of parameters. The first one is $\mathcal{I}$, which represents the locations of change points for each topic; while the second one is Φ, which represents the meanings of topics. Therefore, we study the posterior consistency for the two parameters, respectively.

We first focus on $\mathcal{I}$. Note that it is difficult to directly apply traditional theoretical results (such as Doob's Theorem (Doob, 1949)) on the posterior consistency of $\mathcal{I}$. This is mainly because of two reasons. First, the words $w_{t,d,n}$s across different time points may not be identically distributed because the topic-word distributions represented by $\phi_{k,s}$ can change over time. Second, since the estimate of $\mathcal{I}$ is obtained through Gibbs sampling, it is challenging to derive a function f such that $\mathcal{I} = f(\mathcal{W})$, which is measurable with respect to the sigma-field generated by the sequence $\{w_{t,d,n} : d \to \infty, n \to \infty\}$. These issues complicate the direct application of traditional results on posterior consistency of $\mathcal{I}$. To address this issue, we study the posterior consistency of $\mathcal{I}$ based on McGoff et al. (2022), which establishes a general framework for understanding the asymptotic behavior and Bayesian posterior consistency of Gibbs posterior distributions in dependent processes. To simplify the analysis, we assume all time points have the same number of documents, i.e., $D_t = D$ for $1 \le t \le T$; and all documents have the same number of words, i.e., $N_{t,d} = N$ for $1 \le t \le T, 1 \le d \le D$. Then the posterior consistency of $\mathcal{I}$ is summarized in Theorem 1.

Theorem 1 *Let $\mathcal{W} = \{w_{t,d,n}, 1 \le t \le T, 1 \le d \le D, 1 \le n \le N\}$ be a collection of words generated independently from the TOPIC-PYP model $P_{\mathcal{I}^*,\Phi^*}$, which is parameterized by the true change points locations $\mathcal{I}^* = (i^*_{k,t}) \in \{0,1\}^{KT}$ and the true topic-word distributions $\Phi^* = \{\phi^*_{k,s}, 1 \le k \le K, 1 \le s \le S_k\}$, with the other parameters omitted. Denote by $\mathbf{\Pi}$ the prior distribution of $\mathcal{I}$, with its density specified in (A.1) in Appendix A.2. For a given $\varepsilon > 0$, define the set $B(\mathcal{I}^*, \varepsilon) = \{\mathcal{I} \in \{0,1\}^{KT} : \sum_k \sum_t \mathbb{I}(i_{k,t} \ne i^*_{k,t}) < \varepsilon\}$, where $\mathbb{I}(\cdot)$ denotes the indicator function. Then with a fixed time span T, for any neighborhood of $\mathcal{I}^*$, we have the asymptotic result:*

$$\lim_{D,N\to\infty} \mathbf{\Pi}\{\mathcal{I} \in B(\mathcal{I}^*, \varepsilon)|\mathcal{W}\} = 1 \quad a.s. \quad [P_{\mathcal{I}^*,\Phi^*}].$$

The proof of Theorem 1 can be found in Appendix A.5. This theorem illustrates that as the number of documents and the number of words increase, the posterior distribution of the change points will almost surely concentrate around the true change points. This implies that as D and N increase, the estimate of the change points $\mathcal{I}$ will converge to the true change points $\mathcal{I}^*$ with high probability.

Next, we focus on the posterior consistency of Φ. Let $\Omega^{k-1} = \{\mathbf{x} \in \mathbb{R}^k : 0 \le x_i \le 1, \sum_{i=1}^k x_i = 1\}$ denote the $(k-1)$-dimensional probability simplex. Then each topic $\phi_{k,s}$ $(1 \le k \le K, 1 \le s \le S_k)$ is a vector in the probability simplex Ω^{V-1}. Our primary interest is to study the posterior consistency of each individual topic parameter $\phi_{k,s}$. However,

in practice, topics are often correlated and tend to share keywords. This is particularly true in our case, since the topic-word probability vectors associated with the same topic across different segments often have some similarity, making them not fully identifiable. To tackle this problem, we follow the practice of Tang et al. (2014) and Nguyen (2015) to study the convergence of *topic polytope*, which is defined as the convex hull of the topic parameters, i.e., $G = \text{conv}(\Phi) = \text{conv}(\phi_{1,1}, \cdots, \phi_{1,S_1}, \cdots, \phi_{K,1}, \cdots, \phi_{K,S_K})$. Then G has at most $\sum_{k=1}^{K} S_k$ vertices (i.e., extreme points) among $\{\phi_{1,1}, \cdots, \phi_{1,S_1}, \cdots, \phi_{K,1}, \cdots, \phi_{K,S_K}\}$. Assume the observed words $\mathcal{W}$ are generated according to the TOPIC-PYP model $P_{\mathcal{I}^*,\Phi^*}$ with the other parameters omitted. Let $G^* = \text{conv}(\Phi^*)$ be the true topic polytope. Under the TOPIC-PYP model, Φ is endowed with a prior distribution $\mathbf{\Psi}$ based on the Pitman-Yor process. Then Theorem 2 characterizes the contraction behavior of the posterior distribution $\mathbf{\Psi}(G|\mathcal{W})$, as the numbers of documents D and words N go to infinity.

Theorem 2 *Assume G^* is in the support of prior $\mathbf{\Psi}$. Let $p = \min(\sum_{k=1}^{K} S_k - 1, V)$. Under the assumptions (A1)-(A3) in Appendix A.6 and with a fixed time length T, as $D \to \infty$, $N \to \infty$ and D, N satisfying $\log(TD) \leq N$, for some sufficiently large constant C independent of D and N, we have*

$$\mathbf{\Psi}(d_{\mathcal{M}}(G^*, G) \geq C\delta_{T,D,N}|\mathcal{W}) \to 0 \quad a.s. \quad [P_{\mathcal{I}^*,\Phi^*}],$$

where $\delta_{T,D,N} = \{(\log TD)/TD + (\log N)/N + (\log N)/TD\}^{1/(2p)}$ is the posterior concentration rate, and $d_{\mathcal{M}}$ is some distance measure.

The proof of Theorem 2 is provided in Appendix A.6. In this theorem, we employ the Euclidean distance metric, specifically the "minimum-matching" distance $d_{\mathcal{M}}(G^*, G)$, to measure the dissimilarity between the topic polytope G and the true topic polytope G^*. A precise definition of this metric is provided in Appendix A.6. Since the minimum-matching distance is determined solely by the extreme points of the polytopes, the convergence of the convex polytope G in Theorem 2 ensures the convergence of all extreme points in the polytope. Consequently, Theorem 2 establishes an explicit upper bound for the asymptotic rate at which the posterior distribution of the topic-word parameters Φ concentrates around their true values Φ^*. This derived rate is contingent upon a condition that governs the required "thickness" of the prior support for the marginal densities of the data $P_{\mathcal{W}|G}$, in addition to an upper bound on the entropy of the space of such densities. The upper bound $\delta_{T,D,N}$ is shown to deteriorate with respect to p, exhibiting a behavior akin to that of a nonparametric rate. This phenomenon arises because the true number of topics does not need to be pre-specified, nor is there a requirement for the topics to be well-separated. In practice, the total number of segments split by change points (i.e., $\sum_{k=1}^{K} S_k$) is typically smaller than the vocabulary size V. In such cases, the convergence rate deteriorates rapidly as the total number of segments $\sum_{k=1}^{K} S_k$ used in TOPIC-PYP increases. Therefore, to guarantee the statistical efficiency of TOPIC-PYP, we do not expect the change points occur too frequently, and meanwhile avoid selecting an excessively large number of topics for the model.

4. Experiments on Synthetic Data

4.1 Experimental setup and evaluation

In this section, we first evaluate the change point detection performance of TOPIC-PYP on synthetic data and then explore the stability of TOPIC-PYP on different hyperparameters. We generally follow the generative process of TOPIC-PYP (as illustrated in Figure 2) to generate the synthetic documents. Specifically, assume the total number of moments to be T. At each time moment, there are $D_t = 100$ documents with $1 \leq t \leq T$. The number of words in each document is fixed as $N_{t,d} = 100$ with $1 \leq d \leq D_t$ and $1 \leq t \leq T$. Assume the whole vocabulary size is $V = 1000$ and the number of topics underlying the whole corpus is $K = 5$. For easy comparison, we assume all topics share the same locations of the change points. Specifically, we consider the following four scenarios.

- Scenario 1. We assume $T = 10$ and there exists one single change point (i.e., $Q_k = 1$ with $1 \leq k \leq K$). All change points are assumed to occur at the fifth moment, i.e., $i_k = 5$
- Scenario 2. Assume $T = 10$ and each topic has two change points (i.e., $Q_k = 2$), which occur at the third and sixth moments, i.e., $i_k = (3, 6)^\top$.
- Scenario 3. Assume $T = 30$ and there are $Q_k = 5$ change points occurring at the moments $i_k = (5, 10, 15, 20, 25)^\top$
- Scenario 4. Assume $T = 100$ and the number of change points is $Q_k = 10$. The moments of change points are $i_k = (9, 18, 27, 36, 45, 54, 63, 72, 81, 90)^\top$.

Among the four scenarios, Scenario 1 and Scenario 2 are simple cases with a small number of change points, while Scenario 3 and Scenario 4 serve as more complicated cases with a larger number of change points. In all settings, the hyperparameters in PYP are set as $a = 0.5$ and $b = 5$. Other hyperparameters used in TOPIC-PYP are set as $\gamma = 0.1$, $\alpha = 0.2$, and $\lambda = \{\lambda_0, \lambda_1\} = \{2, 5\}$. We repeat the above data generation process for $B = 50$ times.

After generating the synthetic documents, we apply the proposed TOPIC-PYP model to perform topic change point detection. For comparison purposes, we compare it with five methods. The first one is a unified state-of-the-art method Topic-CD (Lu et al., 2022). It combines topic modeling and change point detection in a unified Bayesian framework. By defining change points based on shifts in hyperparameters affecting topic-word distributions, Topic-CD can detect significant changes occurring for all topics. The other four methods are all two-stage methods, among which the first stage is to obtain the dynamic sequences of topic-word distributions, and the second stage is to detect change points for each topic using the resulting topic-word distributions. As for the first stage, we consider four methods. They are, respectively, (1) the dynamic topic model (DTM, Blei and Lafferty, 2006b), (2) the neural dynamic topic model (D-ETM, Dieng et al., 2019), (3) the Rolling LDA method (Rieger et al., 2022), and (4) the aligned neural topic model (ANTM, Rahimi et al., 2024). After obtaining the dynamic sequences of topic-word distributions in the first stage, we try to detect the change points for each topic. To this end, we first use the cosine similarity (Bruggermann et al., 2016) and the Jensen-Shannon divergence (Lau et al., 2012;

Wang and Goutte, 2018) to measure the distance between any two topic-word probability distributions in adjacent periods. Then three methods are applied to detect the change points for each topic, including: (1) the threshold method (Bruggermann et al., 2016), (2) the dynamic programming method (Truong et al., 2020), and (3) the binary segmentation method (Truong et al., 2020). In total, there are $2 \times 3 = 6$ methods to detect the change points in the second stage. To implement different methods, we assume the true number of topics K is already known. In addition, since the number of change points should be predefined for two-stage methods, we assume the true number of change points is already known for the two-stage methods.

We evaluate the performance of TOPIC-PYP from two perspectives. The first one is the topic modeling performance, while the second one is the change point detection performance. To evaluate the topic modeling performance, we use the coherence score (CS) (Newman et al., 2010), which is popularly used to measure the quality of generated topics. The basic idea of the coherence score is that words belonging to the same topic should have a higher probability to co-occur within the same document. Specifically, let $\{w_1, w_2, ..., w_L\}$ be the first L words with the highest probabilities for topic k. Define $F(w)$ to be the number of documents including word w, and $F(w, w')$ to be the number of documents including both words w and w'. Then the coherence score of topic k is defined as:

$$\mathrm{CS}_k = \sum_{l=2}^{L} \sum_{l'=1}^{l-1} \log \frac{F(w_l, w_{l'}) + 1}{F(w_{l'})}.$$

The higher the coherence score, the better the quality of the topic. After calculating the topic coherence for each topic, we average among all topics to obtain the final topic coherence score.

The second perspective is the change point detection performance. To this end, we generally follow Lu et al. (2022) and use the following evaluation metrics. Let $\widehat{Q}_k^{(b)}$ be the estimated number of change points for topic k with $1 \le k \le K$ in the bth experiment. Denote $\widehat{\Omega}_k^{(b)} = \{\widehat{t}_{1k}^{(b)} \dots \widehat{t}_{\widehat{Q}_k^{(b)}}^{(b)}\}$ to be the estimated locations of the estimated change points for topic k. Similarly, define Q_k and Ω_k to be the true number and locations of change points for topic k. Define $M_k^{(b)} = \sum_{q=1}^{Q_k} I\{\widehat{t}_{qk}^{(b)} \in (t_{q-h,k}, t_{q+h,k})\}$ as the number of correctly detected locations for topic k, where $I(\cdot)$ is the indicator function and h is the bandwidth. Then, the precision and recall of detected change points are defined as follows,

$$\text{Precision} = \frac{1}{BK} \sum_{b=1}^{B} \sum_{k=1}^{K} \left(|M_k^{(b)}| / |\widehat{\Omega}_k^{(b)}| \right), \text{Recall} = \frac{1}{BK} \sum_{b=1}^{B} \sum_{k=1}^{K} \left(|M_k^{(b)}| / \Omega_k | \right).$$

Here $|\cdot|$ is a counting function. Except for precision and recall, we also adopt the commonly used P score and WindowDiff to measure the change point detection performance. Both two measures use a moving window with bandwidth $h = T/2$ to check whether the partitions split by the change points are correct. The lower the two measures, the better the performance of change point detection. See Pevzner and Hearst (2002) and Lu et al. (2022) for the detailed definition of P score and WindowDiff.

4.2 Main comparison results

We employ the Gibbs sampling method to estimate TOPIC-PYP. In all scenarios, each experiment runs for 50 iterations, which is sufficient to achieve convergence. A detailed convergence check is provided in Appendix B.1. After model estimation, we compare the performance of TOPIC-PYP with the other methods. We first focus on the performance of topic modeling, which is evaluated by the coherence score. To compute this measure, we set $L = 10$ for illustration purpose. Table B.5 presents the coherence score results of different methods. As shown, the TOPIC-PYP model consistently achieves the highest coherence score across all scenarios, indicating that it generates topics of the highest quality. In addition, we find both unified methods (i.e., TOPIC-PYP and Topic-CD) have achieved significantly better results than the two-stage methods. This finding suggests that, when change points indeed exist in dynamic documents, incorporating them into the topic modeling process (i.e., the unified methods) would result in higher-quality topics than ignoring their presence (i.e., the two-stage methods).

Table 1: The comparison results of coherence score for different methods in four scenarios.

Method		Scenario 1	Scenario 2	Scenario 3	Scenario 4
Unified	TOPIC-PYP	29.065	22.640	15.418	6.754
	Topic-CD	22.192	18.349	15.179	6.150
Two-Stage	DTM	0.721	0.658	0.658	0.289
	D-ETM	1.285	0.859	0.674	0.538
	Rolling LDA	3.826	3.342	2.916	2.857
	ANTM	0.479	0.338	0.781	0.064

Next, we compare the change point detection performance of different methods. To denote the two-stage methods, we use "CS" and "JS" to represent the distance measures cosine similarity and Jensen-Shannon, and use "BS", "DP", and "T" for the three offline change point detection methods, i.e., the dynamic programming, binary segmentation, and threshold method, respectively. It results in a total of six versions for each two-stage method. It is notable that, ANTM employs a clustering approach to documents. Then topics are identified by summarizing the meanings of each cluster. As a fully unsupervised method, it does not allow for predefining the number of topics. Consequently, the number of topics may vary across different time points. Therefore, to find the change point among topics, it is required to manually identify the change points by comparing the extracted topics across different time points. Due to this limitation, we exclude ANTM from the evaluation of change point detection performance in experiments on syngenetic datasets.

We first focus on the change point detection performance evaluated by the P score and WindowDiff. Table B.3 summarizes the experimental results. As shown, TOPIC-PYP achieves substantially lower P score and WindowDiff than the other methods in nearly all scenarios. These results suggest that TOPI-PYP has a better performance of change point detection. It is important to note that, two-stage methods require prior knowledge of the number of change points. In our experiments, we assume the two-stage methods already know the true number of change points. However, our TOPIC-PYP method does

not require the number of change points to be determined in advance. Similarly, Topic-CD also does not require prior knowledge of the number of change points. This distinction highlights the advantage of unified methods over two-stage methods.

Table 2: The comparison results of P score (PS) and WindowDiff (WD) for different methods in four scenarios.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Method		PS	WD	PS	WD	PS	WD	PS	WD
TOPIC-PYP		0.010	0.070	0.001	0.073	0.025	0.038	0.038	0.036
Topic-CD		0.092	0.243	0.318	0.350	0.050	0.050	0.143	0.287
DTM	_CS_DP	0.367	0.367	0.017	0.217	0.250	0.250	0.250	0.275
	_CS_BS	0.367	0.367	0.058	0.342	0.225	0.268	0.288	0.338
	_CS_T	0.233	0.233	0.083	0.242	0.125	0.111	0.161	0.238
	_JS_DP	0.342	0.342	0.033	0.267	0.215	0.215	0.275	0.275
	_JS_BS	0.342	0.342	0.058	0.292	0.220	0.220	0.294	0.319
	_JS_T	0.251	0.251	0.051	0.151	0.133	0.133	0.097	0.156
D-ETM	_CS_DP	0.300	0.300	0.001	0.001	0.325	0.275	0.275	0.317
	_CS_BS	0.200	0.200	0.375	0.500	0.050	0.050	0.333	0.319
	_CS_T	0.250	0.250	0.450	0.525	0.200	0.200	0.264	0.264
	_JS_DP	0.300	0.300	0.001	0.001	0.050	0.050	0.195	0.236
	_JS_BS	0.300	0.300	0.150	0.200	0.125	0.125	0.195	0.250
	_JS_T	0.100	0.100	0.450	0.525	0.150	0.050	0.275	0.275
Rolling LDA	_CS_DP	0.500	0.500	0.075	0.100	0.050	0.050	0.250	0.275
	_CS_BS	0.500	0.500	0.075	0.100	0.220	0.220	0.362	0.350
	_CS_T	0.575	0.575	0.375	0.375	0.075	0.075	0.131	0.183
	_JS_DP	0.200	0.200	0.375	0.500	0.222	0.222	0.275	0.300
	_JS_BS	0.200	0.200	0.375	0.500	0.050	0.093	0.294	0.319
	_JS_T	0.625	0.625	0.625	0.750	0.222	0.175	0.235	0.250

We further compare the change point detection performance using precision and recall. We consider different bandwidths as $h = 0, 1, 2$. Note that when $h = 0$, the precision and recall correspond to the performance of "accurately detected" change point locations. When $h > 0$, the detected change points are allowed in a flexible interval centered by the true locations. Table 3 summarizes the results under Scenario 4, and the corresponding results under the other three scenarios are present in Appendix B.2 for brevity. In general, the results under different scenarios are similar, leading to the following conclusions. First, when $h = 0$, the TOPIC-PYP model achieves significantly higher precision and recall compared to all the other methods. Second, when $h > 0$, TOPIC-PYP maintains strong change point detection performance, with precision and recall consistently exceeding 0.9. Third, as h increases, all methods show improved precision and recall. Throughout this process, some competitors do outperform TOPIC-PYP occasionally. Finally, when $h = 2$, all methods achieve excellent results, as the conditions for detecting change points become very relaxed.

Last, we discuss the computational efficiency of TOPIC-PYP. To this end, we calculate the running time for different methods. For the unified methods, including our proposed TOPIC-PYP and Topic-CD, we measure the total runtime for the entire process. In contrast, for the two-stage methods, we only account for the runtime of topic modeling (i.e.,

Table 3: The comparison results of precision and recall for different methods with $h = 0, 1, 2$ under Scenario 4.

		Precision			Recall		
Method		$h = 0$	$h = 1$	$h = 2$	$h = 0$	$h = 1$	$h = 2$
TOPIC-PYP		0.709	0.973	0.973	0.850	0.920	1.000
Topic-CD		0.625	0.950	0.950	0.763	0.925	0.950
DTM	_CS_DP	0.275	0.675	1.000	0.275	0.581	0.950
	_CS_BS	0.300	0.725	1.000	0.300	0.725	1.000
	_CS_T	0.450	0.800	1.000	0.450	0.725	1.000
	_JS_DP	0.275	0.650	0.975	0.275	0.650	1.000
	_JS_BS	0.350	0.725	0.975	0.350	0.800	1.000
	_JS_T	0.347	0.656	1.000	0.347	0.656	1.000
D-ETM	_CS_DP	0.450	0.875	1.000	0.450	0.875	1.000
	_CS_BS	0.450	0.725	1.000	0.450	0.725	1.000
	_CS_T	0.625	0.875	1.000	0.550	0.750	1.000
	_JS_DP	0.500	0.725	0.975	0.500	0.650	1.000
	_JS_BS	0.450	0.656	0.900	0.200	0.656	0.950
	_JS_T	0.500	0.950	1.000	0.500	1.000	1.000
Rolling LDA	_CS_DP	0.450	0.800	1.000	0.450	0.800	1.000
	_CS_BS	0.225	0.625	1.000	0.225	0.625	1.000
	_CS_T	0.243	0.865	1.000	0.275	0.725	1.000
	_JS_DP	0.250	0.725	0.975	0.375	0.725	1.000
	_JS_BS	0.500	0.800	0.975	0.500	0.750	1.000
	_JS_T	0.450	0.950	1.000	0.625	1.000	1.000

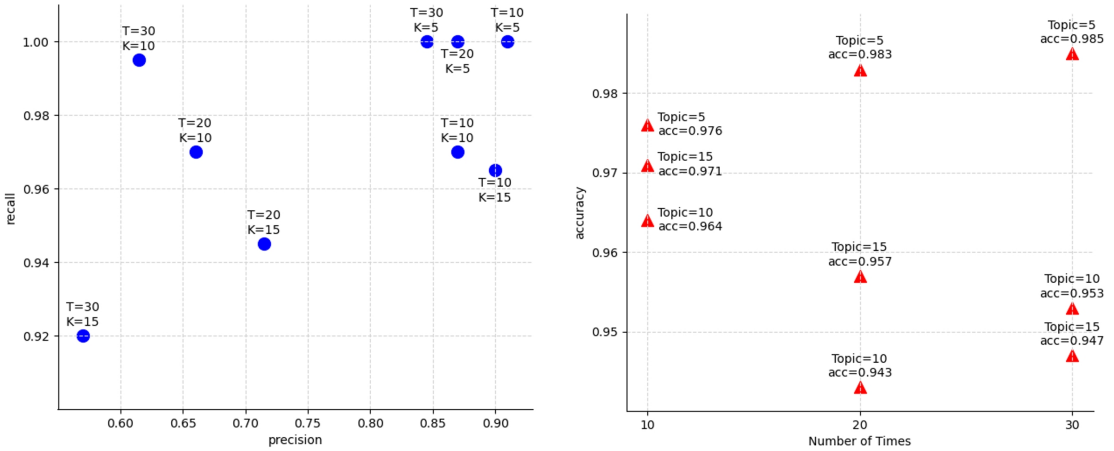
the first stage), as the second stage of change point detection is extremely fast. The implementation details and results can be found in Appendix B.3. The main findings are as follows. Compared to two-stage methods, the unified methods (including both TOPIC-PYP and Topic-CD) have significantly longer running time. This is because the unified methods combine topic modeling and change point detection, making the model structures more complicated. In comparison to Topic-CD, the running time of TOPIC-PYP is generally comparable, although slightly slower. However, TOPIC-PYP allows for change point detection for each individual topic, whereas Topic-CD can only detect change points that are shared across all topics. This is the trade-off TOPIC-PYP makes in terms of computational efficiency.

4.3 Influence of T and K

In this section, we explore the influence of period T and topic number K on the performance of TOPIC-PYP. We basically follow the generation process described in Section 4.1 to generate the synthetic documents. We then vary the period as $T = [10, 20, 30]$ and the number of topics as $K = [5, 10, 15]$. This leads to a total of $3 \times 3 = 9$ experimental settings. To generate the document-topic distributions, we set $\alpha = 0.2$ for $K = 5$, while $\alpha = 0.1$ for $K = 10$ and 15. For illustration purposes, we consider two change points for each topic, i.e., $Q_k = 2$. Specifically, we set $i_k = (3, 6)^\top$ in the case of $T = 10$. That is, at the third

and sixth moments, all topics have change points. In cases of $T = 20$ and $T = 30$, we set $i_k = (6, 12)^\top$ and $i_k = (10, 20)^\top$, respectively.

The averaged precision and recall in each experimental setup are reported in Figure 3. Except for precision and recall, we also compute the accuracy, which is defined as $accuracy = \sum_{k,t} I(\widehat{i}_{k,t} = i_{k,t})/T$, where $\widehat{i}_{k,t}$ is the true counterpart of $i_{k,t}$. As shown, the accuracy of the TOPIC-PYP model is always above 94%, and the recall of TOPIC-PYP can also be as high as 92%. These results indicate that the estimated locations of change points can better cover the true ones. The good performance of TOPIC-PYP is also robust to different periods and the number of topics. In addition, as the period T increases, the precision of TOPIC-PYP decreases, but the recall can always remain higher than 90%. This finding indicates that TOPIC-PYP has a strong recognition ability for the true change points, but it might find more change points when the period is relatively long.



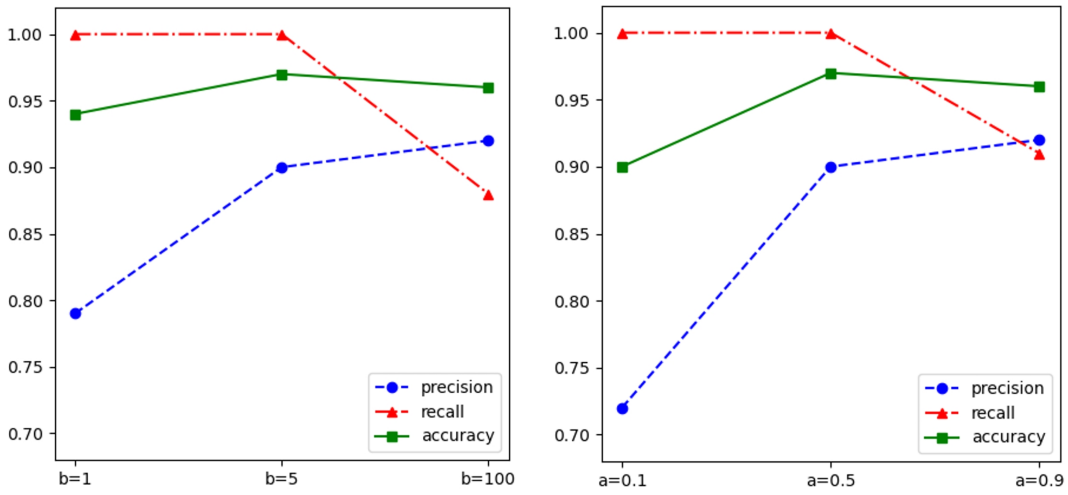
(a) (b)

Figure 3: The left panel presents the precision and recall for different T and K, and the right panel presents the accuracy for different T and K.

4.4 Inference of hyperparameters

We examine the impact of hyperparameters on the performance of the TOPIC-PYP model. To this end, we consider Scenario 2 for illustration, which assumes the presence of two change points within a time period of $T = 10$. We generally follow the experimental settings in Section 4.1. For the hyperparameters, we set $a = 0.5$, $b = 5$, $\lambda = \{2, 5\}$, $\alpha = 0.1$, and $\gamma = 0.1$ for illustration. We refer to this setup of hyperparameters as "baseline". Next, we vary the value of each hyperparameter while keeping the others fixed. We mainly focus on the hyperparameters of PYP, since the Pitman-Yor process is an important step in our proposed model. There are two hyperparameters a and b in this process. Thus we are interested to investigate the influence of a and b. The detailed results in different settings are reported in Figure 4. As shown, the accuracy of change point detection is always above 90%. In addition, in the case of $b = 5$, as a increases, we find both precision and accuracy

gradually increase, but the recall decreases. In the case of $a = 0.5$, when b increases, the precision gradually increases, the recall decreases, and the accuracy remains nearly unchanged. The investigation of other hyperparameters in TOPIC-PYP (i.e., λ, α, and γ) can be found in Appendix B.4. Results show that varying the values of λ, α, and γ has little effect on the change point detection performance of the TOPIC-PYP model.



(a) fix $a = 0.5$ (b) fix $b = 5$

Figure 4: The precision, recall, and accuracy for different combinations of (a, b) used in the data generation process.

Last, we investigate the robustness of model performance when using the incorrect hyperparameters in the PYP process. We take the case with $a = 0.5$ and $b = 5$ in the data generation process as an example. Table 4 presents the used a and b when estimating the TOPIC-PYP model and also lists the corresponding results. As shown, when using the incorrect hyperparameters, the high accuracy of topic change point detection can still be achieved, since the accuracies of all setups are above 96%. The values of recall are also close to 100%. These findings indicate that TOPIC-PYP is robust to the used hyperparameters of PYP in model estimation.

Table 4: The precision, recall, and accuracy under different combinations of (a, b) used in the estimation of TOPIC-PYP.

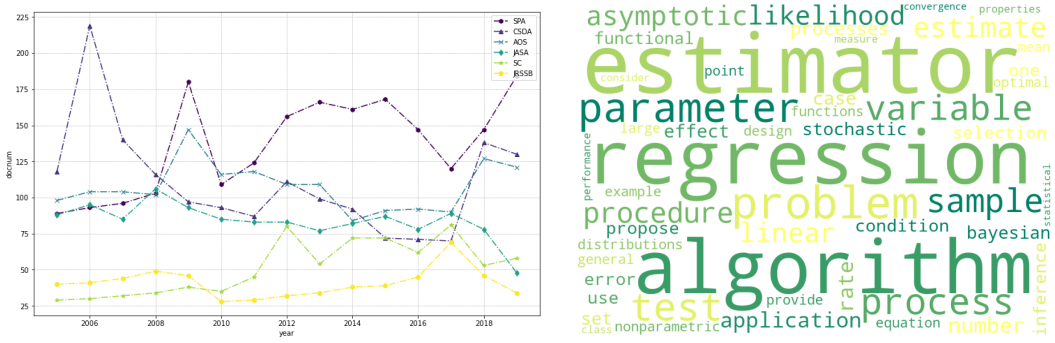
	Case 1	Case 2	Case 3	Case 4	True
a_{used}	0.1	0.9	0.5	0.5	0.5
b_{used}	5	5	1	100	5
Precision	0.84	0.93	0.9	0.97	0.92
Recall	1.00	1.00	0.98	1.00	1.00
Accuracy	0.96	0.98	0.97	0.99	0.99

5. Experiments on the Journal Dataset

5.1 Data description

Research papers published in top-tier journals often explore cutting-edge topics. Consequently, identifying the change points within the stream of published papers can effectively capture the evolution patterns within a particular discipline. For illustration purposes, we take the statistical journal dataset as an example. Specifically, according to the number of journal citations in the Web of Science (i.e., the Journal Citation Report in 2019), ten journals with the highest number of citations are selected under the category of Statistics and Probability. To get relatively focused topics, we drop two journals on applied statistics and two journals on economic statistics. The final dataset contains six statistical journals, including *Stochastic Processes and their Applications* (SPA), *Computational Statistics & Data Analysis* (CSDA), *The Annals of Statistics* (AOS), *Journal of the American Statistical Association* (JASA), *Statistics and Computing* (SC), and *Journal of the Royal Statistical Society: Series B* (JRSSB), following a decreasing order of number of citations. Among them, SPA mainly publishes papers related to the stochastic process; CSDA and SC mainly focus on computational statistics; while AOS, JASA, and JRSSB are widely recognized theoretical research journals in statistics. For each journal, we collect its published papers from 2005 to 2019. This leads to 7954 papers in total. For each paper, we collect its title, abstract, authors, and publication year.

After data collection, we apply TOPIC-PYP on the Journal dataset to explore the change points underlying the published papers. We mainly focus on the title and abstract of each paper, since they are the summary of the paper's content. We conduct some preprocessing steps before topic modeling. Specifically, we first merge the title and abstract for each paper. Then we use the *nltk* library in Python to remove numbers, punctuations, stop words, and words with lower frequency than 20 or appearing in less than 0.0005% of the whole documents. This leads to a vocabulary of 5034 words and a paper corpus of 7954 documents. The average number of words in each document is about 70.



(a) annual number of papers

(b) wordcloud of Top 50 words

Figure 5: The left panel presents the number of papers for each journal per year in the journal dataset, and the right panel is the wordcloud of the top 50 words with the highest frequency.

The left panel in Figure 5 presents the number of papers for each journal per year. Generally, the number of papers ranges from 460 to 600, with no obvious temporal pattern. As for different journals, we find the annual number of papers published by SPA is the largest (around 150), while those of CSDA, AOS, and JASA are all close to 100. The annual numbers of papers published by SC and JRSSB are relatively small (below 75). We then explore the textual content of the published papers. The right panel in Figure 5 presents the wordcloud of the top fifty words with the highest frequency in the dataset. As shown, the words "estimator", "regression" and "algorithm" enjoy high frequency in the Journal dataset.

5.2 Change point detection

To detect the change point in the Journal dataset, we apply the TOPIC-PYP model. With some preliminary analysis, we set the number of topics as $K = 20$. As for the hyperparameters, we set $\gamma = 0.1$, $a = 0.5$, $b = 5$, $\alpha = 0.1$, and $\lambda = \{2, 5\}$ for illustration purpose. After estimation of TOPIC-PYP, we find Topic 1 and Topic 15 have one change point in the year 2016, while the other topics have no change points. In topic models, the meaning of topics is often characterized by its high probability words in the topic-word distributions, which we refer to as the *representative words*. To explore the changing meanings of Topic 1 and Topic 15, we present the fifteen representative words with the highest probabilities for each topic before and after the change point. The detailed results are summarized in Figure 6. As for the other eighteen topics, their topic meanings are summarized in Appendix B.5.

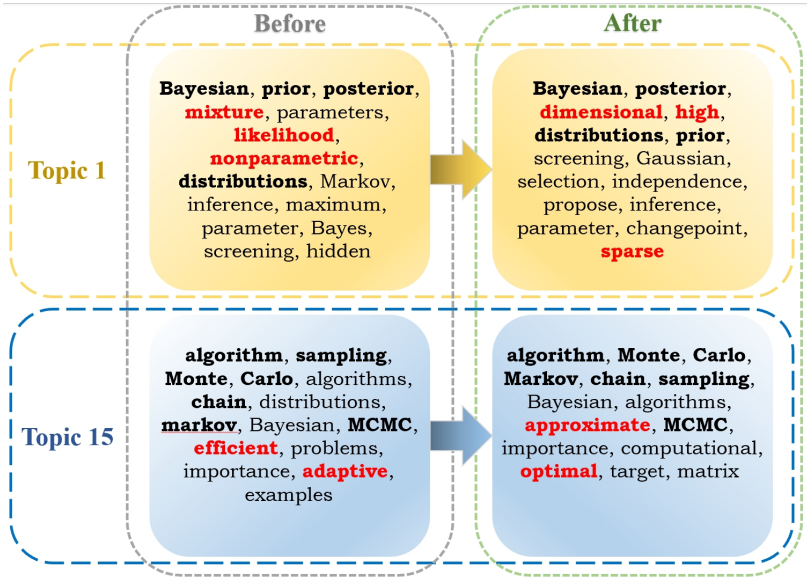


Figure 6: The representative words under Topic 1 and Topic 15 before and after the change point occurring at year 2016. Words in bold are shared by both periods before and after the change point, while words in red are typical to one period.

As shown, Topic 1 mainly discusses Bayesian analysis, since some representative words with high probabilities include "Bayesian", "prior", "posterior", and "distributions". Some typical representative words before the change point include "mixture", "likelihood", and "nonparametric", indicating the research interests mainly focus on nonparametric Bayesian methods and mixture Bayesian methods. After the change point, some typical words "high", "dimensional", and "sparse" appear, indicating Bayesian research has turned its focus on high-dimensional Bayesian methods and sparse Bayesian methods.

As for Topic 15, its representative words include "algorithm", "sampling", "month", "Carlo", "Markov", "chain", and "MCMC", which indicate this topic mainly discusses the Markov chain Monte Carlo (MCMC) algorithm. Before the change point, some typical words under this topic include "effective" and "adaptive", indicating that the MCMC method focuses on efficiency and adaptability of the algorithm in the early stage. After the change point, some typical words "approximate" and "optimal" have appeared. These results show that the MCMC algorithm puts more emphasis on optimization and approximation solutions for complex problems. Note that MCMC algorithms are often used for Bayesian models. Therefore the two topics share the same location of change point. This consistency also verifies the accuracy of the TOPIC-PYP model to some extent.

To further display the changes in topic meanings for Topic 1 and Topic 15, we investigate the frequency trend of some representative words across time. For each topic, we select four representative words. Figure 7 shows the annual frequency of documents containing each selected word. As shown, the word "Bayesian" in Topic 1, and the word "MCMC" in Topic 15 show no obvious dynamic patterns over time. This is because these two words indicate the key meanings of the two topics. The frequency of the word "nonparametric" in Topic 1 gradually decreases, while the frequencies of "dimensional" and "sparse" increase over time. A similar phenomenon is obtained for Topic 15. For example, the words "optimal" and "approximate" show increasing frequency over time.

6. Experiments on the Twitter Dataset

6.1 Data description

We apply TOPIC-PYP on a Twitter dataset to demonstrate its change point detection performance. Compared with the Journal dataset, texts in the Twitter dataset are relatively shorter. To collect the Twitter dataset, we choose the brand Burger King as an example. We use the Python package *scweet* to obtain daily tweets containing the keyword "Burger King" or its other variants. The collection period is from March 15, 2016, to April 11, 2016. This leads to a total of 28 time periods. During this period, Burger King released a new product (the Angriest Whopper) on March 29, 2016. There also existed other hot events related to Burger King from March 15 to April 10, which are summarized in Table 5. Therefore, the goal of analyzing this dataset is to explore change points related to the new product release and the hot events.

Before model building, we conduct pre-processing steps on the Twitter dataset, which are similar to those on the Journal dataset. After pre-processing, the dataset contains 24187 tweets with a vocabulary of $V = 8756$. By some preliminary analysis, we find the average length per tweet is about 10 words. The daily average number of tweets is around 750. The dynamic trend of the number of tweets is shown in the left panel of Figure 8. As shown,

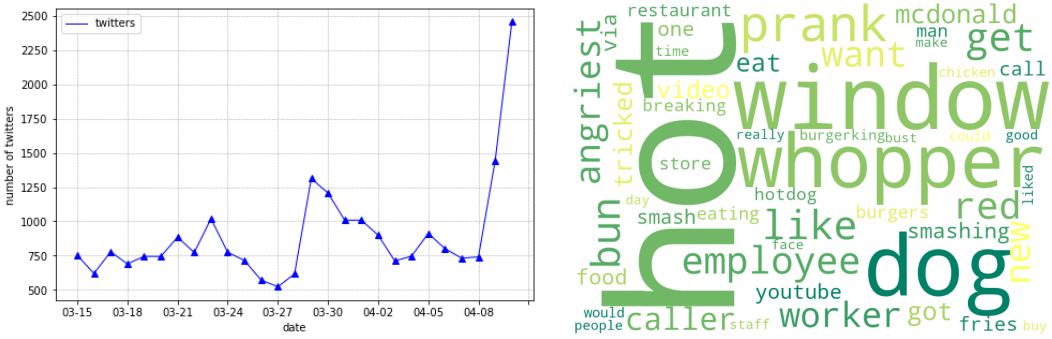
Figure 7: The annual frequencies of some representative words under Topic 1 and Topic 15. The left panel represents four words (i.e., "Bayesian", "nonparametric", "dimensional", and "sparse") selected from the meaning of Topic 1; while the right panel represents four words (i.e., "MCMC", "optimal", "approximate", and "adaptive") selected from the meaning of Topic 15.

Table 5: Five hot events in the Tweeter dataset.

Date	Hot Events	Indix
March 22	People are rallying around a Burger King ad after the Brussels terrorist attacks.	E1
March 23	FullContact buys Brewsterś technology after team got acqui-hired by Burger King owner RBI.	E2
March 24	A customer overheard a Burger King worker mock a fallen police officer.	E3
April 1	Gay man brutally attacked at a Miami Beach Burger King for kissing his boyfriend.	E4
April 9	Prank caller convinces Burger King employees to smash their windows.	E5

on March 29th, when the new product was released, the number of tweets increased to 1250. Subsequently, the number of tweets began to decline to around 750. Then around the hot event E5, there was a noticeable increase in the number of tweets, indicating a hot

discussion about this event. As shown in the right panel of Figure 8, the frequently used words include "hot" and "whopper". The uninformative words "burger king" and "http" have been removed from the figure.



(a) daily number of tweets

(b) wordcloud of Top 50 words

Figure 8: The left panel is the daily number of tweets in the Twitter dataset, and the right panel is the wordcloud of the top 50 words with the highest frequency.

6.2 Change point detection

After pre-processing, we conduct TOPIC-PYP on the Twitter dataset to find topic change points. We set the number of topics as $K = 5$, since the whole Twitter dataset focuses on Burger King and the tweets are relatively short. The hyperparameters are set as the same as those used in the Journal dataset. By TOPIC-PYP, we find a total of twenty-seven change points for five topics. The specific locations of these change points are shown in Figure 9.

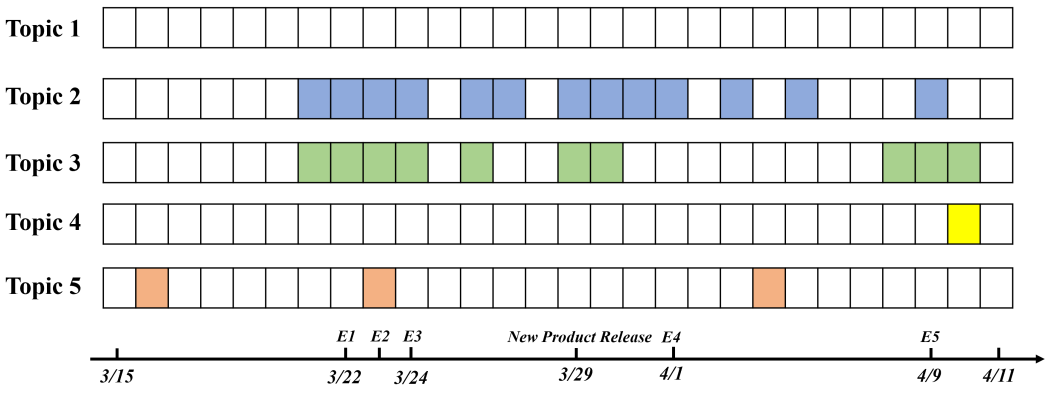


Figure 9: The location of change points (marked by the colored boxes) for five topics in the Tweeter dataset.

As shown, for Topic 1, we do not find any change points. Topic 2 has 13 change points and Topic 3 has 10 change points. In addition, Topic 2 and Topic 3 share some same locations of change points. The numbers of change points associated with Topic 4 and Topic 5 are relatively small. In general, the detected change points are consistent with the occurring time of the new product release and the five hot events.

Next, we conduct a deeper analysis of the detected change points of each topic. Figure 10 presents the representative words with high frequency for five topics in each period segmented by the change points. In general, we find our TOPIC-PYP successfully identifies topic-level transitions in the Twitter dataset that align with the real-life events. These topical shifts illustrate the model's effectiveness in identifying meaningful change points and uncovering temporally coherent patterns. For instance, the release time of the new Angriest Whopper product has been detected as a change point by Topic 2 and Topic 3. Note that, Angriest Whopper is a spicy red bread burger. Tweets discussing Angriest Whopper often use words such as "red bun" and "hot". Event E2 (occurred on March 23) aligns with a topic shift in Topic 2 identified by TOPIC-PYP. This shift highlights the words "brewster" and "technology", coinciding with FullContact acquiring Brewster's technology after the team joined Burger King's owner. Similarly, Event E3 (occurred on March 24) matches another change point identified by Topic 2, as words like "officer", "police", and "fallen" rise to prominence after a Burger King worker mocked a fallen police officer. Other events show similar patterns. For example, the introduction of the Angriest Whopper (late March to early April) leads to shifts in Topics 2 and 3, with words like "angriest", "red" and "bun" becoming prominent. The Belgian terrorist attacks (Event E1 on March 22) are reflected in Topic 3 through words like "brussels", "attacks" and "rallying". Violent incidents, such as the assault of a gay man at a Miami Beach Burger King (Event E4 on April 1), trigger a change in Topic 2, surfacing words like "gay", "attacked", and "kissing". Finally, the prank involving employees smashing store windows (Event E5 on April 9) generates new change points in Topics 2 and 3, with words like "prank", "windows" and "smash" appearing suddenly. In summary, we find TOPIC-PYP can capture topic changes in everyday discussions. It also adapts dynamically to breaking news and unexpected events. This ability creates a temporal narrative that mirrors real-world developments.

6.3 Model comparison

For comparison purposes, we also analyze the Twitter dataset using the other models except TOPIC-PYP. Consistent with the experiments on synthetic datasets, we consider Topic-CD (Lu et al., 2022), DTM (Blei and Lafferty, 2006b), D-ETM (Dieng et al., 2019), Rolling LDA (Rieger et al., 2022), and ANTM (Rahimi et al., 2024) as five competitors. Then we compare TOPIC-PYP with these methods from two perspectives. The first perspective is the topic modeling performance, which is evaluated by the coherence score. As shown by Table 6, TOPIC-PYP achieves the highest coherence score among all methods, indicating that the topics generated by TOPIC-PYP are of the highest quality.

The second perspective is the performance of change point detection. Note that in real data analysis, the true change points are typically unknown. To address this issue, we attempt to leverage the real-world context of the data. As we mentioned before, during the time period for analysis, Burger King launched a new product, and five other hot events

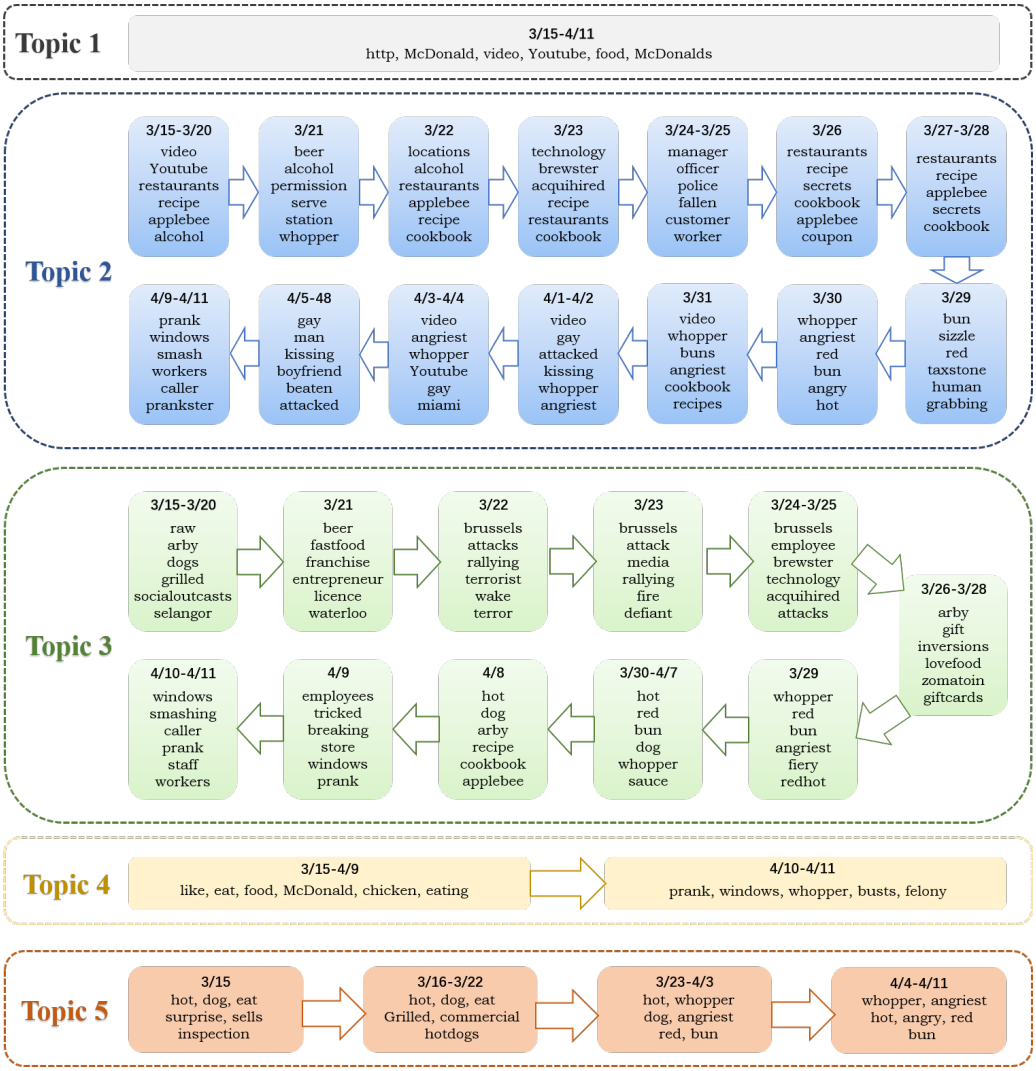


Figure 10: The representative words of five topics extracted in the Twitter dataset. For each topic, we show the changes of representative words in each period split by the detected change points.

occurred. These six events are likely to cause shifts in public opinion and can therefore be treated as potential change points. Therefore, we regard the six events as the true change points and compare the detection accuracy of different methods in catching these six change points. The results are shown in Table 7. It is noteworthy that, the DTM, D-ETM, and Rolling LDA methods are combined with a second stage of change point detection algorithms. As for ANTM, we have to manually identify the change points by comparing the extracted topics at different time points. For each method, we consider an event to be successfully detected, if at least one topic identifies the occurrence time of the event as a change point. As shown, both TOPIC-PYP and Topic-CD successfully detect the six events

as change points, achieving a detection accuracy of 100%. However, Topic-CD identifies a total of 16 change points and forces all five generated topics to include these 16 change points. It is because of this strict requirement, Topic-CD results in poorer topic quality compared to TOPIC-PYP, as evident from the results shown in Table 6. The detection performances of the other methods (including DTM, D-ETM, Rolling LDA, and ANTM) are all very poor. For example, none of these methods identifies the new product release as a change point, which is a highly significant event within the analysis time period.

Table 6: The coherence score of different methods on the Twitter dataset.

TOPIC-PYP	Topic-CD	DTM	D-ETM	Rolling LDA	ANTM
6.990	1.689	0.220	2.118	0.078	0.036

Table 7: The detection results of different methods for the new product release and five hot events. The total detection accuracy is also reported.

Method		New Product	E1	E2	E3	E4	E5	Accuracy
TOPIC-PYP		✓	✓	✓	✓	✓	✓	100%
Topic-CD		✓	✓	✓	✓	✓	✓	100%
DTM	_CS_DP		✓		✓		✓	50%
	_CS_BS		✓		✓		✓	50%
	_CS_T		✓	✓	✓		✓	67%
	_JS_DP		✓		✓	✓	✓	67%
	_JS_BS		✓		✓	✓	✓	67%
	_JS_T		✓	✓	✓		✓	67%
D-ETM	_CS_DP		✓		✓		✓	50%
	_CS_BS		✓		✓		✓	50%
	_CS_T		✓		✓		✓	50%
	_JS_DP		✓		✓		✓	50%
	_JS_BS		✓		✓		✓	50%
	_JS_T		✓		✓		✓	50%
Rolling LDA	_CS_DP		✓		✓		✓	50%
	_CS_BS		✓		✓		✓	50%
	_CS_T		✓	✓			✓	50%
	_JS_DP		✓		✓		✓	50%
	_JS_BS		✓		✓	✓	✓	67%
	_JS_T		✓	✓			✓	50%
ANTM			✓				✓	33%

7. Conclusion and Discussion

In this work, we present a novel change point detection model called TOPIC-PYP, which uses the Pitman-Yor process to model the changing meanings of each topic over time. This approach combines topic modeling and topic-level change point detection in a unified framework and is evaluated using a series of experiments on both synthetic and two real

datasets. Compared with the state-of-the-art methods in topic change point detection, we demonstrate the effectiveness of TOPIC-PYP in detecting fine-grained changes for each topic over time.

There are several directions for future work. First, one can explore other Bayesian nonparametric processes for change point detection in topic models, such as the Polya tree process, the hierarchical Dirichlet process, and the Indian buffet process. Second, as a probabilistic topic model, TOPIC-PYP has some limitations compared to neural topic models, particularly in capturing complex relationships and handling large-scale dynamic documents. Thus we aim to extend TOPIC-PYP by incorporating modern deep learning techniques in the future. For instance, we can embed words into a semantic embedding space using LLMs or leverage neural network architectures (e.g., variational autoencoders) to improve the modeling performance. Third, although we have established the posterior consistency properties for $\mathcal{I}$ and Φ, deeper asymptotic analysis for other parameters is of great interest and could offer additional insights. This is a promising direction for future research. Last, due to the complex structure of TOPIC-PYP, its computational efficiency is not very high. This constitutes one of the main limitations of TOPIC-PYP. Therefore, further improving the computational efficiency of TOPIC-PYP should be an important area for future research.

Conflict of Interest

The authors declare they have no conflict of interest.

Acknowledgments

Xiaoling Lu (xiaolinglu@ruc.edu.cn) is the corresponding author. We thank Jie Song (at Capital University of Economics and Business) and Huiyun Tang (at Renmin University of China) for their helpful discussions. We also sincerely thank the Action Editor and three reviewers for their thoughtful comments and constructive suggestions. This work is supported by National Natural Science Foundation of China (No. 72371241, 72171229), the MOE Project of Key Research Institute of Humanities and Social Sciences (No. 22JJD110001), and the Big Data and Responsible Artificial Intelligence for National Governance, Renmin University of China.

Appendix A. Technical Materials for TOPIC-PYP

A.1 Understanding PYP using the Chinese restaurant process

The Pitman-Yor process can also be described from the perspective of the Chinese restaurant process (CRP). CRP is a sampling perspective on the Pitman-Yor process. It explains how samples (e.g., words) are drawn given the distribution from the Pitman-Yor process. Below, we provide a detailed description.

Imagine that we need to select a sequence of words $w_1, w_2, ...$ For each word w_i, we already know it represents topic k, i.e., $z_i = k$. Then under a segment s, w_i follows a multinomial distribution with parameter $\phi_{k,s}$. Thus the target here is to select a specific value v_i for w_i from the vocabulary using the above multinomial distribution. To describe PYP using the Chinese restaurant process, we assume each word w_i refers to a new customer entering the restaurant, its corresponding topic indicator z_i refers to a table, and the selected value v_i from the vocabulary refers to the dish enjoyed by the customer. Recall that, each time a new customer (word) enters the restaurant, it should make two choices. First, it should choose a table (topic). In this step, it either chooses an existing table (topic) or a new table (topic). After sitting at the table, it should choose a dish (value in vocabulary) on the table. If the customer joins an existing table, it automatically selects the dish on that table. Note that each table can only have one dish. If the customer opens a new table, it selects a dish from the menu h_k. Let N be the total number of customers. Let $\mathcal{M}$ contain the indices of customers choosing to open new tables, and M be the corresponding count. Thus M is also the count of tables. For the mth table with $m \in \mathcal{M}$, let v_m be the dish on the table. Then let n_{v_m} be the count of dish v_m selected by N customers. To assist the mathematical derivations for model estimation, we further denote τ_{v_m} to be the count of dish v_m selected but by new tables. In other words, τ_{v_m} is the count of new tables associated with the dish v_m. Note that $\tau_{v_m} \geq 1$, since a dish can be selected by more than one table. We should also have $M = \sum_{m \in \mathcal{M}} \tau_{v_m}$. In addition, since a dish v_m can be selected by more than one customer, we should have $\tau_{v_m} \leq n_{v_m}$.

Based on the above notations, the Pitman-Yor process can be described in the following steps; see also Figure A.1 for an illustration.

Step 1 When the first customer w_1 walks into the restaurant, he/she decides to open a new table and then select a dish v_1 from the menu h_k. Thus, in this step, we have $N = 1$ and $M = 1$. For dish v_1, we have $n_{v_1} = 1$ and $\tau_{v_1} = 1$.

Step 2 When the second customer w_2 walks into the restaurant, he/she sees that there already exists one table. Assume he/she decides to open a new table with the probability of $(b + a)/(b + 1)$, and selects the dish v_2 for this table from the menu h_k. Then in this step, we have $N = 2, M = 2$, $n_{v_1} = 1, \tau_{v_1} = 1$ for dish v_1, and $n_{v_2} = 1, \tau_{v_2} = 1$ for dish v_2.

Step 3 When the third customer w_3 enters the restaurant, he/she sees two tables, and he/she can open the new table with the probability $(b + 2a)/(b + 2)$, or sits next to the customer w_1 or w_2 with the probability $(1 - a)/(b + 2)$. Assume that the customer w_3 sits at the table that customer w_1 selected, and then automatically selects the dish v_1

on that table. Then in this step, we have $N = 3$ and $M = 2$, $n_{v_1} = 2, \tau_{v_1} = 1$ for dish v_1, and $n_{v_2} = \tau_{v_2} = 1$ for dish v_2.

Generally, when the $(N+1)$-th customer w_{N+1} enters the restaurant, he/she observes M existing tables, which opened by customers in the set $\mathcal{M}$. For the table selected by the mth customer with $m \in \mathcal{M}$, it is associated with a dish v_m from the menu h_k. Note that a dish v_m may be selected by more than one customer on more than one table. The total number of customers selecting the dish v_m is n_{v_m}. Then for the $(N+1)$-th customer, the probability of joining an existing table m is proportional to $n_{v_m} - a$, where a is the discount parameter; while the probability of opening a new table is proportional to $b + aM$, where b is the concentration parameter. Formally, the probabilities of choosing tables are:

$$P(\text{join table } m) = \frac{n_{v_m} - a}{b + N}, \quad P(\text{open new table}) = \frac{b + aM}{b + N}.$$

After sitting at the selected table, the customer further selects a dish. If the customer joins an existing table m, he/she automatically selects the dish v_m on the table. If the customer opens a new table, he/she selects a dish v_{N+1} from the menu h_k, i.e., from the multinomial distribution defined by h_k.

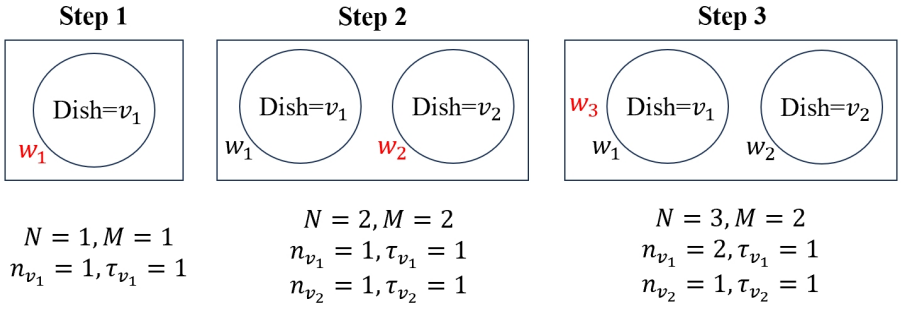


Figure A.1: Illustration of the Chinese restaurant process for PYP

A.2 The derivation details of joint posterior distribution

In this section, we give the detailed derivations of the joint posterior distribution. Recall the joint posterior distribution of all variables $\{\Pi, \mathcal{H}, \mathcal{I}, \Phi, \Theta, \mathcal{Z}, \Delta\}$ is given below.

$$\begin{aligned} &f(\Pi, \mathcal{I}, \mathcal{H}, \Phi, \Theta, \mathcal{Z}, \Delta | \mathcal{W}, \Xi) \\ \propto &f(\Pi \mid \lambda_0, \lambda_1) f(\mathcal{I} \mid \Pi) f(\mathcal{H} \mid \gamma) f(\Phi \mid a, b, \mathcal{H}) f(\Theta \mid \alpha) f(\mathcal{Z} \mid \Theta) f(\mathcal{W}, \Delta \mid \Phi, \mathcal{Z}, \mathcal{I}). \end{aligned}$$

We aim to simplify the posterior distribution. To this end, we take the following steps by integrating out Π, Θ, Φ and $\mathcal{H}$.

- Step 1: Integrate out Π.

Note that the variable Π only appears in $f(\Pi \mid \lambda_0, \lambda_1)$ and $f(\mathcal{I} \mid \Pi)$. Each π_k follows the Beta distribution with parameters λ_0 and λ_1 and each $i_{k,t}$ follows the Bernoulli distribution

with parameter π_k. Then we have

$$f(\Pi \mid \lambda_0, \lambda_1) = \prod_{k=1}^{K} \frac{1}{\text{Beta}(\lambda_0, \lambda_1)} \pi_k^{\lambda_0 - 1} (1 - \pi_k)^{\lambda_1 - 1}, \text{where } \pi_k \in (0, 1).$$
$$f(\mathcal{I} \mid \Pi) = \prod_{k=1}^{K} \prod_{t=1}^{T} \pi_k^{i_{k,t}} (1 - \pi_k)^{(1 - i_{k,t})}, \text{where } i_{k,t} \in \{0, 1\}.$$

Since the Beta distribution and Bernoulli distribution are conjugate, we can integrate out Π and have

$$f(\mathcal{I}|\lambda_0, \lambda_1) = \int f(\Pi \mid \lambda_0, \lambda_1) f(\mathcal{I} \mid \Pi) d\Pi = \prod_{k}^{K} \frac{\text{Beta}(\lambda_0 + \sum_{t=1}^{T} i_{k,t}, \lambda_1 + T - \sum_{t=1}^{T} i_{k,t})}{\text{Beta}(\lambda_0, \lambda_1)}. \tag{A.1}$$

- Step 2: Integrate out Θ.

Note that Θ appears in $f(\Theta \mid \alpha)$ and $f(\mathcal{Z} \mid \Theta)$. The prior distribution for each $\theta_{t,d}$ is Dirichlet with parameter α and the probability distribution for each $z_{t,d,n}$ is multinomial with parameter $\theta_{t,d}$. Then we have

$$f(\Theta \mid \alpha) = \frac{1}{\text{Beta}_K(\alpha)} \prod_{t=1}^{T} \prod_{d=1}^{D_t} \prod_{k=1}^{K} (\theta_{t,d,k})^{\alpha - 1},$$
$$f(\mathcal{Z} \mid \Theta) = \prod_{t=1}^{T} \prod_{d=1}^{D_t} \prod_{k=1}^{K} (\theta_{t,d,k})^{\sum_v m_{k,t,d,v}},$$

where $m_{k,t,d,v}$ represents the number of word v under topic k in document d at time t. Then $\sum_v m_{k,t,d,v}$ represents the summation of $m_{k,t,d,v}$ over all V words in the dictionary. We can easily verify that $\sum_v m_{k,t,d,v} = \sum_{n=1}^{N_{t,d}} I(z_{t,d,n} == k)$. That is, $\sum_v m_{k,t,d,v}$ equals to the total number of words representing topic k in document d at time t. Then by integrating out Θ, we have

$$f(\mathcal{Z} \mid \alpha) = \int f(\Theta \mid \alpha) f(\mathcal{Z} \mid \Theta) d\Theta = \prod_{t=1}^{T} \prod_{d=1}^{D_t} \frac{\text{Beta}_K\left(\sum_v m_{t,d,v} + \alpha\right)}{\text{Beta}_K(\alpha)}, \tag{A.2}$$

where $\sum_v m_{t,d,v} = (\sum_v m_{1,t,d,v}, \ldots, \sum_v m_{K,t,d,v})^\top$ is a vector with K dimension.

- Step 3: Integrate out Φ.

Given each $\phi_{k,s}$ follows the Pitman-Yor process and also involves in the likelihood of words, it is not easy to integrate out Φ. To this end, we refer to Theorem 17 in Buntine and Hutter (2010). Specifically, we need to involve an augmentation variable $\mathcal{T}$. Recall in the Chinese restaurant process described in Appendix A.1, we define the count variables τ_{v_m} and n_{v_m} for dish v_m. We then extend this definition to the setup of TOPIC-PYP. Let $\tau_{k,s,v}$ be the count of new tables with the selected word v representing topic k in segment s. Then

let $\mathcal{T}_{k,s} = \sum_v \tau_{k,s,v}$, and $\mathcal{T} = \{\tau_{k,s,v}, k = 1, \ldots, K, 1 \leq s \leq S_k, 1 \leq v \leq V\}$. Let $n_{k,s,v}$ be the count of selected word v representing topic k in segment s. Then let $N_{k,s} = \sum_v n_{k,s,v}$.

Based on Theorem 17 in Buntine and Hutter (2010), we have the following joint conditional distribution by integrating out Φ, i.e.,

$$\begin{aligned} f(\mathcal{W}, \mathcal{T} \mid \mathcal{Z}, a, b, \mathcal{I}, \mathcal{H}) &= \int f(\Phi \mid a, b, \mathcal{H}) f(W, \mathcal{T} \mid \Phi, \mathcal{Z}, \mathcal{I}) d\Phi \\ &= \left(\prod_{k=1}^{K} \prod_{s=1}^{S_k} \frac{(b \mid a)_{\mathcal{T}_{k,s}}}{(b)_{N_{k,s}}} \prod_{v=1}^{V} S^{n_{k,s,v}}_{\tau_{k,s,v},a} \right) \times \left(\prod_{k=1}^{K} \prod_{v=1}^{V} h_{k,v}^{\sum_s \tau_{k,s,v}} \right). \end{aligned} \tag{A.3}$$

Here $(x)_N$ denotes the Pochhammer symbol, i.e., $x(x+1)\ldots(x+N-1)$ and $(x|y)_N$ denotes the Pochhammer symbol with increment y, i.e., $x(x+y)\ldots(x+(N-1)y)$. $S^N_{M,a}$ is a Stiring number of the second kind with a linear recursion, i.e., $S^{N+1}_{M,a} = S^N_{M-1,a} + (N - Ma)S^N_{M,a}$. Please see Lemma 14 and Lemma 16 in Buntine and Hutter (2010) for more detailed definitions.

- Step 4: Integrate out $\mathcal{H}$.

Note that we have $f(\mathcal{H} \mid \gamma)$ by definition, i.e.,

$$\begin{aligned} f(\mathcal{H} \mid \gamma) &= \prod_{k=1}^{K} f(h_k \mid \gamma) = \prod_{k=1}^{K} f(h_k \mid \gamma) = \prod_{k=1}^{K} \frac{1}{\text{Beta}_V(\gamma)} \prod_{v=1}^{V} h_{k,w}^{\gamma-1} \\ &= \left(\prod_{k=1}^{K} \frac{1}{\text{Beta}_V(\gamma)} \right) \times \left(\prod_{k=1}^{K} \prod_{v=1}^{V} h_{k,w}^{\gamma-1} \right). \end{aligned} \tag{A.4}$$

Then combining (A.3) and (A.4), we can integrate out $\mathcal{H}$ and have the following

$$\begin{aligned} f(\mathcal{W}, \mathcal{T} \mid \mathcal{Z}, a, b, \mathcal{I}, \gamma) &= \int f(\mathcal{W}, \mathcal{T} \mid \mathcal{Z}, a, b, \mathcal{I}, \mathcal{H}) f(\mathcal{H} \mid \gamma) d\mathcal{H} \\ &= \left(\prod_{k=1}^{K} \prod_{s=1}^{S_k} \frac{(b \mid a)_{\mathcal{T}_{k,s}}}{(b)_{N_{k,s}}} \prod_{v=1}^{V} S^{n_{k,s,v}}_{\tau_{k,s,v},a} \right) \times \left(\prod_{k=1}^{K} \frac{\text{Beta}_V\left(\sum_s \tau_{k,s,v} + \gamma\right)}{\text{Beta}_V(\gamma)} \right). \end{aligned} \tag{A.5}$$

- Step 5: Replace $\mathcal{T}$ by Δ.

Based on the above four steps, we have the following

$$f(\mathcal{I}, \mathcal{Z}, \mathcal{T} | \mathcal{W}, \Xi) \propto f(\mathcal{I} \mid \lambda_0, \lambda_1) f(\mathcal{Z} \mid \alpha) f(\mathcal{W}, \mathcal{T} \mid \mathcal{Z}, a, b, \mathcal{I}, \gamma).$$

Note that $f(\mathcal{I} \mid \lambda_0, \lambda_1)$, $f(\mathcal{Z} \mid \alpha)$, and $f(\mathcal{W}, \mathcal{T} \mid \mathcal{Z}, a, b, \mathcal{I}, \gamma)$ are given in (A.1), (A.2), and (A.5) respectively. Then we combine (A.2) and (A.5), which leads to

$$\begin{aligned} f(\mathcal{W}, \mathcal{Z}, \mathcal{T} \mid a, b, \mathcal{I}, \gamma, \alpha) &= f(\mathcal{W}, \mathcal{T} \mid \mathcal{Z}, a, b, \mathcal{I}, \gamma) f(\mathcal{Z} \mid \alpha) \\ &= \left(\prod_{k=1}^{K} \prod_{s=1}^{S_k} \frac{(b \mid a)_{\mathcal{T}_{k,s}}}{(b)_{N_{k,s}}} \prod_{v=1}^{V} S^{n_{k,s,v}}_{\tau_{k,s,v},a} \right) \times \left(\prod_{k=1}^{K} \frac{\text{Beta}_V\left(\sum_s \tau_{k,s,v} + \gamma\right)}{\text{Beta}_V(\gamma)} \right) \\ &\times \left(\prod_{t=1}^{T} \prod_{d=1}^{D_t} \frac{\text{Beta}_K\left(\sum_w m_{t,d,w} + \alpha\right)}{\text{Beta}_K(\alpha)} \right). \end{aligned} \tag{A.6}$$

Note that $\mathcal{T}$ indicates the sum of tables and Δ serves as table indicator for each word. Theorem 1 in Chen et al. (2011) gives the relationship between $\mathcal{T}$ and Δ, i.e.,

$$f(\mathcal{W},\mathcal{T})=\prod_{k=1}^{K}\prod_{s=1}^{S_k}\prod_{v=1}^{V}\frac{n_{k,s,v}!}{\tau_{k,s,v}!(n_{k,s,v}-\tau_{k,s,v})!}f(\mathcal{W},\Delta). \tag{A.7}$$

Then combining (A.6) and (A.7), we have the following

$$\begin{aligned} f(\mathcal{W},\mathcal{Z},\Delta\mid\mathcal{I},a,b,\gamma,\alpha)=&\left(\prod_{k=1}^{K}\prod_{s=1}^{S_k}\frac{(b\mid a)_{\mathcal{T}_{k,s}}}{(b)_{N_{k,s}}}\prod_{v=1}^{V}S^{n_{k,s,v}}_{\tau_{k,s,v},a}\frac{\tau_{k,s,v}!(n_{k,s,v}-\tau_{k,s,v})!}{n_{k,s,v}!}\right)\\ \times&\left(\prod_{k=1}^{K}\frac{\mathrm{Beta}_V\left(\sum_s\tau_{k,s,v}+\gamma\right)}{\mathrm{Beta}_V(\gamma)}\right)\left(\prod_{t=1}^{T}\prod_{d=1}^{D_t}\frac{\mathrm{Beta}_K\left(\sum_v m_{t,d,v}+\alpha\right)}{\mathrm{Beta}_K(\alpha)}\right). \end{aligned} \tag{A.8}$$

- Step 6: Obtain the final joint posterior distribution.

Finally, by combining (A.1) and (A.8), we can obtain the joint posterior distribution $f(\mathcal{I},\Delta,\mathcal{Z}\mid\mathcal{W},\Xi)$ as follows.

$$\begin{aligned} &f(\mathcal{I},\Delta,\mathcal{Z}\mid\mathcal{W},\Xi)\propto f(\mathcal{W},\mathcal{Z},\Delta,\mid\mathcal{I},a,b,\gamma,\alpha)f(\mathcal{I}\mid\lambda_0,\lambda_1)\\ \propto&\left(\prod_{k=1}^{K}\prod_{s=1}^{S_k}\frac{(b\mid a)_{\mathcal{T}_{k,s}}}{(b)_{N_{k,s}}}\prod_{v=1}^{V}S^{n_{k,s,v}}_{\tau_{k,s,v},a}\frac{\tau_{k,s,v}!(n_{k,s,v}-\tau_{k,s,v})!}{n_{k,s,v}!}\right)\\ \times&\left(\prod_{k=1}^{K}\frac{\mathrm{Beta}_V\left(\sum_s\tau_{k,s,v}+\gamma\right)}{\mathrm{Beta}_V(\gamma)}\right)\left(\prod_{t=1}^{T}\prod_{d=1}^{D_t}\frac{\mathrm{Beta}_K\left(\sum_v m_{t,d,v}+\alpha\right)}{\mathrm{Beta}_K(\alpha)}\right)\\ \times&\left(\prod_{k=1}^{K}\frac{\mathrm{Beta}(\lambda_0+\sum_{t=1}^{T}i_{k,t},\lambda_1+T-\sum_{t=1}^{T}i_{k,t})}{\mathrm{Beta}(\lambda_0,\lambda_1)}\right). \end{aligned} \tag{A.9}$$

A.3 Derivation details of $f(\Delta,\mathcal{Z}|\mathcal{I},\mathcal{W},\Xi)$

Given the joint posterior distribution $f(\mathcal{I},\Delta,\mathcal{Z}\mid\mathcal{W},\Xi)$ in (A.9), we can obtain the posterior distribution of $\{\Delta,\mathcal{Z}\}$ as follows:

$$\begin{aligned} &f(\mathcal{Z},\Delta\mid\mathcal{I},\mathcal{W},\Xi)\propto f(\mathcal{I},\Delta,\mathcal{Z}\mid\mathcal{W},\Xi)\\ \propto&\left(\prod_{k=1}^{K}\prod_{s=1}^{S_k}\frac{(b\mid a)_{\mathcal{T}_{k,s}}}{(b)_{N_{k,s}}}\prod_{v=1}^{V}S^{n_{k,s,v}}_{\tau_{k,s,v},a}\frac{\tau_{k,s,v}!(n_{k,s,v}-\tau_{k,s,v})!}{n_{k,s,v}!}\right)\\ \times&\left(\prod_{k=1}^{K}\frac{\mathrm{Beta}_V\left(\sum_s\tau_{k,s,v}+\gamma\right)}{\mathrm{Beta}_V(\gamma)}\right)\left(\prod_{t=1}^{T}\prod_{d=1}^{D_t}\frac{\mathrm{Beta}_K\left(\sum_v m_{t,d,v}+\alpha\right)}{\mathrm{Beta}_K(\alpha)}\right). \end{aligned} \tag{A.10}$$

Next, assume the number of word count variables and table count variables are those excluding the n-th word in document d in the t-th moment (i.e., the word $w_{t,d,n}$). Then we can derive the posterior distribution of $(\mathcal{Z},\Delta)$ without $(z_{t,d,n},\delta_{t,d,n})$, which is similar with (A.10). We denote it by $f\left(\mathcal{Z}_{(-t,d,n)},\Delta_{(-t,d,n)}\mid\mathcal{I},\mathcal{W},\Xi\right)$. Further assume the value of $w_{t,d,n}$

is v in the dictionary, the topic indicators $(z_{t,d,n} = k,\ \delta_{t,d,n} = \delta)$, and the corresponding segment of time t is $s_k = s$. Then we can derive the joint distribution of $(\mathcal{Z}_{(-t,d,n)}, \Delta_{(-t,d,n)})$ and $(z_{t,d,n} = k, \delta_{t,d,n} = \delta)$ as follows.

$$
\begin{aligned}
&P\left(z_{t,d,n} = k, \delta_{t,d,n} = \delta, \mathcal{Z}_{(-t,d,n)}, \Delta_{(-t,d,n)} \mid \mathcal{I}, \mathcal{W}, \Xi\right) \\
&= \left(\prod_{k^*=1}^{K} \prod_{s^* \neq s} \frac{(b \mid a)_{\mathcal{T}_{k^*,s^*}}}{(b)_{N_{k^*,s^*}}} \prod_{v^*=1}^{V} S^{n_{k^*,s^*,v^*}}_{\tau_{k^*,s^*,v^*},a} \frac{\tau_{k^*,s^*,v^*}!\,(n_{k^*,s^*,v^*} - \tau_{k^*,s^*,v^*})!}{n_{k^*,s^*,v^*}!} \right) \\
&\times \frac{(b \mid a)_{\mathcal{T}'_{k,s}}}{(b)_{N'_{k,s}}} \left(\prod_{v^* \neq v} S^{n_{k,s,v^*}}_{\tau_{k,s,v^*},a} \frac{\tau_{k,s,v^*}!\,(n_{k,s,v^*} - \tau_{k,s,v^*})!}{n_{k,s,v^*}!} \right) \times S^{n'_{k,s,v}}_{\tau'_{k,s,v},a} \frac{\tau'_{k,s,v}!(n'_{k,s,v} - \tau'_{k,s,v})!}{n'_{k,s,v}!} \\
&\times \left(\prod_{k^* \neq k} \frac{\mathrm{Beta}_V\left(\sum_{s^*=1}^{S_k} \tau_{k^*,s^*,v^*} + \gamma\right)}{\mathrm{Beta}_V(\gamma)} \right) \times \left(\prod_{t^*=1}^{T} \prod_{(t^*,d^*) \neq (t,d)}^{D_t} \frac{\mathrm{Beta}_K\left(\sum_{v^*=1}^{V} m_{t^*,d^*,v^*} + \alpha\right)}{\mathrm{Beta}_K(\alpha)} \right) \\
&\times \frac{\mathrm{Beta}_V\left(\sum_{s^*=1}^{S_k} \tau'_{k,s^*,v^*} + \gamma\right)}{\mathrm{Beta}_V(\gamma)} \frac{\mathrm{Beta}_K\left(\sum_{v^*} m'_{t,d,v^*} + \alpha\right)}{\mathrm{Beta}_K(\alpha)}.
\end{aligned}
\tag{A.11}
$$

Note that, in the above equation, the notations τ_{k^*,s^*,v^*}, n_{k^*,s^*,v^*} and m_{k^*,t^*,d^*,v^*} for any k^*, s^*, v^* represent the corresponding values without the word $w_{t,d,n}$. However, the notations $\tau'_{k,s,v}$, $n'_{k,s,v}$ and $m'_{k,t,d,v}$ are the corresponding counterparts adding the information of word $w_{t,d,n}$. It is noteworthy that, $z_{t,d,n} \in \{1, 2, ..., K\}$ and $\delta_{t,d,n} \in \{0, 1\}$. We then derive the posterior distributions for $\{z_{t,d,n} = k, \delta_{t,d,n} = 0\}$ and $\{z_{t,d,n} = k, \delta_{t,d,n} = 1\}$ with $1 \leq k \leq K$ separately. When taking into account $\{z_{t,d,n} = k, \delta_{t,d,n} = 0\}$, the two groups of variables $(\tau'_{k,s,v}, n'_{k,s,v}, m'_{k,t,d,v})$ (including word $w_{t,d,n}$) and $(\tau_{k,s,v}, n_{k,s,v}, m_{k,t,d,v})$ (excluding word $w_{t,d,n}$) will satisfy the following relationships.

$$
\tau'_{k,s,v} = \tau_{k,s,v}, \quad n'_{k,s,v} = n_{k,s,v} + 1, \quad m'_{k,t,d,v} = m_{k,t,d,v} + 1. \tag{A.12}
$$

On the contrary, when we have $\{z_{t,d,n} = k, \delta_{t,d,n} = 1\}$, the two groups of variables will satisfy the following relationships.

$$
\tau'_{k,s,v} = \tau_{k,s,v} + 1, \quad n'_{k,s,v} = n_{k,s,v} + 1, \quad m'_{k,t,d,v} = m_{k,t,d,v} + 1. \tag{A.13}
$$

Based on the relationships in (A.12) and (A.13), the conditional probabilities of $\{z_{t,d,n} = k, \delta_{t,d,n} = 0\}$ and $\{z_{t,d,n} = k, \delta_{t,d,n} = 1\}$ can be derived as follows.

$$
\begin{aligned}
f_{k0} &= P\left(z_{t,d,n} = k, \delta_{t,d,n} = 0 \mid \mathcal{Z}_{(-t,d,n)}, \Delta_{(-t,d,n)}, \mathcal{I}, \mathcal{W}, \Xi\right) \\
&= \frac{P\left(z_{t,d,n} = k, \delta_{t,d,n} = 0, \mathcal{Z}_{(-t,d,n)}, \Delta_{(-t,d,n)} \mid \mathcal{I}, \mathcal{W}, \Xi\right)}{f\left(\mathcal{Z}_{(-t,d,n)}, \Delta_{(-t,d,n)} \mid \mathcal{I}, \mathcal{W}, \Xi\right)} \\
&= \frac{1}{b + N_{k,s}} \frac{S^{n_{k,s,v}+1}_{\tau_{k,s,v},a}}{S^{n_{k,s,v}}_{\tau_{k,s,v},a}} \frac{n_{k,s,v} + 1 - \tau_{k,s,v}}{n_{k,s,v} + 1} \times \frac{\alpha + \sum_{v^*} m_{k,d,v^*}}{\sum_{k^*} \left(\sum_{v^*} m_{k^*,d,v^*} + \alpha\right)}.
\end{aligned}
\tag{A.14}
$$

$$
\begin{aligned}
f_{k1} =& P\left(z_{t,d,n}=k, \delta_{t,d,n}=1 \mid \mathcal{Z}_{(-t,d,n)}, \Delta_{(-t,d,n)}, \mathcal{I}, \mathcal{W}, \Xi\right) \\
=& \frac{P\left(z_{t,d,n}=k, \delta_{t,d,n}=1, \mathcal{Z}_{(-t,d,n)}, \Delta_{(-t,d,n)} \mid \mathcal{I}, \mathcal{W}, \Xi\right)}{f\left(\mathcal{Z}_{(-t,d,n)}, \Delta_{(-t,d,n)} \mid \mathcal{I}, \mathcal{W}, \Xi\right)} \\
=& \frac{b+a\mathcal{T}_{k,s}}{b+N_{k,s}} \frac{S^{n_{k,s,v}+1}_{\tau_{k,s,v}+1,a}}{S^{n_{k,s,v}}_{\tau_{k,s,v},a}} \frac{\tau_{k,s,v}+1}{n_{k,s,v}+1} \times \frac{\alpha+\sum_{v^*} m_{k,t,d,v^*}}{\sum_{k^*}\left(\sum_{v^*} m_{k^*,t,d,v^*}+\alpha\right)} \frac{\sum_{s^*=1}^{S_k} \tau_{k,s^*,v}+\gamma}{\sum_{v^*}\left(\sum_{s^*=1}^{S_k} \tau_{k,s^*,v^*}+\gamma\right)}
\end{aligned}
\tag{A.15}
$$

Based on (A.14) and (A.15), we can compute the posterior values f_{k0} and f_{k1} for all $1 \le k \le K$. Then the obtained f_{k0} and f_{k1} are normalized to get the posterior probabilities, denoted by P_{k0} and P_{k1}. Note that we should have $\sum_{k=1}^{K}(P_{k0}+P_{k1})=1$. Then $(z_{t,d,n}, \delta_{t,d,n})$ can be sampled from a posterior multinomial distribution with parameters $\{P_{k0}, P_{k1}\}$ with $1 \le k \le K$.

A.4 Derivation details of $f(\mathcal{I}|\Delta, \mathcal{Z}, \mathcal{W}, \Xi)$

Based on the joint posterior distribution (A.9), we can derive the posterior distribution of $\mathcal{I}$ as follows

$$
f(\mathcal{I}|\mathcal{W}, \mathcal{Z}, \Delta, \Xi) = \prod_{k}^{K} \frac{\mathrm{Beta}(\lambda_0 + \sum_{t=1}^{T} i_{k,t}, \lambda_1 + T - \sum_{t=1} i_{k,t})}{\mathrm{Beta}(\lambda_0, \lambda_1)}.
$$

For easy illustration, denote $c_{k,1} = Q_k = \sum_{t=1}^{T} i_{k,t}$ to be the number of change points for topic k and $c_{k,0} = T - c_{k,1}$ to be the number of moments without change points. Note that $i_{k,t}$ is a dummy variable. To update $i_{k,t}$, we only need to compute the posterior values $f(i_{k,t}=1|\cdot)$ and $f(i_{k,t}=0|\cdot)$, and then normalize the values to get the posterior probabilities. Below, we derive the posterior values $f(i_{k,t}=1|\cdot)$ and $f(i_{k,t}=0|\cdot)$ as follows.

$$
\begin{aligned}
& f(i_{k,t}=0 \mid \mathcal{I}_{(-k,t)}, \mathcal{Z}, \Delta, \mathcal{I}, \mathcal{W}, \Xi) \\
\propto & \frac{\lambda_1 + c_{k,0} - 1}{\lambda_0 + c_{k,1} + \lambda_1 + c_{k,0} - 1} \mathrm{Beta}_V\left(\sum_{s=1}^{S_k} \tau_{k,s} + \gamma\right) \frac{(b \mid a)_{\mathcal{T}_{k,s_m}}}{(b)_{N_{k,s_m}}} \prod_{v=1}^{V} S^{n_{k,s_m,v}}_{\tau_{k,s_m,v},a}, \\
& f(i_{k,t}=1 \mid \mathcal{I}_{(-k,t)}, \mathcal{Z}, \Delta, \mathcal{I}, \mathcal{W}, \Xi) \\
\propto & \frac{\lambda_0 + c_{k,1} - 1}{\lambda_0 + c_{k,1} + \lambda_1 + c_{k,0} - 1} \mathrm{Beta}_V\left(\sum_{s=1}^{S_k} \tau_{k,s} + \gamma\right) \prod_{s \in \{s_l, s_r\}} \frac{(b \mid a)_{\mathcal{T}_{k,s}}}{(b)_{N_{k,s}}} \prod_{v=1}^{V} S^{n_{k,s,v}}_{\tau_{k,s,v},a}.
\end{aligned}
\tag{A.16}
$$

For $i_{k,t}=0$, it means no change point occurs. Then assume the t-th moment belongs to the segment s_m. For $i_{k,t}=1$, it means a change point occurs at the t-th moment. Then assume t and $t-1$ belong to the segments s_r and s_l, respectively. Note that $i_{k,t}$ represents whether a change point occurs and it determines the segmentation of moments. Then it is not easy to compute the posterior values $f(i_{k,t}=1|\cdot)$ and $f(i_{k,t}=0|\cdot)$ during the Gibbs sampling procedure, since we need to consider both cases that a change point occurs or not. To address this issue, we use the *merge and split algorithm* (Lan et al., 2013), which is introduced as follows.

(1) **The Merge Algorithm**

Let $i_{k,t}^{(b)}$ denote the value of $i_{k,t}$ in the bth iteration during the Gibbs sampling procedure. If $i_{k,t}^{(b)} = 1$, there is a change point in the bth iteration. Then we can easily compute the posterior probability of $i_{k,t}^{(b+1)} = 1$. However, when calculating the posterior value of $i_{k,t}^{(b+1)} = 0$, we should use the *merge algorithm*. Specifically, assume in the bth iteration, $i_{k,t}^{(b)}$ belongs to the segment s_r and $i_{k,t-1}^{(b)}$ belongs to the segment s_l. Then in the $(b+1)$th iteration, $i_{k,t}^{(b+1)} = 0$ implies the two segments are merged into one, denoted by s_m; see the top panel in Figure A.2 for illustration of this case.

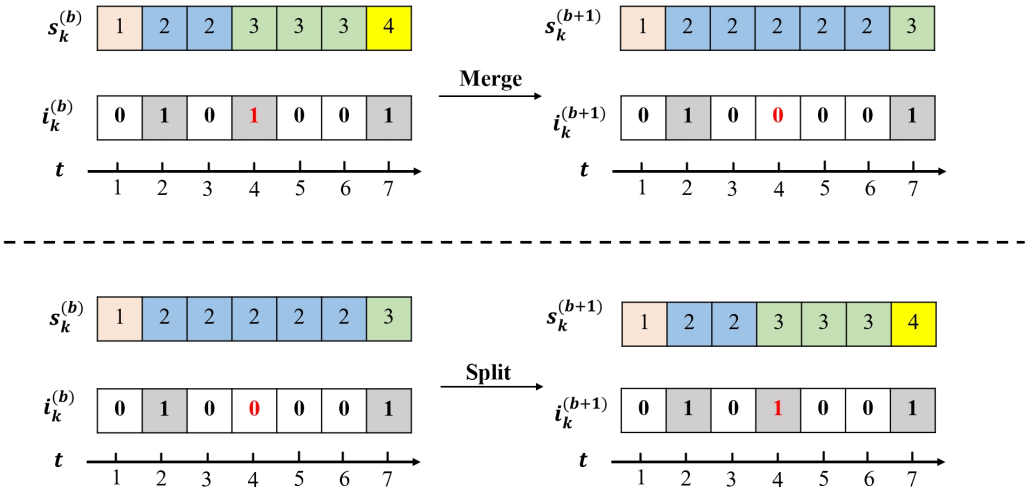


Figure A.2: The top panel illustrates the case when $i_{k,t}^{(b)} = 1$ but $i_{k,t}^{(b+1)} = 0$ at $t = 4$, and the bottom panel illustrates the case when $i_{k,t}^{(b)} = 0$ but $i_{k,t}^{(b+1)} = 1$ at $t = 4$.

Then to compute the posterior probability for $i_{k,t}^{(b+1)} = 0$, we need to calculate $\tau_{k,s_m,v}$ and $n_{k,s_m,v}$ for each word v in the vocabulary. As for $n_{k,s_m,v}$, we always have $n_{k,s_m,v} = n_{k,s_l,v} + n_{k,s_r,v}$. As for $\tau_{k,s_m,v}$, we basically have $\tau_{k,s_m,v} = \tau_{k,s_l,v} + \tau_{k,s_r,v}$ except for one case. That is, if in segments s_l and s_r, we have $n_{k,s,v} \geq 0$. Meanwhile, in s_l or s_r, we have $\tau_{k,s,v} = 1$. Then the calculation of $\tau_{k,s_m,v}$ can randomly select (a) or (b), where (a) $\tau_{k,s_m,v} = \tau_{k,s_l,v} + \tau_{k,s_r,v}$ and (b) $\tau_{k,s_m,v} = \tau_{k,s_l,v} + \tau_{k,s_r,v} - 1$.

(2) **The Split Algorithm**

Consider another case with $i_{k,t}^{(b)} = 0$, which means moments $t-1$ and t belong to the same segment. When $i_{k,t}^{(b+1)} = 1$, the original segment would split into two parts in moment t. Thus we need to use the split algorithm to calculate the posterior probability for $i_{k,t}^{(b+1)} = 1$. Specifically, assume in the bth iteration, $i_{k,t}^{(b)}$ belongs to the segment s_m. Further assume s_m splits into two parts s_l and s_r; see the bottom panel in Figure A.2 for an illustration of

this case. In this case, based on the topic assignment $z_{t,d,n}$ in s_l and s_r, we need to update $\delta_{t,d,n}$ in these segments. Below we give the process to sample $\delta_{t,d,n}$ with $z_{t,d,n} = k$.

Step 1. Sample $\delta_{t,d,n}$ from a Bernoulli distribution with parameter $\rho = \tau_{k,s_m,v}/n_{k,s_m,v}$.

Step 2. If $\delta_{t,d,n} = 1$, we have $\tau_{k,s_m,v} = \tau_{k,s_m,v} - 1$ and $n_{k,s_m,v} = n_{k,s_m,v} - 1$. If $\delta_{t,d,n} = 0$, we have $\tau_{k,s_m,v} = \tau_{k,s_m,v}$ and $n_{k,s_m,v} = n_{k,s_m,v} - 1$.

Step 3. Repeat the first two steps, until all the words have δ updated.

After resampling δ, there might exist some cases not satisfy the constraint $\tau_{k,s,v} = 0$ if and only if $n_{k,s,v} = 0$. If $\tau_{k,s,v} = 0$ and $n_{k,s,v} > 0$ in segments s_l or s_r, we set $\tau_{k,s,v} = 1$ and re-sample the corresponding δ for all the words in the vocabulary. Then the posterior probability of $i_{k,t} = 1$ is the same as (A.16), but the corresponding value of τ needs to be updated by recounting δ.

A.5 The proof of Theorem 1

The proof of Theorem 1 basically follows Theorem 3 in McGoff et al. (2022). We begin by applying the main result of Theorem 3, considering the posterior distribution of the data generated by the model $P_{\mathcal{I}^*,\Phi^*}$, and particularly examining how the posterior distribution concentrates in the parameter space around the true change point set $\mathcal{I}^*$.

Firstly, we assume that the data generation process $P_{\mathcal{I}^*,\Phi^*}$ is correctly specified, meaning that the model accurately reflects the true underlying process that generates the data. Based on the prior density given in (A.1) in Appendix A.2, the corresponding prior distribution $\mathbf{\Pi}$ of $\mathcal{I}$ is fully supported. Specifically, for any set $A \subset \Theta_{\mathcal{I}}$ within the parameter space $\Theta_{\mathcal{I}} = \{0,1\}^{KT}$, we have $\mathbf{\Pi}(A) > 0$. The full support of the prior guarantees that all potential parameter values are assigned non-zero probability, thereby preventing prior bias. In addition, assume the updating function f, i.e., the posterior distribution of $\mathcal{I}$ provided in Appendix A.4, is regular and satisfies the following conditions.

(i) There exists a measurable function $f^* : \boldsymbol{\mathcal{W}} \to \mathbb{R}$ such that for all $\mathcal{W} \in \boldsymbol{\mathcal{W}}$, we have $\sup_{\mathcal{I}} |f(\mathcal{I}, \mathcal{W})| \leq f^*(\mathcal{W})$.

(ii) For each $\delta > 0$, there exists a measurable function $\rho_\delta : \boldsymbol{\mathcal{W}} \to (0, \infty)$ such that for each $\mathcal{W} \in \boldsymbol{\mathcal{W}}$, and $\lim_{\delta \to 0^+} \int \rho_\delta d\nu = 0$, we have

$$\sup\{|f(\mathcal{I}, \mathcal{W}) - f(\mathcal{I}', \mathcal{W})| : d(\mathcal{I}, \mathcal{I}') \leq \delta\} \leq \rho_\delta(\mathcal{W}),$$

where $d(\cdot, \cdot)$ represents the Hamming distance. This condition ensures that the variation of the updating function is smooth and tends to zero as δ decreases.

Then according to Theorem 2 in McGoff et al. (2022), it can be guaranteed that the Gibbs posterior distribution of $\mathcal{I}$ concentrates around a set $\Theta_{\mathcal{I},\min}$, which is characterized as the solution set of a variational problem. This variational characterization is crucial in demonstrating that $\Theta_{\mathcal{I},\min}$ corresponds to the identifiability class $[\mathcal{I}^*]$. Here $[\mathcal{I}^*]$ is the set of parameters $\mathcal{I}'$ satisfying $\mu_{\mathcal{I}'} = \mu_{\mathcal{I}^*}$, where μ is the unique Gibbs measure associated with f. In this context, we rely on specific arguments related to the problem at hand, particularly a foundational result from Bowen's thermodynamic formalism (Bowen, 1975), which asserts the uniqueness of equilibrium states for Hölder continuous potentials on a mixing Subshift of Finite Type (SFT). This result guarantees that there exists a unique equilibrium distribution corresponding to $\mathcal{I}^*$. As a result, the posterior distribution will concentrate around $\mathcal{I}^*$ as the number of documents D and the number of words N increase.

A.6 The proof of Theorem 2

Given the change points, document d at time t uniquely corresponds to a word probability vector $\boldsymbol{\eta}_{t,d} = \sum_{k=1}^{K} \theta_{t,d,k}\phi_{k,s_{k,t}} \in \Omega^{V-1}$, where $s_{k,t} \in \{1, 2, \cdots, S_k\}$ denotes the index of segment of topic k at time t and $\sum_{k=1}^{K} \theta_{t,d,k} = 1$. To simplify the analysis, we assume all time points have the same number of documents D and all documents have the same number of words N. Then the words of document d at time t, i.e. $\mathcal{W}_{t,d} = \{w_{t,d,n}\}_{n=1}^{N}$, are i.i.d. samples from a multinomial distribution parameterized by this probability vector $\boldsymbol{\eta}_{t,d}$. Given the topic-word parameters Φ, this induces a probability distribution $P_{\boldsymbol{\eta}_{t,d}|G}$, whose support is the convex set G. The joint distribution of the full data set $\mathcal{W}$, denoted by $P_{\mathcal{W}|G}$, is the product distribution of all single document distributions: $P_{\mathcal{W}|G} = \prod_{t=1}^{T}\prod_{d=1}^{D} P_{\mathcal{W}_{t,d}|G}(\mathcal{W}_{t,d})$, where $p_{\mathcal{W}_{t,d}|G}(\mathcal{W}_{t,d}) = \int_G \prod_{n=1}^{N}\prod_{l=1}^{V} \eta_{t,d,l}^{\mathbb{I}(w_{t,d,n}=l)} dP_{\boldsymbol{\eta}_{t,d}|G}(\boldsymbol{\eta}_{t,d})$.

Assume the observed words $\mathcal{W}$ are generated according to the TOPIC-PYP process given the fixed change points and the true parameters $\Phi^* = (\phi^*_{1,1}, \cdots, \phi^*_{1,S_1}, \cdots, \phi^*_{K,1}, \cdots, \phi^*_{K,S_K})^\top$. Let $G^* = \text{conv}(\Phi^*)$ be the true topic polytope. Under the TOPIC-PYP process, Φ is endowed with a prior distribution $\boldsymbol{\Psi}$ based on the Pitman-Yor process. The main question here is the contraction behavior of the posterior distribution $\boldsymbol{\Psi}(G|\mathcal{W})$ under a fixed T, as the number of documents D and the number of words N go to infinity. To address this problem, we first provide some needed assumptions.

Assumption 1 *$\boldsymbol{\Psi}$ is a prior distribution on Φ such that the following assumptions hold for the relevant parameters that reside in the support of $\boldsymbol{\Psi}$.*

(A1) Geometric properties (A1) and (A2) listed in Section 3 in Nguyen (2015) are satisfied uniformly for all G in the support of $\boldsymbol{\Psi}$.

(A2) Each topic vector $\phi_{k,s}(1 \le K, 1 \le s \le S_k)$ is bounded away from the boundary of Ω^{V-1}.

(A3) For any small ϵ, $\boldsymbol{\Psi}(\|\phi_{k,s} - \phi^*_{k,s}\| \le \epsilon) \ge c\epsilon^{V\sum_{k=1}^{K} S_k}$, $\forall\ 1 \le k \le K, 1 \le s \le S_k$.

Assumptions (A1)-(A3) are mild assumptions observed in practice. It is noteworthy that, conditional on the change points, the generation of documents and words can be framed within a general hierarchical model: $G|\mathcal{I} \sim \boldsymbol{\Psi}$, $\boldsymbol{\eta}_{t,d}|G \sim P_{\boldsymbol{\eta}_{t,d}|G}$, $\mathcal{W}_{t,d}|\boldsymbol{\eta}_{t,d} \sim P_{\mathcal{W}_{t,d}|\boldsymbol{\eta}_{t,d}}$ for $t \in \{1, \cdots, T\}$ and $d \in \{1, \cdots, D\}$. Note that the model satisfies the conditions concerned with entropy condition for certain sets in the support of the prior, the "thickness" of the prior as measured by the Kullback–Leibler distance, and the conditions related to the Hellinger information function. Nguyen (2015) presented an abstract posterior contraction theorem for hierarchical models of this nature. Building upon Theorem 4 of Nguyen (2015), it suffices to verify the conditions of this theorem for its applicability to our specific setting.

Denote $\mathcal{G}$ to be the support of the prior distribution $\boldsymbol{\Psi}$ of G. Then define the "minimum-matching" Euclidean distance between two topic polytope as $d_{\mathcal{M}}(G, G') = \max_{\phi \in \text{extr}\, G} \min_{\phi' \in \text{extr}\, G'} \|\phi - \phi'\| \wedge \max_{\phi' \in \text{extr}\, G'} \min_{\phi \in \text{extr}\, G} \|\phi' - \phi\|$. Define the Hausdorff metric as $d_{\mathcal{H}}(G, G') = \min\{\varepsilon \ge 0 | G \subset G'_\varepsilon; G' \subset G_\varepsilon\} = \max_{\phi \in G} d(\phi, G') \wedge \max_{\phi' \in G'} d(\phi', G)$, where $G_\varepsilon = G + B_V(\mathbf{0}, \varepsilon) = \{\phi + e | \phi \in G, e \in \mathbb{R}^V, \|e\| \le 1\}$ and $d(\phi, G') = \inf\{\|\phi - \phi'\|, \phi' \in G'\}$. Let $N(\varepsilon, \mathcal{G}, d_{\mathcal{H}})$ and $D(\varepsilon, \mathcal{G}, d_{\mathcal{H}})$ denote the covering number and packing number of $\mathcal{G}$ in Hausdorff metric

$d_{\mathcal{H}}$, respectively. Define the Hausdorff ball as $B_{d_{\mathcal{H}}}(G_1, \delta) = \{G \in \Omega^{V-1} : d_{\mathcal{H}}(G_1, G) \leq \delta\}$. A useful quantity for proving posterior concentration theorems is the Hellinger information of Hausdorff metric for a given set, which is a real-valued function defined as $\Psi_{\mathcal{G}}(\delta) = \inf_{G \in \mathcal{G}; d_{\mathcal{H}}(G^*, G) \geq \delta/2} h^2(p_{\mathcal{W}_{t,d}|G^*}, p_{\mathcal{W}_{t,d}|G})$, where $h^2(p, q) = (1/2)\int(\sqrt{p} - \sqrt{q})^2$ represents the Hellinger distance between two densities p and q. Define $\Phi_{\mathcal{G}} : \mathbb{R}_+ \to \mathbb{R}$ to be an arbitrary non-negative valued function on the positive reals such that for any $\delta > 0$, $\sup_{G, G' \in \mathcal{G}; d_{\mathcal{H}}(G, G') \leq \Phi_{\mathcal{G}}(\delta)} h^2(p_{\mathcal{W}_{t,d}|G}, p_{\mathcal{W}_{t,d}|G'}) \leq \Psi_{\mathcal{G}}(\delta)/4$. Since $\theta_{t,d} \sim \mathrm{Dir}(\alpha)$, meaning that $\{\theta_{t,d,1}, \cdots, \theta_{t,d,K}\}$ are exchangeable, then we have $\Phi_{\mathcal{G}}(\delta) = \frac{c_0}{4TNC_0}\Psi_{\mathcal{G}}(\delta)$. Define the neighborhood of the prior support around G^* in terms of Kullback–Leibler distance of the marginal densities $p_{\mathcal{W}_{t,d}|G}$: $B_K(G^*, \delta) = \{G | K(p_{\mathcal{W}_{t,d}|G^*}, p_{\mathcal{W}_{t,d}|G}) \leq \delta^2; K_2(p_{\mathcal{W}_{t,d}|G^*}, p_{\mathcal{W}_{t,d}|G}) \leq \delta^2\}$, where $K(p, q)$ denotes the Kullback–Leibler divergence and $K_2 = \int p[\log(p/q)]^2$.

Starting with the entropy condition, we note that

$$\log D\left\{\varepsilon/2, \mathcal{G} \cap B_{\mathcal{H}}\left(G^*, 2\varepsilon\right), d_{\mathcal{H}}\right\} \leq \log N\left\{\varepsilon/4, \mathcal{G} \cap B_{\mathcal{H}}\left(G^*, 2\varepsilon\right), d_{\mathcal{H}}\right\} = \mathrm{O}(1).$$

Since $\Psi_{\mathcal{G}}(\varepsilon) \geq \left[c_1(\varepsilon/2)^p - 6(V+1)\mathrm{e}^{-N\varepsilon^2/32(V+1)}\right]^2$, we obtain that $\Psi_{\mathcal{G}}(\varepsilon) \geq c\varepsilon^{2p}$ as long as $c_1(\varepsilon/2)^p \geq 12(V+1)\exp[-N\varepsilon^2/32(V+1)]$. Set $\varepsilon_{T,D,N} = (\log(TD)/(TD))^{1/2} + (\log N/(TD))^{1/2} + (\log N/N)^{1/2}$. This is satisfied because ε is bounded from below by a large multiple of $\varepsilon_{T,D,N} > (\log N/N)^{1/2}$. Recall that $\Phi_{\mathcal{G}}(\delta) = \frac{c_0}{4TNC_0}\Psi_{\mathcal{G}}(\delta)$. Then we have

$$\begin{aligned}
&\log D\left\{c_0\Psi_{\mathcal{G}}(\varepsilon)/\left(4TNC_0\right), \mathcal{G} \cap B_{\mathcal{H}}\left(G_1, \varepsilon/2\right), d_{\mathcal{H}}\right\} \\
&\quad \leq \log N\left\{c_0 c\varepsilon^{2p}/\left(4TNC_0\right), \mathcal{G} \cap B_{\mathcal{H}}\left(G_1, \varepsilon/2\right), d_{\mathcal{H}}\right\} \\
&\quad \lesssim \log\left(N^{V\sum_{k=1}^K S_k} \times \varepsilon^{-(2p-1)V\sum_{k=1}^K S_k}\right) \leq TD\varepsilon^2.
\end{aligned}$$

Thus the entropy condition is established. To verify the Hellinger information conditions, we note that for some constant $c > 0$,

$$\begin{aligned}
&\exp\left(2TD\varepsilon_{T,D,N}^2\right) \sum_{j \geq M_m} \exp\left[-D\Psi_{\mathcal{G}}\left(j\varepsilon_{T,D,N}\right)/8\right] \\
&\quad \leq \exp\left(2TD\varepsilon_{T,D,N}^2\right) \sum_{j \geq M_m} \exp\left[-cD\left(j\varepsilon_{T,D,N}\right)^{2p}/8\right] \\
&\quad \lesssim \exp\left(2TD\varepsilon_{T,D,N}^2\right) \exp\left[-cD\left(M_m\varepsilon_{T,D,N}\right)^{2p}/8\right],
\end{aligned}$$

where the right side of the above vanishes if $(M_m\varepsilon_{T,D,N})^p$ is a sufficiently large multiple of $\varepsilon_{T,D,N}$. This holds if we choose $M_m = M\varepsilon_{T,D,N}^{-(p-1)/p}$ for a large constant M.

It remains to verify the "thickness" property of the prior distribution. It is easy to abtain that as long as $N \gtrsim \log(1/\varepsilon_{T,D,N})$, then

$$\begin{aligned}
\log \mathbf{\Pi}\left(G \in B_K\left(G_0, \varepsilon_{T,D,N}\right)\right) &\geq c\left(c_0\right) \log\left(\varepsilon_{T,D,N}^2/N^3\right)^{V\sum_{k=1}^K S_k} \\
&= c\left(c_0\right) V\left(\sum_{k=1}^K S_k\right)\left(2\log\varepsilon_{T,D,N} - 3\log N\right).
\end{aligned}$$

Based on the above analysis, we can apply Theorem 4 of Nguyen (2015) to obtain a posterior contraction rate $M_m\varepsilon_{T,D,N} \asymp \varepsilon_{T,D,N}^{1/p}$, which leads to Theorem 2 in this work.

Appendix B. Additional Empirical Results

B.1 Checking the convergence of TOPIC-PYP

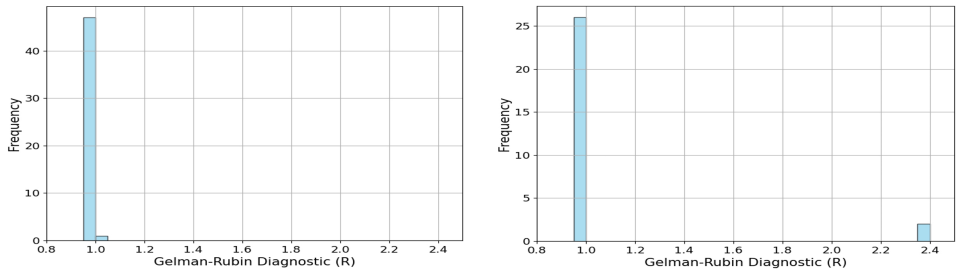
We apply the Gibbs sampling method to estimate TOPIC-PYP, which is a well-established Markov chain Monte Carlo (MCMC) technique. We have shown in Section 3.3 that TOPIC-PYP satisfies the necessary conditions for posterior convergence required by Gibbs sampling. In this section, we empirically demonstrate the posterior convergence of TOPIC-PYP in the simulation experiments on syngenetic datasets. To this end, we evaluate the posterior convergence using the Gelman-Rubin diagnostic, which is also known as the $\widehat{R}$ statistic (Gelman and Rubin, 1992; Gelman et al., 2013). It is a widely used convergence assessment method for MCMC algorithms. The key idea of Gelman-Rubin diagnostic is to examine multiple independent chains initialized from diverse starting points and to determine whether all chains have converged to the same target posterior distribution. If different chains have not converged, their sample means and variances will differ substantially. Once they approach the target posterior distribution, they will exhibit similar behavior, with their means and variances stabilizing approximately at the same values. Based on this idea, the Gelman-Rubin diagnostic $\widehat{R}$ statistic is computed by comparing the within-chain variance to the between-chain variance. A value of $\widehat{R}$ close to 1.0 suggests that the chains have mixed well and reached the target distribution, signaling convergence.

To compute the $\widehat{R}$ statistic, we ran 5 chains initialized from diverse starting points. Each chain is run for 50 iterations, which are sufficient to achieve convergence. Then we collect the last 10 posterior samples from each chain to compute the $\widehat{R}$ statistic. Not that TOPIC-PYP involves a lot of parameters to estimate. Specifically, $\Delta = \{\delta_{t,d,n}\}$, $\mathcal{Z} = \{z_{t,d,n}\}$, where $1 \leq t \leq T, 1 \leq d \leq D_t, 1 \leq n \leq N_{t,d}$, and $\mathcal{I} = \{i_{k,t}\}$, where $1 \leq k \leq K, 1 \leq t \leq T$. Thus we separately compute the $\widehat{R}$ statistic for each parameter. Then we compute the averaged $\widehat{R}$ statistic for the groups Δ, $\mathcal{Z}$, and $\mathcal{I}$ separately. The corresponding results are shown in Table B.1. It is obvious that, the averaged $\widehat{R}$ statistics are all close to 1 in the four scenarios. Particularly, even for the complicated scenarios (i.e., Scenario 3 and Scenario 4), the averaged $\widehat{R}$ statistics behave very well.

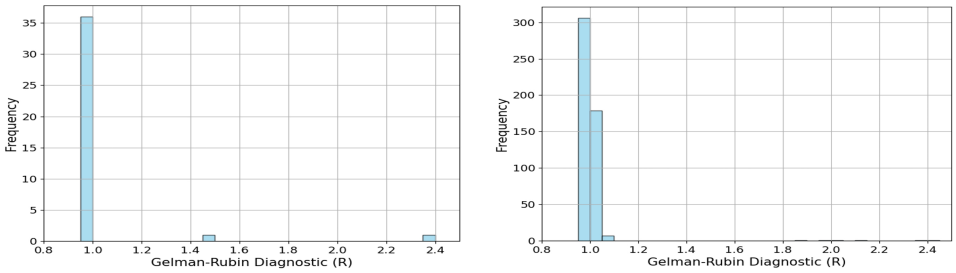
Table B.1: The averaged value of $\widehat{R}$ statistic in different scenarios.

Group	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Δ	1.010	1.014	1.010	1.018
$\mathcal{Z}$	1.171	1.155	1.140	1.136
$\mathcal{I}$	1.158	1.017	1.097	1.054

To provide a more comprehensive understanding of the convergence diagnostics, we further present the distribution plots of the $\widehat{R}$ statistics. To save space, we use $\mathcal{I}$ as an example, as it represents the locations of change points and serves as one of the key parameter groups in TOPIC-PYP. Figure B.1 presents the histograms of the $\widehat{R}$ statistics for all parameters in the group $\mathcal{I}$ across the four scenarios. It is obvious that, nearly all parameters in $\mathcal{I}$ have $\widehat{R}$ statistics close to 1. For the parameter groups Δ and $\mathcal{Z}$, we observe similar results. All these findings empirically demonstrate that TOPIC-PYP has achieved satisfactory convergence.



(a) Scenario 1 (b) Scenario 2



(c) Scenario 3 (d) Scenario 4

Figure B.1: The histograms of $\widehat{R}$ statistics in the parameter group $\mathcal{I}$ under four scenarios.

B.2 The results of precision and recall under Scenarios 1-3

Table B.2: The comparison results of precision and recall under Scenario 1.

		Precision			Recall		
Method		$h=0$	$h=1$	$h=2$	$h=0$	$h=1$	$h=2$
TOPIC-PYP		0.950	0.950	0.950	0.950	0.950	0.950
Topic-CD		0.925	0.925	0.925	0.950	0.950	0.950
DTM	_CS_DP	0.300	0.300	1.000	0.300	0.300	1.000
	_CS_BS	0.300	0.300	1.000	0.300	0.300	1.000
	_CS_T	0.500	1.000	1.000	0.500	1.000	1.000
	_JS_DP	0.300	0.300	0.950	0.300	0.300	1.000
	_JS_BS	0.300	0.300	0.950	0.300	0.300	1.000
	_JS_T	0.300	0.300	0.950	0.300	0.300	1.000
D-ETM	_CS_DP	0.500	0.500	1.000	0.500	0.500	1.000
	_CS_BS	0.300	0.300	1.000	0.300	0.300	1.000
	_CS_T	0.500	1.000	1.000	0.500	1.000	1.000
	_JS_DP	0.500	1.000	1.000	0.500	1.000	1.000
	_JS_BS	0.500	1.000	1.000	0.500	1.000	1.000
	_JS_T	0.500	1.000	1.000	0.500	1.000	1.000
Rolling LDA	_CS_DP	0.500	0.500	0.950	0.500	0.500	1.000
	_CS_BS	0.300	0.300	1.000	0.300	0.300	1.000
	_CS_T	0.500	1.000	1.000	0.500	1.000	1.000
	_JS_DP	0.300	0.300	1.000	0.300	0.300	1.000
	_JS_BS	0.500	1.000	1.000	0.500	1.000	1.000
	_JS_T	0.300	0.450	0.800	0.300	1.000	1.000

Table B.3: The comparison results of precision and recall under Scenario 2.

		Precision			Recall		
Method		$h=0$	$h=1$	$h=2$	$h=0$	$h=1$	$h=2$
TOPIC-PYP		0.933	0.933	0.933	1.000	1.000	1.000
Topic-CD		0.833	0.933	0.933	0.900	1.000	1.000
DTM	_CS_DP	0.500	1.000	1.000	0.500	1.000	1.000
	_CS_BS	0.500	1.000	1.000	0.500	1.000	1.000
	_CS_T	1.000	1.000	1.000	1.000	1.000	1.000
	_JS_DP	0.500	1.000	1.000	0.500	1.000	1.000
	_JS_BS	0.500	0.500	1.000	0.500	0.500	1.000
	_JS_T	1.000	1.000	1.000	1.000	1.000	1.000
D-ETM	_CS_DP	0.400	0.900	1.000	0.400	0.900	1.000
	_CS_BS	0.200	0.900	1.000	0.200	0.900	1.000
	_CS_T	0.600	1.000	1.000	0.600	1.000	1.000
	_JS_DP	0.300	0.800	1.000	0.300	0.800	1.000
	_JS_BS	0.300	1.000	1.000	0.300	1.000	1.000
	_JS_T	0.500	1.000	1.000	0.500	1.000	1.000
Rolling LDA	_CS_DP	0.600	1.000	1.000	0.600	1.000	1.000
	_CS_BS	0.300	1.000	1.000	0.300	1.000	1.000
	_CS_T	0.700	1.000	1.000	0.700	1.000	1.000
	_JS_DP	0.500	0.600	1.000	0.500	0.600	1.000
	_JS_BS	0.400	0.600	1.000	0.400	0.600	1.000
	_JS_T	0.300	0.900	1.000	0.300	0.900	1.000

Table B.4: The comparison results of precision and recall under Scenario 3.

		Precision			Recall		
Method		$h=0$	$h=1$	$h=2$	$h=0$	$h=1$	$h=2$
TOPIC-PYP		0.775	1.000	1.000	0.945	1.000	1.000
Topic-CD		0.683	0.975	0.975	0.790	1.000	1.000
DTM	_CS_DP	0.500	0.950	1.000	0.500	0.950	1.000
	_CS_BS	0.450	0.875	0.950	0.450	0.875	0.950
	_CS_T	0.575	1.000	1.000	0.575	1.000	1.000
	_JS_DP	0.500	0.975	1.000	0.500	0.950	0.950
	_JS_BS	0.450	0.875	1.000	0.450	0.875	0.875
	_JS_T	0.575	1.000	1.000	0.575	1.000	1.000
D-ETM	_CS_DP	0.400	0.800	1.000	0.400	0.800	0.975
	_CS_BS	0.200	1.000	1.000	0.200	1.000	1.000
	_CS_T	0.600	1.000	1.000	0.600	1.000	1.000
	_JS_DP	0.400	1.000	1.000	0.400	1.000	1.000
	_JS_BS	0.400	1.000	1.000	0.400	1.000	1.000
	_JS_T	0.600	1.000	1.000	0.600	1.000	1.000
Rolling LDA	_CS_DP	0.500	0.950	0.950	0.500	0.950	0.950
	_CS_BS	0.450	0.945	0.975	0.450	0.945	0.975
	_CS_T	0.750	0.975	1.000	0.675	0.975	0.975
	_JS_DP	0.500	0.950	1.000	0.500	0.950	1.000
	_JS_BS	0.500	1.000	1.000	0.500	0.950	0.950
	_JS_T	0.750	1.000	1.000	0.750	1.000	1.000

B.3 Comparison of computational efficiency

We compare the computational efficiency of TOPIC-PYP with other methods in this section. To this end, we calculate the running time for different methods. All methods are implemented on a server with 8 CPUs and 16 GB memory. For the competing methods, the experiments are conducted using publicly available codes or programs provided by the authors, with experimental settings adhering to their default configurations. Specifically, the codes of Topic-CD provided by the authors are available at `https://github.com/ffair/Topic-CD`. The DTM method is implemented using the Python library *gensim*. The D-ETM method is implemented by the codes accompanying the original paper (Dieng et al., 2019), which are available at `https://github.com/adjidieng/DETM`. The Rolling LDA method is implemented by the R package *rollinglda*; see `https://github.com/JonasRieger/rollinglda` for details. Last, the codes for the ANTM method are available at `https://github.com/hamedR96/antm`.

The detailed runtime results (in seconds) are shown in Table B.5, from which we can draw the following conclusions. First, compared to two-stage methods, both unified methods (including TOPIC-PYP and Topic-CD) have significantly longer running time and thus behave less computationally efficient. This is because the unified methods combine topic modeling and change point detection, making the model structures more complex. Additionally, some existing two-stage models (e.g., DTM and Rolling LDA) can be implemented via well-developed packages and thus run very fast. Second, in comparison to Topic-CD, the running time of TOPIC-PYP is generally comparable, although slightly slower. This is because TOPIC-PYP allows for change point detection for each individual topic, whereas Topic-CD can only detect change points that are shared across all topics. That is the price TOPIC-PYP should pay in terms of computational efficiency. Last, the runtime of TOPIC-PYP increases as the model becomes more complex (i.e., with longer time span T or a larger number of change points Q_k). Basically, he runtime of TOPIC-PYP empirically exhibits a linear growth relationship with both T and Q_k. Given the computational limitations of TOPIC-PYP, enhancing its computational efficiency should be an important area of future research.

Table B.5: The comparison results of running time (in seconds) for different methods in four scenarios.

Method		Scenario 1	Scenario 2	Scenario 3	Scenario 4
Unified	TOPIC-PYP	4692	4692	48715	217357
	Topic-CD	4317	4317	43624	212884
Two-Stage	DTM	1093	1077	9726	22477
	D-ETM	486	353	6247	44333
	Rolling LDA	15	15	42	108
	ANTM	2112	2112	17226	51099

B.4 Influence of other hyperparameters

We investigate the influence of other hyperparameters in TOPIC-PYP (i.e., $\lambda = \{\lambda_0, \lambda_1\}$, α, and γ). We consider Scenario 2 for illustration, which assumes the presence of two

change points and $T = 10$. We generally follow the experimental settings in Section 4.1. For the baseline setup, let $a = 0.5$, $b = 5$, $\lambda = \{2, 5\}$, $\alpha = 0.1$, and $\gamma = 0.1$. Then we vary the value of each hyperparameter while keeping the others fixed. Table B.6 summarizes the hyperparameter value and the corresponding results of TOPIC-PYP evaluated by precision, recall, and accuracy. We find that, varying the values of λ, α, and γ has little effect on the change point detection performance of the TOPIC-PYP model. This result implies that TOPIC-PYP is quite robust to these hyperparameters.

Table B.6: The precision, recall, and accuracy of TOPIC-PYP by changing the value of hyperparameters.

Hyperparameter		Precision	Recall	Accuracy
Baseline		0.83	1.00	0.89
λ	$\{2, 4\}$	0.83	1.00	0.89
	$\{2, 6\}$	0.83	1.00	0.89
α	0.05	0.81	1.00	0.89
	0.2	0.83	1.00	0.89
γ	0.05	0.85	1.00	0.91
	0.2	0.83	0.97	0.87

B.5 Topic meanings in the Journal dataset

For the Journal dataset, we extract $K = 20$ topics, among which Topic 1 and Topic 15 share one change point at year $t = 2016$. The other eighteen topics do not have changing topic meanings. Table B.7 summarizes the meanings of these eighteen topics using the representative words. As shown, we find all eighteen topics have an explicit meaning. For example, Topics 4, 6, and 10 are related to the stochastic process. Among them, Topic 4 focuses on the Poisson process and jump process, Topic 6 focuses on the Brownian process and Levy process, while Topic 10 focuses on the theory of stochastic process. Other topics also have their thematic meanings, such as linear regression represented by Topic 11, and hypothesis test represented by Topic 17.

Table B.7: The meanings of eighteen topics without change points in the Journal dataset.

Number	Meaning	Representative words
2	Quantile Regression	regression, functional, quantile, functions, nonparametric, principal, components, component
3	Mixed Effect	effects, mixture, multivariate, mixed, longitudinal, latent, generalized, parameters
4	Stochastic Process	process, processes, point, Poisson, stochastic, population, jump, state
5	Causal Inference, Graphical Analysis	treatment, effects, covariates, effect, inference, network, graphs, variables, cluster
6	Brownian Process	stochastic, equations, differential, Brownian, solutions, processes, lévy, process
7	Survival Analysis	error, survival, measurement, assumptions, sensitivity, effects, estimates, hazard
8	System Analysis	system, state, sequential, optimal, PCA, framework, cost, structure
9	Optimal Design	designs, optimal, design, observations, change, volatility, threshold, influence, minimum
10	Stochastic Process Theorem	process, limit, large, theorem, asymptotic, prove, stationary, convergence
11	Linear Regression	regression, linear, covariance, matrix, highdimensional, correlation, sample, variables
12	Classification Method	class, functions, measure, general, representation, particular, Dirichlet, measures
13	Probability	density, matrix, covariance, Gaussian, functions, matrices, kernel, convergence
14	Feature Selection	selection, regression, adaptive, dimension, sparse, reduction, lasso, loss
16	Maximum Likelihood Estimator	estimator, likelihood, selection, asymptotic, regression, bootstrap, empirical, consistency
17	Hypothesis Eest	test, tests, testing, statistics, null, power, hypothesis, statistic
18	Parameter Estimator Inference	algorithm, estimator, likelihood, variance, robust, maximum, confidence, error
19	Time Series Analysis	series, statistical, performance, structure, predictive, applied, provide, multivariate
20	Online Learning	spatial, online, classification, statistical, multiple, modeling, clustering, Bayesian

References

Amr Ahmed and Eric P. Xing. 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, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA*, pages 219–230, 2008.

Amr Ahmed and Eric P. Xing. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, pages 20–29, 2010.

Jushan Bai. Estimation of a change point in multiple regression models. *Review of Economics & Statistics*, 79(4):551–563, 1997.

D. Blei and J. D. Mcauliffe. Supervised topic models. *Advances in Neural Information Processing Systems*, 3:327–332, 2008.

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

David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006a.

David M. Blei and John D. Lafferty. Dynamic topic models. pages 113–120, New York, NY, USA, 2006b. Association for Computing Machinery. ISBN 1595933832.

Rufus Bowen. Equilibrium states and the ergodic theory of anosov diffeomorphisms. *Lecture notes in mathematics*, 470:11–25, 1975.

Daniel Bruggermann, Yannik Hermey, Carsten Orth, Darius Schneider, Stefan Selzer, and Gerasimos Spanakis. Storyline detection and tracking using dynamic latent Dirichlet allocation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016), EMNLP 2016*, pages 9–19, 11 2016.

Wray Buntine and Marcus Hutter. A bayesian view of the poisson-dirichlet process, 2010.

Wray L Buntine and Swapnil Mishra. Experiments with non-parametric topic models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 881–890, 2014.

Changyou Chen, Lan Du, and Wray Buntine. Sampling table configurations for the hierarchical poisson-dirichlet process. In *ECML PKDD'11: Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases*, pages 296–311, 2011.

Siddhartha Chib. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241, 1998.

Kostadin Cvejoski, Ramsés J Sánchez, and César Ojeda. Neural dynamic focused topic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12719–12727, 2023.

A. B. Dieng, Fjr Ruiz, and D. M. Blei. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545, 2019*, 2019.

Joseph L Doob. Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pages 23–27, 1949.

Thomas S. Ferguson. Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics*, 2(4):615 – 629, 1974. doi: 10.1214/aos/1176342752. URL `https://doi.org/10.1214/aos/1176342752`.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC press, Boca Raton, FL, 3rd edition, 2013.

Zoubin Ghahramani and Thomas Griffiths. Infinite latent feature models and the indian buffet process. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/2ef35a8b78b572a47f56846acbeef5d3-Paper.pdf.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1):5228–5235, 2004.

Jianjie Guo, Lin Guo, Wenchao Xu, and Haibin Zhang. Hidden markov model with pitman-yor priors for probabilistic topic model. *Communications in Statistics-Theory and Methods*, pages 1–15, 2024.

Pankaj Gupta, Yatin Chaudhary, Florian Buettner, and Hinrich Schuetze. textTOvec: Deep Contextualized Neural Autoregressive Topic Models of Language with Distributed Compositional Prior. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkgoyn09KQ.

Florian Holz and Sven Teresniak. Towards automatic detection and tracking of topic change. In *The 11th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 327–339, 2010.

Olga Kellert and Md Mahmud Uz Zaman. Using neural topic models to track context shifts of words: a case study of COVID-related terms before and after the lockdown in April 2020. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 131–139, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.14. URL https://aclanthology.org/2022.lchange-1.14.

Stanley I. M. Ko, Terence T. L. Chong, and Pulak Ghosh. Dirichlet process hidden Markov multiple change-point model. *Bayesian Analysis*, 10(2):275–296, 2015.

D. Lan, W. Buntine, and M. Johnson. Topic segmentation with a structured topic model. In *Proceedings of NAACL-HLT*, pages 190–200, 2013.

J Lau, N Collier, and T Baldwin. On-line trend analysis with topic models: Twitter trends detection topic model online. In *Proceedings of COLING*, pages 1519–1534, 2012.

Michael Lavine. Some Aspects of Polya Tree Distributions for Statistical Modelling. *The Annals of Statistics*, 20(3):1222 – 1235, 1992. doi: 10.1214/aos/1176348767. URL https://doi.org/10.1214/aos/1176348767.

Kar Wai Lim, Wray Buntine, Changyou Chen, and Lan Du. Nonparametric bayesian topic modelling with the hierarchical pitman–yor processes. *International Journal of Approximate Reasoning*, 78:172–191, 2016.

Robert Lindsey, William Headden, and Michael Stipicevic. A phrase-discovering topic model using hierarchical pitman-yor processes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 214–222, 2012.

J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

Xiaoling Lu, Yuxuan Guo, Jiayi Chen, and Feifei Wang. Topic change point detection using a mixed bayesian model. *Data Min. Knowl. Discov.*, 36(1):146–173, 2022.

Kevin McGoff, Sayan Mukherjee, and Andrew B Nobel. Gibbs posterior convergence and the thermodynamic formalism. *The Annals of Applied Probability*, 32(1):461–496, 2022.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/miao16.html`.

Nozomu Miyamoto, Masaru Isonuma, Sho Takase, Junichiro Mori, and Ichiro Sakata. Dynamic structured neural topic model with self-attention mechanism. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5916–5930, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.366. URL `https://aclanthology.org/2023.findings-acl.366`.

Ramesh M. Nallapati, Susan Ditmore, John D. Lafferty, and Kin Ung. Multiscale topic tomography. In *ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*, pages 520–529, 2007.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108, 2010.

Xuanlong Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, pages 618–646, 2015.

Lev Pevzner and Marti Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28:1–19, 2002.

Jim Pitman and Marc Yor. The two parameter poisson-dirichlet distribution derived from a stable subordinator. Annals of Probability, 25:855–900, 1995.

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann. *ANTM: Aligned Neural Topic Models for Exploring Evolving Topics*, pages 76–97. Springer Berlin Heidelberg, Berlin, Heidelberg, 2024. ISBN 978-3-662-69603-3. doi: 10.1007/978-3-662-69603-3_3. URL https://doi.org/10.1007/978-3-662-69603-3_3.

Jonas Rieger, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch. Dynamic change detection in topics based on rolling ldas. In *Text2Story@ ECIR*, pages 5–13, 2022.

Gareth O Roberts and Jeffrey S Rosenthal. General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*, 2012.

Issei Sato and Hiroshi Nakagawa. Topic models with power-law using pitman-yor process. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–682, 2010.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International conference on machine learning*, pages 190–198. PMLR, 2014.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. doi: 10.1198/016214506000000302. URL https://doi.org/10.1198/016214506000000302.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167(107299), 2020.

Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking lda: why priors matter. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, page 1973–1981, Red Hook, NY, USA, 2009. Curran Associates Inc. ISBN 9781615679119.

Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*, 2012.

Rui Wang, Deyu Zhou, and Yulan He. Atm: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098, 2019.

Xuerui Wang and Andrew Mccallum. Topics over time: a non-markov continuous-time model of topical trends. In *Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*, 2006.

Yunli Wang and Cyril Goutte. Real-time change point detection using on-line topic models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2505–2515. Association for Computational Linguistics, 2018.

Xiaobao Wu, Xinshuai Dong, Liangming Pan, Thong Nguyen, and Anh Tuan Luu. Modeling dynamic topics in chain-free fashion by evolution-tracking contrastive learning and unassociated word exclusion. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3088–3105, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.183. URL https://aclanthology.org/2024.findings-acl.183.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2), January 2024b. ISSN 1573-7462. doi: 10.1007/s10462-023-10661-7. URL http://dx.doi.org/10.1007/s10462-023-10661-7.

Delvin Ce Zhang and Hady Lauw. Dynamic topic models for temporal document networks. In *International Conference on Machine Learning*, pages 26281–26292. PMLR, 2022.

Y. Zhang, H. Chen, J. Lu, and G. Zhang. Detecting and predicting the topic change of knowledge-based systems: A topic-based bibliometric analysis from 1991 to 2016. *Knowledge-Based Systems*, 133:255–268, 2017.

N. Zhong and D. A. Schweidel. Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science*, 39(4):827–846, 2020.

Y. Zhu, X. Lu, J. Hong, and F Wang. Joint dynamic topic model for recognition of lead-lag relationship in two text corpora. *Data Mining and Knowledge Discovery*, 2022.